

Flux Correction Tools for Finite Elements

D. Kuzmin and S. Turek

Institute of Applied Mathematics, LS III University of Dortmund, Vogelpothsweg 87

D-44227, Dortmund, Germany

E-mail: kuzmin@math.uni-dortmund.de

Received March 26, 2001; revised October 1, 2001

Flux correction in the finite element context is addressed. 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 antidiffusive 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. © 2002 Elsevier Science (USA)

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 [6]. Methods based on flux (or slope) limiting are nonlinear and quite costly, but at the same time they are very robust and yield nonoscillatory results with sharp resolution of discontinuities. There exists a variety of such schemes, most of which are 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 [36].

The past decade has witnessed the advent of discontinuous Galerkin methods, which effectively incorporate the concepts of numerical fluxes and slope limiters into the finite element framework (see [11] and references therein). DG methods possess the local conservation property and admit approximations of arbitrarily high formal order. In fact, they represent a generalization of finite volume methods which enjoys the main advantages of the finite element approach. Discontinuous approximations lead to block diagonal mass matrices, whereby the blocks can be easily inverted by hand. If Legendre polynomials are chosen as basis functions, the mass matrix is diagonal from the outset. DG methods lend themselves to h - p adaptivity, since there are no continuity constraints at interelement boundaries and no special treatment is required for “hanging nodes.” However, the number of degrees of freedom is increased as compared to continuous finite element methods. Furthermore, the inherently explicit Runge–Kutta DG schemes [9] are subject to a CFL condition which becomes increasingly restrictive for high-order methods. Another open problem is the choice of an empirical constant, which was introduced to make the “minmod” limiter less diffusive. In addition, DG methods were originally developed for hyperbolic conservation laws and experience fundamental difficulties with handling diffusion. A straightforward use of the discontinuous formulation for elliptic or parabolic equations results in a variational crime, so that it is necessary to rewrite the original problem as a first-order system for the unknown function and its derivative [2, 10] or enforce the consistency of the discrete scheme by adding extra boundary terms [3, 4].

The design of high-resolution finite element schemes based on continuous approximations is even more challenging 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, 26]. This nonrigorous extension of 1D concepts to multidimensions works well in practice but, generally, 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.* [28] 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 continuous finite element discretizations of high and low order in a mass-conserving fashion is a nontrivial task.

An elegant finite element methodology which circumvents the above difficulties was introduced by Löhner *et al.* [24, 25]. 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 [22]. 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 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. STANDARD FEM–FCT PROCEDURE

Consider a generic time-dependent conservation law

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

where u is the scalar quantity to be transported, \mathbf{v} is an externally specified velocity field, and ϵ is a diffusion coefficient. This problem is notoriously difficult to treat numerically in convection-dominated cases, i.e., when ϵ is small.

Before embarking on the development of high-resolution finite element schemes using a flux-based representation of (anti-) diffusive terms, let us recall the FEM–FCT procedure due to Löhner *et al.* [24]. Building on an earlier paper by Parrott and Christie [27], it has established the framework for implementation of Zalesak’s multidimensional limiter [36] 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 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 of reducing the error by adding compensating antidiffusion in regions where the solution is smooth and where the Taylor series expansion makes sense. The standard FEM–FCT procedure as proposed by Löhner *et al.* [24, 25] 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 nonoscillatory 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

$$u_i^{n+1} = u_i^L + \sum_e \alpha_e F_{e,i}, \quad 0 \leq \alpha_e \leq 1. \quad (2)$$

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

$$M_C \Delta u = R, \quad (3)$$

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.* [24, 25] 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 (3) clearly satisfies

$$M_L \Delta u^H = R + (M_L - M_C) \Delta u^H, \quad \Delta u^H = u^H - u^n. \quad (4)$$

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 [17]. 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.* [24] perform mass lumping and add explicit mass diffusion to transform the high-order method into a low-order one,

$$M_L \Delta u^L = R + c_d (M_C - M_L) u^n, \quad \Delta u^L = u^L - u^n, \quad (5)$$

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 [14, 31]

$$M_L u^L = M_C u^n + R, \quad (6)$$

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

$$|v| \leq \sqrt{\frac{2}{3}}.$$

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 [6]. 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 nonoscillatory results.

Antidiffusive Element Contributions

Note that if we subtract (5) from (4), 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

$$F_e = M_L^{-1} \Big|_e (\hat{M}_L - \hat{M}_C) (c_d \hat{u}^n + \Delta \hat{u}^H). \quad (7)$$

The above notation is to be understood in the following sense. The local antidiffusion operator $\hat{M}_L - \hat{M}_C$ is constructed for *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 [6] and in the old solution u^n [36]. Löhner *et al.* [24, 25] estimate the solution bounds $u^{\max/\min}$ by the following three-step algorithm:

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

$$u_i^* = \max_{\min} \{u_i^L, u_i^n\}. \quad (8)$$

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

$$u_e^{**} = \max_{\min} u_i^*, \quad i \in N_e. \quad (9)$$

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

$$u_i^{\max/\min} = \max_{\min} u_e^{**}, \quad e \in E_i. \quad (10)$$

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, for instance, if the velocity field is not divergence-free. In this case, physical extrema may decay with time (see Fig. 5), so that adopting solution bounds from the previous time step would produce an overshoot. Thus, 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 [36]. Six auxiliary quantities are defined for each node:

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

$$P_i^\pm = \sum_{e \in E_i} \max_{\min} \{0, F_{e,i}\}. \quad (11)$$

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

$$Q_i^\pm = u_i^{\max/\min} - u_i^L. \quad (12)$$

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

$$R_i^\pm = \begin{cases} \min\{1, Q_i^\pm / P_i^\pm\}, & \text{if } P_i^\pm \neq 0, \\ 0, & \text{if } P_i^\pm = 0. \end{cases} \quad (13)$$

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

$$\alpha_e = \min_{i \in N_e} \begin{cases} R_i^+, & \text{if } F_{e,i} \geq 0, \\ R_i^-, & \text{if } F_{e,i} < 0. \end{cases} \quad (14)$$

It is safe 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.

3. POSITIVITY AND LED CRITERIA

In order to derive an alternative FEM–FCT formulation applicable to implicit time stepping schemes, we need to introduce some mathematical tools. This section is devoted to the analysis of properties which the discrete scheme must satisfy to maintain positivity of the numerical solution. In subsequent sections we will consider the issue of mass conservation and the structure of discrete diffusion operators, which are also of paramount importance for the development of high-resolution finite element schemes.

Let the governing equation (1) be discretized on an arbitrary (possibly unstructured) mesh. Assume that the semi-discrete problem can be represented in the form

$$\frac{du_i}{dt} = \sum_j c_{ij} u_j, \quad \sum_j c_{ij} = 0, \quad (15)$$

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

$$\frac{du_i}{dt} = \sum_{j \neq i} c_{ij} (u_j - u_i). \quad (16)$$

Furthermore, suppose that all coefficients are nonnegative: $c_{ij} \geq 0, j \neq i$. Then such a 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 [18, 19] 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

$$TV(u) = \int_{-\infty}^{+\infty} \left| \frac{\partial u}{\partial x} \right| dx \quad (17)$$

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

$$TV(u) = 2 \left(\sum \max u - \sum \min u \right). \quad (18)$$

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 nonoscillatory schemes. The advantage of the LED principle as compared to TVD concepts is its applicability to multidimensional problems on both structured and unstructured meshes.

Recall that Eqs. (15) and (16) 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

$$\frac{u_i^{n+1} - u_i^n}{\Delta t} = \theta \sum_{j \neq i} c_{ij} (u_j^{n+1} - u_i^{n+1}) + (1 - \theta) \sum_{j \neq i} c_{ij} (u_j^n - u_i^n), \quad 0 \leq \theta \leq 1. \quad (19)$$

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 1. *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 Eq. (19) 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

$$u_k^{n+1} = u_k^n + \Delta t \sum_{j \neq k} c_{kj} (u_j^{n+1} - u_k^{n+1}). \quad (20)$$

By inductive assumption, the old solution u^n must be nonnegative everywhere. The coefficients c_{kj} are also nonnegative 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 (19) will preserve positivity if the explicit term satisfies the inequality

$$u_i^n + \Delta t(1 - \theta) \sum_{j \neq i} c_{ij} (u_j^n - u_i^n) \geq 0 \quad \forall i. \quad (21)$$

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

$$1 + \Delta t(1 - \theta) \min_i c_{ii} \geq 0, \quad (22)$$

where $c_{ii} = -\sum_{j \neq i} c_{ij}$ are the diagonal elements of the original coefficient matrix defined by Eq. (15). This condition provides the desired positivity criterion, which can be used for the time step control. ■

If the discrete scheme is not local extremum diminishing (e.g., due to compressibility or the presence of reactive terms), then a weaker positivity criterion should be applied. It is based on the concept of an *M-matrix* as elucidated below.

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

If A is strictly diagonally dominant and $a_{ii} > 0$, while $a_{ij} \leq 0$ for $i \neq j$, then A is an *M-matrix* (see, e.g., [29]). Note that for *M-matrices* $Ax \geq 0$ implies that $x \geq 0$. This motivates the following generalization of the positivity theorem:

POSITIVITY THEOREM 2. *Let the numerical scheme be represented in abstract matrix operator form as*

$$Lu^{n+1} = Ru^n.$$

A sufficient condition for such scheme to preserve positivity is that L be an M-matrix and all entries of R be nonnegative ($R \geq 0$).

Proof. The inverse of L is nonnegative due to the *M-matrix* property. Thus, $u^{n+1} = L^{-1}Ru^n \geq 0$, as long as $u^n \geq 0$. ■

Remark 1. Conditions of the Theorem are sufficient (but not necessary) to guarantee that the numerical solution satisfies the discrete maximum principle.

Remark 2. It seems expedient to require that the steady-state counterpart of L be an *M-matrix* as well. Otherwise, nonphysical ripples might emerge even though the solution remains positive.

Remark 3. As before, the time step can affect the sign of matrix entries, so that the condition $R \geq 0$ yields a CFL-like upper bound for explicit schemes.

The above positivity criteria lay the groundwork for the construction of high-resolution numerical schemes. The desired properties of discrete operators can be realized by the introduction of artificial diffusion or by the use of upwind biasing. 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. We have already seen how this can be accomplished in the framework of the FEM–FCT methodology.

4. MASS CONSERVATION

Conservation of mass is crucial to the design of numerical methods for the bulk of transport problems [17]. 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. Nonconservative 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 Eq. (1) reads

$$\int_{\Omega} \left[\frac{\partial u}{\partial t} + \nabla \cdot (\mathbf{v}u) - \nabla \cdot (\epsilon \nabla u) \right] w \, d\mathbf{x} = 0, \quad \forall w. \quad (23)$$

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

$$\frac{d}{dt} \int_{\Omega} u \, d\mathbf{x} = - \int_S (\mathbf{v}u - \epsilon \nabla u) \cdot \mathbf{n} \, ds, \quad (24)$$

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 and discontinuous Galerkin methods apply formulation (24) directly to each element of the triangulation, so that mass conservation is enforced not only globally but also locally. Flux correction for such discontinuous approximations is fairly straightforward. The objective of this paper is to extend the available FCT machinery to continuous (linear and multilinear) finite elements.

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 nonconservative 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 postprocessing step for the recovery of the lost mass [21].

5. 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 justified in different ways. The concept of streamline diffusion was introduced by Brooks and Hughes [7] and employed within a consistent Petrov–Galerkin formulation. A similar approach was followed by Johnson [20] and his collaborators. The least-squares formulation [8] 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 [13].

- 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. As we have seen, mass diffusion can be used for the construction of low-order finite element schemes to be combined with high-order ones in the FCT framework.

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

$$d_{ij} = d_{ji} \tag{25}$$

and zero row/column sums

$$\sum_j d_{ij} = \sum_i d_{ij} = 0 \tag{26}$$

(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

$$(Du)_i = \sum_j d_{ij} u_j = \sum_{j \neq i} d_{ij} (u_j - u_i) \quad (27)$$

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

$$(Du)_i = \sum_{j \neq i} f_{ij}, \quad f_{ji} = -f_{ij}. \quad (28)$$

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.

6. NEW FEM-FCT PROCEDURE

Now we are in a position to derive a flux-based FCT formulation as an alternative to the element-based approach of Löhner *et al.* [24, 25]. 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. The flux-based decomposition of (anti-) diffusive terms introduced above appears to be much more flexible and efficient. High-resolution finite element schemes of this type were proposed in [14, 31, 32]. 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. We will follow a similar approach while using rigorous positivity criteria to develop both explicit and implicit FCT schemes. The main ideas behind the new methodology are presented below.

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 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 [27]. However, upwinding is rather cumbersome and unnatural in the finite element context, which has led Löhner *et al.* [24] 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 chosen

is 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.* [16], 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 such algorithm can be expected to work well only on very regular meshes and may fail to preserve positivity. Another attempt to improve on Löhner’s method (without referring to the original paper) was recently published in [33]. In contrast to [16], a constant artificial diffusion coefficient was used to enforce the M-matrix property. In what follows, we present a new approach, which does reconcile the conflicting demands for accuracy and positivity of the low-order scheme.

If the flow is incompressible, Eq. (1) can be written in the nonconservative 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, we employ mass lumping to remove the implicit antidiffusion intrinsic to the consistent mass matrix. The resulting scheme can be cast into the form (16) and would possess the LED property if all coefficients k_{ij}^H , $j \neq i$ were nonnegative. This suggests the 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:

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

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 nonnegative 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 Eq. (1) is not divergence-free, a local accumulation of the conserved quantity can occur. Therefore, the formation of physical extrema must be reckoned with. Nevertheless, the left-hand side operator will remain an M-matrix if the low-order operator K^L is constructed as above. According to the second theorem, positivity will be preserved under a proper CFL-like condition. Depending on the sign of $\nabla \cdot \mathbf{v}$, 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 $\nabla \cdot \mathbf{v} \ll 0$. However, this is very unlikely to happen for any practical flows of interest.

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

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

that is

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

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 first positivity theorem, the backward Euler time discretization of a LED scheme is unconditionally positive, while other time stepping schemes preserve positivity as long as

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

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 (6), 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

$$\frac{\partial u}{\partial t} + v \frac{\partial u}{\partial x} = 0 \quad (37)$$

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

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

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

$$\frac{du_i}{dt} = -v \frac{u_{i+1} - u_{i-1}}{2\Delta x}. \quad (39)$$

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

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

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

$$\frac{du_i}{dt} = -v \frac{u_i - u_{i-1}}{\Delta x}. \quad (41)$$

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

$$v \frac{\Delta t}{\Delta x} \leq \frac{1}{1 - \theta}. \quad (42)$$

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.* [16], 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 does not matter, so that finite element matrices resulting from the discretization of 1D, 2D, and 3D problems can be treated in exactly the same way.

Flux-Based FEM–FCT Formulation

Our next step is to reformulate the FEM–FCT procedure in terms of internodal fluxes and to extend it to implicit time stepping schemes, so as to eliminate or alleviate severe time step restrictions inherent to explicit methods. Of primary interest are the backward Euler and the Crank–Nicolson scheme. Both of them are unconditionally stable and can be used as a high-order method in conjunction with the Galerkin spatial discretization. No extra stabilization of convective terms is required in this case. At the same time, the fully explicit Galerkin method needs to be stabilized. A classical way to accomplish this is to add a proper amount of artificial diffusion in the streamline direction. In this paper, we use the Lax–Wendroff method, whereby the streamline diffusion corresponds to 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 higher order Taylor–Galerkin methods, the reader is referred to [13].

The high-order transport operator can be transformed into a low-order one as explained above. The resulting methods discretized in time by the standard θ -scheme are related by the formula

$$(M_L - \theta \Delta t K^L)u^H = (M_L + (1 - \theta)\Delta t K^L)u^n + F(u^H, u^n), \quad (43)$$

where the antidiffusion responsible for high spatial accuracy is given by

$$F(u^H, u^n) = -(M_C - M_L)\Delta u^H - \Delta t(K^L - K^H)[\theta u^H + (1 - \theta)u^n] + \Delta t S u^n. \quad (44)$$

Here S stands for the streamline diffusion operator which is present only in the fully explicit scheme. If the antidiffusive term $F(u^H, u^n)$ is omitted, then the positive low-order scheme is recovered, whereas retaining it yields the original high-order method.

It can readily be seen that all the matrices in (44) represent discrete (anti-) diffusion operators and thereby lend themselves to decomposition into fluxes

$$\begin{aligned} f_{ij} &= -m_{ij}(\Delta u_j^H - \Delta u_i^H) - \Delta t d_{ij}[\theta(u_j^H - u_i^H) + (1 - \theta)(u_j^n - u_i^n)] \\ &\quad + \Delta t s_{ij}(u_j^n - u_i^n), \quad f_{ji} = -f_{ij}, \quad i < j. \end{aligned} \quad (45)$$

These raw antidiffusive fluxes offset the errors induced by mass lumping, “upwinding,” and first-order time discretization (for the explicit scheme). Coefficients m_{ij} , d_{ij} , and s_{ij} denote the entries of the consistent mass matrix, artificial diffusion, and streamline diffusion operators, respectively.

In light of the above, the flux-corrected version of (43) can be written in the form

$$m_i u_i^{n+1} - \theta \Delta t \sum_j k_{ij}^L u_j^{n+1} = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad \alpha_{ji} = \alpha_{ij}, \quad (46)$$

where α_{ij} denote the correction factors (to be defined below), while \tilde{u} represents the positivity-preserving solution to the explicit subproblem

$$m_i \tilde{u}_i = m_i u_i^n + (1 - \theta)\Delta t \sum_j k_{ij}^L u_j^n. \quad (47)$$

In essence, \tilde{u} corresponds to an intermediate solution computed at the time instant $t^{n+1-\theta}$ by the explicit low-order scheme. It reduces to u^n for the backward Euler method ($\theta = 1$) and to u^L for the forward Euler method ($\theta = 0$).

The newly introduced family of FCT schemes distinguishes itself in that it is applicable to explicit and implicit time discretizations alike. The fully explicit scheme is consistent with the standard FCT methodology. Note that implicit schemes require solving *two* non-symmetric linear systems per time step: one for the high-order solution (which is needed to compute the antidiffusive fluxes) and one for the final solution. Nevertheless, implicit methods are typically more efficient than explicit ones because larger time steps can be taken. If iterative solvers are employed, the high-order solution provides a reasonable initial guess for the final solution.

It is obvious that the success of the FCT algorithm depends on the positivity of the provisional solution \tilde{u} and on the choice of correction factors α_{ij} . For \tilde{u} to be positive, the time step must satisfy the CFL-like condition (36) unless the scheme is fully implicit ($\tilde{u} = u^n$). As long as the left-hand side operator is an M-matrix, our positivity theorems guarantee that the scheme (46) can be rendered positive by tuning the correction factors. 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 for explicit and implicit schemes.

Limiting Strategy

Let us start with an optional but important component of the FCT limiter. It turns out that explicit FCT schemes can benefit from canceling all antidiffusive fluxes directed down the gradient of \tilde{u} :

$$f_{ij} := 0, \quad \text{if } f_{ij}(\tilde{u}_i - \tilde{u}_j) < 0. \tag{48}$$

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 [12].

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 of steepen the gradient, while the effect of (48) 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 step is missing in the FEM-FCT procedure of Löhner *et al.* [24, 25] 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 [12] 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 \max_{\min} limiter but is derived and interpreted in a quite different way. As before, we denote by u_i^{\max} the maximum and minimum solution values at the stencil S_i which consists of the node i and its nearest neighbors:

$$u_i^{\max} = \max_{\min} \tilde{u}_j, \quad j \in S_i. \tag{49}$$

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.

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

$$\alpha_{ij} = 0, \quad \text{if } \tilde{u}_i = u_i^{\max}, \quad f_{ij} > 0 \quad \text{or} \quad \tilde{u}_i = u_i^{\min}, \quad f_{ij} < 0. \quad (50)$$

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 scheme (46) admits the representation

$$RHS = m_i \tilde{u}_i + \sum_{j \neq i} \alpha_{ij} f_{ij} = m_i \tilde{u}_i + c_i Q_i, \quad c_i = \frac{\sum_{j \neq i} \alpha_{ij} f_{ij}}{Q_i}, \quad (51)$$

where the multiplier Q_i is chosen to be

$$Q_i = \begin{cases} Q_i^+ = u_i^{\max} - \tilde{u}_i, & \text{if } \sum_{j \neq i} \alpha_{ij} f_{ij} > 0, \\ Q_i^- = u_i^{\min} - \tilde{u}_i, & \text{if } \sum_{j \neq i} \alpha_{ij} f_{ij} < 0, \\ 1, & \text{if } \sum_{j \neq i} \alpha_{ij} f_{ij} = 0. \end{cases} \quad (52)$$

By virtue of (50), we have $Q_i \neq 0$, so that no division by zero takes place. Furthermore, the coefficient, c_i is always nonnegative. Let the local extremum u_i^{\max} be attained at a node k adjacent to the node i . Then the antidiffusive term exhibits a LED structure, and we obtain

$$RHS = m_i \tilde{u}_i + c_i (\tilde{u}_k - \tilde{u}_i) = (m_i - c_i) \tilde{u}_i + c_i \tilde{u}_k, \quad c_i \geq 0. \quad (53)$$

According to the positivity theorems, the FEM–FCT scheme (46) 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

$$P_i^\pm = \frac{1}{m_i} \sum_{j \neq i} \max_{\min} \{0, f_{ij}\}, \quad R_i^\pm = \begin{cases} \min\{1, Q_i^\pm / P_i^\pm\}, & \text{if } P_i^\pm \neq 0, \\ 0, & \text{if } P_i^\pm = 0. \end{cases} \quad (54)$$

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

$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij} \geq 0, \\ \min\{R_j^+, R_i^-\} & \text{if } f_{ij} < 0. \end{cases} \quad (55)$$

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 (50) 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:

$$\sum_{j \neq i} \alpha_{ij} f_{ij} \leq \sum_{j \neq i} \alpha_{ij} \max\{0, f_{ij}\} \leq m_i R_i^+ P_i^+ \leq m_i Q_i^+. \quad (56)$$

In much the same way, it can be verified that

$$\sum_{j \neq i} \alpha_{ij} f_{ij} \geq \sum_{j \neq i} \alpha_{ij} \min\{0, f_{ij}\} \geq m_i R_i^- P_i^- \geq m_i Q_i^-. \tag{57}$$

This proves that the corrected antidiffusive fluxes satisfy the constraint $m_i \geq c_i$.

Defect Correction

Many practical applications are described by *nonlinear* conservation laws. A typical example is given by 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 for implicit schemes. 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. If we consider an abstract nonlinear system of the form

$$A(u)u = b, \tag{59}$$

then the basic nonlinear iteration can be formulated as

$$u^{(l+1)} = u^{(l)} - [C(u^{(l)})]^{-1} (A(u^{(l)})u^{(l)} - b), \tag{60}$$

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

For a nonlinear problem of the form (43), it is reasonable to use the low-order operator as the preconditioner:

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

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

$$(M_L - \theta \Delta t K^L(u^{(l)}))u^{(l+1)} = (M_L + (1 - \theta) \Delta t K^L(u^n))u^n + F(u^{(l)}, u^n). \tag{61}$$

The last term is composed from the (limited) antidiffusive fluxes. 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.

Even if the problem at hand is linear, it might be worthwhile to equip implicit schemes with an outer defect correction loop. At large time steps, the matrix of the high-order system degenerates, so that an iterative method may fail to converge. This can be rectified by resorting to defect correction, which approximates the original matrix by a well-behaved preconditioner. The M-matrix properties of the low-order operator make it particularly amenable to iterative solution, which results in a very robust solver.

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 where 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 [5]. 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 [13]. Inclusion of similar integrals into the Galerkin least squares formulation also yields the desired effect [15].

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 on 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 nonvanishing 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

$$d_{ii} = - \sum_{k \neq i} d_{ik}, \quad d_{ij} = \max \{0, -k_{ij}^H\}, \quad d_{ji} = 0 \quad (62)$$

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

$$\hat{D} = \frac{v}{2} \begin{bmatrix} -1 & 1 \\ 0 & 0 \end{bmatrix} \Rightarrow \hat{K}^L = \frac{v}{2} \begin{bmatrix} 0 & 0 \\ 1 & -1 \end{bmatrix}, \quad (63)$$

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 nonnegative, 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 one of the nodes belongs to the outflow boundary, while the other one resides in the interior. 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 new 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 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 Fig. 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

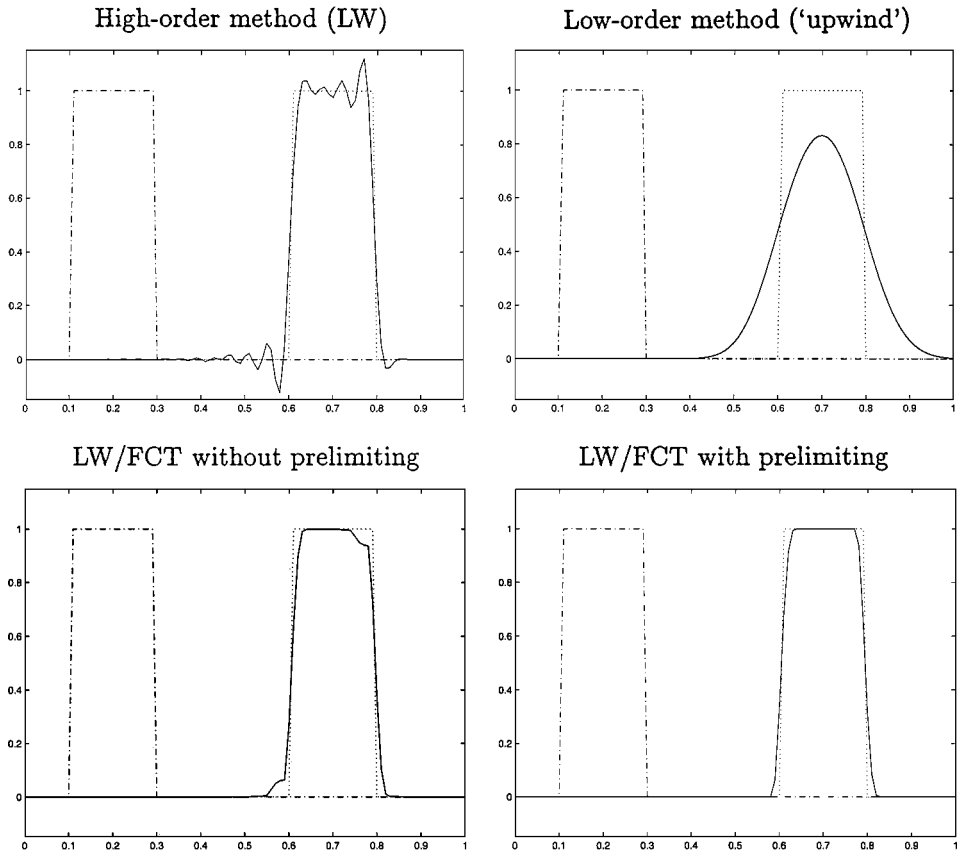


FIG. 1. Convection of a step function. Lax–Wendroff/FCT scheme, $t = 0.5$.

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 Fig. 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

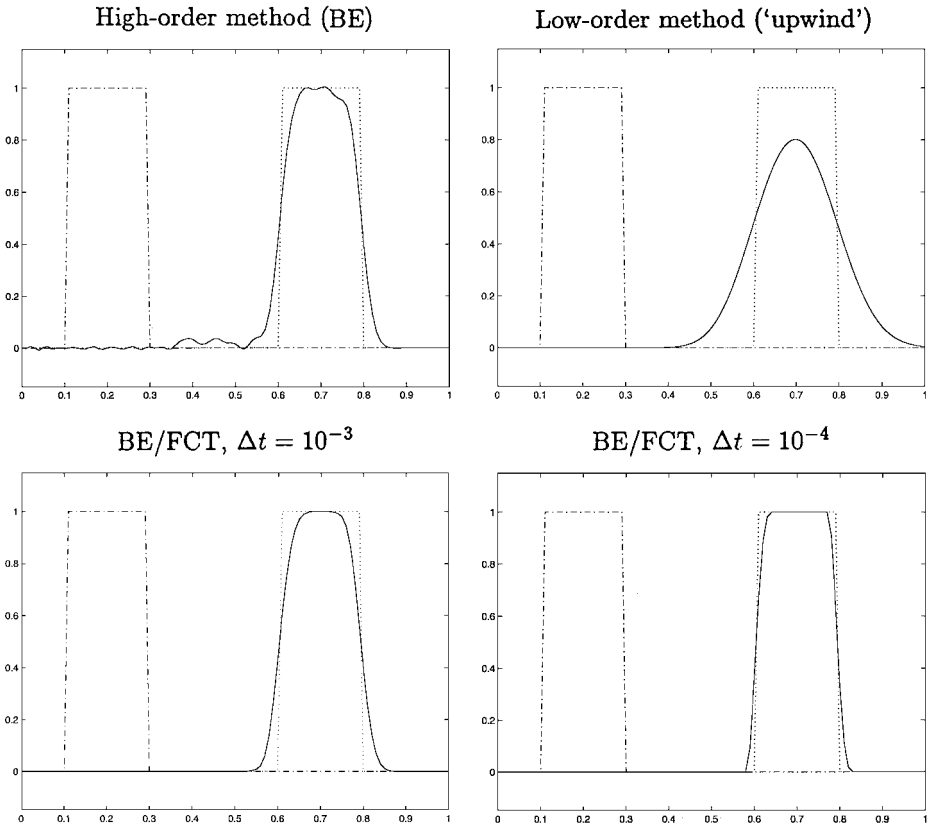


FIG. 2. Convection of a step function. Backward Euler/FCT scheme, $t = 0.5$.

by the FEM–FCT schemes are displayed in Fig. 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. This indicates that realistic applications can be efficiently simulated. 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 Fig. 4. Hence, flux correction is superfluous in this case. However, it is often 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 Fig. 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.

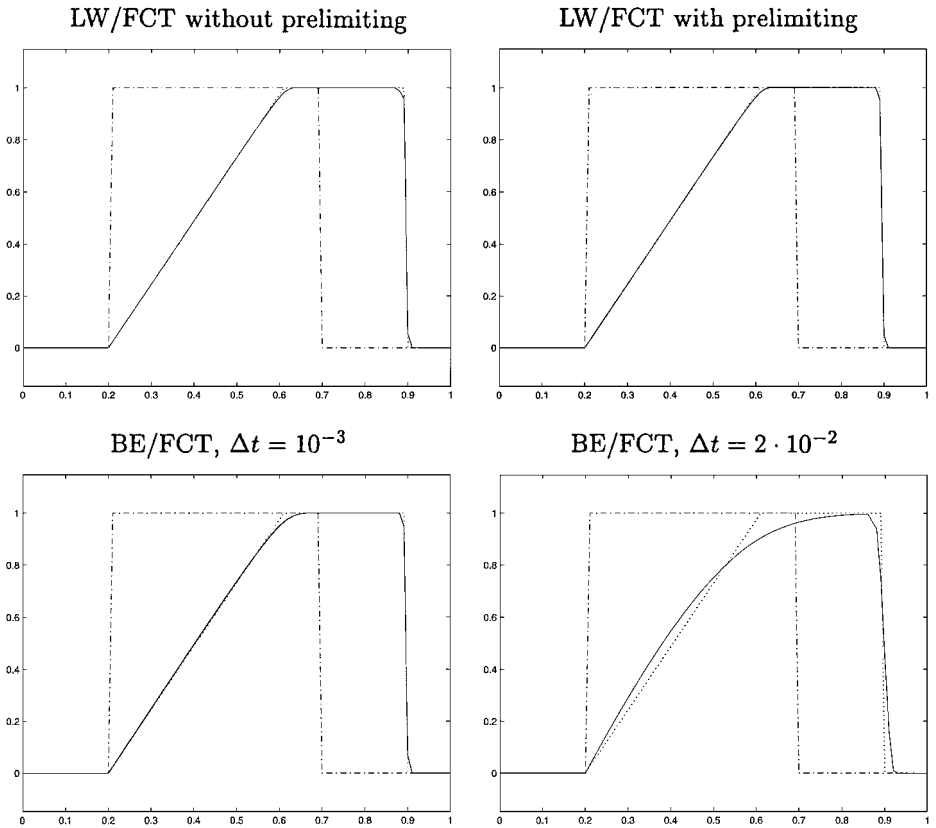


FIG. 3. Inviscid Burgers equation. Solution at $t = 0.4$.

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 nonuniform 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 Fig. 5. In this case, both LW/FCT and BE/FCT deliver nonoscillatory 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 Fig. 6, where the velocity is taken to be

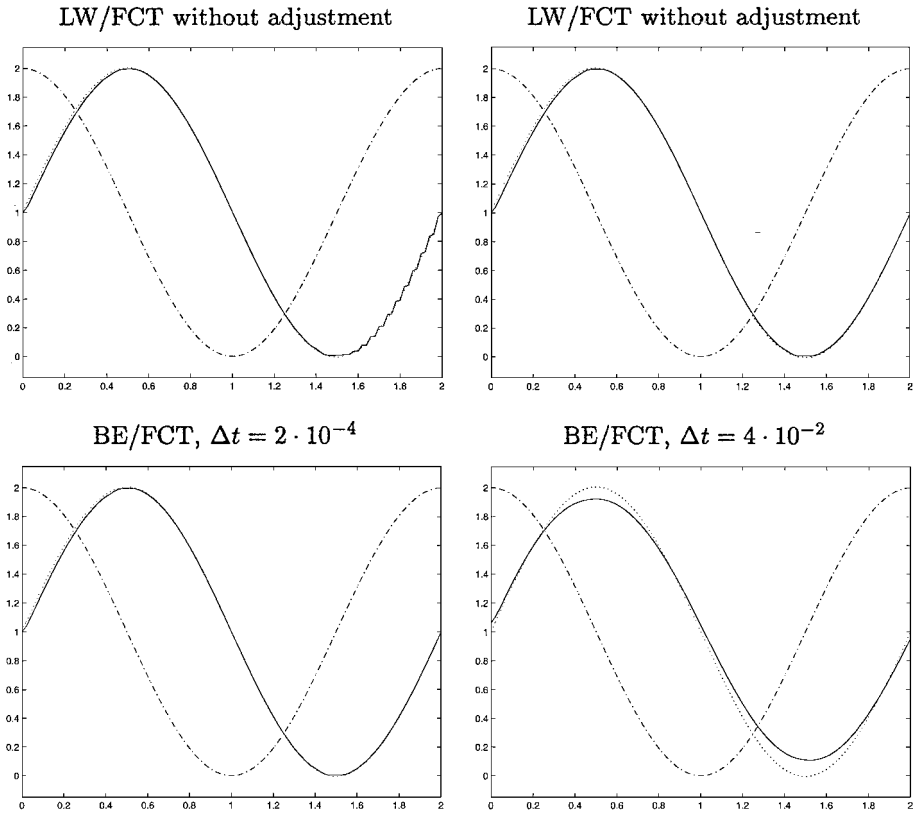


FIG. 4. Convection of a cosine wave. Solution at $t = 0.5$.

$v = 1 - x$. 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 following 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

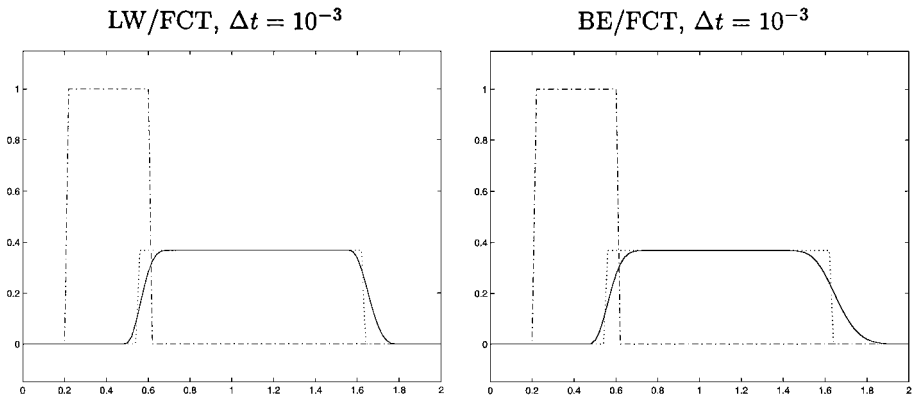


FIG. 5. Stretching by the variable velocity field $v = x$. Solution at $t = 1.0$.

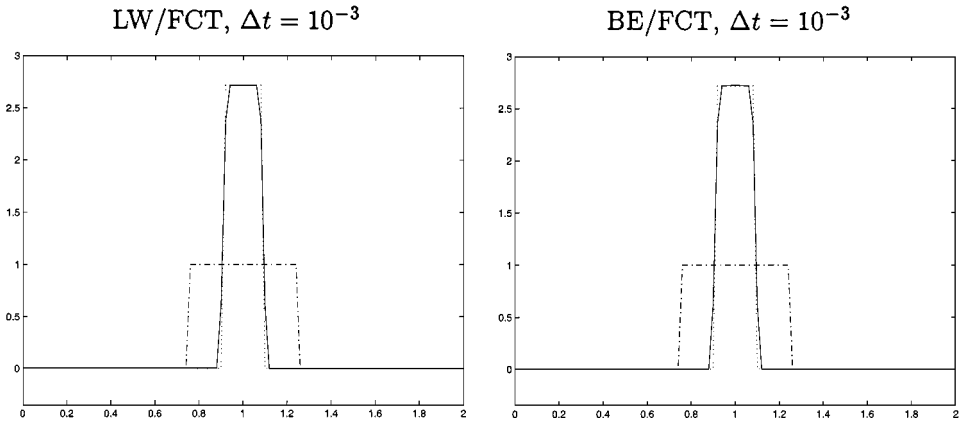


FIG. 6. Compression by the variable velocity field $v = 1 - x$. Solution at $t = 1.0$.

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 Fig. 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 [34]. Thus, no artificial steepening of the profile takes place.

Convection of the same function with the variable velocity $v = x$ is investigated in Fig. 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.

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

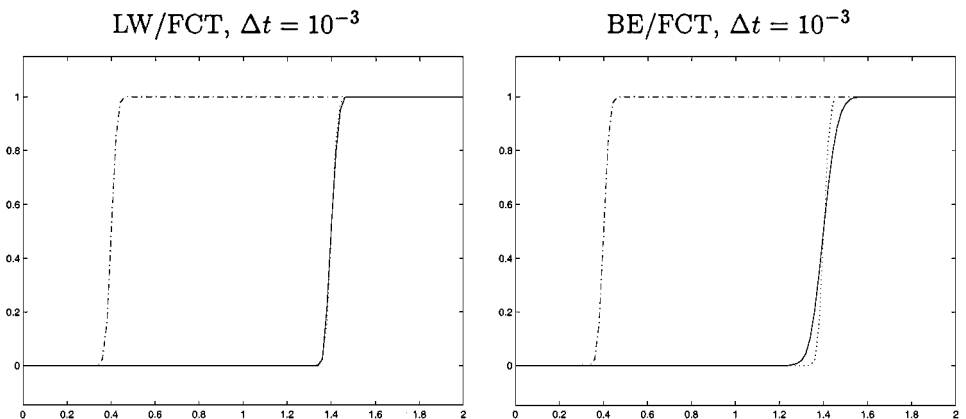


FIG. 7. Convection of a monotone profile with $v = 1$. Solution at $t = 1.0$.

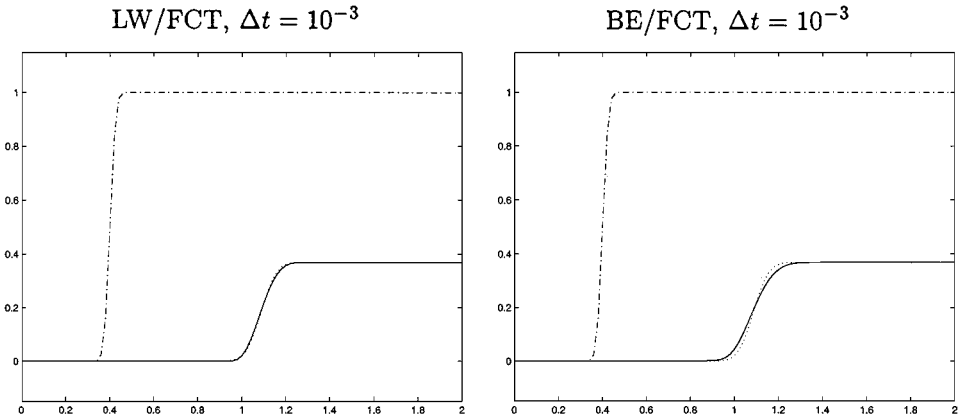


FIG. 8. Convection of a monotone profile with $v = x$. Solution at $t = 1.0$.

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 [5]. 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 noncompetitive, 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

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

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 Fig. 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.

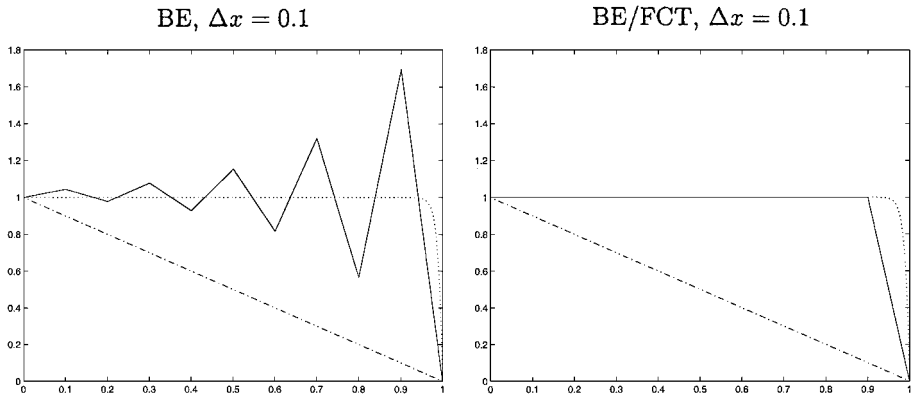


FIG. 9. Steady-state convection–diffusion in 1D, $\epsilon = 10^{-2}$.

Convection of a Discontinuous Profile in 2D

The first two-dimensional example shown in Fig. 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 on 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

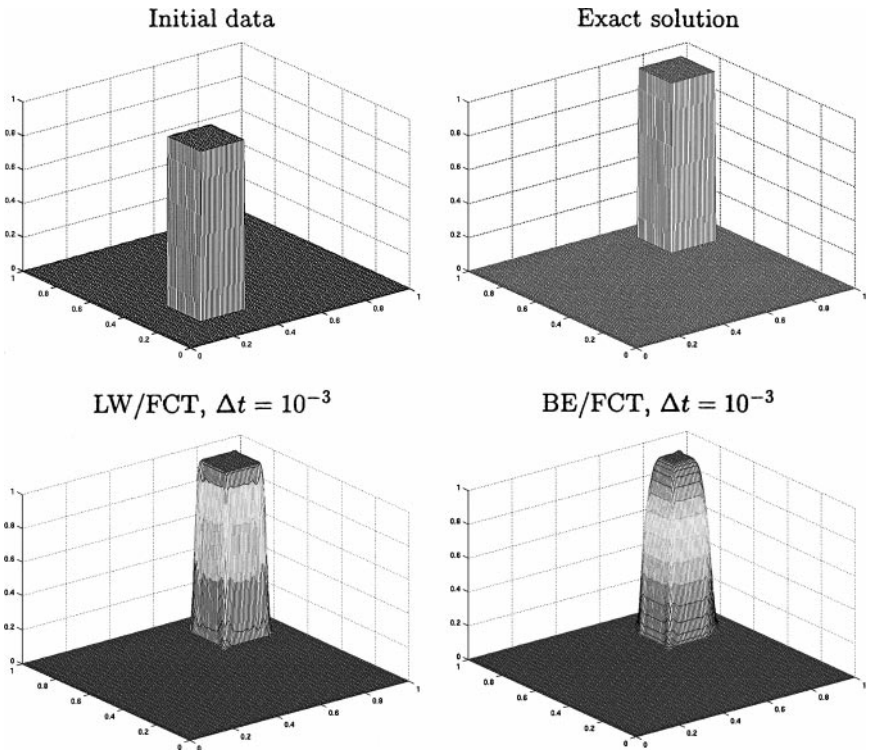


FIG. 10. Convection of a discontinuous profile. Initial data and solution at $t = 0.5$.

examination. Both methods succeed in the elimination of wiggles and preserve the steepness of the profile fairly well, unlike the underlying low-order scheme. What might look like an overshoot at the leading corner is just an optical effect due to lateral erosion (actually $0 \leq u \leq 1$ everywhere). The temporal error induced by the backward Euler time stepping is still nonnegligible 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.

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 Fig. 11. All remarks regarding the treatment of outflow boundaries remain valid

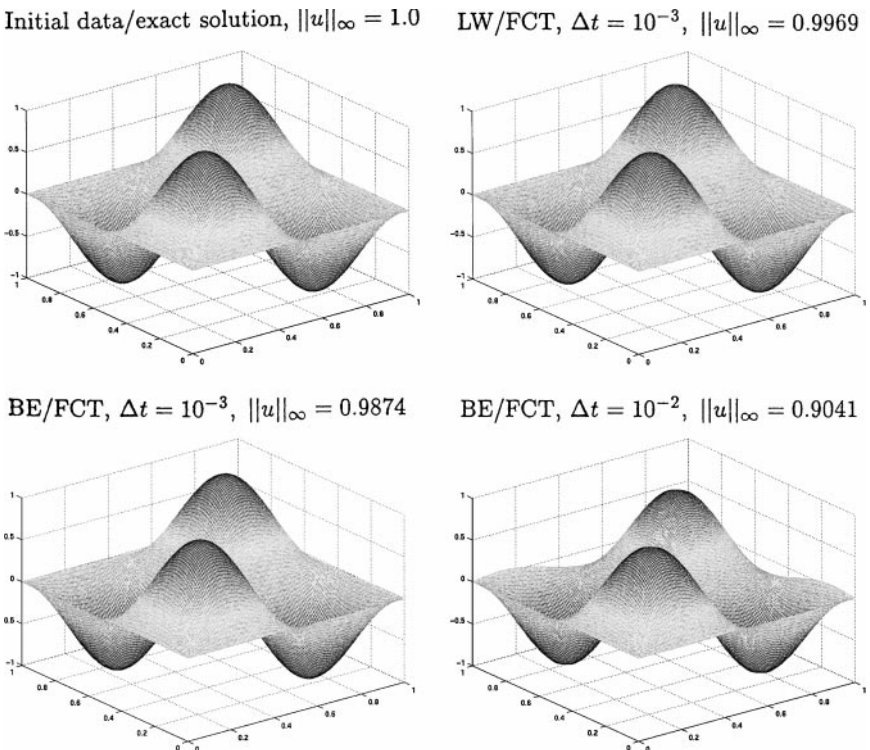


FIG. 11. Convection of a smooth profile. Initial data and solution at $t = 1.0$.

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.

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.* [25], Zalesak [36], 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

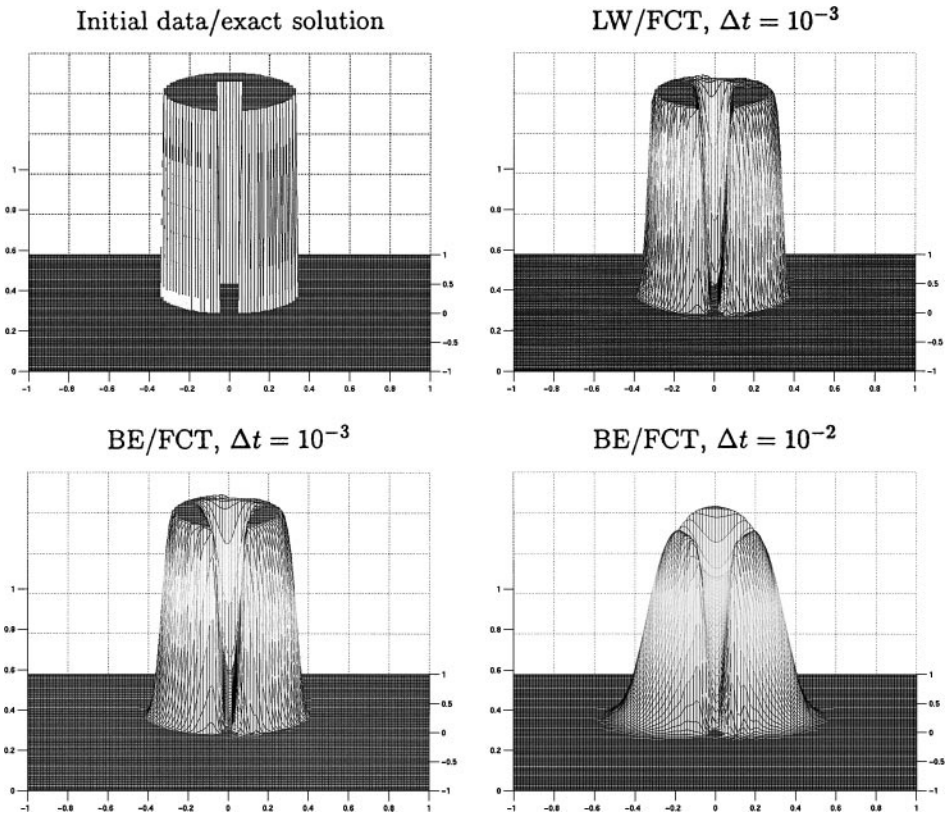


FIG. 12. Rotation of a cylinder with a slot. Initial data and solution at $t = 2\pi$.

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.

Steady-State Convection–Diffusion in 2D

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

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

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

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

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 Fig. 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 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.

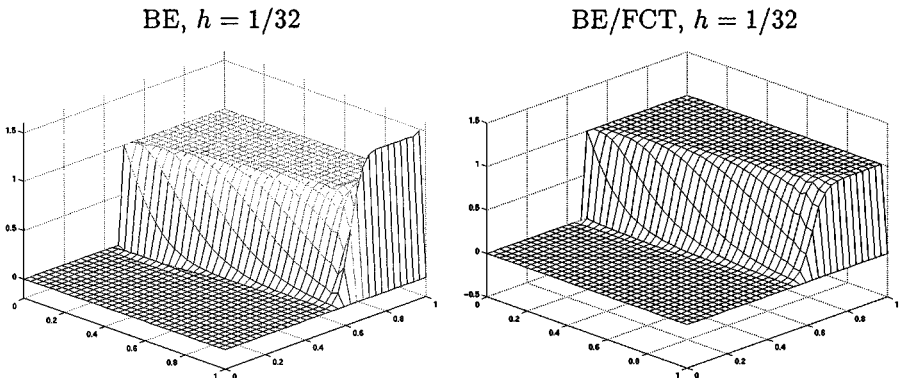


FIG. 13. Steady-state convection–diffusion in 2D, $\epsilon = 10^{-3}$.

8. CONCLUSIONS AND OUTLOOK

A new approach to flux correction for finite elements was presented. Its major highlights include: the novel technique for the construction of nonoscillatory 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 nonuniform 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 of gas–liquid reactors [23]. 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 [34]. 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 [35] and illustrated by representative benchmark computations in [30], 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 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 a more 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 [30]. Otherwise the advantages of unconditional stability cannot be realized due to a disproportionately 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. F. Bassi and S. Rebay, A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier–Stokes equations, *J. Comput. Phys.* **131**, 267 (1997).
3. C. E. Baumann and J. T. Oden, A discontinuous hp finite element method for convection-diffusion problems, *Comput. Meth. Appl. Mech. Eng.* **175**, 311 (1999).
4. R. Becker and P. Hansbo, Discontinuous Galerkin methods for convection-diffusion problems with arbitrary Peclet number, in *Proceedings of the 3rd European Conference on Numerical Mathematics and Advanced Applications* (World Scientific, Singapore, 2000), pp. 100–109.
5. H. Blank, M. Rudgyard, and A. Wathen, Stabilised finite element methods for steady incompressible flow, *Comput. Meth. Appl. Mech. Eng.* **174**, 91 (1999).
6. J. P. Boris and D. L. Book, Flux-corrected transport. I. SHASTA, A fluid transport algorithm that works, *J. Comput. Phys.* **11**, 38 (1973).
7. 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. Meth. Appl. Mech. Eng.* **32**, 199 (1982).
8. G. F. Carey and B. N. Jiang, Least-squares finite elements for first-order hyperbolic systems, *Int. J. Numer. Meth. Fluids* **26**, 81 (1988).
9. B. Cockburn and C.-W. Shu, The Runge–Kutta discontinuous Galerkin method for conservation laws. V: Multidimensional systems, *J. Comput. Phys.* **141**, 199 (1998).
10. B. Cockburn and C.-W. Shu, The local discontinuous Galerkin method for time-dependent convection-diffusion systems, *SIAM J. Numer. Anal.* **35**, 2440 (1998).
11. B. Cockburn, G. E. Karniadakis, and C.-W. Shu, The development of discontinuous Galerkin methods, in *Discontinuous Galerkin methods. Theory, computation and applications*, Lecture Notes Computational Science Engineering, Springer-Verlag, New York (2000), Vol. 11, pp. 3–50.
12. C. R. DeVore, *An Improved Limiter for Multidimensional Flux-Corrected Transport*, NASA Technical Report AD-A360122 (1998).
13. J. Donea, L. Quartapelle, and V. Selmin, An analysis of time discretization in the finite element solution of hyperbolic problems, *J. Comput. Phys.* **70**, 463 (1987).
14. J. Donea, V. Selmin, and L. Quartapelle, Recent developments of the Taylor–Galerkin method for the numerical solution of hyperbolic problems, in *Numerical Methods for Fluid Dynamics* (Oxford Univ. Press., London, 1988), Vol. III., pp. 171–185.

15. J.-J. Droux and T. J. R. Hughes, A boundary integral modification of the Galerkin least squares formulation for the Stokes problem, *Comput. Meth. Appl. Mech. Eng.* **113**, 173 (1994).
16. G. E. Georghiou, R. Morrow, and A. C. Metaxas, An improved finite-element flux-corrected transport algorithm, *J. Comput. Phys.* **148**, 605 (1999).
17. P. Hansbo, Aspects of conservation in finite element flow computations, *Comput. Meth. Appl. Mech. Eng.* **117**, 423 (1994).
18. A. Jameson, Computational algorithms for aerodynamic analysis and design, *Appl. Numer. Math.* **13**, 383 (1993).
19. A. Jameson, Positive schemes and shock modelling for compressible flows, *Int. J. Numer. Meth. Fluids* **20**, 743 (1995).
20. C. Johnson, The characteristic streamline diffusion finite element method, *Mat. Appli. Comp.* **10**, 229 (1991).
21. D. Kuzmin, A high-resolution finite element scheme for convection-dominated transport, *Commun. Numer. Meth. Eng.* **16**, 215 (2000).
22. D. Kuzmin, Positive finite element schemes based on the flux-corrected transport procedure, in *Computational Fluid and Solid Mechanics* (Elsevier, Amsterdam/New York, 2001), pp. 887–888.
23. 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, 2000), pp. 99–104.
24. 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**, 1093 (1987).
25. R. Löhner, K. Morgan, M. Vahdati, J. P. Boris, and D. L. Book, FEM–FCT: combining unstructured grids with high resolution, *Commun. Appl. Numer. Meth.* **4**, 717 (1988).
26. 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**, 827 (1994).
27. 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, in *Proceedings ICFD Conference on Numerical Methods in Fluid Dynamics* (Oxford, Univ. Press, London, 1986), pp. 609–619.
28. 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* (Oxford Univ. Press, London, 1993), Vol. IV, pp. 221–239.
29. A. Quarteroni and A. Valli, *Numerical Approximation of Partial Differential Equations* (Springer-Verlag, Berlin/New York, 1993).
30. M. Schäfer and S. Turek (with support of F. Durst, E. Krause, R. Rannacher), Benchmark computations of laminar flow around cylinder, in *Flow Simulation with High-Performance Computers II*, Notes on Numerical Fluid Mechanics, edited by E. H. Hirschel (Vieweg, Weisbaden, 1996), pp. 547–566.
31. V. Selmin, Finite element solution of hyperbolic equations. I. One-dimensional case, *INRIA Research Report* **655** (1987).
32. V. Selmin, Finite element solution of hyperbolic equations. II. Two-dimensional case, *INRIA Research Report* **708** (1987).
33. T. W. H. Sheu and C. C. Fang, High resolution finite-element analysis of shallow water equations in two dimensions, *Comput. Meth. Appl. Mech. Eng.* **190**, 2581 (2001).
34. A. Sokolichin, *Mathematische Modellbildung und numerische Simulation von Gasflüssigkeits-Blasenströmungen* (Habilitationsschrift, Universität Stuttgart, 2001).
35. S. Turek, *Efficient Solvers for Incompressible Flow Problems: An Algorithmic and Computational Approach*, LNCSE 6 (Springer-Verlag, Berlin/New York, 1999).
36. S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, *J. Comput. Phys.* **31**, 335 (1979).