On the design of higher-order FEM satisfying the discrete maximum principle

Dmitri Kuzmin

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

Abstract

A fully algebraic approach to the design of nonlinear high-resolution schemes is revisited and extended to quadratic finite elements. The matrices resulting from a standard Galerkin discretization are modified so as to satisfy sufficient conditions of the discrete maximum principle for nodal values. In order to provide mass conservation, the perturbation terms are assembled from skew-symmetric internodal fluxes which are redefined as a combination of first- and second-order divided differences. The new approach to the construction of artificial diffusion operators is combined with a node-oriented limiting strategy. The resulting algorithm is applied to P_1 and P_2 approximations of stationary convection-diffusion equations in 1D/2D.

Key words: finite elements, discrete maximum principle, M-matrix, flux correction

1 Introduction

Discrete maximum principles (DMP) play an important role in the analysis and design of finite element methods. Most of the proofs and algorithms are based on a well-known set of sufficient conditions which provide the *M-matrix* property and are easy to verify. Algebraic flux correction [5,6] makes it possible to enforce these conditions *a posteriori* by adding a suitably designed artificial diffusion operator to an *a priori* unstable Galerkin discretization. In the case of low-order finite elements, all the necessary information is inferred from the original stiffness matrix. However, the use of higher-order basis functions may entail a loss of accuracy/consistency if the DMP constraint is enforced in the usual way. In this paper, we focus on the peculiarities of P_2 elements and explain how to overcome some of the difficulties that arise from their use.

Email address: kuzmin@math.uni-dortmund.de (Dmitri Kuzmin).

2 Design criteria

A finite element discretization of the form $Au = b$ satisfies the discrete maximum principle for nodal values under the following constraints (see, e.g., [3])

(i) all diagonal coefficients of A are positive $\quad a_{ii} > 0, \quad \forall i, \qquad (1)$

(ii) there are no positive off-diagonal entries $\quad a_{ij} \leq 0, \quad \forall j \neq i, \qquad (2)$

(iii) A is strictly diagonally dominant $\quad \sum_j a_{ij} > 0, \quad \forall i. \qquad (3)$

These sufficient (but not necessary) conditions ensure that $A = \{a_{ij}\}$ is an M-matrix and has a nonnegative inverse. Inequalities (1)–(3) are easy to verify but may impose severe restrictions on the shape of finite elements and on the choice of basis functions. In particular, any discretization of convective/diffusive terms which satisfies the M-matrix criterion *a priori* is doomed to be first-/second-order accurate, respectively (see, e.g., [4], pp. 119-120).

In order to circumvent the above-mentioned order barriers, the matrix coefficients may need to be adjusted *a posteriori* so as to take the local solution behavior into account. This idea forms the basis for the development of algebraic flux correction schemes [5,6]. The first step is to construct an artificial diffusion operator $D = \{d_{ij}\}$ such that $\tilde{A} = A + D$ is an M-matrix and the vector Du can be decomposed into a sum of internodal fluxes [5]

$$(Du)_i = \sum_{j \neq i} f_{ij}. \qquad f_{ji} = -f_{ij}. \tag{4}$$

In the case of low-order finite elements, which produce stiffness matrices with a compact stencil, the minimum amount of artificial diffusion is given by

$$f_{ij} = d_{ij}(u_j - u_i), \qquad d_{ij} = -\max\{a_{ij}, 0, a_{ji}\} \tag{5}$$

but the design of artificial diffusion operators and of the corresponding fluxes for P_2 finite elements is more involved, as explained in the next section.

3 Artificial diffusion operators in 1D

In order to illustrate the differences between linear and quadratic FEM approximations, let us consider the one-dimensional convection-diffusion equation

$$v\frac{\mathrm{d}u}{\mathrm{d}x} - \varepsilon\frac{\mathrm{d}^2u}{\mathrm{d}x^2} = 0, \qquad v > 0, \quad \varepsilon > 0. \tag{6}$$

The standard Galerkin discretization of the convective and diffusive terms on a single P_1 element $e = (x_i, x_{i+1})$ of length Δx yields the element matrices

$$C|_e = \frac{v}{2}\begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}, \qquad S|_e = \frac{\varepsilon}{\Delta x}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}. \tag{7}$$

Note that the entries of the symmetric operator $S|_e$ have the right sign and do not need to be modified. However, the discrete transport operator $C|_e$ has a negative diagonal entry and a positive off-diagonal one, which can be rectified by adding a symmetric perturbation matrix with zero row and column sums

$$\tilde{C}|_e = C|_e + \frac{v}{2}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} = v\begin{bmatrix} 0 & 0 \\ -1 & 1 \end{bmatrix}, \qquad \tilde{S}|_e = S|_e. \tag{8}$$

After the global matrix assembly, the upwind difference scheme is recovered

$$v\ \frac{u_i - u_{i-1}}{\Delta x} - \varepsilon \frac{u_{i-1} - 2u_i + u_{i+1}}{(\Delta x)^2} = 0. \tag{9}$$

By virtue of (8), the replacement of $C|_e$ by $\tilde{C}|_e$ generates an error which can be decomposed into a sum of skew-symmetric numerical fluxes [5]

$$f_{i+1/2} = \frac{v}{2}(u_i - u_{i+1}) = -\frac{v\Delta x}{2}\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)_i + O(\Delta x^2), \tag{10}$$

where $f_{i+1/2}$ is the shorthand notation for $f_{ij},\ j = i+1$. The Taylor series expansion of u_{i+1} about u_i reveals that the numerical diffusion coefficient is proportional to Δx, which corresponds to a consistent first-order perturbation.

The element matrices for a P_2 discretization on $e = (x_{i-1}, x_{i+1})$ are given by

$$C|_e = \frac{v}{6}\begin{bmatrix} -3 & 4 & -1 \\ -4 & 0 & 4 \\ 1 & -4 & 3 \end{bmatrix}, \qquad S|_e = \frac{\varepsilon}{6\Delta x}\begin{bmatrix} 7 & -8 & 1 \\ -8 & 16 & -8 \\ 1 & -8 & 7 \end{bmatrix}. \tag{11}$$

In this case, even the operator $S|_e$ does not pass the M-matrix test and needs to be modified. However, artificial diffusion of the form (5) would introduce a zeroth-order perturbation error rendering the scheme inconsistent. This is why, it is appropriate to apply a discrete diffusion operator of 4th order

$$\tilde{S}|_e = S|_e + \frac{\varepsilon}{6\Delta x}\begin{bmatrix} -1 & 2 & -1 \\ 2 & -4 & 2 \\ -1 & 2 & -1 \end{bmatrix} = \frac{\varepsilon}{\Delta x}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 2 & -1 \\ 0 & -1 & 1 \end{bmatrix}. \tag{12}$$

Both farmost off-diagonal entries of $C|_e$ also need to be nullified, since their treatment in accordance with (5) would result in a significant loss of accuracy. Hence, it is worthwhile to reduce the stencil of the matrix as follows

$$\hat{C}|_e = C|_e + \frac{v}{6}\begin{bmatrix} 1 & -2 & 1 \\ 0 & 0 & 0 \\ -1 & 2 & -1 \end{bmatrix} = \frac{v}{3}\begin{bmatrix} -1 & 1 & 0 \\ -2 & 0 & 2 \\ 0 & -1 & 1 \end{bmatrix}. \tag{13}$$

Due to the fact that $C|_e$ is skew-symmetric, this modification affects only the elements of the first and last row, while the middle row remains unchanged.

The remaining positive off-diagonal entries are eliminated in the usual way using $d_{12} = \frac{2}{3}v = d_{23}$ based on the coefficients of the original operator

$$\tilde{C}|_e = \hat{C}|_e + d_{12}\begin{bmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix} + d_{23}\begin{bmatrix} 0 & 0 & 0 \\ 0 & 1 & -1 \\ 0 & -1 & 1 \end{bmatrix} = \frac{v}{3}\begin{bmatrix} 1 & -1 & 0 \\ -4 & 4 & 0 \\ 0 & -3 & 3 \end{bmatrix}. \tag{14}$$

In fact, it is tempting to take $d_{12} = \frac{v}{3}$ which would produce a lower triangular matrix with $\tilde{c}_{12} = 0$. However, using $d_{12} \neq d_{23}$ would result in a first-order perturbation of the middle row, whereas the above definition corresponds to a second-order perturbation and yields a consistent first-order scheme. Indeed, after the matrix assembly and division by the diagonal entries of the (row-sum) lumped mass matrix $M_L = \mathrm{diag}\{\ldots, \frac{2\Delta x}{3}, \frac{4\Delta x}{3}, \frac{2\Delta x}{3}, \ldots\}$ we obtain

$$\left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)_{i-1} \approx \frac{-3u_{i-2} + 4u_{i-1} - u_i}{2\Delta x}, \qquad \left(\frac{\mathrm{d}u}{\mathrm{d}x}\right)_{i} \approx \frac{u_i - u_{i-1}}{\Delta x}, \quad \ldots \tag{15}$$

By construction, the fluxes into the midpoint node x_i can be expressed in terms of first- and second-order divided differences

$$f_{i-1/2} = \frac{2}{3}v(u_i - u_{i-1}) - \frac{1}{6}\left[v - \frac{2\varepsilon}{\Delta x}\right](u_{i-1} - 2u_i + u_{i+1}) = O(\Delta x), \tag{16}$$

$$f_{i+1/2} = \frac{2}{3}v(u_i - u_{i+1}) + \frac{1}{6}\left[v + \frac{2\varepsilon}{\Delta x}\right](u_{i-1} - 2u_i + u_{i+1}) = O(\Delta x), \tag{17}$$

which follows from (12)–(14). The so-defined fluxes $f_{i\pm1/2}$ preserve consistency even though the coefficients of $S|_e$ are inversely proportional to Δx.

4 Artificial diffusion operators in 2D

In the 2D case, the coefficients of the discrete convection and diffusion operators C and S are given by $c_{ij} = \int_\Omega \varphi_i(\mathbf{v}\cdot\nabla\varphi_j)\,\mathrm{d}\mathbf{x}$ and $s_{ij} = \varepsilon\int_\Omega \nabla\varphi_i\cdot\nabla\varphi_j\,\mathrm{d}\mathbf{x}$. As a rule, the element matrix $C|_e$ must be generated using numerical integration but $S|_e$ can be obtained in analytical form. For the six-node triangle depicted in Fig. 1, the entries of $S|_e$ depend on the cotangents α, β and γ of interior angles opposite the edges $\mathbf{x}_3\mathbf{x}_5$, $\mathbf{x}_5\mathbf{x}_1$ and $\mathbf{x}_1\mathbf{x}_3$, respectively [2]

$$S|_e = \varepsilon\begin{bmatrix} 3(\beta+\gamma) & -4\gamma & \gamma & 0 & \beta & -4\beta \\ -4\gamma & 8\sigma & -4\gamma & -8\beta & 0 & -8\alpha \\ \gamma & -4\gamma & 3(\gamma+\alpha) & -4\alpha & \alpha & 0 \\ 0 & -8\beta & -4\alpha & 8\sigma & -4\alpha & -8\gamma \\ \beta & 0 & \alpha & -4\alpha & 3(\alpha+\beta) & -4\beta \\ -4\beta & -8\alpha & 0 & -8\gamma & -4\beta & 8\sigma \end{bmatrix}, \quad (18)$$

where $\sigma = \alpha + \beta + \gamma$. Hence, the off-diagonal coefficients are of variable sign.

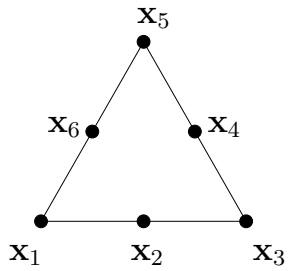


Fig. 1. Degrees of freedom for a six-node triangle.

In order to enforce the M-matrix property and reduce the stencil of the stiffness matrix following the strategy developed in 1D, all off-diagonal coefficients which correspond to vertex-vertex connections should be eliminated regardless of their sign. In a practical implementation, the global stiffness matrix S or the element matrices $S|_e$ are modified edge-by-edge. If $\mathbf{x}_k$ is the midpoint of the edge $\mathbf{x}_i\mathbf{x}_j$, then the coefficients s_{ij} and s_{ji} can be nullified as follows

$$\begin{aligned} &s_{ii} := s_{ii} + d_{ij}, \quad && s_{ik} := s_{ik} - 2d_{ij}, \quad && s_{ij} := s_{ij} + d_{ij}, \\ &s_{ki} := s_{ki} - 2d_{ij}, \quad && s_{kk} := s_{kk} + 4d_{ij}, \quad && s_{kj} := s_{kj} - 2d_{ij}, \\ &s_{ji} := s_{ji} + d_{ij}, \quad && s_{jk} := s_{jk} - 2d_{ij}, \quad && s_{jj} := s_{jj} + d_{ij}, \end{aligned} \quad (19)$$

where $d_{ij} = -s_{ij}$, cf. equation (12). The corresponding fluxes are proportional to a second-order divided difference along the edge

$$-f_{ki} = f_{ik} = d_{ij}(u_i - 2u_k + u_j) = f_{jk} = -f_{kj}. \quad (20)$$

It is instructive to consider the reduced-stencil counterpart of $S|_e$ which reads

$$\tilde{S}|_e = 2\varepsilon \begin{bmatrix} \beta+\gamma & -\gamma & 0 & 0 & 0 & -\beta \\ -\gamma & 4\sigma-2\gamma & -\gamma & -4\beta & 0 & -4\alpha \\ 0 & -\gamma & \gamma+\alpha & -\alpha & 0 & 0 \\ 0 & -4\beta & -\alpha & 4\sigma-2\alpha & -\alpha & -4\gamma \\ 0 & 0 & 0 & -\alpha & \alpha+\beta & -\beta \\ -\beta & -4\alpha & 0 & -4\gamma & -\beta & 4\sigma-2\beta \end{bmatrix}. \tag{21}$$

If all angles are less than or equal to $\pi/2$, i.e., if the triangulation is of *nonobtuse* type [3], then $\alpha > 0,\ \beta > 0,\ \gamma > 0$. Hence, the remaining off-diagonal entries are nonpositive and no further modifications of $\tilde{S}|_e$ are required.

As in the 1D case, the stencil of the discrete transport operator C also needs to be reduced. The pair of matrix entries c_{ij} and c_{ji} associated with vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ can be eliminated using the artificial diffusion coefficients

$$d_{ij} = -\frac{c_{ij}+c_{ji}}{2} = d_{ji}, \qquad d'_{ij} = -\frac{c_{ij}-c_{ji}}{2} = -d'_{ji} \tag{22}$$

which represent the symmetric and skew-symmetric part of C, respectively. It is worth mentioning that d_{ij} is usually very small or zero, unless the velocity field is strongly nonuniform and/or at least one of the nodes is located on the boundary. The symmetric part is removed using the matrix update

$$\begin{aligned} c_{ii} &:= c_{ii} - d_{ij}, \quad & c_{ij} &:= c_{ij} + d_{ij}, \\ c_{ji} &:= c_{ji} + d_{ij}, \quad & c_{jj} &:= c_{jj} - d_{ij}, \end{aligned} \tag{23}$$

while the skew-symmetric one calls for the second-order perturbation

$$\begin{aligned} c_{ii} &:= c_{ii} + d'_{ij}, \quad & c_{ik} &:= c_{ik} - 2d'_{ij}, \quad & c_{ij} &:= c_{ij} + d'_{ij}, \\ c_{ji} &:= c_{ji} - d'_{ij}, \quad & c_{jk} &:= c_{jk} + 2d'_{ij}, \quad & c_{jj} &:= c_{jj} - d'_{ij}. \end{aligned} \tag{24}$$

The net internodal flux associated with these modifications is given by

$$f_{ij} = d_{ij}(u_j - u_i) + d'_{ij}(u_i - 2u_k + u_j) = -f_{ji} \tag{25}$$

or, equivalently, by the pair of fluxes $f_{ki} = -f_{ij}$ and $f_{kj} = f_{ij}$ into node k.

Furthermore, the matrix entries associated with an edge midpoint $\mathbf{x}_i$ and a vertex $\mathbf{x}_j$ located opposite the edge are also nonvanishing. The symmetric part (if any) can be eliminated via (22)–(23) and the skew-symmetric one

using fluxes of the form $f_{ij} = d'_{ij}(u_i - u_k - u_l + u_j) = -f_{ji}$, where u_k and u_l are the solution values at the midpoints of adjacent edges. Thus, we set

$$\begin{aligned} &c_{ii} := c_{ii} + d'_{ij}, \quad c_{ik} := c_{ik} - d'_{ij}, \quad c_{il} := c_{il} - d'_{ij}, \quad c_{ij} := c_{ij} + d'_{ij}, \\ &c_{ji} := c_{ji} - d'_{ij}, \quad c_{jk} := c_{jk} + d'_{ij}, \quad c_{jl} := c_{jl} + d'_{ij}, \quad c_{jj} := c_{jj} - d'_{ij}. \end{aligned} \tag{26}$$

Finally, the remaining positive off-diagonal entries of C are eliminated using symmetric perturbations of the form (23). The artificial diffusion coefficient

$$d_{ij} = -\max\{c_{ij}, 0, c_{ji}\} = d_{ji} \tag{27}$$

is to be employed for all midpoint pairs. On the other hand, if $\mathbf{x}_k$ is a midpoint located between the vertices $\mathbf{x}_i$ and $\mathbf{x}_j$, then it is advisable to synchronize the artificial diffusion coefficients d_{ik} and d_{ki} as in the 1D case. Therefore, let

$$d_{ik} = -\max\{c_{ik}, c_{ki}, 0, c_{kj}, c_{jk}\} = d_{jk} \tag{28}$$

which results in a second-order perturbation of row k.

5 Limited antidiffusive correction

The linear M-matrix $\tilde{A} = A + D$ constructed as explained in the previous section gives rise to a first-order discretization error, so that the use of quadratic basis functions does not pay off. In order to recover the high accuracy of the underlying Galerkin scheme, excessive artificial diffusion needs to be removed. By construction, the perturbed and unperturbed matrices satisfy the relation

$$Au = b \quad \Leftrightarrow \quad \tilde{A}u = b + Du, \qquad (Du)_i = \sum_{j \neq i} f_{ij}. \tag{29}$$

Thus, the artificial diffusion built into $\tilde{A}$ can be canceled by adding the sums of raw antidiffusive fluxes f_{ij} to the corresponding rows of b. Some fluxes are harmless but others may need to be limited in order to suppress spurious undershoots and overshoots. To this end, each flux is multiplied by a correction factor α_{ij} and inserted into the right-hand side of the perturbed system

$$\tilde{A}u = \tilde{b}, \qquad \tilde{b}_i = b_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \qquad 0 \leq \alpha_{ij} \leq 1. \tag{30}$$

Setting all correction factors equal to zero, we obtain the low-order scheme, while the original Galerkin scheme is recovered if no limiting is performed.

The discrete maximum principle for nodal values is satisfied if the sum of limited antidiffusive fluxes can be represented in the form

$$\sum_{j\neq i}\alpha_{ij}f_{ij}=\sum_{j\neq i}q_{ij}(u)(u_j-u_i),\quad q_{ij}\geq 0,\quad \forall j\neq i. \tag{31}$$

This constraint can be enforced by tuning the correction factors α_{ij} which should be as close to 1 as possible for accuracy reasons. The resulting nonlinear algebraic system can be solved, e.g., using iterative defect correction [5].

The limiting strategy developed in [5,6] is based on the following generic algorithm, whereby positive and negative antidiffusive fluxes are treated separately

1. Compute the sums of positive and negative antidiffusive fluxes

$$P_i^+=\sum_{j\neq i}\max\{0,f_{ij}\},\qquad P_i^-=\sum_{j\neq i}\min\{0,f_{ij}\}. \tag{32}$$

2. Pick a set of coefficients $q_{ij}\geq 0$ and define the upper/lower bounds

$$Q_i^+=\sum_{j\neq i}q_{ij}\max\{0,u_j-u_i\},\qquad Q_i^-=\sum_{j\neq i}q_{ij}\min\{0,u_j-u_i\}. \tag{33}$$

3. Evaluate the *nodal correction factors* for positive/negative fluxes

$$R_i^+=\min\{1,Q_i^+/P_i^+\},\qquad R_i^-=\min\{1,Q_i^-/P_i^-\}. \tag{34}$$

4. Multiply the raw antidiffusive flux f_{ij} by the minimum of $R_i^\pm$ and $R_j^\mp$

$$f_{ij}:=\alpha_{ij}f_{ij},\qquad \alpha_{ij}=\begin{cases}\min\{R_i^+,R_j^-\}, & \text{if } f_{ij}>0,\\ \min\{R_i^-,R_j^+\}, & \text{otherwise.}\end{cases} \tag{35}$$

The above definition of α_{ij} corresponds to a symmetric flux limiter which is used by default. Upwind-biased limiters of TVD type [5,6] take advantage of the fact that one of the off-diagonal coefficients, say $\tilde{a}_{ji}<\tilde{a}_{ij}\leq 0$, is strictly negative and preserves its sign as long as the magnitude of f_{ij} is bounded by that of $\tilde{a}_{ji}(u_j-u_i)$. In order to enforce this condition, the original flux f_{ij} may need to be replaced by $f_{ij}:=\text{minmod}(f_{ij},\tilde{a}_{ji}(u_j-u_i))$, where

$$\text{minmod}(a,b)=\begin{cases}\min\{a,b\}, & \text{if } a>0,\ b>0,\\ \max\{a,b\}, & \text{if } a<0,\ b<0,\\ 0, & \text{otherwise.}\end{cases} \tag{36}$$

This manipulation renders the flux $f_{ji}:=-f_{ij}$ harmless, so that it suffices to increment the sum $P_i^\pm$ and apply the correction factor $\alpha_{ij}=R_i^\pm$. The upwind-

biased limiting strategy is appropriate for fluxes of the form $f_{ij} = d_{ij}(u_i - u_j)$ which are responsible for mass exchange between nearest neighbors.

The part of the artificial diffusion operator which is associated with stencil reduction admits a nonunique flux decomposition. On the one hand, fluxes of the form (25) can be redirected into an edge midpoint by setting

$$f_{ki} := f_{ki} - f_{ij}, \qquad f_{kj} := f_{kj} + f_{ij}, \qquad f_{ij} := 0. \tag{37}$$

On the other hand, they can be treated as such and limited separately by algorithm (32)–(35) based on another set of $P_i^{\pm}$, $Q_i^{\pm}$ and $R_i^{\pm}$. In this case, the symmetric limiting strategy is in order, since both off-diagonal coefficients of the low-order operator are equal to zero by construction. It is worth mentioning that fluxes given by (20) with $d_{ij} < 0$ preserve the sign of $\tilde{a}_{ki}$ and $\tilde{a}_{kj}$. Hence, they should be limited using the nodal correction factors $R_i^{\pm}$ and/or $R_j^{\pm}$.

The upper/lower bounds $Q_i^{\pm}$ are assembled from edge contributions of the form $q_{ij}(u_j - u_i)$ using the unperturbed coefficients of C and S to define

$$q_{ij} := \max\{c_{ij}, 0, c_{ji}\} + |s_{ij}|. \tag{38}$$

If nodes i and j are not nearest neighbors, then $q_{ij}(u_j - u_i)$ should be omitted or replaced by $2q_{ij}(u_m - u_i)$, where $u_m = u_k$ for vertex-vertex connections and $u_m = \frac{u_k + u_l}{2}$ for vertex-midpoint connections. This definition is motivated by the fact that $2(u_m - u_i) = u_j - u_i = 2(u_j - u_m)$ for a linear function.

6 Numerical examples

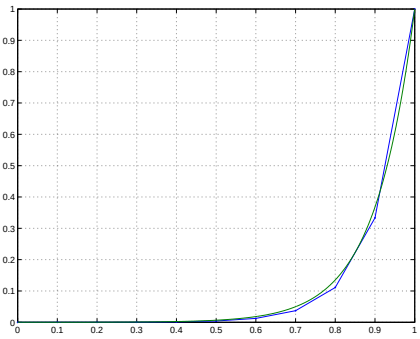
As a standard test problem, consider the one-dimensional convection-diffusion equation (6) to be solved in the domain $\Omega = (0, 1)$ subject to the boundary conditions $u(0) = 0$ and $u(1) = 1$. Fig. 2 displays the analytical solution

$$u(x) = \frac{e^{Pe\,x} - 1}{e^{Pe} - 1}, \quad \text{where} \quad Pe = \frac{v}{\varepsilon} \tag{39}$$

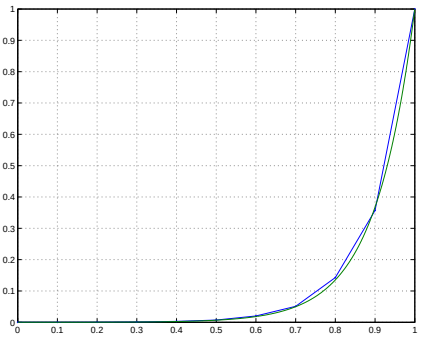
as compared to the piecewise-linear interpolants of nodal values computed using algebraic flux correction for P_1 and P_2 finite elements. In the latter case, equations (16)–(17) were used to define the raw antidiffusive fluxes.

If the Peclet number Pe is relatively small, then the analytical solution is smooth and the correction factors returned by the flux limiter approach 1. Hence, the accuracy of the underlying high-order discretization is decisive.

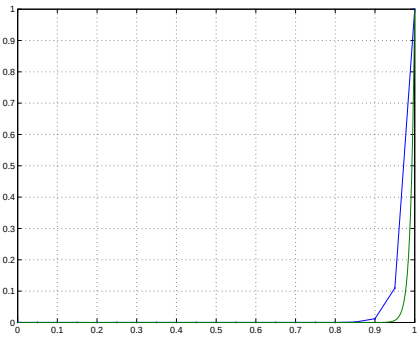
(a) Pe = 10, ten P_1 elements



(b) Pe = 10, five P_2 elements



(c) Pe = 100, twenty P_1 elements



(d) Pe = 100, ten P_2 elements

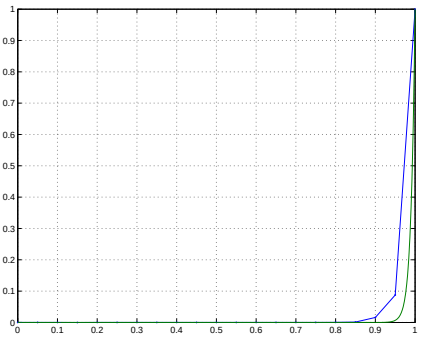


Fig. 2. Convection-diffusion in 1D: linear vs. quadratic elements.

The numerical solution for $Pe = 10$ produced by the P_1 version on a uniform mesh of 10 elements (Fig. 2a) is slightly underdiffusive, whereas the nodal values computed using 5 quadratic elements are almost exact, see Fig. 2b.

At large Peclet numbers, the standard Galerkin method tends to produce spurious oscillations, whereas flux-limited solutions are uniformly bounded by the boundary values 0 and 1, as required by the discrete maximum principle. Fig. 2c and Fig. 2d reveal that the approximations computed using twenty P_1 and ten P_2 elements for $Pe = 100$ are of comparable quality. As a rule of thumb, the use of higher-order basis functions pays off only for smooth data.

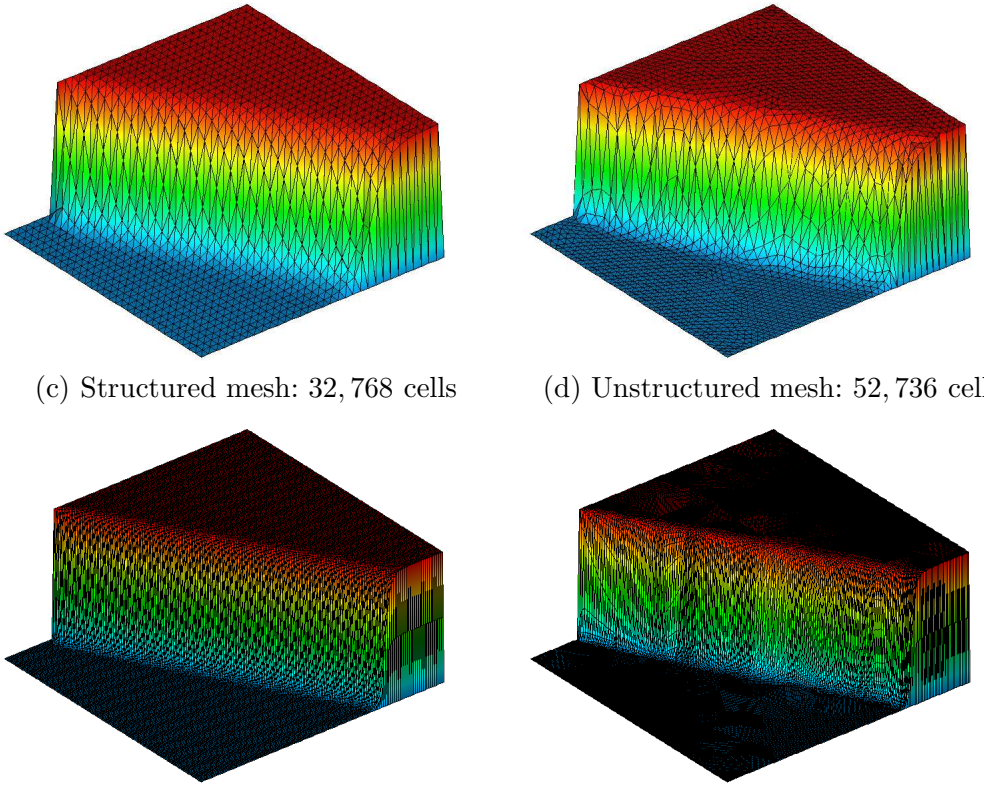
The second example illustrates the performance of the new algorithm as applied to the P_2 discretization of the 2D convection-diffusion equation

$$\mathbf{v} \cdot \nabla u - \varepsilon \Delta u = 0 \quad \text{in} \quad \Omega = (0,1) \times (0,1), \tag{40}$$

where $\mathbf{v} = (\cos 30°, \sin 30°)$ and $\varepsilon = 10^{-8}$. The boundary conditions read

$$u(x,0) = 0, \qquad \frac{\partial u}{\partial y}(x,1) = 0, \qquad u(0,y) = \begin{cases} 1 & \text{if } y \geq 0.2, \\ 0 & \text{otherwise.} \end{cases} \tag{41}$$
$$u(1,y) = 0,$$

(a) Structured mesh: 2,048 cells (b) Unstructured mesh: 3,296 cells



(c) Structured mesh: 32,768 cells (d) Unstructured mesh: 52,736 cells

Fig. 3. Convection-diffusion in 2D: flux-limited P_2 solution.

The standard Galerkin method would produce nonphysical oscillations in the vicinity of the line $x = 1$, where the Dirichlet boundary condition makes the solution gradient very steep. Fig. 3 displays the linear interpolants of the flux-limited solutions computed using P_2 elements on 4 different meshes. Two of them are structured, whereas the other two are unstructured. In either case, the finer mesh (Fig. 3c-d) is constructed from the coarser one (Fig. 3a-b) using two quad-tree refinements. The discrete maximum principle for nodal values is satisfied regardless of the size, shape and orientation of mesh elements.

7 Conclusions

This paper was intended to address some frequently asked questions regarding the applicability of algebraic flux correction to quadratic finite elements. The shortcomings of straightforward extensions were exposed and a new approach to the design of artificial diffusion operators was introduced. The corresponding internodal fluxes were defined in terms of first- and second-order divided differences, and a multidimensional limiting strategy was outlined. The proposed algorithm was applied to stationary convection-diffusion equations in

1D and 2D. The resulting solutions were shown to be nonoscillatory and, in some cases, superior to those computed on the basis of a P_1 discretization. On the other hand, the involved programming effort and the overhead cost are quite significant, whereas the improvements are often marginal, if any. In fact, flux limiters of TVD type tend to be slightly overdiffusive for P_1 and P_2 elements alike, so that the higher accuracy of quadratic FEM approximations cannot be recovered to the full extent. Hence, the overall performance depends strongly on the quality of the limiter and on the nonunique flux decomposition.

An extension of the above methodology to time-dependent problems is complicated by the fact that row-sum lumping (the only lumping technique that does conserve mass [1]) stores the whole mass at edge midpoints, whereas there is no mass associated with vertices. In fact, the lumped-mass Galerkin discretization can be interpreted as a finite volume scheme with edge-centered degrees of freedom [7]. Therefore, it is necessary to distinguish between vertex-vertex, vertex-midpoint and midpoint-midpoint interactions which should be treated differently. As of this writing, an optimal way to define and limit the corresponding internodal fluxes is yet to be found. In summary, algebraic flux correction for quadratic finite elements seems to be feasible but gives rise to many challenging open problems. It is hoped that this paper sheds some light on the difficulties to be dealt with and introduces some useful tools.

References

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

[2] W. Höhn and H.-D. Mittelmann, Some remarks on the discrete maximum-principle for finite elements of higher order. *Computing* **27** (1981) 145–154.

[3] I. Farago, R. Horvath and S. Korotov, Discrete maximum principle for linear parabolic problems solved on hybrid meshes. *Appl. Numer. Math.* **53** (2005), no. 2-4, 249–264.

[4] W. Hundsdorfer and J.G.Verwer, *Numerical Solution of Time-Dependent Advection-Diffusion-Reaction Equations*. Springer, 2003.

[5] D. Kuzmin and M. Möller, Algebraic flux correction I. Scalar conservation laws. In: D. Kuzmin, R. Löhner and S. Turek (eds.) *Flux-Corrected Transport: Principles, Algorithms, and Applications*. Springer, 2005, 155-206.

[6] D. Kuzmin, On the design of general-purpose flux limiters for implicit FEM with a consistent mass matrix. *J. Comput. Phys.* **219** (2006) 513-531.

[7] L. Postma and J.-M. Hervouet, Compatibility between finite volumes and finite elements using solutions of shallow water equations for substance transport. *Int. J. Numer. Meth. Fluids* **53** (2007) 1495-1507.