NUMERICAL SOLUTION OF CONVECTION–DIFFUSION EQUATIONS USING UPWINDING TECHNIQUES SATISFYING THE DISCRETE MAXIMUM PRINCIPLE

PETR KNOBLOCH

Abstract. We discuss the application of the finite element method to the numerical solution of scalar two-dimensional steady convection–diffusion equations with the emphasis on upwinding techniques satisfying the discrete maximum principle. Numerical experiments in convection–dominated case indicate that the improved Mizukami–Hughes method is the best choice for solving the mentioned class of problems using conforming piecewise linear triangular finite elements.

Key words. Convection–diffusion equation, Stabilized FEM, SUPG method, DCCD method, Upwinding, Mizukami–Hughes method, Discrete maximum principle, Numerical experiments

AMS subject classifications. 65N30

1. Introduction. This paper is devoted to the application of the finite element method to the numerical solution of the convection–diffusion equation

\[-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u = f \quad \text{in } \Omega, \quad u = u_b \quad \text{on } \partial \Omega,\]

where \(\Omega\) is a bounded two-dimensional domain with a polygonal boundary \(\partial \Omega\), \(f\) is a given outer source of the unknown scalar quantity \(u\), \(\varepsilon > 0\) is the constant diffusivity, \(\mathbf{b}\) is the flow velocity and \(u_b\) is a given function. The Dirichlet boundary condition is considered only for brevity.

We are interested in the strongly convection–dominated case \(\varepsilon \ll |\mathbf{b}|\) in which the solution of (1.1) typically contains narrow inner and boundary layers. It is well known that the application of the classical Galerkin finite element method is inappropriate in this case since the discrete solution is usually globally polluted by spurious oscillations. Therefore, various stabilization strategies have been developed during the last three decades, see e.g. the recent review paper by John and Knobloch [5]. In the present paper, we shall describe some of these methods and compare them numerically. Particularly, we shall concentrate on upwinding techniques satisfying the discrete maximum principle. Although these techniques were proposed mainly at the beginning of the development of stabilized methods, some of them can still compete with best methods proposed in the recent time when piecewise linear finite elements are used. Let us mention that the discrete maximum principle ensures that no spurious oscillations will appear in the discrete solution, not even in the vicinity of sharp layers. Moreover, it enables to prove uniform convergence results.

The plan of the paper is as follows. First, in the next section, we discuss various approaches to the numerical solution of problem (1.1). Then, in Sections 3–6 we describe several upwinding techniques: the method of Tabata in Section 3, methods based on dual meshes in Section 4, the Mizukami–Hughes method in Section 5 and the
improved Mizukami–Hughes method in Section 6. Finally, in Section 7, we present results of our numerical tests.

2. Finite element discretization of problem (1.1). Let $T_h$ be a triangulation of $\Omega$ consisting of a finite number of open triangular elements $K$. The discretization parameter $h$ in the notation $T_h$ is a positive real number satisfying $\text{diam}(K) \leq h$ for any $K \in T_h$. We assume that $\overline{\Omega} = \bigcup_{K \in T_h} K$ and that the closures of any two different elements of $T_h$ are either disjoint or possess either a common vertex or a common edge. Finally, we assume that the triangulation $T_h$ is of weakly acute type, i.e., the magnitude of all angles of elements $K \in T_h$ is less than or equal to $\pi/2$. This assumption is needed for proving the discrete maximum principle.

The solution $u$ of (1.1) will be approximated by a continuous piecewise linear function $u_h$ from the space

$$V_h = \{ v \in C(\overline{\Omega}) ; \ v|_K \in P_1(K) \quad \forall \ K \in T_h \}.$$ 

Let $a_1, \ldots, a_{M_h}$ be the vertices of $T_h$ lying in $\Omega$ and let $a_{M_h+1}, \ldots, a_{N_h}$ be the vertices of $T_h$ lying on $\partial \Omega$. For any $i \in \{1, \ldots, N_h\}$, let $\varphi_i \in V_h$ be the function satisfying $\varphi_i(a_j) = \delta_{ij}$ for $j = 1, \ldots, N_h$, where $\delta_{ij}$ is the Kronecker symbol. Then $V_h = \text{span}\{\varphi_i\}_{i=1}^{N_h}$.

The solution $u_h$ of the classical Galerkin finite element discretization of problem (1.1) is defined by

\begin{align*}
& (1.1) & u_h \in V_h, \\
& (2.2) & u_h(a_i) = u(a_i), \quad i = M_h + 1, \ldots, N_h, \\
& (2.3) & \varepsilon (\nabla u_h, \nabla \varphi_i) + (b \cdot \nabla u_h, \varphi_i) = (f, \varphi_i), \quad i = 1, \ldots, M_h, 
\end{align*}

where $(\cdot, \cdot)$ is the inner product in $L^2(\Omega)$ or $L^2(\Omega)^2$.

As we already mentioned, the Galerkin discretization is inappropriate if convection dominates diffusion. In this case, Petrov–Galerkin methods are often used, which consist in replacing (2.3) by

\begin{align*}
& (2.4) & \varepsilon (\nabla u_h, \nabla \varphi_i) + (b \cdot \nabla u_h, \tilde{\varphi}_i) = (f, \tilde{\varphi}_i), \quad i = 1, \ldots, M_h, 
\end{align*}

with some suitable weighting functions $\tilde{\varphi}_i$. One of the most efficient methods of this type is the streamline upwind/Petrov–Galerkin (SUPG) method proposed by Brooks and Hughes [2] where $\tilde{\varphi}_i = \varphi_i + \tau b \cdot \nabla \varphi_i$ and $\tau$ is a stabilization parameter. Denoting by $h_K$ the diameter of $K \in T_h$ in the direction of $b$, we set

$$\tau|_K = \frac{h_K}{2|b|} \left( \coth Pe_K - \frac{1}{Pe_K} \right) \quad \text{with} \quad Pe_K = \frac{|b| h_K}{2 \varepsilon},$$

where $Pe_K$ is the local Péclet number. Although the SUPG method produces to a great extent accurate and oscillation–free solutions, it does not preclude small non-physical oscillations localized in narrow regions along sharp layers. Since these oscillations are not permissible in many applications, various terms introducing artificial crosswind diffusion in the neighborhood of layers have been proposed to be added to the SUPG formulation in order to obtain a method which is monotone or which at least reduces the local oscillations, see e.g. the recent review paper by John and Knobloch [5]. This procedure is usually referred to as discontinuity capturing (or shock capturing).
One of the best discontinuity-capturing methods is the discontinuity-capturing crosswind-dissipation (DCCD) method by Codina [3, 8] defined by adding the term

$$(\sigma D \nabla u_h, \nabla \varphi_i) \quad \text{with} \quad D = \begin{cases} I - \frac{b \otimes b}{|b|^2} & \text{if } b \neq 0, \\ 0 & \text{if } b = 0 \end{cases}$$

to the left-hand side of the SUPG discretization (2.4). Here, $\sigma$ is defined by

$$\sigma_K = \frac{1}{2} \max \left\{ 0, \beta - \frac{2 \varepsilon |u_h|_{1,K}}{\|R_h(u_h)\|_{0,K} \text{diam}(K)} \right\} \text{diam}(K) \frac{\|R_h(u_h)\|_{0,K}}{|u_h|_{1,K}} \quad \forall K \in T_h,$$

where $\beta$ is a constant, $R_h(u_h) = b \cdot \nabla u_h - f$ is the residual, $\| \cdot \|_{0,K}$ is the $L^2(K)$ norm and $\| \cdot \|_{1,K}$ is the $H^1(K)$ seminorm. Codina [3] recommends to set $\beta \approx 0.7$. Note that the DCCD method is nonlinear since $\sigma$ depends on the unknown discrete solution $u_h$.

Most of the stabilization methods mentioned in this section involve stabilization parameters but it is not clear how to choose these parameters in an optimal way. From this point of view, upwinding techniques are very attractive since they do not involve any stabilization parameters.

3. **Tabata’s upwind method.** One of the first upwind finite element methods for the numerical solution of problem (1.1) was proposed by Tabata [10]. The basic idea is to assign an upwind element $K_{ij}^{\text{upwind}}$ to each vertex $a_i$ of the triangulation lying in $\Omega$. The upwind element $K_{ij}^{\text{upwind}}$ is any element possessing the vertex $a_i$ such that $-b(a_i)$ points from $a_i$ into the closure of this element, see Fig. 4.1(a). The discrete problem is obtained from (2.1)-(2.3) using the following approximations:

$$(b \cdot \nabla u_h, \varphi_i) \approx (b(a_i) \cdot \nabla u_h)|_{K_{ij}^{\text{upwind}}} (\varphi_i), \quad (f, \varphi_i) \approx (f(a_i), \varphi_i).$$

Then the discrete problem satisfies the discrete maximum principle and uniform convergence of the discrete solution $u_h$ to the solution $u$ of (1.1) can be proved.

4. **Upwinding techniques based on dual meshes.** Many upwinding techniques are based on a dual mesh consisting of mutually disjoint domains $D_i$ assigned to vertices $a_i$ of the triangulation $T_h$, see Fig. 4.1(b). In this way, a new subdivision of $\Omega$ is obtained. We denote by $\Gamma_i$ the boundary of $D_i$, by $n_i$ the unit outward normal vector to $\Gamma_i$ and we set $\Gamma_{ij} = \Gamma_i \cap \Gamma_j$. To derive an upwind discretization of the convective term in (2.3), we first consider the approximation

$$(b \cdot \nabla u_h, \varphi_i) = (\text{div}(b u_h) - u_h \text{div} b, \varphi_i) \approx \int_{D_i} \text{div}(b u_h) - u_h(a_i) \text{div} b \, dx.$$

Consequently,

$$(b \cdot \nabla u_h, \varphi_i) \approx \sum_{j=1}^{N_h} \int_{\Gamma_{ij}} (u_h - u_h(a_i)) b \cdot n_i \, d\sigma.$$

Setting $a_{ij}^{up} = a_i$ if $\int_{\Gamma_{ij}} b \cdot n_i \, d\sigma \geq 0$ and $a_{ij}^{up} = a_j$ if $\int_{\Gamma_{ij}} b \cdot n_i \, d\sigma < 0$, we obtain the upwind approximation

$$(b \cdot \nabla u_h, \varphi_i) \approx \sum_{j=1}^{N_h} (u_h(a_{ij}^{up}) - u_h(a_i)) \int_{\Gamma_{ij}} b \cdot n_i \, d\sigma.$$
Usually, either circumcentric or barycentric dual elements $D_i$ are used. In both cases, if $K$ is an element containing the vertex $a_i$, then $Q^K_i \equiv D_i \cap K$ is a quadrilateral whose vertices are $a_i$, a point $z^K_i \in K$ and the midpoints of the two edges of $K$ containing $a_i$. In the circumcentric case, the two edges of $Q^K_i$ containing $z^K_i$ are perpendicular to the edges of $K$ whereas, in the barycentric case, $z^K_i$ is the barycentre of $K$. Using the described approximation of the convective term and performing a similar modification of the right-hand side of (2.3) as in Tabata’s method, we obtain the method of Kanayama [6] in the circumcentric case and the method of Baba and Tabata [1] in the barycentric case. Let us also mention the partial upwind scheme of Ikeda [4] where upwinding is applied only to a part of the convective term, depending on $\varepsilon$ and $b$. In the strongly convection-dominated case $\varepsilon \ll |b|$ this method reduces to the method of Kanayama. For all three methods mentioned in this section the discrete maximum principle is satisfied and uniform convergence results are available.

5. The Mizukami–Hughes method. An interesting upwinding technique satisfying the discrete maximum principle was introduced by Mizukami and Hughes [9]. It is a Petrov–Galerkin method of the form (2.1), (2.2), (2.4) with weighting functions

$$\tilde{\varphi}_i = \varphi_i + \sum_{K \in T_h, \ a_i \in \overline{K}} C^K_i \chi_K, \quad i = 1, \ldots, M_h,$$

where $C^K_i$ are constants and $\chi_K$ is the characteristic function of $K$ (i.e., $\chi_K = 1$ in $K$ and $\chi_K = 0$ elsewhere). The flow velocity $b$ is considered to be piecewise constant. The idea is to choose the constants $C^K_i$ for any $K \in T_h$ in such a way that

$$C^K_i \geq -\frac{1}{3} \quad \forall \ i \in \{1, \ldots, N_h\}, \ a_i \in \overline{K}, \quad \sum_{i=1}^{N_h} \sum_{a_i \in \overline{K}} C^K_i = 0 \quad (5.1)$$

and that the local convection matrix $A^K$ with entries

$$a^K_{ij} = (b \cdot \nabla \varphi_j, \tilde{\varphi}_i)_K, \quad i = 1, \ldots, M_h, \ j = 1, \ldots, N_h, \ a_i, a_j \in \overline{K},$$

is of nonnegative type (i.e., off-diagonal entries of $A^K$ are nonpositive and the sum of the entries in each row of $A^K$ is nonnegative). As usual, $(\cdot, \cdot)_K$ denotes the inner product in $L^2(K)$.

Let $K$ be any element of the triangulation $T_h$ and let the vertices of $K$ be $a_1$, $a_2$ and $a_3$. For each vertex $a_i$, $i = 1, 2, 3$, we define a vertex zone $VZ_i$ and an edge
zone EZ, whose boundaries consist of lines intersecting the barycentre of K which are parallel to the two edges of K possessing the vertex ai, see Fig. 5.1. The common part of the boundaries of two adjacent zones is included in the respective vertex zone.

Without loss of generality, we may assume that the vertices of K are numbered in such a way that b points into the vertex zone or the edge zone of a1 as depicted in Fig. 5.1. If b ∈ VZ1, then (5.1) holds and \( A^K \) is of nonnegative type for

\[
C^K_1 = \frac{2}{3}, \quad C^K_2 = C^K_3 = -\frac{1}{3}.
\]

On the other hand, if b ∈ EZ1, then it is generally not possible to choose the constants \( C^K_1, C^K_2, C^K_3 \) in such a way that (5.1) holds and \( A^K \) is of nonnegative type. Since \( u \) still solves the equation (1.1) if we replace \( b \) by any function \( \tilde{b} \) such that \( \tilde{b} - b \) is orthogonal to \( \nabla u \), Mizukami and Hughes proposed to define the constants \( C^K_i \) in such a way that the matrix \( A^K \) is of nonnegative type for \( b \) replaced by a function \( \tilde{b} \) pointing into a vertex zone. Since \( \nabla u \) is not known a priori, we obtain a nonlinear problem where the constants \( C^K_i \) depend on the unknown discrete solution \( u_h \).

Let us assume that \( b \in EZ_1 \) and \( b \cdot \nabla u_h|_K \neq 0 \) and let \( w \neq 0 \) be a vector orthogonal to \( \nabla u_h|_K \). Then at least one of the sets

\[
V_k = \{ \alpha \in \mathbb{R} : b + \alpha w \in VZ_k \}, \quad k = 2, 3,
\]

is nonempty. Mizukami and Hughes show that, depending on \( V_2 \) and \( V_3 \), the following values of the constants \( C^K_i \) should be used:

\[
\begin{align*}
(5.2) & \quad V_2 \neq \emptyset \quad & \text{and} \quad V_3 = \emptyset \quad \implies \quad C^K_2 = \frac{2}{3}, \quad C^K_1 = C^K_3 = -\frac{1}{3}, \\
(5.3) & \quad V_2 = \emptyset \quad & \text{and} \quad V_3 \neq \emptyset \quad \implies \quad C^K_3 = \frac{2}{3}, \quad C^K_1 = C^K_2 = -\frac{1}{3}, \\
(5.4) & \quad V_2 \neq \emptyset \quad & \text{and} \quad V_3 \neq \emptyset \quad \implies \quad C^K_1 = -\frac{1}{3}, \quad C^K_2 + C^K_3 = \frac{1}{3}, \\
& \quad C^K_2 \geq -\frac{1}{3}, \quad C^K_3 \geq -\frac{1}{3}.
\end{align*}
\]

In case (5.4), Mizukami and Hughes suggest to set

\[
C^K_i = \frac{b \cdot \nabla \varphi_i}{3|b \cdot \nabla \varphi_i|}, \quad i = 1, 2, 3.
\]

This choice is also considered if \( b \in EZ_1 \) satisfies \( b \cdot \nabla u_h|_K = 0 \). If \( b = 0 \), Mizukami and Hughes set \( C^K_i = 0 \) for \( i = 1, 2, 3 \). The discrete problem satisfies the discrete maximum principle but no convergence results are available.
6. The improved Mizukami–Hughes method. The constants $C^K_i$ of the Mizukami–Hughes method depend on the orientation of both $b$ and $w$ in a discontinuous way. This does not seem to be reasonable and it may deteriorate the quality of the discrete solution and prevent the nonlinear iterative process from converging. Therefore, Knobloch [7] proposed another way how to compute the constants $C^K_i$ for $b$ pointing into an edge zone is not appropriate if $K$ lies in the numerical boundary layer. As a remedy, he proposed to set $C^K_i = \frac{1}{3}$ for all $i$ corresponding to inner vertices. All these changes resulted in the improved Mizukami–Hughes method with constants $C^K_i$ defined according to Fig. 6.1.

In Fig. 6.1, we again consider an element $K \in T_h$ with vertices $a_1, a_2$ and $a_3$. If $b \neq 0$, we assume that $b$ points into the vertex zone or the edge zone of $a_1$ (cf. Fig. 5.1) and we denote

$$s = b \frac{b}{|b|}, \quad v_2 = \frac{a_2 - a_1}{|a_2 - a_1|}, \quad v_3 = \frac{a_3 - a_1}{|a_3 - a_1|}, \quad v = \frac{v_2 + v_3}{|v_2 + v_3|}.$$

Further, we introduce unit vectors $w, v^\perp, v^\perp_2$ and $v^\perp_3$ such that

$$w \cdot \nabla u_h|_K = 0, \quad v^\perp \cdot v = 0, \quad v^\perp_2 \cdot v_2 = 0, \quad v^\perp_3 \cdot v_3 = 0, \quad w \cdot v \geq 0, \quad v^\perp \cdot v^\perp_3 \geq 0.$$

The improved method preserves the general properties of the original Mizukami–Hughes method, particularly, it satisfies the discrete maximum principle.

7. Numerical results. In this section we present numerical results for the convection–diffusion equation (1.1) considered in $\Omega = (0, 1)^2$ with the data $\varepsilon = 10^{-7}$, $b = (\cos(-\pi/3), \sin(-\pi/3))^T$, $f = 0$ and

$$u_0(x, y) = \begin{cases} 0 & \text{for } x = 1 \text{ or } y \leq 0.7, \\ 1 & \text{else}. \end{cases}$$

The solution possesses an interior layer in the direction of the convection starting at $(0, 0.7)$. On the boundary $x = 1$ and on the right part of the boundary $y = 0$, exponential layers are developed. We solved this problem on a uniform triangulation of the type depicted in Fig. 7.1(a) consisting of 800 triangles.

Fig. 7.1(b)–7.1(h) show the results obtained using the methods discussed in this paper. The Galerkin solution was computed for $\varepsilon = 10^{-3}$ since for $\varepsilon \to 0$ the linear system is very difficult to solve. The upwind methods as well as the DCCD method satisfy the discrete maximum principle, but lead to a smearing of the layers. The best solution was computed by the improved Mizukami–Hughes method, see Fig. 7.1(h). The high accuracy of the improved Mizukami–Hughes method was also demonstrated by many other numerical experiments, see e.g. [5, 7].

REFERENCES

IF \( b = \mathbf{0} \) THEN
\[
C_1^K = C_2^K = C_3^K = 0
\]
ELSE IF \( b \in \mathbf{VZ}_1 \) THEN
\[
C_1^K = \frac{2}{3}, \quad C_2^K = C_3^K = -\frac{1}{3}
\]
ELSE IF \( \mathcal{K} \cap \partial \Omega \neq \emptyset \) THEN
\[
C_1^K = C_2^K = C_3^K = -\frac{1}{3}
\]
ELSE IF \( T_h \) is not of the type from Fig. 7.1(a) and all vertices of \( \mathcal{K} \) are connected by edges to vertices on \( \partial \Omega \) THEN
\[
C_1^K = C_2^K = C_3^K = -\frac{1}{3}
\]
ELSE IF \( b \cdot \nabla u_h|_{\mathcal{K}} = 0 \) THEN
\[
C_1^K = -\frac{1}{3}, \quad C_2^K = C_3^K = \frac{1}{6}
\]
ELSE IF \( V_2 \neq \emptyset \) \& \( V_3 = \emptyset \) THEN
\[
C_2^K = \frac{2}{3}, \quad C_1^K = C_3^K = -\frac{1}{3}
\]
ELSE IF \( V_2 = \emptyset \) \& \( V_3 \neq \emptyset \) THEN
\[
C_3^K = \frac{2}{3}, \quad C_1^K = C_2^K = -\frac{1}{3}
\]
ELSE IF \( w \cdot \mathbf{v}^\perp < 0 \) THEN
\[
\begin{align*}
C_1^K &= \min \left\{ 1, \frac{|s \cdot \mathbf{v}_2^\perp|}{|\mathbf{v} \cdot \mathbf{v}_2^\perp|} + 1 - \text{sgn}(b \cdot \mathbf{v}_2) \right\}, \\
\Phi &= \min \left\{ 1, \frac{2|w \cdot \mathbf{v}_2^\perp|}{r_2 \mathbf{v} \cdot \mathbf{v}_2} \right\}, \\
C_2^K &= -\frac{1}{3} + \frac{2}{3} \Phi \left[ 1 + \frac{(v_2 - v_3) \cdot s}{1 - \mathbf{v}_2 \cdot \mathbf{v}_3} \right], \\
C_3^K &= \frac{1}{3} - C_2^K, \\
C_1^K &= -\frac{1}{3}
\end{align*}
\]
ELSE
\[
\begin{align*}
C_1^K &= \min \left\{ 1, \frac{|s \cdot \mathbf{v}_3^\perp|}{|\mathbf{v} \cdot \mathbf{v}_3^\perp|} + 1 - \text{sgn}(b \cdot \mathbf{v}_3) \right\}, \\
\Phi &= \min \left\{ 1, \frac{2|w \cdot \mathbf{v}_3^\perp|}{r_3 \mathbf{v} \cdot \mathbf{v}_3} \right\}, \\
C_2^K &= -\frac{1}{3} + \frac{2}{3} \Phi \left[ 1 + \frac{(v_3 - v_2) \cdot s}{1 - \mathbf{v}_3 \cdot \mathbf{v}_2} \right], \\
C_3^K &= \frac{1}{3} - C_2^K, \\
C_1^K &= -\frac{1}{3}
\end{align*}
\]

Fig. 6.1. Definition of the constants \( C^K_I \) in the improved Mizukami–Hughes method.

(a) Type of triangulation
(b) Galerkin, $\varepsilon = 10^{-3}$
(c) SUPG
(d) DCCD
(e) Tabata
(f) Baba, Tabata
(g) Kanayama
(h) Mizukami, Hughes

Fig. 7.1. Numerical results.