Adaptive Choice of Parameters in Stabilization Methods for Convection-Diffusion Equations

P. Lukáš
Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic.

Abstract. This paper is devoted to the numerical solution of the scalar convection–diffusion equation. We present new results of the adaptive technique in finite element method based on minimizing a functional called error indicator. Particularly, we use more different spaces (space of piecewise constant functions, piecewise linear continuous functions, and piecewise linear discontinuous functions) for the parameter $\tau$ from the SUPG (SDFEM) method.

Introduction

We are concerned with the scalar convection–diffusion problem

$$-\varepsilon \Delta u + b \cdot \nabla u + cu = f \quad \text{in } \Omega, \quad u = u_b \quad \text{on } \Gamma^D, \quad \varepsilon \frac{\partial u}{\partial n} = g \quad \text{on } \Gamma^N.$$  

Here $\Omega \subset \mathbb{R}^2$ is a bounded domain with a polyhedral Lipschitz–continuous boundary $\partial \Omega$ and $\Gamma^D, \Gamma^N$ are disjoint and relatively open subsets of $\partial \Omega$ satisfying $\text{meas} (\Gamma^D) > 0$ and $\Gamma^D \cup \Gamma^N = \partial \Omega$. Furthermore, $n$ is the outward unit normal vector to $\partial \Omega$, $\varepsilon > 0$ is a constant diffusivity, $b \in W^{1,\infty}(\Omega)^2$ is the flow velocity, $c \in L^\infty(\Omega)$ is the reaction coefficient, $f \in L^2(\Omega)$ is a given outer source of the unknown scalar quantity $u$, and $u_b \in H^{1/2}(\Gamma^D)$, $g \in L^2(\Gamma^N)$ are given functions specifying the boundary conditions. We make the usual assumption that

$$c - \frac{1}{2} \text{div } b \geq 0.$$  

A well known aspect of the numerical solution of (1) are spurious oscillations which often appear in the discrete solution when convection dominates diffusion. Many various stabilized methods have been proposed. However, these methods often depend on parameters whose optimal choice is usually not known. John, Knobloch, and Savescu [2011] described how these parameters can be optimized by means of a posteriori error estimates. In this paper we enrich the space of parameters to get new results.

We will use the standard notation for usual function spaces and norms, see, e.g., Ciarlet [1978]. The notation $(\cdot, \cdot)_G$ is used for the inner product in the space $L^2(G)$ or $L^2(G)^d$ and we set $(\cdot, \cdot) = (\cdot, \cdot)_\Omega$.

Weak formulation

Let $\tilde{u}_b \in H^1(\Omega)$ be an extension of $u_b$ (i.e., the trace of $\tilde{u}_b$ equals $u_b$ on $\Gamma^D$) and let

$$V = \{ v \in H^1(\Omega); \ v = 0 \quad \text{on } \Gamma^D \}.$$  

Then the weak formulation of (1) reads: Find $u \in H^1(\Omega)$ such that $u - \tilde{u}_b \in V$ and

$$a(u, v) = (f, v) + (g, v)_{\Gamma^N} \quad \forall \ v \in V,$$  

where $a(u, v)$ is the usual bilinear form

$$a(u, v) = \varepsilon (\nabla u, \nabla v) + (b \cdot \nabla u, v) + (cu, v).$$  

From the assumption (2) it follows that the weak formulation has a unique solution.

Galerkin finite element discretization

Let $\{T_h\}$ be a family of triangulations of $\Omega$ parametrized by positive parameters $h$ whose only accumulation point is zero. The triangulations $T_h$ are assumed to consist of a finite number of open polyhedral subsets $K$ of $\Omega$ such that $\Omega = \bigcup_{K \in T_h} K$ and the closures of any two different sets in $T_h$ are either disjoint or possess either a common vertex or a common edge. Further, we assume that any edge of $T_h$ which lies on $\partial \Omega$ is contained either in $\Gamma^D$ or in $\Gamma^N$. 

\textcopyright MATFYZPRESS}
For each $h$, we introduce a finite element space $W_h \subset H^1(\Omega)$ defined on $\mathcal{T}_h$ and approximating the space $H^1(\Omega)$ in the usual sense, see Ciarlet [1978]. Furthermore, for each $h$, we introduce a function $\tilde{u}_h \in W_h$ whose trace on $\Gamma^D$ approximates $u_0$. Finally, we set $V_h = W_h \cap V$. Then the Galerkin discretization of (1) reads: Find $u_h \in W_h$ such that $u_h - \tilde{u}_h \in V_h$ and

$$a(u_h, v_h) = (f, v_h) + (g, v_h)_{\Gamma^N} \quad \forall v_h \in V_h. \quad (4)$$

Again, this problem is uniquely solvable.

**SUPG stabilization**

It is well known that the Galerkin discretization (4) is inappropriate if convection dominates diffusion since then the discrete solution is usually globally polluted by spurious oscillations. An improvement can be achieved by adding a stabilization term to the Galerkin discretization. One of the most efficient procedures of this type is the streamline upwind/Petrov–Galerkin (SUPG or SDFEM) method, Brooks and Hughes [1982], which is frequently used because of its stability properties and higher–order accuracy.

The SUPG stabilization depends on a stabilization parameter which will be denoted by $\tau_h$ in the following. We assume that all admissible stabilization parameters form a set $Y_h \subset L^\infty(\Omega)$. The SUPG discretization of (1) reads: Find $u_h \in W_h$ such that $u_h - \tilde{u}_h \in V_h$ and

$$a(u_h, v_h) + s_h(\tau_h; u_h, v_h) = (f, v_h) + (g, v_h)_{\Gamma^N} + r_h(\tau_h; v_h) \quad \forall v_h \in V_h, \quad (5)$$

where

$$s_h(\tau_h; u_h, v_h) = (-\varepsilon \Delta_h u_h + b \cdot \nabla u_h + c u_h, \tau_h b \cdot \nabla v_h),$$

$$r_h(\tau_h; v_h) = (f, \tau_h b \cdot \nabla v_h).$$

The SUPG method requires that the functions from $W_h$ are $H^2$ on each element of $\mathcal{T}_h$, which is satisfied for common finite element spaces. The notation $\Delta_h$ denotes the Laplace operator defined elementwise.

The parameter $\tau_h$ is often defined, on an element $K \in \mathcal{T}_h$, by the formula

$$\tau_h|_K = \frac{h_K}{2|b|} \xi(\text{Pe}_K) \quad \text{with} \quad \xi(\alpha) = \coth \alpha - \frac{1}{\alpha}, \quad \text{Pe}_K = \frac{|b| h_K}{2 \varepsilon}, \quad (6)$$

where $h_K$ is the element diameter in the direction of the convection vector $b$ and $\text{Pe}_K$ is the local Péclet number which determines whether the problem is locally (i.e., within a particular element) convection dominated or diffusion dominated.

**Optimization of parameters**

Let $D_h \subset Y_h$ be an open set such that, for any $\tau_h \in D_h$, the SUPG method (5) has a unique solution $u_h \in W_h$. To emphasize that $u_h$ depends on $\tau_h$, we shall write $u_h(\tau_h)$ instead of $u_h$ in the following. Let $I_h : W_h \to \mathbb{R}$ be an error indicator, i.e.,

$$\Phi_h(\tau_h) := I_h(u_h(\tau_h))$$

represents a measure of the error of the discrete solution $u_h(\tau_h)$ corresponding to a given parameter $\tau_h$. We use two different indicators in our tests proposed by John, Knobloch, and Savescu [2011]. The first indicator is defined as

$$I_h(w_h) = \sum_{K \in \mathcal{T}_h, K \cap \partial \Omega = \emptyset} h_K^2 \| - \varepsilon \Delta w_h + b \cdot \nabla w_h + c w_h - f \|_{0,K}^2 \quad \forall w_h \in W_h \quad (7)$$

and the second one, also referred to as “indicator with crosswind derivative control term”, is

$$I_h(w_h) = \sum_{K \in \mathcal{T}_h, K \cap \partial \Omega = \emptyset} (\| - \varepsilon \Delta w_h + b \cdot \nabla w_h + c w_h - f \|_{0,K}^2 + \| \phi([b^\perp \cdot \nabla w_h]) \|_{0,1,K}) \quad \forall w_h \in W_h, \quad (8)$$

where

$$b^\perp(x) = \begin{cases} \frac{[b(x) \cdot \nabla w_h(x)]}{|b(x)|} & \text{if } b(x) \neq 0, \\ 0 & \text{if } b(x) = 0, \quad \forall x \in \Omega, \quad \phi(t) = \begin{cases} \sqrt{t} & \text{if } t \geq 1, \\ 0.5(5t^2 - 3t^3) & \text{if } t < 1. \end{cases} \end{cases} \quad (9)$$
For the derivative of the indicator (7) we have
\[
\langle \partial_l I_h(\tilde{u}_h(\tau_h)), v_h \rangle = \sum_{K \in \mathcal{T}_h, K \cap \Gamma = \emptyset} h_K^2 2 \left( \mathcal{L} u_h(\tau_h) - f, \mathcal{L} v_h \right)_K \quad \forall \ v_h \in V_h, \tag{10}
\]
where \( \mathcal{L} = -\varepsilon \Delta + b \cdot \nabla + c \) and \( I_h, \tilde{u}_h \) are defined in the next section. For Indicator (8) we have
\[
\langle \partial_l I_h(\tilde{u}_h(\tau_h)), v_h \rangle = \sum_{K \in \mathcal{T}_h, K \cap \Gamma = \emptyset} 2 \left( \mathcal{L} u_h(\tau_h) - f, \mathcal{L} v_h \right)_K + \phi'(\langle b \cdot \nabla u_h(\tau_h) \rangle) \| b \cdot \nabla v_h \|_{0,1,K} \quad \tag{11}
\]
for all \( v_h \in V_h \). Our aim is to compute a parameter \( \tau_h \in D_h \) for which \( \Phi_h \) attains its minimum. To this end, it is convenient to compute effectively the Fréchet derivative of \( \Phi_h \).

**Fréchet derivative of \( \Phi_h \)**

For any \( \tau_h \in D_h \), we have \( u_h(\tau_h) = \tilde{u}_h(\tau_h) + \tilde{u}_h \) with \( \tilde{u}_h : D_h \to V_h \). Thus, we do not have to consider the space \( W_h \) but can work only with the space \( V_h \), which is more convenient.

We denote \( I_h(w_h) = I_h(w_h + \tilde{u}_h) \) for any \( w_h \in V_h \). Then \( I_h : V_h \to \mathbb{R} \) and
\[
\Phi_h(\tau_h) = I_h(\tilde{u}_h(\tau_h)) \quad \forall \ \tau_h \in D_h. \tag{12}
\]
We define the “residue” operator \( R_h : V_h \times Y_h \to V_h' \) by
\[
\langle R_h(w_h, \tau_h), v_h \rangle = a(w_h + \tilde{u}_h, v_h) + s_h(\tau_h; w_h + \tilde{u}_h, v_h) - \langle f, v_h \rangle, \quad \forall \ v_h, w_h \in V_h, \tau_h \in Y_h. \tag{13}
\]
Then it holds
\[
R_h(\tilde{u}_h(\tau_h), \tau_h) = 0 \quad \forall \ \tau_h \in Y_h. \tag{14}
\]
Let us assume that the mappings \( R_h = R_h(w_h, \tau_h), I_h = I_h(w_h) \) and \( \tilde{u}_h = \tilde{u}_h(\tau_h) \) are Fréchet-differentiable. We denote the respective Fréchet derivatives by \( \partial_w R_h, \partial_{\tau_h} R_h, \partial_l I_h, \) and \( \partial_l \tilde{u}_h \). Then the Fréchet derivative of \( \Phi_h \) exists and is given by
\[
D\Phi_h(\tau_h) = D\tilde{I}_h(\tilde{u}_h(\tau_h)) D\tilde{u}_h(\tau_h). \tag{15}
\]
However, this formula is not suitable for computing \( D\Phi_h(\tau_h) \) since the computation of \( D\tilde{u}_h(\tau_h) \) requires the solution of \( \dim Y_h \) systems of \( \dim V_h \) algebraic equations. We circumvent this difficulty by introducing an auxiliary mapping \( \psi_h : D_h \to V_h \) which solves the adjoint problem, see [Giles and Pierce [2000]],
\[
\langle \partial_w R_h \rangle'(\tilde{u}_h(\tau_h), \tau_h) \psi_h(\tau_h) = D\tilde{I}_h(\tilde{u}_h(\tau_h)). \tag{16}
\]
The operator \( \langle \partial_w R_h \rangle'(w_h, \tau_h) \) is defined by
\[
\langle \left( \partial_w R_h \right)'(w_h, \tau_h), v_h \rangle = \langle \left( \partial_w R_h \right)(w_h, \tau_h) v_h, v_h \rangle \quad \forall \ v_h, \tilde{v}_h \in V_h \tag{17}
\]
and \( \langle \partial_{\tau_h} R_h \rangle'(w_h, \tau_h) \) is defined in a similar way. According to (12), we have
\[
\partial_w R_h(\tilde{u}_h(\tau_h), \tau_h) D\tilde{u}_h(\tau_h) + \partial_{\tau_h} R_h(\tilde{u}_h(\tau_h), \tau_h) = 0 \tag{18}
\]
and hence we deduce that
\[
D\Phi_h(\tau_h) = -\langle \partial_{\tau_h} R_h \rangle'(\tilde{u}_h(\tau_h), \tau_h), \psi_h(\tau_h) \rangle. \tag{19}
\]
This can also be written in the form
\[
(D\Phi_h(\tau_h), \tilde{v}_h) = -\langle \left( \partial_{\tau_h} R_h \right)(\tilde{u}_h(\tau_h), \tau_h) v_h, \psi_h(\tau_h) \rangle \quad \forall \ \tilde{v}_h \in Y_h. \tag{20}
\]
To clarify the approach, we would like to give its algebraic (finite-dimensional) version, described carefully by [John, Knabche, and Sawaescu [2011]]. Let \( \tau_h \in D_h \) be given and denote by \( D\Phi_h \in \mathbb{R}^{1 \times \dim Y_h} \) and \( D\tilde{I}_h \in \mathbb{R}^{1 \times \dim Y_h} \) the vectors representing the derivatives \( D\Phi_h(\tau_h) \) and \( D\tilde{I}_h(\tilde{u}_h(\tau_h)) \), respectively. Furthermore, let \( D\tilde{u}_h \in \mathbb{R}^{\dim V_h \times \dim Y_h}, \partial_w R_h \in \mathbb{R}^{\dim V_h \times \dim V_h}, \) and \( \partial_{\tau_h} R_h \in \mathbb{R}^{\dim V_h \times \dim Y_h} \) be the matrices which represent the derivatives \( D\tilde{u}_h(\tau_h), \partial_w R_h(\tilde{u}_h(\tau_h), \tau_h), \) and \( \partial_{\tau_h} R_h(\tilde{u}_h(\tau_h), \tau_h) \), respectively. Then, equation (13) holds true if and only if
\[
D\Phi_h y = D\tilde{I}_h D\tilde{u}_h y \quad \forall \ y \in \mathbb{R}^{\dim Y_h}. \tag{21}
\]
Relation (15) is equivalent to
\[ \mathbf{v}^T \partial_u R_h D \tilde{u}_h \mathbf{y} = - \mathbf{v}^T \partial_r R_h \mathbf{y} \quad \forall \mathbf{v} \in \mathbb{R}^{\dim \mathbf{V}_h}. \] (18)
The goal of the adjoint approach consists in reformulating the right-hand side of (17). To this end, choose \( \mathbf{v} \) in (18) such that \( \mathbf{v}^T \partial_u R_h = DI_h \), i.e., \( \mathbf{v} = (\partial_u R_h)^{-T} D I_h^T \), which is the algebraic version of (14). Inserting this \( \mathbf{v} \) into (17) and using (18) gives us
\[ D \Phi_h \mathbf{y} = \psi^T \partial_u R_h D \tilde{u}_h \mathbf{y} = - \psi^T \partial_r R_h \mathbf{y} \quad \forall \mathbf{y} \in \mathbb{R}^{\dim \mathbf{Y}_h}, \]
which is equivalent to \( D \Phi_h = - \psi^T \partial_r R_h \) and corresponding to (16).

**Results**

For the SUPG method (5), we have
\[ \langle (\partial_u R_h)(w_h, \tau_h), v_h \rangle = a(\tilde{u}_h, v_h) + s_h(\tau_h; \tilde{u}_h, v_h), \]
\[ \langle (\partial_r R_h)(w_h, \tau_h), v_h \rangle = s_h(\tau_h; w_h + \tilde{u}_h, v_h) - r_h(\tilde{u}_h; v_h) \]
for any \( \tau_h, \tilde{u}_h \in \mathbf{Y}_h \) and \( v_h, \tilde{u}_h, w_h \in \mathbf{V}_h \). Thus, for any \( \tau_h \in \mathbf{D}_h \), the auxiliary function \( \psi_h(\tau_h) \in \mathbf{V}_h \) is the solution of the equation
\[ a(v_h, \psi_h(\tau_h)) + s_h(\tau_h; v_h, \psi_h(\tau_h)) = \langle DI_h(\tilde{u}_h(\tau_h)), v_h \rangle \quad \forall v_h \in \mathbf{V}_h \]
and the Fréchet derivative of \( \Phi_h \) is given by
\[ \langle D \Phi_h(\tau_h), \tilde{u}_h \rangle = -s_h(\tau_h; w_h(\tau_h), \psi_h(\tau_h)) + r_h(\tilde{u}_h; \psi_h(\tau_h)) \quad \forall \tilde{u}_h \in \mathbf{Y}_h. \]

In this paper we consider three different spaces \( \mathbf{Y}_h \) of parameter \( \tau_h \). In addition to the classical choice of \( P_0(K) \) for all \( K \in \mathbf{T}_h \), which is the space of piecewise constant functions, we consider \( P_1(\Omega) \), the space of continuous piecewise linear functions, and \( P_1(K) \) for all \( K \in \mathbf{T}_h \), the space of discontinuous piecewise linear functions.

The parameter \( \tau_h \) is optimized by the L-BFGS nonlinear minimization method described in *Nocedal and Wright* [2006] using exactly the setup from Lukáš [2011]. The method starts with the values given by (6). The reason why we choose the L-BFGS method among all other nonlinear minimization methods is that it performs well in comparison with other methods, see Lukáš [2011].

As an example, let us consider the convection-diffusion equation (1) in \( \Omega = (0, 1)^2 \) with \( \varepsilon = 10^{-8} \), \( \mathbf{b} = (-y, x)^T \), \( c = 0 \), \( f = 0 \), Neumann condition \( \partial \mathbf{u} / \partial n = 0 \) on \( x = 0 \), and \( u_h = \begin{cases} 1 & \text{if } \frac{1}{3} \leq x \leq \frac{2}{3} \text{ and } y = 0, \\ 0 & \text{otherwise, if } x \neq 0. \end{cases} \) (19)

We use linear conforming (\( P_1 \)) type of finite elements. This defines the space \( \mathbf{W}_h \) and together with (19) also the space \( \mathbf{V}_h \). The solution possesses two interior characteristic layers in the direction of the convection starting at \( (\frac{1}{3}, 0) \) and \( (\frac{2}{3}, 0) \). We provide results on an unstructured mesh with 885 elements. The Péclet number from (6) in this example is in the interval from \( 8 \cdot 10^5 \) to \( 2 \cdot 10^6 \). This example was used by Knopp, Lube, and Rapin [2002]. We use Indicator (8) in the example.

The Lagrange interpolation in \( \mathbf{W}_h \) of exact solution \( u(x, y) \) of this problem is depicted in Figure 1a. In Figure 1b we can see the solution of the SUPG method (5). In Figure 1c we show the discrete solution with piecewise linear continuous parameter \( \tau_h \) (\( \tau_h \in P_1(\Omega) \)). The discrete solutions with piecewise constant and piecewise linear discontinuous parameter \( \tau_h \) are almost the same. Therefore, we provide only the figure of discrete solution with piecewise linear discontinuous parameter \( \tau_h \) (\( \tau_h \in P_1(K) \)) for all \( K \in \mathbf{T}_h \), see Figure 1d.

In Figure 1e there is the grayscale graph of values of the parameter \( \tau_h \) from the SUPG method (5), \( \tau_h \in P_0(K) \) for all \( K \in \mathbf{T}_h \), which is also the initial setup for the L-BFGS minimization method. The fact that the results with piecewise linear discontinuous and piecewise constant parameters are very similar can be seen also in Figure 1f and 1g, where the values of parameter \( \tau_h \) for these spaces are shown. We provide in Figure 1f also the values of \( \tau_h \) after the minimization process with piecewise linear continuous parameter \( \tau_h \).

In Figure 1i are depicted the cross-sections of resulting discrete solutions for all spaces of parameter \( \tau_h \). In Figure 1j, we can see the fact that even the minimization speed is almost equivalent for piecewise linear discontinuous and piecewise constant parameters \( \tau_h \).
Figure 1. In (a), (b), (c), and (d), there is a comparison of different discrete solutions. In (e), (f), (g), and (h), there are grayscale maps of values of $\tau_h$ for all spaces of $\tau_h$. In (i) there are the cross-sections at $x = 0$. (j) Minimizing process according to Indicator (7) for all spaces of $\tau_h$, on the vertical axis there is the $L^2$ residue according to Indicator (7) and on the horizontal axis there is time in seconds.
Conclusions

From the numerical tests we have done so far it comes out that for piecewise linear continuous ($P_1$) finite elements using the piecewise linear discontinuous space of functions for parameter $\tau_h$ has no essential improvement effect over using the standard piecewise constant space of functions for parameter $\tau_h$. This holds also for other indicators we used in the adaptive method.

Our future interest is to implement and test higher order finite elements for the space $W_h$ and higher order spaces for the parameter $\tau_h$.

References


