Modified SUPG method on oriented meshes

Jan Lamač

Abstract We consider the SUPG method for the numerical solution of the scalar steady convection-diffusion equation using conforming simplicial piecewise linear finite elements. We change the convective vector in the SUPG stabilizing term and adjust the triangulation so that the discrete maximum principle is satisfied. Then the error analysis is performed and the method is tested on several numerical examples.

1 Introduction and the idea of the method

Let us solve the convection-diffusion equation

$$-\varepsilon\Delta u(x) + \mathbf{b}(x)\nabla u(x) = f(x) \qquad \text{in } \Omega \subset \mathbb{R}^n\,, \tag{1}$$

$$u(x) = 0 \qquad \text{on } \partial\Omega\,, \tag{2}$$

where $n \in \mathbb{N}$, Ω is a bounded polytopic domain with Lipschitz-continuous boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\mathbf{b} \in W^{1,\infty}(\Omega)^n$ is a given convective field and $f \in L^2(\Omega)$ is an outer force. Further, we assume that the boundary $\partial\Omega$ is divided into three subsets $\Gamma_+ = \{x \in \partial\Omega,\, \mathbf{b}(x)\mathbf{n}(x) > 0\}$, $\Gamma_0 = \{x \in \partial\Omega,\, \mathbf{b}(x)\mathbf{n}(x) = 0\}$ and $\Gamma_- = \{x \in \partial\Omega,\, \mathbf{b}(x)\mathbf{n}(x) < 0\}$ satisfying

$$\overline{\partial\Omega} = \overline{\Gamma}_+ \cup \overline{\Gamma}_0 \cup \overline{\Gamma}_- \qquad \text{and} \qquad \Gamma_+ \cap \Gamma_0 = \Gamma_0 \cap \Gamma_- = \Gamma_- \cap \Gamma_+ = \emptyset.$$

Here, a vector $\mathbf{n}(x)$ denotes a unit outer normal to the boundary $\partial\Omega$. Let us also emphasize that in the whole article any two vectors are always multiplied by the standard inner product.

Jan Lamač
Charles University in Prague, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 186 75 Praha 8, Czech Republic, e-mail: jan.lamac@centrum.cz

As $\varepsilon \to 0$, the equation (1) becomes singularly perturbed and near the boundary Γ_+ the finite element solution often contains spurious oscillations. We call this region exponential boundary layer. In order to diminish the oscillations at the exponential boundary layers, one may use the SUPG method [1]. However, the SUPG method does not diminish all the oscillations, in particular, at the parabolic (characteristic) boundary layers. These regions usually appear near the boundary Γ_0, but also along interior layers that propagate from discontinuous boundary conditions at Γ_- .

Apart from the SUPG method, one can also use the method of Mizukami and Hughes [3]. Unlike the SUPG method, the Mizukami-Hughes method satisfies the discrete maximum principle and therefore it diminishes all the spurious oscillations at the layers. The drawback of the Mizukami-Hughes method is its nonlinearity and the absence of an error analysis. In order to eliminate this drawback we construct a special mesh, which is well-aligned with the vector field $\mathbf{b}$. The created linear method then enjoys both positive properties of the Mizukami-Hughes method and the SUPG method – it satisfies the discrete maximum principle and we can apply an error analysis analogous to the SUPG method.

Since ε is considered to be very small, the exact solution at any point $x \in \Omega$ in fact depends almost only on the values in the direction $-\mathbf{b}(x)$. It means that the discretization of the convective term should use only the upwind values. To achieve this, we construct a special mesh $\mathscr{T}_h$. Each element of such a mesh should have one of its edges oriented in the direction of the vector $\mathbf{b}$. Then, if $\mathbf{b}_K$ is a constant approximation of $\mathbf{b}$ on the element $K \in \mathscr{T}_h$ parallel to one of its edges and if we use simplicial finite elements with linear basis functions $\{\varphi_{K,i}\}_{i=1}^{n+1}$, only *two* values of $\mathbf{b}_K \nabla \varphi_{K,i}$, $i \in \{1,2,\dots,n+1\}$, are nonzero. This property can be used for characterization of a good mesh.

Due to their length we omit proofs of all lemmas and theorems. They can be found in the future work [2].

2 Derivation of the method

At the beginning of any finite element discretization, we derive the weak formulation of the respective problem. Let us therefore multiply (1) by the function $\varphi \in H_0^1(\Omega)$ and integrate over the whole domain Ω. Using the Green's theorem the weak formulation of (1) reads: Find $u \in H_0^1(\Omega)$ such that

$$\varepsilon(\nabla u, \nabla\varphi)_\Omega + (\mathbf{b}\nabla u, \varphi)_\Omega = (f, \varphi)_\Omega \qquad \forall \varphi \in H_0^1(\Omega). \tag{3}$$

Further, let us define a triangulation $\mathscr{T}_h$ of the domain Ω. It consists of the finite number of open simplicial elements K. We assume that $\overline{\Omega} = \cup_{K \in \mathscr{T}_h} \overline{K}$ and that the closures of any two different elements $K, \widetilde{K} \in \mathscr{T}_h$ are either disjoint or possess a common d-dimensional simplex ($d \in \{0,1,\dots,n-1\}$). We also denote by $\mathscr{M}_h$ the set of nodes of $\mathscr{T}_h$ and by $\mathscr{N}_h \subset \mathscr{M}_h$ the set of all inner nodes of $\mathscr{T}_h$.

To derive the Galerkin's finite element discretization of (1), we define a finite element space $X_h = \{v_h \in C(\Omega), v_h|_K \in P_1(K), \forall K \in \mathscr{T}_h\}$ and a space of test functions $V_h = X_h \cap H_0^1(\Omega)$. The barycentric coordinates $\{\varphi_{K,j}\}_{j=1}^{n+1}$ of the element $K \in \mathscr{T}_h$ then form a basis of $P_1(K)$ and we reorder them so that

$$\int_K \frac{\mathbf{b}\nabla\varphi_{K,j}}{|\nabla\varphi_{K,j}|}\,\mathrm{d}\mathbf{x} \;\leq\; \int_K \frac{\mathbf{b}\nabla\varphi_{K,j+1}}{|\nabla\varphi_{K,j+1}|}\,\mathrm{d}\mathbf{x}, \quad \text{for } j = 1, 2, \ldots, n. \tag{4}$$

Remark 1. Since $\sum_{j=1}^{n+1} \int_K \mathbf{b}\nabla\varphi_{K,j}\,\mathrm{d}\mathbf{x} = 0$ for each $K \in \mathscr{T}_h$, then if one of the previous expressions is nonzero we obtain $\int_K \mathbf{b}\nabla\varphi_{K,1}\,\mathrm{d}\mathbf{x} < 0$ and $\int_K \mathbf{b}\nabla\varphi_{K,n+1}\,\mathrm{d}\mathbf{x} > 0$.

Further, we assume that for each $K \in \mathscr{T}_h$ the barycentric coordinates $\{\varphi_{K,j}\}_{j=1}^{n+1}$ satisfy

$$(\varphi_{K,j}, \varphi_{K,i})_K \;\leq\; 0 \quad \text{whenever } i \neq j. \tag{5}$$

In 2D this assumption is satisfied for triangulations not containing obtuse triangles.

The SUPG method adds weighted residuals $R(u) = -\varepsilon\Delta u + \mathbf{b}\nabla u - f$ to the usual Galerkin's finite element method. Since $R(u)$ vanishes for the exact solution, we can add any multiple of $R(u)$ to the weak formulation. Unlike the original SUPG method, which adds the residual multiplied by the streamline derivative of v, we add the residual multiplied on each $K \in \mathscr{T}_h$ by derivative of v in the direction $P_{K,n+1} - C_K$. Here C_K are the barycentres of K and $P_{K,j}$, $j = 1, 2, \ldots, n+1$, are the vertices of K satisfying $\varphi_{K,i}(P_{K,j}) = \delta_{ij}$ for $1 \leq i, j \leq n+1$.

Thus, the solution $u \in H_0^1(\Omega) \cap H^2(\Omega)$ of the problem (3) satisfies also for all $\varphi \in H_0^1(\Omega)$

$$a(u, \varphi) \;=\; F(\varphi), \tag{6}$$

where

$$F(\varphi) = \sum_{K \in \mathscr{T}_h} (f, \varphi + (P_{K,n+1} - C_K)\nabla\varphi)_K \qquad \text{and} \tag{7}$$

$$a(u, \varphi) \;=\; \varepsilon(\nabla u, \nabla\varphi)_\Omega + (\mathbf{b}\nabla u, \varphi)_\Omega + \sum_{K \in \mathscr{T}_h} \Big(-\,\varepsilon\Delta u + \mathbf{b}\nabla u, (P_{K,n+1} - C_K)\nabla\varphi \Big)_K. \tag{8}$$

If we now apply the finite element method using the continuous piecewise linear finite elements, the spurious oscillations unfortunately persist (analogous to the original SUPG method). The reason is the presence of the positive off-diagonal entries in the matrix obtained by the discretization of the last two terms in (8) resulting in the non-fulfillment of the discrete maximum principle.

In order to eliminate these positive entries, we define $\mathbf{d}_K = P_{K,n+1} - P_{K,1}$ and consider the element-wise constant approximation $\mathbf{b}_K$ of the vector field $\mathbf{b}$ by vectors that are parallel with $\mathbf{d}_K$ on each element K. More precisely, first of all we consider that our mesh is ”well-aligned” with respect to the vector field $\mathbf{b}$ and then on each element K we construct a constant approximation $\mathbf{b}_K$ of $\mathbf{b}$. This ”well-alignment” provides following assumptions.

($\mathscr{A}1$) The ordering given by (4) on each $K \in \mathscr{T}_h$ uniquely defines the vector $\mathbf{d}_K = P_{K,n+1} - P_{K,1}$. We assume that if any edge e of $\mathscr{T}_h$ corresponds to $\mathbf{d}_K$ of some K, then e corresponds to $\mathbf{d}_K$ for each K containing e. We denote by $\mathscr{E}_h$ the set of such edges.

($\mathscr{A}2$) Each inner node P of $\mathscr{T}_h$ is the endpoint of exactly two edges of $\mathscr{E}_h$.

Remark 2. Let us call a *discrete streamline* any set of edges $\mathscr{S} \subset \mathscr{E}_h$ such that for each $e \in \mathscr{S}$ there exists $e' \in \mathscr{S}$ such that

$$e' \neq e \quad \& \quad e \cap e' \neq \emptyset. \tag{9}$$

The discrete streamline $\mathscr{S}$ is *closed* if for each $e \in \mathscr{S}$ there exist exactly two different edges e' and e'' satisfying (9). Consequently, the assumptions $(\mathscr{A}1)-(\mathscr{A}2)$ do not allow closed discrete streamlines in 2D. Indeed, if there is a closed discrete streamline then there exists a node ("inside" the closed streamline) which does not satisfy $(\mathscr{A}2)$. The mesh satisfying $(\mathscr{A}1)-(\mathscr{A}2)$ can be, for instance, constructed by approximation of streamlines by linear spline functions. This will be the subject of future work. Further assumptions on the structure of the mesh will be given by the inequalities (23) and (25).

It remains to define the piecewise constant approximation of $\mathbf{b}$. On each element $K \in \mathscr{T}_h$ it is defined in the following way

$$\mathbf{b}_K \;=\; -\frac{1}{|K|}\left(\int_K \mathbf{b}\nabla\varphi_{K,1}\,\mathrm{d}\mathbf{x}\right)\mathbf{d}_K. \tag{10}$$

Finally, we apply the finite element method and the new method reads:
Find $u_h \in V_h$ such that for all $\varphi_h \in V_h$ holds

$$a_h(u_h,\varphi_h) = F_h(\varphi_h), \tag{11}$$

where

$$\begin{aligned} a_h(u,\varphi) = {}& \varepsilon(\nabla u,\nabla\varphi)_\Omega + \sum_{K\in\mathscr{T}_h}(\mathbf{b}_K\nabla u,\varphi)_K + \\ & + \sum_{K\in\mathscr{T}_h}\Big(-\varepsilon\Delta u + \mathbf{b}_K\nabla u,(P_{K,n+1}-C_K)\nabla\varphi\Big)_K, \end{aligned} \tag{12}$$

$$F_h(\varphi) = \sum_{K\in\mathscr{T}_h}\Big(f,\varphi + (P_{K,n+1}-C_K)\nabla\varphi\Big)_K \tag{13}$$

and the vectors $\mathbf{b}_K$ are defined by (10). The stability of this method then results from the following remark.

Remark 3. Instead of adding stabilization terms to the weak formulation (3) one can equivalently define the method (11) by changing the test functions $\varphi_{K,j}$, $j \in \{1,2,\dots,n+1\}$, $K \in \mathscr{T}_h$, to

$$\widetilde{\varphi}_{K,j} \;=\; \varphi_{K,j} + (P_{K,n+1}-C_K)\nabla\varphi_{K,j}. \tag{14}$$

Then for all $j = 1, 2, \ldots, n$ we obtain $\widetilde{\varphi}_{K,j} = \varphi_{K,j} - \frac{1}{n+1}$ whereas $\widetilde{\varphi}_{K,n+1} = \varphi_{K,n+1} + \frac{n}{n+1}$. This choice of test functions is the same as in the Mizukami-Hughes method [3]. It means that the derived method satisfies the discrete maximum principle.

3 Coercivity

3.1 Technical lemmas

Since $\int_K v_h - v_h(C_K)\,\mathrm{d}\mathbf{x} = 0$ for all $v_h \in V_h$, we can write

$$\Big(\mathbf{b}_K \nabla u_h,\, v_h + (P_{K,n+1} - C_K)\nabla v_h\Big)_K = \Big(\mathbf{b}_K \nabla u_h,\, v_h + v_h(P_{K,n+1}) - v_h(C_K)\Big)_K =$$
$$= \Big(\mathbf{b}_K \nabla u_h, v_h(P_{K,n+1})\Big)_K = |K|\,\frac{|\mathbf{b}_K|}{|\mathbf{d}_K|}\Big(u_h(P_{K,n+1}) - u_h(P_{K,1})\Big) v_h(P_{K,n+1}).$$

Consequently, for the bilinear form a_h holds

$$a_h(v_h, v_h) = \varepsilon |v_h|^2_{1,\Omega} + \sum_{K \in \mathscr{T}_h} |K|\,\frac{|\mathbf{b}_K|}{|\mathbf{d}_K|}\Big(v_h(P_{K,n+1}) - v_h(P_{K,1})\Big) v_h(P_{K,n+1}). \tag{15}$$

Thus, when proving coercivity of the bilinear form a_h, it is necessary to estimate the second term on the right-hand side of (15). For this purpose we use the following lemmas.

Lemma 1. *Let $N \in \mathbb{N}$, $0 < \rho_j < 1$, $j = 1, 2, \ldots, N$, and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying $q_j/q_{j-1} \le \rho_{j-1}$ for $j = 2, 3, \ldots, N$. Then for all $v_j \in \mathbb{R}$, $j = 2, 3, \ldots, N+1$, holds*

$$q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1}^2 - v_j v_{j+1}) \ \ge\ \frac{1}{2} \sum_{j=1}^{N} (1 - \rho_j) q_j v_{j+1}^2. \tag{16}$$

Lemma 2. *Let $N \in \mathbb{N}$, $N \ge 8$, $0 \le \delta < 4$ and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying $q_j/q_{j-1} \le 1 + \delta/N^2$ for $j = 2, 3, \ldots, N$. Then for all $v_j \in \mathbb{R}$, $j = 2, 3, \ldots, N+1$, holds*

$$q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1}^2 - v_j v_{j+1}) \ \ge\ \frac{4 - \delta}{2N^2} \sum_{j=1}^{N} q_j v_{j+1}^2. \tag{17}$$

Remark 4. If we take $\delta = 0$ in Lemma 2, we obtain factor $\frac{2}{N^2}$ on the right-hand side of (17). It can be shown that the constant 2 in this estimate is not optimal, nevertheless, the order is optimal $\left(\frac{1}{N^2}\right)$.

The upper bound $\delta < 4$ is not optimal as well. However, if we consider $N = 5$, $\frac{q_j}{q_{j-1}} = 1 + \frac{25}{3}\frac{1}{N^2} = \frac{4}{3}$ for $j = 2, 3, 4, 5$, then $q_1 v_2^2 + \sum_{j=2}^{5} q_j \left(v_{j+1}^2 - v_j v_{j+1}\right) = 0$ for

$(v_2, v_3, v_4, v_5, v_6) = (16, 24, 24, 18, 9)$. Hence, the optimal upper bound for δ is not greater than $\frac{25}{3}$.

Lemma 3. *Let $N \in \mathbb{N}$ and let q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying $q_j/q_{j-1} \le 1$ for $j = 2, 3, \ldots, N$. Then for all $v_j \in \mathbb{R}$, $j = 2, 3, \ldots, N+1$, holds*

$$q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1}^2 - v_j v_{j+1}) \ \ge\ \frac{1}{2} \left\{ q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1} - v_j)^2 \right\}. \tag{18}$$

Lemma 4. *Let $N \in \mathbb{N}$, $N \ge 8$, $0 \le \delta < 4$ and q_j, $j = 1, 2, \ldots, N$, are positive numbers satisfying $q_j/q_{j-1} \le 1 + \delta/N^2$ for $j = 2, 3, \ldots, N$. Then for all $v_j \in \mathbb{R}$, $j = 2, 3, \ldots, N+1$, holds*

$$q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1}^2 - v_j v_{j+1}) \ \ge\ \frac{4-\delta}{8} \left\{ q_1 v_2^2 + \sum_{j=2}^{N} q_j (v_{j+1} - v_j)^2 \right\}. \tag{19}$$

3.2 Coercivity estimates

We use the properties of the mesh oriented in the flow direction and derive the coercivity estimates in suitable norms. Again, we observe that the mesh whose edges are oriented along **b** has a special property: For each mesh node P_0^s lying on the boundary Γ_- there exists a sequence of nodes $\{P_j^s\}_{j=1}^{N_s}$ which lay on the same discrete streamline given by the vector field **b**.

Thus, each node P_j^s of the mesh can be characterised by two numbers - the number denoting the streamline (s) and the number determining the order of the node on this streamline (j). For each node P_j^s we can further define the following sets: a patch $\Omega_j^s = \cup_{P_j^s \subset \overline{K}} K$, a cluster $\mathscr{C}_j^s = \cup_{P_{j-1}^s, P_j^s \subset \overline{K}} K$ and a complementary set $\Omega_{0,j}^s = \Omega_j^s \backslash \left(\mathscr{C}_j^s \cup \mathscr{C}_{j+1}^s\right)$ (see Figure 1).

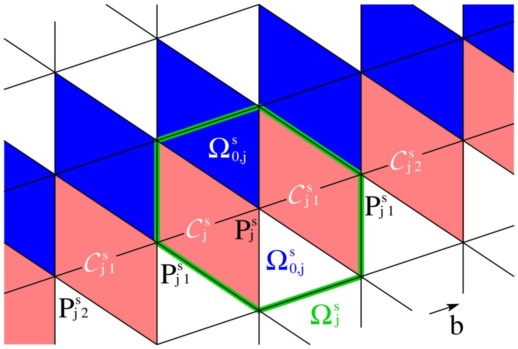

Fig. 1 Definition of the splitting of the domain Ω_j^s.

Definition 1. For each cluster $\mathscr{C}_j^s$ let us define the quantities $h_j^s = |\mathbf{d}_K|, K \subset \mathscr{C}_j^s$, $\beta_j^s = \frac{1}{|\mathscr{C}_j^s|}\sum_{K\subset\mathscr{C}_j^s}|\mathbf{b}_K||K|$ and $q_j^s = -\sum_{K\subset\mathscr{C}_j^s}\int_K \mathbf{b}\nabla\varphi_{K,1}\,\mathrm{d}\mathbf{x} = \beta_j^s|\mathscr{C}_j^s|/h_j^s > 0$.
For each element $K \in \mathscr{T}_h$ let us also define the mesh parameters θ_K and ν_K by

$$\theta_K \;=\; \frac{1}{|K|}\max\left\{\max_{2\le i\le n}\left|\int_K \mathbf{b}\nabla\varphi_{K,i}\,\mathrm{d}\mathbf{x}\right|,\left|\sum_{i=2}^{n}\int_K \mathbf{b}\nabla\varphi_{K,i}\,\mathrm{d}\mathbf{x}\right|\right\} \quad \text{and} \tag{20}$$

$$\nu_K = \max_{1\le i\le n+1}\sigma(P_{K,i}) \quad \text{with} \quad \sigma(P_j^s) = \frac{|\mathscr{C}_j^s| + |\Omega_{0,j}^s|}{|\mathscr{C}_j^s|}\,(h_j^s N_s)^2\,\frac{\|\mathbf{b}\|_{\infty,\Omega_j^s}}{\beta_j^s}\,\frac{\max_{K\subset\Omega_j^s} h_K}{h_j^s}. \tag{21}$$

In the previous definition the quantity h_j^s represents the length of the cluster in the direction of $\mathbf{d}_K$, β_j^s is the weighted average value of $|\mathbf{b}_K|$ on $\mathscr{C}_j^s$ and q_j^s are flows which we use in the Lemmas 1 – 4.

The mesh parameters θ_K vanish whenever $\mathbf{b}$ is parallel to $\mathbf{b}_K$ in K and therefore we use them for characterization of a good mesh. Finally, in the case of constant vector $\mathbf{b}$, $h_j^s N_s = L$ and for a mesh consisting of regular simplices it holds

$$\sigma(P_j^s) \;=\; \frac{|K|n! + |K|(n-1)n!}{|K|n!}L^2 \;=\; nL^2 \qquad \text{for all possible } j,s. \tag{22}$$

For more general data we obtain different values of $\sigma(P_j^s)$ or ν_K, however, the value (22) is still a good approximation, in particular, for quasi-equidistant meshes.

We use these quantities together with the Lemmas 1 – 4 and prove the coercivity of the method with respect to two types of energy norms (see Definition 2).

Theorem 1 **($\operatorname{div}\mathbf{b} < 0$).** *Let there exists $\omega > 0$ such that $\operatorname{div}\mathbf{b} \le -\omega < 0$ in Ω and let for each $K \in \mathscr{T}_h$ holds*

$$\theta_K \;\le\; \frac{\omega}{n+1}. \tag{23}$$

Further, let there exists the constant κ independent of h (and ε) such that $\frac{|\mathscr{C}_{j+1}^s|}{|\Omega_j^s|} \ge \kappa$ for all $s = 1,2,\ldots,\mathscr{P}$ and $j = 1,2,\ldots,N_s$, then

$$a_h(v_h,v_h) \;\ge\; \frac{1}{2}\left\{\varepsilon|v_h|_{1,\Omega}^2 \;+\; \frac{\omega\kappa}{2}\|v_h\|_{0,\Omega}^2 \;+\; \sum_{K\in\mathscr{T}_h}\frac{h_K}{2|\mathbf{b}_K|}\|\mathbf{b}_K\nabla v_h\|_{0,K}^2\right\}. \tag{24}$$

Remark 5 ($\operatorname{div}\mathbf{b} = 0$). If there exists $\delta \ge 0$ such that

$$\theta_K \;\le\; \frac{\delta}{\nu_K}\|\mathbf{b}\|_{\infty,K}\,h_K \qquad \text{for each } K \in \mathscr{T}_h, \tag{25}$$

then in the case when $\operatorname{div}\mathbf{b} = 0$ in Ω we obtain (due to the Lemma 4) the estimate

$$a_h(v_h,v_h) \;\ge\; \varepsilon|v_h|_{1,\Omega}^2 \;+\; \frac{4-\delta}{4}\sum_{K\in\mathscr{T}_h}\frac{h_K}{2|\mathbf{b}_K|}\|\mathbf{b}_K\nabla v_h\|_{0,K}^2. \tag{26}$$

Theorem 2 (div $\mathbf{b}=0$)**.** *Let the assumption (25) be fulfilled. Further, let there exist positive numbers* κ, L, R, β *such that for all* $s=1,2,\dots,\mathscr{P}$ *and* $j=1,2,\dots,N_s$ *holds*

$$|\mathscr{C}_j^s| \geq \kappa |\Omega_j^s|, \qquad N_s h_j^s \leq L, \qquad \max_{K\subset\Omega_j^s} h_K \leq R h_j^s, \qquad \textit{and} \qquad \beta_j^s \geq \beta. \tag{27}$$

Then $a_h(v_h,v_h) \geq \varepsilon |v_h|_{1,\Omega}^2 + \frac{(4-\delta)\kappa\beta}{2L^2R}(n+1)\sum_{K\in\mathscr{T}_h} h_K \|v_h\|_{0,K}^2$ *for each* $v_h \in V_h$.

4 Error estimates

Definition 2. We estimate the error of the presented method in the following types of energy norms $\left(\text{using } C_2^* = \frac{(4-\delta)\kappa\beta}{2L^2R}(n+1) \text{ and } C_b^* = \frac{4-\delta}{4}\right)$

$$|||v|||_b^2 = \varepsilon |v|_{1,\Omega}^2 + \frac{\omega\kappa}{2}\|v\|_{0,\Omega}^2 + \sum_{K\in\mathscr{T}_h} \frac{h_K}{2|\mathbf{b}_K|}\|\mathbf{b}_K \nabla v\|_{0,K}^2, \tag{28}$$

$$|||v|||_{b,*}^2 = \varepsilon |v|_{1,\Omega}^2 + C_2^* \sum_{K\in\mathscr{T}_h} h_K \|v\|_{0,K}^2 + C_b^* \sum_{K\in\mathscr{T}_h} \frac{h_K}{2|\mathbf{b}_K|}\|\mathbf{b}_K \nabla v\|_{0,K}^2. \tag{29}$$

We follow the error analysis applied in [4] using the coercivity estimates and the Galerkin's quasi-orthogonality resulting from the consistency of the method.

Theorem 3. *Let there exists constant* κ *independent of* h *(and* ε*) such that* $\frac{|\mathscr{C}_{j+1}^s|}{|\Omega_j^s|} \geq \kappa$ *for all* $s=1,2,\dots,\mathscr{P}$ *and* $j=1,2,\dots,N_s$*, constant* $\omega > 0$ *such that* $\operatorname{div}\mathbf{b} \leq -\omega < 0$ *in* Ω *and let for each* $K\in\mathscr{T}_h$ *holds*

$$\theta_K \leq \min\left\{\frac{\omega}{n+1}, |\mathbf{b}|_{1,\infty,K}\sqrt{\omega}\max\left\{\frac{h_K}{\varepsilon^{1/2}}, \frac{2}{|\mathbf{b}_K|}\varepsilon^{1/2}\right\}\right\}. \tag{30}$$

If the solution u *of (1) satisfies* $u\in H^2(\Omega)$*, then there exists constant* $C>0$ *independent of* h *and* ε *such that for the solution obtained by the method (11) it holds*

$$|||u-u_h|||_b \leq C\left(\sum_{K\in\mathscr{T}_h} \min\left\{h_K^2, \max\left\{\frac{h_K^4}{\varepsilon}, \varepsilon h_K^2\right\}\right\}\left(|u|_{2,K}^2 + |u|_{1,K}^2\right)\right)^{1/2}, \tag{31}$$

i.e., for $h_K \geq \varepsilon^{1/2}$ *the order of the convergence is 1, for* $\varepsilon \leq h_K \leq \varepsilon^{1/2}$ *the order increases to* $3/2$*, whereas for* $h_K \leq \varepsilon$ *the order decreases back to* 1.

Theorem 4. *Let* $\operatorname{div}\mathbf{b} = 0$ *and let there exists* $\delta \in (0,4)$ *such that (25) is satisfied. Further, let there exist positive numbers* κ, L, R *and* β *such that for all* $s=1,2,\dots,\mathscr{P}$ *and* $j=1,2,\dots,N_s$ *holds (27).*

If the solution u *of (1) satisfies* $u\in H^2(\Omega)$*, then there exists constant* $C^*>0$ *independent of* h *and* ε *such that for the solution obtained by the method (11) it holds*

$$|||u-u_h|||_{b,*} \le C^* \left(\sum_{K\in\mathscr{T}_h} \min\left\{ h_K, \max\left\{ \frac{h_K^4}{\varepsilon}, \varepsilon h_K^2 \right\} \right\} \left(|u|_{2,K}^2 + |u|_{1,K}^2 \right) \right)^{1/2}, \quad (32)$$

i.e., for $h_K \ge \varepsilon^{1/3}$ *the order of the convergence is* $1/2$*, for* $\varepsilon \le h_K \le \varepsilon^{1/3}$ *the order increases to* $3/2$*, whereas for* $h_K \le \varepsilon$ *the order decreases from* $3/2$ *to* 1.

5 Numerical experiments

Example 1

Let us consider $\Omega \subset \mathbb{R}^n$ and let $P = [P_1, P_2, \dots, P_n] \in \mathbb{R}^n$ be any point such that $P \notin \overline{\Omega}$. Further, let us choose any constant $\omega > 0$ and define $\mathbf{b}(\mathbf{x}) = \frac{\omega}{n}(P-\mathbf{x})$, i.e., $b_i(\mathbf{x}) = \frac{\omega}{n}(P_i - x_i)$, where $\mathbf{x} = [x_1, x_2, \dots, x_n]$. Then $\operatorname{div}\mathbf{b} = -\omega$ and the streamlines of $\mathbf{b}$ are rays ending at the point P. Moreover, it can be shown, that for each element K with one edge lying on the streamline holds $\theta_K = \frac{n-1}{n}\frac{\omega}{n+1} < \frac{\omega}{n+1}$. Thus, the condition (23) is always satisfied, however, one cannot improve it to (25) or (30) by mesh-refinig.

To be more specific, let us consider the equation (1) with $n = 2$, $\Omega = (0, 0.9)^2$, $\mathbf{b} = \frac{1}{2}(1-x, 1-y)^T$ and $\varepsilon = 10^{-6}$. The right-hand side f and the boundary condition are given by Figure 2 while the computed solution is depicted in Figure 3.

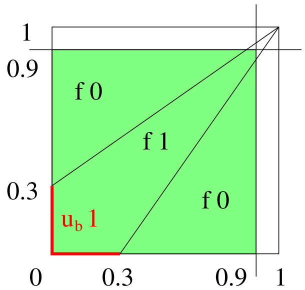

Fig. 2 Definition of the Example 1.

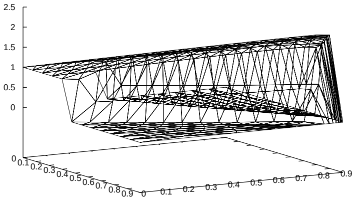

Fig. 3 Solution of the Example 1.

Example 2

Let us consider the equation (1) with $n = 2$, $\Omega = (0.05, 0.5)^2$, $\varepsilon = 10^{-6}$ and $\mathbf{b} = \left(1/\sqrt{x^2+y^2}\right)(-y, x)^T$. The right-hand side f and the boundary condition are given by Figure 4. The computed solution is depicted in Figure 5.

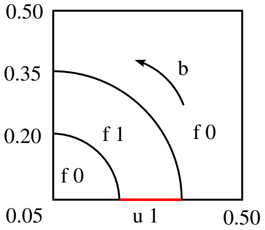

Fig. 4 Definition of the Example 1.

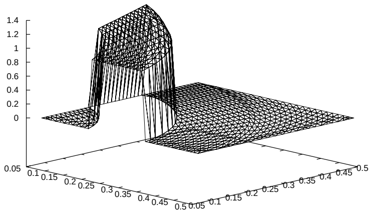

Fig. 5 Solution of the Example 2.

In order to demonstrate the discrete maximum principle property, we have compared the new method with the SUPG method (Figures 6 and 7).

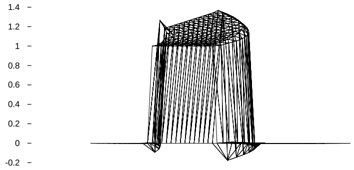

Fig. 6 The SUPG Solution of the Example 2.

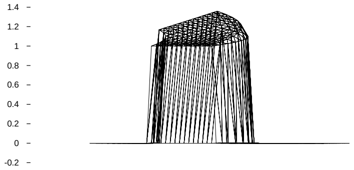

Fig. 7 Solution obtained by the new method.

6 Conclusion

We have constructed a new method for solving singularly perturbed problems: we added another stabilization term than in the SUPG method and adjusted the mesh so that the discrete maximum principle is satisfied. We also derived error estimates in appropriate energy norms. In spite of using first order finite elements it is also possible to extend the method to finite elements of higher orders. This extension and the construction of a suitable mesh generator will be the subject of the future research.

References

1. Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods in Appl. Mech. Engrg. **32**, 199–259 (1982).
2. Lamač, J.: Adaptive methods for singularly perturbed partial differential equations. Doctoral Thesis. Faculty of Mathematics and Physics, Charles University in Prague, expected in 2015.
3. Mizukami, A., Hughes, T.J.R.: A Petrov-Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle. Comput. Methods in Appl. Mech. Engrg. **50**, 181–193 (1985).
4. Roos, H.-G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. 2nd edition. Berlin: Springer-Verlag 2008.