INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS
Int. J. Numer. Meth. Fluids 2000; **00**:1–6 *Prepared using* fldauth.cls *[Version: 2002/09/18 v1.01]*

On an efficient solution strategy of Newton type for implicit finite element schemes based on algebraic flux correction

M. Möller*

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

SUMMARY

A discrete Newton approach is applied to implicit flux limiting schemes based on the concept of *algebraic flux correction* (AFC). The Jacobian matrix is approximated by divided differences and assembled edge-by-edge. The use of a nodal flux limiter leads to an extended stencil which can be constructed *a priori*. Numerical examples for two-dimensional benchmark problems are presented to compare the performance of the algebraic Newton method to the defect correction approach. Copyright © 2000 John Wiley & Sons, Ltd.

KEY WORDS: algebraic Newton method; high-resolution schemes; flux-corrected transport

1. INTRODUCTION

Modern high-resolution schemes combine discretizations of high and low order in a nonlinear fashion so as to recover the high accuracy in regions of smooth solutions without generating nonphysical oscillations in the vicinity of discontinuities. In this paper, we adopt an algebraic approach to the design of flux correction schemes which consists of imposing mathematical constraints on discrete operators so as to achieve certain matrix properties. All conservative matrix manipulations are performed on the discrete level so that the application of Newton's method is a nontrivial task. We present an algebraic approach to the construction of the Jacobian matrix which makes use of an edge-based representation of the artificial nonlinearity.

2. ALGEBRAIC FLUX CORRECTION

Let the time-dependent conservation law $\frac{\partial u}{\partial t}+\nabla\cdot\mathbf{f}(u)=0$ for a scalar quantity u be discretized by the Galerkin finite element method which yields a DAE system for the vector of nodal values

$$M_C\frac{\mathrm{d}u}{\mathrm{d}t}=Ku \tag{1}$$

*Correspondence to: M. Möller, Institute of applied Mathematics (LS III), University of Dortmund, Vogelpothsweg 87, Dortmund D-44227, Germany, E-mail: matthias.moeller@math.uni-dortmund.de

Received 3 July 1999
Revised 18 September 2002

Here, $M_C = \{m_{ij}\}$ is the consistent mass matrix and $K = \{k_{ij}\}$ denotes the discrete transport operator. It is well known that the standard Galerkin method requires some stabilization in order to prevent the formation of spurious oscillations in the vicinity of steep gradients. In this paper we adopt an algebraic approach to render K *local extremum diminishing* (LED) by adding an artificial diffusion operator $D = \{d_{ij}\}$. It is designed to eliminate all negative off-diagonal entries from the convective operator K which would lead to nonphysical undershoots and overshoots otherwise. For (bi-)linear finite elements the optimal choice reads [1]

$$d_{ij} = \max\{-k_{ij}, 0, -k_{ji}\} = d_{ji} \tag{2}$$

so that the off-diagonal coefficients of the low-order operator $l_{ij} := k_{ij} + d_{ij}$ are nonnegative. Without loss of generality, we orient the edges of the sparsity graph so that $l_{ji} \geq l_{ij}$ for the edge ij which implies that node i is located 'upwind' and corresponds to the row number of the eliminated negative entry. Due to the fact that D is a discrete diffusion operator which is defined as a symmetric matrix with zero row and column sums the diagonal entry of the monotone low-order operator is given by $l_{ii} = k_{ii} - \sum_{j\neq i} d_{ij}$. In addition, row-sum mass lumping is performed so as to remove the antidiffusion from the consistent mass matrix which may violate the LED property so that the semi-discrete low-order scheme reads

$$M_L \frac{\mathrm{d}u}{\mathrm{d}t} = Lu, \qquad L = K + D \tag{3}$$

According to the Godunov theorem, linear monotonicity preserving schemes can be at most first order accurate. This order barrier can be overcome by blending the discretizations of high and low order in an adaptive fashion. The difference between the residual terms $f = (M_L - M_C)\frac{\mathrm{d}u}{\mathrm{d}t} + (K - L)u$ can be decomposed into sums of skew-symmetric internodal fluxes which are associated with the edges of the underlying sparsity graph

$$f_i = \sum_{j\neq i} f_{ij}, \qquad f_{ij} = \left[m_{ij}\frac{\mathrm{d}}{\mathrm{d}t} + d_{ij}\right](u_i - u_j) = -f_{ji} \tag{4}$$

Their limited counterparts are applied to the right-hand side of the monotone low-order discretization (3) so as to obtain the following nonlinear high-resolution scheme

$$M_L \frac{\mathrm{d}u}{\mathrm{d}t} = Lu + f^*(u), \qquad f_i^* = \sum_{j\neq i} \alpha_{ij} f_{ij}, \qquad \alpha_{ij} \in [0,1] \tag{5}$$

Setting all correction factors equal to zero, one obtains the low-order discretization (3) while the Galerkin scheme (1) is recovered if no limiting is performed. The task of the flux limiter is to find correction factors α_{ij} as close to 1 as possible without generating spurious oscillations. In the framework of algebraic flux correction [1, 2] the following algorithm is adopted:

1. Compute the sums of positive and negative antidiffusive fluxes

$$P_i^+ = \sum_{j\neq i} \max\{0, f_{ij}\}, \qquad P_i^- = \sum_{j\neq i} \min\{0, f_{ij}\} \tag{6}$$

2. Define the upper/lower bounds for a set of coefficients $q_{ij} \geq 0$

$$Q_i^+ = \sum_{j\neq i} q_{ij} \max\{0, u_j - u_i\}, \qquad Q_i^- = \sum_{j\neq i} q_{ij} \min\{0, u_j - u_i\} \tag{7}$$

3. Evaluate the correction factor for positive/negative fluxes

$$R_i^{\pm} = \min\left\{1, \frac{Q_i^{\pm}}{P_i^{\pm}}\right\}, \qquad \alpha_{ij} = \left\{ \begin{array}{ll} \min\{R_i^+, R_j^-\}, & \text{if } f_{ij} \geq 0 \\ \min\{R_i^-, R_j^+\}, & \text{if } f_{ij} < 0 \end{array} \right. \tag{8}$$

For symmetric limiters of FCT type, $q_{ij} = m_{ij}/\Delta t$ in the definition of upper and lower bounds while the correction factors α_{ij} are determined as the minimum of $R_i^{\pm}$ and $R_j^{\mp}$. For the computation of steady state flows, the use of an upwind-biased limiting strategy is advisable. In this case, $q_{ij} = d_{ij}$ and the sums in (6) only extend over the set of downwind neighbors $\mathcal{J}_i = \{j \neq i \,|\, 0 = l_{ij} < l_{ji}\}$. Moreover, the nodal correction factor of the upwind node is employed in the limiting procedure, that is, $\alpha_{ij} = R_i^{\pm}$ depending on the sign of the flux.

To simplify the presentation, the contribution of the mass matrix is neglected in what follows so that the raw flux given in (4) is redefined as $f_{ij} = d_{ij}(u_i - u_j) = -f_{ji}$.

3. NONLINEAR SOLUTION STRATEGIES

After an implicit time discretization, we end up with a nonlinear algebraic system of the form

$$F(u^{n+1}) := M_L \frac{u^{n+1} - u^n}{\Delta t} - \theta N(u^{n+1}) - (1-\theta)N(u^n) = 0, \qquad 0 < \theta \leq 1 \tag{9}$$

The operator $N(u) = Lu + f^*(u)$ is made up from the monotone transport operator which is nonlinear only if the governing equation is and the antidiffusive correction $f^*(u)$. The solution at time $t^{n+1} = t^n + \Delta t$ can be computed by the fixed-point iteration

$$u^{(m+1)} = u^{(m)} - C^{-1}F(u^{(m)}), \quad u^{(0)} = u^n, \quad m = 0, 1, \ldots \tag{10}$$

where C is a suitable 'preconditioner' to be defined below. The iteration process is terminated for sufficiently small solution increments and/or for a required reduction of the nonlinear residual. In a practical implementation, the 'inversion' of C is also performed by solving a sequence of linear subproblems for the solution increment which is applied to the last iterate

$$\begin{array}{rclr} C\Delta u^{(m+1)} & = & F(u^{(m)}) & m = 0, 1, \ldots \\ u^{(m+1)} & = & u^{(m)} + \Delta u^{(m+1)} & u^{(0)} = u^n \end{array} \tag{11}$$

For very small time steps, the lumped mass matrix yields a usable preconditioner $C = M_L/\Delta t$ which can be inverted directly. If larger time steps should be employed the low-order operator

$$C = M_L/\Delta t - \theta L \tag{12}$$

can be used to turn equation (10) into a fixed-point defect correction scheme. By construction, the preconditioner defined in (12) is an M-matrix and hence exhibits amenable matrix properties [1]. The linear problem (11) is solved by a Krylov subspace method which can be preconditioned by an incomplete LU factorization of the operator C. It is worth mentioning that for an M-matrix the ILU decomposition unconditionally exists and is unique.

Another attractive algorithm for the solution of the nonlinear algebraic system (9) is Newton's method which is recovered from (10) by defining the global preconditioner as follows

$$C = M_L/\Delta t - \theta T, \qquad T = \frac{\partial N(u)}{\partial u} \tag{13}$$

In each step, the Jacobian matrix $T = \{t_{ij}\}$ which corresponds to the nonlinear operator $N(u)$ needs to be evaluated making use of the solution $u^{(m)}$ from the last iteration.

4. APPROXIMATION OF JACOBIAN MATRIX

Recall that the operator $N(u)$ is constructed from the high-order transport operator K by adding artificial diffusion D and applying some portion of admissible antidiffusion afterwards. All conservative manipulations are performed on the algebraic level so that no continuous counterpart exists. Consequently, the Jacobian matrix can only be approximated by divided differences. To this end, let us introduce the operator $\mathcal{D}_k$ for a generic function $g(u)$

$$\mathcal{D}_k[g] := \frac{g(u+he_k)-g(u-he_k)}{2h} \tag{14}$$

where e_k denotes the kth unit vector. The perturbation parameter h should be sufficiently small to obtain a good approximation to the derivative, e.g., $h=\sqrt{\epsilon_{\text{mach}}}$. If the accuracy of the function g is known to be limited, $h=((1+\|u\|)\epsilon_{\text{mach}})^{1/3}$ is a reasonable choice.

By virtue of definition (14) each entry of the Jacobian T can be approximated as follows

$$t_{ik} = \mathcal{D}_k[N(u)_i] + \mathcal{O}(h^2) \tag{15}$$

Let us substitute the definition $N(u)=Ku+Du+f^*(u)$ in the above relation and decompose the net diffusive contribution $(Du)_i+f_i^*$ to each node into sums of internodal fluxes [3]

$$\mathcal{D}_k[N(u)_i] = \hat{k}_{ik} + \sum_j \mathcal{D}_k[k_{ij}]u_j + \sum_{j\neq i} \mathcal{D}_k[(1-\alpha_{ij})d_{ij}(u_j-u_i)] \tag{16}$$

Here, the average of the high-order coefficient results from the two-sided perturbation of u

$$\hat{K}=\{\hat{k}_{ij}\}, \qquad \hat{k}_{ij} = \frac{k_{ij}(u+he_k)+k_{ij}(u-he_k)}{2} \tag{17}$$

If the velocity field does not depend on the solution then the averaged transport operator $\hat{K}$ reduces to K whereas its derivative vanishes. Nonetheless, the approximate Jacobian of the (anti-)diffusive term $Du+f^*(u)$ needs to be evaluated since it contains some artificial nonlinearity engendered by the use of a solution dependent flux correction algorithm.

Without going into detail let us investigate the sparsity pattern of the resulting Jacobian matrix. To this end, let $\mathcal{G}=\langle K\rangle$ denote the connectivity graph of the standard finite element matrix and $\mathcal{Z}=\langle T\rangle$ represent that of the Jacobian for the nonlinear operator $N(u)$. It is easy to verify that $g_{ii}\neq 0, \forall i$ and $g_{ij}\neq 0, \forall i, \forall j\neq i$ if and only if there exists an edge ij. Obviously, the same structure carries over to the matrix T. In addition, the correction factors α_{ij} defined in (8) give rise to an extended coupling of nodes which are not directly connected by an edge. The situation is illustrated in Figure 1 where a small perturbation is applied to node k. Consequently, the nodal quantities $P_i^{\pm}$, $Q_i^{\pm}$ and $R_i^{\pm}$ need to be updated for k and for all of its direct neighbors i for which there exists an edge ik. Since the correction factors α_{ij} are defined as a combination of $R_i^{\pm}$ and $R_j^{\mp}$ the perturbation of the nodal solution value u_k may lead to a nonvanishing contribution $\mathcal{D}_k[(1-\alpha_{ij})d_{ij}(u_j-u_i)]$ to node i and also to j but with opposite sign. In other words, the connectivity graph of T needs to be extended such that $g_{jk}\neq 0$, $\forall j$, $\forall k\neq j$ if there exist two edges ij and ik which allow for an indirect interaction of nodes j and k via their common node neighbor i. The above derivation of the extended connectivity graph can be readily turned into an algorithm for generating $\mathcal{Z}=\mathcal{G}^2$ by means of a matrix-matrix multiplication and filtering the nonzero entries. Interestingly enough, the use of an edge-oriented FEM stabilization techniques yields a similar sparsity pattern of the system matrix so that construction procedures can be carried over.

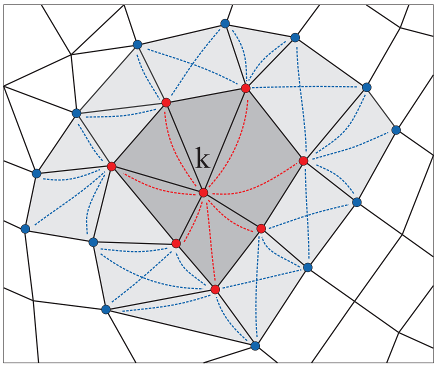


Figure 1. Stencil of the Jacobian matrix for column k.

5. NUMERICAL EXAMPLES

As a first test problem, consider the 2D convection-diffusion equation $\mathbf{v}\cdot\nabla u - d\Delta u = 0$ to be solved in the domain $\Omega = (0,1)\times(0,1)$, whereby the velocity field $\mathbf{v} = (\cos 10°, \sin 10°)$ is constant and $d = 10^{-3}$. The concomitant boundary conditions read

$$\begin{array}{lll} u(x,0)=0, & \dfrac{\partial u}{\partial y}(x,1)=0, & u(0,y)=\left\{\begin{array}{ll} 1 & \text{if } y\geq 0.5 \\ 0 & \text{otherwise}\end{array}\right. \\ u(1,y)=0, & & \end{array} \tag{18}$$

which lead to a sharp front next to the line $x = 1$. A common practice in the computation of steady-state problems is to adopt the fully implicit backward Euler method ($\theta = 1$) and perform pseudo-time stepping. This strategy can be interpreted as applying an individual under-relaxation factor to each nodal equation. The nonlinear algebraic equation (9) is solved repeatedly until $\|F(u)\| \leq 10^{-12}$, whereby only one nonlinear iteration is performed per pseudo-time step. The numerical solution depicted in Figure 2 (a) is computed by the FEM-TVD method on a uniform mesh of 128×128 bilinear elements. The resolution of the thin boundary layer is very crisp and moreover the profile is free of spurious oscillations.

The nonlinear convergence behavior of the algebraic Newton approach (13) and the standard defect correction scheme (12) is presented in Figure 2 (b) for three successively refined meshes making use of a moderate pseudo-time step $\Delta t = 1.0$. Interestingly enough, the convergence of the defect correction scheme improves if the mesh is refined. This may be attributed to the fact, that the nonlinearity is artificially introduced by the flux limiter so that the system becomes 'less nonlinear' if the mesh is refined and the perturbation by artificial (anti-)diffusion becomes smaller. Nonetheless, the algebraic Newton method outperforms the defect correction scheme by a factor of 4-16 with respect to the number of nonlinear iterations.

As a second example consider the 1D inviscid Burgers' equation $u\frac{\partial u}{\partial x} + \frac{\partial u}{\partial t} = 0$ solved in the space time domain $\Omega = (0,1)\times(0,0.5)$. This corresponds to computing the solution for all time levels simultaneously instead of doing it step by step. The staircase profile

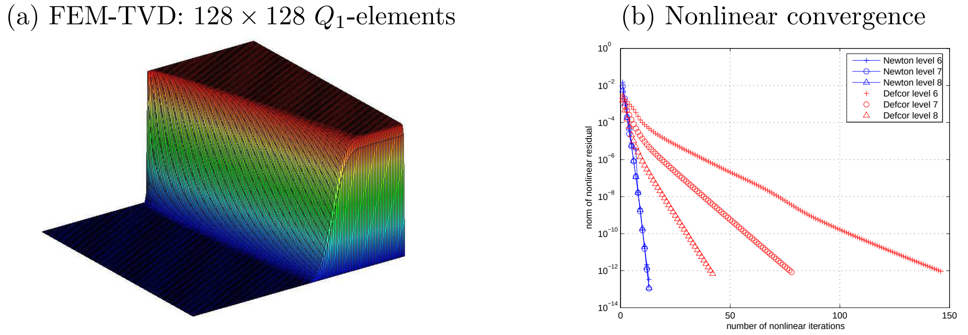
$$u(x,t) = \left\{\begin{array}{lll} 1 & \text{if} & 0\leq x<0.4 \wedge t=0, \\ 0.5 & \text{if} & 0.4\leq x\leq 0.8 \wedge t=0, \\ 0 & \text{if} & 0.8<x\leq 1 \wedge t=0 \quad \vee \quad x=0 \wedge 0\leq t\leq 0.5 \end{array}\right. \tag{19}$$

is prescribed at the ‘inlet’ of Ω which leads to a rarefaction wave and two shock waves. All flow features are captured accurately by the FEM-TVD method as depicted in Figure 3 (a).

For this simulation, the pseudo-time step is chosen adaptively from the interval $[0.01, 1.0]$ by means of a PID controller. Moreover, a maximum number of 10 nonlinear iterations is allowed in each time-step. The nonlinear convergence behavior of the defect correction scheme and the algebraic Newton approach is presented in Figure 3 (b). Newton’s method requires a constant number of 38-39 outer iterations, whereas more than 1000 cycles are necessary for the defect correction approach in order to converge on the finest grid. Throughout all computations, the CPU time required by the defect correction approach exceeds that of Newton’s method by a factor of 3. For both test problems, the linear systems are solved by means of a BiCGSTAB method making use of an ILU(0) decomposition of the low-order operator (12).

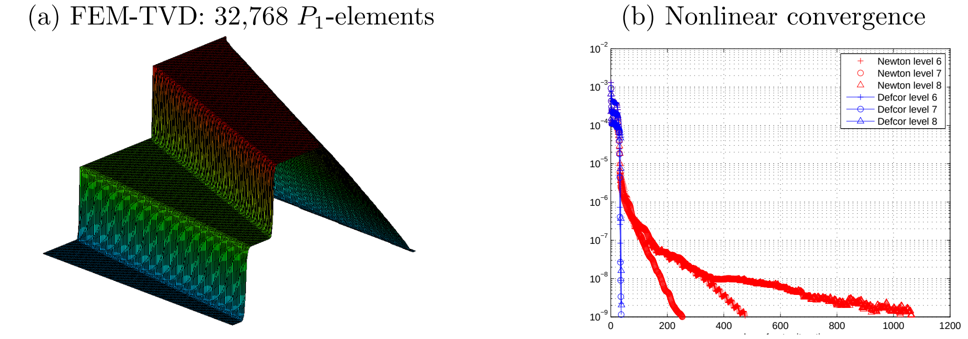
REFERENCES

1. Kuzmin D, Möller M. Algebraic flux correction I. Scalar conservation laws. In *Flux-Corrected Transport: Principles, Algorithms, and Applications*, Kuzmin D, Löhner R, Turek S (eds). Springer, 2005; 155-206.
2. Kuzmin D. On the design of general-purpose flux limiters for implicit FEM with a consistent mass matrix. *Journal of Computational Physics* 2006; **219**:513-531.
3. Möller M. Efficient solution techniques for implicit finite element schemes with flux limiters. *International Journal for Numerical Methods in Fluids* 2007; in press.


(a) FEM-TVD: 128×128 Q_1-elements

(b) Nonlinear convergence

Figure 2. Convection-diffusion in 2D.


(a) FEM-TVD: 32,768 P_1-elements

(b) Nonlinear convergence

Figure 3. Burgers’ equation in space-time.