Flux correction tools for finite elements

D. Kuzmin[1] and S. Turek

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

Abstract

Peculiarities of flux correction in the finite element context are investigated. Criteria for positivity of the numerical solution are formulated, and the low-order transport operator is constructed from the discrete high-order operator by adding modulated dissipation so as to eliminate negative off-diagonal entries. The corresponding anti-diffusive terms can be decomposed into a sum of genuine fluxes (rather than element contributions) which represent bilateral mass exchange between individual nodes. Thereby essentially one-dimensional flux correction tools can be readily applied to multidimensional problems involving unstructured meshes. The proposed methodology guarantees mass conservation and makes it possible to design both explicit and implicit FCT schemes based on a unified limiting strategy. Numerical results for a number of benchmark problems illustrate the performance of the algorithm.

Key Words: high resolution; finite elements; flux correction; positivity; mass conservation; unconditional stability

1 Introduction

Many CFD problems involve transport of scalar quantities (e. g. density, temperature, concentrations, turbulent kinetic energy and its dissipation rate) which must remain positive for physical reasons. An algorithm which fails to enforce the positivity constraint may produce very poor numerical results. Classical upwind methods are positive but notoriously diffusive. At the same time, high-order methods with streamline-diffusion-like stabilization of convective terms tend to produce spurious undershoots and overshoots in regions with steep gradients. Therefore, some extra artificial diffusion has to be added locally in order to suppress the nonphysical oscillations. However, a straightforward implementation of this idea introduces a free parameter which depends on the solution and is difficult to determine. Artificial viscosity methods are inevitably confronted with a tradeoff between positivity and accuracy, whereby neither property can be guaranteed.

Most of the modern high-resolution schemes for convection dominated transport problems blend high- and low-order discretizations, so as to eliminate the numerical ripples. This fundamental approach can be traced back to the concepts of flux-corrected-transport (FCT), which were established by Boris and Book in their renowned paper [4]. Methods based on flux (or slope) limiting are nonlinear and quite costly, but at the same time they are very robust and yield non-oscillatory results with sharp resolution of discontinuities. There exists a variety of such schemes (e. g. TVD, MUSCL, LED), most of which are

[1]Correspondence to: *kuzmin@math.uni-dortmund.de*

amenable to finite difference and finite volume discretizations but constitute a challenge to a finite element practitioner. Many popular schemes are limited to one-dimensional problems or Cartesian grids with directional splitting. A notable exception is the genuinely multidimensional formulation of the FCT algorithm proposed by Zalesak [30].

The design of high-resolution finite element schemes is difficult for a number of reasons. The consistent mass matrix introduces considerable implicit antidiffusion which cannot be curtailed by explicit TVD-like methods. Therefore, mass lumping is commonly employed, which results in the loss of (fourth-order) accuracy offered by the finite element method. Inherently one-dimensional flux limiters are applied edge-by-edge using solution values at the associated 'ghost' nodes [1], [21]. This non-rigorous extension of 1D concepts to multidimensions works well in practice but, strictly speaking, such schemes are not positive and should be classified as artificial viscosity methods. Furthermore, transition to an edge-based data structure as proposed by Peraire *et al.* [23] can be performed only for simplicial elements with linear basis functions which have a constant gradient. In addition, the physical fluxes have to be approximated by their linear interpolants. In general, differential operators resulting from the Galerkin discretization cannot be represented as a sum of fluxes from one node into another. Therefore, combining finite element discretizations of high and low order in a mass-conserving fashion is a nontrivial task. Even upwinding is anything but natural in the finite element context. The first-order upwind scheme of Baba and Tabata [2] is, in fact, a node-centered finite volume method rather than a finite element one.

An elegant finite element methodology which circumvents the above difficulties was introduced by Löhner *et al.* [19], [20]. It is based on Zalesak's edition of the FCT algorithm with antidiffusive element contributions in lieu of fluxes. The FEM-FCT procedure preserves the consistent mass matrix and is applicable to arbitrary unstructured meshes. However, a closer look reveals that some important issues remain unresolved. The low-order scheme is constructed by adding constant 'mass diffusion' to the high-order method, and may cease to be positive for large Courant numbers. Furthermore, the antidiffusive element contributions redistribute the mass inside the whole element rather than between individual nodes. This results in a stronger coupling between the nodal values, so that it is no longer possible to carry out an extra prelimiting step which is present in the monotone finite difference FCT schemes. Consequently, the limiter may fail to preclude the arising of spurious ripples in some cases. Last but not least, the original FEM-FCT procedure is suitable only for explicit time discretizations which are subject to a restrictive CFL condition. If the local Courant number does not exhibit strong variations, then the time step is constrained by accuracy considerations, so that the use of explicit time-stepping is justified. At the same time, the stability limitation makes explicit methods extremely inefficient for problems with strongly varying velocities and/or mesh sizes. Therefore, unconditionally stable implicit schemes are preferable for this class of applications. Likewise, the solution of steady-state problems by 'time marching' calls for a fully implicit time discretization. Indeed, high temporal accuracy is irrelevant in this case, whereas larger (artificial) time steps reduce the computational cost.

In this paper, we formulate sufficient conditions for positivity of the numerical solution and provide guidelines for enforcing them in the framework of finite element FCT schemes. The low-order operator is constructed at the discrete level using a technique

which is equivalent to upwinding in 1D and emulates it in multidimensions. The difference between the high- and low-order terms admits decomposition into a sum of fluxes which represent the mass exchange between two nodes sharing the same element. In the case of simplex elements, the fluxes can be associated with edges of the finite element mesh. At the same time, interacting nodes of multilinear elements do not have to be connected by an edge. The comeback of a flux-based representation makes it possible to apply a prelimiting of antidiffusive fluxes, which contributes greatly to elimination of numerical ripples. Furthermore, we analyze Zalesak's limiter from the viewpoint of the postulated positivity criteria, and provide a new interpretation which enables us to derive a family of implicit FEM-FCT schemes. The one based on the backward Euler time discretization is unconditionally stable and positive. To our knowledge, no other implicit high-resolution finite element schemes are available to date. The proposed algorithms preserve positivity, conserve mass and provide a sharp resolution of discontinuities as demonstrated by the numerical results for one- and two-dimensional test problems.

2 Positivity and LED criteria

Consider a generic time-dependent conservation law

$$\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) = \nabla \cdot (\epsilon \nabla u), \tag{1}$$

where u is the scalar quantity to be transported, $\mathbf{v}$ is an externally specified velocity field, and ϵ is a diffusion coefficient. The equation at hand is discretized on an arbitrary (possibly unstructured) mesh. Assume that the semi-discrete problem can be represented in the form

$$\frac{du_i}{dt} = \sum_j c_{ij} u_j, \qquad \sum_j c_{ij} = 0, \tag{2}$$

where u_i are the nodal values, and c_{ij} are some coefficients depending on the procedure employed for spatial discretization. In particular, the lumped-mass Galerkin finite element discretization with basis functions which sum to unity at each point is seen to admit such representation if the flow is incompressible ($\nabla \cdot \mathbf{v} = 0$).

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

$$\frac{du_i}{dt} = \sum_{j \neq i} c_{ij}(u_j - u_i). \tag{3}$$

Furthermore, suppose that all coefficients are non-negative: $c_{ij} \geq 0,\ j \neq i$. Then such scheme is stable in the L_∞-norm. Indeed, if u_i is a maximum, then $u_j - u_i \leq 0,\ \forall j$, so that $\frac{du_i}{dt} \leq 0$. Hence, a maximum cannot increase, and similarly a minimum cannot decrease. As a rule, coefficient matrices are sparse, so that $c_{ij} = 0$ unless i and j are adjacent nodes. Arguing as above, one can show that in this case a *local* maximum cannot increase, and a *local* minimum cannot decrease. Schemes which possess this property will be called local extremum diminishing (LED).

The LED criterion was introduced by Jameson [14], [15] as a convenient tool for the design of high-resolution schemes on unstructured meshes. It implies positivity, since if the

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. In the one-dimensional case, it guarantees that the total variation of the solution defined as

$$TV(u) = \int_{-\infty}^{+\infty} \left| \frac{\partial u}{\partial x} \right| dx \tag{4}$$

does not increase. For the sake of simplicity, consider homogeneous Dirichlet boundary conditions at both endpoints. Then the total variation is given by

$$TV(u) = 2 \left(\sum \max u - \sum \min u \right). \tag{5}$$

Thus, a one-dimensional LED scheme is necessarily total variation diminishing (TVD). This is a highly advantageous property, which has formed the basis for the development of a whole class of non-oscillatory schemes. The advantage of the LED principle as compared to TVD concepts is its applicability to multidimensional problems on both structured and unstructured meshes.

The LED property can be realized by the introduction of artificial diffusion or by the use of upwind biasing in the discrete scheme. However, it was shown by Godunov that no linear discretization method of order higher than first can guarantee monotonicity of the numerical solution. In practice, this means that the results produced by such schemes are overly diffusive. Superior approximations to convection-dominated transport problems can be obtained only by means of sophisticated nonlinear methods with coefficients depending on the solution. The discretization process is typically controlled by flux or slope limiters which adaptively switch between high- and low-order methods. A high-order approximation is used in regions where the solution is smooth, whereas the order is reduced in the vicinity of discontinuities so as to dampen nonphysical undershoots and overshoots.

Recall that equations (2) and (3) correspond to a semi-discrete convection-diffusion problem. Let us now investigate the conditions under which a LED scheme will remain positive after the time discretization. If a standard one-step θ-scheme is employed, the fully discretized equation reads

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \theta \sum_{j \neq i} c_{ij}(u_j^{n+1} - u_i^{n+1}) + (1-\theta) \sum_{j \neq i} c_{ij}(u_j^n - u_i^n), \quad 0 \leq \theta \leq 1. \tag{6}$$

The choice of 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 $\theta = \frac{1}{2}$ is known as the Crank-Nicolson scheme, which is second-order accurate. Furthermore, the following theorem holds:

Positivity Theorem.

A local extremum diminishing scheme discretized in time by the backward Euler method is unconditionally positive. Other time-stepping schemes $(0 \leq \theta < 1)$ *preserve positivity under an appropriate CFL-like condition.*

PROOF.
Let us first prove the unconditional positivity of the backward Euler method. In this case, the time discretization is fully implicit, so that the last term in the right-hand side of equation (6) vanishes. Assume that the discrete solution u^{n+1} is negative at some nodes and denote by k the node at which the global minimum is attained. The new solution at this node satisfies

$$u_k^{n+1} = u_k^n + \Delta t \sum_{j \neq k} c_{kj}(u_j^{n+1} - u_k^{n+1}). \tag{7}$$

By the inductive assumption, the old solution u^n must be non-negative everywhere. The coefficients c_{kj} are also non-negative due to the LED property, so $u_k^{n+1} < 0$ implies that $u_j^{n+1} - u_k^{n+1} < 0$ for some j. However, this leads to a contradiction, since u_k^{n+1} was chosen to be the global minimum. Hence, the positivity of u^n is inherited by u^{n+1}.

Now let us tackle $\theta < 1$. The above considerations for the implicit term show that the discrete scheme (6) will preserve positivity if the explicit term satisfies the inequality

$$u_i^n + \Delta t(1-\theta) \sum_{j \neq i} c_{ij}(u_j^n - u_i^n) \geq 0 \qquad \forall i. \tag{8}$$

As long as $u_i^n \geq 0$ and $c_{ij} \geq 0$, it is sufficient to require that

$$1 + \Delta t(1-\theta) \min_i c_{ii} \geq 0, \tag{9}$$

where $c_{ii} = -\sum_{j \neq i} c_{ij}$ are the diagonal elements of the original coefficient matrix defined by equation (2). This condition provides the desired positivity criterion, which can be used for the time step control. □

In essence, the above theorem represents a generalization of the discrete maximum principle to time-dependent convection-diffusion problems. It lays the groundwork for the construction of positivity-preserving numerical schemes, and we will see shortly how this can be accomplished in the framework of the FEM-FCT methodology.

3 Mass conservation

Conservation of mass is crucial to the design of numerical methods for the bulk of transport problems [13]. In particular, a failure of the algorithm to conserve mass may cause shocks to propagate with wrong speed if nonlinear conservation laws (e. g. the inviscid Burgers equation) are considered. Non-conservative numerical schemes can produce unacceptable results also in many other cases, so they should be typically avoided.

The conventional Galerkin finite element discretization conserves mass in an integral sense. Indeed, the weak formulation of equation (1) reads

$$\int_\Omega \left[\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) - \nabla \cdot (\epsilon \nabla u) \right] w \, d\mathbf{x} = 0, \qquad \forall w. \tag{10}$$

The associated semi-discrete system is obtained by using an approximation of u in a suitable finite-dimensional space and applying the basis functions φ_i in lieu of w. For

customary finite elements, we have $\sum_i \varphi_i \equiv 1$, so that the sum of all equations yields the original conservation law in the integral form:

$$\frac{d}{dt}\int_\Omega u \; d\mathbf{x} = -\int_S (\mathbf{v}u - \epsilon\,\nabla u)\cdot\mathbf{n}\; ds, \tag{11}$$

where $\mathbf{n}$ is the unit outward normal. It can be seen that the total mass in Ω changes only due to convective and diffusive fluxes through the boundary.

Finite volume methods apply formulation (11) directly to each element of the triangulation, so that mass conservation is enforced not only globally but also locally. This corresponds to a piecewise-constant finite element discretization (discontinuous Galerkin methods). Flux correction in the finite volume framework is straightforward. The objective of this paper is to extend the available FCT machinery to linear and multilinear finite element approximations.

While the standard Galerkin discretization is conservative, this favorable property may be lost in the quest to get rid of nonphysical oscillations which contaminate numerical solutions to convection-dominated problems. For instance, the most straightforward and inexpensive algorithm 'inspired' by the FCT procedure would be:

1. Solve the transport equation by a high-order scheme prone to oscillate.
2. Estimate the upper and lower solution bounds using some *a-priori* knowledge and/or numerical results produced by a monotone low-order scheme.
3. 'Trim' the high-order solution so as to make it stay within the bounds.

Unfortunately, this approach is not to be recommended for an obvious reason: it doesn't conserve mass. This is a quite instructive example, since any other non-conservative limiting technique is equally unreliable but almost certainly more expensive. If the above algorithm is to be employed, it should be complemented by an extra postproccessing step for the recovery of the lost mass [17].

4 Structure of diffusion operators

It is well known that the Galerkin discretization is unstable for pure convection problems. Therefore, the discrete scheme must contain enough dissipation (of physical or numerical origin) to damp out the instabilities. Furthermore, properly tuned artificial diffusion is the key tool for rendering a numerical scheme positive and local extremum diminishing. The structure of the involved diffusive terms is of primary importance for subsequent considerations, so it is worthwhile to study it in some detail. The most common discrete diffusion operators encountered in finite element schemes for transport problems are:

- **The discrete Laplacian operator**

$$d_{ij}^{\Delta} = \int_\Omega \nabla\varphi_i \cdot \nabla\varphi_j \; d\mathbf{x},$$

  which typically results from the discretization of physical diffusion terms. It is also referred to as the 'stiffness matrix'.

- **The streamline diffusion operator**

$$d_{ij}^s = \int_\Omega \mathbf{v} \cdot \nabla\varphi_i \; \mathbf{v} \cdot \nabla\varphi_j \; d\mathbf{x},$$

which represents artificial diffusion in the streamline direction added in order to stabilize the convective terms. It can be arrived at in different ways.

The concept of streamline diffusion was introduced by Brooks and Hughes [5] and employed within a consistent Petrov-Galerkin formulation. A similar approach was followed by Johnson [16] and his collaborators. The least-squares formulation [6] also gives rise to a streamline diffusion operator of the form above. Furthermore, streamline diffusion terms can be attributed to higher-order temporal approximations afforded by Taylor-Galerkin methods [8].

- **The mass diffusion operator**

$$d_{ij}^m = \int_\Omega \varphi_i(\varphi_j - \delta_{ij}) \; d\mathbf{x},$$

which is given by the difference between the consistent mass matrix M_C and its diagonal counterpart M_L obtained by the row-sum mass lumping. Mass diffusion has proved to be particularly useful for the construction of low-order finite element schemes to be combined with high-order ones in the FCT framework (see below).

In spite of their different nature and appearance, discrete diffusion operators possess some common features. The most important ones are the symmetry

$$d_{ij} = d_{ji} \tag{12}$$

and zero row/column sums

$$\sum_j d_{ij} = \sum_i d_{ij} = 0 \tag{13}$$

(for basis functions satisfying $\sum_i \varphi_i \equiv 1$). A tensor D having these properties can be treated as a generalized diffusion operator and constructed so as to provide an appropriate modification of the numerical scheme.

The application of a discrete diffusion operator to the vector of nodal values yields

$$(Du)_i = \sum_j d_{ij} u_j = \sum_{j \neq i} d_{ij}(u_j - u_i) \tag{14}$$

due to the zero row sum property. Let us define the flux f_{ij} from node j into node i as $f_{ij} = d_{ij}(u_j - u_i)$. Then

$$(Du)_i = \sum_{j \neq i} f_{ij}, \qquad f_{ji} = -f_{ij}. \tag{15}$$

Hence, diffusive terms can be decomposed into a sum of numerical fluxes similar to those encountered in conservative finite difference schemes. Each node receives contributions from all nodes sharing an element with it. Mass conservation is guaranteed, since the fluxes representing mass transfer from one node into another are equal in magnitude and opposite in sign. Consequently, it is safe to limit such fluxes, and this can be done in an essentially one-dimensional fashion.

5 Standard FEM-FCT procedure

Before embarking on the development of high-resolution finite element schemes using the flux-based representation of (anti-) diffusive terms, let us recall the FEM-FCT procedure due to Löhner *et al.* [19]. Building on an earlier paper by Parrott and Christie [22], it has established the framework for implementation of Zalesak's multidimensional limiter [30] for finite element approximations on unstructured meshes. FEM-FCT employs constant mass diffusion to construct the low-order scheme and delegates the role of antidiffusive fluxes to element contributions. This constitutes a viable approach but, as we are about to see, there is some room for improvement.

The process of flux correction starts with introducing a strong artificial diffusion into a high-order scheme, so as to enforce positivity of the numerical solution. According to the Godunov theorem, this inevitably degrades the accuracy of the method to first order. The crux of the FCT approach consists in reducing the error by adding a compensating antidiffusion in regions where the solution is smooth and the Taylor series expansion makes sense. The standard FEM-FCT procedure as proposed by Löhner *et al.* [19], [20] involves six algorithmic steps which can be summarized as follows:

1. Discretize the governing equation using an explicit high-order finite element method with an appropriate stabilization of convective terms.

2. Perform mass lumping and insert a discrete diffusion operator into the high-order scheme to construct a non-oscillatory low-order method.

3. Invoke the low-order scheme to compute a provisional solution u^L which is supposed to preserve positivity.

4. Compute the antidiffusive element contributions F_e needed to recover the high accuracy of the original method.

5. Limit the antidiffusive element contributions so as to preclude the formation of new and the enhancement of existing extrema.

6. Apply the corrected antidiffusive element contributions to u^L in order to obtain the end-of-step solution u^{n+1}.

The limiting strategy employed in step 5 is crucial to the performance of the method. It amounts to multiplying the antidiffusive element contributions by certain correction factors which vary between zero and unity. The final solution u^{n+1} is given by

$$u_i^{n+1} = u_i^L + \sum_e \alpha_e F_{e,i}, \qquad 0 \leq \alpha_e \leq 1. \tag{16}$$

Here $F_{e,i}$ denotes the antidiffusive contribution of element e to node i. The control of artificial dissipation is executed by monitoring the smoothness of the solution and adaptively selecting the correction factors so as to switch between the diffusive low-order solution ($\alpha_e = 0$) and the oscillatory high-order solution ($\alpha_e = 1$). The objective of the flux limiter is to utilize the antidiffusive element contributions to the greatest extent possible without generating nonphysical wiggles and violating the positivity constraint. The ins and outs of the FEM-FCT algorithm are elucidated below.

High-order scheme

The governing equation discretized in space and time by an explicit high-order method can be cast into the form

$$M_C \Delta u = R, \tag{17}$$

where M_C denotes the consistent mass matrix, $\Delta u = u^{n+1} - u^n$ is the vector of nodal increments, and the load vector R comprises the convective and diffusive terms evaluated at the old time level. Löhner *et al.* employed a two-step Taylor-Galerkin method of the Lax-Wendroff type. However, any other explicit finite element scheme is feasible.

The solution to problem (17) clearly satisfies

$$M_L \Delta u^H = R + (M_L - M_C)\Delta u^H. \tag{18}$$

Here the superscript H refers to the high-order scheme, and M_L is the (row-sum) lumped mass matrix, which is known to possess the conservation property [13]. The second term in the right-hand side represents the antidiffusion built into the consistent mass matrix, which makes it possible to obtain time-accurate solutions to transient problems albeit at the expense of solving a (well-conditioned) linear system at each time step.

Low-order scheme

The accuracy offered by the consistent mass matrix has to be foregone by linear positivity-preserving schemes. Löhner *et al.* [19] perform mass lumping and add explicit mass diffusion to transform the high-order method into a low-order one:

$$M_L \Delta u^L = R + c_d(M_C - M_L)u^n, \tag{19}$$

where the superscript L denotes the low-order scheme, and c_d is some constant diffusion coefficient. In particular, the choice $c_d = 1$ yields [10], [25]

$$M_L u^L = M_C u^n + R, \tag{20}$$

which corresponds to the high-order method with mass lumping carried out only in the left-hand side. This technique converts the one-dimensional Lax-Wendroff method into a scheme which is stable and monotone for Courant numbers $|\nu| \leq \sqrt{\frac{2}{3}}$. This is more restrictive than the CFL condition for the classical upwind discretization. Furthermore, no information is available about the behavior of the solution in more general settings.

Adding sufficiently large constant diffusion to achieve monotonicity can be traced back to the original SHASTA scheme of Boris and Book [4]. While this approach has been used successfully by many authors, it may fail in some cases. Hence, the diffusion coefficient c_d and the time step Δt should be chosen with care to obtain non-oscillatory results.

Antidiffusive element contributions

Note that if we subtract (19) from (18), the unwieldy term R vanishes. Furthermore, $\Delta u^H - \Delta u^L = u^H - u^L$, so that the antidiffusive element contributions are given by

$$F_e = M_L^{-1}\Big|_e (\hat{M}_L - \hat{M}_C)(c_d \hat{u}^n + \Delta \hat{u}^H). \tag{21}$$

The above notation is to be understood in the following sense. The local antidiffusion operator $\hat{M}_L - \hat{M}_C$ is constructed from *element* mass matrices and acts upon the function values at the nodes of the element. This results in a vector with length equal to the number of local degrees of freedom. Finally, its elements are divided by the corresponding diagonal entries of the *global* matrix M_L to yield the antidiffusive element contributions.

Solution bounds

The admissible solution range is determined by searching for local extrema in the low-order solution u^L [4] and in the old solution u^n [30]. Löhner *et al.* estimate the solution bounds $u_i^{\max\limits_{\min}}$ by the following three-step algorithm:

1. Assemble u^* from the nodal values of u^L or u^n, whichever is greater/smaller:

$$u_i^* = \max_{\min}\{u_i^L, u_i^n\}. \tag{22}$$

2. Compute the maximum/minimum value of u^* within each element:

$$u_e^{**} = \max_{\min} u_i^*, \qquad i \in N_e. \tag{23}$$

3. Pick the maximum/minimum value of u^{**} over all elements containing the node:

$$u_i^{\max\limits_{\min}} = \max_{\min} u_e^{**}, \qquad e \in E_i. \tag{24}$$

Thus, the unknown solution u^{n+1} at any node should be bounded by the maximum and minimum values of u^L and u^n at the stencil associated with this node.

Screening the old solution along with the low-order one was proposed by Zalesak to alleviate 'peak clipping' inherent to the SHASTA scheme. This was shown to yield a considerable improvement for a number of test configurations. However, this generalization may produce numerical ripples for other problems, e.g. those involving a variable velocity. Therefore, it is prudent to set $u^* \equiv u^L$ as in the original method of Boris and Book.

Limiting strategy

The limiting process is based on Zalesak's multidimensional flux correction algorithm [30]. Six auxiliary quantities are defined for each node:

- $P_i^{\pm}$, the sum of all positive/negative antidiffusive element contributions to node i:

$$P_i^{\pm} = \sum_{e \in E_i} \max_{\min}\{0, F_{e,i}\}. \tag{25}$$

- $Q_i^{\pm}$, the maximum/minimum admissible increment for node i:

$$Q_i^{\pm} = u_i^{\max\limits_{\min}} - u_i^L. \tag{26}$$

- $R_i^{\pm}$, the least upper bound for the correction factors which guarantees no overshoot/undershoot at node i:

$$R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{if } P_i^{\pm} \neq 0, \\ 0, & \text{if } P_i^{\pm} = 0. \end{cases} \tag{27}$$

The correction factors must be chosen so that the antidiffusive element contributions acting in concert are unable to create nonphysical extrema. A suitable limiter is given by

$$\alpha_e = \min_{i \in N_e} \begin{cases} R_i^{+}, & \text{if } F_{e,i} \geq 0, \\ R_i^{-}, & \text{if } F_{e,i} < 0. \end{cases} \tag{28}$$

It is conservative enough to guarantee that the constraint $u_i^{\min} \leq u_i^{n+1} \leq u_i^{\max}$ is satisfied at all nodes. Hence, the final solution will preserve positivity if the low-order one does. However, numerical ripples of low amplitude can and do occur in some cases.

6 Alternative FEM-FCT procedure

The representation of antidiffusion in terms of element contributions restricts the choice of artificial diffusion operators and prevents the use of some inherently one-dimensional flux correction tools. An alternative formulation is offered by the flux-based decomposition of (anti-) diffusive terms introduced above. High-resolution finite element schemes of this type were proposed in [10], [25], [26]. The structure of the (constant) mass diffusion operator was utilized to develop artificial viscosity, FCT, and TVD-like methods building on the concept of modulated dissipation. In this section, we will follow a similar approach while using the rigorous LED criteria to develop both explicit and implicit FCT schemes.

Low-order scheme

The quality of the low-order method is of great importance for the overall performance of an FCT algorithm. If the low-order solution ceases to be positive, oscillatory results will certainly ensue. Furthermore, a perfect low-order scheme should contain just as much artificial diffusion as is necessary to enforce the positivity. This would facilitate the task of limiting and preclude excessive smearing. For finite difference or finite volume discretizations, the best candidate for the low-order scheme is clearly the upwind method. An example of a finite element FCT algorithm using upwind as the low-order scheme can be found in the paper of Parrott and Christie [22]. However, upwinding is rather cumbersome and unnatural in the finite element context, which has led Löhner *et al.* [19] to replace it by mass diffusion with a constant coefficient.

Adding the same amount of diffusion everywhere is computationally efficient, but the resulting method is not optimal as far as accuracy is concerned. If the free parameter is chosen too large, the scheme is overdiffusive, and the stability range is reduced. At the same time, insufficient artificial diffusion may lead to the arising of spurious extrema which are transmitted to the final solution. These shortcomings were recognized by Georghiou *et al.* [12], who attempted to design variable 'optimal' diffusion coefficients depending on the local Courant number as in the upwind finite difference method. This seems to be a

poor remedy, since the 'improved' FEM-FCT algorithm can be expected to work well only on very regular meshes and may fail to preserve positivity. In what follows, we will pursue the same goals as Georghiou *et al.* but construct the low-order scheme in a different way, which does reconcile the conflicting demands for accuracy and positivity.

If the flow is incompressible, equation (1) can be written in the non-conservative form:

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \nabla u = \nabla \cdot (\epsilon \nabla u). \tag{29}$$

Let the spatial discretization be performed by the standard Galerkin finite element method. This yields a semi-discrete problem of the form

$$M_C \frac{du}{dt} = K^H u, \tag{30}$$

where M_C is the consistent mass matrix, and K^H is the discrete transport operator, which has zero row sum, so that

$$(K^H u)_i = \sum_{j \neq i} k_{ij}^H (u_j - u_i). \tag{31}$$

In general, the Galerkin scheme (30) is not local extremum diminishing, which manifests itself in the tendency to oscillate (especially in convection-dominated cases). However, the LED criteria at our disposal reveal what measures need to be taken in order to obtain a usable low-order method.

First of all, we employ mass lumping to remove the implicit antidiffusion intrinsic to the consistent mass matrix. The resulting scheme can be cast into the form (3) and would possess the LED property if all coefficients $k_{ij}^H,\ j \neq i$ were non-negative. This suggests the following rule for the construction of the low-order transport operator:

$$K^L = K^H + D, \tag{32}$$

where D is a tensor of modulated dissipation. It is designed so as to eliminate all negative off-diagonal entries of the high-order operator:

$$d_{ii} = -\sum_{k \neq i} d_{ik}, \qquad d_{ij} = d_{ji} = \max\{0, -k_{ij}^H, -k_{ji}^H\}, \qquad \forall\ i < j. \tag{33}$$

In essence, this corresponds to applying one-dimensional diffusion operators associated with the (fictitious) segments connecting the adjacent nodes. The global matrix assembly is performed in a standard way. It is easy to verify that D has zero row- and column sums, and thus enjoys all properties of generalized diffusion operators including mass conservation. Note that if physical diffusion is strong enough, so that the coefficients are non-negative from the outset, then no artificial diffusion is added. Hence, in diffusion-dominated cases the matrices K^H and K^L are identical.

If the velocity field in equation (1) is not divergence-free, a local accumulation of the conserved quantity can occur. Therefore, the formation of physical extrema must be reckoned with. For compressible flows, the low-order operator K^L constructed as above consists of a LED part K_1 and a diagonal residual part K_2. The Positivity Theorem can be

readily extended to such schemes, whereby the extra 'reactive' term affects only the upper bound for the time step. Hence, the numerical algorithm will be positivity-preserving under a proper CFL-like condition. Depending on the sign of $\nabla \cdot \mathbf{v}$, the admissible time steps may be greater or smaller than those for the incompressible case. As a matter of fact, the fully implicit scheme may become conditionally positive for $\nabla \cdot \mathbf{v} << 0$. However, this is very unlikely to happen for any practical flows of interest.

In any event, the semi-discrete low-order scheme reads

$$M_L \frac{du}{dt} = (K^H + D)u = K^L u, \tag{34}$$

that is

$$m_i \frac{du_i}{dt} = \sum_j k_{ij}^H u_j + \sum_{j \neq i} d_{ij}(u_j - u_i) = \sum_j k_{ij}^L u_j, \tag{35}$$

where m_i denote the diagonal entries of the lumped mass matrix. It is notable that the difference between the high- and low-order discretization of the transport terms admits decomposition into fluxes.

According to the Positivity Theorem, the backward Euler time discretization of this problem is unconditionally positive (at least for weakly compressible flows), while other time-stepping schemes preserve positivity as long as

$$\Delta t \leq \frac{1}{1-\theta} \min_i \{-m_i / k_{ii}^L \mid k_{ii}^L < 0\}. \tag{36}$$

This positivity condition gives a practical estimate of the maximum admissible time step. It is influenced by the degree of implicitness θ and by the ratio m_i / k_{ii}^L. Hence, excessive artificial diffusion not only degrades the accuracy of the method but also requires taking smaller time steps. This is exemplified by the scheme (20), whereby the Lax-Wendroff method was augmented by mass diffusion of *constant* magnitude.

Example

Let us illustrate the construction of low-order operators by a one-dimensional example. Consider the pure convection equation

$$\frac{\partial u}{\partial t} + v \frac{\partial u}{\partial x} = 0 \tag{37}$$

discretized on a uniform mesh of linear elements. For the sake of simplicity assume that the velocity v is constant and positive. The involved element matrices have the form

$$\hat{M}_L = \frac{\Delta x}{2} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \qquad \hat{K}^H = \frac{v}{2} \begin{bmatrix} 1 & -1 \\ 1 & -1 \end{bmatrix}. \tag{38}$$

After the global matrix assembly, the central difference approximation of the convective term is recovered at interior nodes:

$$\frac{du_i}{dt} = -v \, \frac{u_{i+1} - u_{i-1}}{2 \, \Delta x}. \tag{39}$$

The minimum amount of artificial dissipation sufficient to enforce positivity is proportional to $\hat{d}_{12} = v/2$. The corresponding discrete diffusion operator restricted to one element is given by

$$\hat{D} = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \quad \Rightarrow \quad \hat{K}^L = v\begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}. \tag{40}$$

The resulting low-order scheme is seen to be equivalent to the upwind finite difference method in the interior:

$$\frac{du_i}{dt} = -v\,\frac{u_i - u_{i-1}}{\Delta x}. \tag{41}$$

Obviously, this is the least diffusive linear scheme which preserves positivity. The associated CFL condition reads:

$$v\frac{\Delta t}{\Delta x} \leq \frac{1}{1-\theta}. \tag{42}$$

In particular, the fully explicit scheme is positive for Courant numbers up to unity.

To summarize, our technique for the construction of positive low-order operators reduces to standard upwinding for pure convection in one dimension and, unlike the *ad hoc* algorithm of Georghiou *et al.* [12], it is applicable to arbitrary meshes and multidimensional problems. Moreover, the resulting scheme is less diffusive than the upwind method in the presence of physical diffusion. A distinct advantage of the proposed approach is that the artificial diffusion operator is assembled at the discrete level and depends only on the location and magnitude of negative off-diagonal entries. The origin of discrete transport operators doesn't matter, so that finite element matrices resulting from the discretization of 1D, 2D and 3D problems can be treated in exactly the same way.

Explicit FEM-FCT formulation

As already mentioned above, it is worthwhile to reformulate the FEM-FCT procedure in terms of internodal fluxes. Let us first consider the fully explicit time-stepping. In this case, the Galerkin method lacks stability for large Peclet numbers, so some stabilization is required for the convective terms. A suitable candidate for the high-order scheme is the Taylor-Galerkin method proposed by Donea *et al.* [9].

The governing equation (1) is discretized in time using a Taylor series expansion in the time step Δt up to the second order:

$$u^{n+1} = u^n + \Delta t\, u_t^n + \frac{(\Delta t)^2}{2}\, u_{tt}^n. \tag{43}$$

The first-order time derivative is provided directly by the original conservation law. If the flow is incompressible and the diffusion coefficient is constant, we have

$$u_t^n = -\mathbf{v}\cdot\nabla u^n + \epsilon\nabla^2 u^n. \tag{44}$$

Assuming that the velocity field is stationary or 'frozen', the second-order time derivative can be calculated as follows [9]:

$$\begin{aligned} u_{tt}^n &= -\nabla\cdot(\mathbf{v}u_t^n) + \epsilon\nabla^2 u_t^n \\ &= (\mathbf{v}\cdot\nabla)^2 u^n + \epsilon\nabla^2(u_t^n - \epsilon\nabla^2 u^n) + \epsilon\nabla^2 u_t^n \\ &= (\mathbf{v}\cdot\nabla)^2 u^n + 2\epsilon\nabla^2\left(\frac{u^{n+1}-u^n}{\Delta t}\right) + \mathcal{O}(\Delta t, \epsilon^2). \end{aligned} \tag{45}$$

After the substitution of these expressions into the Taylor series (43), one obtains the time-discretized scheme

$$[1 - \Delta t\epsilon\nabla^2]\frac{u^{n+1} - u^n}{\Delta t} = -\mathbf{v}\cdot\nabla u^n + \epsilon\nabla^2 u^n + \frac{\Delta t}{2}(\mathbf{v}\cdot\nabla)^2 u^n. \tag{46}$$

Its global accuracy is $\mathcal{O}((\Delta t)^2, \epsilon^2\Delta t)$, which implies that it is of second-order in time, since usually $\epsilon^2 \leq \Delta t$. For pure convection problems, this method is identical to the standard Lax-Wendroff finite element scheme. At the same time, it features good stability properties in diffusion-dominated cases.

The Galerkin spatial discretization applied to the weak formulation of equation (46) yields a linear system of the form

$$M_C \Delta u^H = \Delta t K^H u^n + \frac{(\Delta t)^2}{2} D^S u^n. \tag{47}$$

For notational simplicity, the consistent mass matrix was redefined to comprise the implicit diffusive contribution. The extra term $\frac{(\Delta t)^2}{2} D^S u^n$ results from the second-order time discretization and caters for proper stabilization. Integration by parts is employed to relieve it from the second-order spatial derivatives:

$$\begin{aligned} d^s_{ij} &= \int_\Omega \varphi_i \mathbf{v}\cdot\nabla(\mathbf{v}\cdot\nabla\varphi_j)\, d\mathbf{x} = -\int_\Omega \mathbf{v}\cdot\nabla\varphi_i\ \mathbf{v}\cdot\nabla\varphi_j\, d\mathbf{x} \\ &- \int_\Omega \varphi_i \nabla\cdot\mathbf{v}\ \mathbf{v}\cdot\nabla\varphi_j\, d\mathbf{x} + \int_{S_{\text{out}}} \varphi_i \mathbf{v}\cdot\mathbf{n}\ \mathbf{v}\cdot\nabla\varphi_j\, ds. \end{aligned} \tag{48}$$

The first term in the right-hand side is seen to be a streamline diffusion operator. In contrast to other methods of streamline diffusion type, no artificial parameter needs to be fitted. The amount of stabilization is naturally fixed by the coefficient of the second-order term in the Taylor series expansion. An investigation of Lax-Wendroff schemes by means of the modified equation method 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 Lax-Wendroff and Taylor-Galerkin methods the reader is referred to [8].

The matrix D^S can be turned into a generalized diffusion operator by neglecting the last two integrals in the right-hand side of (48). The former one vanishes for divergence-free velocity fields. The latter one switches off stabilization terms normal to the boundary. This modification was found to preclude the arising of instabilities and spurious pressure boundary layers at the outlet [3]. However, we will neglect boundary correction for the time being and discuss it in some detail a bit later.

The presented Taylor-Galerkin method is only conditionally stable. Moreover, the consistent mass matrix leads to a reduced stability domain as compared to the associated finite difference scheme [8]. Second-order temporal accuracy can also be achieved e. g. by handling the second-order term implicitly while retaining the explicit treatment for the first-order term [29]:

$$u^{n+1} - \frac{(\Delta t)^2}{2} u^{n+1}_{tt} = u^n + \Delta t\, u^n_t. \tag{49}$$

This scheme is unconditionally stable for pure convection problems and can be used as the high-order method. Note that the computational overhead connected with the implicit treatment of the streamline diffusion term is insignificant, since a consistent mass matrix problem has to be solved anyway. In addition, the matrix at hand remains symmetric. At the same time, the unconditional stability of this method cannot be duly utilized in the FEM-FCT framework. As long as the convective term is discretized explicitly, the time step will be constrained by the positivity condition for the low-order scheme.

There exist many other promising high-order methods, but a detailed analysis and comparison of their characteristics would be beyond the scope of this paper. Instead, we will demonstrate how the simple methods presented above can be combined with the affiliated low-order scheme of 'upwind' type. The discrete system (47) implies that

$$M_L \Delta u^H = \Delta t K^L u^n - (M_C - M_L)\Delta u^H - \Delta t (K^L - K^H) u^n + \frac{(\Delta t)^2}{2} D^S u^n. \tag{50}$$

The low-order transport operator K^L is constructed as explained in the previous section. It can readily be seen that the difference between the high- and low-order scheme is represented by the last three terms in the right-hand side. They originate from discrete (anti-) diffusion operators and lend themselves to decomposition into fluxes. Hence, the flux-corrected end-of-step solution u^{n+1} is sought in the form:

$$m_i u_i^{n+1} = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \qquad \alpha_{ji} = \alpha_{ij}, \tag{51}$$

where $\tilde{u}$ denotes the provisional low-order solution computed from

$$m_i \tilde{u}_i = m_i u_i^n + \Delta t \sum_j k_{ij}^L u_j^n. \tag{52}$$

The involved antidiffusive fluxes f_{ij} are given by

$$\begin{aligned} f_{ij} &= -m_{ij}(\Delta u_j^H - \Delta u_i^H) - \Delta t\, d_{ij}(u_j^n - u_i^n) + \tfrac{(\Delta t)^2}{2} d_{ij}^s (u_j^n - u_i^n), \\ f_{ji} &= -f_{ij}, \qquad i < j. \end{aligned} \tag{53}$$

They offset the error induced by mass lumping, 'upwinding', and the first-order time discretization. If the semi-implicit Taylor-Galerkin method (49) is employed, the streamline diffusion contribution should be evaluated using the high-order solution u^H instead of u^n. The selection of correction factors α_{ij} will be addressed below.

Implicit FEM-FCT formulation

In order to eliminate or alleviate severe stability restrictions due to the explicit-time stepping, let us explore the potential of implicit methods. Of primary interest are the backward Euler and the Crank-Nicolson scheme. Both of them are unconditionally stable and can be used as the high-order method in conjunction with the Galerkin spatial discretization. No extra stabilization of convective terms is required in this case.

Let the implicit schemes of high- and low-order be related by the formula

$$(M_L - \theta \Delta t K^L)\Delta u^H = \Delta t K^L u^n - (M_C - M_L)\Delta u^H - \Delta t (K^L - K^H)[\theta u^H + (1-\theta) u^n]. \tag{54}$$

The discrete antidiffusion operators responsible for the high-order accuracy can be easily identified. If they are omitted, the positive low-order scheme is obtained. The proposed FCT algorithm is based on the following representation of the end-of-step solution:

$$m_i u_i^{n+1} - \theta\Delta t \sum_j k_{ij}^L u_j^{n+1} = m_i \tilde{u}_i + \sum_{j\neq i} \alpha_{ij} f_{ij}, \qquad \alpha_{ji} = \alpha_{ij}, \tag{55}$$

where $\tilde{u}$ stands for the positivity-preserving solution to the explicit subproblem

$$m_i \tilde{u}_i = m_i u_i^n + (1-\theta)\Delta t \sum_j k_{ij}^L u_j^n. \tag{56}$$

The backward Euler method corresponds to $\theta = 1$, so that $\tilde{u} \equiv u^n$. In case of the Crank-Nicolson scheme, $\tilde{u}$ is seen to be an intermediate solution at the time instant $t^n + \Delta t/2$ computed by the explicit low-order scheme.

According to the relation (54), the antidiffusive fluxes are defined by

$$\begin{gathered} f_{ij} = -m_{ij}(\Delta u_j^H - \Delta u_i^H) - \theta\Delta t\, d_{ij}(u_j^H - u_i^H) - (1-\theta)\Delta t\, d_{ij}(u_j^n - u_i^n), \\ f_{ji} = -f_{ij}, \qquad i < j. \end{gathered} \tag{57}$$

The computation of the high-order solution u^H requires solving a non-symmetric linear system. Furthermore, the scheme (55) is implicit for $\theta > 0$, so that another algebraic system with the matrix of the low-order operator has to be solved *after* the flux correction step. If iterative solvers are employed, the computed high-order solution provides a reasonable initial approximation to the final solution.

A remark is in order concerning the application of implicit schemes to nonlinear problems like the inviscid Burgers equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = 0, \tag{58}$$

which constitutes a one-dimensional prototype of the Euler and Navier-Stokes equations. In this case, the matrices K^H and K^L depend on the unknown solution, so that additional outer iterations are necessary. It will be noted that the linearization of the problem using a constant extrapolation in time can entail a loss of mass and alter the shock speed.

The simplest iterative treatment of nonlinearities is afforded by a fixed point defect correction method. In each time step, the approximate solution and the transport operator are successively updated as follows:

$$u^{(l+1)} = u^{(l)} - C^{-1}[M_C u^{(l)} - \theta\Delta t K^H(u^{(l)})u^{(l)} - M_C u^n - (1-\theta)\Delta t K^H(u^n)u^n], \tag{59}$$

where l is the outer iteration counter, and C is a suitably chosen 'preconditioner'. The iteration process is terminated when the residual is small enough or l exceeds a given limit. As a rule, the 'inversion' of matrix 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 equation does not have to be solved very accurately at each outer iteration. A moderate improvement of the residual is sufficient to obtain a good overall accuracy.

Setting $C = M_L - \theta\Delta t K^L(u^{(l)})$, that is using the low-order operator as the preconditioner, we obtain the following nonlinear iteration scheme:

$$M_L u^{(l+1)} - \theta\Delta t K^L(u^{(l)})u^{(l+1)} = M_L u^n + (1-\theta)\Delta t K^L(u^n)u^n + F(u^{(l)}, u^n), \tag{60}$$

where $F(u^{(l)}, u^n)$ comprises all pertinent antidiffusive terms (cf. equation (54)), which can be decomposed into fluxes as described above. 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 secured.

Limiting strategy

Now that the difference between the high- and low-order solutions to the linear or nonlinear transport equation is available as a sum of raw antidiffusive fluxes f_{ij}, the algorithm for the selection of correction factors α_{ij} comes into play. The flux limiter is a key element of the FEM-FCT procedure, which needs to be adapted to the new formulation. Below we work out a unified limiting strategy applicable to both explicit and implicit schemes.

Explicit FCT schemes can benefit from canceling all antidiffusive fluxes directed down the gradient of $\tilde{u}$:

$$f_{ij} := 0, \quad \text{if} \quad f_{ij}(\tilde{u}_i - \tilde{u}_j) < 0. \tag{61}$$

This test should be applied *before* the flux correction step. Its purpose is to ensure that the flux does not smooth the low-order solution. To put it another way, an antidiffusive flux is not allowed to be diffusive. When this happens, small-scale numerical ripples can be produced even though the solution remains positive. Hence, the limiter is positivity- but not monotonicity-preserving [7].

The prelimiting of antidiffusive fluxes can be traced back to the celebrated SHASTA scheme. Zalesak also mentioned this approach in passing but did not promote its regular use. He argued that the majority of antidiffusive fluxes act to steepen the gradient, while the effect of (61) is minimal and cosmetic in nature. This remark has discouraged the use of prelimiting in FCT algorithms based on Zalesak's multidimensional limiter. Apparently, this is not the sole reason why this optional step is missing in the FEM-FCT procedure of Löhner *et al.* The replacement of antidiffusive fluxes by element contributions makes the prelimiting impossible to carry out for multidimensional problems. Only the restitution of a flux-based formulation enables us to apply this technique in the finite element context.

DeVore [7] has rediscovered the preprocessing of antidiffusive fluxes as a way to achieve monotonicity and demonstrated that it can lead to a dramatic qualitative improvement of dynamic simulation results. Even for simple test problems with discontinuous solutions, remarkable 'esthetic' improvements are observed (see the numerical examples below). Therefore, the prelimiting step is to be included in explicit FCT algorithms. In our experience, it remains relevant also for the implicit schemes introduced in this paper.

Let us proceed to the algorithm for selection of correction factors. It is largely equivalent to Zalesak's limiter but is derived and interpreted in a quite different way. As before, we denote by $u_i^{\substack{\max\\ \min}}$ the maximum and minimum solution values at the stencil S_i which consists of the node i and its nearest neighbors:

$$u_i^{\substack{\max\\ \min}} = \begin{matrix}\max\\ \min\end{matrix} \tilde{u}_j, \qquad j \in S_i. \tag{62}$$

It should be borne in mind that the positivity-preserving auxiliary solution $\tilde{u} = u^L(t^{n+1-\theta})$ depends on the concrete time-stepping scheme. The old solution u^n is no longer used in the computation of local extrema for the reasons which will become clear shortly.

In accordance with the FCT theory, all antidiffusive fluxes which try to accentuate a local maximum or minimum must be completely canceled:

$$\alpha_{ij} = 0, \quad \text{if} \quad \tilde{u}_i = u_i^{\max}, \quad f_{ij} > 0 \quad \text{or} \quad \tilde{u}_i = u_i^{\min}, \quad f_{ij} < 0. \tag{63}$$

If this applies to all fluxes into the node i, we are done. Otherwise, the remaining fluxes have to be limited so as to comply with the positivity constraint. It is noteworthy that the right-hand side of our schemes (51) and (55) admits the following representation:

$$RHS = m_i\tilde{u}_i + \sum_{j\neq i}\alpha_{ij}f_{ij} = m_i\tilde{u}_i + c_iQ_i, \qquad c_i = \frac{\sum_{j\neq i}\alpha_{ij}f_{ij}}{Q_i}, \tag{64}$$

where the multiplier Q_i is chosen to be

$$Q_i = \begin{cases} Q_i^+ = u_i^{\max} - \tilde{u}_i, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} > 0, \\ Q_i^- = u_i^{\min} - \tilde{u}_i, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} < 0, \\ 1, & \text{if } \sum_{j\neq i}\alpha_{ij}f_{ij} = 0. \end{cases} \tag{65}$$

By virtue of (63), we have $Q_i \neq 0$, so that no division by zero takes place. Furthermore, the coefficient c_i is always non-negative. Let the local extremum $u_i^{\max\atop\min}$ be attained at a node k adjacent to the node i. Then the antidiffusive term exhibits a LED structure, and we obtain

$$RHS = m_i\tilde{u}_i + c_i(\tilde{u}_k - \tilde{u}_i) = (m_i - c_i)\tilde{u}_i + c_i\tilde{u}_k, \qquad c_i \geq 0. \tag{66}$$

In light of the above, the proposed FEM-FCT schemes will preserve positivity provided that $m_i \geq c_i$. This important observation frames a general rule for the selection of correction factors α_{ij}.

It remains to show that Zalesak's limiter does possess the desired properties. Let us restate it for our flux-based formulation. The quantities $P_i^{\pm}$ and $R_i^{\pm}$ are redefined as

$$P_i^{\pm} = \frac{1}{m_i}\sum_{j\neq i}\begin{matrix}\max\\ \min\end{matrix}\{0, f_{ij}\}, \qquad R_i^{\pm} = \begin{cases} \min\{1, Q_i^{\pm}/P_i^{\pm}\}, & \text{if } P_i^{\pm} \neq 0, \\ 0, & \text{if } P_i^{\pm} = 0. \end{cases} \tag{67}$$

Since now the nodes exchange mass on a bilateral basis, the flux limiter is given by

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\}, & \text{if } f_{ij} \geq 0, \\ \min\{R_j^+, R_i^-\}, & \text{if } f_{ij} < 0. \end{cases} \tag{68}$$

It is independent of the number of spatial dimensions and can be easily implemented as a 'black-box' routine which computes the correction factors given an array of antidiffusive fluxes for each pair of neighboring nodes.

The condition (63) is automatically satisfied, since $Q_i^{\pm} = 0$ spells $R_i^{\pm} = 0$ and $\alpha_{ij} = 0$. Hence, any enhancement of local extrema is neutralized by the limiter. Furthermore, the following estimate holds:

$$\sum_{j\neq i}\alpha_{ij}f_{ij} \leq \sum_{j\neq i}\alpha_{ij}\max\{0, f_{ij}\} \leq m_iR_i^+P_i^+ \leq m_iQ^+. \tag{69}$$

In much the same way, it can be verified that

$$\sum_{j\neq i}\alpha_{ij}f_{ij}\geq\sum_{j\neq i}\alpha_{ij}\min\{0,f_{ij}\}\geq m_iR_i^-P_i^-\geq m_iQ_i^-\,. \tag{70}$$

This proves that the corrected antidiffusive fluxes satisfy the constraint $m_i\geq c_i$. Recall that the left-hand sides of our FEM-FCT schemes pose no hazard to positivity. According to the Positivity Theorem, the backward Euler method with flux correction is unconditionally positive for incompressible flows. For highly compressible flows, the time step may depend on the divergence of the velocity as explained above. The Crank-Nicolson scheme is subject to a positivity condition for the auxiliary problem (56), but the admissible Courant numbers are twice as large as those for the explicit Lax-Wendroff scheme.

Treatment of outflow boundaries

Let us make some final remarks regarding the treatment of outflow boundaries. It turns out that FCT schemes can malfunction when applied to problems with smooth solutions (i.e. in situations when flux correction is actually redundant). This major deficiency manifests itself in spurious ripples emanating from the outflow boundary and propagating into the computational domain. A typical example will be presented below. The wiggles can be cured by (local) mesh refinement, but it is necessary to understand their origins in order to find a better remedy. It goes without saying that a failure to cope with smooth solutions seriously compromises the practical utility of the method even if it provides an excellent resolution of shocks and contact discontinuities.

The pathological behavior of the FCT algorithm apparently occurs due to the lack of proper boundary adjustment. Similar problems are observed when Petrov-Galerkin methods are applied without boundary correction for the streamline diffusion terms [3]. At the same time, consistent Lax-Wendroff and Taylor-Galerkin schemes do incorporate the necessary modification. It is given by the surface integral which arises naturally from integration by parts of the second order term (see above). Inclusion of similar integrals into the Galerkin least squares formulation also yields the desired effect [11].

In most cases, streamline diffusion methods without boundary modification still produce acceptable solutions. However, boundary anomalies can be considerably aggravated by flux correction. This can be attributed to a nonphysical natural boundary condition implied by the low-order scheme. For simplicity, consider a one-dimensional pure convection problem and recall that in this case the boundary condition is to be prescribed only at the inflow boundary, i.e. at the endpoint where the velocity is directed into the domain. The positivity of the low-order scheme is enforced by adding strong discrete diffusion to the underlying high-order scheme. This is equivalent to solving a parabolic convection-diffusion equation with *homogeneous* Neumann boundary condition at the outlet. Hence, the low-order solution will exhibit a kink whenever the exact solution has a non-vanishing derivative at the outflow boundary. At the same time, high-order methods handle smooth profiles with ease and provide a much better approximation to the exact solution at the boundary. This discrepancy seems to be the reason why FCT schemes sometimes produce saw-like profiles given smooth initial data.

In fact, the homogeneous Neumann boundary condition is a direct consequence of the conservation property of discrete diffusion operators. If we add artificial dissipation while requiring strict mass conservation, the numerical solution will be forced to bend so as to prevent any nonphysical diffusive flux through the boundary. Hence, it is worthwhile to reconsider the concept of mass conservation and endorse the outflow of mass due to *numerical* diffusion. The aforementioned boundary integrals represent in essence numerical fluxes which cater for a consistent boundary treatment.

A feasible strategy motivated by the above considerations is to construct the discrete low-order transport operator so as to leave the rows corresponding to outflow boundary nodes unchanged. To this end, we can replace formula (33) by

$$d_{ii} = -\sum_{k \neq i} d_{ik}, \qquad d_{ij} = \max\{0, -k_{ij}^H\}, \qquad d_{ji} = 0 \tag{71}$$

if i is an interior node and j is a node on the outflow boundary. Note that the symmetry of antidiffusive fluxes $f_{ij} = -f_{ji}$ is lost for boundary nodes, so that the limiter and the assembly process have to be modified appropriately.

For our one-dimensional example, we obtain

$$\hat{D} = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \quad \Rightarrow \quad \hat{K}^L = \frac{v}{2}\begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}, \tag{72}$$

which is equivalent to adding the missing boundary integral. It is noteworthy that all off-diagonal entries of the low-order transport operator are still non-negative, so that the positivity of the low-order solution is guaranteed. This will also be the case e.g. for bilinear elements provided the velocity and mesh size do not exhibit abrupt changes in proximity to the outflow boundary. A proof for the case of a uniform mesh and a constant velocity is available. It is quite straightforward and will not be presented here.

Another simple way to get rid of ripples is to abstain from adding any artificial diffusion in the boundary layer, i.e. set $d_{ij} = d_{ji} = 0$ if i or j belongs to the outflow boundary. This approach preserves the symmetry of f_{ij} and is probably to be preferred because of its lower complexity. Boundary adjustment should not be applied to convection-diffusion problems with Dirichlet boundary conditions prescribed at the outlet.

7 Numerical examples

Let us substantiate the proposed FEM-FCT methodology by a number of one- and two-dimensional examples. The Lax-Wendroff and Crank-Nicolson schemes are second-order accurate in time and produce virtually identical numerical results. Hence, it suffices to examine the behavior of the Lax-Wendroff (LW/FCT) and backward Euler (BE/FCT) methods. Unless otherwise indicated, the 1D solutions were obtained on a uniform mesh of 100 linear elements, whereas a Cartesian mesh of 128×128 (due to the quadtree data structure for the mesh) bilinear elements was employed for the 2D examples. The time step was chosen rather small in most cases in order to reduce the temporal error for the first-order accurate backward Euler method. However, some solutions for Courant numbers exceeding unity are also presented.

Convection of a step function

As a classical one-dimensional test problem, consider pure convection of a discontinuous step function with unit velocity. The time step Δt is set equal to 10^{-3} which corresponds to the Courant number $\nu = 0.1$. The first method to be evaluated is the explicit FEM-FCT scheme based on the Lax-Wendroff time-stepping. The numerical results at $t = 0.5$ are depicted in Figure 1. Here and below, the dash-dotted line stands for the initial data, and the dotted line designates the analytical solution.

As expected, the high-order LWFE method entails undershoots and overshoots of considerable amplitude, while the low-order solution is monotone but corrupted by excessive numerical diffusion. Flux correction brings about a dramatic improvement, but the solution exhibits some imperfections if the prelimiting step is omitted. By far the most accurate results are produced by the FEM-FCT method equipped with prelimiting. This serves as an evidence that the preprocessing of antidiffusive fluxes is a valuable complement to the FCT procedure.

Let us compare these results with those obtained by the fully implicit BE/FCT scheme (see Figure 2). Even though the Courant number is rather small, the backward Euler method is seen to be diffusive because of the first-order time discretization. At the same time, it is not as oscillatory as the LWFE scheme. The implicit 'upwind' method yields essentially the same results as its explicit counterpart. It is evident that the implicit FEM-FCT algorithm also does a very good job in combining the advantages of high- and low-order schemes. The nonphysical oscillations are filtered out completely, while the slope of the profile remains the same. As the time step is refined, the accuracy approaches that of the explicit LW/FCT scheme.

Inviscid Burgers equation

The inviscid Burgers equation (58) is a standard model problem for nonlinear convection in one dimension. It is frequently employed to assess the ability of numerical methods to deal with formation and propagation of shocks. Let us start with a discontinuous initial profile and simulate its evolution up to the time $t = 0.4$. The numerical solutions produced by the FEM-FCT schemes are displayed in Figure 3. The nonlinearity was treated by the fixed point defect correction method as described above.

It turns out that the effect of the prelimiting step is not so pronounced in this setting. Furthermore, the LW/FCT and BE/FCT yield solutions of comparable quality. At the same time, the fully implicit scheme is unconditionally positive and can be applied at Courant numbers greater than unity. An example for $\Delta t = 2\Delta x$ demonstrates that large time steps degrade the accuracy, but the numerical solution still looks quite reasonable. Note that in all cases the shock propagates with correct speed, which implies that the mass is conserved.

Convection of a cosine wave

Let us come back to linear convection problems with constant velocity $v = 1$. If the initial data is smooth enough, then the conventional Galerkin method performs remarkably well. As a matter of fact, it was used to compute the dotted reference solution for the cosine profile in Figure 4. Hence, flux correction is superfluous in this case. However, it is often

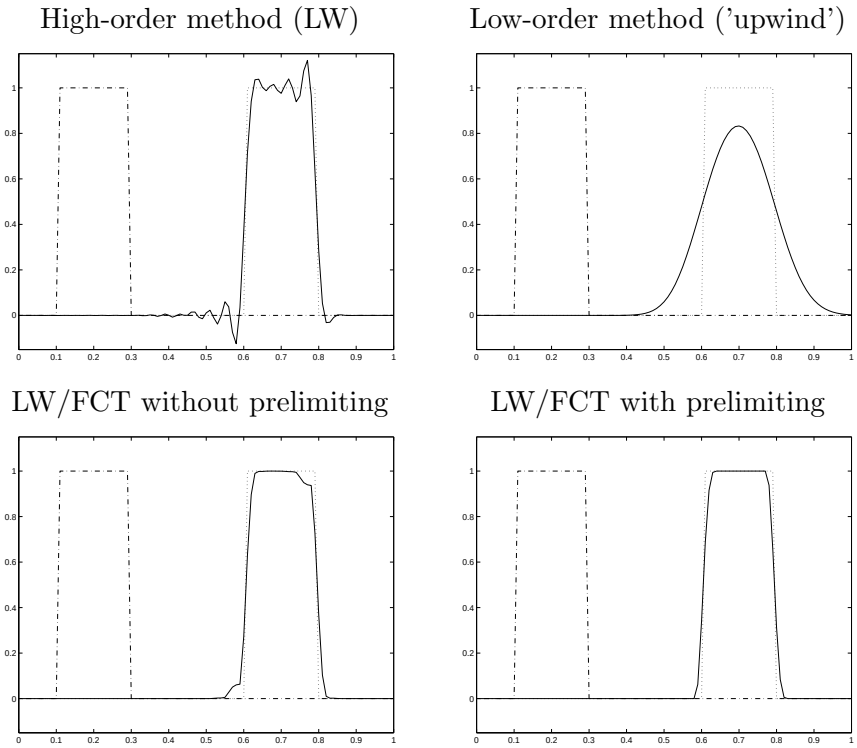


Figure 1. Convection of a step function. Lax-Wendroff/FCT scheme, $t = 0.5$.

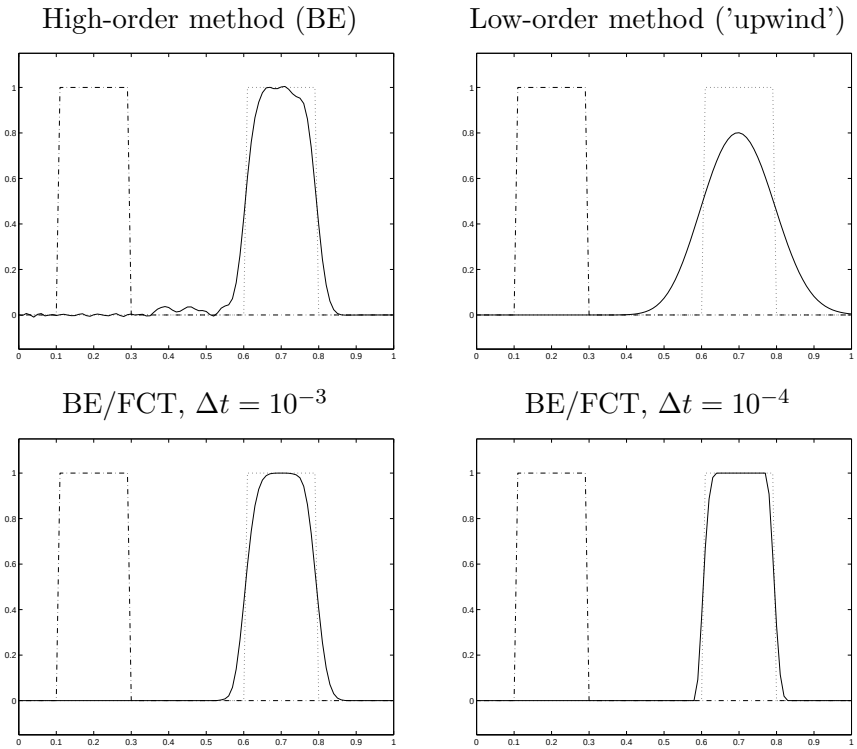


Figure 2. Convection of a step function. Backward Euler/FCT scheme, $t = 0.5$.

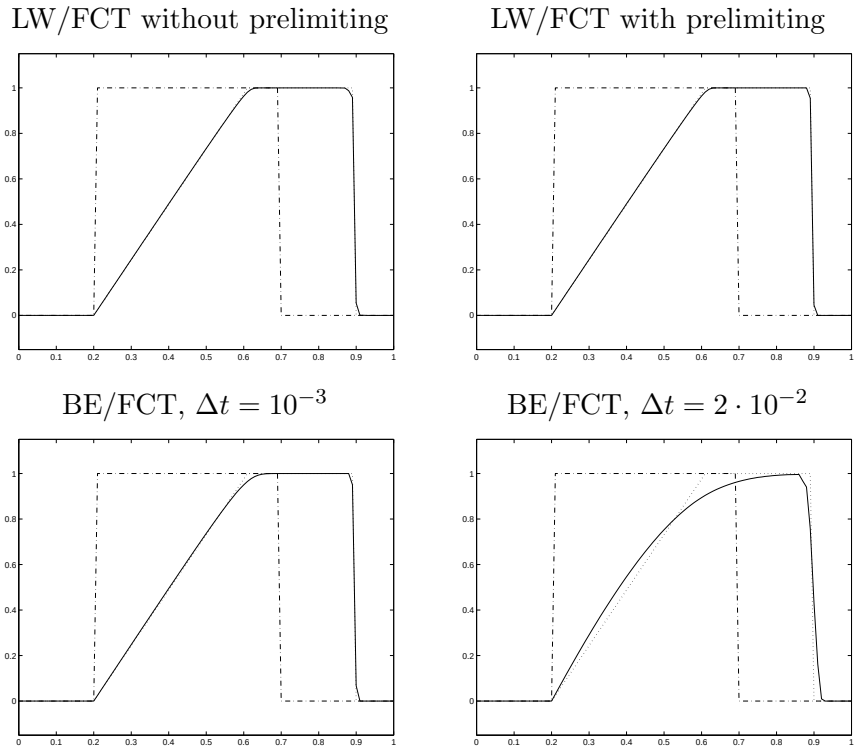


Figure 3. Inviscid Burgers equation. Solution at $t = 0.4$.

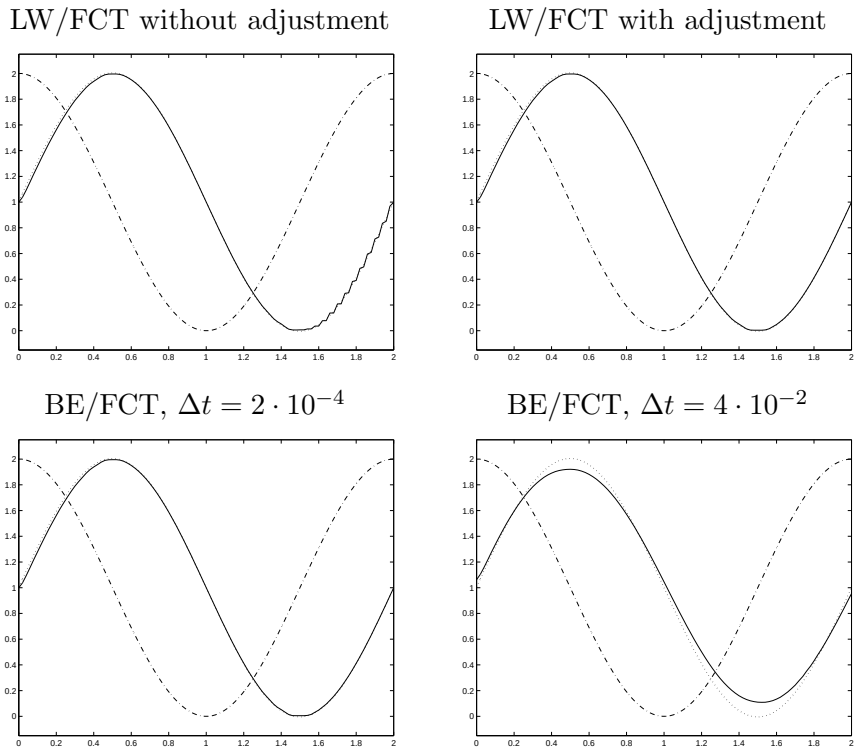


Figure 4. Convection of a cosine wave. Solution at $t = 0.5$.

impossible to detect such situations *a priori*. For most practical CFD applications, the smoothness of the unknown solution varies in space and time. Therefore, the numerical method should be capable of handling both smooth and discontinuous data.

The first plot in Figure 4 reveals that the FCT algorithm in its original form can pollute the high-order solution by spurious ripples which can be traced back to the outflow boundary. The time step was deliberately chosen very small in this example, since this was found to amplify the perturbations. Any of the techniques for boundary correction proposed above makes it possible to restore the smoothness of the solution and obtain accurate results. The BE/FCT scheme remains stable and positive for Courant numbers beyond unity, although the amplitude of the wave is dampened appreciably.

Stretching/compression by a variable velocity

The next two examples illustrate the performance of our FEM-FCT schemes for linear convection problems with velocity depending on the spatial coordinate. The non-uniform velocity field is intended to expose the behavior of the methods under circumstances when a physical growth or decay of extrema occurs. It is important to ascertain that the flux limiter is able to distinguish between physical and nonphysical extrema.

Consider a step function which is convected and spread by the variable velocity field $v = x$ as shown in Figure 5. In this case, both LW/FCT and BE/FCT deliver non-oscillatory but quite diffusive numerical results. Note that the left border of the profile is resolved considerably better than the right one, since the Courant number increases with x. It should be emphasized that the observed smoothing is not a deficiency of flux correction. In fact, the high-order method produces an equally diffusive solution with oscillations superimposed on it.

If the transported profile undergoes compression rather than stretching, the algorithm performs much better. This is exemplified by Figure 6, where the velocity is taken to be $v = 1 - x$. In this case, the mass gradually accumulates in the center of the computational domain. The solutions obtained by the LW/FCT and BE/FCT schemes are virtually identical and exhibit superb accuracy.

Convection of monotone profiles

The last one-dimensional test problem deals with the convection of monotone data. Let the initial profile be a smooth approximation to the Heavyside step function. The front is chosen to be rather steep, so that flux correction is required to preclude the arising of undershoots and overshoots.

The numerical solutions produced by the FEM-FCT schemes in the case of constant velocity $v = 1$ are compared with each other and with the exact solution in Figure 7. The explicit LW/FCT scheme provides an excellent resolution of the front, while the implicit BE/FCT scheme is moderately diffusive for 'large' time steps. It can be seen that both methods are free of false antidiffusion inherent e.g. to the popular superbee limiter [27]. Thus, no artificial steepening of the profile takes place.

Convection of the same function with the variable velocity $v = x$ is investigated in Figure 8. The qualitative behavior of the methods is essentially the same as in the case of constant velocity. It is noteworthy that, in contrast to the stretching of a discontinuous pulse, no pronounced extra smearing is observed.

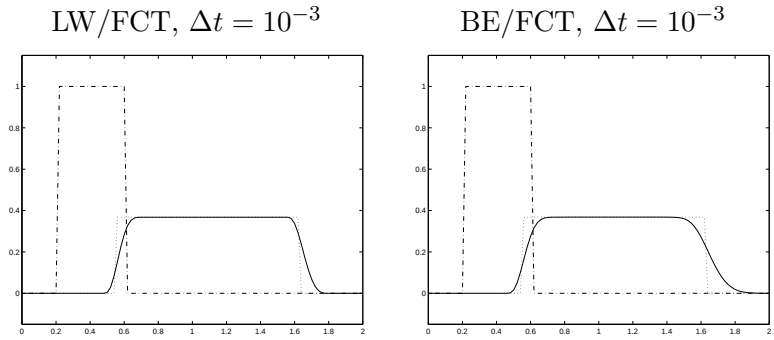


Figure 5. Stretching by the variable velocity field $v = x$. Solution at $t = 1.0$.

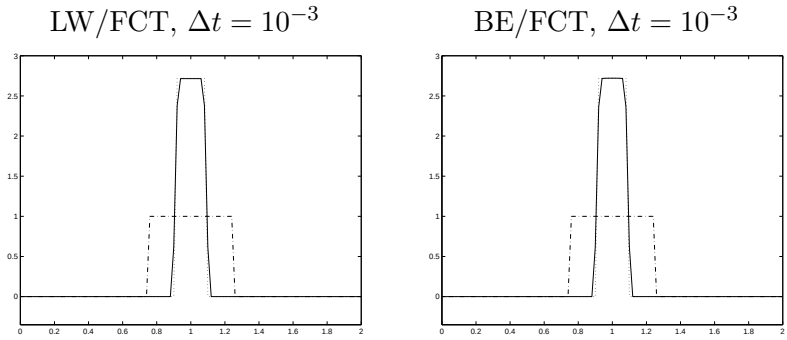


Figure 6. Compression by the variable velocity field $v = 1 - x$. Solution at $t = 1.0$.

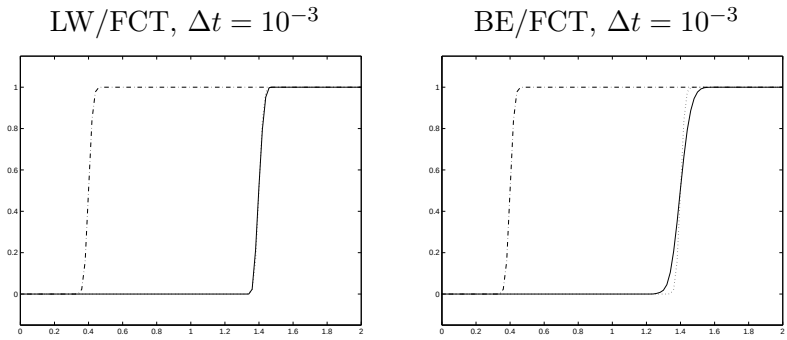


Figure 7. Convection of a monotone profile with $v = 1$. Solution at $t = 1.0$.

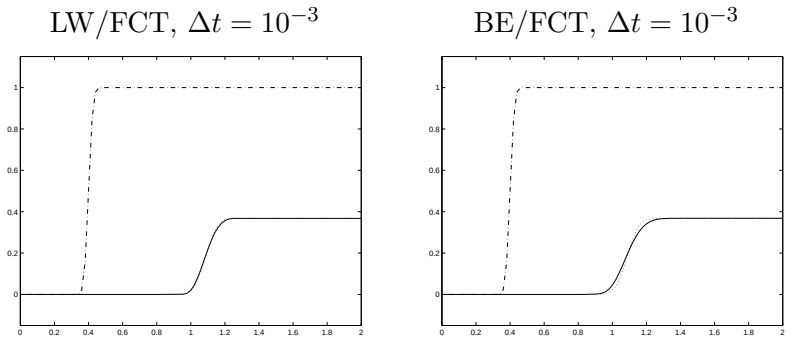


Figure 8. Convection of a monotone profile with $v = x$. Solution at $t = 1.0$.

Steady-state convection-diffusion in 1D

As we have seen, the fully implicit BE/FCT scheme is quite diffusive for transient convection problems. At the same time, it appears to be very attractive as an iterative solver for (quasi-) steady-state convection-diffusion equations. Indeed, the steady-state solution can be obtained by applying a FEM-FCT method to the associated time-dependent problem. Possible nonlinearities can be treated in the same iterative loop. The temporal accuracy of the method does not matter in this case, since the time step is merely an artificial parameter which determines the convergence rates. In fact, local time-stepping can be employed [3]. As long as the accuracy of the converged solution depends entirely on the spatial discretization, it is expedient to choose the time steps as large as possible, so as to reduce the computational cost. This makes explicit schemes non-competitive, since they are subject to a restrictive CFL condition. Moreover, the numerical solution produced e.g. by the Lax-Wendroff method is affected by the streamline diffusion depending on the artificial time step. Hence, steady-state problems call for an implicit treatment.

Consider the one-dimensional stationary convection-diffusion equation

$$v\frac{\partial u}{\partial x} - \epsilon\frac{\partial^2 u}{\partial x^2} = 0, \qquad u(0) = 1, \qquad u(1) = 0$$

for $v = 1$ and $\epsilon = 10^{-2}$, which corresponds to the Peclet number $Pe = 100$. This is a singularly perturbed elliptic problem, which is characterized by the presence of a sharp front next to the outflow boundary $x = 1$. The boundary layer develops because the solution of the reduced problem ($\epsilon = 0$) does not satisfy the homogeneous Dirichlet boundary condition imposed for the full problem.

Let us discretize the domain by a uniform mesh of 10 linear elements and compare the results produced by the backward Euler scheme without and with flux correction. As an initial guess, we take the straight line $u^0 = 1 - x$. The obtained solutions are displayed in Figure 9. The standard Galerkin method reduces to the central difference approximation, which is seen to be oscillatory for the coarse mesh under consideration. Remarkably, the flux-corrected steady-state solution is nodally exact. Actually, even the 'low-order' method yields excellent results in this case. Recall that the tensor of artificial dissipation is constructed in such a way that it just compensates the lack of physical diffusion. If any physical diffusion is present, then less artificial diffusion is required to enforce positivity. Thus, for $\epsilon > 0$ the low-order scheme is less diffusive than the classical upwind method.

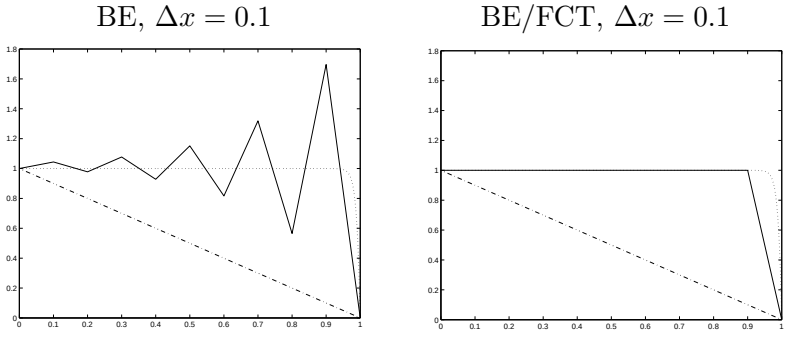


Figure 9. Steady-state convection-diffusion in 1D, $\epsilon = 10^{-2}$.

Convection of a discontinuous profile in 2D

Let us proceed to the two-dimensional examples. The first one shown in Figure 10 is a direct generalization of the 1D problem dealing with the uniform convection of a step function. In the 2D case, the computational domain is a unit square. The velocity is constant and equal to unity in each coordinate direction: $\mathbf{v} = (1, 1)$. Homogeneous Dirichlet boundary conditions are prescribed at the inflow boundaries $x = 0$ and $y = 0$. A discontinuous initial profile is transported along the streamlines, which are parallel to the diagonal $y = x$.

The numerical solutions at the time instant $t = 0.5$ obtained by the prelimited LW/FCT and BE/FCT schemes corroborate the diagnosis made on the basis of the one-dimensional examination. Both methods succeed in the elimination of nonphysical wiggles and preserve the steepness of the profile fairly well, unlike the underlying low-order scheme. However, the temporal error induced by the backward Euler time-stepping is still non-negligible for the employed time step $\Delta t = 10^{-3}$. It is evident that the second-order LW/FCT scheme outperforms the first-order BE/FCT scheme when it comes to the time-accurate solution of transient convection problems.

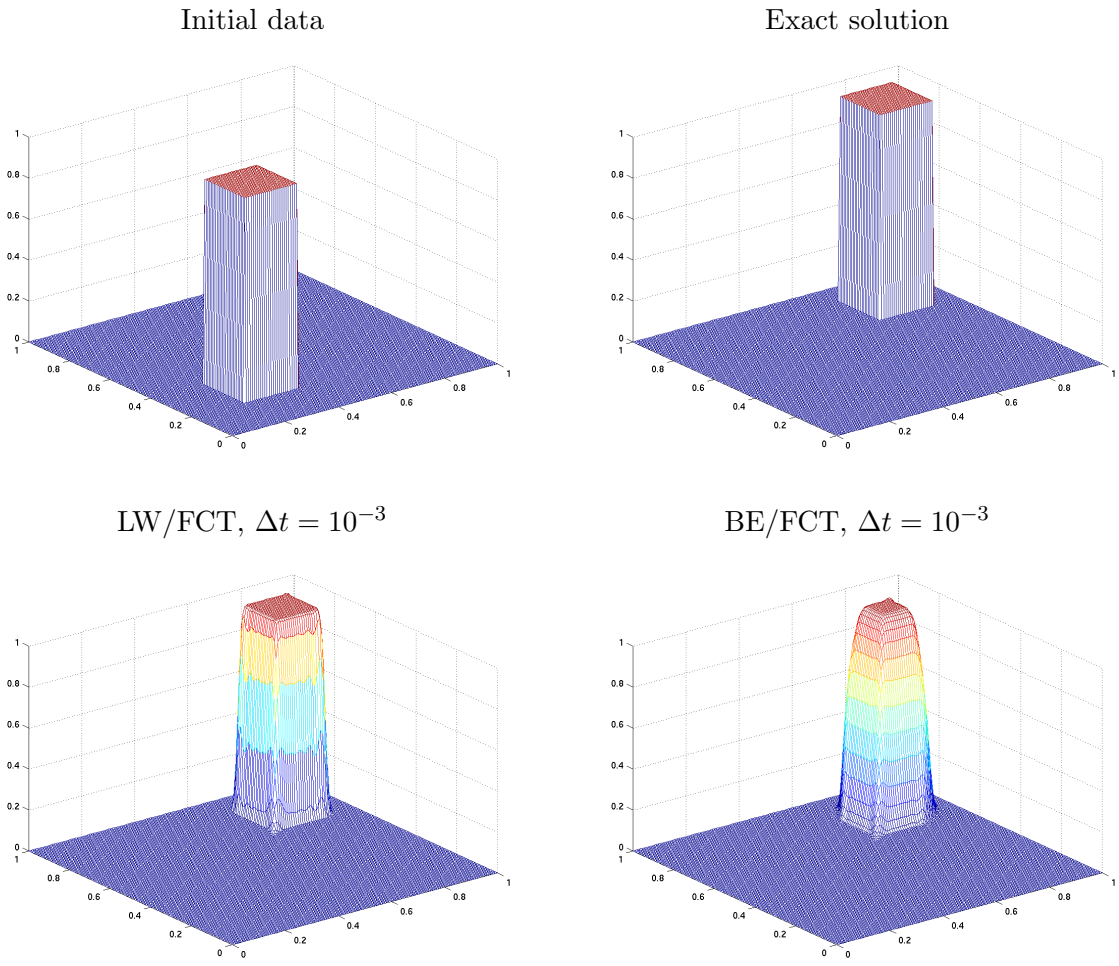


Figure 10. Convection of a discontinuous profile. Initial data and solution at $t = 0.5$.

Convection of a smooth profile in 2D

Our next test problem deals with the evolution of a sinusoidal profile. Consider the same computational domain and velocity field as in the previous example and let the initial condition be given by

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

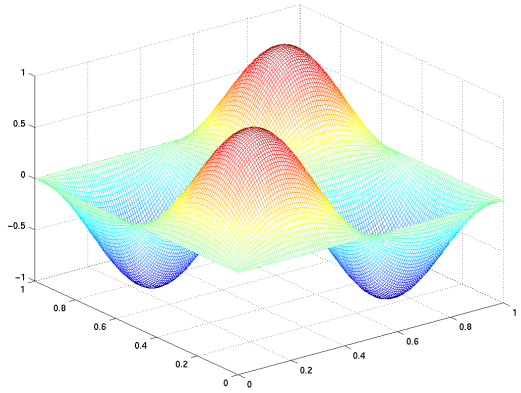
The prescribed boundary conditions are

$$u(0, y, t) = -\sin(2\pi t) \cdot \sin(2\pi(y - t)),$$
$$u(x, 0, t) = -\sin(2\pi t) \cdot \sin(2\pi(x - t)),$$

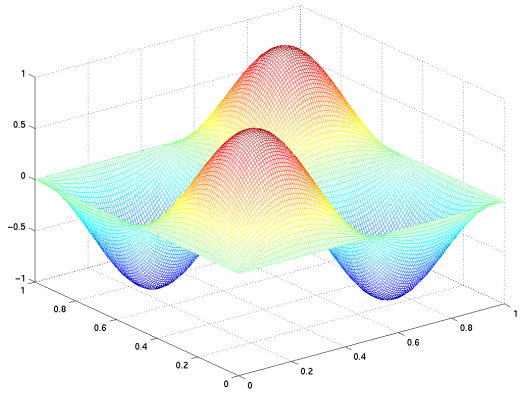
so that the initial data matches the exact solution at the time $t = 1.0$.

The numerical results produced by the FEM-FCT schemes with boundary correction are displayed in Figure 11. All remarks regarding the treatment of outflow boundaries remain valid in two dimensions. The maximum norm of the solution quoted in the diagrams serves as an indicator of numerical damping. The diffusive nature of the BE/FCT method is excused to some extent by its ability to operate with larger time steps.

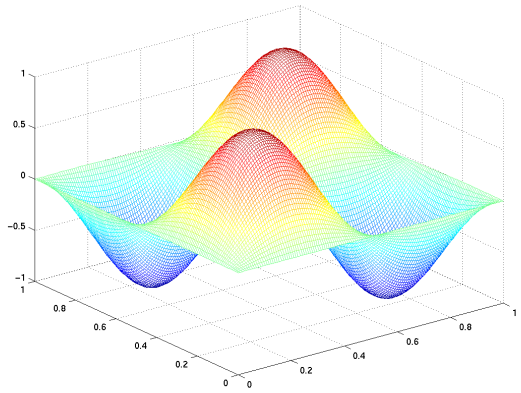
Initial data/exact solution, $||u||_\infty = 1.0$



LW/FCT, $\Delta t = 10^{-3}$, $||u||_\infty = 0.9969$



BE/FCT, $\Delta t = 10^{-3}$, $||u||_\infty = 0.9874$



BE/FCT, $\Delta t = 10^{-2}$, $||u||_\infty = 0.9041$

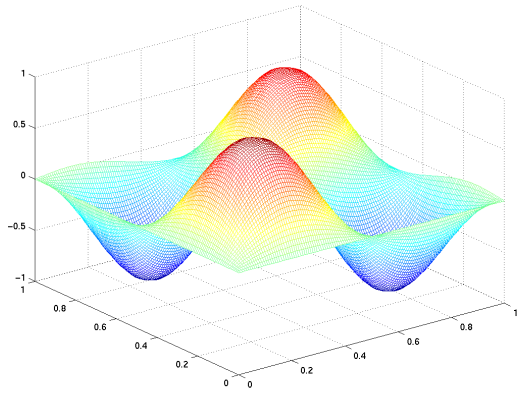


Figure 11. Convection of a smooth profile. Initial data and solution at $t = 1.0$.

Rotation of a cylinder with a slot

Let us turn to the investigation of a solid body rotation in a nonuniform velocity field $\mathbf{v} = (-y, x)$. The counterclockwise rotation takes place about the center of the square domain $(-1, 1) \times (-1, 1)$. The initial data is a cylinder with a slot defined by

$$u(x, y, 0) = \begin{cases} 1, & R < 1/3 \text{ and } (|x| > 0.05 \text{ or } y > 0.5), \\ 0, & \text{otherwise}, \end{cases}$$

where $R = \sqrt{x^2 + (y - 1/3)^2}$. This challenging two-dimensional benchmark problem was considered by Löhner *et al.* [20], Zalesak [30] and many others.

Figure 12 demonstrates that both LW/FCT and BE/FCT produce excellent results as long as the time step is small enough. The prelimiting of antidiffusive fluxes has proved to be expedient for this problem. If it is omitted, the numerical solution is contaminated by innocuous but ugly ripples. The last diagram illustrates the behavior of the implicit scheme at large Courant numbers. Since the velocity increases with distance from the origin, the slot is resolved considerably better than the rear of the cylinder.

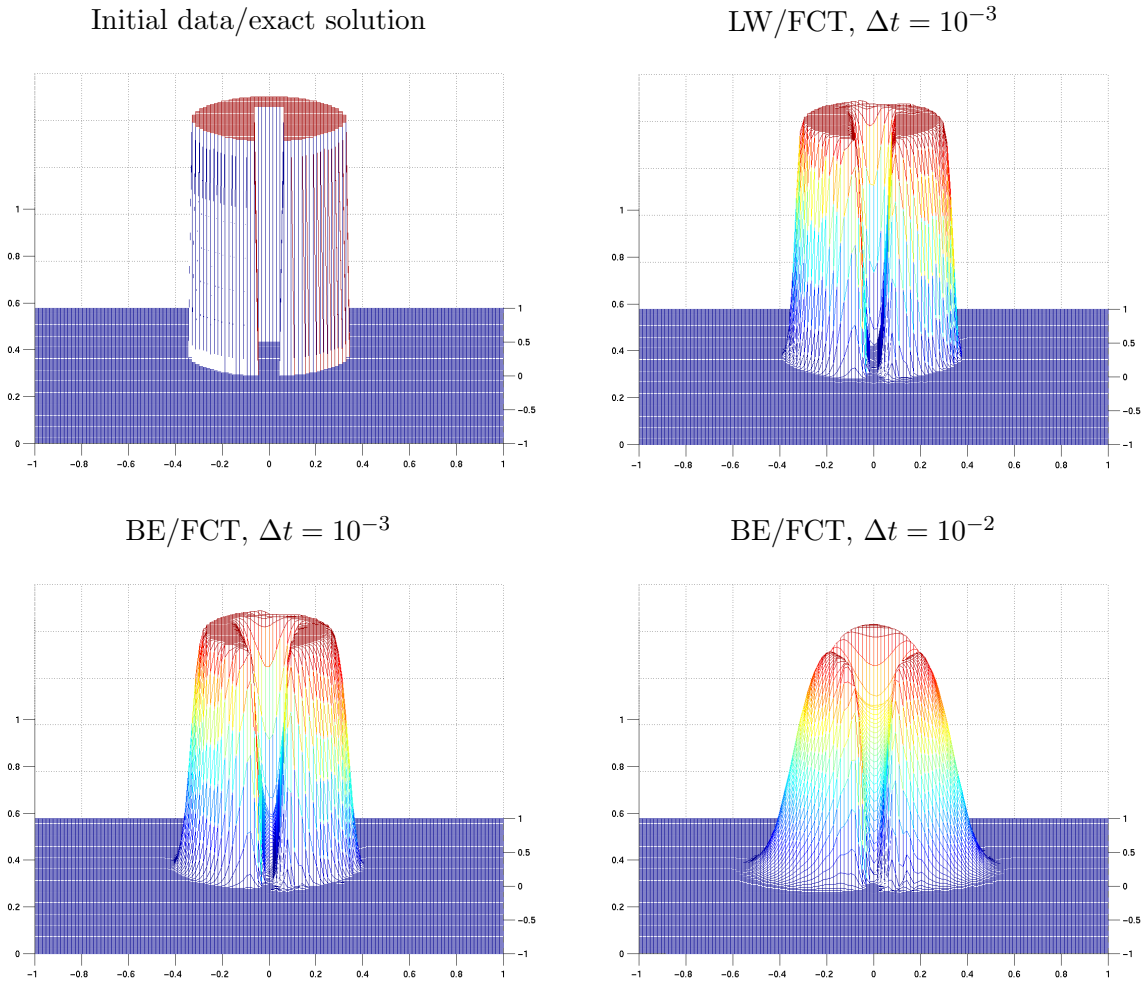


Figure 12. Rotation of a cylinder with a slot. Initial data and solution at $t = 2\pi$.

Steady-state convection-diffusion in 2D

Finally, let us illustrate the advantages of the BE/FCT method by a two-dimensional steady-state example. The convection-diffusion equation at hand reads

$$\mathbf{v} \cdot \nabla u - \epsilon \Delta u = 0 \qquad \text{in } \Omega = (0,1) \times (0,1),$$

where $\mathbf{v} = (\cos 10^{\text{o}}, \sin 10^{\text{o}})$ and $\epsilon = 10^{-3}$. The concomitant boundary conditions are:

$$\frac{\partial u}{\partial y}(x,1) = 0, \qquad u(x,0) = u(1,y) = 0, \qquad u(0,y) = \begin{cases} 1, & y \geq 0.5, \\ 0, & y < 0.5. \end{cases}$$

A reasonable initial approximation is given by

$$u^0(x,y) = \begin{cases} 1-x, & y \geq 0.5, \\ 0, & y < 0.5. \end{cases}$$

For practical applications, it is worthwhile to compute the stationary low-order solution using any direct or iterative solver, and then activate the time-dependent FEM-FCT algorithm. In this case, the cost of flux correction is minimized, since the initial guess should be close enough to the steady-state limit. Furthermore, the use of the consistent mass matrix is not justified for stationary problems, so that mass lumping is appropriate also for the high-order scheme.

The numerical solutions obtained by the BE/Galerkin and BE/FCT schemes on a uniform mesh of 32×32 bilinear elements are depicted in Figure 13. It is observed that the Galerkin method without flux correction gives rise to spurious oscillations in the boundary layer. This is obviously not the case for the flux-corrected solution, which is is highly accurate and satisfies the discrete maximum principle. It follows that BE/FCT is a promising solver for convection-dominated (quasi-) steady-state problems, which makes up for its low temporal accuracy exposed in the previous examples.

BE, $h = 1/32$ BE/FCT, $h = 1/32$

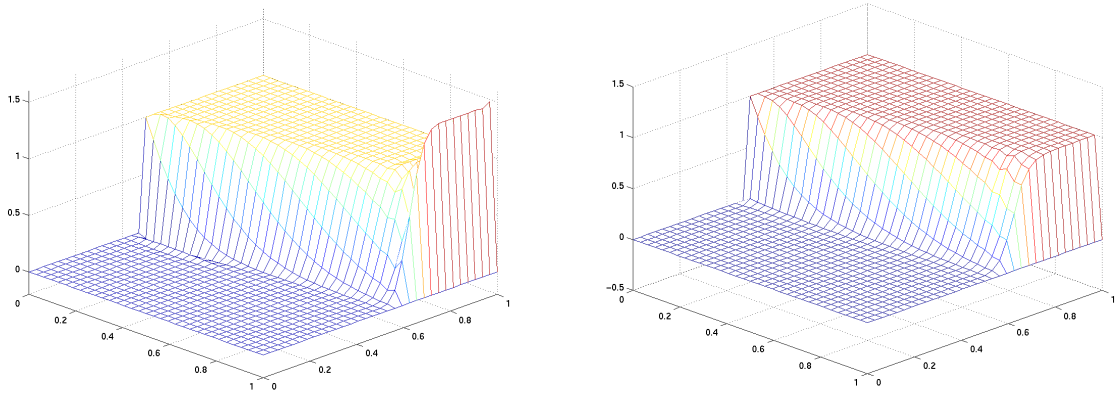


Figure 13. Steady-state convection-diffusion in 2D, $\epsilon = 10^{-3}$.

8 Conclusions and outlook

A new approach to flux correction for finite elements was presented. Its major highlights are: the novel technique for the construction of non-oscillatory low-order schemes, the flux-based representation of antidiffusive terms, and the extension of the FEM-FCT methodology to implicit time discretizations. The low-order transport operator was constructed by elimination of all negative off-diagonal entries of the discrete high-order operator. A decisive advantage of this strategy is its applicability to arbitrary finite element matrices and the fact that it yields the least diffusive positivity-preserving method which is superior to the upwind discretization if any physical diffusion is present. The structure of the discrete antidiffusion operator was exploited to decompose it into a sum of internodal fluxes which can be processed in much the same way as their finite difference counterparts. In particular, an extra prelimiting step was reintroduced to get rid of spurious ripples which are generated otherwise. The flux-based algorithm is readily portable to higher dimensions, so that the same subroutines can be used in 1D, 2D and 3D implementations. The mechanisms underlying flux correction were analyzed on the basis of rigorous positivity criteria, and an implicit version of the FEM-FCT procedure was elaborated. A unified flux limiter was devised for explicit and implicit schemes. It was proved that the fully implicit backward Euler method is unconditionally positive, whereas other schemes are subject to a CFL-like condition. The upper bound for the time step is easily computable and can be used to steer adaptive time-stepping.

The behavior of the proposed schemes was studied numerically for both evolutionary and steady-state problems. Encouraging results were obtained for a wide range of one- and two-dimensional examples. The best transient solutions were produced by the second-order schemes of Lax-Wendroff and Crank-Nicolson type. The backward Euler method is first-order accurate in time, but it constitutes an excellent solver for steady state-problems. In addition, the implicit treatment is appropriate if a non-uniform distribution of Courant numbers (due to adaptive mesh refinement or strongly varying velocities) makes the CFL condition too restrictive. In other cases, explicit or semi-implicit time-stepping should be employed for accuracy reasons. Hence, both explicit and implicit FEM-FCT schemes belong in a CFD toolbox for convection-dominated transport problems.

Apart from the simple test problems considered in this paper, we have successfully applied the new FEM-FCT algorithms to scalar transport equations governing the evolution of phase holdups and concentrations of species in gas-liquid reactors [18]. Such coupled multiphase flow problems described by two-fluid models are especially sensitive to non-physical oscillations and excessive numerical diffusion, so that the use of high-resolution schemes is indispensable [27]. One of the feasible directions for further research is the integration of flux limiters into incompressible flow solvers for the Navier-Stokes equations in the medium and high Reynolds number regime. Even though the presence of the viscous term makes the velocity less susceptible to undershoots and overshoots, linear high-order methods of the streamline diffusion type sometimes yield unsatisfactory results (e.g. in the case of strongly anisotropic meshes). Since the cost of flux correction is rather high, it might be used interchangeably with cheaper artificial viscosity methods. The latter ones can be based on the same high- and low-order transport operators but use some heuristic sensors (e.g. the local Reynolds number) to determine the blending factors.

As elucidated in the monograph [28] and illustrated by representative benchmark computations in [24], unconditionally stable implicit schemes appear to be particularly attractive for the treatment of the incompressible Navier-Stokes equations. On one hand, explicit schemes for the Burgers equation do not require any advanced linear algebra tools, since the consistent mass matrix can be efficiently 'inverted' e.g. by just a few Jacobi-like iterations using the lumped mass matrix as a preconditioner. On the other hand, the inherent Pressure Poisson Equation represents an ill-conditioned elliptic problem which has to be solved at each time step. Consequently, the CFL condition may become a formidable bottleneck, so that an implicit approach is to be preferred.

It should be emphasized that implicit schemes including those with flux correction stipulate the use of optimized multigrid techniques [24]. Otherwise the advantages of unconditional stability cannot be realized due to a disproportionally high computational cost per time step. Therefore, the development of properly tuned linear multigrid solvers is one of our top priorities. Other aspects to be investigated include the application of FEM-FCT schemes to systems of equations and locally refined unstructured grids, combination with adaptive error control mechanisms in space and time, as well as the extension to nonconforming finite elements and higher order approximations. These issues are currently under research and will be addressed in forthcoming papers.

References

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

[2] K. Baba and M. Tabata, On a conservative upwind finite element scheme for convective diffusion equations. *RAIRO Numerical Analysis* **15** (1981) 3–25.

[3] H. Blank, M. Rudgyard and A. Wathen, Stabilised finite element methods for steady incompressible flow. *Comput. Methods Appl. Mech. Engrg.* **174** (1999), no. 1-2, 91-105.

[4] 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.

[5] A. N. Brooks and T. J. R. Hughes, Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.* **32** (1982) 199-259.

[6] G. F. Carey and B. N. Jiang, Least-squares finite elements for first-order hyperbolic systems. *Int. J. Numer. Meth. Fluids* **26** (1995) 81–93.

[7] C. R. DeVore, An improved limiter for multidimensional flux-corrected transport. *NASA Technical Report* AD-A360122 (1998).

[8] 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.

[9] J. Donea, B. Roig and A. Huerta, High-order accurate time-stepping schemes for convection-diffusion problems. Barcelona: International Center for Numerical Methods in Engineering, *CIMNE Monograph* **42** (1998).

[10] J. Donea, V. Selmin and L. Quartapelle, Recent developments of the Taylor-Galerkin method for the numerical solution of hyperbolic problems. *Numerical methods for fluid dynamics III*, Oxford, 171-185 (1988).

[11] J.-J. Droux and T. J. R. Hughes, A boundary integral modification of the Galerkin least squares formulation for the Stokes problem. *Comput. Methods Appl. Mech. Engrg.* **113** (1994) 173–182.

[12] G. E. Georghiou, R. Morrow and A. C. Metaxas, An improved finite-element flux-corrected transport algorithm. *J. Comput. Phys.* **148** (1999) 605–620.

[13] P. Hansbo, Aspects of conservation in finite element flow computations. *Comput. Methods Appl. Mech. Engrg.* **117** (1994) 423-437.

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

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

[16] C. Johnson, The characteristic streamline diffusion finite element method. *Mat. Aplic. Comp.* **10** (1991), no. 3, 229–242.

[17] D. Kuzmin, A high-resolution finite element scheme for convection-dominated transport. *Commun. Numer. Meth. Engrg.* **16** (2000), no. 3, 215-223.

[18] D. Kuzmin and S. Turek, Efficient numerical techniques for flow simulation in bubble column reactors. In: *Preprints of the 5th German-Japanese Symposium on Bubble Columns*, VDI/GVC, 99-104, 2000.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

[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).

[24] M. Schäfer and S. Turek (with support of F. Durst, E. Krause, R. Rannacher), Benchmark computations of laminar flow around cylinder. In: E.H. Hirschel (editor), *Flow Simulation with High-Performance Computers II*, Volume 52 of *Notes on Numerical Fluid Mechanics*, 547–566, Vieweg, 1996.

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

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

[27] A. Sokolichin, *Mathematische Modellbildung und numerische Simulation von Gas-Flüssigkeits-Blasenströmungen*. Habilitationsschrift, Universität Stuttgart, 2000.

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

[29] W. W. Tworzydlo, J. T. Oden and E. A. Thornton, Adaptive implicit/explicit finite element method for compressible viscous flows. *Comput. Methods Appl. Mech. Engrg.* **95** (1992), no.3, 397-440.

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