Proceedings of the Czech–Japanese Seminar in Applied Mathematics 2006
Czech Technical University in Prague, September 14-17, 2006
pp. 137–147

APPLICATION OF THE MIZUKAMI–HUGHES METHOD TO BILINEAR FINITE ELEMENTS*

PETR KNOBLOCH[1]

Abstract. This paper is devoted to the numerical solution of scalar two–dimensional steady convection–diffusion equations using the Mizukami–Hughes method. The Mizukami–Hughes method is a Petrov–Galerkin finite element method satisfying the discrete maximum principle and providing very accurate discrete solutions in convection–dominated regime. However, up to now, this method was available only for conforming triangular linear finite elements. The aim of this paper is to present an extension of the method to bilinear quadrilateral finite elements.

Key words. Stabilized FEM, convection–diffusion equations, Petrov–Galerkin method, discrete maximum principle.

AMS subject classifications. 65N30, 65N12

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

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega. \tag{1.1}$$

Here Ω is a bounded two–dimensional domain with a polygonal boundary $\partial\Omega$, f is a given outer source of the unknown scalar quantity u (e.g., temperature or concentration), $\varepsilon > 0$ is the diffusivity, which is assumed to be constant, and $\boldsymbol{b}$ is the flow velocity. The equation (1.1) is equipped with boundary conditions

$$u = u_b \quad \text{on } \Gamma^D, \qquad \varepsilon\,\frac{\partial u}{\partial \boldsymbol{n}} = g \quad \text{on } \Gamma^N, \tag{1.2}$$

where Γ^D and Γ^N are disjoint and relatively open subsets of the boundary $\partial\Omega$ satisfying $\mathrm{meas}_1(\Gamma^D) > 0$ and $\overline{\Gamma^D \cup \Gamma^N} = \partial\Omega$, $\boldsymbol{n}$ is the outward unit normal vector to $\partial\Omega$ and u_b, g are given functions.

Numerical solution of this seemingly very simple problem has been the subject of an extensive research for several decades but remains still a challenge if convection strongly dominates diffusion. The reason is that, in the convection–dominated regime, the solution of (1.1), (1.2) typically contains narrow inner and boundary layers which cannot be resolved properly unless the used mesh is extremely fine. 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.

To enhance the stability and accuracy of the Galerkin discretization of (1.1), (1.2) in convection–dominated regime, various stabilization strategies have been developed during the last three decades. One of the most efficient procedures for solving convection–dominated equations is the streamline upwind/ Petrov–Galerkin (SUPG)

*The work is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Charles University in Prague under the grant No. 344/2005/B–MAT/MFF.

[1]Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 186 75 Praha 8, Czech Republic

method [2] which consistently introduces numerical diffusion along streamlines. Although this method produces to a great extent accurate and oscillation–free solutions, it does not preclude small nonphysical 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 (cf. e.g. [1, 3, 4, 6, 9, 12, 14] and the references there). This procedure is usually referred to as discontinuity capturing (or shock capturing). A basic problem of most of these methods is the design of appropriate stabilization parameters which lead to sufficiently small nonphysical oscillations without compromising accuracy.

One of the first monotone methods for solving (1.1), (1.2) was introduced by Mizukami and Hughes [13] for linear triangular finite elements. In contrast to the most discontinuity–capturing methods, the solutions of the Mizukami–Hughes method always satisfy the discrete maximum principle, which ensures that no spurious oscillations will appear, not even in the vicinity of sharp layers. Further, as a method of upwind type, it does not contain any stabilization parameters, which also is a great advantage in comparison with the most other stabilized methods. Moreover, it is conservative and since it is a Petrov–Galerkin method, it is consistent. Last but not least, the Mizukami–Hughes method is based on a clear and simple idea whereas many discontinuity–capturing methods are derived using heuristic ad hoc arguments. Like many discontinuity–capturing methods for solving (1.1), (1.2), the Mizukami–Hughes method depends on the unknown discrete solution and hence it is nonlinear. Recently some improvements and extensions of the Mizukami–Hughes method were introduced by Knobloch [11].

The Mizukami–Hughes method in [13, 11] is defined only for conforming triangular linear finite elements and it is not obvious how to apply the method to other types of finite elements. The aim of the present paper is to propose an extension of the method to bilinear quadrilateral finite elements. First, in the next section, we introduce the concept of Petrov–Galerkin discretizations and a general form of the Mizukami–Hughes method considered in this paper. The properties of the method depend on the definitions of four constants on each element of the triangulation. The choice of these constants will be discussed in Sections 3. Since this will take several pages, we briefly summarize the definitions of the constants in Section 4. Finally, in Section 5, we show that our definition of the Mizukami–Hughes method for bilinear finite elements enables to fulfil the discrete maximum principle.

2. A Petrov–Galerkin method for convection–diffusion equations. Let $\mathcal{T}_h$ be a triangulation of Ω consisting of a finite number of open convex quadrilateral elements K. The discretization parameter h in the notation $\mathcal{T}_h$ is a positive real number satisfying $\operatorname{diam}(K) \le h$ for any $K \in \mathcal{T}_h$. We assume that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h} \overline{K}$ and that the closures of any two different elements of $\mathcal{T}_h$ are either disjoint or possess either a common vertex or a common edge. Further, we assume that any edge of an element $K \in \mathcal{T}_h$ which lies on $\partial\Omega$ is contained either in $\overline{\Gamma^D}$ or in $\overline{\Gamma^N}$.

Let $\widehat{K} = (-1,1)^2$ be the reference square and let $Q_1(\widehat{K})$ be the space of bilinear functions on $\widehat{K}$. For any convex quadrilateral $K \in \mathcal{T}_h$, there exists a regular mapping $F_K \in [Q_1(\widehat{K})]^2$ which maps $\widehat{K}$ onto K. The solution u of (1.1), (1.2) will be approximated by a function u_h from the space

$$V_h = \{v \in C(\overline{\Omega})\,;\; v \circ F_K \in Q_1(\widehat{K}) \;\; \forall\, K \in \mathcal{T}_h\}\,.$$

The space V_h is independent of the choice of the mappings F_K. Let $a_1, \dots, a_{M_h}$ be the vertices of $\mathcal{T}_h$ lying in $\Omega \cup \Gamma^N$ and let $a_{M_h+1}, \dots, a_{N_h}$ be the vertices of $\mathcal{T}_h$ lying on $\overline{\Gamma^D}$. For any $i \in \{1, \dots, N_h\}$, let $\varphi_i \in V_h$ be the function satisfying $\varphi_i(a_j) = \delta_{ij}$ for $j = 1, \dots, N_h$, where δ_{ij} is the Kronecker symbol. Then $V_h = \operatorname{span}\{\varphi_i\}_{i=1}^{N_h}$.

Let us recall that, in the classical Galerkin method, the discrete solution u_h satisfies

$$\varepsilon\,(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}\cdot\nabla u_h, \varphi_i) = (f, \varphi_i) + (g, \varphi_i)_{\Gamma^N}\,, \quad i = 1, \dots, M_h\,, \tag{2.1}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^2$ and $(\cdot,\cdot)_{\Gamma^N}$ is the inner product in $L^2(\Gamma^N)$. Similarly, we shall denote by $(\cdot,\cdot)_K$ the inner product in $L^2(K)$ or $L^2(K)^2$, where K is an element of $\mathcal{T}_h$. It is well known that the Galerkin discretization is inappropriate in the convection–dominated regime since it gives unphysically oscillating solutions. As a remedy, various authors proposed to add a weighted residual stabilization term of the form

$$\sum_{K\in\mathcal{T}_h} (-\varepsilon\,\Delta u_h + \boldsymbol{b}\cdot\nabla u_h - f, \psi_i)_K \tag{2.2}$$

to the left–hand side of (2.1) (cf. e.g. [2, 7, 8]). Of course, the properties of such a stabilization strongly depend on the choice of the functions ψ_i. According to our numerical experiences, the term $-\varepsilon\,\Delta u_h$ is negligible in the convection–dominated case and therefore, we shall not consider it in the following (note that $\Delta u_h = 0$ on any rectangular element K). Thus, adding (2.2) to (2.1) and introducing the weighting functions $\widetilde{\varphi}_i = \varphi_i + \psi_i$, we obtain

$$\varepsilon\,(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}\cdot\nabla u_h, \widetilde{\varphi}_i) = (f, \widetilde{\varphi}_i) + (g, \varphi_i)_{\Gamma^N}\,, \quad i = 1, \dots, M_h\,.$$

Further, we introduce approximations of the terms containing $\widetilde{\varphi}_i$. To this end, we define the mapping $\pi_h : L^1(\Omega) \to L^1(\Omega)$ by

$$(\pi_h\, v)|_K = \frac{1}{|K|}\int_K v\,\mathrm{d}x \qquad \forall\, v \in L^1(\Omega),\ K \in \mathcal{T}_h\,,$$

where $|K|$ denotes the area of K. Using this mapping, the weighting function $\widetilde{\varphi}_i$ will be replaced by the piecewise constant function $\pi_h\,\widetilde{\varphi}_i$. In addition, we replace the convective field $\boldsymbol{b}$ by a piecewise constant function $\boldsymbol{b}_h$ (e.g., $\boldsymbol{b}_h = \pi_h\,\boldsymbol{b}$). We shall also use the notation $\boldsymbol{b}_K \equiv \boldsymbol{b}_h|_K$.

Now we can introduce the Petrov–Galerkin method considered in this paper: we define the discrete solution of the problem (1.1), (1.2) as a function u_h satisfying

$$\begin{aligned}
&u_h \in V_h\,,\\
&\varepsilon\,(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}_h\cdot\nabla u_h, \pi_h\,\widetilde{\varphi}_i) = (f, \pi_h\,\widetilde{\varphi}_i) + (g, \varphi_i)_{\Gamma^N}\,, \quad i = 1, \dots, M_h\,,\\
&u_h(a_i) = u_b(a_i)\,, \quad i = M_h+1, \dots, N_h\,.
\end{aligned}$$

Like in [13], we define the weighting functions $\widetilde{\varphi}_i$ by

$$\widetilde{\varphi}_i = \varphi_i + \sum_{\substack{K\in\mathcal{T}_h,\\ a_i\in\overline{K}}} C_i^K\,\chi_K\,, \qquad i = 1, \dots, M_h\,,$$

where C_i^K are constants which will be determined in the next section and χ_K is the characteristic function of K (i.e., $\chi_K = 1$ in K and $\chi_K = 0$ elsewhere).

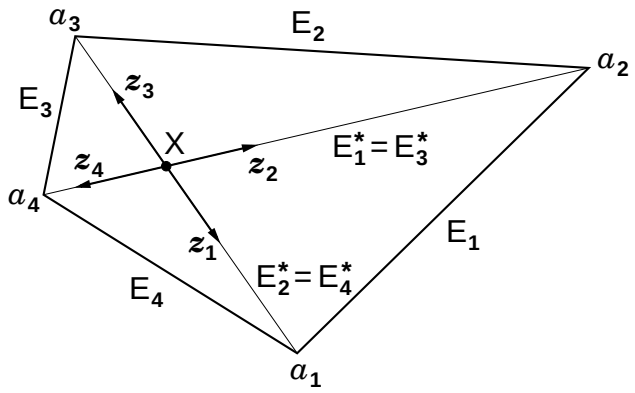

FIG. 3.1. *Notation for vertices, edges and vectors related to an element K.*

3. Choice of the constants C_i^K. In this section we shall discuss the choice of the constants C_i^K in the definition of the weighting functions $\widetilde{\varphi}_i$. We shall apply some of the ideas used in the triangular case treated in [11].

For any element $K \in \mathcal{T}_h$, let us denote

$$B_i^K = (\pi_h \varphi_i)|_K \qquad \forall\, i \in \{1, \ldots, N_h\}\,,\ a_i \in \overline{K}\,. \tag{3.1}$$

Then

$$B_i^K \in (0,1) \qquad \forall\, i \in \{1, \ldots, N_h\}\,,\ a_i \in \overline{K}\,, \qquad \sum_{\substack{i=1 \\ a_i \in \overline{K}}}^{N_h} B_i^K = 1\,.$$

Note also that $B_i^K = \frac{1}{4}$ if K is a parallelepiped. We require that

$$C_i^K \geq -B_i^K \qquad \forall\, i \in \{1, \ldots, N_h\}\,,\ a_i \in \overline{K}\,, \qquad \sum_{\substack{i=1 \\ a_i \in \overline{K}}}^{N_h} C_i^K = 0 \tag{3.2}$$

and that the local convection matrix A^K with entries

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

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, cf. [5]). The matrix A^K has four columns and at most four rows and it will be of nonnegative type as soon as $a_{ij}^K \leq 0$ for $i \neq j$. The latter condition in (3.2) implies that u_h satisfies a discrete mass conservation law if the data in (1.1), (1.2) satisfy $\Gamma^N = \partial\Omega$, $g = 0$ and $\boldsymbol{b} = \text{const.}$, cf. [10].

Consider any $K \in \mathcal{T}_h$ and denote the vertices of K by a_1, a_2, a_3 and a_4 (see Fig. 3.1). For $i = 1, \ldots, 4$, we denote by E_i the edge of K with vertices a_i, a_{i+1} and by T_i the triangle with vertices a_{i-1}, a_i, a_{i+1} (here and in the following all indices are considered modulo 4). Moreover, we denote by E_i^* the edge of the triangle T_i opposite the vertex a_i and we set

$$\boldsymbol{z}_i = \frac{a_i - a_{i+2}}{|a_i - a_{i+2}|}\,, \qquad i = 1, \ldots, 4\,. \tag{3.3}$$

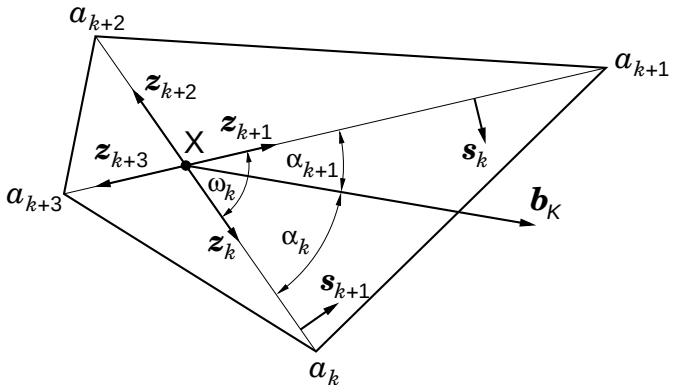

FIG. 3.2. *Definition of the angles ω_k, α_k and α_{k+1}.*

Of course, $\boldsymbol{z}_1 = -\boldsymbol{z}_3$ and $\boldsymbol{z}_2 = -\boldsymbol{z}_4$. Further, for $i = 1, \dots, 4$, we set

$$\boldsymbol{s}_i = \int_K \nabla\varphi_i \,\mathrm{d}x = \int_{\partial K} \varphi_i \,\boldsymbol{n}_{\partial K} \,\mathrm{d}\sigma = \frac{1}{2} \int_{E_{i-1}\cup E_i} \boldsymbol{n}_{\partial T_i} \,\mathrm{d}\sigma = -\frac{|E_i^*|}{2}\, \boldsymbol{n}_{\partial T_i}|_{E_i^*}\,, \tag{3.4}$$

where $\boldsymbol{n}_{\partial K}$ and $\boldsymbol{n}_{\partial T_i}$ denote the outer unit normal vectors to the boundary of K and T_i, respectively, and $|E_i^*|$ is the length of the edge E_i^*. Obviously,

$$\boldsymbol{s}_i \cdot \boldsymbol{z}_{i+1} = 0\,, \qquad \boldsymbol{s}_i = -\boldsymbol{s}_{i+2}\,, \qquad i = 1, \dots, 4\,. \tag{3.5}$$

Using the vectors $\boldsymbol{s}_i$, we can write the entries of the local convection matrix A^K in the form

$$a_{ij}^K = \boldsymbol{b}_K \cdot \boldsymbol{s}_j \,(B_i^K + C_i^K)\,,$$

which is convenient for discussing the choice of the constants C_i^K.

First let us assume that $\boldsymbol{b}_K = \alpha\,\boldsymbol{z}_k$ for some $\alpha > 0$ and $k \in \{1, \dots, 4\}$. Then, in view of (3.5), A^K is of nonnegative type and (3.2) holds if and only if

$$C_k^K = 1 - B_k^K \qquad \text{and} \qquad C_i^K = -B_i^K \quad \text{for } i \neq k,\ i = 1, \dots, 4\,. \tag{3.6}$$

Now, on the contrary, let the vectors $\boldsymbol{b}_K$ and $\boldsymbol{z}_i$ be linearly independent for any $i \in \{1, \dots, 4\}$. Then there exists $k \in \{1, \dots, 4\}$ such that

$$\boldsymbol{b}_K \cdot \boldsymbol{s}_k > 0\,, \quad \boldsymbol{b}_K \cdot \boldsymbol{s}_{k+1} > 0\,, \quad \boldsymbol{b}_K \cdot \boldsymbol{s}_{k+2} < 0\,, \quad \boldsymbol{b}_K \cdot \boldsymbol{s}_{k+3} < 0\,. \tag{3.7}$$

Thus, $\boldsymbol{b}_K$ points from the point $X = E_1^* \cap E_2^*$ into the convex subset of $\mathbb{R}^2$ the boundary of which consists of the half–lines $\{X + \alpha\,\boldsymbol{z}_k\,;\ \alpha > 0\}$ and $\{X + \alpha\,\boldsymbol{z}_{k+1}\,;\ \alpha > 0\}$ (cf. Fig. 3.2). It is obvious that the matrix A^K is of nonnegative type if and only if $C_i^K = -B_i^K$ for all $i = 1, \dots, 4$, which is not permitted by (3.2) (in this case all entries of A^K vanish). However, using an idea of Mizukami and Hughes [13], we can define the constants C_i^K in such a way that (3.2) holds and the coefficients of the discrete solution with respect to the basis of V_h solve a linear system with a matrix of nonnegative type. Such a definition of the constants C_i^K is based on the observation that u still solves the equation (1.1) if we replace $\boldsymbol{b}$ by any function $\tilde{\boldsymbol{b}}$ such that $\tilde{\boldsymbol{b}} - \boldsymbol{b}$ is orthogonal to ∇u. This suggests to define the constants C_i^K in such a way that the

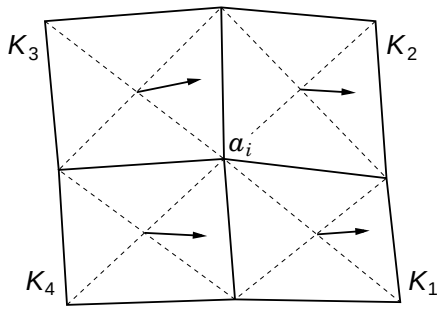

FIG. 3.3. *Notation for demonstrating the upwind character of the method (vectors indicate the directions of* $\boldsymbol{b}$*).*

matrix A^K is of nonnegative type for $\boldsymbol{b}_K$ replaced by a function $\tilde{\boldsymbol{b}}_K$ being a multiple of the vector $\boldsymbol{z}_k$ for some $k \in \{1, \dots, 4\}$. Since ∇u is not known a priori, we obtain a nonlinear discrete problem where the constants C_i^K depend on the discrete solution u_h which we want to compute.

Let us assume that $(\boldsymbol{b}_K, \nabla u_h)_K \neq 0$ and let $\boldsymbol{w}_K$ be a unit vector orthogonal to $(\pi_h \nabla u_h)|_K$. Then $\boldsymbol{w}_K$ and $\boldsymbol{b}_K$ are linearly independent and it is possible to find $\beta \in \mathbb{R}$ such that the vector $\tilde{\boldsymbol{b}}_K \equiv \boldsymbol{b}_K + \beta\, \boldsymbol{w}_K$ satisfies $\tilde{\boldsymbol{b}}_K = \alpha\, \boldsymbol{z}_k$ or $\tilde{\boldsymbol{b}}_K = \alpha\, \boldsymbol{z}_{k+1}$ for some $\alpha > 0$ (we still consider the case (3.7)). In the former case, we denote the respective value of β by β_k, in the latter case by β_{k+1}. If, for some $l \in \{k, k+1\}$, there is no $\beta \in \mathbb{R}$ such that $\tilde{\boldsymbol{b}}_K = \alpha\, \boldsymbol{z}_l$ with $\alpha > 0$, we set $\beta_l = 0$. Note that, for $l \in \{k, k+1\}$,

$$\beta_l \neq 0 \qquad \Longleftrightarrow \qquad (\boldsymbol{b}_K + \beta_l\, \boldsymbol{w}_K) \cdot \boldsymbol{s}_l > 0, \quad (\boldsymbol{b}_K + \beta_l\, \boldsymbol{w}_K) \cdot \boldsymbol{s}_{l+1} = 0\,.$$

Consequently, for $l \in \{k, k+1\}$, we have

$$\begin{aligned} (3.8) \qquad \beta_l \neq 0 \qquad \Longleftrightarrow \qquad & (\boldsymbol{w}_K \cdot \boldsymbol{s}_k)(\boldsymbol{w}_K \cdot \boldsymbol{s}_{k+1}) < 0 \quad \text{or} \\ & |(\boldsymbol{w}_K \cdot \boldsymbol{s}_l)(\boldsymbol{b}_K \cdot \boldsymbol{s}_{l+1})| < |(\boldsymbol{b}_K \cdot \boldsymbol{s}_l)(\boldsymbol{w}_K \cdot \boldsymbol{s}_{l+1})|\,. \end{aligned}$$

The above considerations lead us to the following values of the constants C_i^K, $i = 1, \dots, 4$:

$$\begin{aligned} (3.9) \qquad \beta_k \neq 0 \ \ \& \ \ \beta_{k+1} = 0 \qquad \Longrightarrow \qquad & C_k^K = 1 - B_k^K\,, \\ & C_i^K = -B_i^K \quad \forall\ i \neq k\,, \end{aligned}$$

$$\begin{aligned} (3.10) \qquad \beta_k = 0 \ \ \& \ \ \beta_{k+1} \neq 0 \qquad \Longrightarrow \qquad & C_{k+1}^K = 1 - B_{k+1}^K\,, \\ & C_i^K = -B_i^K \quad \forall\ i \neq k+1\,, \end{aligned}$$

$$\begin{aligned} (3.11) \qquad \beta_k \neq 0 \ \ \& \ \ \beta_{k+1} \neq 0 \qquad \Longrightarrow \qquad & C_i^K = -B_i^K \ \forall\ i \neq k,\, k+1\,, \\ & C_k^K \text{ and } C_{k+1}^K \text{ satisfy (3.2)}\,. \end{aligned}$$

We shall see in Section 5 that this choice of the constants C_i^K enables to prove that the discrete solution u_h satisfies the discrete maximum principle. Note also that in all three cases (3.9)–(3.11), we set $C_{k+2}^K = -B_{k+2}^K$ and $C_{k+3}^K = -B_{k+3}^K$, which gives rise to an upwind effect. Indeed, if we consider the situation depicted in Fig. 3.3, we have $C_i^{K_1} = -B_i^{K_1}$, $C_i^{K_2} = -B_i^{K_2}$ and hence

$$(\boldsymbol{b}_h \cdot \nabla u_h, \pi_h\, \widetilde{\varphi}_i) = (\boldsymbol{b}_h \cdot \nabla u_h, \pi_h\, \widetilde{\varphi}_i)_{K_3 \cup K_4}\,.$$

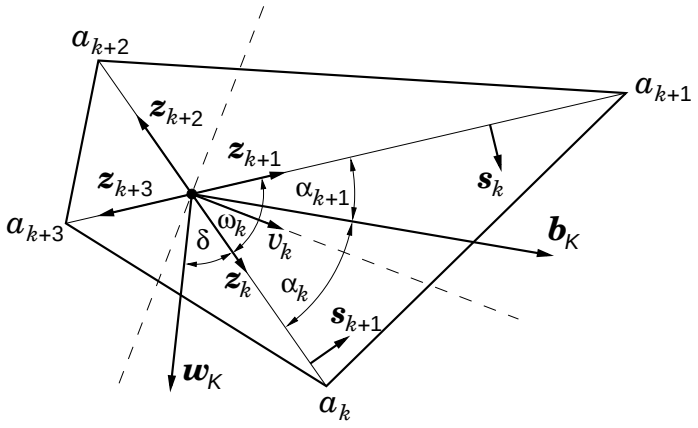

FIG. 3.4. *Definition of the angle δ and the vector $\boldsymbol{v}_k$.*

It remains to decide how to define the constants C_k^K and C_{k+1}^K in case (3.11). In view of (3.6), this definition should satisfy

$$\begin{aligned} \alpha_k \to 0 \qquad &\Longrightarrow \qquad C_k^K \to 1 - B_k^K\,, \quad C_{k+1}^K \to -B_{k+1}^K\,, \\ \alpha_{k+1} \to 0 \qquad &\Longrightarrow \qquad C_k^K \to -B_k^K\,, \quad C_{k+1}^K \to 1 - B_{k+1}^K\,, \end{aligned}$$

where α_k and α_{k+1} are the angles formed by the vectors $\boldsymbol{b}_K$, $\boldsymbol{z}_k$ and $\boldsymbol{b}_K$, $\boldsymbol{z}_{k+1}$, respectively, see Fig. 3.2 or 3.4. We denote by ω_k the angle formed by the vectors $\boldsymbol{z}_k$ and $\boldsymbol{z}_{k+1}$. Then $\alpha_k + \alpha_{k+1} = \omega_k$. It seems to be reasonable to require that

$$C_k^K = \tfrac{1}{2}\,[1 - \xi_k(\alpha_k)] - B_k^K\,, \qquad C_{k+1}^K = \tfrac{1}{2}\,[1 + \xi_k(\alpha_k)] - B_{k+1}^K\,, \tag{3.12}$$

where $\xi_k : [0, \omega_k] \to [-1, 1]$ is a monotone continuous function of α_k which is odd with respect to the value $\omega_k/2$ and satisfies

$$\xi_k(0) = -1\,, \qquad \xi_k(\omega_k) = 1\,.$$

The simplest choice is to set $\xi_k(\alpha_k) = 2\,\alpha_k/\omega_k - 1 = (\alpha_k - \alpha_{k+1})/(\alpha_k + \alpha_{k+1})$. However, to make the computation cheaper, we use

$$\xi_k(\alpha_k) = \frac{\sin[\frac{1}{2}\,(\alpha_k - \alpha_{k+1})]}{\sin[\frac{1}{2}\,(\alpha_k + \alpha_{k+1})]} = \frac{\cos\alpha_{k+1} - \cos\alpha_k}{1 - \cos(\alpha_k + \alpha_{k+1})}\,. \tag{3.13}$$

It follows from (3.8) that the case (3.11) applies if and only if $q_k \equiv (\boldsymbol{w}_K \cdot \boldsymbol{s}_k)(\boldsymbol{w}_K \cdot \boldsymbol{s}_{k+1}) < 0$. However, as soon as the sign of q_k changes, the constants C_i^K change to values given by (3.9) or (3.10). Consequently, the constants C_i^K depend on the orientation of $\boldsymbol{w}_K$ in a discontinuous way, which may prevent the nonlinear iterative process from converging. Therefore, it is desirable to modify the formulas (3.12) taking into account the orientation of $\boldsymbol{w}_K$.

Thus, let $q_k < 0$ and let $l, m \in \{k, k+1\}$, $l \neq m$, be such that $\boldsymbol{w}_K \cdot \boldsymbol{s}_l > 0$ and $\boldsymbol{w}_K \cdot \boldsymbol{s}_m < 0$ ($l = k$ in Fig. 3.4). We shall need some additional notation which is introduced in Fig. 3.4. Here, the dashed lines are axes of the angles formed by the vectors $\boldsymbol{z}_k$, $\boldsymbol{z}_{k+1}$ and $\boldsymbol{z}_{k+1}$, $\boldsymbol{z}_{k+2}$. The magnitude of the former angle is equal to ω_k and we introduce a unit vector $\boldsymbol{v}_k$ in the direction of the axis of this angle pointing

as in Fig. 3.4. Without loss of generality, we may assume that $\boldsymbol{w}_K \cdot \boldsymbol{v}_k \geq 0$ and we denote by δ the angle between $\boldsymbol{w}_K$ and $\boldsymbol{z}_l$. It suffices to discuss the choice of C_l^K since $C_m^K = 1 - (B_l^K + C_l^K) - B_m^K$. Obviously, $\alpha_l \in (0, \omega_k)$ and $\delta \in (0, \kappa]$ with $\kappa = \frac{\pi}{2} - \frac{\omega_k}{2}$. We shall require the following values of C_l^K in the limit cases:

$$\begin{aligned}
\delta = \kappa \quad &\Longrightarrow \quad C_l^K \text{ is determined by (3.12)}\,,\\
\alpha_l \to 0\,, \quad \delta \not\to 0 \quad &\Longrightarrow \quad C_l^K \text{ is determined by (3.12)} \\
&\qquad\qquad\qquad\qquad (\Rightarrow\ C_l^K \to 1 - B_l^K)\,,\\
\delta \to 0\,, \quad \alpha_l \not\to 0 \quad &\Longrightarrow \quad C_l^K \to -B_l^K\,.
\end{aligned}$$

Denoting by $\overline{C}_l^K$ the value of C_l^K determined by (3.12), we set

$$C_l^K = \Phi(\alpha_l, \delta)\, \overline{C}_l^K - [1 - \Phi(\alpha_l, \delta)]\, B_l^K\,,$$

where $\Phi : ([0, \omega_k] \times [0, \kappa]) \setminus (0,0) \to [0,1]$ is a continuous function. The above requirements imply that

$$\Phi(\alpha_l, \kappa) = \Phi(0, \delta) = 1\,, \quad \Phi(\alpha_l, 0) = 0 \qquad \forall\ \alpha_l \in (0, \omega_k],\ \delta \in (0, \kappa]\,.$$

The function Φ can be defined in various ways and we set

$$\Phi(\alpha_l, \delta) = \min\left\{1, \frac{2\,\sin\delta}{r_l\,\sin\kappa}\right\}\,, \tag{3.14}$$

where

$$r_l = \begin{cases} \dfrac{\sin\alpha_l}{\sin\frac{\omega_k}{2}} & \text{if } \alpha_l < \frac{\omega_k}{2}, \\ 1 & \text{if } \alpha_l \geq \frac{\omega_k}{2}. \end{cases}$$

The computation of (3.13) and (3.14) is inexpensive as we shall see in the next section.

Up to now, we have assumed that $(\boldsymbol{b}_K, \nabla u_h)_K \neq 0$. If $\boldsymbol{b}_K = \boldsymbol{0}$, we set $C_i^K = 0$ for all $i = 1, \ldots, 4$. If $\boldsymbol{b}_K \neq \boldsymbol{0}$ and $(\boldsymbol{b}_K, \nabla u_h)_K = 0$, we use average values of those ones defined in (3.9) and (3.10) since these values are used as soon as the direction of $(\pi_h \nabla u_h)|_K$ slightly changes.

4. Summary of the definitions of the constants C_i^K. In this section we summarize the definitions of the constants C_i^K introduced in the previous section and we rewrite them in a form appropriate for implementation. Let us consider any element $K \in \mathcal{T}_h$ and let a_1, a_2, a_3 and a_4 be its vertices, see Fig. 3.1. We shall use the constants B_i^K and the vectors $\boldsymbol{z}_i$ and $\boldsymbol{s}_i$, $i = 1 \ldots, 4$, defined in (3.1), (3.3) and (3.4), respectively. If $\boldsymbol{b}_K \neq \boldsymbol{0}$, we denote by $k \in \{1, \ldots, 4\}$ the uniquely determined index satisfying

$$\boldsymbol{b}_K \cdot \boldsymbol{s}_k > 0\,, \qquad \boldsymbol{b}_K \cdot \boldsymbol{s}_{k+1} \geq 0\,.$$

Further, we set

$$\boldsymbol{s} = \frac{\boldsymbol{b}_K}{|\boldsymbol{b}_K|}\,, \qquad \boldsymbol{v}_k = \frac{\boldsymbol{z}_k + \boldsymbol{z}_{k+1}}{|\boldsymbol{z}_k + \boldsymbol{z}_{k+1}|}$$

IF $\boldsymbol{b}_K \neq \boldsymbol{0}$ THEN

$$C^K_{k+2} = -B^K_{k+2}\,, \quad C^K_{k+3} = -B^K_{k+3}$$

IF $\boldsymbol{b}_K = \boldsymbol{0}$ THEN

$$C^K_1 = C^K_2 = C^K_3 = C^K_4 = 0$$

ELSE IF $(\boldsymbol{b}_K, \nabla u_h)_K = 0$ THEN

$$C^K_k = \tfrac{1}{2} - B^K_k\,, \quad C^K_{k+1} = \tfrac{1}{2} - B^K_{k+1}$$

ELSE IF $q_k \geq 0$ & $|(\boldsymbol{w}_K \cdot \boldsymbol{s}_k)(\boldsymbol{b}_K \cdot \boldsymbol{s}_{k+1})| < |(\boldsymbol{b}_K \cdot \boldsymbol{s}_k)(\boldsymbol{w}_K \cdot \boldsymbol{s}_{k+1})|$ THEN

$$C^K_k = 1 - B^K_k\,, \quad C^K_{k+1} = -B^K_{k+1}$$

ELSE IF $q_k \geq 0$ THEN

$$C^K_k = -B^K_k\,, \quad C^K_{k+1} = 1 - B^K_{k+1}$$

ELSE IF $\boldsymbol{w}_K \cdot \boldsymbol{s}_k > 0$ THEN

$$r_k = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{z}^\perp_k|}{|\boldsymbol{v}_k \cdot \boldsymbol{z}^\perp_k|} + 1 - \operatorname{sgn}(\boldsymbol{b}_K \cdot \boldsymbol{z}_k)\right\},$$

$$\Phi = \min\left\{1, \frac{2\,|\boldsymbol{w}_K \cdot \boldsymbol{z}^\perp_k|}{r_k\,\boldsymbol{v}_k \cdot \boldsymbol{z}_k}\right\},$$

$$C^K_k = -B^K_k + \frac{1}{2}\,\Phi\left[1 + \frac{(\boldsymbol{z}_k - \boldsymbol{z}_{k+1}) \cdot \boldsymbol{s}}{1 - \boldsymbol{z}_k \cdot \boldsymbol{z}_{k+1}}\right],$$

$$C^K_{k+1} = B^K_{k+2} + B^K_{k+3} - C^K_k$$

ELSE

$$r_{k+1} = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{z}^\perp_{k+1}|}{|\boldsymbol{v}_k \cdot \boldsymbol{z}^\perp_{k+1}|} + 1 - \operatorname{sgn}(\boldsymbol{b}_K \cdot \boldsymbol{z}_{k+1})\right\},$$

$$\Phi = \min\left\{1, \frac{2\,|\boldsymbol{w}_K \cdot \boldsymbol{z}^\perp_{k+1}|}{r_{k+1}\,\boldsymbol{v}_k \cdot \boldsymbol{z}_{k+1}}\right\},$$

$$C^K_{k+1} = -B^K_{k+1} + \frac{1}{2}\,\Phi\left[1 + \frac{(\boldsymbol{z}_{k+1} - \boldsymbol{z}_k) \cdot \boldsymbol{s}}{1 - \boldsymbol{z}_k \cdot \boldsymbol{z}_{k+1}}\right],$$

$$C^K_k = B^K_{k+2} + B^K_{k+3} - C^K_{k+1}\,.$$

FIG. 4.1. *Definition of the constants* C^K_i.

and we introduce unit vectors $\boldsymbol{w}_K$, $\boldsymbol{z}^\perp_k$ and $\boldsymbol{z}^\perp_{k+1}$ such that

$$\boldsymbol{w}_K \cdot (\pi_h \nabla u_h)|_K = 0\,, \quad \boldsymbol{w}_K \cdot \boldsymbol{v}_k \geq 0\,, \quad \boldsymbol{z}^\perp_k \cdot \boldsymbol{z}_k = 0\,, \quad \boldsymbol{z}^\perp_{k+1} \cdot \boldsymbol{z}_{k+1} = 0\,.$$

Finally, we set

$$q_k = (\boldsymbol{w}_K \cdot \boldsymbol{s}_k)(\boldsymbol{w}_K \cdot \boldsymbol{s}_{k+1})\,.$$

Then the constants $C^K_1, \ldots, C^K_4$ are determined according to the algorithm in Fig. 4.1.

5. Validity of the discrete maximum principle. Let K be any element of the triangulation $\mathcal{T}_h$ with vertices denoted by a_1, a_2, a_3 and a_4 like in the preceding sections. On this element, the discrete solution introduced in Section 2 can be written in the form $u_h|_K = \sum_{i=1}^4 u_i\,\varphi_i$ and we define the vector $U = (u_1, u_2, u_3, u_4)$. Our aim is to show that, defining the constants $C_1^K, \ldots, C_4^K$ as described in Section 3, one can find a matrix $\tilde{A}^K$ which has the same size as the local convection matrix A^K, is of nonnegative type and satisfies

$$A^K\,U = \tilde{A}^K\,U\,. \tag{5.1}$$

Of course, if $(\boldsymbol{b}_K, \nabla u_h)_K = 0$ or if $\boldsymbol{b}_K$ points in the direction of some of the vectors $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_4$, we can set $\tilde{A}^K = 0$ or $\tilde{A}^K = A^K$, respectively. Thus, let us assume that $(\boldsymbol{b}_K, \nabla u_h)_K \neq 0$ and that (3.7) holds with some $k \in \{1, \ldots, 4\}$. If $\beta_l \neq 0$ for some $l \in \{k, k+1\}$, we define the matrix $\tilde{A}^{K,l}$ having the entries

$$\tilde{a}_{ij}^{K,l} = (\boldsymbol{b}_K + \beta_l\,\boldsymbol{w}_K)\cdot\boldsymbol{s}_j\,(B_i^K + C_i^K)\,, \qquad i,j = 1,\ldots,4,\ \ a_i \in \Omega \cup \Gamma^N,$$

with C_i^K's from (3.9) if $l = k$ and with C_i^K's from (3.10) if $l = k+1$. As we have seen in Section 3, the matrix $\tilde{A}^{K,l}$ is of nonnegative type. Let us assume that β_k or β_{k+1} is equal to zero and let $l \in \{k, k+1\}$ be such that $\beta_l \neq 0$. Then, for $i = 1, \ldots, 4$ (with $a_i \in \Omega \cup \Gamma^N$), we have

$$\begin{aligned}(A^K\,U)_i &= (\boldsymbol{b}_K, \nabla u_h)_K\,(B_i^K + C_i^K)\\ &= (\boldsymbol{b}_K + \beta_l\,\boldsymbol{w}_K, \nabla u_h)_K\,(B_i^K + C_i^K) = (\tilde{A}^{K,l}\,U)_i\,.\end{aligned}$$

Thus, (5.1) holds with $\tilde{A}^K = \tilde{A}^{K,l}$. In case (3.11), we have

$$A^K = (B_k^K + C_k^K)\,A^{K,k} + (B_{k+1}^K + C_{k+1}^K)\,A^{K,k+1}\,,$$

where $A^{K,k}$ and $A^{K,k+1}$ are matrices defined like A^K but using C_i^K's from (3.9) and (3.10), respectively. Consequently, (5.1) holds with

$$\tilde{A}^K = (B_k^K + C_k^K)\,\tilde{A}^{K,k} + (B_{k+1}^K + C_{k+1}^K)\,\tilde{A}^{K,k+1}\,.$$

To establish the validity of the discrete maximum principle, the triangulation $\mathcal{T}_h$ has to be such that the discrete maximum principle holds for the pure diffusion problem (i.e., for $\boldsymbol{b} = \boldsymbol{0}$). For this it is sufficient to assume that, for any element $K \in \mathcal{T}_h$, the local diffusion matrix $\{(\nabla\varphi_j, \nabla\varphi_i)_K\}_{i,j=1}^4$ is of nonnegative type. Denoting by D the matrix having the entries $d_{ij} = (\nabla\varphi_j, \nabla\varphi_i)$, $i = 1, \ldots, M_h$, $j = 1, \ldots, N_h$, and by $\tilde{A}$ the $M_h \times N_h$ matrix made up of the local matrices $\tilde{A}^K$, we see that the vector of coefficients of the discrete solution u_h with respect to the basis $\{\varphi_i\}_{i=1}^{N_h}$ of the space V_h is the solution of a linear system with the matrix $C \equiv \varepsilon D + \tilde{A}$ where all three matrices are of nonnegative type. Moreover, since the matrix $\{d_{ij}\}_{i,j=1}^{M_h}$ is nonsingular, the matrix $\{c_{ij}\}_{i,j=1}^{M_h}$ also is nonsingular. This implies that u_h satisfies the discrete maximum principle (see e.g. [6]). Thus, for any $G \subset \overline{\Omega}$ being a union of closures of elements of $\mathcal{T}_h$, we have

$$\begin{aligned}(f, \widetilde{\varphi}_i) \le 0 \quad \forall\ a_i \in \operatorname{int} G \quad &\Rightarrow \quad \max_G u_h = \max_{\partial G} u_h\,,\\ (f, \widetilde{\varphi}_i) \ge 0 \quad \forall\ a_i \in \operatorname{int} G \quad &\Rightarrow \quad \min_G u_h = \min_{\partial G} u_h\,,\end{aligned}$$

which shows that the discrete solution does not contain any spurious oscillations.

REFERENCES

[1] R.C. Almeida and R.S. Silva, *A Stable Petrov–Galerkin Method for Convection–Dominated Problems*, Comput. Methods Appl. Mech. Engrg., **140** (1997), 291–304.

[2] A.N. Brooks and T.J.R. Hughes, *Streamline Upwind/Petrov–Galerkin Formulations for Convection Dominated Flows with Particular Emphasis on the Incompressible Navier–Stokes Equations*, Comput. Methods Appl. Mech. Engrg., **32** (1982), 199–259.

[3] E. Burman and P. Hansbo, *Edge Stabilization for Galerkin Approximations of Convection–Diffusion–Reaction Problems*, Comput. Methods Appl. Mech. Engrg., **193** (2004), 1437–1453.

[4] E.G.D. do Carmo and G.B. Alvarez, *A New Upwind Function in Stabilized Finite Element Formulations, Using Linear and Quadratic Elements for Scalar Convection–Diffusion Problems*, Comput. Methods Appl. Mech. Engrg., **193** (2004), 2383–2402.

[5] P.G. Ciarlet and P.–A. Raviart, *Maximum Principle and Uniform Convergence for the Finite Element Method*, Comput. Methods Appl. Mech. Engrg., **2** (1973), 17–31.

[6] R. Codina, *A Discontinuity–Capturing Crosswind–Dissipation for the Finite Element Solution of the Convection–Diffusion Equation*, Comput. Methods Appl. Mech. Engrg., **110** (1993), 325–342.

[7] L.P. Franca, S.L. Frey, and T.J.R. Hughes, *Stabilized Finite Element Methods. I.: Application to the Advective–Diffusive Model*, Comput. Methods Appl. Mech. Engrg., **95** (1992), 253–276.

[8] T.J.R. Hughes, L.P. Franca, and G.M. Hulbert, *A New Finite Element Formulation for Computational Fluid Dynamics. VIII. The Galerkin/least–squares method for advective–diffusive equations*, Comput. Methods Appl. Mech. Engrg., **73** (1989), 173–189.

[9] T.J.R. Hughes, M. Mallet, and A. Mizukami, *A New Finite Element Formulation for Computational Fluid Dynamics. II. Beyond SUPG*, Comput. Methods Appl. Mech. Engrg., **54** (1986), 341–355.

[10] T. Ikeda, *Maximum Principle in Finite Element Models for Convection–Diffusion Phenomena*, Lecture Notes in Numerical and Applied Analysis, Vol. 4. North–Holland, Amsterdam (1983).

[11] P. Knobloch, *Improvements of the Mizukami–Hughes Method for Convection–Diffusion equations*, Comput. Methods Appl. Mech. Engrg., **196** (2006), 579–594.

[12] T. Knopp, G. Lube, and G. Rapin, *Stabilized Finite Element Methods with Shock Capturing for Advection–Diffusion Problems*, Comput. Methods Appl. Mech. Engrg., **191** (2002), 2997–3013.

[13] A. Mizukami and T.J.R. Hughes, *A Petrov–Galerkin Finite Element Method for Convection–Dominated Flows: An Accurate Upwinding Technique for Satisfying the Maximum Principle*, Comput. Methods Appl. Mech. Engrg., **50** (1985), 181–193.

[14] Y.–T. Shih and H.C. Elman, *Modified Streamline Diffusion Schemes for Convection–Diffusion Problems*, Comput. Methods Appl. Mech. Engrg. **174** (1999), 137–151.