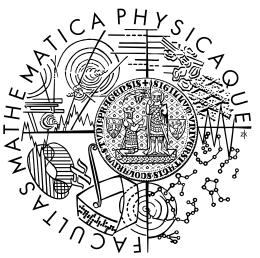


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Finite Element Techniques for Convection–Diffusion Problems

doc. Mgr. Petr Knobloch, Dr.

A thesis submitted to the Czech Academy of Sciences
in partial fulfilment of the requirements for the scientific degree of
Doctor Scientiarum

Prague, January 2017

To Ikar

Preface

Convection and diffusion are basic physical mechanisms which influence or even determine many various processes in the nature, science and technology. A classical example is the distribution of the temperature or the concentration of a substance, e.g., a pollutant. Mathematical models describing processes envolving convective and diffusive effects are usually too complicated to be solved analytically. Therefore, it is necessary to approximate the respective unknown quantities by means of numerical methods. However, in typical applications where convection dominates diffusion, standard numerical techniques fail since the approximate solutions are usually polluted by spurious oscillations. This is connected with the fact that the solutions of convection-dominated problems typically contain so-called layers, which are narrow regions where the solution changes abruptly.

To understand the origins of various undesirable effects that are encountered when convection-dominated problems are solved numerically, it is reasonable to study simplified model problems. The simplest possible model problem is the scalar convection–diffusion equation which describes just the convection and diffusion. Often also a reaction term is added which may be needed in some applications and, moreover, often simplifies mathematical considerations. Investigations of numerical techniques for this model problem are crucial for a successful development of accurate, robust and efficient approaches for the numerical solution of more complicated problems arising in applications. In addition, the convection–diffusion–reaction equation itself is often used (alone or as a part of many mathematical models) for computing distributions of various physical quantities. Therefore the development of numerical methods for convection–diffusion–reaction equations is important also at its own.

Numerical techniques for convection–diffusion problems have been intensively developed and studied for more than four decades, but despite the huge amount of the literature on this topic, one has to state that the numerical solution of a convection-dominated scalar convection–diffusion equation is still a challenge in general. Although a considerable progress has been made and successful techniques were designed for particular problems, there is still no efficient and accurate numerical method which would be successfully applicable at least to a sufficiently large set of test problems.

This thesis represents a collection of my selected publications and reflects my research on finite element techniques for convection–diffusion problems during the past twelve years. It consists of six chapters. The first chapter contains an introduction to the field of numerical methods for convection-dominated problems and

comments on the publications collected in the remaining five chapters. The second chapter is devoted to improvements of the Mizukami–Hughes method which is a nonlinear method of upwind type satisfying the discrete maximum principle. The third chapter contains a review and analysis of spurious oscillations at layers diminishing (SOLD) methods that are again mostly nonlinear. Although it may be surprising that nonlinear techniques are applied to the solution of a linear convection–diffusion equation, it seems that this is unavoidable for obtaininig accurate numerical results on relatively coarse meshes. Since problems in applications are often nonlinear, the nonlinearity of the considered techniques is of minor importance. The fourth chapter contains a novel technique for choosing the stabilization parameter in the SUPG method and proposes an adaptive approach for choosing stabilization parameters in both linear and nonlinear discretizations. The fifth chapter is devoted to the local projection stabilization that may be viewed as a simplification of the SUPG method. In particular, it is shown that the Galerkin finite element method is more stable than expected and the local projection stabilization then stabilizes just the unstable part of the Galerkin solution. This chapter also introduces a new variant of the local projection stabilization and analyzes nonlinear stabilizations defined using local projections. Whereas all the approaches mentioned so far are based on variational formulations, the last chapter studies algebraic flux correction schemes which are approaches modifying the discrete problem on the algebraic level. The publications in this chapter contain the first rigorous analysis of these techniques.

Many of the results contained in this thesis would not have appeared without the support of several grant agencies that enabled me a collaboration and exchange of ideas with my colleagues abroad. For this I would like to thank the Czech Science Foundation (grants No. 201/07/J033, 201/08/0012, P201/11/1304, P201/13/00522S, and 16-03230S), the Grant Agency of the Czech Academy of Science (grants No. IAA100190505 and IAA100190804), the Grant Agency of the Charles University (grant No. 344/2005/B–MAT/MFF) and the Ministry of Education, Youth and Sports of the Czech Republic (project MSM 0021620839).

Prague, January 2017 Petr Knobloch

Contents

1 Comments on the collection of publications **1**
1.1 Introduction 1
1.2 Mizukami–Hughes method (comments on Chapter 2) 4
1.3 SOLD methods (comments on Chapter 3) 5
1.4 Choice of stabilization parameters (comments on Chapter 4) 7
1.5 Local projection stabilization (comments on Chapter 5) 9
1.6 Algebraic flux correction (comments on Chapter 6) 12
1.7 Concluding remarks 14
References 15

2 Mizukami–Hughes method **21**
2.1 Improvements of the Mizukami–Hughes method for convection–diffusion equations 23
2.2 Numerical solution of convection–diffusion equations using a nonlinear method of upwind type 39

3 SOLD methods **57**
3.1 On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review 59
3.2 On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements 79
3.3 On the performance of SOLD methods for convection–diffusion problems with interior layers 97

4 Choice of stabilization parameters **111**
4.1 On the choice of the SUPG parameter at outflow boundary layers . . 113
4.2 On the definition of the SUPG parameter 134
4.3 A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I 148
4.4 Adaptive computation of parameters in stabilized methods for convection–diffusion problems 163

5 Local projection stabilization **173**
5.1 On the stability of finite-element discretizations of convection–diffusion–reaction equations 175

5.2 Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach 193
5.3 A generalization of the local projection stabilization for convection–diffusion–reaction equations 211
5.4 A local projection stabilization finite element method with nonlinear crosswind diffusion for convection–diffusion–reaction equations 233

6 Algebraic flux correction 265
6.1 Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension 267
6.2 Analysis of algebraic flux correction schemes 295
6.3 An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes 321

Chapter 1

Comments on the collection of publications

In this chapter, first a brief introduction to the field of numerical methods for convection-dominated problems is presented and then comments on the publications collected in this work are given. Each of the following chapters is covered by a separate section. The chapter is finished by concluding remarks and references.

1.1 Introduction

The distribution of physical quantities in various physical, technical, biological and other processes is driven by basic physical mechanisms which are diffusion, convection, and reaction. Often, the diffusion is very small in comparison with the convection or reaction. This causes that the distribution of the respective quantity comprises so-called layers, which are narrow regions where the quantity changes abruptly. It is well known that standard discretizations then provide approximate solutions polluted by spurious oscillations unless the underlying mesh resolves the layers, see, e.g., the monograph [66]. Consequently, special discretization techniques (so-called stabilized methods) have to be applied which always introduce a certain amount of artificial diffusion that should suppress the spurious oscillations but also typically increases the smearing of the layers. Therefore, it is usually still a challenge to obtain an accurate approximate solution, despite the huge amount of research on appropriate discretizations during the last four decades.

The simplest model for the above-mentioned class of problems is a scalar steady-state convection–diffusion–reaction equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = f \quad \text{in } \Omega\,, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{1,2,3\}$, is a bounded domain, $\varepsilon > 0$ is a constant diffusion parameter, $\boldsymbol{b}$ is a convection field, c is a reaction coefficient, and f is a term describing sources and sinks. The unknown function u represents, e.g., the temperature in modeling the energy balance, or the concentration or mass fraction in modeling mass balances. To obtain a well-posed problem, (1) has to be equipped with appropriate boundary conditions.

To solve the equation (1) numerically, various methods can be applied: the finite difference method, finite volume method, finite element method, discontinuous Galerkin method, or spectral method, to name the most common ones. For each of these methods, many contributions on its application to the numerical solution of (1) can be found in the literature. This work is devoted exclusively to the application of the finite element method which we prefer because of its flexibility in treating complex geometries, easy incorporation of natural boundary conditions and suitability for theoretical investigations due to its functional analytical setting based usually on Hilbert spaces.

It should be emphasized that the model (1) as a stand-alone equation is considered because, on the one hand, it comprises the effects of diffusion, convection, and reaction and, on the other hand, it simplifies the analysis of numerical techniques for its solution. Nevertheless, also for this simplest available model, there are many discretizations the analysis of which still remains an open problem. In applications, equations of type (1) are often a part of complex systems of equations. For example, they may be coupled with the Navier–Stokes equations describing the convection $\boldsymbol{b}$ which is, in turn, influenced by the temperature or concentrations determined by equations of type (1).

This work is mainly devoted to studies of numerical techniques for the equation (1) in the convection-dominated regime characterized by the conditions $\varepsilon \ll \|\boldsymbol{b}\|_{L^\infty}$ and $\|c\|_{L^\infty} \lesssim \|\boldsymbol{b}\|_{L^\infty}$, which is the case usually encountered in applications. As already mentioned, the main feature of solutions in this regime is the appearance of layers, i.e., narrow regions with large gradients of the solution. Then the standard Galerkin finite element method applied to (1), which corresponds to central finite differencing for constant data and suitable meshes, leads to heavily oscillating solutions unless the layers are resolved by the respective mesh. Therefore, much research has been devoted to the development of numerical methods using anisotropic layer-adapted meshes. Such meshes can be defined either a priori (see, e.g., [59, 66]) or a posteriori by means of adaptive techniques (see, e.g., [1, 67]). Nevertheless, since the layer width is proportional to $\sqrt{\varepsilon}$ or even ε (depending on the type of layer), the geometric resolution of the layers is often not feasible due to high memory and computational time requirements. Therefore, it is important to develop numerical methods providing sufficiently accurate results also on meshes which are coarse in comparison with the width of the layers. This is the main aim of this work.

To suppress the oscillations present in Galerkin solutions obtained on coarse meshes, various stabilized methods have been developed, see, e.g., [66, 65, 39] for reviews. The stabilizing effect of these approaches can be characterized by the artificial diffusion they add to the underlying Galerkin discretization. To diminish the spurious oscillations to a sufficient extent, the artificial diffusion has to be sufficiently large. However, to avoid an excessive smearing of the layers, the artificial diffusion is not allowed to be too large. Consequently, the design of a proper stabilization is very difficult. Despite more than four decades of research, there is so far no efficient discretization for (1) available which would produce accurate numerical solutions (in particular, with sharp layers at correct positions) without unphysical features (e.g.,

negative concentrations). This statement is supported by theoretical and numerical studies in, e.g., [3, 32, 34, 37, 38].

One of the most successful linear stabilizations is the streamline upwind Petrov–Galerkin (SUPG) finite element method [28, 16] which consistently introduces artificial diffusion along streamlines. It combines good stability properties with a high accuracy away from layers. Because this method will be frequently discussed throughout this work, it will be now formulated for the equation (1) in detail.

At this point, one has to specify the boundary conditions for u on the boundary $\partial\Omega$ of Ω. To simplify the presentation in this chapter, we shall consider

$$u = 0 \quad \text{on } \partial\Omega\,. \tag{2}$$

More general boundary conditions can be found in the publications contained in the following chapters. The Galerkin finite element discretization of the problem (1), (2) defines an approximate solution u_h from a finite element space V_h approximating the Sobolev space $H_0^1(\Omega)$ as the solution of the variational problem

$$a(u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,, \tag{3}$$

where

$$a(u_h, v_h) = \varepsilon\,(\nabla u_h, \nabla v_h) + (\boldsymbol{b}\cdot\nabla u_h, v_h) + (c\,u_h, v_h)$$

and $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. The SUPG method adds a weighted residual of (1) to the Galerkin method and defines the approximate solution $u_h \in V_h$ by

$$a(u_h, v_h) + \sum_{T\in\mathscr{T}_h} (-\varepsilon\,\Delta u_h + \boldsymbol{b}\cdot\nabla u_h + c\,u_h - f, \tau\,\boldsymbol{b}\cdot\nabla v_h)_T = (f, v_h) \quad \forall\ v_h \in V_h\,, \tag{4}$$

where $\mathscr{T}_h$ is a triangulation of Ω used for defining the finite element space V_h, τ is a nonnegative stabilization parameter (typically constant on each $T \in \mathscr{T}_h$) and $(\cdot,\cdot)_T$ denotes the inner product in $L^2(T)$ or $L^2(T)^d$. The additional term is written as a sum of local contributions since the operator Δ usually cannot be applied to u_h globally. The parameter τ determines the amount of the artificial diffusion added by the SUPG method to the Galerkin discretization. For linear or bilinear finite elements, it is often defined, on any element $T \in \mathscr{T}_h$, by the formula

$$\tau|_T = \frac{h_T}{2\,|\boldsymbol{b}|}\left(\coth Pe_T - \frac{1}{Pe_T}\right) \quad \text{with} \quad Pe_T = \frac{|\boldsymbol{b}|\,h_T}{2\,\varepsilon}\,, \tag{5}$$

which originates from the one-dimensional case. The notation Pe_T is used for the Péclet number, which determines whether the problem is locally convection-dominated or diffusion-dominated, and h_T is the element diameter in the direction of the convection vector $\boldsymbol{b}$. Throughout this chapter, we shall assume that $\sigma := c - \frac{1}{2}\operatorname{div}\boldsymbol{b} \geq 0$. Then, if τ satisfies suitable assumptions, one can prove the stability and an error estimate for (4) with respect to the norm

$$\|v\|_{SUPG} = \left(\varepsilon\,|v|_{1,\Omega}^2 + \|\sigma^{1/2}\,v\|_{0,\Omega}^2 + \|\tau^{1/2}\,\boldsymbol{b}\cdot\nabla v\|_{0,\Omega}^2\right)^{1/2}, \tag{6}$$

where $|\cdot|_{1,\Omega}$ is the usual seminorm in $H_0^1(\Omega)$ and $\|\cdot\|_{0,\Omega}$ is the norm in $L^2(\Omega)$. The SUPG method represents a significant improvement in comparison with the Galerkin method, nevertheless, since it is not a monotone method, it may compute solutions suffering from spurious oscillations in layer regions.

1.2 Mizukami–Hughes method (comments on Chapter 2)

The Mizukami–Hughes method is an interesting approach proposed in [64] for a two-dimensional convection–diffusion equation (i.e., (1) with $c = 0$ and $d = 2$) discretized using a finite element space V_h consisting of continuous piecewise linear functions over a triangular mesh. To formulate the method, we denote by $\varphi_1, \ldots, \varphi_M$ the standard piecewise linear basis functions of the space V_h. Then the Galerkin discretization (3) can be written in the form

$$\varepsilon\,(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}\cdot\nabla u_h, \varphi_i) = (f, \varphi_i)\,, \qquad i = 1, \ldots, M\,.$$

The Mizukami–Hughes method replaces the test functions φ_i by functions $\widetilde{\varphi}_i$ obtained by adding suitable constants to φ_i on the triangles forming its support. Then the approximate solution $u_h \in V_h$ is defined by

$$\varepsilon\,(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}\cdot\nabla u_h, \widetilde{\varphi}_i) = (f, \widetilde{\varphi}_i)\,, \qquad i = 1, \ldots, M\,.$$

Thus, it is a Petrov–Galerkin method like the SUPG method. It is assumed that $\boldsymbol{b}$ is constant on each element of the triangulation; in practice, $\boldsymbol{b}$ is replaced by a piecewise constant approximation.

The idea of the Mizukami–Hughes method is to define the constants in the definition of the test functions $\widetilde{\varphi}_i$ in such a way that the local finite element matrices corresponding to the convective term are of nonnegative type, i.e., their row sums are nonnegative and off-diagonal entries are nonpositive. Whether this is possible depends on the orientation of $\boldsymbol{b}$ with respect to the given element of the triangulation. However, Mizukami and Hughes made the important observation that u still solves the equation (1) if one replaces $\boldsymbol{b}$ by any function $\tilde{\boldsymbol{b}}$ such that $\tilde{\boldsymbol{b}} - \boldsymbol{b}$ is orthogonal to ∇u. This suggests to define the constants in the definition of the functions $\widetilde{\varphi}_i$ in such a way that the local convection matrix is of nonnegative type for $\boldsymbol{b}$ replaced by a suitable function $\tilde{\boldsymbol{b}}$, which is always possible. Since ∇u is not known a priori, one obtains a nonlinear problem where the constants in the definition of $\widetilde{\varphi}_i$ depend on the unknown approximate solution u_h.

The Mizukami–Hughes method is probably the first nonlinear method for (1) satisfying the discrete maximum principle. Like for many other methods proposed later, see, e.g., [18, 19, 4], this property is proved only for weakly acute meshes, i.e., the magnitude of all angles in the triangles of the mesh is less than or equal to $\pi/2$. Nevertheless, it is also possible to derive methods for which the discrete maximum principle holds on arbitrary meshes, see Section 1.6. The discrete maximum principle is an important property which ensures that no spurious oscillations will appear, not even in the vicinity of sharp layers. In contrast to many methods satisfying the discrete maximum principle, the Mizukami–Hughes method does not lead to a pronounced smearing of layers and it often provides very accurate results.

However, we observed that, in some cases, the Mizukami–Hughes method does not lead to correct solutions. Moreover, sometimes it is very difficult to solve the nonlinear problem with a prescribed accuracy. Therefore, in [40] (pp. 23–38 in this

work), we proposed several improvements of the method which correct the mentioned shortcomings and keep its quality in cases in which it works well. This was achieved by a more careful definition of the constants in the test functions $\widetilde{\varphi}_i$. In particular, a continuous dependence of these constants on the orientation of $\boldsymbol{b}$ and ∇u_h was introduced. Moreover, the method was extended to convection–diffusion–reaction equations and to the three-dimensional case. It was shown that the improved method still satisfies the discrete maximum principle and its high accuracy was demonstrated by many numerical results. The superiority of the improved Mizukami–Hughes method to linear upwinding finite element methods satisfying the discrete maximum principle was clearly demonstrated in [41].

Both the Mizukami–Hughes method in [64] and its improved variant in [40] were designed for the strongly convection-dominated case $\varepsilon \ll |\boldsymbol{b}|$. In [49] (pp. 39–55 in this work), the method was extended to the whole range of the diffusion parameter and it was proved that the extended method satisfies the discrete maximum principle. The favourable properties of the new method were illustrated by means of numerical experiments.

A drawback of both the original and the improved versions of the Mizukami–Hughes method is that no existence, uniqueness and convergence results are available. Moreover, it seems to be rather difficult to generalize the method to more complicated problems or to other types of finite elements. So far, only a variant for bilinear finite elements is available, see [42].

1.3 SOLD methods (comments on Chapter 3)

The SUPG method formulated at the end of Section 1.1 (like many other approaches adding a linear stabilization term to the Galerkin discretization, see, e.g., [17, 22, 27, 63]) significantly reduces the spurious oscillations present in Galerkin solutions but does not preclude small over- and undershoots in the vicinity of layers. Although the remaining nonphysical oscillations are often small in magnitude, they are not permissible in many applications. An example are chemically reacting flows where it is essential to guarantee that the concentrations of all species are nonnegative. Another example are free-convection computations where temperature oscillations create spurious sources and sinks of momentum that effect the computation of the flow field. The small spurious oscillations may also deteriorate the solution of nonlinear problems, e.g., in two-equations turbulence models or in numerical simulations of compressible flow problems, where the solution may develop discontinuities (shocks) whose poor resolution may effect the global stability of the numerical calculations.

The above-mentioned spurious oscillations in SUPG solutions indicate that using the streamlines as upwind direction is not always sufficient. Therefore, as a remedy, various nonlinear 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, at least in some model cases, or which at least reduces the local oscillations. This procedure is often referred to as discontinuity capturing or shock capturing, nevertheless, we prefer to call these methods *spurious oscillations*

at layers diminishing (SOLD) methods, which we regard as more apposite.

It may be surprising that nonlinear methods are applied to the numerical solution of the linear equation (1). However, for the limit $\varepsilon = 0$, the famous Godunov theorem [25] states that a linear monotone discretization is at most of first order convergence so that applying linear methods limits the accuracy if one insists on the monotonicity. We are not aware of an analogous mathematical theorem for $\varepsilon > 0$, but numerical experience suggests that the situation is similar for the case of small ε.

A typical SOLD term added to the left-hand side of (4) is of the form

$$(\widetilde{\varepsilon}(u_h)\,\nabla u_h, \nabla v_h) \qquad \text{or} \qquad (\widetilde{\varepsilon}(u_h)\,D\,\nabla u_h, \nabla v_h)\,, \tag{7}$$

where $\widetilde{\varepsilon}(u_h)$ is a nonnegative solution-dependent artificial diffusion parameter and D is the projection onto the line or plane orthogonal to $\boldsymbol{b}$. Thus, the first term in (7) introduces an isotropic artificial diffusion whereas the second one adds a crosswind artificial diffusion. An example of $\widetilde{\varepsilon}(u_h)$ is a modification of the artificial diffusion parameter by Codina [21] proposed in [32], which is given by

$$\widetilde{\varepsilon}(u_h)|_T = \max\left\{0, \eta\,\frac{\operatorname{diam}(T)\,|R_h(u_h)|}{2\,|\nabla u_h|} - \varepsilon\right\} \tag{8}$$

on any element T of the triangulation. Here, $\operatorname{diam}(T)$ is the diameter of T, $\eta > 0$ is a user-chosen parameter (e.g., $\eta = 0.7$ for linear finite elements) and

$$R_h(u_h) = -\varepsilon\,\Delta u_h + \boldsymbol{b}\cdot\nabla u_h + c\,u_h - f$$

is the residual.

The literature on SOLD methods is rather extended but the various numerical tests published in the literature do not allow to draw a clear conclusion concerning their advantages and drawbacks. Therefore, in [32] (pp. 59–77 in this work), we presented a review of the most published SOLD methods, discussed the motivations of their derivation, proposed some alternative choices of parameters and classified them. The review was followed by a numerical comparison of the considered SOLD methods at two test problems whose solutions possess characteristic features of solutions of (1). The numerical results gave a first systematic insight into the behaviour of the SOLD methods and showed that the Mizukami–Hughes method was always the best method if the nonlinear iterations converged. Among the other SOLD methods, no one could be preferred in all cases but several methods were identified that should not be applied.

The studies in [32] were followed by a second part published in [34] (pp. 79–96 in this work) where the most promising SOLD methods from the first part were investigated in more detail for linear and bilinear finite elements. Analytical and numerical studies showed that SOLD methods without user-chosen parameters are in general not able to remove the spurious oscillations of the solution obtained with the SUPG discretization. For methods with a free parameter, like the one in (7), (8), values of the parameter could be derived in two examples such that the spurious oscillations were almost removed. It turned out that a spatially constant choice of the parameters was not sufficient in general and that the optimal parameters depended on the data of the problem and on the mesh. In addition, an example was

presented for which none of the investigated methods provided a qualitatively correct approximate solution. The iterative solution of the nonlinear discrete problems was also studied. It was shown that the number of iterations or the convergence of the iterative process depend again on the problem, the mesh and the parameters of the SOLD methods. It could be observed that the convergence is often strongly influenced by the choice of an appropriate damping factor and a strategy was proposed for an automatic and dynamic computation of this factor. The studies in this paper revealed that it is in general completely open how to obtain oscillation-free solutions using the considered classes of methods.

The papers [32, 34] were supplemented by numerical studies for a convection–diffusion problem with a nonconstant convection field whose solution possesses interior layers in [33] (pp. 97–110 in this work). This setting is closer to problems one encounters in applications than the test problems considered in the two previous publications. The conclusions were similar as in [34]. Further comparisons of various SOLD methods can be found in [30, 31].

1.4 Choice of stabilization parameters (comments on Chapter 4)

The studies summarized in Section 1.3 showed that it is in general not clear how to design SOLD methods which would suppress the spurious oscillations present in SUPG solutions to a sufficient extent (without smearing the layers considerably). One possibility how to circumvent this problem is to try to improve the definition of the SUPG stabilization parameter. The formula (5) leads to nodally exact solutions in the one-dimensional case under simplifying assumptions, but in two and three dimensions it is not optimal in general. The choice of the stabilization parameter at characteristic layers has only a limited influence on the spurious oscillations appearing in these regions (cf., e.g., [61]), but there is a hope of improvement at outflow boundary layers.

One possibility how to define the SUPG stabilization parameter at outflow boundary layers was proposed in [46] (pp. 113–133 in this work) for linear triangular finite elements. To present this definition, let us first denote by $G_h \subset \Omega$ the set consisting of triangles intersecting the outflow boundary of Ω (i.e., the part of $\partial\Omega$ where the product of $\boldsymbol{b}$ and the outward normal vector to $\partial\Omega$ is positive). Then, by analogy to (5), the parameter τ is defined, on any triangle $T \subset G_h$, by

$$\tau|_T = \tau_0|_T \left(\coth Pe_T - \frac{1}{Pe_T} \right) \quad \text{with} \quad Pe_T = \frac{|\boldsymbol{b}_T|\, h_T}{2\,\varepsilon}\,,$$

where τ_0 is a piecewise constant function satisfying

$$\int_{G_h} \varphi_i + \tau_0\, \boldsymbol{b} \cdot \nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0, \qquad i = 1, \ldots, M\,, \tag{9}$$

$\boldsymbol{b}_T$ is the mean value of $\boldsymbol{b}$ on T, and φ_i are the same basis functions of V_h as in Section 1.2. On triangles $T \not\subset G_h$, the parameter τ is defined by (5) with $\boldsymbol{b}$ replaced

by $\boldsymbol{b}_T$. It was shown in [46] that a piecewise constant function τ_0 satisfying (9) exists and an algorithm how to construct it was given. Numerical results in [46] demonstrate a significant reduction of spurious oscillations in approximate solutions in comparison to the standard choice of τ given by (5) while accuracy away from layers is preserved. For simple model problems, even nodally exact solutions are obtained. Whereas all definitions of stabilization parameters published in the literature so far were based on local information on a given element of the triangulation, the results of [46] show that this local information is not sufficient for obtaining oscillation–free SUPG solutions in general.

The choice of the SUPG stabilization parameter τ introduced in [46] was further discussed in [43] (pp. 134–147 in this work). It was demonstrated that a combination of this choice of τ and the variant of the SOLD method (7), (8) adding crosswind artificial diffusion provides fairly satisfactory approximations of solutions to (1). The results of [43] also show that it is essential to define both the parameter τ and the mesh in such a way that the spurious oscillations in the SUPG solution are as small as possible. Otherwise the addition of a SOLD term cannot be expected to lead to an oscillation-free solution. Numerical tests in [43] illustrate how small modifications of the mesh may significantly improve the quality of SUPG solutions.

The above discussion revealed that a basic problem of most of the stabilized methods is the design of appropriate stabilization parameters which would lead to sufficiently small nonphysical oscillations without compromising accuracy. As it follows from the publications discussed in the preceding section, 'optimal' parameters depend on the data of the problem and the used mesh in a complicated way so that, in general, one cannot expect to be able to define them a priori. Therefore, in [36] (pp. 148–161 in this work), we proposed to compute the stabilization parameters a posteriori by minimizing a target functional characterizing the quality of the approximate solution. This is a nonlinear constraint optimization problem that has to be solved iteratively. A key component of this approach consists in the efficient computation of the Fréchet derivative of the functional with respect to the stabilization parameter. This was achieved by utilizing an adjoint problem with an appropriate right-hand side, which led to a new general framework for the optimization of parameters in stabilized methods for convection–diffusion equations. Benefits of this approach were demonstrated on its application to the optimization of a piecewise constant parameter τ in the SUPG method.

In [35] (pp. 163–171 in this work), the methodology proposed in [36] was applied to the optimization of the parameters in a SOLD method. Since one of the most promising approaches among the SOLD methods seems to be the modified method of Codina (7), (8), we considered the SUPG method enriched by the crosswind artificial diffusion term from (7) with

$$\widetilde{\varepsilon}(u_h)|_T = \eta\,\frac{\operatorname{diam}(T)\,|R_h(u_h)|}{2\,|\nabla u_h|} \qquad \forall\, T \in \mathscr{T}_h\,.$$

Both the parameters τ and η were optimized as piecewise constant functions. In this way very accurate numerical results with steep layers and negligible spurious oscillations could be obtained. The only drawback of this approach is the increased computational cost connected with the solution of the optimization problem.

1.5 Local projection stabilization (comments on Chapter 5)

The enhanced stability of the SUPG method (4) in comparison with the Galerkin method (3) originates from the term $(\boldsymbol{b}\cdot\nabla u_h, \tau\,\boldsymbol{b}\cdot\nabla v_h)$. For several reasons, which will be mentioned below, it would be convenient to consider only this term instead of the whole weighted residual stabilization term in (4). Then, however, the resulting method would not be consistent and the accuracy of the method would considerably deteriorate. A possible remedy is to consider only a small-scale part of $\boldsymbol{b}\cdot\nabla u_h$ defined using local projections into large-scale spaces. If the local projection spaces are chosen appropriately, the stability of the SUPG method is preserved without compromising the accuracy.

The local projection stabilization (LPS) was originally proposed in [10] as a technique for stabilizing discretizations of the Stokes problem in which both the pressure and the velocity components are approximated using the same finite element space. Later, the local projection method was extended to stabilization of convection dominated problems [11] and applied to various types of incompressible flow problems (see the review article [15]) and to convection–diffusion–reaction problems, see, e.g., [23, 45, 63]. To define a local projection stabilization of the Galerkin discretization (3), one introduces a second division $\mathscr{M}_h$ of Ω which typically consists of macroelements, i.e., unions of elements of $\mathscr{T}_h$. For each $M\in\mathscr{M}_h$, one introduces a finite dimensional space $D_M\subset L^2(M)$ and defines an orthogonal L^2 projection π_M of $L^2(M)$ onto D_M. It is assumed that there is a positive constant β independent of h such that

$$\sup_{v\in V_M}\frac{(v,q)_M}{\|v\|_{0,M}}\geq\beta\,\|q\|_{0,M}\qquad\forall\ q\in D_M,\ M\in\mathscr{M}_h\,, \tag{10}$$

where $V_M=\{v\in V_h;\ v=0 \text{ in } \Omega\setminus M\}$. This inf–sup condition is crucial for proving both optimal error estimates and improved stability results, cf. [62, 45, 47, 52]. Finally, it is convenient to introduce a constant approximation $\boldsymbol{b}_M$ of $\boldsymbol{b}$ on each set M. Then, denoting by $\kappa_M := id-\pi_M$ the so-called fluctuation operator (where id is the identity operator on $L^2(M)$), the local projection discretization of (1), (2) defines an approximate solution $u_h\in V_h$ satisfying

$$a(u_h,v_h)+s_h(u_h,v_h)=(f,v_h)\quad\forall\ v_h\in V_h$$

with

$$s_h(u_h,v_h)=\sum_{M\in\mathscr{M}_h}\tau_M\,(\kappa_M(\boldsymbol{b}_M\cdot\nabla u_h),\kappa_M(\boldsymbol{b}_M\cdot\nabla v_h))_M\,, \tag{11}$$

where τ_M is a nonnegative stabilization parameter. It is also possible to use the full gradient instead of $\boldsymbol{b}_M\cdot\nabla$ in the stabilization term, i.e.,

$$s_h(u_h,v_h)=\sum_{M\in\mathscr{M}_h}\tau_M\,(\kappa_M\nabla u_h,\kappa_M\nabla v_h)_M\,, \tag{12}$$

where κ_M is applied to the vector-valued functions componentwise. The parameter τ_M in (11) can be defined analogously as in the SUPG method (cf. (5)); the parameter in (12) should be additionally multiplied by $\|\boldsymbol{b}\|^2_{L^\infty(M)}$. Let us mention that a standard choice is to use $\boldsymbol{b}$ instead of $\boldsymbol{b}_M$ in (11). However, we demonstrated in [45] that then it is generally not possible to obtain optimal convergence results if τ_M scales with respect to the data like in (5).

The advantage of the LPS method compared to the SUPG method is that it does not require the costly computation of second order derivatives and can be easily applied to non-steady problems. Moreover, when applied to systems of partial differential equations, it is possible to avoid undesirable couplings between various components of the solution. A further advantage of these techniques is that they are symmetric. Therefore, if they are applied to optimization problems, the operations 'discretization' and 'optimization' commute [12, 14].

The action of the operator π_M onto a function can be interpreted as extracting its large-scale part. Then the fluctuation operator κ_M provides the small-scale part (fluctuations around the large-scale part). The LPS method can be also interpreted as a variational multiscale method where the influence of the unresolved scales is modeled by the stabilization term determined by the small scales.

A natural norm for the LPS method is given by

$$\|v\|_{LPS} = \left(\varepsilon\,|v|^2_{1,\Omega} + \|\sigma^{1/2}\,v\|^2_{0,\Omega} + s_h(v,v)\right)^{1/2}, \tag{13}$$

which is clearly weaker than the SUPG norm defined in (6). For a long time, it was not clear whether the LPS method is less stable than the SUPG method. The first contribution to clarifying this question was made in [52] (pp. 175–192 in this work), where it was shown that the LPS method is stable in the sense of an inf–sup condition with respect to the norm

$$|||v||| = \left(\varepsilon\,|v|^2_{1,\Omega} + \|\sigma^{1/2}\,v\|^2_{0,\Omega} + s_h(v,v) + \sum_{M\in\mathscr{M}_h} \delta_M\,\|\Pi_M(\mathbf{b}\cdot\nabla v)\|^2_{0,M}\right)^{1/2}, \tag{14}$$

where Π_M is the orthogonal L^2 projection of $L^2(M)$ onto V_M and δ_M is defined analogously as the SUPG parameter in (5). It was proved that, under certain simplifying assumptions, this norm can be bounded from below by a norm analogous to the SUPG norm, which implies, roughly speaking, that the LPS method is as stable as the SUPG method. For the stabilization term (11), the norm $|||\cdot|||$ could be bounded by an analogue of the SUPG norm also from above. The stability of the LPS method with respect to the norm (14) holds true also for $\tau_M = 0$, i.e., the results of [52] show that the Galerkin finite element method (3) is more stable than usually believed. It was demonstrated in [52] that this result implies that certain types of oscillating solutions are not allowed by the Galerkin method; basically, only a small-scale part of the Galerkin solution has to be stabilized – and this is exactly what the LPS method does.

Originally, the LPS method was designed as a two-level approach where the mesh $\mathscr{T}_h$ was obtained by a refinement of a triangulation $\mathscr{M}_h$ of Ω. A crucial property of these refinements is that they always create an additional vertex in the interior of any

refined element of $\mathscr{M}_h$. Later, in [62], the one-level approach was introduced where $\mathscr{M}_h = \mathscr{T}_h$ and the validity of the inf–sup condition (10) was assured by defining V_h as a finite element space enriched using higher-order polynomial bubble functions. In [51] (pp. 193–209 in this work), a critical comparison of the two approaches, both computational and analytical, was given, which showed that there are no convincing arguments for preferring one of these approaches.

A drawback of both variants of the LPS method is that they require more degrees of freedom than the SUPG method since the finite element space is either defined on a refined mesh or enriched by additional functions. Therefore, in [48] and [47] (pp. 211–232 in this work), we introduced a generalization of the LPS method which avoids these drawbacks by allowing to use overlapping macroelements. The error analysis for this generalized LPS method with respect to the norm (13) was presented in [48] for both stabilization terms (11) and (12). In [47], the results of [52] were improved in the sense that the stability of the LPS method defined using (11) with respect to the SUPG norm was shown without any simplifying assumptions. Another stability result with respect to the SUPG norm was established in [44] by defining the local projection operators using a weighted L^2 inner product.

Like the SUPG method, the LPS does not remove the spurious oscillations present in Galerkin solutions completely and some of them still remain in the vicinity of layers. Therefore, in [5] (pp. 233–264 in this work), we combined the LPS method defined using (11) with the SOLD term

$$\sum_{M \in \mathscr{M}_h} (\widetilde{\varepsilon}_M(u_h)\, \kappa_M(D_M \nabla u_h), \kappa_M(D_M \nabla v_h))_M \ ,$$

where

$$\widetilde{\varepsilon}_M(u_h) = \eta\, h_M\, |\boldsymbol{b}_M|\, |\kappa_M(D_M \nabla u_h)| \tag{15}$$

or

$$\widetilde{\varepsilon}_M(u_h) = \eta\, h_M\, |\boldsymbol{b}_M|\, \frac{h_M^{d/2}\, |\kappa_M(D_M \nabla u_h)|}{|u_h|_{1,M}} \,, \tag{16}$$

h_M is the diameter of M, $\eta > 0$ is a user-chosen parameter, and $D_M : \mathbb{R}^d \to \mathbb{R}^d$ is the projection onto the line or plane orthogonal to the vector $\boldsymbol{b}_M$ (cf. (7), (8)). In this paper, also the transient convection–diffusion–reaction equation

$$u_t - \varepsilon\, \Delta u + \boldsymbol{b} \cdot \nabla u + c\, u = f \quad \text{in } (0,T] \times \Omega$$

equipped with initial and boundary conditions was considered. The data $\boldsymbol{b}$, c, and f were assumed to vary on the time interval $[0,T]$. A one-step θ-scheme was applied as temporal discretization whereas the discretization with respect to the space variables was performed as in the steady-state case. For both the steady-state and transient cases, the solvability, uniqueness (for the variant (15) or a sufficiently small time step) and error estimates were proved. In the transient case, both the fully nonlinear scheme and a linearized variant were considered. Promising numerical results were also obtained for $\widetilde{\varepsilon}_M(u_h)$ defined by replacing the fraction in (16) by its square. The corresponding analysis for the steady-state and transient cases was performed in [6] and [50], respectively.

1.6 Algebraic flux correction (comments on Chapter 6)

As we have discussed in the previous sections, most of the methods developed for the numerical solution of convection-dominated problems either do not suppress spurious oscillations in layer regions sufficiently, or introduce too much artificial diffusion and lead to a pronounced smearing of layers. However, there is one class of methods that seems not to suffer from these two deficiencies: the algebraic flux correction (AFC) schemes. These schemes are designed to satisfy the discrete maximum principle by construction (so that spurious oscillations cannot appear) and provide sharp approximations of layers, cf. the numerical results in, e.g., [3, 26, 38, 55]. Like many of the schemes discussed above, the AFC schemes are nonlinear. A drawback of these schemes is that they have been applied successfully only for lowest order finite elements, which limits the accuracy of the computed solutions.

The basic philosophy of flux correction schemes was formulated already in the 1970s in [13, 68]. Later, the idea was applied in the finite element context, e.g., in [2, 60]. In the last fifteen years, these methods have been further intensively developed by Dmitri Kuzmin and his coworkers, see, e.g., [53, 54, 55, 56, 58]. Despite the attractiveness of AFC schemes, there was no rigorous numerical analysis for this class of methods for a long time. To the best of our knowledge, our results in [7, 8, 9] represent the first contributions in this direction.

In contrast to the methods discussed in the preceding sections, which are all based on variational formulations, the idea of the AFC schemes is to modify the algebraic system corresponding to a discrete problem. As this underlying discrete problem, we use the Galerkin discretization (3) with a finite element space V_h consisting of continuous piecewise linear functions with respect to a simplicial triangulation of Ω and assume that $\operatorname{div} \boldsymbol{b} = 0$ and $c \geq 0$. We shall formulate the AFC scheme in a form which can be used also with nonhomogeneous Dirichlet boundary conditions for u. To this end, we denote by $x_1, \ldots, x_M$ the interior vertices of $\mathscr{T}_h$ and by $x_{M+1}, \ldots, x_N$ the vertices of $\mathscr{T}_h$ lying on $\partial\Omega$. Then, a continuous piecewise linear approximate solution u_h can be represented by the vector $U \equiv (u_1, \ldots, u_N)$ of its values at the vertices $x_1, \ldots, x_N$, and the Galerkin discretization (3) can be equivalently written as a linear system

$$\sum_{j=1}^{N} a_{ij}\, u_j = f_i\,, \qquad i = 1, \ldots, M\,, \tag{17}$$

where the values $u_{M+1}, \ldots, u_N$ are determined by the Dirichlet boundary condition on $\partial\Omega$; in our case, they all vanish. Now, the matrix of (17) is extended to a matrix $(a_{ij})_{i,j=1}^{N}$ (typically, one uses the finite element matrix corresponding to the equation (1) with homogeneous Neumann boundary conditions) and one defines a symmetric artificial diffusion matrix $(d_{ij})_{i,j=1}^{N}$ with the entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\, i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j \neq i} d_{ij}\,.$$

Using this artificial diffusion matrix, the linear system (17) is rewritten in a form with a M-matrix on the left-hand side and a sum of antidiffusive fluxes on the right-hand side. Those of these fluxes that are responsible for a violation of the discrete maximum principle are limited using solution-dependent correction factors. In this way, the linear system (17) is replaced by the nonlinear problem

$$(18) \qquad \sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij}(U))\, d_{ij}\, (u_j - u_i) = f_i\,, \qquad i = 1, \ldots, M\,,$$

with $\alpha_{ij}(U) \in [0,1]$, $i,j = 1,\ldots,N$. The limiter functions α_{ij} are to be chosen in such a way that the AFC scheme (18) satisfies the discrete maximum principle.

In [7] (pp. 267–294 in this work), the AFC scheme (18) was investigated in the one-dimensional case for a limiter defined in [54]. In contrast to the common application of AFC schemes, it was not assumed that $\alpha_{ij} = \alpha_{ji}$, which may cause a lack of conservation. It was proved that the scheme satisfies a discrete maximum principle if a solution exists. However, examples were constructed which show that this scheme does not necessarily have a solution. A modification of the scheme was proposed for which the existence of a solution and a weak variant of the discrete maximum principle were proved.

In [8] (pp. 295–319 in this work), the AFC scheme (18) with limiters satisfying the symmetry condition $\alpha_{ij} = \alpha_{ji}$ was analyzed for general linear boundary value problems in any space dimension. Under a continuity assumption on the limiters, the existence of a solution was proved. As a consequence, the unique solvability of the linearized problem (18) (i.e., with α_{ij} independent of U) was obtained, which is useful for computing the solution of (18) numerically using a fixed-point iteration. Furthermore, the AFC scheme was formulated in a variational form and an abstract error estimate was derived. As usual for stabilized methods, the norm for which the error estimate is given contains a contribution from the flux correction term in (18). Then the abstract theory was applied to a discretization of the convection–diffusion–reaction equation (1) and an error estimate was derived. Numerical results in [8] show that, under the minimal assumptions on the limiters used in the analysis, the derived error estimate is sharp. Finally, for the limiter of [54], the AFC scheme (18) was proved to satisfy the discrete maximum principle on Delaunay meshes.

The limiter of [54] investigated in [7, 8] can be regarded as a standard limiter for steady-state problems. However, apart from the fact that is does not guarantee the discrete maximum principle on general meshes, its further drawback is that it is not linearity preserving in general. This property demands that the AFC term vanishes if the solution is a polynomial of degree 1 (at least locally). This restriction, which can be interpreted as a weak consistency requirement, is believed to lead to improved accuracy in regions where the solution is smooth. In fact, in previous works, linearity preservation was linked to good convergence properties for diffusion problems (see, e.g., [29, 57]). In addition, it has been observed in different works (see, e.g., [20] and, especially, the introduction in [24] for a discussion) that linearity preservation improves the quality of the approximate solution on distorted meshes.

The above considerations were a motivation for our recent publication [9] (pp. 321–344 in this work). Here we specified rather weak assumptions on the lim-

iters that are sufficient for proving the discrete maximum principle. Then a limiter was designed that fulfills these assumptions by modifying the algorithm proposed in [55]. The linearity preservation was assured by introducing an explicit geometric information about the mesh into the definition of the limiter. Numerical studies in [9] support the analytical results and indicate that the linearity preservation is important for an optimal convergence of the AFC scheme. To the best of our knowledge, the method presented in [9] is the first AFC scheme for a convection–diffusion–reaction equation that satisfies both the discrete maximum principle and linearity preservation on general simplicial meshes.

1.7 Concluding remarks

This work presents several contributions to the numerical solution of convection–diffusion problems made by the author during the past twelve years. The most important ones include:

- an improved version of the Mizukami–Hughes method (Chapter 2);
- a systematic comparison of SOLD methods (Chapter 3);
- a new definition of the SUPG stabilization parameter at outflow boundaries (Chapter 4);
- an adaptive choice of parameters in stabilized methods (Chapter 4);
- improved results on the stability of finite element discretizations (Chapter 5);
- a generalization of the local projection stabilization (Chapter 5);
- the first analysis of algebraic flux correction schemes (Chapter 6).

These results contributed to a better understanding of numerical techniques for the solution of convection-dominated problems and some of them were also applied to numerical simulations in the engineering literature.

The present work shows that there are still many open questions and a wide potential for improvement in the field of discretization techniques for convection–diffusion problems. In particular, the algebraic flux correction seems to be a promising approach which deserves deeper investigations and we plan to continue our research in this area in the near future. For example, it would be interesting to analyze the time-dependent case or to extend the analysis to anisotropic meshes, to derive a posteriori error estimates and to develop adaptive techniques, or to improve the efficiency of the solution of the nonlinear algebraic systems.

References

[1] F. Alauzet and A. Loseille. A decade of progress on anisotropic mesh adaptation for computational fluid dynamics. *Computer-Aided Design* 72: 13–39, 2016.

[2] P. Arminjon and A. Dervieux. Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. *J. Comput. Phys.* 106 (1): 176–198, 1993.

[3] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla. An assessment of discretizations for convection-dominated convection–diffusion equations. *Comput. Methods Appl. Mech. Engrg.* 200 (47-48): 3395–3409, 2011.

[4] S. Badia and A. Hierro. On monotonicity-preserving stabilized finite element approximations of transport problems. *SIAM J. Sci. Comput.* 36 (6): A2673–A2697, 2014.

[5] G.R. Barrenechea, V. John, and P. Knobloch. A local projection stabilization finite element method with nonlinear crosswind diffusion for convection–diffusion–reaction equations. *ESAIM: Math. Model. Numer. Anal.* 47 (5): 1335–1366, 2013.

[6] G.R. Barrenechea, V. John, and P. Knobloch. A nonlinear local projection stabilization for convection–diffusion–reaction equations. In A. Cangiani, R.L. Davidchack, E. Georgoulis, A.N. Gorban, J. Levesley, and M.V. Tretyakov, editors, *Numerical Mathematics and Advanced Applications 2011, Proceedings of ENUMATH 2011*, pp. 237–245. Springer-Verlag, Berlin, 2013.

[7] G.R. Barrenechea, V. John, and P. Knobloch. Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension. *IMA J. Numer. Anal.* 35 (4): 1729–1756, 2015.

[8] G.R. Barrenechea, V. John, and P. Knobloch. Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.* 54 (4): 2427–2451, 2016.

[9] G.R. Barrenechea, V. John, and P. Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.* 27 (3): doi:10.1142/S0218202517500087, 2017.

[10] R. Becker and M. Braack. A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo* 38: 173–199, 2001.

[11] R. Becker and M. Braack. A two-level stabilization scheme for the Navier–Stokes equations. In M. Feistauer, V. Dolejší, P. Knobloch, and K. Najzar, editors, *Numerical Mathematics and Advanced Applications*, pp. 123–130. Springer-Verlag, Berlin, 2004.

[12] R. Becker and B. Vexler. Optimal control of the convection–diffusion equation using stabilized finite element methods. *Numer. Math.* 106: 349–367, 2007.

[13] J.P. Boris and D.L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.* 11 (1): 38–69, 1973.

[14] M. Braack. Optimal control in fluid mechanics by finite elements with symmetric stabilization. *SIAM J. Control Optim.* 48: 672–687, 2009.

[15] M. Braack and G. Lube. Finite elements with local projection stabilization for incompressible flow problems. *J. Comput. Math.* 27: 116–147, 2009.

[16] 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 (1-3): 199–259, 1982.

[17] E. Burman and P. Hansbo. Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems. *Comput. Methods Appl. Mech. Engrg.* 193: 1437–1453, 2004.

[18] E. Burman and A. Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation. *Comput. Methods Appl. Mech. Engrg.* 191 (35): 3833–3855, 2002.

[19] E. Burman and A. Ern. Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence. *Math. Comp.* 74: 1637–1652, 2005.

[20] L.A. Catalano, P. De Palma, M. Napolitano, and G. Pascazio. A critical analysis of multi-dimensional upwinding for the Euler equations. *Comput. & Fluids* 25 (1): 29–38, 1996.

[21] R. Codina. A discontinuity-capturing crosswind–dissipation for the finite element solution of the convection–diffusion equation. *Comput. Methods Appl. Mech. Engrg.* 110: 325–342, 1993.

[22] 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: 253–276, 1992.

[23] S. Ganesan and L. Tobiska. Stabilization by local projection for convection–diffusion and incompressible flow problems. *J. Sci. Comput.* 43: 326–342, 2010.

[24] Z. Gao and J. Wu. A linearity-preserving cell-centered scheme for the heterogeneous and anisotropic diffusion equations on general meshes. *Internat. J. Numer. Methods Fluids* 67 (12): 2157–2183, 2011.

[25] S.K. Godunov. *Different Methods for Shock Waves.* PhD thesis, Moscow State University, 1954.

[26] M. Gurris, D. Kuzmin, and S. Turek. Implicit finite element schemes for the stationary compressible Euler equations. *Internat. J. Numer. Methods Fluids* 69 (1): 1–28, 2012.

[27] 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: 173–189, 1989.

[28] T.J.R. Hughes and A. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In *Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979)*, volume 34 of *AMD*, pp. 19–35. Amer. Soc. Mech. Engrs. (ASME), New York, 1979.

[29] W. Hundsdorfer and C. Montijn. A note on flux limiting for diffusion discretizations. *IMA J. Numer. Anal.* 24 (4): 635–642, 2004.

[30] V. John and P. Knobloch. A computational comparison of methods diminishing spurious oscillations in finite element solutions of convection–diffusion equations. In J. Chleboun, K. Segeth, and T. Vejchodský, editors, *Proceedings of the International Conference Programs and Algorithms of Numerical Mathematics 13* pp. 122–136. Academy of Science of the Czech Republic, Prague, 2006.

[31] V. John and P. Knobloch. On discontinuity-capturing methods for convection–diffusion equations. In A. Bermúdez de Castro, D. Gómez, P. Quintela, and P. Salgado, editors, *Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2005*, pp. 336–344. Springer–Verlag, Berlin, 2006.

[32] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.* 196 (17-20): 2197–2215, 2007.

[33] V. John and P. Knobloch. On the performance of SOLD methods for convection–diffusion problems with interior layers. *Int. J. Comput. Sci. Math.* 1 (2-4): 245–258, 2007.

[34] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.* 197 (21-24): 1997–2014, 2008.

[35] V. John and P. Knobloch. Adaptive computation of parameters in stabilized methods for convection–diffusion problems. In A. Cangiani, R.L. Davidchack, E.H. Georgoulis, A. Gorban, J. Levesley, and M.V. Tretyakov, editors, *Numerical Mathematics and Advanced Applications 2011, Proceedings of ENUMATH 2011*, pp. 275–283. Springer–Verlag, Berlin, 2013.

[36] V. John, P. Knobloch, and S.B. Savescu. A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I. *Comput. Methods Appl. Mech. Engrg.* 200 (41-44): 2916–2929, 2011.

[37] V. John and J. Novo. On (essentially) non-oscillatory discretizations of evolutionary convection–diffusion equations. *J. Comput. Phys.* 231 (4): 1570–1586, 2012.

[38] V. John and E. Schmeyer. Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. *Comput. Methods Appl. Mech. Engrg.* 198 (3-4): 475–494, 2008.

[39] M. Kadalbajoo and V. Gupta. A brief survey on numerical methods for solving singularly perturbed problems. *Appl. Math. Comput.* 217: 3641–3716, 2010.

[40] P. Knobloch. Improvements of the Mizukami–Hughes method for convection–diffusion equations. *Comput. Methods Appl. Mech. Engrg.* 196 (1-3): 579–594, 2006.

[41] P. Knobloch. Numerical solution of convection–diffusion equations using upwinding techniques satisfying the discrete maximum principle. In M. Beneš, M. Kimura, and T. Nakaki, editors, *Proceedings of Czech–Japanese Seminar in Applied Mathematics 2005*, volume 3 of *COE Lecture Note*, pp. 69–76. Faculty of Mathematics, Kyushu University, 2006.

[42] P. Knobloch. Application of the Mizukami–Hughes method to bilinear finite elements. In M. Beneš, M. Kimura, and T. Nakaki, editors, *Proceedings of Czech–Japanese Seminar in Applied Mathematics 2006*, volume 6 of *COE Lecture Note*, pp. 137–147. Faculty of Mathematics, Kyushu University, 2007.

[43] P. Knobloch. On the definition of the SUPG parameter. *Electron. Trans. Numer. Anal.* 32: 76–89, 2008.

[44] P. Knobloch. On a variant of the local projection method stable in the SUPG norm. *Kybernetika* 45 (4): 634–645, 2009.

[45] P. Knobloch. On the application of local projection methods to convection–diffusion–reaction problems. In A.F. Hegarty, N. Kopteva, E. O'Riordan, and M. Stynes, editors, *BAIL 2008 – Boundary and Interior Layers*, volume 69 of *Lect. Notes Comput. Sci. Eng.*, pp. 183–194. Springer-Verlag, Berlin, 2009.

[46] P. Knobloch. On the choice of the SUPG parameter at outflow boundary layers. *Adv. Comput. Math.* 31 (4): 369–389, 2009.

[47] P. Knobloch. A generalization of the local projection stabilization for convection–diffusion–reaction equations. *SIAM J. Numer. Anal.* 48 (2): 659–680, 2010.

[48] P. Knobloch. Local projection method for convection–diffusion–reaction problems with projection spaces defined on overlapping sets. In G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva, editors, *Numerical Mathematics and Advanced Applications 2009, Proceedings of ENUMATH 2009*, pp. 497–505. Springer-Verlag, Berlin, 2010.

[49] P. Knobloch. Numerical solution of convection–diffusion equations using a nonlinear method of upwind type. *J. Sci. Comput.* 43 (3): 454–470, 2010.

[50] P. Knobloch. Error estimates for a nonlinear local projection stabilization of transient convection–diffusion–reaction equations. *Discrete Contin. Dyn. Syst. Ser. S* 8 (5): 901–911, 2015.

[51] P. Knobloch and G. Lube. Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach. *Appl. Numer. Math.* 59 (12): 2891–2907, 2009.

[52] P. Knobloch and L. Tobiska. On the stability of finite-element discretizations of convection–diffusion–reaction equations. *IMA J. Numer. Anal.* 31 (1): 147–164, 2011.

[53] D. Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. *J. Comput. Phys.* 219 (2): 513–531, 2006.

[54] D. Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. In M. Papadrakakis, E. Oñate, and B. Schrefler, editors, *Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering*, pp. 1–5. CIMNE, Barcelona, 2007.

[55] D. Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *J. Comput. Appl. Math.* 236 (9): 2317–2337, 2012.

[56] D. Kuzmin and M. Möller. Algebraic flux correction I. Scalar conservation laws. In D. Kuzmin, R. Löhner, and S. Turek, editors, *Flux-Corrected Transport. Principles, Algorithms, and Applications*, pp. 155–206. Springer-Verlag, Berlin, 2005.

[57] D. Kuzmin, M.J. Shashkov, and D. Svyatskiy. A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems. *J. Comput. Phys.* 228 (9): 3448–3463, 2009.

[58] D. Kuzmin and S. Turek. High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. *J. Comput. Phys.* 198 (1): 131–158, 2004.

[59] T. Linß. *Layer-adapted meshes for reaction–convection–diffusion problems.* Springer-Verlag, Berlin, 2010.

[60] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati. Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations. *Int. J. Numer. Methods Fluids* 7 (10): 1093–1109, 1987.

[61] N. Madden and M. Stynes. Linear enhancements of the streamline diffusion method for convection–diffusion problems. *Comput. Math. Appl.* 32 (10): 29–42, 1996.

[62] G. Matthies, P. Skrzypacz, and L. Tobiska. A unified convergence analysis for local projection stabilizations applied to the Oseen problem. *M2AN Math. Model. Numer. Anal.* 41: 713–742, 2007.

[63] G. Matthies, P. Skrzypacz, and L. Tobiska. Stabilization of local projection type applied to convection–diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.* 32: 90–105, 2008.

[64] A. Mizukami and T. 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: 181–193, 1985.

[65] H.G. Roos. Robust numerical methods for singularly perturbed differential equations: a survey covering 2008-2012. *ISRN Applied Mathematics* 2012: article ID 379547, 2012.

[66] H.G. Roos, M. Stynes, and L. Tobiska. *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems. 2nd ed.*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2008.

[67] R. Schneider. A review of anisotropic refinement methods for triangular meshes in FEM. In T. Apel, editor, *Advanced finite element methods and applications*, volume 66 of *Lect. Notes Appl. Comput. Mech.*, pp. 133–152. Springer-Verlag, Berlin, 2013.

[68] S.T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* 31 (3): 335–362, 1979.

Chapter 2

Mizukami–Hughes method

This chapter consists of the following publications:

P. Knobloch: Improvements of the Mizukami–Hughes method for convection–diffusion equations, *Computer Methods in Applied Mechanics and Engineering* 196 (1-3): 579–594, 2006. p. 23

P. Knobloch: Numerical solution of convection–diffusion equations using a nonlinear method of upwind type, *Journal of Scientific Computing* 43 (3): 454–470, 2010. p. 39

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 196 (2006) 579–594
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

Improvements of the Mizukami–Hughes method for convection–diffusion equations

Petr Knobloch

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

Received 13 September 2005; received in revised form 30 May 2006; accepted 9 June 2006

Abstract

We consider the Mizukami–Hughes method for the numerical solution of scalar two-dimensional steady convection–diffusion equations using conforming triangular piecewise linear finite elements. We propose several modifications of this method to eliminate its shortcomings. The improved method still satisfies the discrete maximum principle and gives very accurate discrete solutions in convection-dominated regime, which is illustrated by several numerical experiments. In addition, we show how the Mizukami–Hughes method can be applied to convection–diffusion–reaction equations and to three-dimensional problems.

Keywords: Stabilized FEM; Convection–diffusion; Convection–diffusion–reaction; Petrov–Galerkin method; Discrete maximum principle

1. Introduction

In this paper we propose several improvements of the Mizukami–Hughes method introduced in [14] for solving the convection–diffusion equation

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega. \tag{1}$$

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, $\varepsilon > 0$ is the diffusivity, which is assumed to be constant, and $\boldsymbol{b}$ is the flow velocity. Eq. (1) is equipped with boundary conditions

$$u = u_b \quad \text{on } \Gamma^D, \quad \varepsilon\frac{\partial u}{\partial \boldsymbol{n}} = g \quad \text{on } \Gamma^N, \tag{2}$$

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

Despite the apparent simplicity of problem (1) and (2), its numerical solution is by no means an easy task since convection often dominates diffusion and hence the solution of (1) and (2) 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.

To enhance the stability and accuracy of the Galerkin discretization of (1) and (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) 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. [1,3–6,8,9,13,15] and the references there). This procedure

E-mail address: knobloch@karlin.mff.cuni.cz

0045-7825/$ - see front matter
doi:10.1016/j.cma.2006.06.004

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.

An interesting monotone method for solving (1) and (2) was introduced by Mizukami and Hughes [14] for linear triangular finite elements. Although it is not clear how to generalize this method to other types of finite elements, it deserves some attention since it seems to give very accurate solutions and possesses many nice properties. First of all, in contrast to the most discontinuity-capturing methods, the solutions 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) and (2), the Mizukami–Hughes method depends on the unknown discrete solution and hence it is nonlinear.

Although the Mizukami–Hughes discrete solutions are often very accurate, we observed that, in some cases, they are not correct. Moreover, sometimes it was very difficult to solve the nonlinear problem with a prescribed accuracy. Therefore, in this paper, we propose some improvements of the method which correct the mentioned shortcomings and keep its quality in cases in which it works well. We will be interested in the strongly convection-dominated case characterized by the condition $\varepsilon \ll |\boldsymbol{b}|$, where $|\boldsymbol{b}|$ is the Euclidean norm of $\boldsymbol{b}$.

A drawback of both the original and the improved versions of the Mizukami–Hughes method is that no existence, uniqueness and convergence results are available. Moreover, it seems to be rather difficult to generalize the method to more complicated problems. Nevertheless, we shall show that the method can be extended to convection–diffusion–reaction equations and to the three-dimensional case.

The plan of the paper is as follows. First, in the next section, we describe and comment the original Mizukami–Hughes method published in [14]. Then, in Sections 3–5, we discuss shortcomings of this method and propose some modifications to eliminate them. Since this will take several pages, we briefly summarize the improved method in Section 6. Section 7 contains our numerical results which illustrate the high accuracy of the improved method. In Section 8, we deal with a generalization of the Mizukami–Hughes method to convection–diffusion–reaction equations and, finally, in Section 9, we discuss the application of the Mizukami–Hughes method to the three-dimensional case.

2. The Mizukami–Hughes method

Let $\mathscr{T}_h$ be a triangulation of Ω consisting of a finite number of open triangular elements K. The discretization parameter h in the notation $\mathscr{T}_h$ is a positive real number satisfying $\operatorname{diam}(K) \leqslant h$ for any $K \in \mathscr{T}_h$. We assume that $\overline{\Omega} = \bigcup_{K \in \mathscr{T}_h} \overline{K}$ and that the closures of any two different elements K, $\widetilde{K} \in \mathscr{T}_h$ are either disjoint or possess either a common vertex or a common edge. Further, we assume that any edge of an element $K \in \mathscr{T}_h$ which lies on $\partial\Omega$ is contained either in $\overline{\Gamma^{\mathrm{D}}}$ or in $\overline{\Gamma^{\mathrm{N}}}$. Finally, we assume that the triangulation $\mathscr{T}_h$ is of weakly acute type, i.e., the magnitude of all angles of elements $K \in \mathscr{T}_h$ is less than or equal to $\pi/2$. This property will be used for proving the discrete maximum principle.

The solution u of (1) and (2) will be approximated by a continuous piecewise linear function u_h from the space

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

Let $a_1, \ldots, a_{M_h}$ be the vertices of $\mathscr{T}_h$ lying in $\Omega \cup \Gamma^{\mathrm{N}}$ and let $a_{M_h+1}, \ldots, a_{N_h}$ be the vertices of $\mathscr{T}_h$ lying on $\overline{\Gamma^{\mathrm{D}}}$. 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 δ_{ij} is the Kronecker symbol. Then $V_h = \operatorname{span}\{\varphi_i\}_{i=1}^{N_h}$. The Mizukami–Hughes method is a Petrov–Galerkin method with weighting functions

$$\tilde{\varphi}_i = \varphi_i + \sum_{\substack{K \in \mathscr{T}_h, \\ a_i \in \overline{K}}} C_i^K \chi_K, \quad i = 1, \ldots, M_h,$$

where C_i^K are constants to be determined later and χ_K is the characteristic function of K (i.e., $\chi_K = 1$ in K and $\chi_K = 0$ elsewhere). The discrete solution u_h of (1) and (2) is defined by

$$\begin{aligned}
&u_h \in V_h,\\
&\varepsilon(\nabla u_h, \nabla \varphi_i) + (\boldsymbol{b} \cdot \nabla u_h, \tilde{\varphi}_i) = (f, \tilde{\varphi}_i) + (g, \varphi_i)_{\Gamma^{\mathrm{N}}},\\
&\quad i = 1, \ldots, M_h,\\
&u_h(a_i) = u_{\mathrm{b}}(a_i), \quad i = M_h + 1, \ldots, N_h,
\end{aligned}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ and $(\cdot,\cdot)_{\Gamma^{\mathrm{N}}}$ is the inner product in $L^2(\Gamma^{\mathrm{N}})$. Moreover, here and in the following, the flow velocity $\boldsymbol{b}$ is considered to be piecewise constant (equal to the original function $\boldsymbol{b}$ at barycentres of elements of $\mathscr{T}_h$).

It remains to define the constants C_i^K, which is the key point of the method. Mizukami and Hughes require for any $K \in \mathscr{T}_h$ that

$$C_i^K \geqslant -\tfrac{1}{3}\ \forall i \in \{1, \ldots, N_h\},\ a_i \in \overline{K}, \quad \sum_{\substack{i=1 \\ a_i \in \overline{K}}}^{N_h} C_i^K = 0 \tag{3}$$

and that the local convection matrix A^K with entries

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

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, cf. [7]). As usual, $(\cdot,\cdot)_K$ denotes the inner product in $L^2(K)$. The latter condition in (3) implies that u_h satisfies a discrete mass conservation law if the data in (1) and (2) satisfy $\Gamma^N = \partial\Omega$, $g = 0$ and $\boldsymbol{b} = \text{const.}$, cf. [11].

The matrix A^K has three columns and at most three rows and it will be of nonnegative type as soon as $a_{ij}^K \leqslant 0$ for $i \neq j$. Note that

$$a_{ij}^K = (\boldsymbol{b}\cdot\nabla\varphi_j)|_K \int_K \tilde{\varphi}_i \,\mathrm{d}x = (\boldsymbol{b}\cdot\nabla\varphi_j)|_K \mathrm{meas}_2(K)\left(\tfrac{1}{3} + C_i^K\right).$$

Let K be any element of the triangulation $\mathscr{T}_h$ 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_i whose boundaries consist of lines intersecting the barycentre of K which are parallel to the two edges of K possessing the vertex a_i, see Fig. 1. The common part of the boundaries of two adjacent zones is included in the respective vertex zone. To avoid misunderstandings, we shall later also use the notation EZ_i^K instead of EZ_i.

Without loss of generality, we may assume that the vertices of K are numbered in such a way that $\boldsymbol{b}$ points into the vertex zone or the edge zone of a_1 as depicted in Fig. 1. Then

$$\boldsymbol{b} \in \mathrm{VZ}_1 \Longleftrightarrow \boldsymbol{b}\cdot\nabla\varphi_1 > 0, \quad \boldsymbol{b}\cdot\nabla\varphi_2 \leqslant 0, \quad \boldsymbol{b}\cdot\nabla\varphi_3 \leqslant 0,$$
$$\boldsymbol{b} \in \mathrm{EZ}_1 \Longleftrightarrow \boldsymbol{b}\cdot\nabla\varphi_1 < 0, \quad \boldsymbol{b}\cdot\nabla\varphi_2 > 0, \quad \boldsymbol{b}\cdot\nabla\varphi_3 > 0,$$

where we write $\nabla\varphi_i$ instead of $\nabla\varphi_i|_K$ for simplicity.

If $\boldsymbol{b} \in \mathrm{VZ}_1$, then (3) holds and A^K is of nonnegative type for

$$C_1^K = \tfrac{2}{3}, \qquad C_2^K = C_3^K = -\tfrac{1}{3}.$$

If A^K has three rows, this is the only possibility how to choose these constants. On the other hand, if $\boldsymbol{b} \in \mathrm{EZ}_1$, then it is generally not possible to choose the constants C_1^K, C_2^K, C_3^K in such a way that (3) holds and A^K is of nonnegative type. However, Mizukami and Hughes made the important observation that u still solves Eq. (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 matrix A^K is of nonnegative type for $\boldsymbol{b}$ replaced by a function $\tilde{\boldsymbol{b}}$ pointing into a vertex zone. Since ∇u is not known a priori, we obtain a nonlinear problem where the constants C_i^K depend on the discrete solution u_h which we want to compute.

Let us assume that $\boldsymbol{b}\cdot\nabla u_h|_K \neq 0$ and let $\boldsymbol{w} \neq \boldsymbol{0}$ be a vector orthogonal to $\nabla u_h|_K$. Then there exists $\alpha \in \mathbb{R}$ such that $\boldsymbol{b} + \alpha\boldsymbol{w} \in \mathrm{VZ}_2$ or $\boldsymbol{b} + \alpha\boldsymbol{w} \in \mathrm{VZ}_3$. The dashed and dotted arcs in Fig. 2 indicate to which part of the plane the vector $\boldsymbol{w}$ should point from the barycentre of K if the first or the second possibility should arrive. To simplify the presentation, let us introduce the sets

$$V_k = \{\alpha \in \mathbb{R}; \boldsymbol{b} + \alpha\boldsymbol{w} \in \mathrm{VZ}_k\}, \quad k = 2,3.$$

Mizukami and Hughes show that, depending on V_2 and V_3, the following values of the constants C_i^K should be used:

$$V_2 \neq \emptyset \quad \text{and} \quad V_3 = \emptyset$$
$$\Rightarrow \quad C_2^K = \tfrac{2}{3}, \quad C_1^K = C_3^K = -\tfrac{1}{3}, \tag{4}$$
$$V_2 = \emptyset \quad \text{and} \quad V_3 \neq \emptyset$$
$$\Rightarrow \quad C_3^K = \tfrac{2}{3}, \quad C_1^K = C_2^K = -\tfrac{1}{3}, \tag{5}$$
$$V_2 \neq \emptyset \quad \text{and} \quad V_3 \neq \emptyset$$
$$\Rightarrow \quad C_1^K = -\tfrac{1}{3}, \quad C_2^K + C_3^K = \tfrac{1}{3},$$
$$C_2^K \geqslant -\tfrac{1}{3}, \quad C_3^K \geqslant -\tfrac{1}{3}. \tag{6}$$

If, for some $k \in \{2,3\}$, the set V_k is nonempty, we choose $\alpha_k \in V_k$ and define the matrix $\tilde{A}^{K,k}$ with entries

$$\tilde{a}_{ij}^{K,k} = ((\boldsymbol{b} + \alpha_k\boldsymbol{w})\cdot\nabla\varphi_j, \tilde{\varphi}_i)_K, \quad i,j = 1,2,3 \ \ (a_i \in \Omega \cup \Gamma^N),$$

where $\tilde{\varphi}_i$ are defined using C_i^K's from (4) if $k = 2$ and using C_i^K's from (5) if $k = 3$. As we have seen above, the matrix

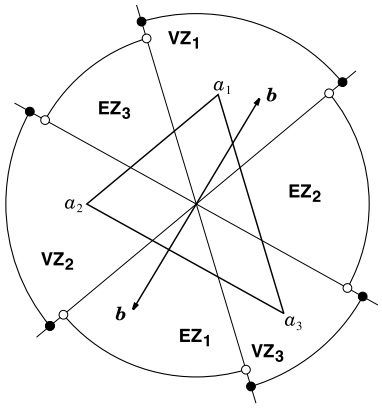


Fig. 1. Definition of edge zones and vertex zones.

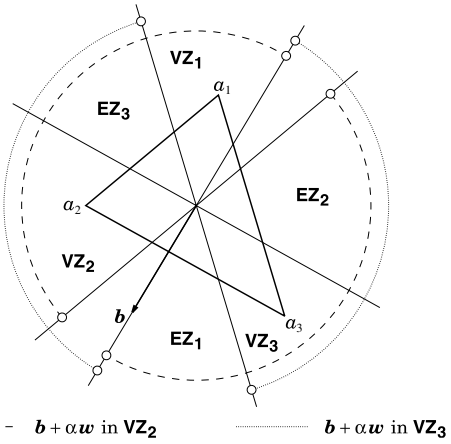


Fig. 2. Orientations of $\boldsymbol{w}$ for which $\boldsymbol{b} + \alpha\boldsymbol{w} \in \mathrm{VZ}_2$ or $\boldsymbol{b} + \alpha\boldsymbol{w} \in \mathrm{VZ}_3$.

$\widetilde{A}^{K,k}$ is of nonnegative type. Let us assume that V_2 or V_3 is empty and let V_k be the nonempty set. Since $u_h|_K = u_1\varphi_1 + u_2\varphi_2 + u_3\varphi_3$, the vector $U = (u_1, u_2, u_3)$ satisfies for $i = 1, 2, 3$ (with $a_i \in \Omega \cup \Gamma^{\mathrm{N}}$)

$$(A^K U)_i = (\boldsymbol{b} \cdot \nabla u_h, \tilde{\varphi}_i)_K = ((\boldsymbol{b} + \alpha_k \boldsymbol{w}) \cdot \nabla u_h, \tilde{\varphi}_i)_K = (\widetilde{A}^{K,k} U)_i.$$

In case (6), we have

$$A^K = (C_2^K + \tfrac{1}{3})A^{K,2} + (C_3^K + \tfrac{1}{3})A^{K,3},$$

where $A^{K,2}$ and $A^{K,3}$ are matrices defined like A^K but using C_i^K's from (4) and (5), respectively. Consequently, for $i = 1, 2, 3$ (with $a_i \in \Omega \cup \Gamma^{\mathrm{N}}$), we obtain

$$(A^K U)_i = (C_2^K + \tfrac{1}{3})(\widetilde{A}^{K,2} U)_i + (C_3^K + \tfrac{1}{3})(\widetilde{A}^{K,3} U)_i.$$

Thus, in all three cases (4)–(6), the discrete solution satisfies

$$(A^K U)_i = (\widetilde{A}^K U)_i, \quad i = 1, 2, 3 \ \ (a_i \in \Omega \cup \Gamma^{\mathrm{N}}), \tag{7}$$

where $\widetilde{A}^K$ is a matrix of nonnegative type. In case (6), Mizukami and Hughes suggest to set

$$C_i^K = \frac{\boldsymbol{b} \cdot \nabla \varphi_i}{3|\boldsymbol{b} \cdot \nabla \varphi_1|}, \quad i = 1, 2, 3. \tag{8}$$

This choice is also considered if $\boldsymbol{b} \in \mathrm{EZ}_1$ satisfies $\boldsymbol{b} \cdot \nabla u_h|_K = 0$. If $\boldsymbol{b} = \boldsymbol{0}$, Mizukami and Hughes set $C_i^K = 0$ for $i = 1, 2, 3$.

The above choice of the constants C_i^K assures that the discrete solution always satisfies (7) with a matrix $\widetilde{A}^K$ 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 $\widetilde{A}$ the $M_h \times N_h$ matrix made up of the local matrices $\widetilde{A}^K$, 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 $C \equiv \varepsilon D + \widetilde{A}$. Since the triangulation $\mathscr{T}_h$ is of weakly acute type, it is easily seen that the matrix $\{(\nabla\varphi_j, \nabla\varphi_i)_K\}_{i,j=1}^3$ is of nonnegative type. Consequently, the matrices D and C also 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. [8]). Thus, for any $G \subset \overline{\Omega}$ being a union of closures of elements of $\mathscr{T}_h$, we have

$$(f, \tilde{\varphi}_i) \leqslant 0 \quad \forall a_i \in \operatorname{int} G \Rightarrow \quad \max_G u_h = \max_{\partial G} u_h, \tag{9}$$

$$(f, \tilde{\varphi}_i) \geqslant 0 \quad \forall a_i \in \operatorname{int} G \Rightarrow \quad \min_G u_h = \min_{\partial G} u_h, \tag{10}$$

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

3. Improvement of the Mizukami–Hughes method in boundary layer regions

The Mizukami–Hughes method often provides accurate and oscillation-free discrete solutions, see the examples in [14,9]. However, in some cases, we observed that the discrete solution was not correct. We shall demonstrate this on a simple example which was also considered in [14].

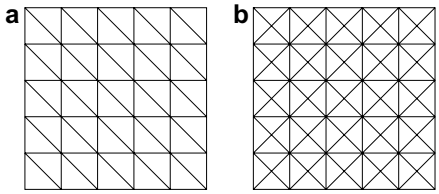


Fig. 3. Considered types of triangulations.

Let $\Omega = (0, 1)^2$ and, like in [14], let us consider uniform triangulations $\mathscr{T}_h$ of Ω of the type depicted in Fig. 3(a), which consist of $2(N \times N)$ equal right-angled isosceles triangles ($N = 5$ in Fig. 3(a)). Let $N = 10$ and let us consider the problem (1) and (2) with

$$\varepsilon = 10^{-7}, \quad \boldsymbol{b} = (1, 0), \quad f = 1, \quad \Gamma^{\mathrm{D}} = \partial\Omega, \quad u_{\mathrm{b}} = 0. \tag{11}$$

The discrete solution obtained using the Mizukami–Hughes method is indistinguishable from the discrete solution corresponding to $\varepsilon \to 0$. For $\varepsilon \to 0$, we easily find that the discrete solution is nodally exact, i.e.,

$$u_h(x, y) = x \quad \text{for } (x, y) \in [0, 0.9] \times [0.1, 0.9]. \tag{12}$$

Changing $\boldsymbol{b}$ to $\boldsymbol{b} = (1, \alpha)$ with $|\alpha| \ll 1$, we expect that the discrete solution basically remains the same. However, Fig. 4(a) corresponding to $\alpha = -0.0001$ shows that the discrete solution changes dramatically. The reason is that the small change of $\boldsymbol{b}$ causes a significant change of the constants C_i^K for elements $K \in \mathscr{T}_h$ having an edge at the upper part of the boundary of Ω, see Fig. 5(a) and (b). Note that we can set $\boldsymbol{w} = (1, 0)$ for these elements K since $u_h = 0$ on $\partial\Omega$. Let us mention that Fig. 4(a) does not show a violation of the discrete maximum principle since $(f, \tilde{\varphi}_i) > 0$ for all $i \in \{1, \ldots, M_h\}$ and the right-hand side of (10) is satisfied for any admissible set G.

It is obvious that a small change of $\boldsymbol{b}$ should only lead to a small change of the constants C_i^K and hence a first idea to improve the behaviour of the method might be to use the vertex-zone definition of C_i^K's also for $\boldsymbol{b}$ which is not contained in a vertex zone but is very near to it. However, the problems also appear for vectors $\boldsymbol{b}$ which cannot be considered to lie near a vertex zone, e.g. for $\alpha \in [-0.5, -0.1]$. For such α, a nodally exact solution (again for $\varepsilon \to 0$) should satisfy

$$u_h(x, y) = x \quad \text{for } (x, y) \in [0, 0.9] \times [0.1, 0.2]. \tag{13}$$

Let us assume that

(A1) the constants C_i^K are defined as described in Section 2 if $\boldsymbol{b}$ lies in a vertex zone;
(A2) $C_j^K = -\frac{1}{3}$ if $\boldsymbol{b} \in \mathrm{EZ}_j^K$ for some index j.

Then, for $\varepsilon = 0$, it is easy to show that the necessary condition for the validity of (13) is that, for any element K having the vertices $(x, 0)$, $(x, 0.1)$, $(x - 0.1, 0.1)$ with $x \in$

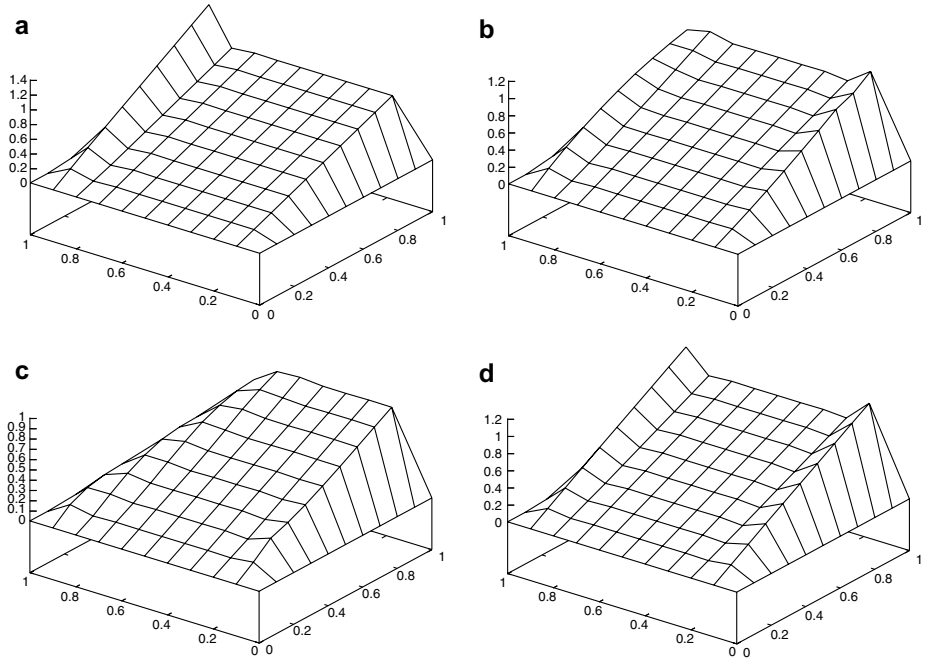


Fig. 4. Mizukami–Hughes discrete solution for data (11) with $\boldsymbol{b}$ replaced by the indicated vectors: (a) $\boldsymbol{b}=(1,-0.0001)$, (b) $\boldsymbol{b}=(1,-0.1)$, (c) $\boldsymbol{b}=(1,-0.4)$ and (d) $\boldsymbol{b}=(1,0)$, $\mathscr{T}_h$ from Fig. 3(b).

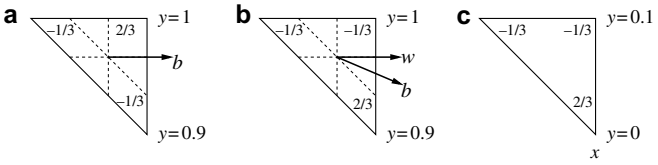


Fig. 5. Values of the constants C_i^K: (a) $\boldsymbol{b}=(1,0)$, (b) $\boldsymbol{b}=(1,\alpha)$, $\alpha\in(-1,0)$ and (c) optimal values.

$\{0.1,0.2,\ldots,0.9\}$, the constants C_i^K are equal to the values depicted in Fig. 5(c). Since $u_h(x,0)=0$, we have $\nabla u_h=(1,10x)$ and we can set $\boldsymbol{w}=(1,-0.1/x)$. Hence the Mizukami–Hughes method gives the optimal values of C_i^K's only if $x>0.1/|\alpha|$. Thus, for $\alpha=-0.1$, the discrete solution u_h is wrong along the whole lower part of the boundary of Ω (cf. Fig. 4(b)) whereas, for $\alpha=-0.4$, the values of C_i^K's are correct for elements with $x>0.25$ and hence u_h is better although still wrong (cf. Fig. 4(c)).

The problems observed above also appear for the data (11) if we consider a triangulation of Ω of the type depicted in Fig. 3(b) which consists of $4(N\times N)$ equal right-angled isosceles triangles ($N=5$ in Fig. 3(b)). For $N=10$, the discrete solution corresponding to the Mizukami–Hughes method is shown in Fig. 4(d) and, as wee see, it is wrong (the solution is visualized using its values at the same points as in Fig. 4(a)–(c)). For $\varepsilon=0$ and under the assumptions (A1) and (A2), the discrete solution satisfies (12) only if, on elements K with vertices $(x,0)$, $(x,0.1)$, $(x-0.05,0.05)$ or $(x,0.9)$, $(x,1)$, $(x-0.05,0.95)$ where $x\in\{0.1,0.2,\ldots,0.9\}$, we set $C_i^K=-\frac{1}{3}$ for i corresponding to $(x,0.1)$ or $(x,0.9)$. Whereas, for the examples mentioned above, we could think of redefining C_i^K's employing the relation between $\boldsymbol{b}$ and $\boldsymbol{w}$ in some more sophisticated way, now this is not possible since $\boldsymbol{w}=\boldsymbol{b}$. Moreover, the direction of ∇u_h on K also cannot be employed since it changes if $f=-1$ is used instead of $f=1$ whereas the values of C_i^K's have to remain the same.

In view of the above discussed and many other numerical experiments, we conclude that the definition of C_i^K's for $\boldsymbol{b}$ lying in an edge zone is not appropriate if K lies in the numerical boundary layer. The only remedy we have found is to set $C_i^K=-\frac{1}{3}$ for all i corresponding to inner vertices. This leads us to the following requirement:

(A3) $C_i^K=-\frac{1}{3}$ for all $i=1,2,3$ if $\overline{K}\cap\overline{\Gamma^{\mathrm{D}}}\neq\emptyset$ and if $\boldsymbol{b}\in\mathrm{EZ}_j^K$ for some $j\in\{1,2,3\}$.

Note that the constants C_i^K corresponding to vertices $a_i \in \overline{\Gamma^D}$ do not influence the discrete solution so that we could also define them in such a way that (3) is formally satisfied.

The requirement (A3) is not sufficient to avoid wrong discrete solutions on a triangulation of the type from Fig. 3(b) if $\boldsymbol{b} = (1, \alpha)$ with $\alpha \neq 0$. In this case we require that

(A3*) $C_i^K = -\frac{1}{3}$ for all $i = 1, 2, 3$ if all vertices of K are connected by edges to vertices on $\overline{\Gamma^D}$ and if $\boldsymbol{b} \in \mathrm{EZ}_j^K$ for some $j \in \{1, 2, 3\}$.

For $\boldsymbol{b} = (1, 0)$, this stronger requirement is not needed on a triangulation of the type from Fig. 3(b): for $N = 10$, there exists a unique $u_h \in V_h$ satisfying (12) and such that $\boldsymbol{b} \cdot \nabla u_h = f$ on any element of $\mathscr{T}_h$ with vertices of the type (x, y), $(x, y + 0.1)$, $(x + 0.05, y + 0.05)$ and on any element having an edge on the boundary of $(0, 0.9) \times (0.1, 0.9)$. Assuming (A3), it is easy to verify that this u_h solves the discrete problem with $\varepsilon = 0$. However, generally, (A3*) is a necessary condition for obtaining a nodally exact solution.

4. Continuous dependence of C_i^K's on the orientation of the convection $\boldsymbol{b}$

Let us consider the situation depicted in Fig. 5(a). Since $\boldsymbol{b}$ lies in a vertex zone, the values of the constants C_i^K are independent of the discrete solution u_h. Now, like in the preceding section, let us change $\boldsymbol{b}$ to $\boldsymbol{b} = (1, \alpha)$ with $\alpha < 0$, $|\alpha| \ll 1$. Then $\boldsymbol{b}$ lies in an edge zone which we denote EZ_1 and the constants C_i^K are determined according to (4)–(6). Assuming that both V_2 and V_3 are nonempty, the formula (8) replaces the value $\frac{2}{3}$ in Fig. 5(a) by $\frac{1+\alpha}{3}$ and the value $-\frac{1}{3}$ at the vertex with $y = 0.9$ by $-\frac{\alpha}{3}$. Thus, the definition of the constants C_i^K is discontinuous with respect to the orientation of $\boldsymbol{b}$. This does not seem to be reasonable and our numerical experiments show that it may deteriorate the quality of the discrete solution. Therefore, in this section, we propose another way how to compute the constants C_i^K in case (6).

Let us again consider an element K with vertices a_1, a_2 and a_3. If $\boldsymbol{b} \in \mathrm{VZ}_2$, then $C_2^K = \frac{2}{3}$, $C_3^K = -\frac{1}{3}$ and, if $\boldsymbol{b} \in \mathrm{VZ}_3$, then $C_2^K = -\frac{1}{3}$, $C_3^K = \frac{2}{3}$. Thus, if $\boldsymbol{b} \in \mathrm{EZ}_1$, it is sensible to set

$$C_2^K = F\left(\frac{\alpha_3}{\alpha_2 + \alpha_3}\right), \quad C_3^K = F\left(\frac{\alpha_2}{\alpha_2 + \alpha_3}\right),$$

where α_2 and α_3 are the angles depicted in Fig. 6 and $F : [0, 1] \to \left[-\frac{1}{3}, \frac{2}{3}\right]$ is a continuous monotone function satisfying $F(0) = -\frac{1}{3}$ and $F(1) = \frac{2}{3}$. It is convenient to replace F by the function

$$G(x) = 2F\left(\frac{x+1}{2}\right) - \frac{1}{3}.$$

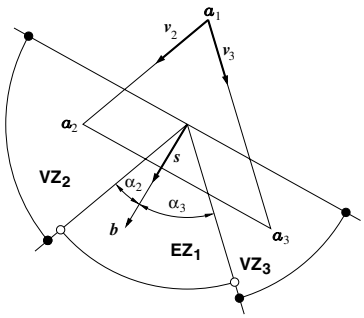


Fig. 6. Definition of angles α_2 and α_3 and of vectors $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ and $\boldsymbol{s}$.

Then

$$C_2^K = \frac{1}{6} + \frac{1}{2} G\left(\frac{\alpha_3 - \alpha_2}{\alpha_2 + \alpha_3}\right), \quad C_3^K = \frac{1}{6} + \frac{1}{2} G\left(\frac{\alpha_2 - \alpha_3}{\alpha_2 + \alpha_3}\right)$$

and G is a continuous monotone function satisfying

$$G : [-1, 1] \to [-1, 1], \quad G(-1) = -1, \quad G(1) = 1. \tag{14}$$

Moreover, (6) implies that G is odd.

To make the computation of the constants C_i^K cheaper, we use the approximation

$$\frac{\alpha_3 - \alpha_2}{\alpha_2 + \alpha_3} \approx \frac{\sin\left[\frac{1}{2}(\alpha_3 - \alpha_2)\right]}{\sin\left[\frac{1}{2}(\alpha_2 + \alpha_3)\right]} = \frac{\cos\alpha_2 - \cos\alpha_3}{1 - \cos(\alpha_2 + \alpha_3)},$$

which is certainly acceptable for $\alpha_2 + \alpha_3 \leqslant \frac{\pi}{2}$. Note that, denoting by $\boldsymbol{v}_2$ and $\boldsymbol{v}_3$ unit vectors pointing from a_1 to a_2 and a_3, respectively, and by $\boldsymbol{s}$ the unit vector in the direction of $\boldsymbol{b}$ (cf. Fig. 6), we have

$$\frac{\cos\alpha_2 - \cos\alpha_3}{1 - \cos(\alpha_2 + \alpha_3)} = \frac{(\boldsymbol{v}_2 - \boldsymbol{v}_3) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}.$$

Thus, we arrive at the formulas

$$C_2^K = \frac{1}{6} + \frac{1}{2} G\left(\frac{(\boldsymbol{v}_2 - \boldsymbol{v}_3) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}\right), \quad C_3^K = \frac{1}{3} - C_2^K, \tag{15}$$

where G is a continuous monotone odd function satisfying (14). We performed a lot of numerical experiments which revealed that a good choice for the function G is to simply set

$$G(x) = x.$$

5. Continuous dependence of C_i^K's on the orientation of ∇u_h

Let us consider the situation depicted in Fig. 2, i.e., $\boldsymbol{b} \in \mathrm{EZ}_1$. If the vector $\boldsymbol{w}$ points from the barycentre of K into the part of EZ_1 marked by the dashed arc, then the constants C_i^K are determined by (4) and hence $C_2^K = \frac{2}{3}$ and $C_3^K = -\frac{1}{3}$. However, as soon as $\boldsymbol{w}$ comes into the interior of VZ_3, the values of these constants change to values given

by (15). Consequently, the constants C_i^K depend on the orientation of $\boldsymbol{w}$ (and hence of ∇u_h) in a discontinuous way. Our numerical experiences show that, in some cases, this prevents the nonlinear iterative process from converging. Therefore, in the following, we describe a modification of the formula (15) taking into account the orientation of $\boldsymbol{w}$. We assume that $\boldsymbol{b}\cdot\nabla u_h|_K \neq 0$.

We shall need some additional notation which is introduced in Fig. 7. Here, the straight dashed lines are axes of the angles between the two lines which cross at the barycentre of K and are parallel to the edges of K containing the vertex a_1. One of these angles is the same as the angle ω_1 of K at a_1 and we introduce a unit vector $\boldsymbol{v}$ in the direction of the axis of this angle pointing as in Fig. 7. Without loss of generality, we may assume that $|\boldsymbol{w}| = 1$ and that $\boldsymbol{w}\cdot\boldsymbol{v} \geqslant 0$. Therefore, the dashed and dotted arcs in Fig. 7, which have the same meaning as in Fig. 2, are restricted to the corresponding half plane. We denote by δ the angle between $\boldsymbol{w}$ and the part of the boundary of EZ_1 which is 'nearer' to $\boldsymbol{w}$ (cf. Fig. 7). Like in Fig. 6, we introduce the angles α_2 and α_3 and the unit vectors $\boldsymbol{v}_2$, $\boldsymbol{v}_3$ and $\boldsymbol{s}$.

If $\boldsymbol{w} \in \overline{\mathrm{EZ}}_1$, the constants C_i^K are uniquely determined by (4) and (5). Thus, let us consider the case (6) and let $j,k \in \{2,3\}$, $j \neq k$, be such that $\boldsymbol{w} \in \mathrm{VZ}_j \cup \mathrm{EZ}_k$ ($j = 3$ in Fig. 7). It suffices to discuss the choice of C_j^K since $C_1^K = -\frac{1}{3}$ and $C_k^K = \frac{1}{3} - C_j^K$. Obviously, $\alpha_j \in (0,\omega_1)$ and $\delta \in (0,\kappa]$ with $\kappa = \frac{\pi}{2} - \frac{\omega_1}{2}$. We shall require the following values of C_j^K in the limit cases:

$\delta = \kappa \Rightarrow C_j^K$ is determined by (15),

$\alpha_j \to 0,\ \delta \nrightarrow 0 \Rightarrow C_j^K$ is determined by (15) ($\Rightarrow C_j^K \to \frac{2}{3}$),

$\delta \to 0,\ \alpha_j \nrightarrow 0 \Rightarrow C_j^K \to -\frac{1}{3}$.

If $\alpha_j \to 0$, $\delta \to 0$, then $\boldsymbol{b}\cdot\nabla u_h|_K \approx 0$ and hence the choice of C_i^K's is not important since $A^K U \approx 0$ in (7). Denoting by $\overline{C}_j^K$ the value of C_j^K determined by (15), we set

$$C_j^K = \overline{C}_j^K \Phi(\alpha_j,\delta) - \tfrac{1}{3}[1 - \Phi(\alpha_j,\delta)],$$

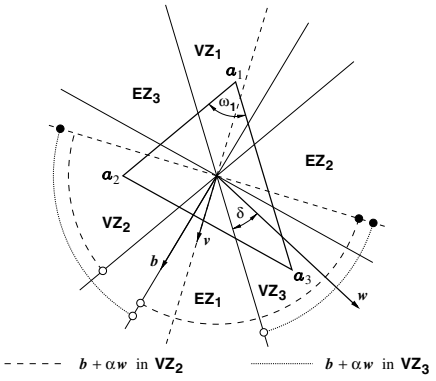


Fig. 7. Definition of angles ω_1 and δ and of the vector $\boldsymbol{v}$.

where $\Phi:([0,\omega_1]\times[0,\kappa])\backslash(0,0) \to [0,1]$ is a continuous function. The above requirements imply that

$$\Phi(\alpha_j,\kappa) = \Phi(0,\delta) = 1, \quad \Phi(\alpha_j,0) = 0 \quad \forall \alpha_j \in (0,\omega_1],\ \delta \in (0,\kappa].$$

Since the direction of $\boldsymbol{w}$ may strongly vary during the nonlinear iterative process, the constant C_j^K should be mainly determined by (15) and the orientation of $\boldsymbol{w}$ should influence C_j^K only if δ/κ is smaller than α_j/ω_1. Therefore, we set

$$\Phi(\alpha_j,\delta) = \min\left\{1, \frac{2\sin\delta}{r_j \sin\kappa}\right\}, \tag{16}$$

where

$$r_j = \begin{cases} \dfrac{\sin\alpha_j}{\sin\frac{\omega_1}{2}} & \text{if } \alpha_j < \dfrac{\omega_1}{2}, \\ 1 & \text{if } \alpha_j \geqslant \dfrac{\omega_1}{2}. \end{cases}$$

Of course, many other formulas for $\Phi(\alpha_j,\delta)$ can also be used. Let us mention that the computation of (16) is inexpensive since, denoting by $\boldsymbol{v}_j^\perp$ a unit vector orthogonal to $\boldsymbol{v}_j$, we have

$$\sin\kappa = \boldsymbol{v}\cdot\boldsymbol{v}_j, \quad \sin\frac{\omega_1}{2} = |\boldsymbol{v}\cdot\boldsymbol{v}_j^\perp|, \quad \sin\delta = |\boldsymbol{w}\cdot\boldsymbol{v}_j^\perp|,$$
$$\sin\alpha_j = |\boldsymbol{s}\cdot\boldsymbol{v}_j^\perp|.$$

Remark 1. The dependence of the constants C_i^K on $\boldsymbol{w}$ is also discontinuous if the orientation of $\boldsymbol{w}$ passes the direction of $\boldsymbol{s}$ (i.e., of $\boldsymbol{b}$). However, this does not seem to be important since, if $\boldsymbol{w} \approx \boldsymbol{s}$, we have $\boldsymbol{b}\cdot\nabla u_h|_K \approx 0$ and hence $A^K U \approx 0$ in (7).

6. Summary of the improved Mizukami–Hughes method

In this section we summarize the definitions of the constants C_i^K introduced in the previous sections. Let us consider any element $K \in \mathscr{T}_h$ and let a_1, a_2 and a_3 be its vertices. If $\boldsymbol{b} \neq \boldsymbol{0}$, we assume that $\boldsymbol{b}$ points into the vertex zone or the edge zone of a_1 (cf. Fig. 1) and we denote

$$\boldsymbol{s} = \frac{\boldsymbol{b}}{|\boldsymbol{b}|}, \quad \boldsymbol{v}_2 = \frac{a_2 - a_1}{|a_2 - a_1|}, \quad \boldsymbol{v}_3 = \frac{a_3 - a_1}{|a_3 - a_1|}, \quad \boldsymbol{v} = \frac{\boldsymbol{v}_2 + \boldsymbol{v}_3}{|\boldsymbol{v}_2 + \boldsymbol{v}_3|}.$$

Further, we introduce unit vectors $\boldsymbol{w}$, $\boldsymbol{v}^\perp$, $\boldsymbol{v}_2^\perp$ and $\boldsymbol{v}_3^\perp$ such that

$$\boldsymbol{w}\cdot\nabla u_h|_K = 0, \quad \boldsymbol{v}^\perp\cdot\boldsymbol{v} = 0, \quad \boldsymbol{v}_2^\perp\cdot\boldsymbol{v}_2 = 0, \quad \boldsymbol{v}_3^\perp\cdot\boldsymbol{v}_3 = 0,$$
$$\boldsymbol{w}\cdot\boldsymbol{v} \geqslant 0, \quad \boldsymbol{v}^\perp\cdot\boldsymbol{v}_3 \geqslant 0.$$

Finally, we recall the spaces V_2 and V_3 introduced in Section 2. Then the constants C_1^K, C_2^K and C_3^K are determined according to the algorithm in Fig. 8. It is obvious that the improved method preserves the general properties of the original Mizukami–Hughes method, particularly, it satisfies the discrete maximum principle discussed at the end of Section 2.

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

$$C_1^K = C_2^K = C_3^K = 0$$

ELSE IF $\boldsymbol{b} \in \mathrm{VZ}_1$ THEN

$$C_1^K = \tfrac{2}{3}, \qquad C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\overline{K} \cap \overline{\Gamma^D} \neq \emptyset$ THEN

$$C_1^K = C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\mathscr{T}_h$ is not of the type from Fig. 3(a) and all vertices of K are connected by edges to vertices on $\overline{\Gamma^D}$ THEN

$$C_1^K = C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\boldsymbol{b} \cdot \nabla u_h|_K = 0$ THEN

$$C_1^K = -\tfrac{1}{3}, \quad C_2^K = C_3^K = \tfrac{1}{6}$$

ELSE IF $V_2 \neq \emptyset$ & $V_3 = \emptyset$ THEN

$$C_2^K = \tfrac{2}{3}, \quad C_1^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $V_2 = \emptyset$ & $V_3 \neq \emptyset$ THEN

$$C_3^K = \tfrac{2}{3}, \quad C_1^K = C_2^K = -\tfrac{1}{3}$$

ELSE IF $\boldsymbol{w} \cdot \boldsymbol{v}^\perp < 0$ THEN

$$r_2 = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{v}_2^\perp|}{|\boldsymbol{v} \cdot \boldsymbol{v}_2^\perp|} + 1 - \mathrm{sgn}(\boldsymbol{b} \cdot \boldsymbol{v}_2)\right\},$$

$$\Phi = \min\left\{1, \frac{2\,|\boldsymbol{w} \cdot \boldsymbol{v}_2^\perp|}{r_2\, \boldsymbol{v} \cdot \boldsymbol{v}_2}\right\},$$

$$C_2^K = -\frac{1}{3} + \frac{1}{2}\,\Phi\left[1 + \frac{(\boldsymbol{v}_2 - \boldsymbol{v}_3) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}\right],$$

$$C_3^K = \frac{1}{3} - C_2^K, \qquad C_1^K = -\tfrac{1}{3}$$

ELSE

$$r_3 = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{v}_3^\perp|}{|\boldsymbol{v} \cdot \boldsymbol{v}_3^\perp|} + 1 - \mathrm{sgn}(\boldsymbol{b} \cdot \boldsymbol{v}_3)\right\},$$

$$\Phi = \min\left\{1, \frac{2\,|\boldsymbol{w} \cdot \boldsymbol{v}_3^\perp|}{r_3\, \boldsymbol{v} \cdot \boldsymbol{v}_3}\right\},$$

$$C_3^K = -\frac{1}{3} + \frac{1}{2}\,\Phi\left[1 + \frac{(\boldsymbol{v}_3 - \boldsymbol{v}_2) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}\right],$$

$$C_2^K = \frac{1}{3} - C_3^K, \qquad C_1^K = -\tfrac{1}{3}.$$

Fig. 8. Definition of the constants C_i^K in the improved Mizukami–Hughes method.

7. Numerical results

In this section we demonstrate the properties of the improved Mizukami–Hughes method by means of several standard test problems formulated in Examples 1–7 below and taken (in a slightly modified form) from [9,12,14]. In all these examples we consider $\varepsilon = 10^{-7}$ and, except for Example 6, $\Omega = (0,1)^2$. Unless otherwise specified, we use a triangulation of the type depicted in Fig. 3(a). The number of elements will be determined by the parameter N introduced at the beginning of Section 3. In Examples 1–3, the convection vector $\boldsymbol{b}$ is defined using an angle θ which is assumed to satisfy $\theta \in (0, \pi/2)$. To simplify the definitions of various parts of $\partial\Omega$, we introduce the sets

$$\Gamma_1 = (\{0\} \times (0,1]) \cup ([0,1) \times \{1\}),$$
$$\Gamma_2 = (\{0\} \times (0.7,1]) \cup ([0,1) \times \{1\}).$$

In the captions of figures we denote by MH the original Mizukami–Hughes method [14] and by IMH the improved Mizukami–Hughes method introduced in this paper. Let us mention that the discrete solutions obtained using the SUPG method [2] contain spurious oscillations for all the examples except for Example 6.

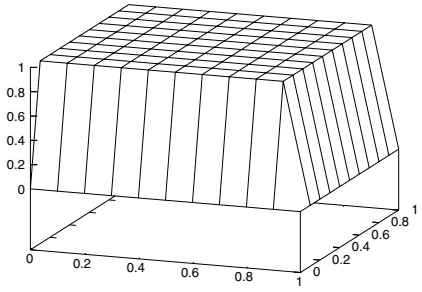


Fig. 9. Example 1, IMH, $N = 10$.

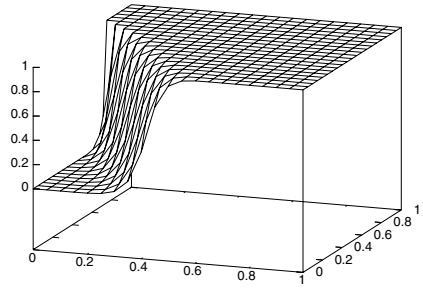


Fig. 10. Example 2, $\theta = \frac{\pi}{3}$, IMH, $N = 20$.

Example 1 (*Convection skew to the mesh with boundary layers*)

$$\boldsymbol{b} = (\cos\theta, -\sin\theta), \quad f = 0, \quad \Gamma^{\mathrm{D}} = \partial\Omega,$$
$$u_{\mathrm{b}} = 1 \quad \text{on } \Gamma_1, \quad u_{\mathrm{b}} = 0 \quad \text{on } \Gamma^{\mathrm{D}} \setminus \Gamma_1.$$

Both the original and the improved Mizukami–Hughes method give the same discrete solution which is nodally exact (cf. Fig. 9). This easily follows from the definition of the constants C_i^K. However, for $\theta \neq \pi/4$, it is rather difficult to compute the discrete solution of the original Mizukami–Hughes method due to the discontinuous dependence of C_i^K's on the orientation of ∇u_h. On the other hand, the computation of the discrete solution of the improved Mizukami–Hughes method needs only a few nonlinear iterations.

Example 2 (*Convection skew to the mesh with an inner layer*)

$$\boldsymbol{b} = (\cos\theta, -\sin\theta), \quad f = 0, \quad g = 0, \quad \Gamma^{\mathrm{D}} = \Gamma_1,$$
$$u_{\mathrm{b}} = 1 \quad \text{on } \Gamma_2, \quad u_{\mathrm{b}} = 0 \quad \text{on } \Gamma^{\mathrm{D}} \setminus \Gamma_2.$$

For $\theta = \pi/4$, the vector $\boldsymbol{b}$ points into vertex zones in all elements of the triangulation and it is easy to see that, for both methods, the discrete solution is constant along the diagonals in Fig. 3(a) if $\varepsilon \to 0$. Consequently, both the original and the improved Mizukami–Hughes method give the same nodally exact discrete solution. If $\theta \neq \pi/4$, the discrete solutions are not nodally exact but they are similar for both methods. Fig. 10 shows the discrete solution for $\theta = \pi/3$ and $N = 20$ obtained using the improved Mizukami–Hughes method. Fig. 11 compares the outflow profiles along the x-axis for the two methods and the exact solution of the hyperbolic limit of (1). The solution of the improved method seems to be slightly better. Like for the previous example, the discrete solution is much more difficult to compute for the original Mizukami–Hughes method.

Example 3 (*Convection skew to the mesh with inner and boundary layers*)

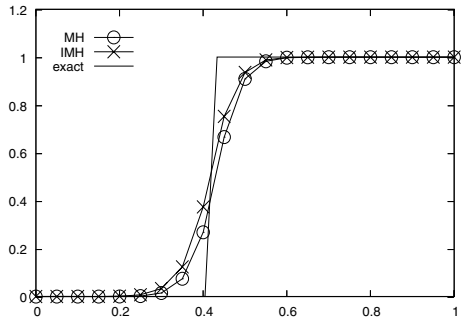


Fig. 11. Example 2, $\theta = \frac{\pi}{3}$, $N = 20$, MH, IMH and exact solution on $y = 0$.

$$\boldsymbol{b} = (\cos\theta, -\sin\theta), \quad f = 0, \quad \Gamma^{\mathrm{D}} = \partial\Omega,$$
$$u_{\mathrm{b}} = 1 \quad \text{on } \Gamma_2, \quad u_{\mathrm{b}} = 0 \quad \text{on } \Gamma^{\mathrm{D}} \setminus \Gamma_2.$$

This test problem is more complicated than the previous one since, in addition to the inner layer, it also involves one or two boundary layers. The relation between the original and the improved Mizukami–Hughes method is similar as in the previous example. Fig. 12 shows the discrete solution obtained using the improved Mizukami–Hughes method for $\theta = \pi/3$ and $N = 20$.

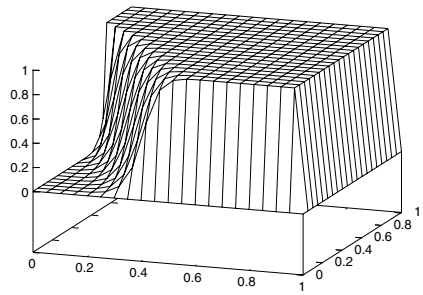


Fig. 12. Example 3, $\theta = \frac{\pi}{3}$, IMH, $N = 20$.

Example 4 (*Convection with a constant source term*)

$$\boldsymbol{b} = (1, \alpha), \quad \alpha \in (-0.5, 0.5), \quad f = 1, \quad \Gamma^{\mathrm{D}} = \partial\Omega, \quad u_{\mathrm{b}} = 0.$$

This problem was already considered in Section 3 where we have seen that the original Mizukami–Hughes method gives wrong discrete solutions (cf. Fig. 4). As we see from Fig. 13, the discrete solutions of the improved Mizukami–Hughes method seem to be correct in all cases considered in Section 3. Moreover, the improved Mizukami–Hughes method gives the discrete solutions shown in Fig. 13(a)–(c) also if we use a triangulation of the type depicted in Fig. 3(b).

Example 5 (*Convection with a nonconstant source term*)

$$\boldsymbol{b} = (1, \alpha), \quad \alpha \in (-0.5, 0.5), \quad \Gamma^{\mathrm{D}} = \partial\Omega, \quad u_{\mathrm{b}} = 0,$$
$$f = 1 \quad \text{in } (0, \tfrac{1}{2}) \times (0, 1), \quad f = -1 \quad \text{in } (\tfrac{1}{2}, 1) \times (0, 1).$$

Like in the previous example, both methods coincide and give a nodally exact solution for $\alpha = 0$ and a triangulation of the type depicted in Fig. 3(a). This is no longer true if we use $\alpha \neq 0$ or a triangulation of the type depicted in Fig. 3(b). Figs. 14 and 15 demonstrate that the original Mizukami–Hughes method generally gives wrong discrete solutions whereas the solutions of the improved Mizukami–Hughes method seem to be correct.

Example 6 (*Donut problem*). We consider $\Omega = (0,1)^2 \backslash \Gamma$ with $\Gamma = \{\frac{1}{2}\} \times (0, \frac{1}{2})$. The convection field $\boldsymbol{b}$ is defined by

$$\boldsymbol{b}(x, y) = (-y + \tfrac{1}{2}, x - \tfrac{1}{2})$$

so that it represents a vortex around the midpoint of the unit square in the counter-clockwise direction. Therefore, Γ represents an inflow boundary denoted by Γ^{in} if we approach Γ from the right but it also represents an outflow boundary denoted by Γ^{out} if we approach it from the left. We set

$$f = 0, \quad g = 0, \quad \Gamma^{\mathrm{D}} = \Gamma^{\mathrm{in}} \cup \partial[(0,1)^2], \quad \Gamma^{\mathrm{N}} = \Gamma^{\mathrm{out}},$$
$$u_{\mathrm{b}} = 0 \quad \text{on } \Gamma^{\mathrm{D}} \setminus \Gamma^{\mathrm{in}}, \quad u_{\mathrm{b}}(\tfrac{1}{2}, y) = \sin(\pi(1 - 2y)) \quad \text{for } y \in (0, \tfrac{1}{2}).$$

For this problem, an almost nodally exact discrete solution can be obtained using the SUPG method and it is interesting to see to what extent the discrete solution deteriorates if other stabilized methods are used. The solution of the improved Mizukami–Hughes discretization is shown in Fig. 16 and is similar to the solution obtained using the original Mizukami–Hughes method. Fig. 17 shows a comparison of the discrete solutions of the two Mizukami–Hughes methods and the exact solution of the hyperbolic limit of (1) by means of cuts through graphs of the solutions along the line $x = 1/2$. It seems that the improved

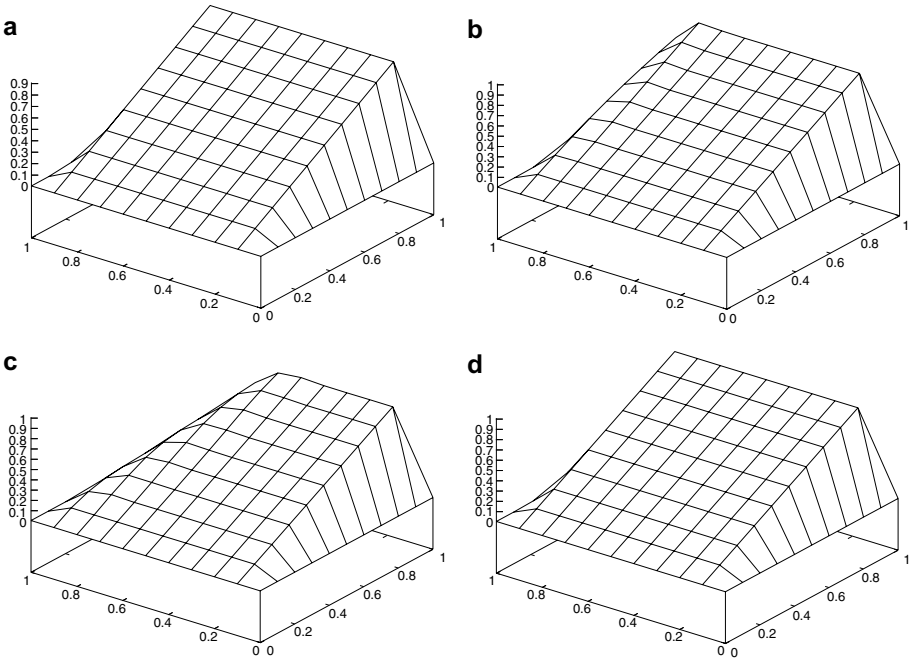


Fig. 13. Example 4, IMH, $N = 10$: (a) $\alpha = -0.0001$, (b) $\alpha = -0.1$, (c) $\alpha = -0.4$ and (d) $\alpha = 0$, $\mathscr{T}_h$ from Fig. 3(b).

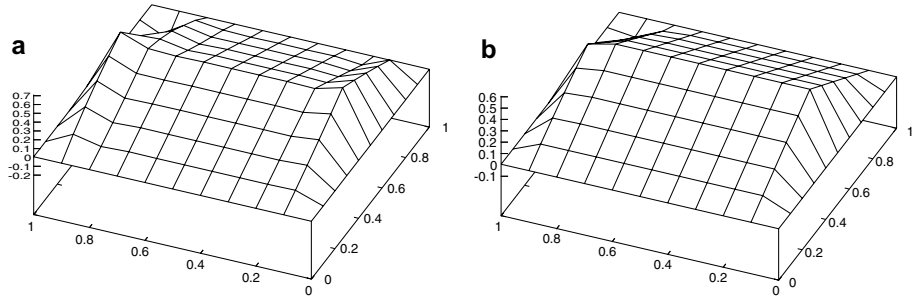


Fig. 14. Example 5, $\alpha = -0.1$, $N = 10$: (a) MH and (b) IMH.

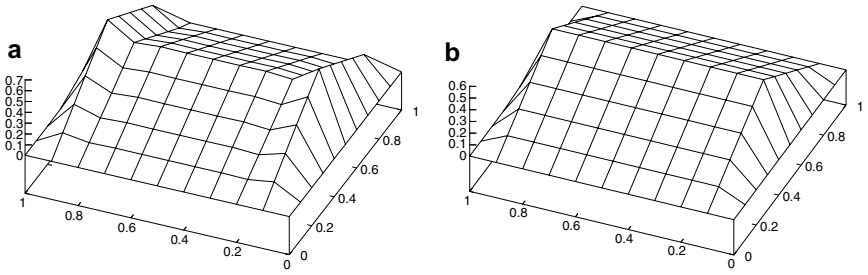


Fig. 15. Example 5, $\alpha = 0$, $\mathscr{T}_h$ from Fig. 3(b), $N = 10$: (a) MH and (b) IMH.

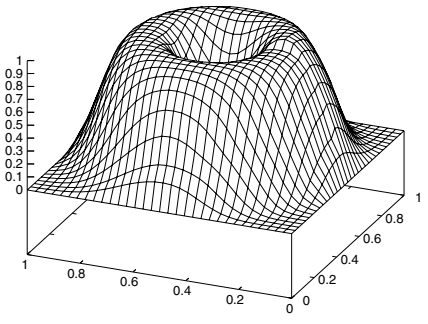


Fig. 16. Example 6, IMH, $N = 32$.

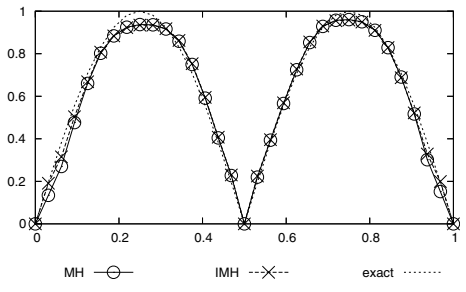


Fig. 17. Example 6, $N = 32$, MH, IMH and exact solution on $x = 1/2$.

Mizukami–Hughes method gives a slightly better solution. The two discrete solutions are comparable with best discrete solutions obtained using discontinuity-capturing methods mentioned in the introduction.

Example 7 (*Problem with known exact solution*)

$$\boldsymbol{b} = (2, 3), \quad \Gamma^{\mathrm{D}} = \partial\Omega.$$

The functions f and u_{b} are chosen in such a way that

$$u(x,y) = xy^2 - y^2 \exp\left(\frac{2(x-1)}{\varepsilon}\right) - x \exp\left(\frac{3(y-1)}{\varepsilon}\right) + \exp\left(\frac{2(x-1)+3(y-1)}{\varepsilon}\right)$$

is the exact solution of (1) and (2).

The function u contains two typical exponential boundary layers and hence this example represents a suitable tool for gauging the accuracy of numerical methods for the solution of convection–diffusion problems. The discrete solution obtained using the improved Mizukami–Hughes method for $N = 20$ can be seen in Fig. 18. Fig. 19 shows the discrete solution computed using the SUPG method [2] with the so-called optimal definition of the stabilization parameter and element size defined as the element diameter in the direction of the flow. We consider the SUPG method here since it is known to approximate solutions with layers on non-layer-adapted meshes at least outside the layers

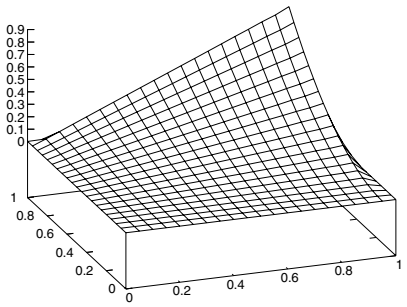


Fig. 18. Example 7, IMH, $N = 20$.

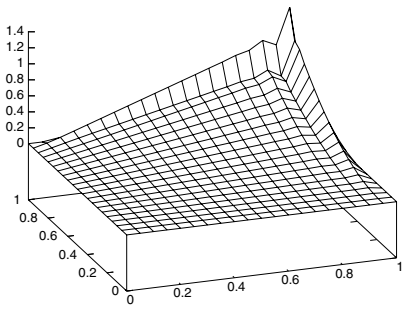


Fig. 19. Example 7, SUPG, $N = 20$.

very precisely. Therefore, it is interesting to compare the accuracy of the SUPG method with the accuracy of the improved Mizukami–Hughes method. We measure the errors of the discrete solutions by means of the norm in $L^2(\Omega)$ denoted by $\|\cdot\|_{0,\Omega}$ and using the discrete L^∞ norm on Ω denoted by $\|\cdot\|_{0,\infty,h}$ and defined as the maximum of the absolute values of errors at vertices of the triangulation. In addition, we consider this type of norms on the domain $\Omega^* \equiv (0, 0.8)^2$ which does not contain a neighborhood of the layers. The respective norms are denoted by $\|\cdot\|^*_{0,\Omega}$ and $|\cdot|^*_{0,\infty,h}$. Finally, we evaluate the $H^1(\Omega^*)$ seminorm denoted by $|\cdot|^*_{1,\Omega}$. Because of the boundary layers it makes no sense to show the $H^1(\Omega)$ seminorm. Like in the previous examples, the bilinear forms of the discrete problems were computed exactly whereas the right-hand sides were evaluated using quadrature formulas which are exact for piecewise cubic f. The evaluation of the L^2 norms (respectively, the H^1 seminorm) was exact for piecewise quadratic (respectively, cubic) functions. The obtained results are shown in Tables 1 and 2 and we see that, outside the layers, both methods converge with optimal convergence orders. On fine meshes, the SUPG method is more precise in Ω^* than the modified Mizukami–Hughes method, particularly, with respect to the discrete L^∞ norm. However, on the whole domain Ω, the SUPG solution does not converge in the discrete L^∞ norm since the magnitude

Table 1
Example 7, errors of the improved Mizukami–Hughes method

N	$\|\cdot\|_{0,\Omega}$	$\|\cdot\|_{0,\infty,h}$	$\|\cdot\|^*_{0,\Omega}$	$\|\cdot\|^*_{1,\Omega}$	$\|\cdot\|^*_{0,\infty,h}$
20	$5.91-2$	$7.02-3$	$3.68-4$	$2.05-2$	$2.15-3$
40	$4.20-2$	$3.93-3$	$1.13-4$	$1.02-2$	$6.71-4$
80	$2.98-2$	$2.07-3$	$3.14-5$	$5.06-3$	$1.87-4$
160	$2.11-2$	$1.05-3$	$8.30-6$	$2.52-3$	$4.94-5$
Order	0.50	0.98	1.92	1.01	1.92

Table 2
Example 7, errors of the SUPG method

N	$\|\cdot\|_{0,\Omega}$	$\|\cdot\|_{0,\infty,h}$	$\|\cdot\|^*_{0,\Omega}$	$\|\cdot\|^*_{1,\Omega}$	$\|\cdot\|^*_{0,\infty,h}$
20	$4.91-2$	$5.08-1$	$3.33-4$	$2.49-2$	$9.37-3$
40	$3.51-2$	$5.70-1$	$3.95-5$	$1.00-2$	$2.32-4$
80	$2.50-2$	$6.02-1$	$9.80-6$	$4.99-3$	$7.06-6$
160	$1.78-2$	$6.18-1$	$2.45-6$	$2.49-3$	$1.74-6$
Order	0.49	−0.04	2.00	1.00	2.02

of the spurious oscillations visible in Fig. 19 does not decrease for decreasing h as long as h is significantly larger than the width of the boundary layers. On the other hand, the solution of the modified Mizukami–Hughes method converges on the whole domain Ω with first order of accuracy in the discrete L^∞ norm and does not contain any spurious oscillations as we can also see from Fig. 18.

8. Application of the Mizukami–Hughes method to convection–diffusion–reaction equations

In this section we extend the Mizukami–Hughes method described in the preceding sections to convection–diffusion–reaction equations

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + cu = f \quad \text{in } \Omega, \tag{17}$$

where c is a given function. Our aim again is to derive a numerical method satisfying the discrete maximum principle and hence we shall assume that $c \geqslant 0$ since otherwise no maximum principle generally holds for Eq. (17). Again, we consider the singularly perturbed case, i.e., $\varepsilon \ll |\boldsymbol{b}| + c$.

The discrete solution u_h of (17), (2) is defined by

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

Like for $\boldsymbol{b}$, we assume that c is piecewise constant.

For any $K \in \mathscr{T}_h$, the local reaction matrix R^K has entries

$$r^K_{ij} = (c\varphi_j, \tilde{\varphi}_i)_K = \tfrac{1}{3}c|_K \mathrm{meas}_2(K)\left(\tfrac{1}{4} + C^K_i + \tfrac{1}{4}\delta_{ij}\right)$$

(with $i = 1, \ldots, M_h$, $j = 1, \ldots, N_h$, $a_i, a_j \in \overline{K}$), where δ_{ij} is the Kronecker symbol. We define the matrix $S^K \equiv A^K + R^K$ (with entries s^K_{ij}), where A^K is the local convection matrix

introduced in Section 2. Like before, we want to define the constants C_i^K in such a way that the matrix S^K is of nonnegative type or at least satisfies an analogue of (7), i.e.,

$$S^K U = \widetilde{S}^K U, \tag{18}$$

where $\widetilde{S}^K$ is a matrix of nonnegative type. Note that

$$\sum_{\substack{j=1\\ a_j\in\overline{K}}}^{N_h} s_{ij}^K = \sum_{\substack{j=1\\ a_j\in\overline{K}}}^{N_h} r_{ij}^K = c|_K \mathrm{meas}_2(K)\left(\tfrac{1}{3}+C_i^K\right)$$
$$\forall i\in\{1,\dots,M_h\},\ \ a_i\in\overline{K},$$

and hence the first condition in (3) is necessary for S^K to be of nonnegative type.

On the other hand, the second condition in (3) cannot be fulfilled in general. To see this, let us denote the vertices of K by a_1, a_2, a_3 and let us assume that $\boldsymbol{b}\in \mathrm{VZ}_1$ (from now on we shall write $\boldsymbol{b}$, c, $\nabla\varphi_i$ instead of $\boldsymbol{b}|_K$, $c|_K$, $\nabla\varphi_i|_K$, respectively). Then s_{21}^K and s_{31}^K may be nonpositive only if $C_2^K < -\frac{1}{4}$ and $C_3^K < -\frac{1}{4}$. A necessary condition for s_{12}^K and s_{13}^K to be nonpositive is

$$\tfrac{2}{3}c\left(\tfrac{1}{4}+C_1^K\right) \leqslant \boldsymbol{b}\cdot\nabla\varphi_1\left(\tfrac{1}{3}+C_1^K\right).$$

If $C_1^K \in (\frac{1}{2},\frac{2}{3}]$, which is necessary for the validity of (3), this inequality will not be satisfied for $c \geqslant 2\boldsymbol{b}\cdot\nabla\varphi_1$. Hence the validity of the second condition in (3) cannot be generally required.

Fortunately, the second condition in (3) is not needed to assure that (7) holds with a matrix $\widetilde{A}^K$ of nonnegative type. It is easy to check that (7) still holds if those constants in the definition of A^K, for which larger values than $-\frac{1}{3}$ are prescribed, are replaced by any values from the interval $[-\frac{1}{3},\infty)$. Thus, our idea is first to compute the constants C_i^K according to the algorithm in Fig. 8 and then possibly to decrease some of the constants in such a way that (18) holds with a matrix $\widetilde{S}^K$ of nonnegative type. Since, for $c>0$, the matrix R^K is of nonnegative type if and only if all the constants C_i^K are from the interval $\left[-\frac{1}{3},-\frac{1}{4}\right]$, a constant C_i^K provided by the algorithm in Fig. 8 will not be decreased if $C_i^K \leqslant -\frac{1}{4}$. If $C_i^K > -\frac{1}{4}$, it is never necessary to decrease this constant below the value $-\frac{1}{4}$.

Now let us describe the new definition of the constants C_i^K in detail. We again denote the vertices of K by a_1, a_2, a_3 and assume that $\boldsymbol{b}\in\mathrm{VZ}_1$. Then, according to Fig. 8, $C_2^K = C_3^K = -\frac{1}{3}$ and hence we only have to assure that s_{12}^K and s_{13}^K are nonpositive, which is the case if and only if

$$36(\boldsymbol{b}\cdot\nabla\varphi_j+\tfrac{1}{3}c)(\tfrac{1}{3}+C_1^K) \leqslant c, \quad j=2,3. \tag{19}$$

Of course, the constant $C_1^K = \frac{2}{3}$ provided by the algorithm in Fig. 8 generally does not satisfy this inequality. Therefore, denoting

$$\xi = 36\max\{0, \boldsymbol{b}\cdot\nabla\varphi_2+\tfrac{1}{3}c,\ \boldsymbol{b}\cdot\nabla\varphi_3+\tfrac{1}{3}c\},$$

we set

$$C_1^K := \min\left\{\frac{2}{3}, -\frac{1}{3}+\frac{c}{\xi}\right\}$$

(if $c=\xi=0$, we define $c/\xi=\infty$). Since $\boldsymbol{b}\cdot\nabla\varphi_j\leqslant 0$ for $j=2,3$, we really have $C_1^K \geqslant -\frac{1}{4}$.

Now let us assume that $\boldsymbol{b}\in\mathrm{EZ}_1$ and that K does not have the properties formulated in (A3) and (A3*) at the end of Section 3. If $\boldsymbol{b}\cdot\nabla u_h|_K = 0$, then $A^K U = 0$ and we set $C_1^K = -\frac{1}{3}$ and $C_2^K = C_3^K = -\frac{1}{4}$, which guarantees that the matrix R^K is of nonnegative type. Let $\boldsymbol{b}\cdot\nabla u_h|_K \neq 0$ and let the vector $\boldsymbol{w}$ be defined like in Section 6. It is convenient to denote for $\alpha\in\mathbb{R}$ and $j,k\in\{1,2,3\}$, $j\neq k$,

$$\xi_j(\alpha) = 36(\boldsymbol{b}\cdot\nabla\varphi_j + \alpha\boldsymbol{w}\cdot\nabla\varphi_j+\tfrac{1}{3}c),$$
$$\xi_{jk}(\alpha) = \max\{0,\xi_j(\alpha),\xi_k(\alpha)\}.$$

Let us first assume that $V_2\neq\emptyset$ and $V_3=\emptyset$. Then $C_1^K = C_3^K = -\frac{1}{3}$ and, like in Section 2, we deduce that (18) holds with $\widetilde{S}^K = \left(\frac{1}{3}+C_2^K\right)\widetilde{A}^{K,2}+R^K$. The matrix $\widetilde{A}^{K,2}$ was defined in Section 2 using an arbitrarily chosen $\alpha_2\in V_2$ and its first and third row consist of zeros. Therefore, we only have to assure that the entries $\widetilde{s}_{21}^K$ and $\widetilde{s}_{23}^K$ of the matrix $\widetilde{S}^K$ are nonpositive *for some* $\alpha_2\in V_2$. Like in (19), we get the condition that, for some $\alpha\in V_2$,

$$\xi_j(\alpha)\left(\tfrac{1}{3}+C_2^K\right) \leqslant c, \quad j=1,3.$$

The set V_2 is a closed interval and hence it is easy to compute

$$\xi = \min_{\alpha\in V_2}\xi_{13}(\alpha).$$

Thus, it suffices to set

$$C_2^K := \min\left\{\frac{2}{3}, -\frac{1}{3}+\frac{c}{\xi}\right\}.$$

Since $\xi_{13}(\alpha)\leqslant 12c$ for any $\alpha\in V_2$, we again have $C_2^K\geqslant -\frac{1}{4}$. The case $V_2=\emptyset$, $V_3\neq\emptyset$ is treated analogously.

If both V_2 and V_3 are nonempty, then $C_1^K = -\frac{1}{3}$ but the constants C_2^K and C_3^K provided by the algorithm in Fig. 8 may be so large that (18) does not hold for any matrix $\widetilde{S}^K$ of nonnegative type. Therefore, like above, we set

$$C_2^K := \min\left\{C_2^K, -\frac{1}{3}+\frac{c}{\xi}\right\}, \quad C_3^K := \min\left\{C_3^K, -\frac{1}{3}+\frac{c}{\xi'}\right\},$$

where

$$\xi = \min_{\alpha\in V_2}\xi_{13}(\alpha), \quad \xi' = \min_{\alpha\in V_3}\xi_{12}(\alpha).$$

Up to now, we have not mentioned the case when $\boldsymbol{b}=\boldsymbol{0}$ and hence $A^K=0$. We set $C_1^K = C_2^K = C_3^K = -\frac{1}{4}$, which leads to a matrix S^K with positive diagonal entries and zero off-diagonal entries.

The above modifications of the constants C_i^K assure that the discrete solution of (17), (2) always satisfies (18) with a matrix $\widetilde{S}^K$ of nonnegative type. Therefore (see the end of Section 2), the discrete solution satisfies the discrete maximum principle and hence it does not contain any spurious oscillations.

Let us illustrate the properties of the improved Mizukami–Hughes method with the above described definition of the constants C_i^K by means of two simple test problems

taken from [10,16]. Like in Section 7, we consider $\varepsilon = 10^{-7}$, $\Omega = (0,1)^2$ and triangulations of the type depicted in Fig. 3(a).

Example 8 (*Reaction without convection*)

$$\boldsymbol{b} = \boldsymbol{0}, \quad c = 1, \quad f = 1, \quad \Gamma^{\mathrm{D}} = \partial\Omega, \quad u_{\mathrm{b}} = 0.$$

Fig. 20 shows a discrete solution computed using the Galerkin discretization (corresponding to the Mizukami–Hughes method with all C_i^K's equal to zero) and we observe significant spurious oscillations along the whole boundary of Ω. On the other hand, the improved Mizukami–Hughes method gives a nodally exact discrete solution, see Fig. 21.

Example 9 (*Reaction with convection*)

$$\boldsymbol{b}(x,y) = (1-y^2, 0), \quad c = 25, \quad f = 0, \quad \Gamma^{\mathrm{D}} = \{0\} \times (0,1), \quad g = 0, \quad u_{\mathrm{b}} = 1.$$

The Galerkin solution (cf. Fig. 22) again exhibits spurious oscillations which become even larger if the SUPG method described in the previous section is applied. The discrete solution obtained using the improved Mizukami–Hughes method with C_i^K's defined by the algorithm in Fig. 8 is comparable with the SUPG solution. However, using the constants C_i^K introduced in this section, we obtain

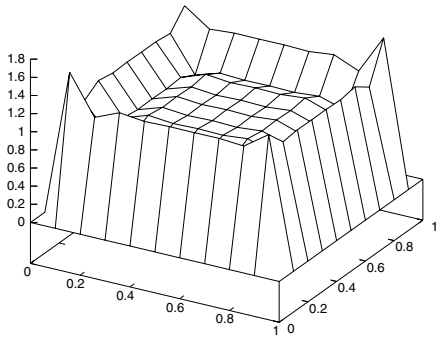


Fig. 20. Example 8, Galerkin, $N = 10$.

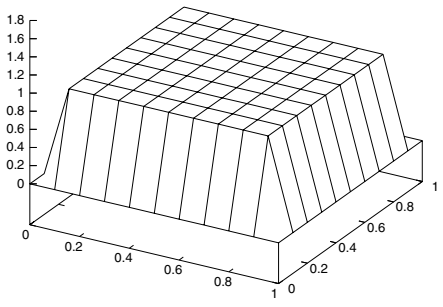


Fig. 21. Example 8, IMH, $N = 10$.

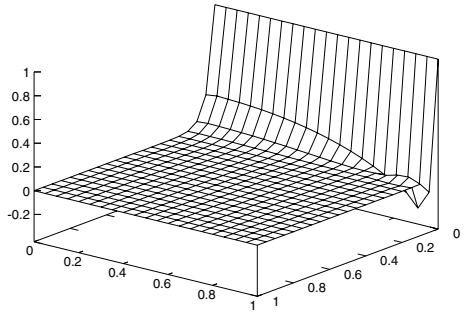


Fig. 22. Example 9, Galerkin, $N = 20$.

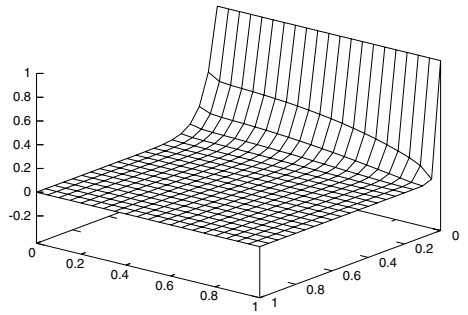


Fig. 23. Example 9, IMH, $N = 20$.

the discrete solution depicted in Fig. 23 where no spurious oscillations are present.

Remark 2. If the assumption that $\varepsilon \ll |\boldsymbol{b}| + c$ is not satisfied and the reaction term dominates the convection term (in particular, if $\boldsymbol{b} = \boldsymbol{0}$), the invalidity of the second condition in (3) may lead to a large error of the discrete solution. Therefore, in this case, instead of requiring that S^K or $\widetilde{S}^K$ be of nonnegative type, one should require this property of the matrix $D^K + S^K$ or $D^K + \widetilde{S}^K$, respectively, where D^K is the local diffusion matrix with entries $d_{ij}^K = \varepsilon(\nabla\varphi_j, \nabla\varphi_i)_K$. Assuming that D^K has three rows (since otherwise the second condition in (3) can always be fulfilled), there is at least one row of D^K whose all entries are different from zero. Therefore, adding D^K to S^K or $\widetilde{S}^K$ always enables to increase at least one of the constants C_i^K. In this way, the second condition in (3) can often be (nearly) satisfied since $d_{ij}^K \approx \varepsilon$ whereas $r_{ij}^K \approx c\ \mathrm{meas}_2(K)$.

9. The Mizukami–Hughes method in three dimensions

In this section, we briefly show how the ideas presented in Section 2 can be applied to the three-dimensional case.

We assume that Ω is a bounded three-dimensional domain with a polyhedral boundary $\partial\Omega$ and that we are given a triangulation $\mathcal{T}_h$ of Ω consisting of a finite number

of open tetrahedral elements K. The notation, assumptions and concepts introduced in Section 1 and at the beginning of Section 2 (by the end of the definition of the discrete solution) can be extended in a natural way to the three-dimensional case and hence we shall not mention them again.

Analogously as in Section 2, the local convection matrices A^K have entries

$$a_{ij}^K = (\boldsymbol{b}\cdot\nabla\varphi_j, \tilde{\varphi}_i)_K = (\boldsymbol{b}\cdot\nabla\varphi_j)|_K \mathrm{meas}_3(K)(\tfrac{1}{4} + C_i^K),$$

$i = 1,\dots,M_h$, $j = 1,\dots,N_h$, $a_i, a_j \in \overline{K}$. Therefore, we shall require that the constants C_i^K satisfy

$$C_i^K \geqslant -\tfrac{1}{4}\ \forall i \in \{1,\dots,N_h\},\ a_i \in \overline{K}, \quad \sum_{\substack{i=1\\ a_i\in\overline{K}}}^{N_h} C_i^K = 0. \tag{20}$$

Let K be any element of the triangulation $\mathscr{T}_h$ and let the vertices of K be a_1, a_2, a_3 and a_4. We divide the space $\mathbb{R}^3$ into 14 sets whose boundaries are formed by the four planes containing the barycentre a_c of K which are parallel to the faces of K. We denote these sets as vertex zones VZ_i, face zones FZ_i and edge zones EZ_{ij}, $i,j \in I$, $i<j$, where we used the index set $I = \{1,2,3,4\}$ for brevity. Precisely, the sets are defined in the following way:

$$\begin{aligned}
\mathrm{VZ}_i &= \{x\in\mathbb{R}^3; (x-a_c)\cdot\nabla\varphi_i > 0, (x-a_c)\cdot\nabla\varphi_k \leqslant 0 \\
&\qquad \forall k \in I\setminus\{i\}\},\\
\mathrm{FZ}_i &= \{x\in\mathbb{R}^3; (x-a_c)\cdot\nabla\varphi_i < 0, (x-a_c)\cdot\nabla\varphi_k \geqslant 0 \\
&\qquad \forall k \in I\setminus\{i\},\\
&\qquad \exists l \in I\setminus\{i\} : (x-a_c)\cdot\nabla\varphi_k > 0\ \forall k \in I\setminus\{i,l\}\},\\
\mathrm{EZ}_{ij} &= \{x\in\mathbb{R}^3; (x-a_c)\cdot\nabla\varphi_i > 0, (x-a_c)\cdot\nabla\varphi_j > 0,\\
&\qquad (x-a_c)\cdot\nabla\varphi_k < 0\ \forall k \in I\setminus\{i,j\}\}.
\end{aligned}$$

Again, we write $\nabla\varphi_i$ instead of $\nabla\varphi_i|_K$ for simplicity. Note that

$$\left(\bigcup_{i\in I}\mathrm{VZ}_i\right) \cup \left(\bigcup_{i\in I}\mathrm{FZ}_i\right) \cup \left(\bigcup_{i,j\in I, i<j}\mathrm{EZ}_{ij}\right) = \mathbb{R}^3\setminus\{a_c\}$$

and that all the 14 sets are mutually disjoint.

To get a better impression of the form of these sets, we introduce the points

$$u_{ij} = \frac{3a_i + a_j}{4}, \quad v_{ij} = \frac{a_i + \sum_{k\in I\setminus\{j\}} a_k}{4}, \quad i,j\in I,\ i\neq j.$$

Obviously, a point u_{ij} lies on the edge of K with end points a_i, a_j and a point v_{ij} lies on the face of K opposite the vertex a_j. It is easy to verify that the closure of $K\cap\mathrm{VZ}_i$ is a parallelepiped whose eight vertices are a_c, a_i, u_{ik}, v_{ik}, $k\in I\setminus\{i\}$, the closure of $K\cap\mathrm{FZ}_i$ is a tetrahedron whose four vertices are a_c, v_{ki}, $k\in I\setminus\{i\}$, and the closure of $K\cap\mathrm{EZ}_{ij}$ is a polyhedron with seven vertices a_c, u_{ij}, u_{ji}, v_{ik}, v_{jk}, $k\in I\setminus\{i,j\}$. Examples of an edge zone, a face zone and a vertex zone can be seen in Fig. 24. Note that, for any $k\in I$, all the nine points u_{ik} and v_{ij} with $i,j\in I\setminus\{k\}$, $i\neq j$, are contained in the plane containing a_c and being parallel to the face of K opposite a_k.

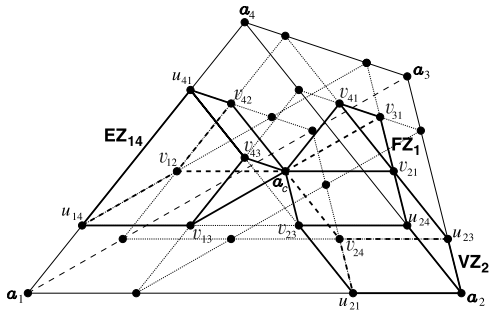


Fig. 24. Definition of edge zones, face zones and vertex zones.

Now let us discuss the definition of the constants $C_1^K,\dots,C_4^K$. If $\boldsymbol{b}$ points into a vertex zone, say VZ_j, $j\in I$, then (20) holds and A^K is of nonnegative type for

$$C_j^K = \tfrac{3}{4}, \quad C_k^K = -\tfrac{1}{4} \quad \forall k \in I\setminus\{j\}.$$

If A^K has four rows, this is the only possibility how to choose these constants.

Now let us assume that $\boldsymbol{b}$ does not point into any of the vertex zones. Then the constants C_i^K cannot be generally defined in such a way that (20) holds and the matrix A^K is of nonnegative type. Therefore, like in Section 2, we shall try to find such constants C_i^K, that the coefficient vector $U\in\mathbb{R}^4$ of $u_h|_K$ with respect to the basis $\{\varphi_i|_K\}_{i\in I}$ satisfies

$$A^K U = \widetilde{A}^K U, \tag{21}$$

where $\widetilde{A}^K$ is a matrix of nonnegative type. This is trivially satisfied if $\boldsymbol{b}\cdot\nabla u_h|_K = 0$ and hence we shall assume that $\boldsymbol{b}\cdot\nabla u_h|_K \neq 0$ in the following. Similarly as in Section 2, we introduce the sets

$$V_k = \{\tilde{\boldsymbol{b}}\in\mathbb{R}^3; (\tilde{\boldsymbol{b}}-\boldsymbol{b})\cdot\nabla u_h|_K = 0,\ a_c + \tilde{\boldsymbol{b}} \in \mathrm{VZ}_k\}, \quad k\in I.$$

If $\boldsymbol{b}$ points into the face zone FZ_j, $j\in I$, there exists $k\in I\setminus\{j\}$ such that $V_k\neq\emptyset$ and we may consider any constants C_i^K satisfying (20) and the following requirements:

$$V_k \neq \emptyset \quad \forall k\in I\setminus\{j\} \quad\Rightarrow\quad C_j^K = -\tfrac{1}{4}, \tag{22}$$

$$\exists k\in I\setminus\{j\} : V_k = \emptyset \quad\text{and}\quad V_l\neq\emptyset \quad \forall l\in I\setminus\{j,k\} \quad\Rightarrow\quad C_j^K = C_k^K = -\tfrac{1}{4}, \tag{23}$$

$$\exists k\in I\setminus\{j\} : V_k \neq \emptyset \quad\text{and}\quad V_l = \emptyset \quad \forall l\in I\setminus\{j,k\} \quad\Rightarrow\quad C_l^K = -\tfrac{1}{4} \quad \forall l\in I\setminus\{k\}. \tag{24}$$

If $\boldsymbol{b}$ points into the edge zone EZ_{jk}, $j,k\in I$, $j<k$, then $V_j\cup V_k\neq\emptyset$ and we consider any constants C_i^K satisfying (20) and the following requirements:

$$V_j\neq\emptyset \quad\text{and}\quad V_k\neq\emptyset \quad\Rightarrow\quad C_l^K = -\tfrac{1}{4} \quad \forall l\in I\setminus\{j,k\}, \tag{25}$$

$$V_j\neq\emptyset \quad\text{and}\quad V_k = \emptyset \quad\Rightarrow\quad C_l^K = -\tfrac{1}{4} \quad \forall l\in I\setminus\{j\}, \tag{26}$$

$$V_j = \emptyset \quad\text{and}\quad V_k\neq\emptyset \quad\Rightarrow\quad C_l^K = -\tfrac{1}{4} \quad \forall l\in I\setminus\{k\}. \tag{27}$$

Let us assume that the constants C_i^K are defined according to (20) and (22)–(27) and let us introduce vectors $\tilde{\boldsymbol{b}}_1, \ldots, \tilde{\boldsymbol{b}}_4$ such that, for any $i \in I$,

$$\tilde{\boldsymbol{b}}_i = \boldsymbol{b} \quad \text{if } C_i^K = -\tfrac{1}{4}, \quad \tilde{\boldsymbol{b}}_i \in V_i \quad \text{if } C_i^K > -\tfrac{1}{4}.$$

We define a matrix $\widetilde{A}^K$ with entries

$$\tilde{a}_{ij}^K = (\tilde{\boldsymbol{b}}_i \cdot \nabla \varphi_j)|_K \mathrm{meas}_3(K)(\tfrac{1}{4} + C_i^K), \quad i, j \in I, \ a_i \in \Omega \cup \Gamma^{\mathrm{N}}.$$

Then $\widetilde{A}^K$ is of nonnegative type and (21) holds.

There are many possibilities how to satisfy the requirements (20) and (22)–(27) and, since (21) always holds with a matrix $\widetilde{A}^K$ of nonnegative type, the discrete solution u_h always satisfies the discrete maximum principle. However, not every choice of the constants C_i^K satisfying (20) and (22)–(27) is appropriate and we may encounter similar difficulties like those ones discussed in Sections 3–5. The derivation of suitable formulas for the constants C_i^K will be a subject of our further research.

10. Conclusions

In this paper we introduced several improvements of the Mizukami–Hughes method for the numerical solution of two-dimensional steady convection–diffusion equations. We have shown that the improved method satisfies the discrete maximum principle and we demonstrated by means of various numerical results that it gives very accurate discrete solutions with no spurious oscillations. Moreover, our extensive numerical tests (which will be published in a separate paper) revealed that none of the discontinuity-capturing methods mentioned in the introduction can be regarded as superior to the improved Mizukami–Hughes method. Therefore, the improved Mizukami–Hughes method seems to be one of the best choices for solving the problem (1) and (2) using conforming piecewise linear triangular finite elements if convection strongly dominates diffusion. We have also shown that the Mizukami–Hughes method can be extended to convection–diffusion–reaction equations and to the three-dimensional case but here further research is necessary.

Acknowledgements

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.

References

[1] R.C. Almeida, 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, 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, A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation, Comput. Methods Appl. Mech. Engrg. 191 (2002) 3833–3855.
[4] E. Burman, A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence, Math. Comput. 74 (2005) 1637–1652.
[5] E. Burman, P. Hansbo, Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems, Comput. Methods Appl. Mech. Engrg. 193 (2004) 1437–1453.
[6] E.G.D. do Carmo, 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.
[7] P.G. Ciarlet, P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, Comput. Methods Appl. Mech. Engrg. 2 (1973) 17–31.
[8] 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.
[9] T.J.R. Hughes, M. Mallet, A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Methods Appl. Mech. Engrg. 54 (1986) 341–355.
[10] S. Idelsohn, N. Nigro, M. Storti, G. Buscaglia, A Petrov–Galerkin formulation for advection–reaction–diffusion problems, Comput. Methods Appl. Mech. Engrg. 136 (1996) 27–46.
[11] 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.
[12] V. John, J.M. Maubach, L. Tobiska, Nonconforming streamline-diffusion-finite-element-methods for convection–diffusion problems, Numer. Math. 78 (1997) 165–188.
[13] T. Knopp, G. Lube, G. Rapin, Stabilized finite element methods with shock capturing for advection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 191 (2002) 2997–3013.
[14] A. Mizukami, 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.
[15] Y.-T. Shih, H.C. Elman, Modified streamline diffusion schemes for convection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 174 (1999) 137–151.
[16] T.E. Tezduyar, Y.J. Park, Discontinuity-capturing finite element formulations for nonlinear convection–diffusion–reaction equations, Comput. Methods Appl. Mech. Engrg. 59 (1986) 307–325.

J Sci Comput (2010) 43: 454–470
DOI 10.1007/s10915-008-9260-2

Numerical Solution of Convection–Diffusion Equations Using a Nonlinear Method of Upwind Type

Petr Knobloch

Received: 30 October 2007 / Revised: 21 November 2008 / Accepted: 2 December 2008 /
Published online: 17 December 2008

Abstract This paper is devoted to the numerical solution of convection–diffusion equations using the Mizukami–Hughes method which is a nonlinear method of upwind type using conforming piecewise linear triangular finite elements. We extend this method to the whole range of the diffusion parameter whereas the original method was introduced for the convection-dominated regime only. We prove that the extended method satisfies the discrete maximum principle and illustrate its properties by means of numerical results.

Keywords Finite element method · Convection–diffusion equations · Upwinding · Mizukami–Hughes method · Discrete maximum principle

1 Introduction

This paper is devoted to the numerical solution of the scalar convection–diffusion problem

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega, \qquad u = u_b \quad \text{on } \Gamma^D, \qquad \varepsilon\frac{\partial u}{\partial \boldsymbol{n}} = g \quad \text{on } \Gamma^N. \tag{1}$$

Here Ω is a bounded two-dimensional domain with a polygonal boundary $\partial\Omega$ and Γ^D, Γ^N are disjoint and relatively open subsets of $\partial\Omega$ satisfying $\mathrm{meas}_1(\Gamma^D) > 0$ and $\overline{\Gamma^D \cup \Gamma^N} = \partial\Omega$. Further, $\boldsymbol{n}$ is the outward unit normal vector to $\partial\Omega$, $\varepsilon > 0$ is a constant diffusivity, $\boldsymbol{b}$ is the flow velocity, f is a given outer source of the unknown scalar quantity u, and u_b, g are given functions.

It is well known that the numerical solution of (1) is a challenging task since convection often dominates diffusion and hence the solution of (1) typically contains narrow inner and boundary layers. Discrete solutions of (1) are then often polluted by spurious oscillations. Therefore, many various stabilized methods have been developed during the past decades

P. Knobloch (✉)
Faculty of Mathematics and Physics, Department of Numerical Mathematics, Charles University, Sokolovská 83, 186 75 Prague 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

J Sci Comput (2010) 43: 454–470 455

and, in the context of finite element methods, these approaches can be usually interpreted as the addition of artificial diffusion to the standard Galerkin discretization, which loses its stability in the convection-dominated regime. The basic difficulty is that the amount of the artificial diffusion should be not too small to remove spurious oscillations but also not too large to avoid excessive smearing of the discrete solution.

Among stabilized finite element methods for the numerical solution of (1), upwinding techniques are attractive since they often satisfy the discrete maximum principle. The first upwind finite element method for which the discrete maximum principle and convergence were proved was developed in [18]. A drawback of upwind finite element methods is that they usually introduce too much artificial diffusion. An exception is the Mizukami–Hughes method [16] which uses conforming triangular piecewise linear finite elements and provides very accurate discrete solutions in the convection-dominated case. Numerical computations indicate that it is more accurate than many other stabilization approaches, see, e.g., [10, 11, 14]. The price we pay for the high accuracy of the Mizukami–Hughes method is that the method is nonlinear.

The original approach by Mizukami and Hughes was further improved in [13] where also extensions to convection–diffusion–reaction equations and to three space dimensions were presented. An extension of the Mizukami–Hughes method to bilinear finite elements was proposed in [15].

All variants of the Mizukami–Hughes method were designed for the strongly convection-dominated regime. Consequently, if the convection is not so dominant, the method usually introduces too much artificial diffusion. The aim of the present paper is to correct this shortcoming, which leads to a method that can be viewed as a partial upwind scheme, see [7]. Moreover, we give a rigorous proof of the discrete maximum principle, which is not available in the preceding papers on the Mizukami–Hughes method.

First, in the next section, we introduce some notation, formulate the SUPG method (which will be compared with the Mizukami–Hughes method in Sect. 5) and discuss the numerical solution of problem (1) by means of stabilized finite element methods. In particular, this section shows the insufficiency of many common approaches for the numerical solution of (1). Then, in Sect. 3, the original method by Mizukami and Hughes is formulated and, in Sect. 4, the improvements introduced in [13] are summarized. In Sect. 5, we discuss the upwinding properties of the Mizukami–Hughes method and introduce a modification of the method which does not change its properties in the strongly convection-dominated regime but improves the accuracy if the ratio between convection and diffusion effects is moderate. In addition, we prove in Sect. 5 that the discrete maximum principle still holds. Section 6 contains numerical results comparing both versions of the improved Mizukami–Hughes method and, finally, in Sect. 7, we present our conclusions.

2 SUPG Method and SOLD Methods

Let $\mathcal{T}_h$ be a triangulation of Ω consisting of a finite number of open triangular elements K. 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 $\mathcal{T}_h$ which lies on $\partial\Omega$ is contained either in $\overline{\Gamma^D}$ or in $\overline{\Gamma^N}$.

We shall discretize the problem (1) using the finite element space

$$V_h = \{v \in C(\overline{\Omega});\ v|_K \in P_1(K)\ \forall K \in \mathcal{T}_h\}$$

consisting of continuous piecewise linear functions. Let $a_1, \ldots, a_{M_h}$ be the vertices of $\mathcal{T}_h$ lying in $\Omega \cup \Gamma^N$ and let $a_{M_h+1}, \ldots, a_{N_h}$ be the vertices of $\mathcal{T}_h$ lying in $\overline{\Gamma^D}$. 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 δ_{ij} is the Kronecker symbol. Then $V_h = \operatorname{span}\{\varphi_i\}_{i=1}^{N_h}$.

It is well known that the standard Galerkin finite element discretization of the convection–diffusion problem (1) 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) method developed by Brooks and Hughes [1] which is frequently used because of its stability properties and higher-order accuracy, see, e.g., [17]. Let us mention that, when using the above space V_h, many other stabilization approaches like the Galerkin/least-squares method [6] or the subgrid scale method [5] are equivalent to the SUPG method if the data of the problem (1) are constant.

The SUPG method is a Petrov–Galerkin method with weighting functions

$$\overline{\varphi}_i = \varphi_i + \tau \boldsymbol{b} \cdot \nabla \varphi_i, \quad i = 1, \ldots, M_h,$$

where $\tau \in L^\infty(\Omega)$ is a nonnegative stabilization parameter. The SUPG solution u_h of (1) is defined by

$$u_h \in V_h, \tag{2}$$

$$\varepsilon(\nabla u_h, \nabla \varphi_i) + (\boldsymbol{b} \cdot \nabla u_h, \overline{\varphi}_i) = (f, \overline{\varphi}_i) + (g, \varphi_i)_{\Gamma^N}, \quad i = 1, \ldots, M_h, \tag{3}$$

$$u_h(a_i) = u_b(a_i), \quad i = M_h + 1, \ldots, N_h, \tag{4}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ and $(\cdot, \cdot)_{\Gamma^N}$ is the inner product in $L^2(\Gamma^N)$. The choice of the stabilization parameter τ may dramatically influence the accuracy of the discrete solution and therefore it has been a subject of an extensive research over the last three decades, see, e.g., the review in [10]. Unfortunately, a general optimal definition of τ is still not known. Often, the parameter τ is defined, on any element $K \in \mathcal{T}_h$, by the formula

$$\tau|_K = \frac{h_K}{2|\boldsymbol{b}|} \xi_0(Pe_K) \quad \text{with } \xi_0(\alpha) = \coth \alpha - \frac{1}{\alpha}, \ Pe_K = \frac{|\boldsymbol{b}| h_K}{2\varepsilon}, \tag{5}$$

where h_K is the element diameter in the direction of the convection vector $\boldsymbol{b}$ and 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. Note that, generally, the parameters h_K, Pe_K and $\tau|_K$ are functions of the points $\mathbf{x} \in K$. Sometimes, the so-called upwind function ξ_0 is approximated by a simpler expression, e.g., one can use

$$\tau|_K = \frac{h_K}{2|\boldsymbol{b}|} \xi_1(Pe_K) \quad \text{with } \xi_1(\alpha) = \max\left\{0, 1 - \frac{1}{\alpha}\right\}. \tag{6}$$

It is interesting that, in [4], this formula was also obtained from an analysis of the structure of eigenvalues of the SUPG stiffness matrix.

Formula (5) originates from the one-dimensional case of (1) with constant data and Dirichlet boundary conditions, where it leads to a nodally exact SUPG solution if continuous piecewise linear finite elements on a uniform division of Ω are used, cf. [2]. In two dimensions, this property is generally lost and since the SUPG method is not monotone, a

J Sci Comput (2010) 43: 454–470 457

discrete solution satisfying (2)–(4) usually contains spurious oscillations localized in narrow regions along sharp layers. A possible remedy is to add a suitable artificial diffusion term to the SUPG method, see the review in [10], where such approaches are called spurious oscillations at layers diminishing (SOLD) methods. Other names which can be found in the literature are discontinuity-capturing or shock-capturing methods. The artificial diffusion added by the SOLD methods typically depends on the unknown discrete solution and hence the SOLD methods are usually nonlinear. Recently, these methods were tested in a number of papers, see [8–12], and it was observed that many of the SOLD methods often substantially reduce the oscillations appearing in the SUPG solutions without an excessive smearing of the layers. However, it also turned out that the properties of the SOLD methods strongly depend on various factors like the data of the problem, the computational mesh or values of parameters and that, also for simple problems, any of the SOLD methods can give a solution with nonnegligible spurious oscillations. In fact, oscillation-free discrete solutions with sharp layers can be generally obtained only using SOLD methods containing free parameters. However, it is generally not known how these parameters should be defined.

The above discussion shows that the only reliable way to obtain oscillation-free discrete solutions of the problem (1) is to apply methods satisfying the discrete maximum principle which do not involve any free parameters. A promising representative of such methods is the Mizukami–Hughes method which will be discussed in the following sections.

3 Mizukami–Hughes Method

In addition to the assumptions made at the beginning of Sect. 2, we shall assume in the following that the triangulation $\mathcal{T}_h$ is of weakly acute type, i.e., the magnitude of angles in all elements of $\mathcal{T}_h$ is less than or equal to $\pi/2$. We shall use the notation introduced in Sect. 2.

The method proposed by Mizukami and Hughes in [16] is a Petrov–Galerkin method with weighting functions

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

Here χ_K are characteristic functions of elements K (i.e., $\chi_K = 1$ in K and $\chi_K = 0$ elsewhere) and C_i^K are constants which will be determined later. The discrete solution u_h of (1) is defined by

$$u_h \in V_h, \tag{7}$$

$$\varepsilon(\nabla u_h, \nabla\varphi_i) + (\boldsymbol{b}_h\cdot\nabla u_h, \widetilde{\varphi}_i) = (f,\widetilde{\varphi}_i) + (g,\varphi_i)_{\Gamma^N}, \quad i = 1,\dots,M_h, \tag{8}$$

$$u_h(a_i) = u_b(a_i), \quad i = M_h+1,\dots,N_h, \tag{9}$$

where $\boldsymbol{b}_h$ is a piecewise constant approximation of $\boldsymbol{b}$. We shall also use the notation $\boldsymbol{b}_K \equiv \boldsymbol{b}_h|_K$ for $K\in\mathcal{T}_h$. In our computations, we set $\boldsymbol{b}_h$ equal to the values of $\boldsymbol{b}$ at barycentres of elements of $\mathcal{T}_h$.

The constants C_i^K in the definition of the weighting functions are required to satisfy, for any $K\in\mathcal{T}_h$,

$$C_i^K \ge -\frac{1}{3} \quad \forall i\in\{1,\dots,N_h\}, \quad a_i\in\overline{K}, \qquad \sum_{i=1, a_i\in\overline{K}}^{N_h} C_i^K = 0. \tag{10}$$

Moreover, the idea of Mizukami and Hughes was to choose the constants C_i^K in such a way that the local convection matrix A^K with entries

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

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. [3]). As usual, $(\cdot,\cdot)_K$ denotes the inner product in $L^2(K)$.

To characterize the direction of the convection vector $\boldsymbol{b}_K$, we decompose any triangle K into vertex zones and edge zones by drawing lines parallel to the edges of K which all intersect at the barycentre of K, see Fig. 1. Denoting the vertices of K by a_1, a_2 and a_3, the set containing the vertex a_i, $i = 1, 2, 3$, will be called vertex zone VZ_i. The remaining three sets are called edge zones and the edge zone opposite the vertex a_i will be denoted by EZ_i. The common part of the boundaries of two adjacent zones is included in the respective vertex zone. The fact that the vector $\boldsymbol{b}_K$ points from the barycentre of K into VZ_i or EZ_i will be shortly expressed by $\boldsymbol{b}_K \in \mathrm{VZ}_i$ or $\boldsymbol{b}_K \in \mathrm{EZ}_i$, respectively. Without loss of generality, we may assume that the vertices of K are numbered in such a way that $\boldsymbol{b}_K \in \mathrm{VZ}_1$ or $\boldsymbol{b}_K \in \mathrm{EZ}_1$ as depicted in Fig. 1.

If $\boldsymbol{b}_K \in \mathrm{VZ}_1$, then (10) holds and A^K is of nonnegative type for

$$C_1^K = \frac{2}{3}, \qquad C_2^K = C_3^K = -\frac{1}{3}.$$

On the other hand, if $\boldsymbol{b}_K \in \mathrm{EZ}_1$, then it is generally not possible to choose the constants C_1^K, C_2^K, C_3^K in such a way that (10) holds and A^K is of nonnegative type. However, these requirements can be satisfied if $\boldsymbol{b}_K$ is replaced by a vector $\tilde{\boldsymbol{b}}_K$ pointing into a vertex zone and preserving the product $\boldsymbol{b}_K \cdot \nabla u_h|_K$. This is motivated by the fact that, in the continuous case (1), the solution u does not change if $\boldsymbol{b}$ is replaced by $\tilde{\boldsymbol{b}}$ such that $\tilde{\boldsymbol{b}} \cdot \nabla u = \boldsymbol{b} \cdot \nabla u$. Note that the local convection matrix A^K will be still defined using $\boldsymbol{b}_K$ and the vector $\tilde{\boldsymbol{b}}_K$ is used only for defining the constants C_i^K. Since the constants C_i^K depend through $\tilde{\boldsymbol{b}}_K$ on the unknown discrete solution u_h, the resulting discrete problem is nonlinear, like the SOLD methods mentioned in the preceding section.

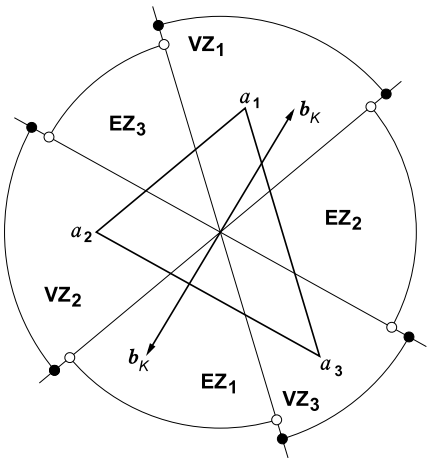


Fig. 1 Definition of edge zones and vertex zones

J Sci Comput (2010) 43: 454–470 459

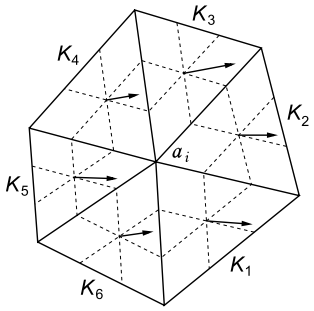


Fig. 2 Notation for demonstrating the upwind character of the Mizukami–Hughes method (*vectors* indicate the directions of $\boldsymbol{b}_h$)

Let us assume that $\boldsymbol{b}_K \in \mathrm{EZ}_1$ and $\boldsymbol{b}_K \cdot \nabla u_h|_K \neq 0$ and let $\boldsymbol{w} \neq \boldsymbol{0}$ be a vector orthogonal to $\nabla u_h|_K$. We introduce the sets

$$V_k = \{\alpha \in \mathbb{R};\, \boldsymbol{b}_K + \alpha \boldsymbol{w} \in \mathrm{VZ}_k\}, \quad k = 2, 3.$$

The vectors $\boldsymbol{b}_K + \alpha \boldsymbol{w}$ play the role of $\tilde{\boldsymbol{b}}_K$ mentioned above. It is easy to see that $V_2 \cup V_3 \neq \emptyset$. Mizukami and Hughes show that, depending on V_2 and V_3, the following values of the constants C_i^K should be used:

$$V_2 \neq \emptyset \quad \& \quad V_3 = \emptyset \quad \Longrightarrow \quad C_2^K = \frac{2}{3}, \quad C_1^K = C_3^K = -\frac{1}{3}, \tag{11}$$

$$V_2 = \emptyset \quad \& \quad V_3 \neq \emptyset \quad \Longrightarrow \quad C_3^K = \frac{2}{3}, \quad C_1^K = C_2^K = -\frac{1}{3}, \tag{12}$$

$$V_2 \neq \emptyset \quad \& \quad V_3 \neq \emptyset \quad \Longrightarrow \quad C_1^K = -\frac{1}{3}, \quad C_2^K + C_3^K = \frac{1}{3},$$
$$C_2^K > -\frac{1}{3}, \quad C_3^K > -\frac{1}{3}. \tag{13}$$

In case (13), Mizukami and Hughes suggest to set

$$C_i^K = \frac{\boldsymbol{b}_K \cdot \nabla \varphi_i|_K}{3|(\boldsymbol{b}_K \cdot \nabla \varphi_1|_K)|}, \quad i = 1, 2, 3.$$

This choice is also considered if $\boldsymbol{b}_K \in \mathrm{EZ}_1$ satisfies $\boldsymbol{b}_K \cdot \nabla u_h|_K = 0$. If $\boldsymbol{b}_K = \boldsymbol{0}$, Mizukami and Hughes set $C_i^K = 0$ for $i = 1, 2, 3$. Although the matrix A^K is generally not of non-negative type if $\boldsymbol{b}_K \in \mathrm{EZ}_1$, it can be proved that the solution of (7)–(9) satisfies the discrete maximum principle.

Note that the above definitions of the constants C_i^K give rise to an upwind effect. Indeed, if we consider the configuration depicted in Fig. 2, we have $C_i^{K_1} = C_i^{K_2} = C_i^{K_3} = -\frac{1}{3}$ and hence

$$(\boldsymbol{b}_h \cdot \nabla u_h, \widetilde{\varphi}_i) = (\boldsymbol{b}_h \cdot \nabla u_h, \widetilde{\varphi}_i)_{K_4 \cup K_5 \cup K_6}.$$

4 Improved Mizukami–Hughes Method

We showed in [13] that the above definition of the constants C_i^K for $\boldsymbol{b}_K$ pointing into an edge zone is not appropriate if K lies in a numerical boundary layer. Moreover, the definition of

the constants C_i^K is discontinuous with respect to the orientation of both $\boldsymbol{b}_K$ and $\nabla u_h|_K$. The latter discontinuity often prevents the nonlinear iterative process from converging. Therefore, in [13], we proposed several improvements of the Mizukami–Hughes method which correct the mentioned shortcomings. Here, we only give the resulting definitions of the constants and refer to [13] for details.

Let us consider any element $K \in \mathcal{T}_h$ and let a_1, a_2 and a_3 be its vertices. If $\boldsymbol{b}_K \neq \boldsymbol{0}$, we assume that $\boldsymbol{b}_K$ points into the vertex zone or the edge zone of a_1 (cf. Fig. 1) and we denote

$$\boldsymbol{s} = \frac{\boldsymbol{b}_K}{|\boldsymbol{b}_K|}, \qquad \boldsymbol{v}_2 = \frac{a_2 - a_1}{|a_2 - a_1|}, \qquad \boldsymbol{v}_3 = \frac{a_3 - a_1}{|a_3 - a_1|}, \qquad \boldsymbol{v} = \frac{\boldsymbol{v}_2 + \boldsymbol{v}_3}{|\boldsymbol{v}_2 + \boldsymbol{v}_3|}.$$

Further, we introduce unit vectors $\boldsymbol{w}$, $\boldsymbol{v}^\perp$, $\boldsymbol{v}_2^\perp$ and $\boldsymbol{v}_3^\perp$ such that

$$\boldsymbol{w} \cdot \nabla u_h|_K = 0, \qquad \boldsymbol{v}^\perp \cdot \boldsymbol{v} = 0, \qquad \boldsymbol{v}_2^\perp \cdot \boldsymbol{v}_2 = 0, \qquad \boldsymbol{v}_3^\perp \cdot \boldsymbol{v}_3 = 0,$$
$$\boldsymbol{w} \cdot \boldsymbol{v} \geq 0, \qquad \boldsymbol{v}^\perp \cdot \boldsymbol{v}_3 \geq 0.$$

Finally, we recall the spaces V_2 and V_3 introduced in Sect. 2. Then the constants C_1^K, C_2^K and C_3^K in the improved Mizukami–Hughes method are determined according to the algorithm in Fig. 5.

5 Suppression of the Upwind Character of the Mizukami–Hughes Method

Let us consider the problem (1) with $\Omega = (0, 1)^2$ and let $\boldsymbol{b}$ and f be constant. Moreover, let $\boldsymbol{b} = (b, 0)$ and let the boundary conditions be such that the solution u does not change in the vertical direction. If we consider a uniform triangulation of Ω of the type depicted in Fig. 3, then, for any element of the triangulation, $\boldsymbol{b}$ points into a vertex zone and hence all constants C_i^K are independent of u_h. Thus, in this special case, the Mizukami–Hughes method is linear and the constants C_i^K have the same values for all elements having the same orientation. Although it is easy to find the six values of these constants, because of further considerations, it is advantageous to denote them by $C_1, \ldots, C_6$, see Fig. 4. Let us assume

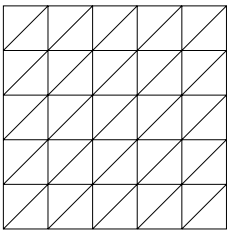


Fig. 3 A uniform triangulation

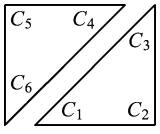


Fig. 4 Constants C_i^K for the two orientations of the elements from Fig. 3

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

$$C_1^K = C_2^K = C_3^K = 0$$

ELSE IF $\boldsymbol{b}_K \in \mathrm{VZ}_1$ THEN

$$C_1^K = \tfrac{2}{3}, \qquad C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\overline{K} \cap \overline{\Gamma^D} \neq \emptyset$ THEN

$$C_1^K = C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\mathcal{T}_h$ is not of the type from Fig. 3 and all vertices of K are connected by edges to vertices on $\overline{\Gamma^D}$ THEN

$$C_1^K = C_2^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $\boldsymbol{b}_K \cdot \nabla u_h|_K = 0$ THEN

$$C_1^K = -\tfrac{1}{3}, \quad C_2^K = C_3^K = \tfrac{1}{6}$$

ELSE IF $V_2 \neq \emptyset$ & $V_3 = \emptyset$ THEN

$$C_2^K = \tfrac{2}{3}, \quad C_1^K = C_3^K = -\tfrac{1}{3}$$

ELSE IF $V_2 = \emptyset$ & $V_3 \neq \emptyset$ THEN

$$C_3^K = \tfrac{2}{3}, \quad C_1^K = C_2^K = -\tfrac{1}{3}$$

ELSE IF $\boldsymbol{w} \cdot \boldsymbol{v}^{\perp} < 0$ THEN

$$r_2 = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{v}_2^{\perp}|}{|\boldsymbol{v} \cdot \boldsymbol{v}_2^{\perp}|} + 1 - \operatorname{sgn}(\boldsymbol{b}_K \cdot \boldsymbol{v}_2)\right\},$$

$$\Phi = \min\left\{1, \frac{2|\boldsymbol{w} \cdot \boldsymbol{v}_2^{\perp}|}{r_2 \boldsymbol{v} \cdot \boldsymbol{v}_2}\right\},$$

$$C_2^K = -\tfrac{1}{3} + \tfrac{1}{2}\Phi\left[1 + \frac{(\boldsymbol{v}_2 - \boldsymbol{v}_3) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}\right],$$

$$C_3^K = \tfrac{1}{3} - C_2^K, \qquad C_1^K = -\tfrac{1}{3}$$

ELSE

$$r_3 = \min\left\{1, \frac{|\boldsymbol{s} \cdot \boldsymbol{v}_3^{\perp}|}{|\boldsymbol{v} \cdot \boldsymbol{v}_3^{\perp}|} + 1 - \operatorname{sgn}(\boldsymbol{b}_K \cdot \boldsymbol{v}_3)\right\},$$

$$\Phi = \min\left\{1, \frac{2|\boldsymbol{w} \cdot \boldsymbol{v}_3^{\perp}|}{r_3 \boldsymbol{v} \cdot \boldsymbol{v}_3}\right\},$$

$$C_3^K = -\tfrac{1}{3} + \tfrac{1}{2}\Phi\left[1 + \frac{(\boldsymbol{v}_3 - \boldsymbol{v}_2) \cdot \boldsymbol{s}}{1 - \boldsymbol{v}_2 \cdot \boldsymbol{v}_3}\right],$$

$$C_2^K = \tfrac{1}{3} - C_3^K, \qquad C_1^K = -\tfrac{1}{3}.$$

Fig. 5 Definition of the constants C_i^K in the improved Mizukami–Hughes method

that, like the solution u, the discrete solution u_h does not change in the vertical direction and let u_i be the value of u_h at nodes with the horizontal coordinate ih, $i = 0, \dots, N$, where $h \equiv 1/N$ is the constant mesh width in both the horizontal and the vertical directions. Denoting $\delta = C_2 + C_3 + C_4$, the properties (10) imply that $C_1 + C_5 + C_6 = -\delta$ and $\delta \in [-1, 1]$ and (8) can be written in the form

$$\varepsilon(-u_{i-1} + 2u_i - u_{i+1}) + \frac{1}{2}bh[(1+\delta)(u_i - u_{i-1}) + (1-\delta)(u_{i+1} - u_i)] = fh^2, \qquad (14)$$

where $i = 1, \dots, N-1$. Thus, the value of δ determines how strong upwind effect is introduced by the Mizukami–Hughes method. Since the second term on the left-hand side of (14) can be written in the form

$$\frac{1}{2}bh(u_{i+1} - u_{i-1}) + \frac{1}{2}\delta bh(-u_{i-1} + 2u_i - u_{i+1}),$$

we can also say that the value of δ determines the amount of artificial diffusion in the Mizukami–Hughes method (note that $\delta b \geq 0$).

Equations (14) are identical with the difference equations corresponding to the SUPG method (3) in the one-dimensional case provided that $\delta = 2b\tau/h$. Since the formula (5) is optimal in the one-dimensional case, we see that the optimal choice of δ is

$$\delta = \xi_0(Pe)\,\operatorname{sgn} b \quad \text{with } Pe = \frac{|b|h}{2\varepsilon}. \qquad (15)$$

However, if we compute the value of δ corresponding to the Mizukami–Hughes method described in the preceding two sections, we obtain $\delta = 1$ if $b > 0$ and $\delta = -1$ if $b < 0$. Thus, independently of the value of the Péclet number, the Mizukami–Hughes method corresponds to the discretization of the convective term by standard upwind differencing. This is appropriate if the Péclet number is large (since then $\xi_0(Pe) \approx 1$), however, for small Péclet numbers, such a discretization leads to a low accuracy since two much artificial diffusion is introduced. Therefore, in the following, we shall modify the constants C_i^K introduced in the preceding sections in order to suppress the upwind character of the Mizukami–Hughes method for small Péclet numbers.

Let us consider any $K \in \mathcal{T}_h$. The definition of the constants C_i^K in Sect. 3 was based on the requirement that the local convection matrix A^K be of nonnegative type. Since the local diffusion matrix D^K having the entries

$$d_{ij}^K = \varepsilon(\nabla\varphi_j, \nabla\varphi_i)_K, \quad i = 1, \dots, M_h,\ j = 1, \dots, N_h, \quad a_i, a_j \in \overline{K},$$

is of nonnegative type, a natural way to reduce the absolute values of the constants C_i^K is to require that the sum $D^K + A^K$ be of nonnegative type but not necessarily A^K. Since the sum of the entries in each row of $D^K + A^K$ vanishes, it suffices to assure that $d_{ij}^K + a_{ij}^K \leq 0$ for $i \neq j$.

Let us again denote the vertices of K by a_1, a_2 and a_3 and let us assume that $\boldsymbol{b}_K \in \mathrm{VZ}_1$. Then

$$\boldsymbol{b}_K \cdot \nabla\varphi_1 > 0, \qquad \boldsymbol{b}_K \cdot \nabla\varphi_2 \leq 0, \qquad \boldsymbol{b}_K \cdot \nabla\varphi_3 \leq 0,$$

where we write $\nabla\varphi_i$ instead of $\nabla\varphi_i|_K$ for simplicity. Since

$$d_{ij}^K + a_{ij}^K = \operatorname{meas}_2(K)\left[\varepsilon\nabla\varphi_j \cdot \nabla\varphi_i + \boldsymbol{b}_K \cdot \nabla\varphi_j\left(\frac{1}{3} + C_i^K\right)\right], \qquad (16)$$

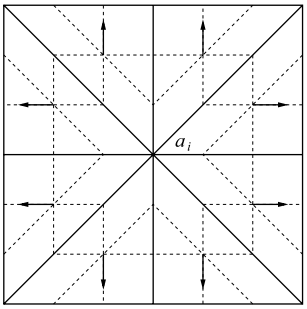


Fig. 6 A configuration for which all entries of the i-th row of the matrix $D + A$ vanish (*vectors* indicate the directions of $\boldsymbol{b}_h$)

the matrix $D^K + A^K$ is of nonnegative type if

$$\boldsymbol{b}_K \cdot \nabla\varphi_1 \left(\frac{1}{3} + C_i^K \right) \leq -\varepsilon \nabla\varphi_1 \cdot \nabla\varphi_i \quad \text{for } i = 2, 3.$$

This suggests to set

$$C_i^K = \min \left\{ 0, -\frac{1}{3} - \frac{\varepsilon \nabla\varphi_1 \cdot \nabla\varphi_i}{\boldsymbol{b}_K \cdot \nabla\varphi_1} \right\}, \quad i = 2, 3, \qquad C_1^K = -C_2^K - C_3^K. \tag{17}$$

If we now return to the special setting considered at the beginning of this section and compute the value of δ corresponding to this new definition of the constants C_i^K, we obtain

$$\delta = \max \left\{ \frac{1}{3}, 1 - \frac{1}{Pe} \right\} \operatorname{sgn} b.$$

This is a good approximation of the optimal relation (15), similar to the approximation (6) of (5).

Unfortunately, using the definition (17) of the constants C_i^K, it can happen that the $M_h \times M_h$ matrix corresponding to the Mizukami–Hughes method is singular. To see this, let us consider a triangulation where the elements are arranged around a given vertex a_i as depicted in Fig. 6. For simplicity, we again assume that the mesh width in both the horizontal and the vertical directions is constant and equal to h. The convection $\boldsymbol{b}_h$ is oriented as shown in Fig. 6. Let K be any of the eight elements depicted in Fig. 6 and let us assume that $\varepsilon < |\boldsymbol{b}_K| h/3$. Then the constant C_i^K corresponding to the vertex a_i is given by

$$C_i^K = -\frac{1}{3} + \frac{\varepsilon}{|\boldsymbol{b}_K| h}.$$

Since $\boldsymbol{b}_K = -|\boldsymbol{b}_K| h \nabla\varphi_i|_K$, it follows from (16) that $d_{ij}^K + a_{ij}^K = 0$ for any j. Thus, denoting by D and A the $M_h \times N_h$ matrices obtained by assembling the local matrices D^K and A^K, respectively, all entries of the i-th row of the matrix $D + A$ vanish. Therefore, the relations (17) have to be modified. We choose $\tilde{\varepsilon} \in (0, \varepsilon)$, preferably near to ε, and define matrices $\widetilde{D}^K$ by replacing ε by $\tilde{\varepsilon}$ in the definition of D^K. Then we require that the sum $\widetilde{D}^K + A^K$ be of nonnegative type, which implies that now $\tilde{\varepsilon}$ will be used instead of ε in (17). Thus, we have

$$C_i^K = \min \left\{ 0, -\frac{1}{3} - \frac{\tilde{\varepsilon} \nabla\varphi_1 \cdot \nabla\varphi_i}{\boldsymbol{b}_K \cdot \nabla\varphi_1} \right\}, \quad i = 2, 3, \qquad C_1^K = -C_2^K - C_3^K. \tag{18}$$

The global matrix $D + A$ is then a sum of the matrix $\widetilde{D} + A$, which is of nonnegative type, and the matrix $D - \widetilde{D}$, which is of nonnegative type and of full rank. The $M_h \times M_h$ submatrix of $D + A$, whose columns correspond to basis functions $\varphi_1, \ldots, \varphi_{M_h}$, is then nonsingular as it follows from the following theorem.

Theorem 5.1 *Let B and C be real $n \times n$ matrices of nonnegative type and let the matrix B be nonsingular. Then the matrix $B + C$ is nonsingular as well.*

Proof Let $u \in \mathbb{R}^n$ be such that $(B + C)u = 0$. Let

$$s = \max\{u_i;\ i = 1, \ldots, n\}, \qquad J = \{i \in \{1, \ldots, n\};\ u_i = s\}$$

and let $J \neq \{1, \ldots, n\}$. Since B and C are of nonnegative type, we have

$$\sum_{j \in J} b_{ij} \geq 0, \quad \sum_{j \in J} c_{ij} \geq 0 \quad \forall i \in J. \tag{19}$$

We shall prove that

$$\exists k \in J: \quad \mu_k \equiv \sum_{j \in J} (b_{kj} + c_{kj}) > 0. \tag{20}$$

Let (20) do not hold. Then (19) implies that $\sum_{j \in J} b_{ij} = 0$ for any $i \in J$ and hence also $b_{ij} = 0$ for any $i \in J$ and $j \notin J$. Thus, the matrix $(b_{ij})_{i,j \in J}$ is singular and hence there exist real numbers $\{v_i\}_{i \in J}$, not all equal to zero, such that $\sum_{i \in J} b_{ij} v_i = 0$ for $j = 1, \ldots, n$. Consequently, the matrix B is singular which contradicts the assumptions of the theorem. Therefore, (20) holds and hence

$$\begin{aligned} s\mu_k &= \sum_{j \in J} (b_{kj} + c_{kj}) u_j = \sum_{j \notin J} (-b_{kj} - c_{kj}) u_j \\ &\leq \max_{j \notin J}\{u_j\} \sum_{j \notin J} (-b_{kj} - c_{kj}) \leq \max_{j \notin J}\{u_j\} \mu_k. \end{aligned}$$

This implies that $s \leq \max_{j \notin J}\{u_j\}$, which is a contradiction with the definition of J. Hence $u_i = s$ for $i = 1, \ldots, n$. If $s > 0$, then all components of the vectors Bu and Cu are nonnegative and hence $Bu = Cu = 0$. The same holds if $s < 0$. Thus, since the matrix B is nonsingular, we conclude that $s = 0$ and hence $u = 0$. □

Up to now, we have only discussed the choice of the constants C_i^K in the case when $\boldsymbol{b}_K \in \mathrm{VZ}_1$ and we proposed to define these constants by (18). Now let us turn our attention to the case when $\boldsymbol{b}_K \in \mathrm{EZ}_1$. If, for some $k \in \{2, 3\}$, the set V_k is nonempty, we denote by α_k the element of V_k satisfying

$$(\boldsymbol{b}_K + \alpha_k \boldsymbol{w}) \cdot \nabla \varphi_k = \min_{\alpha \in V_k} [(\boldsymbol{b}_K + \alpha \boldsymbol{w}) \cdot \nabla \varphi_k]$$

and, by analogy with (18), we set

$$C_i^{K,k} = \min\left\{0, -\frac{1}{3} - \frac{\tilde{\varepsilon} \nabla \varphi_k \cdot \nabla \varphi_i}{(\boldsymbol{b}_K + \alpha_k \boldsymbol{w}) \cdot \nabla \varphi_k}\right\}, \quad i \neq k, \qquad C_k^{K,k} = -\sum_{i \neq k} C_i^{K,k}.$$

Now, denoting by $\bar{C}_i^K$ the constants C_i^K defined by the algorithm in Fig. 5, we set

$$C_i^K = \left(\frac{1}{3} + \bar{C}_2^K\right) C_i^{K,2} + \left(\frac{1}{3} + \bar{C}_3^K\right) C_i^{K,3}, \quad i = 1, 2, 3, \tag{21}$$

except the cases with $\bar{C}_1^K = \bar{C}_2^K = \bar{C}_3^K = -\frac{1}{3}$, where we simply replace the value $-\frac{1}{3}$ by the largest nonpositive value such that the matrix $\widetilde{D}^K + A^K$ is of nonnegative type. Note that, if $V_2 \neq \emptyset$ and $V_3 \neq \emptyset$, the constants defined by the algorithm in Fig. 5 satisfy the relations from (13). It is easy to verify that, in all the cases, we have $C_i^K \to \bar{C}_i^K$ for $\varepsilon/|\boldsymbol{b}_K| \to 0$ and $i = 1, 2, 3$. Thus, if convection strongly dominates diffusion, the modifications introduced in (18) and (21) will not change the properties of the method.

Since the matrix $\widetilde{D}^K + A^K$ is generally not of nonnegative type, we shall now construct a matrix $\widetilde{A}^K$ such that $\widetilde{D}^K + \widetilde{A}^K$ is of nonnegative type and

$$\widetilde{A}^K U^K = A^K U^K, \tag{22}$$

where U^K is the coefficient vector of $u_h|_K$ with respect to the basis functions $\varphi_1, \varphi_2, \varphi_3$. First, for $k \in \{2, 3\}$ such that the set V_k is nonempty, we introduce the matrix $\widetilde{A}^{K,k}$ having the entries

$$\widetilde{a}_{ij}^{K,k} = \operatorname{meas}_2(K)(\boldsymbol{b}_K + \alpha_k \boldsymbol{w}) \cdot \nabla\varphi_j \left(\frac{1}{3} + C_i^{K,k}\right),$$

where $i, j = 1, 2, 3$ ($a_i \in \Omega \cup \Gamma^N$). By the definition of the constants $C_i^{K,k}$, the matrix $\widetilde{D}^K + \widetilde{A}^{K,k}$ is of nonnegative type. Now, we set

$$\widetilde{A}^K = \left(\frac{1}{3} + \bar{C}_2^K\right) \widetilde{A}^{K,2} + \left(\frac{1}{3} + \bar{C}_3^K\right) \widetilde{A}^{K,3}.$$

Since $\bar{C}_2^K + \bar{C}_3^K = \frac{1}{3}$, the matrix $\widetilde{D}^K + \widetilde{A}^K$ also is of nonnegative type. Moreover,

$$\left(\frac{1}{3} + \bar{C}_2^K\right)\left(\frac{1}{3} + C_i^{K,2}\right) + \left(\frac{1}{3} + \bar{C}_3^K\right)\left(\frac{1}{3} + C_i^{K,3}\right) = \frac{1}{3} + C_i^K, \quad i = 1, 2, 3,$$

and since $\boldsymbol{w} \cdot \nabla u_h|_K = 0$, we obtain (22).

The above considerations show that, in all cases treated by the algorithm in Fig. 5 and for any $K \in \mathcal{T}_h$, the discrete solution of the Mizukami–Hughes method with constants C_i^K modified by (18) and (21) satisfies (22) with a matrix $\widetilde{A}^K$ such that the matrix $\widetilde{D}^K + \widetilde{A}^K$ is of nonnegative type. Denoting by $\widetilde{A}$ the $M_h \times N_h$ matrix made up of the local matrices $\widetilde{A}^K$, we see that the vector of coefficients of the discrete solution u_h with respect to the basis $\{\varphi_i\}_{i=1}^{N_h}$ is the solution of a linear system with the matrix $D + \widetilde{A} = (D - \widetilde{D}) + (\widetilde{D} + \widetilde{A})$, where the matrices $D + \widetilde{A}$, $D - \widetilde{D}$ and $\widetilde{D} + \widetilde{A}$ are of nonnegative type. In view of Theorem 5.1, the $M_h \times M_h$ submatrix of $D + \widetilde{A}$, whose columns correspond to basis functions $\varphi_1, \ldots, \varphi_{M_h}$, is nonsingular. Therefore, according to the following theorem, the discrete maximum principle holds.

Theorem 5.2 *Let A be a real $m \times n$ matrix of nonnegative type with $m < n$ and let the matrix $(a_{ij})_{i,j=1}^m$ be nonsingular. Let $u \in \mathbb{R}^n$ and $f \in \mathbb{R}^m$ satisfy $Au = f$ and let $f_i \leq 0$ $\forall i = 1, \ldots, m$. Then*

$$\max\{u_i;\ i = 1, \ldots, m\} \leq \max\{u_i;\ i = m + 1, \ldots, n\}.$$

Proof Let

$$s = \max\{u_i;\ i = 1, \ldots, n\}, \qquad J = \{i \in \{1, \ldots, n\};\ u_i = s\}$$

and let $J \subset \{1, \ldots, m\}$. We shall prove that then

$$\exists k \in J: \quad \mu_k \equiv \sum_{j \in J} a_{kj} > 0. \tag{23}$$

Let (23) do not hold. Then we deduce like in the proof of Theorem 5.1 that $\sum_{j \in J} a_{ij} = 0$ for any $i \in J$ and $a_{ij} = 0$ for any $i \in J$ and $j \notin J$. Consequently, like in the proof of Theorem 5.1, it follows that the matrix $(a_{ij})_{i,j=1}^{m}$ is singular, which contradicts the assumptions of the theorem. Thus, (23) holds and hence, denoting

$$r = \max\{u_i; i = 1, \ldots, n, i \notin J\},$$

we obtain

$$s\mu_k = \sum_{j \in J} a_{kj} u_j = f_k - \sum_{j \notin J} a_{kj} u_j \le f_k + r \sum_{j \notin J} (-a_{kj}) \le r\mu_k.$$

Therefore, $s \le r$, which is a contradiction with the definition of J. Hence there exists $i \in \{m + 1, \ldots, n\}$ such that $u_i = s$. □

6 Numerical Results

In this section we present results of numerical computations for two simple problems with known solutions. Since we are interested in the properties of the Mizukami–Hughes method for small and moderate Péclet numbers, the solutions do not possess any layers. Numerical results for problems with layers can be found in [13]. Apart from results of the Mizukami–Hughes method, we shall also present results obtained using the Galerkin method which usually leads to a high precision for small Péclet numbers. All results discussed in this section were computed on the unstructured triangulation of $(0, 1)^2$ showed in Fig. 7 which consists of 856 acute triangles. The parameter $\tilde{\varepsilon}$ introduced in the preceding section was equal to 0.95ε.

Example 1 We consider the problem (1) in $\Omega = (0, 1)^2$ with $\Gamma^D = \partial\Omega$, $\varepsilon > 0$, $\boldsymbol{b} = (1, 0)$, $u_b = 0$ and a right-hand side f such that the solution of (1) is given by $u(x, y) = 16x(1 - x)y(1 - y)$.

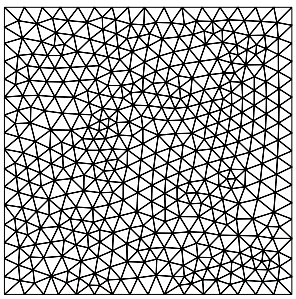


Fig. 7 Triangulation used for the numerical experiments in Sect. 6

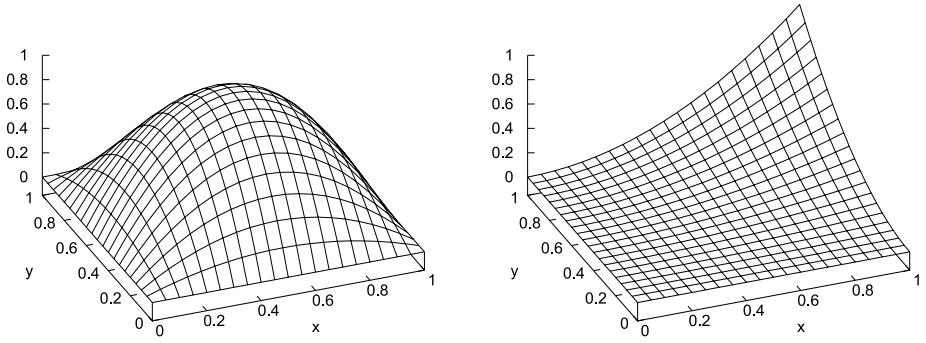


Fig. 8 Solutions of Example 1 (*left*) and Example 2 (*right*)

Table 1 Example 1, errors $\|u - u_h\|_{0,\Omega}$

ε	Galerkin	IMH	New IMH
1−5	7.93−3	4.07−3	4.07−3
1−4	3.78−3	4.03−3	4.02−3
1−3	1.96−3	3.98−3*	3.82−3
1−2	1.58−3	9.25−3*	5.94−3
1−1	2.13−3	2.28−2	2.22−3
1	2.39−3	2.18−2	2.39−3

Table 2 Example 1, errors $|u - u_h|_{1,\Omega}$

ε	Galerkin	IMH	New IMH
1−5	5.60−1	1.99−1	1.99−1
1−4	2.70−1	1.97−1	1.97−1
1−3	1.66−1	1.86−1*	1.84−1
1−2	1.56−1	1.76−1*	1.64−1
1−1	1.55−1	2.08−1	1.55−1
1	1.55−1	2.15−1	1.55−1

Table 3 Example 1, errors $\|u - u_h\|_{0,\infty,h}$

ε	Galerkin	IMH	New IMH
1−5	4.42−2	2.51−2	2.51−2
1−4	2.55−2	2.50−2	2.47−2
1−3	7.02−3	2.38−2*	2.27−2
1−2	3.35−3	1.52−2*	1.11−2
1−1	2.55−3	3.30−2	2.20−3
1	2.63−3	3.87−2	2.63−3

The solution of Example 1 is depicted in Fig. 8. Tables 1–3 show errors of the discrete solutions computed using the Galerkin method, the improved Mizukami–Hughes method (IMH) of [13] and the improved Mizukami–Hughes method with the modifications introduced in the present paper (new IMH) for various values of ε. The errors are measured in

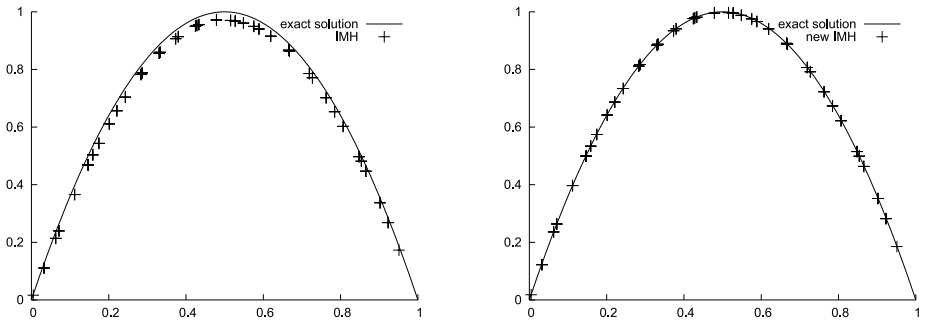


Fig. 9 Example 1 with $\varepsilon = 0.1$: comparison of the exact solution with the IMH solution (*left*) and new IMH solution (*right*) along the line $y = 0.5$

the $L^2(\Omega)$ norm $\|\cdot\|_{0,\Omega}$, the $H^1(\Omega)$ seminorm $|\cdot|_{1,\Omega}$ and the discrete L^∞ norm $\|\cdot\|_{0,\infty,h}$ defined as the maximum absolute value at vertices of the triangulation. The notation $r-n$ used in the tables means $r \cdot 10^{-n}$.

We observe that for large Péclet numbers (i.e., small ε), the accuracy of the Galerkin method deteriorates and is worse than for the IMH method. On the other hand, if the Péclet numbers decrease (i.e., ε increases), the Galerkin method outperforms the IMH method. The new IMH method introduced in the present paper provides results with the same accuracy as the IMH method if the Péclet numbers are large and with a better accuracy than the IMH method if the Péclet numbers decrease. For $\varepsilon \approx 1$, the accuracy of the new IMH method is similar as for the Galerkin method. The improvement of the accuracy for the new IMH method can be also seen in Fig. 9 where the solutions of the IMH method and the new IMH method are compared with the exact solution u along the line $y = 0.5$ for $\varepsilon = 0.1$. The crosses represent the values of the discrete solutions at intersections of the line $y = 0.5$ with edges of the triangulation. The mark * at some errors of the IMH method in Tables 1–3 indicates that the fixed-point iterative process used for computing the solution of the nonlinear discrete problem did not converge. The convergence for the new IMH method was fast in all cases. Thus, the modifications introduced in the present paper improve not only the accuracy of the discrete solution but also the convergence of the nonlinear iterative solver, which was also observed in other numerical tests.

Example 2 We consider the problem (1) in $\Omega = (0, 1)^2$ with $\Gamma^D = \partial\Omega$, $\varepsilon > 0$ and $\boldsymbol{b} = (2, 1)$. The Dirichlet boundary condition u_b and the right-hand side f are such that the solution of (1) is given by $u(x, y) = x^2y^2$.

The solution of Example 2 is depicted in Fig. 8. The main difference to Example 1 is that now inhomogeneous Dirichlet boundary conditions are considered. Tables 4–6 show errors of the discrete solutions computed using the same methods as for Example 1 and the discussion to Example 1 also applies here. Figure 10 shows a comparison of the IMH solution and the new IMH solution with the exact solution u along the line $y = 0.5$ for $\varepsilon = 1$. Again we observe that the IMH method adds too much artificial diffusion whereas the solution of the new IMH method coincides very well with the exact solution.

Table 4 Example 2, errors $\|u - u_h\|_{0,\Omega}$

ε	Galerkin	IMH	New IMH
1−5	8.71−3	7.34−4	7.34−4
1−4	1.41−3	7.28−4	7.27−4
1−3	4.87−4	6.71−4	6.64−4
1−2	2.94−4	4.77−4	4.38−4
1−1	3.13−4	2.22−3	3.81−4
1	4.12−4	5.07−3	4.12−4

Table 5 Example 2, errors $|u - u_h|_{1,\Omega}$

ε	Galerkin	IMH	New IMH
1−5	6.08−1	4.31−2	4.31−2
1−4	1.03−1	4.29−2	4.29−2
1−3	4.71−2	4.16−2	4.15−2
1−2	3.57−2	3.61−2	3.59−2
1−1	3.45−2	4.00−2	3.46−2
1	3.45−2	4.71−2	3.45−2

Table 6 Example 2, errors $\|u - u_h\|_{0,\infty,h}$

ε	Galerkin	IMH	New IMH
1−5	4.90−2	5.81−3	5.81−3
1−4	9.37−3	5.75−3	5.75−3
1−3	4.88−3	5.20−3	5.18−3
1−2	3.06−3	2.19−3	2.08−3
1−1	8.44−4	4.49−3	1.06−3
1	4.67−4	8.00−3	4.67−4

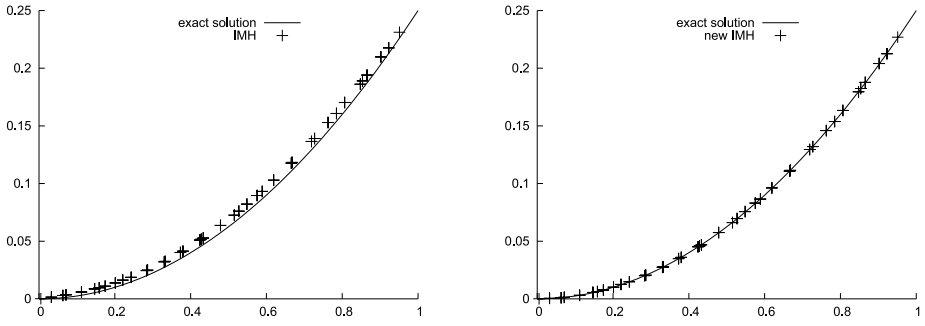


Fig. 10 Example 2 with $\varepsilon = 1$: comparison of the exact solution with the IMH solution (*left*) and new IMH solution (*right*) along the line $y = 0.5$

7 Conclusions

In this paper we discussed the properties of the improved Mizukami–Hughes method applied to scalar steady convection–diffusion equations in small and moderate Péclet number

regimes. We showed that, in this case, the method introduces two much artificial diffusion, which may decrease the accuracy of the discrete solution. Therefore, we proposed modifications of the Mizukami–Hughes method which preserve its favourable properties for large Péclet numbers and reduce its upwind character if the Péclet number is small. Numerical results justify the proposed modifications.

Acknowledgements This work is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Czech Republic under the grant No. 201/08/0012.

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 Appl. Mech. Eng. **32**, 199–259 (1982)
2. Christie, I., Griffiths, D.F., Mitchell, A.R., Zienkiewicz, O.C.: Finite element methods for second order differential equations with significant first derivatives. Int. J. Numer. Methods Eng. **10**, 1389–1396 (1976)
3. Ciarlet, P.G., Raviart, P.-A.: Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Eng. **2**, 17–31 (1973)
4. Fischer, B., Ramage, A., Silvester, D.J., Wathen, A.J.: On parameter choice and iterative convergence for stabilised discretisations of advection–diffusion problems. Comput. Methods Appl. Mech. Eng. **179**, 179–195 (1999)
5. Hughes, T.J.R.: Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods. Comput. Methods Appl. Mech. Eng. **127**, 387–401 (1995)
6. Hughes, T.J.R., Franca, L.P., Hulbert, G.M.: A new finite element formulation for computational fluid dynamics, VIII. The Galerkin/least-squares method for advective–diffusive equations. Comput. Methods Appl. Mech. Eng. **73**, 173–189 (1989)
7. Ikeda, T.: Maximum Principle in Finite Element Models for Convection–Diffusion Phenomena. Lecture Notes in Numerical and Applied Analysis, vol. 4. North-Holland, Amsterdam (1983)
8. John, V., Knobloch, P.: On discontinuity-capturing methods for convection–diffusion equations. In: Bermúdez de Castro, A., Gómez, D., Quintela, P., Salgado, P. (eds.) Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2005, pp. 336–344. Springer, Berlin (2006)
9. John, V., Knobloch, P.: A computational comparison of methods diminishing spurious oscillations in finite element solutions of convection–diffusion equations. In: Chleboun, J., Segeth, K., Vejchodský, T. (eds.) Proceedings of the International Conference Programs and Algorithms of Numerical Mathematics, vol. 13, pp. 122–136. Academy of Science of the Czech Republic, Prague (2006)
10. John, V., Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I—A review. Comput. Methods Appl. Mech. Eng. **196**, 2197–2215 (2007)
11. John, V., Knobloch, P.: On the performance of SOLD methods for convection–diffusion problems with interior layers. Int. J. Comput. Sci. Math. **1**, 245–258 (2007)
12. John, V., Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II—Analysis for P_1 and Q_1 finite elements. Comput. Methods Appl. Mech. Eng. **197**, 1997–2014 (2008)
13. Knobloch, P.: Improvements of the Mizukami–Hughes method for convection–diffusion equations. Comput. Methods Appl. Mech. Eng. **196**, 579–594 (2006)
14. Knobloch, P.: Numerical solution of convection–diffusion equations using upwinding techniques satisfying the discrete maximum principle. In: Beneš, M., Kimura, M., Nakaki, T. (eds.) Proceedings of Czech–Japanese Seminar in Applied Mathematics 2005, pp. 69–76. COE Lecture Note, vol. 3, Faculty of Mathematics, Kyushu University (2006)
15. Knobloch, P.: Application of the Mizukami–Hughes method to bilinear finite elements. In: Beneš, M., Kimura, M. Nakaki, T. (eds.) Proceedings of Czech–Japanese Seminar in Applied Mathematics 2006, pp. 137–147. COE Lecture Note, vol. 6, Faculty of Mathematics, Kyushu University (2007)
16. 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 Appl. Mech. Eng. **50**, 181–193 (1985)
17. Roos, H.-G., Stynes, M., Tobiska, L.: Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems, 2nd edn. Springer, Berlin (2008)
18. Tabata, M.: A finite element approximation corresponding to the upwind finite differencing. Memoirs Numer. Math. **4**, 47–63 (1977)

Chapter 3

SOLD methods

This chapter consists of the following publications:

V. John, P. Knobloch: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review, *Computer Methods in Applied Mechanics and Engineering* 196 (17-20): 2197–2215, 2007. p. 59

V. John, P. Knobloch: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements, *Computer Methods in Applied Mechanics and Engineering* 197 (21-24): 1997–2014, 2008. p. 79

V. John, P. Knobloch: On the performance of SOLD methods for convection–diffusion problems with interior layers, *International Journal of Computing Science and Mathematics* 1 (2-4): 245–258, 2007. p. 97

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 196 (2007) 2197–2215
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review

Volker John [a,*], Petr Knobloch [b]

[a] Universität des Saarlandes, Fachbereich 6.1 – Mathematik, Postfach 15 11 50, 66041 Saarbrücken, Germany
[b] Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 18675 Praha 8, Czech Republic

Received 31 October 2005; received in revised form 13 November 2006; accepted 21 November 2006

Abstract

An unwelcome feature of the popular streamline upwind/Petrov–Galerkin (SUPG) stabilization of convection-dominated convection–diffusion equations is the presence of spurious oscillations at layers. Since the mid of the 1980s, a number of methods have been proposed to remove or, at least, to diminish these oscillations without leading to excessive smearing of the layers. The paper gives a review and state of the art of these methods, discusses their derivation, proposes some alternative choices of parameters in the methods and categorizes them. Some numerical studies which supplement this review provide a first insight into the advantages and drawbacks of the methods.

Keywords: Convection–diffusion equations; Streamline upwind/Petrov–Galerkin (SUPG) method; Spurious oscillations at layers diminishing (SOLD) methods

1. Introduction

This paper is devoted to the numerical solution of the scalar convection–diffusion equation

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega, \qquad u = u_b \quad \text{on } \partial\Omega, \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a bounded domain with a polygonal (resp. polyhedral) boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$ is a given convective field satisfying the incompressibility condition $\operatorname{div}\boldsymbol{b} = 0$, $f \in L^2(\Omega)$ is an outer source of u, and $u_b \in H^{1/2}(\partial\Omega)$ represents the Dirichlet boundary condition. In our numerical tests we shall also consider less regular functions u_b.

Problem (1) describes the stationary distribution of a physical quantity u (e.g., temperature or concentration) determined by two basic physical mechanisms, namely the convection and diffusion. The broad interest in solving problem (1) is caused not only by its physical meaning just explained but also (and perhaps mainly) by the fact that it is a simple model problem for convection–diffusion effects which appear in many more complicated problems arising in applications (e.g. in various fluid flow problems).

Despite the apparent simplicity of problem (1), its numerical solution is still a challenge when convection is strongly dominant (i.e., when $\varepsilon \ll |\boldsymbol{b}|$). The basic difficulty is that, in this case, the solution of (1) typically possesses interior and boundary layers, which are small subregions where the derivatives of the solution are very large. The widths of these layers are usually significantly smaller than the mesh size and hence the layers cannot be resolved properly. This leads to unwanted spurious (nonphysical) oscillations in the numerical solution, the attenuation of which has been the subject of extensive research for more than three decades.

In this paper, we concentrate on the solution of (1) using the finite element method which proved to be a very efficient tool for the numerical solution of various boundary value problems in science and engineering. Unfortunately,

* Corresponding author.
E-mail addresses: john@math.uni-sb.de (V. John), knobloch@karlin.mff.cuni.cz (P. Knobloch).

0045-7825/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.cma.2006.11.013

the classical Galerkin formulation of (1) is inappropriate since, in case of dominant convection, the discrete solution is usually globally polluted by spurious oscillations causing a severe loss of accuracy and stability. This is not surprising since, in simple settings, the standard Galerkin finite element method is equivalent to a central finite difference discretization and it is well known that central difference approximations of the convective term give rise to spurious oscillations in convection dominated regimes (cf. e.g. Roos et al. [58]).

To enhance the stability and accuracy of the Galerkin discretization of (1) in the convection dominated regime, various stabilization strategies have been developed. Initially, these approaches imitated the upwind finite difference techniques. An important contribution to this development was made by Christie et al. [17], who showed that, in the one-dimensional case, a stabilization can be achieved using asymmetric test functions in a weighted residual finite element formulation. Choosing these test functions in a suitable way, they recovered the usual one-sided differences used for the approximation of the convective term in the finite difference method. Two-dimensional upwind finite element discretizations were derived by Heinrich et al. in [32,33] and by Tabata [62]. Many other finite element discretizations of upwind type have been proposed later.

Like in the finite difference method, the upwind finite element discretizations remove the unwanted oscillations but the accuracy attained is often poor since too much numerical diffusion is introduced. In addition, if the flow field $\boldsymbol{b}$ is directed skew to the mesh, an excessive artificial diffusion perpendicular to the flow (crosswind diffusion) can be observed. A further important drawback is that these methods are not consistent, i.e., the solution of (1) is no longer a solution to the variational problem as it is the case for a Galerkin formulation. Consequently, the accuracy is limited to first order. Moreover, non-consistent formulations are also known to produce inaccurate or wrong solutions when f (or the time derivative in case of transient problems) is significant. It can even happen that the discrete solution is then less accurate than that one produced by the Galerkin method (cf. e.g. Brooks and Hughes [9] for a discussion on shortcomings of upwind methods).

A significant improvement came with the streamline upwind/Petrov–Galerkin (SUPG) method developed by Brooks and Hughes [9] which substantially eliminates almost all the difficulties mentioned above. In contrast with upwind methods proposed earlier, the SUPG method introduces numerical diffusion along streamlines only and hence it possesses no spurious crosswind diffusion. Moreover, the streamline diffusion is added in a consistent manner. Consequently, stability is obtained without compromising accuracy and convergence results may be derived for a wide class of finite elements. In view of its stability properties and higher-order accuracy, the SUPG method is regarded as one of the most efficient procedures for solving convection-dominated equations.

An alternative to the SUPG method is the Galerkin/least-squares method introduced by Hughes et al. [35] who observed that stabilization terms can be obtained by minimizing the square of the equation residual. A variant to this method was proposed by Franca et al. [26] using the idea of Douglas and Wang [23] to change the sign of the Laplacian in the test function. Since the SUPG method is the most popular approach, we shall restrict ourselves to this method in the following.

The SUPG method produces accurate and oscillation-free solutions in regions where no abrupt changes in the solution of (1) occur but it does not preclude spurious oscillations (overshooting and undershooting) localized in narrow regions along sharp layers. It was observed by Almeida and Silva [3] that these oscillations can even be amplified if high-order finite elements are used in these regions. This indicates that using the streamlines as upwind direction is not always sufficient. Although the remaining nonphysical oscillations are usually small in magnitude, they are not permissible in many applications. An example are chemically reacting flows where it is essential to guarantee that the concentrations of all species are nonnegative. Another example are free-convection computations where temperature oscillations create spurious sources and sinks of momentum that effect the computation of the flow field. The small spurious oscillations may also deteriorate the solution of nonlinear problems, e.g., in two-equations turbulence models or in numerical simulations of compressible flow problems, where the solution may develop discontinuities (shocks) whose poor resolution may effect the global stability of the numerical calculations.

The oscillations along sharp layers are caused by the fact that the SUPG method is neither monotone nor monotonicity preserving. Therefore, various, often nonlinear, 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, at least in some model cases, or which at least reduces the local oscillations. This procedure is referred to as discontinuity capturing or shock capturing. However, these names are not really appropriate in our opinion for several reasons. First, the solution of (1) does not possess shocks or discontinuities because of the presence of diffusion. Instead, steep but continuous layers are formed. Second, the position of these layers is in general already captured well by the SUPG formulation. And third, a confusion might arise with shock capturing methods which are used in the numerical simulation of compressible flows. For these reasons, we propose to call the methods *spurious oscillations at layers diminishing (SOLD) methods* and this name is used throughout the paper.

The literature on SOLD methods is rather extended but the various numerical tests published in the literature do not allow to draw a clear conclusion concerning their advantages and drawbacks. Therefore, the main goal of the present paper is to provide a review of the most published SOLD methods, to discuss the motivations of their

derivation, to present some alternative choices of parameters and to classify them. This review is followed by a numerical comparison of these methods at two test problems whose solutions possess characteristic features of solutions of (1). The numerical results will only give a first insight into the behavior of the SOLD methods and they serve as a pre-selection to identify those SOLD methods which deserve further numerical studies. Comprehensive numerical studies will be presented in the second part of the paper. In order to keep the paper in a reasonable length, we do not consider a reaction term in Eq. (1) since special techniques are necessary if this term is dominant.

A basic problem of all SOLD methods is to find the proper amount of artificial diffusion which leads to sufficiently small nonphysical oscillations (requiring that the artificial diffusion is not 'too small') and to a sufficiently high accuracy (requiring that the artificial diffusion is not 'too large'). Since the artificial diffusion is the sum of the contributions coming from the SUPG term and the SOLD term, the definition of both terms will be thoroughly presented and discussed in this paper.

Sometimes, it is claimed that the SUPG method applied on adaptively refined meshes should be preferred to SOLD methods. However, if convection strongly dominates diffusion, the spurious oscillations of the SUPG method disappear only if extremely fine meshes are used along inner and boundary layers. This leads to a high computational cost which further increases if systems of equations or transient problems are considered. The numerical comparison of the SUPG method on adaptively refined grids and several SOLD methods will be a topic of the second part of the paper. Let us also mention that a further reason for using SOLD methods is that they try to preserve the inverse monotonicity property of the continuous problem.

The plan of the paper is as follows. In the next section, we describe the usual Galerkin discretization of (1) and, in Section 3, we introduce the SUPG method. The accuracy of the SUPG method is greatly influenced by the choice of the stabilizing parameter, which is discussed in Section 4. Then, a detailed review of SOLD methods follows in Section 5. Results of our numerical tests with the SOLD methods at two typical examples are reported in Section 6. Finally, the paper is closed by Section 7 containing our conclusions and an outlook.

Throughout the paper, we use the standard notations $L^p(\Omega)$, $W^{k,p}(\Omega)$, $H^k(\Omega) = W^{k,2}(\Omega)$, $C(\overline{\Omega})$, etc. for the usual function spaces, see e.g. Ciarlet [18]. The norm and seminorm in the Sobolev space $H^k(\Omega)$ will be denoted by $\|\cdot\|_{k,\Omega}$ and $|\cdot|_{k,\Omega}$, respectively. The inner product in the space $L^2(\Omega)$ or $L^2(\Omega)^d$ will be denoted by $(\cdot,\cdot)$. For a vector $\boldsymbol{a} \in \mathbb{R}^d$, the symbol $|\boldsymbol{a}|$ stands for its Euclidean norm.

2. Galerkin's finite element discretization

The starting point of defining any finite element discretization is a weak (or variational) formulation of the respective problem. Denoting by $\tilde{u}_b \in H^1(\Omega)$ an extension of u_b, a natural weak formulation of the convection–diffusion equation (1) reads:

Find $u \in H^1(\Omega)$ such that $u - \tilde{u}_b \in H_0^1(\Omega)$ and

$$a(u,v) = (f,v) \quad \forall v \in H_0^1(\Omega), \tag{2}$$

where

$$a(u,v) = \varepsilon(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v).$$

Since $a(v,v) = \varepsilon|v|_{1,\Omega}^2$ for any $v \in H_0^1(\Omega)$, it easily follows from the Lax–Milgram theorem that this weak formulation has a unique solution (cf. e.g. Ciarlet [18]).

To define a finite element discretization of (1), we introduce a triangulation $\mathscr{T}_h$ of the domain Ω consisting of a finite number of open polygonal resp. polyhedral elements K. The discretization parameter h in the notation $\mathscr{T}_h$ is a positive real number satisfying $\operatorname{diam}(K) \leqslant h$ for any $K \in \mathscr{T}_h$. We assume that $\overline{\Omega} = \bigcup_{K\in\mathscr{T}_h} \overline{K}$ and that the closures of any two different elements K, $\tilde{K} \in \mathscr{T}_h$ are either disjoint or possess either a common vertex or a common edge or, if $d = 3$, a common face. In what follows, we shall confine ourselves to simplicial elements and to elements which are images of a d-dimensional cube under a d-linear mapping (these are general convex quadrilaterals for $d = 2$ and suitable convex hexahedra for $d = 3$). In order to prevent the elements from degenerating when h tends to zero, the elements have to satisfy certain shape-regularity assumptions.

The Galerkin finite element discretization of (1) is now obtained by replacing the space $H_0^1(\Omega)$ in (2) by a finite element subspace V_h (cf. e.g. Ciarlet [18]). In addition, we approximate the function $\tilde{u}_b$ by a finite element interpolate $\tilde{u}_{bh}$. Thus, we may say that $u_h \in H^1(\Omega)$ is a discrete solution of (1) if $u_h - \tilde{u}_{bh} \in V_h$ and

$$a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h.$$

Again, the discrete problem is uniquely solvable.

3. The SUPG method

Since the Galerkin method lacks stability if convection dominates diffusion, we enrich it by a stabilization term proposed by Brooks and Hughes [9] yielding the SUPG method (also called streamline diffusion finite element method, SDFEM). For doing this, we change the assumptions on the space V_h. First, to introduce the SUPG method, the functions from V_h have to be at least of class H^2 inside each element $K \in \mathscr{T}_h$. To simplify further considerations, we shall assume that they are infinitely smooth inside each element, which can be justified by the fact that typical finite element functions are piecewise polynomial. Second, we shall not require that the functions from V_h are continuous across element edges (resp. faces), in order to include nonconforming finite element spaces into the formulation below. Thus, from now on, we assume that V_h is a finite-dimensional space satisfying

$$V_h \subset \{v \in L^2(\Omega); v|_K \in C^\infty(\overline{K})\ \forall K \in \mathscr{T}_h\}.$$

Defining the discrete operators ∇_h and Δ_h by

$$(\nabla_h v)|_K = \nabla(v|_K), \qquad (\Delta_h v)|_K = \Delta(v|_K) \quad \forall K \in \mathscr{T}_h,$$

the bilinear form

$$a_h(u,v) = \varepsilon(\nabla_h u, \nabla_h v) + (\boldsymbol{b}\cdot\nabla_h u, v)$$

and the residual

$$R_h(u) = -\varepsilon\Delta_h u + \boldsymbol{b}\cdot\nabla_h u - f$$

are well defined for $u, v \in V_h$.

Then, the streamline upwind/Petrov–Galerkin (SUPG) method of Brooks and Hughes [9] reads:

Find $u_h \in L^2(\Omega)$ such that $u_h - \tilde{u}_{bh} \in V_h$ and

$$a_h(u_h, v_h) + (R_h(u_h), \tau\boldsymbol{b}\cdot\nabla_h v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{3}$$

where $\tau \in L^\infty(\Omega)$ is a nonnegative stabilization parameter.

For the SUPG method, many theoretical results have been derived, starting with the fundamental work by Nävert [52] and subsequently continued, e.g., by Johnson et al. [43]. Since the analysis of the SUPG method is not the subject of this paper, we shall not present any details and only refer to the monograph by Roos et al. [58].

4. Choice of the SUPG stabilization parameter

An important drawback of many stabilized methods (including the SUPG method) is that they contain stabilization parameters for which a general 'optimal' choice is not known. Since the SUPG method attracted a considerable attention over the last two decades, much research has also been devoted to the choice of the parameter τ. Theoretical investigations of the SUPG method provide certain bounds for τ for which the SUPG method is stable and leads to (quasi-)optimal convergence of the discrete solution u_h. However, it has been reported many times that the choice of τ inside these bounds may dramatically influence the accuracy of the discrete solution. Since most of the SOLD methods considered in this paper are based on the SUPG method, the choice of the stabilization parameter τ plays also a vital role for the results of the SOLD methods. Therefore, possible choices of τ will be discussed in some detail in this section.

It follows from the results of Christie et al. [17] that, for the one-dimensional case of (1) with constant data, the SUPG solution with continuous piecewise linear finite elements on a uniform division of Ω is nodally exact if

$$\tau = \frac{h}{2|b|}\xi_0(Pe) \quad \text{with } \xi_0(\alpha) = \coth\alpha - \frac{1}{\alpha}, \quad Pe = \frac{|b|h}{2\varepsilon}. \tag{4}$$

Here, h is the element length, ξ_0 is the so-called upwind function and Pe is the local Péclet number which determines whether the problem is locally (i.e., within a particular element) convection dominated or diffusion dominated. The parameter τ is often called 'intrinsic time scale' since $h/(2|b|)$ is the time for a particle to travel the distance $h/2$ at a speed equal to $|b|$. Since $\xi_0(\alpha) \to 1$ for $\alpha \to \infty$ and $\xi_0(\alpha)/\alpha \to 1/3$ for $\alpha \to 0+$ (and the SUPG stabilization is not necessary for $\alpha \to 0+$), the function ξ_0 is often approximated by

$$\xi_1(\alpha) = \max\left\{0, 1 - \frac{1}{\alpha}\right\} \quad \text{or} \quad \xi_2(\alpha) = \min\left\{1, \frac{\alpha}{3}\right\}.$$

Brooks and Hughes [9] call these functions 'critical' and 'doubly asymptotic' approximations of ξ_0, respectively. If the right-hand side of (1) is not constant, the choice (4) generally does not lead to a nodally exact discrete solution. Nevertheless, our numerical tests (not reported in this paper) indicate that, in the most cases, the function ξ_0 leads to better results than ξ_1 and ξ_2. However, it should be stressed that, for large values of Pe, the results for these three upwind functions are very close. This is particularly true for ξ_0 and ξ_1, for which $|\xi_0(\alpha) - \xi_1(\alpha)|/\xi_0(\alpha) < 10^{-3}$ for $\alpha > 4$ and $|\xi_0(\alpha) - \xi_1(\alpha)|/\xi_0(\alpha) < 10^{-10}$ for $\alpha > 12$ so that the corresponding discrete solutions are virtually indistinguishable for $Pe > 10$.

Many researchers have tried to find a suitable generalization of (4) to the multidimensional case and to more general finite element spaces V_h. For linear and d-linear finite elements, this generalization usually takes the form

$$\tau|_K \equiv \tau_K = \frac{h_K}{2\|\boldsymbol{b}\|_K}\xi(Pe_K) \quad \text{with } Pe_K = \frac{\|\boldsymbol{b}\|_K h_K}{2\varepsilon}, \tag{5}$$

where K is any element of the triangulation $\mathscr{T}_h$, h_K is a characteristic dimension of K (also called 'local length scale' or 'element length'), $\|\boldsymbol{b}\|_K$ is a suitable norm of $\boldsymbol{b}$, ξ is an upwind function (such that $\xi(\alpha)/\alpha$ is bounded for $\alpha \to 0+$) and Pe_K is the local Péclet number. This generalization seems to be reasonable since, for linear or d-linear finite elements on certain uniform meshes aligned with a constant velocity $\boldsymbol{b}$, the discrete problem corresponds to the one-dimensional case and hence the formula for τ should reduce to (4). For higher order finite elements, the values of Pe_K and τ_K should decrease with increasing polynomial degree on K, see, e.g., Codina et al. [20], Almeida and Silva [3] and Galeão et al. [28]. However, since our numerical tests in Section 6 are performed for linear elements only, we confine ourselves to a discussion of the choice of τ for first order finite elements.

The mentioned correspondence between the one-dimensional and d-dimensional cases particularly implies that, if K is a rectangle and $\boldsymbol{b}$ is constant on K and aligned with one of its edges, one should choose $\|\boldsymbol{b}\|_K = |(\boldsymbol{b}|_K)|$ and h_K equal to the length of the edge $\boldsymbol{b}$ is aligned with. The same holds if K is a right triangle and the vector $\boldsymbol{b}$ is aligned with one of its legs.

Another hint for choosing $\|\boldsymbol{b}\|_K$ and h_K follows from the necessary conditions for uniform convergence of $\|u - u_h\|_{0,\Omega}$ of order greater than 1/2 introduced by Stynes and Tobiska [61]. Let $d = 2$, $\boldsymbol{b} = (b, b)$ with some constant $b \in \mathbb{R}$ and let $\mathscr{T}_h$ be a uniform triangulation of $\Omega = (0,1)^2$ consisting of equal squares or of equal right triangles with hypotenuses in the direction (1,1). Then, for (bi)linear finite elements, the necessary conditions are satisfied if and only if

$$\tau_K = \frac{\mathrm{diam}(K)}{2|\boldsymbol{b}|}\xi_0(Pe_K/2) \quad \text{with } Pe_K = \frac{|\boldsymbol{b}|\mathrm{diam}(K)}{2\varepsilon},$$

where $\mathrm{diam}(K) \equiv \sup\{|\mathbf{x}-\mathbf{y}|; \mathbf{x},\mathbf{y} \in K\}$ is the diameter of K (see Stynes and Tobiska [61] and Shih and Elman [60] for details). The necessary conditions of Stynes and Tobiska were designed for the convection dominated case where $\xi_0(Pe_K/2) \approx \xi_0(Pe_K)$. This suggests to set $\|\boldsymbol{b}\|_K = |\boldsymbol{b}|$ and $h_K = \mathrm{diam}(K)$.

In view of the above considerations, it seems to be reasonable to define h_K as the diameter of K in the direction of the convection $\boldsymbol{b}$. Generally, given a vector $\boldsymbol{s} \in \mathbb{R}^d$, $\boldsymbol{s} \neq \mathbf{0}$, the diameter of K in the direction of $\boldsymbol{s}$ is defined by

$$\mathrm{diam}(K,\boldsymbol{s}) = \sup\{|\mathbf{x}-\mathbf{y}|; \mathbf{x},\mathbf{y} \in K, \mathbf{x}-\mathbf{y} = \alpha\boldsymbol{s}, \alpha \in \mathbb{R}\}.$$

This value may be sometimes difficult to compute and therefore we consider a slightly different definition which was used by Tezduyar and Park [64].

Let N_K be the number of vertices of K and let $\varphi_1,\ldots,\varphi_{N_K}$ be the usual basis functions of $P_1(K)$ (if K is a simplex) or of $Q_1([0,1]^d)$ mapped onto K (if K is a quadrilateral or a hexahedron). We set

$$\mathrm{diam}^*(K,\boldsymbol{s}) = \frac{2|\boldsymbol{s}|}{\sum_{i=1}^{N_K}|\boldsymbol{s}\cdot\nabla\varphi_i(C_K)|},$$

where C_K is the barycentre of K. Then $\mathrm{diam}^*(K,\boldsymbol{s}) = \mathrm{diam}(K,\boldsymbol{s})$ if K is a simplex or a parallelogram. If K is a hexahedron, then generally $\mathrm{diam}^*(K,\boldsymbol{s}) \neq \mathrm{diam}(K,\boldsymbol{s})$ (even not for a cube), but the value of $\mathrm{diam}^*(K,\boldsymbol{s})$ is still reasonable. If $\boldsymbol{s} = \mathbf{0}$, we set $\mathrm{diam}^*(K,\boldsymbol{s}) = \mathrm{diam}(K)$. Using this notation, we define

$$h_K = \mathrm{diam}^*(K,\boldsymbol{b}). \tag{6}$$

The norm $\|\boldsymbol{b}\|_K$ will be defined as the Euclidean norm of $\boldsymbol{b}$, i.e.,

$$\|\boldsymbol{b}\|_K = |\boldsymbol{b}|. \tag{7}$$

Note that, in view of (5)–(7), the parameters h_K, $\|\boldsymbol{b}\|_K$ and, consequently, Pe_K and τ_K are generally functions of the points $\mathbf{x} \in K$.

Usually, the criterion for choosing τ is the accuracy of the discrete solution measured in some suitable norm. Nevertheless, it is also possible to look for τ such that the stiffness matrix corresponding to the discrete problem is well conditioned and enables an efficient application of iterative solvers. This idea was followed by Fischer et al. [27] and Ramage [55,56]. In these papers, Q_1-discretizations of model problems in both two and three dimensions were investigated and it was observed that there is a close relationship between 'best' solution approximation and fast convergence of iterative methods. Particularly, for constant $\boldsymbol{b}$ aligned with a uniform mesh consisting of squares with side length h, an analysis of the structure of eigenvalues of the stiffness matrix reveals that one should choose $\tau = h/(2|\boldsymbol{b}|)$ for $h/\varepsilon \to \infty$ and provides the formula $\tau = h/(2|\boldsymbol{b}|)\xi_1(Pe)$ with $\mathrm{Pe} = |\boldsymbol{b}|h/(2\varepsilon)$ as a significant value with respect to the changes in the eigenvalue structure. In the general case, the choice of h_K as element size in the direction of $\boldsymbol{b}$ is advocated.

In [24], Elman and Ramage examined how the choice of τ influences the oscillations in a bilinear discrete solution and demonstrated that, generally, τ cannot be chosen in such a way that the discrete solution is simultaneously oscillation-free and accurate. The analysis gives a theoretical justification to the formula for τ given by (5)–(7) with $\xi = \xi_1$.

In Harari et al. [31], a formula for τ was found by requiring that the bilinear discrete solution on a uniform mesh is nodally exact for Eq. (1) with $\boldsymbol{b} = \mathrm{const.}$, $f = 0$ and $\Omega = \mathbb{R}^d$. It is interesting to note that, for $\boldsymbol{b}$ aligned with the element diagonals and $h/\varepsilon \to \infty$, the formula of Harari et al. gives only 2/5 of the value obtained from (5)–(7). However, due to the absence of boundary conditions, the investigations of Harari et al. do not seem to be relevant for problems with boundary layers, which is the type of problems the SUPG method was designed for.

The relations (5)–(7) with $\xi = \xi_0$ represent the complete definition of the stabilization parameter τ used in our numerical tests in Section 6. Let us stress that this definition mostly relies on heuristic arguments and the 'best' way of choosing τ for general convection–diffusion problems is not known. Also, many other ways of computing τ have been proposed in the literature. Let us briefly mention a few of them.

Tezduyar and Osawa [63] proposed to compute stabilization parameters using element-level matrices and vectors. In this way, the local length scales, convection field and Péclet number are automatically taken into account. A similar idea was also used by Mizukami [50] for linear finite elements. A comparison of various definitions of local length scales and stabilization parameters can be found in Akin et al. [2]. Let us also mention the work of Akin and Tezduyar [1] where a comparative investigation of various ways of calculating the advective limit of τ is performed.

Roos et al. [58] propose to set

$$\tau_K = \begin{cases} \tau_0 h_K & \text{if } Pe_K > 1, \\ \tau_1 h_K^2/\varepsilon & \text{if } Pe_K \leqslant 1, \end{cases}$$

where τ_0 and τ_1 are appropriate positive constants. This definition of τ leads to the best possible convergence rate of the discrete solution with respect to the streamline diffusion norm. However, an 'optimal' choice of the constants τ_0 and τ_1 is unsolved.

Another possibility of defining the parameter τ is based on the observation that adding bubbles to the finite element space and eliminating them from the Galerkin discretization by static condensation is equivalent to the addition of a stabilizing term of streamline diffusion type. In this way, the question how to define τ is transformed into the question how to define suitable bubbles (cf. e.g. Brezzi and Russo [8]). This question was partially answered by introducing the concept of residual-free bubbles, see e.g. Brezzi et al. [6–8]. Using a similar approach in the

framework of multiscale methods, an analytical formula for τ in terms of element Green's function was derived by Hughes [34]. Another method for stabilizing convection-dominated problems was proposed by Oñate [54], who introduced higher order terms into the continuous problem using the concept of flow balance over a finite domain. Applying the Galerkin method, the SUPG method can be recovered, which also provides a formula for computing the stabilization parameter τ.

5. A review of SOLD methods

In this section, we review most of the SOLD methods introduced during the last two decades to diminish the oscillations arising in the solution of the SUPG discretization (3). Let us recall that these oscillations appear along sharp layers of the solution to the continuous problem (1) due to the fact that the SUPG method is neither monotone nor monotonicity preserving. Therefore, many researchers tried to design such SOLD terms that the resulting discretization satisfies the discrete maximum principle, at least in some model cases. Since linear monotone methods can be at most first-order accurate, it is natural to look for SOLD terms which depend on the discrete solution in a nonlinear way. However, linear SOLD terms applicable to first-order finite elements have also been developed. Let us mention that the discrete maximum principle is an important property of a numerical scheme since it ensures monotonicity and that no spurious oscillations will appear, not even in the vicinity of sharp layers. Moreover, it enables to prove uniform convergence and pointwise stability estimates.

The SOLD methods presented in this section will be divided into five classes. These are upwinding techniques, SOLD methods which add isotropic additional diffusion to (3), SOLD methods which add the additional diffusion to (3) only orthogonally to the streamlines, SOLD methods which rely upon (3) and an edge stabilization, and, finally, SOLD methods based on other ideas.

5.1. Upwinding techniques

One of the first successful monotone methods for solving (1) was introduced by Mizukami and Hughes [51] for conforming linear triangular finite elements. This method is based on the observation that the convection vector $\boldsymbol{b}$ in (1) can be changed in a direction perpendicular to ∇u without affecting the solution u of (1). This suggests that the streamline may not always be the appropriate upwind direction, an idea which has also been used to derive other SOLD methods later. Mizukami and Hughes used this idea to introduce a Petrov–Galerkin method which, due to the arbitrariness in $\boldsymbol{b}$, can be viewed as a method satisfying the discrete maximum principle. In contrast with other upwinding methods for conforming linear triangular finite elements satisfying the discrete maximum principle published earlier (cf. Tabata [62], Kanayama [45], Baba and Tabata [4], Ikeda [37]), the Mizukami–Hughes method adds much less numerical diffusion and provides rather accurate discrete solutions in the most cases. Recently, some improvements of the Mizukami–Hughes method were introduced by Knobloch [46]. Unfortunately, it is not clear how to generalize the Mizukami–Hughes method to other types of finite elements.

At the time as the Mizukami–Hughes scheme was published, Rice and Schnipke [57] proposed another monotone method which is based on a direct streamline upwind approximation to the convective term, rather than modifying the weighting function. This method was developed for bilinear finite elements and again a generalization does not seem to be easy.

5.2. SOLD terms adding isotropic artificial diffusion

Hughes et al. [36] came with the idea to change the upwind direction in the SUPG term of (3) by adding a multiple of the function

$$\boldsymbol{b}_h^{\parallel} = \begin{cases} \dfrac{(\boldsymbol{b}\cdot\nabla u_h)\nabla u_h}{|\nabla u_h|^2} & \text{if } \nabla u_h \neq \boldsymbol{0}, \\ \boldsymbol{0} & \text{if } \nabla u_h = \boldsymbol{0}, \end{cases}$$

which corresponds to the direction in which oscillations in SUPG solutions are observed. This leads to the additional term

$$(R_h(u_h), \sigma \boldsymbol{b}_h^{\parallel}\cdot\nabla_h v_h) \tag{8}$$

on the left-hand side of (3), where σ is a nonnegative stabilization parameter. This additional term controls the derivatives in the direction of the solution gradient, thus increasing the robustness of the SUPG method in the presence of sharp layers. Since $\boldsymbol{b}_h^{\parallel}$ depends on the unknown discrete solution u_h, the resulting method is nonlinear.

Of course, the key point here and in many other SOLD methods is how to choose the parameter σ. Unfortunately, due to the large number of various SOLD methods and the comparatively small amount of theoretical research on them, the correct choice of the respective stabilization parameters is even less clear than for the SUPG method. Often, the definition of these parameters is related to the choice of the parameter τ in the SUPG stabilization. Therefore, it is convenient to introduce the notation $\tau(\boldsymbol{b}^\star)$ representing τ determined by (5)–(7) with $\boldsymbol{b}$ replaced by some function $\boldsymbol{b}^\star$. Note that $\boldsymbol{b}^\star$ influences the value of $\tau_K(\boldsymbol{b}^\star)$ not only through the norm $\|\boldsymbol{b}^\star\|_K$ but also through the definition of h_K.

Now let us return to the choice of σ from (8). One could think of using the value $\tau(\boldsymbol{b}_h^{\parallel})$ but this would lead to a doubling of the SUPG stabilization if $\boldsymbol{b}_h^{\parallel} = \boldsymbol{b}$. Therefore, Hughes et al. [36] proposed to set

$$\sigma = \max\{0, \tau(\boldsymbol{b}_h^{\parallel}) - \tau(\boldsymbol{b})\}. \tag{9}$$

Although, for linear triangular finite elements, the method does not attain the precision of the Mizukami–Hughes

scheme mentioned above (see Hughes et al. [36]), it has the important property that it is applicable to general finite elements.

Tezduyar and Park [64] proposed to redefine $\tau(\boldsymbol{b}_h^{\parallel})$, which leads to

$$\sigma = \frac{h_K^{\parallel}}{2|\boldsymbol{b}_h^{\parallel}|}\eta\left(\frac{|\boldsymbol{b}_h^{\parallel}|}{|\boldsymbol{b}|}\right) \tag{10}$$

with

$$h_K^{\parallel} = \mathrm{diam}^*(K, \boldsymbol{b}_h^{\parallel}), \qquad \eta(\alpha) = 2\alpha(1-\alpha). \tag{11}$$

This definition assures that the SUPG effect is not doubled if $\boldsymbol{b}_h^{\parallel} = \boldsymbol{b}$ and hence an ad hoc correction like (9) is not needed. Tezduyar and Park also observed that the SOLD term (8) with the above definitions of σ depends only on the direction of ∇u_h but not on its magnitude. Since the SOLD term is required only along steep gradients of the solution, they suggested to use

$$\sigma = \frac{h_K^{\parallel}}{2|\boldsymbol{b}_h^{\parallel}|}\eta\left(\frac{|\boldsymbol{b}_h^{\parallel}|}{|\boldsymbol{b}|}\right) h_K^{\parallel}\frac{|\nabla u_h|}{u_0}, \tag{12}$$

where u_0 is a global scaling value for u_h.

An approach related to the above-described method of Hughes et al. [36] was used by de Sampaio and Coutinho [59], who introduced the concept of the effective transport velocity $\boldsymbol{b}^{\parallel}$ defined on the continuum level analogously as $\boldsymbol{b}_h^{\parallel}$ (i.e., with u instead of u_h). Before performing a discretization, the convective field $\boldsymbol{b}$ in (1) is replaced by $\tilde{\boldsymbol{b}} = \gamma\boldsymbol{b} + (1-\gamma)\boldsymbol{b}^{\parallel}$ with $\gamma \in [0,1]$. Then, an application of a standard discretization technique like the Galerkin/least-squares or, in our case, SUPG method yields a 'Petrov–Galerkin method containing a SOLD term. The method uses only one stabilization parameter (defined using the discrete counterpart of $\tilde{\boldsymbol{b}}$) and hence an alignment of $\boldsymbol{b}$ and ∇u does not create the undesirable doubling effect discussed above. However, it is not clear how to choose the parameter γ and, therefore, the value $\gamma = 0.5$ is recommended except for regions where $\nabla u = \boldsymbol{0}$.

Now, let us return to the SOLD term (8) which can be written in the form

$$(\tilde{\varepsilon}\nabla_h u_h, \nabla_h v_h) \tag{13}$$

with

$$\tilde{\varepsilon} = \begin{cases} \sigma\dfrac{R_h(u_h)\boldsymbol{b}\cdot\nabla u_h}{|\nabla u_h|^2} & \text{if } \nabla u_h \neq \boldsymbol{0}, \\ 0 & \text{if } \nabla u_h = \boldsymbol{0}. \end{cases} \tag{14}$$

Galeão and do Carmo [29] observed that, when $f \neq 0$ in (1), this SOLD term does not prevent localized oscillations in the discrete solution. The reason is that this term introduces a negative artificial diffusion $\tilde{\varepsilon}$ if $R_h(u_h)\boldsymbol{b}\cdot\nabla u_h < 0$. As a remedy, Galeão and do Carmo proposed to replace the flow velocity $\boldsymbol{b}$ in the SUPG stabilization term by an approximate upwind direction

$$\boldsymbol{b}_h^{up} = \alpha_1\boldsymbol{b} + \alpha_2\boldsymbol{b}_h,$$

where $\boldsymbol{b}_h$ is an approximate streamline direction such that, for any $K \in \mathscr{T}_h$, the discrete solution u_h satisfies

$$-\varepsilon\Delta u_h + \boldsymbol{b}_h\cdot\nabla u_h = f \quad \text{in } K. \tag{15}$$

Of course, such $\boldsymbol{b}_h$ generally does not exist at those points of K at which $\nabla u_h = \boldsymbol{0}$. Therefore, we replace (15) by

$$(-\varepsilon\Delta u_h + \boldsymbol{b}_h\cdot\nabla u_h - f)|\nabla u_h| = 0 \quad \text{in } K. \tag{16}$$

A reasonable choice of $\boldsymbol{b}_h$ is $\boldsymbol{b}_h = \boldsymbol{b} - \boldsymbol{z}_h$ with

$$\boldsymbol{z}_h = \begin{cases} \dfrac{R_h(u_h)\nabla u_h}{|\nabla u_h|^2} & \text{if } \nabla u_h \neq \boldsymbol{0}, \\ \boldsymbol{0} & \text{if } \nabla u_h = \boldsymbol{0}, \end{cases}$$

since it minimizes $|\boldsymbol{b}_h - \boldsymbol{b}|$ in any $K \in \mathscr{T}_h$ among all functions $\boldsymbol{b}_h$ satisfying (16). Defining the SUPG stabilization using the approximate upwind direction $\boldsymbol{b}_h^{up}$, we obtain the discretization (3) with the additional term

$$(R_h(u_h), \sigma\boldsymbol{z}_h\cdot\nabla_h v_h) \tag{17}$$

on the left-hand side. The parameter $\tau \equiv \alpha_1 + \alpha_2$ is defined as before and the choice of $\sigma \equiv -\alpha_2$ will be discussed in the following. The SOLD term (17) can be written in the form (13) with

$$\tilde{\varepsilon} = \begin{cases} \sigma\dfrac{|R_h(u_h)|^2}{|\nabla u_h|^2} & \text{if } \nabla u_h \neq \boldsymbol{0}, \\ 0 & \text{if } \nabla u_h = \boldsymbol{0}, \end{cases} \tag{18}$$

and hence it again introduces an isotropic artificial diffusion.

If $f = 0$ and $\Delta_h u_h = 0$ (which holds for (bi,tri)linear finite elements), we have $\boldsymbol{z}_h = \boldsymbol{b}_h^{\parallel}$. Hence, the terms (8) and (17) are the same provided that the parameters σ are defined appropriately. Galeão and do Carmo [29] used (17) with

$$\sigma = \max\{0, \tau(\boldsymbol{z}_h) - \tau(\boldsymbol{b})\}, \tag{19}$$

which is identical with (9) if $\boldsymbol{z}_h = \boldsymbol{b}_h^{\parallel}$. Do Carmo and Galeão [16] proposed to simplify (19) to

$$\sigma = \tau(\boldsymbol{b})\max\left\{0, \frac{|\boldsymbol{b}|}{|\boldsymbol{z}_h|} - 1\right\}, \tag{20}$$

which assures that the term (17) is added only if $|\boldsymbol{b}| > |\boldsymbol{z}_h|$, i.e., only if the above-introduced vector $\boldsymbol{b}_h$ satisfies the natural requirement $\boldsymbol{b}\cdot\boldsymbol{b}_h > 0$.

For problems with regular solutions, it was observed that the SOLD term (17) adds an undesirable crosswind diffusion and that the discrete solution is less accurate than for the SUPG method. Therefore, do Carmo and Galeão [16] introduced a feedback function which should minimize the influence of the SOLD term (17) in regions where the solution of (1) is smooth. Since the definition of the feedback function is rather involved, we only refer to [16].

The intricacy of the feedback approach of do Carmo and Galeão [16] motivated do Carmo and Alvarez [14] to introduce a simpler expression for the parameter σ. For this, the following parameters are used on any element $K \in \mathscr{T}_h$:

$$\alpha_K = \frac{|\boldsymbol{z}_h|}{|\boldsymbol{b}|}, \qquad \beta_K = \min\{1, h_K\}^{1-\alpha_K^2},$$

$$\gamma_K = \min\{\beta_K, \tfrac{1}{2}(\alpha_K + \beta_K)\},$$

$$\lambda_K = \frac{\max\{\alpha_K, |R_h(u_h)|\}^{3+\alpha_K/2+\alpha_K^2}}{\gamma_K^{\max\{1/2,1/4+\alpha_K\}}},$$

$$\kappa_K = |2 - \lambda_K|^{\frac{1-\lambda_K}{1+\alpha_K}} - 1, \qquad \omega_K = \frac{\alpha_K^2 \gamma_K^{2-\alpha_K^2}}{\tau_K(\boldsymbol{b})}.$$

Now, denoting by $\bar{\sigma}$ the value of σ defined by (19), do Carmo and Alvarez [14] consider (17) with

$$\sigma = \varrho\bar{\sigma}, \tag{21}$$

where

$$\varrho|_K = \begin{cases} 1 & \text{if } \alpha_K \geqslant 1 \text{ or } \lambda_K \geqslant 1, \\ [\omega_K \bar{\sigma}]^{\kappa_K} & \text{if } \alpha_K < 1 \text{ and } \lambda_K < 1 \end{cases} \quad \forall K \in \mathscr{T}_h. \tag{22}$$

Like the above-mentioned feedback function, the parameter ϱ should suppress the addition of the artificial diffusion in regions where the solution of (1) is smooth.

In [15], do Carmo and Alvarez introduced a finer tuning of the parameters τ and σ by multiplying them by a factor τ_0 on those elements $K \in \mathscr{T}_h$ whose boundary intersects the outflow part of the boundary of Ω. The value of τ_0 on an element K depends on the geometry of K and the polynomial degree of shape functions on K. Based on numerical experiments, do Carmo and Alvarez set $\tau_0 = 1$ for bilinear shape functions on quadrilaterals, $\tau_0 = 0.5$ for biquadratic shape functions on quadrilaterals or linear shape functions on triangles and $\tau_0 = 0.25$ for quadratic shape functions on triangles.

A remedy for the above-mentioned loss of accuracy which appears when (17) with (19) or (20) is used was also proposed by Almeida and Silva [3], who conjectured that this loss of accuracy was mainly caused by the incapability of the formulas (19) and (20) to avoid the doubling effect. They observed that, setting $v_h = u_h$, the SUPG term in (3) becomes

$$(R_h(u_h), \tau\boldsymbol{b}\cdot\nabla_h u_h) = (R_h(u_h), \tau\vartheta_h \boldsymbol{z}_h \cdot \nabla_h u_h)$$

with

$$\vartheta_h = \frac{\boldsymbol{b}\cdot\nabla_h u_h}{R_h(u_h)}.$$

Therefore, they proposed to replace (20) by

$$\sigma = \tau(\boldsymbol{b}) \max\left\{0, \frac{|\boldsymbol{b}|}{|\boldsymbol{z}_h|} - \zeta_h\right\} \quad \text{with } \zeta_h = \max\left\{1, \frac{\boldsymbol{b}\cdot\nabla_h u_h}{R_h(u_h)}\right\}, \tag{23}$$

which provides a reduction of the amount of artificial diffusion along the $\boldsymbol{z}_h$ direction, which is the direction of the approximate solution gradient.

In order to be able to prove some theoretical results on SOLD methods of the above type, Knopp et al. [47] suggested to replace the isotropic artificial diffusion in (13) by

$$\tilde{\varepsilon}|_K = \sigma_K(u_h)|Q_K(u_h)|^2 \quad \forall K \in \mathscr{T}_h \tag{24}$$

with some appropriate constants $\sigma_K(u_h) \geqslant 0$ (e.g., defined by (19) or (20)) and

$$Q_K(u_h) = \frac{\|R_h(u_h)\|_{0,K}}{S_K + \|u_h\|_{1,K}}, \tag{25}$$

S_K being some constants (equal to 1 in numerical experiments of [47]).

The SOLD term (13) was also used by Johnson [41], who proposed to set

$$\tilde{\varepsilon}|_K = \max\{0, \alpha[\mathrm{diam}(K)]^\nu |R_h(u_h)| - \varepsilon\} \quad \forall K \in \mathscr{T}_h \tag{26}$$

with some constants α and $\nu \in (3/2, 2)$. He suggested to take $\nu \sim 2$. Johnson [42] replaced α by $\beta/\max_\Omega |u_h|$ and proposed to set $\beta = 0.1$. A similar approach was also used by Johnson et al. [44]. A priori and a posteriori error estimates for this type of SOLD discretizations can be found in the papers by Johnson [41] and Eriksson and Johnson [25]. The mentioned papers [42,44] contain convergence results for space–time elements.

5.3. SOLD terms adding crosswind artificial diffusion

An alternative approach to the above SOLD methods is to modify the SUPG discretization (3) by adding artificial diffusion in the crosswind direction only as considered by Johnson et al. [43] for the two-dimensional case with $\boldsymbol{b} = (1, 0)$ and $u_b = 0$. A straightforward generalization of this approach leads to the additional term

$$(\tilde{\varepsilon} D \nabla_h u_h, \nabla_h v_h) \tag{27}$$

on the left-hand side of (3), where

$$\tilde{\varepsilon}|_K = \max\{0, |\boldsymbol{b}| h_K^{3/2} - \varepsilon\} \quad \forall K \in \mathscr{T}_h \tag{28}$$

and D is the projection onto the line or plane orthogonal to $\boldsymbol{b}$ defined by

$$D = \begin{cases} I - \dfrac{\boldsymbol{b}\otimes\boldsymbol{b}}{|\boldsymbol{b}|^2} & \text{if } \boldsymbol{b} \neq \boldsymbol{0}, \\ 0 & \text{if } \boldsymbol{b} = \boldsymbol{0}, \end{cases}$$

I being the identity tensor. The value $h_K^{3/2}$ was motivated by a careful analysis of the numerical crosswind spread in the discrete problem, i.e., of the maximal distance in which the right-hand side f significantly influences the discrete solution. The resulting method is linear but non-consistent and hence it is restricted to finite elements of first order of accuracy. For the two-dimensional case with $\boldsymbol{b} = (1, 0)$, $u_b = 0$ and a reaction term in (1), Johnson et al. [43] proved pointwise error estimates of order $O(h^{5/4})$ in regions of smoothness and a global L^1-estimate of order $O(h^{1/2})$. Later, these results were improved by Niijima [53], Zhou and Rannacher [66] and Zhou [65]. Note that, in the two-dimensional case, the SOLD term (27) can be written in the form

$$(\tilde{\varepsilon}\boldsymbol{b}^\perp \cdot \nabla_h u_h, \boldsymbol{b}^\perp \cdot \nabla_h v_h) \quad \text{with } \boldsymbol{b}^\perp = \frac{(-b_2, b_1)}{|\boldsymbol{b}|}. \tag{29}$$

Shih and Elman [60] considered the SUPG discretization (3) with the additional term (29) for $\Omega = (0,1)^2$ and a constant vector $\boldsymbol{b}$. They used bilinear finite elements on a uniform triangulation of Ω and proposed two choices of the parameters τ and $\tilde{\varepsilon}$ based on the requirement that the necessary conditions for uniform convergence of $\|u - u_h\|_{0,\Omega}$ of order greater than 1/2 introduced by Stynes and Tobiska [61] hold. However, both methods of Shih and Elman reduce to the SUPG discretization (3) whenever the flow vector $\boldsymbol{b}$ is aligned with the mesh, which indicates that the methods generally cannot work properly. Therefore, we do not consider them in our numerical tests.

Codina [19] proposed to set the amount of the artificial crosswind diffusion $\tilde{\varepsilon}$ in (27), for any $K \in \mathscr{T}_h$, to

$$\tilde{\varepsilon}|_K = \frac{1}{2}\max\left\{0, C - \frac{2\varepsilon}{|\boldsymbol{b}_h^{\parallel}|\,\mathrm{diam}(K)}\right\}\mathrm{diam}(K)\frac{|R_h(u_h)|}{|\nabla u_h|} \quad (30)$$

(if $\nabla u_h \neq \boldsymbol{0}$), where C is a suitable constant. Codina [19] reports that two-dimensional numerical experiments suggest to set $C \approx 0.7$ for (bi)linear finite elements and $C \approx 0.35$ for (bi)quadratic finite elements. The design of (30) is based on investigations of the validity of the discrete maximum principle for several simple model problems and on the requirements that $\tilde{\varepsilon}$ should be small in regions where $|\boldsymbol{b}\cdot\nabla u_h|$ is small (to avoid excessive overdamping) and proportional to the element residual (to guarantee consistency).

Knopp et al. [47] proposed to use (27) with $\tilde{\varepsilon}$ defined, for any $K \in \mathscr{T}_h$, by

$$\tilde{\varepsilon}|_K = \frac{1}{2}\max\left\{0, C - \frac{2\varepsilon}{Q_K(u_h)\mathrm{diam}(K)}\right\}\mathrm{diam}(K)Q_K(u_h), \quad (31)$$

where $Q_K(u_h)$ is given by (25). This was also motivated by a posteriori error estimates which show that the action of the SOLD stabilization should be restricted to regions where the local residual is not small. Like in case of (24) with (25), this definition of $\tilde{\varepsilon}$ satisfies assumptions enabling Knopp et al. [47] to perform a priori and a posteriori error analyses of a rather general class of nonlinear discretizations of (1) which include SOLD discretizations with stabilizing terms defined by (27), (31), (25) or (13), (24), (25).

Combining the above two definitions of $\tilde{\varepsilon}$, we further propose to use (27) with $\tilde{\varepsilon}$ defined by (31) where

$$Q_K(u_h) = \frac{|R_h(u_h)|}{|\nabla u_h|} \quad \text{if } \nabla u_h \neq \boldsymbol{0}. \quad (32)$$

This is equivalent to (30) if $f = 0$ and $\Delta_h u_h = 0$. Another possibility is to set

$$Q_K(u_h) = \frac{\|R_h(u_h)\|_{0,K}}{|u_h|_{1,K}}.$$

For the computations considered in this paper (P_1 finite element, constant data $\boldsymbol{b}$ and f in (1)), this value is identical with (32).

It was proposed by Codina and Soto [21] to add both isotropic and crosswind artificial diffusion terms to the left-hand side of (3). Denoting the parameters in (13) and (27) by $\tilde{\varepsilon}^{\mathrm{iso}}$ and $\tilde{\varepsilon}^{\mathrm{cross}}$, respectively, the parameter choice from [21] is

$$\tilde{\varepsilon}^{\mathrm{iso}} = \max\{0, \tilde{\varepsilon}^{\mathrm{dc}} - \tau(\boldsymbol{b})|\boldsymbol{b}|^2\}, \qquad \tilde{\varepsilon}^{\mathrm{cross}} = \tilde{\varepsilon}^{\mathrm{dc}} - \tilde{\varepsilon}^{\mathrm{iso}},$$

with $\tilde{\varepsilon}^{\mathrm{dc}}$ defined similarly to (31) and $Q_K(u_h)$ given in (32). We found in our numerical tests that the results are very similar to those obtained with $\tilde{\varepsilon}$ defined by (27), (31) and (32) (method denoted by KLR02_3 below). For this reason, numerical results for the method from [21] will not be presented.

Burman and Ern [11] derived formulas for $\tilde{\varepsilon}$ in (27) and (13) that guarantee a discrete maximum principle for strictly acute meshes and linear simplicial finite elements. However, they observed that, from a numerical viewpoint, the stronger one wishes to enforce a discrete maximum principle, the more ill behaved the nonlinear discrete equations become. Therefore, they slightly changed the formulas implied by the theoretical investigations and recommended to use (27) with $\tilde{\varepsilon}$ defined, on any $K \in \mathscr{T}_h$, by

$$\tilde{\varepsilon}|_K = \frac{\tau(\boldsymbol{b})|\boldsymbol{b}|^2|R_h(u_h)|}{|\boldsymbol{b}||\nabla_h u_h| + |R_h(u_h)|} \times \frac{|\boldsymbol{b}||\nabla_h u_h| + |R_h(u_h)| + \tan\alpha_K|\boldsymbol{b}||D\nabla_h u_h|}{|R_h(u_h)| + \tan\alpha_K|\boldsymbol{b}||D\nabla_h u_h|} \quad (33)$$

($\tilde{\varepsilon} = 0$ if one of the denominators vanishes). The parameter α_K is equal to $\pi/2 - \beta_K$ where β_K is the largest angle of K if K is a triangle and β_K is the largest angle among the six pairs of faces of K if K is a tetrahedron. If $\beta_K = \pi/2$ (and hence the strictly acute condition is violated), it is recommended to set $\alpha_K = \pi/6$. Further, to improve the convergence of the nonlinear iterations, it is recommended to replace the absolute value $|x|$ of a real number x by the regularized expression $|x|_{\mathrm{reg}} \equiv x\tanh(x/\varepsilon_{\mathrm{reg}})$. We apply this regularization only to $|R_h(u_h)|$ and set $\varepsilon_{\mathrm{reg}} = 2$.

Our numerical experiments in Section 6 indicate that the above artificial diffusion $\tilde{\varepsilon}$ is too large and therefore we also consider (27) with $\tilde{\varepsilon}$ defined by

$$\tilde{\varepsilon} = \frac{\tau(\boldsymbol{b})|\boldsymbol{b}|^2|R_h(u_h)|}{|\boldsymbol{b}||\nabla_h u_h| + |R_h(u_h)|}. \quad (34)$$

In this case, we do not apply any regularization of the absolute values.

An apparently similar simplification of (33) given, on any $K \in \mathscr{T}_h$, by

$$\tilde{\varepsilon}|_K = \frac{\tau(\boldsymbol{b})|\boldsymbol{b}|^2|R_h(u_h)|}{\sqrt{|R_h(u_h)|^2 + (\tan\alpha_K)^2|\boldsymbol{b}|^2|D\nabla_h u_h|^2}} \quad (35)$$

was also proposed by Burman and Ern [11]. Like in (33), we regularize $|R_h(u_h)|$ to improve the convergence of the nonlinear iterations. We shall see that (34) and (35) lead to qualitatively different results.

5.4. Edge stabilization methods

Another SOLD strategy for linear simplicial finite elements was introduced by Burman and Hansbo [13]. The SOLD term to be added to the left-hand side of (3) is defined by

$$\sum_{K\in\mathscr{T}_h}\int_{\partial K}\Psi_K(u_h)\mathrm{sign}(\boldsymbol{t}_{\partial K}\cdot\nabla(u_h|_K))\boldsymbol{t}_{\partial K}\cdot\nabla(v_h|_K)\mathrm{d}\sigma, \tag{36}$$

where $\boldsymbol{t}_{\partial K}$ is a tangent vector to the boundary ∂K of K,

$$\Psi_K(u_h)=\mathrm{diam}(K)(C_1\varepsilon+C_2\mathrm{diam}(K))\max_{E\subset\partial K}|[|\boldsymbol{n}_E\cdot\nabla u_h|]_E|, \tag{37}$$

$\boldsymbol{n}_E$ are normal vectors to edges E of K, $[|v|]_E$ denotes the jump of a function v across the edge E and C_1, C_2 are appropriate constants (note that C_2 has to be proportional to $|\boldsymbol{b}|$). Burman and Hansbo proved that, using an edge stabilization instead of the SUPG term, the discrete maximum principle is satisfied provided that $C_1\geqslant 1/2$ and C_2 is sufficiently large. In their numerical tests with $|\boldsymbol{b}|=1$, they used $C_2=10$. To improve the convergence of the nonlinear iterative process, they further regularize the sign operator in (36) by replacing it by the hyperbolic tangent.

Burman and Ern [12] proposed to use the SOLD term (36) with $\Psi_K(u_h)$ defined by

$$\Psi_K(u_h)|_E=C|\boldsymbol{b}|[\mathrm{diam}(K)]^2|[|\nabla u_h|]_E| \qquad \forall E\subset\partial K, \tag{38}$$

where C is a suitable constant. For linear simplicial finite elements on weakly acute triangulations satisfying a local quasi-uniformity property, they proved the validity of the discrete maximum principle. Another definition of $\Psi_K(u_h)$ proposed in [12] is

$$\Psi_K(u_h)=C|R_h(u_h)|. \tag{39}$$

Let us mention that establishing a discrete maximum principle for higher order stabilized Galerkin methods still remains an open problem.

5.5. Further SOLD methods

At the end of this review of SOLD methods, we will mention some further approaches for reducing spurious oscillations in SUPG solutions. Lube [49] presented an asymptotically fitted variant of the SUPG method which suppresses oscillations along boundary layers. This method consists in replacing the Dirichlet boundary conditions on the downstream (if $\varepsilon<Ch$) and characteristic (if $\varepsilon<h^{3/2}$) parts of the boundary by homogeneous Neumann's conditions. Existence, stability and convergence results are proved for (1) containing a suitable reaction term. Burman [10] and Hughes and Bazilevs [5] demonstrated numerically that using weakly imposed Dirichlet boundary conditions reduces spurious oscillations at outflow boundaries considerably. The consequence of this approach is, however, that the Dirichlet values of the discrete solution will in general not coincide with the given boundary condition.

If $f=0$ in (1), the maximum principle yields a lower bound $u_{\min}$ and an upper bound $u_{\max}$ for the solution u. Layton and Polman [48] proposed to add the nonlinear term

$$ch^{-\alpha}\left[\min_{\text{grid points}}\{u_h(x,y)-u_{\min},0\}+\max_{\text{grid points}}\{u_h(x,y)-u_{\max},0\}\right]$$

to the left-hand side of the SUPG Eq. (3), e.g., with $c=1$, $\alpha=1$. This term penalizes the violation of the discrete maximum principle. However, if $f\neq 0$ or if other types of boundary conditions are used, it is hard to obtain the bounds and this method is not generally applicable. Even for the examples presented in Section 6, it was never among the best methods (results not explicitly reported in this paper).

Guermond [30] studied stabilized schemes based on the minimization of the residual in $L^1(\Omega)$ for first order partial differential equations. Since the second order derivatives are small in a convection-dominated convection–diffusion equation, its solution has similar features as the solution of a first order transport equation, for instance steep layers on the one hand and shocks on the other hand. In Example 4.5 in [30] it is demonstrated that the $L^1(\Omega)$ minimization approach can be used also for convection–diffusion equations.

6. Numerical studies

This section presents results of two numerical examples which are defined in a two-dimensional domain and which are discretized by conforming piecewise linear finite elements. The only criterion for the evaluation of the SOLD methods will be the quality of the computed solution. This evaluation is twofold: the suppressing of spurious oscillations and the smearing of layers will be rated. Since spurious oscillations are far more undesirable than moderately smeared layers, the results concerning spurious oscillations will be weighted higher. We would like to note that the evaluation of the many computational results is rather complicated. The difficulty is that not errors to a known solution are of interest but the size of oscillations and the extent of smearing of layers. Measuring the size of oscillations is only easy if the solution should be constant on both sides of the layer. Often, pictures of the computed solutions give a good impression of their quality. However, due to the considerable potential length of the paper, it is not possible to support each computation with one or even more pictures. Several measures for evaluating the results were tested in our numerical studies. We found out that the measures used below are appropriate ones.

The numerical results presented in this paper give only a first impression of the capabilities of the SOLD methods. They will serve as a pre-selection of those methods which are worth to be studied in detail, also with respect to other properties like the convergence of the solution in various norms or the speed of convergence of the nonlinear iteration process. Comprehensive numerical studies of these methods will be postponed to the second part of this paper. For some additional numerical studies, we refer to [38,39].

We shall test most of the SOLD methods considered in Section 5. A summary of these methods, introducing also their abbreviations which will be used in the evaluation of the numerical examples, is presented in Table 1. The underlying SUPG method (3) was applied with τ defined by (5)–(7) using the upwind function ξ_0 from (4). The nonlinear problems were solved accurately, up to a norm of the residual lower than 10^{-10}. Methods which worked best in our opinion are printed boldly in the tables. Italic is used for methods which also produced acceptable results but which were clearly worse than the best methods. All numerical results have been double-checked by computing them with two different codes, one of them was *MooNMD*, [40].

Table 1
Summary of SOLD methods considered in the numerical tests

Name	Citation	Add. diffusion	Method param.	User param.
MH85	[46]	upwind	–	–
HMM86	[36]	iso. (13)	(14), (9)	–
TP86_1	[64]	iso. (13)	(14), (10), (11)	–
TP86_2	[64]	iso. (13)	(14), (11), (12)	u_0
GdC88	[29]	iso. (13)	(18), (19)	–
dCG91	[16]	iso. (13)	(18), (20)	–
dCA03	[14]	iso. (13)	(18), (21), (22)	–
AS97	[3]	iso. (13)	(18), (23)	–
KLR02_1	[47]	iso. (13)	(24), (19), (25)	S_K
J90	[41]	iso. (13)	(26)	α, ν
JSW87	[43]	orth. (27)	(28)	–
C93	[19]	orth. (27)	(30)	C
KLR02_2	[47]	orth. (27)	(31), (25)	C, S_K
KLR02_3	[47], here	orth. (27)	(31), (32)	C
BE02_1	[11]	orth. (27)	(33)	α_K
BE02_2	[11], here	orth. (27)	(34)	–
BE02_3	[11]	orth. (27)	(35)	α_K
BH04	[13]	edge (36)	(37)	C_1, C_2
BE05_1	[12]	edge (36)	(38)	C
BE05_2	[12]	edge (36)	(39)	C

Example 1 (*Solution with parabolic and exponential boundary layers*). We consider the convection–diffusion equation (1) in $\Omega = (0,1)^2$ with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1,0)^{\mathrm{T}}$, $f = 1$ and $u_b = 0$. The solution $u(x,y)$ of this problem, see Fig. 1, possesses an exponential boundary layer at $x = 1$ and parabolic boundary layers at $y = 0$ and $y = 1$. In the interior grid points, the solution $u(x,y)$ is very close to x.

The numerical tests were performed on a regular and on an unstructured triangular grid, see Fig. 2 for the initial

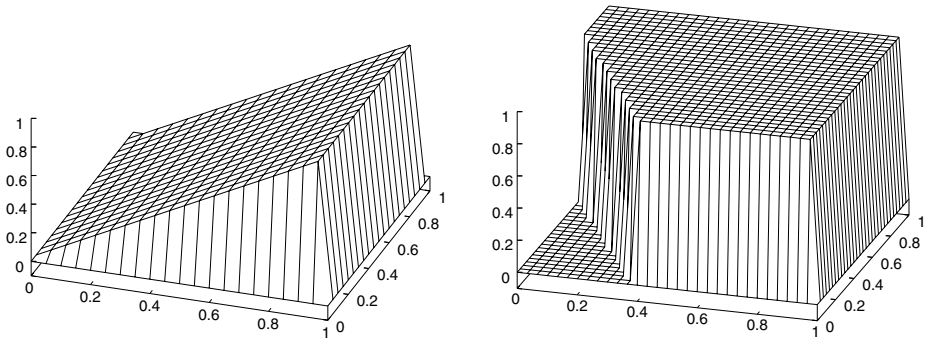


Fig. 1. Solution of Example 1 (left) and of Example 2 (right).

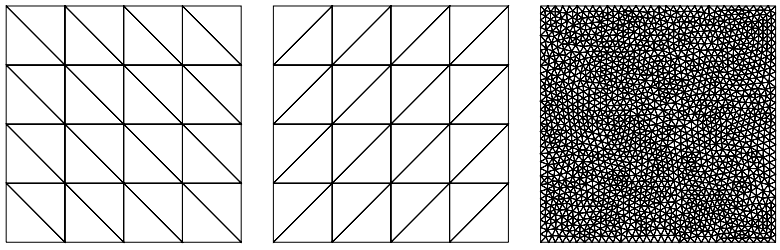


Fig. 2. The grids used in the computations: Grid 1, Grid 2 and Grid 3 (left to right). The structured grids are refined till the length of the legs of the triangles is 1/64.

regular grid (Grid 1) and the final unstructured grid (Grid 3). The latter was obtained using the anisotropic mesh adaptation technique of [22].

First, we present computations on Grid 1 where the length of the legs of the triangles was 1/64. Thus, from (6) follows $h_K = 1/64$ and the Péclet number is $Pe_K = 10^8/128 = 781{,}250$. The number of degrees of freedom is 4225 (including Dirichlet nodes).

For this special example, the stabilization parameter τ used in this paper is optimal along lines $y = \text{const}$ outside the parabolic layers. Applying the SUPG method on Grid 1, one finds that there are no oscillations at the exponential layer. However, there are still strong oscillations at the parabolic layers and for this reason we will concentrate on these layers in the evaluation of the SOLD methods on Grid 1. Particularly, we consider the cut line $x = 0.5$ and the values

$$osc := \max_{y \in \left\{\frac{1}{64}, \frac{2}{64}, \ldots, \frac{63}{64}\right\}} \{u_h(0.5, y) - u_h(0.5, 0.5)\}, \tag{40}$$

$$smear := \max_{y \in \left\{\frac{1}{64}, \frac{2}{64}, \ldots, \frac{63}{64}\right\}} \{u_h(0.5, 0.5) - u_h(0.5, y)\}. \tag{41}$$

The first value measures the oscillations in the parabolic layers. In the case that the oscillations are suppressed to the most part, the second value measures the smearing of these layers. The computational results are given in Table 2 and Fig. 3. To simplify their evaluation and the ranking of the methods, we scored each result. The scores are as follows:

$osc \in$	Score	$smear \in$	Score
[0,1e−3)	4	[0,1e−5)	2
[1e−3,1e−2)	2	[1e−5,1e−3)	1
[1e−2,1e−1)	0	[1e−3,1e−1)	0
[1e−1,1)	−4	[1e−1,1)	−2

Values which are close to the interval with the next higher score will get an intermediate score.

Clearly the best method is MH85. Good results were computed also with dCG91, AS97, KLR02_3 and BE02_2. All other methods, save JSW87, still exhibit non-negligible spurious oscillations at the parabolic layers. These layers are smeared considerably in the solution computed with JSW87. In addition, we want to note that the solutions obtained with J90, BH04 and BE05_2 show, in contrast to all other methods, a smearing of the exponential boundary layer.

Table 4 and Figs. 5 and 6 present results obtained on the unstructured Grid 3 from Fig. 2. This grid possesses 3312 triangles and 1721 vertices (degrees of freedom). Introducing the sets

$$\Omega_1 = \Omega_2 \cup \Omega_3, \quad \Omega_2 = (0, 0.9) \times (0, 0.1],$$
$$\Omega_3 = (0, 0.9) \times [0.9, 1), \quad \Omega_4 = [0.9, 1) \times (0.1, 0.9),$$

see Fig. 4, the spurious oscillations are measured by

$$osc_{\text{para}(1)} := \max_{(x,y) \in \Omega_1} (u_h(x, y) - x), \tag{42}$$

Table 2
Example 1, Grid 1, *osc* and *smear* defined in (40) and (41)

Name	$osc \in$	Score	$smear \in$	Score	Total
SUPG	1.340e−1	−4	–	–	−4
MH85	0	4	5.280e−6	2	6
HMM86	8.737e−2	0	1.141e−2	0	0
TP86_1	1.150e−1	−4	–	–	−4
TP86_2; $u_0 = 1$	1.312e−1	−4	–	–	−4
GdC88	2.179e−3	2	4.860e−2	0	2
dCG91	5.992e−4	4	4.515e−2	0	4
dCA03	1.316e−2	1	4.387e−2	0	1
AS97	4.742e−4	4	4.494e−2	0	4
KLR02_1; $S_K = 1$	1.241e−1	−4	–	–	−4
J90; $\alpha = 0.5, v = 2$	4.273e−3	2	1.540e−3	0	2
JSW87	1.479e−6	4	2.743e−1	−2	2
C93; $C = 0.6$	7.816e−2	0	8.076e−4	1	1
KLR02_2; $C = 0.6, S_K = 1$	9.654e−2	0	2.383e−2	0	0
KLR02_3; $C = 0.6$	2.469e−4	4	3.680e−2	0	4
BE02_1; $\alpha_K = \pi/6$	1.528e−2	1	9.184e−2	0	1
BE02_2	6.942e−4	4	4.729e−2	0	4
BE02_3; $\alpha_K = \pi/6$	6.406e−3	2	2.496e−2	0	2
BH04; $C_1 = 0.5, C_2 = 0.01$	2.477e−3	2	2.168e−1	−2	0
BE05_1; $C = 0.05$	6.765e−3	2	7.212e−2	0	2
BE05_2; $C = 5e{-}5$	2.826e−3	2	1.489e−1	−2	0

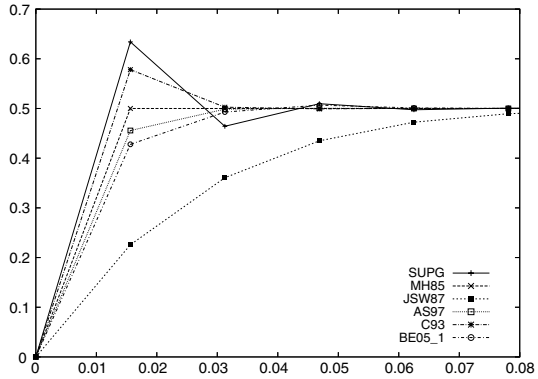


Fig. 3. Example 1, Grid 1, the parabolic boundary layer computed with different schemes.

$$osc_{\text{para}(2)} := \max\left\{ \max_{(x_s,y_s) \in \Omega_2} \left(-\frac{\partial u_h(x_s, y_s)}{\partial y} \right), \max_{(x_s,y_s) \in \Omega_3} \frac{\partial u_h(x_s, y_s)}{\partial y} \right\}, \tag{43}$$

$$osc_{\text{exp}} := \max_{(x_s,y_s) \in \Omega_4} \frac{\partial u_h(x_s, y_s)}{\partial x}, \tag{44}$$

where (x, y) are the nodes in Ω_1 and (x_s, y_s) are the coordinates of the barycentres of the triangles. The optimal value of $osc_{\text{para}(2)}$ is zero and of osc_{exp} is one. The larger these values are, the stronger are the oscillations in the parabolic and exponential layer, respectively. For evaluating the extent of the global smearing, the value

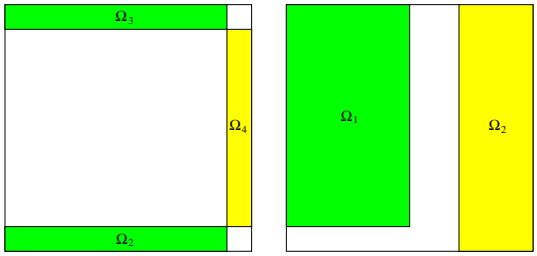


Fig. 4. The subdomains $\Omega_1, \ldots, \Omega_4$ from Example 1 (left) and Ω_1, Ω_2 from Example 2 (right).

$$smear := \left(\sum_{\text{interior nodes } (x,y)} (\min\{0, u_h(x,y) - x\})^2 \right)^{1/2} \qquad (45)$$

is computed. The rating of the results is given in Table 3.

Again, intermediate scores will be given if values are close to the interval with the next higher score. Since there are two criteria for the oscillations in the parabolic layers, the score of each is half of the score of osc_{exp}.

For MH85 and HMM86, we were not able to solve the nonlinear problems. It is remarkable that only the edge stabilization schemes BH04, BE05_1 and BE05_2 and the method J90 were able to compute solutions almost without spurious oscillations at the exponential layer, see Table 4 and Fig. 5. The results at the exponential layer obtained with the most other methods are similar to the result of KLR02_3 in the middle of Fig. 5. However, the edge stabilization schemes lead to a larger smearing of layers, see Fig. 6 for the parabolic layer at $y = 0$. The method J90 produces much larger spurious oscillations in the parabolic layers than BH04, BE05_1 and BE05_2. Altogether, BH04, BE05_1 and BE05_2 worked best on the unstructured Grid 3 since these methods suppressed the spurious oscillations at the exponential layer well and they worked also relatively well in the parabolic layers. A second group of methods, GdC88, dCG91, KLR02_3 and BE02_2, computed good results outside the exponential layer.

Example 2 (*Solution with interior layer and exponential boundary layer*). The convection–diffusion equation (1) is considered in $\Omega = (0,1)^2$ with the data $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))^{\mathrm{T}}$, $f = 0$ and

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

The solution, see Fig. 1, possesses an interior layer in the direction of the convection starting at (0,0.7). On the

Table 3
Definition of scores for the results in Table 4

$osc_{\text{para(1)}} \in$	Score	$osc_{\text{para(2)}} \in$	Score	$osc_{\text{exp}} \in$	Score	$smear \in$	Score
[0,1e−3)	2	[0,1e−1)	2	[1,1.25)	4	[0,1.25)	2
[1e−3, 1e−2)	1	[1e−1,3e−1)	1	[1.25,2)	2	[1.25,2)	1
[1e−2, 1e−1)	0	[3e−1,1)	0	[2,3)	0	[2,3)	0
[1e−1,1)	−2	[1,10)	−2	[3,5)	−4	[3,5)	−2

Table 4
Example 1, Grid 3, the measures for evaluating the oscillations and the smearing are defined in (42)–(45), the parameters in the SOLD methods are the same as in Table 2

Name	$osc_{\text{para(1)}}$	Score	$osc_{\text{para(2)}}$	Score	osc_{exp}	Score	*smear*	Score	Total
SUPG	1.545e−1	−2	7.883e+0	−2	4.972	−4	8.550e−1	2	−6
MH85	No conv.		–		–		–		
HMM86	No conv.		–		–		–		
TP86_1	9.225e−2	0	3.612e+0	−2	2.771	0	9.164e−1	2	0
TP86_2	1.291e−1	−2	6.369e+0	−2	2.968	0	9.125e−1	2	−2
GdC88	7.103e−3	1	2.679e−1	1	2.702	0	1.711e+0	1	3
dCG91	7.048e−3	1	2.746e−1	1	2.675	0	1.846e+0	1	3
dCA03	1.191e−2	0.5	5.550e−1	0	2.695	0	1.720e+0	1	1.5
AS97	8.961e−3	1	4.336e−1	0	2.876	0	1.849e+0	1	2
KLR02_1	1.313e−1	−2	6.786e+0	−2	4.563	−4	9.508e−1	2	−6
J90	3.245e−2	0	1.205e+0	−1	1.156	4	2.833e+0	0	3
JSW87	6.167e−4	2	2.002e−2	2	2.250	0	4.247e+0	−2	2
C93	2.416e−2	0	8.591e−1	0	2.823	0	1.131e+0	2	2
KLR02_2	9.862e−2	0	4.741e+0	−2	2.420	0	1.047e+0	2	0
KLR02_3	2.829e−3	1.5	1.112e−1	1.5	2.823	0	1.549e+0	1	4
BE02_1	5.336e−3	1	2.189e−1	1	3.224	−2	2.177e+0	0	0
BE02_2	2.604e−3	1.5	1.030e−1	1.5	2.320	0	1.826e+0	1	4
BE02_3	7.142e−3	1	3.074e−1	0.5	3.285	−2	1.858e+0	1	0.5
BH04	8.941e−3	1	3.549e−1	0.5	1.086	4	2.309e+0	0	5.5
BE05_1	5.431e−3	1	1.998e−1	1	1.075	4	2.211e+0	0	6
BE05_2	8.367e−3	1	3.417e−1	0.5	1.080	4	2.013e+0	0.5	6

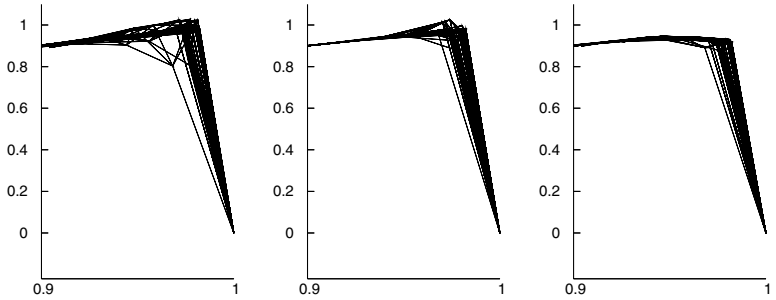


Fig. 5. Example 1, the exponential boundary layer computed with SUPG, KLR02_3 and BH04 (left to right) on Grid 3, $(x,y) \in [0.9,1] \times [0.1,0.9]$.

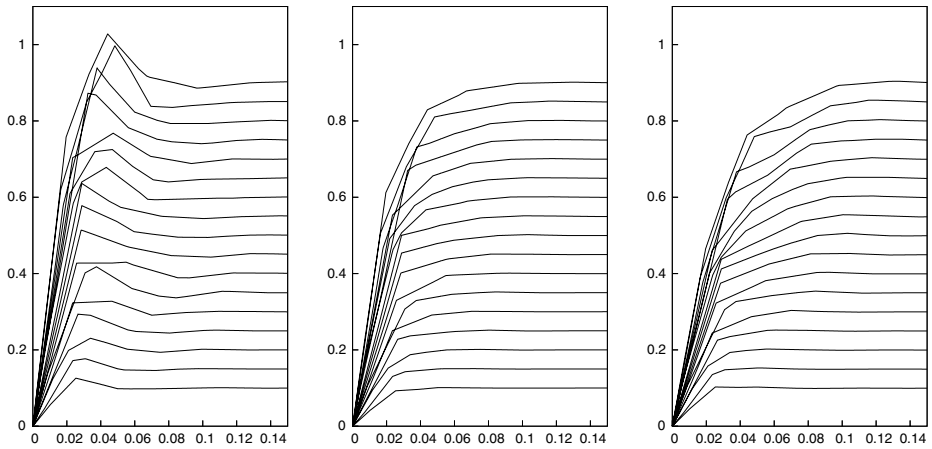


Fig. 6. Example 1, the parabolic boundary layer at $y = 0$ computed with SUPG, KLR02_3 and BH04 (left to right) on Grid 3, cuts of the solution at $x \in \{0.1, 0.15, 0.2, \ldots, 0.9\}$.

boundary $x = 1$ and on the right part of the boundary $y = 0$, exponential layers are developed. This example has been used, e.g., in [36].

The computations were performed on Grid 1, Grid 2 and Grid 3, see Fig. 2. For the regular triangular Grid 1 and Grid 2, the convection is skew to the grid lines. The grid size in the computations was chosen to be 1/64 (length of the legs of the triangles) such that the Péclet number is $Pe_K = 781{,}250$ and the number of degrees of freedom 4225. The features of Grid 3 have been mentioned already in Example 1. Since the right-hand side of (1) vanishes, the following methods are the identical ones: HMM86 and GdC88; dCG91 and AS97; C93 and KLR02_3. The choice of the SUPG parameter τ can be regarded as optimal on Grid 1 since the SUPG solution is nodally exact outside the inner layer and the boundary layer at $x = 1$. Denoting

$$\Omega_1 = \{(x,y) \in \Omega; x \leqslant 0.5, y \geqslant 0.1\},$$
$$\Omega_2 = \{(x,y) \in \Omega; x \geqslant 0.7\},$$

see Fig. 4, the following quantities are considered for assessing the computational results:

$$osc_{\text{int}} := \left(\sum_{(x,y)\in\Omega_1} (\min\{0, u_h(x,y)\})^2 + (\max\{0, u_h(x,y) - 1\})^2 \right)^{1/2}, \tag{46}$$

$$osc_{\text{exp}} := \left(\sum_{(x,y)\in\Omega_2} (\max\{0, u_h(x,y) - 1\})^2 \right)^{1/2}, \tag{47}$$

$$smear_{\text{int}} := x_2 - x_1, \tag{48}$$

$$smear_{\text{exp}} := \left(\sum_{(x,y)\in\Omega_2} (\min\{0, u_h(x,y) - 1\})^2 \right)^{1/2}, \tag{49}$$

where x_1 is the x-coordinate of the first point on the cut line $(x, 0.25)$ with $u_h(x_1, 0.25) \geqslant 0.1$ and x_2 is the x-coordinate of the first point with $u_h(x_2, 0.25) \geqslant 0.9$. Thus, (48) gives a measure for the thickness of the interior layer. The evaluation of x_1 and x_2 used a grid with mesh width 10^{-5} on the cut line. The summations are performed over the nodes (x,y) of the meshes.

Results of the computations on Grid 1 are presented in Table 7. The scoring of the results is given in Table 5. Again, intermediate scores are used.

The method MH85 gives an almost perfect result. Only the interior layer is smeared somewhat. Quite good results are obtained also with dCG91, AS97 and BE02_2. We observed for all SOLD methods that there are no spurious oscillations in the exponential layer at $y = 0$ on Grid 1, see also Fig. 7.

Comparing the results on Grid 1 on the one hand and Grid 2 and Grid 3 on the other hand, one finds that the results on Grid 2 and Grid 3 are considerably worse, see Tables 7–9. Because of this, the conditions for rating the results on Grid 2 and Grid 3 are relaxed somewhat, see Table 6. To obtain a better classification of the methods, intermediate values are used as in the other tests.

The results for Grid 2 are presented in Table 8. The only method which worked still very good was MH85. Only the smearing of the interior layer became somewhat larger in comparison to Grid 1. None of the other SOLD schemes produced a satisfactory solution with respect to all criteria of evaluation. It is remarkable that methods which worked well on Grid 1 completely failed on Grid 2, see Fig. 7 for dCG91 and AS97. Two other results are presented in Fig. 8. It can be seen that the solution computed with HMM86, GdC88 has a big oscillation at the starting point of the interior layer and another one in a vicinity of the corner (1,0) of Ω. The smearing of the layers which led to bad scores for BE05_2 is clearly visible in the right picture of Fig. 8.

A reason for the bad results obtained with the SOLD methods on Grid 2 can be found, in our opinion, already in the underlying SUPG stabilization. Since the SUPG method gives on Grid 2 considerably worse results than on Grid 1, there is not sufficient diffusion introduced in the streamline direction. However, the SOLD methods introduce additional diffusion above all orthogonally to the streamlines and rely upon the assumption that the SUPG method has done a good job in the streamline direction. If this is not the case, the SOLD methods give rather poor results as this example shows.

The results on the unstructured Grid 3, Table 9, show a similar tendency like the results on Grid 2. Again, only MH85 produced a satisfactory solution. All other SOLD schemes are on the one hand clearly worse than MH85 but on the other hand, most of them improved the SUPG solution considerably. We think that the reason for the SUPG-based SOLD methods being far away from a perfect solution is the same as given for Grid 2.

6.1. Summary of the numerical studies

The numerical tests were performed in a two-dimensional domain using the conforming P_1 finite element. Under these conditions, the upwind method MH85 was always the best method if the nonlinear iterations converged. Among the other SOLD methods, no one could be preferred in all cases. The methods dCG91 and BE02_2 were often among the best ones. However, even the best other SOLD methods gave sometimes rather unsatisfactory results. There are also some

Table 5
Definition of scores for the results in Table 7

$osc_{int} \in$	Score	$osc_{exp} \in$	Score	$smear_{int} \in$	Score	$smear_{exp} \in$	Score
[0,1e−4)	4	[0,1e−5)	4	[0,4e−2)	2	[0,1e−4)	2
[1e−4, 1e−2)	2	[1e−5, 1e−3)	2	[4e−2, 6e−2)	1	[1e−4, 1e−2)	1
[1e−2, 1e−1)	0	[1e−3, 1e−1)	0	[6e−2, 8e−2)	0	[1e−2, 5e−1)	0
[1e−1,1)	−4	[1e−1, 10)	−4	[8e−2, 1)	−2	[5e−1, 10)	−2

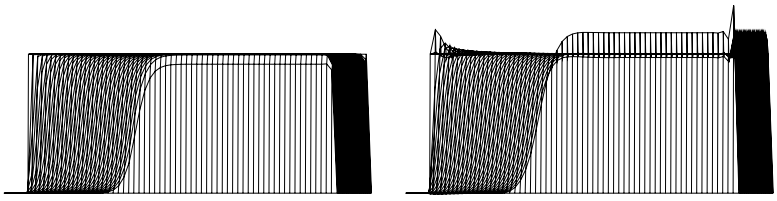


Fig. 7. Example 2, solutions obtained with dCG91 (AS97); left: on Grid 1, right: on Grid 2.

Table 6
Definition of scores for the results in Tables 8 and 9

$osc_{int} \in$	Score	$osc_{exp} \in$	Score	$smear_{int} \in$	Score	$smear_{exp} \in$	Score
[0,1e−3)	4	[0,1e−3)	4	[0,5e−2)	2	[0,1e−4)	2
[1e−3, 1e−2)	2	[1e−3, 2.5e−1)	2	[5e−2, 8e−2)	1	[1e−4, 1e−2)	1
[1e−2, 1e−1)	0	[2.5e−1, 1)	0	[8e−2, 1.1e−1)	0	[1e−2, 5e−1)	0
[1e−1,1)	−4	[1, 10)	−4	[1.1e−1, 1)	−2	[5e−1, 10)	−2

Table 7
Example 2, Grid 1 from Fig. 2, the measures for evaluating the oscillations and the smearing are defined in (46)–(49), the parameters in the SOLD methods are the same as in Table 2

Name	osc_{int}	Score	osc_{exp}	Score	$smear_{int}$	Score	$smear_{exp}$	Score	Total
SUPG	5.891e−1	−4	2.124e+0	−4	3.747e−2	2	5.666e−1	−1	−7
MH85	6.081e−13	4	0	4	5.792e−2	1	1.083e−5	2	11
HMM86, GdC88	1.185e−1	−2	3.010e−2	0	5.927e−2	1	2.921e−3	1	0
TP86_1	2.038e−1	−4	2.581e−6	4	4.020e−2	1.5	5.445e−1	−1	0.5
TP86_2	4.700e−1	−4	5.972e−2	0	3.852e−2	2	4.768e−1	0	−2
dCG91, AS97	1.248e−5	4	1.482e−10	4	7.090e−2	0	6.479e−1	−1	7
dCA03	1.299e−1	−2	3.019e−2	0	6.074e−2	0.5	3.220e−3	1	−0.5
KLR02_1	5.256e−1	−4	1.589e+0	−4	3.852e−2	2	4.118e−1	0	−6
J90	8.798e−2	0	4.157e−2	0	5.714e−2	1	3.058e+0	−2	−1
JSW87	5.440e−11	4	1.007e−4	2	1.473e−1	−2	2.656e−1	0	4
C93, KLR02_3	4.278e−3	2	1.959e−5	3	6.677e−2	0	9.042e−1	−2	3
KLR02_2	2.990e−1	−4	6.240e−1	−4	4.247e−2	1	2.292e−1	0	−7
BE02_1	1.083e−2	1	9.488e−4	2	7.527e−2	0	2.274e+0	−2	1
BE02_2	2.470e−8	4	2.546e−5	3	7.132e−2	0	6.723e−1	−1	6
BE02_3	1.558e−2	1	1.239e−3	1	6.795e−2	0	2.444e+0	−2	0
BH04	1.754e−2	1	5.063e−1	−4	7.106e−2	0	3.793e−1	0	−3
BE05_1	4.906e−3	2	1.904e+0	−4	9.685e−2	−2	4.520e−1	0	−4
BE05_2	4.580e−3	2	1.648e−4	2	7.930e−2	0	3.867e+0	−2	2

Table 8
Example 2, Grid 2 from Fig. 2, the measures for evaluating the oscillations and the smearing are defined in (46)–(49), the parameters in the SOLD methods are the same as in Table 2

Name	osc_{int}	Score	osc_{exp}	Score	$smear_{int}$	Score	$smear_{exp}$	Score	Total
SUPG	6.925e−1	−4	3.847e+0	−4	6.206e−2	1	1.698e+0	−2	−9
MH85	0	4	0	4	1.024e−1	0	1.161e−5	2	10
HMM86, GdC88	2.176e−1	−3	1.279e−1	2	1.037e−1	0	2.480e−3	1	0
TP86_1	2.719e−1	−3	6.713e−1	0	7.424e−2	1	4.586e−2	0	−2
TP86_2	5.509e−1	−4	5.489e−1	0	6.498e−2	1	1.952e−1	0	−3
dCG91, AS97	2.971e−1	−3	1.406e+0	−4	8.544e−2	0	2.114e−1	0	−7
dCA03	2.204e−1	−3	1.279e−1	2	1.060e−1	0	2.527e−3	1	0
KLR02_1	6.629e−1	−4	2.681e+0	−4	6.309e−2	1	1.080e+0	−2	−9
J90	2.939e−1	−3	5.950e−2	2	7.978e−2	1	2.681e+0	−2	−2
JSW87	2.444e−1	−3	2.133e+0	−4	1.117e−1	−1	5.005e−1	0	−8
C93, KLR02_3	1.386e−1	−2	3.606e−1	0	9.750e−2	0	3.126e−2	0	−2
KLR02_2	5.125e−1	−4	1.773e+0	−4	6.671e−2	1	5.941e−1	−1	−8
BE02_1	1.496e−1	−2	4.306e−1	0	1.034e−1	0	3.651e−1	0	−2
BE02_2	2.214e−1	−3	1.396e+0	−4	8.634e−2	0	2.102e−1	0	−7
BE02_3	1.453e−1	−2	3.839e−1	0	9.682e−2	0	5.169e−1	−0.5	−2.5
BH04	9.224e−2	0	1.548e+0	−4	9.966e−2	0	1.408e−1	0	−4
BE05_1	6.153e−3	2	3.514e+0	−4	1.528e−1	−2	1.402e+0	−2	−6
BE05_2	6.470e−3	2	2.163e−3	3	1.435e−1	−2	3.411e+0	−2	1

methods which never produced good results, e.g., TP86_1 and TP86_2 introduce in general not enough artificial diffusion to damp the oscillations sufficiently or JSW87 and J90 are very diffusive and smear the layers considerably. Altogether, there are still many open questions to be answered which will be started in the second part of this paper.

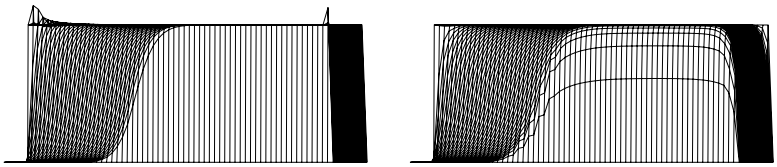


Fig. 8. Example 2, solutions obtained on Grid 2 with HMM86, GdC88 (left) and BE05_2 (right).

Table 9
Example 2, Grid 3 from Fig. 2, the measures for evaluating the oscillations and the smearing are defined in (46)–(49), the parameters in the SOLD methods are the same as in Table 2

Name	osc_{int}	Score	osc_{exp}	Score	$smear_{int}$	Score	$smear_{exp}$	Score	Total
SUPG	5.933e−1	−4	1.526e+0	−4	5.520e−2	1.5	4.070e−1	0	−6.5
MH85	4.940e−15	4	1.785e−14	4	9.717e−2	0	5.302e−2	0.5	8.5
HMM86, GdC88	1.127e−1	−2	1.961e−1	2	9.535e−2	0	1.944e−1	0	0
TP86_1	2.066e−1	−3	1.421e−1	2	6.364e−2	1	4.597e−1	0	0
TP86_2	4.295e−1	−4	1.830e−1	2	5.890e−2	1.5	4.118e−1	0	−0.5
dCG91, AS97	8.229e−2	0	1.282e−1	2	9.701e−2	0	5.998e−1	−1	1
dCA03	1.243e−1	−2	1.993e−1	2	9.553e−2	0	1.955e−1	0	0
KLR02_1	5.199e−1	−4	1.103e+0	−1	5.752e−2	1.5	2.851e−1	0	−3.5
J90	1.412e−1	−2	5.814e−2	2	8.310e−2	0.5	2.390e+0	−2	−1.5
JSW87	3.188e−3	2	1.035e−1	2	1.865e−1	−2	5.364e−1	−1	1
C93, KLR02_3	7.758e−2	0	5.422e−2	3	9.081e−2	0	7.956e−1	−2	1
KLR02_2	3.759e−1	−4	4.975e−1	0	6.190e−2	1	2.167e−1	0	−3
BE02_1	1.857e−2	1	1.877e−2	3	1.086e−1	0	1.703e+0	−2	2
BE02_2	6.342e−2	0	8.718e−2	3	9.803e−2	0	5.964e−1	−1	2
BE02_3	2.276e−2	1	1.166e−2	3	9.902e−2	0	1.818e+0	−2	2
BH04	3.565e−2	0	5.917e−1	0	9.226e−2	0	2.238e−1	0	0
BE05_1	9.236e−3	2	8.335e−1	0	1.264e−1	−2	2.333e−1	0	0
BE05_2	6.425e−2	0	2.212e−2	3	9.056e−2	0	1.036e+0	−2	1

7. Conclusions and outlook

A characteristic feature of numerical solutions of scalar convection-dominated convection–diffusion equations computed with the popular SUPG stabilization is the presence of quite large spurious oscillations at layers. The main goal of SOLD methods consists in suppressing these oscillations without an excessive smearing of the layers. The present paper gave a review of the state of the art of SOLD methods. Most of these methods can be classified into methods adding isotropic diffusion, methods adding diffusion orthogonally to the streamlines and into edge stabilization methods. Some numerical studies gave a first impression of the behavior of the SOLD methods.

Comprehensive numerical studies which will explore the limits of the capabilities of the available SOLD methods will be the subject of the second part of the paper.

Acknowledgments

The research of Petr Knobloch is a part of the project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Academy of Sciences of the Czech Republic under the grant No. IAA100190505.

References

[1] J.E. Akin, T.E. Tezduyar, Calculation of the advective limit of the SUPG stabilization parameter for linear and higher-order elements, Comput. Methods Appl. Mech. Engrg. 193 (2004) 1909–1922.
[2] J.E. Akin, T.E. Tezduyar, M. Ungor, S. Mittal, Stabilization parameters and Smagorinski turbulence model, J. Appl. Mech. 70 (2003) 2–9.
[3] R.C. Almeida, R.S. Silva, A stable Petrov–Galerkin method for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 140 (1997) 291–304.
[4] K. Baba, M. Tabata, On a conservative upwind finite element scheme for convective diffusion equations, RAIRO, Anal. Numer. 15 (1981) 3–25.
[5] Y. Bazilevs, T.J.R. Hughes, Weak imposition of Dirichlet boundary conditions in Fluid Mechanics, ICES Report 05-25, University of Texas at Austin, 2005.
[6] F. Brezzi, L.P. Franca, A. Russo, Further considerations on residual-free bubbles for advective–diffusive equations, Comput. Methods Appl. Mech. Engrg. 166 (1998) 25–33.
[7] F. Brezzi, D. Marini, E. Süli, Residual-free bubbles for advection–diffusion problems: The general error analysis, Numer. Math. 85 (2000) 31–47.
[8] F. Brezzi, A. Russo, Choosing bubbles for advection–diffusion problems, Math. Models Methods Appl. Sci. 4 (1994) 571–587.
[9] A.N. Brooks, 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.
[10] E. Burman, A unified analysis for conforming and nonconforming stabilized finite element methods using interior penalty, SIAM J. Numer. Anal. 43 (2005) 2012–2033.
[11] E. Burman, A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation, Comput. Methods Appl. Mech. Engrg. 191 (2002) 3833–3855.
[12] E. Burman, A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence, Math. Comput. 74 (2005) 1637–1652.
[13] E. Burman, P. Hansbo, Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems, Comput. Methods Appl. Mech. Engrg. 193 (2004) 1437–1453.
[14] E.G.D. do Carmo, G.B. Alvarez, A new stabilized finite element formulation for scalar convection–diffusion problems: the streamline and approximate upwind/Petrov–Galerkin method, Comput. Methods Appl. Mech. Engrg. 192 (2003) 3379–3396.
[15] E.G.D. do Carmo, 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.
[16] E.G.D. do Carmo, A.C. Galeão, Feedback Petrov–Galerkin methods for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 88 (1991) 1–16.

[17] I. Christie, D.F. Griffiths, A.R. Mitchell, O.C. Zienkiewicz, Finite element methods for second order differential equations with significant first derivatives, Int. J. Numer. Methods Engrg. 10 (1976) 1389–1396.
[18] P.G. Ciarlet, Basic error estimates for elliptic problems, in: P.G. Ciarlet, J.L. Lions (Eds.), Handbook of Numerical Analysis, Finite Element Methods (pt. 1), vol. 2, North-Holland, Amsterdam, 1991, pp. 17–351.
[19] 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.
[20] R. Codina, E. Oñate, M. Cervera, The intrinsic time for the streamline upwind/Petrov–Galerkin formulation using quadratic elements, Comput. Methods Appl. Mech. Engrg. 94 (1992) 239–262.
[21] R. Codina, O. Soto, Finite element implementation of two-equation and algebraic stress turbulence models for steady incompressible flows, Int. J. Numer. Meth. Fluids 30 (1999) 309–333.
[22] V. Dolejší, Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes, Comput. Vis. Sci. 1 (1998) 165–178.
[23] J. Douglas jun, J. Wang, An absolutely stabilized finite element method for the Stokes problem, Math. Comput. 52 (1989) 495–508.
[24] H.C. Elman, A. Ramage, An analysis of smoothing effects of upwinding strategies for the convection–diffusion equation, SIAM J. Numer. Anal. 40 (2002) 254–281.
[25] K. Eriksson, C. Johnson, Adaptive streamline diffusion finite element methods for stationary convection–diffusion problems, Math. Comput. 60 (1993) 167–188.
[26] L.P. Franca, S.L. Frey, T.J.R. Hughes, Stabilized finite element methods. I.: Application to the advective–diffusive model, Comput. Methods Appl. Mech. Engrg. 95 (1992) 253–276.
[27] B. Fischer, A. Ramage, D.J. Silvester, A.J. Wathen, On parameter choice and iterative convergence for stabilised discretisations of advection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 179 (1999) 179–195.
[28] A.C. Galeão, R.C. Almeida, S.M.C. Malta, A.F.D. Loula, Finite element analysis of convection dominated reaction–diffusion problems, Appl. Numer. Math. 48 (2004) 205–222.
[29] A.C. Galeão, E.G.D. do Carmo, A consistent approximate upwind Petrov–Galerkin method for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 68 (1988) 83–95.
[30] J.L. Guermond, A finite element technique for solving first-order pdes in L^p, SIAM J. Numer. Anal. 42 (2004).
[31] I. Harari, L.P. Franca, S.P. Oliveira, Streamline design of stability parameters for advection–diffusion problems, J. Comput. Phys. 171 (2001) 115–131.
[32] J.C. Heinrich, P.S. Huyakorn, O.C. Zienkiewicz, A.R. Mitchell, An 'upwind' finite element scheme for two-dimensional convective transport equation, Int. J. Numer. Methods Engrg. 11 (1977) 131–143.
[33] J.C. Heinrich, O.C. Zienkiewicz, Quadratic finite element schemes for two-dimensional convective-transport problems, Int. J. Numer. Methods Engrg. 11 (1977) 1831–1844.
[34] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995) 387–401.
[35] T.J.R. Hughes, L.P. Franca, 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.
[36] T.J.R. Hughes, M. Mallet, A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Methods Appl. Mech. Engrg. 54 (1986) 341–355.
[37] 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.
[38] V. John, P. Knobloch, On discontinuity-capturing methods for convection–diffusion equations, in: A. Bermúdez de Castro, D. Gómez, P. Quintela, P. Salgado (Eds.), Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2005, Springer, Berlin, 2006, pp. 336–344.
[39] V. John, P. Knobloch, A computational comparison of methods diminishing spurious oscillations in finite element solutions of convection–diffusion equations, in: Proceedings of the Conference Programs and Algorithms of Numerical Mathematics, vol. 13, Prague, May 28–31, 2006 (to appear).
[40] V. John, G. Matthies, MooNMD – a program package based on mapped finite element methods, Comput. Visual. Sci. 6 (2004) 163–170.
[41] C. Johnson, Adaptive finite element methods for diffusion and convection problems, Comput. Methods Appl. Mech. Engrg. 82 (1990) 301–322.
[42] C. Johnson, A new approach to algorithms for convection problems which are based on exact transport + projection, Comput. Methods Appl. Mech. Engrg. 100 (1992) 45–62.
[43] C. Johnson, A.H. Schatz, L.B. Wahlbin, Crosswind smear and pointwise errors in streamline diffusion finite element methods, Math. Comput. 49 (1987) 25–38.
[44] C. Johnson, A. Szepessy, P. Hansbo, On the convergence of shock-capturing streamline diffusion finite element methods for hyperbolic conservation laws, Math. Comput. 54 (1990) 107–129.
[45] H. Kanayama, Discrete models for salinity distribution in a bay: conservation law and maximum principle, Theoretical Appl. Mech. 28 (1978) 559–579.
[46] P. Knobloch, Improvements of the Mizukami–Hughes method for convection–diffusion equations, Comput. Methods Appl. Mech. Engrg. 196 (2006) 579–594.
[47] T. Knopp, G. Lube, G. Rapin, Stabilized finite element methods with shock capturing for advection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 191 (2002) 2997–3013.
[48] W. Layton, B. Polman, Oscillation absorption finite element methods for convection–diffusion problems, SIAM J. Sci. Comput. 17 (1996) 1328–1346.
[49] G. Lube, An asymptotically fitted finite element method for convection dominated convection–diffusion–reaction problems, Z. Angew. Math. Mech. 72 (1992) 189–200.
[50] A. Mizukami, An implementation of the streamline-upwind/Petrov–Galerkin method for linear triangular elements, Comput. Methods Appl. Mech. Engrg. 49 (1985) 357–364.
[51] A. Mizukami, 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.
[52] U. Nävert, A finite element method for convection–diffusion problems, Ph.D. Thesis, Chalmers University of Technology Göteborg, 1982.
[53] K. Niijima, Pointwise error estimates for a streamline diffusion finite element scheme, Numer. Math. 56 (1990) 707–719.
[54] E. Oñate, Derivation of stabilized equations for numerical solution of advective–diffusive transport and fluid flow problems, Comput. Methods Appl. Mech. Engrg. 151 (1998) 233–265.
[55] A. Ramage, A note on parameter choice and iterative convergence for stabilised discretisations of advection–diffusion problems in three dimensions, Mathematics Research Report 32/98, University of Strathclyde, July 1998.
[56] A. Ramage, A multigrid preconditioner for stabilised discretisations of advection–diffusion problems, J. Comput. Appl. Math. 110 (1999) 187–203.
[57] J.G. Rice, R.J. Schnipke, A monotone streamline upwind finite element method for convection-dominated flows, Comput. Methods Appl. Mech. Engrg. 48 (1985) 313–327.
[58] H.-G. Roos, M. Stynes, L. Tobiska, Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion and Flow problems, Springer, Berlin, 1996.

[59] P.A.B. de Sampaio, A.L.G.A. Coutinho, A natural derivation of discontinuity capturing operator for convection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 190 (2001) 6291–6308.
[60] Y.-T. Shih, H.C. Elman, Modified streamline diffusion schemes for convection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 174 (1999) 137–151.
[61] M. Stynes, L. Tobiska, Necessary L^2-uniform convergence conditions for difference schemes for two-dimensional convection–diffusion problems, Comput. Math. Appl. 29 (1995) 45–53.
[62] M. Tabata, A finite element approximation corresponding to the upwind finite differencing, Mem. Numer. Math. 4 (1977) 47–63.
[63] T.E. Tezduyar, Y. Osawa, Finite element stabilization parameters computed from element matrices and vectors, Comput. Methods Appl. Mech. Engrg. 190 (2000) 411–430.
[64] T.E. Tezduyar, Y.J. Park, Discontinuity-capturing finite element formulations for nonlinear convection–diffusion–reaction equations, Comput. Methods Appl. Mech. Engrg. 59 (1986) 307–325.
[65] G. Zhou, How accurate is the streamline diffusion finite element method? Math. Comput. 66 (1997) 31–44.
[66] G. Zhou, R. Rannacher, Pointwise superconvergence of the streamline diffusion finite-element method, Numer. Methods Partial Differ. Equations 12 (1996) 123–145.

Available online at www.sciencedirect.com
ScienceDirect
Comput. Methods Appl. Mech. Engrg. 197 (2008) 1997–2014
Computer methods in applied mechanics and engineering
www.elsevier.com/locate/cma

On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements

Volker John [a], Petr Knobloch [b,*]

[a] Universität des Saarlandes, Fachbereich 6.1 – Mathematik, Postfach 15 11 50, 66041 Saarbrücken, Germany
[b] Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 186 75 Praha 8, Czech Republic

Received 27 September 2007; received in revised form 20 December 2007; accepted 27 December 2007
Available online 15 January 2008

Abstract

An unwelcome feature of the popular streamline upwind/Petrov–Galerkin (SUPG) stabilization of convection-dominated convection–diffusion equations is the presence of spurious oscillations at layers. A review and a comparison of the most methods which have been proposed to remove or, at least, to diminish these oscillations without leading to excessive smearing of the layers are given in Part I, [V. John, P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review, Comput. Methods Appl. Mech. Engrg. 196 (2007) 2197–2215]. In the present paper, the most promising of these SOLD methods are investigated in more detail for P_1 and Q_1 finite elements. In particular, the dependence of the results on the mesh, the data of the problems and parameters of the methods are studied analytically and numerically. Furthermore, the numerical solution of the nonlinear discrete problems is discussed and the capability of adaptively refined grids for reducing spurious oscillations is examined. Our conclusion is that, also for simple problems, any of the SOLD methods generally provides solutions with non-negligible spurious oscillations.

Keywords: Convection–diffusion equations; Streamline upwind/Petrov–Galerkin (SUPG) method; Spurious oscillations at layers diminishing (SOLD) methods

1. Introduction

This paper is a continuation of [26], in the following cited as Part I, which was devoted to a review and a comparison of finite element techniques developed to diminish spurious oscillations in discrete solutions of convection-dominated problems. Like in Part I, we consider the steady scalar convection–diffusion equation

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega, \qquad u = u_b \quad \text{on } \partial\Omega. \tag{1}$$

We assume that Ω is a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\boldsymbol{b} \in W^{1,\infty}(\Omega)^2$ is a given convective field, $f \in L^2(\Omega)$ is an outer source of u, and $u_b \in H^{1/2}(\partial\Omega)$ represents the Dirichlet boundary condition. In our numerical tests we shall also consider less regular functions u_b.

A popular finite element discretization technique for (1) is the streamline upwind/Petrov–Galerkin (SUPG) method which is frequently used because of its stability properties and higher-order accuracy. Since, in the convection-dominated regime, the SUPG solutions typically contain oscillations in layer regions, various stabilizing terms have been proposed to be added to the SUPG discretization in order to obtain discrete solutions in which the local oscillations are suppressed. In Part I, we called such techniques *spurious oscillations at layers diminishing (SOLD)* methods.

Part I presented a review of most SOLD methods published in the literature, discussed their derivation, proposed some alternative choices of parameters in the methods and categorized them. Some numerical studies gave a

* Corresponding author. Tel.: +420 737314937; fax: +420 224811036.
E-mail addresses: john@math.uni-sb.de (V. John), knobloch@karlin.mff.cuni.cz (P. Knobloch).

0045-7825/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.cma.2007.12.019

first impression of the behavior of the SOLD methods. These numerical tests were performed in a two-dimensional domain using the conforming P_1 finite element and it was observed that there are large differences between the SOLD methods. In some cases, the SOLD methods were able to significantly improve the SUPG solution and to provide a discrete solution with negligible spurious oscillations and without an excessive smearing of layers. However, it was not possible to identify a method which could be preferred in all the test cases. There are some methods which never produced good results since they either do not suppress the oscillations sufficiently or they are very diffusive and smear the layers considerably.

The aim of the present paper is to perform deeper investigations of those SOLD methods which gave acceptable results in Part I. We shall formulate the SOLD methods in the two-dimensional case and for conforming linear and bilinear finite elements. Formulations valid also in the three-dimensional case and for more general finite element spaces can be found in Part I. We do not consider the Mizukami–Hughes method [35,33] investigated in Part I since its applicability is rather limited. We shall investigate how strongly the methods depend on the computational mesh and the data of the problem. For methods containing parameters, we shall seek their optimal values and study the dependence of the results on the parameters. Since most of the SOLD methods are nonlinear, we shall also address algorithms for computing the discrete solution. Finally, the question will be studied whether adaptively refined grids help to suppress the spurious oscillations in SUPG solutions.

Our investigations will be performed on academic test examples whose solutions possess characteristic features of solutions of convection–diffusion equations. These academic problems allow to study the SOLD methods analytically, at least in the limit $\varepsilon \to 0+$. The analysis enables us to identify clearly those methods which can be expected to suppress the spurious oscillations and to study the dependence of the results on parameters in some of the methods.

The analysis presented in this paper will include the consideration of moderately anisotropic grids. Using such grids might not be reasonable for the considered examples since these grids are not adapted to the layers of the solution. Our motivation for looking at moderately anisotropic grids comes from applications. First, the meshing of complicated domains leads easily to anisotropic elements with moderate aspect ratio. Second, convection–diffusion equations are often just a part of a coupled system of equations, like in the $k-\varepsilon$ turbulence model [36] or in the simulation of precipitation processes [29]. For such problems, an adaptation of the grid is performed rather with respect to other equations in the system, for instance with respect to the Navier–Stokes equations in the mentioned examples. Thus, one has to face the situation that the grids might be not particularly well adapted with respect to the convection–diffusion equation but the SOLD methods still should provide satisfactory results.

The paper is organized in the following way. In the next section, we formulate the usual Galerkin discretization of (1) and introduce the SUPG method. In Section 3, the SOLD methods investigated in this paper are briefly reviewed. Then, in Section 4, we shall investigate the properties of the SOLD methods for three model problems. Section 5 is devoted to the computation of the discrete solution and, in Section 6, the usefulness of adaptively refined grids for the suppression of spurious oscillations is studied. Finally, Section 7 presents our conclusions.

Throughout the paper, we use the standard notations P_1, Q_1, $L^2(\Omega)$, $H^1(\Omega) = W^{1,2}(\Omega)$, etc. for the usual function spaces, see, e.g., Ciarlet [9]. The inner product in the space $L^2(\Omega)$ or $L^2(\Omega)^2$ will be denoted by $(\cdot,\cdot)$. For a vector $\boldsymbol{a} \in \mathbb{R}^2$, the symbol $|\boldsymbol{a}|$ stands for its Euclidean norm.

2. The Galerkin method and the SUPG method

To define a finite element discretization of (1), we introduce a triangulation $\mathscr{T}_h$ of the domain Ω consisting of a finite number of open elements K. We shall assume that all elements of $\mathscr{T}_h$ are either triangles or convex quadrilaterals. The discretization parameter h in the notation $\mathscr{T}_h$ is a positive real number satisfying $\operatorname{diam}(K) \leqslant h$ for any $K \in \mathscr{T}_h$. We assume that $\overline{\Omega} = \bigcup_{K\in\mathscr{T}_h} \overline{K}$ and that the closures of any two different elements of $\mathscr{T}_h$ are either disjoint or possess either a common vertex or a common edge.

We introduce the finite element space

$$V_h = \{v \in H_0^1(\Omega); v|_K \in R(K) \quad \forall K \in \mathscr{T}_h\},$$

where $R(K) = P_1(K)$ if K is a triangle and $R(K) = Q_1(K)$ if K is a rectangle. If K is a general convex quadrilateral, then $R(K)$ is defined by transforming the space $Q_1((0,1)^2)$ onto K by means of a bilinear one-to-one mapping, see, e.g., Ciarlet [9]. Finally, let $u_{bh} \in H^1(\Omega)$ be a function whose trace approximates the boundary condition u_b. Then the usual Galerkin finite element discretization of the convection–diffusion equation (1) reads:

Find $u_h \in H^1(\Omega)$ such that $u_h - u_{bh} \in V_h$ and

$$a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_h,$$

where

$$a(u,v) = \varepsilon(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v).$$

It is well known that this discretization 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) method developed by Brooks and Hughes [3]. To formulate this method, we define the residual

$$R_h(u) = -\varepsilon\Delta_h u + \boldsymbol{b}\cdot\nabla u - f,$$

where Δ_h is the Laplace operator defined elementwise, i.e., $(\Delta_h v)|_K = \Delta(v|_K)$ for any $K \in \mathscr{T}_h$ and any piecewise smooth function v. Then the SUPG method reads:

Find $u_h \in H^1(\Omega)$ such that $u_h - u_{bh} \in V_h$ and

$$a(u_h, v_h) + (R_h(u_h), \tau \boldsymbol{b} \cdot \nabla v_h) = (f, v_h) \quad \forall v_h \in V_h, \tag{2}$$

where $\tau \in L^\infty(\Omega)$ is a nonnegative stabilization parameter. The choice of τ may dramatically influence the accuracy of the discrete solution and therefore it has been a subject of an extensive research over the last three decades, see, e.g., the review in Part I. Unfortunately, a general optimal definition of τ is still not known. In our computations, we define τ, on any element $K \in \mathscr{T}_h$, by the formula

$$\tau|_K = \frac{h_K}{2|\boldsymbol{b}|}\left(\coth Pe_K - \frac{1}{Pe_K}\right) \quad \text{with} \quad Pe_K = \frac{|\boldsymbol{b}| h_K}{2\varepsilon}, \tag{3}$$

where h_K is the element diameter in the direction of the convection vector $\boldsymbol{b}$. We refer to Part I for various justifications of this formula and for a precise definition of h_K. If convection strongly dominates diffusion in Ω and hence the local Péclet numbers Pe_K are very large, the parameter τ is basically given by

$$\tau|_K = \frac{h_K}{2|\boldsymbol{b}|} \quad \forall K \in \mathscr{T}_h. \tag{4}$$

Note that, generally, the parameters h_K, Pe_K and $\tau|_K$ are functions of the points $\mathbf{x} \in K$.

An alternative to the SUPG method is the Galerkin/least-squares method introduced by Hughes et al. [21] or its modification proposed by Franca et al. [16]. A similar stabilization can also be obtained using the subgrid scale method of Hughes [20]. In addition, for transient problems, stabilization terms of the discussed type also result by applying the characteristic Galerkin method of Douglas and Russell [15] or the Taylor–Galerkin method of Donéa [14]. See also Codina [11] for a comparison of these methods. However, all these methods are identical to the SUPG method (up to the choice of the stabilization parameter) if problem (1) has constant coefficients and is discretized using linear triangular or bilinear rectangular finite elements. Since this will be the case in all the model problems discussed in this paper, we confine ourselves to the SUPG method in the following.

3. Spurious oscillations at layers diminishing methods

Because the SUPG method is not monotone, a discrete solution satisfying (2) usually still contains spurious oscillations. Although these oscillations are localized in narrow regions along sharp layers, they are often not negligible and they are not permissible in many applications. A possible remedy is to add a suitable artificial diffusion term to the SUPG method. In Part I, methods of this type are called *spurious oscillations at layers diminishing* (*SOLD*) methods. Here, we describe these methods only very briefly and refer to the review in Part I for details. To make similarities and differences between the methods better visible, we shall formulate the methods in a slightly different way than in Part I.

There are three basic classes of SOLD methods: methods adding isotropic artificial diffusion, methods adding crosswind artificial diffusion, and methods where the additional artificial diffusion stems from an edge stabilization. The amount of the artificial diffusion in these methods typically depends on the unknown discrete solution u_h. Thus, the resulting methods are nonlinear (although the original problem (1) is linear).

The methods of the first class add the isotropic artificial diffusion term

$$(\tilde{\varepsilon} \nabla u_h, \nabla v_h) \tag{5}$$

to the left-hand side of the SUPG discretization (2). The parameter $\tilde{\varepsilon}$ is nonnegative and usually depends on u_h. For the first time, a SOLD term which can be written in the form (5) was introduced by Hughes et al. [22]. Further approaches were proposed by Tezduyar and Park [38] and Galeão and do Carmo [17]. According to the criteria and tests in Part I (and according to further numerical experiments we have performed in [24,25]), one of the best choices of $\tilde{\varepsilon}$ in (5) is to set

$$\tilde{\varepsilon} = \max\left\{0, \frac{\tau|\boldsymbol{b}||R_h(u_h)|}{|\nabla u_h|} - \tau\frac{|R_h(u_h)|^2}{|\nabla u_h|^2}\right\}, \tag{6}$$

as proposed by do Carmo and Galeão [8], abbreviated with dCG91 in Part I. Here and in the following, we always assume that $\tilde{\varepsilon} = 0$ if the denominator of a formula defining $\tilde{\varepsilon}$ vanishes. Almeida and Silva [1] suggested to multiply the negative term in (6) by

$$\zeta_h = \max\left\{1, \frac{\boldsymbol{b} \cdot \nabla u_h}{R_h(u_h)}\right\},$$

which is method AS97 in Part I. However, in our tests, we often observed no significant differences to the results obtained with (6). Another $\tilde{\varepsilon}$, motivated by assumptions needed for theoretical investigations, can be found in Knopp et al. [34]. Further modifications of the above approaches were proposed by do Carmo and Galeão [8] and do Carmo and Alvarez [7], who introduced rather complicated definitions of $\tilde{\varepsilon}$ which should suppress the addition of the artificial diffusion in regions where the solution of (1) is smooth. The SOLD term (5) was also used by Johnson [30], who proposed to set

$$\tilde{\varepsilon}|_K = \max\{0, C[\mathrm{diam}(K)]^2|R_h(u_h)| - \varepsilon\} \quad \forall K \in \mathscr{T}_h, \tag{7}$$

where C is a nonnegative parameter (method J90 in Part I).

Johnson et al. [32] modified the SUPG discretization (2) by adding artificial diffusion in the crosswind direction only. This corresponds to the additional term

$$(\tilde{\varepsilon} \boldsymbol{b}^\perp \cdot \nabla u_h, \boldsymbol{b}^\perp \cdot \nabla v_h) \quad \text{with} \quad \boldsymbol{b}^\perp = \frac{(-b_2, b_1)}{|\boldsymbol{b}|} \tag{8}$$

on the left-hand side of (2). In [32], the parameter $\tilde{\varepsilon}$ was defined by

$$\tilde{\varepsilon}|_K = \max\{0, |\boldsymbol{b}| h_K^{3/2} - \varepsilon\} \quad \forall K \in \mathscr{T}_h \tag{9}$$

so that the resulting method (JSW87 in Part I) is linear but non-consistent and hence it is restricted to finite elements of first order of accuracy. Moreover, the numerical tests from Part I show that this method is very diffusive.

Codina [10] proposed to define $\tilde{\varepsilon}$ in (8), for any $K \in \mathscr{T}_h$, by

$$\tilde{\varepsilon}|_K = \max\left\{0, C\frac{\operatorname{diam}(K)|R_h(u_h)|}{2|\nabla u_h|} - \varepsilon\frac{|R_h(u_h)|}{|\boldsymbol{b}\cdot\nabla u_h|}\right\}, \tag{10}$$

where C is a suitable constant, and he recommended to set $C \approx 0.7$ for (bi)linear finite elements. This is method C93 in Part I. For $f \neq 0$, we observed that, in some cases, this choice of $\tilde{\varepsilon}$ does not lead to a reduction of the oscillations (see the discussion to Example 1 in the next section). Therefore, in Part I, we replaced (10) by

$$\tilde{\varepsilon}|_K = \max\left\{0, C\frac{\operatorname{diam}(K)|R_h(u_h)|}{2|\nabla u_h|} - \varepsilon\right\}, \tag{11}$$

called method KLR02_3 in Part I. Here, we shall also call this method *modified method of Codina*. If $f = 0$ and $\Delta_h u_h = 0$, it is equivalent to the original method (10). A modification of (10), leading to properties convenient for theoretical investigations, was proposed by Knopp et al. [34].

For triangulations consisting of weakly acute triangles, Burman and Ern [4] proposed to use (8) with $\tilde{\varepsilon}$ defined, on any $K \in \mathscr{T}_h$, by

$$\tilde{\varepsilon}|_K = \frac{\tau|\boldsymbol{b}||R_h(u_h)|}{|\nabla u_h|}\frac{|\boldsymbol{b}||\nabla u_h|}{|\boldsymbol{b}||\nabla u_h| + |R_h(u_h)|} \times \frac{|\boldsymbol{b}||\nabla u_h| + |R_h(u_h)| + \tan\alpha_K|\boldsymbol{b}||\boldsymbol{b}^\perp\cdot\nabla u_h|}{|R_h(u_h)| + \tan\alpha_K|\boldsymbol{b}||\boldsymbol{b}^\perp\cdot\nabla u_h|}. \tag{12}$$

The parameter α_K is equal to $\pi/2 - \beta_K$ where β_K is the largest angle of K. If $\beta_K = \pi/2$, it is recommended in [4] to set $\alpha_K = \pi/6$. To improve the convergence of the nonlinear iterations, we replaced in Part I $|R_h(u_h)|$ by $|R_h(u_h)|_{\text{reg}}$ with $|x|_{\text{reg}} \equiv x\tanh(x/2)$ as proposed already in [4]. The resulting method was called BE02_1.

In Part I, we also introduced a simplification of (12), called BE02_2, defined by

$$\tilde{\varepsilon} = \frac{\tau|\boldsymbol{b}||R_h(u_h)|}{|\nabla u_h|}\frac{|\boldsymbol{b}||\nabla u_h|}{|\boldsymbol{b}||\nabla u_h| + |R_h(u_h)|}, \tag{13}$$

which adds less artificial diffusion than (12). In (13), we do not apply any regularization of the absolute values. We call this method *modified method of Burman and Ern*. Based on the evaluation of the numerical studies in Part I and [24,25], in our opinion, this method and the modified method of Codina are the best methods among the methods adding crosswind artificial diffusion.

It is also possible to add both isotropic and crosswind artificial diffusion terms to the left-hand side of (2). Denoting the parameters in (5) and (8) by $\tilde{\varepsilon}^{\text{iso}}$ and $\tilde{\varepsilon}^{\text{cross}}$, respectively, Codina and Soto [12] proposed to set

$$\tilde{\varepsilon}^{\text{iso}} = \max\{0, \tilde{\varepsilon}^{\text{dc}} - \tau|\boldsymbol{b}|^2\}, \quad \tilde{\varepsilon}^{\text{cross}} = \tilde{\varepsilon}^{\text{dc}} - \tilde{\varepsilon}^{\text{iso}},$$

where $\tilde{\varepsilon}^{\text{dc}}$ is defined by a formula similar to (11). However, in the numerical tests we have performed up to now, we have not observed an advantage in using this approach instead of (8) with $\tilde{\varepsilon}$ given by (11).

There are some similarities between the definitions of $\tilde{\varepsilon}$ in (6), (7) and (10)–(13). Particularly, the presence of a term of the type $h|R_h(u_h)|/|\nabla u_h|$ seems to be important. Indeed, if convection is strongly dominant (and hence (4) approximately holds), we have in (6), (12) and (13)

$$\frac{\tau|\boldsymbol{b}||R_h(u_h)|}{|\nabla u_h|} \approx \frac{h_K|R_h(u_h)|}{2|\nabla u_h|}. \tag{14}$$

Remark 1. The recently published $YZ\beta$ scheme for scalar convection–diffusion equations [2], originally proposed by Tezduyar [37] for compressible flows, gives for $\beta = 1$ exactly the parameter (14) if, in contrast to [2], in the definition of the local element length the convection is used instead of the gradient of the solution. Using the latter replaces h_K by the element size orthogonal to the convection, see the discussion of this choice in Section 4.

The third class of SOLD methods is based on so-called edge stabilizations, which add the term

$$\sum_{K\in\mathscr{T}_h}\int_{\partial K}\Psi_K(u_h)\operatorname{sign}\left(\frac{\partial u_h}{\partial \boldsymbol{t}_{\partial K}}\right)\frac{\partial v_h}{\partial \boldsymbol{t}_{\partial K}}\,\mathrm{d}\sigma \tag{15}$$

to the left-hand side of (2), $\boldsymbol{t}_{\partial K}$ being a tangent vector to the boundary ∂K of K. Various choices of the nonnegative function Ψ_K were proposed by Burman and Hansbo [6] and Burman and Ern [5]. To make the convergence of the nonlinear iterative process possible, the sign operator is regularized by replacing it by the hyperbolic tangent as recommended in [6]. Our numerical tests in Part I and in [27] indicate that some SOLD methods based on edge stabilizations work comparatively well on unstructured grids with acute triangles, but still away from being perfect. In general, these methods lead to a more pronounced smearing of layers in comparison with the best methods of the previous two classes. The best edge stabilization method in the numerical studies of Part I is defined by $\Psi_K(u_h) = \gamma|(R_h(u_h)|_K)|$, where γ is a nonnegative parameter. This method was called BE05_2 in Part I. We shall see in the next section that the parameter γ should be proportional to the area $|K|$ of the respective element K, i.e., $\gamma|_K = C|K|$ with some $C \geqslant 0$. Then (15) can be written in the form

$$\sum_{K\in\mathscr{T}_h}|K|\int_{\partial K} C\frac{|R_h(u_h)|_K|}{\left|\frac{\partial u_h}{\partial \boldsymbol{t}_{\partial K}}\right|}\frac{\partial u_h}{\partial \boldsymbol{t}_{\partial K}}\frac{\partial v_h}{\partial \boldsymbol{t}_{\partial K}}\,\mathrm{d}\sigma, \tag{16}$$

which has a similar structure like many of the SOLD terms discussed above.

4. Properties of SOLD methods for model problems

In this section, we shall investigate the properties of the SOLD methods described in the previous section by applying them to three model problems whose solutions possess characteristic features of solutions of (1), in particular, parabolic and exponential boundary layers and interior layers. The goal of these investigations consists in understanding why the methods work well or not. All numerical results have been double-checked by computing them with two different codes, one of them was *MooNMD*, [28].

In all model problems, we shall consider (1) with

$$\Omega = (0,1)^2 \quad \text{and} \quad \varepsilon = 10^{-8}. \tag{17}$$

Moreover, we shall confine ourselves to the two types of triangulations depicted in Fig. 1. To characterize these triangulations, we shall use the notion '$N_1 \times N_2$ mesh' where N_1 and N_2 are the numbers of vertices in the horizontal and vertical directions, respectively. The corresponding mesh widths will be denoted by h_1 and h_2, i.e., $h_1 = 1/(N_1 - 1)$ and $h_2 = 1/(N_2 - 1)$.

Example 1 (*Solution with parabolic and exponential boundary layers*). We consider the convection–diffusion equation (1) with (17) and

$$\boldsymbol{b} = (1,0)^{\mathrm{T}}, \quad f = 1, \quad u_b = 0.$$

The solution $u(x,y)$ of this problem, see Fig. 2a, possesses an exponential boundary layer at $x = 1$ and parabolic (characteristic) boundary layers at $y = 0$ and $y = 1$. Outside the layers, the solution $u(x,y)$ is very close to x.

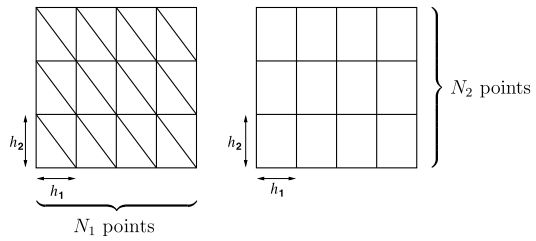


Fig. 1. Triangulations used in Section 4.

This test problem was used, e.g., by Mizukami and Hughes [35].

For this special example, the stabilization parameter τ given in (3) is optimal along lines $y = \text{const.}$ outside the parabolic layers. Therefore, for both the P_1 and Q_1 finite elements, the SUPG method gives a nodally exact solution outside the parabolic layers. However, there are strong oscillations at the parabolic layers, see Fig. 2b, which shows a SUPG solution for the Q_1 finite element. For the P_1 finite element, the solution is similar. To measure the quality of a discrete solution u_h at the parabolic layers, we define the values

$$osc := \max_{y \in [0,1]} \{u_h(0.5,y) - u_h(0.5,0.5)\}, \tag{18}$$

$$smear := \max_{y \in [h_2, 1-h_2]} \{u_h(0.5,0.5) - u_h(0.5,y)\}, \tag{19}$$

see also Part I. The first value measures the oscillations at the parabolic layers. In the case that the oscillations are suppressed to the most part, the second value measures the smearing of these layers.

To investigate the optimality of the definitions of $\tilde{\varepsilon}$ presented in the previous section, we introduce a parameter η such that, for any $K \in \mathscr{T}_h$,

$$\tilde{\varepsilon}|_K = \eta \frac{\operatorname{diam}(K)|R_h(u_h)|}{2|\nabla u_h|} \quad \text{if } \nabla u_h \neq \mathbf{0}. \tag{20}$$

This ansatz is based on the similarities between the SOLD methods discussed at the end of Section 3. The relation (20) can be satisfied provided that $\tilde{\varepsilon} = 0$ if $R_h(u_h) = 0$, which is true in all the cases except for (9). Of course, η generally depends on u_h, $\mathscr{T}_h$ and the data of (1). Nevertheless, we can also consider $\tilde{\varepsilon}$ defined by (20) with a constant value of η, which resembles the first term of (10) and (11). Fig. 3 shows how the value of η influences the oscillations and smearing along the line $x = 0.5$ in a discrete solution of Example 1 defined using the crosswind artificial diffusion term (8). We observe that there is a clear optimal value of η which, however, depends on the used triangulation. We also see that the optimal values of η are nearly the same for both the P_1 and Q_1 finite elements. Using (20) together with the isotropic artificial diffusion term (5), the curves and the optimal values of η are very similar to those in Fig. 3.

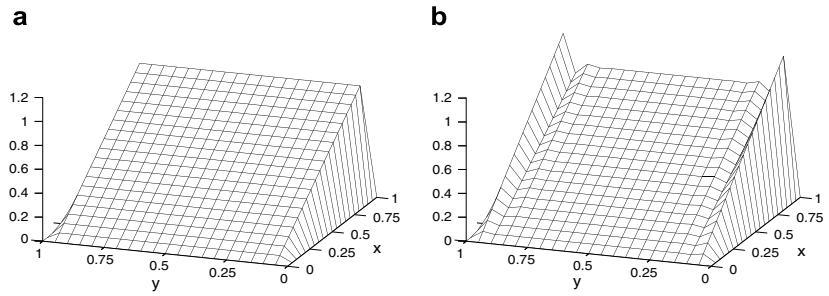


Fig. 2. Example 1: (a) solution u and (b) discrete solution u_h obtained using the SUPG method with the Q_1 finite element on a 21×21 mesh.

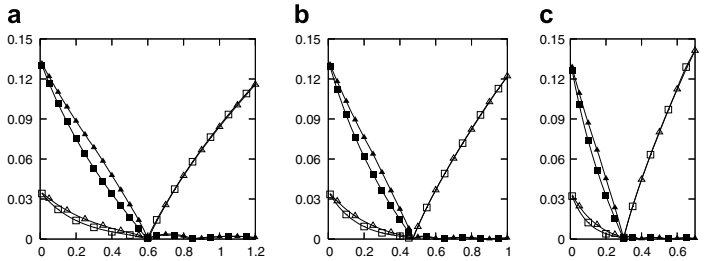


Fig. 3. Example 1, discretization with a crosswind SOLD term given by (8) and (20), dependence of the measures for oscillations (■ P_1, ▲ Q_1) and smearing (□ P_1, △ Q_1) defined by (18) and (19), respectively, on the parameter η: (a) 65×33 mesh; (b) 65×65 mesh; (c) 33×65 mesh.

The optimal values of η from Fig. 3 correspond to discrete solutions which are nodally exact along the line $x = 0.5$. We would like to derive now an analytic expression for the optimal value of η by requiring that the discrete solution be nodally exact outside the exponential boundary layer. For simplicity, we shall consider the case $\varepsilon \to 0+$ so that the nodally exact discrete solution satisfies $u_h(x,y) = x$ for $(x,y) \in [0, 1-h_1] \times [h_2, 1-h_2]$, where h_1 and h_2 are defined in Fig. 1. By the definition of the SOLD methods, we have, for any $v_h \in V_h$,

$$(R_h(u_h), v_h + \tau \boldsymbol{b} \cdot \nabla v_h) + (\tilde{\varepsilon} \nabla u_h, \nabla v_h) = 0 \tag{21}$$

or

$$(R_h(u_h), v_h + \tau \boldsymbol{b} \cdot \nabla v_h) + (\tilde{\varepsilon} \boldsymbol{b}^{\perp} \cdot \nabla u_h, \boldsymbol{b}^{\perp} \cdot \nabla v_h) = 0. \tag{22}$$

In what follows, we shall assume that $\operatorname{supp} v_h \subset [0, 1-h_1] \times [0,1]$. Then it is easy to verify that, for both the P_1 and Q_1 finite elements, the nodally exact discrete solution satisfies $(R_h(u_h), \tau \boldsymbol{b} \cdot \nabla v_h) = 0$ provided that τ is independent of x (for the P_1 finite element, this is true even for any $\tau \in L^{\infty}(\Omega)$ and it follows from the fact that, for any $K \in \mathscr{T}_h$, either $R_h(u_h)|_K = 0$ or $\boldsymbol{b} \cdot \nabla v_h|_K = 0$ – see below). Therefore, the optimal value of η is independent of the choice of τ. It also shows that the SUPG method alone is not able to provide an oscillation-free solution.

Let us consider the P_1 finite element. Then for elements K lying in $[0, 1-h_1] \times [h_2, 1-h_2]$ or having exactly one vertex at the boundary $y = 0$ or $y = 1$, we have $\boldsymbol{b} \cdot \nabla u_h|_K = 1$ and hence $R_h(u_h)|_K = 0$. Thus, the only elements K in $[0, 1-h_1] \times [0,1]$ which may lead to non-vanishing parameters $\tilde{\varepsilon}|_K$ are elements with two vertices at $y = 0$ or $y = 1$. If K is such an element, we may assume that the vertex of K not lying on $y = 0$ or $y = 1$ has the coordinates (ih_1, h_2) or $(ih_1, 1-h_2)$ with $i \in \{1, \ldots, N_1 - 3\}$ since the two elements which have all three vertices on the boundary of $[0, 1-h_1] \times [0,1]$ do not have to be considered. Then $\nabla u_h|_K = (0, \pm ih_1/h_2)$ and, consequently, for any η, we get $(\tilde{\varepsilon} \nabla u_h, \nabla v_h) = (\tilde{\varepsilon} \boldsymbol{b}^{\perp} \cdot \nabla u_h, \boldsymbol{b}^{\perp} \cdot \nabla v_h)$ so that we do not have to distinguish between (21) and (22). If v_h equals 1 at the interior vertex of K and vanishes at all other vertices of the triangulation, the conditions (21) and (22) reduce to

$$(R_h(u_h), v_h)_K + (\tilde{\varepsilon} \nabla u_h, \nabla v_h)_K = 0,$$

where $(\cdot,\cdot)_K$ denotes the inner product in $L^2(K)$ or $L^2(K)^2$. Since $(\nabla u_h \cdot \nabla v_h)|_K = ih_1/h_2^2$ and $R_h(u_h)|_K = -f = -1$, we deduce that the optimal value of $\tilde{\varepsilon}$ is

$$\tilde{\varepsilon}_{\mathrm{opt}}|_K = \frac{h_2^2}{3ih_1}$$

and that the optimal value of η is

$$\eta_{\mathrm{opt}} = \frac{2}{3\sqrt{1 + \left(\dfrac{h_1}{h_2}\right)^2}}. \tag{23}$$

This formula is in a very good agreement with the optimal values of η observed in Fig. 3. Note also that η_{opt} does not depend on K and it depends on the used triangulation only through the aspect ratio of the elements of the triangulation defined by

$$v := \frac{h_1}{h_2}. \tag{24}$$

The graphs in Fig. 3 indicate that a SOLD term of the form (5) or (8) can be expected to lead to an oscillation-free solution only if, on any element $K \subset [0, 1-h_1] \times [0,1]$ with two vertices at $y = 0$ or $y = 1$, the value of $\tilde{\varepsilon}$ corresponding to the nodally exact discrete solution u_h is at least $\tilde{\varepsilon}_{\mathrm{opt}}$. Inserting u_h into the formulas (6), (7) and (10)–(13) from Section 3, we obtain the following relations between $\tilde{\varepsilon}$ and $\tilde{\varepsilon}_{\mathrm{opt}}$ (we drop the notation for restriction to K):

$$(6): \tilde{\varepsilon} = \frac{3}{2}\left(v - \frac{1}{i}\right)\tilde{\varepsilon}_{\mathrm{opt}},$$

$$(7): \tilde{\varepsilon} = 3iv\left(Ch_2(1+v^2) - \frac{\varepsilon}{h_2}\right)\tilde{\varepsilon}_{\mathrm{opt}},$$

$$(9): \tilde{\varepsilon} = 3iv^2\left(\sqrt{h_1} - \frac{\varepsilon}{h_1}\right)\tilde{\varepsilon}_{\mathrm{opt}},$$

$$(10): \tilde{\varepsilon} = 0 \quad \text{since } \boldsymbol{b} \cdot \nabla u_h = 0,$$

$$(11): \tilde{\varepsilon} = \left(C\frac{3}{2}\sqrt{1+v^2} - \frac{3iv\varepsilon}{h_2}\right)\tilde{\varepsilon}_{\mathrm{opt}},$$

$$(12): \tilde{\varepsilon} = \frac{3iv^2}{2(1+iv)}\,\frac{\sqrt{3} + iv(1+\sqrt{3})}{\sqrt{3}+iv}\,\tilde{\varepsilon}_{\mathrm{opt}},$$

$$(13): \tilde{\varepsilon} = \frac{3iv^2}{2(1+iv)}\tilde{\varepsilon}_{\text{opt}}.$$

These relations have to be understood in the way that a right-hand side is replaced by zero if it is negative. As we see, $\tilde{\varepsilon}$ of the original method by Codina defined by (10) cannot be expected to lead to an oscillation-free discrete solution since, for the nodally exact discrete solution, we have $\tilde{\varepsilon} = 0$ on any element in $[0, 1-h_1] \times [0,1]$. On the other hand, using $C = \eta_{\text{opt}}$ in the modified method of Codina with $\tilde{\varepsilon}$ given by (11), we have $\tilde{\varepsilon} \approx \tilde{\varepsilon}_{\text{opt}}$ (provided that the ε-dependent term can be neglected) and hence we obtain nearly the nodally exact solution. The methods with $\tilde{\varepsilon}$ defined by (7) and (9) do not seem to be practical since the ratio $\tilde{\varepsilon}/\tilde{\varepsilon}_{\text{opt}}$ decreases when refining the mesh while keeping the aspect ratio fixed. The remaining three definitions of $\tilde{\varepsilon}$, i.e., (6), (12) and (13), enable to satisfy the condition $\tilde{\varepsilon} \geqslant \tilde{\varepsilon}_{\text{opt}}$ for sufficiently large aspect ratios, in particular, for $v \geqslant 5/3$, $v \geqslant 0.9$ and $v \geqslant (1+\sqrt{7})/3$, respectively.

In the quadrilateral case, it is not possible to derive simple formulas for $\tilde{\varepsilon}_{\text{opt}}$ and η_{opt}, but the results in Fig. 3 suggest that the optimal values of η do not differ much from (23). Therefore, conditions for obtaining an oscillation-free solution can be derived by requiring that the parameters $\tilde{\varepsilon}$ in (5) and (8) satisfy

$$\tilde{\varepsilon}|_K \geqslant \eta_{\text{opt}} \frac{\text{diam}(K)|R_h(u_h)|}{2|\nabla u_h|} = \frac{h_2}{3}\frac{|R_h(u_h)|}{|\nabla u_h|} \quad \forall K \in \mathscr{T}_h \qquad (25)$$

for any function u_h. The resulting relations also apply to the P_1 finite element but are less sharp than above. It is obvious that, for the method of do Carmo and Galeão and for the modified method of Burman and Ern, i.e., for $\tilde{\varepsilon}$ given by (6) or (13), respectively, the inequality (25) may hold only if

$$\tau|\boldsymbol{b}| > \frac{h_2}{3}, \qquad (26)$$

which is equivalent to $v > 2/3$. If $v \leqslant 2/3$, we have to expect spurious oscillations in the discrete solution as it is demonstrated in Fig. 4. The inequality (26) suggests to define τ in (6) and (13) using the element diameter $h_K^{\perp}$ in the direction orthogonal to the convection vector $\boldsymbol{b}$ instead of using h_K. For instance, in the convection-dominated case, we can use the formula

$$\tau|_K = \frac{h_K^{\perp}}{2|\boldsymbol{b}|} \quad \forall K \in \mathscr{T}_h, \qquad (27)$$

which in fact removes the spurious oscillations visible in Fig. 4. For $\tilde{\varepsilon}$ given by (12), the necessary condition obtained from (25) is weaker than (26) but, for a 41×21 mesh, we get a similar discrete solution as in Fig. 4 (slightly better for the P_1 finite element and slightly worse for the Q_1 finite element). On the other hand, if we use $\tilde{\varepsilon}$ given by (11), spurious oscillations should not appear for $C > 2/3 > \eta_{\text{opt}}$, which is particularly satisfied by the value $C \approx 0.7$ recommended in [10]. However, for certain triangulations, the layers can be smeared as Fig. 3 indicates.

As we already showed, $\tilde{\varepsilon}$ defined by (10) is not appropriate in case of the P_1 finite element. The situation is different for the Q_1 finite element for which similar results can be obtained as with (11) provided that the term (8) is evaluated using a quadrature formula with nodes which are not 'too near' to the boundary of Ω.

Finally, let us mention a further drawback of $\tilde{\varepsilon}$ defined by (7). If the functions f and u_b in (1) are multiplied by a constant α, then the solution u changes to αu. For the SOLD methods defined using the terms (5) and (8), this property is valid if and only if the value of $\tilde{\varepsilon}$ does not change after replacing u_h, f by αu_h, αf, respectively. This is true for most of the definitions of $\tilde{\varepsilon}$ mentioned in Section 3, however not for the formula (7). Let us assume that, for a given mesh, the parameter C in (7) is defined in such a way that the corresponding discrete solution is a good approximation to the solution of Example 1. Now, replacing $f = 1$ by $f = \alpha$, we typically obtain with (7) either an oscillatory solution (if $|\alpha| < 1$) or a solution excessively smearing the layers (if $|\alpha| > 1$). This shows that the formula (7) cannot be expected to lead to a qualitatively correct discrete solution unless C depends on u_h or the data of problem (1). This was probably also recognized by Johnson [31] who proposed to set $C = \beta/\max_{\Omega}|u_h|$ in (7) where β is a constant. However, a constant value of β allows to remove spurious oscillations only at the price of a significant smearing of the layers and hence the method does

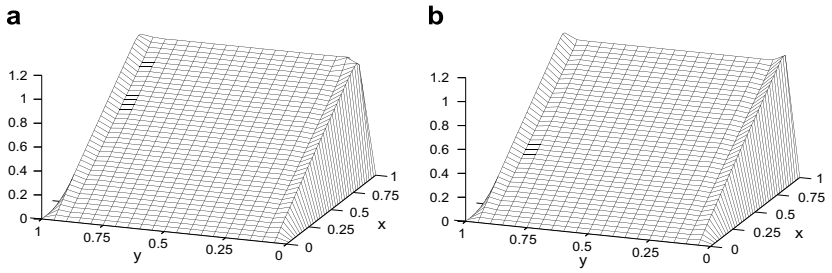


Fig. 4. Example 1, discrete solutions on 41×21 meshes: (a) P_1 finite element, isotropic artificial diffusion given by (6) and (b) Q_1 finite element, crosswind artificial diffusion given by (13).

not attain the quality of the best SOLD methods (see also Part I).

For the edge stabilization term (16) and both the P_1 and Q_1 finite elements, it is easy to derive that the function $u_h(x,y) = x$ satisfies the respective discrete problem for $\varepsilon \to 0+$ and test functions $v_h \in V_h$ with $\operatorname{supp} v_h \subset [0, 1-h_1] \times [0,1]$ if $C = 1/6$. However, in practice, the discrete solution is slightly worse at the parabolic boundary layers due to the regularization of the sign operator. Moreover, in contrast to the modified method of Codina, the discrete solution is significantly smeared along the exponential boundary layer. A sharp approximation of this layer requires to set $C = 0$ in this region.

To summarize the discussion to Example 1, among the SOLD methods adding the isotropic diffusion term (5) or the crosswind diffusion term (8), the only SOLD method which gives satisfactory results seems to be the modified method of Codina defined by (8) and (11), but only with an appropriately chosen constant C. The edge stabilization (16) enables to compute a satisfactory solution if the parameter C is layer-adapted.

Example 2 (*Solution with interior layer and exponential boundary layers*). We consider the convection–diffusion equation (1) with (17) and

$$\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))^{\mathrm{T}}, \quad f = 0,$$
$$u_b(x,y) = \begin{cases} 0 & \text{for } x = 1 \text{ or } y \leqslant 0.7, \\ 1 & \text{else.} \end{cases}$$

The solution, see Fig. 5a, possesses an interior (characteristic) 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. This example was used, e.g., by Hughes et al. [22].

The position of spurious oscillations in the solutions obtained with the SUPG method depends on h_1 and h_2. If the mesh is constructed such that

$$h_1 b_2 + h_2 b_1 < 0, \tag{28}$$

then, for both the P_1 and Q_1 finite elements, the SUPG solution contains oscillations along the interior layer and along the boundary layer at $x = 1$. However, there are no oscillations along the boundary layer at $y = 0$ and this layer is not smeared. This is illustrated in Fig. 5b which shows a SUPG solution for the P_1 finite element. For the Q_1 finite element the discrete solution is very similar. If $h_1 b_2 + h_2 b_1 > 0$, then the SUPG solution contains oscillations along the interior layer and along the boundary layer at $y = 0$ but no oscillations and no smearing occur along the boundary layer at $x = 1$. For shortness of presentation, we shall consider only the case (28) in the following.

For a nodally exact solution, the SUPG term will not vanish in Example 2 (in contrast to Example 1). Thus, for obtaining a nodally exact solution with a SOLD method, the choice of the SUPG parameter τ will be of importance, too. The chosen parameter has to ensure that there is no smearing of layers since smeared layers cannot be corrected with SOLD methods. With the approach presented in Section 2, the SUPG parameter in Example 2 will be the same on each element. We found that the choice (3) is optimal in the class of globally constant parameters in the sense that any larger value leads to a smearing of the layer at $y = 0$ and any smaller value results in spurious oscillations at this layer and increases the oscillations at $x = 1$.

Let us first investigate the quality of the approximation of the interior layer. For simplicity, we shall confine ourselves to the P_1 finite element unless stated otherwise. To measure the oscillations of a discrete solution u_h at the interior layer, we define the value

$$osc_{\mathrm{int}} := \max\left\{ \max_{(x,y)\in G} u_h(x,y) - 1, \left| \min_{(x,y)\in G} u_h(x,y) \right| \right\}, \tag{29}$$

where (x,y) are the nodes in $G := [0, 0.5] \times [0.25, 1]$. Let us again consider SOLD methods defined using the term (5) or (8) with $\tilde{\varepsilon}$ given by (20). Numerical tests show that the value of osc_{int} is a non-increasing function of η on a given mesh. Given an integer m, we define

$$\eta_m := \min\{\eta \in \mathbb{R}_0^+; osc_{\mathrm{int}}(\eta) \leqslant 10^{-m}\}.$$

This value depends on the aspect ratio v defined in (24). In view of (28), we have $v > \sqrt{3}/3$. Fig. 6 presents the dependence of η_2, η_3 and η_4 on the aspect ratio for both the isotropic and the crosswind artificial diffusion and for $h_1 = 1/64$. Of course, h_2 and consequently the number of degrees of freedom is different for different aspect ratios.

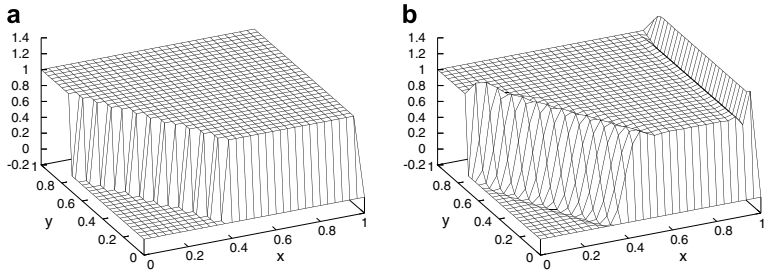


Fig. 5. Example 2: (a) solution u and (b) discrete solution u_h obtained using the SUPG method with the P_1 finite element on a 31×31 mesh.

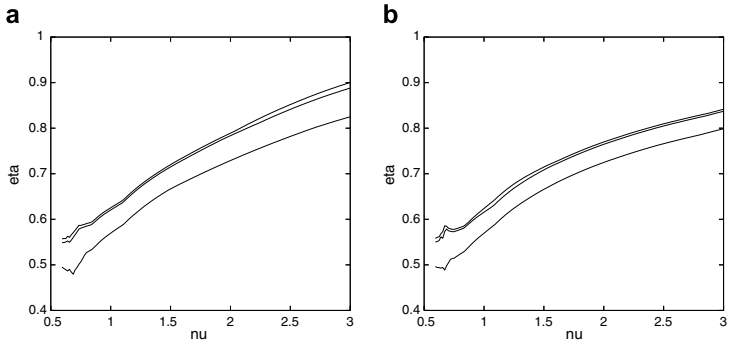


Fig. 6. Example 2, dependence of η_2, η_3 and η_4 on v (from bottom to top) for the P_1 finite element and meshes with $h_1 = 1/64$: (a) isotropic artificial diffusion (5) and (b) crosswind artificial diffusion (8).

We checked with several values for h_1 that the results presented in Fig. 6 depend only on v. Thus, one would get the same results for a fixed number of degrees of freedom with varying h_1 and h_2. Fig. 6 shows that the smallest value of η assuring that oscillations will not exceed a given tolerance increases with increasing aspect ratio. Qualitatively, the results for the Q_1 finite element are the same as for the P_1 finite element: increasing aspect ratios require increasing parameters η to suppress the oscillations below given thresholds.

For small ε, formula (20) for $\tilde{\varepsilon}$ is the main part of the method of Codina given by (8) and (10). Particularly, the results in Fig. 6 show that, in contrast to Example 1, the recommended value $C \approx 0.7$ does not generally lead to sufficiently small spurious oscillations.

Now let us turn our attention to the method of do Carmo and Galeão given by (5) and (6) and the modified method of Burman and Ern given by (8) and (13). Comparing the formulas (6) and (13) with (20), one finds, using (4), that for obtaining comparable results as for $\tilde{\varepsilon}$ defined by (20) with a given value of η, the condition

$$\eta \leqslant \frac{h_K}{\mathrm{diam}(K)} = \frac{2}{\sqrt{3}\sqrt{1+v^2}} \tag{30}$$

should be satisfied. The investigations of Example 1 suggested to define τ in (6) and (13) by (27). Since an interior layer is a characteristic layer, it is natural to ask whether this modification is reasonable also in the present example. Then, instead of (30), we obtain the condition

$$\eta \leqslant \frac{h_K^{\perp}}{\mathrm{diam}(K)} = \frac{2v}{(\sqrt{3}+v)\sqrt{1+v^2}}. \tag{31}$$

Fig. 7 compares the curves $\eta_2 = \eta_2(v)$ for both the isotropic and the crosswind artificial diffusion with the functions on the right-hand sides of (30) and (31). Values of the right-hand sides of (30) and (31) below the curves of $\eta_2(v)$ indicate that the values of (6) and (13) are too small to suppress the oscillations at the interior layer below the value 10^{-2}. Thus, Fig. 7 shows that the method

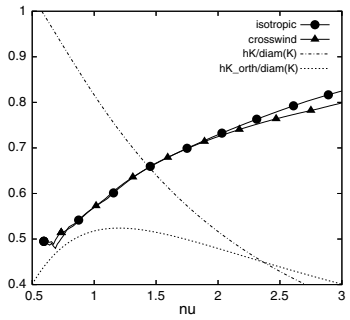


Fig. 7. Example 2, dependence on v of η_2 for the isotropic and the crosswind artificial diffusion and of the functions from the right-hand sides of (30) and (31).

of do Carmo and Galeão and the modified method of Burman and Ern will generally lead to non-negligible spurious oscillations at the interior layer of Example 2. Replacing h_K by $h_K^{\perp}$ in the definition of τ used in (6) and (13), oscillations of size at least 10^{-2} should appear for any aspect ratio and they should be mostly even larger than for τ defined using h_K. Thus, in contrast to Example 1, τ in (6) and (13) should be defined rather using h_K for small aspect ratios ($v \lesssim 1.5$) and using even a measure larger than h_K, for instance $\mathrm{diam}(K)$, for larger aspect ratios.

Next, the usefulness of the curves presented in Fig. 7 will be demonstrated. Considering, e.g., $v = 2$, one expectation is that the method of Codina given by (8) and (10) with $C = 0.7$, whose parameter $\tilde{\varepsilon}$ corresponds to the solid lines, leads to a solution with small spurious oscillations at the interior layer (less than 10^{-2}). In contrast, the methods of do Carmo and Galeão, (5) and (6), and of Burman and Ern, (8) and (13), whose parameters correspond to the dash-dot line, should produce solutions with larger oscillations at the interior layer. Fig. 8 shows numerical examples which confirm both expectations. For the methods (5), (6) and (8), (13), the results obtained

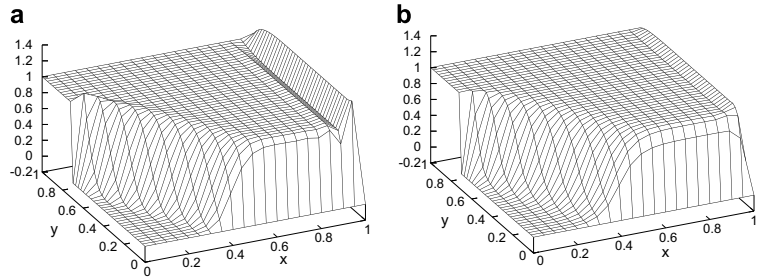


Fig. 8. Example 2, discrete solution u_h obtained on 21×41 meshes using (a) the method of do Carmo and Galeão and the Q_1 finite element and (b) the method of Codina with $C = 0.7$ and the P_1 finite element.

with both the P_1 and the Q_1 finite elements are similar. In particular, these solutions possess non-negligible spurious oscillations at the beginning of the interior layer. Considering the method of Codina and the Q_1 finite element, the violation of the discrete maximum principle at the beginning of the interior layer is larger and mainly in form of undershoots. For the method of Burman and Ern given by (8) and (12), the results are similar as for the method of Codina but slightly worse with respect to the spurious oscillations.

As pointed out above, the results of a SOLD method depend not only on the definition of $\tilde{\varepsilon}$ but also on the definition of τ in the SUPG term. In addition, we explained that the formula (3) is optimal with respect to the boundary layer at $y = 0$. Neglecting for the moment the quality of the solution at this boundary layer, one can ask whether increasing τ can help to reduce the spurious oscillations at the characteristic layer. However, the expectations are rather low because, in case of a characteristic layer, the influence of the choice of τ is usually weak since the SUPG method stabilizes in the streamline direction which is nearly perpendicular to the direction in which oscillations appear. Fig. 9 shows a comparison of η_4 for both the isotropic and the crosswind artificial diffusion and for two choices of τ. One choice of τ is the same as before and the other one is given by the formula (3) where h_K is replaced by $\mathrm{diam}(K)$. The use of the element diameter in the definition of τ is quite common in practice. It can be seen that increasing the amount of the streamline diffusion provided by the SUPG method requires to introduce more crosswind diffusion by the SOLD term if larger aspect ratios are used to reduce the oscillations at the characteristic layer below 10^{-4}. In summary, generally, the spurious oscillations at the interior layer present in the solution of a SOLD method cannot be expected to become smaller if higher values of the SUPG parameter τ are used.

Let us now consider the boundary layers. One can observe in Fig. 8 that the boundary layer at $y = 0$ is slightly smeared and that oscillations appear along the boundary layer at $x = 1$. The smearing is not surprising since the SUPG solution approximates the boundary layer

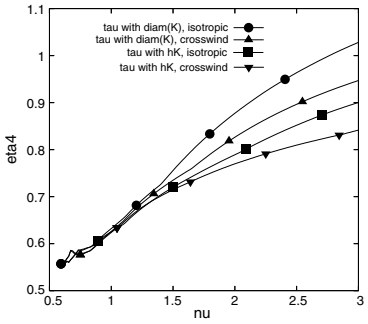


Fig. 9. Example 2, dependence of η_4 on ν for the isotropic and the crosswind artificial diffusion and for the SUPG parameter τ defined either by (3) or by (3) with h_K replaced by $\mathrm{diam}(K)$.

at $y = 0$ nodally exactly for $\varepsilon \to 0+$. Thus, along the boundary layer at $y = 0$, the optimal choice of $\tilde{\varepsilon}$ in a SOLD term is $\tilde{\varepsilon} = 0$, i.e., $\eta_{\mathrm{opt}} = 0$ in (20). To investigate the optimality of $\tilde{\varepsilon}$ for the boundary layer at $x = 1$ with $y \in [h_2, 1]$, let us again consider $\varepsilon \to 0+$ and $\tilde{\varepsilon}$ given by (20). The optimal solution has the values $u_h = 1$ at the nodes with $x = 1 - h_1$. A straightforward computation reveals that the value of η for obtaining this optimal solution is

$$\eta_{\mathrm{opt}} = \frac{h_1 b_2 + h_2 b_1}{\mathrm{diam}(K) b_2}$$

for the isotropic artificial diffusion (5) and

$$\eta_{\mathrm{opt}} = \frac{(h_1 b_2 + h_2 b_1)|\boldsymbol{b}|^2}{\mathrm{diam}(K) b_2^3}$$

for the crosswind artificial diffusion (8). These formulas hold for both the P_1 and the Q_1 finite elements. One can see that the optimal choice of η depends not only on the aspect ratio of the elements of the triangulation but also on the direction of the convection vector $\boldsymbol{b}$. The most important conclusion is that different values of η should be used in different regions of the computational domain.

To find a universal formula for the optimal value of η is very difficult or even impossible. This will be demonstrated by studying a limit case of Example 2 where the limit is approached in two different ways. First, consider the limit case $b_1 \to 0+$ and $b_2 \to -1$, $|\boldsymbol{b}| = 1$. Then, for both SOLD terms (5) and (8), we get

$$\eta_{\text{opt}} = \frac{h_1}{\text{diam}(K)} \quad \text{along the boundary } x = 1 \text{ if } \boldsymbol{b} = (0,-1).$$

On the other hand, consider $\boldsymbol{b} = (0,-1)$, the boundary conditions of Example 2 and a constant right-hand side $f > 0$ of (1). The optimal solution on the mesh line at $x = 1 - h_1$ has the form $u(x,y) = f(1-y) + 1$ (away from the lower boundary). Now, using the considerations leading to (23) gives

$$\eta_{\text{opt}} = \frac{2h_1}{3\text{diam}(K)} \quad \text{along the boundary } x = 1 \text{ if } \boldsymbol{b} = (0,-1) \tag{32}$$

independently of the choice of f. In particular, (32) holds for $f \to 0+$ and hence we obtained two different limit values of η_{opt}.

For the edge stabilization term (16) and both the P_1 and Q_1 finite elements, one can show similarly as above that the optimal value of the parameter C at $x = 1$ is

$$C_{\text{opt}} = \frac{h_1 b_2 + h_2 b_1}{4 h_1 b_2}.$$

For $\boldsymbol{b} = (0,-1)$, the limit values of the optimal C at $x = 1$ are 1/6 for Example 1 and 1/4 for Example 2 and hence they also differ by the factor 2/3. Choosing $C = C_{\text{opt}}$ in Example 2 still leads to oscillations at the interior layer. These can be suppressed by increasing the value of C in this region. This shows once again that different values of the parameter should be used in different regions of the computational domain to obtain a globally satisfactory solution.

The above discussion supports our conclusion to Example 1 that the best SOLD methods are the modified method of Codina and the edge stabilization (16), however, only if the parameter C is chosen appropriately, i.e., layer-adapted. Nevertheless, one generally cannot expect that the discrete solutions will be without any spurious oscillations.

Example 3 (*Solution with two interior layers*). We consider the convection–diffusion equation (1) with (17) and

$$\boldsymbol{b} = (1,0)^{\mathrm{T}}, \quad u_b = 0,$$

$$f(x,y) = \begin{cases} 16(1-2x) & \text{for } (x,y) \in [0.25, 0.75]^2, \\ 0 & \text{else.} \end{cases}$$

The solution, see Fig. 10a, possesses two interior (characteristic) layers at $(0.25, 0.75) \times \{0.25\}$ and $(0.25, 0.75) \times \{0.75\}$. In $(0.25, 0.75)^2$, the solution $u(x,y)$ is very close to the quadratic function $(4x-1)(3-4x)$. This example was first considered by John and Knobloch [25].

This is an example of a problem for which all the SOLD methods mentioned in Section 3 fail. Note that, in contrast to Example 2, the data of Example 3 satisfy the requirements for defining the standard weak formulation of (1). Moreover, the solution of Example 3 belongs to $H^2(\Omega)$, cf. Grisvard [18].

As expected, the SUPG solution of Example 3 possesses spurious oscillations along the interior layers, see Fig. 10b. To visualize both undershoots and overshoots, we present the SUPG solution at an angle for which the plane $z = 0$ reduces to a line. Applying the modified method of Codina with $C = 0.7$, the spurious oscillations present in the SUPG solution are significantly suppressed, however, the solution is wrong in the region $(0.75, 1) \times (0, 1)$, see Fig. 11. Very similar results are obtained for *any* of the SOLD methods mentioned in Section 3 and for both the P_1 and Q_1 finite elements.

Note that, in view of the discontinuous right-hand side f, the SOLD methods should be implemented using quadrature formulas whose nodes do not lie on the edges of the triangulations. However, such nodes cannot be avoided when evaluating the edge stabilization term (16), which complicates the implementation of this method.

To measure the spurious oscillations of a discrete solution u_h to Example 3, we define the values

$$min := -\min_{0.4 \leqslant x \leqslant 0.6} u_h(x,y), \quad diff := \max_{x \geqslant 0.8} u_h(x,y) - \min_{x \geqslant 0.8} u_h(x,y), \tag{33}$$

where $y \in [0,1]$ and $\min u_h$ and $\max u_h$ are computed using values of u_h at the vertices of $\mathscr{T}_h$. Tables 1 and 2 show the values of *min* and *diff*, respectively, for the P_1 finite element, most of the SOLD methods discussed above and several meshes. The abbreviations denoting the methods can be

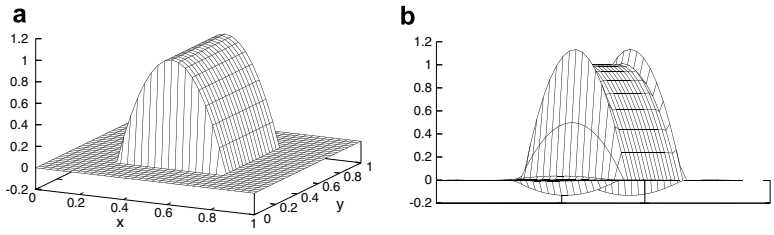


Fig. 10. Example 3: (a) solution u and (b) discrete solution u_h obtained using the SUPG method with the P_1 finite element on a 33×33 mesh.

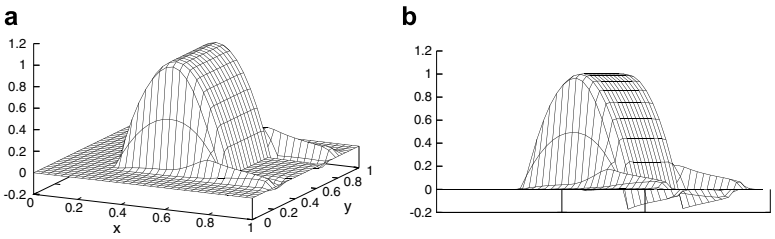


Fig. 11. Example 3, discrete solution u_h obtained on a 33×33 mesh using the modified method of Codina with $C = 0.7$ and the P_1 finite element: (a) view as in Fig. 10a and (b) view as in Fig. 10b.

Table 1
Example 3, values of *min* defined in (33) obtained for the P_1 finite element using the methods from Sections 2 and 3

Method	Mesh			
	17×17	33×33	65×65	129×129
SUPG	1.31e−1	1.33e−1	1.34e−1	1.34e−1
dCG91	2.37e−2	1.27e−2	2.42e−3	n.c.
KLR02_3, $C = 0.4714$	1.93e−2	1.88e−2	1.22e−2	6.85e−3
KLR02_3, $C = 0.7$	8.52e−3	1.38e−3	2.65e−4	n.c.
BE02_1	1.37e−2	9.33e−3	n.c.	n.c.
BE02_2	1.85e−2	7.74e−3	1.20e−3	n.c.
BE05_2, $C = 1/6$	1.06e−2	6.77e−3	3.98e−3	2.04e−3
BE05_2, $C = 0.4$	2.79e−3	1.59e−3	8.24e−4	n.c.

Table 2
Example 3, values of *diff* defined in (33) obtained for the P_1 finite element using the methods from Sections 2 and 3

Method	Mesh			
	17×17	33×33	65×65	129×129
SUPG	3.30e−3	9.52e−5	3.83e−5	1.53e−4
dCG91	2.62e−1	2.95e−1	2.81e−1	n.c.
KLR02_3, $C = 0.4714$	2.88e−1	3.24e−1	3.37e−1	3.37e−1
KLR02_3, $C = 0.7$	2.82e−1	2.74e−1	2.42e−1	n.c.
BE02_1	3.77e−1	4.36e−1	n.c.	n.c.
BE02_2	2.78e−1	2.94e−1	2.76e−1	n.c.
BE05_2, $C = 1/6$	2.76e−1	3.05e−1	3.25e−1	3.36e−1
BE05_2, $C = 0.4$	2.53e−1	2.56e−1	2.43e−1	n.c.

found in Section 3 and are the same as in Part I. The abbreviation nc means that the nonlinear iterative process did not converge, see the next section. This happens mainly for the finest mesh. Generally, the convergence of the nonlinear iterations deteriorates if the mesh becomes finer or the parameter C in (11) or (16) increases. We consider two values of C for each method. First, since interior layers are characteristic layers, we use the optimal values of C found in the investigations of Example 1. For (11), we further use the value $C = 0.7$ recommended in [10]. For (16), the value $C = 0.4$ corresponds to the choice of C in Part I. Table 1 shows that all the SOLD methods significantly reduce the undershoots along the interior layers present in the SUPG solution (the same holds for overshoots). For the considered meshes, the maximal undershoots of the SUPG method are not influenced by the size of the mesh width. In contrast to this, for all the SOLD methods, the undershoots become smaller if the mesh is refined. The undershoots also decrease if the parameter C in (11) or (16) increases. However, for larger values of C, the smearing of the discrete solution is more pronounced and, as we mentioned, the convergence of the nonlinear iterative process deteriorates.

Table 2 shows that the wrong part of the discrete solution in $(0.8, 1) \times (0, 1)$ is of comparable magnitude for all the SOLD methods and does not improve significantly if the mesh is refined or C is increased (both in the range where the nonlinear iterative schemes converge). Therefore, we conclude that, using the SOLD methods described in Section 3, it is not feasible to obtain a qualitatively correct approximation of the solution to Example 3. An open question is whether appropriately defined non-constant parameters in the modified method of Codina (11) or the edge stabilization (16) might lead to satisfactory solutions.

5. The solution of the nonlinear discrete problems

The discrete SOLD problems can be written in the form

$$a_h(u_h; u_h, v_h) = \langle f, v_h \rangle \quad \forall v_h \in V_h,$$

where $a_h(u_h; \cdot, \cdot)$ is a bilinear form and the first argument of a_h enters the definition of a_h through the parameter $\tilde{\varepsilon}$ or the respective term in (16). Thus, it is straightforward to compute the discrete solution by means of the following iterative scheme. Given an approximation u_h^k of the solution of the SOLD system, compute $\tilde{u}_h^{k+1}$ by solving

$$\tilde{u}_h^{k+1}: \quad a_h(u_h^k; \tilde{u}_h^{k+1}, v_h) = \langle f, v_h \rangle \quad \forall v_h \in V_h. \tag{34}$$

The next iterate is defined as

$$u_h^{k+1} := u_h^k + \omega_{k+1}(\tilde{u}_h^{k+1} - u_h^k)$$

with the damping factor $\omega_{k+1} > 0$.

As initial iterate u_h^0, we use the solution obtained with the SUPG method. Thus, apart from the spurious oscillations, the initial iterate coincides already rather well with the solution wished to be obtained with the SOLD methods.

Our experiences are that an appropriate choice of the damping factors $\{\omega_k\}$ is often essential for the convergence

of the iterative process and the number of iterations. Appropriate damping factors depend on the SOLD scheme, the problem and its data, the grid and the choice of parameters in parameter-dependent SOLD schemes and these damping factors might be very different. Since it is not practicable in applications that the user should find every time an appropriate damping factor, it is necessary to use a strategy for an automatic and dynamic choice of this factor.

The dynamic choice of the damping factor which we used in our computations is illustrated with the pseudo code in Fig. 12. Our approach contains a number of parameters, whose values for the results presented in this section are given on lines 1–2. These values seemed reasonable choices in our opinion and we did not try to optimize them for the examples considered in this paper. Our strategy for the dynamic choice of the damping factor is based on the following principles:

- There is an upper bound $\omega_{\max}$ for the damping factor. The upper bound is adjusted dynamically in the course of the iterative process. Initially, we set $\omega_{\max} = 1$, i.e., no damping.
- There is a lower bound $\omega_{\min}$ for the damping factor. This bound is fixed. We used in the computations presented in this paper $\omega_{\min} = 0.01$. Note that very small damping factors lead in general to a very large number of iterations and thus to inefficient schemes.
- The iterate u_h^{k+1} is accepted if the norm $|R_h(u_h^{k+1})|$ of its residual

$$\langle R_h(u_h^{k+1}), v_h\rangle := a_h(u_h^{k+1}; u_h^{k+1}, v_h) - \langle f, v_h\rangle, \quad v_h \in V_h,$$

is smaller than $|R_h(u_h^k)|$ or if ω is not allowed to decrease any more, see the pseudo code presented in Fig. 12, lines 10–14. If $|R_h(u_h^{k+1})| < |R_h(u_h^k)|$ and if there was no rejection of an iterate u_h^{k+1} for a larger value of ω before, the maximal damping factor will be increased, see line l2, and then the damping factor will be increased, too, see line l3.
- If the proposal for the iterate u_h^{k+1} is not accepted, ω will be decreased, see line 16. In addition, if in the step $k+1$ an iterate is rejected the first time, $\omega_{\max}$ will be decreased too, see lines 17–20. Now, a new proposal for u_h^{k+1} is computed with the new value of the damping factor. The acceptance or rejection of this new proposal is checked the same way as for the former damping factor.

```
 1. ω_min := 0.01;  ω_max := 1
 2. c_1 := 1.001;  c_2 := 1.1;  c_3 := 1.001;  c_4 := 0.9
 3. compute SUPG solution u_h^0 and residual r^0
 4. ω := ω_max;  k := 0
 5. while r^k > tolerance do
 6.   compute ũ_h^{k+1} satisfying (34)
 7.   first_damp := 1
 8.   u_h^{k+1} := u_h^k + ω(ũ_h^{k+1} − u_h^k)
 9.   compute residual r^{k+1}
10.   if r^{k+1} < r^k or ω ≤ c_1 ω_min then
11.     if r^{k+1} < r_k and first_damp = 1 then
12.       ω_max := min{1, c_3 ω_max}
13.       ω := min{ω_max, c_2 ω}
14.     endif
15.   else
16.     ω := max{ω_min, ω/2}
17.     if first_damp = 1 then
18.       ω_max := max{ω_min, c_4 ω_max}
19.       first_damp := 0
20.     endif
21.     goto line 8
22.   endif
23.   k := k + 1
24. endwhile
```

Fig. 12. Dynamic choice of the damping factor.

The main features of this approach are as follows:

- The damping factor decreases in general if the residual increases.
- The decrease of the damping factor stops at the threshold $\omega_{\min}$ so that also a non-monotone sequence with respect to the norm of the residual can be computed.
- The damping factor as well as the maximal damping parameter increase if the residual decreases to improve the efficiency of the nonlinear iteration scheme. Thus, a strong damping, which might be necessary only at the beginning of the iterative process, influences the damping factor at the end of the process only slightly.

In the simulations presented in this paper, the linear systems were solved by a sparse direct solver (UMFPACK, [13]). Since the costs for solving the linear systems are always the same, this leads to a fair comparison of the costs of the iterative process for all SOLD schemes by simply giving the number of nonlinear iterations.

In practice, it suffices to solve the linear systems only approximately by a few steps of an iterative method without affecting the convergence of the nonlinear iterative method much. This approach might be faster, depending on the iterative linear system solver. However, different numbers of iterations for solving the linear systems are in general necessary for different SOLD schemes, which makes it harder to perform a fair comparison.

Below, our experiences with respect to the solution of the nonlinear discrete problems corresponding to the examples of Section 4 are reported. Tables with characteristic results are presented, where besides the dynamic approach for computing the damping factor also numbers of iterations with fixed factors are given. The computations were carried out for the P_1 and the Q_1 finite elements on 65×65, 33×65 and 65×33 meshes. The iterative processes were stopped if the l^2-norm of the residual vector was smaller than 10^{-8} or after 100,000 iterations (n.c. = not convergent in the tables). Again, the abbreviations of the SOLD methods given in Section 3 are used.

The numbers of iterations generally depend on the quadrature formula used and this dependence is stronger

for the Q_1 finite element than for the P_1 finite element. All results in this paper were computed using Gaussian quadrature formulas of order 5 (with 7 nodes in case of triangles and 9 nodes in case of rectangles). Of course, for Examples 1 and 2 discretized using the P_1 finite element, the results are independent of the used quadrature formula since all integrands are constant or linear.

We would like to emphasize that analytical results concerning the existence and uniqueness of solutions to the nonlinear discrete problems are not available. Thus, it cannot be excluded that a failure of all used damping strategies has its reason in the non-existence of the solution of the nonlinear discrete problem.

Example 1. The nonlinear discrete problems on the 65×65 and the 65×33 meshes could be usually solved without damping, see Table 3. Apart from dCG91 and BE05_2 with $C = 0.4$, the iterative schemes converged in only few iterations. Solving the problems on the 33×65 mesh required for some SOLD methods considerable damping, see Table 4 for the Q_1 finite element. For the P_1 finite element, the convergence was mostly even worse than in Table 4 and dCG91 did not converge at all. Except the latter case, the dynamic choice of the damping factor was always successful, but often more iterations were needed than with the best fixed damping factor, cf. also the last row in Table 3. In these computations, the dynamic approach proposes many damping factors close to $\omega_{\min}$ because the norm of the residual is slightly oscillating, before finally convergence is achieved. Note that the numbers of iterations for the optimal constant in KLR02_3 are very small on both meshes.

Table 3
Example 1, number of iterations for solving the nonlinear SOLD problems, 65×65 mesh, P_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	472	236	161	169	169
KLR02_3, $C = 0.4714$	71	32	18	9	9
KLR02_3, $C = 0.7$	108	50	32	22	22
BE02_1	76	36	24	28	28
BE02_2	92	44	27	19	19
BE05_2, $C = 1/6$	164	78	50	29	29
BE05_2, $C = 0.4$	1010	506	345	n.c.	10943

Table 4
Example 1, number of iterations for solving the nonlinear SOLD problems, 33×65 mesh, Q_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	394	n.c.	n.c.	n.c.	935
KLR02_3, $C = 0.2981$	73	33	20	13	13
KLR02_3, $C = 0.7$	119	64	63	157	66
BE02_1	235	173	218	n.c.	339
BE02_2	213	380	n.c.	n.c.	353
BE05_2, $C = 1/6$	78	36	23	72	72
BE05_2, $C = 0.4$	n.c.	n.c.	n.c.	n.c.	n.c.

Example 2. The nonlinear discrete SOLD problems in this example were harder to solve than for Example 1, in particular for the P_1 finite element. Even on the equidistant mesh, strong damping was necessary, see Table 5. The dynamic choice of the damping factor always led to the convergence of the iterative process on this mesh. Using the P_1 finite element on the 65×33 mesh, the nonlinear problems could be solved only for KLR02_3 and BE05_2 with sufficiently small parameters. The solution of the discrete problems with the Q_1 finite element was much easier on all grids, see Table 6 for representative results.

Example 3. Using the equidistant 65×65 mesh with the P_1 and Q_1 finite element, the discrete equations could be solved without damping for most of the SOLD methods, see Table 7. Only for BE02_1 and BE05_2 with $C = 0.4$, it was not possible to solve them at all, see also Tables 1 and 2. These tables show also that the solution of the nonlinear problems for the P_1 finite element on the next finer equidistant grid became more difficult. We could obtain convergence only for the parameter-dependent SOLD schemes with sufficiently small parameters. For the P_1 finite element on the 33×65 mesh, the iterative processes was

Table 5
Example 2, number of iterations for solving the nonlinear SOLD problems, 65×65 mesh, P_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	160	n.c.	n.c.	n.c.	340
KLR02_3, $C = 0.7$	194	n.c.	n.c.	n.c.	408
BE02_1	n.c.	n.c.	n.c.	n.c.	389
BE02_2	210	n.c.	n.c.	n.c.	412
BE05_2, $C = 0.4$	362	n.c.	n.c.	n.c.	536

Table 6
Example 2, number of iterations for solving the nonlinear SOLD problems, 65×65 mesh, Q_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	67	36	29	33	33
KLR02_3, $C = 0.7$	102	58	51	60	60
BE02_1	213	275	n.c.	n.c.	203
BE02_2	84	47	39	45	45
BE05_2, $C = 0.4$	689	n.c.	n.c.	n.c.	7520

Table 7
Example 3, number of iterations for solving the nonlinear SOLD problems, 65×65 mesh, P_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	158	86	59	49	49
KLR02_3, $C = 0.4714$	157	74	46	33	33
KLR02_3, $C = 0.7$	199	115	89	115	110
BE02_1	n.c.	n.c.	n.c.	n.c.	n.c.
BE02_2	178	93	65	62	62
BE05_2, $C = 1/6$	173	83	53	37	37
BE05_2, $C = 0.4$	n.c.	n.c.	n.c.	n.c.	n.c.

Table 8
Example 3, number of iterations for solving the nonlinear SOLD problems, 33×65 mesh, Q_1 finite element

Method	$\omega = 0.25$	$\omega = 0.5$	$\omega = 0.75$	$\omega = 1$	Dynamic
dCG91	475	n.c.	n.c.	n.c.	599
KLR02_3, $C = 0.2981$	123	58	36	25	25
KLR02_3, $C = 0.7$	247	168	n.c.	n.c.	345
BE02_1	332	974	n.c.	n.c.	461
BE02_2	317	432	n.c.	n.c.	381
BE05_2, $C = 1/6$	150	72	46	33	33
BE05_2, $C = 0.4$	565	277	184	146	1640

not convergent for dCG91, BE02_2 and KLR02_3 with $C = 0.7$. The results for the Q_1 finite element and the 33×65 mesh are presented in Table 8. Again, the need of damping can be observed as well as the successfulness of the dynamic approach (however, on the expense of somewhat more iterations than for the best fixed damping factors). On 65×33 meshes, the only method which did not converge at all was BE05_2 with $C = 0.4$.

Remark 2. The numerical studies show that even for the academic test problems considered in this paper, it was sometimes difficult to solve the nonlinear SOLD problems. Considering more challenging problems, like the one defined by Hemker [19], the difficulties in the solution of the nonlinear problems became even greater. For instance, convergence for KLR02_3 on reasonably structured grids could be achieved only for rather small constants C.

Remark 3. Another possibility for solving the nonlinear discrete problems is to apply Newton's method. However, it is rather difficult to implement since one deals with non-smooth operators. Therefore, usually it is convenient to use some simplified version of Newton's method. In any case, a good initial approximation is typically needed. Hence a general strategy is first to apply the iterative scheme given above and then to switch to Newton's method, possibly by performing several special iterations assuring a smooth transition between the two iterative processes. An appropriate switching point or transition strategy depend on the solved problem, the SOLD scheme, the grid, etc. If the norm of the residual increases after switching to Newton's method, it is advisable to return to the original iterative process without employing the results of the Newton iterations and to try to switch to Newton's method at a later stage of the iterative process. Applying this alternative strategy instead of the iterative scheme studied in this paper, the numbers of iterations change of course but the ranking of the methods basically remains the same. This can be explained by our observation that the methods with a large number of iterations in Tables 3–8 usually show a slow rate of convergence from the beginning of the iterative process. Thus, these methods also require a large number of iterations for obtaining a good initial approximation for Newton's method.

Our experiences concerning the solution of the nonlinear SOLD problems can be summarized as follows:

- Generally, it was easier to solve the problems for the Q_1 finite element than for the P_1 finite element.
- The larger the constant in the SOLD methods KLR02_3 and BE05_2, the more iterations were needed. If the constant became too large (size depended on the problem, the grid, etc.), the iterative process did not solve the nonlinear problem any more.
- It was often easier to solve the problems arising from the SOLD method BE02_2 than those coming from BE02_1.
- Solving the problems obtained with the edge stabilization BE05_2 required in general somewhat more iterations than solving the problems coming from KLR02_3, if in both SOLD methods reasonable constants with respect to the reduction of the spurious oscillations have been chosen. Moreover, the convergence of BE05_2 was much more sensitive to the choice of the parameter C than it was for the method KLR02_3.
- If the nonlinear discrete problems could be solved at all, the dynamic choice of the damping factor was generally among the successful approaches. If damping was necessary, the dynamic approach needed often more iterations than an appropriately chosen fixed damping factor.

6. Numerical results obtained with adaptive methods

In several discussions with our colleagues about Part I, the question arose whether the application of adaptive methods is useful for the reduction of spurious oscillations. In this section, we shall study this question for the SUPG method and adaptive grids obtained with two residual-based error estimators, which are typically used in applications.

There are different ways of defining criteria for a fair comparison of the results obtained with adaptive methods and with SOLD schemes. One possible criterion is to require that the number of degrees of freedom is roughly the same. A different one might be that the computing times are similar. Since the solution of the nonlinear discrete problems of the SOLD methods often is rather time-consuming (because of the large number of iterations), it is possible to solve the linear problems on adaptive meshes with much more degrees of freedom in the same time. Both criteria might be of interest and thus, we will present results on adaptive meshes starting with a few thousand degrees of freedom up to more than 100,000 degrees of freedom.

Computational studies for Example 2 will be presented. As starting grid for the adaptive refinement, we used the triangular grid from Fig. 1 with $h_1 = h_2 = 1/16$ (289 degrees of freedom). The control of the adaptive refinement process was performed analogously to the way described in Section 4 of [23]. The oscillations at the interior layer were

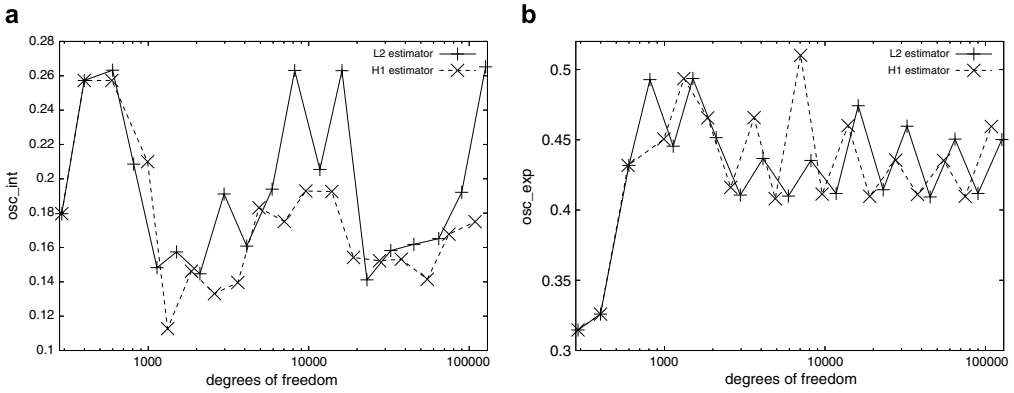


Fig. 13. Example 2, oscillations on adaptively refined grids: (a) interior layer; (b) exponential layer.

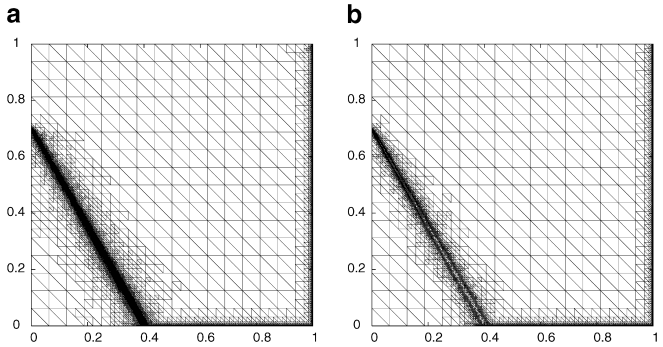


Fig. 14. Example 2, adaptive grids with more than 100,000 degrees of freedom: (a) L^2-error estimator; (b) H^1-seminorm error estimator.

measured with osc_{int} defined in (29) and the oscillations at the exponential boundary layer with

$$osc_{\text{exp}} := \max_{x \geqslant 0.7}(\max\{0, u_h(x,y) - 1\}).$$

We will present results for residual-based error estimators in the H^1-semi norm and the L^2-norm, see [39]. For a detailed description of these estimators and their implementation, we refer to [23]. The gradient indicator and a residual-based error estimator in the energy norm considered in [23] failed to refine the region of the interior layer. This coincides with their behavior observed in Examples 6.4 and 6.5 of [23].

The computational results for osc_{int} and osc_{exp} are presented in Fig. 13 and the final grids for both error estimators in Fig. 14. The meshes match the expectations on the error estimators since the regions of all layers are refined and a deeper refinement occurs at the exponential boundary layers. The graphs in Fig. 13 show that the adaptive refinement of the layer regions neither reduces the spurious oscillations at the interior layer nor at the boundary layers. The adaptively refined meshes are still too coarse in these regions to resolve the layers and to suppress the oscillations.

This section showed exemplarily that a suppression of spurious oscillations cannot be achieved with adaptively refined grids whose elements do not resolve the layers.

7. Conclusions

This paper studied in detail SOLD methods which were identified in Part I as the best ones. In particular, the limits of the available methods were demonstrated. Analytical and numerical studies showed that SOLD methods without user-chosen parameters are in general not able to remove the spurious oscillations of the solution obtained with the SUPG discretization. For the two studied methods involving a parameter, the modified method of Codina (8), (11) and the edge stabilization (16), values of the parameter could be derived in two examples such that the spurious oscillations were almost removed. It turned out that a spatially constant choice of the parameters was not sufficient in general and that the optimal parameters depended on the data of the problem and on the

grid. In addition, an example was presented for which none of the investigated methods provided a qualitatively correct discrete solution.

The iterative solution of the nonlinear discrete problems was also studied. The number of iterations or the convergence of the iterative process depended again on the problem, the grid and the parameters of the SOLD methods. In particular, the convergence of the nonlinear iterative process for the edge stabilization (16) proved to be rather sensitive to this parameter. It could be observed that the convergence is often strongly influenced by the choice of an appropriate damping factor and a strategy was proposed for an automatic and dynamic computation of this factor.

Finally, it was demonstrated that adaptive grid refinement generally does not lead to a suppression of the spurious oscillations of the solutions computed with the SUPG discretization.

Considering the reduction of the spurious oscillations, the sharpness of the layers and the computational overhead for solving the nonlinear discrete problem, the SOLD methods involving parameters, i.e., the modified method of Codina (8), (11) and the edge stabilization method (16), seem to be the only reasonably promising approaches among the studied SOLD methods. However, the appropriate definition of the generally non-constant parameters in these methods will represent a great difficulty in more complicated problems and in applications. Future research should develop an a posteriori algorithm for an automatic choice of these parameters.

The current situation can be summarized as follows: it is in general completely open how to obtain oscillation-free solutions using the considered classes of methods.

Acknowledgments

The research of Petr Knobloch is a part of the Project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Academy of Sciences of the Czech Republic under the Grant No. IAA100190505.

References

[1] R.C. Almeida, R.S. Silva, A stable Petrov–Galerkin method for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 140 (1997) 291–304.
[2] Y. Bazilevs, V.M. Calo, T.E. Tezduyar, T.J.R. Hughes, $YZ\beta$ discontinuity capturing for advection-dominated processes with application to arterial drug delivery, Int. J. Numer. Methods Fluids 54 (2007) 593–608.
[3] A.N. Brooks, 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.
[4] E. Burman, A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation, Comput. Methods Appl. Mech. Engrg. 191 (2002) 3833–3855.
[5] E. Burman, A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence, Math. Comput. 74 (2005) 1637–1652.
[6] E. Burman, P. Hansbo, Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems, Comput. Methods Appl. Mech. Engrg. 193 (2004) 1437–1453.
[7] E.G.D. do Carmo, G.B. Alvarez, A new stabilized finite element formulation for scalar convection–diffusion problems: the streamline and approximate upwind/Petrov–Galerkin method, Comput. Methods Appl. Mech. Engrg. 192 (2003) 3379–3396.
[8] E.G.D. do Carmo, A.C. Galeão, Feedback Petrov–Galerkin methods for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 88 (1991) 1–16.
[9] P.G. Ciarlet, Basic error estimates for elliptic problems, in: P.G. Ciarlet, J.L. Lions (Eds.), Handbook of Numerical Analysis, Finite Element Methods, vol. 2 (pt. 1), North-Holland, Amsterdam, 1991, pp. 17–351.
[10] 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.
[11] R. Codina, Comparison of some finite element methods for solving the diffusion–convection–reaction equation, Comput. Methods Appl. Mech. Engrg. 156 (1998) 185–210.
[12] R. Codina, O. Soto, Finite element implementation of two-equation and algebraic stress turbulence models for steady incompressible flows, Int. J. Numer. Meth. Fluids 30 (1999) 309–333.
[13] T.A. Davis, Algorithm 832: UMFPACK V4.3 – an unsymmetric-pattern multifrontal method, ACM Trans. Math. Software 30 (2004) 196–199.
[14] J. Donéa, A Taylor–Galerkin method for convection transport problems, Int. J. Numer. Methods Engrg. 20 (1984) 101–119.
[15] J. Douglas, T.F. Russell, Numerical methods for convection dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures, SIAM J. Numer. Anal. 19 (1982) 871–885.
[16] L.P. Franca, S.L. Frey, T.J.R. Hughes, Stabilized finite element methods. I: Application to the advective–diffusive model, Comput. Methods Appl. Mech. Engrg. 95 (1992) 253–276.
[17] A.C. Galeão, E.G.D. do Carmo, A consistent approximate upwind Petrov–Galerkin method for convection-dominated problems, Comput. Methods Appl. Mech. Engrg. 68 (1988) 83–95.
[18] P. Grisvard, Elliptic Problems in Nonsmooth Domains, Pitman, 1985.
[19] P.W. Hemker, A singularly perturbed model problem for numerical computation, J. Comput. Appl. Math. 76 (1996) 277–285.
[20] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995) 387–401.
[21] T.J.R. Hughes, L.P. Franca, 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.
[22] T.J.R. Hughes, M. Mallet, A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Methods Appl. Mech. Engrg. 54 (1986) 341–355.
[23] V. John, A numerical study of a posteriori error estimators for convection–diffusion equations, Comput. Methods Appl. Mech. Engrg. 190 (2000) 757–781.
[24] V. John, P. Knobloch, On discontinuity-capturing methods for convection–diffusion equations, in: A. Bermúdez de Castro, D. Gómez, P. Quintela, P. Salgado (Eds.), Numerical Mathematics and Advanced Applications, Proceedings of ENUMATH 2005, Springer-Verlag, Berlin, 2006, pp. 336–344.
[25] V. John, P. Knobloch, A computational comparison of methods diminishing spurious oscillations in finite element solutions of convection–diffusion equations, in: J. Chleboun, K. Segeth, T. Vejchodský (Eds.), Proceedings of the International Conference

Programs and Algorithms of Numerical Mathematics, vol. 13, Academy of Science of the Czech Republic, pp. 122–136.
[26] V. John, P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review, Comput. Methods Appl. Mech. Engrg. 196 (2007) 2197–2215.
[27] V. John, P. Knobloch, On the performance of SOLD methods for convection–diffusion problems with interior layers, Int. J. Comput. Sci. Math. 1 (2007) 245–258.
[28] V. John, G. Matthies, MooNMD – A program package based on mapped finite element methods, Comput. Visual. Sci. 6 (2004) 163–170.
[29] V. John, M. Roland, T. Mitkova, K. Sundmacher, L. Tobiska, A. Voigt, Simulations of population balance systems with one internal coordinate using finite element methods, Technical Report, Universität des Saarlandes, FR. 6.1 – Mathematik, 2007.
[30] C. Johnson, Adaptive finite element methods for diffusion and convection problems, Comput. Methods Appl. Mech. Engrg. 82 (1990) 301–322.
[31] C. Johnson, A new approach to algorithms for convection problems which are based on exact transport + projection, Comput. Methods Appl. Mech. Engrg. 100 (1992) 45–62.
[32] C. Johnson, A.H. Schatz, L.B. Wahlbin, Crosswind smear and pointwise errors in streamline diffusion finite element methods, Math. Comput. 49 (1987) 25–38.
[33] P. Knobloch, Improvements of the Mizukami–Hughes method for convection–diffusion equations, Comput. Methods Appl. Mech. Engrg. 196 (2006) 579–594.
[34] T. Knopp, G. Lube, G. Rapin, Stabilized finite element methods with shock capturing for advection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 191 (2002) 2997–3013.
[35] A. Mizukami, 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.
[36] B. Mohammadi, O. Pironneau, Analysis of the K–Epsilon Turbulence Model, John Wiley & Sons, 1994.
[37] T.E. Tezduyar, Finite element methods for fluid dynamics with moving boundaries and interfaces, in: E. Stein, R. De Borst, T.J.R. Hughes (Eds.), Encyclopedia of Computational Mechanics, Fluids, vol. 3, Wiley, New York, 2004 (Chapter 17).
[38] T.E. Tezduyar, Y.J. Park, Discontinuity-capturing finite element formulations for nonlinear convection–diffusion–reaction equations, Comput. Methods Appl. Mech. Engrg. 59 (1986) 307–325.
[39] R. Verfürth, A Review of a Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques, Wiley and Teubner, 1996.

Int. J. Computing Science and Mathematics, Vol. 1, Nos. 2/3/4, 2007 245

On the performance of SOLD methods for convection–diffusion problems with interior layers

V. John

Universität des Saarlandes,
Fachbereich 6.1 – Mathematik,
Postfach 15 11 50,
66041 Saarbrücken, Germany
E-mail: john@math.uni-sb.de

P. Knobloch*

Charles University,
Faculty of Mathematics and Physics,
Department of Numerical Mathematics,
Sokolovská 83, 18675 Praha 8, Czech Republic
E-mail: knobloch@karlin.mff.cuni.cz
*Corresponding author

Abstract: Numerical solutions of convection–diffusion equations obtained using the Streamline–Upwind Petrov–Galerkin (SUPG) stabilisation typically possess spurious oscillations at layers. Spurious Oscillations at Layers Diminishing (SOLD) methods aim to suppress or at least diminish these oscillations without smearing the layers extensively. In the recent review by John and Knobloch (2007), numerical studies at convection–diffusion problems with constant convection whose solutions have boundary layers led to a pre-selection of the best available SOLD methods with respect to the two goals stated above. The behaviour of these methods is studied in this paper for a convection–diffusion problem with a non-constant convection field whose solution possesses an interior layer.

Keywords: convection–diffusion equations; streamline–upwind Petrov–Galerkin (SUPG) method; spurious oscillations at layers diminishing (SOLD) methods; interior layers.

Reference to this paper should be made as follows: John, V. and Knobloch, P. (2007) 'On the performance of SOLD methods for convection–diffusion problems with interior layers', *Int. J. Computing Science and Mathematics*, Vol. 1, Nos. 2/3/4, pp.245–258.

Biographical notes: Volker John is a Professor of Applied Mathematics at the University of the Saarland. He received his Diploma Degree in 1992 from the University of Halle-Wittenberg (Germany), his PhD Degree in 1997 and his habilitation in 2002 from the Otto-von-Guericke-University Magdeburg (Germany). In 2005 he moved to the University of the Saarland. His research interests include finite element methods in CFD, in particular for turbulent flows, convection-diffusion equations and coupled systems.

Petr Knobloch received his MS Degree from the Charles University in Prague (Czech Republic) in 1993 and his PhD Degree from the Otto-von-Guericke-University Magdeburg (Germany) in 1996. He is currently an Associate Professor at the Department of Numerical Mathematics of the Faculty of Mathematics and Physics at the Charles University in Prague. His research interests are the finite element method in general and numerical solution of fluid flow and singularly perturbed problems in particular.

1 Introduction

Scalar convection–diffusion equations

$$\begin{aligned} -\varepsilon\Delta u + \mathbf{b}\cdot\nabla u &= f \quad \text{in } \Omega, \\ u &= g_D \quad \text{on } \partial\Omega_D, \\ \varepsilon\frac{\partial u}{\partial \mathbf{n}} &= g_N \quad \text{on } \partial\Omega_N, \end{aligned} \tag{1}$$

describe the stationary distribution of a quantity u, like concentration or temperature, determined by the physical mechanisms of convection and diffusion. In equation (1), $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, is a bounded domain with a polygonal or polyhedral boundary $\partial\Omega$ with subsets $\partial\Omega_D$ and $\partial\Omega_N$ satisfying $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ and $\partial\Omega_D \cap \partial\Omega_N = \emptyset$. Further, $\varepsilon \in \mathbb{R}_+$ is a constant diffusion coefficient, $\mathbf{b} \in W^{1,\infty}(\Omega)^d$ is a given convection field satisfying the incompressibility condition $\nabla \cdot \mathbf{b} = 0$, $f \in L^2(\Omega)$ is an outer source of the quantity u, $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$, $g_D \in H^{1/2}(\partial\Omega_D)$ represents Dirichlet boundary conditions and $g_N \in H^{-1/2}(\partial\Omega_N)$ Neumann boundary conditions. The solution of equation (1) is sought in $H^1(\Omega)$.

The interesting case from the practical as well as from the numerical point of view is the convection-dominated one, i.e., $\varepsilon \ll \| \mathbf{b} \|_{L^\infty(\Omega)}$. In this case, the solution of equation (1) typically possesses layers. These are regions where the solution still is continuous but has very large gradients. The width of the layers is in general much smaller than the available mesh width in numerical simulations. Consequently, the layers cannot be resolved. It turns out that standard discretisation approaches, like the Galerkin Finite Element Method (FEM), even lead to solutions that are globally polluted by spurious (unphysical) oscillations.

A dramatical enhancement of the quality of numerical solutions is obtained with stabilised discretisations. In the context of FEM, there are several approaches like upwind techniques (Tabata, 1977), the Streamline–Upwind Petrov–Galerkin (SUPG) method (Brooks and Hughes, 1982), also called Streamline-Diffusion Finite Element Method (SDFEM), or the Galerkin/least-squares method (Hughes et al., 1989), see Roos et al. (1996) for an overview. The most popular one is probably the SUPG method, which will be also considered in this paper.

Applying the SUPG stabilisation, the numerical solutions capture the position of the layers in general quite well and the layers are not smeared. However, spurious oscillations of sometimes considerable magnitude usually appear at the layers. These oscillations are intolerable from the physical point of view since they describe,

for instance, negative concentrations. From the numerical point of view, these oscillations might lead to instabilities in the simulation of coupled systems which involve equations of form (1), for instance if, due to negative spurious oscillations, reactive terms of product form change locally their signs in coupled, non-linear reaction–convection–diffusion equations. Thus, there is an urgent need to remove these spurious oscillations, however, without smearing the layers extensively.

The development of numerical schemes for removing (or at least diminishing) the spurious oscillations of SUPG solutions of equation (1) started around two decades ago. Since then, a number of different approaches have been published, see, e.g., the recent review by John and Knobloch (2007). These schemes were called often *shock capturing methods* or *discontinuity capturing methods;* however, these names do not reflect their real purpose. Therefore, the name spurious oscillations at layers diminishing (SOLD) methods was introduced by John and Knobloch (2007) and this name will be used in this paper, too.

In the review paper by John and Knobloch (2007), numerical tests with constant convection fields and P_1 finite elements are presented to compare most of the published SOLD methods and to obtain a pre-selection of methods that should be studied in detail. The methods were evaluated by means of various criteria, which measure the amount of spurious oscillations and the smearing of the layers in the discrete solution. Thus, if we speak about 'best methods' in the present paper, we always mean with respect to those criteria (if we refer to John and Knobloch (2007)) or with respect to the criteria formulated below. We believe that this procedure is necessary before one should study the error of the discrete solution measured in various norms since such a study makes only sense for methods which substantially reduce the spurious oscillations without an extensive smearing of layers (other methods are not useful in applications). However, based on all our experiences, there is still no method fulfilling this requirement (save (Mizukami and Hughes, 1985) in special cases, see below) and therefore a study of approximation errors is not yet an issue. In our opinion, there are no relations between the measures for evaluating the size of the oscillations and norms in which the approximation error is bounded.

The aim of this paper is to present numerical studies for the best SOLD methods from John and Knobloch (2007) for a problem without boundary layers but with an interior layer created by a non-constant convection field. For such a problem, the localised spurious oscillations of the SUPG solution cannot be significantly influenced by the choice of the stabilisation parameter τ (see below) since the SUPG method does not contain any mechanism for stabilisation perpendicular to streamlines. Let us mention that we do not use layer-adapted meshes (like Bakhvalov or Shishkin type meshes) since our aim is to find methods that can be used in applications, which means in situations where the features of the solution (and hence a layer-adapted mesh) are not known a priori.

The plan of the paper is as follows. In the next section, we formulate the SUPG method and, in Section 3, we review SOLD methods, which were identified as the best ones by John and Knobloch (2007). Then, in Section 4, we present results of our numerical studies. In contrast to John and Knobloch (2007), the Q_1 finite element is used besides the P_1 finite element. The paper ends with our conclusions in Section 5.

2 SUPG method

The SUPG method adds an additional term to the Galerkin FEM to control the derivatives in streamline direction. Let the space $H^1(\Omega)$ in which the solution of equation (1) is sought be approximated by a conforming finite element subspace V_h defined on an admissible triangulation $\mathcal{T}_h$ (Ciarlet, 1991) with elements (mesh cells) K. We introduce a function $g_{D,h} \in V_h$ such that $g_{D,h} \mid_{\partial\Omega_D}$ approximates g_D. Further, we set $V_{0,h} = \{v \in V_h : v \mid_{\partial\Omega_D} = 0\}$. Then, the SUPG method reads as follows: Find $u_h \in V_h$ such that $u_h - g_{D,h} \in V_{0,h}$ and

$$\begin{aligned}(\varepsilon\nabla u_h, \nabla v_h) + (\mathbf{b}\cdot\nabla u_h, v_h) + \sum_{K\in\mathcal{T}_h} (R_h(u_h), \tau\mathbf{b}\cdot\nabla v_h)_K \\ = (f, v_h) + \int_{\partial\Omega_N} g_N v_h \, \mathrm{d}s \quad \forall\, v_h \in V_{0,h},\end{aligned} \tag{2}$$

where $(\cdot,\cdot)_K$ denotes the inner product in $L^2(K)$ or $L^2(K)^d$, $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$,

$$R_h(u_h)|_K = -\varepsilon\Delta(u_h|_K) + (\mathbf{b}\cdot\nabla u_h - f)|_K$$

and $\tau \in L^\infty(\Omega)$ is a non-negative stabilisation parameter. There are several approaches for choosing τ, see John and Knobloch (2007), which lead asymptotically to optimal error estimates. However, they may lead to very different results for a concrete situation and the optimal choice of τ is an open question. We will use in the simulations presented in this paper the following definition:

$$\tau|_K(\mathbf{x}) = \frac{h_{K,\mathbf{b}}(\mathbf{x})}{2\,|\mathbf{b}(\mathbf{x})|}\xi(\mathrm{Pe}_K(\mathbf{x})) \tag{3}$$

with the local Péclet number

$$\mathrm{Pe}_K(\mathbf{x}) = \frac{|\mathbf{b}(\mathbf{x})|\, h_{K,\mathbf{b}}(\mathbf{x})}{2\varepsilon},$$

the upwind function $\xi(\alpha) = \coth(\alpha) - \alpha^{-1}$, $|\mathbf{b}(\mathbf{x})|$ the Euclidean norm of the convection vector in $\mathbf{x} \in K$ and $h_{K,b}(\mathbf{x})$ the diameter of the element K in the direction of $\mathbf{b}(\mathbf{x})$, see John and Knobloch (2007) for a detailed discussion of these choices.

3 SOLD methods

The most SOLD methods considered in the review by John and Knobloch (2007) are defined by adding an artificial diffusion term to the SUPG discretisation (2). The review by John and Knobloch (2007) categorises the available SOLD methods into the following classes:

- SOLD methods adding isotropic artificial diffusion
- SOLD methods adding crosswind artificial diffusion
- SOLD methods based on edge stabilisations
- SOLD methods that are not based on the SUPG method.

Note that the additional terms lead generally to non-linear discrete equations. Below, we formulate the SOLD method(s) from each class, which are the best ones according to the tests and criteria in John and Knobloch (2007) (and also according to further numerical studies we have performed).

SOLD methods adding isotropic artificial diffusion add the term

$$\sum_{K\in\mathcal{T}_h} (\tilde{\varepsilon}\nabla u_h, \nabla v_h)_K$$

to the left-hand side of equation (2). Among the schemes reviewed in John and Knobloch (2007), the best method of this type seems to be that one proposed by do Carmo and Galeão (1991) (dCG91) in which

$$\tilde{\varepsilon} = \tau \max\left\{0, \frac{|\mathbf{b}||R_h(u_h)|}{|\nabla u_h|} - \frac{|R_h(u_h)|^2}{|\nabla u_h|^2}\right\}$$

(we set $\tilde{\varepsilon} = 0$ if $\nabla u_h = 0$). Here, τ is the same as in equation (3).

SOLD methods adding crosswind diffusion introduce an extra term of the form

$$\sum_{K\in\mathcal{T}_h} (\tilde{\varepsilon}D\nabla u_h, \nabla v_h)_K$$

with

$$D = \begin{cases} I - \dfrac{\mathbf{b}\otimes\mathbf{b}}{|\mathbf{b}|^2} & \text{if } \mathbf{b} \neq \mathbf{0} \\ 0 & \text{else} \end{cases}$$

into the SUPG formulation (2). The best results in this class of SOLD methods in John and Knobloch (2007) were obtained with modifications proposed by John and Knobloch (2007) of a parameter suggested by Codina (1993) (C93) and of a parameter by Burman and Ern (2002) (BE02). The parameter of the method C93 is

$$\tilde{\varepsilon}|_K = \max\left\{0, C\frac{h_K|R_h(u_h)|}{2|\nabla u_h|} - \varepsilon\right\}$$

($\tilde{\varepsilon} = 0$ if $\nabla u_h = 0$), where C is a user-chosen parameter and h_K is the diameter of the element K. In the numerical studies in Section 4, the parameter $C = 0.6$ will be used, which is the same value as in John and Knobloch (2007). If $f = 0$ and $\Delta(u_h|_K) = 0$ for any $K \in \mathcal{T}_h$ (which will be the case in Section 4), the above definition of $\tilde{\varepsilon}$ is identical with the original method of Codina (1993). The parameter of BE02 has the form

$$\tilde{\varepsilon}|_K = \frac{\tau|\mathbf{b}|^2|R_h(u_h)|}{|\mathbf{b}||\nabla u_h| + |R_h(u_h)|}.$$

Edge stabilisation methods for linear simplicial finite elements add to the left-hand side of equation (2) the term

$$\sum_{K\in\mathcal{T}_h} \int_{\partial K} \Psi_K(u_h)\operatorname{sign}\left(\frac{\partial u_h}{\partial \mathbf{t}_{\partial K}}\right)\frac{\partial v_h}{\partial \mathbf{t}_{\partial K}}\,\mathrm{d}s,$$

where $\mathbf{t}_{\partial K}$ is a tangential vector on the boundary ∂K of K. The best edge stabilisation method in the numerical studies of John and Knobloch (2007) was proposed by Burman and Ern (2005) (BE05). It has the parameter function

$$\Psi_K(u_h) = C\,|\,R(u_h)|_K|.$$

The same parameter $C = 5 \times 10^{-5}$ as in John and Knobloch (2007) was chosen for the numerical studies presented below.

From the approaches that do not rely on the SUPG method, we will consider an upwind scheme, which was developed by Mizukami and Hughes (1985) (MH85) and recently improved by Knobloch (2006). This upwind scheme is defined only for linear simplicial finite elements and is based on a rather involved geometrical construction, see Knobloch (2006) for details.

For further properties of the SOLD methods, we refer to John and Knobloch (2007).

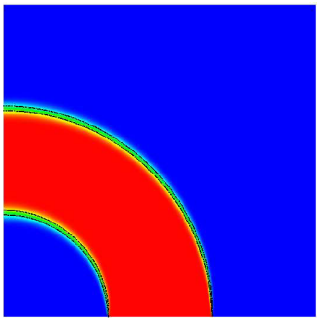
4 Numerical studies

We will study (1) with $\Omega = (0, 1)^2$, $\partial\Omega_N = \{0\} \times (0, 1)$, $f = 0$ and $\mathbf{b}(x, y) = (-y, x)^T$. On the outflow boundary $\partial\Omega_N$, homogeneous conditions $g_N = 0$ are prescribed. The Dirichlet data are discontinuous

$$g_D(x,y) = \begin{cases} 1 & \text{if } (x,y) \in (1/3, 2/3) \times \{0\}, \\ 0 & \text{else on } \partial\Omega_D. \end{cases}$$

The discontinuous Dirichlet boundary condition on $(0, 1) \times \{0\}$ is transported counter-clockwise to the outflow boundary, see Figure 1. The width of the layers, for instance on the outflow boundary, depends on the size of ε. This example was already studied by Knopp et al. (2002). The solution u of the continuous problem does not belong to $H^1(\Omega)$ but, due to the positive diffusion, u is smooth in Ω. Moreover, it is easy to smooth g_D to a function from $H^{1/2}(\partial\Omega_D)$ (which leads to $u \in H^1(\Omega)$) in such a way that the numerical results presented in this paper do not change. We will study the cases of a moderate local Péclet number and of a high local Péclet number. Similarly as in John and Knobloch (2007), the SOLD methods will be evaluated only on measures for the amount of spurious oscillations and layer smearing. The results have been double checked with two different codes, one of them was MooNMD (John and Matthies, 2004).

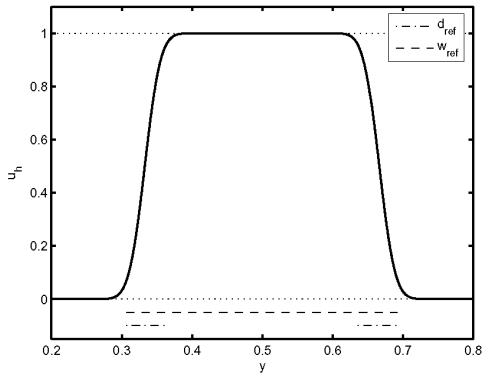
Figure 1 Solution u for $\varepsilon = 10^{-4}$, blue (dark) part is zero, red (light) part is one



4.1 Moderate local mesh Péclet numbers

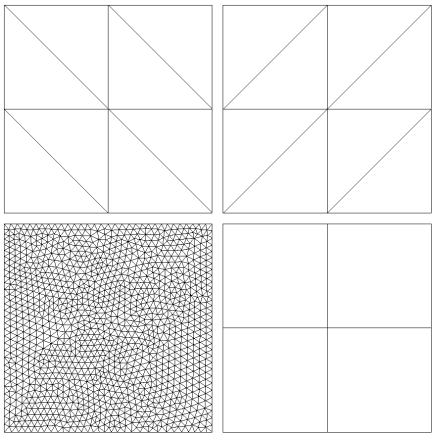
First, we will present computations for $\varepsilon = 10^{-4}$. For this diffusion parameter, we computed a reference solution with the Galerkin FEM (P_2 FEM, 16 785 409 degrees of freedom (dof), $h_K = \sqrt{2}/2048$), see Figure 2, which will be used to evaluate the SOLD methods.

Figure 2 Reference curve for $\varepsilon = 10^{-4}$ on the outflow boundary



The initial regular grids and the unstructured triangular grid are presented in Figure 3. Refining the regular grids till the legs of the triangles or the edges of the squares have the length 1/32 leads to 1089 dof (including Dirichlet nodes). The unstructured grid (Grid 3) has 1244 nodes and was obtained using the anisotropic mesh adaptation technique of Dolejší (1998). The P_1 finite element was used on the simplicial grids (Grid 1–Grid 3) and the Q_1 finite element on the grid consisting of squares (Grid 4). The integrals in the discrete problem were evaluated using quadrature rules which are exact for polynomials of degree 8 (triangles) and 9 (squares).

Figure 3 The grids used in the computations: Grid 1, Grid 2, Grid 3 and Grid 4 (left to right, top to bottom). The structured grids are refined till the length of the legs of the triangles (edges of the squares) is 1/32 in the moderate local Péclet number case and 1/64 in the high local Péclet number case



Since the convection field is not constant, the local mesh Péclet numbers vary in Ω. The lowest Péclet number is on all grids zero (at the corner (0, 0)). The largest Péclet number is around 150 on Grid 2, 240 on Grid 3 and about 300 on Grid 1 and Grid 4. Note that the local mesh Péclet numbers in the regions with the layers are smaller.

Let us denote by $\mathcal{N}_h$ the set of nodes of the triangulation $\mathcal{T}_h$. The measures for evaluating the numerical results are:

- $\text{min} := \left| \min_{(x,y)\in\mathcal{N}_h} u_h(x,y) \right|,$
- $\text{max} := \max_{(x,y)\in\mathcal{N}_h} u_h(x,y) - 1,$
- $\text{min2} := \left(\sum_{(x,y)\in\mathcal{N}_h} (\min\{0, u_h(x,y)\})^2 \right)^{1/2},$
- $\text{max2} := \left(\sum_{(x,y)\in\mathcal{N}_h} (\max\{0, u_h(x,y)-1\})^2 \right)^{1/2},$
- $\text{mino} := \left| \min_{(x,y)\in\mathcal{N}_h\cap\overline{\partial\Omega_N}} u_h(x,y) \right|,$
- $\text{maxo} := \max\{u_h(0,y) - u_h(0,z) : y_0 \le y \le z \le y_{1/2} \text{ or } y_{1/2} \le z \le y \le y_1\},$

  where, denoting by $[\overline{y}_0, \overline{y}_1]$ the interval on the outflow boundary with $u_h(0, y) \ge 0.1$, the point y_0 is such that $\partial_y u_h > 0$ a.e. on $[\overline{y}_0, y_0]$ and $\partial_y u_h(y_0 +) \le 0$. Similarly, $\partial_y u_h < 0$ a.e. on $[y_1, \overline{y}_1]$ and $\partial_y u_h(y_1 -) \ge 0$. Finally, $(0, y_{1/2})$ is the nearest node to $(0, (y_0 + y_1)/2)$. Note that u is non-decreasing on $[\overline{y}_0, 1/2]$ and non-increasing on $[1/2, \overline{y}_1]$ and hence maxo tries to find the largest violation of these monotonicities.

- $smear := (d - d_{\text{ref}})/d_{\text{ref}},$

  where d is the sum of the lengths of the two intervals on the outflow boundary with $u_h(0, y) \in [0.1, 0.9]$ and $d_{\text{ref}} = 0.114518$ is the value of d for the interpolation of the reference curve on an equidistant grid with mesh width 1/32 (cf. Figure 2),

- $width := (w - w_{\text{ref}})/w_{\text{ref}},$

  where w is the length of the interval on the outflow boundary with $u_h(0, y) \ge 0.1$ and $w_{\text{ref}} = 0.385697$ is the value of w for the interpolation of the reference curve on an equidistant grid with mesh width 1/32.

The measures min and max quantify the size of the largest undershoot or overshoot, respectively. An average value for the undershoots and overshoots is obtained with min2 and max2. The oscillations on the outflow boundary are measured with mino and maxo. The values for smear and width describe the smearing of the layers on the outflow boundary.

The results for the moderate Péclet number case are given in Tables 1–4. It can be seen that all SOLD methods considerably reduce the spurious oscillations of the SUPG solution. However, they also increase the smearing of the layers. Concerning the reduction of the oscillations, the best results are obtained with MH85 on

the simplicial meshes and with dCG91 and BE02 on the quadrilateral mesh. The second best method on the simplicial grids was C93. The edge stabilisation method BE05 gives quite poor results on the regular triangular grids. It becomes much better in comparison with the other methods on the unstructured mesh. The good performance of edge stabilisation methods on unstructured grids was observed already by John and Knobloch (2007). It is noteworthy that the results on Grid 2 are worse than on Grid 1 although the local Péclet numbers are smaller on Grid 2. This shows that the orientation of the edges in Grid 1 is better suited to the direction of the convection in this example. Let us mention that, in all computations, the difference of $u_h(0, 0.5)$ to 1 was less than 1%.

Even in the small local Péclet number case, there is no method that worked satisfactorily in all respects.

Table 1 Results for the computations with moderate local mesh Péclet number, Grid 1

	min	*max*	*min2*	*max2*	*mino*	*maxo*	*smear*	*width*
SUPG	1.068 e–1	8.374 e–2	3.137 e–1	2.544 e–1	3.183 e–2	3.933 e–2	2.225 e–1	7.338 e–2
MH85	2.131 e–12	0.000 e+0	5.189 e–12	0.000 e+0	0.000 e+0	0.000 e+0	6.242 e–1	1.330 e–1
dCG91	8.489 e–3	5.589 e–3	9.316 e–3	8.317 e–3	4.150 e–6	0.000 e+0	8.565 e–1	1.672 e–1
C93	1.247 e–4	3.136 e–4	1.266 e–4	3.209 e–4	0.000 e+0	0.000 e+0	6.591 e–1	1.398 e–1
BE02	3.761 e–3	2.663 e–3	3.915 e–3	3.005 e–3	0.000 e+0	0.000 e+0	8.713 e–1	1.693 e–1
BE05	2.544 e–2	1.604 e–2	6.243 e–2	4.403 e–2	4.367 e–3	7.453 e–3	4.504 e–1	1.084 e–1

Table 2 Results for the computations with moderate local mesh Péclet number, Grid 2

	min	*max*	*min2*	*max2*	*mino*	*maxo*	*smear*	*width*
SUPG	1.242 e–1	1.020 e–1	4.808 e–1	4.229 e–1	4.833 e–2	5.879 e–2	6.661 e–1	1.390 e–1
MH85	7.416 e–13	0.000 e+0	1.888 e–12	0.000 e+0	0.000 e+0	0.000 e+0	1.570 e+0	2.769 e–1
dCG91	1.972 e–2	1.833 e–2	7.519 e–2	8.377 e–2	1.175 e–2	0.000 e+0	1.298 e+0	2.334 e–1
C93	5.792 e–3	2.856 e–3	1.297 e–2	7.815 e–3	1.651 e–3	0.000 e+0	1.464 e+0	2.569 e–1
BE02	1.167 e–2	1.333 e–2	3.930 e–2	4.528 e–2	6.099 e–3	0.000 e+0	1.309 e+0	2.354 e–1
BE05	4.335 e–2	3.003 e–2	1.047 e–1	7.845 e–2	7.449 e–3	0.000 e+0	1.263 e+0	2.279 e–1

Table 3 Results for the computations with moderate local mesh Péclet number, Grid 3

	min	*max*	*min2*	*max2*	*mino*	*maxo*	*smear*	*width*
SUPG	7.204 e–2	9.425 e–2	3.452 e–1	3.639 e–1	4.712 e–2	4.682 e–2	2.705 e–1	8.894 e–2
MH85	1.528 e–12	0.000 e+0	3.802 e–12	0.000 e+0	0.000 e+0	0.000 e+0	9.717 e–1	1.877 e–1
dCG91	5.408 e–2	3.589 e–2	9.850 e–2	7.845 e–2	4.952 e–3	0.000 e+0	9.097 e–1	1.877 e–1
C93	2.754 e–2	3.273 e–2	5.494 e–2	5.519 e–2	2.188 e–3	0.000 e+0	6.914 e–1	1.553 e–1
BE02	4.047 e–2	2.966 e–2	6.832 e–2	5.479 e–2	1.688 e–3	0.000 e+0	9.455 e–1	1.924 e–1
BE05	3.111 e–2	2.905 e–2	7.437 e–2	6.821 e–2	6.212 e–3	6.071 e–3	7.072 e–1	1.569 e–1

Table 4 Results for the computations with moderate local mesh Péclet number, Grid 4

	min	*max*	*min2*	*max2*	*mino*	*maxo*	*smear*	*width*
SUPG	1.316 e–1	1.026 e–1	4.205 e–1	3.311 e–1	4.235 e–2	5.191 e–2	3.544 e–1	9.283 e–2
dCG91	1.244 e–2	7.865 e–3	2.381 e–2	1.477 e–2	4.049 e–4	0.000 e+0	9.936 e–1	1.906 e–1
C93	3.311 e–2	2.908 e–2	5.992 e–2	4.414 e–2	1.299 e–5	0.000 e+0	7.823 e–1	1.535 e–1
BE02	1.351 e–2	8.321 e–3	2.267 e–2	1.180 e–2	5.280 e–8	0.000 e+0	1.022 e+0	1.958 e–1

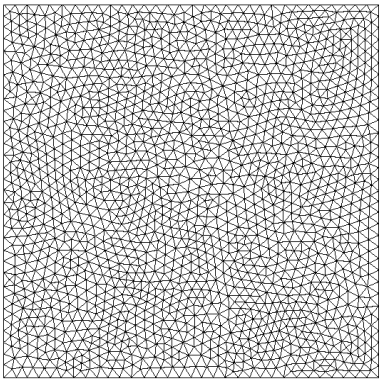
Remark 1: *The results in Tables 1–4 also show that the quality of the solution on the outflow boundary is often better than of the solution inside* Ω *and hence the outflow profile cannot be used as the only measure for an assessment of the considered numerical methods. For the Galerkin discretisation with small* ε*, it can even happen that the inflow profile is almost exactly reproduced on the outflow boundary whereas the solution wildly oscillates inside* Ω*.*

4.2 High local mesh Péclet numbers

We consider the same example as before, however, with the diffusion parameter $\varepsilon = 10^{-8}$. The regular Grids 1, 2 and 4 from Figure 3 are used with edges (legs of the triangles) of length 1/64. The number of degrees of freedom for the P_1, resp. Q_1, discretisation is 4225 (including Dirichlet nodes). The unstructured grid, which was used in the high local mesh Péclet number computations has 1721 nodes and is presented in Figure 4. The largest local mesh Péclet numbers are around 7.7×10^5 for Grid 2, 1.5×10^6 for Grid 1 and Grid 4 and 2.1×10^6 for Grid 5.

Concerning the oscillations, the same measures are used as in the moderate local Péclet number case. The smearing of the layers will be evaluated by means of graphs of the discrete solutions on the outflow boundary. Thus, the measures do not need a reference solution.

Figure 4 Unstructured Grid 5 for the high local Péclet number case



The results concerning the spurious oscillations are collected in Tables 5–8. All SOLD methods give again much better results than the SUPG method. Only on the regular Grids 1 and 2, the fixed-point iterations for solving the non-linear problem of BE05 did not converge (100,000 iterations). On the triangular grids, MH85 was again the best

method. Among the other methods, there is no really a best one. BE02 is slightly better than the other ones on Grid 1 and C93 on Grid 2. On the unstructured Grid 5, the edge stabilisation method BE05 is the second best after MH85. We think that the good performance of MH85 on Grid 3 (with $\varepsilon = 10^{-4}$) and Grid 5 might be caused by the fact that these grids possess only acute triangles. On the quadrilateral Grid 4, all SOLD methods give similar results. Note that the amount of the spurious oscillations is not much different in comparison with the moderate local Péclet number case.

Table 5 Results for the computations with high local mesh Péclet number, Grid 1

	min	*max*	*min2*	*max2*	*mino*	*maxo*
SUPG	1.551 e–1	1.353 e–1	6.602 e–1	5.734 e–1	5.533 e–2	7.004 e–2
MH85	1.176 e–12	6.160 e–13	4.627 e–12	1.406 e–12	3.414 e–13	0.000 e+0
dCG91	5.579 e–3	3.979 e–3	8.559 e–3	6.268 e–3	5.266 e–6	0.000 e+0
C93	5.004 e–3	1.681 e–4	7.299 e–3	1.723 e–4	2.784 e–6	1.602 e–6
BE02	2.198 e–3	1.567 e–3	2.933 e–3	2.119 e–3	5.675 e–8	0.000 e+0
BE05	No convergence					

Table 6 Results for the computations with high local mesh Péclet number, Grid 2

	min	*max*	*min2*	*max2*	*mino*	*maxo*
SUPG	1.655 e–1	1.467 e–1	8.311 e–1	7.438 e–1	6.313 e–2	6.815 e–2
MH85	1.086 e–12	5.294 e–13	3.868 e–12	1.296 e–12	2.500 e–13	0.000 e+0
dCG91	1.405 e–2	1.174 e–2	7.014 e–2	6.853 e–2	7.660 e–3	3.375 e–3
C93	3.948 e–3	1.567 e–3	9.394 e–3	3.714 e–3	4.840 e–4	6.758 e–4
BE02	7.501 e–3	8.482 e–3	3.405 e–2	3.152 e–2	3.951 e–3	9.976 e–4
BE05	No convergence					

Table 7 Results for the computations with high local mesh Péclet number, Grid 4

	min	*max*	*min2*	*max2*	*mino*	*maxo*
SUPG	1.868 e–1	1.642 e–1	8.126 e–1	6.755 e–1	6.618 e–2	6.740 e–2
dCG91	2.861 e–2	2.562 e–2	5.910 e–2	4.747 e–2	1.330 e–4	1.832 e–5
C93	3.898 e–2	3.464 e–2	7.993 e–2	6.243 e–2	7.491 e–6	2.859 e–5
BE02	2.961 e–2	2.740 e–2	5.932 e–2	4.854 e–2	5.965 e–6	1.424 e–5

Table 8 Results for the computations with high local mesh Péclet number, Grid 5

	min	*max*	*min2*	*max2*	*mino*	*maxo*
SUPG	1.156 e–1	9.801 e–2	5.591 e–1	4.947 e–1	7.953 e–2	7.639 e–2
MH85	9.530 e–13	0.000 e+0	2.839 e–12	0.000 e+0	2.029 e–13	0.000 e+0
dCG91	6.264 e–2	7.172 e–2	9.784 e–2	9.648 e–2	6.206 e–3	0.000 e+0
C93	4.470 e–2	4.347 e–2	5.771 e–2	6.197 e–2	2.936 e–4	0.000 e+0
BE02	4.974 e–2	5.225 e–2	6.785 e–2	6.691 e–2	1.995 e–3	0.000 e+0
BE05	3.528 e–2	3.122 e–2	5.767 e–2	5.429 e–2	4.108 e–3	2.622 e–3

Parts of the outflow profiles for selected methods are presented in Figures 5–8. The improvement in comparison with the SUPG solution concerning the spurious oscillations is clearly visible. Likewise, the smearing of the layers in the solutions computed with the SOLD methods can be seen. The smearing is more or less the same for all SOLD methods. Often, the curves are on top of each other. All layers (including the SUPG solution) are extremely smeared on Grid 2. This is a further hint that this grid is less suited for the present example than the other ones. The reason for the stronger smearing of the layers on Grid 5 in comparison with Grids 1 and 4 is the considerably smaller number of degrees of freedom on Grid 5.

The method MH85 practically removes the spurious oscillations in the high local Péclet number computations. The spurious oscillations of the other methods are still not negligible. In addition, all SOLD methods lead to a smearing of the layers.

Figure 5 Solution at the lower part of the outflow boundary for the high local Péclet number case, Grid 1

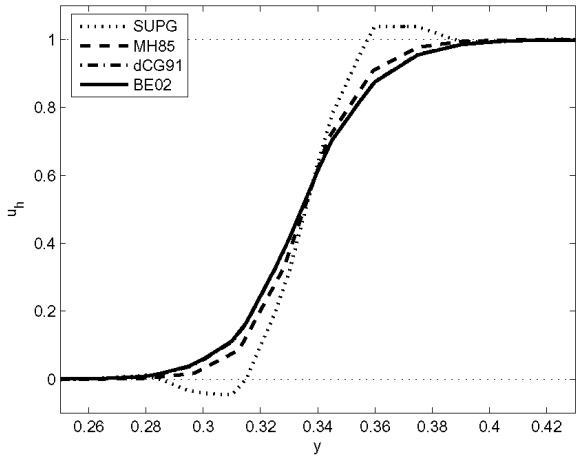


Figure 6 Solution at the lower part of the outflow boundary for the high local Péclet number case, Grid 2

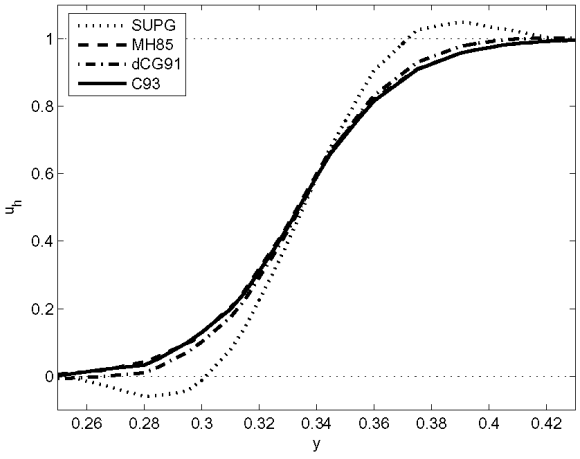


Figure 7 Solution at the lower part of the outflow boundary for the high local Péclet number case, Grid 4

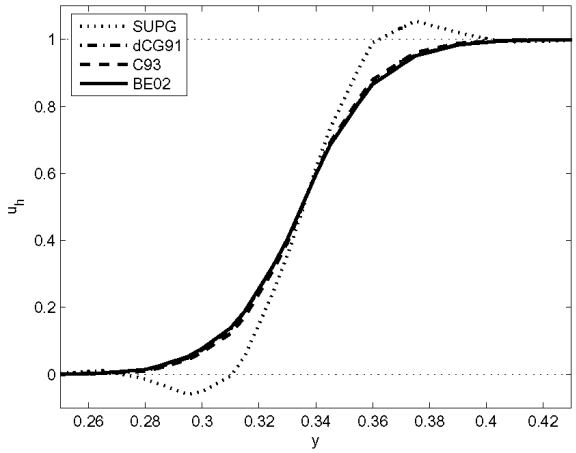
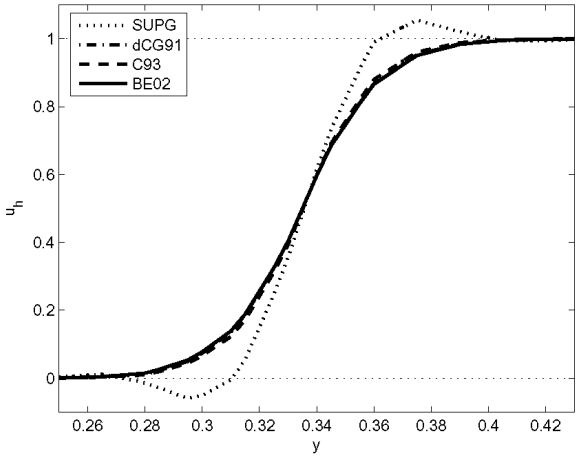


Figure 8 Solution at the lower part of the outflow boundary for the high local Péclet number case, Grid 5



5 Conclusions

The present numerical studies support an observation by John and Knobloch (2007): if the upwind method MH85 can be used, then it is the best method. The edge stabilisation method BE05 worked only properly on the unstructured grids with acute triangles. The differences among the other SOLD methods were small. On the one hand, their results are clearly better than the results of the SUPG method, but on the other hand, the remaining spurious oscillations are still not tolerable in many applications. Combining the results of John and Knobloch (2007) and the present study, one has to conclude that the SOLD methods are still far away from being able to solve convection-dominated problems successfully.

Acknowledgement

The research of Petr Knobloch is a part of the project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Academy of Sciences of the Czech Republic under the Grant No. IAA100190505.

References

Brooks, A.N. and Hughes, T.J.R. (1982) 'Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations', *Comput. Methods Appl. Mech. Eng.*, Vol. 32, pp.199–259.

Burman, E. and Ern, A. (2002) 'Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation', *Comput. Methods Appl. Mech. Eng*., Vol. 191, pp.3833–3855.

Burman, E. and Ern, A. (2005) 'Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence', *Math. Comput*., Vol. 74, pp.1637–1652.

Ciarlet, P.G. (1991) 'Basic error estimates for elliptic problems', in Ciarlet, P.G. and Lions, J.L. (Eds.): *Handbook of Numerical Analysis*, Vol. 2 – Finite Element Methods (pt. 1), North Holland, Amsterdam, pp.17–351.

Codina, R. (1993) 'A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection–diffusion equation', *Comput. Methods Appl. Mech. Eng*., Vol. 110, pp.325–342.

do Carmo, E.G.D. and Galeão, A.C. (1991) 'Feedback Petrov–Galerkin methods for convection-dominated problems', *Comput. Methods Appl. Mech. Eng*., Vol. 88, pp.1–16.

Dolejší, V. (1998) 'Anisotropic mesh adaptation for finite volume and finite element methods on triangular meshes', *Comput. Visual. Sci*., Vol. 1, pp.165–178.

Hughes, T.J.R., Franca, L.P. and Hulbert, G.M. (1989) 'A new finite element formulation for computational fluid dynamics. VIII. The Galerkin/least-squares method for advective–diffusive equations', *Comput. Methods Appl. Mech. Eng*., Vol. 73, pp.173–189.

John, V. and Knobloch, P. (2007) 'On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part I – a review', *Comput. Methods Appl. Mech. Eng*., Vol. 196, pp.2197–2215.

John, V. and Matthies, G. (2004) 'MooNMD – a program package based on mapped finite element methods', *Comput. Visual. Sci*., Vol. 6, pp.163–170.

Knobloch, P. (2006) 'Improvements of the Mizukami–Hughes method for convection–diffusion equations', *Comput. Methods Appl. Mech. Eng*., Vol. 196, pp.579–594.

Knopp, T., Lube, G. and Rapin, G. (2002) 'Stabilized finite element methods with shock capturing for advection–diffusion problems', *Comput. Methods Appl. Mech. Eng*., Vol. 191, pp.2997–3013.

Mizukami, A. and Hughes, T.J.R. (1985) 'A Petrov–Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle', *Comput. Methods Appl. Mech. Eng*., Vol. 50, pp.181–193.

Roos, H-G., Stynes, M. and Tobiska, L. (1996) *Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion and Flow Problems*, Springer-Verlag, Berlin.

Tabata, M. (1977) 'A finite element approximation corresponding to the upwind finite differencing', *Mem. Numer. Math*., Vol. 4, pp.47–63.

Chapter 4

Choice of stabilization parameters

This chapter consists of the following publications:

P. Knobloch: On the choice of the SUPG parameter at outflow boundary layers, *Advances in Computational Mathematics* 31 (4): 369–389, 2009. p. 113

P. Knobloch: On the definition of the SUPG parameter, *Electronic Transactions on Numerical Analysis* 32: 76–89, 2008. p. 134

V. John, P. Knobloch, S.B. Savescu: A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I, *Computer Methods in Applied Mechanics and Engineering* 200 (41-44): 2916–2929, 2011. p. 148

V. John, P. Knobloch: Adaptive computation of parameters in stabilized methods for convection–diffusion problems, in: A. Cangiani, R.L. Davidchack, E.H. Georgoulis, A. Gorban, J. Levesley, M.V. Tretyakov (eds.), Numerical Mathematics and Advanced Applications 2011, Proceedings of ENUMATH 2011, Springer-Verlag, Berlin, 2013, pp. 275–283. p. 163

Adv Comput Math (2009) 31:369–389
DOI 10.1007/s10444-008-9075-6

On the choice of the SUPG parameter at outflow boundary layers

Petr Knobloch

Received: 14 July 2007 / Accepted: 26 February 2008 /
Published online: 26 April 2008

Abstract We consider the Streamline upwind/Petrov–Galerkin (SUPG) finite element method for two–dimensional steady scalar convection–diffusion equations and propose a new definition of the SUPG stabilization parameter along outflow Dirichlet boundaries. Numerical results demonstrate a significant improvement of the accuracy and show that, in some cases, even nodally exact solutions are obtained.

Keywords Convection–diffusion equations · Streamline upwind/Petrov–Galerkin (SUPG) method · Spurious oscillations · Outflow boundary layers · Singularly perturbed problems

Mathematics Subject Classification (2000) 65N30

1 Introduction

This paper is devoted to the application of the finite element method to the numerical solution of a steady scalar convection–diffusion equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u = f \quad \text{in } \Omega\,. \tag{1}$$

We assume that Ω is a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\boldsymbol{b}$ is a given convective field, and f is an outer source of u. The equation (1) has to be equipped with suitable boundary conditions on $\partial\Omega$

Communicated by Martin Stynes.

P. Knobloch ( )
Faculty of Mathematics and Physics, Department of Numerical Mathematics,
Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

which will be specified later. In the convection–dominated case $\varepsilon \ll |\boldsymbol{b}|$, the solution u typically contains interior and boundary layers (which depend on the choice of the boundary conditions). These layers can be divided into characteristic (interior and boundary) layers and outflow boundary layers, see [10].

If the width of layers is smaller than the resolution of the used mesh, discrete solutions of (1) often contain spurious oscillations. The attenuation of these oscillations has been the subject of extensive research for several decades during which a huge number of so–called stabilized methods have been developed. In the context of finite element methods, a very popular stabilization technique is the streamline upwind/Petrov–Galerkin (SUPG) method introduced in [2]. The SUPG method produces accurate and oscillation–free solutions in regions where no abrupt changes in the solution of (1) occur but it does not preclude spurious oscillations localized in narrow regions along sharp layers. The magnitude of these oscillations can be influenced by the choice of the SUPG stabilization parameter and the aim of this paper is to describe a new way in which this parameter can be defined.

We shall confine ourselves to outflow boundary layers where a careful choice of the SUPG parameter can provide a fairly satisfactory approximation of the solution u. The choice of the stabilization parameter at characteristic layers has only a limited influence on the spurious oscillations appearing in these regions (cf., e.g., [8]) and hence an oscillation–free SUPG approximation of a characteristic layer can be generally obtained only by introducing an additional crosswind diffusion [6] or by using a layer–adapted mesh, see, e.g., [9].

Sometimes the question arises whether outflow boundary layers occur in real applications or only in academic problems. Such questions may originate from experiences in computational fluid dynamics (CFD) where outflow boundaries are often artificial boundaries at which no layers occur. However, also in CFD applications, outflow boundary layers may occur when problems with moving boundaries are considered. Moreover, there are many other applications leading to convection–diffusion equations whose solutions possess outflow boundary layers in the sense considered in this paper although the vector $\boldsymbol{b}$ often cannot be interpreted as convection. For example, magnetohydrodynamical pipe flow may lead to the convection–diffusion equation (1) with $\boldsymbol{b} = (1, 0)$ and homogeneous Dirichlet boundary conditions on the whole boundary, cf., e.g., [5]. In this case, Ω is the cross–section of the pipe and the parameter ε is the reciprocal of the Hartmann number so that it can be very small.

The paper is organized in the following way. Sections 2 and 3 are devoted to the formulation of the SUPG method in one and two dimensions, respectively, and to a brief discussion of the optimal choice of the stabilization parameter. Then, in Section 4, the SUPG method is applied to a two–dimensional model problem and the inadequacy of present approaches to the choice of the stabilization parameter is demonstrated. Based on the observations from Section 4, a new definition of the SUPG stabilization parameter at outflow boundary layers is derived in Section 5. Numerical results in Section 6 show the advantages of the new approach and the paper is closed by conclusions in Section 7. Throughout the paper, we use the standard notations $P_1(\Omega)$, $Q_1(\Omega)$, $L^2(\Omega)$, $H^1(\Omega) = W^{1,2}(\Omega)$, etc. for the usual function spaces, see, e.g., [4]. Given a vector $\boldsymbol{a} \in \mathbb{R}^2$, we denote by $|\boldsymbol{a}|$ its Euclidean norm.

2 The SUPG method in one dimension

Let us consider the equation (1) in the one–dimensional case with homogeneous Dirichlet boundary conditions and $\Omega = (0, 1)$:

$$-\varepsilon\, u'' + bu' = f \quad \text{in } (0,1)\,, \qquad u(0) = u(1) = 0\,. \tag{2}$$

For simplicity, let b and f be constants, $b \neq 0$. Then, setting $\alpha = f/b$ and $\beta = b/\varepsilon$, we have

$$u(x) = \alpha\, x - \alpha\, \frac{e^{-\beta(1-x)} - e^{-\beta}}{1 - e^{-\beta}}\,, \quad x \in [0,1]\,.$$

Thus, if $\varepsilon \ll |b|$, the solution u contains a boundary layer. More precisely, if $b > 0$, we see that $u(x) \approx \alpha\, x$ on most of $[0, 1)$ and a boundary layer occurs at $x = 1$. Similarly, if $b < 0$, we have $u(x) \approx \alpha\,(x - 1)$ on most of $(0, 1]$ and a boundary layer occurs at $x = 0$.

Let N be a positive integer and let us set $h = 1/N$ and define the nodes $x_i = i\,h$, $i = 0, 1, \ldots, N$. We introduce the finite element space

$$V_h = \{v \in C([0,1])\,;\ \ v|_{[x_{i-1},x_i]} \in P_1([x_{i-1}, x_i]),\ \ i = 1, \ldots, N,\ \ v(0) = v(1) = 0\}$$

consisting of continuous piecewise linear functions. Then the SUPG method for approximating the solution of (2) reads: Find $u_h \in V_h$ such that

$$\varepsilon\,(u_h', v_h') + (bu_h', v_h + \tau\, bv_h') = (f, v_h + \tau\, bv_h') \qquad \forall\ v_h \in V_h\,, \tag{3}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(0, 1)$ and τ is a nonnegative stabilization parameter. This problem has a unique solution which is determined by the values $u_i \equiv u_h(x_i)$, $i = 0, \ldots, N$. If τ is constant in $(0, 1)$, then (3) can be equivalently written in the form

$$-\left(\varepsilon + \tau\, b^2 + \frac{1}{2}\, bh\right) u_{i-1} + 2\,(\varepsilon + \tau\, b^2)\, u_i - \left(\varepsilon + \tau\, b^2 - \frac{1}{2}\, bh\right) u_{i+1} = f\, h^2, \tag{4}$$

where $i = 1, \ldots, N - 1$.

It is well known that the parameter τ can be chosen in such a way that the solution of (3) is nodally exact [3]. Indeed, setting

$$\tau = \frac{h}{2\,|b|}\left(\coth Pe - \frac{1}{Pe}\right) \qquad \text{with} \qquad Pe = \frac{|b|\,h}{2\,\varepsilon}\,, \tag{5}$$

it is easy to verify that $u_i = u(x_i)$, $i = 0, \ldots, N$. The quantity Pe is the local Péclet number which determines whether the problem is locally (i.e., within a particular subinterval) convection dominated or diffusion dominated.

If $\boldsymbol{b}$ or f in (2) are not constant, then τ defined by (5) does not in general lead to a nodally exact discrete solution. Nevertheless, the discrete solution is significantly better than the wildly oscillating solution of the standard Galerkin discretization (defined by (3) with $\tau = 0$).

3 The SUPG method in two dimensions

Let $\mathcal{T}_h$ be a triangulation of the domain Ω consisting of a finite number of open elements K. For simplicity, we shall assume that all elements of $\mathcal{T}_h$ are either triangles or rectangles. Furthermore, 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.

We define the finite element space

$$W_h = \{v \in H^1(\Omega)\,;\ v|_K \in R(K)\ \ \forall\ K \in \mathcal{T}_h\}\,,$$

where $R(K) = P_1(K)$ if K is a triangle and $R(K) = Q_1(K)$ if K is a rectangle. Furthermore, we introduce a test function space $V_h \subset W_h$ taking into account the boundary conditions prescribed for the solution of (1). For example, denoting by $\partial\Omega^D$ and $\partial\Omega^N$ disjoint subsets of $\partial\Omega$ satisfying $\overline{\partial\Omega^D \cup \partial\Omega^N} = \partial\Omega$, by $\boldsymbol{n}$ the outward unit normal vector to $\partial\Omega$ and by u_b a scalar function on $\partial\Omega^D$, the boundary conditions

$$u = u_b \quad \text{on } \partial\Omega^D\,, \qquad \frac{\partial u}{\partial \boldsymbol{n}} = 0 \quad \text{on } \partial\Omega^N \tag{6}$$

lead to the space

$$V_h = \{v \in W_h\,;\ v = 0 \ \text{ on } \partial\Omega^D\}\,.$$

Of course, the triangulation $\mathcal{T}_h$ should be defined in such a way that any boundary edge is a subset of $\overline{\partial\Omega^D}$ or $\overline{\partial\Omega^N}$.

Denoting by $u_{bh} \in W_h$ a function whose trace approximates the boundary condition u_b, the SUPG method for the convection–diffusion equation (1) equipped with the boundary conditions (6) reads:

Find $u_h \in W_h$ such that $u_h - u_{bh} \in V_h$ and

$$\varepsilon\,(\nabla u_h, \nabla v_h) + (\boldsymbol{b}\cdot\nabla u_h,\, v_h + \tau\,\boldsymbol{b}\cdot\nabla v_h) = (f,\, v_h + \tau\,\boldsymbol{b}\cdot\nabla v_h) \quad \forall\ v_h \in V_h\,, \tag{7}$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^2$ and τ is a nonnegative stabilization parameter.

The choice of τ significantly influences the quality of the discrete solution and therefore it has been the subject of extensive research over the last three decades, see, e.g., the review in [6]. Nevertheless, the definitions of τ mostly rely on heuristic arguments and a general 'optimal' way of choosing τ is still not known. Often, by analogy with the one–dimensional formula (5), the parameter τ is defined, on any element $K \in \mathcal{T}_h$, by

$$\tau|_K = \frac{h_K}{2\,|\boldsymbol{b}|}\left(\coth Pe_K - \frac{1}{Pe_K}\right) \quad \text{with} \quad Pe_K = \frac{|\boldsymbol{b}|\,h_K}{2\,\varepsilon}\,, \tag{8}$$

where h_K is the element diameter in the direction of the convection vector $\boldsymbol{b}$. Various justifications of this formula can be found in [6] (see also the next section). Note that, generally, the parameters h_K, Pe_K and $\tau|_K$ are functions of the points $\mathbf{x} \in K$.

4 Application of the SUPG method to a model problem

Let $\Omega = (0,1)^2$ and let us consider the equation (1) with constant data f and $\boldsymbol{b} \equiv (b_1, b_2)$ satisfying $b_1 \neq 0$ and with the following boundary conditions:

$$u(0,y) = u(1,y) = 0 \qquad \forall\ y \in (0,1)\,, \tag{9}$$

$$u(x,0) = u(x,1)\,, \quad \frac{\partial u}{\partial y}(x,0) = \frac{\partial u}{\partial y}(x,1) \qquad \forall\ x \in (0,1)\,. \tag{10}$$

This problem has a unique solution. Moreover, the solution is independent of y and satisfies (2) with $b = b_1$.

First, we shall confine ourselves to the three types of triangulations depicted in Fig. 1. The nodes are equidistant in both the x– and y–directions and the corresponding mesh widths are denoted by h_1 and h_2, respectively. The test function finite element space is

$$V_h = \{v \in W_h;\ v(0,y) = v(1,y) = 0\ \forall\ y \in (0,1),\\ v(x,0) = v(x,1)\ \forall\ x \in (0,1)\}$$

and the SUPG solution of the considered problem is a function $u_h \in V_h$ satisfying (7). Again, this discrete solution is uniquely determined and, if τ is constant, it does not depend on the y–coordinate. For both the triangular and the rectangular triangulations, the discrete solution then satisfies the one–dimensional scheme (4) with $b = b_1$ and $h = h_1$. Thus, in view of (5), an optimal choice of the stabilization parameter τ in (7) is

$$\tau = \frac{h_1}{2\,|b_1|}\left(\coth Pe - \frac{1}{Pe}\right) \qquad \text{with} \qquad Pe = \frac{|b_1|\,h_1}{2\,\varepsilon}\,. \tag{11}$$

In this case the SUPG solution is nodally exact.

In the triangular case, the optimal one–dimensional scheme can be recovered also for piecewise constant τ. It suffices when τ has the same value on elements whose

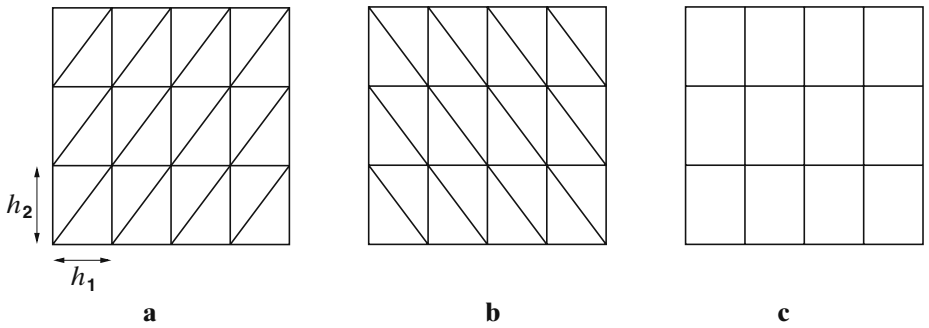


Fig. 1 Types of triangulations considered in Section 4 (**a**–**c**)

barycentres have the same x–coordinate and when, for any two elements K, K' sharing a 'diagonal' edge, we have

$$\frac{1}{2}\left(\tau|_K + \tau|_{K'}\right) = \frac{h_1}{2\,|b_1|}\left(\coth Pe - \frac{1}{Pe}\right) \qquad \text{with} \qquad Pe = \frac{|b_1|\,h_1}{2\,\varepsilon}\,. \tag{12}$$

Then the SUPG solution is again nodally exact.

If the convection vector $\boldsymbol{b}$ points in the x–direction (i.e., $b_2 = 0$), then $h_K = h_1$ and $|\boldsymbol{b}| = |b_1|$ so that the formula (8) provides the optimal value of τ determined by (11). This may be viewed as a justification of using (8) and, in particular, of defining h_K as the diameter of K in the direction of $\boldsymbol{b}$ and not as the real diameter of K.

Now, we shall investigate the following setting of the problem discussed in this section:

Example 1 We consider the equation (1) in $\Omega = (0, 1)^2$ with the boundary conditions (9) and (10) and with $\varepsilon = 10^{-4}$, $\boldsymbol{b} = (1, 0)$, and $f = 1$.

Let us solve Example 1 on a triangulation of the type depicted in Fig. 1a or 1b. Then, as we know, the solution of the SUPG method with τ defined by (8) is nodally exact. Often, a triangulation of a domain with a simple geometry is constructed by refining a coarse triangulation. If all triangles of the triangulations from Fig. 1a or 1b are divided into four equal triangles by connecting midpoints of edges, we obtain triangulations of the same type as in Fig. 1a or 1b, respectively, and hence the corresponding SUPG solutions are again nodally exact. If however we divide all triangles of the triangulations from Fig. 1a or 1b into four equal triangles by applying twice bisection, we obtain the triangulation depicted in Fig. 2 and the corresponding SUPG solution significantly differs from the nodally exact solution, see Fig. 3. Note that the triangulation in Fig. 2 contains the same type of triangles as the two triangulations in Fig. 1a and 1b and also that the orientation of the triangles with respect to the convection vector $\boldsymbol{b}$ is the same as in Fig. 1a and 1b. This shows that the information available on a particular element of the triangulation is not sufficient to define the stabilization parameter τ in an optimal way and that the orientation of the neighbouring elements has to be taken into account.

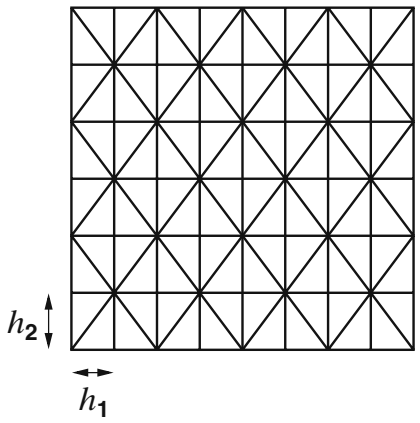


Fig. 2 Triangulation obtained by refining the triangulations from Fig. 1a and 1b

5 A new definition of the SUPG stabilization parameter

The favourable properties of the one–dimensional SUPG method (4) with τ defined by (5) are due to the fact that the upwind character of the method increases with increasing Péclet number. Particularly, for $Pe \gg 1$, we have $\tau \approx h/(2|b|)$ and the coefficient at the downwind node in (4) is

$$-\left(\varepsilon + \tau\, b^2 - \frac{1}{2}\,|b|\,h\right) \approx -\varepsilon\,.$$

Then the SUPG stabilization is basically equivalent to approximating the convective term by classical upwind differencing and the influence of the Dirichlet boundary condition at the outflow boundary node on the values of u_h at interior nodes is significantly suppressed.

In two dimensions, this property is generally lost, which leads to spurious oscillations like in Fig. 3. By analogy with the one–dimensional case, it is natural to ask whether τ can be defined in such a way that, for $\varepsilon \to 0$, the difference scheme corresponding to (7) does not employ the outflow boundary values of u_h. Unfortunately, this is generally impossible. As an example, let us consider a triangulation of $\Omega = (0,1)^2$ of the type in Fig. 1a with $h_1 = h_2 = h$ and let $\varphi_j \in V_h$ be the standard basis function corresponding to a boundary node $\mathbf{x}_j$ lying on the right–hand side of Ω. Let $\varphi_i \in V_h$ be the standard basis function corresponding to the interior node $\mathbf{x}_i$ connected with $\mathbf{x}_j$ by a horizontal edge. Then, for $\boldsymbol{b} = (2,3)$,

$$(\boldsymbol{b}\cdot\nabla\varphi_j,\, \varphi_i + \tau\,\boldsymbol{b}\cdot\nabla\varphi_i) = \frac{h}{6} + \frac{2}{h^2}\int_{\operatorname{supp}\varphi_i \cap \operatorname{supp}\varphi_j} \tau\,\mathrm{d}\mathbf{x} \tag{13}$$

and hence, for any choice of τ, the value of u_h at $\mathbf{x}_j$ contributes to the approximation of the convective term at $\mathbf{x}_i$.

Thus, let us at least investigate whether a suitable choice of τ can remove the oscillations shown in Fig. 3. We denote the outflow Dirichlet boundary by Γ, i.e., $\Gamma = \{1\} \times [0,1]$, and we set

$$G_h = \bigcup_{K\in\mathcal{T}_h,\ \overline{K}\cap\Gamma\neq\emptyset} K\,.$$

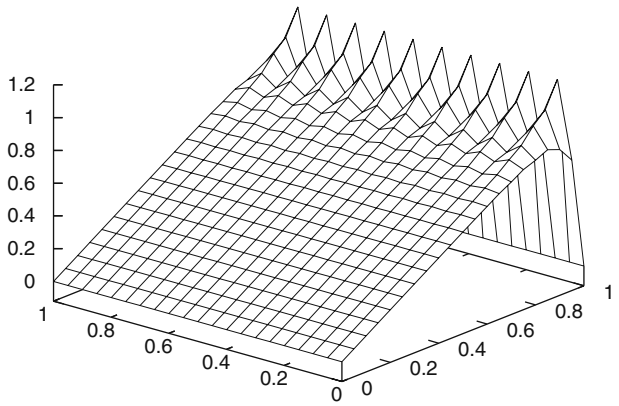


Fig. 3 Example 1, SUPG solution for τ defined by (8) computed on the triangulation from Fig. 2 with $h_1 = h_2 = 1/20$

Furthermore, we denote by $\varphi_1, \ldots, \varphi_{M_h}$ all standard basis functions of V_h that satisfy

$$\operatorname{supp} \varphi_i \cap G_h \neq \emptyset, \qquad i = 1, \ldots, M_h\,. \tag{14}$$

The nodally exact solution of Example 1 on the triangulation of Fig. 2 with $h_1 = h_2 = 1/20$ satisfies $u_h(x, y) \approx x$ in $[0, 1 - h_1] \times [0, 1]$ and hence, neglecting the diffusion term, it satisfies (7) if and only if

$$\int_{G_h} v_h + \tau\, \boldsymbol{b} \cdot \nabla v_h \,\mathrm{d}\mathbf{x} = 0 \qquad \forall\ v_h \in V_h\,.$$

This can be written in the equivalent form

$$\int_{G_h} \varphi_i + \tau\, \boldsymbol{b} \cdot \nabla \varphi_i \,\mathrm{d}\mathbf{x} = 0, \qquad i = 1, \ldots, M_h\,. \tag{15}$$

There are many possible ways of satisfying these relations. Probably the simplest one is to set

$$\tau|_K = \begin{cases} \dfrac{2\,h_1}{3} & \text{if } K \text{ has an edge on } \Gamma, \\ \dfrac{h_1}{3} & \text{otherwise} \end{cases} \qquad \forall\ K \in \mathcal{T}_h,\ K \subset G_h\,. \tag{16}$$

On the remaining elements $K \in \mathcal{T}_h$ we define τ by (8). Then the SUPG solution is a very good approximation of the solution to Example 1 as Fig. 4 shows.

The relations (15) used to define τ were obtained thanks to the fact that the nodally exact solution of Example 1 satisfies

$$\boldsymbol{b} \cdot \nabla u_h - f \approx 0 \quad \text{in } \Omega \setminus \overline{G_h}\,, \qquad \boldsymbol{b} \cdot \nabla u_h - f = const. \quad \text{in } \overline{G_h}\,. \tag{17}$$

For other data or boundary conditions, this will usually not be satisfied but it can be expected that the validity of (15) will diminish the spurious oscillations along an outflow boundary layer.

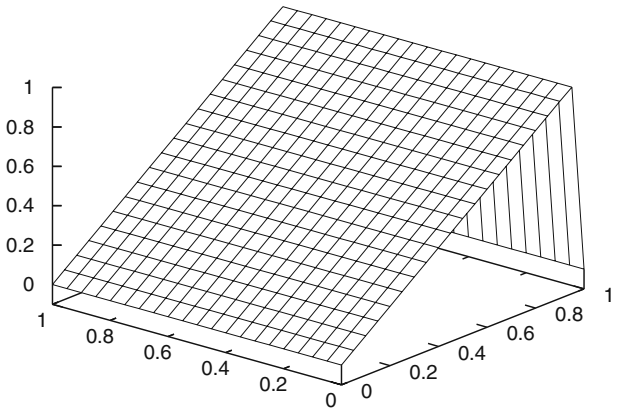


Fig. 4 Example 1, SUPG solution for τ defined by (8) and (16) computed on the triangulation from Fig. 2 with $h_1 = h_2 = 1/20$

Let us now investigate whether (15) can be satisfied for a general polygonal domain Ω and a triangulation $\mathcal{T}_h$ consisting of triangles. Using the notation $\partial\Omega^D$ for the part of $\partial\Omega$ where Dirichlet boundary conditions are prescribed, we again introduce the outflow Dirichlet boundary

$$\Gamma = \overline{\{\mathbf{x} \in \partial\Omega^D\,;\ (\boldsymbol{b}\cdot\boldsymbol{n})(\mathbf{x}) > 0\}}\,.$$

For simplicity, we assume that Γ is connected and consists of whole boundary edges of $\mathcal{T}_h$. Like above, we set

$$G_h = \bigcup_{K\in\mathcal{G}_h} K \qquad \text{where} \qquad \mathcal{G}_h = \{K \in \mathcal{T}_h\,;\ \overline{K}\cap\Gamma \neq \emptyset\}\,.$$

Furthermore, we set

$$\mathcal{G}_h^1 = \{K \in \mathcal{G}_h\,;\ K \text{ has only one vertex on } \Gamma\}\,, \qquad \mathcal{G}_h^2 = \mathcal{G}_h \setminus \mathcal{G}_h^1\,.$$

For any vertex $\mathbf{z} \in \Gamma$, we denote by

$$\mathcal{G}_h^1(\mathbf{z}) = \{K \in \mathcal{G}_h^1\,;\ \mathbf{z} \in \overline{K}\}$$

the set of all elements possessing the vertex $\mathbf{z}$ with no other vertex lying on Γ. For any $K \in \mathcal{T}_h$, we set

$$\boldsymbol{b}_K = \frac{1}{|K|}\int_K \boldsymbol{b}\,\mathrm{d}\mathbf{x}\,.$$

If $K \in \mathcal{G}_h^1$, we assume that at the vertex $\overline{K}\cap\Gamma$, the vector $\boldsymbol{b}_K$ points outwards from $\overline{K}$. If $K \in \mathcal{G}_h^2$ has exactly two vertices on Γ, we denote by $\boldsymbol{n}_E$ the outward normal vector to the edge E connecting these two vertices and assume that $\boldsymbol{b}_K\cdot\boldsymbol{n}_E > 0$.

We again denote by $\varphi_1, \ldots, \varphi_{M_h}$ all standard basis functions of V_h satisfying (14). For $i = 1, \ldots, M_h$, let $\mathbf{x}_i$ be the vertex associated with the basis function φ_i, i.e., $\varphi_i(\mathbf{x}_i) = 1$ and $\varphi_i(\mathbf{x}) = 0$ for any vertex $\mathbf{x} \neq \mathbf{x}_i$. We set

$$\mathcal{N}_h = \{\mathbf{x}_1, \ldots, \mathbf{x}_{M_h}\}\,,$$
$$\mathcal{N}_h^2 = \{\mathbf{x} \in \mathcal{N}_h\,;\ \exists\, K \in \mathcal{G}_h^2 :\ \mathbf{x} \in \overline{K}\}\,, \qquad \mathcal{N}_h^1 = \mathcal{N}_h \setminus \mathcal{N}_h^2\,.$$

The example leading to (13) shows that it is in general impossible to satisfy (15) elementwise. Nevertheless, we can use the fact that each vertex $\mathbf{x}_i$, $i = 1, \ldots, M_h$, can be easily assigned to an element $K\in\mathcal{G}_h$ (in a one–to–one way) such that $\boldsymbol{b}_K\cdot\nabla\varphi_i|_K<0$. This follows from the following results.

Lemma 1 *For any $K \in \mathcal{G}_h^1$ satisfying* $\mathrm{card}(\overline{K}\cap\mathcal{N}_h) = 2$, *there exists $i \in \{1, \ldots, M_h\}$ such that $\mathbf{x}_i$ is a vertex of K and $\boldsymbol{b}_K\cdot\nabla\varphi_i|_K < 0$.*

Proof Let us assume that the lemma is false. Then there exist $K \in \mathcal{G}_h^1$ and $j, k \in \{1, \ldots, M_h\}$ such that $j \neq k$ and

$$\mathbf{x}_j, \mathbf{x}_k \in \overline{K}\,, \qquad \boldsymbol{b}_K\cdot\nabla\varphi_j|_K \geq 0\,, \qquad \boldsymbol{b}_K\cdot\nabla\varphi_k|_K \geq 0\,.$$

We denote by $\mathbf{z}$ the remaining vertex of K. The vectors $\nabla\varphi_j|_K$ and $\nabla\varphi_k|_K$ are orthogonal to the edges $\mathbf{z}, \mathbf{x}_k$ and $\mathbf{z}, \mathbf{x}_j$, respectively, and point into K. Consequently, $\boldsymbol{b}_K$ points from the vertex $\mathbf{z}$ into $\overline{K}$. This is not possible since $\mathbf{z} \in \Gamma$. □

Lemma 2 *Let $K \in \mathcal{G}_h^2$ satisfy $\overline{K} \cap \mathcal{N}_h \neq \emptyset$ and let $i \in \{1, \ldots, M_h\}$ be such that $\overline{K} \cap \mathcal{N}_h = \{\mathbf{x}_i\}$. Then $\boldsymbol{b}_K \cdot \nabla \varphi_i|_K < 0$.*

Proof Since the vector $\nabla \varphi_i|_K$ is orthogonal to the edge E of K opposite the vertex $\mathbf{x}_i$ and points into K, the lemma follows immediately from the assumptions on $\boldsymbol{b}_K$. □

Lemma 3 *Let $\mathbf{z} \in \Gamma$ be any vertex other than the end points of Γ and let* $\operatorname{card} \mathcal{G}_h^1(\mathbf{z}) \geq 2$. *Let the edges of elements of $\mathcal{G}_h^1(\mathbf{z})$ opposite $\mathbf{z}$ form a connected curve, see Fig. 5. For simplicity, let us assume that there exist $k, l \in \{1, \ldots, M_h\}$ such that $k \leq l$,*

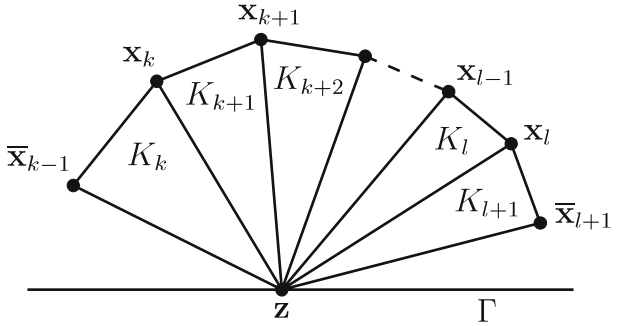
$$\{\mathbf{x}_k, \ldots, \mathbf{x}_l\} = \{\mathbf{x} \in \mathcal{N}_h ;\ \exists\, K, K' \in \mathcal{G}_h^1(\mathbf{z}) : \mathbf{x} \in \overline{K} \cap \overline{K'}\}$$

and $\operatorname{card} \mathcal{G}_h^1(\mathbf{z}) = l - k + 2$, *see Fig. 5. Moreover, if $k < l$, we assume that, for $i = k, \ldots, l-1$, the vertices $\mathbf{x}_i$ and $\mathbf{x}_{i+1}$ are connected by an edge of the triangulation $\mathcal{T}_h$. Finally, we assume that $\boldsymbol{b}$ is constant on the union of elements of $\mathcal{G}_h^1(\mathbf{z})$. We denote the elements of $\mathcal{G}_h^1(\mathbf{z})$ by $K_k, \ldots, K_{l+1}$ in such a way that, for $i = k, \ldots, l$, the elements K_i and K_{i+1} share the vertex $\mathbf{x}_i$. Furthermore, we denote by $\overline{\mathbf{x}}_{k-1}$ and $\overline{\mathbf{x}}_{l+1}$ the remaining vertices of the elements K_k and K_{l+1}, respectively. Since these vertices may lie on $\partial\Omega^D \setminus \Gamma$, we denote the piecewise linear basis functions associated with these vertices by $\overline{\varphi}_{k-1}$ and $\overline{\varphi}_{l+1}$, respectively. Then there exists $j \in \{k, \ldots, l+1\}$ such that*

$$\begin{aligned}
&\boldsymbol{b} \cdot \nabla \varphi_i|_{K_i} < 0\,, && \boldsymbol{b} \cdot \nabla \varphi_i|_{K_{i+1}} \geq 0\,, && i = k, \ldots, j-2\,,\\
&\boldsymbol{b} \cdot \nabla \varphi_i|_{K_i} \geq 0\,, && \boldsymbol{b} \cdot \nabla \varphi_i|_{K_{i+1}} < 0\,, && i = j+1, \ldots, l\,,\\
&\boldsymbol{b} \cdot \nabla \overline{\varphi}_{k-1}|_{K_k} \geq 0\,, && \boldsymbol{b} \cdot \nabla \varphi_{j-1}|_{K_{j-1}} < 0 && \text{if } j > k\,,\\
&\boldsymbol{b} \cdot \nabla \varphi_j|_{K_{j+1}} < 0\,, && \boldsymbol{b} \cdot \nabla \overline{\varphi}_{l+1}|_{K_{l+1}} \geq 0 && \text{if } j \leq l\,.
\end{aligned}$$

Proof For simplicity, we shall write φ_{k-1} and φ_{l+1} instead of $\overline{\varphi}_{k-1}$ and $\overline{\varphi}_{l+1}$, respectively. The vector $\nabla \varphi_k|_{K_k}$ is orthogonal to the edge $\mathbf{z}, \overline{\mathbf{x}}_{k-1}$ and points into K_k. Similarly, $\nabla \varphi_l|_{K_{l+1}}$ is a vector orthogonal to the edge $\mathbf{z}, \overline{\mathbf{x}}_{l+1}$ which points into K_{l+1}. Since $\boldsymbol{b}$ does not point from $\mathbf{z}$ into $\cup_{i=k}^{l+1} \overline{K_i}$, we deduce that $\boldsymbol{b} \cdot \nabla \varphi_k|_{K_k}$ or $\boldsymbol{b} \cdot \nabla \varphi_l|_{K_{l+1}}$ is

Fig. 5 Notation for Lemma 3



negative. Without loss of generality we may assume that $\boldsymbol{b}\cdot\nabla\varphi_k|_{K_k} < 0$. Then there exists $j \in \{k, \dots, l+1\}$ such that

$$\boldsymbol{b}\cdot\nabla\varphi_i|_{K_i} < 0\,, \qquad i = k, \dots, j\,,$$

and, if $j \le l$,

$$\boldsymbol{b}\cdot\nabla\varphi_{j+1}|_{K_{j+1}} \ge 0\,.$$

From the latter inequality, analogously to the beginning of the proof we see that

$$\boldsymbol{b}\cdot\nabla\varphi_i|_{K_{i+1}} < 0\,, \qquad i = j, \dots, l\,.$$

Finally, since the vectors $\nabla\varphi_{i-1}|_{K_i}$ and $\nabla\varphi_{i+1}|_{K_{i+1}}$ have opposite directions for any $i \in \{k, \dots, l\}$, it follows from the above inequalities that

$$\begin{aligned} \boldsymbol{b}\cdot\nabla\varphi_i|_{K_{i+1}} &> 0\,, \qquad i = k-1, \dots, j-2\,,\\ \boldsymbol{b}\cdot\nabla\varphi_i|_{K_i} &> 0\,, \qquad i = j+2, \dots, l+1\,. \end{aligned}$$

□

Lemmas 1–3 enable us to introduce an algorithm for defining the SUPG parameter τ at outflow Dirichlet boundaries. Since the relations (15) correspond to $\varepsilon \to 0$, we denote τ satisfying (15) by τ_0. Thus, we shall construct a piecewise constant function τ_0 on G_h satisfying

$$\int_{G_h} \varphi_i + \tau_0\, \boldsymbol{b}\cdot\nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0, \qquad i = 1, \dots, M_h\,. \tag{18}$$

Then, by analogy to (8), we define the parameter τ, on any element $K \in \mathcal{G}_h$, by

$$\tau|_K = \tau_0|_K \left(\coth Pe_K - \frac{1}{Pe_K}\right) \quad \text{with} \quad Pe_K = \frac{|\boldsymbol{b}_K|\, h_K}{2\,\varepsilon}\,. \tag{19}$$

On elements $K \in \mathcal{T}_h \setminus \mathcal{G}_h$, we define τ by (8) with $\boldsymbol{b}$ replaced by $\boldsymbol{b}_K$.

Let us note that the definition of τ_0 is not important on elements which have all three vertices on the Dirichlet boundary since all functions from V_h vanish on these elements. Therefore, we shall not mention such elements in what follows.

It is advantageous to start defining τ_0 on elements of $\mathcal{G}_h^1$. First, for any vertex $\mathbf{z} \in \Gamma$ we construct the set $\mathcal{G}_h^1(\mathbf{z})$. If this set consists of one element K, the value of τ_0 on K can be defined arbitrarily. If $\overline{K} \cap \mathcal{N}_h = \{\mathbf{x}_i\}$ for some $i \in \{1, \dots, M_h\}$ and $\boldsymbol{b}_K\cdot\nabla\varphi_i|_K \ge 0$, we set $\tau_0|_K = h_K/(2|\boldsymbol{b}_K|)$ like in (8). If $\boldsymbol{b}_K\cdot\nabla\varphi_i|_K < 0$, we can define τ_0 on K in such a way that

$$\int_K \varphi_i + \tau_0\, \boldsymbol{b}\cdot\nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0\,.$$

However, the value of τ_0 determined from this relation tends to infinity if the vector $\boldsymbol{b}_K$ approaches the direction of the edge of K opposite $\mathbf{x}_i$. Therefore, we introduce a positive parameter α_{min} (e.g., $\alpha_{min} = 0.1$) and set

$$\tau_0|_K = \frac{1}{\max\{-3\,\boldsymbol{b}_K\cdot\nabla\varphi_i|_K,\ \alpha_{min}\,|\boldsymbol{b}_K|/h_K\}}\,. \tag{20}$$

If $\overline{K} \cap \mathcal{N}_h = \{\mathbf{x}_i, \mathbf{x}_j\}$ for some $i, j \in \{1, \ldots, M_h\}$, $i \neq j$, we set

$$\tau_0|_K = -\frac{1}{3\min\{\boldsymbol{b}_K \cdot \nabla\varphi_i|_K,\ \boldsymbol{b}_K \cdot \nabla\varphi_j|_K\}}\,. \tag{21}$$

This value of τ_0 is positive by Lemma 1 and, if $\mathbf{z}$ is different from the end points of Γ, it is bounded by a constant depending on the minimal angle θ in the elements of $\mathcal{T}_h$. More precisely, it can be shown that $\tau_0|_K \le h_K/\left(3\min\left\{\frac{1}{2}, \sin^2\theta\right\}|\boldsymbol{b}_K|\right)$. The bound $h_K/(3\sin^2\theta|\boldsymbol{b}_K|)$ corresponds to $\boldsymbol{b}_K$ aligned with Γ so that the values of τ_0 are smaller in practice. It is easy to see that in all three cases discussed above we have

$$\int_K \varphi_i + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_i\,\mathrm{d}\mathbf{x} \ge 0 \qquad \forall\ \mathbf{x}_i \in \overline{K} \cap \mathcal{N}_h\,. \tag{22}$$

Now let $\operatorname{card}\mathcal{G}_h^1(\mathbf{z}) \ge 2$ and let $\mathbf{z}$ be different from the end points of Γ. If necessary, we decompose $\mathcal{G}_h^1(\mathbf{z})$ into several sets satisfying the assumptions of Lemma 3 or consisting of one element and we treat these sets separately. The treatment of single elements was discussed in the preceding paragraph and hence it suffices to consider the case when $\mathcal{G}_h^1(\mathbf{z})$ satisfies the assumptions of Lemma 3. This lemma was formulated for a constant vector $\boldsymbol{b}$ but if $\boldsymbol{b}$ is non–constant, the assertion remains true provided that the triangulation $\mathcal{T}_h$ is fine enough with respect to variations of $\boldsymbol{b}$. An alternative is to modify the discrete problem (7) in such a way that $\boldsymbol{b}$ is replaced on the elements of $\mathcal{G}_h^1(\mathbf{z})$ by its mean value. Thus, let us consider the notation of Lemma 3 and let j be the integer introduced in the assertion of this lemma. We define τ_0 on K_j in the same way as in the case $\operatorname{card}\mathcal{G}_h^1(\mathbf{z}) = 1$ discussed above. To fix ideas, let us assume that $j \in \{k+1, \ldots, l\}$. Then we compute τ_0 on K_{j-1} and on K_{j+1} from the relations

$$\int_{K_{j-1}\cup K_j} \varphi_{j-1} + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_{j-1}\,\mathrm{d}\mathbf{x} = 0\,, \quad \int_{K_j\cup K_{j+1}} \varphi_j + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_j\,\mathrm{d}\mathbf{x} = 0\,. \tag{23}$$

Since $\tau_0|_{K_j}$ is given by (21) with $K = K_j$ and $i = j-1$, the inequality in (22) holds with $K = K_j$ and $i = j-1$, j. Thus, it follows from Lemma 3 that the relations (23) determine both $\tau_0|_{K_{j-1}}$ and $\tau_0|_{K_{j+1}}$ uniquely and that both these values are positive. To determine τ_0 on the remaining elements of $\mathcal{G}_h^1(\mathbf{z})$, we require

$$\int_{K_i\cup K_{i+1}} \varphi_i + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_i\,\mathrm{d}\mathbf{x} = 0 \quad \text{for } i = k, \ldots, j-2 \ \text{ and } \ i = j+1, \ldots, l\,.$$

According to Lemma 3, the respective values of τ_0 can be easily computed and are positive. The cases $j = k$ and $j = l+1$ can be viewed as particular cases of the above procedure. Note also that, if $\bar{\mathbf{x}}_{k-1} = \mathbf{x}_{k-1} \in \mathcal{N}_h$ or $\bar{\mathbf{x}}_{l+1} = \mathbf{x}_{l+1} \in \mathcal{N}_h$, we have respectively

$$\int_{K_k} \varphi_{k-1} + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_{k-1}\,\mathrm{d}\mathbf{x} \ge 0 \quad \text{or} \quad \int_{K_{l+1}} \varphi_{l+1} + \tau_0\,\boldsymbol{b}\cdot\nabla\varphi_{l+1}\,\mathrm{d}\mathbf{x} \ge 0\,. \tag{24}$$

Indeed, if $\boldsymbol{b}_K \cdot \nabla\varphi_{k-1}|_{K_k} < 0$, we have $j = k$ in view of Lemma 3. As we explained above, the inequality in (22) is satisfied for $K = K_j$ and $i = j-1$ and hence the first inequality in (24) holds. The validity of the second inequality in (24) follows analogously.

If $\operatorname{card} \mathcal{G}_h^1(\mathbf{z}) \geq 2$ and $\mathbf{z}$ is an end point of Γ, we can often proceed in the same way as above. However, generally, it is not possible to guarantee the existence of τ_0 satisfying (18) for $\mathbf{x}_i$ connected by an edge with this $\mathbf{z}$. On elements $K \in \mathcal{G}_h^1(\mathbf{z})$ such that $\boldsymbol{b}_K \cdot \nabla\varphi_i|_K \geq 0$ for any $\mathbf{x}_i \in \overline{K} \cap \mathcal{N}_h$, we set $\tau_0|_K = h_K/(2|\boldsymbol{b}_K|)$ like in (8). If the above procedure leads to a negative value of τ_0 on some $K \in \mathcal{G}_h^1(\mathbf{z})$, we set $\tau_0|_K = 0$.

The above definition of τ_0 on elements of $\mathcal{G}_h^1$ ensures that (18) holds for any $i \in \{1, \ldots, M_h\}$ such that $\mathbf{x}_i \in \mathcal{N}_h^1$, with the possible exception of some $\mathbf{x}_i$ that are connected by an edge to an end point of Γ. Moreover, for $\mathbf{x}_i \in \mathcal{N}_h^2$, setting

$$\mathcal{G}_h^{1,i} = \{K \in \mathcal{G}_h^1 ;\ \mathbf{x}_i \in \overline{K} \ \text{ and } \ \forall\, K' \in \mathcal{G}_h^2 : \mathbf{x}_i \in \overline{K'} \ \Rightarrow\ \overline{K} \cap \overline{K'} = \mathbf{x}_i\},$$

we have (again with the possible exception of some $\mathbf{x}_i$ that are connected to an end point of Γ)

$$\sum_{K \in \mathcal{G}_h^{1,i}} \int_K \varphi_i + \tau_0\, \boldsymbol{b} \cdot \nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0 \qquad \forall\ \mathbf{x}_i \in \mathcal{N}_h^2 .$$

Therefore, to satisfy (18), we may define τ_0 on any $K \in \mathcal{G}_h^2$ with $\overline{K} \cap \mathcal{N}_h = \{\mathbf{x}_i\}$ by

$$\sum_{\substack{K' \in \mathcal{G}_h^1 \cup \{K\},\\ \operatorname{meas}_1(\overline{K} \cap \overline{K'}) \neq 0}} \int_{K'} \varphi_i + \tau_0\, \boldsymbol{b} \cdot \nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0\,. \tag{25}$$

Note that, in (25), we integrate over a set consisting of K and elements of $\mathcal{G}_h^1$ sharing an edge with K. According to Lemma 2 and the inequalities (22) and (24), the value

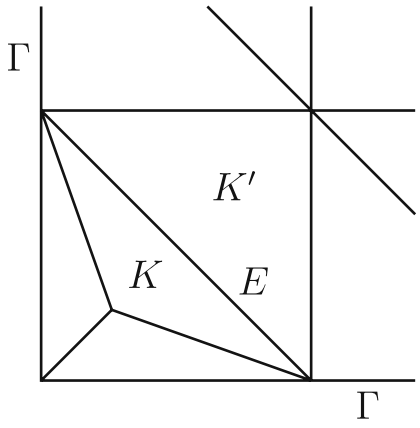


Fig. 6 Element K not satisfying the assumption $\boldsymbol{b}_K \cdot \boldsymbol{n}_E > 0$

of $\tau_0|_K$ is determined uniquely by (25) and is positive. This completes the definition of τ_0 on G_h. For clarity, we summarize the whole algorithm in Fig. 7.

Remark 1 If we apply the above definition of τ_0 to Example 1 on the meshes from Figs. 1a, 1b and 2, we obtain for τ_0 the values given in (16). Thus, in view of (12), the

```
for K ∈ T_h \ G_h do
    τ_0|_K := h_K/(2|b_K|)
enddo
for vertices z ∈ Γ do
    if G_h^1(z) = {K} then
        if K̄ ∩ N_h = {x_i} then
            if b_K · ∇φ_i|_K ≥ 0 then
                τ_0|_K := h_K/(2|b_K|)
            else
                use (20)
            endif
        else
            use (21)
        endif
    else
        decompose G_h^1(z) into subsets satisfying the assumptions
        of Lemma 3 or consisting of one element
        if subset = {K} then
            define τ_0|_K as for G_h^1(z) = {K}
        else
            find j from Lemma 3 and define τ_0|_{K_j} as for G_h^1(z) = {K}
            for i ∈ {k, ..., l} successively determine τ_0 to satisfy
                ∫_{K_i ∪ K_{i+1}} φ_i + τ_0 b · ∇φ_i dx = 0
        endif
    endif
enddo
for K ∈ G_h^2 do
    determine τ_0|_K from (25)
enddo
for K ∈ T_h do
    compute τ|_K from (19)
enddo
REMARKS:
if K ∈ G_h and b_K · ∇φ_i|_K ≥ 0  ∀ x_i ∈ K̄ ∩ N_h then
    do not use the above procedure and set τ_0|_K := h_K/(2|b_K|)
after computing any new τ_0|_K set
    τ_0|_K := min{max{τ_0|_K, 0}, h_K/(α_min|b_K|)}
```

Fig. 7 New definition of the SUPG parameter

definition of τ given in (19) leads to a nodally exact SUPG solution of Example 1 on meshes of the type depicted in Fig. 1a and 1b.

Remark 2 The proposed approach will not lead to satisfactory results if $(\boldsymbol{b}\cdot\boldsymbol{n})/|\boldsymbol{b}| \to 0$ on Γ. This can be also deduced from the fact that, in this case, the value of τ_0 determined from (25) tends to infinity. The algorithm may also fail if the triangulation contains an element $K \in \mathcal{G}_h^2$ of the type depicted in Fig. 6. Then the assumption $\boldsymbol{b}_K \cdot \boldsymbol{n}_E > 0$ is typically not satisfied. The simplest remedy is to bisect the elements K, K' sharing the edge E. Note also that the triangulation should be constructed in such a way that the part of the boundary of the strip $\overline{G_h}$ lying in Ω copies the outflow boundary Γ. This helps to satisfy approximately the second relation in (17).

Remark 3 As we shall see in the next section, the above definition of τ_0 sometimes removes completely the spurious oscillations present in the SUPG solution when τ is defined by (8). Nevertheless, this does not mean that the discrete problem (7) then satisfies the discrete maximum principle. As an example, let us consider $\Omega = (0,1)^2$, $\boldsymbol{b} = (1,0)$ and a triangulation of Ω of the type depicted in Fig. 1a with $h_1 = h_2 = h$. Then, for $\varepsilon < h/6$ and for any nonnegative $\tau \in L^\infty(\Omega)$, the entries of the stencil corresponding to any row of the stiffness matrix of (7) have the following signs:

$$\begin{bmatrix} 0 & - & + \\ - & + & \pm \\ - & + & 0 \end{bmatrix}.$$

Thus, independently of the choice of τ, the discrete maximum principle does not hold in general.

Remark 4 For simple model problems, a piecewise constant function τ_0 such that (18) holds can be defined also in the quadrilateral case, but in general the existence of a nonnegative piecewise constant τ_0 satisfying (18) cannot be guaranteed. A remedy could be to use a non–constant τ_0 on some elements but this is not very convenient from a practical point of view. A further drawback of the quadrilateral case is that the definition of τ_0 is nonlocal. Therefore, it is advantageous to divide the quadrilaterals intersecting Γ into triangles and to use continuous piecewise linear functions in $\overline{G_h}$ together with τ_0 defined by the algorithm in Fig. 7.

Remark 5 An alternative technique for reducing spurious oscillations at outflow boundaries is a weak imposition of Dirichlet boundary conditions, see, e.g., [1]. For $\varepsilon \to 0$, this technique leads to a problem without any boundary condition at the outflow boundary, in contrast to our approach, which aims at suppressing the influence of the outflow boundary condition on the discrete solution in the interior of the computational domain.

6 Numerical results

In this section we present some of our numerical results illustrating the properties of the approach proposed in the preceding section. We start with the following very simple model problem.

Example 2 We consider the equation (1) and the boundary conditions (6) with $\Omega = (0,1)^2$, $\partial\Omega^D = \partial\Omega$, $\partial\Omega^N = \emptyset$, $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (\cos(\pi/3), -\sin(\pi/3))$, $f = 0$, and

$$u_b(x,y) = \begin{cases} 0 & \text{if } x = 1 \text{ or } y = 0, \\ 1 & \text{otherwise.} \end{cases}$$

We use a triangulation of the type in Fig. 1a with $h_1 = h_2 = 1/20$. The SUPG solution with τ defined by (8), see Fig. 8a, contains large spurious oscillations along both outflow boundary layers. On the other hand, if we define τ by the algorithm in Fig. 7, we obtain a nodally exact solution, see Fig. 8b.

On the triangulations considered above, the nodally exact solution u_h of Example 2 is constant in $\Omega \setminus \overline{G_h}$ if $\varepsilon \to 0$. Moreover, $\boldsymbol{b}$ and f are constant and the set $\overline{G_h}$ can be decomposed into subsets on which u_h is linear and (18) holds. Thus, repeating the considerations from the beginning of Section 5, we can easily deduce that the nodally exact solution really solves (7) for $\varepsilon \to 0$. In the subsequent examples, such simple considerations will not be possible.

Example 3 We consider the equation (1) and the boundary conditions (6) with $\Omega = (0,1)^2$, $\partial\Omega^D = \partial\Omega$, $\partial\Omega^N = \emptyset$, $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (\cos(\pi/3), -\sin(\pi/3))$, $f = 0$, and

$$u_b(x,y) = \begin{cases} 0 & \text{if } x = 1 \text{ or } y = 0, \\ \sin\dfrac{(b_2\,x - b_1\,y)\,\pi}{b_2 - b_1} & \text{otherwise.} \end{cases}$$

Note that now u_b is continuous. We shall consider the triangulation of Fig. 2 with $h_1 = h_2 = 1/20$. Figure 9 shows that the SUPG solution obtained for τ defined by (8) contains large spurious oscillations whereas a good approximation of the exact solution is obtained for τ defined by the algorithm in Fig. 7.

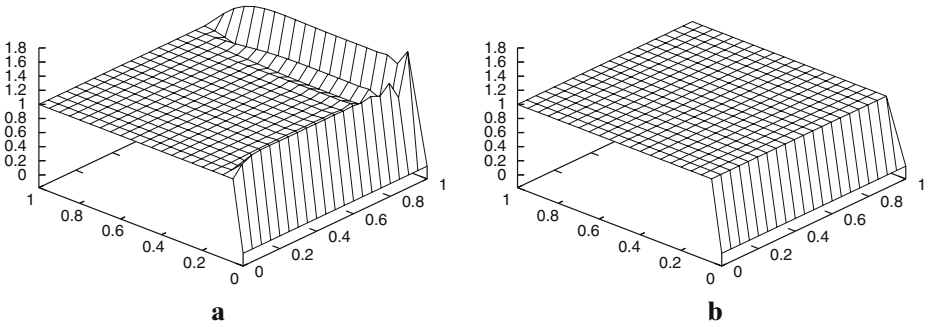


Fig. 8 Example 2, SUPG solutions computed on the triangulation from Fig. 1a with $h_1 = h_2 = 1/20$: **a** τ defined by (8) **b** τ defined by Fig. 7

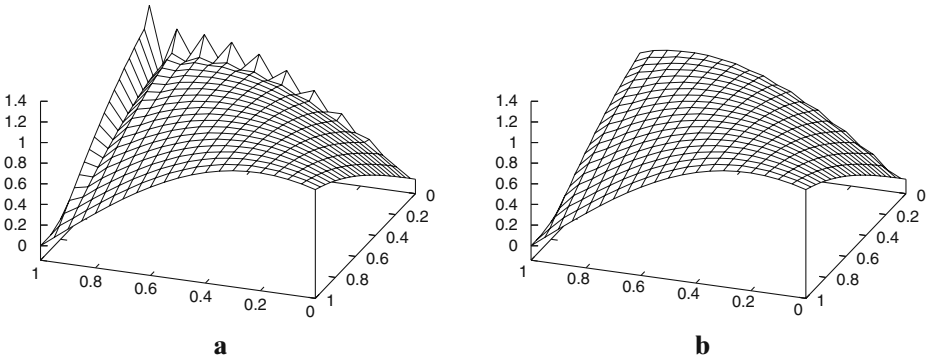


Fig. 9 Example 3, SUPG solutions computed on the triangulation from Fig. 2 with $h_1 = h_2 = 1/20$: **a** τ defined by (8) **b** τ defined by Fig. 7

Example 4 We consider the equation (1) and the boundary conditions (6) with $\Omega = (0, 1)^2$, $\partial\Omega^D = \partial\Omega$, $\partial\Omega^N = \emptyset$, $\varepsilon = 10^{-7}$, $u_b = 0$, and

$$\boldsymbol{b}(x, y) = (-y^3 + 2\,y + 1,\ 2\,x^2 - 3\,x + 2)\,, \qquad f(x, y) = \frac{\cos(x - y)}{1 + x + y}\,.$$

Using a triangulation of the type in Fig. 1b with $h_1 = h_2 = 1/20$, we obtain the discrete solutions depicted in Fig. 10. The solution corresponding to τ defined by (8) again contains large spurious oscillations. These oscillations almost completely disappear if τ is defined by the algorithm in Fig. 7 although u, $\boldsymbol{b}$ and f are nonlinear in $\mathbf{x}$. To compute τ from (8), we replaced $\boldsymbol{b}|_K$ by its value at the barycentre of K. The terms from the discrete problem (7) were evaluated by means of quadrature formulas which were exact for piecewise linear $\boldsymbol{b}$ and piecewise cubic f. A more precise integration does not lead to any visible difference in the computed solution.

Example 5 We consider the equation (1) and the boundary conditions (6) with

$$\Omega = \left\{(x, y) \in (-1, 1) \times (0, 1)\,;\ x^2 + (y - 1)^2 > \tfrac{1}{4}\right\}\,,$$
$$\partial\Omega^N = \left\{\left(-1, -\tfrac{1}{2}\right) \cup \left(\tfrac{1}{2}, 1\right)\right\} \times \{1\}\,, \qquad \partial\Omega^D = \partial\Omega \setminus \partial\Omega^N\,,$$

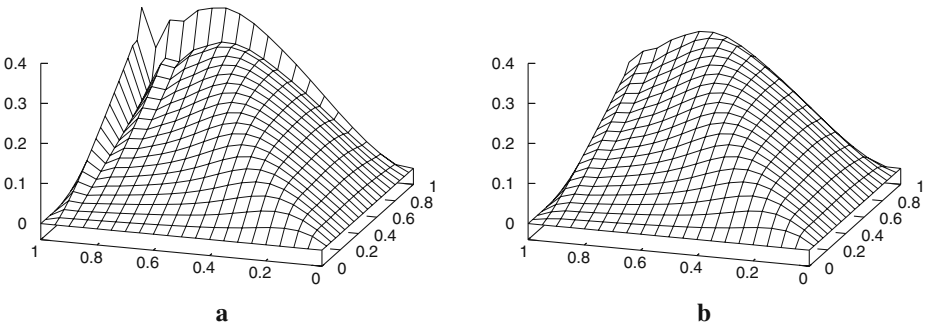


Fig. 10 Example 4, SUPG solutions computed on the triangulation from Fig. 1b with $h_1 = h_2 = 1/20$: **a** τ defined by (8) **b** τ defined by Fig. 7

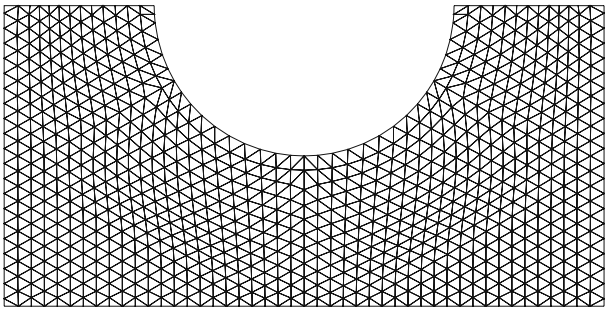


Fig. 11 Triangulation used in Example 5

$\varepsilon = 10^{-7}$, $\boldsymbol{b} = (0, 1)$, $f = 0$, and

$$u_b(x, y) = \begin{cases} 1 & \text{if } x^2 + (y-1)^2 = \frac{1}{4}, \\ 0 & \text{otherwise.} \end{cases}$$

We use the triangulation depicted in Fig. 11. This example demonstrates that the algorithm in Fig. 7 can be successfully applied also when the outflow boundary is curved. The SUPG solution shown in Figs. 12b and 13b is not completely oscillation–free but the spurious oscillations are significantly smaller than for τ defined by (8), see Figs. 12a and 13a.

Example 6 We consider the equation (1) and the boundary conditions (6) with $\Omega = (0, 1)^2$, $\partial\Omega^D = \partial\Omega$, $\partial\Omega^N = \emptyset$, $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (2, 3)$, and the function f chosen in such a way that

$$u(x, y) = x\,y^2 - y^2 \exp\left(\frac{2\,(x-1)}{\varepsilon}\right) - x \exp\left(\frac{3\,(y-1)}{\varepsilon}\right) + \exp\left(\frac{2\,(x-1) + 3\,(y-1)}{\varepsilon}\right)$$

is the exact solution of (1), and with $u_b = u|_{\partial\Omega}$.

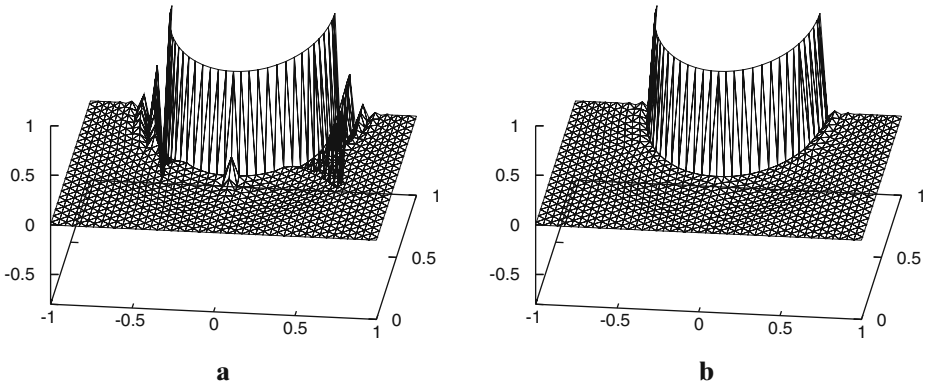


Fig. 12 Example 5, SUPG solutions computed on the triangulation from Fig. 11: **a** τ defined by (8) **b** τ defined by Fig. 7

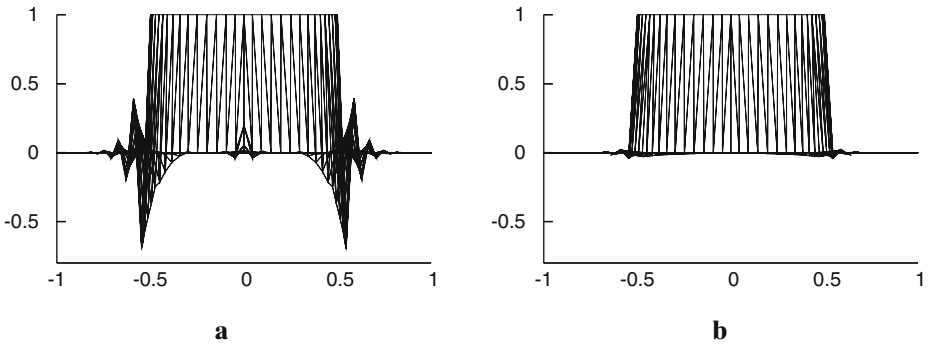


Fig. 13 Example 5, another view of the solutions from Fig. 12 (**a**, **b**)

The function u contains two typical outflow boundary layers and hence this example represents a suitable tool for gauging the accuracy of numerical methods for the solution of convection–diffusion problems. In [7], we used this example for investigating the SUPG method with τ defined by (8) on a sequence of triangulations of the type depicted in Fig. 1b with $h_1 = h_2 \equiv h$. The accuracy of the discrete solutions u_h was measured in various norms and it turned out that away from the boundary layers the discrete solutions are rather accurate and converge to u with the usual optimal convergence rates. However, along the outflow boundary layers, the discrete solutions contain large spurious oscillations (see Fig. 14a for $h = 1/20$) and the magnitude of these oscillations does not decrease for decreasing h as long as $h \gg \varepsilon$. This can be deduced from Table 1 where the second column contains values of the discrete maximum norm $\|u - u_h\|_{0,\infty,h}$ defined as the maximum of the absolute values of the error $u - u_h$ at the vertices of the triangulation. We observe that $\|u - u_h\|_{0,\infty,h}$ even increases slightly if the triangulations are refined.

Defining τ by the algorithm in Fig. 7, the discrete solutions have the same accuracy away from layers as for τ defined by (8), provided that the triangulations are sufficiently fine so that no spurious oscillations occur in the region on which norms of the errors of the discrete solutions are computed. However, in contrast to discrete

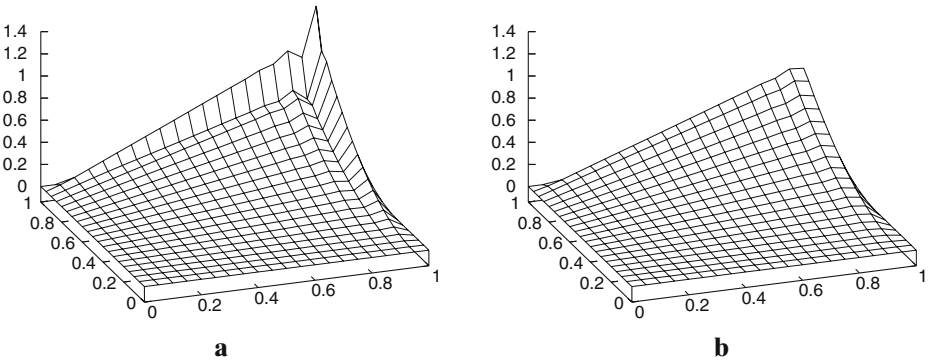


Fig. 14 Example 6, SUPG solutions computed on the triangulation from Fig. 1b with $h_1 = h_2 = 1/20$: **a** τ defined by (8) **b** τ defined by Fig. 7

Table 1 Example 6, errors of SUPG solutions u_h computed for τ defined by (8) or by Fig. 7 on triangulations from Fig. 1b with $h_1 = h_2 \equiv h$

	$\|u - u_h\|_{0,\infty,h}$		$\|u - u_h\|^*_{0,\infty,h}$	
τ	(8)	Fig. 7	(8)	Fig. 7
h=5.000−2	5.08−1	5.48−2	9.37−3	2.45−3
h=2.500−2	5.70−1	2.90−2	2.32−4	6.28−5
h=1.250−2	6.02−1	1.49−2	7.06−6	6.97−6
h=6.250−3	6.18−1	7.54−3	1.74−6	1.74−6
h=3.125−3	6.27−1	3.80−3	4.35−7	4.35−7
Conv. order	−0.02	0.99	2.00	2.00

solutions obtained for τ defined by (8), the discrete maximum norm $\|u - u_h\|_{0,\infty,h}$ now converges linearly to zero for decreasing h even when $h \gg \varepsilon$, see the third column of Table 1. The values of $\|u - u_h\|_{0,\infty,h}$ indicate that the large oscillations visible in Fig. 14a are not present in the SUPG solution obtained for τ defined by Fig. 7, see also Fig. 14b.

In Table 1 we also show values of the discrete maximum norm $\|u - u_h\|^*_{0,\infty,h}$ defined as the maximum of $|u - u_h|$ at those vertices of the triangulation contained in the set $[0, 0.8]^2$. This set does not include a neighbourhood of the layers. As mentioned above, for both choices of τ, the values of $\|u - u_h\|^*_{0,\infty,h}$ are the same if h is sufficiently small and the convergence rate is the optimal second order. The convergence orders in Table 1 are computed from the values for $h = 6.25 \cdot 10^{-3}$ and $h = 3.125 \cdot 10^{-3}$.

7 Conclusions

In this paper we discussed the properties of the SUPG finite element method applied to the numerical solution of two–dimensional steady scalar convection–diffusion equations. We concentrated on the choice of the SUPG stabilization parameter τ along outflow Dirichlet boundaries where the exact solution typically contains boundary layers. Our discussion concentrated on the case of conforming piecewise linear triangular finite elements. We demonstrated that in general an oscillation–free SUPG solution cannot be obtained if, on each triangle of the triangulation, the definition of τ uses only the information available on that triangle. Therefore, we proposed a new approach for defining τ on triangles intersecting an outflow Dirichlet boundary. On any such triangle K, the value $\tau|_K$ depends not only on K and the convection vector $\boldsymbol{b}|_K$ but also on the shape and orientation of triangles K' and convection vectors $\boldsymbol{b}|_{K'}$ in a neighbourhood of K. Numerical results show a significant reduction of spurious oscillations in discrete solutions in comparison to standard choices of τ, while accuracy away from layers is preserved. For simple model problems, even nodally exact solutions are obtained.

Acknowledgements This research is a part of the project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Academy of Sciences of the Czech Republic under the grant No. IAA100190505.

References

1. Bazilevs, Y., Hughes, T.J.R.: Weak imposition of Dirichlet boundary conditions in fluid mechanics. Comput. & Fluids **36**, 12–26 (2007)
2. 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 Appl. Mech. Engrg. **32**, 199–259 (1982)
3. Christie, I., Griffiths, D.F., Mitchell, A.R., Zienkiewicz, O.C.: Finite element methods for second order differential equations with significant first derivatives. Internat. J. Numer. Methods Engrg. **10**, 1389–1396 (1976)
4. Ciarlet, P.G.: Basic error estimates for elliptic problems. In: Ciarlet, P.G., Lions, J.L. (eds.) Handbook of Numerical Analysis, vol. 2, pp. 17–351. Finite Element Methods (pt. 1). North–Holland, Amsterdam (1991)
5. Grasman, J.: On the Birth of Boundary Layers. Mathematical Centre Tracts, vol. 36. Mathematical Centre, Amsterdam (1971)
6. John, V., Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part I – a review. Comput. Methods Appl. Mech. Engrg. **196**, 2197–2215 (2007)
7. Knobloch, P.: Improvements of the Mizukami–Hughes method for convection–diffusion equations. Comput. Methods Appl. Mech. Engrg. **196**, 579–594 (2006)
8. Madden, N., Stynes, M.: Linear enhancements of the streamline diffusion method for convection–diffusion problems. Comput. Math. Appl. **32**, 29–42 (1996)
9. Madden, N., Stynes, M.: Efficient generation of oriented meshes for solving convection–diffusion problems. Internat. J. Numer. Methods Engrg. **40**, 565–576 (1997)
10. Roos, H.–G., Stynes, M., Tobiska, L.: Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion and Flow Problems, 2nd edn. Springer, Berlin (2008)

Electronic Transactions on Numerical Analysis.
Volume 32, pp. 76-89, 2008.

ISSN 1068-9613.

ETNA
Kent State University
etna@mcs.kent.edu

ON THE DEFINITION OF THE SUPG PARAMETER*

PETR KNOBLOCH†

Abstract. We consider the SUPG finite element method for two–dimensional steady scalar convection–diffusion equations and discuss a recently introduced definition of the SUPG stabilization parameter along outflow Dirichlet boundaries for problems containing interior layers.

Key words. convection–diffusion equations, singularly perturbed problems, boundary layers, spurious oscillations, streamline upwind/Petrov–Galerkin (SUPG) method, SOLD methods

AMS subject classifications. 65N30

1. Introduction.

In many applications, transport processes are the main mechanism determining distributions of the observed physical quantities. Often, the distributions of some of the quantities are not smooth and contain narrow regions where the quantities change abruptly. Depending on the application, one speaks about layers, shocks or discontinuities. When approximating such quantities numerically, the width of the regions where shocks or layers occur is often much smaller than the resolution of the used mesh. Consequently, the shocks or layers cannot be resolved properly, which usually leads to unwanted spurious (nonphysical) oscillations in the numerical solution. The attenuation of these oscillations has been the subject of extensive research for several decades during which a huge number of so–called stabilized methods have been developed. The stabilizing effect can be often interpreted as the addition of some artificial diffusion to a standard (unstable) numerical scheme. On the one hand, this artificial diffusion should damp the oscillations but, on the other hand, it should not smear the numerical solution. Therefore, the design of a proper stabilization is a very difficult task.

In the context of finite element methods, a very popular stabilization technique is the streamline upwind/Petrov–Galerkin (SUPG) method. This method was introduced by Brooks and Hughes [1] for advection–diffusion equations and incompressible Navier–Stokes equations. Later this technique has been applied to various other problems, e.g., coupled multidimensional advective–diffusive systems [8], first–order linear hyperbolic systems [12] or first–order hyperbolic systems of conservation laws [9]. Because of its structural simplicity, generality and the quality of numerical solutions, the SUPG method has attracted considerable attention over the last two decades and many theoretical and computational results have been published. It is not the aim of this paper to provide a review of these results and we only refer to the monograph [16].

For simplicity, we shall confine ourselves to a steady scalar convection–diffusion equation

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

We assume that Ω is a bounded domain in $\mathbb{R}^2$ with a polygonal boundary $\partial\Omega$, $\varepsilon > 0$ is the constant diffusivity, $\boldsymbol{b}$ is a given convective field, f is an outer source of u, and u_b represents the Dirichlet boundary condition. In the convection–dominated case $\varepsilon \ll |\boldsymbol{b}|$, the solution u

*Received January 7, 2008. Accepted for publication June 23, 2008. Published online on January 20, 2009. Recommended by A. Rösch. This work is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Czech Republic under the grant No. 201/08/0012.

†Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 186 75 Praha 8, Czech Republic (knobloch@karlin.mff.cuni.cz).

ETNA
Kent State University
etna@mcs.kent.edu

typically contains interior and boundary layers. These layers can be divided into characteristic (interior and boundary) layers and outflow boundary layers; see [16].

The SUPG method produces accurate and oscillation–free solutions in regions where no abrupt changes in the solution of (1.1) occur but it does not preclude spurious oscillations (overshooting and undershooting) localized in narrow regions along sharp layers. The magnitude of these oscillations strongly depends on the SUPG stabilization parameter. Unfortunately, a general 'optimal' choice of this parameter is not known. Theoretical investigations of model problems only provide asymptotic behaviour of this parameter (with respect to the mesh width) and certain bounds for which the SUPG method is stable and leads to (quasi–) optimal convergence of the discrete solution. However, it has been reported many times that the choice of the stabilization parameter inside these bounds may dramatically influence the accuracy of the discrete solution.

Recently, a new definition of the SUPG stabilization parameter on elements intersecting an outflow Dirichlet boundary was proposed in [13]. In contrast to other approaches, the parameter on a given element depends on the shape and orientation of neighbouring elements and the convection vector $\boldsymbol{b}$ on these elements. Numerical results in [13] show a significant reduction of spurious oscillations in SUPG solutions in comparison to usual choices of the stabilization parameter while accuracy away from layers is preserved. For simple model problems, even nodally exact solutions are obtained.

The aim of this paper is to discuss the application of the new stabilization parameter to problems involving both boundary and interior layers. Since the choice of the stabilization parameter at interior layers has only a limited influence on the spurious oscillations appearing in these regions (see, e.g., [14]), we shall also apply the discontinuity–capturing crosswind–dissipation method [6] as an additional stabilization. We shall demonstrate that the combination of the new definition of the SUPG stabilization parameter and the discontinuity–capturing crosswind–dissipation method provide fairly satisfactory approximations of solutions to (1.1). Furthermore, we shall show how the quality of a SUPG solution can be improved by small modifications of the mesh.

The plan of the paper is as follows. Section 2 formulates the SUPG method and Section 3 describes the discontinuity–capturing crosswind–dissipation method as an example of spurious oscillations at layers diminishing (SOLD) methods. In Section 4 the SUPG stabilization parameter of [13] is briefly introduced. Section 5 compares this definition of the stabilization parameter with an approach by Madden and Stynes. Finally, various numerical results for problems involving interior layers are presented in Sections 6 and 7. The paper is closed by conclusions in Section 8. Throughout the paper, we use the standard notations $P_1(\Omega)$, $L^2(\Omega)$, $H^1(\Omega) = W^{1,2}(\Omega)$, etc., for the usual function spaces; see, e.g., [5]. For a vector $\boldsymbol{a} \in \mathbb{R}^2$, we denote by $|\boldsymbol{a}|$ its Euclidean norm.

2. The SUPG method. Let $\mathcal{T}_h$ be a triangulation of the domain Ω consisting of a finite number of open triangular elements K. Further, 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.

We define the finite element spaces

$$W_h = \{v \in H^1(\Omega);\ v|_K \in P_1(K) \ \ \forall\, K \in \mathcal{T}_h\}\,, \qquad V_h = W_h \cap H_0^1(\Omega)\,.$$

Denoting by $u_{bh} \in W_h$ a function whose trace approximates the boundary condition u_b, the SUPG method for the convection–diffusion equation (1.1) reads:

Find $u_h \in W_h$, such that $u_h - u_{bh} \in V_h$ and

$$\varepsilon\,(\nabla u_h, \nabla v_h) + (\boldsymbol{b}\cdot\nabla u_h, v_h + \tau\,\boldsymbol{b}\cdot\nabla v_h) = (f, v_h + \tau\,\boldsymbol{b}\cdot\nabla v_h) \qquad \forall\, v_h \in V_h\,, \tag{2.1}$$

ETNA
Kent State University
etna@mcs.kent.edu

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^2$ and τ is a nonnegative stabilization parameter.

The choice of τ significantly influences the quality of the discrete solution and therefore it has been a subject of an extensive research over the last three decades; see, e.g., the review in the recent paper [10]. Nevertheless, the definitions of τ mostly rely on heuristic arguments and a general 'optimal' way of choosing τ is still not known. Often, the parameter τ is defined, on any element $K \in \mathcal{T}_h$, by

$$\tau|_K = \frac{h_K}{2\,|\boldsymbol{b}|}\left(\coth Pe_K - \frac{1}{Pe_K}\right), \qquad \text{with} \qquad Pe_K = \frac{|\boldsymbol{b}|\,h_K}{2\,\varepsilon}, \tag{2.2}$$

where h_K is the element diameter in the direction of the convection vector $\boldsymbol{b}$. Various justifications of this formula can be found in [10]. Note that, generally, the parameters h_K, Pe_K and $\tau|_K$ are functions of the points $\mathbf{x} \in K$.

3. SOLD methods. In the convection–dominated regime, the SUPG solutions typically contain oscillations in layer regions. Therefore, various stabilizing terms have been proposed to be added to the SUPG discretization in order to obtain discrete solutions in which the local oscillations are suppressed. In [10, 11], such techniques are called spurious oscillations at layers diminishing (SOLD) methods. Other names are shock–capturing methods or discontinuity–capturing methods.

A review of most SOLD methods published in the literature can be found in [10]. According to the numerical and analytical studies in [10, 11], one of the best SOLD methods is a modification of the discontinuity–capturing crosswind–dissipation method by Codina [6] proposed in [10]. This method adds the term

$$(\widetilde{\varepsilon}\,\boldsymbol{b}^\perp \cdot \nabla u_h, \boldsymbol{b}^\perp \cdot \nabla v_h), \qquad \text{with} \qquad \boldsymbol{b}^\perp = \frac{(-b_2, b_1)}{|\boldsymbol{b}|} \tag{3.1}$$

to the left–hand side of the SUPG discretization (2.1) and hence introduces an additional artificial diffusion in the crosswind direction (in the three–dimensional case, the operator $\boldsymbol{b}^\perp \cdot \nabla$ is replaced by the projection of ∇ into the plane orthogonal to $\boldsymbol{b}$). The parameter $\widetilde{\varepsilon}$ is defined, for any $K \in \mathcal{T}_h$, by

$$\widetilde{\varepsilon}|_K = \max\left\{0, C\,\frac{\operatorname{diam}(K)\,|R_h(u_h)|}{2\,|\nabla u_h|} - \varepsilon\right\}, \tag{3.2}$$

where $R_h(u) = \boldsymbol{b} \cdot \nabla u - f$ is the residual and C is a suitable constant. Codina [6] recommended to set $C \approx 0.7$ for linear finite elements and this value was also used in the computations presented in Sections 6 and 7. If $\nabla u_h = 0$ in (3.2), we set $\widetilde{\varepsilon} = 0$. For $f = 0$ (which will be the case in the examples presented in this paper), $\widetilde{\varepsilon}$ is equal to the parameter proposed by Codina [6]. Note that $\widetilde{\varepsilon}$ depends on the unknown discrete solution u_h and hence the resulting method is nonlinear.

There are also SOLD terms for which the validity of the discrete maximum principle can be proved; see, e.g., [2, 3]. Unfortunately, such methods do not attain the quality of the above mentioned method by Codina since they usually lead to considerable smearing of layers; cf., e.g., [10]. Moreover, it is often very difficult to compute the solution of the nonlinear discrete problem.

Numerical tests in [10] revealed that the SOLD methods significantly improve the quality of a SUPG solution only if the SUPG method adds enough artificial diffusion in the streamline direction. This showed the necessity to reconsider the definition of the SUPG stabilization parameter.

ETNA
Kent State University
etna@mcs.kent.edu

4. SUPG stabilization parameter defined using patches of elements. It was demonstrated in [13] that the information available on a particular element of the triangulation is not sufficient for defining the stabilization parameter τ in an optimal way and that the orientation of the neighbouring elements has to be taken into account. Therefore, a new definition of τ appropriate for elements lying at an outflow Dirichlet boundary was proposed in [13] employing information on patches of elements.

Let us mention that an appropriate definition of τ at outflow Dirichlet boundaries is important also in real–life applications although sometimes it is claimed that outflow boundary layers are mainly encountered in academic problems. Of course, it is true that, in computational fluid dynamics (CFD), outflow boundaries are often artificial boundaries at which no layers occur. However, also in CFD applications, outflow boundary layers may occur when problems with moving boundaries are considered. Moreover, there are many other applications leading to convection–diffusion equations whose solutions possess outflow boundary layers in the sense considered in this paper although the vector $\boldsymbol{b}$ often cannot be interpreted as convection. For example, magnetohydrodynamical pipe flow may lead to the convection–diffusion equation (1.1) with $\boldsymbol{b} = (1,0)$ and $u_b = 0$; cf., e.g., [7]. In this case, Ω is the cross–section of the pipe and the parameter ε is the reciprocal of the Hartmann number so that it can be very small.

Let us introduce the outflow Dirichlet boundary

$$\Gamma = \overline{\{\mathbf{x} \in \partial\Omega\,;\ (\boldsymbol{b}\cdot\boldsymbol{n})(\mathbf{x}) > 0\}}\,,$$

where $\boldsymbol{n}$ is the unit outward normal vector to $\partial\Omega$. For simplicity, we assume that Γ is connected and consists of whole boundary edges of $\mathcal{T}_h$. We set

$$G_h = \text{interior} \bigcup_{K\in\mathcal{G}_h} \overline{K}, \qquad \text{where} \qquad \mathcal{G}_h = \{K \in \mathcal{T}_h\,;\ \overline{K}\cap\Gamma \neq \emptyset\},$$

and denote by $\varphi_1, \ldots, \varphi_{M_h}$ all standard basis functions of V_h satisfying

$$\operatorname{supp}\varphi_i \cap G_h \neq \emptyset, \qquad i = 1, \ldots, M_h\,.$$

For $i = 1, \ldots, M_h$, let $\mathbf{x}_i$ be the vertex associated with the basis function φ_i, i.e., $\varphi_i(\mathbf{x}_i) = 1$ and $\varphi_i(\mathbf{x}) = 0$ for any vertex $\mathbf{x} \neq \mathbf{x}_i$.

The idea of defining τ is to require that

$$\int_{G_h} v_h + \tau\,\boldsymbol{b}\cdot\nabla v_h \,\mathrm{d}\mathbf{x} = 0 \qquad \forall\ v_h \in V_h\,,$$

which can be equivalently written in the form

$$\int_{G_h} \varphi_i + \tau\,\boldsymbol{b}\cdot\nabla\varphi_i \,\mathrm{d}\mathbf{x} = 0, \qquad i = 1, \ldots, M_h\,.$$

This suppresses the influence of the Dirichlet boundary condition onto the values of the SUPG solution at interior vertices near Γ. In other words, it increases the upwind character of the method near Γ. The efficiency of this approach also depends on the used triangulation, in particular, on its alignment with the boundary.

To obtain a method which is applicable also for small values of the Péclet number, we set, on any element $K \in \mathcal{G}_h$,

$$\tau|_K = \tau_0|_K \left(\coth Pe_K - \frac{1}{Pe_K}\right), \qquad \text{with} \qquad Pe_K = \frac{|\boldsymbol{b}_K|\,h_K}{2\,\varepsilon}\,, \tag{4.1}$$

ETNA
Kent State University
etna@mcs.kent.edu

where

$$\boldsymbol{b}_K = \frac{1}{|K|} \int_K \boldsymbol{b} \, \mathrm{d}\mathbf{x}$$

and τ_0 is a piecewise constant function on G_h satisfying

$$\int_{G_h} \varphi_i + \tau_0 \, \boldsymbol{b} \cdot \nabla \varphi_i \, \mathrm{d}\mathbf{x} = 0, \qquad i = 1, \ldots, M_h \,. \tag{4.2}$$

On elements $K \in \mathcal{T}_h \setminus \mathcal{G}_h$, we define τ by (2.2) with $\boldsymbol{b}$ replaced by $\boldsymbol{b}_K$.

The relations (4.2) do not determine τ_0 uniquely. However, it was shown in [13] that there always exists τ_0, such that (4.2) holds at least for vertices $\mathbf{x}_i$ which are not contained in elements sharing with Γ only the end points of Γ. A detailed algorithm for computing τ_0 applicable to general triangulations can be found in [13]. Here we mention only the basic idea.

Since it is generally not possible to fulfil (4.2) elementwise, we first determine τ_0 on elements having only one vertex on Γ. Consider any vertex $\mathbf{z} \in \Gamma$ and let $G_{\mathbf{z}}$ be the union of all elements sharing with Γ only the vertex $\mathbf{z}$. If $\mathbf{z}$ is not an end point of Γ, then it is possible to define τ_0 on $G_{\mathbf{z}}$ in such a way that

$$\int_{G_{\mathbf{z}}} \varphi_i + \tau_0 \, \boldsymbol{b} \cdot \nabla \varphi_i \, \mathrm{d}\mathbf{x} = 0$$

at least for all $\mathbf{x}_i$ which are not contained in elements sharing two vertices with Γ. Now it is easy to define τ_0 on elements sharing two vertices with Γ in such a way that (4.2) holds.

5. Comparison with the approach by Madden and Stynes. In some cases the parameter τ defined in the preceding section coincides with the stabilization parameter introduced by Madden and Stynes [14]. In this section, we compare these two choices for the following very simple model problem.

EXAMPLE 5.1. We consider the problem (1.1) with

$$\Omega = (0,1)^2, \qquad \varepsilon = 10^{-8}, \qquad \boldsymbol{b} = (\cos(\pi/3), -\sin(\pi/3)), \qquad f = 0, \tag{5.1}$$

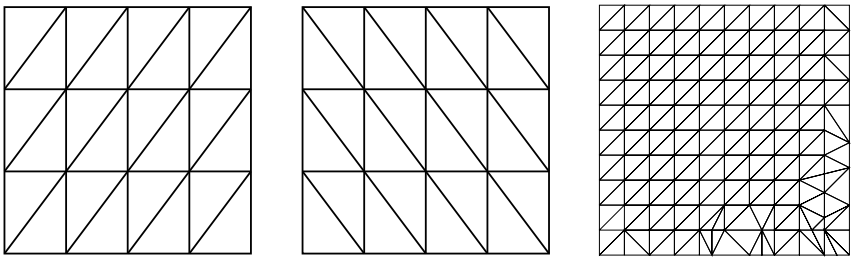
and

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

If we use a triangulation of the type from Figure 5.1(a) with the same mesh width h in both the horizontal and the vertical directions, the SUPG solution with τ defined by (2.2) contains large spurious oscillations along both outflow boundary layers; see [13]. On the other hand, if we define τ as in the preceding section, we obtain a nodally exact solution. This also can be verified by simple theoretical considerations.

Usually, there are many possibilities how to define a piecewise constant function τ_0 satisfying (4.2). Particularly, in the present example, we can use τ_0 which is constant for $x < 1 - 2\,h$ and for $y > 2\,h$. Then $\tau_0 = \frac{1}{2} h/|b_2| = h/\sqrt{3}$ in the former case and $\tau_0 = \frac{1}{2} h/b_1 = h$ in the latter case. These values also can be obtained by the approach of Madden and Stynes [14] who adjusted the SUPG parameter in boundary layer regions in such a way that the artificial diffusion added by the SUPG method in the normal direction to an outflow boundary equals to the optimal value known from the one–dimensional case. Consequently, the approach of Madden and Stynes leads to a discrete solution which is nodally exact except in a small neighbourhood of the corner $(1,0)$.

ETNA
Kent State University
etna@mcs.kent.edu

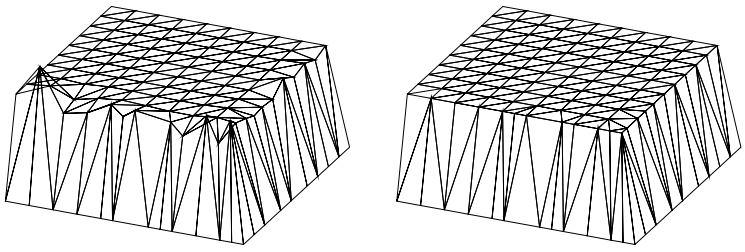


(a) (b) (c)

FIG. 5.1. *Triangulations of the unit square.*

If we use a triangulation which is irregular along the outflow boundary, simple approaches like the one of Madden and Stynes typically do not work properly. As an example, let us consider the triangulation of Figure 5.1(c). Figure 5.2(a) shows that the approach of Madden and Stynes does not give a satisfactory solution, which is due to the fact that the irregular triangulation does not allow to locally reduce the problem to the one–dimensional case. Nevertheless, the solution in Figure 5.2(a) is much better than for τ defined by (2.2). The discrete solution corresponding to τ defined in Section 4 is still nodally exact; see Figure 5.2(b).

A tuning of the SUPG parameter on elements intersecting an outflow boundary was also proposed by do Carmo and Alvarez [4]. However, on uniform triangulations like in Figure 5.1(a), the parameter τ would have the same value on all elements intersecting the outflow boundary, which does not enable the computation of both boundary layers of Example 5.1 sharply.



(a) (b)

FIG. 5.2. *Example 5.1, SUPG solutions computed on the triangulation from Figure 5.1(c): (a) τ defined according to Madden and Stynes [14], (b) τ defined by (4.1).*

6. Example with an interior layer originating from a discontinuous boundary condition. In this section, we investigate the problem of Example 5.1 with another discontinuous boundary condition:

EXAMPLE 6.1. We consider the problem (1.1) with (5.1) and

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

ETNA
Kent State University
etna@mcs.kent.edu

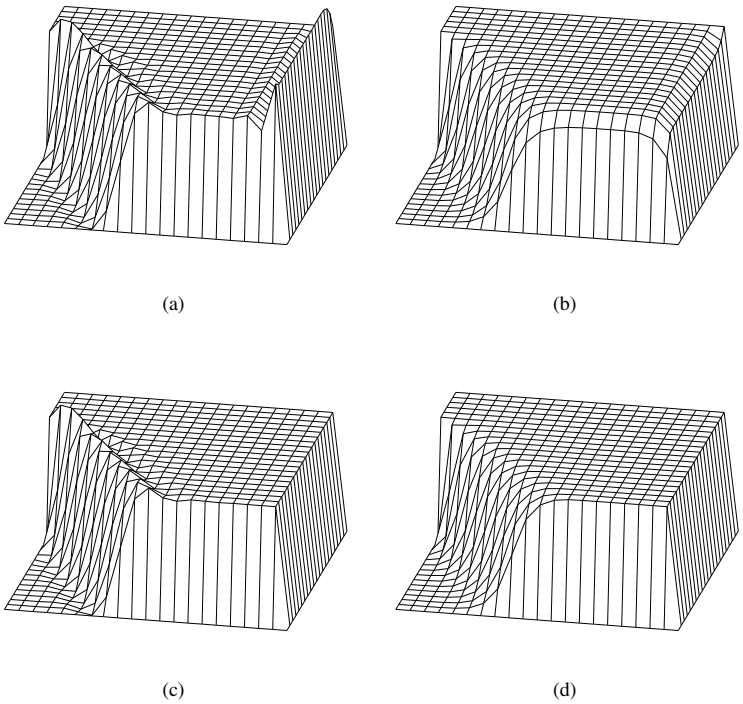


FIG. 6.1. *Example 6.1, 21×21 triangulation of the type from Figure 5.1(b): (a) SUPG method with τ from (2.2), (b) SOLD method with τ from (2.2) and SOLD term (3.1), (3.2), (c) SUPG method with τ defined by (4.1), (d) SOLD method with τ defined by (4.1) and SOLD term (3.1), (3.2) applied away from the boundary layers.*

The solution u possesses an interior (characteristic) layer in the direction of the convection starting at $(0, 0.7)$. On the boundary $x = 1$ and on the right-hand part of the boundary $y = 0$, exponential layers are developed.

To discretize this example, we use a triangulation of Ω of the type shown in Figure 5.1(b) containing 21×21 vertices. If we solve Example 6.1 using the SUPG method with the stabilization parameter τ defined by (2.2), we obtain a solution with spurious oscillations along the interior layer and along the boundary layer at $x = 1$; see Figure 6.1(a). One possibility to suppress these oscillations is to apply a SOLD method. If we add the SOLD term (3.1) with $\widetilde{\varepsilon}$ defined by (3.2) to the SUPG discretization just applied, we obtain the solution depicted in Figure 6.1(b). This solution is oscillation–free, however, the boundary layers are smeared. It is possible to adjust the constant C in (3.2) in such a way that this smearing is avoided; cf. [11]. But, in general, the appropriate value of C is not known. On the other hand, if we apply the SUPG method with τ defined in Section 4, the discrete solution possesses sharp oscillation–free boundary layers; see Figure 6.1(c). Of course, along the interior layer, the solution is the same as in Figure 6.1(a) since we use the same values of τ in this region. The oscillations along the interior layer can be suppressed by using the additional SOLD term (3.1), (3.2). However, since we now know that the parameter τ from Section 4 suppresses oscillations along boundary layers, it suffices to add the SOLD term only on elements which do not intersect the outflow boundary. Then we obtain the oscillation–free solution depicted in Figure 6.1(d) with sharp boundary layers and an acceptable smearing of

ETNA
Kent State University
etna@mcs.kent.edu

the interior layer.

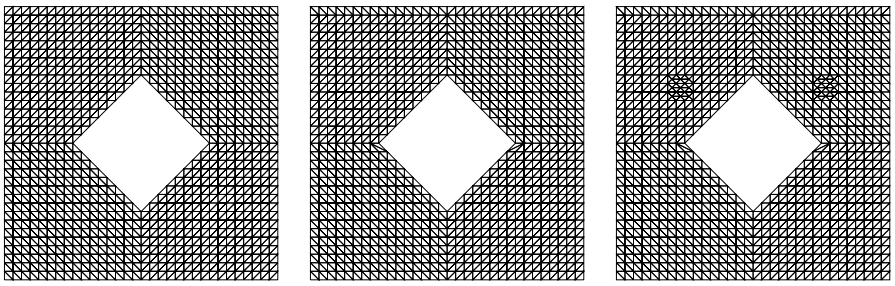
For more general problems and triangulations, we can not guarantee that the SUPG method with τ from Section 4 completely removes oscillations at boundary layers. Nevertheless, numerical results in [13] show that the oscillations are significantly suppressed. Therefore, the SOLD term would be applied also in the boundary layer region but with a much smaller parameter C than in the interior of the computational domain.

Let us mention that it cannot be generally expected that the oscillations along interior (or more generally characteristic) layers will be significantly suppressed by an appropriate choice of the stabilization parameter τ. Indeed, characteristic layers follow the streamlines and the SUPG method contains no mechanism for stabilization in the direction perpendicular to streamlines where spurious oscillations occur. Therefore, an oscillation–free SUPG approximation of a characteristic layer can be obtained only by introducing an additional crosswind diffusion like above or by using a layer–adapted mesh; see, e.g., [15].

7. Examples with interior layers behind an obstacle. In this section, we shall consider the computational domain

$$\Omega = \{(x,y) \in (-1,1)^2 ;\ |x| + |y| > \tfrac{1}{2}\} .$$

Three structured triangulations of Ω which will be discussed in this section are depicted in Figure 7.1. The square hole in Ω can be viewed as an obstacle inside the computational domain. We shall start with the following setting.



(a) (b) (c)

FIG. 7.1. *Structured triangulations of the domain Ω from Section 7.*

EXAMPLE 7.1. We consider the problem (1.1) with the above domain Ω and $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1,2)$, $f = 0$, and

$$u_b(x,y) = \begin{cases} 1 & \text{for } |x| + |y| = \frac{1}{2}, \\ 0 & \text{else.} \end{cases}$$

In view of the boundary conditions, the obstacle inside the flow field gives rise to two interior layers. Moreover, there is a boundary layer at the front part of the obstacle (with respect to the flow) and a boundary layer at a part of the boundary of $(-1,1)^2$ behind the obstacle.

We shall first consider the triangulation depicted in Figure 7.1(a). If we compute an approximation of the solution to Example 7.1 using the SUPG method with τ defined by (2.2), we obtain a solution with spurious oscillations at all four layers; see Figure 7.2(a). An application of the SOLD method (3.1), (3.2) is now not able to suppress the oscillations at $y = 1$

ETNA
Kent State University
etna@mcs.kent.edu

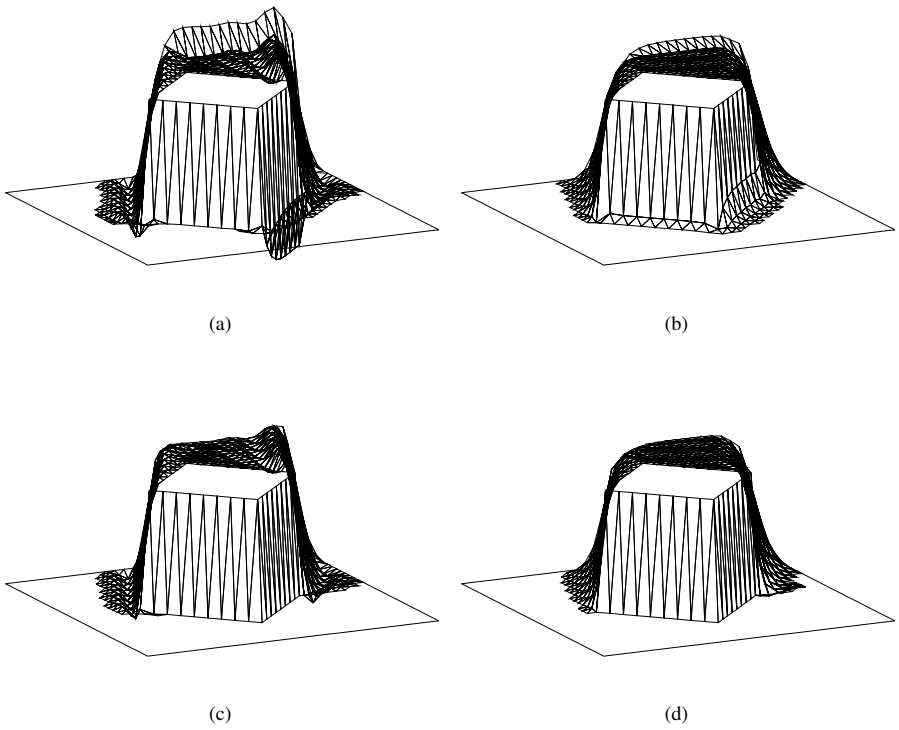


FIG. 7.2. *Example 7.1, triangulation from Figure 7.1(a): (a) SUPG method with τ from (2.2), (b) SOLD method with τ from (2.2) and SOLD term (3.1), (3.2), (c) SUPG method with τ defined by (4.1), (d) SOLD method with τ defined by (4.1) and SOLD term (3.1), (3.2).*

sufficiently; see Figure 7.2(b). At the remaining three layers the oscillations are removed. On the other hand, if we apply the SUPG method with τ defined in Section 4, we obtain sharp approximations of both boundary layers without any oscillations; see Figure 7.2(c). An addition of the SOLD term (3.1), (3.2) removes to a large extent also the oscillations at the interior layers; see Figure 7.2(d). We observe that both boundary layers are approximated significantly better than in case of the solutions from Figures 7.2(a) and 7.2(b).

The results in Figure 7.2 demonstrate that it is essential to define the parameter τ (and also the mesh as we shall see in the following) in such a way that the spurious oscillations in the SUPG solution are as small as possible. Otherwise the addition of a SOLD term cannot be expected to lead to an oscillation–free solution (unless we use a very diffusive method, which typically leads to an excessive smearing of the layers). This is true for all the SOLD methods reviewed and investigated in [10, 11]. Note also that it is generally not possible to remove spurious oscillations at outflow Dirichlet boundaries by simply increasing the parameter τ since the oscillations are influenced not only by the magnitude of τ but also by the relation between values of τ on neighbouring elements. Moreover, such simple approaches are usually not able to suppress spurious oscillations without smearing the layers. Therefore, more complicated definitions of τ, such as the one described in Section 4, seem to be unavoidable.

ETNA
Kent State University
etna@mcs.kent.edu

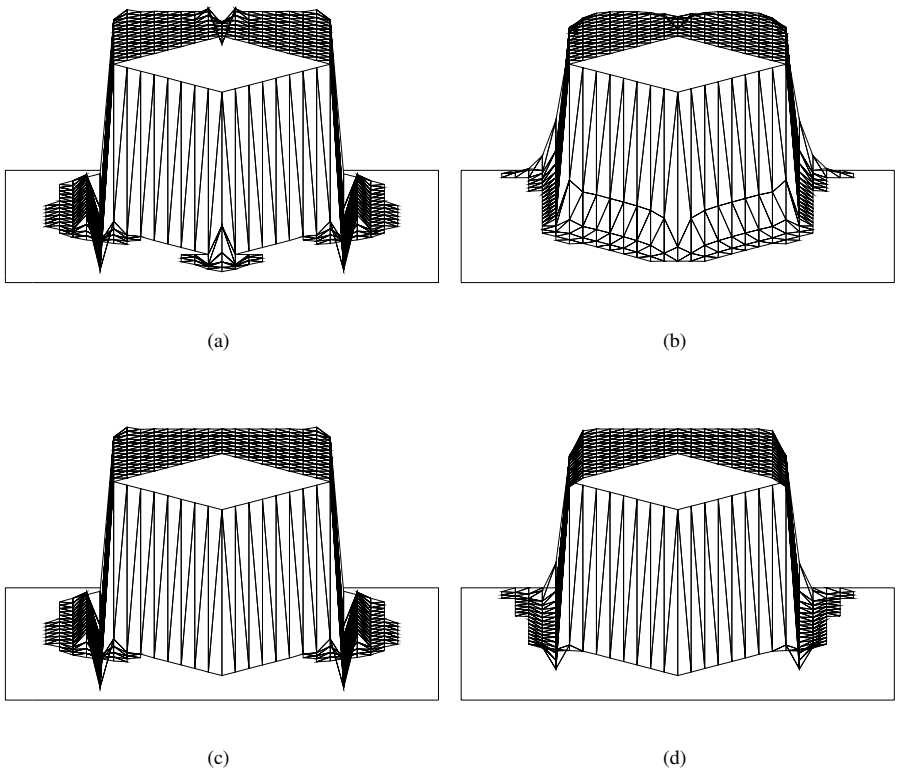


FIG. 7.3. *Example 7.2, triangulation from Figure 7.1(a): (a) SUPG method with τ from (2.2), (b) SOLD method with τ from (2.2) and SOLD term (3.1), (3.2), (c) SUPG method with τ defined by (4.1), (d) SOLD method with τ defined by (4.1) and SOLD term (3.1), (3.2).*

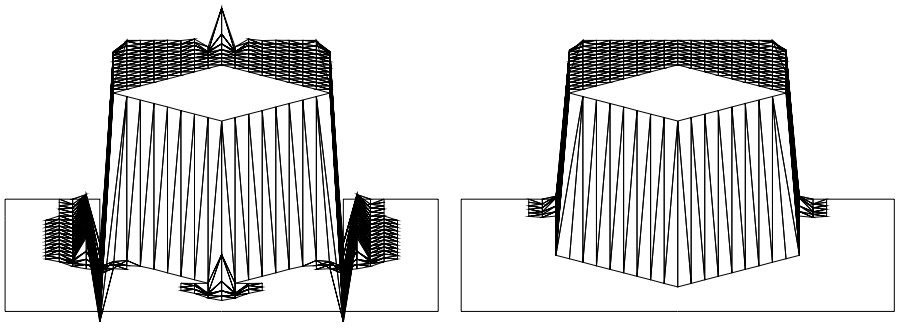
Now we investigate the following simpler situation.

EXAMPLE 7.2. We consider the problem (1.1) with the same data as in Example 7.1 except for $\boldsymbol{b}$ which is defined by $\boldsymbol{b} = (0, 1)$.

In this case the convection and hence also the interior layers are aligned with the mesh and we may expect better properties of discrete solutions. Indeed, the SUPG solution for τ defined by (2.2) (see Figure 7.3(a)) approximates the boundary layers much better and all oscillations can be removed by introducing the SOLD term considered above (see Figure 7.3(b)). The SUPG solution with τ defined by (4.1) now differs from the SUPG solution with τ defined by (2.2) only by better approximations in the middle of boundary layers; see Figure 7.3(c). However, after adding the SOLD term, the difference between the two choices of τ is much larger; cf. Figures 7.3(b) and 7.3(d). For τ defined by (4.1), the approximation of boundary layers is much better but, at the same time, the suppression of oscillations at the beginning of the interior layers is worse. This is probably connected with the fact that the definition of τ from Section 4 tries to assure that the piecewise linear interpolate of u is (at least locally) the solution of the SUPG discretization, which is not allowed by the used triangulation.

Let us now look closer at the oscillations in the SUPG solutions along the interior layers.

ETNA
Kent State University
etna@mcs.kent.edu



(a) (b)

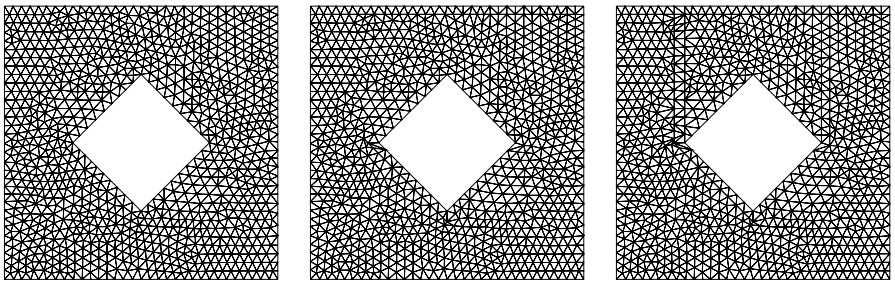
FIG. 7.4. *Example 7.2, triangulation from Figure 7.1(b), SUPG method: (a) τ defined by (2.2), (b) τ defined by (4.1).*

We denote by $\mathbf{x}_\pm = (\pm\frac{1}{2}, 0)$ the two vertices of the obstacle where the interior layers begin and by Γ^o the part of the boundary of the obstacle consisting of points with a nonpositive vertical coordinate. Then Γ^o represents an outflow Dirichlet boundary and $\mathbf{x}_\pm$ are the end points of Γ^o. Furthermore, we denote by G_h^o the set G_h corresponding to Γ^o; see Section 4. A careful inspection of elements near the vertices $\mathbf{x}_\pm$ shows that the interpolant of u cannot be the SUPG solution for any choice of τ. Moreover, the conditions (4.2) cannot be satisfied for some vertices $\mathbf{x}_i$ connected by an edge with $\mathbf{x}_+$ or $\mathbf{x}_-$. To improve the quality of the SUPG solution in a neighbourhood of $\mathbf{x}_+$ (say), we proceed in the following way. First, we denote by $\mathbf{x}_+^i$ the neighbouring vertex of $\mathbf{x}_+$ lying on Γ^o and by $\mathbf{x}_+^e$ the remaining vertex of the element possessing the vertices $\mathbf{x}_+$ and $\mathbf{x}_+^i$. Then we go through the vertices on $\partial G_h^o \setminus \Gamma^o$ in the order in which they are connected by edges, starting with the vertex connected with $\mathbf{x}_+$ by an edge lying on ∂G_h^o, and we find the first vertex $\overline{\mathbf{x}}$ for which the open triangle with the vertices $\mathbf{x}_+$, $\mathbf{x}_+^i$, $\overline{\mathbf{x}}$ lies in Ω and satisfies the required minimal angle condition. If $\overline{\mathbf{x}} \neq \mathbf{x}_+^e$, we change the triangulation in such a way that we delete the edges connecting the vertex $\mathbf{x}_+$ with the vertex $\mathbf{x}_+^e$ and the vertices on ∂G_h^o between $\mathbf{x}_+^e$ and $\overline{\mathbf{x}}$ and we introduce new edges which connect the vertex $\mathbf{x}_+^i$ with the vertex $\overline{\mathbf{x}}$ and the vertices on ∂G_h^o between $\mathbf{x}_+^e$ and $\overline{\mathbf{x}}$. The elements containing any of the vertices lying on ∂G_h^o between $\mathbf{x}_+$ and $\overline{\mathbf{x}}$ are removed from the definition of the set G_h^o. Analogously, we proceed for the vertex $\mathbf{x}_-$. Then, using the algorithm from [13], we can compute a piecewise constant function τ_0 on G_h^o, such that the requirement (4.2) is satisfied. In case of the triangulation from Figure 7.1(a), the described changes of the triangulation concern two elements at each of the vertices $\mathbf{x}_+$ and $\mathbf{x}_-$; see the modified triangulation in Figure 7.1(b). Note that such modifications of the triangulation can be performed a priori in the framework of a computer code.

Further improvements of the SUPG solution can be achieved a posteriori at places where the computer code detects that an interior layer meets a boundary layer. In the present case this happens at the boundary $y = 1$. Here it is desirable to change the direction of the 'diagonal' edges. For simplicity, we made this change along the whole boundary $y = 1$ although it would be sufficient only in a neighbourhood of the interior layer; see again Figure 7.1(b).

On the modified triangulation shown in Figure 7.1(b), the solution of the SUPG method with τ defined by (4.1) is almost nodally exact; see Figure 7.4(b). The only discrepancies appear in the neighbourhood of points where the interior layers meet the boundary $y = 1$. In

ETNA
Kent State University
etna@mcs.kent.edu



(a) (b) (c)

FIG. 7.5. *Unstructured triangulations of the domain* Ω *from Section 7.*

fact, the definition of τ could be modified in such a way that the discrete solution is nodally exact also in these regions, however, such modifications cannot be easily performed in an automatic way in the framework of a computer code. Let us also mention that the SUPG solution for τ defined by (2.2) is worse on the triangulation from Figure 7.1(b) than on the triangulation from Figure 7.1(a); see Figure 7.4(a). Moreover, the SOLD method (3.1), (3.2) is not able to remove the overshoot in the neighbourhood of the point $(0, 1)$.

It should be emphasized that a SUPG solution like in Figure 7.4(b) can be obtained only for special triangulations. As soon as the interior layers will cross elements of the triangulation, like in case of the triangulation in Figure 7.1(c), spurious oscillations will appear and the application of a SOLD method will be necessary.

Finally, let us discuss the application of the techniques treated in this paper on unstructured meshes. We shall consider the triangulation depicted in Figure 7.5(a) which contains approximately the same number of elements as the structured triangulation in Figure 7.1(a). Figures 7.6(a) and 7.6(b) show the SUPG solutions of Example 7.2 for τ defined by (2.2) and (4.1), respectively, computed on this unstructured triangulation and we observe that both solutions contain unacceptable spurious oscillations. In case of τ defined by (4.1), the oscillations are partially caused by the fact that the unstructured triangulation does not satisfy the assumptions used in [13] for deriving the conditions (4.2). More precisely, the triangulation should be constructed in such a way that the part of the boundary of the set G_h lying in Ω copies the outflow boundary Γ. This requirement can be easily satisfied by shifting some of the vertices of the triangulation shown in Figure 7.5(a). In addition, we modify the triangulation in the neighbourhoods of the vertices $\mathbf{x}_{\pm}$ as described above, which leads to the triangulation depicted in Figure 7.5(b). The corresponding SUPG solution with τ defined by (4.1) approximates very well the boundary layers but possesses still spurious oscillations along the interior layers as we can observe in Figure 7.6(c). These oscillations can be removed by aligning edges of the triangulation with the interior layers; see Figures 7.5(c) and 7.6(d). The quality of the triangulation in Figure 7.5(c) could be improved but our aim was only to show that simple shifting of vertices of the triangulation leads to an almost perfect SUPG solution. Let us mention that, for τ defined by (2.2), the magnitude of spurious oscillations in the SUPG solution even increases if the triangulation from Figure 7.5(a) is replaced by the triangulations from Figures 7.5(b) or 7.5(c).

The above results show that the construction or adaptation of the triangulation is very important for the quality of the discrete solution. Although small deviations from an optimal mesh alignment do not lead to a dramatic deterioration of the discrete solution, it is difficult

ETNA
Kent State University
etna@mcs.kent.edu

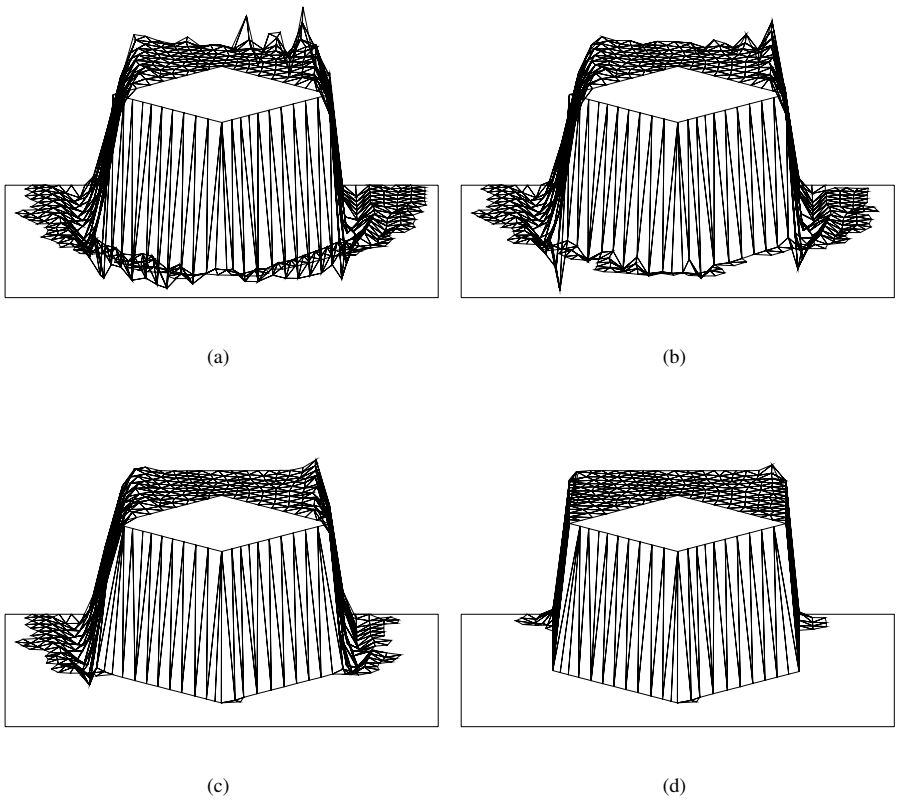


FIG. 7.6. *Example 7.2, SUPG method: (a) τ defined by (2.2), triangulation from Figure 7.5(a), (b) τ defined by (4.1), triangulation from Figure 7.5(a), (c) τ defined by (4.1), triangulation from Figure 7.5(b), (d) τ defined by (4.1), triangulation from Figure 7.5(c).*

to quantify the sensitivity of the discrete solution to the mesh since spurious oscillations are significantly influenced by mutual orientation of neighbouring elements of the triangulation.

Finally, let us mention that it is completely open to what extent the presented techniques can be extended to the three–dimensional case.

8. Conclusions. In this paper, we discussed properties of the SUPG finite element method applied to the numerical solution of two–dimensional steady scalar convection–diffusion equations. We demonstrated that the choice of the SUPG stabilization parameter proposed in [13] together with an application of the discontinuity–capturing crosswind–dissipation method [6] leads to satisfactory discrete solutions in the convection–dominated case. Further numerical results show that the quality of the SUPG solution can be significantly improved if an appropriate mesh is used.

REFERENCES

[1] 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), pp. 199–259.

ETNA
Kent State University
etna@mcs.kent.edu

[2] E. BURMAN AND A. ERN, *Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3833–3855.

[3] E. BURMAN AND A. ERN, *Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652.

[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. Eng., 193 (2004), pp. 2383–2402.

[5] P. G. CIARLET, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, v. 2 – Finite Element Methods (pt. 1), P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.

[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), pp. 325–342.

[7] J. GRASMAN, *On the Birth of Boundary Layers*, Mathematical Centre Tracts 36, Mathematical Centre, Amsterdam, 1971.

[8] T. J. R. HUGHES AND M. MALLET, *A new finite element formulation for computational fluid dynamics: III. The generalized streamline operator for multidimensional advective–diffusive systems*, Comput. Methods Appl. Mech. Engrg., 58 (1986), pp. 305–328.

[9] T. J. R. HUGHES AND T. E. TEZDUYAR, *Finite element methods for first–order hyperbolic systems with particular emphasis on the compressible Euler equations*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 217–284.

[10] V. JOHN AND P. KNOBLOCH, *On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2197–2215.

[11] ———, *On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements*, Comput. Methods Appl. Mech. Engrg., 197 (2008), pp. 1997–2014.

[12] C. JOHNSON, U. NÄVERT, AND J. PITKÄRANTA, *Finite element methods for linear hyperbolic problems*, Comput. Methods Appl. Mech. Engrg., 45 (1984), pp. 285–312.

[13] P. KNOBLOCH, *On the choice of the SUPG parameter at outflow boundary layers*, Adv. Comput. Math., DOI 10.1007/s10444-008-9075-6.

[14] N. MADDEN AND M. STYNES, *Linear enhancements of the streamline diffusion method for convection–diffusion problems*, Comput. Math. Appl., 32 (1996), pp. 29–42.

[15] N. MADDEN AND M. STYNES, *Efficient generation of oriented meshes for solving convection–diffusion problems*, Internat. J. Numer. Methods Engrg., 40 (1997), pp. 565–576.

[16] H.–G. ROOS, M. STYNES, AND L. TOBISKA, *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems*, 2nd edition, Springer–Verlag, Berlin, 2008.

Comput. Methods Appl. Mech. Engrg. 200 (2011) 2916–2929

Contents lists available at ScienceDirect

Comput. Methods Appl. Mech. Engrg.

journal homepage: www.elsevier.com/locate/cma

A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I

Volker John [a,b,*], Petr Knobloch [c,1], Simona B. Savescu [a,2]

[a] Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany
[b] Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany
[c] Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 18675 Praha 8, Czech Republic

ARTICLE INFO

Article history:
Received 15 July 2010
Received in revised form 1 April 2011
Accepted 16 April 2011
Available online 27 April 2011

Keywords:
Stabilized finite element methods
Parameter optimization by minimizing a target functional
SUPG method

ABSTRACT

Stabilized finite element methods for convection-dominated problems require the choice of appropriate stabilization parameters. From numerical analysis, often only their asymptotic values are known. This paper presents a general framework for optimizing stabilization parameters with respect to the minimization of a target functional. Exemplarily, this framework is applied to the SUPG finite element method and the minimization of a residual-based error estimator, an error indicator, and a functional including the crosswind derivative of the computed solution. Benefits of the basic approach are demonstrated by means of numerical results.

1. Introduction

The numerical solution of challenging problems in various engineering applications is in general not possible with standard methods that are based, e.g., on central finite differences or the Galerkin finite element method. More sophisticated schemes become necessary that are designed to tackle the special difficulties of the underlying problem.

An example, that will be considered in this paper, are scalar convection-dominated convection–diffusion equations. Solutions of these equations exhibit very fine structures, so-called layers, which cannot be resolved on meshes that are not extremely fine, at least locally. Standard discretizations lead to solutions that are globally polluted by large spurious oscillations. In practice, stabilized methods are used. These methods introduce artificial diffusion. The difficulty consists now in defining the correct amount of diffusion at the correct positions in the correct directions (anisotropic diffusion) such that numerical solutions with sharp layers and without spurious oscillations are obtained. A method that is optimal with respect to all criteria does not exist yet. Many proposed stabilized methods include so-called stabilization parameters. Often, the asymptotic choice of these parameters is known, e.g., that they should be proportional to the local mesh width. However, in practice, the proportionality factor has to be chosen. There is the experience that different choices of such factors might lead to considerably different numerical solutions. Moreover, the asymptotic choice of the stabilization parameters is based on global stability and convergence analysis. Local features of solutions, like layers, are not taken into account in this analysis.

We would like to mention a second example that demonstrates the difficulties of choosing parameters in numerical simulations – Large Eddy Simulation (LES) of turbulent flows. Turbulent flow simulations require the use of some turbulence model. An often used, so-called eddy viscosity model, is the Smagorinsky model [40]. This model is based on some insight into the physics of turbulent flows and it finally introduces a nonlinear viscosity into the discrete equations. It is rather easy to implement and very well understood from the point of view of mathematical analysis [32]. The derivation of the Smagorinsky model is based on some proportionality relations such that at the end a proportionality factor occurs. Experience shows that the use of a constant for this factor does not lead to good results. Instead, this factor has to be adapted to the local features of the turbulent flow field. An approach in this direction is the dynamic Smagorinsky model [12,33]. Despite all

* Corresponding author at: Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany.
E-mail addresses: volker.john@wias-berlin.de (V. John), knobloch@karlin.mff.cuni.cz (P. Knobloch), simona.b.savescu@wias-berlin.de (S.B. Savescu).

[1] The work of P. Knobloch was supported in part by the Grant Agency of the Academy of Sciences of the Czech Republic under the Grant No. IAA100190804, by the Grant Agency of the Czech Republic under the Grant No. P201/11/1304, and by the Ministry of Education, Youth and Sports of the Czech Republic in the framework of the research project MSM 0021620839.
[2] The work of S.B. Savescu was supported by the Deutsche Forschungsgemeinschaft (DFG), Grant No. Jo 329/9-1.

0045-7825/$ - see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.cma.2011.04.016

drawbacks, e.g., see [24], the dynamic Smagorinsky model is one of the most often used and most successful LES models. Nowadays, there is another approach to control the influence of the Smagorinsky model – Variational Multiscale (VMS) methods. These methods try to select appropriate scales to which this model is applied [20,15,25,26]. Turbulent flow simulations are a typical example where principal forms of models are known but the results obtained with these models depend on the correct setting of parameters. There are many more numerical methods that require the choice of parameters and for which an a posteriori choice would greatly improve the ability to use them in applications. The a posteriori choice of parameters seems to be a widely open and challenging task in scientific computing.

The idea of choosing parameters in numerical methods a posteriori is not new, the dynamic Smagorinsky model was already mentioned. In essence, this method computes two (or more) discrete solutions in different ways and the parameter choice is based on comparing them. This idea was recently carried over to scalar convection–diffusion equations in [1], based on the work from [35]. In this approach, the different solutions are computed on coarser mesh(es). On the coarser meshes, information on the respective stabilization parameters are derived which are used to update the stabilization parameters on the fine mesh. A severe drawback of this approach is that the dimension of the parameter space is not allowed to exceed the dimension of the respective test function space. Therefore, the approach cannot be applied to the optimization of stabilization parameters in discretizations with first order finite elements as considered in this paper. Moreover, the methodology seems to be only simple for a few globally constant parameters, which is explicitly not the goal of our approach. Another method which determines the stabilization parameter on the basis of two solutions was presented in [36]. In this method, the residuals and their derivatives are used to compute a characteristic length scale which enters the formula for the stabilization parameter. The computations of the stabilization parameters in [36] are restricted to convection–diffusion equations in one dimension and a generalization to more dimensions is not obvious. A method for hyperbolic conservation laws in one dimension can be found in [10]. In this paper, the streamline-diffusion stabilization parameter and an adaptively refined grid are computed a posteriori. The adaptive algorithm uses the Dual Weighted Residual (DWR) approach [2,3] with a backward-in-time dual problem. An iterative procedure based on equilibrating components of the error estimator is used to compute the stabilization parameters and the grids. This method was extended to one-dimensional nonlinear convection–diffusion–reaction equations in [18].

The present paper considers the Streamline-Upwind Petrov–Galerkin (SUPG) finite element method for scalar convection-dominated convection–diffusion equations introduced in [21,4]. Although a number of other stabilized finite element methods have been developed in the past decades, the SUPG method is still the standard approach. In essence, this method adds numerical diffusion in streamline direction. The amount of diffusion depends on local stabilization parameters. There are different formulae for these parameters whose asymptotics are the same, see [27] for a discussion of parameter choices. The properties of solutions obtained with the SUPG method are well known: sharp layers at the correct positions are computed, but non-negligible spurious oscillations occur in a vicinity of layers. These oscillations make the use of the SUPG method in applications difficult as they correspond in general to unphysical situations, like negative concentrations. There have been a large number of attempts to improve the SUPG method in order to get rid of these oscillations while preserving its good properties. However, none of these so-called Spurious Oscillations at Layers Diminishing (SOLD) methods turned out to be entirely successful [27,28].

To improve the solutions obtained with the SUPG method, the present paper pursues a different approach than the SOLD methods. It relies on the optimization of the stabilization parameter, however, in contrast to [1,10,18,36,35], the parameter optimization is formulated as minimization of some functional. This is a nonlinear constrained optimization problem that has to be solved iteratively. A key component of this approach consists in the efficient computation of the Fréchet derivative of the functional with respect to the stabilization parameter. This is achieved by utilizing an adjoint problem with an appropriate right-hand side. The aim of the present paper is to provide a new general framework for the optimization of parameters in stabilized methods for convection–diffusion equations and to demonstrate exemplarily the benefits of this approach. A comprehensive discussion of the choice of appropriate target functionals is postponed to the second part of this paper.

The paper is organized as follows. Section 2 presents the equation and the SUPG method. A general approach for computing the Fréchet derivative of a functional that depends on the numerical solution with respect to parameters in the numerical method is presented in Section 3. This approach is applied to the SUPG method in Section 4. Section 5 contains a proof of concept. It is demonstrated that errors to known solutions can be reduced by using as functional the error in some norm. For problems with unknown solutions, Section 6 illustrates the application of the a posteriori parameter choice based on the minimization of a residual-based a posteriori error estimator, an error indicator, and a functional that includes the crosswind derivative of the computed solution. The most important conclusions, open problems, and an outlook are presented in Section 7. Throughout the paper, standard notations are used for usual function spaces and norms, see, e.g., [6]. The notation $(\cdot,\cdot)_G$ with a set $G \subset \mathbb{R}^d,\ d = 1,2,3$, is used for the inner product in the space $L^2(G)$ or $L^2(G)^d$, with $(\cdot,\cdot) = (\cdot,\cdot)_\Omega$.

2. The convection–diffusion problem and its SUPG stabilization

Consider the scalar convection–diffusion problem

$$-\varepsilon\Delta u + \mathbf{b}\cdot\nabla u + cu = f \text{ in } \Omega, \quad u = u_b \text{ on } \Gamma^D, \quad \varepsilon\frac{\partial u}{\partial \mathbf{n}} = g \text{ on } \Gamma^N. \tag{1}$$

Here, $\Omega \subset \mathbb{R}^d,\ d = 2,3$, is a bounded domain with a polyhedral Lipschitz–continuous boundary $\partial\Omega$ and Γ^D, Γ^N are disjoint and relatively open subsets of $\partial\Omega$ satisfying $\text{meas}_{d-1}(\Gamma^D) > 0$ and $\overline{\Gamma^D \cup \Gamma^N} = \partial\Omega$. Furthermore, $\mathbf{n}$ is the outward unit normal vector to $\partial\Omega$, $\varepsilon > 0$ is a constant diffusivity, $\mathbf{b} \in W^{1,\infty}(\Omega)^d$ 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. The usual assumption that

$$c - \frac{1}{2}\text{div}\mathbf{b} \geqslant c_0 \geqslant 0 \tag{2}$$

with a constant c_0 is made. Moreover, it is assumed that

$$\{\mathbf{x} \in \partial\Omega; (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) < 0\} \subset \Gamma^D, \tag{3}$$

i.e., the inflow boundary is a part of the Dirichlet boundary Γ^D.

This paper studies finite element methods for the numerical solution of (1). To this end, (1) is transformed into a variational 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 Γ^D) and let

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

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

$$a(u,v) = (f,v) + (g,v)_{\Gamma^N} \quad \forall v \in V, \tag{4}$$

where

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

In view of (2) and (3), the weak formulation (4) has a unique solution.

Let $\{\mathcal{T}_h\}_h$ be a family of triangulations of Ω parameterized by positive parameters h whose only accumulation point is zero. The triangulations $\mathcal{T}_h$ are assumed to consist of a finite number of open (mapped) polyhedral subsets K of Ω such that $\overline{\Omega} = \bigcup_{K\in\mathcal{T}_h}\overline{K}$ and the closures of any two different sets in $\mathcal{T}_h$ are either disjoint or possess either a common vertex or a common edge or (if $d = 3$) a common face. Further, it is assumed that any edge (face) of $\mathcal{T}_h$ which lies on $\partial\Omega$ is contained either in $\overline{\Gamma^D}$ or in $\overline{\Gamma^N}$.

For each h, 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 is introduced, see, e.g., [6]. Furthermore, for each h, let $\tilde{u}_{bh} \in W_h$ be a function whose trace on Γ^D approximates u_b. Finally, we set $V_h = W_h \cap V$. Then, the Galerkin discretization of (1) reads: Find $u_h \in W_h$ such that $u_h - \tilde{u}_{bh} \in V_h$ and

$$a(u_h, v_h) = (f, v_h) + (g, v_h)_{\Gamma^N} \quad \forall v_h \in V_h. \tag{5}$$

Again, this problem is uniquely solvable. As discussed in the introduction, the Galerkin discretization (5) is inappropriate if convection dominates diffusion since in this case 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 SUPG method [21,4] that is frequently used because of its stability properties, its higher-order accuracy in appropriate norms, and its easy implementation, see, e.g., [37].

The SUPG stabilization depends on a stabilization parameter that will be denoted by y_h in the following. It is assumed that all admissible stabilization parameters are contained in a finite-dimensional space $Y_h \subset L^\infty(\Omega)$. For example, Y_h can consist of piecewise constant functions with respect to the triangulation $\mathcal{T}_h$.

The SUPG discretization of (1) reads: Find $u_h \in W_h$ such that $u_h - \tilde{u}_{bh} \in V_h$ and

$$a(u_h, v_h) + s_h(y_h; u_h, v_h) = (f, v_h) + (g, v_h)_{\Gamma^N} + r_h(y_h; v_h) \quad \forall v_h \in V_h, \tag{6}$$

where

$$s_h(y_h; u_h, v_h) = (-\varepsilon\Delta_h u_h + \mathbf{b}\cdot\nabla u_h + cu_h, y_h\mathbf{b}\cdot\nabla v_h),$$
$$r_h(y_h; v_h) = (f, y_h\mathbf{b}\cdot\nabla v_h).$$

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

A detailed discussion of ways that are used in practice for choosing the stabilization parameter y_h in the case of first order finite elements can be found in [27]. Modifications for higher order finite elements are discussed, e.g., in [7,9,11]. A common choice is, for any mesh cell $K \in \mathcal{T}_h$,

$$y_h|_K = \frac{h_K}{2p|\mathbf{b}|}\xi(Pe_K) \quad \text{with} \quad \xi(\alpha) = \coth\alpha - \frac{1}{\alpha}, \quad Pe_K = \frac{|\mathbf{b}|h_K}{2p\varepsilon}, \tag{7}$$

where h_K is the cell diameter in the direction of the convection vector $\mathbf{b}$, p is the polynomial degree of the local finite element space, and Pe_K is the local Péclet number which determines whether the problem is locally (i.e., within a particular mesh cell) convection dominated or diffusion dominated. Note that, generally, the parameters h_K, Pe_K and $y_h|_K$ are functions of the points $\mathbf{x} \in K$. The evaluation of the cell diameter in the direction of the convection is discussed also in [27].

If (2) holds with $c_0 > 0$, a sufficient condition for the ellipticity of the bilinear form on the left-hand side of (6) in a standard SUPG norm is

$$0 \leqslant y_h(\mathbf{x}) \leqslant \frac{1}{2}\min\left\{\frac{(\mathrm{diam}(K))^2}{\varepsilon c_{\mathrm{inv}}^2}, \frac{c_0}{\|c\|_{0,\infty,K}^2}\right\}, \quad \mathbf{x}\in K, \tag{8}$$

see [37], where $\mathrm{diam}(K)$ denotes the diameter of K, c_{inv} is a constant from the inverse inequality

$$\|\Delta v_h\|_{0,K} \leqslant c_{\mathrm{inv}}[\mathrm{diam}(K)]^{-1}|v_h|_{1,K} \quad \forall v_h \in V_h,$$

and $\|\cdot\|_{0,\infty,K}$ denotes the $L^\infty(K)$ norm. The first term in the minimum in (8) does not appear for P_1 finite elements and for Q_1 finite elements on rectangles since in these cases $\Delta_h v_h = 0$ for all $v_h \in V_h$.

An important class of convection–diffusion problems possesses the properties $\mathrm{div}\mathbf{b} = 0$, e.g., if $\mathbf{b}$ is the velocity field of an incompressible fluid, and $c = 0$. Hence, (2) holds only with $c_0 = 0$. For this class of problems, one can prove the ellipticity of the SUPG bilinear form (in a weaker SUPG norm than for $c_0 > 0$) if

$$0 \leqslant y_h(\mathbf{x}) \leqslant \frac{(\mathrm{diam}(K))^2}{\varepsilon c_{\mathrm{inv}}^2}, \quad \mathbf{x}\in K. \tag{9}$$

For the same reason as above, the bound on the right-hand side of (9) is not needed if P_1 finite elements or Q_1 finite elements on rectangles are used.

In the special case of a constant convection field and a uniform grid, the stabilization parameter given by (7) is the same in all mesh cells, independently of local features of the solution, like layers. This does not seem to be an optimal choice. This paper will present and study an approach for choosing the values of the stabilization parameter locally, based on the minimization of a functional that measures or estimates the accuracy of the computed solution.

3. Optimization of parameters in numerical methods with respect to the minimization of a functional

Let us assume that a numerical method for the solution of (1) is given and let the method depend on a parameter $y_h \in Y_h$. An example is the SUPG method (6). Let $D_h \subset Y_h$ be an open set such that, for any $y_h \in D_h$, the considered method has a unique solution $u_h \in W_h$. To emphasize that u_h depends on y_h, we shall write $u_h(y_h)$ instead of u_h in the following. Let $I_h : W_h \to \mathbb{R}$ be a functional such that

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

represents a measure of the error of the discrete solution $u_h(y_h)$ corresponding to a given parameter y_h. The aim is to compute a parameter $y_h \in D_h$ for which Φ_h attains a minimum on D_h or is near to a minimum (or the infimum) of Φ_h on D_h. This nonlinear minimization problem has to be solved iteratively. Reasonable iterative schemes require at least information on how Φ_h changes if the parameter y_h is changed, i.e., on the Fréchet derivative of Φ_h. An efficient way to compute this derivative is needed. Such a way will be explained in this section.

For any $y_h \in D_h$, it holds $u_h(y_h) = \tilde{u}_h(y_h) + \tilde{u}_{bh}$ with $\tilde{u}_h : D_h \to V_h$. Thus, one does not need to consider the space W_h in the optimization process but can work with the space V_h, which is more convenient.

Denote $\widetilde{I}_h(w_h) = I_h(w_h + \tilde{u}_{bh})$ for any $w_h \in V_h$. Then $\widetilde{I}_h : V_h \to \mathbb{R}$ and

$$\Phi_h(y_h) = \widetilde{I}_h(\tilde{u}_h(y_h)) \quad \forall y_h \in D_h.$$

Let us assume that the mappings $\widetilde{I}_h = \widetilde{I}_h(w_h)$ and $\tilde{u}_h = \tilde{u}_h(y_h)$ are Fréchet–differentiable. The Fréchet derivatives are denoted by $D\widetilde{I}_h : V_h \to V_h'$ and $D\tilde{u}_h : D_h \to \mathcal{L}(Y_h, V_h)$. Then, the Fréchet derivative $D\Phi_h : D_h \to Y_h'$ of Φ_h exists and it is given by

$$D\Phi_h(y_h) = D\widetilde{I}_h(\tilde{u}_h(y_h))D\tilde{u}_h(y_h) \quad \forall y_h \in D_h. \tag{10}$$

The naive way of using this formula for computing $D\Phi_h(y_h)$ is very inefficient as the computation of $D\tilde{u}_h(y_h)$ requires the solution of dim Y_h systems of dim V_h algebraic equations.

The problem of efficiently evaluating a derivative of form (10) is well known, e.g., from optimal control of partial differential equations. There is a way for obtaining this derivative that is based on an appropriate adjoint problem, e.g., see [42]. This way will be applied to the situation considered in this paper. The minimization of Φ_h occurs under the condition that $u_h(y_h)$ should fulfill the discretized partial differential equation (6), i.e., for a residual operator $R_h : V_h \times Y_h \to V_h'$ holds

$$R_h(\tilde{u}_h(y_h), y_h) = 0 \quad \forall y_h \in D_h. \tag{11}$$

For the SUPG method (6), the operator R_h is given by

$$\begin{aligned}\langle R_h(w_h, y_h), v_h\rangle = {} & a(w_h + \tilde{u}_{bh}, v_h) + s_h(y_h; w_h + \tilde{u}_{bh}, v_h) \\ & - (f, v_h) - (g, v_h)_{\Gamma^N} - r_h(y_h; v_h) \\ & \forall v_h, w_h \in V_h,\ y_h \in Y_h.\end{aligned}$$

Differentiating (11) with respect to y_h leads to

$$\partial_w R_h(\tilde{u}_h(y_h), y_h)D\tilde{u}_h(y_h) + \partial_y R_h(\tilde{u}_h(y_h), y_h) = 0 \quad \forall y_h \in D_h, \tag{12}$$

provided that the mapping $R_h = R_h(w_h, y_h)$ is Fréchet–differentiable. Note that $\partial_w R_h : V_h \times Y_h \to \mathcal{L}(V_h, V_h')$ and $\partial_y R_h : V_h \times Y_h \to \mathcal{L}(Y_h, V_h')$. Assume that there is a mapping ψ_h: $D_h \to V_h$ such that

$$\langle D\Phi_h(y_h), \tilde{y}_h\rangle = -\langle(\partial_y R_h)(\tilde{u}_h(y_h), y_h)\tilde{y}_h, \psi_h(y_h)\rangle \quad \forall y_h \in D_h,\ \tilde{y}_h \in Y_h. \tag{13}$$

Then, according to (12), one obtains

$$\begin{aligned}\langle D\Phi_h(y_h), \tilde{y}_h\rangle &= \langle(\partial_w R_h)(\tilde{u}_h(y_h), y_h)D\tilde{u}_h(y_h)\tilde{y}_h, \psi_h(y_h)\rangle \\ &= \langle(\partial_w R_h)'(\tilde{u}_h(y_h), y_h)\psi_h(y_h), D\tilde{u}_h(y_h)\tilde{y}_h\rangle \\ &\quad \forall y_h \in D_h,\ \tilde{y}_h \in Y_h,\end{aligned}$$

where the adjoint operator is defined by

$$\begin{aligned}\langle(\partial_w R_h)'(w_h, y_h)v_h, \tilde{v}_h\rangle = {} & \langle(\partial_w R_h)(w_h, y_h)\tilde{v}_h, v_h\rangle \\ & \forall v_h, \tilde{v}_h, w_h \in V_h,\ y_h \in Y_h.\end{aligned}$$

On the other hand, from (10) follows that

$$\begin{aligned}\langle D\Phi_h(y_h), \tilde{y}_h\rangle &= \langle D\widetilde{I}_h(\tilde{u}_h(y_h))D\tilde{u}_h(y_h), \tilde{y}_h\rangle \\ &= \langle D\widetilde{I}_h(\tilde{u}_h(y_h)), D\tilde{u}_h(y_h)\tilde{y}_h\rangle \qquad \forall y_h \in D_h,\ \tilde{y}_h \in Y_h.\end{aligned}$$

The two representations of $D\Phi_h(y_h)$ suggest to define $\psi_h(y_h)$ as the solution of the adjoint problem, cf., e.g., [13,38],

$$(\partial_w R_h)'(\tilde{u}_h(y_h), y_h)\psi_h(y_h) = D\widetilde{I}_h(\tilde{u}_h(y_h)) \quad \forall y_h \in D_h. \tag{14}$$

Then $\psi_h(y_h)$ satisfies (13) and hence the Fréchet derivative of Φ_h is given by

$$D\Phi_h(y_h) = -(\partial_y R_h)'(\tilde{u}_h(y_h), y_h)\psi_h(y_h) \quad \forall y_h \in D_h. \tag{15}$$

The adjoint operator is defined by

$$\begin{aligned}\langle(\partial_y R_h)'(w_h, y_h)v_h, \tilde{y}_h\rangle = {} & \langle(\partial_y R_h)(w_h, y_h)\tilde{y}_h, v_h\rangle \\ & \forall v_h, w_h \in V_h,\ y_h, \tilde{y}_h \in Y_h.\end{aligned}$$

To clarify the approach, we would like to give its algebraic version. All operators and functionals are defined using finite-dimensional spaces, such that their Fréchet derivatives can be represented by matrices and vectors. Let $y_h \in D_h$ be given and denote by $D\Phi_h \in \mathbb{R}^{1\times \dim Y_h}$ and $D\widetilde{I}_h \in \mathbb{R}^{1\times \dim V_h}$ the vectors representing the derivatives $D\Phi_h(y_h)$ and $D\widetilde{I}_h(\tilde{u}_h(y_h))$, 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 $\partial_y R_h \in \mathbb{R}^{\dim V_h \times \dim Y_h}$ be the matrices representing the derivatives $D\tilde{u}_h(y_h)$, $\partial_w R_h(\tilde{u}_h(y_h), y_h)$, and $\partial_y R_h(\tilde{u}_h(y_h), y_h)$, respectively. Then, equation (10) holds true if and only if

$$D\Phi_h\mathbf{y} = D\widetilde{I}_h D\tilde{u}_h\mathbf{y} \quad \forall\ \mathbf{y} \in \mathbb{R}^{\dim Y_h}. \tag{16}$$

Relation (12) is equivalent to

$$\mathbf{v}^T\partial_w R_h D\tilde{u}_h\mathbf{y} = -\mathbf{v}^T\partial_y R_h\mathbf{y} \quad \forall\ \mathbf{v} \in \mathbb{R}^{\dim V_h}. \tag{17}$$

The goal of the adjoint approach consists in reformulating the right-hand side of (16). To this end, choose $\mathbf{v}$ in (17) such that $\mathbf{v}^T\partial_w R_h = D\widetilde{I}_h$, i.e.,

$$\boldsymbol{\psi} := \mathbf{v} = (\partial_w R_h)^{-T}D\widetilde{I}_h^T,$$

which is the algebraic version of (14). Inserting $\boldsymbol{\psi}$ into (16) and using (17) gives

$$D\Phi_h\mathbf{y} = \boldsymbol{\psi}^T\partial_w R_h D\tilde{u}_h\mathbf{y} = -\boldsymbol{\psi}^T\partial_y R_h\mathbf{y} \quad \forall\ \mathbf{y} \in \mathbb{R}^{\dim Y_h}.$$

This is equivalent to

$$D\Phi_h = -\boldsymbol{\psi}^T\partial_y R_h,$$

that is the algebraic version of (15).

4. Application to the SUPG method

For the SUPG method (6), there are

$$\begin{aligned}&\langle(\partial_w R_h)(w_h, y_h)\tilde{v}_h, v_h\rangle = a(\tilde{v}_h, v_h) + s_h(y_h; \tilde{v}_h, v_h), \\ &\langle(\partial_y R_h)(w_h, y_h)\tilde{y}_h, v_h\rangle = s_h(\tilde{y}_h; w_h + \tilde{u}_{bh}, v_h) - r_h(\tilde{y}_h; v_h)\end{aligned}$$

for any $y_h, \tilde{y}_h \in Y_h$ and $v_h, \tilde{v}_h, w_h \in V_h$. Thus, for any $y_h \in D_h$, the auxiliary function $\psi_h(y_h) \in V_h$ is the solution of

$$a(v_h, \psi_h(y_h)) + s_h(y_h; v_h, \psi_h(y_h)) = \langle D\widetilde{I}_h(\tilde{u}_h(y_h)), v_h\rangle \quad \forall v_h \in V_h \tag{18}$$

and the Fréchet derivative of Φ_h is given by

$$\langle D\Phi_h(y_h), \tilde{y}_h\rangle = -s_h(\tilde{y}_h; u_h(y_h), \psi_h(y_h)) + r_h(\tilde{y}_h; \psi_h(y_h)) \quad \forall \tilde{y}_h \in Y_h.$$

We define Y_h as the space of piecewise constant functions. After having solved (18) for a given stabilization parameter y_h, the Fréchet derivative of Φ_h at y_h with respect to the stabilization parameter is available.

The most popular [34] quasi-Newton method for solving a nonlinear minimization problem is the BFGS (Broyden, Fletcher, Goldfarb, Shanno) method [5,8,14,39]. This method requires only the gradient of the functional with respect to the stabilization parameter. By measuring the changes of the gradients, it constructs a model for the functional that delivers information to obtain superlinear convergence. The cost consists in the storage of the gradients, which are piecewise constant finite element functions. For practical reasons, this can be done only for a limited number of gradients. The resulting algorithm is called limited memory BFGS or L-BFGS, see Algorithm 7.5 in [34]. This algorithm is used in the simulations presented below. We could observe a dramatic improvement of efficiency compared with the application of the steepest descent method which was used in preliminary numerical studies.

The L-BFGS method proposes a search direction for updating the stabilization parameter in the kth iteration, $k \geqslant 0$. In addition, a step length $\alpha^{(k)}$ is needed. In our implementation of the method, the step length is determined such that the decrease of the functional I_h is locally maximized. To this end, the initial guess for each step length $\alpha^{(k)}$ is a value α_{ini}. If the application of α_{ini} leads to a reduction of the target functional, the step length will be doubled. This step is repeated as long as the target functional decreases. If the application of α_{ini} does not lead to a reduction of the value of the target functional, α_{ini} will be divided by 2. The reduction of α_{ini}

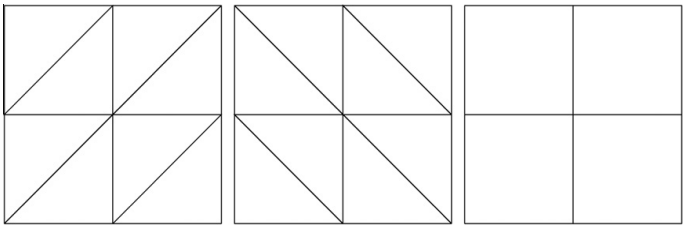


Fig. 1. Types of triangulations used in the computations (left to right): Grid 1, Grid 2, and Grid 3 (level 1).

will be stopped if either a step length is found that leads to a decrease of I_h or if a minimal value $\alpha_{\min}$ for the step length is obtained. The iteration stops either after reaching a prescribed maximal number of iterations $k_{\max}$ or if the decrease of the target functional becomes too slow. The concrete test for the last stopping criterion is

$$\frac{\Phi_h\left(y_h^{(k-10)}\right) - \Phi_h\left(y_h^{(k)}\right)}{\Phi_h\left(y_h^{(k-10)}\right)} \leqslant d_{\min}, \quad k \geqslant 10.$$

If the computation of a search direction with L-BFGS is not successful or if the proposed step length becomes smaller than $\alpha_{\min}$, L-BFGS is restarted.

Of course, before the solution with a proposed stabilization parameter $y_h^{(k+1)}$ is computed, the stabilization parameter is always restricted to admissible values according to (8) and (9). The values from [17] are used for c_{inv} in (8) and (9).

In our numerical tests, the initial step length parameter was set similarly to a proposal from [34]

$$\alpha_{\text{ini}}^{(0)} = 10^{-6},$$
$$\alpha_{\text{ini}}^{(k)} = \max\left\{\min\left\{1, \frac{D\Phi_h(y_h^{(k-1)})^T \Delta y_h^{(k-1)}}{D\Phi_h(y_h^{(k)})^T \Delta y_h^{(k)}}\right\}, 10^{-6}\right\}, \quad k \geqslant 1,$$

where $\Delta y_h^{(k)}$ is the search direction proposed by the L-BFGS method in the kth step. The minimal step length parameter was set to be $\alpha_{\min} = 10^{-12}$, the maximal number of iterations was prescribed with $k_{\max} = 10000$ (which was never reached), at most 100 gradients in the L-BFGS method were stored, and the parameter in the stopping criterion was set to be $d_{\min} = 10^{-4}$. The stabilization parameter was initialized with the standard choice (7). Most of the computations were performed and double-checked with two codes, one of them MooNMD [29].

5. Proof of concept: parameter optimization with respect to errors

A common approach for supporting error estimates consists in prescribing a solution of (1), that defines also the right-hand side and the boundary conditions of (1), and measuring errors of the numerical solution in certain norms. If errors can be measured, it should be possible with the proposed methodology to compute a SUPG stabilization parameter such that these errors are reduced compared with the standard choice of the SUPG parameter (7). This section studies this topic.

Numerical studies with respect to the error in the $L^2(\Omega)$ norm and the $H^1(\Omega)$ semi norm were performed. For shortness, the detailed presentation will be restricted to the error in the $L^2(\Omega)$ norm

$$I_h(w_h) = \|u - w_h\|_{0,\Omega}^2.$$

Then, the right-hand side of the adjoint problem (18) becomes

$$\langle D\widetilde{I}_h(\tilde{u}_h(y_h)), v_h\rangle = -2(u - u_h(y_h), v_h). \tag{19}$$

At the end of this section, some remarks will be given on the error in the $H^1(\Omega)$ semi norm.

A difficulty consists in finding or defining examples that, on the one hand, have a known solution and, on the other hand, possess typical features of solutions of convection-dominated problems, in particular layers. Below, results obtained with two examples defined in [30] will be presented. The solutions of these examples depend on the diffusion coefficient ε, and so the right-hand sides do. As already noticed in [31], high order quadrature rules are necessary to keep the quadrature error for the right-hand side small in the case of small ε. For this reason, the diffusion coefficient was chosen only three or four orders of magnitude smaller than the convection in these examples.

Both examples are defined on the unit square. In the computations, triangular grids (Grid 1 in Fig. 1) with P_1, P_2, P_3 finite elements and square grids (Grid 3) with Q_1, Q_2, Q_3 finite elements were used. Level 0 of Grid 1 consists of two triangles and level 0 of Grid 3 of one square. The grids were regularly refined using so-called red refinement. A quadrature rule that is exact for polynomials of degree 19 was used on triangles and a Gaussian quadrature rule that is exact for polynomials of degree 17 on squares.

Example 5.1 (*Example with interior layer*). This example is given by $\Omega = (0,1)^2$, $\Gamma^D = \partial\Omega$, $\varepsilon = 10^{-4}$, $\mathbf{b} = (2,3)^T$, $c = 2$. The right-hand side f and the Dirichlet boundary condition u_b are prescribed such that

$$u(x,y) = 16x(1-x)y(1-y) \times \left(\frac{1}{2} + \frac{\arctan[2\varepsilon^{-1/2}(0.25^2 - (x-0.5)^2 - (y-0.5)^2)]}{\pi}\right)$$

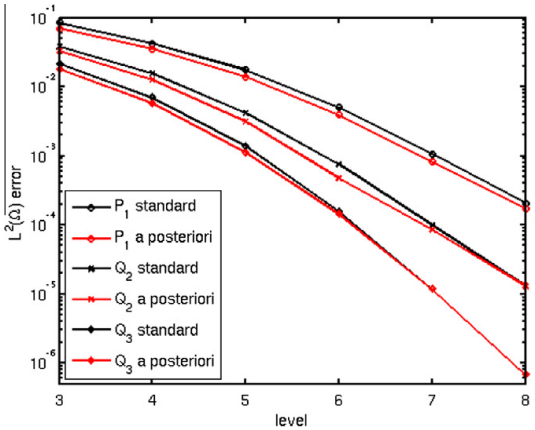


Fig. 2. Example 5.1, $L^2(\Omega)$ errors for different finite elements, comparison of standard parameter choice (7) and the a posteriori choice based on minimizing the $L^2(\Omega)$ error.

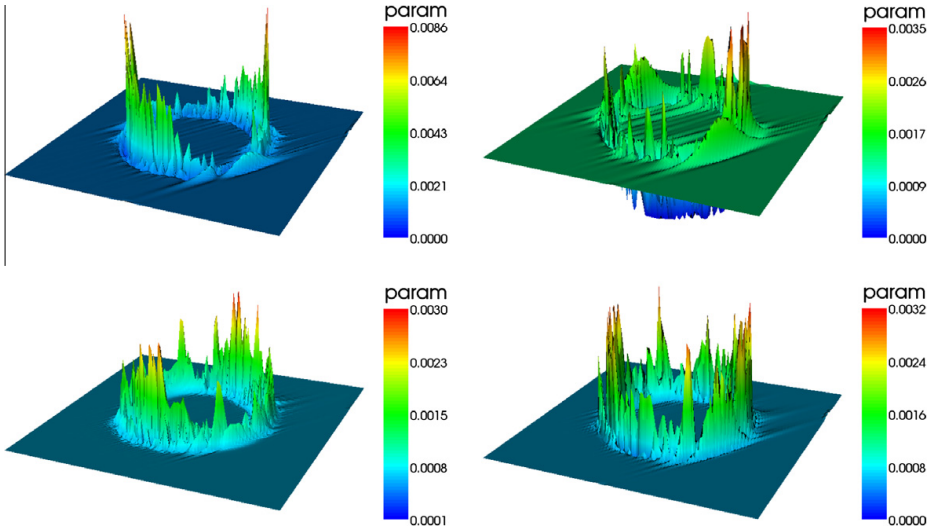


Fig. 3. Example 5.1, a posteriori defined stabilization parameters; top left: P_1, standard parameter $y_h = 1.294391e-3$; top right: Q_1, standard parameter $y_h = 1.294391e-3$; bottom left: P_2, standard parameter $y_h = 6.433494e-4$; bottom right: Q_2, standard parameter $y_h = 6.433494e-4$; all level 7 (visualization by projection to P_1 or Q_1 finite element).

is the solution of (1). The solution has the form of a circular hump in the center of the domain.

A comparison of the $L^2(\Omega)$ errors obtained with the standard parameter choice (7) and the a posteriori choice based on the adjoint problem with right-hand side (19) is presented in Fig. 2. It can be observed that the a posteriori parameter choice leads in fact to solutions with smaller $L^2(\Omega)$ error. Naturally, on finer grids, where the stabilization looses importance, the error reduction becomes smaller. Since the convection is constant, the standard parameter (7) is constant on a uniform grid, too. Fig. 3 shows the distribution of the stabilization parameter for different finite elements on certain grid levels. The corresponding standard parameters are given in the caption. It can be seen that the a posteriori methodology changes the parameter in the layer, which is not surprising since the stabilization is needed in the layer. On many mesh cells at the layer, the parameter is increased considerably. A large stabilization parameter can be observed at the front and at the back (with respect to the direction of the convection) of the hump. Note, in few mesh cells at the layer, a reduction of the stabilization parameter is proposed. This reduction is in general much smaller than the increase of the parameter in other mesh cells and therefore it is only visible in the picture for the Q_1 finite element. In summary, the main mechanism to reduce the $L^2(\Omega)$ error was always a significant increase of the stabilization parameter within the layer.

Example 5.2 (*Example with boundary layer*). This example is defined by $\Omega = (0,1)^2$, $\Gamma^D = \partial\Omega$, $\varepsilon = 10^{-3}$, $\mathbf{b} = (2,3)^T$, and $c = 1$. The prescribed solution

$$u(x,y) = xy^2 - y^2 \exp\left(\frac{2(x-1)}{\varepsilon}\right) - x\exp\left(\frac{3(y-1)}{\varepsilon}\right) + \exp\left(\frac{2(x-1)+3(y-1)}{\varepsilon}\right)$$

defines the right-hand side f and the Dirichlet boundary condition u_b. It possesses boundary layers at $x = 1$ and $y = 1$, see Fig. 7.

Fig. 4 presents comparisons of the $L^2(\Omega)$ errors obtained with the standard and the a posteriori parameter choices. Clearly, the a posteriori parameter choice leads always to a reduction of the $L^2(\Omega)$ errors. However, a higher order of convergence cannot be observed. A posteriori computed parameters are presented in Fig. 5. It can be noticed that the optimization of the $L^2(\Omega)$ error reduces the stabilization parameters in the layers.

Concerning the a posteriori parameter choice based on the error in the $H^1(\Omega)$ semi norm, we could observe essentially the same behavior as for the $L^2(\Omega)$ norm: the $H^1(\Omega)$ semi norm error becomes always smaller than for the solution with the standard parameter (7). However, sometimes the error reduction is very

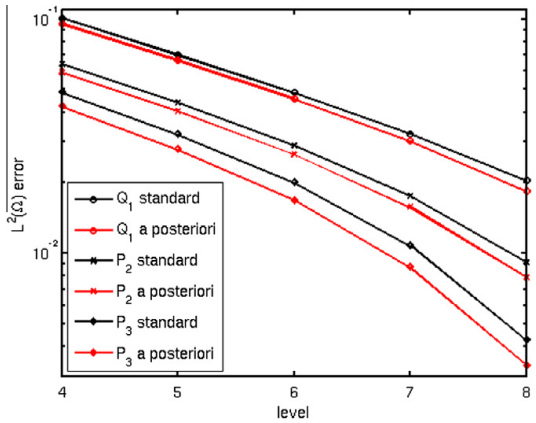


Fig. 4. Example 5.2, $L^2(\Omega)$ errors for different finite elements, comparison of standard parameter choice (7) and the a posteriori choice based on minimizing the $L^2(\Omega)$ error.

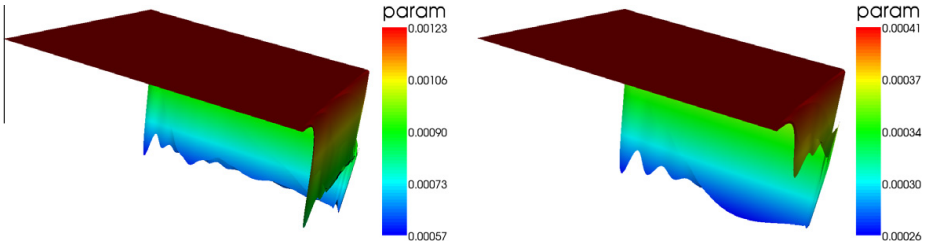


Fig. 5. Example 5.2, a posteriori defined stabilization parameters; left: Q_1, standard parameter $y_h = 1.225160e-3$; right: Q_2, standard parameter $y_h = 5.741186e-4$; both level 7 (visualization by projection to Q_1 finite element).

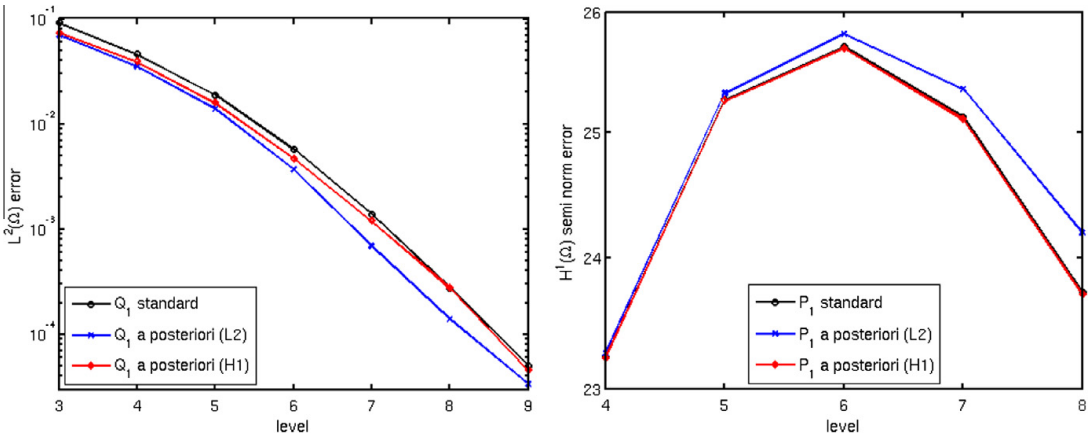


Fig. 6. Left: Example 5.1, Q_1 finite element and $L^2(\Omega)$ error; right: Example 5.2, P_1 finite element and $H^1(\Omega)$ semi norm error; comparison of the different parameter choices.

small. Because of the unresolved layers, in particular in Example 5.2, the error in the $H^1(\Omega)$ semi norm, computed with the above mentioned quadrature rules, even grows on coarse grids, compare Fig. 6.

Considering all three parameter choices (standard, a posteriori based on $L^2(\Omega)$ error, a posteriori based on $H^1(\Omega)$ semi norm error), one can observe that the optimization with respect to the error in one norm might reduce the error in the other norm, too, compared with the standard parameter choice. But the other error might also increase, see Fig. 6. Fig. 7 shows stabilization parameters and corresponding solutions with respect to the optimization of errors in different norms. Whereas the optimization of the $L^2(\Omega)$ error reduces the parameter in the boundary layers, the optimization with respect to the error in the $H^1(\Omega)$ semi norm increases the parameter in the layer at $x = 1$. The different effects on the computed solutions are clearly visible. In the $L^2(\Omega)$ error optimized solution, considerable spurious oscillations can be observed in the layers. They are even larger than in the solution computed with the standard parameter (7). The solution with $H^1(\Omega)$ semi norm error optimization looks much better. This comparison demonstrates already the importance of using an appropriate measure upon which the a posteriori selection of the parameter is based.

Altogether, the results presented in this section demonstrate that the proposed methodology is able to compute a stabilization parameter in the SUPG method in an a posteriori way such that solutions with reduced errors are obtained.

6. Parameter optimization with respect to functionals which are candidates for describing the quality of computed solutions

Generally, the evaluation of errors is not possible as the solution of (1) is not known. In this situation, other functionals are necessary to measure or estimate the quality of computed solutions.

On the first glance, a posteriori error estimators might be an appropriate choice. The construction of reliable error estimators with respect to global norms for convection-dominated problems is difficult. As demonstrated, e.g., in [23], the application of standard estimators for elliptic problems does not lead to reliable error predictions. The numerical studies presented below will consider a residual-based error estimator from [43]

$$I_h(w_h) = \sum_{K\in T_h} \alpha_K^2 \| -\varepsilon\Delta w_h + \mathbf{b}\cdot\nabla w_h + c w_h - f\|_{0,K}^2 + \sum_{E\subset\partial K} \varepsilon^{-1/2}\alpha_E \|R_E(w_h)\|_{0,E}^2 \quad \forall w_h \in W_h \tag{20}$$

with

$$R_E(w_h) = \begin{cases} -[|\varepsilon \mathbf{n}_E\cdot\nabla w_h|]_E, & \text{if } E \not\subset \partial\Omega, \\ g - \varepsilon \mathbf{n}_E\cdot\nabla w_h, & \text{if } E \subset \Gamma^N, \\ 0, & \text{if } E \subset \Gamma^D, \end{cases}$$

and

$$\alpha_K = \min\left\{\operatorname{diam}(K)\varepsilon^{-1/2}, c_0^{-1/2}\right\}, \quad \alpha_E = \min\left\{\operatorname{diam}(E)\varepsilon^{-1/2}, c_0^{-1/2}\right\}.$$

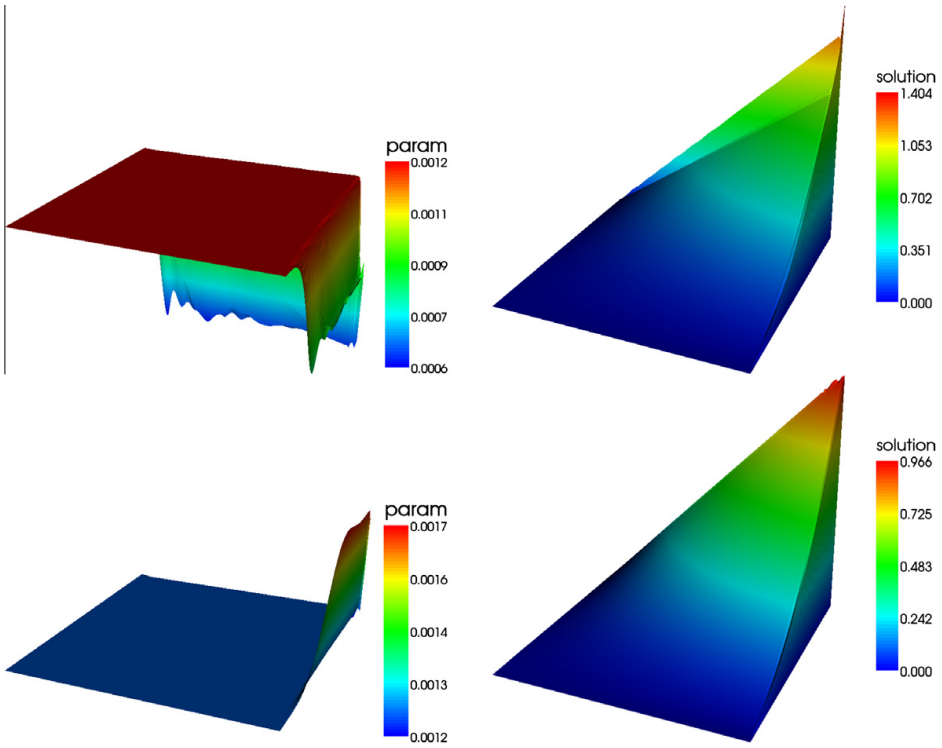


Fig. 7. Example 5.2, a posteriori defined stabilization parameters and computed solutions with the P_1 finite element; top: optimization with respect to the $L^2(\Omega)$ error; bottom: optimization with respect to the $H^1(\Omega)$ semi norm error; both at level 7 (parameter: visualization by projection to P_1 finite element).

Here, diam(K) and diam(E) denote the diameters of the mesh cell K and the face E, respectively, $\mathbf{n}_E$ is a unit normal on E, and c_0 is defined in (2). The jump of a function across the face E is denoted by $[|\cdot|]_E$. This error estimator is robust in a norm that is a sum of the standard energy norm and a dual norm of the convective derivative, see [43].

The right-hand side of the adjoint problem for the functional (20) is given by

$$\begin{aligned}\langle D\widetilde{I}_h(\tilde{u}_h(y_h)), v_h\rangle = \sum_{K\in\mathcal{T}_h} & 2\alpha_K^2(-\varepsilon\Delta u_h(y_h) + \mathbf{b}\cdot\nabla u_h(y_h) \\ & + cu_h(y_h) - f, -\varepsilon\Delta v_h + \mathbf{b}\cdot\nabla v_h + cv_h)_K \\ & + \sum_{E\subset\partial K} 2\varepsilon^{-1/2}\alpha_E(R_E(u_h(y_h)), \widetilde{R}_E(v_h))_E,\end{aligned}$$

where $\widetilde{R}_E(v_h)$ is $R_E(v_h)$ with $g = 0$.

Applying the estimator (20) as functional for the parameter optimization, it turns out that the global errors are dominated by the local contributions from the mesh cells in layers at the Dirichlet boundary. This effect comes from the nature of the underlying problem. For a local error estimate to be small, in particular the strong residual on a mesh cell (first term in (20)) has to be small. This cannot be achieved in mesh cells with boundary layers since the layers are not resolved. Even a nodally exact numerical solution leads to a large residual in those mesh cells. Thus, a significant reduction of the residual in such mesh cells is not possible. As the optimization algorithm concentrates on the reduction of the dominating errors, consequently, the errors in mesh cells away from the Dirichlet boundary are also not reduced notably. For this reason, an error indicator that excludes the mesh cells at the Dirichlet boundary will be considered, too. Furthermore, we could observe that the influence of the residuals on the edges in (20) is negligible. One obtains practically the same results with and without using these terms. Thus, besides (20), the error indicator

$$I_h(w_h) = \sum_{K\in\mathcal{T}_h, \overline{K}\cap\overline{\Gamma^D}=\emptyset} \alpha_K^2 \| -\varepsilon\Delta w_h + \mathbf{b}\cdot\nabla w_h + cw_h - f\|_{0,K}^2 \quad \forall w_h \in W_h \tag{21}$$

will be considered. Note, the mesh cells at the Dirichlet boundary do not contribute to the error indicator, but the stabilization parameter in these cells is still included into the optimization process.

The most serious drawback of using the SUPG method are the spurious oscillations that might appear in a vicinity of the layers. An optimization of the stabilization parameter should try above all to reduce them. These oscillations are connected to large derivatives of the computed solutions in crosswind direction. For this reason, a third functional that contains, besides the residual, also a control of the crosswind derivative will be included into the studies

$$\begin{aligned}I_h(w_h) = \sum_{K\in\mathcal{T}_h, \overline{K}\cap\overline{\Gamma^D}=\emptyset} & \Big(\| -\varepsilon\Delta w_h + \mathbf{b}\cdot\nabla w_h + cw_h - f\|_{0,K}^2 \\ & + \|\phi(|\mathbf{b}^\perp\cdot\nabla w_h|)\|_{0,1,K}\Big) \quad \forall w_h \in W_h,\end{aligned} \tag{22}$$

where

$$\mathbf{b}^\perp(\mathbf{x}) = \begin{cases} \frac{(b_2(\mathbf{x}), -b_1(\mathbf{x}))}{|\mathbf{b}(\mathbf{x})|}, & \text{if } \mathbf{b}(\mathbf{x}) \neq 0, \\ 0, & \text{if } \mathbf{b}(\mathbf{x}) = 0, \end{cases}$$

and

$$\phi(x) = \begin{cases} \sqrt{x}, & \text{if } x \geqslant 1, \\ 0.5(5x^2 - 3x^3), & \text{if } x < 1. \end{cases}$$

The special choice of $\phi(x)$ ensures that this functional is Fréchet differentiable. Its derivative can be computed in the usual way.

The numerical studies will consider a standard example, defined on the unit square, that is often used for the evaluation

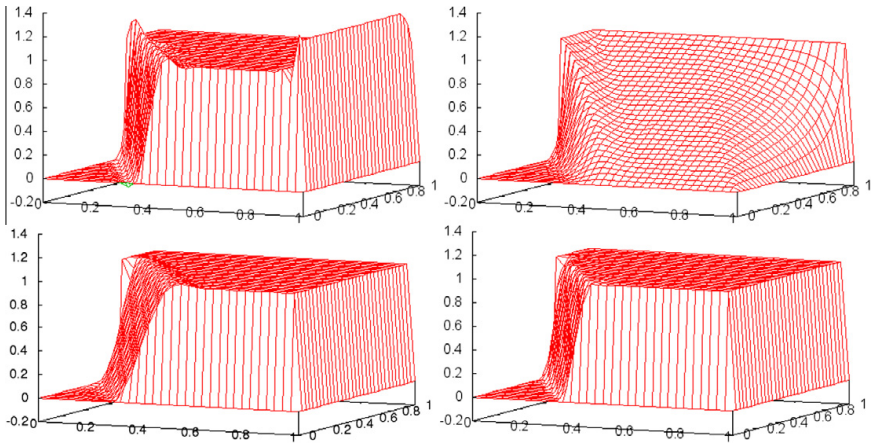


Fig. 8. Example 6.1: P_1, level 5 (1089 d.o.f.), solution with standard parameter (7), minimization of (20), minimization of (21), and minimization of (22), left to right, top to bottom.

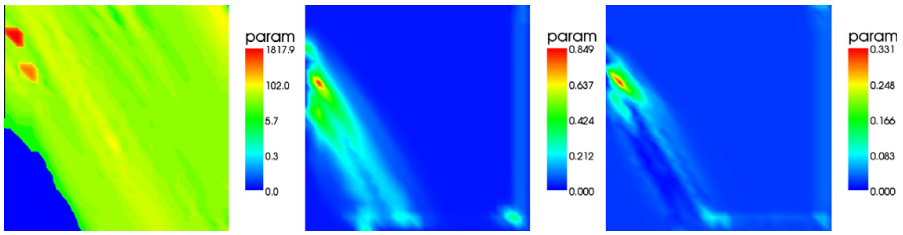


Fig. 9. Example 6.1: P_1, level 5 (1089 d.o.f.), stabilization parameter (standard parameter (7) $y_h = 0.018042$), minimization of (20) (logarithmic scale), minimization of (21), and minimization of (22), left to right (visualization by projection to P_1 finite element).

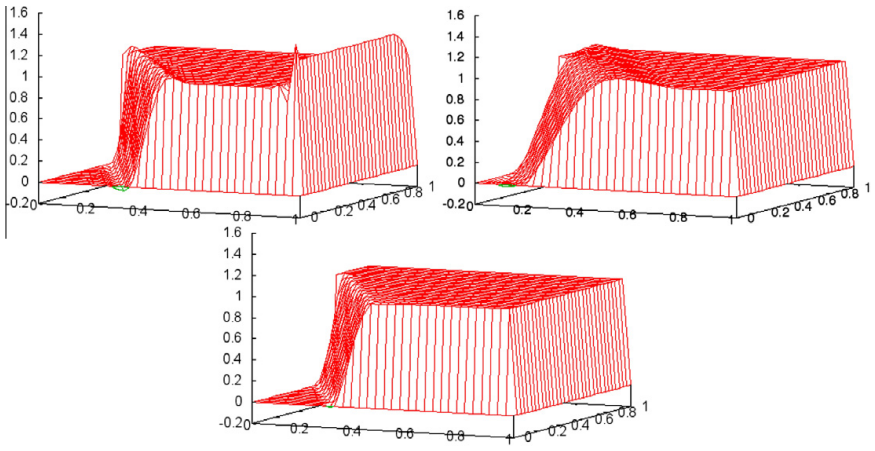


Fig. 10. Example 6.1: Q_1, level 5 (1089 d.o.f.), solution with standard parameter (7), minimization of (21), and minimization of (22), left to right, top to bottom.

of stabilized methods, and an example in a more complicated domain that attracted some attention in the past years. Both examples have the properties $\mathrm{div}\mathbf{b} = 0$, $c = 0$, such that the upper bound (9) for the stabilization parameter applies.

Example 6.1 (*Example with interior and exponential boundary layers*). This example was proposed in [22]. It is given by $\Omega = (0,1)^2$, $\Gamma^D = \partial\Omega$, with the data $\varepsilon = 10^{-8}$, $\mathbf{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = 0$, $f = 0$, and

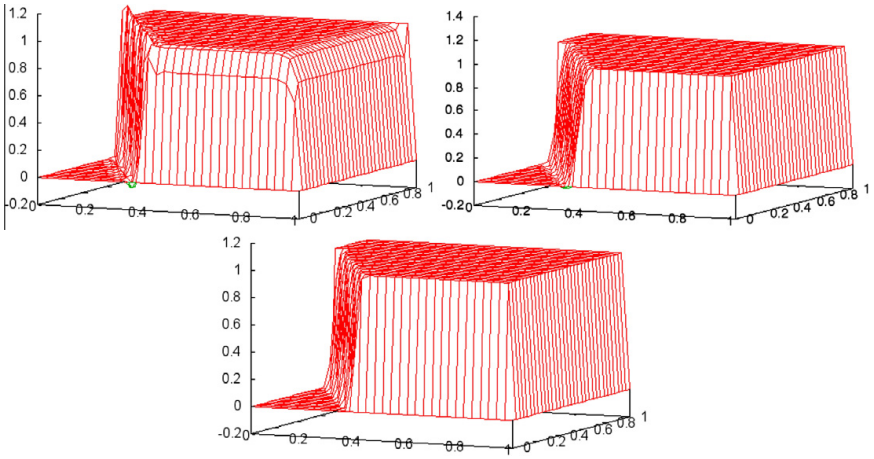


Fig. 11. Example 6.1: P_2, level 5 (4225 d.o.f.), solution with standard parameter (7), minimization of (21), and minimization of (22), left to right, top to bottom (visualization by projection to P_1 finite element).

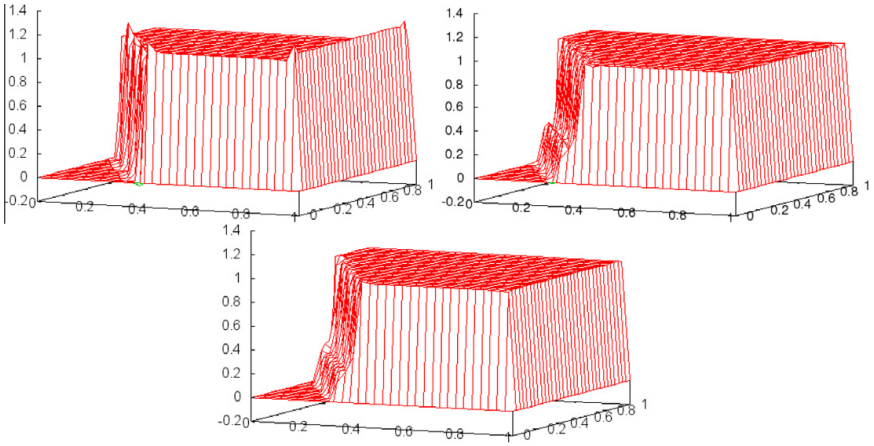


Fig. 12. Example 6.1: P_3, level 5 (9409 d.o.f.), solution with standard parameter (7), minimization of (21), and minimization of (22), left to right, top to bottom (visualization by projection to P_1 finite element).

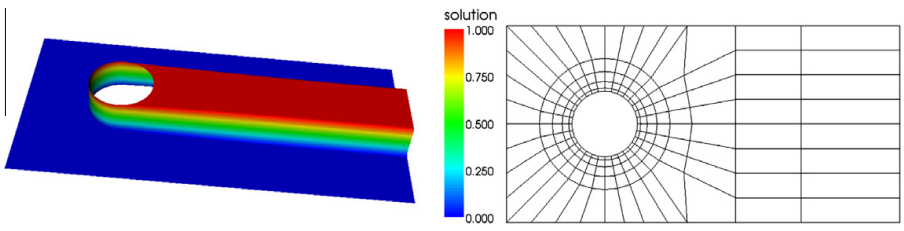


Fig. 13. Solution and initial grid for Example 6.2.

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

The simulations were performed on Grid 2 and Grid 3 from Fig. 1. For shortness of presentation, only results on a rather coarse mesh are shown in Figs. 8–12. We could observe that the principal behavior for P_k and Q_k finite elements, with the same k, was always similar.

As already mentioned above, the minimization of (20) does not lead to useful results. This is demonstrated exemplarily for the P_1 finite element in Fig. 8. It can be seen that the spurious oscillations are removed but the layers are extremely smeared. The plot of the stabilization parameter in Fig. 9 shows that this is caused by very large values of this parameter (note the logarithmic scale in this picture). Minimizing (21) instead of (20) leads to a considerable improvement with respect to the extreme smearing. However, a notable smearing of the interior layer can still be observed. The reason is the prediction of a rather large stabilization parameter in this layer, see Fig. 9. Nearly perfect solutions are obtained with the parameter choice based on minimizing (22). The spurious oscillations are almost removed, only around 2% are left. A large stabilization parameter is proposed in all layers, but its maximal value is smaller than the maximal value of the parameter computed with minimizing (21). Only the interior layer in the solution with the P_3 finite element is somewhat smeared. We think, the reason is the use of a piecewise constant stabilization parameter in this case. This polynomial degree of the parameter might not be sufficiently flexible for the changes of the finite element solution within a mesh cell that occur for higher order finite elements.

Example 6.2 (*The Hemker example*). This example was defined in [19]. The simulations were performed with $\Omega = \{(-3,8) \times (-3,3)\} \backslash \{(x,y); x^2 + y^2 \leqslant 1\}$, $\varepsilon = 10^{-6}$, $\mathbf{b} = (1,0)^T$, and $c = f = 0$. At

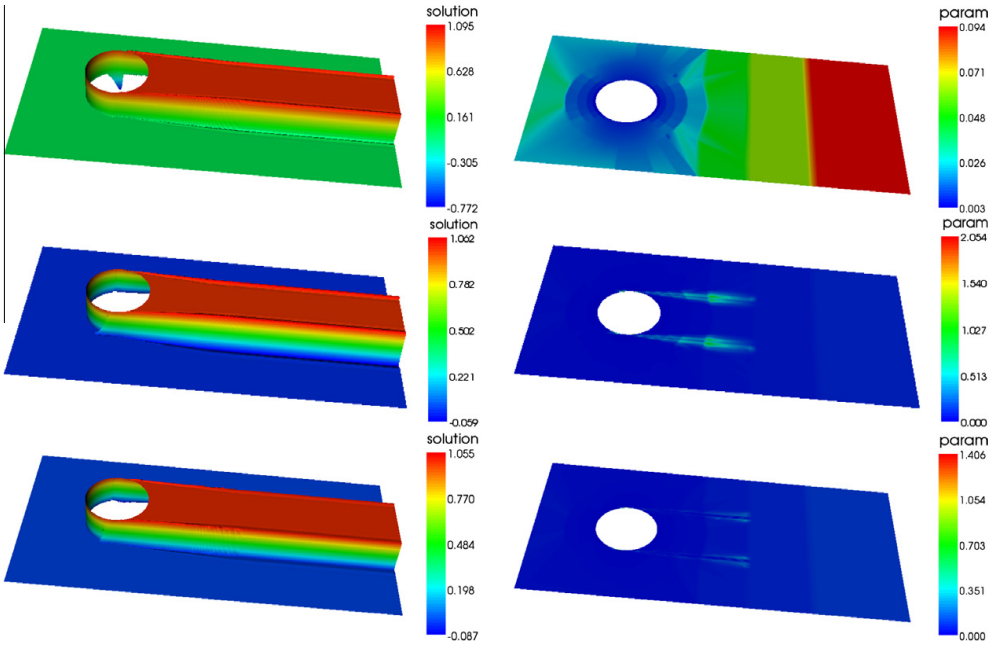


Fig. 14. Example 6.2: Q_1, level 4 (47 664 d.o.f.), standard parameter (7), minimization of (21), and minimization of (22), top to bottom (stabilization parameter visualization by projection to Q_1 finite element).

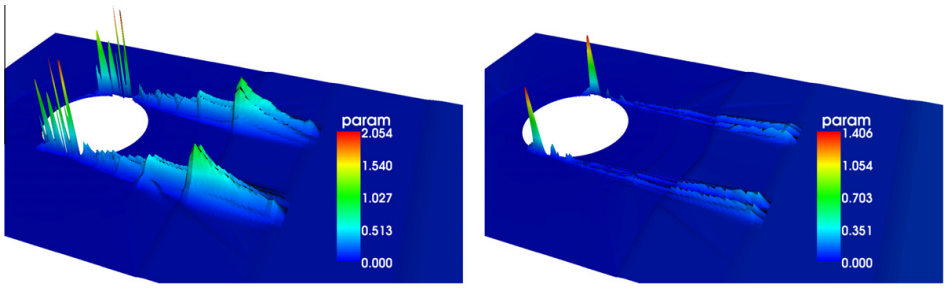


Fig. 15. Example 6.2: Q_1, level 4 (47 664 d.o.f.), details of the stabilization parameter, minimization of (21), and minimization of (22), left to right (visualization by projection to Q_1 finite element).

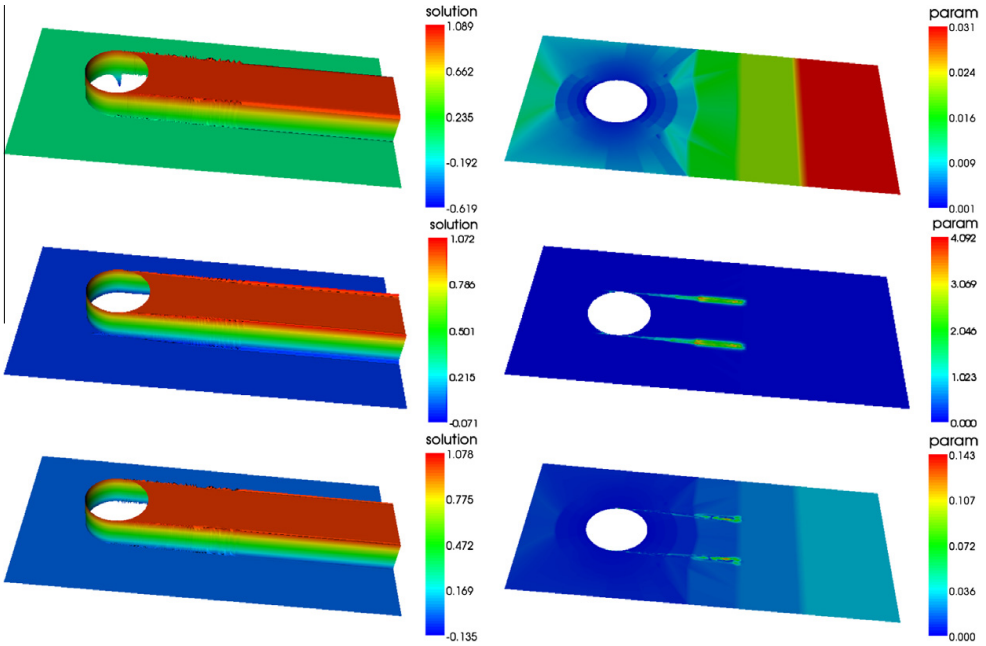


Fig. 16. Example 6.2: Q_3, level 4 (425 616 d.o.f.), standard parameter (7), minimization of (21), and minimization of (22), top to bottom (visualization by projection to Q_1 finite element).

the inlet $x = -3$, a homogeneous Dirichlet boundary condition is prescribed, at the circle there is $u = 1$, and on all other parts of $\partial\Omega$, homogeneous Neumann boundary conditions are given.

This example attracted recently some interest [16,41] since it is considered to be closer to situations arising in applications than many usual test examples. It can be interpreted as a model of heat transfer from a hot column in the direction of the convection.

Results of numerical simulations are presented for the Q_1, Q_2, and Q_3 finite elements in Figs. 14–17. The initial grid (level 0) is shown in Fig. 13. Isoparametric finite elements were used to approximate the curved boundary. It can be seen that the most difficult regions for computing a correct solution are the starting points of the interior layers on top and on bottom of the circle. The SUPG method was applied with the standard parameter choice (7). Considerable negative spurious oscillations at the starting points of the interior layers can be observed for the solutions computed with this choice. Note that the values of the standard parameter are rather small in the vicinity of the circle due to small diameters of mesh cells in this region. Solutions obtained with the minimization of the error estimator (20) are not shown. Similarly as in the previous example, the layers are smeared very much, in particular the layer in front of the circle. For the Q_1 finite element, the minimization of the functional (21) reduces the negative spurious oscillations considerably, compared with the solution obtained with the standard parameter choice, see also Fig. 17. However, the solutions which are based on the minimization of this functional possess the wrong feature that the interior layers start somewhat before the top and bottom of the circle. This feature was reduced or even removed by minimizing (22) for the determination of the stabilization parameter. It can be observed that both, the minimization of (21) and the minimization of (22), lead to an increase of the parameter in the region of the interior layers, in particular at the starting points of the interior layers, cf. Fig. 15.

Since the large undershoots are a distinguished bad feature of the standard SUPG approach, Fig. 17 shows the size of the undershoots obtained in the simulations. For the Q_1 finite element, the parameter choices based on the minimization of (21) and (22) reduce these undershoots on all levels considerably. The situation is different for the Q_2 and Q_3 finite element, where only the minimization of (22) leads to smaller undershoots on most levels. A reason for not observing this on all levels might be the insufficient flexibility of using a piecewise constant stabilization parameter for a higher order finite element, see the discussion at the end of Example 6.1. The overshoots are much less pronounced than the undershoots. They are similar for all simulations, between 0.05 and 0.15.

Altogether, the parameter choice based on the minimization of (22) gave the best results among the considered approaches. However, these results are not yet optimal.

In the optimization process, always a fast decrease of the functionals within the first steps could be observed. To fulfill the stopping criterion formulated in Section 4, in general some dozens to a few hundred L-BFGS steps were necessary. As could be seen in the presented examples, the values of the stabilization parameter have very little effect on the solution in smooth regions and hence varying them has also little influence on the target functional. This observation offers a way for a possible improvement of the efficiency of the algorithm by identifying in the first few steps the values of the stabilization parameter which are important for the decrease of the functional and then restricting the optimization process to those values.

7. Summary and outlook

This paper presented a general framework for optimizing parameters in stabilized finite element methods for convection–diffusion

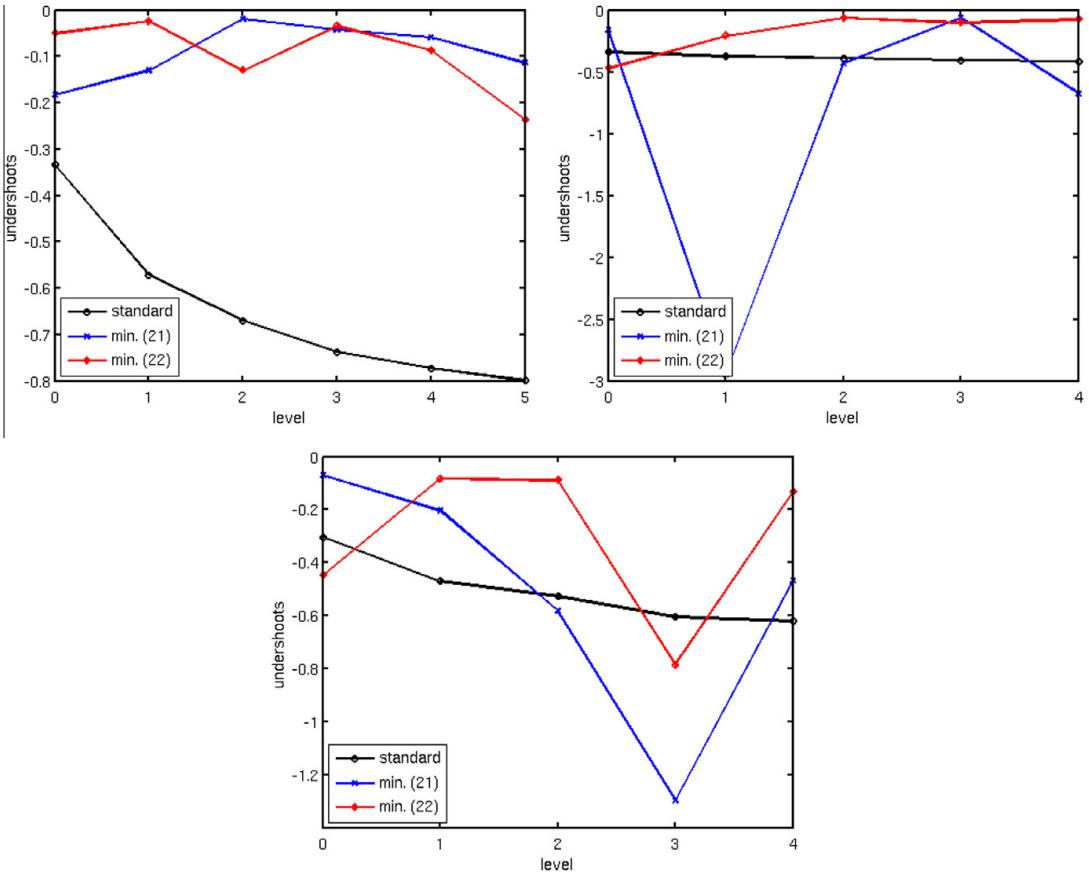


Fig. 17. Example 6.2: Undershoots of the computed solutions, Q_1 finite element, Q_2 finite element, and Q_3 finite element, left to right, top to bottom.

problems. The optimization is based on minimizing a target functional that indicates the quality of the computed solution. The L-BFGS method is used to solve the arising constrained optimization problem. Key of the algorithm is the efficient evaluation of the derivative of the target functional with respect to the stabilization parameter that utilizes the solution of an appropriate adjoint problem. Benefits and difficulties of this basic approach were studied exemplarily at the SUPG finite element method and three different functionals. A main observation is that a straightforward choice, a residual-based a posteriori error estimator, is not appropriate for measuring the quality of computed solutions. A better functional could be found, (22), but the results obtained with this functional are not yet optimal.

Important next steps in the exploration and improvement of the parameter optimization are as follows:

- A very important goal consists in identifying better functionals than used in this manuscript. The chosen functional is the main component of the algorithm that determines the quality of the computed solutions.
- It is known that the introduction of diffusion in streamline direction only, as in the SUPG method, is often not sufficient to obtain satisfactory numerical solutions. Some diffusion orthogonal to the streamlines (in crosswind direction) might be necessary, as it is done by SOLD methods [27]. A new aspect in the application of the general framework to SOLD methods consists in the optimization of two stabilization parameters.
- Algorithmic improvements are possible. These include, e.g., the restriction of the optimization to important values of the stabilization parameter as discussed at the end of Section 6.
- The considerable decrease of the functionals within the first few optimization steps suggest that the improvement of the solutions occurs mainly also within these steps. This effect will be studied in detail, leading hopefully to an efficient method for just improving (but not optimizing) standard SUPG solutions.

References

[1] I. Akkerman, K.G. van der Zee, S.J. Hulshoff, A variational Germano approach for stabilized finite element methods, Comput. Methods Appl. Mech. Engrg. 199 (2010) 502–513.
[2] W. Bangerth, R. Rannacher, Adaptive Finite Element Methods for Differential Equations, Lectures in Mathematics, ETH Zürich. Birkhäuser, Basel, 2003.
[3] R. Becker, R. Rannacher, An optimal control approach to a posteriori error estimation in finite element methods, in: A. Iserles (Ed.), Acta Numerica, Cambridge University Press, 2004, pp. 1–102.
[4] A.N. Brooks, 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.

[5] C.G. Broyden, The convergence of a class of double-rank minimization algorithms 1. General considerations, J. Inst. Math. Appl. 6 (1970) 76–90.
[6] P.G. Ciarlet, Basic error estimates for elliptic problems, in: P.G. Ciarlet, J.L. Lions (Eds.), Handbook of Numer. Anal., v. 2 – Finite Elem. Methods (pt. 1), North-Holland, Amsterdam, 1991, pp. 17–351.
[7] R. Codina, E. Oñate, M. Cervera, The intrinsic time for the streamline upwind/Petrov–Galerkin formulation using quadratic elements, Comput. Methods Appl. Mech. Engrg. 94 (1992) 239–262.
[8] R. Fletcher, A new approach to variable metric algorithms, Comput. J. 13 (1970) 317–322.
[9] L.P. Franca, S.L. Frey, T.J.R. Hughes, Stabilized finite element methods: I. Application to the advective–diffusive model, Comput. Methods Appl. Mech. Engrg. 95 (1992) 253–276.
[10] C. Führer, R. Rannacher, An adaptive streamline-diffusion finite element method for hyperbolic conservation laws, East–West J. Numer. Math. 5 (1997) 145–162.
[11] A.C. Galeão, R.C. Almeida, S.M.C. Malta, A.F.D. Loula, Finite element analysis of convection dominated reaction–diffusion problems, Appl. Numer. Math. 48 (2004) 205–222.
[12] M. Germano, U. Piomelli, P. Moin, W. Cabot, A dynamic subgrid-scale eddy viscosity model, Phys. Fluids A 3 (1991) 1760–1765.
[13] M.B. Giles, N.A. Pierce, An introduction to the adjoint approach to design, Flow, Turbulence Combust. 65 (2000) 393–415.
[14] D. Goldfarb, A family of variable metric updates derived by variational means, Math. Comput. 24 (1970) 23–26.
[15] J.-L. Guermond, Stabilization of Galerkin approximations of transport equations by subgrid modeling, M2AN 33 (1999) 1293–1316.
[16] H. Han, Z. Huang, R.B. Kellogg, A tailored finite point method for a singular perturbation problem on an unbounded domain, J. Sci. Comput. 36 (2008) 243–261.
[17] I. Harari, T.J.R. Hughes, What are c and h?: inequalities for the analysis and design of finite element methods, Comput. Methods Appl. Mech. Engrg. 97 (1992) 157–192.
[18] F.K. Hebeker, R. Rannacher, An adaptive finite element method for unsteady convection-dominated flows with stiff source terms, SIAM J. Sci. Comput. 21 (1999) 799–818.
[19] W.P. Hemker, A singularly perturbed model problem for numerical computation, J. Comput. Appl. Math. 76 (1996) 277–285.
[20] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid-scale models, bubbles and the origin of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995) 387–401.
[21] T.J.R. Hughes, A.N. Brooks, A multidimensional upwind scheme with no crosswind diffusion, in: T.J.R. Hughes (Ed.), Finite Element Methods for Convection Dominated Flows, AMD, vol. 34, ASME, New York, 1979, pp. 19–35.
[22] T.J.R. Hughes, M. Mallet, A. Mizukami, A new finite element formulation for computational fluid dynamics: II. Beyond SUPG, Comput. Methods Appl. Mech. Engrg. 54 (1986) 341–355.
[23] V. John, A numerical study of a posteriori error estimators for convection-diffusion equations, Comput. Methods Appl. Mech. Engrg. 190 (2000) 757–781.
[24] V. John, Large eddy simulation of turbulent incompressible flows, analytical and numerical results for a class of LES models, Lecture Notes in Computational Science and Engineering, vol. 34, Springer-Verlag, Berlin, Heidelberg, New York, 2004.
[25] V. John, S. Kaya, A finite element variational multiscale method for the Navier–Stokes equations, SIAM J. Sci. Comput. 26 (2005) 1485–1503.
[26] V. John, A. Kindl, A variational multiscale method for turbulent flow simulation with adaptive large scale space, J. Comput. Phys. 229 (2010) 301–312.
[27] V. John, P. Knobloch, A comparison of spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part I – a review, Comput. Methods Appl. Mech. Engrg. 196 (2007) 2197–2215.
[28] V. John, P. Knobloch, A comparison of spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: part II – analysis for P_1 and Q_1 finite elements, Comput. Methods Appl. Mech. Engrg. 197 (2008) 1997–2014.
[29] V. John, G. Matthies, MooNMD – a program package based on mapped finite element methods, Comput. Visual. Sci. 6 (2004) 163–170.
[30] V. John, J.M. Maubach, L. Tobiska, Nonconforming streamline-diffusion-finite-element-methods for convection-diffusion problems, Numer. Math. 78 (1997) 165–188.
[31] V. John, E. Schmeyer, Stabilized finite element methods for time-dependent convection-diffusion-reaction equations, Comput. Methods Appl. Mech. Engrg. 198 (2008) 475–494.
[32] O.A. Ladyzhenskaya, New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary value problems for them, Proc. Steklov Inst. Math. 102 (1967) 95–118.
[33] D.K. Lilly, A proposed modification of the Germano subgrid-scale closure method, Phys. Fluids A 4 (1992) 633–635.
[34] J. Nocedal, S.J. Wright, Numerical optimization, in: Springer Series in Operations Research and Financial Engineering, second ed., Springer, 2006.
[35] A.A. Oberai, J. Wanderer, A dynamic approach for evaluating parameters in a numerical method, Int. J. Numer. Methods Engng. 62 (2005) 50–71.
[36] E. Oñate, Derivation of stabilized equations for numerical solutions of advective-diffusive transport and fluid flow problems, Comput. Methods Appl. Mech. Engrg. 151 (1998) 233–265.
[37] H.-G. Roos, M. Stynes, L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations, Convection–Diffusion–Reaction and Flow Problems, second ed., Springer-Verlag, Berlin, 2008.
[38] R. Schneider, Applications of the Discrete Adjoint Method in Computational Fluid Dynamics, PhD thesis, University of Leeds, School of Computing, 2006.
[39] D.F. Shanno, Conditioning of quasi-Newton methods for function minimization, Math. Comput. 24 (1970) 647–655.
[40] J.S. Smagorinsky, General circulation experiments with the primitive equations, Mon. Weather Rev. 91 (1963) 99–164.
[41] P. Sun, L. Chen, J. Xu, Numerical studies of adaptive finite element methods for two dimensional convection-dominated problems, J. Sci. Comput. 43 (2010) 24–43.
[42] F. Tröltzsch, Optimale Steuerung partieller Differentialgleichungen - Theorie, Verfahren und Anwendungen, second ed., Vieweg+Teubner Verlag, 2009.
[43] R. Verfürth, Robust a posteriori error estimates for stationary convection-diffusion equations, SIAM J. Numer. Anal. 43 (2005) 1766–1782.

Adaptive Computation of Parameters in Stabilized Methods for Convection-Diffusion Problems

V. John and P. Knobloch

Abstract Stabilized finite element methods for convection-dominated problems contain parameters whose optimal choice is usually not known. This paper presents techniques for computing stabilization parameters in an adaptive way by minimizing a target functional characterizing the quality of the approximate solution. This leads to a constrained nonlinear optimization problem. Numerical results obtained for various target functionals are presented. They demonstrate that a posteriori optimization of parameters can significantly improve the quality of solutions obtained using stabilized methods.

1 Introduction

This paper is devoted to the numerical solution of a steady scalar convection-diffusion equation

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + c\, u = f \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega \tag{1}$$

by means of the finite element method. In (1), $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, is a bounded domain with a polygonal (resp. polyhedral) Lipschitz-continuous boundary $\partial\Omega$,

V. John
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39,
10117 Berlin, Germany

Department of Mathematics and Computer Science, Free University of Berlin, Arnimallee 6,
14195 Berlin, Germany
e-mail: volker.john@wias-berlin.de

P. Knobloch (✉)
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University,
Sokolovská 83, 186 75 Praha 8, Czech Republic
e-mail: knobloch@karlin.mff.cuni.cz

A. Cangiani et al. (eds.), *Numerical Mathematics and Advanced Applications 2011*,
DOI 10.1007/978-3-642-33134-3_30, © Springer-Verlag Berlin Heidelberg 2013

$\varepsilon > 0$ is constant, $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $u_b \in H^{1/2}(\partial\Omega)$. The Dirichlet boundary condition is used for the sake of simplicity only. In the numerical computations presented in this paper also more general boundary conditions were used.

Problem (1) is a simple model problem for convection-diffusion effects appearing in many more complicated applications. Therefore, it is important to be able to solve this problem numerically in a satisfactory way. However, this is by no means easy if convection dominates diffusion, i.e., $\varepsilon \ll |\mathbf{b}|$, since then the solution of (1) contains so-called layers, which are narrow regions where the solution changes abruptly. It is well known that the standard Galerkin finite element method provides approximate solutions that are globally polluted by spurious oscillations unless the computational mesh is sufficiently fine, i.e., $\varepsilon \gtrsim |\mathbf{b}|\, h$ where h is the mesh parameter.

To suppress the spurious oscillations, there are basically two options. Either one can use a layer-adapted mesh (e.g., a piecewise uniform mesh or a mesh obtained by an anisotropic adaptive refinement strategy) or one can consider a relatively coarse mesh and employ a modification of the standard discretization. There are various modifications that can be found in the literature: special discretizations of the convective term (upwinding), introduction of additional terms (stabilization) or manipulations at the algebraic level (e.g., FEMTVD schemes). In this paper, we shall be interested in stabilization techniques applied on relatively coarse meshes.

A common feature of stabilized finite element methods is that they contain parameters whose values significantly influence the quality of the approximate solution but whose optimal choice is usually not known. The aim of the present paper is to describe techniques that make it possible to compute stabilization parameters in an adaptive way by minimizing a functional characterizing the quality of the approximate solution. This leads to a constrained nonlinear optimization problem. The paper is a continuation of our previous work published in [3] where basic ideas of the optimization of stabilization parameters were presented.

The plan of the paper is as follows. In the next two sections we discuss linear and nonlinear stabilization approaches for finite element discretizations of (1). Then, in Sect. 4, we describe our approach of parameter optimization and explain how the Fréchet derivative of the target functional can be computed in an efficient way. Finally, in Sect. 5, we construct several target functionals and illustrate their properties by means of numerical results.

2 Linear Stabilized Methods

Let W_h be a finite element space approximating the space $H^1(\Omega)$ and set $V_h := W_h \cap H_0^1(\Omega)$. Let $u_{bh} \in W_h$ be a function whose trace approximates the function u_b. The simplest finite element discretization of (1) is the Galerkin method that reads: Find $u_h \in W_h$ such that $u_h = u_{bh}$ on $\partial\Omega$ and

$$a(u_h, v_h) = (f, v_h) \qquad \forall\, v_h \in V_h\,,$$

where $a(u,v) = \varepsilon(\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (c\,u, v)$ and $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. As we mentioned in the introduction, the Galerkin discretization is not appropriate if convection dominates diffusion and, as a remedy, a stabilization of the Galerkin method will be considered.

A stabilized finite element method for the numerical solution of (1) can be obtained from the Galerkin method by adding a stabilization term. We shall consider methods that read: Find $u_h \in W_h$ such that $u_h = u_{bh}$ on $\partial\Omega$ and

$$a(u_h, v_h) + \sum_{K\in\mathcal{T}_h} \tau_K\, s_K(u_h, v_h) = (f, v_h) \qquad \forall\, v_h \in V_h\,.$$

Here $\mathcal{T}_h$ is the triangulation used for constructing the finite element space W_h, τ_K is a nonnegative stabilization parameter, and s_K is a local form whose arguments are functions defined on the set $K \in \mathcal{T}_h$. The form s_K is always linear in the second argument and, if $f = 0$, it is also linear in the first argument. There are examples of s_K which are bilinear for any f. The parameter τ_K determines the artificial diffusion added by the stabilization term and it should be not 'too small' to remove oscillations but also not 'too large' to avoid excessive smearing. Consequently, it is very difficult to find appropriate values of τ_K a priori.

One of the most popular finite element approaches for convection-dominated problems is the SUPG method for which

$$s_K(u,v) = (\mathcal{L}_h\, u - f, \mathbf{b}\cdot\nabla v)_K$$

with the differential operator $\mathcal{L}_h = -\Delta_h + \mathbf{b}\cdot\nabla + c$ where the subscript h indicates that the Laplace operator is applied elementwise. The stabilization parameter is often defined by

$$\tau_K = \frac{h_K}{2\,|\mathbf{b}|}\left(\coth \mathrm{Pe}_K - \frac{1}{\mathrm{Pe}_K}\right) \quad \text{with} \quad \mathrm{Pe}_K = \frac{|\mathbf{b}|\,h_K}{2\,\varepsilon}\,, \tag{2}$$

where h_K is the diameter of K in the direction of $\mathbf{b}$.

3 Nonlinear Stabilized Methods

Since solutions of linear stabilized methods usually possess spurious oscillations in layer regions, the so-called SOLD (spurious oscillations at layers diminishing) methods have been developed. These methods add an additional stabilization term to the left-hand side of a linear stabilized method. Typical examples of this term are $(\widetilde{\varepsilon}\nabla u_h, \nabla v_h)$ adding isotropic artificial diffusion and $(\widetilde{\varepsilon}\,P\,\nabla u_h, P\,\nabla v_h)$ with the orthogonal projection P onto the plane orthogonal to $\mathbf{b}$, adding crosswind artificial diffusion. The parameter $\widetilde{\varepsilon}$ usually depends on the unknown approximate solution u_h and hence the resulting method is nonlinear.

In the literature, many proposals for the parameter $\widetilde{\varepsilon}$ can be found and we refer to [1, 2] for a review and computational comparison. One of the most successful formulas is

$$\widetilde{\varepsilon}|_K = \eta \frac{\operatorname{diam}(K)\,|\mathscr{L}_h u_h - f|}{2\,|\nabla u_h|} \qquad \forall\ K \in \mathscr{T}_h\,,$$

where η is a user-chosen parameter. From now on, the notion 'SOLD method' will mean that the crosswind diffusion term $(\widetilde{\varepsilon} P \nabla u_h, P \nabla v_h)$ together with this choice of $\widetilde{\varepsilon}$ is used. In the framework of parameter optimization, the parameter η will be considered piecewise constant. If an optimization of η is not considered, we set $\eta = 0.7$.

4 A Posteriori Optimization of Stabilization Parameters

In this section, we describe basic ideas of our approach to a posteriori optimization of stabilization parameters. For clarity of the presentation, we shall restrict ourselves to $u_b = 0$.

Let us write a linear or nonlinear stabilized method in the abstract form:

Given a stabilization parameter $y_h \in Y_h$, find $u_h \in V_h$ such that $R_h(u_h, y_h) = 0$.

Here, Y_h is a finite-dimensional space of functions on Ω and the operator R_h maps the space $V_h \times Y_h$ into the dual space V_h'. For example, for the SUPG method introduced in Sect. 2, we have

$$\langle R_h(u_h, y_h), v_h\rangle = a(u_h, v_h) + (\mathscr{L}_h u_h - f, y_h\, \mathbf{b}\cdot\nabla v_h) - (f, v_h)$$

and Y_h can be the space of piecewise constant functions on Ω. To emphasize that the approximate solution u_h depends on the choice of the stabilization parameter $y_h \in Y_h$, we shall write $u_h(y_h)$ instead of u_h in the following.

We introduce a functional $I_h : V_h \to \mathbb{R}$ such that $I_h(u_h(y_h))$ represents a measure of the error or the quality of $u_h(y_h)$. We assume that the solution $u_h(y_h)$ improves if the functional $\Phi_h(y_h) := I_h(u_h(y_h))$ decreases. Thus, our aim is to find $y_h \in Y_h$ such that $\Phi_h(y_h)$ is 'small'. This is a constrained nonlinear optimization problem since y_h has to be nonnegative and smaller than some upper bound. For example, for the SUPG method,

$$0 \le y_h|_K \le 10\,\tau_K \qquad \forall\ K \in \mathscr{T}_h\,, \tag{3}$$

where τ_K is defined by (2). The factor 10 can be changed to another value but numerical experiments indicate that the factor should not differ too much from 10.

Common minimization algorithms require at least the knowledge of the derivative of the function which should be minimized. Thus, we have to compute the Fréchet derivative of the functional Φ_h. Using the chain rule, we obtain

$$D\Phi_h(y_h) = DI_h(u_h(y_h))\,Du_h(y_h)\,.$$

However, it is not efficient to compute $D\Phi_h(y_h)$ using this formula since it requires the solution of dim Y_h linear problems of the size of the original discrete problem. Therefore, we first define the adjoint problem: Find $\psi_h(y_h) \in V_h$ such that

$$(\partial_u R_h)'(u_h(y_h), y_h)\, \psi_h(y_h) = DI_h(u_h(y_h))\,,$$

where $\langle(\partial_u R_h)'(w_h, y_h)v_h, \tilde{v}_h\rangle = \langle(\partial_u R_h)(w_h, y_h)\tilde{v}_h, v_h\rangle \quad \forall\, v_h, \tilde{v}_h, w_h \in V_h$, $y_h \in Y_h$. Since $R_h(u_h(y_h), y_h) = 0$, we have $\partial_u R_h(u_h(y_h), y_h)Du_h(y_h) + \partial_y R_h(u_h(y_h), y_h) = 0$. Thus, combining the above relations, we deduce that

$$D\Phi_h(y_h) = -(\partial_y R_h)'(u_h(y_h), y_h)\psi_h(y_h)\,,$$

where $\langle(\partial_y R_h)'(w_h, y_h)v_h, \tilde{y}_h\rangle = \langle(\partial_y R_h)(w_h, y_h)\tilde{y}_h, v_h\rangle \quad \forall\, v_h, w_h \in V_h, y_h, \tilde{y}_h \in Y_h$. Note that, for the SUPG method, the function $\psi_h(y_h)$ solves

$$a(v_h, \psi_h(y_h)) + (\mathscr{L}_h v_h, y_h\, \mathbf{b}\cdot\nabla\psi_h(y_h)) = \langle DI_h(u_h(y_h)), v_h\rangle \quad \forall\, v_h \in V_h$$

and the Fréchet derivative of Φ_h is given by

$$\langle D\Phi_h(y_h), \tilde{y}_h\rangle = -(\mathscr{L}_h u_h(y_h) - f, \tilde{y}_h\, \mathbf{b}\cdot\nabla\psi_h(y_h))\,.$$

5 Choice of the Functional I_h

In this section, we propose various choices of the functional I_h introduced in the previous section and present numerical results illustrating the properties of these functionals.

All numerical results were computed for $\Omega = (0, 1)^2$ and, in all cases, we considered a triangulation $\mathscr{T}_h$ of Ω constructed by dividing Ω into 32×32 equal squares and each square into two triangles by drawing a diagonal from bottom right to top left. The space W_h consisted of continuous piecewise linear functions. The functional Φ_h was minimized using the BFGS method [4]. The SUPG parameter was initialized by (2) and the SOLD parameter by 0. The SUPG parameter satisfied the constraints (3) and the SOLD parameter was required to be in the interval $[0, 1]$.

In each iteration of the BFGS method, one has to solve once the adjoint problem and several times the discrete problem for various values of the stabilization parameter. Consequently, the cost of the computation of an optimized SUPG stabilization parameter is significantly higher than the computation of the SUPG solution for a prescribed stabilization parameter. Comparing the cost of the optimization with the cost of the solution of a nonlinear SOLD method, the difference is not so large. We believe that the higher computational cost of the parameter optimization is justified by the quality of the resulting approximate solution, cf. the examples in this section.

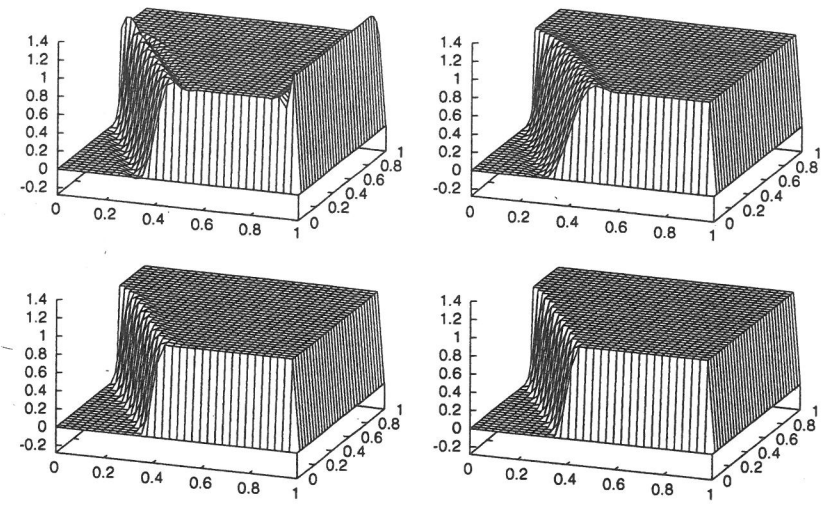


Fig. 1 Example 1: SUPG standard (*top left*), SUPG optimized using I_h^{res} (*top right*), SUPG optimized using $I_h^{\text{res}} + \alpha\, I_h^{\text{cross}}$ (*bottom left*), SOLD optimized using $I_h^{\text{res}} + \alpha\, I_h^{\text{cross}}$ (*bottom right*)

We denote by $\Gamma^+ = \overline{\{\mathbf{x} \in \partial\Omega\,;\ (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) > 0\}}$, $\Gamma^0 = \overline{\{\mathbf{x} \in \partial\Omega\,;\ (\mathbf{b}\cdot\mathbf{n})(\mathbf{x}) = 0\}}$ the outflow and characteristic boundaries of Ω, respectively. Furthermore, we set

$$G_h = \bigcup_{K \in \mathscr{G}_h} \overline{K} \qquad \text{with} \qquad \mathscr{G}_h = \{K \in \mathscr{T}_h\,;\ \overline{K} \cap \Gamma^+ \neq \emptyset \ \text{ or } \ \overline{K} \cap \Gamma^0 \neq \emptyset\}\,.$$

Note that G_h represents a strip along Γ^+ and Γ^0 made up of elements of $\mathscr{T}_h$ having at least one vertex on these parts of the boundary. A functional characterizing the quality of an approximate solution u_h of (1) can be now defined by

$$I_h^{\text{res}}(u_h) = \|\mathscr{L}_h\, u_h - f\|^2_{0,\Omega\setminus G_h}\,.$$

We exclude the strip G_h since even a nodally exact solution has a large error in G_h. Let us apply the functional I_h^{res} to the numerical solution of the following example.

Example 1 (Solution with an interior layer and two exponential boundary layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = f = 0$, $u_b(x,y) = 0$ for $x = 1$ or $y \leq 0.7$, and $u_b(x,y) = 1$ else. The function u_b could also be replaced by a function from $H^{1/2}(\partial\Omega)$ leading to the same numerical results as presented in this paper.

Figure 1 (top left) shows the SUPG solution computed with the stabilization parameter τ_K given by (2). If we optimize the stabilization parameter using the functional I_h^{res}, the spurious oscillations along the exponential boundary layer are

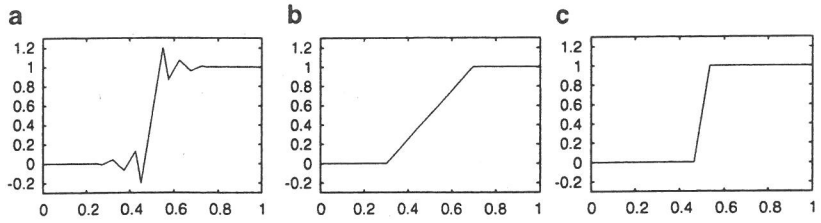


Fig. 2 Idealized cuts through approximate solutions across an interior layer

removed but those along the interior layer are not suppressed sufficiently. Moreover, the interior layer is smeared, see Fig. 1 (top right).

If we observe a cut through the solution in Fig. 1, top left, across the interior layer, we shall see a curve like in Fig. 2a. We would like to compute a solution without spurious oscillations corresponding to Fig. 2b or c. A candidate for a functional which prefers a solution without spurious oscillations is $\int_0^1 |u'|^p \mathrm{d}x$, where u represents the functions in Fig. 2. Denoting by d the width of the layer in Fig. 2b or c, the integral equals d^{1-p}. Since we prefer the curve c, we have to use $p < 1$. Thus, we may consider the functional

$$I_h^{\text{cross}}(u_h) = \int_{\Omega \setminus G_h} \sqrt{|\mathbf{b}^\perp \cdot \nabla u_h|}\, \mathrm{d}\mathbf{x}\,,$$

where $\mathbf{b}^\perp$ is a unit vector orthogonal to $\mathbf{b}$. In our implementation, the square root is regularized near 0, see [3] for details. If we now optimize the SUPG stabilization parameter using a combination of I_h^{res} and I_h^{cross}, the solution improves considerably, see Fig. 1 (bottom left). Finally, if we perform the optimization with the same functional but for the SOLD method, we obtain a solution without any visible spurious oscillations and with steep layers, see Fig. 1 (bottom right).

Example 2 (Solution with one exponential and two parabolic boundary layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b} = (1,0)^T$, $c = 0$, $f = 1$, and $u_b = 0$.

For this example, a comparison of the SUPG solution without parameter optimization and an optimized SOLD solution is given in Fig. 3. It can be observed, that the parameter optimization leads to an almost nodally exact solution.

Example 3 (Solution with two interior layers). We consider the convection-diffusion equation (1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b}(x,y) = (-y,x)^T$, and $c = f = 0$. On $\Gamma^N := \{0\} \times (0,1)$, we prescribe a homogeneous Neumann boundary condition whereas the Dirichlet boundary condition is considered only on $\Gamma^D := \partial\Omega \setminus \overline{\Gamma^N}$ with $u_b(x,y) = 1$ for $(x,y) \in (1/3, 2/3) \times \{0\}$ and $u_b(x,y) = 0$ else on Γ^D.

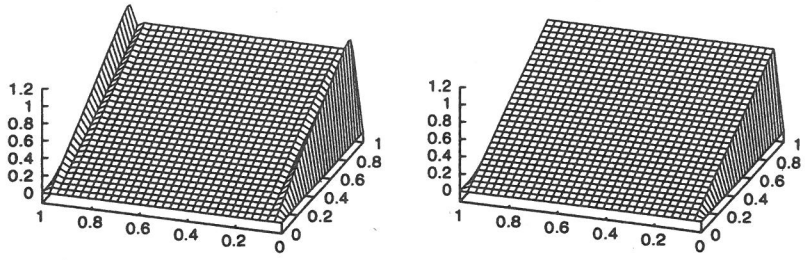


Fig. 3 Example 2: SUPG standard (*left*), SOLD optimized using I_h^{res} (*right*)

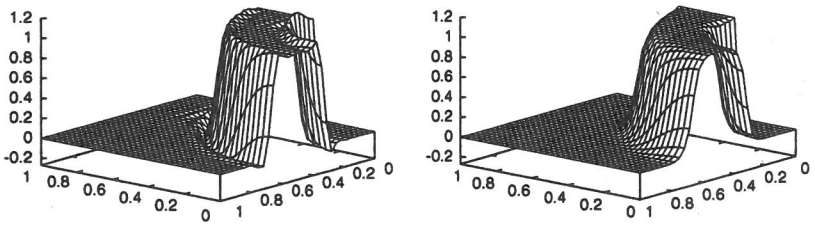


Fig. 4 Example 3: SUPG standard (*left*), SOLD standard (*right*)

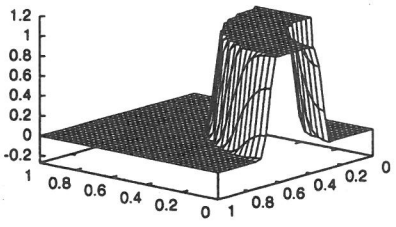


Fig. 5 Example 3: SOLD optimized using I_h^{cross}

Figure 4 shows results for this example obtained without parameter optimization. We see that the SOLD method suppresses the oscillations present in the SUPG solution but leads to a slight smearing of the layers. The quality of the SOLD solution obtained using parameter optimization is much better, see Fig. 5.

Acknowledgements The work of P. Knobloch is a part of the research project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Czech Republic under the grant No. 201/08/0012.

References

1. V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.*, 196:2197–2215, 2007.
2. V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.*, 197:1997–2014, 2008.
3. V. John, P. Knobloch, and S. B. Savescu. A posteriori optimization of parameters in stabilized methods for convection–diffusion problems – Part I. *Comput. Methods Appl. Mech. Engrg.*, 200:2916–2929, 2011.
4. J. Nocedal and S. J. Wright. *Numerical Optimization*. Springer Series in Operations Research and Financial Engineering. Springer, 2nd edition, 2006.

Chapter 5

Local projection stabilization

This chapter consists of the following publications:

P. Knobloch, L. Tobiska: On the stability of finite-element discretizations of convection–diffusion–reaction equations, *IMA Journal of Numerical Analysis* 31 (1): 147–164, 2011. p. 175

P. Knobloch, G. Lube: Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach, *Applied Numerical Mathematics* 59 (12): 2891–2907, 2009. p. 193

P. Knobloch: A generalization of the local projection stabilization for convection–diffusion–reaction equations, *SIAM Journal on Numerical Analysis* 48 (2): 659–680, 2010. p. 211

G.R. Barrenechea, V. John, P. Knobloch: A local projection stabilization finite element method with nonlinear crosswind diffusion for convection–diffusion–reaction equations, *ESAIM: Mathematical Modelling and Numerical Analysis* 47 (5): 1335–1366, 2013. p. 233

IMA Journal of Numerical Analysis (2011) **31**, 147–164
doi:10.1093/imanum/drp020
Advance Access publication on August 25, 2009

On the stability of finite-element discretizations of convection–diffusion–reaction equations

PETR KNOBLOCH†
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic

AND

LUTZ TOBISKA‡
Institute for Analysis and Computational Mathematics, Faculty of Mathematics, Otto von Guericke University Magdeburg, PF 4120, 39016 Magdeburg, Germany

[Received on 13 June 2008; revised on 22 April 2009]

A priori error estimates for the local projection (LP) stabilization applied to convection–diffusion–reaction equations are generally based on the coercivity of the underlying bilinear form with respect to the LP norm. We show that the bilinear form of the LP stabilization satisfies an inf–sup condition in a stronger norm that is equivalent to that of the streamline upwind/Petrov–Galerkin method. As a consequence, we get some insight into the stabilization mechanism of Galerkin discretizations of higher order.

Keywords: finite-element method; convection–diffusion–reaction equation; stability; inf–sup condition; stabilization; SUPG method; local projection.

1. Introduction

Let $\Omega \subset \mathbb{R}^d$, where $d \geqslant 1$, be a bounded domain with a polyhedral Lipschitz continuous boundary $\partial\Omega$ and let us consider the convection–diffusion–reaction equation

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + c\, u = f \quad \text{in } \Omega, \qquad u = u_b \quad \text{on } \partial\Omega. \tag{1.1}$$

We assume that ε is a positive constant and $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$ and $u_b \in H^{1/2}(\partial\Omega)$ are given functions satisfying

$$\sigma := c - \frac{1}{2} \operatorname{div} \mathbf{b} \geqslant \sigma_0 > 0, \tag{1.2}$$

where σ_0 is a constant. It is well known that, under the assumption (1.2), the boundary-value problem (1.1) has a unique solution in $H^1(\Omega)$. Standard Galerkin finite-element discretizations of (1.1) become unstable for $\varepsilon \to 0$, which was the origin of the development of stabilized finite-element discretizations (see, e.g., Roos *et al.*, 2008).

In this paper we concentrate on local projection (LP) stabilizations, which have been intensively studied during recent years (cf., e.g., Braack & Burman, 2006; Matthies *et al.*, 2007, 2008; Ganesan &

†Email: knobloch@karlin.mff.cuni.cz
‡Corresponding author. Email: tobiska@mathematik.uni-magdeburg.de

Tobiska, 2008; Knobloch & Lube, 2009; Rapin *et al.*, 2008). Our aim is to show that the LP methods are more stable than their coercivity suggests. For this we shall prove that the bilinear form of LP methods satisfies an inf–sup condition with respect to a norm that is stronger than the usual LP norm. Under additional assumptions, we shall show that this norm is equivalent to the streamline upwind/Petrov–Galerkin (SUPG) norm, which implies that LP methods are as stable as the SUPG method. As a particular case of our general results, we shall also establish improved stability properties of the standard Galerkin method.

The idea of deriving inf–sup conditions with respect to SUPG-like norms that are stronger than norms for which the coercivity holds can also be found in the analysis of other stabilized methods like the orthogonal subgrid scale (OSS) method (cf., Codina & Blasco, 2002; Badia & Codina, 2006) or the edge stabilization method (cf., Burman & Hansbo, 2004). The difference between the LP method and the OSS method is that the projection operator used for defining the stabilization term is constructed locally and not globally. The proof of the inf–sup condition by Codina *et al.* (2002) is based on this global OSS projection operator. However, the operator defining the LP method cannot be applied for this purpose. Therefore we introduce another locally acting projection operator that, in combination with the original LP operator, leads to stability with respect to the SUPG norm. It is by no means obvious that such an operator exists.

The paper is organized in the following way. First, in Section 2 we formulate several finite-element discretizations of (1.1) and discuss the coercivity of the corresponding bilinear forms. Then, in Section 3 we prove the main result of this paper, which is an inf–sup condition for a general bilinear form corresponding to LP methods. Section 4 investigates the equivalence between the norm used in the inf–sup condition and the SUPG norm. Furthermore, an improved stability of the standard Galerkin method for higher-order finite elements is established and discussed in Section 5. Finally, in Section 6 we compare the SUPG method with LP methods by means of numerical computations. Throughout the paper we use standard notation for usual function spaces and norms (see, e.g., Ciarlet, 1991). Given a vector $\mathbf{a} \in \mathbb{R}^d$, we denote by $|\mathbf{a}|$ its Euclidean norm.

2. Stabilized finite-element discretizations of convection–diffusion–reaction equations

Let $\mathcal{T}_h$ be a triangulation of Ω consisting of shape-regular cells K possessing the usual compatibility properties. We set $h_K = \operatorname{diam}(K)$ for any $K \in \mathcal{T}_h$ and assume that $h_K \leqslant h$ for all $K \in \mathcal{T}_h$. Using the triangulation $\mathcal{T}_h$, we define a finite-element space $W_h \subset H^1(\Omega)$ (see, e.g., Ciarlet, 1991), and we set $V_h = W_h \cap H_0^1(\Omega)$. In addition, we introduce a function $\widetilde{u}_{bh} \in W_h$ such that its trace approximates the boundary condition u_b. We shall discuss finite-element discretizations of the problem (1.1) that have the following form.

Find $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$\tilde{a}(u_h, v_h) = \langle \tilde{f}, v_h \rangle \quad \forall\, v_h \in V_h,$$

where $\tilde{a} : W_h \times W_h \to \mathbb{R}$ is a bilinear form and $\tilde{f} : V_h \to \mathbb{R}$ is a linear functional. We shall be particularly interested in the coercivity of $\tilde{a}$ on V_h, i.e., in the validity of the inequality

$$\tilde{a}(v_h, v_h) \geqslant C \|v_h\|^2 \quad \forall\, v_h \in V_h,$$

where $\|\cdot\|$ is a suitable norm on V_h and C is a positive constant that is independent of h and the data of the problem (1.1). Note that this stability property immediately implies that the discrete problem has a unique solution.

The simplest finite-element discretization of (1.1) is the Galerkin discretization, which is obtained by replacing the space $H_0^1(\Omega)$ in the weak formulation of (1.1) by its subspace V_h. This leads to a discrete problem with $\tilde{a} = a^G$ and $\tilde{f} = f$, where

$$a^G(u, v) = \varepsilon\,(\nabla u, \nabla v) + (\mathbf{b} \cdot \nabla u, v) + (c\,u, v)$$

and $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. The statement $\tilde{f} = f$ is understood in the sense that $\langle \tilde{f}, v_h \rangle = (f, v_h)$ for any $v_h \in V_h$. The bilinear form a^G is coercive on V_h with respect to the norm $|||\cdot|||_G$ defined by

$$|||v|||_G = (\varepsilon\,|v|_{1,\Omega}^2 + \|\sigma^{1/2}\,v\|_{0,\Omega}^2)^{1/2}.$$

In fact, integrating by parts, we even derive

$$a^G(v, v) = |||v|||_G^2 \quad \forall\, v \in H_0^1(\Omega). \tag{2.1}$$

It is well known that the Galerkin discretization is inappropriate if convection dominates diffusion since then the discrete solution is usually globally polluted by spurious oscillations (cf., e.g., Roos *et al.*, 2008). To enhance the stability and accuracy of the Galerkin discretization of (1.1) in the convection-dominated regime, various stabilization strategies have been developed. One of the most popular approaches is the SUPG method proposed by Brooks & Hughes (1982), which is given by $\tilde{a} = a_h^{\mathrm{SUPG}}$ and $\tilde{f} = f_h^{\mathrm{SUPG}}$, where

$$a_h^{\mathrm{SUPG}}(u, v) = a^G(u, v) + \sum_{K \in \mathcal{T}_h} (-\varepsilon\,\Delta u + \mathbf{b} \cdot \nabla u + cu, \delta \mathbf{b} \cdot \nabla v)_K,$$

$$\langle f_h^{\mathrm{SUPG}}, v \rangle = (f, v + \delta \mathbf{b} \cdot \nabla v)$$

and $\delta \in L^\infty(\Omega)$ is a non-negative stabilization parameter. As usual, $(\cdot, \cdot)_K$ denotes the inner product in $L^2(K)$ or $L^2(K)^d$. If

$$0 \leqslant \delta|_K \leqslant \min\left\{ \frac{\sigma_0}{2\,\|c\|_{0,\infty,K}^2}, \frac{h_K^2}{2\,\varepsilon\,\mu^2} \right\} \quad \forall\, K \in \mathcal{T}_h, \tag{2.2}$$

where μ is a constant from the inverse inequality

$$\|\Delta v_h\|_{0,K} \leqslant \mu\, h_K^{-1} |v_h|_{1,K} \qquad \forall\, v_h \in V_h, \quad K \in \mathcal{T}_h,$$

then the bilinear form a_h^{SUPG} is coercive on V_h with respect to the norm

$$|||v|||_{\mathrm{SUPG}} = (|||v|||_G^2 + \|\delta^{1/2}\,\mathbf{b} \cdot \nabla v\|_{0,\Omega}^2)^{1/2}. \tag{2.3}$$

Thus, if $\delta > 0$, then the SUPG method possesses a stronger stability in the streamline direction than the Galerkin discretization. The choice of δ significantly influences the accuracy of the SUPG solution and therefore extensive research has been devoted to the development of suitable formulas for this stabilization parameter (see, e.g., the review in John & Knobloch (2007)). Unfortunately, a general optimal definition of δ is still unknown. For finite elements of first order of accuracy, the parameter δ is often defined on any $K \in \mathcal{T}_h$ by the formula

$$\delta|_K = \frac{h_{K,\mathbf{b}}}{2\,|\mathbf{b}|}\,\xi_0(Pe_K) \qquad \text{with } \xi_0(\alpha) = \coth\alpha - \frac{1}{\alpha}, \quad Pe_K = \frac{|\mathbf{b}|\,h_{K,\mathbf{b}}}{2\varepsilon}, \tag{2.4}$$

where $h_{K,\mathbf{b}}$ is the diameter of K in the direction of the convection vector $\mathbf{b}$. Note that, generally, the parameters $h_{K,\mathbf{b}}$, Pe_K and $\delta|_K$ are functions of the points $\mathbf{x} \in K$. Under some simplifying assumptions, the formula (2.4) is optimal in the one-dimensional case (see Christie *et al.*, 1976). The function ξ_0 is sometimes approximated by (cf., e.g., Brooks & Hughes, 1982)

$$\xi_1(\alpha) = \max\left\{0, 1 - \frac{1}{\alpha}\right\} \quad \text{or} \quad \xi_2(\alpha) = \min\left\{1, \frac{\alpha}{3}\right\}. \tag{2.5}$$

Note that $\xi_1(\alpha) \leqslant \xi_0(\alpha) \leqslant \xi_2(\alpha)$ for any $\alpha \geqslant 0$. For higher-order finite elements the values of Pe_K and $\delta|_K$ should decrease with increasing polynomial degree on K (see, e.g., Codina *et al.*, 2002; Almeida & Silva, 1997; Galeão *et al.*, 2004). For a review and comparison of various (nonlinear) extensions of the SUPG method, we refer to John & Knobloch (2007, 2009).

During the last decade, stabilization techniques based on LPs have become very popular (see, e.g., Becker & Braack, 2004; Braack & Burman, 2006; Matthies *et al.*, 2007). To formulate an LP method we introduce a discontinuous finite-element space $D_h \subset L^2(\Omega)$ and denote by π_h the orthogonal L^2 projection of $L^2(\Omega)$ onto D_h, and of $L^2(\Omega)^d$ onto D_h^d. Furthermore, we define the so-called fluctuation operator $\kappa_h = \mathrm{id} - \pi_h$, where id is the identity operator on $L^2(\Omega)$, and on $L^2(\Omega)^d$. Then the discretization of (1.1) is given by $\tilde{a} = a_h^{\mathrm{LP1}}$ or $\tilde{a} = a_h^{\mathrm{LP2}}$ and $\tilde{f} = f$, where

$$a_h^{\mathrm{LP1}}(u, v) = a^G(u, v) + (\kappa_h(\mathbf{b} \cdot \nabla u), \tau\, \kappa_h(\mathbf{b} \cdot \nabla v)), \tag{2.6}$$

$$a_h^{\mathrm{LP2}}(u, v) = a^G(u, v) + (\kappa_h \nabla u, \tau \kappa_h \nabla v) \tag{2.7}$$

and $\tau \in L^\infty(\Omega)$ is a non-negative stabilization parameter. These bilinear forms are coercive on V_h with respect to the norms

$$|||v|||_{\mathrm{LP1}} = (|||v|||_G^2 + \|\tau^{1/2}\, \kappa_h(\mathbf{b} \cdot \nabla v)\|_{0,\Omega}^2)^{1/2},$$

$$|||v|||_{\mathrm{LP2}} = (|||v|||_G^2 + \|\tau^{1/2} \kappa_h \nabla v\|_{0,\Omega}^2)^{1/2},$$

respectively.

Let us describe the typical finite-element spaces W_h and D_h used in the LP method. We shall assume that the triangulation $\mathcal{T}_h$ consists of simplices or mapped rectangles and, for simplicity, that all cells K of $\mathcal{T}_h$ are affine equivalent to a reference cell $\widehat{K}$. We denote by $F_K : \widehat{K} \to K$ the respective affine regular mappings such that $F_K(\widehat{K}) = K$. Let $\widehat{W} \subset H^1(\widehat{K})$ and $\widehat{D} \subset L^2(\widehat{K})$ be finite-dimensional spaces, and for any $K \in \mathcal{T}_h$ let us set

$$W(K) = \{\widehat{v} \circ F_K^{-1}; \widehat{v} \in \widehat{W}\}, \quad D(K) = \{\widehat{v} \circ F_K^{-1}; \widehat{v} \in \widehat{D}\}. \tag{2.8}$$

Now we define the spaces W_h and D_h by

$$W_h = \{v \in H^1(\Omega); v|_K \in W(K) \quad \forall K \in \mathcal{T}_h\}, \tag{2.9}$$

$$D_h = \{v \in L^2(\Omega); v|_K \in D(K) \quad \forall K \in \mathcal{T}_h\}. \tag{2.10}$$

Originally, the LP method was designed as a two-level approach (see Becker & Braack, 2001), since the space W_h was constructed on a mesh obtained by refining the triangulation $\mathcal{T}_h$ used for constructing the space D_h (cf., Fig. 1 for $d = 2$). This corresponds to refining the reference cell $\widehat{K}$ into cells

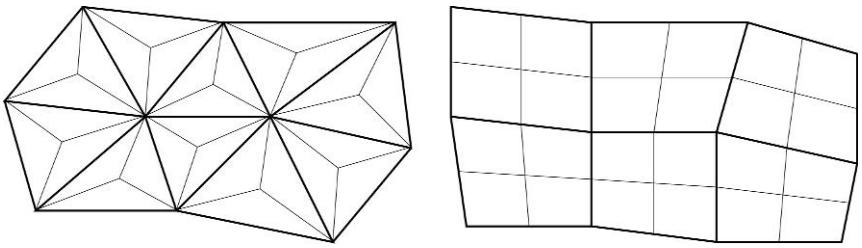


FIG. 1. Refinements of the triangulation $\mathcal{T}_h$ used for constructing the space W_h in the two-level approach.

$\widehat{K}_1, \ldots, \widehat{K}_n$. A crucial property of these refinements is that they always create an additional vertex in the interior of $\widehat{K}$. Given a positive integer r, the space $\widehat{W}$ is defined by

$$\widehat{W} = \{\widehat{v} \in H^1(\widehat{K}); \widehat{v}|_{\widehat{K}_i} \in R_r(\widehat{K}_i), i = 1, \ldots, n\},$$

where $R_r(\widehat{K}_i) = P_r(\widehat{K}_i)$ in the simplicial case and $R_r(\widehat{K}_i) = Q_r(\widehat{K}_i)$ in the mapped rectangular case. The space $\widehat{D}$ can be defined as $P_{r-1}(\widehat{K})$ in both cases. In the mapped rectangular case we can also set $\widehat{D} = Q_{r-1}(\widehat{K})$. Note that the space

$$\widehat{B} = \widehat{W} \cap H_0^1(\widehat{K}) \tag{2.11}$$

always has a positive dimension. Moreover, it is easy to verify that

$$\sup_{\widehat{v} \in \widehat{W} \cap H_0^1(\widehat{K})} \frac{(\widehat{v}, \widehat{q})_{\widehat{K}}}{\|\widehat{v}\|_{0,\widehat{K}}} \geqslant \beta_{\mathrm{LP}} \|\widehat{q}\|_{0,\widehat{K}} \quad \forall \widehat{q} \in \widehat{D} \tag{2.12}$$

with a positive constant β_{LP}.

Recently, a one-level approach was introduced by Matthies *et al.* (2007) based on using higher-order polynomials rather than refining the triangulation. Using the same spaces $\widehat{D}$ as for the two-level approach, we set

$$\widehat{W} = R_r(\widehat{K}) + \hat{b} \cdot \widehat{D},$$

where $\hat{b} \in H_0^1(\widehat{K}) \setminus \{0\}$ is a polynomial of the lowest possible degree. Thus, again, the space $\widehat{B}$ defined in (2.11) has a positive dimension and the inf–sup condition (2.12) holds.

The space W_h of both the one-level and the two-level approaches can be considered as an enriched finite-element space. In the one-level approach higher-order polynomials are added, whereas in the two-level approach piecewise polynomials on a refinement of the triangulation $\mathcal{T}_h$ are used. This view of the two-level approach differs from other papers, for example, Becker & Braack (2001, 2004), Braack & Burman (2006), Matthies *et al.* (2007) and Rapin *et al.* (2008), in which the space W_h is defined on an 'unrefined' triangulation $\mathcal{T}_h$ and the projection space D_h is defined on a triangulation $\mathcal{T}_{2h}$ obtained by coarsening the triangulation $\mathcal{T}_h$.

We refer to Matthies *et al.* (2007) for a detailed description of various pairs of finite-element spaces W_h and D_h that are applicable to the LP method. Let us remark that the inf–sup condition (2.12) is equivalent to inf–sup conditions

$$\sup_{v_h \in W(K) \cap H_0^1(K)} \frac{(v_h, q_h)_K}{\|v_h\|_{0,K}} \geqslant \beta_{\mathrm{LP}} \|q_h\|_{0,K} \qquad \forall q_h \in D(K), \quad K \in \mathcal{T}_h, \tag{2.13}$$

which were introduced in the general theory of Matthies *et al.* (2007) as a requirement on any pair of spaces W_h and D_h used in the LP method. Concerning the choice of the stabilization parameter τ, we refer to Tobiska (2009), where optimal parameters were derived for the one-level LP method in the one-dimensional case with constant coefficients.

The above examples show that, in both the one-level and the two-level approaches of the LP method, the space W_h contains on each element $K \in \mathcal{T}_h$ a nontrivial bubble space $B(K) = W(K) \cap H_0^1(K)$. This will be crucial for our considerations in the next section.

Comparing the norms $|||\cdot|||_{\mathrm{LP1}}$ and $|||\cdot|||_{\mathrm{SUPG}}$, we see that the SUPG norm is stronger and we would expect that the SUPG method is more stable than the LP method. However, coercivity of the bilinear form in a certain norm is not necessary for the method to be stable in that norm. It is enough that an inf–sup condition is satisfied. In the following section we will show that an inf–sup condition holds for the LP methods in a norm stronger than those in which coercivity can be proven. This enables us to recover essentially the same stability and convergence properties for the LP method as we have for the SUPG method.

3. Inf–sup condition

In this section we consider the bilinear forms a_h^{LP1} and a_h^{LP2} as special cases of the general form

$$a_h(u,v) = a^G(u,v) + \sum_{K\in\mathcal{T}_h} \tau_K s_K(u,v). \tag{3.1}$$

Here $s_K : H^1(K) \times H^1(K) \to \mathbb{R}$ are non-negative bilinear forms such that

$$s_K(u,u) \leqslant \gamma_K(\mathbf{b})|u|_{1,K}^2, \quad s_K(u,v) \leqslant \sqrt{s_K(u,u)}\sqrt{s_K(v,v)} \quad \forall u,v \in H^1(K), \tag{3.2}$$

where $\gamma_K(\mathbf{b}) = \|\mathbf{b}\|_{0,\infty,K}^2$ if the bilinear form is given by (2.6) and $\gamma_K(\mathbf{b}) = 1$ if the bilinear form is given by (2.7). Moreover, we assume that $|\cdot|_{s_K} := \sqrt{s_K(\cdot,\cdot)}$ is a seminorm on $H^1(K)$. For simplicity, the stabilization parameter τ is now considered piecewise constant and denoted by τ_K on each cell $K \in \mathcal{T}_h$.

In view of (2.1), the bilinear form a_h is obviously coercive on V_h with respect to the norm

$$|||v|||_{\mathrm{LP}} = \left(|||v|||_G^2 + \sum_{K\in\mathcal{T}_h} \tau_K\, s_K(v,v) \right)^{1/2}, \tag{3.3}$$

generalizing the norms $|||\cdot|||_{\mathrm{LP1}}$ and $|||\cdot|||_{\mathrm{LP2}}$. Our aim is to show that the bilinear form a_h satisfies an inf–sup condition in a stronger norm (cf., (3.4) below). For this we shall assume that there exists a space $B_h \subset V_h$ such that

$$B_h = \bigoplus_{K\in\mathcal{T}_h} B(K) \ \text{ with } \ B(K) \subset H_0^1(K).$$

We have seen in the preceding section that such a nontrivial space B_h exists for typical finite-element spaces V_h used in both the one-level and the two-level approaches of the LP method.

For any $K \in \mathcal{T}_h$ let Π_K be the orthogonal L^2 projection of $L^2(K)$ onto $B(K)$. Combining and modifying the norms $|||\cdot|||_{\mathrm{SUPG}}$ and $|||\cdot|||_{\mathrm{LP}}$, we introduce the norm

$$|||v||| = \left(|||v|||_G^2 + \sum_{K\in\mathcal{T}_h} \left\{ \delta_K \,\|\Pi_K(\mathbf{b}\cdot\nabla v)\|_{0,K}^2 + \tau_K\, s_K(v,v) \right\} \right)^{1/2}, \tag{3.4}$$

where we now consider a piecewise constant parameter δ_K. We assume that

$$0 \leqslant \delta_K \leqslant C_1 h_K^2 (\max\{\varepsilon, h_K \|\mathbf{b}\|_{0,\infty,K}\})^{-1}, \tag{3.5}$$

$$0 \leqslant \tau_K \gamma_K(\mathbf{b}) \leqslant C_2 \max\{\varepsilon, \delta_K \|\mathbf{b}\|_{0,\infty,K}^2\}. \tag{3.6}$$

The assumption (3.5) is in agreement with the relations (2.4), (2.5) and, partially, (2.2). The upper bound in (3.6) corresponds to the usual choice of δ_K and τ_K in the convection-dominated limit where $\delta_K \|\mathbf{b}\|_{0,\infty,K} \sim h_K$ and $\tau_K \gamma_K(\mathbf{b}) \sim h_K \|\mathbf{b}\|_{0,\infty,K}$ (cf., Roos *et al.*, 2008). Note that we always have

$$\delta_K \|\mathbf{b}\|_{0,\infty,K} \leqslant C_1 h_K, \quad \tau_K \gamma_K(\mathbf{b}) \delta_K \leqslant C_1 C_2 \max\{1, C_1\} h_K^2. \tag{3.7}$$

We shall also need the inverse inequality

$$|v_h|_{1,K} \leqslant C_3 h_K^{-1} \|v_h\|_{0,K} \quad \forall K \in \mathcal{T}_h, v_h \in W_h. \tag{3.8}$$

The constants C_1, C_2 and C_3 are assumed to be independent of K, h and all data of the problem (1.1).

THEOREM 3.1 The bilinear form a_h satisfies

$$\sup_{v_h \in V_h} \frac{a_h(u_h, v_h)}{|||v_h|||} \geqslant \beta |||u_h||| \quad \forall u_h \in V_h \tag{3.9}$$

with a positive constant β that is independent of h and ε.

Proof. Using (2.1), for any $u_h \in V_h$ we obtain

$$a_h(u_h, u_h) = |||u_h|||_G^2 + \sum_{K \in \mathcal{T}_h} \tau_K s_K(u_h, u_h). \tag{3.10}$$

We see that the term $\sum_{K \in \mathcal{T}_h} \delta_K \|\Pi_K(\mathbf{b} \cdot \nabla u_h)\|_{0,K}^2$ is missing on the right-hand side of (3.10), which does not allow us to conclude the coercivity of a_h on V_h with respect to $|||\cdot|||$. Therefore, given $u_h \in V_h$, we shall construct a function $v_h \in V_h$ such that

$$a_h(u_h, v_h) \geqslant |||u_h|||^2 \quad \text{and} \quad |||u_h||| \geqslant \beta |||v_h|||. \tag{3.11}$$

The inequalities (3.11) immediately imply the inf–sup condition (3.9).

First, we introduce a function $z_h \in B_h$ by

$$z_h|_K = \delta_K \Pi_K(\mathbf{b} \cdot \nabla u_h) \quad \forall K \in \mathcal{T}_h.$$

An important property of this function is that

$$(\mathbf{b} \cdot \nabla u_h, z_h)_K = \delta_K \|\Pi_K(\mathbf{b} \cdot \nabla u_h)\|_{0,K}^2 \quad \forall K \in \mathcal{T}_h.$$

Consequently, we have

$$a_h(u_h, z_h) = \sum_{K \in \mathcal{T}_h} \delta_K \|\Pi_K(\mathbf{b} \cdot \nabla u_h)\|_{0,K}^2 + \varepsilon(\nabla u_h, \nabla z_h) + (c\, u_h, z_h) + \sum_{K \in \mathcal{T}_h} \tau_K s_K(u_h, z_h). \tag{3.12}$$

Employing the relations (3.8), (3.5), (3.7) and (3.2), we obtain for any $K \in \mathcal{T}_h$, that

$$\varepsilon\,|z_h|_{1,K}^2 \leqslant C_3^2\,\varepsilon\,h_K^{-2}\,\|z_h\|_{0,K}^2 \leqslant C_1\,C_3^2\,\delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2, \tag{3.13}$$

$$\|z_h\|_{0,K} \leqslant \delta_K\,\|\mathbf{b}\|_{0,\infty,K}\,|u_h|_{1,K} \leqslant C_1\,C_3\,\|u_h\|_{0,K}, \tag{3.14}$$

$$\begin{aligned}\tau_K\,s_K(z_h,z_h) &\leqslant \tau_K\,\gamma_K(\mathbf{b})|z_h|_{1,K}^2 \leqslant C_3^2\,\tau_K\,\gamma_K(\mathbf{b})\,\delta_K^2\,h_K^{-2}\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2\\ &\leqslant C_1\,C_2\,C_3^2\max\{1,C_1\}\,\delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2.\end{aligned} \tag{3.15}$$

Therefore, applying the inequality $a\,b \leqslant a^2 + \frac{1}{4}\,b^2$, which is valid for any $a, b \in \mathbb{R}$, we derive

$$\begin{aligned}&|\varepsilon\,(\nabla u_h,\nabla z_h)_K + (c\,u_h,z_h)_K + \tau_K\,s_K(u_h,z_h)|\\ &\quad\leqslant \varepsilon\,|u_h|_{1,K}\,|z_h|_{1,K} + \|c\|_{0,\infty,K}\,\|u_h\|_{0,K}\,\|z_h\|_{0,K} + \tau_K\,\sqrt{s_K(u_h,u_h)}\,\sqrt{s_K(z_h,z_h)}\\ &\quad\leqslant \zeta(\varepsilon\,|u_h|_{1,K}^2 + \|\sigma^{1/2}u_h\|_{0,K}^2 + \tau_K s_K(u_h,u_h)) + \frac{1}{2}\,\delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2\end{aligned} \tag{3.16}$$

with

$$\zeta = C_1\,C_3^2 + C_1\,C_2\,C_3^2\max\{1,C_1\} + C_1\,C_3\,\|c\|_{0,\infty,\Omega}\,\sigma_0^{-1}.$$

Summing (3.16) over all cells $K \in \mathcal{T}_h$ and using (3.12) and (3.10), we get

$$a_h(u_h,z_h) \geqslant \frac{1}{2}\sum_{K\in\mathcal{T}_h}\delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2 - \zeta\,a_h(u_h,u_h).$$

Thus $v_h \in V_h$, given by

$$v_h := 2\,z_h + (1+2\,\zeta)u_h,$$

satisfies the first inequality in (3.11). To establish the second inequality in (3.11) it suffices to show that

$$|||z_h|||^2 \leqslant C_1^2\,C_3^2\big(1 + C_2 + \|\sigma\|_{0,\infty,\Omega}\,\sigma_0^{-1}\big)\sum_{K\in\mathcal{T}_h}\big(\varepsilon\,|u_h|_{1,K}^2 + \|\sigma^{1/2}u_h\|_{0,K}^2 + \delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2\big). \tag{3.17}$$

Using (3.8), for any $K \in \mathcal{T}_h$ we obtain

$$|z_h|_{1,K} \leqslant C_3\,\delta_K\,h_K^{-1}\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}. \tag{3.18}$$

This implies, in view of the first inequality in (3.7), that

$$|z_h|_{1,K} \leqslant C_3\,\delta_K\,\|\mathbf{b}\|_{0,\infty,K}\,h_K^{-1}\,|u_h|_{1,K} \leqslant C_1\,C_3\,|u_h|_{1,K}, \tag{3.19}$$

$$\|\Pi_K(\mathbf{b}\cdot\nabla z_h)\|_{0,K} \leqslant \|\mathbf{b}\|_{0,\infty,K}\,|z_h|_{1,K} \leqslant C_1\,C_3\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}. \tag{3.20}$$

Finally, applying (3.2), (3.6), (3.19), (3.18) and (3.7), we get

$$\tau_K\,s_K(z_h,z_h) \leqslant \tau_K\,\gamma_K(\mathbf{b})\,|z_h|_{1,K}^2 \leqslant C_1^2\,C_2\,C_3^2\max\big\{\varepsilon\,|u_h|_{1,K}^2, \delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla u_h)\|_{0,K}^2\big\}. \tag{3.21}$$

Now the inequality (3.17) follows from (3.19)–(3.21) and (3.14). □

4. Stability of LP methods

In this section we estimate the norm $|||\cdot|||$ defined in (3.4) from below by a norm similar to the SUPG norm $|||\cdot|||_{\mathrm{SUPG}}$. As a conclusion, the LP stabilization controls not only the fluctuations $\|\kappa_h(\mathbf{b}\cdot\nabla u_h)\|_{0,K}$ or $\|\kappa_h\nabla u_h\|_{0,K}$, but also $\|\mathbf{b}\cdot\nabla u_h\|_{0,K}$, i.e., the derivatives in the streamline direction. Roughly speaking, the LP method is as stable as the SUPG method.

We shall assume that all cells K of the triangulation $\mathcal{T}_h$ are affine equivalent to a reference cell $\widehat{K}$ and we again use the notation F_K for an affine regular mapping that maps $\widehat{K}$ onto K. We introduce finite-dimensional spaces $\widehat{B} \subset H_0^1(\widehat{K})$, $\widehat{D} \subset L^2(\widehat{K})$ and $\widehat{W} \subset H^1(\widehat{K})$ and assume that

$$B(K) = \left\{\widehat{v} \circ F_K^{-1}; \widehat{v} \in \widehat{B}\right\} \quad \forall\, K \in \mathcal{T}_h.$$

Furthermore, we assume that the approximation and projection spaces are given by (2.8)–(2.10). For any $K \in \mathcal{T}_h$ we denote by π_K the orthogonal L^2 projection of $L^2(K)$ onto $D(K)$, and of $L^2(K)^d$ onto $D(K)^d$, and we set $\kappa_K = \mathrm{id} - \pi_K$, where now id is the identity operator on $L^2(K)$, and on $L^2(K)^d$. We shall consider two bilinear forms s_K corresponding to the LP method, i.e.,

$$s_K(u, v) = (\kappa_K(\mathbf{b}\cdot\nabla u), \kappa_K(\mathbf{b}\cdot\nabla v))_K \tag{4.1}$$

or

$$s_K(u, v) = (\kappa_K\nabla u, \kappa_K\nabla v)_K. \tag{4.2}$$

We shall start with the case (4.1), where the results are more satisfactory.

LEMMA 4.1 Suppose that $\widehat{D}\cap\widehat{B}^\perp = \{0\}$, where $\widehat{B}^\perp$ denotes the orthogonal complement of $\widehat{B}$ in $L^2(\widehat{K})$. If the bilinear forms s_K are given by (4.1) then the norm $|||\cdot|||$ defined in (3.4) satisfies

$$|||v_h||| \leqslant \left(|||v_h|||_G^2 + \sum_{K\in\mathcal{T}_h}(\delta_K + \tau_K)\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2\right)^{1/2} \quad \forall\, v_h \in W_h. \tag{4.3}$$

If, in addition, $\mathbf{b}$ is piecewise polynomial, i.e., $\mathbf{b}|_K \in P_q(K)^d$ for some fixed $q \in \mathbb{N}_0$ and any $K \in \mathcal{T}_h$, then we also have

$$|||v_h||| \geqslant \left(|||v_h|||_G^2 + C\sum_{K\in\mathcal{T}_h}\min\{\delta_K, \tau_K\}\,\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2\right)^{1/2} \quad \forall\, v_h \in W_h, \tag{4.4}$$

where C is positive and depends only on $\widehat{B}$, $\widehat{D}$, $\widehat{W}$ and q. If $\mathbf{b}$ is not piecewise polynomial then there exists $h_0 > 0$ such that, for $0 < h \leqslant h_0$, we have

$$|||v_h||| \geqslant \left(\varepsilon|v_h|_{1,\Omega}^2 + \frac{\sigma_0}{2}\|v_h\|_{0,\Omega}^2 + C\sum_{K\in\mathcal{T}_h}\min\{\delta_K, \tau_K\}\,\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2\right)^{1/2} \quad \forall\, v_h \in W_h, \tag{4.5}$$

where C is a positive constant that is independent of h and the data of the problem (1.1). If $\mathbf{b} \in W^{2,\infty}(\Omega)^d$ or $\mathbf{b} \neq \mathbf{0}$ in $\overline{\Omega}$ then h_0 does not depend on ε.

Proof. The inequality (4.3) easily follows from the fact that Π_K and π_K are orthogonal L^2 projections. Let us prove the inequality (4.4). Consider any $v_h \in W_h$ and any $K \in \mathcal{T}_h$. Then $(\nabla v_h) \circ F_K \in \widehat{Z}^d$, where $\widehat{Z} = \{\mathbf{a} \cdot \widehat{\nabla} \widehat{v}; \mathbf{a} \in \mathbb{R}^d, \widehat{v} \in \widehat{W}\}$. Let us define $w = (\mathbf{b} \cdot \nabla v_h)|_K$ and $\widehat{w} = w \circ F_K$ and assume that $\mathbf{b}|_K \in P_q(K)^d$. Then $\widehat{w} \in \widehat{X} := \mathrm{span}\{\widehat{p}\,\widehat{z}; \widehat{p} \in P_q(\widehat{K}), \widehat{z} \in \widehat{Z}\}$ and we obtain

$$\|\Pi_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2 = \|\Pi_K\, w\|_{0,K}^2 + \|\kappa_K\, w\|_{0,K}^2 = \frac{|K|}{|\widehat{K}|}\left(\|\widehat{\Pi}\,\widehat{w}\|_{0,\widehat{K}}^2 + \|\widehat{\kappa}\,\widehat{w}\|_{0,\widehat{K}}^2\right),$$

where $|K|$ and $|\widehat{K}|$ are the volumes of K and $\widehat{K}$, respectively, and $\widehat{\Pi} : L^2(\widehat{K}) \to \widehat{B}$ and $\widehat{\kappa} : L^2(\widehat{K}) \to L^2(\widehat{K})$ are defined analogously as Π_K and κ_K, respectively. If both $\widehat{\Pi}\,\widehat{w} = 0$ and $\widehat{\kappa}\,\widehat{w} = 0$ then we deduce that $\widehat{w} \in \widehat{B}^\perp$ and $\widehat{w} \in \widehat{D}$. Thus the assumption that $\widehat{D} \cap \widehat{B}^\perp = \{0\}$ implies that $\widehat{w} = 0$. Consequently, the functional $(\|\widehat{\Pi} \cdot\|_{0,\widehat{K}}^2 + \|\widehat{\kappa} \cdot\|_{0,\widehat{K}}^2)^{1/2}$ is a norm on $L^2(\widehat{K})$. Since all norms are equivalent on the finite-dimensional space $\widehat{X}$, there exists a constant C_4 such that

$$\|\Pi_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2 \geqslant C_4\, \frac{|K|}{|\widehat{K}|}\, \|\widehat{w}\|_{0,\widehat{K}}^2 = C_4\, \|\mathbf{b} \cdot \nabla v_h\|_{0,K}^2. \tag{4.6}$$

This proves the inequality (4.4). To simplify the proof of the inequality (4.5) we first set

$$A_K(\mathbf{b}, v_h) = \|\Pi_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{b} \cdot \nabla v_h)\|_{0,K}^2.$$

Let $\mathbf{b} \in W^{q+1,\infty}(\Omega)^d$ for some $q \in \mathbb{N}_0$, and for any $K \in \mathcal{T}_h$ let us denote by $\mathbf{b}_K$ the orthogonal L^2 projection of $\mathbf{b}|_K$ onto $P_q(K)$. Using the triangular inequality, (4.6) with $\mathbf{b} = \mathbf{b}_K$, and the L^2 stability of Π_K and π_K, we derive

$$\begin{aligned}
A_K(\mathbf{b}, v_h) &\geqslant \frac{1}{2}\, A_K(\mathbf{b}_K, v_h) - A_K(\mathbf{b} - \mathbf{b}_K, v_h) \geqslant \frac{1}{2}\, C_4\, \|\mathbf{b}_K \cdot \nabla v_h\|_{0,K}^2 - A_K(\mathbf{b} - \mathbf{b}_K, v_h) \\
&\geqslant \frac{1}{4}\, C_4\, \|\mathbf{b} \cdot \nabla v_h\|_{0,K}^2 - \frac{1}{2}\, C_4\, \|(\mathbf{b} - \mathbf{b}_K) \cdot \nabla v_h\|_{0,K}^2 - A_K(\mathbf{b} - \mathbf{b}_K, v_h) \\
&\geqslant \frac{1}{4}\, C_4\, \|\mathbf{b} \cdot \nabla v_h\|_{0,K}^2 - \left(2 + \frac{1}{2}\, C_4\right) \|\mathbf{b} - \mathbf{b}_K\|_{0,\infty,K}^2\, |v_h|_{1,K}^2.
\end{aligned}$$

Therefore, using the inverse inequality (3.8), we obtain

$$\begin{aligned}
|||v_h|||^2 \geqslant |||v_h|||_G^2 - C_5 \max_{K \in \mathcal{T}_h} (\delta_K\, h_K^{-2}\, \|\mathbf{b} - \mathbf{b}_K\|_{0,\infty,K}^2)\, \|v_h\|_{0,\Omega}^2 \\
+ \frac{1}{4}\, C_4 \sum_{K \in \mathcal{T}_h} \min\{\delta_K, \tau_K\}\, \|\mathbf{b} \cdot \nabla v_h\|_{0,K}^2
\end{aligned}$$

with $C_5 = \left(2 + \frac{1}{2}\, C_4\right) C_3^2$. Since $\|\mathbf{b} - \mathbf{b}_K\|_{0,\infty,K} \leqslant C_6\, h_K^{q+1}\, |\mathbf{b}|_{q+1,\infty,K}$ and $\|\mathbf{b}_K\|_{0,\infty,K} \leqslant C_7\, \|\mathbf{b}\|_{0,\infty,K}$ with constants C_6 and C_7 depending only on q, d and the shape regularity of K (see, e.g., Ciarlet, 1991), we derive using (3.5) that

$$\delta_K\, h_K^{-2}\, \|\mathbf{b} - \mathbf{b}_K\|_{0,\infty,K}^2 \leqslant C_1 \min\left\{2\, C_6\, C_7\, h_K^q\, |\mathbf{b}|_{q+1,\infty,K}, \frac{C_6^2\, h_K^{2q+2}\, |\mathbf{b}|_{q+1,\infty,K}^2}{\max\{\varepsilon, h_K\, \|\mathbf{b}\|_{0,\infty,K}\}}\right\}.$$

Thus, setting $C_8 = C_1\, C_6\, \max\{C_6, 2\, C_7\}$, we get

$$\delta_K\, h_K^{-2}\, \|\mathbf{b}-\mathbf{b}_K\|_{0,\infty,K}^2 \leqslant C_8 \max\left\{|\mathbf{b}|_{q+1,\infty,\Omega}, |\mathbf{b}|_{q+1,\infty,\Omega}^2\right\} \min\left\{h^q, h^{2\,q+2}\,\varepsilon^{-1}, h^{2\,q+1}\left(\min_{\overline{\Omega}}\,|\mathbf{b}|\right)^{-1}\right\}.$$

Since $|||v_h|||_G^2 \geqslant \varepsilon|v_h|_{1,\Omega}^2 + \sigma_0\|v_h\|_{0,\Omega}^2$, we obtain (4.5) for sufficiently small h. □

For s_K defined by (4.2), an estimate like (4.3) does not hold. Nevertheless, we can still prove analogous lower bounds as in (4.4) and (4.5) from which stability follows with respect to the norms given by the right-hand sides of (4.4) and (4.5).

LEMMA 4.2 Suppose that $\widehat{D}\cap\widehat{B}^{\perp} = \{0\}$, where $\widehat{B}^{\perp}$ denotes the orthogonal complement of $\widehat{B}$ in $L^2(\widehat{K})$. If the bilinear forms s_K are given by (4.2) and $\mathbf{b}$ is constant, where $\mathbf{b} \neq \mathbf{0}$, then

$$|||v_h||| \geqslant \left(|||v_h|||_G^2 + C\sum_{K\in\mathcal{T}_h}\min\left\{\delta_K, \frac{\tau_K}{|\mathbf{b}|^2}\right\}\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2\right)^{1/2} \qquad \forall\, v_h \in W_h, \tag{4.7}$$

where C is positive and depends only on $\widehat{B}$, $\widehat{D}$ and $\widehat{W}$. If $\mathbf{b}$ is not constant then there exists $h_0 > 0$ such that, for $0 < h \leqslant h_0$, we have

$$|||v_h||| \geqslant \left(\varepsilon|v_h|_{1,\Omega}^2 + \frac{\sigma_0}{2}\|v_h\|_{0,\Omega}^2 + C\sum_{K\in\mathcal{T}_h}\min\left\{\delta_K, \frac{\tau_K}{\|\mathbf{b}\|_{0,\infty,K}^2}\right\}\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2\right)^{1/2} \qquad \forall\, v_h \in W_h, \tag{4.8}$$

where C is a positive constant that is independent of h and the data of the problem (1.1). If $\mathbf{b} \neq \mathbf{0}$ in $\overline{\Omega}$ then h_0 does not depend on ε.

Proof. Let us assume that $\mathbf{b}$ is constant and $\mathbf{b} \neq \mathbf{0}$. Consider any $v_h \in W_h$ and any $K \in \mathcal{T}_h$. Then

$$\delta_K\|\Pi_K(\mathbf{b}\cdot\nabla v_h)\|_{0,K}^2 + \tau_K\|\kappa_K\nabla v_h\|_{0,K}^2 \geqslant \min\left\{\delta_K, \frac{\tau_K}{|\mathbf{b}|^2}\right\}\left(\|\Pi_K(\mathbf{b}\cdot\nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{b}\cdot\nabla v_h)\|_{0,K}^2\right)$$

and (4.7) follows using (4.6). If $\mathbf{b}$ is not constant then we consider any $\mathbf{a} \in \mathbb{R}^d \setminus \{\mathbf{0}\}$ and use the estimates

$$\begin{aligned}
&\delta_K\|\Pi_K(\mathbf{b}\cdot\nabla v_h)\|_{0,K}^2 + \tau_K\|\kappa_K\nabla v_h\|_{0,K}^2 \\
&\quad\geqslant \min\left\{\delta_K, \frac{\tau_K}{|\mathbf{a}|^2}\right\}\left(\|\Pi_K(\mathbf{b}\cdot\nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{a}\cdot\nabla v_h)\|_{0,K}^2\right) \\
&\quad\geqslant \min\left\{\delta_K, \frac{\tau_K}{|\mathbf{a}|^2}\right\}\left(\frac{1}{2}\|\Pi_K(\mathbf{a}\cdot\nabla v_h)\|_{0,K}^2 + \|\kappa_K(\mathbf{a}\cdot\nabla v_h)\|_{0,K}^2\right) - \delta_K\,\|\Pi_K((\mathbf{b}-\mathbf{a})\cdot\nabla v_h)\|_{0,K}^2 \\
&\quad\geqslant \frac{1}{4}\,C_4\min\left\{\delta_K, \frac{\tau_K}{|\mathbf{a}|^2}\right\}\|\mathbf{b}\cdot\nabla v_h\|_{0,K}^2 - \left(1+\frac{1}{2}\,C_4\right)\delta_K\,\|(\mathbf{b}-\mathbf{a})\cdot\nabla v_h\|_{0,K}^2,
\end{aligned}$$

where we have applied (4.6) to derive the last inequality. If $\mathbf{b}_{0K} := (\mathbf{b}, 1)_K/|K| \neq \mathbf{0}$ then we may set $\mathbf{a} = \mathbf{b}_{0K}$. If $\mathbf{b}_{0K} = \mathbf{0}$ then we set $\mathbf{a} = \mathbf{e}\,\|\mathbf{b}\|_{0,\infty,K}$, where $\mathbf{e} \in \mathbb{R}^d$ is any unit vector. Then, in both cases, $|\mathbf{a}| \leqslant \|\mathbf{b}\|_{0,\infty,K}$ and $\|\mathbf{b}-\mathbf{a}\|_{0,\infty,K} \leqslant 2\,\|\mathbf{b}-\mathbf{b}_{0K}\|_{0,\infty,K}$. Now, as in the case of (4.5), the estimate (4.8) follows using the estimate $\|\mathbf{b}-\mathbf{b}_{0K}\|_{0,\infty,K} \leqslant C\,h_K\,|\mathbf{b}|_{1,\infty,K}$ with a constant C depending only on d and the shape regularity of K (see, e.g., Ciarlet, 1991), applying the inverse inequality (3.8) and taking into consideration the estimate (3.5) for δ_K. □

REMARK 4.3 We have seen in Section 2 that, for typical finite-element spaces W_h and B_h used in both the one-level and the two-level approaches of the LP method, the spaces $\widehat{W}$ and $\widehat{D}$ satisfy the inf–sup condition (2.12). Thus, if we set $\widehat{B} = \widehat{W} \cap H_0^1(\widehat{K})$, then we have $\widehat{D} \cap \widehat{B}^\perp = \{0\}$, and hence Lemmas 4.1 and 4.2 hold.

REMARK 4.4 If the parameters δ_K and τ_K satisfy

$$\tau_K \, \gamma_K(\mathbf{b}) \approx \delta_K \, \|\mathbf{b}\|_{0,\infty,K}^2,$$

which is allowed by the assumption (3.6), then the inequalities (4.4), (4.5), (4.7) and (4.8) can be replaced by

$$|||v_h||| \geqslant C \, |||v_h|||_{\mathrm{SUPG}} \quad \forall\, v_h \in W_h,$$

where the SUPG norm $|||\cdot|||_{\mathrm{SUPG}}$ is defined by (2.3) with $\delta|_K = \delta_K$ for all $K \in \mathcal{T}_h$. Thus, under the assumptions of this section, the inf–sup condition (3.9) also holds with respect to the SUPG norm.

We conclude this section with a brief discussion of convergence results for the LP discretization. Find $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$a_h(u_h, v_h) = (f, v_h) \quad \forall\, v_h \in V_h,$$

where a_h is defined by (3.1) with s_K given by (4.1) or (4.2). The spaces D_h and W_h are defined as at the beginning of this section and we assume that the inf–sup conditions (2.13) hold. For simplicity, we further assume that $\mathbf{b} \neq \mathbf{0}$ in $\overline{\Omega}$. Error estimates for u_h with respect to the standard LP norm $|||\cdot|||_{\mathrm{LP}}$ defined in (3.3) can be found in, for example, Matthies *et al.* (2007, 2008) and Knobloch (2009). A careful analysis suggests that we set (see Knobloch, 2009)

$$\tau_K \sim \min\left\{ \frac{h_K}{\|\mathbf{b}\|_{0,\infty,K}}, \frac{h_K^2}{\varepsilon} \right\} \frac{\|\mathbf{b}\|_{0,\infty,K}^2}{\gamma_K(\mathbf{b})}.$$

Let us assume that the spaces D_h and W_h have the usual approximation properties of order $r \in \mathbb{N}$ with respect to the norms $\|\cdot\|_{0,\Omega}$ and $\|\cdot\|_{1,\Omega}$, respectively, that the approximation $\widetilde{u}_{bh}$ of the Dirichlet boundary condition is sufficiently accurate and that the solution u of (1.1) belongs to $H^{r+1}(\Omega)$. Moreover, if s_K is given by (4.1) then we assume that $\mathbf{b} \in W^{r,\infty}(\Omega)$. Then

$$|||u - u_h|||_{\mathrm{LP}} \leqslant C \, (\varepsilon^{1/2} + h^{1/2}) \, h^r \, \|u\|_{r+1,\Omega}, \tag{4.9}$$

where, for simplicity, only the dependence on ε and h is shown explicitly. Now let us set

$$\delta_K \sim \min\left\{ \frac{h_K}{\|\mathbf{b}\|_{0,\infty,K}}, \frac{h_K^2}{\varepsilon} \right\},$$

as in (2.4) and (2.5), and consider the norm $|||\cdot|||$ from (3.4). We assume that the space B_h that is hidden in the definition of this norm is determined by the reference space $\widehat{B} = \widehat{W} \cap H_0^1(\widehat{K})$. Using the inf–sup condition (3.9), the stronger *a priori* estimate

$$|||u - u_h||| \leqslant C \, (\varepsilon^{1/2} + h^{1/2}) \, h^r \, \|u\|_{r+1,\Omega}$$

follows analogously to (4.9). Finally, in view of Lemmas 4.1 and 4.2 and Remarks 4.3 and 4.4, we also obtain the convergence of the LP method in the SUPG norm that is stronger than the LP1 norm:

$$|||u - u_h|||_{\mathrm{SUPG}} \leqslant C\,(\varepsilon^{1/2} + h^{1/2})\,h^r\,\|u\|_{r+1,\Omega}.$$

If $\mathbf{b}$ is not piecewise polynomial, or not constant in the case of s_K defined by (4.2), then this estimate holds for sufficiently small h.

5. Improved stability properties of the standard Galerkin method

It follows from Section 3 by setting $\tau_K = 0$ that the standard Galerkin method is stable with respect to the norm

$$|||v|||_0 = \left(|||v|||_G^2 + \sum_{K\in\mathcal{T}_h} \delta_K\,\|\Pi_K(\mathbf{b}\cdot\nabla v)\|_{0,K}^2 \right)^{1/2}, \tag{5.1}$$

which is stronger than the usual norm $|||\cdot|||_G$ provided that $B_h \neq \{0\}$. Let us consider a standard Galerkin discretization of (1.1) based on the space W_h of continuous piecewise polynomials of degree r on a simplicial triangulation $\mathcal{T}_h$ of Ω. If $r \geqslant d+1$ then this space contains a nontrivial bubble subspace B_h generated elementwise by $\widehat{B} = \hat{b}\cdot P_{r-d-1}(\widehat{K})$, where $\hat{b}$ is the product of the barycentric coordinates on $\widehat{K}$. We shall show in Lemma 5.1 that W_h contains a stable subspace defined by

$$S_h := \{v \in H^1(\Omega);\ v|_K \in P_{r-d}(K) \quad \forall\, K \in \mathcal{T}_h\}.$$

If the space W_h is constructed on a triangulation obtained by refining the simplicial triangulation $\mathcal{T}_h$ as in the two-level approach of the LP method (see Section 2), then a nontrivial bubble subspace B_h of W_h exists for any $r \geqslant 1$. In this case we set $\widehat{B} = \hat{b}\cdot P_{r-1}(\widehat{K})$, where $\hat{b} \in H_0^1(\widehat{K})$ is a nonvanishing function that is piecewise linear with respect to the refinement of $\widehat{K}$. Then a stable subspace of W_h is given by

$$S_h := \{v \in H^1(\Omega);\ v|_K \in P_r(K) \quad \forall\, K \in \mathcal{T}_h\}.$$

LEMMA 5.1 Let $\mathbf{b}$ be constant or h be sufficiently small. Let the spaces W_h and B_h be constructed in one of the ways mentioned above and let S_h be the corresponding subspace of W_h. Then, on the space S_h, the norm $|||\cdot|||_0$ given by (5.1) is equivalent to the SUPG norm $|||\cdot|||_{\mathrm{SUPG}}$ defined by (2.3) with $\delta|_K = \delta_K$ for any $K \in \mathcal{T}_h$.

Proof. Obviously, $|||v|||_0 \leqslant |||v|||_{\mathrm{SUPG}}$ for any $v \in H^1(\Omega)$. To prove that $|||v_h|||_0 \geqslant C\,|||v_h|||_{\mathrm{SUPG}}$ for $v_h \in S_h$, let us set $\widehat{D} = P_{r-d-1}(\widehat{K})$ for the one-level space W_h and $\widehat{D} = P_{r-1}(\widehat{K})$ for the two-level space W_h. Then $\widehat{D} \cap \widehat{B}^{\perp} = \{0\}$ and $\kappa_K \nabla v_h = \mathbf{0}$ for any $v_h \in S_h$. Consequently, the lemma follows immediately from Lemma 4.2 with $\tau_K = \delta_K \|\mathbf{b}\|_{0,\infty,K}^2$. □

There is an interesting consequence of Lemma 5.1 that we discuss now, for simplicity, for $d = 1$ and a constant function b. Using continuous piecewise polynomials of degree $r \geqslant 2$ and the standard Galerkin method, the subspace consisting of piecewise polynomials of degree $r-1$ is already controlled by the SUPG norm. Only the SUPG norm of the highest-degree polynomials have to be still controlled. If we add a stabilization term of the form

$$\sum_{K\in\mathcal{T}_h} \delta_K\,(\kappa_h(bu_h'), \kappa_h(bv_h'))_K,$$

where $\kappa_h = \mathrm{id} - \pi_h$ and π_h is the orthogonal L^2 projection onto the space of discontinuous piecewise polynomials of degree $r-2$, then just the highest-degree polynomials will be controlled by this LP term. In this sense the LP method represents a minimal stabilization. An alternative choice would be the term

$$\sum_{K\in\mathcal{T}_h} \delta_K h_K^{2r-2}\big((bu_h')^{(r-1)}, (bv_h')^{(r-1)}\big)_K,$$

which is closely related to the method studied in Tobiska (2006).

As an example, let us consider the problem (1.1) in one space dimension with $\Omega = (0,1)$ and $b = c = 1$. We introduce a uniform decomposition of $\Omega = (0,1)$ with nodes $x_i = i\,h$, where $i = 0,\dots,N$ and $h = 1/N$, and apply the Galerkin method with the space W_h consisting of continuous piecewise linear functions with respect to this decomposition. Furthermore, we introduce a second decomposition of $\Omega = (0,1)$ with the nodes $\widetilde{x}_j$, for $j = 0,\dots,J$, where $\widetilde{x}_0 = 0$ and $\widetilde{x}_J = 1$. To define the interior nodes we fix a positive integer $k \ll N$, choose $i_0 \in \{1,\dots,k\}$ and set $\widetilde{x}_j = x_{i_0+(j-1)\,k}$, where $j = 1,\dots,J-1$. We assume that $N - k \leqslant i_0 + (J-2)\,k < N$. For $k = 1$ the second decomposition is identical to the original one, but for $k \geqslant 2$ it is a coarser decomposition. Let us denote by v_h a globally oscillating function from W_h that is linear on each interval $(\widetilde{x}_{j-1}, \widetilde{x}_j)$, where $j = 1,\dots,J$, and satisfies $v_h(x_0) = v_h(x_N) = 0$ and $v_h(\widetilde{x}_j) = (-1)^j$ for $j = 0,\dots,J$. The stability of the standard Galerkin approach guarantees the boundedness of the discrete solution with respect to the norm $|||\cdot|||_G$. However,

$$|||v_h|||_G \leqslant \sqrt{\frac{6\,\varepsilon}{h^2} + \frac{1}{3}} \sim \varepsilon^{1/2} h^{-1} + 1,$$

and hence, for ε small enough, oscillating functions such as v_h are not excluded by the boundedness in the norm $|||\cdot|||_G$. On the other hand, considering $k \geqslant 2$ and denoting by $\mathcal{T}_h$ the coarse decomposition of Ω, the space W_h contains a nontrivial bubble space B_h with respect to $\mathcal{T}_h$. Hence the Galerkin solution is also bounded in the stronger norm $|||\cdot|||_0$ defined by (5.1). It is natural to define B_h as the subspace of W_h consisting of functions vanishing at the nodes of $\mathcal{T}_h$. Then, for any interval $K \in \mathcal{T}_h$ away from the boundary of Ω, we have $\|\Pi_K\,1\|_{0,K}^2 = \alpha\,h$ with a constant α that is independent of K and h. Consequently, using a constant value δ of δ_K, we get

$$\sum_{K\in\mathcal{T}_h} \delta\,\|\Pi_K v_h'\|_{0,K}^2 \geqslant \alpha\,\delta\,\|v_h'\|_{0,(\widetilde{x}_1,\widetilde{x}_{J-1})}^2 = \frac{4\,\alpha\,\delta}{k^2\,h^2}\,(\widetilde{x}_{J-1} - \widetilde{x}_1) \sim h^{-1} \quad \text{for } \delta \sim h.$$

Thus our improved stability results show that all of the oscillating functions defined for $k \geqslant 2$ cannot appear among the Galerkin solutions for small h. Only the most rapidly oscillating function (obtained for $k = 1$) is not excluded by the boundedness in the norm $|||\cdot|||_0$, as indicated in Fig. 2(a). Indeed, in this case v_h' is L^2 orthogonal to all functions from W_h vanishing in $(0,h) \cup (1-h,1)$, and hence the additional term in (5.1) provides only very little information about the function v_h.

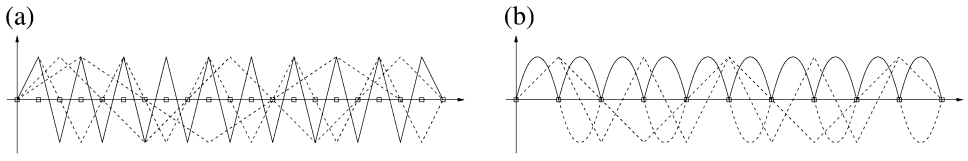


FIG. 2. Globally oscillating modes that are suppressed (dashed lines) and not suppressed (solid line) by the boundedness in the norm $|||\cdot|||_0$: (a) piecewise linear elements and (b) piecewise quadratic elements.

The above considerations correspond to the two-level approach of the LP method. If the space W_h contains higher-order polynomials then we can also proceed as in the one-level approach. To fix the ideas let us denote by $\mathcal{T}_h$ the decomposition of Ω with the nodes $x_0, x_1, \ldots, x_N$ and let us assume that the space W_h consists of continuous piecewise quadratic functions with respect to $\mathcal{T}_h$. For any $K \in \mathcal{T}_h$ the operator Π_K is the orthogonal L^2 projection of $L^2(K)$ onto the space span$\{\varphi_K\}$, where φ_K is the quadratic function vanishing at the end points of the interval K and that equals 1 at the midpoint of K. The solution of the standard Galerkin approach with the space W_h is then bounded in the norm $|||\cdot|||_0$. For the above-introduced piecewise linear function v_h satisfying $v_h(x_i) = (-1)^i$, where $i = 1, \ldots, N-1$, the additional term in the norm $|||\cdot|||_0$ can be easily evaluated and we get

$$\sum_{K\in\mathcal{T}_h} \delta \|\Pi_K v_h'\|_{0,K}^2 = \frac{5}{6}\,\delta\,\|v_h'\|_{0,\Omega}^2 = \frac{5\,\delta}{3\,h}\left(\frac{2}{h} - 3\right) \sim h^{-1} \quad \text{for } \delta \sim h.$$

Since $\Pi_K \varphi_K' = 0$ for any $K \in \mathcal{T}_h$, the improved stability property of the higher-order approximation excludes discrete solutions whose piecewise linear part oscillates like the function v_h, as indicated in Fig. 2(b).

The boundedness with respect to the norm $|||\cdot|||_0$ defined using the one-level projections does not exclude globally oscillating higher-order modes. Indeed, defining a function $w_h \in W_h$ in such a way that $w_h|_K = (-1)^i\,\varphi_K$ if $K = (x_{i-1}, x_i)$, where $i = 1, \ldots, N$, we again have $|||w_h|||_G \sim \varepsilon^{1/2} h^{-1} + 1$ but $\sum_{K\in\mathcal{T}_h} \delta \|\Pi_K w_h'\|_{0,K}^2 = 0$. However, a different conclusion is obtained if we consider the stability originating from the two-level approach. In this case we denote by $\mathcal{T}_h$ the decomposition of Ω with the nodes $\widetilde{x}_0, \widetilde{x}_1, \ldots, \widetilde{x}_J$ defined for $i_0 = k = 2$ that can be viewed as a product of coarsening the decomposition on which the space W_h is defined. Then, on each $K \in \mathcal{T}_h$ (except for $K = (\widetilde{x}_{J-1}, \widetilde{x}_J) = (1-h, 1)$ if N is odd), we have a one-dimensional piecewise linear bubble space $B(K)$. The corresponding operator Π_K gives $\|\Pi_K w_h'\|_{0,K}^2 = 1/(6\,h)$, and hence $\sum_{K\in\mathcal{T}_h} \delta \|\Pi_K w_h'\|_{0,K}^2 \sim h^{-1}$ for $\delta \sim h$. This excludes the function w_h if h is small (cf., Fig. 2(b), dashed piecewise quadratic curve).

Nevertheless, some globally oscillating higher-order modes are still allowed, as shown by the piecewise quadratic curve in Fig. 2(b). If we set $w_h|_K = \varphi_K$ for any $K = (x_{i-1}, x_i)$, where $i = 1, \ldots, N$, then we deduce that w_h' is L^2 orthogonal to the space $W_h \cap H_0^1(\Omega)$, and hence $|||w_h|||_0 = |||w_h|||_G$ for any choice of the projection operators Π_K. In order to also exclude this type of mode we can apply a one-level LP method with operators π_K projecting onto constant functions. Then $\kappa_K \varphi_K' = \varphi_K'$ for any $K \in \mathcal{T}_h$, and hence the improved stability of the LP method excludes globally oscillating higher-order modes such as the function w_h.

The boundedness of a finite-element solution u_h with respect to a norm equivalent to the SUPG norm suppresses spurious oscillations, which was also pointed out by F. Schieweck (2007, private communication). In the standard Galerkin method it turns out that this boundedness can already be guaranteed for certain subspaces of a finite-element space, and thus only some high-frequency modes have to be stabilized.

6. Numerical results

In this section we present numerical results for the following setting of the problem (1.1).

EXAMPLE 6.1 We consider the problem (1.1) in $\Omega = (0, 1)^2$ with

$$\varepsilon = 10^{-8}, \quad \mathbf{b} = (1, 0), \quad c = 1, \quad f = 1, \quad u_b = 0.$$

The solution of Example 6.1 possesses an exponential boundary layer at $x = 1$ and parabolic boundary layers at $y = 0$ and $y = 1$. Outside the layers the solution is very close to the function $u_0(x, y) = 1 - \mathrm{e}^{-x}$.

Fig. 3 shows the SUPG solution of Example 6.1 computed using the Q_2 element, a triangulation $\mathcal{T}_h$ consisting of 20×20 equal squares and a stabilization parameter defined by

$$\delta|_K = \frac{1}{4} \min\left\{ \frac{h_K}{\|\mathbf{b}\|_{0,\infty,K}}, \frac{h_K^2}{6\,\varepsilon} \right\} \quad \forall\, K \in \mathcal{T}_h$$

(see Codina *et al.*, 2002). The lines in Fig. 3 connect the values of the SUPG solution at vertices, midpoints of edges and centres of elements of $\mathcal{T}_h$. We observe that the SUPG solution contains spurious oscillations along the parabolic layers. At the exponential layer no oscillations are present, which is caused by the fact that, outside the parabolic layers, the discrete problem reduces to the one-dimensional case. For other convection fields $\mathbf{b}$ and/or other types of triangulations, spurious oscillations also have to be generally expected at exponential layers, unless a special tuning of the stabilization parameter is performed.

Now let us consider LP stabilizations. We shall present results for a one-level method and a two-level method. The one-level method is defined using the bilinear form a_h^{LP2} from (2.7), the same triangulation $\mathcal{T}_h$ as above and a space W_h constructed using the Q_2 element enriched by three bubble functions on each element K of $\mathcal{T}_h$. Choosing a nonvanishing function $b_K \in Q_2(K) \cap H_0^1(K)$, these three bubble functions are $b_K\, x$, $b_K\, y$ and $b_K\, x\, y$. The two-level method uses the bilinear form a_h^{LP1} from (2.6) and a triangulation $\mathcal{T}_h$ consisting of 10×10 equal squares. The finite-element space W_h is defined using the Q_2 element on a triangulation obtained by refining $\mathcal{T}_h$ as explained in Section 2. Thus the space W_h is the same as for the SUPG method. For both methods the projection space D_h is defined by (2.10) with $D(K) = Q_1(K)$ and the stabilization parameter by

$$\tau|_K = \frac{1}{15} \min\left\{ \frac{h_K}{\|\mathbf{b}\|_{0,\infty,K}}, \frac{h_K^2}{6\,\varepsilon} \right\} \frac{\|\mathbf{b}\|_{0,\infty,K}^2}{\gamma_K(\mathbf{b})} \quad \forall\, K \in \mathcal{T}_h.$$

The discrete solutions obtained are depicted in Fig. 4. The one-level solution is visualized without the additional bubbles, so that the corresponding function belongs to the space W_h of the remaining

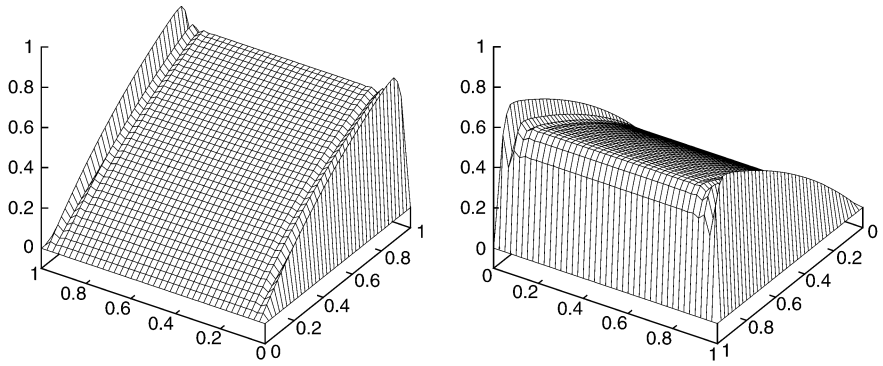


FIG. 3. Two views of the SUPG solution of Example 6.1.

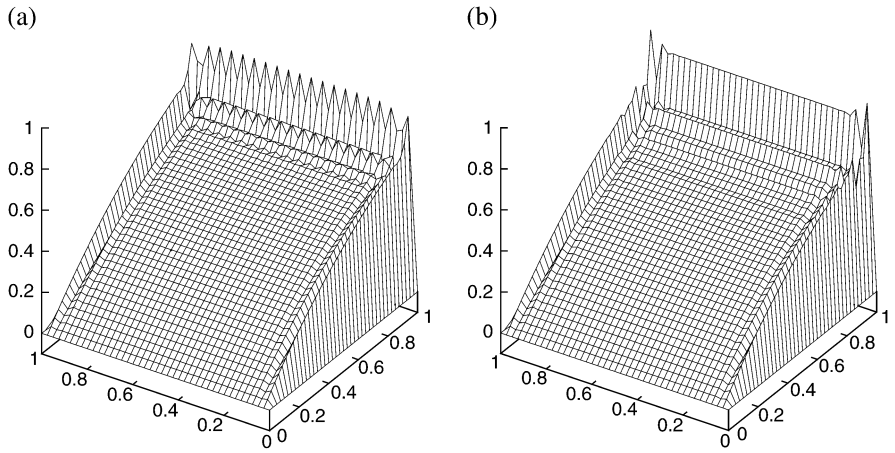


FIG. 4. LP solutions of Example 6.1: (a) the one-level approach with a_h^{LP2} and (b) the two-level approach with a_h^{LP1}.

two methods. At the parabolic boundary layers both solutions contain spurious oscillations that are almost indistinguishable from those of the SUPG solution. The LP methods also contain oscillations along the exponential boundary layers. However, it is important that these oscillations are localized. The localization is more pronounced for the one-level method, but note that this method employed more degrees of freedom than the two-level method in the presented computations. Away from boundary layers, for example, in the domain $(0, 2/3) \times (1/3, 2/3)$, all three discrete solutions are very close to the function u_0 introduced below Example 6.1.

The numerical results of this section show that neither the SUPG method nor the LP method removes spurious oscillations completely. This is due to the fact that neither method results in a linear system with an inverse monotone matrix. The maximum amplitude of oscillations in the layer region is larger for the LP method than for the SUPG method. This demonstrates that the stability of different discretizations with respect to the same norm does not necessarily mean that the corresponding numerical solutions will be of similar accuracy. Nevertheless, the LP method preserves the important property of the SUPG method that oscillations are localized to the layer regions.

Funding

Bilateral Czech–German project financed by the Grant Agency of the Czech Republic (201/07/J033); German Research Foundation (To 143/10); Ministry of Education, Youth and Sports of the Czech Republic (MSM 0021620839) to P.K.

REFERENCES

ALMEIDA, R. C. & SILVA, R. S. (1997) A stable Petrov–Galerkin method for convection-dominated problems. *Comput. Methods Appl. Mech. Eng.*, **140**, 291–304.

BADIA, S. & CODINA, R. (2006) Analysis of a stabilized finite element approximation of the transient convection–diffusion equation using an ALE framework. *SIAM J. Numer. Anal.*, **44**, 2159–2197.

BECKER, R. & BRAACK, M. (2001) A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo*, **38**, 173–199.

BECKER, R. & BRAACK, M. (2004) A two-level stabilization scheme for the Navier–Stokes equations. *Numerical Mathematics and Advanced Applications* (M. Feistauer, V. Dolej, P. Knobloch & K. Najzar eds). Berlin: Springer, pp. 123–130.

BRAACK, M. & BURMAN, E. (2006) Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.*, **43**, 2544–2566.

BROOKS, A. N. & HUGHES, T. J. R. (1982) Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Eng.*, **32**, 199–259.

BURMAN, E. & HANSBO, P. (2004) Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems. *Comput. Methods Appl. Mech. Eng.*, **193**, 1437–1453.

CHRISTIE, I., GRIFFITHS, D. F., MITCHELL, A. R. & ZIENKIEWICZ, O. C. (1976) Finite element methods for second order differential equations with significant first derivatives. *Int. J. Numer. Methods Eng.*, **10**, 1389–1396.

CIARLET, P. G. (1991) Basic error estimates for elliptic problems. *Handbook of Numerical Analysis, Volume II: Finite Element Methods* (*part* 1) (P. G. CIARLET & J. L. LIONS eds). Amsterdam: North-Holland, pp. 17–351.

CODINA, R. & BLASCO, J. (2002) Analysis of a stabilized finite element approximation of the transient convection–diffusion–reaction equation using orthogonal subscales. *Comput. Vis. Sci.*, **4**, 167–174.

CODINA, R., OÑATE, E. & CERVERA, M. (2002) The intrinsic time for the streamline upwind/Petrov–Galerkin formulation using quadratic elements. *Comput. Methods Appl. Mech. Eng.*, **94**, 239–262.

GALEÃO, A. C., ALMEIDA, R. C., MALTA, S. M. C. & LOULA, A. F. D. (2004) Finite element analysis of convection dominated reaction–diffusion problems. *Appl. Numer. Math.*, **48**, 205–222.

GANESAN, S. & TOBISKA, L. (2008) Stabilization by local projection for convection–diffusion and incompressible flow problems. *J. Sci. Comput.* DOI: 10.1007/s10915-008-9259-8.

JOHN, V. & KNOBLOCH, P. (2007) On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part I—a review. *Comput. Methods Appl. Mech. Eng.*, **196**, 2197–2215.

JOHN, V. & KNOBLOCH, P. (2009) On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: part II—analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Eng.*, **197**, 1997–2014.

KNOBLOCH, P. (2009) On the application of local projection methods to convection-diffusion-reaction problems. *Boundary and Interior Layers* (A. F. Hegarty, N. Kopteva, E. O' Riordan & M. Stynes eds). Lect. Notes Comput. Sci. Eng., Vol. 69. Berlin: Springer, pp. 183–194.

KNOBLOCH, P. & LUBE, G. (2009) Local projection stabilization for advection-diffusion-reaction problems: one level vs. two-level approach. *Appl. Numer. Math.* DOI: 10.1016/j.apnum.2009.06.004.

MATTHIES, G., SKRZYPACZ, P. & TOBISKA, L. (2007) A unified convergence analysis for local projection stabilisations applied to the Oseen problem. *Math. Model. Numer. Anal. M2AN*, **41**, 713–742.

MATTHIES, G., SKRZYPACZ, P. & TOBISKA, L. (2008) Stabilization of local projection type applied to convection–diffusion problems with mixed boundary conditions. *Electron. Trans. Numer. Anal.*, **32**, 90–105.

RAPIN, G., LUBE, G. & LÖWE, J. (2008) Applying local projection stabilization to inf–sup stable elements. *Numerical Mathematics and Advanced Applications* (K. Kunisch, G. Of & O. Steinbach eds). Berlin: Springer, pp. 521–528.

ROOS, H.-G., STYNES, M. & TOBISKA, L. (2008) *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection–Diffusion–Reaction and Flow Problems*, 2nd edn. Berlin: Springer.

TOBISKA, L. (2006) Analysis of a new stabilized higher order finite element method for advection–diffusion equations. *Comput. Methods Appl. Mech. Eng.*, **196**, 538–550.

TOBISKA, L. (2009) On the relationship of local projection stabilization to other stabilized methods for one-dimensional advection–diffusion equations. *Comput. Methods Appl. Mech. Eng.*, **198**, 831–837.

Applied Numerical Mathematics 59 (2009) 2891–2907

Contents lists available at ScienceDirect

Applied Numerical Mathematics

www.elsevier.com/locate/apnum

Local projection stabilization for advection–diffusion–reaction problems: One-level vs. two-level approach

Petr Knobloch [a], Gert Lube [b,*]

[a] Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 186 75 Praha 8, Czech Republic
[b] Georg-August-Universität Göttingen, Mathematische Fakultät, NAM, Lotzestrasse 16-18, D-37083 Göttingen, Germany

ARTICLE INFO

Article history:
Received 8 July 2008
Received in revised form 23 April 2009
Accepted 23 June 2009
Available online 30 June 2009

Keywords:
Local projection stabilization
Finite element method
Advection–diffusion–reaction problem
Advection-dominated problem
SUPG method

ABSTRACT

Local projection stabilization (LPS) of finite element methods is a new technique for the numerical solution of transport-dominated problems. The main aim of this paper is a critical discussion and comparison of the one- and two-level approaches to LPS for the linear advection–diffusion–reaction problem. Moreover, the paper contains several other novel contributions to the theory of LPS. In particular, we derive an error estimate showing not only the usual error dependence on the mesh width but also on the polynomial degree of the finite element space. Based on this error estimate, we propose a definition of the stabilization parameter depending on the data of the solved problem. Unlike other papers on LPS methods, we observe that the consistency error may deteriorate the convergence order. Finally, we explain the relation between the LPS method and residual-based stabilization techniques for simplicial finite elements.

1. Introduction

Consider the stationary advection–diffusion–reaction problem

$$Lu := -\varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u + \sigma u = f \quad \text{in } \Omega; \qquad u = 0 \quad \text{on } \partial\Omega \tag{1}$$

for the scalar field u in a bounded domain $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, with given source term f, advection field $\boldsymbol{b}$ and constant data $\varepsilon > 0, \sigma \geqslant 0$. Problem (1) is a basic model in fluid mechanics and many other applications.

The Galerkin finite element (FE) approximation of (1) may suffer from dominating advection, i.e., $\varepsilon \ll \|\boldsymbol{b}\|_{[L^\infty(\Omega)]^d} h$, and/or dominating reaction, i.e., $\varepsilon \ll \sigma h^2$, where h denotes the mesh width. The traditional way to cope with this problem is the application of residual-based stabilization (RBS) techniques. The basic approach is the streamline-upwind/Petrov–Galerkin (SUPG) method [8] or related variants. An overview about RBS methods and further stabilization techniques for problem (1) can be found in [24].

The class of RBS techniques is still quite popular since they are robust and easy to implement. Nevertheless, they have severe drawbacks stemming from the non-symmetric form of the stabilization terms and the occurrence of second-order derivatives in the residual $Lu - f$. Therefore, other stabilization techniques appeared recently, in particular, the edge-stabilization method [9,7] and variational multiscale (VMS) methods [17,18,15,10]. We emphasize that almost all stabilization methods can be interpreted as special VMS methods. The key idea of VMS methods is a separation of scales: large scales, small scales and unresolved scales. The influence of the unresolved scales on the other scales has to be modelled. Mostly, it is assumed that the unresolved scales do not influence the large scales.

* Corresponding author.
E-mail addresses: knobloch@karlin.mff.cuni.cz (P. Knobloch), lube@math.uni-goettingen.de (G. Lube).

0168-9274/$30.00 © 2009 IMACS. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.apnum.2009.06.004

Local projection stabilization (LPS) methods as special VMS-type methods are of current interest [6,21]. Here the influence of the unresolved scales on the small scales is modelled by additional artificial diffusion terms for the small scales. LPS methods belong to the class of symmetric stabilization techniques [7]. One major advantage of such methods applied to optimization problems with partial differential equations is that the operations 'discretization' and 'optimization' commute [4].

Let us mention the main novel contributions of this paper. There are currently two basic variants of LPS methods: a two-level approach [6,21,23] and a one-level approach [21,26,22,12]. One goal of this paper is a detailed computational comparison of both variants and of the SUPG method. Another goal is a critical review of the numerical analysis (based on energy estimates). In particular, we consider the error estimates in terms of both the mesh width and the polynomial degree of the finite element space. Balancing terms in the error estimate, we come to a formula for the stabilization parameter which scales correctly with respect to $\boldsymbol{b}$. For neighborhoods of subregions with a vanishing advection field $\boldsymbol{b}$ we show that a deterioration of the convergence order can occur. Finally, we show that the LPS approach is very close to RBS methods like the algebraic subgrid scale stabilization [16,10] or the 'unusual' Galerkin/least-squares method [11]. The latter result is new for higher-order finite elements and is established simultaneously for both variants of the LPS approach.

The outline of the paper is as follows. The basic Galerkin FEM and its stabilization via local projection is discussed in Section 2. In Section 3 of this paper, we present a unified theory of local projection methods for problem (1) based on energy estimates. In contrast to other papers, the dependence on the polynomial degree of the finite element method is considered. In Section 4, examples of finite element spaces satisfying the assumptions of Section 3 are presented and, in Section 5, a comparison of both variants of LPS methods is performed by means of simple numerical experiments. Section 6 is devoted to the relationship between simplicial LPS methods and residual-based stabilization methods.

Throughout this paper, standard notations for Lebesgue and Sobolev spaces are used. The L^2 inner product in a domain G is denoted by $(\cdot,\cdot)_G$. Moreover, we use the notation $a \lesssim b$ if there is a constant $C > 0$ independent of all relevant parameters like mesh size, polynomial degree or coefficients of L.

2. Variational formulation and stabilization

Here, the basic Galerkin finite element formulation of problem (1) and its stabilized variants via local projection (LPS) are introduced. Moreover, various technical tools are given.

2.1. Basic Galerkin approximation

The variational formulation for the advection–diffusion–reaction problem (1) reads: Find $u \in V := H_0^1(\Omega)$ such that

$$a(u, v) := (\varepsilon \nabla u, \nabla v)_\Omega + (\boldsymbol{b} \cdot \nabla u + \sigma u, v)_\Omega = (f, v)_\Omega, \quad \forall v \in V. \tag{2}$$

Assumption 1. Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, be a bounded, polyhedral domain. Moreover, assume that $\varepsilon > 0$ is constant, $f \in L^2(\Omega)$, $\boldsymbol{b} \in [L^\infty(\Omega) \cap H^1(\Omega)]^d$ with $\nabla \cdot \boldsymbol{b} = 0$ a.e. in Ω and $\sigma \geqslant 0$ is constant.

Remark 1. Typically, $\boldsymbol{b}$ is a finite element solution of an incompressible flow problem. Then it holds that $(\nabla \cdot \boldsymbol{b}, q_h)_\Omega = 0$ for certain test functions q_h. Hence, $\nabla \cdot \boldsymbol{b}$ is small but does not vanish in general. A simple remedy to ensure coercivity of $a(\cdot,\cdot)$ is to replace the advective term $(\boldsymbol{b} \cdot \nabla u, v)_\Omega$ by $\frac{1}{2}(\boldsymbol{b} \cdot \nabla u, v)_\Omega - \frac{1}{2}(\boldsymbol{b} \cdot \nabla v, u)_\Omega - \frac{1}{2}((\nabla \cdot \boldsymbol{b})u, v)_\Omega$.

Consider a decomposition $\mathcal{T}_h$ of Ω belonging to a family of shape-regular, admissible decompositions of Ω into d-dimensional simplices, quadrilaterals in the two-dimensional case or hexahedra for three dimensions. Let h_T be the diameter of a cell $T \in \mathcal{T}_h$ and h the maximum of all h_T, $T \in \mathcal{T}_h$. Let $\hat{T}$ be a reference element of the decomposition $\mathcal{T}_h$. Let us assume that, for each $T \in \mathcal{T}_h$, there is an affine mapping $F_T : \hat{T} \to T$ which maps $\hat{T}$ onto T. This quite restrictive assumption for quadrilaterals can be weakened to asymptotically affine mappings [1].

Set

$$P_{k,\mathcal{T}_h} := \left\{ v_h \in L^2(\Omega);\ v_h \circ F_T \in P_k(\hat{T}), T \in \mathcal{T}_h \right\}$$

with the space $P_k(\hat{T})$ of complete polynomials of degree k defined on $\hat{T}$ and

$$Q_{k,\mathcal{T}_h} := \left\{ v_h \in L^2(\Omega);\ v_h \circ F_T \in Q_k(\hat{T}), T \in \mathcal{T}_h \right\}$$

with the space $Q_k(\hat{T})$ of all polynomials on $\hat{T}$ with maximal degree k in each coordinate direction. We shall approximate the space V by a finite element space $V_{h,k} \subset V$ such that

$$V_{h,k} \supset P_{k,\mathcal{T}_h} \cap V \quad \text{or} \quad V_{h,k} \supset Q_{k,\mathcal{T}_h} \cap V.$$

Now, the standard Galerkin discretization of problem (1) reads: Find $u_h \in V_{h,k}$ such that

$$a(u_h, v_h) = (f, v_h)_\Omega, \quad \forall v_h \in V_{h,k}. \tag{3}$$

As mentioned in the introduction, the solution u_h of (3) usually suffers from spurious oscillations, which is often cured by introducing a stabilization in (3).

2.2. Local projection stabilization (LPS)

The idea of LPS methods is to split the discrete function spaces into small and large scales and to add stabilization terms of diffusion type acting only on the small scales. Such stabilization terms can be interpreted as models for the influence of the unresolved scales on the small scales. Therefore, LPS methods can be regarded as special variational multiscale (VMS) methods. There are two obvious choices of the space of large scales: a two-level and a one-level approach.

The first, the two-level variant, is to determine the large scales with the help of a coarse mesh. The coarse mesh $\mathcal{M}_h$ is constructed by coarsening the basic mesh $\mathcal{T}_h$ such that each macro-element $M \in \mathcal{M}_h$ is the union of one or more neighboring cells $T \in \mathcal{T}_h$. The diameter of $M \in \mathcal{M}_h$ is denoted by h_M. We assume that the decomposition $\mathcal{M}_h$ of Ω is non-overlapping and shape-regular. Additionally, the interior cells are supposed to be of the same size as the corresponding macro-cell:

$$\exists C > 0: \quad h_M \leqslant C h_T, \quad \forall T \in \mathcal{T}_h,\ M \in \mathcal{M}_h \text{ with } T \subset M, \tag{4}$$

where the constant C is the same for all $\mathcal{T}_h$ belonging to the considered family of decompositions of Ω. Following the approach in [21], we define a discrete space $D_h \subset L^2(\Omega)$ as a discontinuous finite element space defined on the macro-partition $\mathcal{M}_h$. The restriction of D_h on a macro-element $M \in \mathcal{M}_h$ is denoted by $D_h(M) := \{v_h|_M;\ v_h \in D_h\}$.

The next ingredient is a local projection $\pi_M : L^2(M) \to D_h(M)$ which defines the global projection $\pi_h : L^2(\Omega) \to D_h$ by $(\pi_h v)|_M := \pi_M(v|_M)$ for all $M \in \mathcal{M}_h$ and for all $v \in L^2(\Omega)$. A standard variant is the local orthogonal L^2 projection. Denoting the identity on $L^2(\Omega)$ by id, the associated fluctuation operator $\kappa_h : L^2(\Omega) \to L^2(\Omega)$ is defined by $\kappa_h := \mathrm{id} - \pi_h$.

The second approach, the one-level variant, consists in choosing a discontinuous lower order finite element space D_h on the original mesh $\mathcal{T}_h$. The same abstract framework as in the first approach can be used by setting $\mathcal{M}_h = \mathcal{T}_h$.

For both variants, the stabilized discrete formulation reads: find $u_h \in V_{h,k}$ such that

$$a(u_h, v_h) + s_h(u_h, v_h) = (f, v_h)_\Omega, \quad \forall v_h \in V_{h,k}, \tag{5}$$

where the additional stabilization term is given by

$$s_h(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tau_M \big(\kappa_h(\boldsymbol{b} \cdot \nabla u_h), \kappa_h(\boldsymbol{b} \cdot \nabla v_h)\big)_M. \tag{6}$$

Remark 2. The LPS scheme (5) with (6) will be denoted as streamline-derivative-based LPS scheme (SD-based LPS scheme for short below). Another variant is to replace $s_h(\cdot,\cdot)$ with

$$\tilde{s}_h(u_h, v_h) := \sum_{M \in \mathcal{M}_h} \tilde{\tau}_M (\kappa_h \nabla u_h, \kappa_h \nabla v_h)_M.$$

Later on, it will be called gradient-based LPS scheme. We will summarize the corresponding result in Remark 6.

The constants τ_M and $\tilde{\tau}_M$ will be determined later based on an a priori estimate. Please notice that the stabilizations s_h and $\tilde{s}_h$ act solely on the small scales. Of course, there is some more freedom in the choice of s_h, see also [21,6].

In order to control the consistency error of the κ_h-dependent stabilization terms, the space D_h has to be large enough; more precisely:

Assumption 2. The fluctuation operator κ_h satisfies for $0 \leqslant l \leqslant k$ the following approximation property:

$$\exists C_\kappa > 0: \quad \|\kappa_h q\|_{0,M} \leqslant C_\kappa \frac{h_M^l}{k^l} |q|_{l,M}, \quad \forall q \in L^2(\Omega),\ q|_M \in H^l(M),\ \forall M \in \mathcal{M}_h. \tag{7}$$

The subsequent numerical analysis takes advantage of the inverse inequality

$$\exists \mu_{\mathrm{inv}} > 0: \quad |v_h|_{1,T} \leqslant \mu_{\mathrm{inv}} k^2 h_T^{-1} \|v_h\|_{0,T}, \quad \forall T \in \mathcal{T}_h,\ \forall v_h \in V_{h,k} \tag{8}$$

(see [13]) and of the interpolation properties of the finite element space $V_{h,k}$. For the Scott–Zhang quasi-interpolant operator $I_{h,k}$ [27], one obtains for $v \in V$ with $v|_{\omega_T} \in H^r(\omega_T)$, $r \geqslant 1$, on the patches $\omega_T := \bigcup_{\bar{T}' \cap \bar{T} \neq \emptyset} T'$

$$\exists C > 0: \quad \|v - I_{h,k} v\|_{m,T} \leqslant C \frac{h_T^{l-m}}{k^{r-m}} \|v\|_{r,\omega_T}, \quad 0 \leqslant m \leqslant l = \min\{k+1, r\}. \tag{9}$$

The constant C may depend on r. Like in (4), the constants in the inequalities (7)–(9) are the same for all $\mathcal{T}_h$ belonging to the considered family of decompositions of Ω.

2.3. Special interpolation operator

Following [21], we construct a special interpolation $j_h : V \to V_{h,k}$ such that the error $v - j_h v$ is L^2-orthogonal to D_h for all $v \in V$. In order to conserve the standard approximation properties, we additionally assume

Assumption 3. There is a constant $\beta > 0$ such that, for any $M \in \mathcal{M}_h$,

$$\inf_{q_h \in D_h(M)} \sup_{v_h \in Y_h(M)} \frac{(v_h, q_h)_M}{\|v_h\|_{0,M}\|q_h\|_{0,M}} \geqslant \beta \tag{10}$$

where $Y_h(M) := \{v_h|_M;\ v_h \in V_{h,k}, v_h = 0 \text{ on } \Omega \setminus M\}$. The constant β is assumed to be the same for all $\mathcal{M}_h$ belonging to the considered family of macro-decompositions of Ω.

Remark 3. The inf-sup condition (10) implies that the space D_h must not be too rich. On the other hand, D_h must be rich enough to fulfill the approximation property (7). Later we will present several function spaces D_h satisfying (10).

Lemma 1. *Let Assumption* 3 *be satisfied. Then there is an interpolation operator $j_h : V \to V_{h,k}$ such that*

$$(v - j_h v, q_h)_\Omega = 0, \quad \forall q_h \in D_h,\ \forall v \in V, \tag{11}$$

$$\|v - j_h v\|_{0,M} + \frac{h_M}{k^2}|v - j_h v|_{1,M} \lesssim \left(1 + \frac{1}{\beta}\right)\frac{h_M^l}{k^l}\|v\|_{l,\omega_M}, \quad \forall M \in \mathcal{M}_h,\ v \in V \cap H^l(\Omega),\ 1 \leqslant l \leqslant k+1. \tag{12}$$

Proof. We follow the lines of the proof of Theorem 2.2 in [21], but we take into account the dependence of the constants on the polynomial order and the inf-sup constant β.

Consider any $M \in \mathcal{M}_h$ and define the linear continuous operator $B_h : Y_h(M) \to D_h(M)'$ by

$$\langle B_h v_h, q_h\rangle := (v_h, q_h)_M, \quad \forall v_h \in Y_h(M),\ q_h \in D_h(M).$$

Denote $W_h(M) := \mathrm{Ker}(B_h)$ and let $W_h(M)^\perp$ be the orthogonal complement of $W_h(M)$ in $Y_h(M)$ with respect to $(\cdot,\cdot)_M$. The Closed Range Theorem yields via Assumption 3 (cf. [14], p. 58, Lemma 4.1) that B_h is an isomorphism from $W_h(M)^\perp$ onto $D_h(M)'$ with $\beta\|v_h\|_{0,M} \leqslant \|B_h v_h\|_{D_h(M)'}$ for any $v_h \in W_h(M)^\perp$. Therefore, for any $v \in V$, there is a unique $z_h(v, M) \in W_h(M)^\perp$ with $\|z_h(v, M)\|_{0,M} \leqslant \frac{1}{\beta}\|v - I_{h,k}v\|_{0,M}$ such that

$$\big\langle B_h z_h(v, M), q_h\big\rangle = \big(z_h(v, M), q_h\big)_M = (v - I_{h,k}v, q_h)_M, \quad \forall q_h \in D_h(M).$$

Since $\mathcal{M}_h$ is a partition of Ω, we can define an operator $j_h : V \to V_{h,k}$ by $(j_h v)|_M := (I_{h,k}v)|_M + z_h(v, M)$, $M \in \mathcal{M}_h$. Then we immediately obtain the orthogonality property (11). Due to (9) the operator j_h satisfies for $1 \leqslant l \leqslant k+1$ and all $M \in \mathcal{M}_h$, $v \in V \cap H^l(\Omega)$

$$\|v - j_h v\|_{0,M}^2 \leqslant \left(1 + \frac{1}{\beta}\right)^2 \|v - I_{h,k}v\|_{0,M}^2 \leqslant C\left(1 + \frac{1}{\beta}\right)^2 \sum_{\substack{T \subset M \\ T \in \mathcal{T}_h}} \frac{h_T^{2l}}{k^{2l}}\|v\|_{l,\omega_T}^2.$$

To derive an approximation property in the H^1 seminorm, we first use the inverse inequality (8) and the assumption (4), which implies

$$\big|z_h(v, M)\big|_{1,M}^2 \leqslant \sum_{\substack{T \subset M \\ T \in \mathcal{T}_h}} \mu_{\mathrm{inv}}^2 k^4 h_T^{-2}\big\|z_h(v, M)\big\|_{0,T}^2 \lesssim \frac{\mu_{\mathrm{inv}}^2}{\beta^2} k^4 h_M^{-2}\|v - I_{h,k}v\|_{0,M}^2.$$

Then, applying the approximation property (9), we get

$$\begin{aligned} |v - j_h v|_{1,M} &= \big|v - I_{h,k}v - z_h(v, M)\big|_{1,M} \leqslant |v - I_{h,k}v|_{1,M} + \big|z_h(v, M)\big|_{1,M} \\ &\lesssim \left(\frac{1}{k} + \frac{\mu_{\mathrm{inv}}}{\beta}\right)\frac{h_M^{l-1}}{k^{l-2}}\|v\|_{l,\omega_M}. \qquad \square \end{aligned}$$

Remark 4.

(i) The estimate of Lemma 1 is optimal with respect to h_M. The estimate in the seminorm $|\cdot|_{1,M}$ is seemingly sub-optimal regarding k. A discussion of the stability constant β appearing in Lemma 1 is given in [23].

(ii) If $v \in V \cap H^t(\Omega)$ with $t > \frac{3}{2}$, it is possible to replace the Scott–Zhang quasi-interpolant operator $I_{h,k}$ in (9) by a pointwise interpolant, e.g., the Lagrangian interpolant. This allows to replace the sets ω_M in (12) and in the a priori estimates of the next section by the macro-elements M, see [22].

3. A priori analysis

The next goal is an error estimate for the scheme (5). Therefore, further assumptions on the finite element spaces $V_{h,k}$ and D_h are required. We will derive all results for the SD-based LPS scheme. The corresponding results for the gradient-based LPS scheme, see Remark 2, will be summarized in Remark 6.

3.1. Stability

First, the stability of the scheme will be proven in the mesh-dependent norm

$$|||v||| := \big(\varepsilon|v|_{1,\Omega}^2 + \sigma\|v\|_{0,\Omega}^2 + s_h(v,v)\big)^{\frac{1}{2}}, \quad \forall v \in V.$$

The corresponding norm for the gradient-based LPS scheme follows by replacing s_h with $\tilde{s}_h$.

Lemma 2. *The following a priori estimate is valid for the SD-based LPS scheme*

$$\varepsilon|u_h|_{1,\Omega}^2 + \sigma\|u_h\|_{0,\Omega}^2 \leqslant |||u_h|||^2 \leqslant (f, u_h)_\Omega, \tag{13}$$

hence existence and uniqueness of $u_h \in V_{h,k}$ in the scheme (5) follow.

Proof. For any $v \in V$, integration by parts yields $(\boldsymbol{b}\cdot\nabla v, v)_\Omega = -\frac{1}{2}((\nabla\cdot\boldsymbol{b})v, v)_\Omega = 0$ and therefore

$$(a + s_h)(v,v) = \varepsilon|v|_{1,\Omega}^2 + \sigma\|v\|_{0,\Omega}^2 + s_h(v,v) = |||v|||^2, \quad \forall v \in V. \tag{14}$$

This implies (13), hence existence and uniqueness of $u_h \in V_{h,k}$ in the scheme (5). □

3.2. Approximate Galerkin orthogonality

In LPS methods the Galerkin orthogonality is not fulfilled and a careful analysis of the consistency error has to be done.

Lemma 3. *Let $u \in V$ and $u_h \in V_{h,k}$ be the solutions of (2) and of (5), respectively. Then it holds that*

$$a(u - u_h, v_h) = s_h(u_h, v_h), \quad \forall v_h \in V_{h,k}. \tag{15}$$

Proof. The assertion (15) follows by subtracting (5) from (2) with $v = v_h$. □

Now we estimate the consistency error.

Lemma 4. *Let Assumption 2 be fulfilled and let $u \in V$ with $\boldsymbol{b}\cdot\nabla u \in H^l(M)$ for some $l \in \{0,\dots,k\}$ and for all $M \in \mathcal{M}_h$. Then it holds for the SD-based LPS scheme that*

$$\big|s_h(u, v_h)\big| \lesssim \Bigg(\sum_{M\in\mathcal{M}_h} C_M^s \frac{h_M^{2l}}{k^{2l}}|\boldsymbol{b}\cdot\nabla u|_{l,M}^2\Bigg)^{\frac{1}{2}} |||v_h|||, \quad \forall v_h \in V_{h,k}$$

with

$$C_M^s := \min\Bigg\{\tau_M, \frac{(\tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}k^2)^2}{\sigma\, h_M^2}\Bigg\}. \tag{16}$$

Proof. Consider any $M \in \mathcal{M}_h$ and $v_h \in V_{h,k}$. Then the Cauchy–Schwarz inequality and Assumption 2 yield

$$\big(\kappa_h(\boldsymbol{b}\cdot\nabla u), \kappa_h(\boldsymbol{b}\cdot\nabla v_h)\big)_M \lesssim \frac{h_M^l}{k^l}|\boldsymbol{b}\cdot\nabla u|_{l,M}\big\|\kappa_h(\boldsymbol{b}\cdot\nabla v_h)\big\|_{0,M}.$$

Furthermore, we deduce using the L^2 stability of κ_h in Assumption 2, the inverse inequality (8) and the assumption (4) that

$$\big\|\kappa_h(\boldsymbol{b}\cdot\nabla v_h)\big\|_{0,M} \lesssim \|\boldsymbol{b}\|_{[L^\infty(M)]^d}|v_h|_{1,M} \lesssim \|\boldsymbol{b}\|_{[L^\infty(M)]^d}k^2h_M^{-1}\|v_h\|_{0,M}.$$

Thus,

$$\tau_M\big(\kappa_h(\boldsymbol{b}\cdot\nabla u), \kappa_h(\boldsymbol{b}\cdot\nabla v_h)\big)_M \lesssim \sqrt{C_M^s}\,\frac{h_M^l}{k^l}|\boldsymbol{b}\cdot\nabla u|_{l,M}\big(\sigma\|v_h\|_{0,M}^2 + \tau_M\big\|\kappa_h(\boldsymbol{b}\cdot\nabla v_h)\big\|_{0,M}^2\big)^{\frac{1}{2}},$$

which proves the lemma. □

3.3. A priori error estimate

The a priori estimate can be proven using the standard technique of combining the stability and the consistency results of the previous subsections.

Theorem 1. *Let $u \in V$ be the solution of (2) and $u_h \in V_{h,k}$ the solution of (5). We assume that $u \in H^{l+1}(\Omega)$ for some $l \in \{1,\dots,k\}$ and that $\boldsymbol{b}\cdot\nabla u \in H^l(M)$ for all $M \in \mathcal{M}_h$. Furthermore let Assumptions 2 and 3 for the coarse space D_h be satisfied. Then it holds for the SD-based LPS scheme that*

$$|||u-u_h|||^2 \lesssim \sum_{M\in\mathcal{M}_h}\left\{C_M^s\frac{h_M^{2l}}{k^{2l}}|\boldsymbol{b}\cdot\nabla u|_{l,M}^2 + \left(1+\frac{1}{\beta}\right)^2 C_M\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2\right\} \tag{17}$$

with C_M^s defined in (16) and

$$C_M := \varepsilon + \sigma\frac{h_M^2}{k^4} + \frac{h_M^2}{\tau_M k^4} + \tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}^2.$$

Proof. The error is split into $u-u_h = (u-j_hu)+(j_hu-u_h)$. We start with the approximation error $u-j_hu$. Lemma 1 yields

$$|||u-j_hu||| \lesssim \left(1+\frac{1}{\beta}\right)\left(\sum_{M\in\mathcal{M}_h}\left[\varepsilon+\sigma\frac{h_M^2}{k^4}+\tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}^2\right]\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2\right)^{\frac12}.$$

Now we estimate the remaining part $w_h := j_hu-u_h$ using (14)

$$\begin{aligned}|||j_hu-u_h||| &= \frac{(a+s_h)(j_hu-u_h,w_h)}{|||w_h|||}\\ &= \frac{(a+s_h)(u-u_h,w_h)}{|||w_h|||}+\frac{(a+s_h)(j_hu-u,w_h)}{|||w_h|||} =: I+II.\end{aligned}$$

Applying Lemmata 3 and 4, the first term is bounded by

$$I = \frac{s_h(u,w_h)}{|||w_h|||} \lesssim \left(\sum_{M\in\mathcal{M}_h} C_M^s\frac{h_M^{2l}}{k^{2l}}|\boldsymbol{b}\cdot\nabla u|_{l,M}^2\right)^{\frac12}.$$

Now we consider the terms of *II* separately. Integration by parts, the orthogonality property (11) and the estimate (12) yield for $w_h \in V_{h,k}$ that

$$\begin{aligned}a(j_hu-u,w_h) &= \varepsilon\big(\nabla(j_hu-u),\nabla w_h\big)_\Omega - \big(\kappa_h(\boldsymbol{b}\cdot\nabla w_h),j_hu-u\big)_\Omega + \sigma(j_hu-u,w_h)_\Omega\\ &\lesssim \left(1+\frac{1}{\beta}\right)\left(\sum_{M\in\mathcal{M}_h}\left[\varepsilon+\left(\sigma+\frac{1}{\tau_M}\right)\frac{h_M^2}{k^4}\right]\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2\right)^{\frac12}|||w_h|||.\end{aligned}$$

The estimate of the stabilization term follows using (7) and (12)

$$s_h(j_hu-u,w_h) \lesssim \left(1+\frac{1}{\beta}\right)\left(\sum_{M\in\mathcal{M}_h}\tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}^2\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2\right)^{\frac12}|||w_h|||.$$

Summing up all inequalities in this proof gives the assertion. □

3.4. Parameter design

Now we will calibrate the stabilization parameters τ_M with respect to the local mesh size h_M, the polynomial degree k of the discrete ansatz functions and problem data. The parameters τ_M are determined by balancing the terms $\frac{h_M^2}{\tau_M k^4} \sim \tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}^2$ in C_M on the right-hand side of the general a priori error estimate (17), hence

$$\tau_M \sim \frac{h_M}{\|\boldsymbol{b}\|_{[L^\infty(M)]^d}k^2}. \tag{18}$$

Note that the discrete problem is well defined also if $\|\boldsymbol{b}\|_{[L^\infty(M)]^d}=0$ for some $M\in\mathcal{M}_h$ since

$$\big|s_h(v,w)\big| \lesssim \sum_{M\in\mathcal{M}_h}\tau_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}^2|v|_{1,M}|w|_{1,M},\quad \forall v,w\in V.$$

Corollary 1. *If τ_M satisfies* (18)*, then we obtain for the SD-based LPS scheme under the assumptions of Theorem* 1

$$|||u-u_h|||^2 \lesssim \sum_{M\in\mathcal{M}_h}\left\{\frac{h_M^{2l}}{k^{2l}}\min\left\{\frac{h_M}{k^2}\frac{|\boldsymbol{b}\cdot\nabla u|_{l,M}^2}{\|\boldsymbol{b}\|_{[L^\infty(M)]^d}},\frac{|\boldsymbol{b}\cdot\nabla u|_{l,M}^2}{\sigma}\right\}\right.$$
$$\left.+\left(1+\frac{1}{\beta}\right)^2\left[\varepsilon+\sigma\frac{h_M^2}{k^4}+\|\boldsymbol{b}\|_{[L^\infty(M)]^d}\frac{h_M}{k^2}\right]\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2\right\}.$$

Remark 5. This result requires some discussion:

(i) For $l=k$ and $\varepsilon\lesssim h_M$, we obtain for the second right-hand side term in Corollary 1 the optimal convergence rate $\mathcal{O}(h_M^{k+\frac{1}{2}})$ with respect to h_M.
For the first right-hand side term, the optimal rate is obtained if $\boldsymbol{b}\neq\boldsymbol{0}$ in $\bar{\Omega}$. If this is not the case, then in a neighborhood of points where $\boldsymbol{b}$ vanishes, the term $|\boldsymbol{b}\cdot\nabla u|_{l,M}^2/\|\boldsymbol{b}\|_{[L^\infty(M)]^d}$ may tend to infinity for $h\to 0$. If $\sigma>0$, then one gets at least the suboptimal rate $\mathcal{O}(h_M^k)$ but if $\sigma=0$, an additional reduction of the rate may occur. A simple example is the case $(\boldsymbol{b}\cdot\nabla u)(x)=|\boldsymbol{b}(x)|=x_1^l$, $x_1\geqslant 0$, for which $|\boldsymbol{b}\cdot\nabla u|_{l,M}^2/\|\boldsymbol{b}\|_{[L^\infty(M)]^d}\sim h_M^{-l}$ for $M\subset\{x\in\mathbb{R}^d;\ x_1\geqslant 0\}$ intersecting the line $\{x\in\mathbb{R}^d;\ x_1=0\}$.

(ii) The a priori estimate (17) in Theorem 1 seems to be suboptimal with respect to the polynomial degree k. This suboptimal dependence on k is a consequence of the estimate of $|v-j_hv|_{1,M}$ in Lemma 1 which is presumably suboptimal with respect to k as well. In fact, one would expect that j_h satisfies the estimate

$$\frac{h_M}{k}|v-j_hv|_{1,M}\lesssim\left(1+\frac{1}{\beta}\right)\frac{h_M^l}{k^l}\|v\|_{l,\omega_M}.$$

Assuming that this estimate holds true, a careful check of the above proofs reveals that the explicit dependence on k in the a priori estimate (17) becomes optimal and we obtain

$$\tau_M\sim\frac{h_M}{\|\boldsymbol{b}\|_{[L^\infty(M)]^d}k}$$

instead of (18). Unfortunately, in the numerical Example 1 below we could not find significantly different results for both choices of τ_M.

Remark 6. The result for the gradient-based LPS scheme (see Remark 2) corresponding to Corollary 1 reads as follows: Assume that $\boldsymbol{b}\in[W^{1,\infty}(\Omega)]^d$, $\sigma>0$ and $u\in H^{l+1}(\Omega)$ for some $l\in\{1,\dots,k\}$. Moreover, let Assumptions 2 and 3 hold. For $\tilde{\tau}_M\sim h_M\|\boldsymbol{b}\|_{[L^\infty(M)]^d}/k^2$ we obtain for the gradient-based LPS scheme

$$|||u-u_h|||^2\lesssim\left(1+\frac{1}{\beta}\right)^2\sum_{M\in\mathcal{M}_h}\left[\varepsilon+\sigma\frac{h_M^2}{k^4}+\frac{h_M^2|\boldsymbol{b}|_{[W^{1,\infty}(M)]^d}^2}{\sigma}+\|\boldsymbol{b}\|_{[L^\infty(M)]^d}\frac{h_M}{k^2}\right]\frac{h_M^{2l}}{k^{2l-2}}\|u\|_{l+1,\omega_M}^2.$$

For $l=k$, $\varepsilon\lesssim h_M$, we obtain the optimal convergence rate $\mathcal{O}(h_M^{k+\frac{1}{2}})$ with respect to h_M. This estimate is better with respect to h_M than for the SD-based LPS scheme, see Remark 5(i).

On the other hand, in contrast to the SD-based LPS scheme, we cannot allow $\sigma=0$. This is because of the estimate of the term $(\kappa_h(\boldsymbol{b}\cdot\nabla w_h),j_hu-u)_\Omega$ in the proof of Theorem 1. Since the norm $|||w_h|||$ is defined using $\kappa_h\nabla w_h$ and not $\kappa_h(\boldsymbol{b}\cdot\nabla w_h)$, the function $\boldsymbol{b}$ is approximated by a piecewise constant function $\boldsymbol{b}_h$ and the identity

$$\kappa_h(\boldsymbol{b}\cdot\nabla w_h)=\boldsymbol{b}\cdot\kappa_h\nabla w_h-(\boldsymbol{b}-\boldsymbol{b}_h)\cdot\kappa_h\nabla w_h+\kappa_h\big((\boldsymbol{b}-\boldsymbol{b}_h)\cdot\nabla w_h\big)$$

is used. The estimation of the second and third term on the right-hand side of this identity is based on estimating $\|\boldsymbol{b}-\boldsymbol{b}_h\|_{[L^\infty(M)]^d}$ and applying the inverse inequality (8). This leads to $\|w_h\|_{0,\Omega}$ which is present in $|||w_h|||$ only if $\sigma>0$.

Remark 7. The LPS approach was first proposed for the Stokes problem in [3] using local projections of the pressure gradient. Like in the present paper, the stabilization term contains a parameter which has to be chosen by the user. An attractive alternative to this approach is the parameter-free LPS proposed in [5] where not the pressure gradient but the pressure itself is projected. However, it cannot be expected that the approach of [5] could be successfully extended to the advection–diffusion–reaction problem considered in this paper since the stabilization has to depend locally on the relation among diffusion, convection and reaction.

4. Examples of finite element spaces

The paper [21] presents different variants for the choice of the discrete spaces $V_{h,k}$ and D_h using simplicial, quadrilateral and hexahedral elements. There are two basic variants of the LPS methods: the one-level approach for which $\mathcal{M}_h = \mathcal{T}_h$ and the two-level approach for which the mesh $\mathcal{T}_h$ is obtained by refining the mesh $\mathcal{M}_h$, see Fig. 1 for $d = 2$. In what follows, we describe some details of these two approaches.

We shall assume that all macro-elements in $\mathcal{M}_h$ are affine equivalent to the reference element $\hat{T}$ and that $D_h \subset P_{m,\mathcal{M}_h}$ for some $m \in \mathbb{N}_0$. Let us formulate a sufficient condition for the validity of the inf-sup condition (10). We introduce a reference bubble function $\hat{b} \in C(\bar{\hat{T}}) \cap H_0^1(\hat{T})$ satisfying $\hat{b} \geqslant 0$ and $\hat{b} \neq 0$ and, for any $M \in \mathcal{M}_h$, we set $b_M = \hat{b} \circ F_M^{-1}$. Then there is a positive constant α such that

$$(b_M q, q)_M \geqslant \alpha \|q\|_{0,M}^2, \quad \forall q \in D_h(M), M \in \mathcal{M}_h.$$

Thus, it suffices to require that

$$b_M \cdot D_h(M) \subset Y_h(M), \quad \forall M \in \mathcal{M}_h. \tag{19}$$

Then the inf-sup condition (10) holds with $\beta = (\alpha/\|\hat{b}\|_{L^\infty(\hat{T})})^{1/2}$. Note that a necessary condition for the validity of (10) is that $\dim Y_h(M) \geqslant \dim D_h(M)$. Therefore, if $Y_h(M) = b_M \cdot D_h(M)$, then $Y_h(M)$ has the smallest possible dimension.

The one-level approach with $\mathcal{M}_h = \mathcal{T}_h$ starts from a given discontinuous space D_h and uses an enrichment of the spaces $P_{k,\mathcal{T}_h} \cap V$ or $Q_{k,\mathcal{T}_h} \cap V$ to satisfy (19). For simplicial elements, we set

$$D_h := P_{k-1,\mathcal{T}_h}, \qquad V_{h,k} := \big\{ v \in V;\ v|_T \circ F_T \in P_k^{\mathrm{bub}}(\hat{T})\ \forall T \in \mathcal{T}_h \big\},$$

where

$$P_k^{\mathrm{bub}}(\hat{T}) := P_k(\hat{T}) + \hat{b} \cdot P_{k-1}(\hat{T}), \qquad \hat{b}(\hat{x}) := (d+1)^{d+1} \prod_{i=1}^{d+1} \hat{\lambda}_i(\hat{x})$$

with the barycentric coordinates $\hat{\lambda}_i$, $i = 1, \ldots, d+1$. For quadrilateral/hexahedral elements, we can use either $D_h = P_{k-1,\mathcal{T}_h}$ or $D_h = Q_{k-1,\mathcal{T}_h}$. Setting $\hat{D} = P_{k-1}(\hat{T})$ or $\hat{D} = Q_{k-1}(\hat{T})$, respectively, the spaces $V_{h,k}$ are constructed analogously as for simplices with

$$Q_k^{\mathrm{bub}}(\hat{T}) := Q_k(\hat{T}) + \hat{b} \cdot \hat{D}, \qquad \hat{b}(\hat{x}) := \prod_{i=1}^{d} \big(1 - \hat{x}_i^2\big),$$

where $\hat{T} = (-1, 1)^d$. In the numerical experiments presented in the next section, we consider $\hat{D} = Q_{k-1}(\hat{T})$.

Now consider the two-level approach (cf. Fig. 1 for $d = 2$). In the simplicial case, each element $M \in \mathcal{M}_h$ is divided into $d+1$ simplices by connecting the barycentre of M with the vertices of M. For quadrilateral/hexahedral elements, each $M \in \mathcal{M}_h$ is uniformly refined into 2^d subelements. Then, for simplices, we set

$$V_{h,k} := P_{k,\mathcal{T}_h} \cap V, \qquad D_h := P_{k-1,\mathcal{M}_h}$$

and, for quadrilaterals/hexahedra,

$$V_{h,k} := Q_{k,\mathcal{T}_h} \cap V, \qquad D_h := Q_{k-1,\mathcal{M}_h}.$$

Then the condition (19) is obviously satisfied if $\hat{b} \in H_0^1(\hat{T})$ is defined as a non-negative piecewise P_1/Q_1 function with respect to a division of $\hat{T}$ corresponding to the relation between $\mathcal{M}_h$ and $\mathcal{T}_h$. Hence the inf-sup constant β in Assumption 3 is independent of h. Moreover, the β scales like $\mathcal{O}(\sqrt{k})$ for simplicial elements and like $\mathcal{O}(1)$ for quadrilateral elements in the affine case, see [23].

Note that, for the two-level approach based on simplicial finite elements, the space $V_{h,k}$ can be written in the form

$$V_{h,k} = \big\{ v \in V:\ v|_M \circ F_M \in P_k(\hat{T}) \oplus \hat{B}_k\ \forall M \in \mathcal{M}_h \big\},$$

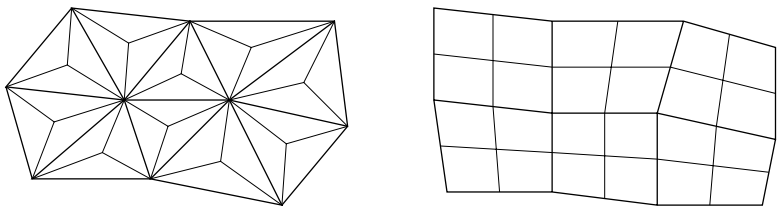


Fig. 1. Relation between the meshes $\mathcal{M}_h$ and $\mathcal{T}_h$ in the two-level approach. The bold lines indicate the mesh $\mathcal{M}_h$, the fine lines $\mathcal{T}_h$.

where $\hat{B}_k \subset H_0^1(\hat{T})$ is a finite-dimensional space consisting of continuous piecewise polynomial functions of degree k. Therefore, the simplicial two-level approach can be regarded as a one-level approach with respect to the mesh $\mathcal{M}_h$. This will be used in Section 6.

5. Comparison of one- and two-level approach

In this section, we provide a comparison of the one- and two-level variants of the LPS method. The following arguments are relevant for the comparison regarding the efficiency and flexibility:

The data structure for the one-level method is much simpler than for the two-level approach. Moreover, adaptive mesh refinement tools can be easier incorporated. On the other hand, for the same fine mesh, the one-level approach requires more degrees of freedom than the two-level approach.

Moreover, there is a formal argument from the regularity point of view against the SD-based variant of the two-level method: The assumption $\boldsymbol{b} \cdot \nabla u \in H^l(M)$ for all $M \in \mathcal{M}_h$ in Theorem 1 implicitly requires that $\boldsymbol{b} \in [H^l(M)]^d$. This is not realistic as $\boldsymbol{b}$ is usually a finite element solution stemming from a flow simulation. Please note that this argument is not valid for the gradient-based variant of the two-level method.

Now we proceed with the comparison by evaluating some numerical experiments for the SD-based LPS scheme. First of all, we emphasize that both, the one-level and the two-level method, perform very well according to the theory of Section 3 for problems with solutions without boundary and interior layers.

Example 1 *(Smooth solution without layers).* Consider in $\Omega = (0,1)^2$ the model problem (1) with $\varepsilon = 10^{-9}$, $\boldsymbol{b} = (-x_2, x_1)^T$ and $\sigma \in \{0, 10^3\}$. The exact solution

$$u(x) = e^{-x_1 x_2} \sin(\pi x_1) \sin(\pi x_2)$$

has no layers and generates the right-hand side $f = Lu$.

Figs. 2 and 3 show exemplarily convergence plots with respect to the norm $(\varepsilon|\cdot|_{1,\Omega}^2 + \|\cdot\|_{0,\Omega}^2)^{\frac{1}{2}}$ on a sequence of equidistant grids with $h \in \{\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}\}$ and polynomial degree $k \in \{1,2,3,4,5,6\}$. Results are only shown for the SUPG method and for the two-level LPS approach as the corresponding results for the one-level method are similar.

The k-convergence is according to our theoretical results and not sensitive with respect to σ. It turns out that the results of the SUPG method are better for $\sigma = 0$. For $\sigma = 10^3$ both methods give similar results. The LPS parameter is defined by $\tau_M = \tau_0 \frac{h_M}{k^2 \|\boldsymbol{b}\|_{[L^\infty(M)]^d}}$ with the rather small (optimized) scaling parameter $\tau_0 = 0.003$ whereas the corresponding formula for the SUPG parameter has the (much larger) scaling parameter $\tau_0 = 1$. The design of the LPS parameter as $\tau_M = \tau_0 \frac{h_M}{k \|\boldsymbol{b}\|_{[L^\infty(M)]^d}}$, see Remark 5(ii), gave no significantly different results.

From here, we concentrate ourselves on the more interesting case of problems with layers. In all numerical experiments, the computational domain Ω is the unit square. We shall consider both one- and two-level approach which will be compared with the SUPG method. The parameter design is $\tau_M = \tau_0 h_M$ for the LPS methods and $\delta_T = \delta_0 h_T$ for the SUPG method with free parameters τ_0 and δ_0. The computations were performed for the one-level method with the Q_1^{bub} and Q_2^{bub} elements on uniform grids consisting of 64×64 and of 32×32 equal square elements, respectively. Similarly, for the SUPG method, we apply the Q_1 and Q_2 elements on uniform grids consisting of 64×64 and of 32×32 equal square elements, respectively. For the two-level approach, we apply the Q_1 and Q_2 elements on uniform grids consisting of 128×128 and of

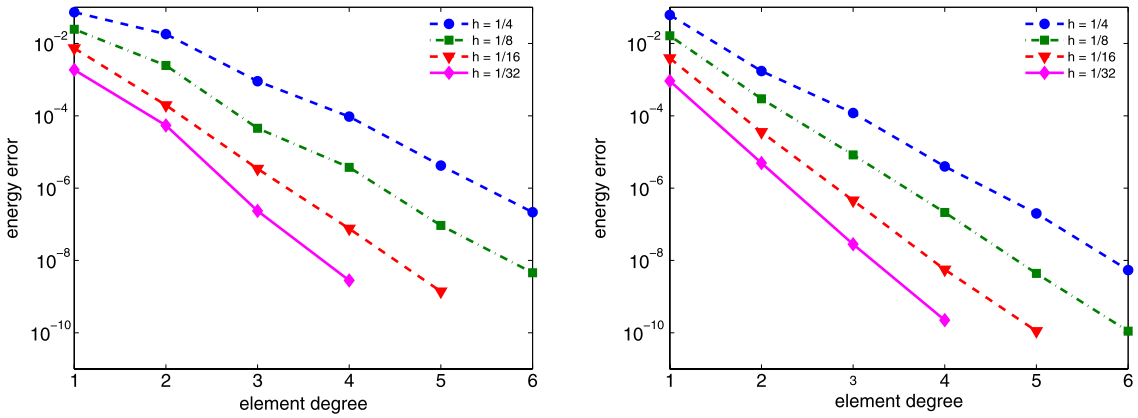


Fig. 2. Dependence of error of two-level LPS scheme (left) and SUPG scheme (right) on h and polynomial degree k for Example 1 with $\sigma = 0$.

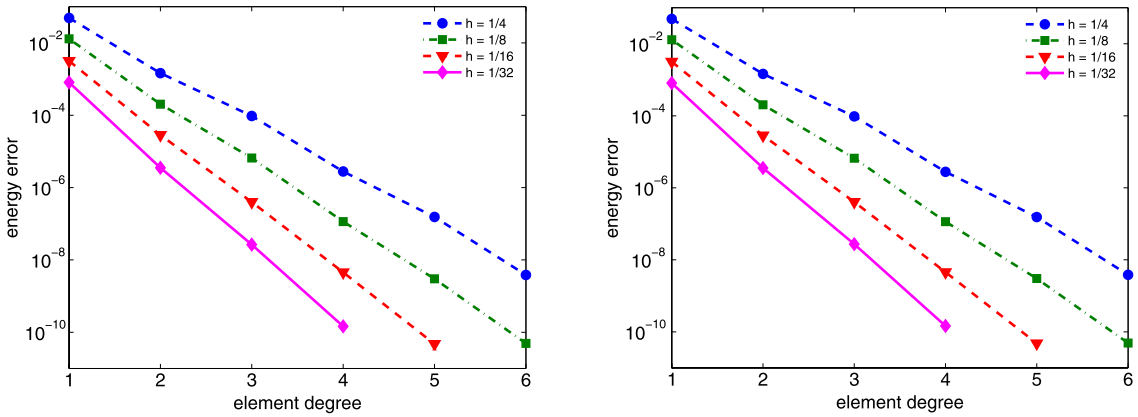


Fig. 3. Dependence of error of two-level LPS scheme (left) and SUPG scheme (right) on h and polynomial degree k for Example 1 with $\sigma = 10^3$.

64×64 equal square elements, respectively. Thus, the corresponding coarse meshes $\mathcal{M}_h$ consist of 64×64 and of 32×32 elements and hence are the same as for the one-level approach. This gives an almost fair comparison of both approaches.

We start with two rather academic problems where the flow field $\boldsymbol{b}$ is aligned with the uniform (Cartesian) mesh in Ω.

Example 2 *(Exponential outflow layer).* (See [22], Example 4.2.) Consider in $\Omega = (0,1)^2$ the model problem (1) with $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (0,2)^T$ and $\sigma = 0$. The exact solution

$$u(x) = (2x_1 - 1)\frac{1 - \exp(-2(1-x_2)/\varepsilon)}{1 - \exp(-2/\varepsilon)}$$

has an exponential boundary layer at the outflow part of the boundary and generates the right-hand side $f = Lu$. On the whole boundary of Ω, a Dirichlet boundary condition determined by u is prescribed. Note that the limit solution $\lim_{\varepsilon \to 0} u(x) = 2x_1 - 1$ can be exactly interpolated by Q_k elements, $k \geqslant 1$.

Fig. 4 provides a comparison of the errors in the L^2 norm, H^1 seminorm and the (discrete) L^∞ norm vs. the scaling parameters τ_0 for the LPS method and δ_0 for the SUPG method. We calculate all (semi)norms on the subdomain Ω_0 which does not contain those elements $M \in \mathcal{M}_h$ which intersect the outflow boundary layer at $x_2 = 1$. In particular, the H^1 seminorm of u on these elements would otherwise dominate the error. For the Q_1^{bub} and Q_2^{bub} elements, we drop the additional bubble functions when computing the errors.

First let us consider the Q_1 and Q_1^{bub} elements in the left column of Fig. 4. For all methods, one observes a global minimum of the errors for some τ_0^* and δ_0^*, which corresponds to the nodally exact solution on Ω resp. Ω_0 in case of the two-level method. The two-level solution possesses a spurious oscillation along $x_2 = 1 - 1/128$ which is in agreement with the one-dimensional theoretical investigations of [25].

The results are less good for the Q_2 and Q_2^{bub} elements in the right column of Fig. 4 as nodally exact discrete solutions cannot be obtained. Nevertheless, a global minimum can be observed for certain values of τ_0^* and δ_0^*. The LPS methods are clearly outperformed by the SUPG method with the optimized parameter δ_0^*. Furthermore, we observe that the one-level method leads to larger errors with respect to all norms than the two-level method. In particular, the one-level method leads to larger oscillations than the two-level method. This is highlighted by Fig. 5 where a cross-section of the discrete solutions at $x_1 = 1 - 1/32$ is shown (here the largest oscillations of the discrete solution can be observed). The solutions are shown only for $x_2 \geqslant 0.7$ since they are nearly constant for $x_2 < 0.7$. It can also be seen that the discrete solutions can be improved if they are replaced by the piecewise bilinear interpolate in case of the one-level method and by the piecewise biquadratic interpolate on the macro-mesh in case of the two-level method. Fig. 5 further shows the SUPG solution which is significantly better than both LPS solutions although much less degrees of freedom are needed.

In the above comparison, the number of degrees of freedom considered for the one-level method is smaller than for the two-level method, which leads to a larger smearing of the boundary layer in case of the one-level method, see Fig. 5. If we apply the one-level method on the fine mesh of the two-level method (and hence the number of degrees of freedom is larger for the one-level method than for the two-level method), than the smearings caused by both LPS methods are comparable but the oscillations of the one-level solutions remain larger than for the two-level method. Also the errors considered in Fig. 4 remain larger for the one-level method.

Example 3 *(Parabolic layers).* (See [22], Example 4.4.) Consider in $\Omega = (0,1)^2$ the model problem (1) with $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (0, 1 + x_1^2)^T$, $\sigma = 0$ and $f = 0$. At the outflow boundary $\Gamma_{\text{out}} = (0,1) \times \{1\}$, a homogeneous Neumann condition is considered

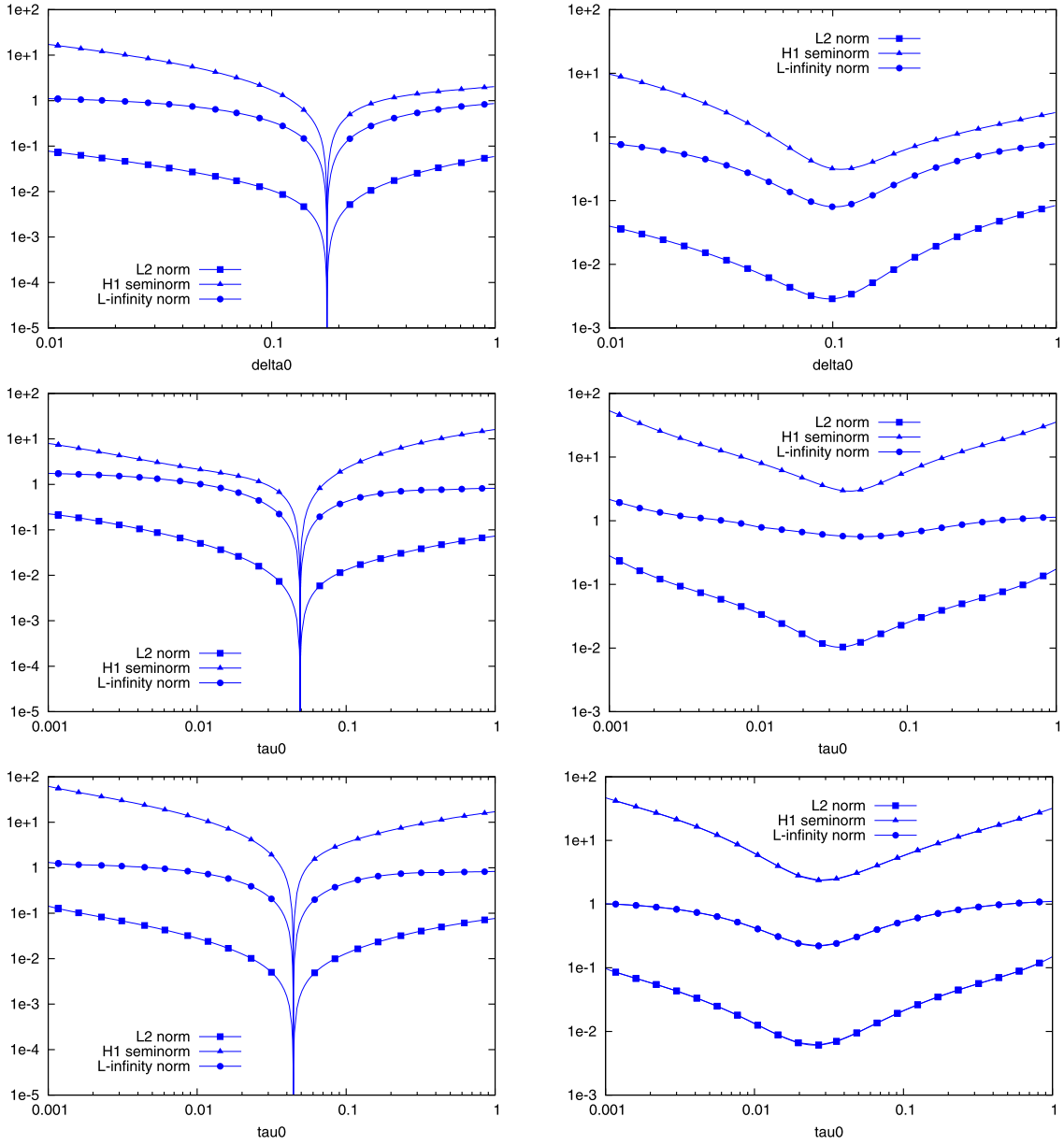


Fig. 4. Dependence of errors on scaling parameters δ_0 and τ_0 for different methods and Example 2: Q_1 elements (left column) and Q_2 elements (right column) for SUPG method (first row), one-level LPS method (second row) and two-level LPS method (third row).

whereas, at $\partial\Omega \setminus \Gamma_{\text{out}}$, an inhomogeneous Dirichlet condition $u(x) = 1 - x_2$ is prescribed. The exact solution exhibits parabolic layers at $x_1 = 0$ and $x_1 = 1$.

As an exact solution is not available, we provide a comparison of cross-sections of the discrete solution at the outflow part of the boundary at $x_2 = 1$ for different values of τ_0.

For this example, the Galerkin method leads to solutions with spurious oscillations localized along the boundary layers, see Fig. 6 left. Moreover, the oscillations depicted in this figure disappear if we represent the discrete solutions by their values at the vertices of the 32×32 mesh, see Fig. 6 right. This nice behaviour is seemingly an effect of the Cartesian

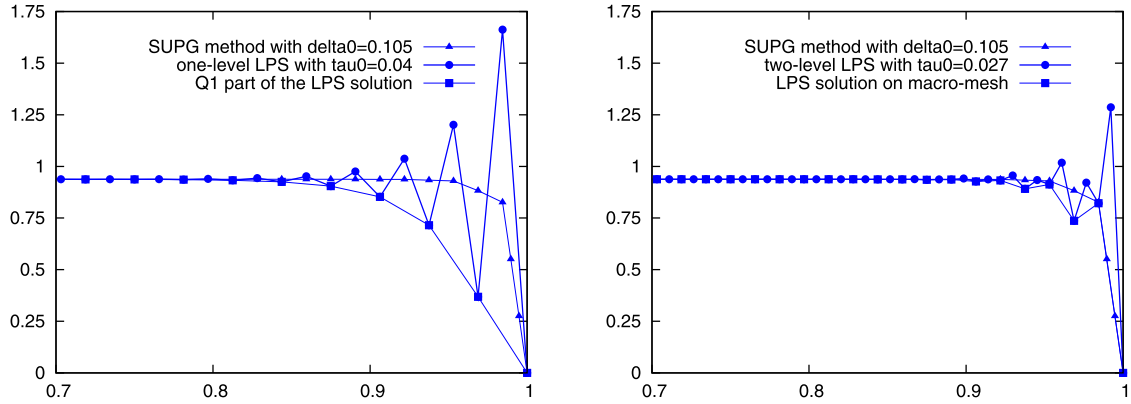


Fig. 5. Cross-section of the discrete solutions for Example 2 at $x_1 = 1 - 1/32$ for one-level method with Q_2^{bub} elements (left) and two-level method with Q_2 elements (right) compared to the SUPG solution.

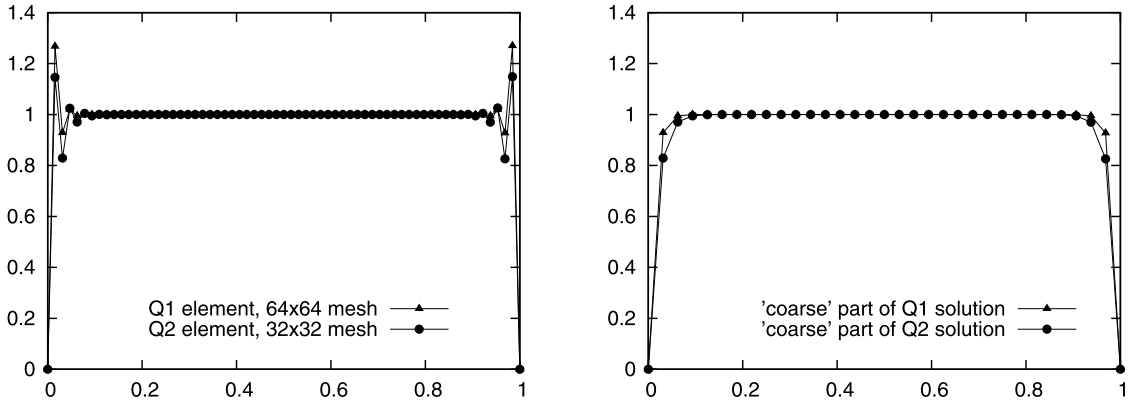


Fig. 6. Outflow profiles for the Galerkin solutions of Example 3.

mesh being aligned with the flow field $\boldsymbol{b}$. In what follows, we shall investigate to what extent the Galerkin solutions can be improved by means of the LPS method. We shall present the outflow profiles only in a neighborhood of the right boundary layer.

For all four LPS methods and $\tau_0 \in (0.01, 1)$, the outflow profiles are very similar to that of the Galerkin method with the Q_1 or Q_2 element on the mesh $\mathcal{T}_h$ of the respective LPS method. For the two-level methods, this is true also for smaller values of τ_0. For the one-level methods, the behaviour for $\tau_0 \in (0, 0.01)$ is different since the Galerkin solutions for the Q_1^{bub} or Q_2^{bub} elements significantly differ from the Galerkin solutions for the Q_1 or Q_2 elements, respectively.

For $\tau_0 > 10^3$, the LPS with the Q_1 element leads to very similar outflow profiles as the LPS with the Q_1^{bub} element, and the LPS with the Q_2 element gives almost the same outflow profiles as the LPS with the Q_2^{bub} element. However, the qualitative behaviour of the first order and the second order LPS methods is different. Whereas, for the second order LPS methods, the outflow profiles are basically independent of $\tau_0 > 10^3$, the first order LPS methods introduce a considerable smearing of the boundary layers which increases with increasing τ_0 and makes the discrete solutions useless.

It remains to discuss the properties of the LPS methods for $\tau_0 \in (1, 10^3)$, see Fig. 7. As we observe, for first order LPS methods, the oscillations decrease with increasing τ_0 but simultaneously the boundary layers are smeared. For second order LPS methods, the oscillations first decrease but soon they again start to increase and, for $\tau_0 = 10^2$, they are already larger than for the Galerkin method. Thus, for first order LPS methods, oscillation-free discrete solutions can be obtained only at the prize of smearing the layers. For second order LPS methods, it seems that, for any choice of τ_0, it is not possible to obtain a discrete solution with sufficiently suppressed spurious oscillations.

An alternative way to suppress the spurious oscillations of the LPS solutions is to consider only a 'coarse' part of the solution like in Fig. 6. However, for the two-level methods, this does not lead to an improvement in comparison with the 'coarse' part of the Galerkin solution. For the one-level methods, a small improvement is possible, nevertheless, it is

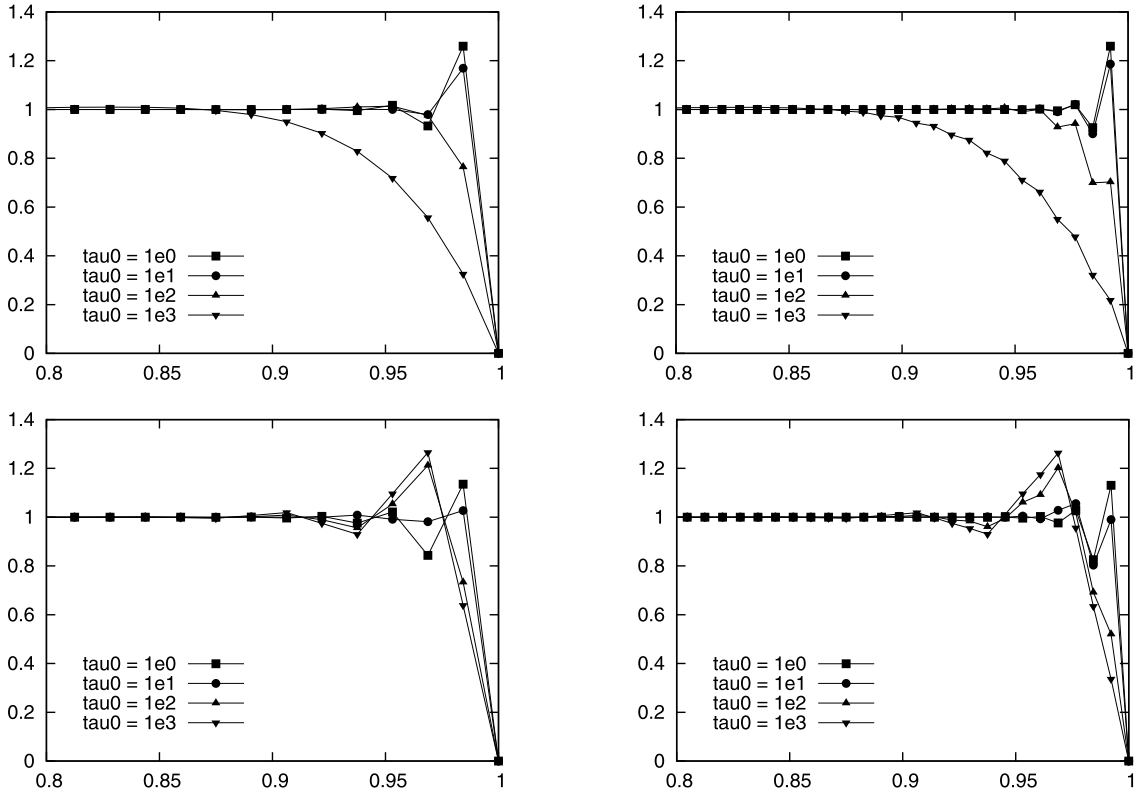


Fig. 7. Outflow profiles for LPS solutions of Example 3 with different values of τ_0: one-level LPS (left column) and two-level LPS (right column) for the Q_1^{bub} and Q_1 elements (first row) and for the Q_2^{bub} and Q_2 elements (second row).

questionable whether this improvement is worth the increased computational cost. Moreover, it is very sensitive to the choice of τ_0.

Note that the restriction of the discrete solution to a coarse grid shows the stabilizing effect of the small scales on the large scales. Eliminating the small scales from the discrete problem would lead to a formulation where the influence of the small (now unresolved) scales is represented by an additional (stabilization) term. Thus, the 'coarse' part of the discrete solution can be also interpreted as a solution of a VMS method.

Finally, we consider an example where the flow field $\boldsymbol{b}$ is not aligned with the uniform (Cartesian) mesh.

Example 4 *(Interior layers).* Consider in $\Omega = (0,1)^2$ the model problem (1) with $\varepsilon = 10^{-7}$, $\boldsymbol{b} = (-x_2, x_1)^T$, $\sigma = 0$ and $f = 0$. At the outflow boundary $\Gamma_{\text{out}} = (0,1) \times \{1\}$, a homogeneous Neumann condition is considered whereas, at $\partial\Omega \setminus \Gamma_{\text{out}}$, an inhomogeneous Dirichlet condition $u(x) = 1$ for $x \in [\frac{1}{3}, \frac{2}{3}] \times \{0\}$ and $u(x) = 0$ elsewhere is prescribed. The exact solution exhibits interior parabolic layers starting from the discontinuities of the inflow profile at $x_2 = 0$.

The solutions of all four LPS methods with optimized parameters τ_0 are comparable, see Fig. 8 where two such solutions are shown. The discrete solutions detect the interior layers well but have local spurious oscillations in this numerical layers. A comparison of the results for the LPS methods to the SUPG method (not shown) clarifies that the LPS methods cannot outperform the SUPG method.

Summarizing, both variants of the LPS method give comparable results for problems with boundary and interior layers and we have not found any convincing arguments for preferring one of these variants. All methods are able to detect boundary and interior layers numerically but they are rather sensitive to the scaling of the stabilization parameter. In general, the LPS methods do not attain the quality of the classical SUPG method. As for the SUPG method, the discrete solutions exhibit local spurious oscillations in layer regions unless the mesh is aligned with the advection direction. A potential remedy in case of boundary layers is the weak imposition of Dirichlet data by using Nitzsche's method, cf., e.g., [2]. Another idea is the implementation of additional (non-linear) stabilization terms which reduce oscillations in crosswind directions around layers, see [19]. Moreover, we refer to the possibility to resolve layers with well-adapted anisotropic finite elements, see, e.g., [20].

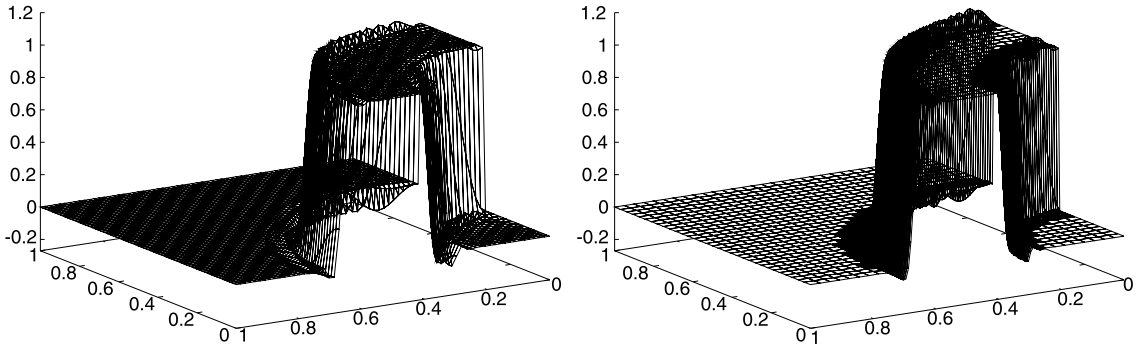


Fig. 8. Plot of the discrete solutions for Example 4 for the one-level method with the Q_1^{bub} element and $\tau_0 = 0.03$ (left) and for the two-level method with the Q_2 element and $\tau_0 = 3$ (right).

6. Relation to residual-based stabilizations

In this section we shall demonstrate that LPS methods based on simplicial meshes are very close to RBS techniques. The dependence on the polynomial degree k will not be considered here.

As we have seen in Section 4, for both the one- and two-level approach, the spaces $V_{h,k}$ and D_h are given by

$$V_{h,k} = \bar{V}_{h,k} \oplus B_{h,k}, \qquad D_h = P_{k-1,\mathcal{M}_h},$$

where

$$\bar{V}_{h,k} := P_{k,\mathcal{M}_h} \cap V, \qquad B_{h,k} := \bigoplus_{M\in\mathcal{M}_h} B_k(M).$$

The spaces $B_k(M)$ are defined using a finite-dimensional space $\hat{B}_k \subset C(\bar{\hat{T}}) \cap H_0^1(\hat{T})$ such that $\hat{B}_k \cap P_k(\hat{T}) = \{0\}$, i.e., for any $M \in \mathcal{M}_h$, we set $B_k(M) := \{\hat{v} \circ F_M^{-1};\ \hat{v} \in \hat{B}_k\}$. Then $B_k(M) \subset H_0^1(M)$ and $B_k(M) \cap P_k(M) = \{0\}$.

Let us consider the gradient-based LPS scheme, i.e., the discrete solution is a function $u_h \in V_{h,k}$ satisfying

$$a(u_h, v_h) + \sum_{M\in\mathcal{M}_h} \tau_M(\kappa_h \nabla u_h, \kappa_h \nabla v_h)_M = (f, v_h)_\Omega, \quad \forall v_h \in V_{h,k}, \tag{20}$$

where we dropped the tilde over τ_M for simplicity. The local projection $\pi_M : L^2(M) \to D_h(M) = P_{k-1}(M)$ used to define the fluctuation operator κ_h is assumed to be the orthogonal L^2 projection of $L^2(M)$ onto $P_{k-1}(M)$. We shall also use the local fluctuation operator $\kappa_M := \mathrm{id} - \pi_M$. Note that, for any $\bar{v}_h \in \bar{V}_{h,k}$, we have $\nabla \bar{v}_h \in [D_h]^d$ and hence $\kappa_h \nabla \bar{v}_h = \mathbf{0}$. Thus, it follows from (20) that

$$a(u_h, \bar{v}_h) = (f, \bar{v}_h)_\Omega, \quad \forall \bar{v}_h \in \bar{V}_{h,k}. \tag{21}$$

We define the bilinear forms

$$a_M(u, v) := \varepsilon(\nabla u, \nabla v)_M + (\boldsymbol{b} \cdot \nabla u, v)_M + \sigma(u, v)_M,$$
$$a_M^\star(u, v) := \varepsilon(\nabla u, \nabla v)_M - (\boldsymbol{b} \cdot \nabla u, v)_M + \sigma(u, v)_M.$$

Then

$$a_M(u, v) = a_M^\star(v, u), \quad \forall u, v \in H_0^1(M),\ M \in \mathcal{M}_h. \tag{22}$$

Denoting

$$L^\star u := -\varepsilon \Delta u - \boldsymbol{b} \cdot \nabla u + \sigma u,$$

we have

$$a_M(u, v) = (Lu, v)_M, \quad \forall u \in H^2(M),\ v \in H_0^1(M), \tag{23}$$
$$a_M(u, v) = (u, L^\star v)_M, \quad \forall u \in H_0^1(M),\ v \in H^2(M). \tag{24}$$

Using the local bilinear forms, we deduce from (20) that, for any $M \in \mathcal{M}_h$, we have

$$a_M(u_h, v_M) + \tau_M(\kappa_M \nabla u_h, \kappa_M \nabla v_M)_M = (f, v_M)_M, \quad \forall v_M \in B_k(M). \tag{25}$$

We denote by $\bar{u}_h \in \bar{V}_{h,k}$ and $u_h^b \in B_{h,k}$ the uniquely determined functions satisfying $\bar{u}_h + u_h^b = u_h$ and set $u_M = u_h^b|_M$ for any $M \in \mathcal{M}_h$. Combining (23) and (25), we derive that

$$a_M(u_M, v_M) + \tau_M(\kappa_M \nabla u_M, \kappa_M \nabla v_M)_M = (f - L\bar{u}_h, v_M)_M, \quad \forall v_M \in B_k(M).$$

We define one-to-one linear operators $A_M,\ A_M^\star : B_k(M) \to B_k(M)$ by

$$a_M(u, v) + \tau_M(\kappa_M \nabla u, \kappa_M \nabla v)_M = (A_M u, v)_M, \quad \forall u, v \in B_k(M),$$
$$a_M^\star(v, u) + \tau_M(\kappa_M \nabla v, \kappa_M \nabla u)_M = \left(u, A_M^\star v\right)_M, \quad \forall u, v \in B_k(M).$$

According to (22), the operator $A_M^\star$ is adjoint to the operator A_M. Clearly,

$$(A_M u_M, v_M)_M = (f - L\bar{u}_h, v_M)_M, \quad \forall v_M \in B_k(M)$$

and hence

$$u_M = A_M^{-1} \varrho_M (f - L\bar{u}_h), \tag{26}$$

where ϱ_M is the orthogonal L^2 projection from $L^2(M)$ onto $B_k(M)$. According to (21), we have

$$a(\bar{u}_h, \bar{v}_h) + \sum_{M \in \mathcal{M}_h} a_M(u_M, \bar{v}_h) = (f, \bar{v}_h)_\Omega, \quad \forall \bar{v}_h \in \bar{V}_{h,k}.$$

Using (24) and (26), we obtain

$$a_M(u_M, \bar{v}_h) = \left(u_M, L^\star \bar{v}_h\right)_M = \left(A_M^{-1} \varrho_M (f - L\bar{u}_h), \varrho_M L^\star \bar{v}_h\right)_M, \quad \forall \bar{v}_h \in \bar{V}_{h,k}$$

and hence we derive that

$$a(\bar{u}_h, \bar{v}_h) + \sum_{M \in \mathcal{M}_h} \left(f - L\bar{u}_h, \left(A_M^\star\right)^{-1} \varrho_M L^\star \bar{v}_h\right)_M = (f, \bar{v}_h)_\Omega, \quad \forall \bar{v}_h \in \bar{V}_{h,k}. \tag{27}$$

Since $(A_M^\star)^{-1}$ maps into $B_k(M)$, it is not necessary to apply the projection ϱ_M to $f - L\bar{u}_h$.

The relation (27) shows that any simplicial LPS method can be interpreted as a residual-based stabilization. The operator $(A_M^\star)^{-1}$ plays the role of a stabilization parameter and we shall investigate in the following how it depends on the LPS parameter τ_M and on the data of the problem (1).

Lemma 5. *There is $\gamma > 0$ such that*

$$\|\kappa_M \nabla v\|_{0,M} \geqslant \gamma \|\nabla v\|_{0,M}, \quad \forall v \in B_k(M),\ M \in \mathcal{M}_h.$$

Proof. Consider any $M \in \mathcal{M}_h$ and $v \in B_k(M)$. Then there is $\hat{v} \in \hat{B}_k$ such that $v = \hat{v} \circ F_M^{-1}$ and we have $\nabla v = (DF_M)^{-T}(\hat{\nabla}\hat{v}) \circ F_M^{-1}$ where DF_M is the Jacobi matrix of F_M. Thus, given any $i \in \{1, \dots, d\}$, there is a vector $\boldsymbol{a} \in \mathbb{R}^d$ such that $(\partial v/\partial x_i) \circ F_M = \boldsymbol{a} \cdot \hat{\nabla}\hat{v}$. Consequently, it suffices to prove the existence of $\gamma > 0$ such that

$$\left\|\hat{\kappa}(\boldsymbol{a} \cdot \hat{\nabla}\hat{v})\right\|_{0,\hat{T}} \geqslant \gamma \|\boldsymbol{a} \cdot \hat{\nabla}\hat{v}\|_{0,\hat{T}}, \quad \forall \boldsymbol{a} \in \mathbb{R}^d,\ \hat{v} \in \hat{B}_k, \tag{28}$$

where $\hat{\kappa} = \mathrm{id} - \hat{\pi}$ and $\hat{\pi}$ is the orthogonal L^2 projection of $L^2(\hat{T})$ onto $P_{k-1}(\hat{T})$. Let us assume that (28) does not hold for any $\gamma > 0$. Then there are sequences $\{\boldsymbol{a}_n\}_{n=1}^\infty \subset \mathbb{R}^d$ and $\{\hat{v}_n\}_{n=1}^\infty \subset \hat{B}_k$ such that $|\boldsymbol{a}_n| = 1$, $\|\hat{\nabla}\hat{v}_n\|_{0,\hat{T}} = 1$ and $\|\hat{\kappa}(\boldsymbol{a}_n \cdot \hat{\nabla}\hat{v}_n)\|_{0,\hat{T}} < (1/n)\|\boldsymbol{a}_n \cdot \hat{\nabla}\hat{v}_n\|_{0,\hat{T}} \leqslant 1/n$ for any $n \in \mathbb{N}$. Since the spaces $\mathbb{R}^d$ and $\hat{B}_k$ are finite-dimensional, there are subsequences $\{\boldsymbol{a}_{n_l}\}$ and $\{\hat{v}_{n_l}\}$ converging to some $\boldsymbol{a} \in \mathbb{R}^d$ and $\hat{v} \in \hat{B}_k$, respectively. Clearly, $|\boldsymbol{a}| = 1$, $\|\hat{\nabla}\hat{v}\|_{0,\hat{T}} = 1$ and $\hat{\kappa}(\boldsymbol{a} \cdot \hat{\nabla}\hat{v}) = 0$. The last relation implies that $\boldsymbol{a} \cdot \hat{\nabla}\hat{v} \in P_{k-1}(\hat{T})$ and hence $\hat{v} \in P_k(\hat{T})$ since $\hat{v} \in C(\bar{\hat{T}}) \cap H_0^1(\hat{T})$. Consequently, $\hat{v} = 0$ as $\hat{B}_k \cap P_k(\hat{T}) = \{0\}$. This is in contradiction with the fact that $\|\hat{\nabla}\hat{v}\|_{0,\hat{T}} = 1$. □

Theorem 2. *There are positive constants C_1 and C_2 such that, for any $M \in \mathcal{M}_h$ and $g \in B_k(M)$, we have*

$$\frac{C_1 h_M^2}{\varepsilon + \tau_M + \|\boldsymbol{b}\|_{[L^\infty(M)]^d} h_M + \sigma h_M^2} \leqslant \frac{\|(A_M^\star)^{-1} g\|_{0,M}}{\|g\|_{0,M}} \leqslant \frac{C_2 h_M^2}{\varepsilon + \tau_M + \sigma h_M^2}. \tag{29}$$

Proof. Consider any $M \in \mathcal{M}_h$ and $g \in B_k(M)$ and set $u = (A_M^\star)^{-1} g$. Then $a_M^\star(u, v) + \tau_M(\kappa_M \nabla u, \kappa_M \nabla v)_M = (g, v)_M$ for any $v \in B_k(M)$. It is well known that

$$C_3 h_M |v|_{1,M} \leqslant \|v\|_{0,M} \leqslant h_M |v|_{1,M}, \quad \forall v \in B_k(M),$$

where C_3 is positive and independent of M and v. Therefore, in view of Lemma 5,

$$h_M|u|_{1,M}\|g\|_{0,M} \geqslant (g,u)_M = \varepsilon|u|_{1,M}^2 + \sigma\|u\|_{0,M}^2 + \tau_M\|\kappa_M\nabla u\|_{0,M}^2$$
$$\geqslant \big(\varepsilon + \gamma^2\tau_M + \sigma C_3^2 h_M^2\big)|u|_{1,M}^2,$$

which implies that

$$\min\big\{1,\gamma^2,C_3^2\big\}\big(\varepsilon+\tau_M+\sigma h_M^2\big)\|u\|_{0,M} \leqslant h_M^2\|g\|_{0,M},$$

thus proving the right-hand side inequality in (29). On the other hand, for any $v \in B_k(M)$, we have

$$(g,v)_M \leqslant \big\{(\varepsilon+\tau_M)C_3^{-1}h_M^{-1} + \|\boldsymbol{b}\|_{[L^\infty(M)]^d} + \sigma h_M\big\}|u|_{1,M}\|v\|_{0,M},$$

where we used the fact that $\|\kappa_M z\|_{0,M}^2 = \|z\|_{0,M}^2 - \|\pi_M z\|_{0,M}^2 \leqslant \|z\|_{0,M}^2$ for any $z \in L^2(M)$. Consequently,

$$C_3^2 h_M^2\|g\|_{0,M} \leqslant \max\{1,C_3\}\big(\varepsilon+\tau_M+\|\boldsymbol{b}\|_{[L^\infty(M)]^d}h_M+\sigma h_M^2\big)\|u\|_{0,M},$$

which completes the proof. □

Remark 8. Let us consider the simplest case $k=1$. Since, for any $M \in \mathcal{M}_h$, the space $B_1(M)$ is one-dimensional, the operator $A_M^\star$ represents a multiplicative factor and we easily obtain

$$\big(A_M^\star\big)^{-1} = \frac{\|b_M\|_{0,M}^2}{(\varepsilon+\tau_M)|b_M|_{1,M}^2 + \sigma\|b_M\|_{0,M}^2},$$

where $b_M = \hat{b}\circ F_M^{-1}$. Moreover, introducing the mean values

$$\boldsymbol{b}_M = \frac{(\boldsymbol{b},b_M)_M}{(1,b_M)_M}, \qquad f_M = \frac{(f,b_M)_M}{(1,b_M)_M}$$

and denoting by x_M the barycentre of M, we derive that

$$\big(f - L\bar{u}_h, \big(A_M^\star\big)^{-1}\varrho_M L^\star\bar{v}_h\big)_M = \delta_M\big(\boldsymbol{b}_M\cdot\nabla\bar{u}_h + \sigma\bar{u}_h - f_M, \boldsymbol{b}_M\cdot\nabla\bar{v}_h - \sigma\bar{v}_h(x_M)\big)_M$$

with

$$\delta_M = \frac{(1,b_M)_M^2}{|M|\{(\varepsilon+\tau_M)|b_M|_{1,M}^2 + \sigma\|b_M\|_{0,M}^2\}},$$

where $|M|$ is the volume of M.

Remark 9. Let us consider the SD-based LPS scheme (5), (6) which we now write in the form

$$a(u_h,v_h) + \sum_{M\in\mathcal{M}_h}\tau_M\big(\kappa_h(\boldsymbol{e_b}\cdot\nabla u_h),\kappa_h(\boldsymbol{e_b}\cdot\nabla v_h)\big)_M = (f,v_h)_\Omega, \quad \forall v_h \in V_{h,k},$$

where $\boldsymbol{e_b} = \boldsymbol{b}/|\boldsymbol{b}|$ ($\boldsymbol{e_b} = \boldsymbol{0}$ if $\boldsymbol{b} = \boldsymbol{0}$). If we assume that $\boldsymbol{b}$ is piecewise constant, we again deduce that the component $\bar{u}_h \in \bar{V}_{h,k}$ of the discrete solution $u_h \in V_{h,k}$ satisfies the relation (27), where the operator $A_M^\star : B_k(M) \to B_k(M)$ is now defined by

$$a_M^\star(v,u) + \tau_M\big(\kappa_M(\boldsymbol{e_b}\cdot\nabla v),\kappa_h(\boldsymbol{e_b}\cdot\nabla u)\big)_M = \big(u, A_M^\star v\big)_M, \quad \forall u,v \in B_k(M).$$

It is easy to check that the statement of Theorem 2 remains valid as well, provided that $\tau_M = 0$ if $\boldsymbol{b}|_M = 0$.

Remark 10. As we see from (29), the limit case $\tau_M \to \infty$ corresponds to the Galerkin discretization (3) with $V_{h,k}$ replaced by $\bar{V}_{h,k}$.

7. Summary

In this paper, we considered the local projection stabilization (LPS) of finite element methods for the linear advection–diffusion–reaction problem. This new technique for the numerical solution of transport-dominated problems preserves the stability of methods with residual-based stabilization but has a symmetric form of the stabilization term. A comparison between the LPS methods and the standard SUPG method showed that results are often comparable but sometimes we obtained better results for the SUPG method. We gave a critical discussion and comparison of the one- and two-level approaches to LPS which showed that there are no convincing arguments for preferring one of these approaches. Moreover, the relation between the LPS method and residual-based stabilization techniques was explained for simplicial elements.

Acknowledgements

The research of Petr Knobloch is a part of the project MSM 0021620839 financed by MSMT and it was partly supported by the Grant Agency of the Academy of Sciences of the Czech Republic under the grant No. IAA100190804. We thank Benjamin Tews and Johannes Löwe for performing numerical experiments for the two-level approach.

References

[1] D.N. Arnold, D. Boffi, R.S. Falk, Approximation by quadrilateral finite elements, Math. Comp. 71 (2002) 909–922.
[2] Y. Bazilevs, T.J.R. Hughes, Weak imposition of Dirichlet boundary conditions in fluid mechanics, Comput. & Fluids 36 (2007) 12–26.
[3] R. Becker, M. Braack, A finite element pressure gradient stabilization for the Stokes equations based on local projections, Calcolo 38 (2001) 173–199.
[4] R. Becker, B. Vexler, Stabilized finite element methods for the generalized Oseen problem, Numer. Math. 106 (2007) 349–367.
[5] P.B. Bochev, C.R. Dohrmann, M.D. Gunzburger, Stabilization of low-order mixed finite elements for the Stokes equations, SIAM J. Numer. Anal. 44 (2006) 82–101.
[6] M. Braack, E. Burman, Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method, SIAM J. Numer. Anal. 43 (2006) 2544–2566.
[7] M. Braack, E. Burman, V. John, G. Lube, Stabilized finite element methods for the generalized Oseen problem, Comput. Methods Appl. Mech. Engrg. 196 (2007) 853–866.
[8] A.N. Brooks, 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.
[9] E. Burmann, P. Hansbo, Edge stabilization for Galerkin approximations of convection–diffusion problems, Comput. Methods Appl. Mech. Engrg. 193 (2004) 1437–1453.
[10] R. Codina, Analysis of a stabilized finite element approximation of the Oseen equations using orthogonal subscales, Appl. Numer. Math. 58 (2008) 264–283.
[11] L.P. Franca, F. Valentin, On an improved unusual stabilized finite element method for the advective–reactive–diffusive equation, Comput. Methods Appl. Mech. Engrg. 190 (2000) 1785–1800.
[12] S. Ganesan, L. Tobiska, Stabilization by local projection for convection–diffusion and incompressible flow problems, J. Sci. Comput., DOI 10.1007/s10915-008-9259-8, in press.
[13] E.H. Georgoulis, Inverse-type estimates on *hp*-finite element spaces and applications, Math. Comput. 77 (2008) 201–219.
[14] V. Girault, P.-A. Raviart, Finite Element Methods for Navier–Stokes Equations, Springer, Heidelberg–Berlin, 1986.
[15] V. Gravemeier, The variational multiscale method for laminar and turbulent flow, Arch. Comput. Methods Engrg. 13 (2006) 249–324.
[16] T.J.R. Hughes, Multiscale phenomena: Green's functions, the Dirichlet-to-Neumann formulation, subgrid scale models, bubbles and the origins of stabilized methods, Comput. Methods Appl. Mech. Engrg. 127 (1995) 387–401.
[17] T.J.R. Hughes, L. Mazzei, K. Janssen, Large eddy simulation and the variational multiscale method, Comput. Vis. Sci. 3 (2000) 47–59.
[18] V. John, S. Kaya, A finite element variational multiscale method for the Navier–Stokes equations, SIAM J. Sci. Comput. 26 (2005) 1485–1503.
[19] V. John, P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review, Comput. Methods Appl. Mech. Engrg. 196 (2007) 2197–2215.
[20] N. Madden, M. Stynes, Efficient generation of oriented meshes for solving convection–diffusion problems, Int. J. Numer. Methods Engrg. 40 (1997) 565–576.
[21] G. Matthies, P. Skrzypacz, L. Tobiska, A unified convergence analysis for local projection stabilizations applied to the Oseen problem, M^2AN Math. Model. Numer. Anal. 41 (2007) 713–742.
[22] G. Matthies, P. Skrzypacz, L. Tobiska, Stabilization of local projection type applied to convection–diffusion problems with mixed boundary conditions, Electron. Trans. Numer. Anal. 32 (2008) 90–105.
[23] G. Rapin, G. Lube, J. Löwe, Applying local projection stabilization to inf-sup stable elements, in: K. Kunisch, G. Of, O. Steinbach (Eds.), Numerical Mathematics and Advanced Applications, Springer-Verlag, Berlin, 2008, pp. 521–528.
[24] H.-G. Roos, M. Stynes, L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations, Springer, Berlin, 2008.
[25] H.-G. Roos, R. Vanselow, A comparison of four- and five-point difference approximations for stabilizing the one-dimensional stationary convection–diffusion equation, Electron. Trans. Numer. Anal. 32 (2008) 63–75.
[26] S. Schmaljohann, Local projection stabilization for the Oseen problem, Master's thesis, Ruhr-Universität Bochum, 2007 (in German).
[27] L.R. Scott, S. Zhang, Finite element interpolation of nonsmooth functions satisfying boundary conditions, Math. Comput. 54 (1990) 483–493.

SIAM J. NUMER. ANAL.
Vol. 48, No. 2, pp. 659–680

A GENERALIZATION OF THE LOCAL PROJECTION STABILIZATION FOR CONVECTION-DIFFUSION-REACTION EQUATIONS*

PETR KNOBLOCH†

Dedicated to Professor Lutz Tobiska on the occasion of his 60th birthday.

Abstract. We introduce a generalization of the local projection stabilization for steady scalar convection-diffusion-reaction equations which allows us to use local projection spaces defined on overlapping sets. This enables us to define the local projection method without the need of a mesh refinement or an enrichment of the finite element space and increases the robustness of the local projection method with respect to the choice of the stabilization parameter. The stabilization term is slightly modified, which leads to an optimal estimate of the consistency error even if the stabilization parameters scale correctly with respect to convection, diffusion, and mesh width. We prove that the bilinear form corresponding to the method satisfies an inf-sup condition with respect to the SUPG norm and establish an optimal error estimate in this norm. The theoretical considerations are illustrated by numerical results.

Key words. finite element method, convection-diffusion-reaction equation, stabilization, local projection, stability, inf-sup condition, error estimates, boundary layers

AMS subject classifications. 65N30, 65N12, 65N15

DOI. 10.1137/090767807

1. Introduction. Let $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a bounded domain with a polyhedral Lipschitz-continuous boundary $\partial\Omega$, and let us consider the convection-diffusion-reaction equation

$$-\varepsilon\,\Delta u + \mathbf{b}\cdot\nabla u + c\,u = f \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega\,. \tag{1.1}$$

We assume that ε is a positive constant and $\mathbf{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $u_b \in H^{1/2}(\partial\Omega)$ are given functions satisfying

$$\sigma := c - \frac{1}{2}\operatorname{div}\mathbf{b} \geq \sigma_0 > 0\,,$$

where σ_0 is a constant. Then the boundary value problem (1.1) has a unique solution in $H^1(\Omega)$.

It is well known that the Galerkin finite element method is not appropriate for solving problem (1.1) numerically since the discrete solution is typically globally polluted by spurious oscillations if convection dominates diffusion (i.e., $|\mathbf{b}| \gg \varepsilon$). To enhance the stability and accuracy of the Galerkin method, various stabilization approaches have been developed; see [19] for an overview. In this paper, we concentrate on stabilization by local projections. This technique was originally proposed for stabilizing discretizations of the Stokes problem in which both the pressure and the velocity

*Received by the editors August 11, 2009; accepted for publication (in revised form) March 15, 2010; published electronically June 2, 2010. This work was supported in part by the Grant Agency of the Czech Republic under grant 201/07/J033 and by the Ministry of Education, Youth and Sports of the Czech Republic in the framework of the research project MSM 0021620839.
http://www.siam.org/journals/sinum/48-2/76780.html
†Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 186 75 Praha 8, Czech Republic, and Institut für Numerische Mathematik, Fakultät Mathematik und Naturwissenschaften, Technische Universität Dresden, D-01062 Dresden, Germany (knobloch@karlin.mff.cuni.cz).

components are approximated using the same finite element space [1]. Later, the local projection method was extended to stabilization of convection dominated problems [2] and applied to various types of incompressible flow problems (see the review article [5]) and to convection-diffusion-reaction problems; see [12, 14, 15, 16, 18]. Local projection stabilizations preserve the stability properties of the popular residual-based stabilizations [19] but do not require the computation of second order derivatives and can be easily applied to nonsteady problems. Moreover, when applied to systems of partial differential equations, it is possible to avoid undesirable couplings between various components of the solution. A further advantage of these techniques is that they are symmetric. Therefore, if they are applied to optimization problems, the operations "discretization" and "optimization" commute [3, 4].

A drawback of all local projection formulations proposed up to now is that they require (significantly) more degrees of freedom than, e.g., residual-based methods. In this paper we remove this drawback by allowing the sets on which local projection spaces are defined to overlap. Although this is a rather simple idea, the corresponding analysis is by no means a straightforward extension of results published before. In contrast to the traditional error analysis, which is based on the construction of a special interpolation operator, we first show that the bilinear form of the local projection method satisfies an inf-sup condition with respect to a norm containing a streamline derivative term. This improved stability of the local projection method enables us to perform the error analysis in a way similar to that for residual-based methods. Of course, it is also important on its own since it shows that the local projection method is more stable than its coercivity suggests. Let us mention that a similar stability result was already established in [16] as a consequence of a more general inf-sup condition, however, only under certain restrictions on the convection field $\mathbf{b}$ or the mesh width h.

Our numerical results show that the overlapping of the sets on which local projection spaces are defined significantly increases the robustness of the local projection method with respect to the choice of the stabilization parameter. Roughly speaking, in nonoverlapping variants, spurious oscillations appear for both "too small" and "too large" stabilization parameters whereas, for the overlapping variant, "large" values of the stabilization parameter lead to a smearing of the discrete solution.

Since local projection stabilizations are not consistent, an important step in the error analysis is an estimation of the consistency error. It was demonstrated in [14] that, for stabilizations based on local projections of streamline derivatives, the consistency error generally deteriorates the convergence order if the stabilization parameter scales correctly with respect to $\mathbf{b}$. As a remedy, we propose to define the local stabilization terms using constant approximations of $\mathbf{b}$, which makes it possible to prove an optimal error estimate with respect to the norm used in the inf-sup condition. Moreover, in contrast to the analyses published before, it is not necessary to assume a higher (often unrealistic) regularity of $\mathbf{b}$.

The plan of the paper is as follows. In the next section, we formulate assumptions on approximation and projection spaces and define the local projection discretization investigated in this paper. Section 3 is devoted to the proof of the inf-sup condition and, in section 4, we derive an optimal error estimate. In section 5, we present examples of finite element spaces satisfying the assumptions of our theory. In particular, we show that standard finite element spaces can be used without the need of a mesh refinement or bubble enrichment. In section 6, we present our numerical results, and we close the paper by our conclusions in section 7. Throughout the paper, we use standard notation for Sobolev spaces and corresponding norms; see, e.g., [8]. Given a

measurable set $M \subset \mathbb{R}^d$, the inner product in $L^2(M)$ or $L^2(M)^d$ is denoted by $(\cdot,\cdot)_M$, and we use the notation $(\cdot,\cdot)$ instead of $(\cdot,\cdot)_\Omega$.

2. A local projection discretization. Given $h > 0$, let $W_h \subset H^1(\Omega)$ be a finite-dimensional space approximating the space $H^1(\Omega)$, and set $V_h = W_h \cap H_0^1(\Omega)$. Furthermore, let $\mathscr{M}_h$ be a set consisting of a finite number of open subsets M of Ω such that $\overline{\Omega} = \cup_{M\in\mathscr{M}_h} \overline{M}$. We assume that

$$\text{card}\{M' \in \mathscr{M}_h\,;\ M \cap M' \neq \emptyset\} \leq C_{\mathscr{M}} \qquad \forall\ M \in \mathscr{M}_h \tag{2.1}$$

and

$$h_M := \text{diam}(M) \leq C'_{\mathscr{M}}\, h \qquad \forall\ M \in \mathscr{M}_h\,, \tag{2.2}$$

where $C_{\mathscr{M}} \geq 1$ and $C'_{\mathscr{M}} \geq 1$ are constants independent of h. Moreover, we assume that, for any $M \in \mathscr{M}_h$, there is a nontrivial space $B_M \subset (W_h|_M) \cap H_0^1(M)$ such that $B_M \subset W_h$ if the functions from B_M are extended by zero outside M. For any $M \in \mathscr{M}_h$, we introduce a finite-dimensional space $D_M \subset L^2(M)$, and we assume that there exists a positive constant β_{LP} independent of h such that

$$\sup_{v\in B_M} \frac{(v,q)_M}{\|v\|_{0,M}} \geq \beta_{LP}\, \|q\|_{0,M} \qquad \forall\ q \in D_M,\ M \in \mathscr{M}_h\,. \tag{2.3}$$

We shall also need the inverse inequality

$$|v_h|_{1,M} \leq C_{inv}\, h_M^{-1}\, \|v_h\|_{0,M} \qquad \forall\ v_h \in W_h,\ M \in \mathscr{M}_h\,, \tag{2.4}$$

where C_{inv} is a constant independent of h.

We have in mind that W_h is a finite element space (see, e.g., [8]) and that the set $\mathscr{M}_h$ is constructed using the triangulation of Ω on which the space W_h is defined. Various possibilities of how the above assumptions can be satisfied will be presented in section 5. Note that the sets $M \in \mathscr{M}_h$ may overlap, which was not allowed in all formulations of the local projection method published up to now.

For any $M \in \mathscr{M}_h$, we denote by π_M a continuous linear projection operator which maps the space $L^2(M)$ onto the space D_M. We assume that

$$\|\pi_M\|_{\mathscr{L}(L^2(M),L^2(M))} \leq C_\pi \qquad \forall\ M \in \mathscr{M}_h\,,$$

where C_π is a constant independent of h. For example, π_M can be the orthogonal L^2 projection for which $C_\pi = 1$. For any $M \in \mathscr{M}_h$, we introduce the so-called fluctuation operator $\kappa_M = id - \pi_M$, where id is the identity operator on $L^2(M)$. Then

$$\|\kappa_M\|_{\mathscr{L}(L^2(M),L^2(M))} \leq C_\kappa \qquad \forall\ M \in \mathscr{M}_h\,, \tag{2.5}$$

where $C_\kappa = 1 + C_\pi$. An application of κ_M to a vector valued function means that κ_M is applied componentwise.

For any $M \in \mathscr{M}_h$, we choose a constant $\mathbf{b}_M \in \mathbb{R}^d$ such that

$$|\mathbf{b}_M| \leq \|\mathbf{b}\|_{0,\infty,M}\,, \qquad \|\mathbf{b} - \mathbf{b}_M\|_{0,\infty,M} \leq C_b\, h_M\, |\mathbf{b}|_{1,\infty,M} \tag{2.6}$$

with a constant C_b independent of h. In addition, we introduce a function $\widetilde{u}_{bh} \in W_h$ such that its trace approximates the boundary condition u_b.

The Galerkin solution of (1.1) is a function $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$a^G(u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,,$$

where

$$a^G(u, v) = \varepsilon\,(\nabla u, \nabla v) + (\mathbf{b}\cdot\nabla u, v) + (c\,u, v)\,.$$

If convection dominates diffusion, the Galerkin discretization has to be stabilized; cf., e.g., [19]. To this end we change the bilinear form a^G to

$$a_h^{LP}(u, v) = a^G(u, v) + s_h(u, v)\,,$$

where $s_h(u, v)$ is a local projection stabilization term given by

$$s_h(u, v) = \sum_{M\in\mathscr{M}_h} \tau_M\, s_M(u, v)\,,$$

τ_M are nonnegative stabilization parameters, and

$$s_M(u, v) = (\kappa_M(\mathbf{b}_M\cdot\nabla u), \kappa_M(\mathbf{b}_M\cdot\nabla v))_M\,. \tag{2.7}$$

Thus, the local projection discretization of (1.1) considered in this paper reads as follows.

Find $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$a_h^{LP}(u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h\,. \tag{2.8}$$

Introducing the norms

$$|||v|||_G = \Big(\varepsilon\,|v|_{1,\Omega}^2 + \|\sigma^{1/2}\,v\|_{0,\Omega}^2\Big)^{1/2}\,, \qquad |||v|||_{LP} = \big(|||v|||_G^2 + s_h(v, v)\big)^{1/2}\,,$$

we obtain

$$a^G(v, v) = |||v|||_G^2\,, \quad a_h^{LP}(v, v) = |||v|||_{LP}^2 \qquad \forall\ v \in H_0^1(\Omega)\,. \tag{2.9}$$

The latter shows that the local projection discretization (2.8) has a unique solution.

Remark 2.1. It was shown in [14] that, for nonoverlapping sets M, the stabilization parameters τ_M should satisfy

$$\tau_M \sim \min\left\{\frac{h_M}{\|\mathbf{b}\|_{0,\infty,M}}, \frac{h_M^2}{\varepsilon}\right\}\,. \tag{2.10}$$

Remark 2.2. A standard choice is to use $\mathbf{b}$ instead of $\mathbf{b}_M$ in (2.7). However, it was demonstrated in [14] that then it is generally not possible to obtain optimal convergence results if τ_M is chosen according to (2.10). We shall see in the next sections that the use of $\mathbf{b}_M$ leads to an optimal error estimate.

3. Stability of the local projection discretization. One of the most popular finite element techniques for the numerical solution of problem (1.1) is the streamline upwind/Petrov–Galerkin (SUPG) method proposed in [7]. An important feature of this method is that it provides stability with respect to the norm

$$|||v|||_{SUPG} = \Big(|||v|||_G^2 + \|\delta^{1/2}\,\mathbf{b}\cdot\nabla v\|_{0,\Omega}^2\Big)^{1/2}\,,$$

where δ is a stabilization parameter satisfying a relation of the type (2.10). In this section, we shall show that the local projection method has similar stability properties. More precisely, we shall prove that it is stable with respect to the norm

$$|||v|||_{LPSD} = \left(|||v|||_G^2 + s_h(v,v) + \sum_{M \in \mathscr{M}_h} \tau_M \, \|\mathbf{b} \cdot \nabla v\|_{0,M}^2 \right)^{1/2} .$$

The letters SD in the notation $|||\cdot|||_{LPSD}$ refer to the streamline derivative term.

THEOREM 3.1. *Let the stabilization parameters τ_M satisfy*

$$0 \le \tau_M \le C_\tau \, h_M^2 \left(\max\{\varepsilon, h_M \, \|\mathbf{b}\|_{0,\infty,M}\} \right)^{-1} \qquad \forall \; M \in \mathscr{M}_h \tag{3.1}$$

with a constant $C_\tau \ge 1$ independent of h and the data of (1.1). *Then the bilinear form a_h^{LP} satisfies*

$$\sup_{v_h \in V_h} \frac{a_h^{LP}(u_h, v_h)}{|||v_h|||_{LPSD}} \ge \beta \, |||u_h|||_{LPSD} \qquad \forall \; u_h \in V_h \, , \tag{3.2}$$

where β is a positive constant independent of h and ε.

Proof. Given $u_h \in V_h$, we shall construct a function $v_h \in V_h$ such that

$$a_h^{LP}(u_h, v_h) \ge |||u_h|||_{LPSD}^2 \qquad \text{and} \qquad |||u_h|||_{LPSD} \ge \beta \, |||v_h|||_{LPSD} \, . \tag{3.3}$$

Inequalities (3.3) immediately imply the inf-sup condition (3.2).

Consider any $M \in \mathscr{M}_h$. In view of the inf-sup conditions (2.3), there exists $z_M \in B_M$ such that (cf., e.g., [10])

$$(z_M, q)_M = \tau_M \, (\mathbf{b} \cdot \nabla u_h, q)_M \qquad \forall \; q \in D_M \, , \tag{3.4}$$

$$\|z_M\|_{0,M} \le \beta_{LP}^{-1} \, \tau_M \, \|\mathbf{b} \cdot \nabla u_h\|_{0,M} \, . \tag{3.5}$$

Consequently,

$$\begin{aligned} (z_M, \pi_M(\mathbf{b} \cdot \nabla u_h))_M &= \tau_M \, (\mathbf{b} \cdot \nabla u_h, \pi_M(\mathbf{b} \cdot \nabla u_h))_M \\ &= \tau_M \, \|\mathbf{b} \cdot \nabla u_h\|_{0,M}^2 - \tau_M \, (\mathbf{b} \cdot \nabla u_h, \kappa_M(\mathbf{b} \cdot \nabla u_h))_M, \end{aligned}$$

and hence

$$\begin{aligned} (z_M, \mathbf{b} \cdot \nabla u_h)_M = \tau_M \, \|\mathbf{b} \cdot \nabla u_h\|_{0,M}^2 &- \tau_M \, (\mathbf{b} \cdot \nabla u_h, \kappa_M(\mathbf{b} \cdot \nabla u_h))_M \\ &+ (z_M, \kappa_M(\mathbf{b} \cdot \nabla u_h))_M \, . \end{aligned}$$

Thus, denoting $z_h = \sum_{M \in \mathscr{M}_h} z_M$ (with $z_M = 0$ in $\Omega \setminus M$), we get

$$\begin{aligned} a_h^{LP}(u_h, z_h) = \sum_{M \in \mathscr{M}_h} \big\{ &\tau_M \, \|\mathbf{b} \cdot \nabla u_h\|_{0,M}^2 - \tau_M \, (\mathbf{b} \cdot \nabla u_h, \kappa_M(\mathbf{b} \cdot \nabla u_h))_M \\ &+ (z_M, \kappa_M(\mathbf{b} \cdot \nabla u_h))_M + \varepsilon \, (\nabla u_h, \nabla z_M)_M + (c \, u_h, z_M)_M \big\} + s_h(u_h, z_h). \end{aligned} \tag{3.6}$$

Using (3.5), (2.4), and (3.1), we derive for any $M \in \mathscr{M}_h$

$$\|z_M\|_{0,M} \le \beta_{LP}^{-1} \, \tau_M \, \|\mathbf{b}\|_{0,\infty,M} \, |u_h|_{1,M} \le C_\tau \, C_{inv} \, \beta_{LP}^{-1} \, \|u_h\|_{0,M}, \tag{3.7}$$

$$\begin{aligned} |z_M|_{1,M} &\le C_{inv} \, h_M^{-1} \, \|z_M\|_{0,M} \le C_{inv} \, h_M^{-1} \, \beta_{LP}^{-1} \, \tau_M \, \|\mathbf{b}\|_{0,\infty,M} \, |u_h|_{1,M} \\ &\le C_\tau \, C_{inv} \, \beta_{LP}^{-1} \, |u_h|_{1,M}, \end{aligned} \tag{3.8}$$

$$\varepsilon^{1/2} \, |z_M|_{1,M} \le \varepsilon^{1/2} \, C_{inv} \, h_M^{-1} \, \|z_M\|_{0,M} \le C_\tau^{1/2} \, C_{inv} \, \beta_{LP}^{-1} \, \tau_M^{1/2} \, \|\mathbf{b} \cdot \nabla u_h\|_{0,M}. \tag{3.9}$$

Moreover, it follows from the triangular inequality and (2.5), (2.6), (2.4), and (3.1) that

$$\begin{aligned}(3.10)\qquad \tau_M \left(\|\kappa_M(\mathbf{b}\cdot\nabla u_h)\|_{0,M} - \|\kappa_M(\mathbf{b}_M\cdot\nabla u_h)\|_{0,M}\right)^2 &\\ \le \tau_M \,\|\kappa_M((\mathbf{b}-\mathbf{b}_M)\cdot\nabla u_h)\|_{0,M}^2 &\\ \le C_\kappa^2\,\tau_M\,\|\mathbf{b}-\mathbf{b}_M\|_{0,\infty,M}^2\,|u_h|_{1,M}^2 &\\ \le 2\,C_\kappa^2\,C_b\,\tau_M\,h_M\,\|\mathbf{b}\|_{0,\infty,M}\,|\mathbf{b}|_{1,\infty,M}\,|u_h|_{1,M}^2 &\\ \le C_1\,\|u_h\|_{0,M}^2\,, &\end{aligned}$$

where $C_1 = 2\,C_\kappa^2\,C_b\,C_\tau\,C_{inv}^2\,|\mathbf{b}|_{1,\infty,\Omega}$. Applying the Schwarz inequality to the terms on the right-hand side of (3.6), using (3.5), (3.10), (3.9), and (3.7) and taking into account that $\beta_{LP}\le 1$, we deduce that

$$\begin{aligned}a_h^{LP}(u_h,z_h) \ge s_h(u_h,z_h) + \sum_{M\in\mathscr{M}_h}\tau_M\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M}^2 - C_2\sum_{M\in\mathscr{M}_h}\|\sigma^{1/2}\,u_h\|_{0,M}^2 &\\ - C_3\sum_{M\in\mathscr{M}_h}\tau_M^{1/2}\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M}\left(\varepsilon\,|u_h|_{1,M}^2 + \|\sigma^{1/2}\,u_h\|_{0,M}^2 + \tau_M\,s_M(u_h,u_h)\right)^{1/2} &\end{aligned}$$

with $C_2 = C_\tau\,C_{inv}\,\|c\|_{0,\infty,\Omega}\,\sigma_0^{-1}\,\beta_{LP}^{-1}$ and $C_3 = (2 + C_\tau^{1/2}\,C_{inv} + 2\,C_1^{1/2}\,\sigma_0^{-1/2})\,\beta_{LP}^{-1}$. In view of (2.1), we obtain

$$(3.11)\qquad \sum_{M\in\mathscr{M}_h}\|\sigma^{1/2}\,u_h\|_{0,M}^2 \le C_{\mathscr{M}}\,\|\sigma^{1/2}\,u_h\|_{0,\Omega}^2\,,\qquad \sum_{M\in\mathscr{M}_h}|u_h|_{1,M}^2 \le C_{\mathscr{M}}\,|u_h|_{1,\Omega}^2,$$

and hence, using the inequality $a\,b\le\frac{1}{4}\,a^2 + b^2$ valid for any $a,b\in\mathbb{R}$, we infer that

$$(3.12)\qquad a_h^{LP}(u_h,z_h) \ge s_h(u_h,z_h) + \frac{3}{4}\sum_{M\in\mathscr{M}_h}\tau_M\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M}^2 - C_4\,|||u_h|||_{LP}^2$$

with $C_4 = (C_2 + C_3^2)\,C_{\mathscr{M}}$. Now let us estimate the term $s_h(u_h,z_h)$. We have

$$(3.13)\qquad s_h(u_h,z_h) \le \sqrt{s_h(u_h,u_h)}\,\sqrt{s_h(z_h,z_h)} \le \sqrt{s_h(z_h,z_h)}\,|||u_h|||_{LP}\,.$$

Using (3.10) with z_h instead of u_h and applying (2.5), we get

$$(3.14)\qquad s_h(z_h,z_h) \le 2\,C_\kappa^2\sum_{M\in\mathscr{M}_h}\tau_M\,\|\mathbf{b}\cdot\nabla z_h\|_{0,M}^2 + 2\,C_1\sum_{M\in\mathscr{M}_h}\|z_h\|_{0,M}^2\,.$$

Furthermore, using (2.1), (2.4), (3.1), and (3.5), we derive

$$
\begin{aligned}
(3.15)\qquad \sum_{M\in\mathscr{M}_h} \tau_M \,\|\mathbf{b}\cdot\nabla z_h\|_{0,M}^2 &\le C_{\mathscr{M}} \sum_{M\in\mathscr{M}_h} \sum_{\substack{M'\in\mathscr{M}_h,\\ M\cap M'\neq\emptyset}} \tau_M \,\|\mathbf{b}\cdot\nabla z_{M'}\|_{0,M}^2 \\
&\le C_{\mathscr{M}} \sum_{\substack{M,\,M'\in\mathscr{M}_h,\\ M\cap M'\neq\emptyset}} \tau_M \,\|\mathbf{b}\|_{0,\infty,M\cap M'}^2 \,|z_{M'}|_{1,M\cap M'}^2 \\
&\le C_\tau\, C_{inv}\, C_{\mathscr{M}} \sum_{\substack{M,\,M'\in\mathscr{M}_h,\\ M\cap M'\neq\emptyset}} \|\mathbf{b}\|_{0,\infty,M\cap M'} \,\|z_{M'}\|_{0,M}\,|z_{M'}|_{1,M\cap M'} \\
&\le C_\tau\, C_{inv}^2\, C_{\mathscr{M}}^2 \sum_{M'\in\mathscr{M}_h} \|\mathbf{b}\|_{0,\infty,M'}\, h_{M'}^{-1}\,\|z_{M'}\|_{0,M'}^2 \\
&\le C_5^2 \sum_{M'\in\mathscr{M}_h} \tau_{M'}\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M'}^2\,,
\end{aligned}
$$

where $C_5 = C_\tau\, C_{inv}\, C_{\mathscr{M}}\, \beta_{LP}^{-1}$. Assumption (2.1) implies that

$$
\sum_{M\in\mathscr{M}_h} \|z_h\|_{0,M}^2 \le C_{\mathscr{M}} \sum_{\substack{M,\,M'\in\mathscr{M}_h,\\ M\cap M'\neq\emptyset}} \|z_{M'}\|_{0,M}^2 \le C_{\mathscr{M}}^2 \sum_{M'\in\mathscr{M}_h} \|z_{M'}\|_{0,M'}^2\,,
$$

which, in view of (3.7) and (3.11), gives

$$
(3.16)\qquad \|z_h\|_{0,\Omega} \le \left(\sum_{M\in\mathscr{M}_h} \|z_h\|_{0,M}^2\right)^{1/2} \le C_5\, C_{\mathscr{M}}^{1/2}\, \sigma_0^{-1/2}\,\|\sigma^{1/2}\,u_h\|_{0,\Omega}\,.
$$

Analogously, using (3.8) instead of (3.7), we derive

$$
(3.17)\qquad |z_h|_{1,\Omega} \le C_5\, C_{\mathscr{M}}^{1/2}\,|u_h|_{1,\Omega}\,.
$$

Substituting (3.15) and (3.16) into (3.14), we obtain

$$
(3.18)\qquad s_h(z_h,z_h) \le 2\,C_5^2\,C_\kappa^2 \sum_{M\in\mathscr{M}_h} \tau_M\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M}^2 + 2\,C_1\,C_5^2\,C_{\mathscr{M}}\,\sigma_0^{-1}\,|||u_h|||_{LP}^2\,.
$$

Combining this inequality with (3.13) and (3.12) and using once again the inequality $a\,b \le \frac{1}{4}\,a^2 + b^2$, we arrive at

$$
a_h^{LP}(u_h,z_h) \ge \frac{1}{2}\sum_{M\in\mathscr{M}_h} \tau_M\,\|\mathbf{b}\cdot\nabla u_h\|_{0,M}^2 - C_6\,|||u_h|||_{LP}^2
$$

with $C_6 = (C_2 + C_3^2)\,C_{\mathscr{M}} + 2\,C_5^2\,C_\kappa^2 + C_5\,(2\,C_1\,C_{\mathscr{M}}\,\sigma_0^{-1})^{1/2}$. Thus, employing (2.9), we see that $v_h \in V_h$ given by $v_h := 2\,z_h + (1 + 2\,C_6)\,u_h$ satisfies the first inequality in (3.3). The second inequality in (3.3) is a simple consequence of (3.15)–(3.18). ∎

4. Error analysis. In this section, we shall investigate the error of the solution of the local projection discretization (2.8) with respect to the norm $|||\cdot|||_{LPSD}$. Our considerations will be based on the following estimate which is similar to Strang's lemmas (see, e.g., [8]).

LEMMA 4.1. *Let $u \in H^1(\Omega)$ be the weak solution of (1.1), and let u_h be the solution of the local projection discretization (2.8). Then, under the assumption of Theorem 3.1, we have*

$$\beta\,|||u-u_h|||_{LPSD} \le \inf_{w_h\in W_h^b}\left\{\beta\,|||u-w_h|||_{LPSD} + \sup_{v_h\in V_h}\frac{a_h^{LP}(u-w_h,v_h)}{|||v_h|||_{LPSD}}\right\} + \sup_{v_h\in V_h}\frac{s_h(u,v_h)}{|||v_h|||_{LPSD}}\,, \tag{4.1}$$

where $W_h^b = \{w_h \in W_h\,;\ w_h - \widetilde{u}_{bh} \in V_h\}$.

Proof. The weak solution of (1.1) satisfies $a^G(u,v) = (f,v)$ for any $v \in H_0^1(\Omega)$. Therefore,

$$a_h^{LP}(u-u_h,v_h) = s_h(u,v_h) \qquad \forall\ v_h \in V_h,$$

and hence also

$$a_h^{LP}(w_h-u_h,v_h) = a_h^{LP}(w_h-u,v_h) + s_h(u,v_h) \qquad \forall\ v_h\in V_h,\ w_h\in W_h^b\,.$$

Now (4.1) follows by applying (3.2) and the triangular inequality. ∎

In the following two lemmas, we establish estimates of the terms on the right-hand side of (4.1).

LEMMA 4.2. *Let the stabilization parameters τ_M satisfy (3.1) and*

$$h_M\,\|\mathbf{b}\|_{0,\infty,M} \le C_\tau\,\max\{\varepsilon,\tau_M\,\|\mathbf{b}\|_{0,\infty,M}^2\} \qquad \forall\ M\in\mathscr{M}_h\,. \tag{4.2}$$

Then there exists a constant C independent of h and ε such that, for any $w\in H^1(\Omega)$,

$$|||w|||_{LPSD} + \sup_{v\in H_0^1(\Omega)}\frac{a_h^{LP}(w,v)}{|||v|||_{LPSD}} \le C\,(\varepsilon + h\,\|\mathbf{b}\|_{0,\infty,\Omega} + h^2\,\|\sigma\|_{0,\infty,\Omega})^{1/2}\left(\sum_{M\in\mathscr{M}_h}\{|w|_{1,M}^2 + h_M^{-2}\,\|w\|_{0,M}^2\}\right)^{1/2}.$$

Proof. Consider any $w\in H^1(\Omega)$ and $v\in H_0^1(\Omega)$. Integrating by parts and using the Schwarz inequality, we obtain

$$\begin{aligned} a_h^{LP}(w,v) &= \varepsilon\,(\nabla w,\nabla v) - (w,\mathbf{b}\cdot\nabla v) + ((c-\operatorname{div}\mathbf{b})\,w,v) + s_h(w,v)\\ &\le -(w,\mathbf{b}\cdot\nabla v) + (1+\sigma_0^{-1}\,\|c-\operatorname{div}\mathbf{b}\|_{0,\infty,\Omega})\,|||w|||_{LP}\,|||v|||_{LP}\,. \end{aligned}$$

Furthermore, in view of (2.5), (2.6), (3.1), and (2.2), we get

$$|||w|||_{LPSD}^2 \le \|\sigma\|_{0,\infty,\Omega}\,\|w\|_{0,\Omega}^2 + 2\,C_\kappa^2\,C_\tau\,C'_{\mathscr{M}}\,(\varepsilon + h\,\|\mathbf{b}\|_{0,\infty,\Omega})\sum_{M\in\mathscr{M}_h}|w|_{1,M}^2\,.$$

It remains to estimate the term $(w,\mathbf{b}\cdot\nabla v)$. We have

$$|(w,\mathbf{b}\cdot\nabla v)| \le \left(\sum_{M\in\mathscr{M}_h} h_M^{-2}\,\|w\|_{0,M}^2\right)^{1/2}\left(\sum_{M\in\mathscr{M}_h} h_M^2\,\|\mathbf{b}\cdot\nabla v\|_{0,M}^2\right)^{1/2}.$$

Consider any $M \in \mathscr{M}_h$. If $h_M \|\mathbf{b}\|_{0,\infty,M} \le C_\tau\,\varepsilon$, then $h_M^2 \|\mathbf{b}\cdot\nabla v\|_{0,M}^2 \le C_\tau^2\,\varepsilon^2\,|v|_{1,M}^2$, and hence also

$$h_M^2 \|\mathbf{b}\cdot\nabla v\|_{0,M}^2 \le C_\tau^2\,(\varepsilon + h_M\|\mathbf{b}\|_{0,\infty,M})\,(\varepsilon\,|v|_{1,M}^2 + \tau_M\|\mathbf{b}\cdot\nabla v\|_{0,M}^2)\,. \tag{4.3}$$

If $h_M \|\mathbf{b}\|_{0,\infty,M} > C_\tau\,\varepsilon$, then $h_M \le C_\tau\,\tau_M\|\mathbf{b}\|_{0,\infty,M}$ due to (4.2), and hence $h_M^2 \le C_\tau^2\,h_M\|\mathbf{b}\|_{0,\infty,M}\,\tau_M$ (since $C_\tau \ge 1$). Therefore, (4.3) holds also in this case, and we deduce using (2.1) and (2.2) that

$$\sum_{M\in\mathscr{M}_h} h_M^2\|\mathbf{b}\cdot\nabla v\|_{0,M}^2 \le C_\tau^2\,C_{\mathscr{M}}\,C'_{\mathscr{M}}\,(\varepsilon + h\|\mathbf{b}\|_{0,\infty,\Omega})\,|||v|||_{LPSD}^2\,,$$

which completes the proof. ∎

LEMMA 4.3. *Let the stabilization parameters τ_M satisfy* (3.1). *Then, for any $u \in H^1(\Omega)$, we have*

$$\sup_{v\in H^1(\Omega)} \frac{s_h(u,v)}{|||v|||_{LPSD}} \le C\,h^{1/2}\,\|\mathbf{b}\|_{0,\infty,\Omega}^{1/2}\left(\sum_{M\in\mathscr{M}_h}\inf_{\mathbf{q}_M\in[D_M]^d}\|\nabla u - \mathbf{q}_M\|_{0,M}^2\right)^{1/2},$$

where $C = C_\kappa\,(C_\tau\,C'_{\mathscr{M}})^{1/2}$.

Proof. For any $u, v \in H^1(\Omega)$, we have

$$s_h(u,v) \le \sqrt{s_h(u,u)}\,\sqrt{s_h(v,v)} \le \sqrt{s_h(u,u)}\,|||v|||_{LPSD},$$

and hence it suffices to estimate $\tau_M\,s_M(u,u)$ with an arbitrary $M \in \mathscr{M}_h$. Consider any $\mathbf{q}_M \in [D_M]^d$. Since $\kappa_M\,\mathbf{q}_M = 0$, we obtain using (2.5) and (2.6)

$$s_M(u,u) \le |\mathbf{b}_M|^2\,\|\kappa_M(\nabla u - \mathbf{q}_M)\|_{0,M}^2 \le C_\kappa^2\,\|\mathbf{b}\|_{0,\infty,M}^2\,\|\nabla u - \mathbf{q}_M\|_{0,M}^2\,.$$

Therefore, applying (3.1) and (2.2), we obtain

$$\tau_M\,s_M(u,u) \le C_\kappa^2\,C_\tau\,C'_{\mathscr{M}}\,h\,\|\mathbf{b}\|_{0,\infty,\Omega}\,\|\nabla u - \mathbf{q}_M\|_{0,M}^2\,,$$

which proves the lemma. ∎

To prove convergence results for the solution of the local projection discretization (2.8), we have to introduce some approximation properties of the spaces W_h and D_M. We shall assume that there exist interpolation operators $i_h \in \mathscr{L}(H^2(\Omega), W_h) \cap \mathscr{L}(H^2(\Omega)\cap H_0^1(\Omega), V_h)$ and $j_M \in \mathscr{L}(H^1(M), D_M)$, $M \in \mathscr{M}_h$, such that, for some constants $l \in \mathbb{N}$ and $C > 0$, we have

$$\left(\sum_{M\in\mathscr{M}_h}\{|v - i_h v|_{1,M}^2 + h_M^{-2}\,\|v - i_h v\|_{0,M}^2\}\right)^{1/2} \le C\,h^k\,|v|_{k+1,\Omega} \qquad \forall\, v \in H^{k+1}(\Omega),\ k = 1,\dots,l, \tag{4.4}$$

and

$$\|q - j_M q\|_{0,M} \le C\,h_M^k\,|q|_{k,M} \qquad \forall\, q \in H^k(M),\ M \in \mathscr{M}_h,\ k = 1,\dots,l\,. \tag{4.5}$$

Now we are in a position to prove an a priori error estimate for the local projection discretization (2.8).

THEOREM 4.4. *Let the stabilization parameters τ_M satisfy* (3.1) *and* (4.2). *Let the spaces W_h and D_M possess the approximation properties* (4.4) *and* (4.5). *Let the weak solution of* (1.1) *satisfy $u \in H^{k+1}(\Omega)$ for some $k \in \{1,\dots,l\}$. Finally, let $\widetilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\widetilde{u}_{bh} = i_h \widetilde{u}_b$. Then the solution u_h of the local projection discretization* (2.8) *satisfies the error estimate*

$$|||u-u_h|||_{LPSD} \le C\,(\varepsilon + h\,\|\mathbf{b}\|_{0,\infty,\Omega} + h^2\,\|\sigma\|_{0,\infty,\Omega})^{1/2}\,h^k\,|u|_{k+1,\Omega}\,, \tag{4.6}$$

where the constant C is independent of h and ε.

Proof. Combining Lemmas 4.1–4.3 and setting $w_h = i_h u$ and $\mathbf{q}_M = j_M(\nabla u)$, $M \in \mathscr{M}_h$, the theorem follows by applying (4.4), (4.5), (2.1), and (2.2). ∎

Remark 4.5. Estimates of the type (4.6) can be proved for various stabilized finite element methods applied to problem (1.1) (e.g., the SUPG method) and are known to be optimal; see, e.g., [19]. If we define the stabilization term $s_h(u,v)$ using $\mathbf{b}$ instead of $\mathbf{b}_M$, then Theorem 3.1 and Lemmas 4.1 and 4.2 still hold, but the consistency error cannot be estimated as in Lemma 4.3. Assuming (3.1) and $\mathbf{b}\cdot\nabla u \in H^k(\Omega)$ with $k \in \{1,\dots,l\}$, we obtain

$$\sup_{v_h\in V_h} \frac{s_h(u,v_h)}{|||v_h|||_{LPSD}} \le C\,h^k \left(\sum_{M\in\mathscr{M}_h} \min\left\{ \frac{|\mathbf{b}\cdot\nabla u|^2_{k,M}}{\sigma_0}, \frac{h_M\,|\mathbf{b}\cdot\nabla u|^2_{k,M}}{\|\mathbf{b}\|_{0,\infty,M}} \right\} \right)^{1/2};$$

see [14, 15]. Thus, if $\mathbf{b} \neq \mathbf{0}$ in $\overline{\Omega}$, the optimal convergence order still can be proved, but if $\mathbf{b}$ is allowed to vanish, we have only the suboptimal convergence order k in general. Moreover, for small σ_0, the accuracy of the discrete solution may be significantly worse than for s_h defined using $\mathbf{b}_M$; see also Example 6.1 in section 6.

Remark 4.6. The assumptions (3.1) and (4.2) are fulfilled for τ_M satisfying (2.10). Another possibility is to use $\tau_M \sim h_M/\|\mathbf{b}\|_{0,\infty,M}$ if $h_M\,\|\mathbf{b}\|_{0,\infty,M} \gtrsim \varepsilon$ and $\tau_M = 0$ otherwise.

5. Examples of spaces W_h and D_M. In this section, we present several examples of spaces W_h and D_M satisfying the assumptions made in sections 2 and 4. For ease of exposition, we confine ourselves to the two-dimensional case. In one or three dimensions, the spaces can be constructed analogously.

First let us recall some basic notions (cf. [6, 8, 10]) which will be used in the following. Given $h > 0$, a set $\mathscr{T}_h$ will be called a triangulation of Ω if it consists of a finite number of mutually disjoint open subsets T of Ω such that $\overline{\Omega} = \cup_{T\in\mathscr{T}_h}\overline{T}$ and $h_T := \mathrm{diam}(T) \le h$ for any $T \in \mathscr{T}_h$. We shall assume that either all elements of $\mathscr{T}_h$ are triangles or all elements of $\mathscr{T}_h$ are convex quadrilaterals, and that the intersection of the closures of any two different elements of $\mathscr{T}_h$ either is empty or consists of a common vertex or a common edge of these two elements. A triangulation $\mathscr{T}_h$ consisting of triangles is shape-regular if

$$\frac{h_T}{\varrho_T} \le C_{\mathscr{T}} \qquad \forall\, T \in \mathscr{T}_h\,, \tag{5.1}$$

where ϱ_T is the diameter of the largest circle inscribed in $\overline{T}$ and $C_{\mathscr{T}}$ is a constant independent of h which is common to the considered family of triangulations. If $\mathscr{T}_h$ consists of quadrilaterals, then it is shape-regular if (5.1) holds for any triangle whose vertices coincide with three vertices of some element of $\mathscr{T}_h$. We denote by $\widehat{T}$ a reference element, which is either a triangle or a square, depending on the type of

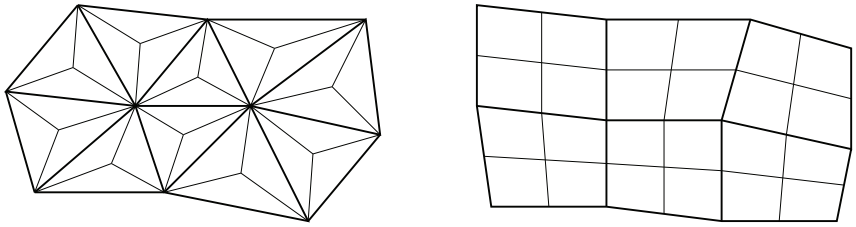


FIG. 5.1. *Relation between the meshes $\mathcal{M}_h$ (bold lines) and $\mathcal{T}_h$ (bold and fine lines) in the two-level method.*

elements in $\mathscr{T}_h$. If $\widehat{T}$ is a triangle, we set $R_l(\widehat{T}) := P_l(\widehat{T})$, where $P_l(\widehat{T})$ is the space of polynomials on $\widehat{T}$ of degree $\leq l$. If $\widehat{T}$ is a square, then we set $R_l(\widehat{T}) := Q_l(\widehat{T})$, where $Q_l(\widehat{T})$ is the space of polynomials on $\widehat{T}$ of degree $\leq l$ in each variable. For any $T \in \mathscr{T}_h$, there exists a one-to-one mapping $F_T \in [R_1(\widehat{T})]^2$ such that $F_T(\widehat{T}) = T$. Given $l \in \mathbb{N}$, we define the space

$$X_{\mathscr{T}_h,l} = \{v \in C(\overline{\Omega})\,;\ v \circ F_T \in R_l(\widehat{T})\ \ \forall\, T \in \mathscr{T}_h\}\,. \tag{5.2}$$

Then $X_{\mathscr{T}_h,l}$ does not depend on the choice of the mappings F_T, we have $X_{\mathscr{T}_h,l} \subset H^1(\Omega)$, and $X_{\mathscr{T}_h,l}$ satisfies the inverse inequality

$$|v_h|_{1,T} \leq C\, h_T^{-1}\, \|v_h\|_{0,T} \qquad \forall\, v_h \in X_{\mathscr{T}_h,l},\, T \in \mathscr{T}_h\,. \tag{5.3}$$

Moreover, the Lagrange interpolation operator $i_{\mathscr{T}_h,l} \in \mathscr{L}(C(\overline{\Omega}), X_{\mathscr{T}_h,l})$ satisfies $i_{\mathscr{T}_h,l} \in \mathscr{L}(C(\overline{\Omega}) \cap H_0^1(\Omega), X_{\mathscr{T}_h,l} \cap H_0^1(\Omega))$ and

$$\begin{aligned} |v - i_{\mathscr{T}_h,l}\, v|_{1,T} + h_T^{-1}\, \|v - i_{\mathscr{T}_h,l}\, v\|_{0,T} &\leq C\, h_T^k\, |v|_{k+1,T} \\ &\forall\, v \in H^{k+1}(T),\ T \in \mathscr{T}_h,\ k = 1, \ldots, l\,. \end{aligned} \tag{5.4}$$

Finally, denoting by $j_{T,l}$ the orthogonal projection of $L^2(T)$ onto $P_l(T)$, we also have

$$\|q - j_{T,l} q\|_{0,T} \leq C\, h_T^k\, |q|_{k,T} \qquad \forall\, q \in H^k(T),\ T \in \mathscr{T}_h,\ k = 0, \ldots, l+1\,. \tag{5.5}$$

In all the inequalities (5.3)–(5.5), the constant C depends only on l and $C_{\mathscr{T}}$ from (5.1).

Now let us discuss the construction of the spaces W_h and D_M. The original local projection stabilization [1, 2] was designed as a two-level method. Given a shape-regular triangulation of Ω, the elements of this triangulation are considered as the set $\mathscr{M}_h$ introduced in section 2. Then this triangulation is refined as depicted in Figure 5.1; i.e., each triangle is divided into three triangles by connecting its vertices with the barycenter, and each quadrilateral is divided into four quadrilaterals by connecting midpoints of opposite edges. Let us denote the resulting triangulation by $\mathscr{T}_h$. Given $l \in \mathbb{N}$, we set

$$W_h = X_{\mathscr{T}_h,l}\,, \qquad D_M = P_{l-1}(M) \quad \forall\, M \in \mathcal{M}_h\,. \tag{5.6}$$

Then in view of Lemma 5.1 and according to what was said above, all the assumptions of section 2 as well as (4.4) and (4.5) are satisfied (note that $X_{\mathscr{M}_h,l} \subset X_{\mathscr{T}_h,l}$ and that triangulations $\mathscr{T}_h$ assigned to a shape-regular family $\{\mathscr{M}_h\}$ also form a shape-regular family).

Another choice of the spaces W_h and D_M (a one-level method) was proposed in [17]. Here, given a shape-regular triangulation of Ω, the elements of this triangulation

are again considered as the set $\mathscr{M}_h$, but the space W_h is constructed on this triangulation $\mathscr{M}_h$ as well. However, the spaces W_h and D_M defined by (5.6) with $\mathscr{T}_h = \mathscr{M}_h$ do not satisfy the inf-sup conditions (2.3) in general. Indeed, the validity of the inf-sup conditions (2.3) would imply that $\dim B_M \geq \dim D_M$, but this cannot be satisfied if M is a triangle or $l < 5$. Therefore, the space $X_{\mathscr{M}_h,l}$ is enriched elementwise by bubble functions. More precisely, introducing a polynomial bubble function $\widehat{\varphi} \in H_0^1(\widehat{T}) \setminus \{0\}$ (cubic if $\widehat{T}$ is a triangle and biquadratic if $\widehat{T}$ is a square), we set

$$W_h = \{v \in C(\overline{\Omega});\ v \circ F_M \in R_l(\widehat{T}) + \widehat{\varphi} \cdot R_{l-1}(\widehat{T})\ \forall\, M \in \mathscr{M}_h\}\,, \tag{5.7}$$

$$D_M = P_{l-1}(M) \qquad \forall\, M \in \mathcal{M}_h\,. \tag{5.8}$$

Then the inf-sup conditions (2.3) hold with $B_M = (W_h|_M) \cap H_0^1(M)$; see [17] or the proof of Lemma 5.1. The remaining assumptions of section 2 as well as (4.4) and (4.5) are clearly satisfied as well.

There are also other possibilities to define the spaces W_h and D_M in both the one-level and two-level frameworks; see [17] for details. A common feature of all these constructions is that they lead to a (significant) increase of the number of degrees of freedom in comparison with applying, e.g., a residual-based stabilization [19] for which we could simply use a finite element space consisting of piecewise polynomials of degree l (in the sense of (5.2)) on a given triangulation. This increase of the number of degrees of freedom, either due to a refinement of the given triangulation or due to an enrichment by bubble functions, is a consequence of the fact that the sets in $\mathcal{M}_h$ are assumed to be nonoverlapping. We shall demonstrate in the following that our theory, which enables us to use sets $\mathcal{M}_h$ consisting of overlapping subsets of Ω, makes it possible to satisfy the assumptions on the spaces W_h and D_M without introducing additional degrees of freedom.

Let $\mathscr{T}_h$ be a shape-regular triangulation of Ω. We shall assume that any element of $\mathscr{T}_h$ has at least one vertex in Ω. Let $x_1, \ldots, x_{N_h}$ be the vertices of $\mathscr{T}_h$ lying in Ω, and let us denote

$$M_i = \operatorname{int} \bigcup_{T \in \mathscr{T}_h,\, x_i \in \overline{T}} \overline{T}\,, \qquad i = 1, \ldots, N_h\,, \tag{5.9}$$

where "int" denotes the interior of the respective polygon. We set

$$\mathscr{M}_h = \{M_i\}_{i=1}^{N_h}\,. \tag{5.10}$$

Then we can define the spaces W_h and D_M as in the two-level method by (5.6). Let us emphasize once more that we use only the triangulation $\mathscr{T}_h$ we were given at the beginning.

Let us discuss the validity of the assumptions on $\mathscr{M}_h$, W_h, and D_M made in this paper. Since the number of elements of $\mathscr{T}_h$ sharing a common vertex is bounded by a constant depending only on $C_{\mathscr{T}}$ from (5.1), assumption (2.1) is satisfied. Moreover, the shape-regularity of $\mathscr{T}_h$ implies that

$$C_{\mathscr{M},\mathscr{T}}\, h_M \leq h_T \leq h_M \qquad \forall\, T \in \mathscr{T}_h,\, M \in \mathscr{M}_h,\, T \cap M \neq \emptyset\,, \tag{5.11}$$

where $C_{\mathscr{M},\mathscr{T}}$ is a positive constant again depending only on $C_{\mathscr{T}}$ from (5.1). The assumption (2.2) obviously holds with $C'_{\mathscr{M}} = 2$. The validity of (2.4) and (4.4) is a direct consequence of (5.3), (5.4), and (5.11). In view of (5.1) and (5.11), any set M_i is star-shaped with respect to the ball with the center x_i and diameter $h_{M_i} C_{\mathscr{M},\mathscr{T}} / C_{\mathscr{T}}$,

and hence (4.5) holds as well; see, e.g., [6]. Finally, the inf-sup conditions (2.3) hold according to the following lemma, and hence all the assumptions on $\mathscr{M}_h$, W_h, and D_M are satisfied.

LEMMA 5.1. *Let $\mathscr{T}_h$ be a triangulation of Ω and let $\mathscr{M}_h$ be given by* (5.9) *and* (5.10). *Consider any $l \in \mathbb{N}$. Then the spaces W_h and D_M defined by* (5.6) *satisfy the* inf-sup *conditions* (2.3) *with $B_M = (W_h|_M) \cap H_0^1(M)$ and a constant β_{LP} depending only on l.*

Proof. Consider any $M \in \mathscr{M}_h$, and let $i \in \{1, \dots, N_h\}$ be such that $M = M_i$. Let $\varphi \in X_{\mathscr{T}_h,1} \cap H_0^1(\Omega)$ satisfy $\varphi(x_i) = 1$ and $\varphi(x_j) = 0$ for any $j \neq i$, $j = 1, \dots, N_h$. For any $T \in \mathscr{T}_h$ such that $T \subset M$, we set $\widehat{\varphi}_T := \varphi \circ F_T$ and $\widehat{J}_T = \det DF_T$, where DF_T is the Jacobi matrix of F_T. Note that $\widehat{\varphi}_T \in R_1(\widehat{T})$ equals 1 at one vertex of $\widehat{T}$ and vanishes at the remaining vertices of $\widehat{T}$ so that $\widehat{\varphi}_T$ is one of three (resp., four) fixed functions on $\widehat{T}$ in the triangular (resp., quadrilateral) case. Setting $\|\widehat{u}\|_* = \|\widehat{\varphi}_T\, \widehat{u}\|_{0,1,\widehat{T}}$, the functional $\|\cdot\|_*$ is a norm on $L^1(\widehat{T})$, and we have

$$\|\widehat{u}\|_* \geq C_l \, \|\widehat{u}\|_{0,1,\widehat{T}} \qquad \forall\, \widehat{u} \in R_l(\widehat{T}) \tag{5.12}$$

with $C_l > 0$ due to the equivalence of norms on finite-dimensional spaces. Now consider any $q \in D_M$. Since $\varphi|_M \in H_0^1(M)$ and $\widehat{q}_T := q \circ F_T \in R_{l-1}(\widehat{T})$ for any $T \in \mathscr{T}_h$ such that $T \subset M$, the function $v := \varphi\, q$ is an element of B_M. Using (5.12), we derive for any $T \subset M$

$$(v, q)_T = \|\varphi\, q^2\|_{0,1,T} = \|\widehat{q}_T^2\, \widehat{J}_T\|_* \geq C_{2l-1} \, \|\widehat{q}_T^2\, \widehat{J}_T\|_{0,1,\widehat{T}} = C_{2l-1} \, \|q\|_{0,T}^2 \,.$$

Moreover, $\|\varphi\|_{0,\infty,M} = 1$ and hence $\|v\|_{0,M} \leq \|q\|_{0,M}$. Thus, (2.3) holds with $\beta_{LP} = C_{2l-1}$. ∎

Remark 5.2. The assumption that any element of $\mathscr{T}_h$ has at least one vertex in Ω is satisfied for common quadrilateral and hexahedral meshes. For simplicial meshes, this assumption can be violated by elements of $\mathscr{T}_h$ lying at vertices of Ω. Since these are typically only a few elements, the validity of the mentioned assumption does not seem to be important for practical computations, at least for low order finite elements.

Remark 5.3. If the sets $M \in \mathcal{M}_h$ are defined as patches of elements of $\mathcal{T}_h$, the stencil of the stiffness matrix corresponding to the local projection discretization (2.8) may be considerably larger than for the Galerkin discretization. Let us demonstrate this for a rectangular triangulation $\mathcal{T}_h$ of a two-dimensional domain Ω. Let $\mathbb{A}^{Gal}$ be the stiffness matrix of the Galerkin discretization with $W_h = X_{\mathscr{T}_h,2}$, and let $\mathbb{A}^{LP2}$ be the stiffness matrix of the two-level local projection discretization with $\mathcal{M}_h$ consisting of nonoverlapping patches of always four elements of $\mathscr{T}_h$ (cf. Figure 5.1) and spaces W_h and D_M defined by (5.6) with $l = 2$. Furthermore, let $\mathbb{A}^{LPo}$ be the stiffness matrix of (2.8) with $\mathcal{M}_h$ consisting of overlapping patches of always four elements of $\mathscr{T}_h$ given by (5.9) and (5.10) and with spaces W_h and D_M again defined by (5.6) with $l = 2$. Then, denoting by $\#\mathbb{A}$ the number of nonzero entries of a matrix $\mathbb{A}$, we have

$$\#\mathbb{A}^{LP2} \approx \frac{9}{4}\, \#\mathbb{A}^{Gal}, \qquad \#\mathbb{A}^{LPo} \approx 4\, \#\mathbb{A}^{Gal}, \tag{5.13}$$

whereas the number of degrees of freedom is the same in all three cases. This is a drawback of the local projection stabilization in comparison to residual-based stabilizations for which the sparsity pattern of the stiffness matrix is the same as that for the Galerkin method. In order to save computer memory and avoid implementational

difficulties, an often used approach is to store only the matrix $\widetilde{\mathbb{A}}^{LP}$ corresponding to the bilinear form

$$\widetilde{a}_h^{LP}(u,v) = a^G(u,v) + \sum_{M\in\mathscr{M}_h} \tau_M \left(\mathbf{b}_M\cdot\nabla u, \mathbf{b}_M\cdot\nabla v\right)_M ,$$

which has the same sparsity pattern as the Galerkin stiffness matrix. If the operators π_M are orthogonal L^2 projections of $L^2(M)$ onto D_M, we have $a_h^{LP} = \widetilde{a}_h^{LP} - \widetilde{s}_h$ with

$$\widetilde{s}_h(u,v) = \sum_{M\in\mathscr{M}_h} \tau_M \left(\pi_M(\mathbf{b}_M\cdot\nabla u), \mathbf{b}_M\cdot\nabla v\right)_M .$$

An application of the matrix corresponding to $\widetilde{s}_h$ is not implemented explicitly but obtained implicitly by evaluating this bilinear form. This is sufficient for computing the solution of (2.8) by an iterative method, e.g., the generalized minimal residual (GMRES) method. The matrix $\widetilde{\mathbb{A}}^{LP}$ or some spectral equivalent approximation of $\widetilde{\mathbb{A}}^{LP}$ can be used as a preconditioner for this iterative method; see [1], where this solution strategy is thoroughly discussed. This approach is very efficient for transient problems since then a good initial guess of the discrete solution is available at each time step. Let us mention that the comparison in (5.13) does not take into account the fact that, for the nonoverlapping variant of the local projection stabilization, a given triangulation has to be refined in general before defining the space W_h so that then the overlapping variant leads to lower computational cost.

For the one-level method with $\mathscr{M}_h = \mathscr{T}_h$ and spaces W_h and D_M defined by (5.7) and (5.8) with $l = 2$, the corresponding stiffness matrix $\mathbb{A}^{LP1}$ satisfies $\#\mathbb{A}^{LP1} \approx (127/64)\,\#\mathbb{A}^{Gal}$. In this case, the local projections do not influence the sparsity pattern of the stiffness matrix, but the higher number of nonzero entries of the stiffness matrix is caused by the increase in the number of degrees of freedom (we have $\dim W_h \approx (7/4)\dim X_{\mathscr{T}_h,2}$). If we use $P_{l-1}(\widehat{T})$ instead of $R_{l-1}(\widehat{T})$ in (5.7), then $\#\mathbb{A}^{LP1} \approx (104/64)\,\#\mathbb{A}^{Gal}$ and $\dim W_h \approx (3/2)\dim X_{\mathscr{T}_h,2}$.

6. Numerical results. In this section, we present numerical results for three test problems illustrating the properties of the methods discussed in the preceding sections. In all computations, the operators π_M are orthogonal L^2 projections of $L^2(M)$ onto D_M. The constant approximations $\mathbf{b}_M$ of $\mathbf{b}$ in M are defined as values of $\mathbf{b}$ at barycenters of M if M are triangles or quadrilaterals. For $M = M_i$ defined by (5.9) we set $\mathbf{b}_{M_i} = \mathbf{b}(x_i)$. These choices of $\mathbf{b}_M$ assure the validity of (2.6). Since local projection methods with spaces W_h of first order approximation properties provide solutions similar to those of the SUPG method if the stabilization parameters are defined appropriately (cf. [12, 15, 18]), we concentrate on second order spaces for which the SUPG method cannot be recovered in general.

Let us first consider the following example showing that it is really important to use $\mathbf{b}_M$ instead of $\mathbf{b}$ in the local projection stabilization term (2.7).

Example 6.1 (problem without layers). We consider the problem (1.1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-12}$, $\mathbf{b}(x,y) = (0,x^2)$, and $c = 10^{-5}$. The functions f and u_b are such that the solution of (1.1) is $u(x,y) = \sin(x+y)$.

We consider triangulations $\mathcal{T}_h$ of the type depicted on the left in Figure 6.1 and set $\mathcal{M}_h = \mathcal{T}_h$. The spaces W_h and D_M are defined by (5.7) and (5.8) with $l = 2$. The stabilization parameters τ_M are defined simply by the right-hand side of (2.10). Table 6.1 shows errors of the solutions of the local projection method (2.8) for various values of h. The errors are measured in the (semi)norms $|||\cdot|||_{LP}$, $|||\cdot|||_{SUPG}$, $\|\cdot\|_{0,\Omega}$,

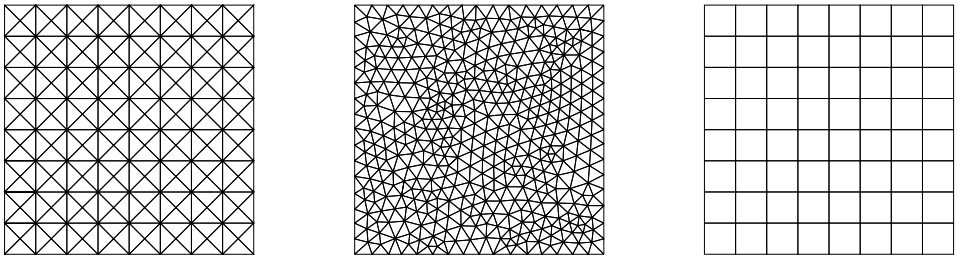


FIG. 6.1. *Triangulations used for computations presented in section* 6.

TABLE 6.1
Example 6.1, *errors for the local projection method* (2.8).

h	$\|\|\|\cdot\|\|\|_{LP}$	$\|\|\|\cdot\|\|\|_{SUPG}$	$\|\|\cdot\|\|_{0,\Omega}$	$\|\cdot\|_{1,\Omega}$	$\|\|\cdot\|\|_{0,\infty,h}$
6.25−2	7.54−6	1.02−4	1.59−5	1.70−3	4.63−5
3.13−2	1.37−6	2.01−5	2.33−6	4.74−4	6.66−6
1.56−2	2.47−7	3.90−6	3.33−7	1.30−4	9.15−7
7.81−3	4.47−8	7.46−7	4.52−8	3.46−5	1.21−7
Conv. order	2.47	2.39	2.88	1.91	2.92

TABLE 6.2
Example 6.1, *errors for the local projection method* (2.8) *with* $\mathbf{b}_M$ *replaced by* $\mathbf{b}$.

h	$\|\|\|\cdot\|\|\|_{LP}$	$\|\|\|\cdot\|\|\|_{SUPG}$	$\|\|\cdot\|\|_{0,\Omega}$	$\|\cdot\|_{1,\Omega}$	$\|\|\cdot\|\|_{0,\infty,h}$
6.25−2	6.68−5	4.46−4	7.07−4	8.56−2	6.06−3
3.13−2	1.65−5	1.17−4	2.55−4	6.23−2	3.11−3
1.56−2	4.13−6	3.05−5	8.91−5	4.41−2	1.57−3
7.81−3	1.04−6	7.88−6	2.95−5	3.02−2	7.85−4
Conv. order	1.99	1.95	1.60	0.55	1.00

$|\cdot|_{1,\Omega}$, and $\|\cdot\|_{0,\infty,h}$, where the SUPG norm is computed with $\delta|_M = \tau_M$ for any $M \in \mathcal{M}_h$ and the discrete L^∞ norm $\|\cdot\|_{0,\infty,h}$ is defined as the maximum absolute value at the points of principal lattices of order 6 on the elements of $\mathcal{T}_h$; see [8]. The convergence orders are computed from the errors on the two finest meshes. The notation $r{-}n$ used in the table means $r\cdot 10^{-n}$. We observe that the convergence orders of errors in the norms $|||\cdot|||_{LP}$ and $|||\cdot|||_{SUPG}$ are near to the value 2.5 predicted by Theorem 4.4. For the (semi)norms $\|\cdot\|_{0,\Omega}$, $|\cdot|_{1,\Omega}$, and $\|\cdot\|_{0,\infty,h}$, the convergence orders are nearly optimal as well. On the other hand, if we use $\mathbf{b}$ instead of $\mathbf{b}_M$ in (2.7), then, as we can see from Table 6.2, all convergence orders are suboptimal, and the errors of the discrete solutions are much larger than in the previous case. The convergence orders with respect to the norms $|||\cdot|||_{LP}$ and $|||\cdot|||_{SUPG}$ are in agreement with Remark 4.5.

Example 6.2 (problem with two interior layers). We consider the problem (1.1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-7}$, $\mathbf{b}(x,y) = (-y,x)$, $c = 0$, and $f = 0$. We set $u_b(x,0) = 1$ for $x \in (\frac{1}{3},\frac{2}{3})$ and $u_b(x,y) = 0$ elsewhere. Moreover, we do not use the Dirichlet boundary condition at the outflow boundary $(0,1)\times\{1\}$, where we prescribe a homogeneous Neumann boundary condition.

The solution of this example exhibits two interior layers starting from the discontinuities of the inflow profile at $y = 0$. We shall consider the unstructured triangulation

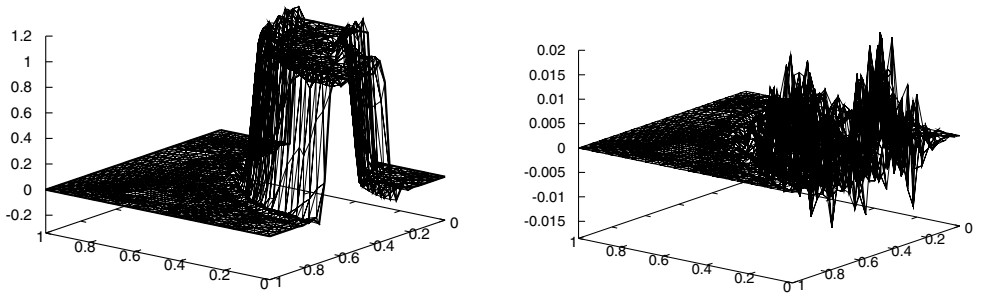


FIG. 6.2. *Example* 2. *Left: LPS solution for* $\tau_0 = 0.03$. *Right: difference between the SUPG solution and the LPS solution.*

$\mathscr{T}_h$ depicted in the middle of Figure 6.1 and the set $\mathscr{M}_h$ given by (5.9) and (5.10). The spaces W_h and D_M are now defined by (5.6) with $l = 2$. Note that if the sets $M \in \mathscr{M}_h$ were not allowed to overlap, we could not use such a space W_h, but, as explained in section 5, the classical approaches would be either to define the space W_h on a refined triangulation or to add additional bubble functions to the space W_h. In both cases, the number of degrees of freedom would significantly increase.

Figure 6.2 (left) shows the solution of the local projection discretization (2.8) for

$$\tau_M = \tau_0 \min\left\{ \frac{h_M}{\|\mathbf{b}\|_{0,\infty,M}}, \frac{h_M^2}{\varepsilon} \right\}, \qquad M \in \mathscr{M}_h, \tag{6.1}$$

with $\tau_0 = 0.03$. It is interesting that, for this value of τ_0, the discrete solution is very similar to the SUPG solution; see Figure 6.2 (right), where the difference between the SUPG solution and the local projection stabilization (LPS) solution is shown. In the SUPG method, we used a stabilization parameter δ given for any $T \in \mathscr{T}_h$ by

$$\delta|_T = \frac{h_{T,\mathbf{b}}}{4\,|\mathbf{b}|} \left(\coth Pe_T - \frac{1}{Pe_T} \right) \quad \text{with} \quad Pe_T = \frac{|\mathbf{b}|\, h_{T,\mathbf{b}}}{4\,\varepsilon},$$

where $h_{T,\mathbf{b}}$ is the diameter of T in the direction of $\mathbf{b}$ (we refer to [9, 11, 13] for details on the definition of δ). Thus, although this example does not satisfy the assumptions of our theory, the local projection method is competitive to the SUPG method.

If the scaling factor τ_0 is increased, then the spurious oscillations visible in Figure 6.2 (left) decrease and the smearing of the discrete solution increases. For the one-level and two-level approaches of the local projection method (see section 5), i.e., for $\mathscr{M}_h$ with nonoverlapping sets M, it is also possible to find values of τ_0 for which the discrete solutions are similar to the SUPG solution. However, if τ_0 is increased (or decreased), the spurious oscillations in the discrete solutions become larger and spread over the whole computational domain. Consequently, for the approaches with nonoverlapping sets M, it is very difficult to find a proper value of τ_0 since both under- and overestimation lead to solutions globally polluted by spurious oscillations. This is a further argument for using the variant with overlapping sets M.

Example 6.3 (problem with parabolic and exponential boundary layers). We consider problem (1.1) with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\mathbf{b} = (1,0)$, $c = 1$, $f = 1$, and $u_b = 0$.

The solution of Example 6.3 possesses an exponential boundary layer at $x = 1$ and parabolic boundary layers at $y = 0$ and $y = 1$. Outside the layers, the solution is

very close to the function

$$u_0(x,y) = 1 - e^{-x}\,. \tag{6.2}$$

We shall consider a structured triangulation $\mathscr{T}_h$ of the type depicted on the right in Figure 6.1 consisting of 16×16 equal squares. Our aim is to compare the following three choices of the spaces W_h and D_M:

- *one-level LPS* with $\mathscr{M}_h = \mathscr{T}_h$ and spaces W_h and D_M defined by (5.7) and (5.8) with $l = 2$;
- *two-level LPS*, where $\mathscr{M}_h$ is a triangulation of Ω of the type depicted on the right in Figure 6.1 consisting of 8×8 equal squares and the spaces W_h and D_M are defined by (5.6) with $l = 2$;
- *overlapping LPS*, where $\mathscr{M}_h$ is given by (5.9) and (5.10) and W_h and D_M are given again by (5.6) with $l = 2$.

Thus, in all three cases, the solution of (2.8) is quadratic along the edges of $\mathscr{T}_h$. In the interiors of the elements of $\mathscr{T}_h$, either the solution is biquadratic or, for the one-level LPS, it belongs to the space of biquadratic functions enriched by three bubble functions. Using two bubble functions, i.e., $P_{l-1}(\widehat{T})$ instead of $R_{l-1}(\widehat{T})$ in (5.7), leads to results almost identical to those for three bubble functions. We refer to Remark 5.3 for the numbers of nonzero matrix entries and numbers of degrees of freedom corresponding to the methods compared in this example.

A stabilized method should be able to provide a good approximation of the solution u away from the boundary layers. Therefore, we shall investigate the quality of the discrete solutions u_h at the points $(x_i, 0.5)$ with $x_i = i/32$, $i = 0, \dots, 24$. These points are all vertices and midpoints of edges of $\mathscr{T}_h$ lying on the line $y = 0.5$ and having their x coordinate in the interval $[0, 0.75]$. To measure the oscillations and accuracy of u_h along $[0, 0.75] \times \{0.5\}$, we define the quantities

$$RTV(u_h) = \frac{\sum_{i=1}^{24} |u_h(x_i, 0.5) - u_h(x_{i-1}, 0.5)|}{\max\limits_{i=1,\dots,24} |u_h(x_i, 0.5)|}\,,$$

$$ERR(u_h) = \sqrt{\sum_{i=1}^{24} \left(u_h(x_i, 0.5) - u_0(x_i, 0.5)\right)^2}\,.$$

The value $RTV(u_h)$ represents an approximation of the relative total variation of u_h along $[0, 0.75] \times \{0.5\}$. Since $u_h(x_0, 0.5) = 0$, we have $RTV(u_h) \geq 1$. The sequence $\{u_h(x_i, 0.5)\}_{i=0}^{24}$ is monotone if and only if $RTV(u_h) = 1$. Large $RTV(u_h)$ indicates that the values $u_h(x_i, 0.5)$ oscillate. The value $ERR(u_h)$ measures the accuracy of u_h by comparing u_h with the limit solution u_0 given in (6.2).

The stabilization parameters τ_M are again defined by (6.1), and we shall discuss how the solutions of the local projection discretization (2.8) are influenced by the choice of the scaling factor τ_0. Figure 6.3 shows the dependence of $RTV(u_h)$ and $ERR(u_h)$ on τ_0 for the three choices of the spaces W_h and D_M described above. First let us consider the one-level and two-level methods. For $\tau_0 \lesssim 10^{-3}$, the solutions possess large oscillations along $[0, 1] \times \{0.5\}$ whose width is $1/16$. Away from the parabolic layers, the two-level solution is independent of y, whereas the one-level solution is periodic in the y direction with the period $1/16$. Along horizontal lines (i.e., lines with a constant y coordinate) crossing midpoints of elements, the width of oscillations is $1/8$. Thus, for both methods, the values at vertices lying on a horizontal line do not oscillate. For the two-level LPS, also the values at midpoints of edges lying on any horizontal line do not oscillate, whereas, for the one-level LPS, this holds only

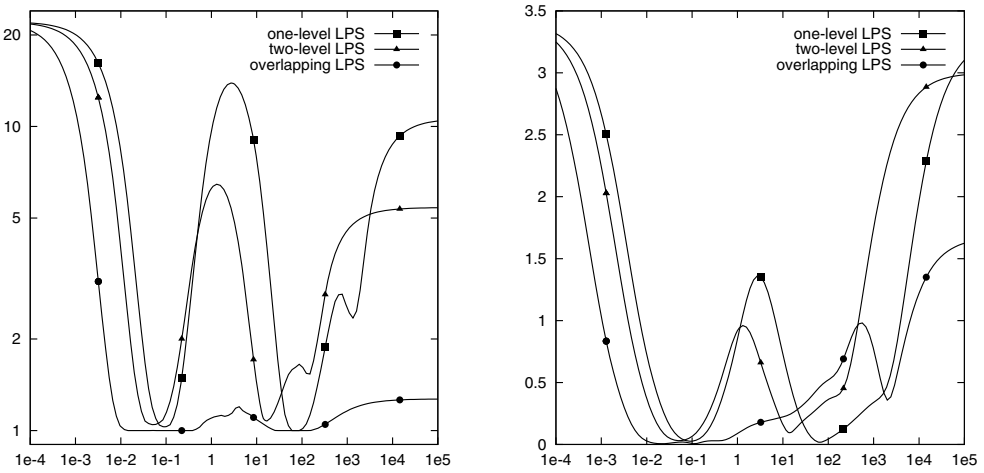


FIG. 6.3. *Example* 6.3*: dependence of* $RTV(u_h)$ *(left) and* $ERR(u_h)$ *(right) on the scaling factor* τ_0*.*

for mesh lines. As we see from Figure 6.3, for increasing τ_0, both the oscillations and errors decrease and a minimum is attained for τ_0 near 0.1. The corresponding solutions are shown in Figure 6.4 (top row). Since the solutions are symmetric with respect to the line $y = 0.5$, we show the solutions only on $[0, 1] \times [0.5, 1]$, which makes oscillations more visible. The lines in the figures connect values at vertices, midpoints of edges, and midpoints of elements. The additional bubble functions of the one-level method are not taken into account in this case. We observe that spurious oscillations are localized along the boundary layers but the width of the numerical boundary layers at the outflow boundary is rather large. Far away from the boundary layers, the quality of the discrete solutions is satisfactory.

If τ_0 is increased above 0.1, then Figure 6.3 indicates that, for both the one-level and the two-level methods, the oscillations and errors increase, reach a maximum around $\tau_0 = 1$, and decrease again towards a second minimum which is reached for τ_0 between 10 and 100. The oscillations for $\tau_0 \in (0.1, 10)$ have a different character than those for $\tau_0 < 10^{-3}$. For both methods, the solutions depend only slightly on y away from the parabolic layers, whereas they possess oscillations in the x direction whose width is 1/16 for the one-level LPS and 1/8 for the two-level LPS. Thus, the width of the oscillations corresponds to the size of the sets $M \in \mathscr{M}_h$. To get an impression of how fast the solutions deteriorate if τ_0 is increased, we show the two wildly oscillating solutions in Figure 6.4 (middle row) obtained for τ_0 that are slightly larger than the "optimal" values near 0.1. Figure 6.4 (bottom row) shows that the values of τ_0 corresponding to the second minima in Figure 6.3 (between 10 and 100) lead to worse discrete solutions than in case of the first minima. If τ_0 further increases, the oscillations in parabolic layers become larger, in particular for the one-level LPS. For very large values of τ_0, the solutions again wildly oscillate in the x direction. The width of the oscillations is 1/8 for the one-level LPS and 1/4 for the two-level LPS. Inside the sets $M \in \mathscr{M}_h$, no oscillations occur.

For the overlapping LPS and $\tau_0 \lesssim 10^{-3}$, the discrete solutions are qualitatively similar to those for the two-level LPS. However, as we can see from Figure 6.3, the dependence of $RTV(u_h)$ on τ_0 is different after the minimal value $RTV(u_h) = 1$ has

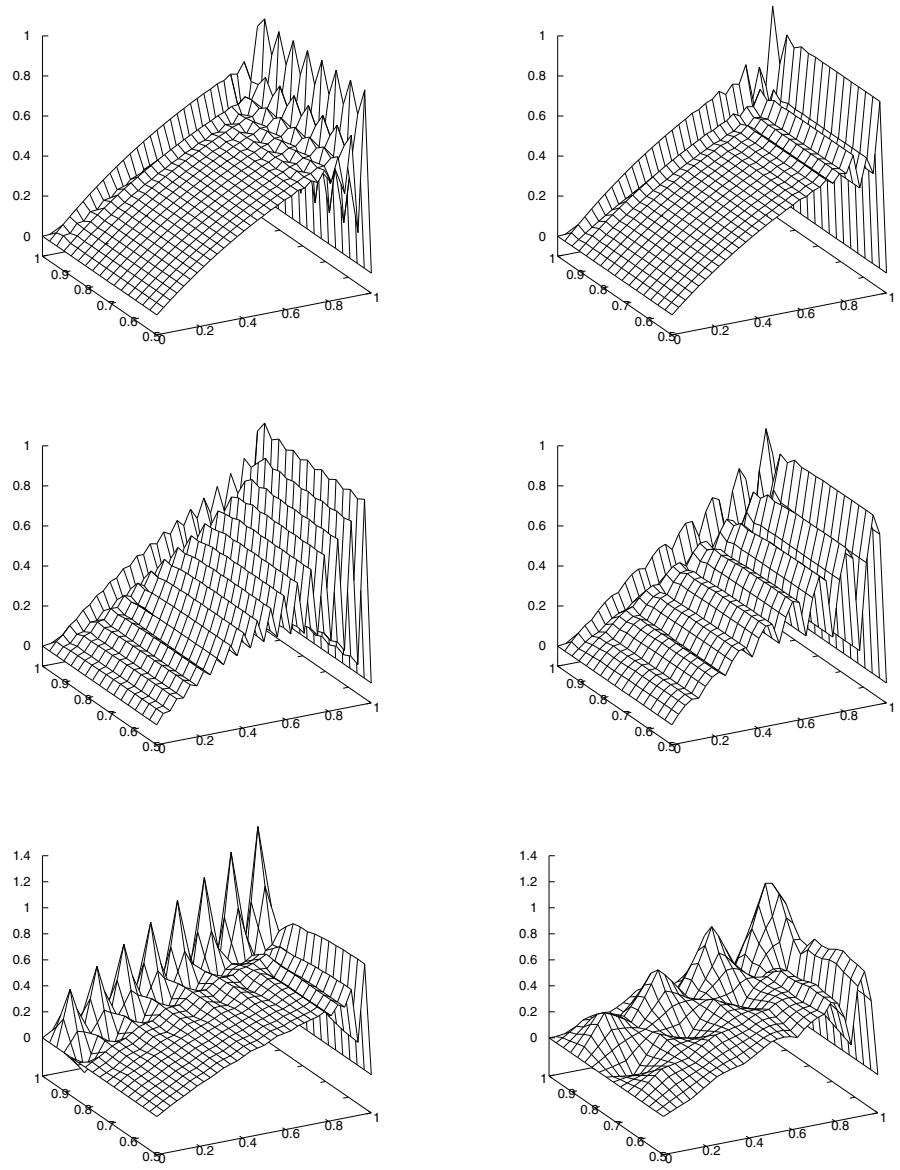


FIG. 6.4. *Example* 6.3. *Left column, top to bottom: Solutions of the one-level LPS for* $\tau_0 = 0.1$, $\tau_0 = 0.5$, *and* $\tau_0 = 65$. *Right column, top to bottom: Solutions of the two-level LPS for* $\tau_0 = 0.05$, $\tau_0 = 0.3$, *and* $\tau_0 = 15$.

been attained (for $\tau_0 \sim 0.014$). In particular, we observe that u_h is monotone along $[0, 0.75] \times \{0.5\}$ for a wide range of τ_0 and that $RTV(u_h) < 1.3$ for any $\tau_0 > 0.01$. A more detailed investigation of the solutions reveals that, for $\tau_0 \gtrsim 0.1$, no oscillations at all occur along $[0, 0.75] \times \{0.5\}$. More precisely, for $\tau_0 \in (0.1, 25)$, the solution u_h has one minimum and two maxima in $(0, 1) \times \{0.5\}$. For $\tau_0 \sim 0.1$, the points where the extrema are attained have their x coordinate larger than 0.75, and hence u_h is monotone along $[0, 0.75] \times \{0.5\}$ (i.e., $RTV(u_h) = 1$). When τ_0 is increased, these points shift towards $x = 0$ and the distances among them become larger so that the

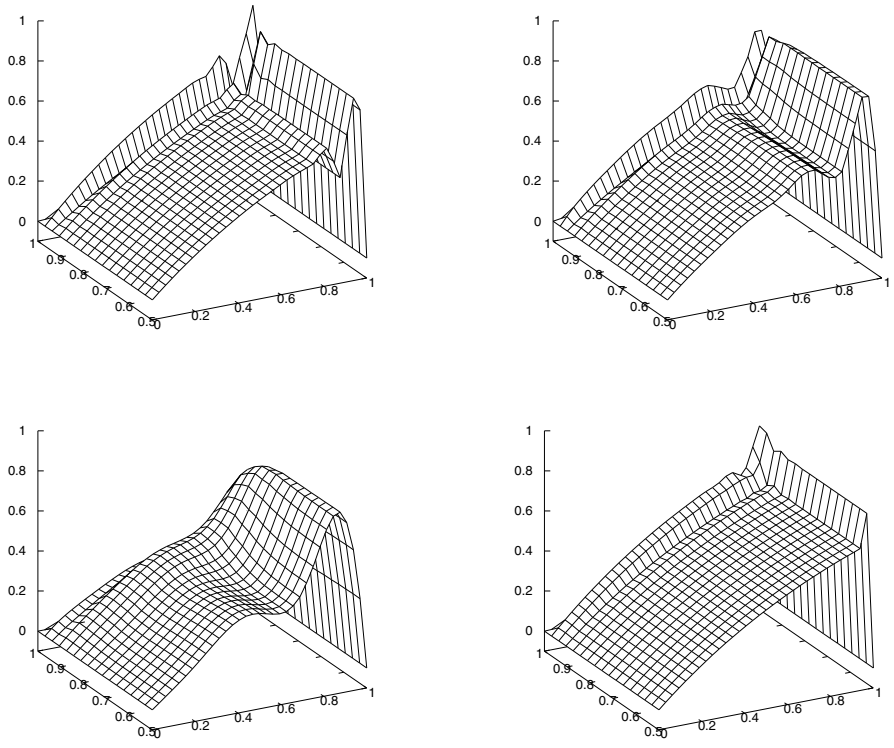


FIG. 6.5. *Example* 6.3*: solutions for the overlapping LPS with* $\tau_0 = 0.1$ *(top left),* $\tau_0 = 1$ *(top right),* $\tau_0 = 10$ *(bottom left), and, finally, with* $\tau_M = 1/320$*,* W_h *from* (6.3)*, and* $\mathscr{M}_h$ *consisting of sets not intersecting the boundary of* $\Omega \setminus G_h$ *(bottom right).*

solutions appear smoother. When at least one extremum point is in $(0, 0.75) \times \{0.5\}$, the solution is not monotone along $[0, 0.75] \times \{0.5\}$ and $RTV(u_h) > 1$; see Figure 6.3. For $\tau_0 \sim 25$, two extrema disappear and u_h has only one maximum and no minimum in $(0, 1) \times \{0.5\}$. The x coordinate of the maximum point is larger than 0.75 so that again $RTV(u_h) = 1$. If τ_0 is further increased, the maximum point moves towards $(0.5, 0.5)$ and the magnitude of u_h decreases. For $\tau_0 = 10^5$, the maximum of u_h in $\overline{\Omega}$ is less than 0.014. The above discussion is illustrated by the solutions for $\tau_0 = 0.1, 1$, and 10 in Figure 6.5 (top left, top right, and bottom left). The value $\tau_0 = 0.1$ represents a good choice with respect to both oscillations and accuracy; see also the graph of $ERR(u_h)$ in Figure 6.3.

The detailed discussion of this example clearly shows the important difference between the one-level and two-level LPS on the one side and the overlapping LPS on the other side which was already briefly mentioned in the discussion of Example 6.2. For the former two methods, there is only a small interval of values of τ_0 which lead to acceptable discrete solutions. Since τ_0 both smaller and larger than these "optimal" values leads to spurious oscillations, it is very difficult to find a proper value of τ_0 numerically, and a small deviation of τ_0 from the "optimal" value may deteriorate the solution considerably. On the other hand, for the overlapping LPS, the properties of the discrete solution depend on τ_0 in a monotone way: for increasing τ_0, oscillations decrease and smearing increases. This is much more convenient from the practical point of view since, in many applications, a moderate smearing is more acceptable than spurious oscillations.

The "optimal" solution in Figure 6.5 (top left) for the overlapping LPS is slightly better than the corresponding solutions for the one-level and two-level LPS in Figure 6.4 (top row). Nevertheless, it is still not satisfactory since the numerical boundary layer at the outflow boundary is rather wide. Theoretical considerations that are outside the scope of this paper and will be the subject of a separate publication show that a considerable improvement can be achieved by increasing the polynomial degree of W_h on elements of $\mathscr{T}_h$ lying at the boundary of Ω. More precisely, denoting

$$\mathscr{G}_h = \{T \in \mathscr{T}_h\,;\, \overline{T} \cap \partial\Omega \neq \emptyset\}\,, \qquad G_h = \operatorname{int} \bigcup_{T \in \mathscr{G}_h} \overline{T}\,,$$

we consider

$$W_h = X_{\mathscr{T}_h,2} + (X_{\mathscr{G}_h,3} \cap H_0^1(G_h)) \tag{6.3}$$

instead of $W_h = X_{\mathscr{T}_h,2}$. The set $\mathcal{M}_h$ consists of the sets M_i defined by (5.9) with $x_i \notin \overline{G_h}$ and of sets $M \subset G_h$ defined by $M = \operatorname{int}(\overline{T} \cup \overline{T'})$, where T, T' are elements of $\mathscr{T}_h$ adjacent to any edge of $\mathscr{T}_h$ contained in G_h (but not in ∂G_h). Thus, the sets in $\mathscr{M}_h$ are not allowed to intersect the boundary of $\Omega \setminus G_h$. Again, $D_M = P_1(M)$ for any $M \in \mathscr{M}_h$. Theoretical considerations provide an optimal value $\tau_M = 1/320$ for M at the outflow boundary of Ω. For simplicity, this value of τ_M is considered for any $M \in \mathscr{M}_h$. The solution is depicted in Figure 6.5 (bottom right), and we observe that now the oscillations at the outflow boundary are localized to one row of elements along this boundary and can be easily removed by postprocessing. Away from this row of elements and the parabolic boundary layers, the discrete solution is very accurate. Although the approximation of the parabolic boundary layers improved a little bit, a substantial improvement cannot be expected since only the streamline derivative is used to define the stabilization term. Here different techniques have to be applied; see, e.g., [13].

7. Conclusions. We have introduced a generalization of the local projection stabilization for finite element discretizations of steady scalar convection-diffusion-reaction equations. The important feature of this generalization is that projection spaces may be defined on overlapping sets. Consequently, we can use standard finite element spaces for approximating the unknown solution, whereas, in the classical local projection method, the definition of approximation spaces requires either a refinement of the triangulation or the introduction of additional bubble functions, both leading to a considerable increase of the number of degrees of freedom. Moreover, numerical results in this paper show that the use of projection spaces defined on overlapping sets significantly enhances the robustness of the method with respect to the choice of the stabilization parameter. The stabilization term has been defined using local constant approximations of the convection field, which has enabled us to prove an optimal convergence result, not available up to now for stabilization parameters with a physically correct scaling with respect to the data. Stability and convergence have been established in the LPSD norm that is stronger than the norm for which the discretization is coercive. Moreover, the LPSD norm is equivalent to the SUPG norm for appropriately chosen δ. In contrast to an earlier paper, no additional assumptions have been necessary for proving these improved results. Finally, we demonstrated that the quality of discrete solutions to problems with exponential boundary layers can be substantially improved if we increase the polynomial degree of the approximation space along the boundary.

Acknowledgment. The author is gratefully indebted to Professor Hans-Görg Roos for many fruitful discussions which inspired this work.

REFERENCES

[1] R. Becker and M. Braack, *A finite element pressure gradient stabilization for the Stokes equations based on local projections*, Calcolo, 38 (2001), pp. 173–199.
[2] R. Becker and M. Braack, *A two-level stabilization scheme for the Navier–Stokes equations*, in Numerical Mathematics and Advanced Applications, M. Feistauer, V. Dolejší, P. Knobloch, and K. Najzar, eds., Springer-Verlag, Berlin, 2004, pp. 123–130.
[3] R. Becker and B. Vexler, *Stabilized finite element methods for the generalized Oseen problem*, Numer. Math., 106 (2007), pp. 349–367.
[4] M. Braack, *Optimal control in fluid mechanics by finite elements with symmetric stabilization*, SIAM J. Control Optim., 48 (2009), pp. 672–687.
[5] M. Braack and G. Lube, *Finite elements with local projection stabilization for incompressible flow problems*, J. Comput. Math., 27 (2009), pp. 116–147.
[6] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, 3rd ed., Texts Appl. Math. 15, Springer, New York, 2008.
[7] 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), pp. 199–259.
[8] P. G. Ciarlet, *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, Vol. II. Finite Element Methods. Part 1, P. G. Ciarlet and J. L. Lions, eds., North–Holland, Amsterdam, 1991, pp. 17–351.
[9] R. Codina, E. Oñate, and M. Cervera, *The intrinsic time for the streamline upwind/Petrov–Galerkin formulation using quadratic elements*, Comput. Methods Appl. Mech. Engrg., 94 (1992), pp. 239–262.
[10] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, New York, 2004.
[11] A. C. Galeão, R. C. Almeida, S. M. C. Malta, and A. F. D. Loula, *Finite element analysis of convection dominated reaction-diffusion problems*, Appl. Numer. Math., 48 (2004), pp. 205–222.
[12] S. Ganesan and L. Tobiska, *Stabilization by local projection for convection-diffusion and incompressible flow problems*, J. Sci. Comput., 43 (2010), pp. 326–342.
[13] V. John and P. Knobloch, *On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part* I—*A review*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2197–2215.
[14] P. Knobloch, *On the application of local projection methods to convection-diffusion-reaction problems*, in BAIL 2008—Boundary and Interior Layers, A. F. Hegarty, N. Kopteva, E. O' Riordan, and M. Stynes, eds., Lecture Notes Comput. Sci. Eng. 69, Springer-Verlag, Berlin, 2009, pp. 183–194.
[15] P. Knobloch and G. Lube, *Local projection stabilization for advection-diffusion-reaction problems: One-level vs. two-level approach*, Appl. Numer. Math., 59 (2009), pp. 2891–2907.
[16] P. Knobloch and L. Tobiska, *On the stability of finite-element discretizations of convection-diffusion-reaction equations*, IMA J. Numer. Anal., to appear (published online August 27, 2009).
[17] G. Matthies, P. Skrzypacz, and L. Tobiska, *A unified convergence analysis for local projection stabilisations applied to the Oseen problem*, M2AN Math. Model. Numer. Anal., 41 (2007), pp. 713–742.
[18] G. Matthies, P. Skrzypacz, and L. Tobiska, *Stabilization of local projection type applied to convection-diffusion problems with mixed boundary conditions*, Electron. Trans. Numer. Anal., 32 (2008), pp. 90–105.
[19] H.-G. Roos, M. Stynes, and L. Tobiska, *Robust Numerical Methods for Singularly Perturbed Differential Equations. Convection-Diffusion-Reaction and Flow Problems*, 2nd ed., Springer-Verlag, Berlin, 2008.

ESAIM: M2AN 47 (2013) 1335–1366
DOI: 10.1051/m2an/2013071

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

A LOCAL PROJECTION STABILIZATION FINITE ELEMENT METHOD WITH NONLINEAR CROSSWIND DIFFUSION FOR CONVECTION-DIFFUSION-REACTION EQUATIONS

Gabriel R. Barrenechea[1], Volker John[2] and Petr Knobloch[3]

Abstract. An extension of the local projection stabilization (LPS) finite element method for convection-diffusion-reaction equations is presented and analyzed, both in the steady-state and the transient setting. In addition to the standard LPS method, a nonlinear crosswind diffusion term is introduced that accounts for the reduction of spurious oscillations. The existence of a solution can be proved and, depending on the choice of the stabilization parameter, also its uniqueness. Error estimates are derived which are supported by numerical studies. These studies demonstrate also the reduction of the spurious oscillations.

Mathematics Subject Classification. 65N30, 65N12, 65N15, 65M60.

Received October 30, 2012. Revised February 8, 2013.
Published online July 30, 2013.

1. Introduction

The solution of convection-dominated convection-diffusion-reaction equations with finite element methods constitutes a very challenging (and open) problem. Over the last three decades, the amount of work devoted to this problem is impressive. The usual way of treating dominating convection, at least in the context of finite element methods, consists in adding extra terms to the standard Galerkin formulation, aimed at enhancing the stability of the discrete solution by means of introducing artificial diffusion. These new terms vary according to the method, and can be residual-based, as in the SUPG/GLS/SDFEM family (see [6,13,14,16,29]), or edge based, such as the CIP method (see [7,9]). For an up-to-date and thorough review of these and other techniques, see [31]. It is striking to notice that, despite the impressive amount of work that has been devoted to this topic, up to now there is not a method that 'ticks all the boxes', *i.e.*, a method that produces sharp layers while avoiding oscillations, see [1] for a recent review and a numerical assessment.

Keywords and phrases. Finite element method, local projection stabilization, crosswind diffusion, convection-diffusion-reaction equation, well posedness, time dependent problem, stability, error estimates.

[1] Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland. gabriel.barrenechea@strath.ac.uk

[2] Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany and Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany. john@wias-berlin.de

[3] Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic. knobloch@karlin.mff.cuni.cz

Article published by EDP Sciences

Among the various stabilized finite element methods, the local projection stabilization (LPS) method has received some attention over the last decade. Originally proposed for the Stokes problem in [2], and extended to the Oseen equations in [4] (see also [5,30]), the LPS method has been also used recently to treat convection-diffusion equations (see [15,24–26]). The basic idea of this method consists in restricting the direct application of the stabilization to so-called fluctuations or resolved small scales, which are defined by local projections. It has several attractive features, such as adding symmetric terms to the formulation and avoiding the computation of second derivatives of the basis functions (thus using only information that is needed for the assembly of the matrices from the standard Galerkin method). Unfortunately, the solutions obtained with the LPS method possess the same deficiency like solutions computed, *e.g.*, with the SUPG method: non-negligible spurious oscillations are often present in a vicinity of layers.

Motivated by the wish of recovering the monotonicity properties of the continuous problem, which might be crucial in applications, a number of so-called Spurious Oscillations at Layers Diminishing (SOLD) methods were proposed. SOLD methods add an extra term to the already stabilized formulation, which usually depends on the discrete solution in a nonlinear way, vanishes for small residuals (thus acting mostly at layers), and adds some extra, but different, diffusivity to the formulation. In particular, methods that add crosswind diffusion, like the one proposed in [11], have been proved to belong to the best SOLD methods in comprehensive studies [17,18]. Although these methods diminish oscillations considerably, no single method succeeds to fully eliminate them [17,18,23]. Also, from a purely mathematical point of view, it is unknown if these methods lead to well-posed problems. In fact, existence of solutions is usually possible to prove, but, to our best knowledge, there is no nonlinear SOLD method that is known to produce a unique solution, see [7,27] for a discussion of this topic.

Based on the previous considerations, this paper has three major objectives, namely:

- to improve the quality of the LPS solution (especially in the vicinity of layers);
- to explore the applicability of SOLD-type strategies within a LPS context; and
- to contribute to the mathematical understanding of nonlinear stabilization techniques for the convection-diffusion equation.

Hence, in this work we propose a LPS method with nonlinear crosswind diffusion for convection-diffusion-reaction equations. Two ways for choosing the parameter in the crosswind diffusion term will be studied. The first choice uses global information obtained from the data of the problem, whereas the second proposal is completely local, employing information of the computed solution instead of the data. For the first approach, which is the simpler one, the existence and the uniqueness of the solution can be proved for the steady-state and time-dependent equations, where the latter is discretized in time with an implicit one-step θ-scheme. To our best knowledge, this is the first nonlinear discretization for convection-diffusion-reaction equations for which both, existence and uniqueness of a solution can be shown. The form of the crosswind term resembles the Smagorinsky Large Eddy Simulation (LES) model which was analyzed in [28]. It involves fluctuations of a term mimicking a p-Laplacian. The crucial analytical property for proving the uniqueness of the solution is the strong monotonicity of the corresponding operator. For the more complicated local definition of the parameter, the analysis will show the existence of a solution and its uniqueness for the time-dependent discretization in the case of sufficiently small time steps.

The analysis is performed for the model problems of linear steady-state and time-dependent convection-diffusion-reaction equations. Applying a nonlinear discretization scheme to a linear problem leads certainly to a considerable complication of the solution process and to an additional numerical cost. This latter aspect can be overcome in the transient regime by using a semi-implicit (linearized) approach that computes the stabilization parameter with the solution from the previous discrete time. With respect to the former aspect, it has to be mentioned that the most important motivation for studying discretizations that reduce spurious oscillations comes from the need to address applications that lead to nonlinear coupled systems of convection-diffusion-reaction equations as in [21]. It was demonstrated in [21] that the locally large spurious oscillations of the SUPG method might lead to a fast blow-up of the simulations, and hence the reduction of the spurious

oscillations is essential to perform simulations at all. Thus, the reduction of the oscillations at layers becomes a priority, even over computational cost. It should be noted that in many applications, like in [21], only interior or characteristic layers are present, such that a method for reducing the oscillations has to work properly in particular for these types of layers. Finally, it is worth mentioning that our final aim is to address applications that lead to such coupled problems. Since these problems are nonlinear, the use of a nonlinear stabilization usually does not result in a notable complication of the solution procedure.

The plan of the paper is as follows. In the remaining part of this introduction, the problems of interest are stated and some basic notations are given. Section 2 will summarize the main abstract hypothesis imposed on the different partitions of the domain and the finite element spaces considered. Section 3 presents the method for the steady-state case, for which well-posedness is analyzed in Section 3.1 and error estimates are proved in Section 3.2. In Section 4, the method for the time-dependent problem is presented. Well-posedness and stability are proved in Section 4.1 and error estimates in Section 4.2. Since the analysis is based on the abstract framework from Section 2, Section 5 presents some concrete examples that fit into this framework. Finally, numerical illustrations that support the analytical results and which demonstrate the reduction of spurious oscillations are presented in Section 6.

Throughout the paper, standard notations are used for Sobolev spaces and corresponding norms, see, *e.g.*, [10]. In particular, given a measurable set $D \subset \mathbb{R}^d$, the inner product in $L^2(D)$ or $L^2(D)^d$ is denoted by $(\cdot,\cdot)_D$ and the notation $(\cdot,\cdot)$ is used instead of $(\cdot,\cdot)_\Omega$. The norm (seminorm) in $W^{m,p}(D)$ will be denoted by $\|\cdot\|_{m,p,D}$ ($|\cdot|_{m,p,D}$), with the convention $\|\cdot\|_{m,D} = \|\cdot\|_{m,2,D}$, and the same notation is used for scalar and vector-valued functions.

1.1. The problems of interest

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded polygonal (polyhedral) domain with a Lipschitz-continuous boundary $\partial\Omega$ and let us consider the steady-state convection-diffusion-reaction equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = f \quad \text{in } \Omega, \qquad\qquad u = u_b \quad \text{on } \partial\Omega. \tag{1}$$

It is assumed that ε is a positive constant and $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, and $u_b \in H^{1/2}(\partial\Omega)$ are given functions satisfying

$$\sigma := c - \frac{1}{2}\nabla\cdot\boldsymbol{b} \geq \sigma_0 > 0 \quad \text{in } \Omega, \tag{2}$$

where σ_0 is a constant. Then the boundary value problem (1) has a unique solution in $H^1(\Omega)$.

The condition $\sigma_0 > 0$ is often used in the analysis of stabilized finite element methods for the numerical solution of (1), see, *e.g.*, [31], but it limits the applications of the theory since many problems of interest involve solenoidal convective velocities and no zero-order terms, which leads to $\sigma_0 = 0$. Unfortunately, it is not known how to prove optimal convergence results even for the underlying linear local projection stabilization without assuming $\sigma_0 > 0$, although numerical results do not indicate any deterioration of the convergence rates when $\sigma_0 = 0$. The analysis of the nonlinear term introduced in this paper does not require this assumption.

Besides the steady-state case, also the time-dependent convection-diffusion-reaction equation

$$\left.\begin{aligned} u_t - \varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u &= f \quad \text{in } (0,T]\times\Omega,\\ u &= u_b \text{ in } [0,T]\times\partial\Omega,\\ u(0,\cdot) &= u_0 \text{ in } \Omega, \end{aligned}\right\} \tag{3}$$

will be considered. In (3), $[0,T]$ is a finite time interval, ε is assumed to be a positive constant, $\boldsymbol{b} \in L^\infty(0,T;W^{1,\infty}(\Omega)^d)$, $c \in L^\infty(0,T;L^\infty(\Omega))$, $f \in L^2(0,T;L^2(\Omega))$, $u_b \in L^2(0,T;H^{1/2}(\partial\Omega))$, and $u_0 \in H^1(\Omega)$ denotes the initial condition. The function σ is defined analogously to (2) and the inequality (2) is assumed to hold for all $t \in [0,T]$. In this case, the condition $\sigma_0 > 0$ can be circumvented by considering instead of (3) an equivalent problem for $v = u\,\mathrm{e}^{-\alpha\,t}$ which satisfies $\sigma_0 > 0$ for sufficiently large α.

2. Assumptions on approximation spaces and the set $\mathscr{M}_h$

From now on, C, $\tilde{C}$ or $\bar{C}$ denote generic constants which may take different values at different occurrences but are always independent of the data ε, $\boldsymbol{b}$, c, f, and u_b, the constant σ_0, and the discretization parameters (h and δt in the following).

Given $h > 0$, let $W_h \subset W^{1,\infty}(\Omega)$ be a finite-dimensional space approximating the space $H^1(\Omega)$ and set $V_h = W_h \cap H_0^1(\Omega)$. Next, let $\mathscr{M}_h$ be a set consisting of a finite number of open subsets M of Ω such that $\overline{\Omega} = \cup_{M\in\mathscr{M}_h} \overline{M}$. It will be supposed that, for any $M \in \mathscr{M}_h$,

$$\operatorname{card}\{M' \in \mathscr{M}_h\,;\ M \cap M' \neq \emptyset\} \leq C, \tag{4}$$

$$h_M := \operatorname{diam}(M) \leq C\,h, \tag{5}$$

$$h_M \leq C\,h_{M'} \qquad \forall\ M' \in \mathscr{M}_h,\ M \cap M' \neq \emptyset, \tag{6}$$

$$h_M^d \leq C \operatorname{meas}_d(M). \tag{7}$$

The space W_h is assumed to satisfy the local inverse inequality

$$|v_h|_{1,M} \leq C\,h_M^{-1}\,\|v_h\|_{0,M} \qquad \forall\ v_h \in W_h,\ M \in \mathscr{M}_h. \tag{8}$$

For any $M \in \mathscr{M}_h$, a finite-dimensional space $D_M \subset L^\infty(M)$ is introduced. It is assumed that there exists a positive constant β_{LP} independent of h such that

$$\sup_{v\in V_M} \frac{(v,q)_M}{\|v\|_{0,M}} \geq \beta_{LP}\,\|q\|_{0,M} \qquad \forall\ q \in D_M,\ M \in \mathscr{M}_h, \tag{9}$$

where $V_M = \{v_h \in V_h\,;\ v_h = 0 \text{ in } \Omega \setminus M\}$. This hypothesis will be needed in what follows for the construction of a special interpolation operator (see Lemma 3.7 below). Concrete examples of spaces W_h and D_M satisfying the assumptions formulated here will be presented in Section 5.

Furthermore, for any $M \in \mathscr{M}_h$, a finite-dimensional space $G_M \subset L^\infty(M)$ with $G_M \supset D_M$ is introduced such that

$$\left.\frac{\partial v_h}{\partial x_i}\right|_M \in G_M \qquad \forall\ v_h \in W_h,\ i = 1,\dots,d,$$

and it is assumed that, for any $p \in [1,\infty]$, there is a constant C such that

$$\|q\|_{0,p,M} \leq C\,h_M^{\frac{d}{p}-\frac{d}{2}}\,\|q\|_{0,M} \qquad \forall\ q \in G_M,\ M \in \mathscr{M}_h. \tag{10}$$

To characterize the approximation properties of the spaces W_h and D_M, it is assumed that there exist interpolation operators $i_h \in \mathscr{L}(C(\overline{\Omega}), W_h) \cap \mathscr{L}(C(\overline{\Omega}) \cap H_0^1(\Omega), V_h)$ and $j_M \in \mathscr{L}(H^1(M), D_M)$, $M \in \mathscr{M}_h$, such that, for some constants $l \in \mathbb{N}$ and $C > 0$ and for any set $M \in \mathscr{M}_h$, it holds

$$|v - i_h v|_{1,M} + h_M^{-1}\,\|v - i_h v\|_{0,M} \leq C\,h_M^k\,|v|_{k+1,M} \qquad \forall\ v \in H^{k+1}(M),\ k = 1,\dots,l, \tag{11}$$

$$\|q - j_M q\|_{0,M} \leq C\,h_M^k\,|q|_{k,M} \qquad \forall\ q \in H^k(M),\ k = 1,\dots,l. \tag{12}$$

In addition, it is assumed that, for any $p \in [1,6]$,

$$|v - i_h v|_{1,p,M} \leq C\,h_M^{k+\frac{d}{p}-\frac{d}{2}}\,|v|_{k+1,M} \qquad \forall\ v \in H^{k+1}(M),\ k = 1,\dots,l. \tag{13}$$

3. A local projection discretization of the steady-state problem

The weak form of problem (1) is: find $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u,v) = (f,v) \qquad \forall\ v \in H_0^1(\Omega), \tag{14}$$

where the bilinear form a is given by

$$a(u,v) := \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + (c\,u, v).$$

As it was mentioned in the introduction, the most often used approach to cure the instabilities of the Galerkin method consists in adding extra terms to the formulation. To build these additional terms for the method studied here, for any $M \in \mathscr{M}_h$, a continuous linear projection operator π_M is introduced which maps the space $L^2(M)$ onto the space D_M. It is assumed that

$$\|\pi_M\|_{\mathscr{L}(L^2(M),L^2(M))} \le C \qquad \forall\ M \in \mathscr{M}_h. \tag{15}$$

E.g., if π_M is the orthogonal L^2 projection, then $C = 1$. Using this operator, the fluctuation operator $\kappa_M := id - \pi_M$ is defined, where id is the identity operator on $L^2(M)$. Then, clearly

$$\|\kappa_M\|_{\mathscr{L}(L^2(M),L^2(M))} \le C \qquad \forall\ M \in \mathscr{M}_h. \tag{16}$$

Since κ_M vanishes on D_M, it follows from (16) and (12) that

$$\|\kappa_M\, q\|_{0,M} \le C\, h_M^k\, |q|_{k,M} \qquad \forall\ q \in H^k(M),\ M \in \mathscr{M}_h,\ k = 0,\dots,l. \tag{17}$$

An application of κ_M to a vector-valued function means that κ_M is applied component-wise.

For any $M \in \mathscr{M}_h$, a constant $\boldsymbol{b}_M \in \mathbb{R}^d$ is chosen such that

$$|\boldsymbol{b}_M| \le \|\boldsymbol{b}\|_{0,\infty,M}, \qquad \|\boldsymbol{b} - \boldsymbol{b}_M\|_{0,\infty,M} \le C\, h_M\, |\boldsymbol{b}|_{1,\infty,M}, \tag{18}$$

where $|\cdot|$ denotes the Euclidean norm in $\mathbb{R}^d$. A typical choice for $\boldsymbol{b}_M$ is the value of $\boldsymbol{b}$ at one point of M, or the integral mean value of $\boldsymbol{b}$ over M. In addition, a function $\widetilde{u}_{bh} \in W_h$ is introduced such that its trace approximates the boundary condition u_b.

We are now ready to present the finite element method to be studied: find $u_h \in W_h$ such that $u_h - \widetilde{u}_{bh} \in V_h$ and

$$a(u_h, v_h) + s_h(u_h, v_h) + d_h(u_h; u_h, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h, \tag{19}$$

where

$$\begin{aligned} s_h(u,v) &= \sum_{M\in\mathscr{M}_h} \tau_M\ \left(\kappa_M(\boldsymbol{b}_M\cdot\nabla u), \kappa_M(\boldsymbol{b}_M\cdot\nabla v)\right)_M, \\ d_h(w;u,v) &= \sum_{M\in\mathscr{M}_h} \left(\tau_M^{\mathrm{sold}}(w)\,\kappa_M(P_M\nabla u), \kappa_M(P_M\nabla v)\right)_M, \end{aligned}$$

and $P_M : \mathbb{R}^d \to \mathbb{R}^d$ is the projection onto the line (plane) orthogonal (crosswind) to the vector $\boldsymbol{b}_M$ defined by

$$P_M = \begin{cases} I - \dfrac{\boldsymbol{b}_M \otimes \boldsymbol{b}_M}{|\boldsymbol{b}_M|^2} & \text{if } \boldsymbol{b}_M \neq \boldsymbol{0}, \\ 0 & \text{if } \boldsymbol{b}_M = \boldsymbol{0}, \end{cases}$$

I being the identity tensor. The stabilization parameters are given by

$$\begin{aligned} \tau_M &= \tau_0\ \min\left\{ \frac{h_M}{\|\boldsymbol{b}\|_{0,\infty,M}}, \frac{h_M^2}{\varepsilon} \right\}, \\ \tau_M^{\mathrm{sold}}(u_h) &= \tilde{\tau}_M(u_h)\,|\kappa_M(P_M\nabla u_h)|, \end{aligned} \tag{20}$$

where τ_0 is a positive constant and $\tilde{\tau}_M$ is a non-negative function of u_h and the data of (1). Note that the crosswind stabilization term is of p-Laplacian type with $p = 3$.

It remains to specify the function $\tilde{\tau}_M$. First, inspired by the definition of s_h, where each term in the sum is bounded by $\tau_0 \, h_M \, |\boldsymbol{b}_M| \, \|\kappa_M \nabla u\|_{0,M} \, \|\kappa_M \nabla v\|_{0,M}$, we set $\tilde{\tau}_M(u_h) = \gamma_M(u_h) \, h_M \, |\boldsymbol{b}_M|$ with a function γ_M still depending on u_h and/or the data of (1). Second, the function γ_M has to be chosen in such a way that the discrete problem preserves the following scaling properties of the problem (1):

- if the data ε, $\boldsymbol{b}$, c, and f are replaced by $\alpha\,\varepsilon$, $\alpha\,\boldsymbol{b}$, $\alpha\,c$, and $\alpha\,f$, respectively, with some constant $\alpha \neq 0$, then the solution of (1) does not change;
- if f and u_b are replaced by $\alpha\,f$ and $\alpha\,u_b$, respectively, then u changes to $\alpha\,u$;
- if Ω is transformed to $F^{-1}(\Omega)$ with $F(x) = x/\alpha$, then $u \circ F$ solves an analog of (1) in $F^{-1}(\Omega)$ with the data $\alpha^2\,\varepsilon$, $\alpha\,\boldsymbol{b} \circ F$, $c \circ F$, $f \circ F$, and $u_b \circ F$.

Note that the discrete problem (19) without the nonlinear term d_h preserves these properties. To preserve the properties also when using the nonlinear term, the function γ_M has to satisfy

$$\begin{aligned}\gamma_M(\varepsilon, \boldsymbol{b}, c, f, u_b, \Omega, u_h) &= \gamma_M(\alpha\,\varepsilon, \alpha\,\boldsymbol{b}, \alpha\,c, \alpha\,f, u_b, \Omega, u_h)\\ &= \alpha\,\gamma_M(\varepsilon, \boldsymbol{b}, c, \alpha\,f, \alpha\,u_b, \Omega, \alpha\,u_h)\\ &= \alpha^{-1}\,\gamma_{F^{-1}(M)}(\alpha^2\,\varepsilon, \alpha\,\boldsymbol{b} \circ F, c \circ F, f \circ F, u_b \circ F, F^{-1}(\Omega), u_h \circ F)\end{aligned}$$

for any admissible data, $\alpha \neq 0$, and $u_h \in W_h$. We shall consider two choices of the scaling function γ_M: a global one independent of u_h and a local one depending on u_h. In the former case, one may set

$$\gamma_M = \gamma_0 \, \mathrm{diam}(\Omega)^{d/2} \left(\frac{\|f\|_{0,\Omega} \, \mathrm{diam}(\Omega)}{\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega} \, \mathrm{diam}(\Omega) + \|c\|_{0,\infty,\Omega} \, \mathrm{diam}(\Omega)^2} + \frac{\|u_b\|_{0,\partial\Omega}}{\mathrm{diam}(\Omega)^{1/2}} \right)^{-1} \tag{21}$$

with a positive constant γ_0. The local scaling can be defined by setting $\gamma_M = \beta \, h_M^{d/2} / |u_h|_{1,M}$ with a positive constant β if $|u_h|_{1,M} \neq 0$. Thus, we arrive at the following two formulas for the function $\tilde{\tau}_M$:

$$\tilde{\tau}_M = \beta \, h_M \, |\boldsymbol{b}_M|, \tag{22}$$

and

$$\tilde{\tau}_M(u_h) = \begin{cases} \dfrac{\beta \, h_M^{1+d/2} \, |\boldsymbol{b}_M|}{|u_h|_{1,M}} & \text{if } |u_h|_{1,M} \neq 0, \\ 0 & \text{if } |u_h|_{1,M} = 0, \end{cases} \tag{23}$$

where β is a positive real number independent of u_h and h. The parameter β depends on the data of (1) in case of (22) (*e.g.*, like γ_M in (21)), but it is independent of the data of (1) in case of (23). For these two choices of $\tilde{\tau}_M$, we shall investigate the properties of the discrete problem (19). Although the local scaling is likely to lead to better numerical results than the global one, we consider both variants since the choice (22) turns out to be more appealing for the analysis.

Remark 3.1.

- If $d = 2$ and $\boldsymbol{b}_M \neq \boldsymbol{0}$, one has $P_M = \boldsymbol{b}_M^\perp \otimes \boldsymbol{b}_M^\perp$ where $\boldsymbol{b}_M^\perp$ is a vector satisfying $\boldsymbol{b}_M^\perp \cdot \boldsymbol{b}_M = 0$ and $|\boldsymbol{b}_M^\perp| = 1$. Thus, in this case, the nonlinear stabilization term can be written in the form

$$d_h(w; u, v) = \sum_{M \in \mathscr{M}_h} (\tau_M^{\mathrm{sold}}(w) \, \kappa_M(\boldsymbol{b}_M^\perp \cdot \nabla u), \kappa_M(\boldsymbol{b}_M^\perp \cdot \nabla v))_M.$$

- It is useful for the analysis of the discrete problem to note that $\kappa_M(\boldsymbol{b}_M \cdot \nabla u) = \boldsymbol{b}_M \cdot \kappa_M \nabla u$ and $\kappa_M(P_M \nabla u) = P_M \kappa_M \nabla u$. Note also that $\|P_M\|_2 = 1$.

- Finally, if $\tilde{\tau}_M$ is defined by (23), then, using the stability of κ_M and $\boldsymbol{b}_M$ (18) and (16), respectively, and $\|P_M\|_2 = 1$, one obtains

$$\|\tau_M^{\text{sold}}(v)\|_{0,M} \leq C\, h_M^{1+d/2}\, \|\boldsymbol{b}\|_{0,\infty,M} \qquad \forall\, v \in H^1(\Omega),\ M \in \mathscr{M}_h. \tag{24}$$

In the analysis, the error will be measured using the following mesh-dependent norm

$$\|v\|_{\text{LPS}} := \Big(\varepsilon\, |v|_{1,\Omega}^2 + \|\sigma^{1/2}\, v\|_{0,\Omega}^2 + s_h(v,v)\Big)^{1/2},$$

and a term involving the crosswind derivative of the error. Note that integrating by parts gives

$$a(v,v) + s_h(v,v) = \|v\|_{\text{LPS}}^2 \qquad \forall\, v \in H_0^1(\Omega). \tag{25}$$

3.1. Well-posedness of the nonlinear discrete problem

This section studies the existence and uniqueness of solutions for the nonlinear discrete problem (19). The results of this section are valid also for $\sigma_0 = 0$.

Let us define the nonlinear operator $T_h : V_h \to V_h$ by

$$(T_h z_h, v_h) = a(z_h + \widetilde{u}_{bh}, v_h) + s_h(z_h + \widetilde{u}_{bh}, v_h) + d_h(z_h + \widetilde{u}_{bh}; z_h + \widetilde{u}_{bh}, v_h) - (f, v_h) \tag{26}$$

for any $z_h, v_h \in V_h$. Then $u_h \in W_h$ is a solution of (19) if and only if $u_h|_{\partial\Omega} = \widetilde{u}_{bh}|_{\partial\Omega}$ and

$$T_h(u_h - \widetilde{u}_{bh}) = 0,$$

or, equivalently, $u_h = \widetilde{u}_h + \widetilde{u}_{bh} \in W_h$ is a solution of (19) if $\widetilde{u}_h \in V_h$ and $T_h(\widetilde{u}_h) = 0$. Thus, our aim is to prove that the operator T_h has a zero in V_h. To this end, the properties of the form d_h shall be investigated first. As these properties are different with respect to the definition of $\tilde{\tau}_M$, we start supposing that $\tilde{\tau}_M$ is given by (22).

Lemma 3.2. *Let $\tilde{\tau}_M$ be defined by (22). Consider any $u, v, z \in W^{1,3}(\Omega)$ and set $w := u - v$. Then*

$$d_h(u;u,w) - d_h(v;v,w) \geq \frac{1}{7} \sum_{M \in \mathscr{M}_h} \tilde{\tau}_M\, \|\kappa_M(P_M \nabla w)\|_{0,3,M}^3 = \frac{1}{7}\, d_h(w;w,w), \tag{27}$$

$$\begin{aligned} |d_h(u;u,z) - d_h(v;v,z)| \leq \sum_{M \in \mathscr{M}_h} \tilde{\tau}_M\, &\big(\|\kappa_M(P_M \nabla u)\|_{0,3,M} + \|\kappa_M(P_M \nabla v)\|_{0,3,M}\big) \\ &\times \|\kappa_M(P_M \nabla w)\|_{0,3,M}\, \|\kappa_M(P_M \nabla z)\|_{0,3,M}. \end{aligned} \tag{28}$$

Proof. Let us denote

$$d_h(u;u,z) - d_h(v;v,z) = \sum_{M \in \mathscr{M}_h} N_M(u,v,z), \tag{29}$$

where

$$N_M(u,v,z) := \big(\tau_M^{\text{sold}}(u)\, \kappa_M(P_M \nabla u) - \tau_M^{\text{sold}}(v)\, \kappa_M(P_M \nabla v), \kappa_M(P_M \nabla z)\big)_M.$$

For $t \in [0,1]$, let $u^t := tu + (1-t)v$ and set

$$g(t) := \tilde{\tau}_M\, |\kappa_M(P_M \nabla u^t)|\, \kappa_M(P_M \nabla u^t), \qquad t \in [0,1].$$

Then

$$N_M(u,v,z) = \big(g(1) - g(0), \kappa_M(P_M \nabla z)\big)_M = \left(\int_0^1 g'(t)\, \mathrm{d}t, \kappa_M(P_M \nabla z)\right)_M.$$

Since

$$g'(t) = \tilde{\tau}_M\, \frac{\kappa_M(P_M \nabla u^t)}{|\kappa_M(P_M \nabla u^t)|}\, \kappa_M(P_M \nabla u^t) \cdot \kappa_M(P_M \nabla w) + \tilde{\tau}_M\, |\kappa_M(P_M \nabla u^t)|\, \kappa_M(P_M \nabla w), \tag{30}$$

one has

$$\begin{aligned}|g'(t)| &\le 2\,\tilde\tau_M\,|\kappa_M(P_M\nabla u^t)|\,|\kappa_M(P_M\nabla w)|\\ &\le 2\,\tilde\tau_M\,(t\,|\kappa_M(P_M\nabla u)| + (1-t)\,|\kappa_M(P_M\nabla v)|)\,|\kappa_M(P_M\nabla w)|,\end{aligned}$$

which implies (28). On the other hand, since multiplication of the first term on the right-hand side of (30) by $\kappa_M(P_M\nabla w)$ gives a non-negative expression, one obtains

$$N_M(u,v,w) \ge \left(\tilde\tau_M \int_0^1 |\kappa_M(P_M\nabla u^t)|\,\mathrm{d}t\,\kappa_M(P_M\nabla w), \kappa_M(P_M\nabla w)\right)_M. \tag{31}$$

Next, clearly

$$\int_0^1 |\kappa_M(P_M\nabla u^t)|\,\mathrm{d}t \ge \max_{i=1,\dots,d}\int_0^1 |t\,\kappa_M(P_M\nabla u)_i + (1-t)\,\kappa_M(P_M\nabla v)_i|\,\mathrm{d}t.$$

Denoting

$$I(a,b) = \int_0^1 |ta + (1-t)b|\,\mathrm{d}t, \qquad a,b\in\mathbb{R},$$

a direct computation gives

$$I(a,b) = \frac{|a|+|b|}{2} \quad \text{if } a\,b \ge 0, \qquad I(a,b) = \frac{1}{2}\,\frac{a^2+b^2}{|a|+|b|} \quad \text{if } a\,b < 0.$$

Thus, for any $a,b\in\mathbb{R}$, it follows

$$I(a,b) \ge \frac{|a|+|b|}{4} \ge \frac{|a-b|}{4}.$$

Consequently,

$$\int_0^1 |\kappa_M(P_M\nabla u^t)|\,\mathrm{d}t \ge \frac{1}{4}\max_{i=1,\dots,d}|\kappa_M(P_M\nabla w)_i| \ge \frac{1}{4\sqrt{d}}\,|\kappa_M(P_M\nabla w)| \ge \frac{1}{7}\,|\kappa_M(P_M\nabla w)|.$$

Combining this estimate with (31) and using (29) gives (27). □

Next, the properties of d_h are explored for the case that $\tilde\tau_M$ is defined by (23).

Lemma 3.3. *Let $\tilde\tau_M$ be defined by (23). Consider any $u,v,z\in W^{1,4}(\Omega)$. Then*

$$|d_h(u;v,z)| \le C\sum_{M\in\mathscr{M}_h} h_M^{1+d/2}\,\|\boldsymbol{b}\|_{0,\infty,M}\,\|\kappa_M(P_M\nabla v)\|_{0,4,M}\,\|\kappa_M(P_M\nabla z)\|_{0,4,M}, \tag{32}$$

$$\begin{aligned}|d_h(u;u,z) - d_h(v;v,z)| \le C\sum_{M\in\mathscr{M}_h} h_M^{1+d/2}\,\|\boldsymbol{b}\|_{0,\infty,M}\,\zeta_M(u,v)\times\\ \times\,(\|\kappa_M(P_M\nabla u)\|_{0,4,M} + \|\kappa_M(P_M\nabla v)\|_{0,4,M})\,\|\kappa_M(P_M\nabla z)\|_{0,4,M},\end{aligned} \tag{33}$$

where

$$\zeta_M(u,v) = \begin{cases}\dfrac{|u-v|_{1,M}}{|u|_{1,M}+|v|_{1,M}} & \text{if } |u|_{1,M}\ne 0 \text{ or } |v|_{1,M}\ne 0,\\ 0 & \text{if } |u|_{1,M} = |v|_{1,M} = 0.\end{cases}$$

Proof. Denoting

$$d_M(u;v,z) = \left(\tau_M^{\text{sold}}(u)\,\kappa_M(P_M\nabla v), \kappa_M(P_M\nabla z)\right)_M,$$

it is easy to realize that

$$d_h(u;v,z) = \sum_{M\in\mathscr{M}_h} d_M(u;v,z).$$

Applying Hölder's inequality yields

$$|d_M(u;v,z)| \le \|\tau_M^{\text{sold}}(u)\|_{0,M}\,\|\kappa_M(P_M\nabla v)\|_{0,4,M}\,\|\kappa_M(P_M\nabla z)\|_{0,4,M},$$

which, using (24), gives

$$|d_M(u;v,z)| \le C\,h_M^{1+d/2}\,\|\boldsymbol{b}\|_{0,\infty,M}\,\|\kappa_M(P_M\nabla v)\|_{0,4,M}\,\|\kappa_M(P_M\nabla z)\|_{0,4,M}, \tag{34}$$

thus proving (32). Now it will be shown that

$$\begin{aligned}|d_M(u;u,z) - d_M(v;v,z)| &\le C\,h_M^{1+d/2}\,\|\boldsymbol{b}\|_{0,\infty,M}\,\zeta_M(u,v)\\ &\quad\times\left(\|\kappa_M(P_M\nabla u)\|_{0,4,M} + \|\kappa_M(P_M\nabla v)\|_{0,4,M}\right)\|\kappa_M(P_M\nabla z)\|_{0,4,M}.\end{aligned} \tag{35}$$

If $|u|_{1,M}=0$ or $|v|_{1,M}=0$, then (35) is a particular case of (34). Thus, it suffices to consider the case $|u|_{1,M}\neq 0$, $|v|_{1,M}\neq 0$. Denoting $\xi(x)=|x|\,x$, one obtains

$$\begin{aligned}d_M(u;u,z) - d_M(v;v,z) &= \frac{\beta\,h_M^{1+d/2}\,|\boldsymbol{b}_M|}{|u|_{1,M}}\left(\xi(\kappa_M(P_M\nabla u)) - \xi(\kappa_M(P_M\nabla v)), \kappa_M(P_M\nabla z)\right)_M\\ &\quad+\beta\,h_M^{1+d/2}\,|\boldsymbol{b}_M|\left(\frac{1}{|u|_{1,M}} - \frac{1}{|v|_{1,M}}\right)\left(\xi(\kappa_M(P_M\nabla v)), \kappa_M(P_M\nabla z)\right)_M.\end{aligned} \tag{36}$$

The integral terms on M possess the same structure as the term $N_M(u,v,z)$ in the proof of Lemma 3.2 (the second term corresponds to $N_M(0,v,z)$). They are estimated using the same technique, only with a different Hölder inequality. Then, (16) is applied to $\|\kappa_M(P_M\nabla(u-v))\|_{0,M}$ resp. $\|\kappa_M(P_M\nabla v)\|_{0,M}$. Furthermore, the first inequality from (18) is employed. To finish the estimate of the second term in (36), the triangle inequality is used. One obtains

$$\begin{aligned}|d_M(u;u,z) - d_M(v;v,z)| &\le C\,h_M^{1+d/2}\,\|\boldsymbol{b}\|_{0,\infty,M}\,\frac{|u-v|_{1,M}}{|u|_{1,M}}\\ &\quad\times\left(\|\kappa_M(P_M\nabla u)\|_{0,4,M} + \|\kappa_M(P_M\nabla v)\|_{0,4,M}\right)\|\kappa_M(P_M\nabla z)\|_{0,4,M}.\end{aligned}$$

The same type of inequality follows by interchanging u and v. Then, using the sharper of these two estimates and $\min\{|u|_{1,M}^{-1}, |v|_{1,M}^{-1}\} \le 2/(|u|_{1,M}+|v|_{1,M})$ gives (35). □

The properties of the operator T_h, namely its monotonicity and local Lipschitz continuity, follow now by the results of the two previous lemmas and the representation of the LPS norm (25).

Lemma 3.4. *If $\tilde\tau_M$ is defined by (22), then the operator T_h defined in (26) is locally Lipschitz-continuous and strongly monotone,* i.e., *it satisfies*

$$(T_h w_h - T_h z_h, w_h - z_h) \ge \|w_h - z_h\|_{\text{LPS}}^2 + \frac{1}{7}\sum_{M\in\mathscr{M}_h}\tilde\tau_M\,\|\kappa_M(P_M\nabla(w_h - z_h))\|_{0,3,M}^3 \tag{37}$$

for all $w_h, z_h \in V_h$. If $\tilde\tau_M$ is defined by (23), then the operator T_h is Lipschitz-continuous and it satisfies

$$(T_h z_h, z_h) \ge \frac{\varepsilon}{2}\,|z_h|_{1,\Omega}^2 - C_0\left(\|\widetilde{u}_{bh}\|_{1,\Omega}^2 + \|f\|_{0,\Omega}^2\right) \tag{38}$$

for all $z_h \in V_h$, where $C_0 > 0$ depends on ε, $\boldsymbol{b}$, and c, but not on z_h, h, and σ_0.

Proof. Let us define the operators $A_h, N_h : V_h \to V_h$ by

$$(A_h z_h, v_h) = a(z_h, v_h) + s_h(z_h, v_h) \qquad \forall\, z_h, v_h \in V_h,$$
$$(N_h z_h, v_h) = d_h(z_h + \widetilde{u}_{bh}; z_h + \widetilde{u}_{bh}, v_h) \qquad \forall\, z_h, v_h \in V_h.$$

Then, for any $w_h, z_h \in V_h$, there holds

$$T_h w_h - T_h z_h = A_h(w_h - z_h) + N_h w_h - N_h z_h.$$

The operator A_h is linear on a finite-dimensional space and hence it is Lipschitz continuous. Thus, the (local) Lipschitz-continuity of T_h follows from (28), (33), and the equivalence of norms on finite-dimensional spaces. The strong monotonicity (37) follows from (25) and (27). Finally, let $\tilde{\tau}_M$ be defined by (23). In view of (25), it holds

$$\begin{aligned}(T_h z_h, z_h) = {} & \|z_h\|^2_{\mathrm{LPS}} + d_h(z_h + \widetilde{u}_{bh}; z_h, z_h) \\ & + a(\widetilde{u}_{bh}, z_h) + s_h(\widetilde{u}_{bh}, z_h) + d_h(z_h + \widetilde{u}_{bh}; \widetilde{u}_{bh}, z_h) - (f, z_h). \end{aligned} \tag{39}$$

Applying (32), (10), (16), (18), (4), and (5), one obtains

$$|d_h(z_h + \widetilde{u}_{bh}; \widetilde{u}_{bh}, z_h)| \le C\, h\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, |\widetilde{u}_{bh}|_{1,\Omega}\, |z_h|_{1,\Omega}.$$

The same estimate also holds for $s_h(\widetilde{u}_{bh}, z_h)$. Using the fact that $d_h(z_h + \widetilde{u}_{bh}; z_h, z_h) \ge 0$ and applying the Cauchy-Schwarz inequality to the third and last term on the right-hand side of (39), one derives

$$(T_h z_h, z_h) \ge \varepsilon\, |z_h|^2_{1,\Omega} - (\varepsilon + C\, \|\boldsymbol{b}\|_{0,\infty,\Omega} + \|c\|_{0,\infty,\Omega})\, \|\widetilde{u}_{bh}\|_{1,\Omega}\, \|z_h\|_{1,\Omega} - \|f\|_{0,\Omega}\, \|z_h\|_{0,\Omega}.$$

Now, employing the Poincaré and Young inequalities, one obtains (38). □

To prove that the discrete problem (19) has at least one solution, we shall use the following simple consequence of Brouwer's fixed-point theorem, whose proof can be found in [32], p. 164, Lemma 1.4.

Lemma 3.5. *Let X be a finite-dimensional Hilbert space with inner product $(\cdot,\cdot)$ and norm $\|\cdot\|$. Let $P : X \to X$ be a continuous mapping and $K > 0$ a real number such that $(Px, x) > 0$ for any $x \in X$ with $\|x\| = K$. Then there exists $x \in X$ such that $\|x\| \le K$ and $Px = 0$.*

Collecting the previous results, the main result of this section can be stated now, namely, the well-posedness of the problem (19).

Theorem 3.6. *If $\tilde{\tau}_M$ is defined by (22) or (23), then the problem (19) has a solution. If $\tilde{\tau}_M$ is defined by (22), the solution of (19) is unique.*

Proof. If $\tilde{\tau}_M$ is defined by (22), then it follows from the strong monotonicity (37) that, for any $z_h \in V_h$,

$$(T_h z_h, z_h) \ge \|z_h\|^2_{\mathrm{LPS}} + (T_h 0, z_h) \ge \varepsilon\, |z_h|^2_{1,\Omega} - \|T_h 0\|_{0,\Omega}\, \|z_h\|_{0,\Omega}.$$

Thus, using Young's inequality and the equivalence of norms in the space V_h one gets

$$(T_h z_h, z_h) \ge C_1\, \|z_h\|^2_{0,\Omega} - C_2,$$

where C_1, C_2 are positive constants that depend on h and the data of (1), but not on z_h and σ_0. According to (38), the same inequality holds if $\tilde{\tau}_M$ is defined by (23). Thus, in view of Lemma 3.5 with any $K > \sqrt{C_2/C_1}$, the operator T_h has a zero and hence the problem (19) has a solution. The uniqueness in the case that $\tilde{\tau}_M$ is defined by (22) follows from the strong monotonicity (37). □

3.2. Error estimates

For the analysis of the methods introduced in Section 3, we will need an appropriate interpolation operator. An important tool for the construction of such an operator is provided by the following result, whose proof can be found in [25], Lemma 1.

Lemma 3.7. *Let us suppose the inf-sup condition (9) to be satisfied. Then, there exists an operator* $\varrho_h : L^2(\Omega) \to V_h$ *such that, for any* $v, w \in L^2(\Omega)$*, the estimates*

$$|(v - \varrho_h v, w)| \le C \sum_{M \in \mathscr{M}_h} \|v\|_{0,M} \, \|\kappa_M w\|_{0,M}, \tag{40}$$

$$|\varrho_h v|^2_{1,M} + h_M^{-2} \, \|\varrho_h v\|^2_{0,M} \le C \sum_{\substack{M' \in \mathscr{M}_h, \\ M \cap M' \neq \emptyset}} h_{M'}^{-2} \, \|v\|^2_{0,M'} \qquad \forall \; M \in \mathscr{M}_h \tag{41}$$

are valid. Consequently, for any $\alpha \in \mathbb{R}$*, it holds*

$$\sum_{M \in \mathscr{M}_h} h_M^{\alpha} \, (|\varrho_h v|^2_{1,M} + h_M^{-2} \, \|\varrho_h v\|^2_{0,M}) \le C \sum_{M \in \mathscr{M}_h} h_M^{\alpha - 2} \, \|v\|^2_{0,M}, \tag{42}$$

where the constant C *is independent of* v *and* h *but can depend on* α*.*

With the operators i_h and ϱ_h, an operator $r_h \in \mathscr{L}(C(\overline{\Omega}), W_h) \cap \mathscr{L}(C(\overline{\Omega}) \cap H_0^1(\Omega), V_h)$ is defined by

$$r_h v := i_h v + \varrho_h (v - i_h v). \tag{43}$$

To formulate the interpolation properties of r_h, it is convenient to introduce the mesh dependent norm

$$\|v\|_{1,h} = \left(\sum_{M \in \mathscr{M}_h} \{ |v|^2_{1,M} + h_M^{-2} \, \|v\|^2_{0,M} \} \right)^{1/2}.$$

Then, using (41), the geometrical hypotheses (4) and (5), and the approximation property of i_h (11), one obtains

$$\|v - r_h v\|_{1,h} \le C \, \|v - i_h v\|_{1,h} \le \tilde{C} \, h^k \, |v|_{k+1,\Omega} \qquad \forall \; v \in H^{k+1}(\Omega), \; k = 1, \ldots, l, \tag{44}$$

and consequently

$$|v - r_h v|_{1,\Omega} + h^{-1} \, \|v - r_h v\|_{0,\Omega} \le C \, h^k \, |v|_{k+1,\Omega} \qquad \forall \; v \in H^{k+1}(\Omega), \; k = 1, \ldots, l. \tag{45}$$

The derivation of the error estimates will be based on the following two lemmas. The first one states an interpolation error estimate and the second one states a bound on the nonlinear form d_h.

Lemma 3.8. *Let* $u \in H^{k+1}(\Omega)$ *for some* $k \in \{1, \ldots, l\}$*, and let* $\eta := u - r_h u$*. Then, for any* $v_h \in V_h \setminus \{0\}$*, the following estimate holds*

$$\begin{aligned} \|\eta\|_{\mathrm{LPS}} + \frac{a(\eta, v_h) + s_h(\eta, v_h) - s_h(u, v_h)}{\|v_h\|_{\mathrm{LPS}}} \\ \le C \left(\varepsilon + h \, \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2 \, \|\sigma\|_{0,\infty,\Omega} + h^2 \, |\boldsymbol{b}|^2_{1,\infty,\Omega} \, \sigma_0^{-1} \right)^{1/2} h^k \, |u|_{k+1,\Omega}. \end{aligned} \tag{46}$$

Proof. Since, in view of (5), (16), (18), and the definition of τ_M (20)

$$\|v\|_{\mathrm{LPS}} \le C \left(\varepsilon + h \, \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2 \, \|\sigma\|_{0,\infty,\Omega} \right)^{1/2} \|v\|_{1,h} \qquad \forall \; v \in H^1(\Omega),$$

it follows from (44) that

$$\|\eta\|_{\mathrm{LPS}} \leq C \left(\varepsilon + h \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2 \|\sigma\|_{0,\infty,\Omega}\right)^{1/2} h^k |u|_{k+1,\Omega}.$$

Next, for any $v_h \in V_h \setminus \{0\}$, integration by parts gives

$$(\boldsymbol{b} \cdot \nabla \eta, v_h) = -(\eta, \boldsymbol{b} \cdot \nabla v_h) - ((\nabla \cdot \boldsymbol{b})\, \eta, v_h).$$

Thus, applying the Cauchy-Schwarz inequality and (45), it follows that

$$a(\eta, v_h) + s_h(\eta, v_h) \leq \left(\|\eta\|_{\mathrm{LPS}} + C\, |\boldsymbol{b}|_{1,\infty,\Omega}\, \sigma_0^{-1/2}\, h^{k+1}\, |u|_{k+1,\Omega}\right) \|v_h\|_{\mathrm{LPS}} - (\eta, \boldsymbol{b} \cdot \nabla v_h).$$

The use of (40), the approximation property of i_h (11), (4), and (5) lead to

$$\begin{aligned}(\eta, \boldsymbol{b} \cdot \nabla v_h) &\leq C \sum_{M \in \mathscr{M}_h} \|u - i_h u\|_{0,M}\, \|\kappa_M(\boldsymbol{b} \cdot \nabla v_h)\|_{0,M} \\ &\leq C\, h^k\, |u|_{k+1,\Omega} \left(\sum_{M \in \mathscr{M}_h} h_M^2\, \|\kappa_M(\boldsymbol{b} \cdot \nabla v_h)\|_{0,M}^2\right)^{1/2}.\end{aligned}$$

Applying (16), (18), (20), and the inverse inequality (8), one derives

$$\begin{aligned}\|\kappa_M(\boldsymbol{b} \cdot \nabla v_h)\|_{0,M} &\leq \|\kappa_M((\boldsymbol{b} - \boldsymbol{b}_M) \cdot \nabla v_h)\|_{0,M} + \|\kappa_M(\boldsymbol{b}_M \cdot \nabla v_h)\|_{0,M} \\ &\leq C\, |\boldsymbol{b}|_{1,\infty,M}\, \|v_h\|_{0,M} + \tau_0^{-1/2} \left(\varepsilon + h_M \|\boldsymbol{b}\|_{0,\infty,M}\right)^{1/2} h_M^{-1}\, \tau_M^{1/2}\, \|\kappa_M(\boldsymbol{b}_M \cdot \nabla v_h)\|_{0,M},\end{aligned}$$

which leads to the estimate

$$(\eta, \boldsymbol{b} \cdot \nabla v_h) \leq C \left(\varepsilon + h \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2 |\boldsymbol{b}|_{1,\infty,\Omega}^2\, \sigma_0^{-1}\right)^{1/2} h^k\, |u|_{k+1,\Omega}\, \|v_h\|_{\mathrm{LPS}}.$$

Finally, using (17), (18), (20), and the geometrical hypotheses (4) and (5), one obtains

$$s_h(u,u) \leq \sum_{M \in \mathscr{M}_h} \tau_M\, |\boldsymbol{b}_M|^2\, \|\kappa_M \nabla u\|_{0,M}^2 \leq C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{2k+1}\, |u|_{k+1,\Omega}^2,$$

and hence

$$s_h(u, v_h) \leq \sqrt{s_h(u,u)}\, \sqrt{s_h(v_h, v_h)} \leq C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}^{1/2}\, h^{k+1/2}\, |u|_{k+1,\Omega}\, \|v_h\|_{\mathrm{LPS}},$$

which completes the proof. □

Lemma 3.9. *For any $w_h \in W_h$ and $u, v \in H^{k+1}(\Omega)$ with $k \in \{1, \ldots, l\}$, it holds*

$$d_h(w_h; r_h u, r_h v) \leq C\, h^{2k-d/2} \left(\max_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(w_h)\|_{0,M}\right) |u|_{k+1,\Omega}\, |v|_{k+1,\Omega}. \tag{47}$$

Proof. The application of Hölder's inequality and (10) lead to

$$\begin{aligned}d_h(w_h; r_h u, r_h v) &\leq \sum_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(w_h)\|_{0,M}\, \|\kappa_M(P_M \nabla(r_h u))\|_{0,4,M}\, \|\kappa_M(P_M \nabla(r_h v))\|_{0,4,M} \\ &\leq C \sum_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(w_h)\|_{0,M}\, h_M^{-d/2}\, \|\kappa_M(P_M \nabla(r_h u))\|_{0,M}\, \|\kappa_M(P_M \nabla(r_h v))\|_{0,M} \\ &\leq C \left(\max_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(w_h)\|_{0,M}\right) \left(\sum_{M \in \mathscr{M}_h} h_M^{-d/2}\, \|\kappa_M(P_M \nabla(r_h u))\|_{0,M}^2\right)^{1/2} \\ &\quad \times \left(\sum_{M \in \mathscr{M}_h} h_M^{-d/2}\, \|\kappa_M(P_M \nabla(r_h v))\|_{0,M}^2\right)^{1/2}.\end{aligned} \tag{48}$$

Let us estimate the term with u; the term with v can be treated analogously. Using (16) and (17), for $u \in H^{k+1}(\Omega)$ with $k \in \{1,\dots,l\}$ there holds

$$\begin{aligned}\|\kappa_M(P_M\nabla(r_h u))\|_{0,M} &\le \|\kappa_M(P_M\nabla u)\|_{0,M} + \|\kappa_M(P_M\nabla(u-r_h u))\|_{0,M}\\ &\le C\,h_M^k\,|u|_{k+1,M} + C\,|u-r_h u|_{1,M}. \end{aligned} \tag{49}$$

According to (42), one has for any $\alpha \in \mathbb{R}$

$$\begin{aligned}\sum_{M\in\mathscr{M}_h} h_M^\alpha\,|u-r_h u|_{1,M}^2 &\le 2\sum_{M\in\mathscr{M}_h} h_M^\alpha\,|u-i_h u|_{1,M}^2 + 2\sum_{M\in\mathscr{M}_h} h_M^\alpha\,|\varrho_h(u-i_h u)|_{1,M}^2\\ &\le C\sum_{M\in\mathscr{M}_h} h_M^\alpha\,(|u-i_h u|_{1,M}^2 + h_M^{-2}\,\|u-i_h u\|_{0,M}^2),\end{aligned}$$

and hence it follows from the approximation property of i_h (11), (4), and (5) that, for $\alpha \ge -2$,

$$\sum_{M\in\mathscr{M}_h} h_M^\alpha\,\|\kappa_M(P_M\nabla(r_h u))\|_{0,M}^2 \le C\,h^{2\,k+\alpha}\,|u|_{k+1,\Omega}^2. \tag{50}$$

Inserting (50) with $\alpha = -d/2$ into (48), the statement of the lemma is proved. □

We are now in position to prove the first error estimate. The following theorem states the error estimate in the case $\tilde\tau_M$ is given by (22).

Theorem 3.10. *Let $\tilde\tau_M$ be defined by (22). Let the weak solution of (1) satisfy $u \in H^{k+1}(\Omega)$ for some $k \in \{1,\dots,l\}$. Let $\widetilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\widetilde{u}_{bh} = i_h\widetilde{u}_b$. Then the solution u_h of the local projection discretization (19) satisfies the error estimate*

$$\begin{aligned}&\|u-u_h\|_{\mathrm{LPS}} + \left(\sum_{M\in\mathscr{M}_h}\tilde\tau_M\,\|\kappa_M(P_M\nabla(u-u_h))\|_{0,3,M}^3\right)^{1/2}\\ &\le C\Big\{\varepsilon + h\,\|\boldsymbol{b}\|_{0,\infty,\Omega}\,(1+\beta\,h^{k-d/2}\,|u|_{k+1,\Omega}) + h^2\,\big(\|\sigma\|_{0,\infty,\Omega} + |\boldsymbol{b}|_{1,\infty,\Omega}^2\,\sigma_0^{-1}\big)\Big\}^{1/2} h^k\,|u|_{k+1,\Omega}.\end{aligned}$$

If $u \in W^{k+1,\infty}(\Omega)$ with $k \in \{1,\dots,l\}$, then

$$\begin{aligned}&\|u-u_h\|_{\mathrm{LPS}} + \left(\sum_{M\in\mathscr{M}_h}\tilde\tau_M\,\|\kappa_M(P_M\nabla(u-u_h))\|_{0,3,M}^3\right)^{1/2}\\ &\le C\Big\{\varepsilon + h\,\|\boldsymbol{b}\|_{0,\infty,\Omega}\,(1+\beta\,h^{k}\,|u|_{k+1,\infty,\Omega}) + h^2\,\big(\|\sigma\|_{0,\infty,\Omega} + |\boldsymbol{b}|_{1,\infty,\Omega}^2\,\sigma_0^{-1}\big)\Big\}^{1/2} h^k\,|u|_{k+1,\Omega}.\end{aligned}$$

Proof. The error $u-u_h$ is split into the interpolation error $\eta := u - r_h u$ and the discrete error $e_h := u_h - r_h u$. Then $e_h \in V_h$ and also $r_h u - \widetilde{u}_{bh} \in V_h$. From the monotonicity (37) it follows with the discrete problem (19) and the continuous problem (14) that

$$\begin{aligned}&\|e_h\|_{\mathrm{LPS}}^2 + \frac{1}{7}\sum_{M\in\mathscr{M}_h}\tilde\tau_M\,\|\kappa_M(P_M\nabla e_h)\|_{0,3,M}^3 \le (T_h(u_h-\widetilde{u}_{bh}) - T_h(r_h u-\widetilde{u}_{bh}), e_h)\\ &= a(u_h,e_h) + s_h(u_h,e_h) + d_h(u_h;u_h,e_h) - (T_h(r_h u-\widetilde{u}_{bh}),e_h)\\ &= (f,e_h) - (T_h(r_h u-\widetilde{u}_{bh}),e_h)\\ &= a(u,e_h) - a(r_h u,e_h) - s_h(r_h u,e_h) - d_h(r_h u;r_h u,e_h)\\ &= a(\eta,e_h) + s_h(\eta,e_h) - s_h(u,e_h) - d_h(r_h u;r_h u,e_h).\end{aligned}$$

The first three terms on the right-hand side can be estimated using (46). To bound the nonlinear term, Hölder's and Young's inequalities are applied to conclude

$$\begin{aligned} d_h(r_h u; r_h u, e_h) &\le \{d_h(r_h u; r_h u, r_h u)\}^{\frac{2}{3}} \{d_h(e_h; e_h, e_h)\}^{\frac{1}{3}} \\ &\le 2\, d_h(r_h u; r_h u, r_h u) + \frac{3}{70}\, d_h(e_h; e_h, e_h). \end{aligned} \tag{51}$$

Then (47), (49), the bound of h_M (5), (18), and (45) yield

$$d_h(r_h u; r_h u, r_h u) \le C\, \beta\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{3\,k+1-d/2}\, |u|^3_{k+1,\Omega}. \tag{52}$$

Therefore,

$$\begin{aligned} &\|e_h\|^2_{\mathrm{LPS}} + \sum_{M\in\mathscr{M}_h} \tilde{\tau}_M\, \|\kappa_M(P_M \nabla e_h)\|^3_{0,3,M} \\ &\le C \left\{ \varepsilon + h\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, (1 + \beta\, h^{k-d/2}\, |u|_{k+1,\Omega}) + h^2\, \|\sigma\|_{0,\infty,\Omega} + h^2\, |\boldsymbol{b}|^2_{1,\infty,\Omega}\, \sigma_0^{-1} \right\} h^{2\,k}\, |u|^2_{k+1,\Omega}. \end{aligned} \tag{53}$$

Next, to estimate the interpolation error, for any $p \in [1,6]$, it follows from the commutation property of κ_M and P_M, the estimate of the $L^p(M)$ norm by the $L^2(M)$ norm (10), (15), and (13) that

$$\begin{aligned} \|\kappa_M(P_M\nabla\eta)\|_{0,p,M} &\le \|\nabla\eta - \pi_M\nabla\eta\|_{0,p,M} \\ &\le \|\nabla(u - i_h u)\|_{0,p,M} + \|\nabla(i_h u - r_h u) - \pi_M\nabla\eta\|_{0,p,M} \\ &\le |u - i_h u|_{1,p,M} + C\, h_M^{\frac{d}{p}-\frac{d}{2}}\, \|\nabla(i_h u - r_h u) - \pi_M\nabla\eta\|_{0,M} \\ &\le |u - i_h u|_{1,p,M} + \tilde{C}\, h_M^{\frac{d}{p}-\frac{d}{2}}\, \left(|\varrho_h(u - i_h u)|_{1,M} + |u - i_h u|_{1,M} \right) \\ &\le \bar{C}\, h_M^{k+\frac{d}{p}-\frac{d}{2}}\, |u|_{k+1,M} + \tilde{C}\, h_M^{\frac{d}{p}-\frac{d}{2}}\, |\varrho_h(u - i_h u)|_{1,M}. \end{aligned} \tag{54}$$

Then, applying (54), (22), (5), (18), (41), (11), (4), and (6), one derives

$$\sum_{M\in\mathscr{M}_h} \tilde{\tau}_M\, \|\kappa_M(P_M\nabla\eta)\|^3_{0,3,M} \le C\, \beta\, h\, \|\boldsymbol{b}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} h_M^{3\,k-d/2}\, |u|^3_{k+1,M}. \tag{55}$$

Thus, combining (53), (55), and (46), the first estimate of the theorem follows.

If $u \in W^{k+1,\infty}(\Omega)$ with $k \in \{1,\dots,l\}$, then local norms of Sobolev spaces with $p = 2$ can be estimated with norms of Sobolev spaces with $p = \infty$, thereby gaining powers of h from the smallness of the local domain: $|u|_{k+1,M} \le C\, h_M^{d/2}\, |u|_{k+1,\infty,M}$ for any $M \in \mathscr{M}_h$. Hence, it follows from (55) and the geometrical hypotheses (4) and (5) that

$$\sum_{M\in\mathscr{M}_h} \tilde{\tau}_M\, \|\kappa_M(P_M\nabla\eta)\|^3_{0,3,M} \le C\, \beta\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{3\,k+1}\, |u|_{k+1,\infty,\Omega}\, |u|^2_{k+1,\Omega}.$$

Furthermore, using (41), (11), and (4), one gets

$$|u - r_h u|_{1,M} \le C \sum_{\substack{M'\in\mathscr{M}_h, \\ M\cap M' \neq \emptyset}} h_{M'}^k\, |u|_{k+1,M'} \le \tilde{C}\, h^{k+d/2}\, |u|_{k+1,\infty,\Omega} \qquad \forall\, M \in \mathscr{M}_h.$$

Therefore, according to (47) and (49),

$$d_h(r_h u; r_h u, r_h u) \le C\, \beta\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{3\,k+1}\, |u|_{k+1,\infty,\Omega}\, |u|^2_{k+1,\Omega}, \tag{56}$$

which implies the second estimate of the theorem. □

Remark 3.11. Theorem 3.10 implies, in particular, the following convergence estimates in the convection-dominated case $\varepsilon < h$: If $u \in H^2(\Omega)$, then

$$\|u - u_h\|_{\mathrm{LPS}} \leq C_0\, h^{2-d/4}\, (h^{(d-2)/4} + |u|_{2,\Omega}^{1/2})\, |u|_{2,\Omega},$$

where C_0 depends on the data of the problem. If $u \in W^{2,\infty}(\Omega)$, then

$$\|u - u_h\|_{\mathrm{LPS}} \leq C_0\, h^{3/2}\, (1 + h^{1/2}\, |u|_{2,\infty,\Omega}^{1/2})\, |u|_{2,\Omega}.$$

If $u \in H^{k+1}(\Omega)$ with $k \in \{2, \ldots, l\}$, then

$$\|u - u_h\|_{\mathrm{LPS}} \leq C_0\, h^{k+1/2}\, (1 + h^{(2\,k-d)/4}\, |u|_{k+1,\Omega}^{1/2})\, |u|_{k+1,\Omega}.$$

Remark 3.12. A situation of practical interest is that the convective field $\boldsymbol{b}$ arises from a finite element approximation of the Navier-Stokes equations. In this case, a necessary condition for a uniform convergence of $\|\boldsymbol{b}\|_{1,\infty,\Omega}$ with respect to h is that the exact velocity is sufficiently regular. This condition might not be fulfilled, *e.g.*, if the domain possesses re-entrant corners, and therefore estimates involving weaker norms of $\boldsymbol{b}$ are also of interest. Changing the arguments in the proof of Lemma 3.8 slightly, one obtains, *e.g.*, the following result

$$\begin{aligned}
&\|u - u_h\|_{\mathrm{LPS}} + \left(\sum_{M \in \mathscr{M}_h} \tilde{\tau}_M\, \|\kappa_M (P_M \nabla (u - u_h))\|_{0,3,M}^3 \right)^{1/2} \\
&\leq C \Big\{ \varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}^2\, \sigma_0^{-1} + h\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, (1 + \beta\, h^{k-d/2}\, |u|_{k+1,\Omega}) \\
&\quad + h^{2-\frac{d}{2}} \max_{M \in \mathcal{M}_h} \|\nabla \cdot \boldsymbol{b}\|_{0,4,M}^2\, \sigma_0^{-1} + h^2\, \|\sigma\|_{0,\infty,\Omega} \Big\}^{1/2} h^k\, |u|_{k+1,\Omega}.
\end{aligned} \tag{57}$$

If the norms of $\boldsymbol{b}$ in (57) are still too strong, one can use the discrete character of a computed convection field $\boldsymbol{b}$ and apply inverse inequalities to derive estimates involving the weaker norms $\|\boldsymbol{b}\|_{1,\Omega}$ and $\|\nabla \cdot \boldsymbol{b}\|_{0,\Omega}$. However, the relaxation of the regularity assumption on $\boldsymbol{b}$ in the error bounds is accompanied with a reduction of the order of convergence, *e.g.*, the order of convergence of (57) is reduced by 1/2 compared with the orders given in the previous remark.

Remark 3.13. The right-hand sides of the estimates in Theorem 3.10 can be stated in terms of local (semi)norms of the data and of the solution on macro-elements multiplied by diameters of the macro-elements. However, due to the use of the interpolation operator r_h, such estimates are more complicated than usually. For example, a counterpart of (52) using local quantities has the form

$$d_h(r_h u; r_h u, r_h u) \leq C\, \beta \sum_{M \in \mathscr{M}_h} \|\boldsymbol{b}\|_{0,\infty,M}\, h_M^{1-d/2} \left(\sum_{\substack{M' \in \mathscr{M}_h, \\ M \cap M' \neq \emptyset}} h_{M'}^{2k}\, |u|_{k+1,M'}^2 \right)^{3/2}.$$

Therefore, for clarity, we decided to state the estimates in terms of global quantities.

We end this section by presenting the error estimate in the case $\tilde{\tau}_M$ is defined by (23).

Theorem 3.14. *Let $\tilde{\tau}_M$ be defined by (23). Let the weak solution of (1) satisfy $u \in H^{k+1}(\Omega)$ for some $k \in \{1, \ldots, l\}$. Let $\widetilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\widetilde{u}_{bh} = i_h \widetilde{u}_b$. Then the solution u_h of the local projection discretization (19) satisfies the error estimate*

$$\begin{aligned}
&\|u - u_h\|_{\mathrm{LPS}} + (d_h(u_h; u - u_h, u - u_h))^{1/2} \\
&\qquad \leq C \left(\varepsilon + h\, \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2\, \|\sigma\|_{0,\infty,\Omega} + h^2\, |\boldsymbol{b}|_{1,\infty,\Omega}^2\, \sigma_0^{-1} \right)^{1/2} h^k\, |u|_{k+1,\Omega}.
\end{aligned}$$

Proof. Set again $\eta := u - r_h u$ and $e_h := u_h - r_h u$. From (19) and (14), it follows that

$$\begin{aligned} a(e_h, e_h) + s_h(e_h, e_h) + d_h(u_h; u_h, e_h) \\ &= a(u_h, e_h) + s_h(u_h, e_h) + d_h(u_h; u_h, e_h) - a(r_h u, e_h) - s_h(r_h u, e_h) \\ &= a(\eta, e_h) + s_h(\eta, e_h) - s_h(u, e_h). \end{aligned}$$

Thus, in view of the representation of the LPS norm (25), one gets

$$\|e_h\|_{\mathrm{LPS}}^2 + d_h(u_h; e_h, e_h) = a(\eta, e_h) + s_h(\eta, e_h) - s_h(u, e_h) - d_h(u_h; r_h u, e_h).$$

The first three terms on the right-hand side can be estimated using (46). To bound the nonlinear term, Hölder's and Young's inequalities are again applied

$$d_h(u_h; r_h u, e_h) \leq \sqrt{d_h(u_h; r_h u, r_h u)}\, \sqrt{d_h(u_h; e_h, e_h)} \leq d_h(u_h; r_h u, r_h u) + \frac{1}{4}\, d_h(u_h; e_h, e_h). \tag{58}$$

Using (47), (24), and (5), one obtains

$$d_h(u_h; r_h u, r_h u) \leq C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{2\,k+1}\, |u|_{k+1,\Omega}^2. \tag{59}$$

Therefore,

$$\|e_h\|_{\mathrm{LPS}}^2 + d_h(u_h; e_h, e_h) \leq C\, \left(\varepsilon + h\, \|\boldsymbol{b}\|_{0,\infty,\Omega} + h^2\, \|\sigma\|_{0,\infty,\Omega} + h^2\, |\boldsymbol{b}|_{1,\infty,\Omega}^2\, \sigma_0^{-1}\right) h^{2\,k}\, |u|_{k+1,\Omega}^2.$$

Note that an application of the triangle inequality gives

$$d_h(u_h; u - u_h, u - u_h) \leq 2\, d_h(u_h; \eta, \eta) + 2\, d_h(u_h; e_h, e_h). \tag{60}$$

It follows from Hölder's inequality, (24), (54), (42) with $\alpha = 0$, (11), (4), and (5), that

$$d_h(u_h; \eta, \eta) \leq \sum_{M \in \mathscr{M}_h} \|\tau_M^{\mathrm{sold}}(u_h)\|_{0,M}\, \|\kappa_M (P_M \nabla \eta)\|_{0,4,M}^2 \leq C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h^{2\,k+1}\, |u|_{k+1,\Omega}^2. \tag{61}$$

Finally, using the triangle inequality and the estimate (46), the statement of the theorem follows. □

Remark 3.15. Theorems 3.10 and 3.14 prove the convergence of the method in the LPS norm plus an extra term involving the crosswind derivative of the error. Hence, these estimates give, essentially, an extra control of the whole gradient of the error.

4. The time-dependent problem

We now move on to the study of the time-dependent problem (3). A weak form of problem (3) reads as follows: find $u \in L^2(0,T; H^1(\Omega)) \cap H^1(0,T; L^2(\Omega))$ such that $u = u_b$ on $[0,T] \times \partial\Omega$, $u(0,\cdot) = u_0$ and

$$(u_t, v) + a(u, v) = (f, v) \quad \forall\, v \in H_0^1(\Omega), \quad \text{for almost every } t \in (0,T]. \tag{62}$$

To avoid technicalities in the analysis, it is assumed that the boundary condition does not depend on time, $u_b(t,\cdot) = u_b$. The initial condition u_0 is assumed to satisfy $u_0|_{\partial\Omega} = u_b$ and it is approximated by a function $u_h^0 \in W_h$ such that $u_h^0 - \widetilde{u}_{bh} \in V_h$.

To perform the discretization of the time derivative, the time interval $[0,T]$ is divided into N_T equidistant strips of length $\delta t = T/N_T$. The constant time step is used only for simplicity of presentation; for variable time steps the same techniques can be applied leading to essentially the same results. The nodes are denoted by $t^n = n\,\delta t$ for $n = 0, 1, \ldots, N_T$ and the abbreviations $u^n := u(t^n, \cdot)$, $f^n := f(t^n, \cdot)$, *etc.* are used. Since this

section studies the LPS method with nonlinear crosswind diffusion in combination with a one-step θ-scheme as temporal discretization, from now on, the superscript $n+\theta$ denotes for all functions which are defined in $[0,T]$ the values at time $t^{n+\theta} := \theta\, t^{n+1} + (1-\theta)\, t^n$ with any $n \in \{0,\dots,N_T-1\}$ and $\theta \in [0,1]$, *e.g.* $\boldsymbol{b}^{n+\theta} = \boldsymbol{b}(t^{n+\theta},\cdot)$. For functions, which are defined only at the discrete times t^n and t^{n+1}, it denotes the linear interpolation, *e.g.* $u_h^{n+\theta} = \theta\, u_h^{n+1} + (1-\theta)\, u_h^n$. Finally, it is convenient to introduce the interpolation operator $\tilde{r}_h^{n+\theta}$ satisfying

$$\tilde{r}_h^{n+\theta} u = \theta\, r_h u^{n+1} + (1-\theta)\, r_h u^n \tag{63}$$

with r_h from (43). Thus, writing α instead of $n+\theta$, functions u^α, u_h^α, $\tilde{r}_h^\alpha u$, etc. are defined for any $\alpha \in [0, N_T]$.

Then, given $\theta \in (0,1]$, the fully discrete problem reads as follows: for $n = 0, 1, \dots, N_T - 1$, find $u_h^{n+1} \in W_h$ such that $u_h^{n+1} - \widetilde{u}_{bh} \in V_h$ and

$$\left(\frac{u_h^{n+1}-u_h^n}{\delta t}, v_h\right) + a^{n+\theta}(u_h^{n+\theta}, v_h) + s_h^{n+\theta}(u_h^{n+\theta}, v_h) + d_h^{n+\theta}(u_h^{n+\theta}; u_h^{n+\theta}, v_h) = (f^{n+\theta}, v_h) \qquad \forall\, v_h \in V_h. \tag{64}$$

For $\theta = 1/2$, the Crank-Nicolson scheme is recovered and for $\theta = 1$, the implicit Euler scheme is obtained.

Remark 4.1. To simplify the notation, we will not explicitly indicate at which time instant the functions $\boldsymbol{b}$ and σ in the definition of the norm $\|\cdot\|_{\mathrm{LPS}}$ are evaluated. This will be implicitly determined from the context or by the argument of the norm. Thus, if we write, *e.g.*, $\|u_h^{n+\theta}\|_{\mathrm{LPS}}$, the norm $\|\cdot\|_{\mathrm{LPS}}$ is defined using $\boldsymbol{b}^{n+\theta}$ and $\sigma^{n+\theta}$.

4.1. Well-posedness and stability

The well-posedness of (64) can be traced back to the well-posedness of the LPS scheme with crosswind diffusion for the steady-state problem. The discretization of the temporal derivative can be written in the form

$$\left(\frac{u_h^{n+1}-u_h^n}{\delta t}, v_h\right) = \frac{1}{\theta}\left(\frac{u_h^{n+\theta}-u_h^n}{\delta t}, v_h\right).$$

The first part of this term has the form of a reaction term for $u_h^{n+\theta}$. Thus, given u_h^n, the equation at the discrete time t^{n+1} is an equation for $u_h^{n+\theta}$ which has the same form as (19) with the data of the problem at $t^{n+\theta}$ and with a reaction coefficient which has a contribution from the temporal derivative. Thus, defining the operator $\tilde{T}_h^{n+\theta} : V_h \to V_h$ by

$$(\tilde{T}_h^{n+\theta} z_h, v_h) = (T_h^{n+\theta} z_h, v_h) + \frac{1}{\theta\,\delta t}(z_h + \widetilde{u}_{bh}, v_h) - \frac{1}{\theta\,\delta t}(u_h^n, v_h) \qquad \forall\, z_h, v_h \in V_h,$$

it follows that $\tilde{T}_h^{n+\theta}(u_h^{n+\theta} - \widetilde{u}_{bh}) = 0$. Therefore, the existence and uniqueness of a solution $u_h^{n+\theta}$ can be proved in the same way as in the steady-state case, see Section 3.1. This fact is stated in the next result.

Corollary 4.2. *Let $n \in \{0,1,\dots,N_T-1\}$ and $u_h^n \in W_h$ with $u_h^n|_{\partial\Omega} = \widetilde{u}_{bh}$ be given. If $\tilde{\tau}_M$ is defined by (22) or (23), then the problem (64) possesses a solution u_h^{n+1}. In the case that $\tilde{\tau}_M$ is defined by (22), the solution of (64) is unique. Furthermore, there is a constant $C > 0$ such that the solution of the scheme (64) with $\tilde{\tau}_M$ given by (23) is unique if $\delta t\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,M} \le C\, h_M$ for any $M \in \mathscr{M}_h$.*

Proof. The only point remaining to prove is the uniqueness in the case $\tilde{\tau}_M$ is given by (23). For this, let $v_h, w_h \in W_h$ and $z_h := v_h - w_h$. Then, applying (33), the estimate of the $L^p(M)$ norm by the $L^2(M)$ norm (10), (16), $\|P_M^{n+\theta}\|_2 = 1$, and the inverse inequality (8), one arrives at

$$|d_h^{n+\theta}(v_h; v_h, z_h) - d_h^{n+\theta}(w_h; w_h, z_h)| \le C \sum_{M\in\mathscr{M}_h} h_M^{-1}\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,M}\, \|z_h\|_{0,M}^2.$$

Thus, if $v_h, w_h \in V_h$, one obtains

$$(\tilde{T}_h^{n+\theta} v_h - \tilde{T}_h^{n+\theta} w_h, z_h) \geq \sum_{M \in \mathscr{M}_h} \left(\frac{\tilde{C}}{\theta\, \delta t} - \frac{C\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,M}}{h_M} \right) \|z_h\|_{0,M}^2 + \|z_h\|_{\mathrm{LPS}}^2.$$

Consequently, for δt small enough, the operator $\tilde{T}_h^{n+\theta}$ is strongly monotone and hence the solution to the discrete problem (64) is unique. □

The next result states the stability of the method.

Lemma 4.3. *Let $\theta \in [1/2, 1]$ be given. Let $\widetilde{u}_h^\alpha := u_h^\alpha - \widetilde{u}_{bh}$ for any $\alpha \in [0, N_T]$. Then any solution of (64) satisfies the following stability estimate for all $N = 1, 2, \ldots, N_T$:*

$$\begin{aligned} &\|\widetilde{u}_h^N\|_{0,\Omega}^2 + (2\,\theta - 1) \sum_{n=0}^{N-1} \|\widetilde{u}_h^{n+1} - \widetilde{u}_h^n\|_{0,\Omega}^2 + \delta t \sum_{n=0}^{N-1} \|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2 \\ &\quad + \delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(\bar{u}_h^{n+\theta}; \widetilde{u}_h^{n+\theta}, \widetilde{u}_h^{n+\theta}) \leq \|\widetilde{u}_h^0\|_{0,\Omega}^2 + C\, \delta t \sum_{n=0}^{N-1} \Big\{ \sigma_0^{-1} \|f^{n+\theta}\|_{0,\Omega}^2 \\ &\quad + \Big[\varepsilon + \sigma_0^{-1} \left(\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}^2 + \|c^{n+\theta}\|_{0,\infty,\Omega}^2 \right) + h\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \Big] \|\widetilde{u}_{bh}\|_{1,\Omega}^2 + \mu_h \Big\}, \end{aligned} \tag{65}$$

where

$$\bar{u}_h^{n+\theta} = \widetilde{u}_h^{n+\theta}, \qquad \mu_h = \beta\, h\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, |\widetilde{u}_{bh}|_{1,3,\Omega}^3 \qquad \text{if } \tilde{\tau}_M \text{ is given by (22)}, \tag{66}$$

$$\bar{u}_h^{n+\theta} = u_h^{n+\theta}, \qquad \mu_h = 0 \qquad \text{if } \tilde{\tau}_M \text{ is given by (23)}. \tag{67}$$

Proof. The proof starts in the usual way by setting $v_h = \widetilde{u}_h^{n+\theta} \in V_h$ in (64) and using that $u_h^{n+1} - u_h^n = \widetilde{u}_h^{n+1} - \widetilde{u}_h^n$, which leads to

$$\begin{aligned} &(\widetilde{u}_h^{n+1} - \widetilde{u}_h^n, \widetilde{u}_h^{n+\theta}) + \delta t\, \|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2 + \delta t\, d_h^{n+\theta}(u_h^{n+\theta}; u_h^{n+\theta}, \widetilde{u}_h^{n+\theta}) \\ &\qquad = \delta t\, (f^{n+\theta}, \widetilde{u}_h^{n+\theta}) - \delta t\, a^{n+\theta}(\widetilde{u}_{bh}, \widetilde{u}_h^{n+\theta}) - \delta t\, s_h^{n+\theta}(\widetilde{u}_{bh}, \widetilde{u}_h^{n+\theta}). \end{aligned} \tag{68}$$

A straightforward computation gives

$$(\widetilde{u}_h^{n+1} - \widetilde{u}_h^n, \widetilde{u}_h^{n+\theta}) = \frac{1}{2} \left(\|\widetilde{u}_h^{n+1}\|_{0,\Omega}^2 - \|\widetilde{u}_h^n\|_{0,\Omega}^2 \right) + \frac{2\,\theta - 1}{2} \|\widetilde{u}_h^{n+1} - \widetilde{u}_h^n\|_{0,\Omega}^2. \tag{69}$$

Next, the application of the Cauchy-Schwarz inequality, the Young inequality, (16), (18), the definition of τ_M (20), and the geometrical hypotheses (4) and (5) yield

$$\begin{aligned} (f^{n+\theta}, \widetilde{u}_h^{n+\theta}) &\leq \frac{1}{\sigma_0} \|f^{n+\theta}\|_{0,\Omega}^2 + \frac{1}{4} \|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2, \\ a^{n+\theta}(\widetilde{u}_{bh}, \widetilde{u}_h^{n+\theta}) &\leq 6 \left[\varepsilon + \sigma_0^{-1} \left(\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}^2 + \|c^{n+\theta}\|_{0,\infty,\Omega}^2 \right) \right] \|\widetilde{u}_{bh}\|_{1,\Omega}^2 + \frac{1}{8} \|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2, \\ s_h^{n+\theta}(\widetilde{u}_{bh}, \widetilde{u}_h^{n+\theta}) &\leq C\, h\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, |\widetilde{u}_{bh}|_{1,\Omega}^2 + \frac{1}{8} \|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2. \end{aligned}$$

If $\tilde{\tau}_M$ is given by (22), then, from (27) and an analog of (51), one obtains

$$\begin{aligned} d_h^{n+\theta}(u_h^{n+\theta}; u_h^{n+\theta}, \widetilde{u}_h^{n+\theta}) &\geq \frac{1}{7} d_h^{n+\theta}(\widetilde{u}_h^{n+\theta}; \widetilde{u}_h^{n+\theta}, \widetilde{u}_h^{n+\theta}) + d_h^{n+\theta}(\widetilde{u}_{bh}; \widetilde{u}_{bh}, \widetilde{u}_h^{n+\theta}) \\ &\geq \frac{1}{10} d_h^{n+\theta}(\widetilde{u}_h^{n+\theta}; \widetilde{u}_h^{n+\theta}, \widetilde{u}_h^{n+\theta}) - 2\, d_h^{n+\theta}(\widetilde{u}_{bh}; \widetilde{u}_{bh}, \widetilde{u}_{bh}). \end{aligned}$$

Furthermore, the use of (10), (16), (18), $\|P_M^{n+\theta}\|_2 = 1$, (4), and (5) leads to

$$d_h^{n+\theta}(\widetilde{u}_{bh};\widetilde{u}_{bh},\widetilde{u}_{bh}) \le C\,\beta \sum_{M\in\mathscr{M}_h} h_M^{1-d/2}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,M}\,|\widetilde{u}_{bh}|_{1,M}^3 \le \tilde{C}\,\beta\,h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\,|\widetilde{u}_{bh}|_{1,3,\Omega}^3.$$

If $\tilde{\tau}_M$ is given by (23), then, using an inequality like (58), one gets

$$\begin{aligned} d_h^{n+\theta}(u_h^{n+\theta};u_h^{n+\theta},\widetilde{u}_h^{n+\theta}) &= d_h^{n+\theta}(u_h^{n+\theta};\widetilde{u}_h^{n+\theta},\widetilde{u}_h^{n+\theta}) + d_h^{n+\theta}(u_h^{n+\theta};\widetilde{u}_{bh},\widetilde{u}_h^{n+\theta}) \\ &\ge \frac{1}{2}\,d_h^{n+\theta}(u_h^{n+\theta};\widetilde{u}_h^{n+\theta},\widetilde{u}_h^{n+\theta}) - \frac{1}{2}\,d_h^{n+\theta}(u_h^{n+\theta};\widetilde{u}_{bh},\widetilde{u}_{bh}). \end{aligned}$$

Applying the Hölder inequality, (24), the estimate of the $L^p(M)$ norm by the $L^2(M)$ norm (10), (16), $\|P_M^{n+\theta}\|_2 = 1$, (4), and (5), one deduces that

$$\begin{aligned} d_h^{n+\theta}(u_h^{n+\theta};\widetilde{u}_{bh},\widetilde{u}_{bh}) &\le C \sum_{M\in\mathscr{M}_h} h_M^{1+d/2}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,M}\,\|\kappa_M(P_M^{n+\theta}\,\nabla\widetilde{u}_{bh})\|_{0,4,M}^2 \\ &\le \tilde{C}\,h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\,|\widetilde{u}_{bh}|_{1,\Omega}^2. \end{aligned}$$

Now, inserting the above relations into (4.1) and using the notation (66) and (67), one obtains

$$\begin{aligned} &\frac{1}{2}\,(\|\widetilde{u}_h^{n+1}\|_{0,\Omega}^2 - \|\widetilde{u}_h^{n}\|_{0,\Omega}^2) + \frac{2\,\theta-1}{2}\,\|\widetilde{u}_h^{n+1}-\widetilde{u}_h^{n}\|_{0,\Omega}^2 + \frac{\delta t}{2}\,\|\widetilde{u}_h^{n+\theta}\|_{\mathrm{LPS}}^2 + \frac{\delta t}{6}\,d_h^{n+\theta}(\bar{u}_h^{n+\theta};\widetilde{u}_h^{n+\theta},\widetilde{u}_h^{n+\theta}) \\ &\le \delta t\,\sigma_0^{-1}\,\|f^{n+\theta}\|_{0,\Omega}^2 + C\,\delta t\,\big\{\varepsilon + \sigma_0^{-1}\,(\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}^2 + \|c^{n+\theta}\|_{0,\infty,\Omega}^2) + h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\big\}\,\|\widetilde{u}_{bh}\|_{1,\Omega}^2 \\ &\quad + C\,\delta t\,\mu_h, \end{aligned}$$

and (65) follows by summing up from $n = 0$ to $N-1$. □

Remark 4.4. The inequality (65) is a proper stability result provided that $\|u_h^0\|_{0,\Omega}$, $\|\widetilde{u}_{bh}\|_{1,\Omega}$ and, if $\tilde{\tau}_M$ is given by (22), also $|\widetilde{u}_{bh}|_{1,3,\Omega}$ are bounded when $h \to 0$. One may set $u_h^0 = I_h u_0$ and $\widetilde{u}_{bh} = I_h\widetilde{u}_b$, where $I_h : H^1(\Omega) \to W_h$ is the Scott-Zhang interpolation operator (*cf.*, *e.g.*, [12]) and $\widetilde{u}_b \in H^1(\Omega)$ is an extension of u_b. Then $\|u_h^0\|_{0,\Omega} \le C\,\|u_0\|_{1,\Omega}$ and $\|\widetilde{u}_{bh}\|_{1,\Omega} \le C\,\|\widetilde{u}_b\|_{1,\Omega}$. If $\widetilde{u}_b \in W^{1,3}(\Omega)$ (requiring the stronger assumption $u_b \in W^{2/3,3}(\partial\Omega)$), then also $|\widetilde{u}_{bh}|_{1,3,\Omega} \le C\,\|\widetilde{u}_b\|_{1,3,\Omega}$. It is important that I_h preserves homogeneous boundary conditions since one has to assure that u_h^0 and $\widetilde{u}_{bh}$ coincide on the boundary of Ω. If $u_0 \in H^2(\Omega)$ and $u_b \in H^{3/2}(\partial\Omega)$, which are the minimal regularity assumptions for deriving the error estimates in the next section, one may use the operator i_h from Section 2 instead of I_h. Now $\widetilde{u}_b \in H^2(\Omega)$ and, according to the approximation properties of i_h (11) and (13), one has $\|u_h^0\|_{0,\Omega} \le C\,\|u_0\|_{2,\Omega}$ and $\|\widetilde{u}_{bh}\|_{1,\Omega} + |\widetilde{u}_{bh}|_{1,3,\Omega} \le C\,\|\widetilde{u}_b\|_{2,\Omega}$.

Remark 4.5. It is worth remarking that, for the homogeneous case $u_b = 0$, instead of the direct proof presented in this manuscript, an analysis completely analogous to the one given in [8], Corollary 7, leads to the following stability result for $\theta \in [1/2, 1]$ and $N < N_T$

$$\frac{1}{2}\,\|u_h^N\|_{0,\Omega}^2 + \delta t \sum_{n=0}^{N-1}\,\{\|u_h^{n+\theta}\|_{\mathrm{LPS}}^2 + d_h^{n+\theta}(u_h^{n+\theta};u_h^{n+\theta},u_h^{n+\theta})\} \le \mathrm{e}^{\frac{T}{T-\delta t}}\,\left\{T\,\delta t \sum_{n=0}^{N-1}\,\|f^{n+\theta}\|_{0,\Omega}^2 + \frac{1}{2}\,\|u_h^0\|_{0,\Omega}^2\right\}. \tag{70}$$

This result, very similar in form to the one in [8] (with the extra control on the nonlinear term, and a slightly smaller right-hand side), is independent of σ_0, and hence represents an improvement over the way Lemma 4.3 is presented. The reason to present the direct proof here lies in the non-homogeneous case, where the presence of u_b is responsible for the dependency of the constant on the right-hand side on σ_0^{-1}. In the non-homogeneous case, both proofs lead to essentially equivalent results, the direct proof presented in this work being more straightforward.

Finally, if u_b would be supposed time dependent, then in the first line of the proof of stability there holds $u_h^{n+1} - u_h^n = \widetilde{u}_h^{n+1} - \widetilde{u}_h^n + \widetilde{u}_{bh}^{n+1} - \widetilde{u}_{bh}^n$, thus creating an extra right-hand side depending on the time derivative of u_b.

4.2. Error estimates

In this section, error estimates are derived for the solution of the discrete problem (64) with $\theta \in [1/2, 1]$. The error will be analyzed essentially in the quantity which is given by the stability estimate (65). Let us denote the error by $e^\alpha := u^\alpha - u_h^\alpha$ with $\alpha \in [0, N_T]$. Furthermore, to simplify the presentation of our results, we introduce the quantities

$$
\begin{aligned}
E^N &= \|e^N\|_{0,\Omega} + \left(\delta t \sum_{n=0}^{N-1} \|e^{n+\theta}\|_{\mathrm{LPS}}^2\right)^{1/2},\\
Q^N &= h\left(|u_0|_{k+1,\Omega} + |u^N|_{k+1,\Omega} + \sigma_0^{-1/2}\,\|u_t\|_{L^2(0,t^N;H^{k+1}(\Omega))}\right) + \left(\delta t \sum_{n=0}^{N-1}\Big(\varepsilon + h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\right.\\
&\quad \left. + h^2\,\|\sigma^{n+\theta}\|_{0,\infty,\Omega} + h^2\,\sigma_0^{-1}\,|\boldsymbol{b}^{n+\theta}|_{1,\infty,\Omega}^2\Big)\Big(|u^n|_{k+1,\Omega}^2 + |u^{n+1}|_{k+1,\Omega}^2\Big)\right)^{1/2},\\
R^N &= \left(\delta t \sum_{n=0}^{N-1} h^{k+1-d/2}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\left(|u^n|_{k+1,\Omega}^3 + |u^{n+1}|_{k+1,\Omega}^3\right)\right)^{1/2},\\
S^N &= \left(\delta t \sum_{n=0}^{N-1} h^{k+1}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\left(|u^n|_{k+1,\infty,\Omega} + |u^{n+1}|_{k+1,\infty,\Omega}\right)\Big(|u^n|_{k+1,\Omega}^2 + |u^{n+1}|_{k+1,\Omega}^2\Big)\right)^{1/2},\\
X^N &= \max_{n=0,\dots,N-1}\left(\varepsilon + h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} + \|\sigma^{n+\theta}\|_{0,\infty,\Omega} + \sigma_0^{-1}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}^2 + \sigma_0^{-1}\,\|c^{n+\theta}\|_{0,\infty,\Omega}^2\right)^{1/2},\\
Y^N &= h^{1/2}\max_{n=0,\dots,N-1}\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}^{1/2},
\end{aligned}
$$

where $N = 1, 2, \dots, N_T$.

Theorem 4.6. *Let $\theta \in [1/2, 1]$ be given. Let the weak solution of (3) satisfy $u, u_t \in L^2(0,T;H^{k+1}(\Omega))$ for some $k \in \{1,\dots,l\}$ and assume $u_{tt} \in L^2(0,T;L^2(\Omega))$. Let $\widetilde{u}_b \in H^2(\Omega)$ be an extension of u_b and let $\widetilde{u}_{bh} = i_h\widetilde{u}_b$. Assume $u_0 \in H^{k+1}(\Omega)$ and let $u_h^0 = i_h u_0$. Let $\{u_h^n\}_{n=0}^{N_T}$ be the solution of the local projection discretization (64). If $\tilde{\tau}_M$ is defined by (22) and $u_t \in L^3(0,T;W^{1,3}(\Omega))$, then the error estimate*

$$
\begin{aligned}
&E^N + \left(\delta t \sum_{n=0}^{N-1}\sum_{M\in\mathscr{M}_h} \tilde{\tau}_M \|\kappa_M(P_M^{n+\theta}\nabla e^{n+\theta})\|_{0,3,M}^3\right)^{1/2}\\
&\qquad \le C\,h^k\,Q^N + C\,\beta\,h^k\,R^N + C\,\delta t\,X^N\,\|u_t\|_{L^2(0,t^N;H^1(\Omega))}\\
&\qquad\quad + C\,\beta\,(\delta t)^{3/2}\,Y^N\,\|u_t\|_{L^3(0,t^N;W^{1,3}(\Omega))}^{3/2} + C\,\delta t\,\sigma_0^{-1/2}\,\|u_{tt}\|_{L^2(0,t^N;L^2(\Omega))}
\end{aligned}
\tag{71}
$$

is satisfied for $N = 1, 2, \dots, N_T$. Moreover, if $\theta = 1/2$, $u_{tt} \in L^3(0,T;W^{1,3}(\Omega))$, and $u_{ttt} \in L^2(0,T;L^2(\Omega))$, then

$$
\begin{aligned}
&E^N + \left(\delta t \sum_{n=0}^{N-1}\sum_{M\in\mathscr{M}_h} \tilde{\tau}_M \|\kappa_M(P_M^{n+\theta}\nabla e^{n+\theta})\|_{0,3,M}^3\right)^{1/2}\\
&\qquad \le C\,h^k\,Q^N + C\,\beta\,h^k\,R^N + C\,(\delta t)^2\,X^N\,\|u_{tt}\|_{L^2(0,t^N;H^1(\Omega))}\\
&\qquad\quad + C\,\beta\,(\delta t)^3\,Y^N\,\|u_{tt}\|_{L^3(0,t^N;W^{1,3}(\Omega))}^{3/2} + C\,(\delta t)^2\,\sigma_0^{-1/2}\,\|u_{ttt}\|_{L^2(0,t^N;L^2(\Omega))}.
\end{aligned}
$$

If $u \in L^2(0,T;W^{k+1,\infty}(\Omega))$, then, in both estimates, R^N can be replaced by S^N.

If $\tilde{\tau}_M$ is defined by (23) and $u_t \in L^4(0,T;W^{1,4}(\Omega))$, then the following error estimate holds

$$E^N + \left(\delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(u_h^{n+\theta}; e^{n+\theta}, e^{n+\theta}) \right)^{1/2} \leq C\, h^k\, Q^N + C\, \delta t\, X^N\, \|u_t\|_{L^2(0,t^N;H^1(\Omega))} + C\, \delta t\, T^{1/4}\, Y^N\, \|u_t\|_{L^4(0,t^N;W^{1,4}(\Omega))} + C\, \delta t\, \sigma_0^{-1/2}\, \|u_{tt}\|_{L^2(0,t^N;L^2(\Omega))}. \tag{72}$$

Moreover, if $\theta = 1/2$, $u_{tt} \in L^4(0,T;W^{1,4}(\Omega))$, and $u_{ttt} \in L^2(0,T;L^2(\Omega))$, then

$$E^N + \left(\delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(u_h^{n+\theta}; e^{n+\theta}, e^{n+\theta}) \right)^{1/2} \leq C\, h^k\, Q^N + C\, (\delta t)^2\, X^N\, \|u_{tt}\|_{L^2(0,t^N;H^1(\Omega))} + C\, (\delta t)^2\, T^{1/4}\, Y^N\, \|u_{tt}\|_{L^4(0,t^N;W^{1,4}(\Omega))} + C\, (\delta t)^2\, \sigma_0^{-1/2}\, \|u_{ttt}\|_{L^2(0,t^N;L^2(\Omega))}.$$

Proof. Analogously to the steady-state case, the error will be split into an interpolation error and a remainder which belongs to the finite element space. The decomposition of the error e^α with any $\alpha \in [0, N_T]$ has the form

$$e^\alpha = \eta^\alpha - e_h^\alpha \qquad \text{with} \qquad \eta^\alpha := u^\alpha - \bar{r}_h^\alpha, \quad e_h^\alpha := u_h^\alpha - \bar{r}_h^\alpha \in V_h,$$

where we use the abbreviation $\bar{r}_h^\alpha = \tilde{r}_h^\alpha u$ with $\tilde{r}_h^\alpha$ given by (63). Using this decomposition, one obtains with the triangle inequality and with (60)

$$\begin{aligned} \|e^N\|_{0,\Omega}^2 + \delta t \sum_{n=0}^{N-1} \|e^{n+\theta}\|_{\mathrm{LPS}}^2 + \delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(\gamma_0^{n+\theta}; e^{n+\theta}, e^{n+\theta}) \\ \leq 4 \left[\|\eta^N\|_{0,\Omega}^2 + \delta t \sum_{n=0}^{N-1} \|\eta^{n+\theta}\|_{\mathrm{LPS}}^2 + \delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(\gamma_1^{n+\theta}; \eta^{n+\theta}, \eta^{n+\theta}) \right] \\ + 4 \left[\|e_h^N\|_{0,\Omega}^2 + \delta t \sum_{n=0}^{N-1} \|e_h^{n+\theta}\|_{\mathrm{LPS}}^2 + \delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(\gamma_2^{n+\theta}; e_h^{n+\theta}, e_h^{n+\theta}) \right], \end{aligned} \tag{73}$$

where $\gamma_0^{n+\theta} = e^{n+\theta}$, $\gamma_1^{n+\theta} = \eta^{n+\theta}$, $\gamma_2^{n+\theta} = e_h^{n+\theta}$ if $\tilde{\tau}_M$ is defined by (22) and $\gamma_0^{n+\theta} = \gamma_1^{n+\theta} = \gamma_2^{n+\theta} = u_h^{n+\theta}$ if $\tilde{\tau}_M$ is defined by (23).

First let us estimate the interpolation errors. The starting point is the identity

$$\eta^{n+\theta} = u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n + \theta\, (u^{n+1} - r_h u^{n+1}) + (1-\theta)\, (u^n - r_h u^n). \tag{74}$$

One has

$$u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n = (1-\theta) \int_{t^n}^{t^{n+\theta}} u_t(t)\, \mathrm{d}t - \theta \int_{t^{n+\theta}}^{t^{n+1}} u_t(t)\, \mathrm{d}t, \tag{75}$$

which, in view of (45), leads to

$$\begin{aligned} \|\eta^{n+\theta}\|_{0,\Omega} &\leq C\, h^{k+1}\, (|u^n|_{k+1,\Omega} + |u^{n+1}|_{k+1,\Omega}) + \sqrt{\delta t}\, \|u_t\|_{L^2(t^n,t^{n+1};L^2(\Omega))}, \\ |\eta^{n+\theta}|_{1,\Omega} &\leq C\, h^k\, (|u^n|_{k+1,\Omega} + |u^{n+1}|_{k+1,\Omega}) + \sqrt{\delta t}\, \|u_t\|_{L^2(t^n,t^{n+1};H^1(\Omega))}. \end{aligned}$$

Using Taylor's formula with integral remainder or applying successively integration by parts gives

$$u^n = u^{n+\theta} - \theta\, \delta t\, u_t^{n+\theta} + \int_{t^{n+\theta}}^{t^n} u_{tt}(t)\, (t^n - t)\, \mathrm{d}t, \tag{76}$$

$$u^{n+1} = u^{n+\theta} + (1-\theta)\, \delta t\, u_t^{n+\theta} + \int_{t^{n+\theta}}^{t^{n+1}} u_{tt}(t)\, (t^{n+1} - t)\, \mathrm{d}t. \tag{77}$$

This may be used to derive improved interpolation estimates with respect to the time step provided that $u_{tt} \in L^2(0,T;H^1(\Omega))$. Indeed,

$$u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n = -(1-\theta) \int_{t^n}^{t^{n+\theta}} u_{tt}(t)\,(t-t^n)\,\mathrm{d}t - \theta \int_{t^{n+\theta}}^{t^{n+1}} u_{tt}(t)\,(t^{n+1}-t)\,\mathrm{d}t, \tag{78}$$

which leads to

$$\begin{aligned}
\|\eta^{n+\theta}\|_{0,\Omega} &\le C\, h^{k+1}\,(|u^n|_{k+1,\Omega} + |u^{n+1}|_{k+1,\Omega}) + (\delta t)^{3/2}\,\|u_{tt}\|_{L^2(t^n,t^{n+1};L^2(\Omega))},\\
|\eta^{n+\theta}|_{1,\Omega} &\le C\, h^{k}\,(|u^n|_{k+1,\Omega} + |u^{n+1}|_{k+1,\Omega}) + (\delta t)^{3/2}\,\|u_{tt}\|_{L^2(t^n,t^{n+1};H^1(\Omega))}.
\end{aligned}$$

Now let us estimate the norms of the interpolation error in (73). In view of (63), (45), (16), (18), and the geometrical hypotheses (5) and (4), one has

$$\begin{aligned}
\|\eta^N\|_{0,\Omega} &= \|u^N - r_h u^N\|_{0,\Omega} \le C\, h^{k+1}\,|u^N|_{k+1,\Omega},\\
\|\eta^{n+\theta}\|_{\mathrm{LPS}} &\le \Big(\varepsilon + C\,h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\Big)^{1/2}\,|\eta^{n+\theta}|_{1,\Omega} + \|\sigma^{n+\theta}\|_{0,\infty,\Omega}^{1/2}\,\|\eta^{n+\theta}\|_{0,\Omega}.
\end{aligned}$$

Furthermore, analogously as in (54), for any $p \in [2,6]$, one obtains

$$\begin{aligned}
\|\kappa_M(P_M^{n+\theta}\nabla\eta^{n+\theta})\|_{0,p,M} \le{}& C\,|u^{n+\theta} - \theta\, i_h u^{n+1} - (1-\theta)\, i_h u^n|_{1,p,M}\\
&+ C\,h_M^{\frac{d}{p}-\frac{d}{2}}\,\big(|\varrho_h(u^n - i_h u^n)|_{1,M} + |\varrho_h(u^{n+1} - i_h u^{n+1})|_{1,M}\big).
\end{aligned} \tag{79}$$

If $\tilde{\tau}_M$ is defined by (22), this inequality implies that

$$d_h^{n+\theta}(\eta^{n+\theta};\eta^{n+\theta},\eta^{n+\theta}) \le C\,\beta\,(I + II),$$

where

$$\begin{aligned}
I &:= h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} |u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n|_{1,3,M}^3,\\
II &:= h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} \big(|u^{n+1} - i_h u^{n+1}|_{1,3,M}^3 + |u^n - i_h u^n|_{1,3,M}^3\big)\\
&\quad + h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} h_M^{-\frac{d}{2}}\,\big(|\varrho_h(u^n - i_h u^n)|_{1,M}^3 + |\varrho_h(u^{n+1} - i_h u^{n+1})|_{1,M}^3\big).
\end{aligned}$$

Using (75) and (78), one obtains

$$I \le C\,h\,(\delta t)^2\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\,\|u_t\|_{L^3(t^n,t^{n+1};W^{1,3}(\Omega))}^3,$$

resp.

$$I \le C\,h\,(\delta t)^5\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\,\|u_{tt}\|_{L^3(t^n,t^{n+1};W^{1,3}(\Omega))}^3.$$

Furthermore, it follows from (13), (41), (11), (6), and (4) that

$$II \le C\,h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} h_M^{3k-d/2}\,(|u^n|_{k+1,M}^3 + |u^{n+1}|_{k+1,M}^3), \tag{80}$$

which implies in view of (4) and (5) that

$$II \le C\,h^{3k+1-d/2}\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\,(|u^n|_{k+1,\Omega}^3 + |u^{n+1}|_{k+1,\Omega}^3).$$

If $u \in L^2(0,T;W^{k+1,\infty}(\Omega))$, the inequality (80) together with (4) and (5) implies that

$$II \le C\, h^{3\,k+1}\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, (|u^n|_{k+1,\infty,\Omega}\, |u^n|^2_{k+1,\Omega} + |u^{n+1}|_{k+1,\infty,\Omega}\, |u^{n+1}|^2_{k+1,\Omega}).$$

If $\tilde{\tau}_M$ is defined by (23), then, proceeding analogously as when deriving (61), but with (79) instead of (54), and applying (13) in addition, one gets

$$d_h^{n+\theta}(u_h^{n+\theta};\eta^{n+\theta},\eta^{n+\theta}) \le C\,\tilde{I} + C\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, h^{2\,k+1}\, (|u^n|^2_{k+1,\Omega} + |u^{n+1}|^2_{k+1,\Omega}),$$

where

$$\tilde{I} := h\,\|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega} \sum_{M\in\mathscr{M}_h} h_M^{d/2}\, |u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n|^2_{1,4,M}.$$

Similarly as above, one obtains

$$\tilde{I} \le C\, h\, (\delta t)^{3/2}\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, \|u_t\|^2_{L^4(t^n,t^{n+1};W^{1,4}(\Omega))},$$

resp.

$$\tilde{I} \le C\, h\, (\delta t)^{7/2}\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, \|u_{tt}\|^2_{L^4(t^n,t^{n+1};W^{1,4}(\Omega))}.$$

Now let us estimate the norms of the discrete part of the error on the right-hand side of (73). To derive an equation for this part of the error, the weak formulation (62) at $t = t^{n+\theta}$ is subtracted from (64) with $v = v_h = e_h^{n+\theta}$. Then, using the fact that $u_h^\alpha = e_h^\alpha + \bar{r}_h^\alpha$, one deduces that

$$\begin{aligned}(e_h^{n+1} - e_h^n, e_h^{n+\theta}) &+ \delta t\, \|e_h^{n+\theta}\|^2_{\mathrm{LPS}} + \delta t\, d_h^{n+\theta}(u_h^{n+\theta};u_h^{n+\theta},e_h^{n+\theta}) \\ &= \delta t \left[\left(u_t^{n+\theta} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t}, e_h^{n+\theta}\right) + a^{n+\theta}(\eta^{n+\theta}, e_h^{n+\theta}) - s_h^{n+\theta}(\bar{r}_h^{n+\theta}, e_h^{n+\theta})\right]. \end{aligned} \tag{81}$$

Furthermore, one obtains

$$d_h^{n+\theta}(u_h^{n+\theta};u_h^{n+\theta},e_h^{n+\theta}) \ge \frac{1}{7}\, d_h^{n+\theta}(\gamma_2^{n+\theta};e_h^{n+\theta},e_h^{n+\theta}) + d_h^{n+\theta}(\gamma_3^{n+\theta};\bar{r}_h^{n+\theta},e_h^{n+\theta}), \tag{82}$$

where $\gamma_3^{n+\theta} = \bar{r}_h^{n+\theta}$ if $\tilde{\tau}_M$ is defined by (22) and $\gamma_3^{n+\theta} = u_h^{n+\theta}$ if $\tilde{\tau}_M$ is defined by (23) ($\gamma_2^{n+\theta}$ was defined below (73)). This estimate follows from (27) if $\tilde{\tau}_M$ is defined by (22) and simply by writing the second argument of $d_h^{n+\theta}$ as $e_h^{n+\theta} + \bar{r}_h^{n+\theta}$ and using the fact that $d_h^{n+\theta}(u_h^{n+\theta};e_h^{n+\theta},e_h^{n+\theta}) \ge 0$ if $\tilde{\tau}_M$ is defined by (23). Since $\theta \ge 1/2$, it follows from (69) with $\widetilde{u}$ replaced by e that

$$(e_h^{n+1} - e_h^n, e_h^{n+\theta}) \ge \frac{1}{2}\, (\|e_h^{n+1}\|^2_{0,\Omega} - \|e_h^n\|^2_{0,\Omega}). \tag{83}$$

Substituting (82) and (83) into (81) and summing up over the discrete times yields an upper bound for the discrete part of the estimate (73)

$$\begin{aligned}\|e_h^N\|^2_{0,\Omega} &+ \delta t \sum_{n=0}^{N-1} \|e_h^{n+\theta}\|^2_{\mathrm{LPS}} + \delta t \sum_{n=0}^{N-1} d_h^{n+\theta}(\gamma_2^{n+\theta};e_h^{n+\theta},e_h^{n+\theta}) \\ &\le \frac{7}{2}\, \|e_h^0\|^2_{0,\Omega} + 7\,\delta t \sum_{n=0}^{N-1} \left[\left(u_t^{n+\theta} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t}, e_h^{n+\theta}\right) + a^{n+\theta}(\eta^{n+\theta}, e_h^{n+\theta})\right. \\ &\quad \left. - s_h^{n+\theta}(\bar{r}_h^{n+\theta}, e_h^{n+\theta}) - d_h^{n+\theta}(\gamma_3^{n+\theta};\bar{r}_h^{n+\theta},e_h^{n+\theta})\right]. \end{aligned} \tag{84}$$

Using (42), the approximation property of i_h (11), (5), and (4), one obtains

$$\|e_h^0\|_{0,\Omega} = \|i_h u^0 - r_h u^0\|_{0,\Omega} = \|\varrho_h(u^0 - i_h u^0)\|_{0,\Omega} \leq C\, h^{k+1}\, |u^0|_{k+1,\Omega}.$$

Applying the Cauchy-Schwarz and Young inequalities gives

$$\left(u_t^{n+\theta} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t}, e_h^{n+\theta}\right) \leq \frac{1}{\sigma_0} \left\| u_t^{n+\theta} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t} \right\|_{0,\Omega}^2 + \frac{1}{4} \|e_h^{n+\theta}\|_{\mathrm{LPS}}^2.$$

The last term can be hidden in the left-hand side of (84). The first term is a mixture of discretization errors in time and space. Elimination of $u^{n+\theta}$ from (76) and (77) yields

$$u_t^{n+\theta} = \frac{u^{n+1} - u^n}{\delta t} - \frac{1}{\delta t} \int_{t^n}^{t^{n+\theta}} u_{tt}(t)\, (t^n - t)\, \mathrm{d}t - \frac{1}{\delta t} \int_{t^{n+\theta}}^{t^{n+1}} u_{tt}(t)\, (t^{n+1} - t)\, \mathrm{d}t.$$

Since interpolation in space and differentiation in time commute, one has

$$u^{n+1} - \bar{r}_h^{n+1} - (u^n - \bar{r}_h^n) = \int_{t^n}^{t^{n+1}} (u_t - r_h u_t)(t)\, \mathrm{d}t.$$

Thus, applying the Cauchy-Schwarz inequality, one derives

$$\left\| u_t^{n+\theta} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t} \right\|_{0,\Omega}^2 \leq \frac{2}{\delta t} \|u_t - r_h u_t\|_{L^2(t^n, t^{n+1}; L^2(\Omega))}^2 + 2\, \delta t\, \|u_{tt}\|_{L^2(t^n, t^{n+1}; L^2(\Omega))}^2.$$

The first term on the right-hand side can be bounded using (45).

Assuming $u_{ttt} \in L^2(0, T; L^2(\Omega))$ and replacing (76) and (77) by

$$u^n = u^{n+\theta} - \theta\, \delta t\, u_t^{n+\theta} + \frac{\theta^2}{2} (\delta t)^2\, u_{tt}^{n+\theta} + \frac{1}{2} \int_{t^{n+\theta}}^{t^n} u_{ttt}(t)\, (t^n - t)^2\, \mathrm{d}t,$$

$$u^{n+1} = u^{n+\theta} + (1-\theta)\, \delta t\, u_t^{n+\theta} + \frac{(1-\theta)^2}{2} (\delta t)^2\, u_{tt}^{n+\theta} + \frac{1}{2} \int_{t^{n+\theta}}^{t^{n+1}} u_{ttt}(t)\, (t^{n+1} - t)^2\, \mathrm{d}t,$$

one obtains

$$\begin{aligned} u_t^{n+\theta} = \frac{u^{n+1} - u^n}{\delta t} &+ \frac{\delta t}{2} \left[\theta^2 - (1-\theta)^2\right] u_{tt}^{n+\theta} \\ &- \frac{1}{2\, \delta t} \int_{t^n}^{t^{n+\theta}} u_{ttt}(t)\, (t^n - t)^2\, \mathrm{d}t - \frac{1}{2\, \delta t} \int_{t^{n+\theta}}^{t^{n+1}} u_{ttt}(t)\, (t^{n+1} - t)^2\, \mathrm{d}t, \end{aligned}$$

which shows that an improved estimate with respect to δt follows for $\theta = 1/2$, *i.e.*, for the Crank-Nicolson scheme. Indeed, one gets

$$\left\| u_t^{n+1/2} - \frac{\bar{r}_h^{n+1} - \bar{r}_h^n}{\delta t} \right\|_{0,\Omega}^2 \leq \frac{2}{\delta t} \|u_t - r_h u_t\|_{L^2(t^n, t^{n+1}; L^2(\Omega))}^2 + (\delta t)^3\, \|u_{ttt}\|_{L^2(t^n, t^{n+1}; L^2(\Omega))}^2.$$

Now let us consider the remaining three terms on the right-hand side of (84). According to (74) and (63), one has

$$\begin{aligned} a^{n+\theta}(\eta^{n+\theta}, e_h^{n+\theta}) - s_h^{n+\theta}(\bar{r}_h^{n+\theta}, e_h^{n+\theta}) &= a^{n+\theta}(u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n, e_h^{n+\theta}) \\ &+ \theta \left[a^{n+\theta}(u^{n+1} - r_h u^{n+1}, e_h^{n+\theta}) - s_h^{n+\theta}(r_h u^{n+1}, e_h^{n+\theta}) \right] \\ &+ (1-\theta) \left[a^{n+\theta}(u^n - r_h u^n, e_h^{n+\theta}) - s_h^{n+\theta}(r_h u^n, e_h^{n+\theta}) \right]. \end{aligned}$$

The last two terms can be estimated by (46) and the estimation of the first term on the right-hand side is performed using

$$\|u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n\|_{1,\Omega}^2 \le \delta t\, \|u_t\|^2_{L^2(t^n,t^{n+1};H^1(\Omega))},$$

resp.

$$\|u^{n+\theta} - \theta\, u^{n+1} - (1-\theta)\, u^n\|_{1,\Omega}^2 \le (\delta t)^3\, \|u_{tt}\|^2_{L^2(t^n,t^{n+1};H^1(\Omega))},$$

which follows from (75), resp. (78). Finally, the last term on the right-hand side of (84) can be estimated analogously as (52), (56), and (59): if $\tilde\tau_M$ is defined by (22), one derives

$$d_h^{n+\theta}(\bar r_h^{n+\theta}; \bar r_h^{n+\theta}, \bar r_h^{n+\theta}) \le C\,\beta\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, h^{3\,k+1-d/2}\, (|u^n|^3_{k+1,\Omega} + |u^{n+1}|^3_{k+1,\Omega}),$$

if, in addition, $u \in L^2(0,T;W^{k+1,\infty}(\Omega))$, then

$$\begin{aligned} &d_h^{n+\theta}(\bar r_h^{n+\theta}; \bar r_h^{n+\theta}, \bar r_h^{n+\theta}) \\ &\qquad \le C\,\beta\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, h^{3\,k+1}\, (|u^n|_{k+1,\infty,\Omega} + |u^{n+1}|_{k+1,\infty,\Omega})(|u^n|^2_{k+1,\Omega} + |u^{n+1}|^2_{k+1,\Omega}), \end{aligned}$$

and, if $\tilde\tau_M$ is defined by (23), then

$$d_h^{n+\theta}(u_h^{n+\theta}; \bar r_h^{n+\theta}, \bar r_h^{n+\theta}) \le C\, \|\boldsymbol{b}^{n+\theta}\|_{0,\infty,\Omega}\, h^{2\,k+1}\, (|u^n|^2_{k+1,\Omega} + |u^{n+1}|^2_{k+1,\Omega}).$$

These estimates together with analogs of (51) and (58) lead to an estimate of the term $d_h^{n+\theta}(\gamma_3^{n+\theta}; \bar r_h^{n+\theta}, e_h^{n+\theta})$.

Collecting all the above estimates proves the theorem. □

At the end of this section, a semi-implicit (linearized) variant of the method (64) will be discussed: for $n = 0, 1, \ldots, N_T - 1$, find $u_h^{n+1} \in W_h$ such that $u_h^{n+1} - \widetilde{u}_{bh} \in V_h$ and

$$\left(\frac{u_h^{n+1} - u_h^n}{\delta t}, v_h\right) + a^{n+\theta}(u_h^{n+\theta}, v_h) + s_h^{n+\theta}(u_h^{n+\theta}, v_h) + d_h^{n+\theta}(u_h^n; u_h^{n+\theta}, v_h) = (f^{n+\theta}, v_h) \qquad \forall\, v_h \in V_h. \quad (85)$$

The advantages of this linearized scheme over (64) in terms of computational complexity are clear. Indeed, for (85) only one linear system needs to be solved per time step. Moreover, the linearized problem is uniquely solvable for any non-negative integrable stabilization parameter τ_M^{sold}. If the parameter $\tilde\tau_M$ is defined by (23), the results of Lemma 4.3 and Theorem 4.6 remain essentially valid; the only difference is that in these results the first argument of $d_h^{n+\theta}$ is now u_h^n. The proofs of Lemma 4.3 and Theorem 4.6 can be repeated without any changes for $\tilde\tau_M$ defined by (23) since the estimates of the nonlinear term $d_h^{n+\theta}$ are based on (24) and hence are independent of the first argument of $d_h^{n+\theta}$. This is not the case if $\tilde\tau_M$ is defined by (22) and, therefore, we were able to prove only suboptimal convergence results and a stability result depending on T in a similar way as in (70). Details of this analysis will be omitted here.

5. Examples of spaces and partitions satisfying the hypotheses

This section is devoted to the presentation of some examples of spaces W_h and D_M and partitions $\mathscr{M}_h$ satisfying the hypotheses from Section 2. For simplicity, the discussion is restricted to the two-dimensional case. In three dimensions, the spaces can be constructed analogously (for details, see [30]). Throughout this section, $\{\mathscr{T}_h\}_{h>0}$ stands for a regular family of triangulations of $\overline{\Omega}$. This family is formed either by closed triangles or by closed convex quadrilaterals K with diameters h_K and one has $h = \max_{K\in\mathscr{T}_h} h_K$. Note that the hypotheses from Section 2, *e.g.*, (4), (6), and (7), do not allow the application of the analysis to anisotropic triangulations. In what follows, $\widehat{K}$ stands for a reference mesh cell, which is either a triangle or a square, depending on the type of elements in $\mathscr{T}_h$. For any $K \in \mathscr{T}_h$, there exists a bijective mapping $F_K : \widehat{K} \to K$ that maps $\widehat{K}$ onto K and is affine if $\widehat{K}$ is a triangle and bilinear if $\widehat{K}$ is a square. For any integer $l \ge 0$, we denote by P_l the space of polynomials of total degree at most l and by Q_l the space of polynomials of degree at most l in each variable. Finally, we set $R_l(\widehat{K}) = P_l(\widehat{K})$ if $\widehat{K}$ is a triangle and $R_l(\widehat{K}) = Q_l(\widehat{K})$ if $\widehat{K}$ is a square.

i) *The two-level approach.* This is the approach considered in the original local projection stabilization method (*cf.* [2,3]). The starting point is $\{\mathscr{M}_h\}_{h>0}$, a shape regular family of triangulations of $\overline{\Omega}$. Then, each triangle is divided into three triangles by connecting its vertices with the barycenter and each quadrilateral is divided into four quadrilaterals by connecting midpoints of opposite edges. The resulting triangulation is denoted by $\mathscr{T}_h$. Finally, given an integer $l \geq 1$, the spaces W_h and D_M are given by

$$W_h := \{v_h \in C(\overline{\Omega})\,;\ v_h|_K \circ F_K \in R_l(\widehat{K})\ \ \forall\, K \in \mathscr{T}_h\}, \qquad D_M := P_{l-1}(M). \tag{86}$$

The inf-sup condition (9) is proved for this pair in [30].

Alternatively, for the quadrilateral case, the space D_M could be defined as the space of mapped polynomials. More precisely, we can present the following two alternative definitions for D_M:

$$D_M^1 := \{v \in L^2(M)\,;\ v \circ F_M \in P_{l-1}(\widehat{M})\},$$
$$D_M^2 := \{v \in L^2(M)\,;\ v \circ F_M \in Q_{l-1}(\widehat{M})\},$$

where $\widehat{M}$ is a reference macro-cell and F_M is the analog of F_K. Both definitions lead to different methods (both different from the one presented so far) and have the advantage that the computations can be done directly on the reference element, leading to simpler implementations. All the approximation and stability assumptions hold for D_M^2, but for D_M^1 the approximation property (12) holds only on uniformly refined meshes (see [31], pp. 345-346 for a discussion on the topic).

ii) *The one-level approach.* This alternative was introduced in [30] and assumes $\mathscr{M}_h = \mathscr{T}_h$. Introducing a polynomial bubble function $b_{\widehat{K}} \in H_0^1(\widehat{K}) \setminus \{0\}$ (cubic if $\widehat{K}$ is a triangle and biquadratic if $\widehat{K}$ is a square), the spaces are given by

$$W_h := \{v_h \in C(\overline{\Omega})\,;\ v_h|_K \circ F_K \in R_l(\widehat{K}) + b_{\widehat{K}} \cdot R_{l-1}(\widehat{K})\ \ \forall\, K \in \mathscr{T}_h\}, \qquad D_M := P_{l-1}(M).$$

The inf-sup condition (9) is proved for this pair in [30].

iii) *The overlapping method.* Let $x_1, \ldots, x_{N_h}$ be the inner vertices of the triangulation $\mathscr{T}_h$, introduce the neighborhoods $M_i := \operatorname{int} \bigcup_{K \in \mathscr{T}_h, x_i \in K} K$ (where 'int' denotes the interior of the respective set), and define $\mathscr{M}_h := \{M_i\}_{i=1}^{N_h}$. The spaces W_h and D_M are given by (86). The inf-sup condition (9) is proved for this pair in [24].

In all of the examples above, i_h can be chosen to be the Lagrange interpolation operator and j_M to be the orthogonal L^2 projection of $L^2(M)$ onto D_M (see, *e.g.*, [12]). The validity of the geometrical hypotheses (4)-(7) follows from the mesh regularity. The inverse inequality (8) arises from a local inverse inequality (*cf.* [12]) and the mesh regularity. Finally, if F_K is linear for any $K \in \mathscr{T}_h$, then the space G_M consists of functions that are polynomial on the mesh cells included in M and the inverse inequality (10) is standard (*cf.* [12]).

Note that if the set $\mathscr{M}_h$ consists of nonoverlapping sets M, which is the case for both the one-level and two-level methods, then (significantly) more degrees of freedom are used for constructing the space W_h than in case of the method with overlapping sets M. This increase of the number of degrees of freedom is either due to an enrichment by bubble functions (in the one-level method) or due to a refinement of the given triangulation (in the two-level method). On the other hand, given a triangulation $\mathscr{T}_h$ of $\overline{\Omega}$ and using $\mathscr{M}_h$ consisting of overlapping sets M, the space W_h can be defined as a standard finite element space consisting of piecewise polynomials of degree l on $\mathscr{T}_h$, like in the Galerkin discretization.

6. Numerical illustrations

In this section, the theory of this paper is illustrated by results of numerical computations performed for both the steady-state problem (1) and the time-dependent problem (3). In addition, the reduction of spurious oscillations by applying the nonlinear crosswind diffusion is demonstrated. From the three possibilities for spaces

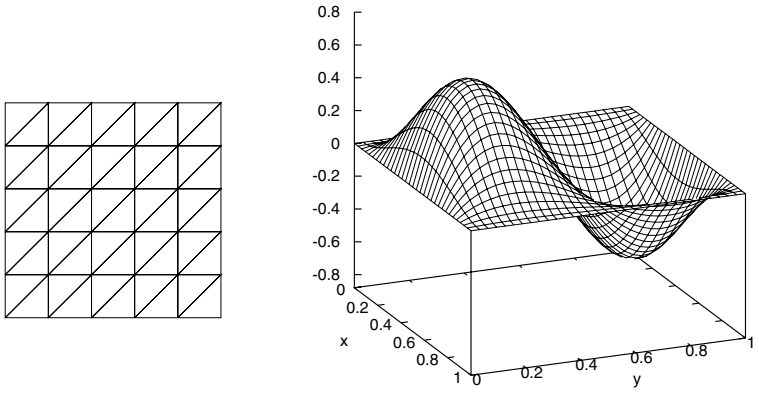


FIGURE 1. Type of the triangulations used in numerical computations (left) and solution for Example 6.1 (right).

and partitions proposed in the preceding section, we have chosen the overlapping version of the LPS method. This is mainly due to the fact that, as shown in [24], the overlapping version is more robust with respect to the stabilization parameter than both the one- and two-level approaches. The overlapping version was applied with triangular meshes and conforming piecewise linear approximation spaces W_h (thus $l = 1$). Both possible definitions (22) and (23) of $\tilde{\tau}_M(u_h)$ were considered. The solution of the nonlinear system was performed using a fixed point iteration: given an initial approximation $u_h^0 \in W_h$ of the solution of (19) satisfying $u_h^0 - \widetilde{u}_{bh} \in V_h$, compute a sequence $\{u_h^k\} \subset W_h$ defined by

$$u_n^k = u_h^{k-1} + \omega\,(\tilde{u}_h^k - u_h^{k-1}), \qquad k = 1, 2, \ldots,$$

where $\omega \in (0, 1]$ is a damping factor and $\tilde{u}_h^k \in W_h$ satisfies $\tilde{u}_h^k - \widetilde{u}_{bh} \in V_h$ and

$$a(\tilde{u}_h^k, v_h) + s_h(\tilde{u}_h^k, v_h) + d_h(u_h^{k-1}; \tilde{u}_h^k, v_h) = (f, v_h) \qquad \forall\ v_h \in V_h.$$

The analysis of the convergence of this scheme remains an open problem. Its proof, based on the properties of the nonlinear operator from Section 3, does not seem an easy task. The actual behavior of the iteration in our numerical studies will be discussed in Example 6.2.

In all examples, $\Omega = (0, 1)^2$ and Friedrichs-Keller triangulations of the type depicted in Figure 1 were used. It is worth mentioning that the mesh is not aligned with the considered convection fields.

Example 6.1. *Smooth polynomial solution [20], support of error estimates.* We considered problem (1) with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (3, 2)^T$, $c = 2$, and $u_b = 0$. The right-hand side f was chosen such that

$$u(x, y) = 100\, x^2\,(1 - x)^2\, y\,(1 - y)\,(1 - 2\,y)$$

is the solution of (1), see Figure 1.

In the stabilization parameters, the values $\tau_0 = 0.02$ and $\beta = 0.1$ were used. Table 1 shows errors of the discrete solutions measured in various norms for various mesh sizes. The notation $\|\cdot\|_{0,\infty,h}$ is used for the discrete L^∞ norm defined as the maximum of the errors at the vertices of the respective triangulation. The convergence orders were computed using values from the two finest triangulations. One can observe that the convergence order with respect to the LPS norm is $3/2$, as predicted by the theory, and that in other norms one obtains the usual optimal convergence orders.

TABLE 1. Example 6.1, errors of the discrete solutions.

	parameter (22)				parameter (23)			
h	$\|\cdot\|_{\mathrm{LPS}}$	$\|\cdot\|_{0,\Omega}$	$\|\cdot\|_{1,\Omega}$	$\|\cdot\|_{0,\infty,h}$	$\|\cdot\|_{\mathrm{LPS}}$	$\|\cdot\|_{0,\Omega}$	$\|\cdot\|_{1,\Omega}$	$\|\cdot\|_{0,\infty,h}$
8.84−2	4.74−2	1.83−2	4.20−1	6.46−2	4.30−2	1.47−2	4.00−1	5.04−2
4.42−2	1.48−2	3.54−3	1.88−1	1.52−2	1.41−2	2.93−3	1.84−1	1.13−2
2.21−2	5.02−3	7.24−4	9.02−2	3.40−3	4.93−3	6.57−4	8.96−2	2.44−3
1.10−2	1.76−3	1.58−4	4.45−2	7.63−4	1.75−3	1.57−4	4.44−2	5.57−4
5.52−3	6.19−4	3.63−5	2.21−2	1.77−4	6.18−4	3.83−5	2.21−2	1.44−4
order	1.50	2.12	1.01	2.11	1.50	2.03	1.01	1.95

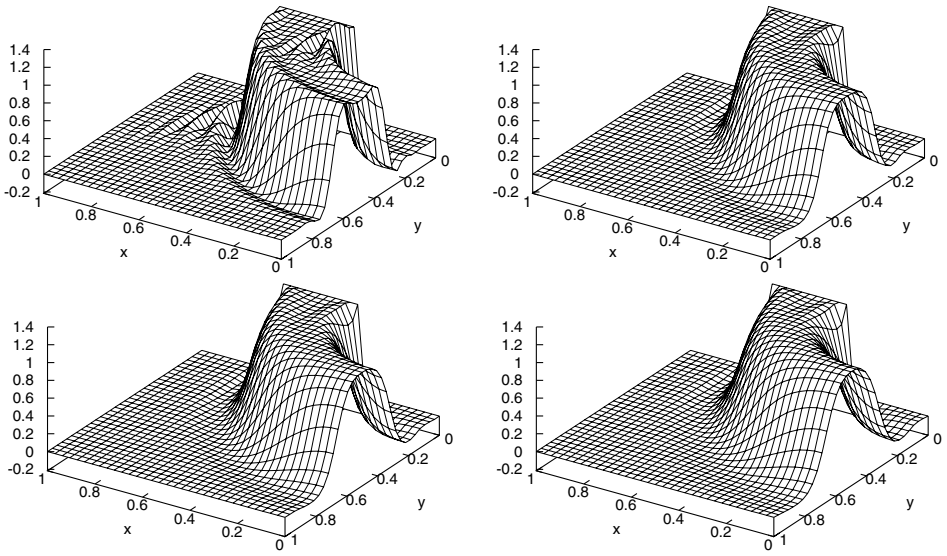


FIGURE 2. Example 6.2: solutions for the parameter (23) with $\tau_0 = 0.02$ and $\beta = 0$, $\beta = 0.03$, $\beta = 0.05$, $\beta = 0.1$, left to right, top to bottom.

Example 6.2. *Solution with two interior layers [27], reduction of spurious oscillations.* Equation (1) was considered with $\varepsilon = 10^{-8}$, $\boldsymbol{b}(x,y) = (-y,x)^T$, $c = f = 0$, and the boundary condition

$$u = u_b \quad \text{on } \Gamma^D, \qquad \frac{\partial u}{\partial \boldsymbol{n}} = 0 \quad \text{on } \Gamma^N,$$

where $\Gamma^N = \{0\} \times (0,1)$, $\Gamma^D = \partial\Omega \setminus \overline{\Gamma^N}$, $\boldsymbol{n}$ is the outward pointing unit normal vector to the boundary of Ω, and

$$u_b(x,y) = \begin{cases} 1 & \text{for } (x,y) \in (1/3, 2/3) \times \{0\}, \\ 0 & \text{else on } \Gamma^D. \end{cases}$$

Results that were obtained on the triangulation having 33×33 vertices are presented. Figure 2 shows solutions computed by means of the LPS method with and without the nonlinear crosswind diffusion term d_h defined using the parameter (23). One can observe that the crosswind diffusion term manages to reduce the oscillations appearing in the solution of the linear LPS method. An increase of the parameter β does not only reduce the oscillations but also increases the smearing appearing at the layers. In this respect, the method behaves

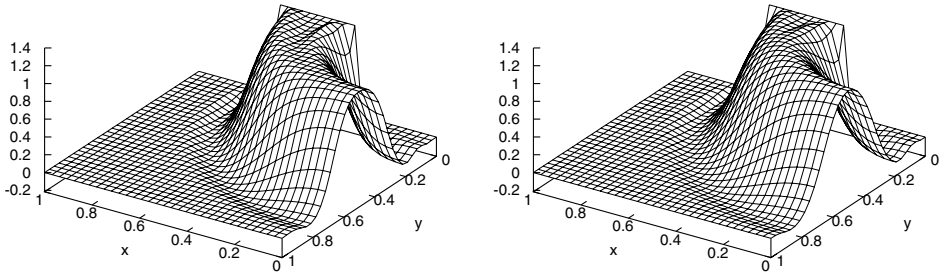


FIGURE 3. Example 6.2: solutions for the parameter (22) with $\tau_0 = 0.02$, $\beta = 0.03$ (left) and $\tau_0 = 0.02$, $\beta = 0.1$ (right).

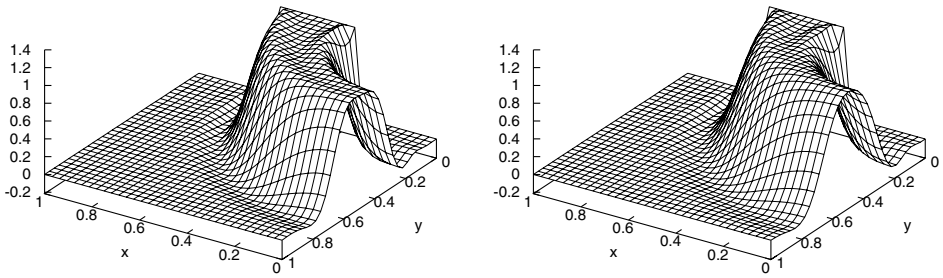


FIGURE 4. Example 6.2: solutions for the parameter (87) with $\tau_0 = 0.02$, $\beta = 0.025$ (left) and $\tau_0 = 0.02$, $\beta = 0.06$ (right).

as expected. Two results obtained for d_h defined using the parameter (22) are shown in Figure 3. A detailed comparison of the results in Figures 2 and 3 reveals that the method with the parameter (22) is less successful in suppressing spurious oscillations whereas it leads to a more pronounced smearing.

It is natural to ask whether similar results as presented above can be obtained using a linear crosswind diffusion term. To this end, the term d_h with

$$\tau_M^{\text{sold}} = \beta\, h_M\, |\boldsymbol{b}_M| \tag{87}$$

was considered. All other settings were the same as above. Since it is difficult to compare various solutions, we first concentrated on the outflow profile, *i.e.*, the solution graph along the line $x = 0$. For $\beta \leq 0.02$, the outflow profile contains overshoots that decrease with increasing β. Figure 4 shows that, for $\beta = 0.025$, the overshoots are not present in the outflow profile but they can be still observed inside the computational domain. For this value of β, the outflow profile does not differ too much from the outflow profile in Figure 2, top right. However, inside the computational domain, both overshoots and undershoots are larger for the linear method. A further increase of β leads to a reduction of the overshoots but also to a smearing of the solution whereas the magnitude of the undershoots does not change significantly. As an example, the solution for $\beta = 0.06$ is shown in Figure 4. The smearing and the undershoots of this solution are more pronounced than in case of all the three solutions of the nonlinear method in Figure 2. This study demonstrates that the method with linear crosswind diffusion was outperformed, with respect to the quality of the computed solution, by the nonlinear method with $\tilde{\tau}_M$ defined by (23).

From the discussion of the preceding paragraphs, the choice of the stabilization parameter β appears as an important issue. A good choice of user-chosen parameters in stabilized finite element methods is an open

TABLE 2. Example 6.2, number of fixed-point iterations.

	parameter (22)				parameter (23)			
	$\beta = 0.01$	$\beta = 0.03$	$\beta = 0.06$	$\beta = 0.10$	$\beta = 0.01$	$\beta = 0.03$	$\beta = 0.06$	$\beta = 0.10$
$\omega = 1.0$	82	163	305	494	16	27	39	51
$\omega = 0.9$	42	58	68	73	12	18	24	29
$\omega = 0.8$	25	30	32	33	12	13	16	19
$\omega = 0.7$	16	17	18	20	16	16	16	16
$\omega = 0.6$	20	20	20	20	21	21	21	21
$\omega = 0.5$	27	27	27	27	27	27	27	27

problem for all methods. In general, the parameters need to be chosen not constant but as functions (see [18] for the construction of an example). A non-constant choice, done automatically like in [19], will be the subject of future research.

Next, the computational cost connected with the solution of the nonlinear discrete problems will be briefly illustrated. Table 2 shows numbers of fixed-point iterations needed to solve Example 6.2 for $\tau_0 = 0.02$ and various values of β and the damping parameter ω. The iterative process was terminated if the Euclidean norm of the residual of the nonlinear algebraic system divided by the Euclidean norm of its right-hand side was smaller than 10^{-8}. The sequences of the residuals were monotonically decreasing, except for some of the computations with the parameter (22) for $\omega \in \{0.9, 1\}$ where oscillations of the residuals appeared at the beginning of the iterative process. One can observe that the number of iterations depends both on β and ω and that this dependence is more pronounced if the parameter $\tilde{\tau}_M$ is defined by (22). Since the optimal value of the damping parameter is usually not known, it can be expected that the numerical effort caused by the nonlinear crosswind diffusion term will be generally smaller if the parameter $\tilde{\tau}_M$ is defined by (23).

Example 6.3. *Smooth time-dependent solution, support of error estimates.* The setup of this example is very similar to Example 6.1 in [22]. Problem (3) was considered in the time interval $[0, 1]$ with $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (3, 2)^T$, $c = 2$, and $u_b = 0$. The right-hand side f and the initial condition u_0 were chosen such that

$$u(x, y, t) = \mathrm{e}^{\sin(2\,\pi\,t)}\, \sin(2\,\pi\,x)\, \sin(2\,\pi\,y)$$

is the solution of (3).

We considered the discrete problem (64) and its linearized variant (85) with $\theta = 1$ (*i.e.*, the backward Euler scheme) for both choices of $\tilde{\tau}_M$. Like in Example 6.1, the values $\tau_0 = 0.02$ and $\beta = 0.1$ were used for the stabilization parameters. According to error estimates (71) and (72), one expects that the quantity E^N tends to zero with the convergence order 3/2 if $\delta t \sim h^{3/2}$ and a nonlinear discretization is used (note the extra power of $h^{1/2}$ in Q^N and R^N). The same convergence behavior is expected for the linearized method if $\tilde{\tau}_M$ is defined by (23), see the discussion at the end of Section 4. These expectations are supported by the results presented in Figure 5. In this figure, level 1 corresponds to the grid with mesh cells of diameter $h = \sqrt{2}\,\tilde{h}$ with $\tilde{h} = 1/8$. Uniform refinement in space was used and the length of the time step was set to be $\delta t = \tilde{h}^{3/2}$. If the final time was not obtained exactly with these time steps, the simulations were terminated at the last discrete time smaller than $T = 1$. It can be observed in Figure 5 that the order of convergence 3/2 was obtained for the error in the l^2-LPS norm for all four methods. We could observe the same order of convergence also for $\|e^N\|_{0,\Omega}$. Using the time step $\delta t = \tilde{h}^2$, the error $\|e^N\|_{0,\Omega}$ showed even second order convergence, whereas the order of convergence of the error in the l^2-LPS norm was still 3/2. This result demonstrates the sharpness of the estimates (71) and (72).

Concerning a comparison of the fully nonlinear and the linearized version of the methods, only very little differences can be seen in this example. On coarser grids, the solutions computed using the parameter (23) were more accurate compared with the solutions obtained using the parameter (22).

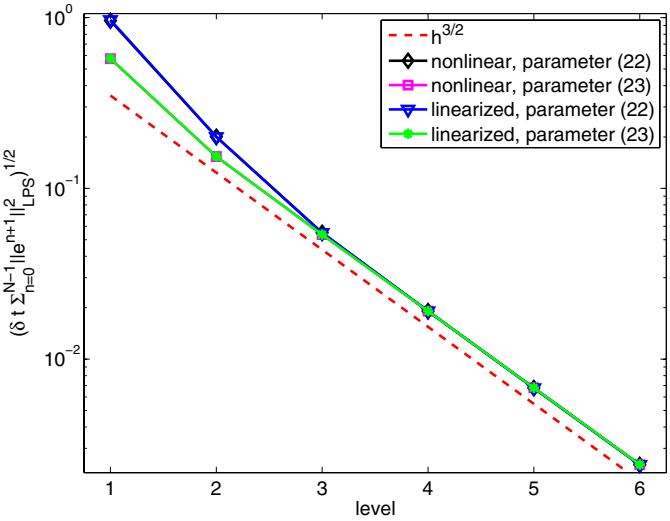


FIGURE 5. Example 6.3: order of convergence for piecewise linear finite elements, the backward Euler scheme, and $\delta t \sim h^{3/2}$. Note that the curves for the linearized methods are on top of the curves of the corresponding nonlinear method.

Acknowledgements. The work of G.R. Barrenechea has been partially funded by The Leverhulme Trust, through the Research Project Grant RPG-2012-483. The work of P. Knobloch is a part of the research project MSM 0021620839 funded by the Ministry of Education, Youth and Sports of the Czech Republic and it was partly supported by the Grant Agency of the Czech Republic under the grant No. P201/11/1304.

REFERENCES

[1] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke and R. Umla, An assessment of discretizations for convection-dominated convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.* **200** (2011) 3395–3409.

[2] R. Becker and M. Braack, A finite element pressure gradient stabilization for the Stokes equations based on local projections. *Calcolo* **38** (2001) 173–199.

[3] R. Becker and M. Braack, A two-level stabilization scheme for the Navier-Stokes equations, Proc. of ENUMATH 2003, *Numerical Mathematics and Advanced Applications*, edited by M. Feistauer, V. Dolejší, P. Knobloch and K. Najzar. Springer-Verlag, Berlin (2004) 123–130.

[4] M. Braack and E. Burman, Local projection stabilization for the Oseen problem and its interpretation as a variational multiscale method. *SIAM J. Numer. Anal.* **43** (2006) 2544–2566.

[5] M. Braack, E. Burman, V. John and G. Lube, Stabilized finite element methods for the generalized Oseen problem. *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 853–866.

[6] 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.

[7] E. Burman and A. Ern, Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence. *Math. Comput.* **74** (2005) 1637–1652.

[8] E. Burman and M.A. Fernández, Finite element methods with symmetric stabilization for the transient convection-diffusion-reaction equation. *Comput. Methods Appl. Mech. Engrg.* **198** (2009) 2508–2519.

[9] E. Burman and P. Hansbo, Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.* **193** (2004) 1437–1453.

[10] P.G. Ciarlet, *The Finite Element Method for Elliptic Problems*. North-Holland, Amsterdam (1978).

[11] 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.

[12] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*. Springer-Verlag, New York (2004).

[13] 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.

[14] L.P. Franca and F. Valentin, On an improved unusual stabilized finite element method for the advective-reactive-diffusive equation. *Comput. Methods Appl. Mech. Engrg.* **190** (2000) 1785–1800.

[15] S. Ganesan and L. Tobiska, Stabilization by local projection for convection-diffusion and incompressible flow problems. *J. Sci. Comput.* **43** (2010) 326–342.

[16] 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.

[17] V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: Part I – A review. *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 2197–2215.

[18] V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.* **197** (2008) 1997–2014.

[19] V. John, P. Knobloch and S.B. Savescu, A posteriori optimization of parameters in stabilized methods for convection-diffusion problems – Part I. *Comput. Methods Appl. Mech. Engrg.* **200** (2011) 2916–2929.

[20] V. John, J.M. Maubach and L. Tobiska, Nonconforming streamline-diffusion-finite-element-methods for convection-diffusion problems. *Numer. Math.* **78** (1997) 165–188.

[21] V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska and A. Voigt, Simulations of population balance systems with one internal coordinate using finite element methods. *Chem. Engrg. Sci.* **64** (2009) 733–741.

[22] V. John and J. Novo, Error analysis of the SUPG finite element discretization of evolutionary convection-diffusion-reaction equations. *SIAM J. Numer. Anal.* **49** (2011) 1149–1176.

[23] V. John and E. Schmeyer, Finite element methods for time-dependent convection-diffusion-reaction equations with small diffusion. *Comput. Methods Appl. Mech. Engrg.* **198** (2008) 475–494.

[24] P. Knobloch, A generalization of the local projection stabilization for convection-diffusion-reaction equations. *SIAM J. Numer. Anal.* **48** (2010) 659–680.

[25] P. Knobloch, Local projection method for convection-diffusion-reaction problems with projection spaces defined on overlapping sets. Proc. of ENUMATH 2009, *Numerical Mathematics and Advanced Applications*, edited by G. Kreiss, P. Lötstedt, A. Målqvist and M. Neytcheva. Springer-Verlag, Berlin (2010) 497–505.

[26] P. Knobloch and G. Lube, Local projection stabilization for advection-diffusion-reaction problems: One-level vs. two-level approach. *Appl. Numer. Math.* **59** (2009) 2891–2907.

[27] 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.

[28] O.A. Ladyzhenskaya, New equations for the description of motion of viscous incompressible fluids and solvability in the large of boundary value problems for them. *Tr. Mat. Inst. Steklova* **102** (1967) 85–104.

[29] G. Lube and G. Rapin, residual-based stabilized higher-order FEM for advection-dominated problems. *Comput. Methods Appl. Mech. Engrg.* **195** (2006) 4124–4138.

[30] G. Matthies, P. Skrzypacz and L. Tobiska, A unified convergence analysis for local projection stabilizations applied to the Oseen problem. *Math. Model. Numer. Anal.* **41** (2007) 713–742.

[31] H.-G. Roos, M. Stynes and L. Tobiska, Robust Numerical Methods for Singularly Perturbed Differential Equations. *Convection-Diffusion-Reaction and Flow Problems,* 2nd ed. Springer-Verlag, Berlin (2008).

[32] R. Temam, Navier-Stokes Equations. Theory and Numerical Analysis North-Holland, Amsterdam (1977).

Chapter 6

Algebraic flux correction

This chapter consists of the following publications:

G.R. Barrenechea, V. John, P. Knobloch: Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension, *IMA Journal of Numerical Analysis* 35 (4): 1729–1756, 2015. p. 267

G.R. Barrenechea, V. John, P. Knobloch: Analysis of algebraic flux correction schemes, *SIAM Journal on Numerical Analysis* 54 (4): 2427–2451, 2016. p. 295

G.R. Barrenechea, V. John, P. Knobloch: An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes, *Mathematical Models and Methods in Applied Sciences* 27 (3): 2017, doi: 10.1142/S0218202517500087, in press. p. 321

IMA Journal of Numerical Analysis (2015) **35**, 1729–1756
doi:10.1093/imanum/dru041
Advance Access publication on October 17, 2014

Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension

Gabriel R. Barrenechea
Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, UK
gabriel.barrenechea@strath.ac.uk

Volker John
Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany and Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany
john@wias-berlin.de

and

Petr Knobloch*
Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic
*Corresponding author: knobloch@karlin.mff.cuni.cz

[Received on 12 December 2013; revised on 29 April 2014]

Algebraic flux correction schemes are nonlinear discretizations of convection-dominated problems. In this work, a scheme from this class is studied for a steady-state convection–diffusion equation in one dimension. It is proved that this scheme satisfies the discrete maximum principle. Also, as it is a nonlinear scheme, the solvability of the linear subproblems arising in a Picard iteration is studied, where positive and negative results are proved. Furthermore, the nonexistence of solutions for the nonlinear scheme is proved by means of counterexamples. Therefore, a modification of the method, which ensures the existence of a solution, is proposed. A weak version of the discrete maximum principle is proved for this modified method.

Keywords: finite element method; convection–diffusion equation; algebraic flux correction; discrete maximum principle; fixed-point iteration; solvability of linear subproblems; solvability of nonlinear problem.

1. Introduction

Scalar convection–diffusion equations model the convective and diffusive transport of a scalar quantity, such as temperature or concentration. Solutions of convection-dominated convection–diffusion equations typically possess layers, which cannot be resolved unless the given mesh is sufficiently fine in layer regions. Standard discretizations, such as central finite differences or the Galerkin finite element method, cannot cope with this situation and the computed solutions are globally polluted with spurious oscillations. It is well known that so-called stabilized discretizations have to be applied. There are many proposals of such discretizations; see the monograph Roos *et al.* (2008) for an extensive review.

In the past few years, comprehensive numerical studies revealed, however, that none of the proposed stabilized discretizations satisfies the following three requirements: accuracy, efficiency and numerical solution without spurious oscillations (discrete maximum principle). This statement holds true for the steady-state equation (John & Knobloch, 2007, 2008; Augustin *et al.*, 2011; Bause & Schwegler, 2012; John & Schumacher, 2014) as well as for the time-dependent equation (Codina, 1998; John & Schmeyer, 2008; John & Novo, 2012). Indeed, most of the methods fail to satisfy a discrete maximum principle. However, this property is particularly important in applications, where numerical results, e.g., with negative concentrations, will be considered to be worthless. Even if such quantities are not of primary interest, spurious oscillations have been shown to lead to blow-ups in the simulation of coupled problems (John *et al.*, 2009). Altogether, the validity of a discrete maximum principle is, in our opinion, of utmost importance for simulations of applications.

There are few discretizations that satisfy a discrete maximum principle, such as the upwind finite difference scheme (Roos *et al.*, 2008), a finite volume scheme on Delaunay meshes Fuhrmann & Langmach (2001) and algebraic flux correction schemes. The first two methods are generally rather inaccurate, while the algebraic flux correction schemes are usually nonlinear discretizations and their application might be time consuming. However, applications often lead to nonlinear models, and then a nonlinear discretization of a linear equation in such a model seems not to be a severe disadvantage. Altogether, from the point of view of applications, algebraic flux correction schemes are very attractive.

The basic philosophy of flux correction schemes was formulated in the 1970s in Boris & Book (1973) and Zalesak (1979). Later, the idea was applied in the finite element context, e.g., in Löhner *et al.* (1987) and Arminjon & Dervieux (1993). In the last decade, the methods have been further developed and refined, in particular in Kuzmin & Turek (2004), Kuzmin & Möller (2005) and Kuzmin (2006, 2007, 2008, 2009, 2012). Until not long ago, two limiting techniques within algebraic flux correction schemes were pursued: so-called flux-corrected transport (FCT) schemes for the time-dependent equation and total variation diminishing (TVD) schemes for the steady-state equation. Finally, a scheme was presented in Kuzmin (2012) that can handle both situations. For the time-dependent problem, a linear variant of an FCT scheme was proposed in Kuzmin (2009).

Despite the attractiveness of algebraic flux correction schemes, there seems to be no rigorous numerical analysis for this class of methods. The main reason lies probably in their construction, which does not allow the usual tools of the analysis of finite element discretizations to be applied. Unlike almost all other stabilized methods, which modify the bilinear form of the discrete problem in some way, algebraic flux correction schemes work on the algebraic level. They manipulate the matrix and the right-hand side of the algebraic system of equations. A few basic properties of these schemes can be deduced immediately from their construction, such as mass conservation or the discrete maximum principle for transport equations (Kuzmin & Möller, 2005).

In this work we study some properties of a nonlinear discrete problem that generalizes the algebraic flux correction method of TVD type from Kuzmin (2007) applied to the one-dimensional steady-state convection–diffusion equation. We present both theoretical and computational results; the latter are obtained by solving the nonlinear discrete problem using a fixed-point iteration. While the linear subproblems in the fixed-point iteration are proved to be well posed, the nonlinear problem is shown to be not solvable in general. However, we prove the solvability for a modified nonlinear discrete problem. To the authors' best knowledge, the results concerning the solvability of the linear subproblems and the nonlinear problem are the first results of this kind for algebraic flux correction schemes. In addition, the present work represents a basis for analysing algebraic flux correction schemes applied to multidimensional problems.

The paper is organized in the following way. First, the algebraic flux correction method will be introduced in Section 2. In Section 3, the one-dimensional model problem will be formulated and its finite element discretization will be presented. The application of the algebraic flux correction method to this problem is the topic of Section 4. It will be shown there that the discrete operator of this scheme can be written as a nonlinear finite difference operator with an artificial diffusion vector whose components are bounded by a data-dependent constant $\tilde{\varepsilon}$. In Section 5, the discrete maximum principle for this operator will be proved for appropriately chosen values of $\tilde{\varepsilon}$. Different choices of $\tilde{\varepsilon}$, for which the discrete maximum principle is satisfied, will be studied numerically in Section 6. The unique solvability of the linear subproblems arising in the fixed-point iteration is studied in Section 7 under more general conditions on the artificial diffusion vector than from the actual method (Kuzmin, 2007). Some positive but also a negative result are proved. Section 8 starts with a number of counterexamples concerning the solvability of the nonlinear discrete problem. Then, the existence of a solution of the nonlinear problem is proved for a modification of the method. A concrete realization of this modification is proposed in Section 9, where a weak form of the discrete maximum principle is proved and numerical results are presented. Finally, a summary and an outlook are given in Section 10.

2. An algebraic flux correction scheme

Consider a linear boundary value problem whose solution is (mainly) determined by convection and for which the maximum principle holds. Let us discretize this problem by the finite element method. Then, the discrete solution can be represented by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of U ($0 < M < N$) correspond to nodes where Dirichlet boundary conditions are prescribed, whereas the first M components of U are computed using the finite element discretization of the underlying partial differential equation. Then $\mathrm{U} \equiv (u_1, \ldots, u_N)$ satisfies a system of linear equations of the form

$$\sum_{j=1}^{N} a_{ij} u_j = g_i, \quad i = 1, \ldots, M, \tag{2.1}$$

$$u_i = u_i^{\mathrm{b}}, \quad i = M+1, \ldots, N. \tag{2.2}$$

We assume that

$$a_{ii} > 0, \quad \sum_{j=1}^{N} a_{ij} = 0, \quad i = 1, \ldots, M, \tag{2.3}$$

which is often the case when incompressible convection fields are considered.

Since the original problem satisfies the maximum principle, it is natural to require that this property is inherited by the discrete problem. Unfortunately, the discrete maximum principle does not hold for many finite element discretizations of convection-dominated problems, in particular, for the Galerkin discretization and most stabilized methods; see, e.g., Roos *et al.* (2008). The aim of algebraic flux correction approaches is to cure this deficiency by manipulating the algebraic system in such a way that the solution satisfies the discrete maximum principle and layers are not excessively smeared.

The starting point of the algebraic flux correction algorithm is the finite element matrix $\mathbb{A} = (a_{ij})_{i,j=1}^{N}$ corresponding to the above finite element discretization in the case where homogeneous natural boundary conditions are used instead of the Dirichlet ones. We introduce the symmetric artificial diffusion

matrix $\mathbb{D} = (d_{ij})_{i,j=1}^{N}$ possessing the entries

$$d_{ij} = -\max\{a_{ij}, 0, a_{ji}\} \ \ \forall i \neq j, \quad d_{ii} = -\sum_{j \neq i} d_{ij}.$$

Then, the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ has nonpositive off-diagonal entries and each of its row sums vanishes. A vector $\mathrm{U} \in \mathbb{R}^N$ being a solution of a linear system with the matrix $\tilde{\mathbb{A}}$ satisfies the discrete maximum principle in the sense that for any $i \in \{1, \ldots, M\}$ the following holds:

$$(\tilde{\mathbb{A}}\mathrm{U})_i \leqslant 0 \ \ \Rightarrow \ \ u_i \leqslant \max_{j \neq i, \tilde{a}_{ij} \neq 0} u_j.$$

This property immediately follows from the fact that, using (2.3), one gets

$$\tilde{a}_{ii} u_i \leqslant -\sum_{j \neq i} \tilde{a}_{ij} u_j = \tilde{a}_{ii} c - \sum_{j \neq i} \tilde{a}_{ij} (u_j - c) \leqslant \tilde{a}_{ii} c \quad \forall\, c \geqslant \max_{j \neq i,\ \tilde{a}_{ij} \neq 0} u_j.$$

Going back to the solution of system (2.1), this system is equivalent to

$$(\tilde{\mathbb{A}}\mathrm{U})_i = g_i + (\mathbb{D}\mathrm{U})_i, \quad i = 1, \ldots, M. \tag{2.4}$$

Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\mathrm{U})_i = \sum_{j \neq i} f_{ij}, \quad i = 1, \ldots, N,$$

where $f_{ij} = d_{ij}(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. Now, the idea of the algebraic flux correction schemes is to limit those antidiffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (2.1) (or, equivalently (2.4)) is replaced by

$$(\tilde{\mathbb{A}}\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad i = 1, \ldots, M, \tag{2.5}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (2.1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem.

The coefficients α_{ij} can be chosen in various ways but their definition is always based on the above fluxes f_{ij}; see Kuzmin (2006, 2007, 2008, 2009, 2012) for examples. In this work we consider coefficients α_{ij} proposed in Kuzmin (2007). This definition relies on the values P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i = 1, \ldots, N$ in the following way. First, one initializes all these quantities with 0. Then one goes through all pairs of indices $i, j \in \{1, \ldots, N\}$ and if $a_{ji} \leqslant a_{ij}$, one performs the updates

$$P_i^+ := P_i^+ + \max\{0, f_{ij}\}, \quad P_i^- := P_i^- - \max\{0, f_{ji}\}, \tag{2.6}$$

$$Q_i^+ := Q_i^+ + \max\{0, f_{ji}\}, \quad Q_i^- := Q_i^- - \max\{0, f_{ij}\}, \tag{2.7}$$

$$Q_j^+ := Q_j^+ + \max\{0, f_{ij}\}, \quad Q_j^- := Q_j^- - \max\{0, f_{ji}\}. \tag{2.8}$$

After having computed the values P_i^+, P_i^-, Q_i^+, Q_i^-, $i = 1, \ldots, N$, one sets

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \ldots, N.$$

Finally, the coefficients α_{ij} are defined by

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ R_i^- & \text{if } f_{ij} < 0, \end{cases} \quad i, j = 1, \ldots, N.$$

3. Finite element discretization of a one-dimensional convection–diffusion equation

To better understand the algebraic flux correction method described in the previous section, we shall apply it to a finite element discretization of a scalar one-dimensional convection–diffusion equation. In this section we formulate the one-dimensional problem, introduce its discretization, and for completeness, we review its main characteristics.

We consider the boundary value problem

$$-\varepsilon u'' + bu' = g \ \text{ in } (0,1), \quad u(0) = u_{\mathrm{L}}, \quad u(1) = u_{\mathrm{R}}, \tag{3.1}$$

where, for simplicity, ε and b are assumed to be positive constants. Moreover, g is supposed to belong to $L^2(0,1)$ and u_{L}, u_{R} are any real numbers. If g is constant, then the solution of (3.1) is given by the formula

$$u(x) = u_{\mathrm{L}} + \frac{g}{b}x + \gamma \frac{e^{-(1-x)b/\varepsilon} - e^{-b/\varepsilon}}{1 - e^{-b/\varepsilon}} \tag{3.2}$$

with $\gamma := u_{\mathrm{R}} - u_{\mathrm{L}} - g/b$. Thus, for $\gamma \neq 0$ and $\varepsilon \ll b$, the solution of (3.1) possesses a boundary layer at the right-hand boundary point.

Let us divide the interval $[0,1]$ into $n+1$ subintervals $[x_i, x_{i+1}]$, $i = 0, \ldots, n$, with $x_i = ih$ and $h = 1/(n+1)$. We define the finite element space

$$W_h = \{v_h \in C([0,1]); v_h|_{[x_i, x_{i+1}]} \in P_1([x_i, x_{i+1}]),\ i = 0, \ldots, n\}$$

consisting of continuous piecewise linear functions and set

$$V_h = \{v_h \in W_h; v_h(0) = v_h(1) = 0\}.$$

Then the Galerkin finite element discretization of (3.1) reads, find $u_h \in W_h$ such that $u_h(0) = u_{\mathrm{L}}$, $u_h(1) = u_{\mathrm{R}}$ and

$$\varepsilon(u_h', v_h') + (bu_h', v_h) = (g, v_h) \quad \forall\, v_h \in V_h, \tag{3.3}$$

where $(\cdot, \cdot)$ denotes the inner product in $L^2(0,1)$.

Let us denote by $\varphi_1, \ldots, \varphi_n \in V_h$ the usual basis functions of V_h, i.e., $\varphi_i(x_j) = \delta_{ij}$ for $i,j = 1, \ldots, n$. We define

$$g_i = \frac{1}{h}(g, \varphi_i), \quad i = 1, \ldots, n.$$

Setting $u_i = u_h(x_i)$, $i = 0, \ldots, n+1$, then (3.3) is equivalent to the system

$$-\varepsilon \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + b\frac{u_{i+1} - u_{i-1}}{2h} = g_i, \quad i = 1, \ldots, n. \tag{3.4}$$

This system can be also obtained by discretizing (3.1) using the central finite difference method. Then, however, $g_i = g(x_i)$.

Let us introduce the Péclet number

$$\mathrm{Pe} = \frac{bh}{2\varepsilon}$$

and let g be constant. If $\mathrm{Pe} = 1$, then (3.4) reduces to

$$b\frac{u_i - u_{i-1}}{h} = g, \quad i = 1, \ldots, n,$$

and hence $u_i = u_\mathrm{L} + (g/b)x_i$, $i = 0, \ldots, n$. Thus, in this case,

$$u_h(x) = u_\mathrm{L} + \frac{g}{b}x, \quad x \in [0, 1-h].$$

If $\mathrm{Pe} \neq 1$, then

$$u_i = \frac{g}{b}x_i + A + B\left(\frac{1+\mathrm{Pe}}{1-\mathrm{Pe}}\right)^i, \quad i = 0, \ldots, n+1, \tag{3.5}$$

where A and B are determined by the conditions $u_0 = u_\mathrm{L}$ and $u_{n+1} = u_\mathrm{R}$. We observe that, for $\mathrm{Pe} < 1$, the discrete solution is the sum of two monotone grid functions but, for $\mathrm{Pe} > 1$, the discrete solution u_i generally possesses spurious oscillations. This shows that the Galerkin discretization is not appropriate for solving (3.1) numerically if $\mathrm{Pe} > 1$.

4. The algebraic flux correction scheme applied to the one-dimensional problem

To suppress the spurious oscillations in the solutions of the Galerkin finite element discretization of (3.1) given by (3.3), we shall apply the algebraic flux correction scheme described in Section 2. We shall assume that $\mathrm{Pe} > 1$, which is the interesting case in practice.

The Galerkin discretization of (3.1) introduced in the previous section corresponds to the system from Section 2 with $N = n + 2$ but with a different numbering of the nodes. The matrices $\mathbb{A}$ and $\mathbb{D}$ are

tridiagonal $(n+2)\times(n+2)$ matrices with entries (cf. (3.4))

$$
\begin{aligned}
&a_{0,0}=\frac{\varepsilon}{h^2}-\frac{b}{2h},\quad a_{0,1}=-\frac{\varepsilon}{h^2}+\frac{b}{2h},\\
&a_{i,i-1}=-\frac{\varepsilon}{h^2}-\frac{b}{2h},\ a_{i,i}=\frac{2\varepsilon}{h^2},\ a_{i,i+1}=-\frac{\varepsilon}{h^2}+\frac{b}{2h},\quad i=1,\dots,n,\\
&a_{n+1,n}=-\frac{\varepsilon}{h^2}-\frac{b}{2h},\quad a_{n+1,n+1}=\frac{\varepsilon}{h^2}+\frac{b}{2h},\\
&d_{i,i+1}=\frac{\varepsilon}{h^2}-\frac{b}{2h},\quad i=0,\dots,n.
\end{aligned}
\tag{4.1}
$$

The vector U in (2.5) is given by $\mathrm{U}=(u_0,u_1,\dots,u_{n+1})^{\mathrm{T}}$. Note that the assumption (2.3) is satisfied.

Now let us compute the values α_{ij} in (2.5). The values α_{ij} are needed only for $i=1,\dots,n$ and $|i-j|=1$, and they are not important if $f_{ij}=0$. Since $f_{ij}\neq 0$ only if $|i-j|=1$, and $a_{i+1,i}<a_{i,i+1}$ for $i=0,\dots,n$, the updates (2.6–2.8) have to be computed only for $j=i+1$, $i=0,\dots,n$. This readily gives

$$
\begin{aligned}
&P_i^+=\max\{0,f_{i,i+1}\},\quad P_i^-=-\max\{0,f_{i+1,i}\},\\
&Q_i^+=\max\{0,f_{i-1,i}\}+\max\{0,f_{i+1,i}\},\quad Q_i^-=-\max\{0,f_{i,i-1}\}-\max\{0,f_{i,i+1}\}
\end{aligned}
$$

for $i=1,\dots,n$. Thus, for $i=1,\dots,n$, one obtains

$$
\alpha_{i,i-1}=\begin{cases}\min\left\{1,\dfrac{\max\{0,f_{i+1,i}\}}{\max\{0,f_{i,i+1}\}}\right\} & \text{if } f_{i,i-1}>0,\\ \min\left\{1,\dfrac{\max\{0,f_{i,i+1}\}}{\max\{0,f_{i+1,i}\}}\right\} & \text{if } f_{i,i-1}<0,\end{cases}
$$

$$
\alpha_{i,i+1}=\begin{cases}\min\left\{1,\dfrac{\max\{0,f_{i-1,i}\}}{f_{i,i+1}}\right\} & \text{if } f_{i,i+1}>0,\\ \min\left\{1,\dfrac{\max\{0,f_{i,i-1}\}}{f_{i+1,i}}\right\} & \text{if } f_{i,i+1}<0.\end{cases}
$$

It is not completely clear how to interpret the definition of $\alpha_{i,i-1}$ when the denominator vanishes. In this case we always set $\alpha_{i,i-1}=1$. This leads to

$$
\begin{aligned}
&\alpha_{i,i-1}=\alpha_{i,i+1}=0\quad \text{if } f_{i,i-1}f_{i,i+1}>0,\\
&\alpha_{i,i-1}=1,\quad \alpha_{i,i+1}=\min\left\{1,\frac{f_{i-1,i}}{f_{i,i+1}}\right\}\quad \text{if } f_{i,i-1}f_{i,i+1}\leqslant 0.
\end{aligned}
$$

Setting

$$
\beta_i=\begin{cases}1 & \text{if } f_{i,i+1}\neq 0 \text{ and } \dfrac{f_{i-1,i}}{f_{i,i+1}}<1,\\ 0 & \text{otherwise,}\end{cases}\qquad i=1,\dots,n,
$$

system (2.5) is equivalent to

$$\begin{aligned} &u_0 = u_{\mathrm{L}}, \\ &(\mathbb{A}\mathrm{U})_i + \beta_i(f_{i,i-1} + f_{i,i+1}) = g_i, \quad i = 1, \dots, n, \\ &u_{n+1} = u_{\mathrm{R}}. \end{aligned}$$

The definition of the coefficients β_i can be written also in the form

$$\beta_i = \begin{cases} 1 & \text{if } u_i \neq u_{i+1} \text{ and } \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < 1, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \dots, n. \tag{4.2}$$

Finally, applying

$$f_{i,i-1} + f_{i,i+1} = \left(\frac{\varepsilon}{h^2} - \frac{b}{2h} \right) (u_{i-1} - 2u_i + u_{i+1}), \quad i = 1, \dots, n,$$

and setting

$$\tilde{\varepsilon} = \frac{bh}{2} - \varepsilon = \varepsilon(\mathrm{Pe} - 1), \tag{4.3}$$

one arrives at the following final version of the algebraic flux correction scheme.

Find $u_0, \dots, u_{n+1}$ such that

$$u_0 = u_{\mathrm{L}}, \quad u_{n+1} = u_{\mathrm{R}}, \tag{4.4}$$

and

$$-(\varepsilon + \beta_i \tilde{\varepsilon}) \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + b \frac{u_{i+1} - u_{i-1}}{2h} = g_i, \quad i = 1, \dots, n. \tag{4.5}$$

Since definitions of β_i other than (4.2) may be convenient also (see the end of this section), we shall analyse the flux correction scheme (4.4), (4.5) for a class of functions β_i satisfying

$$\beta_i \in \{0, 1\}, \quad \beta_i = 1 \ \text{ if } (u_i - u_{i-1})(u_{i+1} - u_i) < 0, \quad i = 1, \dots, n. \tag{4.6}$$

Note that functions β_i defined by (4.2) satisfy (4.6).

REMARK 4.1 Some comments on this method are in order.

1. Condition (4.6) ensures that artificial diffusion is added to the equation at the node x_i whenever the discrete solution has a local extremum at x_i.
2. If $\beta_i = 1$, then the corresponding equation in (4.5) reduces to

$$b \frac{u_i - u_{i-1}}{h} = g_i. \tag{4.7}$$

Thus, in this case the method transforms (locally) the original Galerkin method into an upwinded discretization of the hyperbolic equation $bu' = g$.

3. There are alternative ways to define the matrix $\mathbb{D}$. For example, if it is defined with respect to the convection matrix only, i.e., setting $\varepsilon = 0$ in (4.1), one obtains (4.5) with

$$\tilde{\varepsilon} = \frac{bh}{2}. \tag{4.8}$$

If $\beta_i = 1$, then scheme (4.5) becomes

$$-\varepsilon \frac{u_{i-1} - 2u_i + u_{i+1}}{h^2} + b\frac{u_i - u_{i-1}}{h} = g_i, \tag{4.9}$$

which is the usual upwind discretization of (3.1) at the node x_i. This approach was used, e.g., in Kuzmin (2012). The definition of $\mathbb{D}$ using the whole matrix $\mathbb{A}$, as it was considered in this section, makes the implementation of the method simpler (and more economical) and was used, e.g., in John & Schmeyer (2008) and Augustin *et al.* (2011). Furthermore, another possible alternative to define the matrix $\mathbb{D}$ is to use the sum of the convection matrix and the diffusion matrix multiplied by a constant from the interval $(0, 1)$. This approach leads to (4.5) with $\tilde{\varepsilon} \in ((bh/2) - \varepsilon, bh/2)$, i.e., a method that can be viewed as intermediate with respect to the two upwinding strategies expressed by (4.7) and (4.9).

Let us now present two choices of β_i different from (4.2). For simplicity, we shall assume that $u_i = u(x_i)$, $i = 0, \ldots, n+1$. If u is increasing and strictly convex in $[0, 1]$ or decreasing and strictly concave in $[0, 1]$, then definition (4.2) gives $\beta_i = 1$, $i = 1, \ldots, n$. Thus, artificial diffusion may be added in regions where it is not needed at all, i.e., where no layer occurs. A partial remedy is to set $\beta_i = 1$ only at nodes where the increase or decrease of u sufficiently accelerates. For example, one can set

$$\beta_i = \begin{cases} 1 & \text{if } u_i \neq u_{i+1} \text{ and } \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < L, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n, \tag{4.10}$$

with a constant $L \in (0, 1)$, e.g., $L = 0.5$.

Unfortunately, the relation (4.10) does not prevent the method from adding artificial diffusion in regions where the solution is nearly constant with respect to its global behaviour. For example, for $u(x) = 1 + x^5$ and any $n > 5$, the definition (4.10) with $L = 0.5$ leads to $\beta_1 = \cdots = \beta_5 = 1$ and $\beta_i = 0$ for $i > 5$, i.e., artificial diffusion is added on the interval $[0, x_5]$. However, $u(x) \in [1, 1.001]$ and $u'(x) \in [0, 0.02]$ for $x \in [0, 0.25]$, whereas $u(x) \in [1, 2]$ and $u'(x) \in [0, 5]$ for $x \in [0, 1]$, so that u can be regarded as nearly constant in $[0, 0.25]$. Hence artificial diffusion is not needed at nodes near to 0. This suggests replacing (4.10) by

$$\beta_i = \begin{cases} 1 & \text{if } (u_i - u_{i-1})(u_{i+1} - u_i) < 0, \\ & \text{or } \dfrac{|u_{i+1} - u_i|}{h} > D \text{ and } \dfrac{u_i - u_{i-1}}{u_{i+1} - u_i} < L, \\ 0 & \text{otherwise,} \end{cases} \qquad i = 1, \ldots, n, \tag{4.11}$$

with some suitable threshold D, e.g.,

$$D = \kappa \frac{\Delta u}{\Delta x}, \quad \kappa = 0.5, \tag{4.12}$$

where Δx is a characteristic length scale and Δu a corresponding characteristic variation of u. For the above example of u, one gets $D = 0.5$ and $\beta_i = 0$, $i = 1, \ldots, n$, if $h \leqslant 0.1$. Note that if (4.10) leads to $\beta_i = 0$, then so does (4.11) and if the values of β_i provided by (4.10) and (4.11) differ, then $|u_i - u_{i-1}|/h < DL$.

As another example, let us consider the function $u(x) = e^{-(1-x)b/\varepsilon}$, $x \in [0, 1]$ (cf. (3.2)), which possesses a boundary layer at the point 1 for large values of b/ε. For any $i \in \{1, \ldots, n\}$, one obtains

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = e^{-2\,\mathrm{Pe}}, \quad \frac{u_{i+1} - u_i}{h} = u(x_i)\frac{e^{2\,\mathrm{Pe}} - 1}{h},$$

so that (4.2) gives $\beta_1 = \cdots = \beta_n = 1$. Definition (4.10) gives either the same result or $\beta_1 = \cdots = \beta_n = 0$ if $L \leqslant e^{-2\,\mathrm{Pe}}$. However, using (4.11) with $L = 0.5$ and $D = 0.5$, one always has $\beta_n = 1$ and possibly $\beta_i = 1$ at some further nodes near to 1, depending on ε, b and h. At the remaining nodes, $\beta_i = 0$. In particular, for $n \geqslant 4$, one obtains $\beta_i = 0$ for $i \leqslant (n + 1)/2$. Thus, artificial diffusion is added only near the layer region, as desired.

5. Discrete maximum principle

From the last point in Remark 4.1, one can see that it makes sense to consider (4.5) with any

$$\tilde{\varepsilon} \in \left[\frac{bh}{2} - \varepsilon, \frac{bh}{2}\right]. \tag{5.1}$$

In this section we prove that then the method satisfies the discrete maximum principle and we formulate various consequences of this fact.

THEOREM 5.1 Consider any $\tilde{\varepsilon} \geqslant (bh/2) - \varepsilon$. Then any solution of the nonlinear problem (4.4–4.6) satisfies the discrete maximum principle, i.e., for any $i \in \{1, \ldots, n\}$, one has

$$g_i \leqslant 0 \;\Rightarrow\; u_i \leqslant \max\{u_{i-1}, u_{i+1}\}, \tag{5.2}$$

$$g_i \geqslant 0 \;\Rightarrow\; u_i \geqslant \min\{u_{i-1}, u_{i+1}\}. \tag{5.3}$$

Moreover, for any $k, l \in \{0, 1, \ldots, n + 1\}$ with $k + 1 < l$, one has

$$g_i \leqslant 0,\ i = k + 1, \ldots, l - 1 \;\Rightarrow\; u_i \leqslant \max\{u_k, u_l\},\ i = k, \ldots, l, \tag{5.4}$$

$$g_i \geqslant 0,\ i = k + 1, \ldots, l - 1 \;\Rightarrow\; u_i \geqslant \min\{u_k, u_l\},\ i = k, \ldots, l. \tag{5.5}$$

Proof. Let the values $u_0, u_1, \ldots, u_{n+1}$ satisfy (4.4–4.6). Consider any $i \in \{1, \ldots, n\}$ and let $g_i \leqslant 0$. If $u_i > \max\{u_{i-1}, u_{i+1}\}$, then $\beta_i = 1$ and hence

$$\begin{aligned} 0 \geqslant g_i h^2 &= -\left(\varepsilon + \tilde{\varepsilon} + \frac{bh}{2}\right) u_{i-1} + 2(\varepsilon + \tilde{\varepsilon}) u_i - \left(\varepsilon + \tilde{\varepsilon} - \frac{bh}{2}\right) u_{i+1} \\ &> -\left(\varepsilon + \tilde{\varepsilon} + \frac{bh}{2}\right) u_i + 2(\varepsilon + \tilde{\varepsilon}) u_i - \left(\varepsilon + \tilde{\varepsilon} - \frac{bh}{2}\right) u_i = 0, \end{aligned}$$

which is a contradiction. Therefore, $u_i \leqslant \max\{u_{i-1}, u_{i+1}\}$.

Now consider any $k, l \in \{0, 1, \ldots, n + 1\}$ with $k + 1 < l$ and let $g_i \leqslant 0$ for $i = k + 1, \ldots, l - 1$. Let $j \in \{k, \ldots, l\}$ be such that $u_j \geqslant u_i$ for $i = k, \ldots, l$. If $j \in \{k, l\}$, then the right-hand side of implication (5.4)

holds. Thus, let $k < j < l$. If $u_j > u_{j+1}$, then $u_{j-1} = u_j$ in view of (5.2). If $u_j = u_{j+1}$, then it follows from (4.5) that

$$0 \geqslant g_j = \left(\frac{\varepsilon + \beta_j \tilde{\varepsilon}}{h^2} + \frac{b}{2h} \right) (u_j - u_{j-1}) \geqslant 0$$

and hence again $u_{j-1} = u_j$. Repeating the above argument, one deduces that $u_j = u_{j-1} = \cdots = u_k$ so that the right-hand side of (5.4) is satisfied.

Implications (5.3) and (5.5) follow analogously. □

COROLLARY 5.2 Consider any $\tilde{\varepsilon} \geqslant (bh/2) - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (4.4–4.6) with $g_i \geqslant 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geqslant u_i$, $i = 0, \ldots, n+1$. Then the solution increases monotonically until u_j and, after that, it decreases monotonically, i.e.,

$$u_0 \leqslant u_1 \leqslant \cdots \leqslant u_j, \quad u_j \geqslant u_{j+1} \geqslant \cdots \geqslant u_{n+1}. \tag{5.6}$$

If $g_i = 0$, $i = 1, \ldots, n$, then the solution is monotone, i.e.,

$$u_0 \leqslant u_1 \leqslant \cdots \leqslant u_{n+1} \quad \text{or} \quad u_0 \geqslant u_1 \geqslant \cdots \geqslant u_{n+1}. \tag{5.7}$$

Proof. If $0 < i < j$, then $u_i \geqslant \min\{u_j, u_{i-1}\} = u_{i-1}$. If $j < i < n+1$, then $u_i \geqslant \min\{u_j, u_{i+1}\} = u_{i+1}$. Therefore, (5.6) holds. If $g_i = 0$, $i = 1, \ldots, n$, then $u_j = \max\{u_0, u_{n+1}\}$ according to (5.4) so that (5.7) follows from (5.6). □

COROLLARY 5.3 Consider any $\tilde{\varepsilon} > (bh/2) - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (4.4–4.6) with $g_i \geqslant 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geqslant u_i$, $i = 0, \ldots, n+1$. If $j < n$, $i \in \{j+1, \ldots, n\}$, and $g_i > 0$, then $u_i > u_{i+1}$ and $\beta_i = 1$. If $g_i = 0$ for some $i \in \{1, \ldots, n\}$, then either $u_{i-1} = u_i = u_{i+1}$ or

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} < 1.$$

Finally, if $u_L > u_R$, one obtains

$$g_i = 0, \ i = 1, \ldots, n \ \Rightarrow \ u_0 > u_1 > \cdots > u_{n+1}, \quad \beta_1 = \beta_2 = \cdots = \beta_n = 1.$$

Proof. According to (4.5), one has

$$\left(\varepsilon + \beta_i \tilde{\varepsilon} + \frac{bh}{2} \right) (u_i - u_{i-1}) + \left(\varepsilon + \beta_i \tilde{\varepsilon} - \frac{bh}{2} \right) (u_i - u_{i+1}) = g_i h^2 \tag{5.8}$$

for $i = 1, \ldots, n$. If $i > j$, then $u_{i-1} \geqslant u_i \geqslant u_{i+1}$ due to (5.6) and hence the first term on the left-hand side of (5.8) is nonpositive. Therefore, (5.8) can be satisfied with $g_i > 0$ only if the second term on the left-hand side of (5.8) is positive, which implies that $u_i > u_{i+1}$ and $\beta_i = 1$. Furthermore, for any $i \in \{1, \ldots, n\}$ such that $g_i = 0$ and $u_i \neq u_{i+1}$, one deduces from (5.8) that

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = \frac{\varepsilon + \beta_i \tilde{\varepsilon} - bh/2}{\varepsilon + \beta_i \tilde{\varepsilon} + bh/2} < 1. \tag{5.9}$$

If $g_i = 0$ and $u_i = u_{i+1}$, then obviously also $u_i = u_{i-1}$.

Finally, let $g_i = 0$, $i = 1, \ldots, n$. If $u_k = u_{k+1}$ for some $k \in \{0, \ldots, n\}$, then according to (5.8) with $i = k$ and $i = k+1$, one obtains $u_k = u_{k-1}$ (if $k > 0$) and $u_{k+1} = u_{k+2}$ (if $k < n$). Thus, one deduces that

$u_0 = u_1 = \cdots = u_{n+1}$. Therefore, if $u_L > u_R$, one gets $u_i \neq u_{i+1}$ for $i = 0, \ldots, n$ and hence (5.7) implies that $u_0 > u_1 > \cdots > u_{n+1}$. Consequently, for any $i \in \{1, \ldots, n\}$, the left-hand side of (5.9) is positive and therefore $\beta_i = 1$. □

COROLLARY 5.4 Let $\tilde{\varepsilon} = (bh/2) - \varepsilon$. Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (4.4–4.6) with $g_i \geqslant 0$, $i = 1, \ldots, n$. Let $j \in \{0, \ldots, n+1\}$ satisfy $u_j \geqslant u_i$, $i = 0, \ldots, n+1$. Then either $j \geqslant n$ or $g_{j+1} = \cdots = g_n = 0$ and $u_j = u_{j+1} = \cdots = u_n$.

If $i \in \{1, \ldots, n\}$ and $g_i = 0$, then either $u_{i-1} = u_i = u_{i+1}$ or $u_i = u_{i-1}$ and $\beta_i = 1$. Consequently,

$$g_i = 0,\ i = 1, \ldots, n \ \Rightarrow\ u_i = u_L,\ i = 1, \ldots, n.$$

Proof. Let $j < n$ and $i \in \{j+1, \ldots, n\}$. Then the left-hand side of (5.8) is nonpositive due to (5.6) and hence (5.8) cannot hold with $g_i > 0$. Therefore, $g_{j+1} = \cdots = g_n = 0$. If $g_i = 0$ for some $i \neq j$, then it follows from (5.6) and (5.8) that $u_i = u_{i-1}$, which completes the proof of the first statement of the corollary. If $j \in \{1, \ldots, n\}$ and $g_j = 0$, then $u_j = u_{j-1}$ since otherwise $u_j > u_{j-1}$ and, in view of (5.8), $u_j > u_{j+1}$ and $\beta_j = 0$, which is in contradiction with (4.6). Thus, for any $i \in \{1, \ldots, n\}$ such that $g_i = 0$, one has $u_i = u_{i-1}$ and it follows from (5.8) that $u_i = u_{i+1}$ or $\beta_i = 1$. □

REMARK 5.5 Let $u_L > u_R$ and $g_i = 0$ for $i = 1, \ldots, n$. It follows from Corollaries 5.3 and 5.4 that if a solution of the nonlinear problem (4.4–4.6) exists, then it is determined uniquely. It is the solution of (3.4) with ε replaced by $\varepsilon + \tilde{\varepsilon}$. Thus, the nonlinear problem is solvable if this solution leads to $\beta_1 = \cdots = \beta_n = 1$ in the case $\tilde{\varepsilon} > (bh/2) - \varepsilon$, and to $\beta_n = 1$ in the case $\tilde{\varepsilon} = (bh/2) - \varepsilon$. If $\tilde{\varepsilon} = (bh/2) - \varepsilon$, this means that $\beta_i = 1$ for $u_{i-1} = u_i \neq u_{i+1}$. This is the case for (4.2) and (4.10) but not necessarily for (4.11). If $\tilde{\varepsilon} > (bh/2) - \varepsilon$, the solution is given by (3.5) with $g = 0$ and Pe replaced by

$$\mathrm{Pe}^* = \frac{bh}{2(\varepsilon + \tilde{\varepsilon})}.$$

Then, for any $i \in \{1, \ldots, n\}$,

$$\frac{u_i - u_{i-1}}{u_{i+1} - u_i} = \frac{1 - \mathrm{Pe}^*}{1 + \mathrm{Pe}^*} < \frac{1}{3} \quad \text{for } \tilde{\varepsilon} \in \left(\frac{bh}{2} - \varepsilon, \frac{bh}{2} \right].$$

Thus, the nonlinear problem is solvable if β_i is defined by (4.2) or by (4.10) with $L \in [\frac{1}{3}, 1)$. On the other hand, if β_i satisfies (4.6) and $\beta_i = 0$ for $(u_i - u_{i-1})(u_{i+1} - u_i) \geqslant 0$, then the nonlinear problem is not solvable for any data. Unfortunately, the favourable choice (4.11) does not lead to a solvable nonlinear problem in general either. We shall return to this choice in Section 9, where it will be used for deriving a convenient definition of β_i.

6. The solution of the nonlinear system and the choice of $\tilde{\varepsilon}$

In this section we report some numerical results obtained by solving the nonlinear problem (4.4), (4.5). We start by briefly describing the solution algorithm. Problem (4.4), (4.5) was solved by a fixed-point iteration: one chooses an initial guess $\underline{\boldsymbol{u}}^0$ for the solution $\underline{\boldsymbol{u}} := \{u_i\}_{i=0}^{n+1}$ and computes a sequence $\{\underline{\boldsymbol{u}}^k\}$ where each $\underline{\boldsymbol{u}}^k$ with $k = 1, 2, \ldots$ solves the linearized problem (4.4), (4.5) with β_i determined by means of the already known discrete solution $\underline{\boldsymbol{u}}^{k-1}$. In our case, the initial guess $\underline{\boldsymbol{u}}^0$ was computed as the solution of (4.4), (4.5) with $\beta_i = 1$, $i = 1, \ldots, n$. We shall prove in Section 7 that the linear problems

defining this fixed-point algorithm are well posed. The iteration was stopped if the coefficients β_i did not change.

Since this section focuses on the choice of $\tilde{\varepsilon}$, we shall present results obtained for β_i defined by (4.2) only. To suppress the influence of the rounding errors on the validity of the conditions in (4.2) for setting $\beta_i = 1$, we replaced (4.2) by

$$\beta_i = \begin{cases} 1 & \text{if } u_i + \tau < u_{i+1} \text{ and } 2u_i + \tau < u_{i-1} + u_{i+1} \\ & \text{or } u_i - \tau > u_{i+1} \text{ and } 2u_i - \tau > u_{i-1} + u_{i+1}, \\ 0 & \text{otherwise,} \end{cases} \tag{6.1}$$

with a suitable positive constant τ. In the computations presented in this section, we used $\tau = 10^{-12}$. For $\tau = 0$, the relations (4.2) and (6.1) are equivalent.

As we pointed out in the previous section, any $\tilde{\varepsilon}$ satisfying (5.1) can be used in (4.5). Then, a natural question is which choice of $\tilde{\varepsilon}$ is most convenient. It is well known that if all the coefficients β_i in (4.5) are set to 1, then

$$\tilde{\varepsilon} = \frac{bh}{2}\left(\coth \mathrm{Pe} - \frac{1}{\mathrm{Pe}}\right) \tag{6.2}$$

is optimal in the sense that, for constant g, the discrete solution is nodally exact, i.e., $u_i = u(x_i)$ for $i = 1, \ldots, n$; see Christie *et al.* (1976). On the other hand, in general the parameter $\tilde{\varepsilon}$ cannot be chosen in such a way that the discrete solution is nodally exact if the coefficients β_i are defined by (4.2). However, it is well known that the performance of most stabilized methods is primarily affected by the amount of artificial diffusion introduced near the numerical layers, and quite insensitive to the changes on it far away from them. Thus, since we expect that $\beta_i = 1$ in numerical boundary layers, it may be of advantage to use $\tilde{\varepsilon}$ given by (6.2) also when the coefficients β_i are defined by (4.2). Then what is required is that the exact solution solves scheme (4.5) for the nodes x_i where $\beta_i = 1$. Note that the parameter $\tilde{\varepsilon}$ defined in (6.2) is larger than $\tilde{\varepsilon}$ from (4.3) and smaller than $\tilde{\varepsilon}$ from (4.8).

In what follows, we shall compare solutions of the problem given by (4.4), (4.5), (6.1) for $\tilde{\varepsilon}$ defined by (4.3), (4.8) and (6.2). We shall consider

$$b = g = 1, \quad u_{\mathrm{L}} = u_{\mathrm{R}} = 0, \tag{6.3}$$

and various choices of ε and n.

First, we notice that if $\tilde{\varepsilon}$ is defined by (4.3), it is easy to verify that, for the data (6.3) and any ε and n,

$$u_i = ih, \quad i = 0, \ldots, n, \quad u_{n+1} = 0$$

is a solution of (4.4), (4.5) with β_i given by (6.1) or any β_i satisfying (4.6) (it is the only solution of the respective nonlinear problem). In this case, $\beta_n = 1$ and if β_i is defined by (6.1) or (4.2), one has $\beta_i = 0$ for $i = 1, \ldots, n-1$. Since the discrete solution is independent of ε, one cannot expect a good approximation of the exact solution for the whole range of values of ε. Indeed, according to (3.2), the error in the discrete solution satisfies

$$u_i - u(x_i) = \frac{e^{-(1-x_i)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}, \quad i = 0, \ldots, n, \tag{6.4}$$

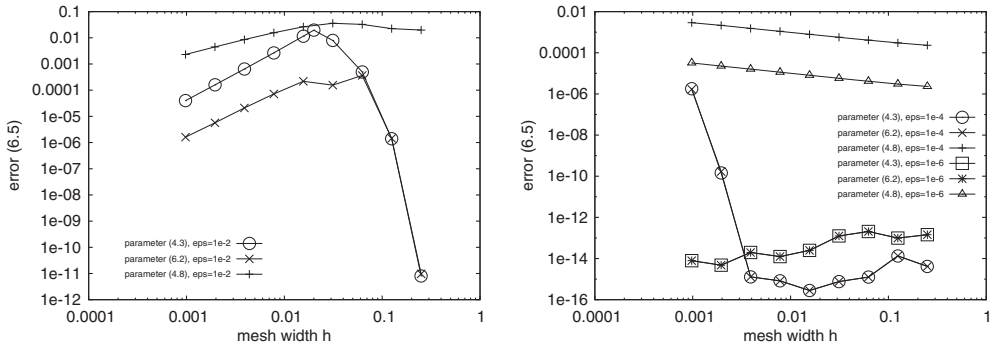


FIG. 1. Dependence of the errors of the solutions of (4.4), (4.5), (6.1) on h for $\tilde{\varepsilon}$ defined by (4.3), (6.2) and (4.8), and for $\varepsilon = 10^{-2}$ (left) and $\varepsilon \in \{10^{-4}, 10^{-6}\}$ (right).

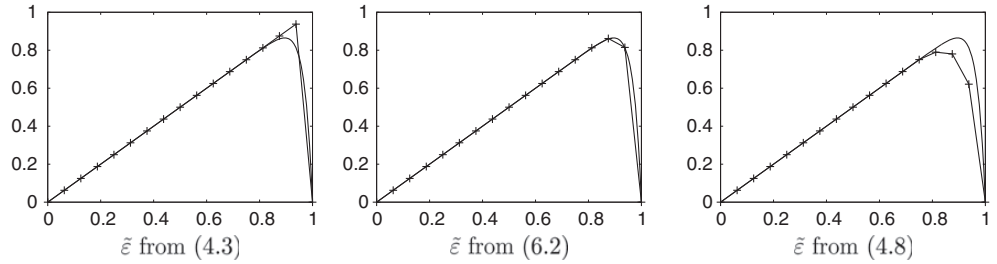


FIG. 2. Comparisons of the exact solution and solutions of (4.4), (4.5), (6.1) for $\varepsilon = 0.03$, $n = 15$, and $\tilde{\varepsilon}$ defined by (4.3), (6.2) and (4.8).

so that the largest error appears at node x_n and, for $\varepsilon \leqslant 0.1$, one has

$$u_n - u(x_n) = \frac{e^{-h/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}} > 0.135 \quad \text{for Pe} \to 1.$$

To see the impact of the nonlinear artificial diffusion in (4.5) on the discrete solutions, we computed the errors

$$\left(\frac{1}{n} \sum_{i=1}^{n} (u(x_i) - u_i)^2 \right)^{1/2} \tag{6.5}$$

for different values of h, the three definitions of $\tilde{\varepsilon}$ (cf. (4.3), (6.2), and (4.8)), and for $\varepsilon \in \{10^{-2}, 10^{-4}, 10^{-6}\}$. If $\tilde{\varepsilon}$ is defined by (4.3), we set $\tilde{\varepsilon} = 0$ for $\text{Pe} \leqslant 1$ (this situation occurs only for $\varepsilon = 10^{-2}$). The results are depicted in Fig. 1, where we observe that the best results are obtained for $\tilde{\varepsilon}$ defined by (6.2). For large Péclet numbers, comparable errors are also obtained for $\tilde{\varepsilon}$ defined by (4.3). The choice (4.8) always adds too much artificial diffusion and leads to the worst results. To further stress this, Fig. 2 depicts the discrete solutions corresponding to $\text{Pe} = \frac{25}{24}$ and clearly demonstrates the differences between the three choices of $\tilde{\varepsilon}$.

One final comment is required for the case where $\tilde{\varepsilon}$ is given by (4.3). In this case, according to (6.4), the error (6.5) is bounded by $e^{-h/\varepsilon}$. This shows that, for $\varepsilon = 10^{-6}$ (and partly also for $\varepsilon = 10^{-4}$), the errors depicted in Fig. 1 are the results of rounding errors and are much larger than the actual values of the errors.

7. Solvability of the linear subproblems

At the beginning of the previous section, the solution of the nonlinear problem (4.4), (4.5) using a fixed-point iteration was described. In this section, we shall discuss under which conditions the corresponding linear subproblems are uniquely solvable.

We shall consider the following more general problem: given positive numbers $d_1, \ldots, d_n$, find $u_1, \ldots, u_n$ such that

$$-d_i(u_{i-1} - 2u_i + u_{i+1}) + u_{i+1} - u_{i-1} = \tilde{g}_i, \quad i = 1, \ldots, n, \tag{7.1}$$

where $u_0 = u_{\mathrm{L}}$ and $u_{n+1} = u_{\mathrm{R}}$. This problem corresponds to (4.5) for $d_i = 2(\varepsilon + \beta_i\tilde{\varepsilon})/(bh)$ and $\tilde{g}_i = 2hg_i/b$.

The following theorem proves the unique solvability of problem (7.1) in the case that the coefficients d_i are allowed to take the values 1 and d with $d > 0$. As a consequence, the unique solvability of the linearized problem (4.5) with $\tilde{\varepsilon}$ given by (4.3) follows.

THEOREM 7.1 Let $d_1, \ldots, d_n \in \{1, d\}$ with an arbitrary $d > 0$. Then problem (7.1) has a unique solution.

Proof. It suffices to show that the homogeneous problem corresponding to (7.1) has only the trivial solution, i.e., that if

$$-d_i(u_{i-1} - 2u_i + u_{i+1}) + u_{i+1} - u_{i-1} = 0, \quad i = 1, \ldots, n, \tag{7.2}$$

with $u_0 = u_{n+1} = 0$, then

$$u_1 = u_2 = \cdots = u_n = 0. \tag{7.3}$$

Let $1 \leqslant K \leqslant L \leqslant n$ and $d_K = d_{K+1} = \cdots = d_L = d$. Multiplying the ith equation in (7.2) by u_i and summing over $i = K, \ldots, L$, one obtains

$$du_K^2 + d\sum_{i=K}^{L-1}(u_i - u_{i+1})^2 + du_L^2 - (1+d)u_{K-1}u_K + (1-d)u_Lu_{L+1} = 0. \tag{7.4}$$

Thus, if $d_1 = d_2 = \cdots = d_n = d$, one may set $K = 1$ and $L = n$, and (7.4) readily implies (7.3). Of course, this result also follows from the equivalence between (3.3) and (3.4) and the fact that (3.3) is uniquely solvable.

It remains to investigate the case when the values of d_i are not all equal. Let $K \in \{1, \ldots, n\}$ be the smallest index such that $d_K = d$ and let $L \in \{K, \ldots, n\}$ be the largest index such that $d_K = d_{K+1} = \cdots = d_L = d$. Then, for any $i \in \{1, \ldots, K-1\}$, one has $d_i = 1$ and hence $u_i = u_{i-1}$. Consequently, $u_i = 0$ for $i = 0, \ldots, K-1$. Furthermore, if $L < n$, then $d_{L+1} = 1$ and hence $u_{L+1} = u_L$, which implies that $du_L^2 + (1-d)u_Lu_{L+1} \geqslant 0$. This inequality is satisfied also if $L = n$ since then $u_{L+1} = 0$. Thus, one deduces from (7.4) that

$$du_K^2 + d\sum_{i=K}^{L-1}(u_i - u_{i+1})^2 \leqslant 0,$$

which gives $0 = u_K = u_{K+1} = \cdots = u_L$. Repeating the above arguments until $L = n$, one obtains (7.3). □

The following theorem proves the unique solvability of (7.1) for a more general choice of $d_1, \ldots, d_n$.

THEOREM 7.2 Let $d_1, \ldots, d_n \in (0, 1]$ or $d_1, \ldots, d_n \in [\delta, 1 + \delta]$ with $\delta \in (0, 1]$. Then problem (7.1) has a unique solution. However, for any $\delta > 0$, there are $d_1, \ldots, d_n \in (0, 1 + \delta]$ such that problem (7.1) is not uniquely solvable.

Proof. We introduce the $n \times n$ matrices

$$\mathbb{B} = \operatorname{diag}(d_1, d_2, \ldots, d_n), \quad \mathbb{C} = \operatorname{tridiag}(-1, 2, -1), \quad \mathbb{E} = \operatorname{tridiag}(-1, 0, 1).$$

Then the matrix corresponding to (7.1) is $\mathbb{B}\mathbb{C} + \mathbb{E}$. This matrix will be transformed by operations which preserve full rank such that it becomes possible to see that its determinant does not vanish.

Let $\mathbb{G} = (g_{ij})_{i,j=1}^n$ be a symmetric matrix given by

$$g_{ij} = (n - i + 1)j, \quad j = 1, \ldots, i, \quad i = 1, \ldots, n.$$

Then $\mathbb{C}\mathbb{G} = (n + 1)\mathbb{I}$, where $\mathbb{I}$ is the identity matrix. Setting $\mathbb{Q} = (\mathbb{B}\mathbb{C} + \mathbb{E})\mathbb{G}$, one obtains a matrix with the entries

$$\begin{aligned} q_{ij} &= -2j + 2(n + 1), \quad \text{for } i = 1, \ldots, j - 1, \\ q_{jj} &= -2j + (n + 1)(1 + d_j), \\ q_{ij} &= -2j, \quad \text{for } i = j + 1, \ldots, n, \end{aligned}$$

where $j = 1, \ldots, n$. Now, let us define the matrix $\mathbb{Z} = (z_{ij})_{i,j=1}^n$ by

$$z_{ij} = \frac{1}{n+1}(q_{ij} - q_{i+1,j}), \quad i = 1, \ldots, n - 1, \quad z_{nj} = \frac{1}{n+1}\left(2q_{nj} + \sum_{i=1}^{n-1} q_{ij}\right),$$

where $j = 1, \ldots, n$. Then $\det(\mathbb{B}\mathbb{C} + \mathbb{E}) \neq 0$ if and only if $\det \mathbb{Z} \neq 0$ and one has

$$\begin{aligned} &z_{ii} = 1 + d_i, \quad z_{i,i+1} = 1 - d_{i+1}, \quad z_{ij} = 0 \text{ for } j \notin \{i, i + 1\}, \quad i = 1, \ldots, n - 1, \\ &z_{nj} = -1 + d_j, \quad j = 1, \ldots, n - 1, \quad z_{nn} = 2d_n. \end{aligned}$$

Let $\mathbb{Z}^{ij}$ be the $(n - 1) \times (n - 1)$ matrix obtained from $\mathbb{Z}$ by removing the ith row and jth column. Then

$$\det \mathbb{Z}^{nj} = \prod_{k=1}^{j-1}(1 + d_k) \prod_{l=j+1}^{n}(1 - d_l). \tag{7.5}$$

Let n be odd and denote

$$\tilde{z}_{nj} = \sum_{\substack{i=1 \\ i \text{ is odd}}}^{n} z_{ij}, \quad j = 1, \ldots, n.$$

Then, for $j = 1, \ldots, n$, one has

$$\tilde{z}_{nj} = 2d_j \ \text{ if } j \text{ is odd}, \quad \tilde{z}_{nj} = 0 \ \text{ if } j \text{ is even}.$$

Thus,

$$\det \mathbb{Z} = 2 \sum_{\substack{j=1 \\ j \text{ is odd}}}^{n} d_j \det \mathbb{Z}^{nj}. \tag{7.6}$$

If $d_1, \ldots, d_n \in (0, 1]$, then $\det \mathbb{Z}^{nj} \geqslant 0$, $j = 1, \ldots, n-1$, and $\det \mathbb{Z}^{nn} > 0$ so that $\det \mathbb{Z} > 0$. If n is even, then

$$\det \mathbb{Z} = (1 + d_1) \det \mathbb{Z}^{11} + (1 - d_1) \det \mathbb{Z}^{n1}. \tag{7.7}$$

Since $\mathbb{Z}^{11}$ has the same structure as $\mathbb{Z}$ and has an odd number of rows and columns, one has $\det \mathbb{Z}^{11} > 0$ for $d_1, \ldots, d_n \in (0, 1]$. Moreover, $\det \mathbb{Z}^{n1} \geqslant 0$ in view of (7.5) and hence again $\det \mathbb{Z} > 0$, which proves the first part of the theorem.

Now let $d_1, \ldots, d_n \in [\delta, 1 + \delta]$ with $\delta \in (0, 1]$. We denote

$$A_s = \prod_{k=1}^{s} (1 + d_k), \quad B_s = \prod_{l=s}^{n} (1 - d_l), \quad s = 1, \ldots, n,$$

and we set $A_0 = 1$. If $B_s < 0$, then, for some $k \in \{1, \ldots, n\}$, we have $|1 - d_k| \leqslant \delta$. Therefore, since $|1 - d_l| \leqslant 1$ for any $l \in \{1, \ldots, n\}$, one gets

$$B_s \geqslant -\delta, \quad s = 1, \ldots, n. \tag{7.8}$$

First, let n be odd and let us prove that, for any odd $m \in \{1, \ldots, n\}$, the matrices $\mathbb{Z}^{nj}$ satisfy

$$\sum_{\substack{j=m \\ j \text{ is odd}}}^{n} d_j \det \mathbb{Z}^{nj} \geqslant d_n A_{m-1}. \tag{7.9}$$

In view of (7.5), this inequality holds for $m = n$. Let us assume that (7.9) holds for a given odd $m \in \{3, \ldots, n\}$. Then, again in view of (7.5),

$$\begin{aligned} \sum_{\substack{j=m-2 \\ j \text{ is odd}}}^{n} d_j \det \mathbb{Z}^{nj} &\geqslant d_n A_{m-1} + d_{m-2} A_{m-3} B_{m-1} \\ &> d_n A_{m-3} + d_{m-2} A_{m-3} [d_n (1 + d_{m-1}) + B_{m-1}] > d_n A_{m-3} \end{aligned}$$

since $d_n(1 + d_{m-1}) > \delta$ and $B_{m-1} \geqslant -\delta$; see (7.8). Thus, (7.9) holds for any odd $m \in \{1, \ldots, n\}$ and hence, setting $m = 1$ and using (7.6), one gets $\det \mathbb{Z} \geqslant 2d_n$. If n is even, then $\det \mathbb{Z}^{11} \geqslant 2d_n$ and hence, according to (7.7), $\det \mathbb{Z} = (1 + d_1) \det \mathbb{Z}^{11} + B_1 > 2d_n + B_1 \geqslant d_n$.

Finally, let us consider any $\delta > 0$ and set

$$d_1 = d_2 = \cdots = d_{n-2} = 1, \quad d_{n-1} = 1 + \delta, \quad d_n = \frac{\delta}{3\delta + 4}.$$

Then $d_1, \ldots, d_n \in (0, 1 + \delta]$ and

$$\det \mathbb{Z} = 2^{n-2} \det \begin{pmatrix} 1 + d_{n-1} & 1 - d_n \\ -1 + d_{n-1} & 2d_n \end{pmatrix} = 0.$$

Consequently, the matrix corresponding to (7.1) is singular and hence problem (7.1) is not uniquely solvable. □

The following corollary states the unique solvability of the linearized problem (4.4), (4.5) for any $\tilde{\varepsilon}$ satisfying (5.1).

COROLLARY 7.3 Consider any $\tilde{\varepsilon} \in [0, bh/2]$ and any $\beta_1, \ldots, \beta_n \in [0, 1]$. Then the linear problem (4.4), (4.5) has a unique solution.

Proof. Since (4.5) is equivalent to (7.1) with $d_1, \ldots, d_n \in [1/\mathrm{Pe}, 1 + 1/\mathrm{Pe}]$, the statement follows immediately from Theorem 7.2. □

8. Solvability of the nonlinear problem

The computations reported in Section 6 were those for which convergence of the fixed-point iteration was achieved. However, some other computations we performed did not converge at all. In some cases, convergence was obtained after changing the value of τ in (6.1) (although we realized that the iterative process was still very sensitive to rounding errors). For some other cases though, we were not able to find any way to achieve convergence and hence no solution at all was found. The ultimate conclusion of these numerical experiments was that the nonlinear problem (4.4–4.6) is not solvable in general. In this section we first describe examples of data for which the nonlinear problem has no solution, thus proving the above claim. This lack of solvability is due to the discontinuous character of the coefficients β_i. As a matter of fact, at the end of the present section we shall prove that problem (4.4), (4.5) is solvable if one considers coefficients β_i depending on the discrete solution in a continuous way.

Let us start with the following remark. If the nonlinear problem (4.4), (4.5) with some functions β_i satisfying (4.6) has a solution, then there are numbers $\bar{\beta}_1, \ldots, \bar{\beta}_n \in \{0, 1\}$ such that, after having computed the solution $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ of (4.4), (4.5) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, n$, one has $\beta_i(\underline{\boldsymbol{u}}) = \bar{\beta}_i$, $i = 1, \ldots, n$. Since there are only 2^n admissible choices of $\bar{\beta}_1, \ldots, \bar{\beta}_n$, one can easily check (at least for small n) whether the nonlinear problem is solvable. In what follows, we shall consider the three choices of $\tilde{\varepsilon}$ tested in Section 6 and, for each of them, we shall present an example of data such that the nonlinear problem (4.4), (4.5) is not solvable for any functions β_i satisfying (4.6) and

$$\beta_i = 0 \quad \text{if } u_i \neq u_{i+1} \quad \text{and} \quad \frac{u_i - u_{i-1}}{u_{i+1} - u_i} > 1. \tag{8.1}$$

These requirements are met by all three choices (4.2), (4.10) and (4.11). In all cases, we shall use

$$n = 4, \quad u_{\mathrm{L}} = u_{\mathrm{R}} = 0. \tag{8.2}$$

First, let us study problem (4.4), (4.5), (4.2) with $\tilde{\varepsilon}$ defined by (4.3). We consider the data

$$\varepsilon = 0.03, \quad b = 1, \quad g_1 = 6, \quad g_2 = -6, \quad g_3 = 3, \quad g_4 = -2. \tag{8.3}$$

As explained above, for each of the 16 possible choices of $\bar{\beta}_1, \ldots, \bar{\beta}_4$, we compute the solution $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{5}$ of (4.4), (4.5) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$. These solutions together with the values of $\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}})$ computed according to (4.2) are shown in Figs 3 and 4. Since $(\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}}))$ always differs from $(\bar{\beta}_1, \ldots, \bar{\beta}_4)$, one concludes that the nonlinear problem (4.4), (4.5), (4.2) does not

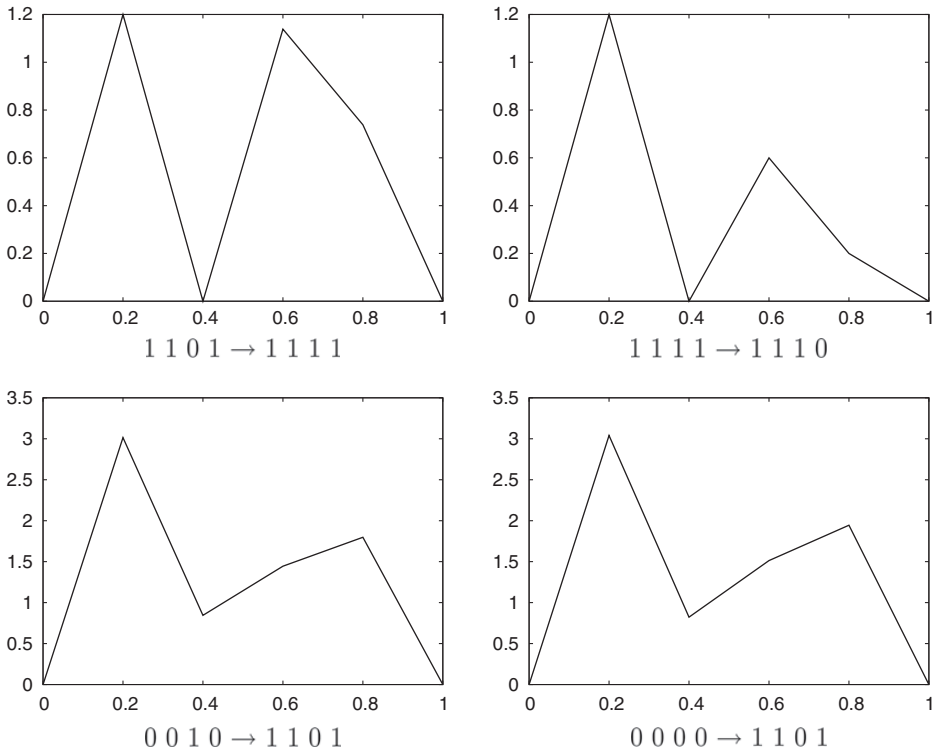


FIG. 3. Solutions $\underline{\boldsymbol{u}}$ of (4.4), (4.5) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$ for the data (8.2), (8.3) and $\tilde{\varepsilon}$ defined by (4.3). The numbers on the left of '→' represent $\bar{\beta}_1, \ldots, \bar{\beta}_4$, the numbers on the right of '→' represent $\beta_1(\underline{\boldsymbol{u}}), \ldots, \beta_4(\underline{\boldsymbol{u}})$ corresponding to the respective solution according to (4.2).

have any solution. Note that, for all choices of $\bar{\beta}_1, \ldots, \bar{\beta}_4$ except $\bar{\beta}_1 = \cdots = \bar{\beta}_4 = 1$, there always exists $j \in \{1, 2, 3, 4\}$ such that $\bar{\beta}_j = 0$ and the solution $\underline{\boldsymbol{u}}$ has an extremum at the node x_j so that $\beta_j(\underline{\boldsymbol{u}}) = 1$ as soon as (4.6) holds. If $\bar{\beta}_1 = \cdots = \bar{\beta}_4 = 1$, one observes that $\beta_4(\underline{\boldsymbol{u}}) = 0$ as soon as (8.1) holds. This shows that problem (4.4), (4.5) is not solvable for any functions β_i satisfying (4.6) and (8.1).

Similar nonexistence studies were performed for the case in which $\tilde{\varepsilon}$ is defined by (6.2) and (4.8). For both cases we were able to find various right-hand sides for which the discrete problem does not have a solution. For example, if $\tilde{\varepsilon}$ is defined by (6.2), then the nonlinear problem with any β_i satisfying (4.6) and (8.1) is not solvable for the following data:

$$\varepsilon = 0.09, \quad b = 1, \quad g_1 = 6, \quad g_2 = g_3 = g_4 = 1. \tag{8.4}$$

Finally, if $\tilde{\varepsilon}$ is defined by (4.8), then the nonlinear problem with any β_i satisfying (4.6) and (8.1) is not solvable, e.g., for

$$\varepsilon = 0.064, \quad b = 1, \quad g_1 = g_2 = g_3 = g_4 = 1. \tag{8.5}$$

We have verified that the nonexistence of a solution to the nonlinear problem (4.4), (4.5) in the cases presented in this section is not caused by rounding errors.

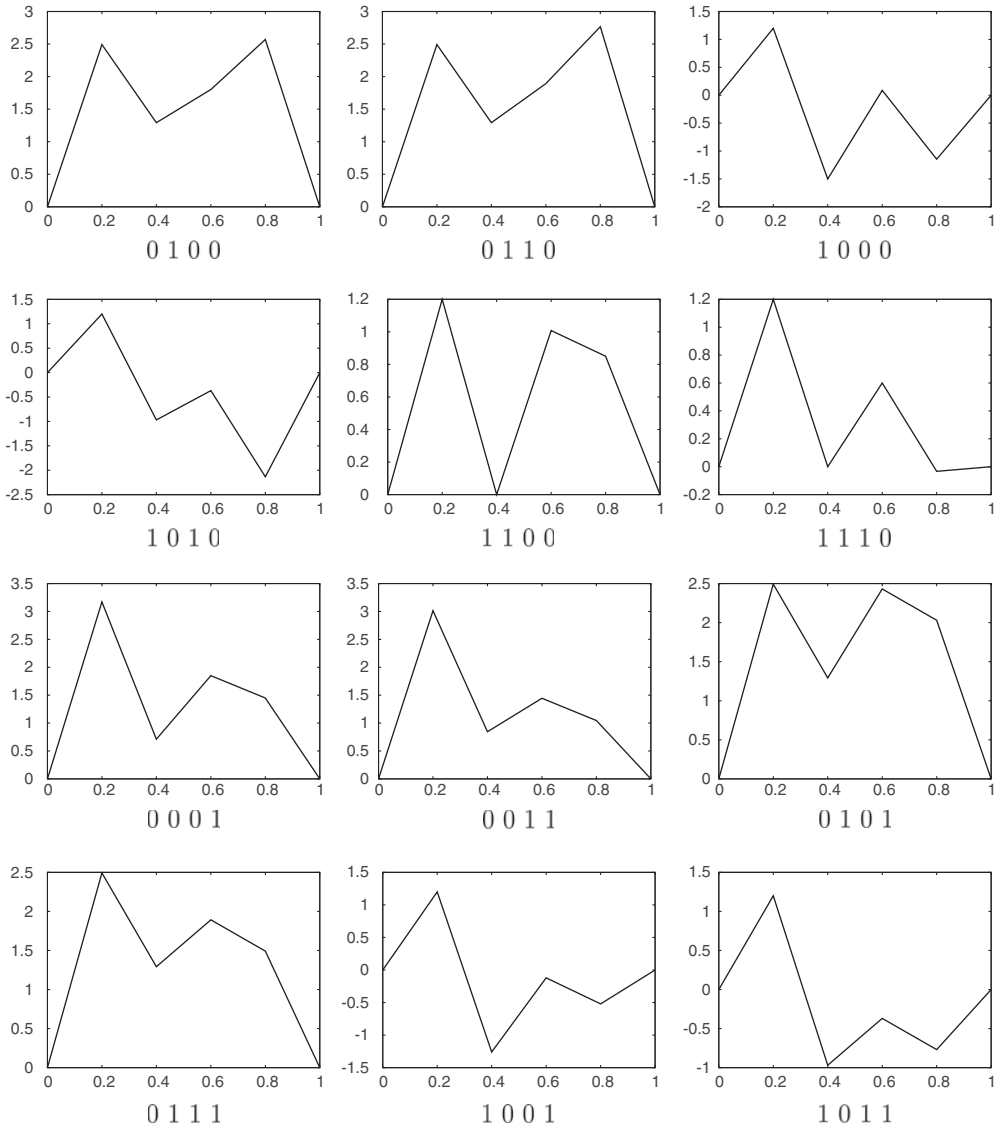


FIG. 4. Solutions $\underline{\boldsymbol{u}}$ of (4.4), (4.5) with $\beta_i = \bar{\beta}_i$, $i = 1, \ldots, 4$ for the data (8.2), (8.3) and $\tilde{\varepsilon}$ defined by (4.3). The numbers below the graphs represent $\bar{\beta}_1, \ldots, \bar{\beta}_4$. For all solutions, formula (4.2) gives $\beta_1(\underline{\boldsymbol{u}}) = \beta_2(\underline{\boldsymbol{u}}) = \beta_3(\underline{\boldsymbol{u}}) = \beta_4(\underline{\boldsymbol{u}}) = 1$.

Now, as we have already stated, we present a result ensuring the solvability of the nonlinear problem (4.4), (4.5) under the hypothesis of continuity of the coefficients β_i.

THEOREM 8.1 Let $\beta_i : \mathbb{R}^{n+2} \to [0, 1]$, $i = 1, \ldots, n$ be continuous functions and let $\tilde{\varepsilon} \in [0, bh/2]$. Then there exists a solution of the nonlinear problem (4.4), (4.5).

Proof. We set $\boldsymbol{\beta}(\underline{\boldsymbol{u}}) := \{\beta_i(\underline{\boldsymbol{u}})\}_{i=1}^{n}$ with $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$. We also denote by $\mathbb{M}(\boldsymbol{\beta}) \in \mathbb{R}^{n\times n}$ the matrix corresponding to system (4.5) for a particular choice of the coefficients $\boldsymbol{\beta} \in \mathbb{R}^n$. Then the nonlinear problem

(4.4), (4.5) can be written as, find $\boldsymbol{u} \equiv \{u_i\}_{i=1}^n$ such that

$$\mathbb{M}(\boldsymbol{\beta}(\underline{\boldsymbol{u}}))\boldsymbol{u} = \tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}), \tag{8.6}$$

where $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ with $u_0 = u_{\mathrm{L}}$, $u_{n+1} = u_{\mathrm{R}}$ and $\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) = \{\tilde{g}_i(\underline{\boldsymbol{u}})\}_{i=1}^n$ with $\tilde{g}_i(\underline{\boldsymbol{u}}) = g_i$ for $i = 2, \ldots, n-1$, and

$$\tilde{g}_1(\underline{\boldsymbol{u}}) = g_1 + (\varepsilon + \beta_1(\underline{\boldsymbol{u}})\tilde{\varepsilon})\frac{u_{\mathrm{L}}}{h^2} + b\frac{u_{\mathrm{L}}}{2h}, \quad \tilde{g}_n(\underline{\boldsymbol{u}}) = g_n + (\varepsilon + \beta_n(\underline{\boldsymbol{u}})\tilde{\varepsilon})\frac{u_{\mathrm{R}}}{h^2} - b\frac{u_{\mathrm{R}}}{2h}.$$

Since $|\beta_i(\underline{\boldsymbol{u}})| \leqslant 1$ for $i = 1, \ldots, n$, one has

$$\|\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}})\| \leqslant \|\boldsymbol{g}\| + \frac{\varepsilon + bh}{h^2}(|u_{\mathrm{L}}| + |u_{\mathrm{R}}|) \quad \forall\, \underline{\boldsymbol{u}} \in \mathbb{R}^{n+2}, \tag{8.7}$$

where $\|\cdot\|$ denotes the Euclidean norm on $\mathbb{R}^n$ and $\boldsymbol{g} = \{g_i\}_{i=1}^n$.

Corollary 7.3 guarantees that the matrix $\mathbb{M}(\boldsymbol{\beta})$ is invertible for all $\boldsymbol{\beta}$ belonging to the hypercube $[0,1]^n$. Then, since the determinant of a matrix is a continuous function of its entries, there exists $\sigma_0 > 0$ such that

$$|\det \mathbb{M}(\boldsymbol{\beta})| \geqslant \sigma_0 \quad \forall\, \boldsymbol{\beta} \in [0,1]^n.$$

Hence, the function $\boldsymbol{\beta} \mapsto [\mathbb{M}(\boldsymbol{\beta})]^{-1}$ is continuous on $[0,1]^n$, and there exists $C > 0$ such that

$$\|[\mathbb{M}(\boldsymbol{\beta})]^{-1}\| \leqslant C \quad \forall\, \boldsymbol{\beta} \in [0,1]^n, \tag{8.8}$$

where we use the matrix norm induced by the Euclidean norm on $\mathbb{R}^n$. Consequently, there exists a constant $C_0 > 0$ such that

$$\forall\, \boldsymbol{\beta} \in [0,1]^n,\ \boldsymbol{v} \in \mathbb{R}^n,\ \underline{\boldsymbol{u}} \in \mathbb{R}^{n+2}: \quad \mathbb{M}(\boldsymbol{\beta})\boldsymbol{v} = \tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) \ \Rightarrow\ \|\boldsymbol{v}\| \leqslant C_0. \tag{8.9}$$

In view of (8.7) and (8.8), the constant C_0 depends on the data of (3.1) and, possibly, on h, but it does not depend on $\underline{\boldsymbol{u}}$.

Now let $T : \mathbb{R}^n \to \mathbb{R}^n$ be the mapping defined by

$$T\boldsymbol{u} := [\mathbb{M}(\boldsymbol{\beta}(\underline{\boldsymbol{u}}))]^{-1}\tilde{\boldsymbol{g}}(\underline{\boldsymbol{u}}) \quad \forall\, \boldsymbol{u} \equiv \{u_i\}_{i=1}^n \in \mathbb{R}^n,$$

where $\underline{\boldsymbol{u}} = \{u_i\}_{i=0}^{n+1}$ with $u_0 = u_{\mathrm{L}}$ and $u_{n+1} = u_{\mathrm{R}}$. Then T is continuous and, according to (8.9), it maps the closed ball $B(0, C_0) := \{\boldsymbol{v} \in \mathbb{R}^n;\ \|\boldsymbol{v}\| \leqslant C_0\}$ into itself. Applying Brouwer's fixed-point theorem, there exists $\boldsymbol{u} \in B(0, C_0)$ such that $T\boldsymbol{u} = \boldsymbol{u}$, i.e., $\boldsymbol{u}$ satisfies (8.6). □

9. An example of continuous β_i and properties of the resulting solvable nonlinear discrete problem

In this section we propose a definition of continuous coefficients β_i that, according to Theorem 8.1, leads to a solvable nonlinear discrete problem, prove a corresponding (weaker) variant of the discrete maximum principle and present a few numerical results.

For $i = 1, \dots, n$, let us denote the derivatives of the discrete solution to the left and to the right of a point x_i by

$$u'_{i-} = \frac{u_i - u_{i-1}}{h}, \quad u'_{i+} = \frac{u_{i+1} - u_i}{h},$$

respectively. If β_i are defined by (4.2), then

$$\beta_i = \begin{cases} 1 & \text{if } u'_{i+} > \max\{0, u'_{i-}\} \text{ or } u'_{i+} < \min\{0, u'_{i-}\}, \\ 0 & \text{if } \min\{0, u'_{i-}\} \leqslant u'_{i+} \leqslant \max\{0, u'_{i-}\}, \end{cases}$$

for $i = 1, \dots, n$; see Fig. 5. Note that β_i is discontinuous along the lines $u'_{i-} = u'_{i+}$ and $u'_{i+} = 0$. Similarly, β_i is discontinuous if it is defined by (4.10) or (4.11).

Our aim is to introduce continuous coefficients β_i to guarantee the solvability of the nonlinear problem (4.4), (4.5). Based on the relation (4.11) and the discussion at the end of Section 4 and in Remark 5.5, we propose to set (cf. Fig. 6)

$$\beta_i = \begin{cases} 1 & \text{if } (u'_{i+} \geqslant \Delta + \max\{0, 2u'_{i-}\} \text{ or } u'_{i+} \leqslant -\Delta + \min\{0, 2u'_{i-}\}), \\ & \text{and } (u'_{i-}, u'_{i+}) \notin (-\Delta, D/2) \times (0, D + \Delta), \\ & \text{and } (u'_{i-}, u'_{i+}) \notin (-D/2, \Delta) \times (-D - \Delta, 0), \\ 0 & \text{if } \min\{0, 2u'_{i-}\} \leqslant u'_{i+} \leqslant \max\{0, 2u'_{i-}\}, \\ & \text{or } (u'_{i-}, u'_{i+}) \in [0, D/2] \times [0, D], \\ & \text{or } (u'_{i-}, u'_{i+}) \in [-D/2, 0] \times [-D, 0], \end{cases} \tag{9.1}$$

with positive parameters $\Delta \leqslant D$. Furthermore, we require that β_i is continuous and that it is linear in each of the eight dark shadow subregions in Fig. 6. These requirements define the function β_i uniquely. The parameters D and Δ should be proportional to a characteristic derivative $\Delta u / \Delta x$; see (4.12).

Unfortunately, with the new definition of the coefficients β_i, we cannot guarantee the validity of the discrete maximum principle formulated in Theorem 5.1. Nevertheless, the following result shows that

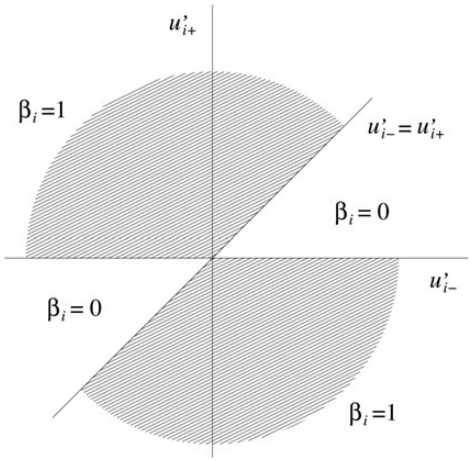


FIG. 5. Dependence of values of β_i from (4.2) on u'_{i-} and u'_{i+}.

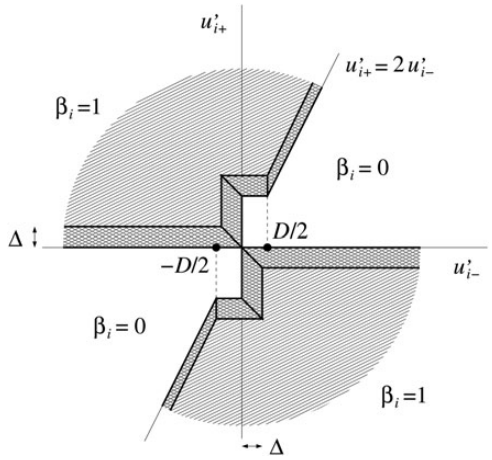


FIG. 6. Definition of continuous β_i according to (9.1).

a possible violation of the discrete maximum principle is not significant if the parameter D or the mesh width h are small. The constant δ in the following theorem is related to the above definition of β_i by $\delta = D + \Delta$.

THEOREM 9.1 Consider any $\tilde{\varepsilon}$ satisfying (5.1). Let $u_0, \ldots, u_{n+1}$ be a solution of the nonlinear problem (4.4), (4.5) with any functions $\beta_1, \ldots, \beta_n \in [0, 1]$ satisfying

$$\beta_i = 1 \quad \text{if} \quad u_i < \min\{u_{i-1}, u_{i+1} - \delta h\} \quad \text{or} \quad u_i > \max\{u_{i-1}, u_{i+1} + \delta h\}$$

for some $\delta > 0$ and $i = 1, \ldots, n$. Then

$$\begin{aligned} g_i \leqslant 0 \;&\Rightarrow\; u_i \leqslant \max\{u_{i-1}, u_{i+1}\} \quad \text{or} \quad u_i \leqslant \min\{u_{i-1}, u_{i+1}\} + \delta h, \\ g_i \geqslant 0 \;&\Rightarrow\; u_i \geqslant \min\{u_{i-1}, u_{i+1}\} \quad \text{or} \quad u_i \geqslant \max\{u_{i-1}, u_{i+1}\} - \delta h, \end{aligned}$$

for $i = 1, \ldots, n$. Moreover, for any $k, l \in \{0, 1, \ldots, n+1\}$ with $k + 1 < l$, one has

$$\begin{aligned} g_i \leqslant 0,\ i = k+1, \ldots, l-1 \;&\Rightarrow\; u_i < \max\{u_k, u_l\} + \delta h, \quad i = k, \ldots, l, \\ g_i \geqslant 0,\ i = k+1, \ldots, l-1 \;&\Rightarrow\; u_i > \min\{u_k, u_l\} - \delta h, \quad i = k, \ldots, l. \end{aligned}$$

Proof. Consider any $i \in \{1, \ldots, n\}$ and let $g_i \leqslant 0$. If $u_i - u_{i+1} \notin [0, \delta h]$, then $u_i \leqslant \max\{u_{i-1}, u_{i+1}\}$ since the proof of Theorem 5.1 can be repeated without any changes. For $u_i - u_{i+1} \in [0, \delta h]$ it will be shown that $u_i \leqslant \min\{u_{i-1}, u_{i+1}\} + \delta h$. To this end, assume that $u_i > \min\{u_{i-1}, u_{i+1}\} + \delta h$. Then

$$u_{i+1} + \delta h \geqslant u_i \geqslant u_{i+1}, \quad u_i > u_{i-1} + \delta h.$$

Therefore, using (5.8) and noting that $(u_i - u_{i+1})$ is estimated either from below or from above depending on the sign of the term in front of it, one derives

$$\begin{aligned} 0 \geqslant g_i h^2 &= \left(\varepsilon + \beta_i\tilde{\varepsilon} + \frac{bh}{2}\right)(u_i - u_{i-1}) + \left(\varepsilon + \beta_i\tilde{\varepsilon} - \frac{bh}{2}\right)(u_i - u_{i+1}) \\ &> \left(\varepsilon + \beta_i\tilde{\varepsilon} + \frac{bh}{2}\right)\delta h + \min\left\{0, \varepsilon + \beta_i\tilde{\varepsilon} - \frac{bh}{2}\right\}\delta h > 0, \end{aligned}$$

which is a contradiction. Therefore, $u_i \leqslant \min\{u_{i-1}, u_{i+1}\} + \delta h$.

Now consider any $k, l \in \{0, 1, \ldots, n+1\}$ with $k+1 < l$ and let $g_i \leqslant 0$ for $i = k+1, \ldots, l-1$. First, we shall prove that, for any $i \in \{k+1, \ldots, l-1\}$, the following implication holds:

$$u_{i-1} \leqslant u_i \quad \text{and} \quad u_i > u_{i+1} \quad \Rightarrow \quad u_k > u_{i+1}. \tag{9.2}$$

Thus, consider any $i \in \{k+1, \ldots, l-1\}$ such that the left-hand side of (9.2) is satisfied. Let $m \in \{k, \ldots, i-1\}$ be such that $u_s \leqslant u_{s+1}$ for $s = m, \ldots, i-1$. We assume that m cannot be further decreased, i.e., either $m = k$ or $u_{m-1} > u_m$. According to (5.8), one has

$$\begin{aligned} 0 &\geqslant \left(\varepsilon + \beta_i\tilde{\varepsilon} + \frac{bh}{2}\right)(u_i - u_{i-1}) + \left(\varepsilon + \beta_i\tilde{\varepsilon} - \frac{bh}{2}\right)(u_i - u_{i+1}) \\ &> \left(\varepsilon + \frac{bh}{2}\right)(u_i - u_{i-1}) - \frac{bh}{2}(u_i - u_{i+1}). \end{aligned} \tag{9.3}$$

If $m < i-1$, then for $s = m+1, \ldots, i-1$, in view of (5.8) one derives

$$\begin{aligned} 0 &\geqslant \left(\varepsilon + \beta_s\tilde{\varepsilon} + \frac{bh}{2}\right)(u_s - u_{s-1}) + \left(-\varepsilon - \beta_s\tilde{\varepsilon} + \frac{bh}{2}\right)(u_{s+1} - u_s) \\ &\geqslant \left(\varepsilon + \frac{bh}{2}\right)(u_s - u_{s-1}) - \varepsilon(u_{s+1} - u_s). \end{aligned} \tag{9.4}$$

Summing inequalities (9.3) and (9.4), one obtains

$$\begin{aligned} 0 &> \left(\varepsilon + \frac{bh}{2}\right)\sum_{s=m+1}^{i}(u_s - u_{s-1}) - \varepsilon\sum_{s=m+1}^{i-1}(u_{s+1} - u_s) - \frac{bh}{2}(u_i - u_{i+1}) \\ &= \left(\varepsilon + \frac{bh}{2}\right)(u_i - u_m) - \varepsilon(u_i - u_{m+1}) - \frac{bh}{2}(u_i - u_{i+1}) \geqslant \frac{bh}{2}(u_{i+1} - u_m). \end{aligned}$$

Therefore, $u_m > u_{i+1}$, which is true also if $m = i-1$ according to (9.3). If $m = k$ or $u_s > u_{s+1}$ for $s = k, \ldots, m-1$, then the right-hand side of (9.2) holds. Otherwise $m \geqslant k+2$ and there is $i' \in \{k+1, \ldots, m-1\}$ for which the left-hand side of (9.2) is satisfied and $u_{i'+1} > u_{i+1}$ holds. Hence the inequality $u_k > u_{i+1}$ follows by induction. For proving the statement of the theorem, let $j \in \{k, \ldots, l\}$ be such that $u_j = \max\{u_k, u_{k+1}, \ldots, u_l\}$ and let $u_j > \max\{u_k, u_l\}$. Then $u_j > u_{j+1}$ since otherwise $u_j = u_{j-1}$ in view of (5.8) and hence $u_j = u_k$ by induction. Thus, one has $u_k > u_{j+1}$ according to (9.2). Finally, applying the first part of the theorem, one obtains $u_j \leqslant \min\{u_{j-1}, u_{j+1}\} + \delta h < u_k + \delta h \leqslant \max\{u_k, u_l\} + \delta h$.

The implications for $g_i \geqslant 0$ follow analogously. □

Theorem 9.1 shows that if the discrete maximum principle is violated then the discrete solution is locally near to a constant function provided that δ or h is sufficiently small. Globally, the violation of the discrete maximum principle is smaller than or equal to δh.

REMARK 9.2 Using a similar construction to above, one could modify definition (4.11) in such a way that the resulting function β_i is continuous and equals 1 whenever the discrete solution attains an extremum at the node x_i. Then the statements of Theorem 9.1 hold with $\delta = 0$. However, the resulting method then adds artificial diffusion of magnitude $\tilde{\varepsilon}$ in regions where the discrete solution is constant, which is not desirable. Moreover, due to rounding errors, an approximation of a constant solution u typically possesses a lot of negligible extrema that also should not lead to adding a significant amount of artificial diffusion. The continuous function β_i defined at the beginning of this section satisfies this requirement.

Now let us report a few numerical results for β_i defined by (9.1). We used $\Delta = D = 0.5$ so that $\delta = 1$. For decreasing δ, we encountered increasing difficulty with the solution of the nonlinear problem, whereas the resulting approximate solution was not affected significantly. We again applied the fixed-point iteration described at the beginning of Section 6 that was terminated if absolute values of all components of the residual vector were smaller than 5×10^{-14}.

First, we repeated the computations of Section 6 and realized that all results are very similar for the continuous β_i, at least for $\mathrm{Pe} \geqslant 1$ (for $\mathrm{Pe} < 1$, a difference stems from using $L = 0.5$ instead of $L = 1$; cf. the end of Section 4). Then, we considered the counterexamples from Section 8 for which the discrete problems with discontinuous β_i were not solvable. Now, solutions could be computed and we obtained the following values of $\beta_1, \ldots, \beta_4$:

$$\begin{aligned}
&\text{data (8.3):} \quad \beta_1 = 1, \quad \beta_2 = 1, \quad \beta_3 = 1, \quad \beta_4 = 0.041172246777;\\
&\text{data (8.4):} \quad \beta_1 = 0, \quad \beta_2 = 0, \quad \beta_3 = 0.016194286589, \quad \beta_4 = 1;\\
&\text{data (8.5):} \quad \beta_1 = 0, \quad \beta_2 = 0, \quad \beta_3 = 0.018436266748, \quad \beta_4 = 1.
\end{aligned}$$

Finally, we investigated numerically a possible violation of the discrete maximum principle by the method (4.4), (4.5) if β_i are defined by (9.1). We used $\tilde{\varepsilon}$ from (6.2) and considered problem (3.1) with the data

$$b = 1, \quad g = 0, \quad u_{\mathrm{L}} = 1, \quad u_{\mathrm{R}} = 0, \tag{9.5}$$

and various values of $\varepsilon > 0$. According to (3.2), the exact solution of this problem is a decreasing function with values in the interval $[0, 1]$. For small ε, the solution is nearly constant except for a small neighbourhood of the right boundary point. Therefore, this problem is suitable for testing the validity of the discrete maximum principle by comparing the maximum value of the approximate solution

$$u_h^{\max} = \max_{i=0,\ldots,n+1} u_i$$

with the value 1. We used several values of ε and, for each of them, we computed approximate solutions for all values of $h \equiv 1/(n+1) \leqslant 0.25$ leading to $\mathrm{Pe} \geqslant 1$. It turns out that it is reasonable to consider moderate Péclet numbers and large Péclet numbers separately. More precisely, we considered $\mathrm{Pe} \in [1, 20)$ and $\mathrm{Pe} \in [20, \infty)$ separately. We denote by MAX the maximum of $u_h^{\max} - 1$ over all h for which the Péclet number belongs to the respective interval, by RMAX the maximum of $(u_h^{\max} - 1)/h$ and by $\mathrm{Pe}_{\mathrm{RMAX}}$ the value of Pe for which the maximum RMAX is attained. The results are summarized in Table 1. We observe that the results are in agreement with Theorem 9.1 and that the largest violations of

TABLE 1 *Violation of the discrete maximum principle for the data* (9.5) *and continuous* β_i *given by* (9.1)

	$\mathrm{Pe} \in [1, 20)$			$\mathrm{Pe} \in [20, \infty)$		
ε	MAX	RMAX	$\mathrm{Pe}_{\mathrm{RMAX}}$	MAX	RMAX	$\mathrm{Pe}_{\mathrm{RMAX}}$
10^{-1}	6.62−3	2.65−2	1.25		No $\mathrm{Pe} \geqslant 20$	
10^{-2}	3.55−3	9.27−2	1.85		No $\mathrm{Pe} \geqslant 20$	
10^{-3}	7.14−4	1.28−1	2.79	4.88−15	4.88−14	25.0
10^{-4}	1.06−4	1.40−1	3.77	5.60−14	9.23−13	21.6
10^{-5}	1.41−5	1.47−1	4.80	4.81−13	5.59−10	21.6
10^{-6}	1.77−6	1.51−1	5.84	6.06−12	6.92−8	22.9

the discrete maximum principle appear for small Péclet numbers, i.e., when the mesh width approaches the thickness of the boundary layer. The numerical results also suggest that the violation of the discrete maximum principle is bounded by $0.2 \min\{h, \varepsilon \ln(1/\varepsilon)\}$ and is often significantly smaller, so that it is negligible in the most cases. The results presented for $\mathrm{Pe} \in [20, \infty)$ are influenced by rounding errors and hence differ from values that would be obtained in exact arithmetic.

10. Conclusions and outlook

An algebraic flux correction scheme of TVD type, generalizing the one proposed in Kuzmin (2007), was studied in this work for one-dimensional steady-state convection–diffusion equations. The discrete operator was reformulated as a nonlinear finite difference operator with a parameter vector. Possible choices of this parameter vector were studied numerically. A fixed-point iteration was used for solving the nonlinear problem. The main results of this work concern properties of the nonlinear problem and the linear subproblems (discrete maximum principle, solvability). The unique solvability of the linear subproblems was studied under rather general conditions on the parameter vector of the scheme. Counterexamples concerning the existence of a solution of the nonlinear problem were provided. A modification of the scheme was proposed for which the existence of a solution and a weak variant of the discrete maximum principle were proved. Numerical experiments suggested that a good choice of the maximum artificial diffusion is $\tilde{\varepsilon} = bh(\coth \mathrm{Pe} - 1/\mathrm{Pe})/2$. Then the modified nonlinear scheme is solvable and, in all numerical experiments, the approximate solutions were not smeared and the violation of the discrete maximum principle was negligible.

Future work will study alternative algebraic flux correction schemes proposed, e.g., in Kuzmin (2012). As a first step, it has to be ensured that a solution of these nonlinear schemes exists. If this point is positively clarified, it makes sense to investigate the (order of) convergence to a solution. Of course, a numerical analysis for multidimensional problems is of utmost interest. From our experience so far, we think that such an analysis should initially consider model problems, simple domains and structured grids.

Acknowledgement

The authors are gratefully indebted to Professor Dmitri Kuzmin for many fruitful discussions on algebraic flux correction schemes.

Funding

The work of G.R.B. has been partially funded by the Leverhulme Trust, through the Research Project Grant RPG-2012-483. The work of P.K. has been partially supported through grant No.13-00522S of the Czech Science Foundation.

References

Arminjon, P. & Dervieux, A. (1993) Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. *J. Comput. Phys.*, **106**, 176–198.

Augustin, M., Caiazzo, A., Fiebach, A., Fuhrmann, J., John, V., Linke, A. & Umla, R. (2011) An assessment of discretizations for convection-dominated convection–diffusion equations. *Comput. Methods Appl. Mech. Eng.*, **200**, 3395–3409.

Bause, M. & Schwegler, K. (2012) Analysis of stabilized higher-order finite element approximation of nonstationary and nonlinear convection–diffusion–reaction equations. *Comput. Methods Appl. Mech. Eng.*, **209–212**, 184–196.

Boris, J. P. & Book, D. L. (1973) Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.*, **11**, 38–69.

Christie, I., Griffiths, D. F., Mitchell, A. R. & Zienkiewicz, O. C. (1976) Finite element methods for second order differential equations with significant first derivatives. *Int. J. Numer. Methods Eng.*, **10**, 1389–1396.

Codina, R. (1998) Comparison of some finite element methods for solving the diffusion–convection–reaction equation. *Comput. Methods Appl. Mech. Eng.*, **156**, 185–210.

Fuhrmann, J. & Langmach, H. (2001) Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. *Appl. Numer. Math.*, **37**, 201–230.

John, V. & Knobloch, P. (2007) On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. I. A review. *Comput. Methods Appl. Mech. Eng.*, **196**, 2197–2215.

John, V. & Knobloch, P. (2008) On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. II. Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Eng.*, **197**, 1997–2014.

John, V., Mitkova, T., Roland, M., Sundmacher, K., Tobiska, L. & Voigt, A. (2009) Simulations of population balance systems with one internal coordinate using finite element methods. *Chem. Eng. Sci.*, **64**, 733–741.

John, V. & Novo, J. (2012) On (essentially) non-oscillatory discretizations of evolutionary convection–diffusion equations. *J. Comput. Phys.*, **231**, 1570–1586.

John, V. & Schmeyer, E. (2008) Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. *Comput. Methods Appl. Mech. Eng.*, **198**, 475–494.

John, V. & Schumacher, L. (2014) A study of isogeometric analysis for scalar convection–diffusion equations. *Appl. Math. Lett.*, **27**, 43–48.

Kuzmin, D. (2006) On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. *J. Comput. Phys.*, **219**, 513–531.

Kuzmin, D. (2007) Algebraic flux correction for finite element discretizations of coupled systems. *Proceedings of the International Conference on Computational Methods for Coupled Problems in Science and Engineering* (M. Papadrakakis, E. Oñate & B. Schrefler eds). Barcelona: CIMNE, pp. 1–5.

Kuzmin, D. (2008) On the design of algebraic flux correction schemes for quadratic finite elements. *J. Comput. Appl. Math.*, **218**, 79–87.

Kuzmin, D. (2009) Explicit and implicit FEM-FCT algorithms with flux linearization. *J. Comput. Phys.*, **228**, 2517–2534.

Kuzmin, D. (2012) Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *J. Comput. Appl. Math.*, **236**, 2317–2337.

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

KUZMIN, D. & TUREK, S. (2004) High-resolution FEM–TVD schemes based on a fully multidimensional flux limiter. *J. Comput. Phys.*, **198**, 131–158.

LÖHNER, R., MORGAN, K., PERAIRE, J. & VAHDATI, M. (1987) Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier–Stokes equations. *Int. J. Numer. Methods Fluids*, **7**, 1093–1109.

ROOS, H.-G., STYNES, M. & TOBISKA, L. (2008) *Robust Numerical Methods for Singularly Perturbed Differential Equations*, 2nd edn. Berlin: Springer.

ZALESAK, S. T. (1979) Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, **31**, 335–362.

SIAM J. NUMER. ANAL.
Vol. 54, No. 4, pp. 2427–2451

ANALYSIS OF ALGEBRAIC FLUX CORRECTION SCHEMES*

GABRIEL R. BARRENECHEA†, VOLKER JOHN‡, AND PETR KNOBLOCH§

Abstract. A family of algebraic flux correction (AFC) schemes for linear boundary value problems in any space dimension is studied. These methods' main feature is that they limit the fluxes along each one of the edges of the triangulation, and we suppose that the limiters used are symmetric. For an abstract problem, the existence of a solution, existence and uniqueness of the solution of a linearized problem, and an a priori error estimate are proved under rather general assumptions on the limiters. For a particular (but standard in practice) choice of the limiters, it is shown that a local discrete maximum principle holds. The theory developed for the abstract problem is applied to convection-diffusion-reaction equations, where in particular an error estimate is derived. Numerical studies show its sharpness.

Key words. algebraic flux correction method, linear boundary value problem, well-posedness, discrete maximum principle, convergence analysis, convection-diffusion-reaction equations

AMS subject classifications. 65N12, 65N30

DOI. 10.1137/15M1018216

1. Introduction.

Many processes from nature and industry can be modeled using (systems of) partial differential equations (PDEs). Usually, these equations cannot be solved analytically. Instead, only numerical approximations can be computed, e.g., by using a finite element method (FEM). The Galerkin FEM replaces just the infinite-dimensional spaces from the variational form of the differential equation with finite-dimensional counterparts. However, if the considered problem contains a wide range of important scales, the Galerkin FEM does not give useful numerical results unless all scales are resolved. For many problems, the resolution of all scales is not affordable because of the huge computational costs (memory, computing time). The remedy consists of modifying the Galerkin FEM in such a way that the effect of small scales is taken into account on grids which do not resolve all scales. This methodology is usually called stabilization. The most common strategy modifies or enriches the Galerkin FEM, e.g., such that the new discrete problem provides additional control of the error in appropriate norms. An alternative approach acts on the algebraic level; i.e., algebraic representations of discrete operators and vectors are modified before computing a numerical solution. This paper studies a method of the latter type.

Applications of algebraically stabilized FEMs can be found in particular for convection-dominated problems. Their construction, e.g., in [18, 16, 17], is performed for transport equations, and they are called flux-corrected transport (FCT) schemes

*Received by the editors April 23, 2015; accepted for publication (in revised form) March 30, 2016; published electronically August 16, 2016. The research of the authors was funded by the Leverhulme Trust under grant RPG-2012-483.

http://www.siam.org/journals/sinum/54-4/M101821.html

†Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, Scotland (gabriel.barrenechea@strath.ac.uk).

‡Weierstrass Institute for Applied Analysis and Stochastics (WIAS), 10117 Berlin, Germany, and Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany (john@wias-berlin.de). The research of this author was partially supported by grant Jo329/10-2 within the DFG priority programme 1679: Dynamic simulation of interconnected solids processes.

§Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University in Prague, 18675 Praha 8, Czech Republic (knobloch@karlin.mff.cuni.cz). The research of this author was partially supported through grant 13-00522S of the Czech Science Foundation.

(see also [7] for their application to compressible flows). These schemes can be used also for the discretization of time-dependent convection-diffusion equations, e.g., as in [4, 11], where the convection-diffusion equations are part of population balance systems. In [11] it is explicitly emphasized that the FCT scheme was preferred over the popular streamline-upwind Petrov–Galerkin (SUPG) stabilization, which adds an additional term to the Galerkin FEM, because of a former bad experience with this stabilization. More precisely, the lack of positivity of the solution provided by SUPG caused blow ups in finite time for some nonlinear coupled problems in chemical engineering (for details, see [10]). Altogether, the advantages of the FCT methods, compared with the majority of other stabilized methods, are as follows. First, their construction relies on the goal of conservation and of satisfying a discrete maximum principle. Second, since this sort of method acts only at the algebraic level, without taking into consideration the weak formulation, their implementation is independent of the space dimension. The importance of these two points for many applications does not need to be emphasized. However, there are also drawbacks. First, for most methods, one has to solve a nonlinear discrete problem, even when the PDE to be solved is linear. This issue is, in our opinion, of minor importance, since in applications one encounters generally nonlinear problems. Second, the FCT methodology has, so far, been applied successfully only for lowest order finite elements, which limits the accuracy of the computed solutions to the best approximation in these spaces (the only exception of this fact being, to the best of our knowledge, the work [15]).

This paper analyzes algebraic stabilizations for linear steady-state boundary value problems. These methods are called algebraic flux correction (AFC) schemes. Apart from the obvious properties of these methods, which are the basis of their construction, there has been no numerical analysis of them until very recently. The first contribution in this field is [2], where some preliminary results on the analysis of an AFC scheme (cf. [14]) for a linear steady-state convection-diffusion-reaction equation in one space dimension were reported. The discretization studied in [2] is in some sense more general than the AFC methodology used in practice. In the methodology of [2], one has to compute limiters $\alpha_{ij} \in [0,1]$ (see below), and in contrast to the common application of AFC schemes, it was not assumed that $\alpha_{ij} = \alpha_{ji}$, which may cause a lack of conservation. Besides other properties, it was proved in [2] that the nonlinear discrete problem might not even possess a solution. Thus, there is an important physical as well as a strong mathematical reason for including the symmetry condition in the scheme, which will be done in this paper.

The first part of the paper (sections 2–6) considers a general linear boundary value problem in several space dimensions. After introducing a nonlinear AFC scheme in section 2, the existence of a solution is proved, and then the existence of a unique solution of the linearized scheme is shown, both in section 3. The symmetry of the limiters, i.e., $\alpha_{ij} = \alpha_{ji}$, the requirement that $\alpha_{ij} \in [0,1]$, and a continuity assumption are the minimal assumptions used in this section. Section 4 considers a concrete choice of the limiters, which is a standard definition found in the literature. It is shown that these limiters satisfy the assumptions made in the preceding analysis, so they lead to discrete problems that have a solution. In section 5 we give a general proof of the discrete maximum principle, since we have not been able to find it in the literature, although the AFC family of methods is built to preserve this property. In section 6, the AFC scheme is formulated in a variational form and an abstract error estimate is derived, with only the same minimal assumptions on the limiters as used in section 3. As usual for stabilized methods, the norm for which the error estimate is given contains a contribution from the stabilization. To the best of our

knowledge, this is the first error estimate for algebraically stabilized FEMs. In the second part of the paper (sections 7–8), the abstract theory is applied to steady-state linear convection-diffusion-reaction equations. In section 7 an error estimate for this kind of equation is derived. Numerical studies are presented in section 8. It is shown that within the minimal assumptions on the limiters used in the analysis, the derived error estimate is sharp. However, applying the definition of the limiters as discussed in section 5, one can observe a higher order of convergence. The orders of convergence for standard norms depend on the concrete grid and are sometimes suboptimal. Finally, in an appendix at the end of the paper a few supplementary results are proved.

2. An algebraic flux correction scheme. Consider a linear boundary value problem for which the maximum principle holds. Let us discretize this problem by the FEM. Then the discrete solution can be represented by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of U ($0 < M < N$) correspond to nodes where Dirichlet boundary conditions are prescribed, whereas the first M components of U are computed using the finite element discretization of the underlying PDE. Then $\mathrm{U} \equiv (u_1, \ldots, u_N)$ satisfies a system of linear equations of the form

$$\sum_{j=1}^{N} a_{ij}\, u_j = g_i\,, \qquad i = 1, \ldots, M\,, \tag{1}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{2}$$

We assume that the matrix $(a_{ij})_{i,j=1}^{M}$ is positive definite, i.e.,

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j > 0 \quad \forall\, (u_1, \ldots, u_M) \in \mathbb{R}^M \setminus \{0\}\,. \tag{3}$$

It is natural to require that the maximum principle also hold for the discrete problem (1), (2). Due to (3), the diagonal entries of the matrix $(a_{ij})_{i,j=1}^{M}$ are positive, and hence, locally, the discrete maximum principle corresponds to the statement

$$\forall\, i \in \{1, \ldots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i,\, a_{ij} \ne 0} u_j \tag{4}$$

or, at least,

$$\forall\, i \in \{1, \ldots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i,\, a_{ij} \ne 0} u_j^+\,, \tag{5}$$

where $u_j^+ = \max\{0, u_j\}$. It can be shown (cf. the appendix), that (4) holds if and only if

$$a_{ij} \le 0 \qquad \forall\, i \ne j,\, i = 1, \ldots, M,\, j = 1, \ldots, N\,, \tag{6}$$

and

$$\sum_{j=1}^{N} a_{ij} = 0\,, \qquad i = 1, \ldots, M\,. \tag{7}$$

The discrete maximum principle (5) holds if and only if (6) is satisfied and

$$\sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1, \ldots, M\,. \tag{8}$$

While conditions (7) or (8) are often satisfied, the property (6) does not hold for many discretizations, in particular, of convection-dominated problems. The aim of the AFC method is to modify the algebraic system (1) in such a way that the necessary conditions for the validity of the discrete maximum principle are satisfied and layers are not excessively smeared.

The starting point of the AFC algorithm is the finite element matrix $\mathbb{A} = (a_{ij})_{i,j=1}^{N}$ corresponding to the above-mentioned finite element discretization in the case where homogeneous natural boundary conditions are used instead of the Dirichlet ones. We introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^{N}$ possessing the entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \quad \forall\, i \neq j\,, \qquad d_{ii} = -\sum_{j \neq i} d_{ij}\,. \tag{9}$$

Then the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ satisfies the necessary conditions for the discrete maximum principle provided that (7) or (8) holds for the matrix $\mathbb{A}$.

Going back to the solution of (1), this system is equivalent to

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + (\mathbb{D}\,\mathrm{U})_i\,, \qquad i = 1, \ldots, M\,. \tag{10}$$

Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\,\mathrm{U})_i = \sum_{j \neq i} f_{ij}\,, \qquad i = 1, \ldots, N\,,$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. Now the idea of the AFC schemes is to limit those antidiffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (1) (or, equivalently, (10)) is replaced by

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij}\, f_{ij}\,, \qquad i = 1, \ldots, M\,, \tag{11}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem. They can be chosen in various ways, but their definition is always based on the above fluxes f_{ij}; see [13, 14, 15, 16, 17] for examples. To guarantee that the resulting scheme is conservative, one should require that the coefficients α_{ij} be symmetric, i.e.,

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i, j = 1, \ldots, N\,. \tag{12}$$

Rewriting (11) using the definition of the matrix $\tilde{\mathbb{A}}$, one obtains the final form of the AFC scheme to be investigated in this paper. It is the following system of nonlinear equations:

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) = g_i\,, \qquad i = 1, \ldots, M\,, \tag{13}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,, \tag{14}$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \ldots, u_N) \in [0, 1]$, $i, j = 1, \ldots, N$, satisfy (12).

3. Solvability of the algebraic flux correction scheme and of its linearized variant. In this section we prove that the nonlinear problem (13), (14) is solvable under a continuity assumption on α_{ij}. As a consequence, we obtain the unique solvability of the linearized problem (13), (14) (with α_{ij} independent of the solution), which is useful for computing the solution of (13), (14) numerically using a fixed-point iteration. The following result will be of great use in the proof of existence of solutions below.

LEMMA 1. *Consider any $\mu_{ij} = \mu_{ji} \le 0$, $i, j = 1, \dots, N$. Then*

$$\sum_{i,j=1}^{N} v_i \, \mu_{ij} \, (v_j - v_i) = - \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} \mu_{ij} \, (v_i - v_j)^2 \ge 0 \quad \forall \, v_1, \dots, v_N \in \mathbb{R} \, .$$

Proof. A quick calculation shows that

$$\begin{aligned} \sum_{i,j=1}^{N} v_i \, \mu_{ij} \, (v_j - v_i) &= \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} v_i \, \mu_{ij} \, (v_j - v_i) + \sum_{\substack{j,\, i \,=\, 1 \\ j \,>\, i}}^{N} v_j \, \mu_{ji} \, (v_i - v_j) \\ &= - \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} \mu_{ij} \, (v_i - v_j)^2 \ge 0 \, , \end{aligned}$$

and the proof is finished. ∎

For proving the solvability of the nonlinear problem, we use the following consequence of Brower's fixed-point theorem, whose proof can be found in [20, Lemma 1.4, p. 164].

LEMMA 2. *Let X be a finite-dimensional Hilbert space with inner product $(\cdot, \cdot)_X$ and norm $\| \cdot \|_X$. Let $T : X \to X$ be a continuous mapping, and let $K > 0$ be a real number such that $(Tx, x)_X > 0$ for any $x \in X$ with $\|x\|_X = K$. Then there exists $x \in X$ such that $\|x\|_X < K$ and $Tx = 0$.*

The following is our main result on existence of solutions for the AFC scheme.

THEOREM 3. *Let (3) hold. For any $i, j \in \{1, \dots, N\}$, let $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ be such that $\alpha_{ij}(u_1, \dots, u_N)(u_j - u_i)$ is a continuous function of $u_1, \dots, u_N$. Finally, let the functions α_{ij} satisfy (12). Then there exists a solution of the nonlinear problem (13), (14).*

Proof. Throughout this proof, we denote by $\widetilde{\mathrm{V}} \equiv (v_1, \dots, v_M)$ the elements of the space $\mathbb{R}^M$ and, if v_i with $i \in \{M+1, \dots, N\}$ occurs, we always assume that $v_i = u_i^b$. To any $\widetilde{\mathrm{V}} \in \mathbb{R}^M$ we assign $\mathrm{V} := (v_1, \dots, v_N)$. Furthermore, we set $\mathrm{G} := (g_1, \dots, g_M)$. We denote by $(\cdot, \cdot)$ the usual inner product in $\mathbb{R}^M$ and by $\| \cdot \|$ the corresponding (Euclidean) norm.

It is easy to show by contradiction that, in view of (3),

$$C_M := \inf_{\|\widetilde{\mathrm{V}}\| = 1} \sum_{i,j=1}^{M} v_i \, a_{ij} \, v_j > 0 \, .$$

Thus, one has

$$\sum_{i,j=1}^{M} v_i \, a_{ij} \, v_j \ge C_M \, \|\widetilde{\mathrm{V}}\|^2 \quad \forall \, \widetilde{\mathrm{V}} \in \mathbb{R}^M \, . \tag{15}$$

Let us define the operator $T : \mathbb{R}^M \to \mathbb{R}^M$ by

$$(T\,\widetilde{\mathrm{V}})_i = \sum_{j=1}^{N} a_{ij}\, v_j + \sum_{j=1}^{N} [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - v_i) - g_i\,, \qquad i = 1, \ldots, M\,.$$

Then U is a solution of the nonlinear problem (13), (14) if and only if $T\,\widetilde{\mathrm{U}} = 0$. The operator T is continuous and, in view of (15), Lemma 1, and Hölder's and Young's inequalities, one derives

$$\begin{aligned}(T\,\widetilde{\mathrm{V}}, \widetilde{\mathrm{V}}) &= \sum_{i,j=1}^{M} v_i\, a_{ij}\, v_j + \sum_{i,j=1}^{N} v_i\, [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - v_i) \\ &\quad + \sum_{i=1}^{M} v_i \sum_{j=M+1}^{N} a_{ij}\, u_j^b - \sum_{i=M+1}^{N} u_i^b \sum_{j=1}^{N} [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - u_i^b) - (\mathrm{G}, \widetilde{\mathrm{V}}) \\ &\geq C_M\, \|\widetilde{\mathrm{V}}\|^2 - C_0 - C_1\, \|\widetilde{\mathrm{V}}\| \geq \frac{C_M}{2}\, \|\widetilde{\mathrm{V}}\|^2 - C_2\,,\end{aligned}$$

where C_0, C_1, and C_2 are positive constants that do not depend on $\widetilde{\mathrm{V}}$. Then for any $\widetilde{\mathrm{V}} \in \mathbb{R}^M$ satisfying $\|\widetilde{\mathrm{V}}\| = \sqrt{3\, C_2 / C_M}$, one has $(T\,\widetilde{\mathrm{V}}, \widetilde{\mathrm{V}}) > 0$, and hence, according to Lemma 2, there exists $\widetilde{\mathrm{U}} \in \mathbb{R}^M$ such that $T\,\widetilde{\mathrm{U}} = 0$. ∎

Corollary 4. *Let* (3) *hold. Consider any* $\alpha_{ij} \in [0, 1]$, $i, j = 1, \ldots, N$, *satisfying* (12). *Then the system* (13), (14) *has a unique solution for any* $g_1, \ldots, g_M \in \mathbb{R}$ *and* $u_{M+1}^b, \ldots, u_N^b \in \mathbb{R}$.

Proof. According to Theorem 3, for any values of $g_1, \ldots, g_M$ and $u_{M+1}^b, \ldots, u_N^b$, there exists a solution of the considered linear system. Consequently, the solutions have to be unique. ∎

Remark 5. The statement of Corollary 4 can be proved directly (without using Theorem 3) by showing that the homogeneous system

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) = 0\,, \qquad i = 1, \ldots, M\,, \tag{16}$$

$$u_i = 0\,, \qquad i = M+1, \ldots, N\,, \tag{17}$$

has only the trivial solution. Indeed, if $U = (u_1, \ldots, u_N)$ solves (16), (17), then according to Lemma 1, one has

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j = -\sum_{i,j=1}^{N} u_i\, (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) \leq 0\,.$$

Therefore, $u_i = 0$, $i = 1, \ldots, M$, in view of (3).

Finally, let us formulate sufficient conditions on the functions α_{ij}, ensuring the validity of the continuity assumption in Theorem 3 for many particular examples of the functions α_{ij} used in practice (cf., e.g., [13, 16, 17]).

Lemma 6. *Consider any* $i, j \in \{1, \ldots, N\}$, *and let* $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ *satisfy*

$$\alpha_{ij}(U) = \frac{A_{ij}(U)}{|u_j - u_i| + B_{ij}(U)} \quad \forall\, U \equiv (u_1, \ldots, u_N) \in \mathbb{R}^N,\ u_i \neq u_j\,, \tag{18}$$

where $A_{ij}, B_{ij} : \mathbb{R}^N \to [0,\infty)$ are nonnegative functions that are continuous at any point $U \in \mathbb{R}^N$ with $u_i \neq u_j$. Then $\Phi_{ij}(U) := \alpha_{ij}(U)(u_j - u_i)$ is a continuous function of $u_1, \dots, u_N$ on $\mathbb{R}^N$. Moreover, if the functions A_{ij}, B_{ij} are Lipschitz-continuous with the constant L in the sets $\{U \in \mathbb{R}^N;\ u_i < u_j\}$ and $\{U \in \mathbb{R}^N;\ u_i > u_j\}$, then the function Φ_{ij} is Lipschitz-continuous on $\mathbb{R}^N$, with the constant $2\,L + \sqrt{2}$.

Proof. Consider any $\bar{U} \equiv (\bar{u}_1, \dots, \bar{u}_N) \in \mathbb{R}^N$. If $\bar{u}_i \neq \bar{u}_j$, then there is a neighbourhood of $\bar{U}$, where the denominator from (18) does not vanish and the functions A_{ij}, B_{ij} are continuous so that α_{ij} is continuous at $\bar{U}$. If $\bar{u}_i = \bar{u}_j$, we employ the fact that $\alpha_{ij} \in [0,1]$, which implies that $|\alpha_{ij}(U)(u_j - u_i)| \le |u_j - u_i| \le \sqrt{2}\,\|U - \bar{U}\|$ for any $U \equiv (u_1, \dots, u_N) \in \mathbb{R}^N$. Thus, $\alpha_{ij}(U)(u_j - u_i)$ is continuous at $\bar{U}$.

To prove the Lipschitz-continuity of Φ_{ij}, consider any U, $\bar{U} \in \mathbb{R}^N$ with $U = (u_1, \dots, u_N)$ and $\bar{U} = (\bar{u}_1, \dots, \bar{u}_N)$. Set $v = u_j - u_i$, $\bar{v} = \bar{u}_j - \bar{u}_i$. If $v\,\bar{v} \le 0$, then

$$|\Phi_{ij}(U) - \Phi_{ij}(\bar{U})| \le |v| + |\bar{v}| = |v - \bar{v}| \le \sqrt{2}\,\|U - \bar{U}\|\,.$$

If $v\,\bar{v} > 0$, then

$$\Phi_{ij}(U) - \Phi_{ij}(\bar{U}) = (A_{ij}(U) - A_{ij}(\bar{U}))\,\frac{\bar{v}}{|\bar{v}| + B_{ij}(\bar{U})} + \alpha_{ij}(U)\,\frac{(B_{ij}(\bar{U}) - B_{ij}(U))\,\bar{v} + (v - \bar{v})\,B_{ij}(\bar{U})}{|\bar{v}| + B_{ij}(\bar{U})},$$

and hence

$$|\Phi_{ij}(U) - \Phi_{ij}(\bar{U})| \le |A_{ij}(U) - A_{ij}(\bar{U})| + |B_{ij}(U) - B_{ij}(\bar{U})| + |v - \bar{v}|\,.$$

This proves the lemma. ∎

4. An example of the choice of α_{ij}. In this section we present a concrete choice of the limiters α_{ij}. This choice is often used in computations, and we show that it satisfies the assumptions of Lemma 6 and hence leads to a solvable nonlinear problem (13), (14).

The definition of the coefficients α_{ij} considered in this section relies on the values P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i = 1, \dots, N$ in the following way. First, one initializes all these quantities by zero. Then one goes through all pairs of indices $i, j \in \{1, \dots, N\}$ and performs the updates

$$\begin{aligned}
&P_i^+ := P_i^+ + \max\{0, f_{ij}\}\,,\ P_i^- := P_i^- - \max\{0, f_{ji}\} \quad \text{if } a_{ji} \le a_{ij}\,,\\
&Q_i^+ := Q_i^+ + \max\{0, f_{ji}\}\,,\ Q_i^- := Q_i^- - \max\{0, f_{ij}\} \quad \text{if } i < j\,,\\
&Q_j^+ := Q_j^+ + \max\{0, f_{ij}\}\,,\ Q_j^- := Q_j^- - \max\{0, f_{ji}\} \quad \text{if } i < j\,,
\end{aligned}$$

where we again use the notation $f_{ij} = d_{ij}\,(u_j - u_i)$. After having computed the values P_i^+, P_i^-, Q_i^+, Q_i^-, $i = 1, \dots, N$, one defines

$$R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \qquad i = 1, \dots, N\,.$$

If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Furthermore, according to [12], these quantities are set to 1 at Dirichlet nodes, i.e.,

$$R_i^+ := 1\,, \quad R_i^- := 1\,, \qquad i = M+1, \dots, N\,.$$

Finally, for any $i, j \in \{1, \ldots, N\}$ such that $a_{ji} \leq a_{ij}$, one sets

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}\,. \tag{19}$$

It is worth mentioning that this algorithm is the one presented in [14] (that originates from the ideas of [22]) to which, following [12], the symmetry condition $\alpha_{ij} = \alpha_{ji}$ has been added.

Note that the quantities P_i^+, P_i^-, Q_i^+, Q_i^- can be expressed in the form

$$P_i^+ = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} f_{ij}^+\,, \qquad P_i^- = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} f_{ij}^-\,, \qquad Q_i^+ = -\sum_{j=1}^{N} f_{ij}^-\,, \qquad Q_i^- = -\sum_{j=1}^{N} f_{ij}^+\,, \tag{20}$$

where $f_{ij}^+ = \max\{0, f_{ij}\}$ and $f_{ij}^- = \min\{0, f_{ij}\}$.

The following result shows that the above coefficients α_{ij} satisfy the hypotheses of Theorem 3, and then that they lead to a solvable nonlinear problem (13), (14).

LEMMA 7. *The above coefficients α_{ij} are such that $\alpha_{ij}(u_1, \ldots, u_N)(u_j - u_i)$ are Lipschitz-continuous functions of $u_1, \ldots, u_N$ on $\mathbb{R}^N$.*

Proof. Consider any $i, j \in \{1, \ldots, N\}$. It suffices to consider the case $\alpha_{ij} \not\equiv 1$ (and hence $d_{ij} \neq 0$). Furthermore, due to (12), one may assume that $a_{ji} \leq a_{ij}$. If $u_i > u_j$, then $f_{ij} > 0$, and hence

$$\alpha_{ij} = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{|f_{ij}| + \widetilde{P}_i^+} \qquad \text{with} \qquad \widetilde{P}_i^+ = \sum_{\substack{k=1 \\ a_{ki} \leq a_{ik},\, k \neq j}}^{N} f_{ik}^+\,.$$

If $u_i < u_j$, then $f_{ij} < 0$ so that

$$\alpha_{ij} = R_i^- = \frac{\min\{-P_i^-, -Q_i^-\}}{|f_{ij}| - \widetilde{P}_i^-} \qquad \text{with} \qquad \widetilde{P}_i^- = \sum_{\substack{k=1 \\ a_{ki} \leq a_{ik},\, k \neq j}}^{N} f_{ik}^-\,.$$

Thus, α_{ij} is of the form (18), with functions A_{ij} and B_{ij} satisfying

$$A_{ij} = \frac{1}{|d_{ij}|} \begin{cases} \min\{-P_i^-, -Q_i^-\} & \text{if } u_i < u_j\,, \\ \min\{P_i^+, Q_i^+\} & \text{if } u_i > u_j\,, \end{cases} \qquad B_{ij} = \frac{1}{|d_{ij}|} \begin{cases} -\widetilde{P}_i^- & \text{if } u_i < u_j\,, \\ \widetilde{P}_i^+ & \text{if } u_i > u_j\,. \end{cases}$$

Since the maximum or minimum of two Lipschitz-continuous functions with constant L is again a Lipschitz-continuous function with constant L, the functions A_{ij} and B_{ij} are Lipschitz-continuous with constant $\sqrt{2}\,(\sum_{k=1}^{N} |d_{ik}|)/|d_{ij}|$ in the sets $\{u_i < u_j\}$ and $\{u_i > u_j\}$. Then the hypotheses of Lemma 6 are satisfied, and the result immediately follows from Lemma 6. ∎

Remark 8. There is an apparent ambiguity in the definition of the coefficients α_{ij} if $a_{ij} = a_{ji}$. However, often $a_{ij} + a_{ji} \leq 0$ (cf. assumption (22) in the next section), and then $a_{ij} = a_{ji} \leq 0$. Thus, if the artificial diffusion matrix is defined by (9), one obtains $d_{ij} = 0$ so that the respective α_{ij} does not occur in the nonlinear problem (13), (14) and can be defined arbitrarily.

5. The discrete maximum principle. In this section we prove several versions of the discrete maximum principle for the case when the coefficients α_{ij} are defined as in the previous section. We start with the main assumptions needed for the proofs, namely,

$$a_{ii} > 0\,, \qquad \sum_{j=1}^{N} a_{ij} \geq 0 \quad \forall\, i = 1, \ldots, M\,, \tag{21}$$

$$a_{kl} + a_{lk} \leq 0 \quad \forall\, k, l = 1, \ldots, N\,,\ k \neq l\,,\ k \leq M,\ \text{or}\ l \leq M\,, \tag{22}$$

and we recall that $d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\}$ for all $i, j = 1, \ldots, N$, $i \neq j$ (cf. (9)). The first condition in (21) is a consequence of (3), and the second is a necessary condition for the validity of the discrete maximum principle in the case of linear problem (1), (2). Note that the row sums are not affected by adding the nonlinear term in (13). Condition (22) is weaker than (6). In section 7, we present a discrete problem for which all the assumptions in (21) and (22) are satisfied.

Also, we present some notation that will be useful in what follows. We denote by

$$\mathrm{Up}_i = \{j \in \{1, \ldots, N\}\,;\ j \neq i,\ a_{ij} < 0\}\,, \qquad i = 1, \ldots, M\,,$$

the sets of upwind nodes, and by

$$\mathrm{Do}_i = \{j \in \{1, \ldots, N\}\,;\ j \neq i,\ a_{ij} > 0\}\,, \qquad i = 1, \ldots, M\,,$$

the sets of downwind nodes. In what follows, we shall tacitly assume that these sets are not empty.

Thanks to (22), for any $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N\}$ such that $i \neq j$ and $d_{ij} \neq 0$, one derives

$$a_{ij} < a_{ji} \quad \Leftrightarrow \quad j \in \mathrm{Up}_i\,, \qquad\qquad a_{ji} \leq a_{ij} \quad \Leftrightarrow \quad j \in \mathrm{Do}_i\,.$$

Therefore, the sums in (20) defining P_i^+ and P_i^- can be written in the form

$$P_i^+ = \sum_{j \in \mathrm{Do}_i} f_{ij}^+\,, \qquad P_i^- = \sum_{j \in \mathrm{Do}_i} f_{ij}^-\,, \qquad i = 1, \ldots, M\,. \tag{23}$$

Moreover, the second term on the left-hand side of (13) can be written as

$$\begin{aligned}\sum_{j=1}^{N} (1 - \alpha_{ij})\, f_{ij} &= \sum_{j=1}^{N} f_{ij} - \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} \alpha_{ij}\, f_{ij} + \sum_{\substack{j=1 \\ a_{ij} < a_{ji}}}^{N} \alpha_{ji}\, f_{ji} \\ &= \sum_{j=1}^{N} f_{ij} - \sum_{j \in \mathrm{Do}_i} \alpha_{ij}\, f_{ij} + \sum_{j \in \mathrm{Up}_i} \alpha_{ji}\, f_{ji}\,.\end{aligned}$$

Furthermore, $\alpha_{ij}\, f_{ij} = R_i^+\, f_{ij}^+ + R_i^-\, f_{ij}^-$ for $i \in \{1, \ldots, M\}$ and $j \in \mathrm{Do}_i$, and consequently, $\alpha_{ji}\, f_{ji} = R_j^+\, f_{ji}^+ + R_j^-\, f_{ji}^-$ if $i \in \{1, \ldots, M\}$ and $j \in \mathrm{Up}_i$. Then since $f_{ji}^+ = -f_{ij}^-$ and $f_{ji}^- = -f_{ij}^+$, one obtains

$$\sum_{j=1}^{N} (1 - \alpha_{ij})\, f_{ij} = \sum_{j=1}^{N} f_{ij} - \sum_{j \in \mathrm{Do}_i} (R_i^+\, f_{ij}^+ + R_i^-\, f_{ij}^-) - \sum_{j \in \mathrm{Up}_i} (R_j^+\, f_{ij}^- + R_j^-\, f_{ij}^+)\,.$$

Finally, denoting $Z_i^+ := 1 - R_i^+$ and $Z_i^- := 1 - R_i^-$, it follows that

$$\sum_{j=1}^{N} (1-\alpha_{ij})\, f_{ij} = \sum_{j\in \mathrm{Do}_i} (Z_i^+ \, f_{ij}^+ + Z_i^- \, f_{ij}^-) + \sum_{j\in \mathrm{Up}_i} (Z_j^+ \, f_{ij}^- + Z_j^- \, f_{ij}^+)\,.$$

Thus, the AFC scheme (13), (14) can be written in the form

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j\in \mathrm{Do}_i} (Z_i^+ \, f_{ij}^+ + Z_i^- \, f_{ij}^-) + \sum_{j\in \mathrm{Up}_i} (Z_j^+ \, f_{ij}^- + Z_j^- \, f_{ij}^+) = g_i\,, \qquad i = 1, \ldots, M\,, \tag{24}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{25}$$

Next, defining

$$A_i = u_i \sum_{j=1}^{N} a_{ij}\,, \tag{26}$$

one derives, for any $i \in \{1, \ldots, M\}$,

$$\sum_{j=1}^{N} a_{ij}\, u_j = \sum_{j=1}^{N} a_{ij}\, (u_j - u_i) + A_i = \sum_{j\in \mathrm{Up}_i} a_{ij}\, (u_j - u_i) + \sum_{j\in \mathrm{Do}_i} a_{ij}\, (u_j - u_i) + A_i\,.$$

In view of (22), one has $a_{ij} = -d_{ij}$ for $j \in \mathrm{Do}_i$, and then

$$\sum_{j=1}^{N} a_{ij}\, u_j = \sum_{j\in \mathrm{Up}_i} a_{ij}\, (u_j - u_i) - \sum_{j\in \mathrm{Do}_i} f_{ij} + A_i\,.$$

Therefore, using that $\sum_{j\in \mathrm{Do}_i} f_{ij} = P_i^+ + P_i^-$ (cf. (23)), (24) is equivalent to

$$A_i - P_i^+ \, R_i^+ - \, P_i^- \, R_i^- + \sum_{j\in \mathrm{Up}_i} (Z_j^+ \, f_{ij}^- + Z_j^- \, f_{ij}^+ + a_{ij}\, (u_j - u_i)) = g_i\,. \tag{27}$$

The following is a preliminary technical result.

LEMMA 9. *Consider any $i \in \{1, \ldots, M\}$, and let $u_i \le u_j$ for all $j \in \mathrm{Up}_i$. Then*

$$A_i - \, P_i^- \, R_i^- + R_i^+ \sum_{j\in \mathrm{Do}_i} a_{ij}\, (u_j - u_i)^- + \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^+ \, d_{ij})\, |u_j - u_i| = g_i\,. \tag{28}$$

On the other hand, if $u_i \ge u_j$ for all $j \in \mathrm{Up}_i$, then

$$A_i - P_i^+ \, R_i^+ + R_i^- \sum_{j\in \mathrm{Do}_i} a_{ij}\, (u_j - u_i)^+ - \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^- \, d_{ij})\, |u_j - u_i| = g_i\,. \tag{29}$$

Proof. Since $f_{ij}^+ = d_{ij}\, (u_j - u_i)^-$, $f_{ij}^- = d_{ij}\, (u_j - u_i)^+$, and $d_{ij} = -a_{ij}$ if $j \in \mathrm{Do}_i$, the lemma follows immediately from (27). ∎

The following result is a quick consequence of the above lemma, whose implications will become apparent in Corollary 11.

COROLLARY 10. *Consider any* $i \in \{1, \dots, M\}$, *and let* $u_i \le u_j$ *for all* $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$. *Then*

$$A_i + \sum_{j \in \mathrm{Up}_i} (a_{ij} + Z_j^+ \, d_{ij}) \, |u_j - u_i| = g_i \, . \tag{30}$$

On the other hand, if $u_i \ge u_j$ *for all* $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, *then*

$$A_i - \sum_{j \in \mathrm{Up}_i} (a_{ij} + Z_j^- \, d_{ij}) \, |u_j - u_i| = g_i \, . \tag{31}$$

Proof. One has $f_{ij}^+ = 0$ for $j = 1, \dots, N$, and hence $Q_i^- = 0$, which gives $P_i^- \, R_i^- = 0$. Then (30) follows from (28). To prove (31) it is enough to note that $f_{ij}^- = 0$ for $j = 1, \dots, N$, which leads to $Q_i^+ = 0$ and $P_i^+ \, R_i^+ = 0$, and then apply (29). ∎

Finally, the following corollary states that if $g_i \le 0$ (≥ 0), then u_i cannot be a strict *positive (negative)* local maximum (minimum).

COROLLARY 11. *Consider any* $i \in \{1, \dots, M\}$. *Then*

$$g_i \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j \quad \textit{for } u_i \ge 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j^+ \, , \tag{32}$$

$$g_i \ge 0 \quad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j \quad \textit{for } u_i \le 0 \quad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j^- \, . \tag{33}$$

Proof. Let $u_i \ge 0$. Then thanks to (21), $A_i \ge 0$ (where A_i is defined in (26)). If $u_i > u_j$ for all $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, then (31) holds with a positive left-hand side. Thus, if $g_i \le 0$, then $u_i \le u_j$ for some $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, which implies (32). The second statement is proved in an analogous way. ∎

Remark 12. It is worth remarking that, if $\sum_{j=1}^N a_{ij} = 0$, then the previous results can be strengthened since Lemma 9 and Corollary 10 hold with $A_i = 0$. Then Corollary 11 is valid without the restriction on the sign of u_i; i.e., for any $i \in \{1, \dots, M\}$, one has

$$g_i \le 0 \qquad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j \, ,$$

$$g_i \ge 0 \qquad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j \, .$$

This is in accordance with the corresponding results for PDEs (see, e.g., [6]).

6. Variational form of the algebraic flux correction scheme and error estimation. In this section we show how the linear system (1), (2) originates from a variational problem representing a finite element discretization and how, in turn, the nonlinear algebraic problem (13), (14) can be put into a variational form. Then the derivation of an error estimate is discussed. It is important to notice that all of the results of this section, and the following one, are valid for limiters α_{ij} that are only required to belong to $[0, 1]$.

Let $\Omega \subset \mathbb{R}^d$, $d \ge 1$, be a bounded domain and let the boundary $\partial\Omega$ of Ω be Lipschitz-continuous and polyhedral (if $d \ge 2$). Let $a : H^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be a bilinear form, let $u_b \in H^{1/2}(\partial\Omega) \cap C(\partial\Omega)$, let $g \in H^{-1}(\Omega)$, and consider the following variational problem:

Find $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u, v) = \langle g, v \rangle \quad \forall \, v \in H_0^1(\Omega) \, . \tag{34}$$

An example of such a variational problem will be presented in the next section.

To solve (34) numerically, let us introduce a finite element space $W_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$ approximating the space $H^1(\Omega)$, and set $V_h := W_h \cap H_0^1(\Omega)$. We denote the basis functions of W_h by $\varphi_1, \ldots, \varphi_N$ and assume that the functions $\varphi_1, \ldots, \varphi_M$ (with $0 < M < N$) form a basis in V_h. In addition, we assume that there are points $x_1, \ldots, x_N \in \overline{\Omega}$ such that $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol, and that $x_{M+1}, \ldots, x_N \in \partial\Omega$ (note that $x_1, \ldots, x_M \in \Omega$). Since constant functions are always required to be contained in W_h, one has $\sum_{i=1}^N \varphi_i = 1$ in Ω. In what follows, for any $u_h \in W_h$ (or v_h, z_h, etc.), we shall denote by $\{u_i\}_{i=1}^N$ (or $\{v_i\}_{i=1}^N$, $\{z_i\}_{i=1}^N$, etc.) the uniquely determined coefficients with respect to the above basis of W_h, i.e.,

$$u_h = \sum_{i=1}^N u_i \, \varphi_i \qquad \left(\text{or} \quad v_h = \sum_{i=1}^N v_i \, \varphi_i \,, \quad z_h = \sum_{i=1}^N z_i \, \varphi_i \,, \quad \text{etc.} \right) .$$

Of course, $u_i = u_h(x_i)$ (or $v_i = v_h(x_i)$, $z_i = z_h(x_i)$, etc.) for any $i \in \{1, \ldots, N\}$.

It is sometimes convenient (cf. section 7) to approximate the bilinear form a by a bilinear form $a_h : W_h \times V_h \to \mathbb{R}$. We assume that a_h is elliptic on the space V_h; i.e., there is a constant $C_a > 0$ such that

$$a_h(v_h, v_h) \geq C_a \, \|v_h\|_a^2 \quad \forall \; v_h \in V_h \,, \tag{35}$$

where $\|\cdot\|_a$ is a norm on the space $H_0^1(\Omega)$ but generally only a seminorm on the space $H^1(\Omega)$.

Now an approximate solution of the variational problem (34) can be introduced as the solution of the following finite-dimensional problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a_h(u_h, v_h) = \langle g, v_h \rangle \quad \forall \; v_h \in V_h \,. \tag{36}$$

We denote

$$a_{ij} = a_h(\varphi_j, \varphi_i) \,, \qquad i, j = 1, \ldots, N \,, \tag{37}$$

$$g_i = \langle g, \varphi_i \rangle \,, \qquad i = 1, \ldots, M \,, \tag{38}$$

$$u_i^b = u_b(x_i) \,, \qquad i = M+1, \ldots, N \,. \tag{39}$$

Then u_h is a solution of the finite-dimensional problem (36) if and only if it satisfies the relations (1) and (2). Moreover, the matrix $(a_{ij})_{i,j=1}^M$ satisfies (3). We denote

$$d_h(w; z, v) = \sum_{i,j=1}^N (1 - \alpha_{ij}(w)) \, d_{ij} \, (z(x_j) - z(x_i)) \, v(x_i) \quad \forall \; w, z, v \in C(\overline{\Omega}) \,,$$

with $\alpha_{ij}(w) := \alpha_{ij}(\{w(x_i)\}_{i=1}^N)$. This implies that

$$d_h(w_h; z_h, v_h) = \sum_{i,j=1}^N (1 - \alpha_{ij}(w_h)) \, d_{ij} \, (z_j - z_i) \, v_i \quad \forall \; w_h, z_h, v_h \in W_h \,,$$

and hence we realize that the corresponding flux correction scheme (13), (14) is equivalent to the following variational problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a_h(u_h, v_h) + d_h(u_h; u_h, v_h) = \langle g, v_h \rangle \quad \forall \; v_h \in V_h \,. \tag{40}$$

For any $w \in C(\overline{\Omega})$, the mapping $d_h(w;\cdot,\cdot) : C(\overline{\Omega}) \times C(\overline{\Omega}) \to \mathbb{R}$ is a nonnegative symmetric bilinear form (cf. Lemma 1), and hence it satisfies Schwarz's inequality

$$|d_h(w;z,v)|^2 \le d_h(w;z,z)\, d_h(w;v,v) \quad \forall\ w,z,v \in C(\overline{\Omega})\,. \tag{41}$$

Thus, for any $w \in C(\overline{\Omega})$, the functional $(d_h(w;\cdot,\cdot))^{1/2}$ is a seminorm on $C(\overline{\Omega})$.

Now let $u_h \in W_h$ be a solution of (40), and let us derive an estimate of the error $u - u_h$. A natural norm on V_h corresponding to the left-hand side of (40) is defined by

$$\|v_h\|_h := \Big(C_a\,\|v_h\|_a^2 + d_h(u_h;v_h,v_h)\Big)^{1/2}, \qquad v_h \in V_h\,.$$

Note that $\|\cdot\|_h$ may be only a seminorm on W_h and that it is not defined on the space $H^1(\Omega)$. We introduce the set

$$W_h^b = \{z_h \in W_h\,;\ z_h(x_i) = u_b(x_i),\ i = M+1,\dots,N\}$$

and consider any $v_h \in V_h$ and $z_h \in W_h^b$. Then, according to (34) and (40), one obtains

$$a_h(u_h - z_h, v_h) + d_h(u_h; u_h - z_h, v_h) = a(u,v_h) - a_h(z_h,v_h) - d_h(u_h;z_h,v_h)\,.$$

Since $u_h - z_h \in V_h$, using (35) and (41) one derives that

$$\|u_h - z_h\|_h \le \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(z_h,v_h)}{\|v_h\|_h} + (d_h(u_h;z_h,z_h))^{1/2}\,.$$

Assuming that $u \in C(\overline{\Omega})$, adding $\|u - z_h\|_h$ to both sides of this estimate and using the triangle inequality, one obtains

$$\|u - u_h\|_h \le \inf_{z_h \in W_h^b} \left\{\|u - z_h\|_h + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(z_h,v_h)}{\|v_h\|_h} + (d_h(u_h;z_h,z_h))^{1/2}\right\}. \tag{42}$$

Let us introduce the Lagrange interpolation operator $i_h : C(\overline{\Omega}) \to W_h$ by

$$i_h v = \sum_{i=1}^N v(x_i)\,\varphi_i\,, \qquad v \in C(\overline{\Omega})\,.$$

Then $i_h u \in W_h^b$, and hence using (42) one gets the estimate

$$\|u - u_h\|_h \le C_a^{1/2}\,\|u - i_h u\|_a + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} + (d_h(u_h;i_h u, i_h u))^{1/2}\,. \tag{43}$$

Thus, as usual, the error of the discrete solution is estimated by an interpolation error and a consistency error. In the following section we estimate these terms for a discretization of a convection-diffusion-reaction equation.

7. Application to a convection-diffusion-reaction equation. Let Ω be as in section 6, and let us consider the steady-state convection-diffusion-reaction equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = g \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega\,, \tag{44}$$

where $\varepsilon \in (0,\varepsilon_0)$ with $\varepsilon_0 < +\infty$ is a constant, and $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, and $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$ are given functions satisfying

$$\nabla\cdot\boldsymbol{b} = 0\,, \qquad c \ge \sigma_0 \ge 0 \qquad \text{in } \Omega\,,$$

where σ_0 is a constant. The weak solution of (44) satisfies (34) with

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + (c\,u, v) \quad \text{and} \quad \langle g, v\rangle = (g, v)\,,$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. It is well known that the weak solution of (44) exists, is unique, and satisfies the maximum principle (cf. [6]).

Let $\mathscr{T}_h$ belong to a regular family of triangulations of Ω consisting of simplices. We consider a space $W_h \subset H^1(\Omega)$ consisting of continuous piecewise linear functions, i.e.,

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in \mathbb{P}_1(T)\ \forall\, T \in \mathscr{T}_h\}\,.$$

The points x_i assigned to the basis functions φ_i introduced in the previous section are vertices of the triangulation $\mathscr{T}_h$.

The matrix corresponding to the reaction term $(c\,u_h, v_h)$ in the Galerkin finite element discretization of (44) has only nonnegative entries, which may cause a violation of the condition (6). In order to overcome this, we replace the matrix corresponding to the reaction term by a simple diagonal approximation:

$$(c\,u_h, v_h) = \sum_{i=1}^{M} (c\,u_h, \varphi_i)\,v_i \approx \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \quad \forall\ u_h \in W_h,\ v_h \in V_h\,. \tag{45}$$

This has the extra impact of making the matrix $\mathbb{D}$ independent of c (see below). An alternative diagonal approximation of the reaction matrix can be defined using a low-order nodal quadrature for the reaction term, in which case the estimation of the associated error follows standard approaches (provided that c has a higher regularity than the one assumed so far). The error incurred by the use of (45) is estimated in the next lemma.

LEMMA 13. *There is a constant C independent of h such that*

$$\left|(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i\right| \le C\,h\,\|c\|_{0,\infty,\Omega}\,|u_h|_{1,\Omega}\,\|v_h\|_{0,\Omega}$$

for all $c \in L^\infty(\Omega)$, $u_h \in W_h$, and $v_h \in V_h$.

Proof. Consider any $c \in L^\infty(\Omega)$, $u_h \in W_h$, and $v_h \in V_h$. Then

$$(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i = \sum_{i=1}^{M} (c\,(u_h - u_i), \varphi_i)\,v_i = \sum_{T\in\mathscr{T}_h} \sum_{\substack{i=1\\ x_i\in T}}^{M} (c\,(u_h - u_i), \varphi_i)_T\,v_i$$

$$\le \|c\|_{0,\infty,\Omega} \sum_{T\in\mathscr{T}_h} \sum_{\substack{i=1\\ x_i\in T}}^{M} \|u_h - u_i\|_{0,1,T}\,|v_i|\,.$$

Next, using the Cauchy–Schwarz inequality one obtains

$$\|u_h - u_i\|_{0,1,T} \le |T|^{1/2}\|u_h - u_i\|_{0,T} \le h_T^{d/2}\|\nabla u_h\cdot(x - x_i)\|_{0,T} \le h_T^{1+d/2}|u_h|_{1,T}\,,$$

where $h_T = \operatorname{diam}(T)$. Consequently,

$$(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \le\ h\,\|c\|_{0,\infty,\Omega} \sum_{T\in\mathscr{T}_h} |u_h|_{1,T}\,h_T^{d/2} \sum_{x_i\in T} |v_h(x_i)|\,.$$

Since $h_T^{d/2} \sum_{x_i \in T} |v_h(x_i)| \leq C \|v_h\|_{0,T}$, the lemma follows by applying Hölder's inequality. ∎

Using the approximation (45), the bilinear form a_h in (36) is given by

$$a_h(u_h, v_h) = \varepsilon (\nabla u_h, \nabla v_h) + (\boldsymbol{b} \cdot \nabla u_h, v_h) + \sum_{i=1}^{M} (c, \varphi_i)\, u_i\, v_i \quad \forall\ u_h \in W_h,\ v_h \in V_h$$

and satisfies (35), with

$$\|v\|_a^2 = \varepsilon\, |v|_{1,\Omega}^2 + \sigma_0\, \|v\|_{0,\Omega}^2,$$

and $C_a > 0$ independent of h and the data of (44). The bilinear form a_h defines the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, whose entries are given by (37). The artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is defined using (9), and thus it is independent of c.

Remark 14. It is easy to verify that the matrix $\mathbb{A}$ satisfies (21). The assumption (22) holds if and only if

$$(\nabla \varphi_k, \nabla \varphi_l) \leq 0 \quad \forall\ k, l = 1, \ldots, N,\ k \neq l\,,\ k \leq M,\ \text{or}\ l \leq M\,. \tag{46}$$

The validity of (46) is guaranteed if the triangulation $\mathscr{T}_h$ is weakly acute, i.e., if the angles between faces in $\mathscr{T}_h$ do not exceed $\pi/2$. In the two-dimensional case, it is sufficient for (46) that $\mathscr{T}_h$ is a Delaunay triangulation, i.e., that the sum of any pair of angles opposite a common edge is less than or equal to π.

Now we can discuss the estimation of the terms on the right-hand side of the error estimate (43). To this end, we assume that $u \in H^2(\Omega)$. Then, standard interpolation estimates (cf. [5]) give

$$\|u - i_h u\|_a \leq C\, (\varepsilon + \sigma_0\, h^2)^{1/2}\, h\, |u|_{2,\Omega}\,. \tag{47}$$

The remaining two terms on the right-hand side of (43) will be estimated in the following two lemmas.

LEMMA 15. *Let $\sigma_0 > 0$. Then there is a constant C independent of h and the data of problem* (44) *such that for any $u \in H^2(\Omega)$,*

$$\sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \leq C\, (\varepsilon + \sigma_0^{-1}\, \{\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\})^{1/2}\, h\, \|u\|_{2,\Omega}\,. \tag{48}$$

If $c \equiv 0$, then

$$\sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \leq C\, (\varepsilon + \varepsilon^{-1}\, \|\boldsymbol{b}\|_{0,\infty,\Omega}^2\, h^2)^{1/2}\, h\, |u|_{2,\Omega}\,. \tag{49}$$

Proof. Consider any $u \in H^2(\Omega)$ and $v_h \in V_h$. Then, in view of Lemma 13,

$$\begin{aligned} a(u, v_h) - a_h(i_h u, v_h) &= \varepsilon\, (\nabla(u - i_h u), \nabla v_h) + (\boldsymbol{b} \cdot \nabla(u - i_h u), v_h) \\ &\quad + (c\, (u - i_h u), v_h) + (c\, i_h u, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\, (i_h u)(x_i)\, v_i \\ &\leq C\, (\varepsilon\, |v_h|_{1,\Omega} + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega} + \|c\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega})\, h\, \|u\|_{2,\Omega}\,. \end{aligned}$$

Therefore, if $\sigma_0 > 0$, one obtains (48). If $c \equiv 0$, one can employ the fact that

$$(\boldsymbol{b} \cdot \nabla(u - i_h u), v_h) = -(u - i_h u, \boldsymbol{b} \cdot \nabla v_h) \leq C\, h^2\, |u|_{2,\Omega}\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, |v_h|_{1,\Omega}\,,$$

which leads to (49). ∎

Lemma 15 shows that, if $\sigma_0 > 0$, one obtains from (43)

$$\|u - u_h\|_h \leq C\, h\, \|u\|_{2,\Omega} + (d_h(u_h; i_h u, i_h u))^{1/2}\,, \tag{50}$$

where C is independent of u, h, and ε. However, if $c \equiv 0$ (hence $\sigma_0 = 0$), one cannot avoid an explicit negative power of ε in the estimate (49) since the seminorm $(d_h(u_h; v_h, v_h))^{1/2}$ cannot be used for estimating v_h due to the possibly vanishing factors $(1 - \alpha_{ij}(u_h))$. The negative power of ε in (49) is somewhat compensated by the presence of h in the numerator. Still, this estimate can be considered fully satisfactory only if $h \lesssim \varepsilon^{1/2}$.

Finally, let us estimate the last term on the right-hand side of (43).

LEMMA 16. *Let the matrix $\mathbb{D}$ be defined by* (9)*. Then there is a constant C independent of h and the data of problem* (44) *such that*

$$d_h(w_h; i_h u, i_h u) \leq C\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)\, |i_h u|_{1,\Omega}^2 \quad \forall\ w_h \in W_h,\ u \in C(\overline{\Omega})\,. \tag{51}$$

Proof. Consider any $i, j \in \{1, \ldots, N\}$ such that $i \neq j$ and $d_{ij} \neq 0$. Then

$$\begin{aligned} |d_{ij}| &\leq \sum_{T \in \mathscr{T}_h,\, x_i, x_j \in T} \left(\varepsilon\, |\varphi_i|_{1,T}\, |\varphi_j|_{1,T} + \|\boldsymbol{b}\|_{0,\infty,T}\, \{|\varphi_i|_{1,T}\, \|\varphi_j\|_{0,T} + |\varphi_j|_{1,T}\, \|\varphi_i\|_{0,T}\}\right) \\ &\leq C \sum_{T \in \mathscr{T}_h,\, x_i, x_j \in T} \left(\varepsilon\, h_T^{d-2} + \|\boldsymbol{b}\|_{0,\infty,T}\, h_T^{d-1}\right) \leq \widetilde{C}\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)\, |x_i - x_j|^{d-2}\,. \end{aligned}$$

Therefore, using Lemma 1, one derives for any $w_h \in W_h$ and $u \in C(\overline{\Omega})$

$$\begin{aligned} d_h(w_h; i_h u, i_h u) &= \sum_{\substack{i,j=1 \\ i<j}}^{N} (1 - \alpha_{ij}(w_h))\, |d_{ij}|\, [u(x_i) - u(x_j)]^2 \\ &\leq \sum_{T \in \mathscr{T}_h} \sum_{x_i, x_j \in T} |d_{ij}|\, [u(x_i) - u(x_j)]^2 \\ &\leq \widetilde{C}\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h) \sum_{T \in \mathscr{T}_h} h_T^{d-2} \sum_{x_i, x_j \in T} [u(x_i) - u(x_j)]^2\,. \end{aligned}$$

Since

$$h_T^{d-2} \sum_{x_i, x_j \in T} [u(x_i) - u(x_j)]^2 \leq C\, |i_h u|_{1,T}^2\,,$$

one obtains the statement of the lemma. ∎

One observes that if $d_h(u_h; i_h u, i_h u)$ in (50) is estimated using Lemma 16, the convergence order is reduced. As a matter of fact, (47), (48), and (51) lead to the following global error estimate.

COROLLARY 17. *Let $u \in H^2(\Omega)$ be the solution of* (44)*, and let u_h be a solution of the discrete problem* (40)*. Then if $\sigma_0 > 0$, there exists a constant $C > 0$ independent of h and the data of* (44) *such that*

$$\begin{aligned} \|u - u_h\|_h \leq\ & C\,(\varepsilon + \sigma_0^{-1}\, \{\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\} + \sigma_0 h^2)^{1/2}\, h\, \|u\|_{2,\Omega} \\ & + C\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)^{1/2}\, |i_h u|_{1,\Omega}\,. \end{aligned}$$

Remark 18. A careful inspection of the proof of Lemma 16 reveals that the convergence order of the term $d_h(u_h; i_h u, i_h u)$ depends on the relation between ε and $\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$ and on properties of the triangulations $\mathscr{T}_h$. For simplicity, the discussion will be restricted to the two-dimensional case, but the same arguments are valid (with minor modifications) in the higher-dimensional case. We distinguish the following cases:

- *convection-dominated regime* ($\varepsilon < \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$): the estimate (51) reduces to

$$d_h(w_h; i_h u, i_h u) \le C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |i_h u|^2_{1,\Omega} \quad \forall\, w_h \in W_h,\, u \in C(\overline{\Omega})\,. \tag{52}$$

This estimate implies an $\mathcal{O}(\sqrt{h})$ error estimate in (50), which will be confirmed by numerical experiments in section 8 for a particular choice of the coefficients α_{ij}.

- *diffusion-dominated regime* ($\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$). In this case, the estimate (51) reduces to

$$d_h(w_h; i_h u, i_h u) \le C\, \varepsilon\, |i_h u|^2_{1,\Omega} \quad \forall\, w_h \in W_h,\, u \in C(\overline{\Omega})\,, \tag{53}$$

which does not imply any convergence of $\|u - u_h\|_h$. However, this result can be improved for suitable types of meshes. To characterize the geometry of a triangulation $\mathscr{T}_h$, we introduce a quantity θ_{ij} for any edge E_{ij} with end points x_i, x_j. If $E_{ij} \subset \partial\Omega$, then θ_{ij} is the angle opposite E_{ij}. If $E_{ij} \not\subset \partial\Omega$, then θ_{ij} is the average of the pair of angles opposite E_{ij}. Finally, we denote by θ_h the maximum of all θ_{ij}. Then we consider the following values of θ_h:

(a) $\theta_h \le \pi/2$, i.e., $\mathscr{T}_h$ is a *Delaunay triangulation* (in particular, $\mathscr{T}_h$ may consist of *weakly acute triangles*, i.e., with all angles $\le \pi/2$). Then the off-diagonal entries of the diffusion matrix are all nonpositive, and hence $|d_{ij}| \le \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h/3$ for $i \ne j$. Thus, the estimate (52) is again valid and leads to an $\mathcal{O}(\sqrt{h})$ in estimate (50).

(b) $\theta_h < \pi/2$, a particular case of (a), satisfied, e.g., for $\mathscr{T}_h$ consisting of *acute triangles* (all angles $< \pi/2$). Then all off-diagonal entries of the diffusion matrix are negative, and hence all off-diagonal entries of the matrix $\mathbb{A}$ are nonpositive in the strongly diffusion-dominated case (precisely, if $\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, (\tan\theta_h)/3$). In this case, all entries of the artificial diffusion matrix $\mathbb{D}$ vanish, and hence the AFC method (40) reduces to the original linear method (36). Consequently, the standard $\mathcal{O}(h)$ error estimate of $\|u - u_h\|_h$ is valid.

(c) $\theta_h = \pi/2$, again a particular case of (a) which may happen, e.g., if $\mathscr{T}_h$ consists of right-angled triangles. Then some off-diagonal entries of the diffusion matrix vanish, and hence the corresponding entries d_{ij} do not vanish in general. Thus, if $\theta_h = \pi/2$ for all $\mathscr{T}_h$ in the family of triangulations, then, in contrast to the previous case, the AFC method (40) does not reduce to the original linear method (36) for $h \to 0$.

(d) $\theta_h > \pi/2$, i.e., $\mathscr{T}_h$ is *not of Delaunay type*, which implies that $\mathscr{T}_h$ contains *obtuse triangles* (with an angle $> \pi/2$). In this case, some off-diagonal entries of the diffusion matrix are positive, and hence the estimate (53) cannot be improved in general. Indeed, if $\theta_{ij} > \pi/2$ and $\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |\tan\theta_{ij}|$, then $|d_{ij}| \ge \varepsilon\, |\cot\theta_{ij}|/3$. Thus, if the mesh is not of Delaunay type, the results presented in this work do not prove convergence of the method, which will be also confirmed by numerical experiments presented in section 8. Note also that, in this case, the results of section 5 are not valid for the AFC scheme considered in this section.

It is worth remarking that these last results are the best that can be obtained using the general approach described in the previous sections, combined with the choice for limiters α_{ij} from section 4. As a matter of fact, the algebraic construction of the method has been carried out using a rather general splitting of the stiffness matrix. Now, for the convection-diffusion equation, the lack of convergence of the method for non-Delaunay meshes can be overcome by changing the way the matrix $\mathbb{D}$ is built. In fact, if instead of using the whole stiffness matrix to build $\mathbb{D}$, we use only the convection matrix to build it, that is,

$$d_{ij} = -\max\{(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i), 0, (\boldsymbol{b}\cdot\nabla\varphi_i,\varphi_j)\} \quad \forall\, i\neq j\,; \tag{54}$$

then the estimate (51) in Lemma 16 becomes

$$d_h(w_h; i_h u, i_h u) \leq C\,\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |i_h u|^2_{1,\Omega} \quad \forall\, w_h\in W_h,\, u\in C(\overline{\Omega})\,.$$

This leads to an $\mathcal{O}(\sqrt{h})$ estimate of $\|u-u_h\|_h$, even on non-Delaunay meshes in the diffusion-dominated regime. An alternative way to solve this would be to change the definition of the limiters α_{ij} to make them more suitable for diffusion problems. Examples of limiters suitable for diffusion problems can be found, e.g., in [8, 19], but their applicability to convection-dominated problems has yet to be explored.

We finally mention that numerical results in section 8 indicate that the estimates of $d_h(w_h; i_h u, i_h u)$ discussed above are sharp. Note, however, that the only properties of the coefficients α_{ij} used in the proof of Lemma 16 were the fact that their values are from the interval $[0,1]$ and that $\alpha_{ij}=\alpha_{ji}$. If the coefficients α_{ij} are defined as in section 4, then in the convection-dominated regime, better convergence rates are observed than those predicted by estimate (52). Some deeper analysis of this choice of α_{ij} might lead to an improved estimate of $d_h(w_h; i_h u, i_h u)$ in the convection-dominated case.

Remark 19. We finish this section by making some comments on the stability of the nonlinear discretization (40) with W_h defined in section 7. Our objective is to show that this formulation can be viewed as a way of adding numerical diffusion to the Galerkin discretization. We restrict our discussion to the two-dimensional case, but the results can be extended to three space dimensions. First, given $u_h\in W_h$, we divide the triangulation $\mathscr{T}_h$ as $\mathscr{T}_h=\mathscr{T}_1\cup\mathscr{T}_2$, where $\mathscr{T}_1$ and $\mathscr{T}_2$ are disjoint and $T\in\mathscr{T}_1$ if and only if for at least two edges of T we have $(1-\alpha_{ij}(u_h))|d_{ij}|>0$. We will denote by α_T the minimum value of these nonzero quantities. Typically, T will belong to $\mathscr{T}_1$ if there is an extremum of u_h in a vertex of T or if u_h has a layer through T. Then from the proof of Lemma 1, and using a scaling argument, it is not difficult to realize that for any $v_h\in W_h$,

$$\begin{aligned} d_h(u_h; v_h, v_h) &= \frac{1}{2}\sum_{i,j=1}^{N} (1-\alpha_{ij}(u_h))\,|d_{ij}|\,(v_i-v_j)^2 \\ &\geq \frac{1}{12}\sum_{T\in\mathscr{T}_1}\sum_{x_i,x_j\in T}\alpha_T\,(v_i-v_j)^2 \geq C\sum_{T\in\mathscr{T}_1}\alpha_T\,|v_h|^2_{1,T}\,. \end{aligned}$$

Note that for simplicity, we used the inequality

$$(v_i-v_j)^2+(v_j-v_k)^2 \geq \frac{1}{3}\left((v_i-v_j)^2+(v_j-v_k)^2+(v_k-v_i)^2\right) \quad \forall\, i,j,k\,.$$

Then, AFC methods add numerical diffusion on certain elements of the triangulation, namely, the elements which contain extrema of the discrete solution or lie in its layer regions.

In addition, we can also compare this last result with a parameter-free stabilized method proposed in [3]. That method is based on rewriting the gradient of the $\mathbb{P}_1$ basis functions in terms of the Nédélec edge FEM. More precisely, the stabilization term added to the Galerkin formulation in [3] reads as follows:

$$Q(u_h, v_h) = (\Theta_h(u_h), \Theta_h(v_h))\,, \tag{55}$$

with

$$\Theta_h(u_h) = \sum_{E\in\mathcal{E}_h} \tilde{\theta}_E\,(u_h(x_{E1}) - u_h(x_{E2}))\,\boldsymbol{N}_E\,, \tag{56}$$

where $\mathcal{E}_h$ stands for the set of edges of the triangulation $\mathscr{T}_h$, x_{E1}, x_{E2} are the end points of an edge E, and $\boldsymbol{N}_E$ stands for the basis function of the Nédélec space associated to E. In (56), $\tilde{\theta}_E$ is a positive parameter depending on the edge Péclet number (for details, see [3, eqs. (2.14) and (2.10)]). With these definitions, the term defined in (55) satisfies

$$\begin{aligned} Q(u_h, u_h) &= \sum_{E,E'\in\mathcal{E}_h} \tilde{\theta}_E\,\tilde{\theta}_{E'}\,(u_h(x_{E1}) - u_h(x_{E2}))\,(u_h(x_{E'1}) - u_h(x_{E'2}))\,(\boldsymbol{N}_E, \boldsymbol{N}_{E'}) \\ &\approx \sum_{E\in\mathcal{E}_h} |E|^{d-2}\big(\tilde{\theta}_E\,(u_h(x_{E1}) - u_h(x_{E2}))\big)^2\,, \end{aligned}$$

where by $\approx$ we mean that both terms bound each other with constants that do not depend on h. Then we see that the method from [3] can be seen as well as a "linearized" version of (40) (where we choose α_{ij} in such a way that $(1-\alpha_{ij}(u_h))|d_{ij}| = \tilde{\theta}_E^2\,|E|^{d-2}$ for every edge E). This also explains the fact that only $\mathcal{O}(\sqrt{h})$ convergence has been obtained in Table 1 (where we choose $\alpha_{ij}(u_h) = 0.5$ for every edge). As a matter of fact, that was the order of convergence proven in [3].

8. Numerical results. This section presents numerical results obtained with the AFC scheme applied to the convection-diffusion-reaction equation (44). For the sake of brevity, the presentation is restricted to studies of the convergence of the method for the following example with smooth solution. Results for an example with layers can be found, e.g., in [1].

Example 20. Problem (44) is considered with $\Omega = (0,1)^2$, with different values of ε, and with $\boldsymbol{b} = (3,2)^T$, $c = 1$, $u_b = 0$, and the right-hand side g chosen such that

$$u(x,y) = 100\,x^2\,(1-x)^2\,y\,(1-y)\,(1-2y)$$

is the solution of (44).

In the numerical simulations, $\mathbb{P}_1$ finite elements were used on triangular grids. Mass lumping (cf. (45)) was performed for the reactive term, but only very small differences could be observed compared to results obtained without mass lumping. If x_i is a Dirichlet node, we set $R_i^+ := 1$, $R_i^- := 1$, leading to $\alpha_{ij} = 1$ if $a_{ji} \le a_{ij}$; see section 4. Concerning the errors in $\|\cdot\|_h$, qualitatively the same results were obtained with and without this definition. However, the errors in other norms of interest were sometimes clearly smaller with this definition, and we decided to present these better

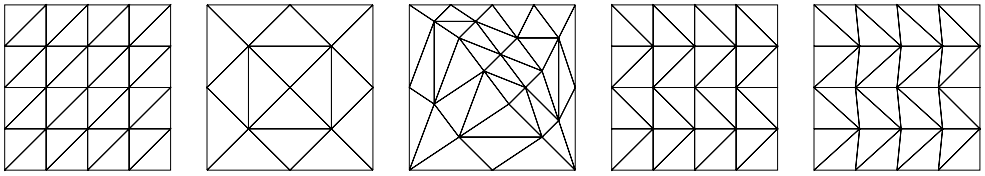


FIG. 1. *Grids* 1–5 *(left to right), level* 0. *The differences between grid* 4 *and grid* 5 *are described in the text.*

results. The nonlinear discrete equations were solved with a fixed-point iteration with Anderson acceleration [21]. The iterations were stopped when the Euclidean norm of the residual vector was smaller than 10^{-9}. All simulations were double-checked by computing them with two different codes, one of which was MooNMD [9].

Simulations were performed on several structured and unstructured grids; see Figure 1 for the coarsest grids (level 0). Grids 1, 2, and 3 were refined uniformly. Grid 4 was obtained from grid 1 by changing the directions of the diagonals in even rows of squares (from below). Grid 5 was obtained from grid 4 by shifting interior nodes to the right by a tenth of the horizontal mesh width on each even horizontal mesh line. Therefore, for any diagonal edge E_{ij} of grid 5, the value θ_{ij} introduced in Remark 18 satisfies $\theta_{ij} > \pi/2$.

Considering a problem without reaction, i.e., with $c = 0$ instead of $c = 1$, and otherwise the same setup, one obtains qualitatively the same results as below. For the sake of brevity, we omit the results for $c = 0$.

8.1. Constant weights α_{ij}. The case of constant weights $\alpha_{ij} = 0.5$ (with the modification at Dirichlet nodes mentioned above) fits into the presented error analysis. Fixing the weights independently of the approximate solution u_h replaces the nonlinear problem (13), (14) by a linear problem, which is essentially a stabilized method adding first-order artificial diffusion to the original problem (1), (2). Then, some suboptimal convergence results are to be expected. Table 1 shows numerical results obtained in the convection-dominated regime for grid 1. In the first row of the table, we use the following notation: l is the grid level, $e_h = u - u_h$, $d_h^{1/2}(u_h) = d_h(u_h; i_h u, i_h u)^{1/2}$, and "ord." denotes experimental convergence orders computed from values in the preceding column. The results in Table 1 indicate that the estimate (52) of $d_h(w_h; i_h u, i_h u)$ and also the estimate for $\|u - u_h\|_h$ given in Corollary 17 are sharp.

TABLE 1
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 5 *and constant weights* α_{ij}.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.622e−2	0.66	7.668e−1	0.11	2.722e−1	0.43	9.666e−2	0.57
4	1.527e−2	0.78	7.021e−1	0.13	1.975e−1	0.46	6.397e−2	0.60
5	8.260e−3	0.89	6.489e−1	0.11	1.415e−1	0.48	4.274e−2	0.58
6	4.295e−3	0.94	6.149e−1	0.08	1.008e−1	0.49	2.912e−2	0.55
7	2.189e−3	0.97	5.956e−1	0.05	7.150e−2	0.50	2.015e−2	0.53
8	1.105e−3	0.99	5.854e−1	0.02	5.065e−2	0.50	1.408e−2	0.52

8.2. Weights computed with the algorithm from section 4. As already mentioned, the computation of the weights as presented in section 4 is a standard choice in practice. For the convection-dominated regime, numerical results are

presented in Tables 2–6. It can be observed that the order of convergence of $\|u-u_h\|_h$ is around two on grid 1 and around one for all other simulations. The errors $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ behave differently on different grids. For grid 1, which is of Friedrichs–Keller type (it consists of three sets of parallel lines), one can see the optimal order of convergence for $\|u-u_h\|_{0,\Omega}$ and also the convergence of $|u-u_h|_{1,\Omega}$ is almost optimal. For grids 2–5, the orders of convergence of $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ are clearly smaller than the optimal order. Moreover, for grids 4 and 5, the convergence order of $|u-u_h|_{1,\Omega}$ tends to zero for $h \to 0$.

TABLE 2
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 1 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	5.457e−3	1.85	2.287e−1	1.10	1.112e−1	0.97	1.163e−2	2.11
4	1.408e−3	1.95	1.074e−1	1.09	5.317e−2	1.06	2.683e−3	2.12
5	3.493e−4	2.01	5.113e−2	1.07	2.472e−2	1.11	6.410e−4	2.07
6	8.652e−5	2.01	2.546e−2	1.01	1.158e−2	1.09	1.633e−4	1.97
7	2.152e−5	2.01	1.321e−2	0.95	5.533e−3	1.07	4.099e−5	1.99
8	5.357e−6	2.01	6.822e−3	0.95	2.685e−3	1.04	1.018e−5	2.01

TABLE 3
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 2 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	8.533e−3	1.86	2.901e−1	1.00	1.236e−1	1.03	1.855e−2	1.91
4	2.516e−3	1.76	1.954e−1	0.57	5.884e−2	1.07	6.065e−3	1.61
5	8.369e−4	1.59	1.380e−1	0.50	2.801e−2	1.07	2.640e−3	1.20
6	2.891e−4	1.53	1.031e−1	0.42	1.356e−2	1.05	1.254e−3	1.07
7	1.103e−4	1.39	7.865e−2	0.39	6.638e−3	1.03	5.938e−4	1.08
8	4.136e−5	1.42	6.524e−2	0.27	3.263e−3	1.02	2.924e−4	1.02
9	1.539e−5	1.43	5.768e−2	0.18	1.618e−3	1.01	1.436e−4	1.03

In summary, in the convection-dominated regime, the numerical studies for the choice of the weights as presented in section 4 show a higher order of error reduction than in the worst case which was considered in the analysis. The difference with respect to the numerical studies of section 8.1 is the behavior of the weights. They do not stay constant but converge in the mean to 1; see Table 7 which shows a representative result for the arithmetic mean value of $\{1-\alpha_{ij}(u_h)\}$. This indicates that the estimate $1-\alpha_{ij}(u_h) \le 1$ used in the proof of Lemma 16 is too rough in some cases.

For the diffusion-dominated regime, numerical results are presented in Tables 8–10. For grid 1, the convergence orders of $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ are again optimal, but for grid 4 only $|u-u_h|_{1,\Omega}$ is still optimal, whereas $d_h(u_h; i_h u, i_h u)^{1/2}$ converges with the order 1/2. For grid 5, no convergence is observed. The observations with respect to convergence orders of $d_h(u_h; i_h u, i_h u)^{1/2}$ on grids 4 and 5 are in accordance with the discussion in Remark 18. If the matrix $\mathbb{D}$ is defined using the convection matrix only (i.e., by (54)), then on grids 1 and 4 the results qualitatively do not change, whereas on grid 5, we observe an analogous behavior as on grid 4; see Table 11.

TABLE 4
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 3 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.125e−3	1.61	3.202e−1	0.71	9.189e−2	1.05	1.569e−2	1.81
4	2.216e−3	1.47	2.244e−1	0.51	4.488e−2	1.03	6.502e−3	1.27
5	9.946e−4	1.16	1.821e−1	0.30	2.224e−2	1.01	3.376e−3	0.95
6	4.993e−4	0.99	1.559e−1	0.22	1.124e−2	0.98	1.802e−3	0.91
7	2.519e−4	0.99	1.375e−1	0.18	5.676e−3	0.98	9.649e−4	0.90
8	1.277e−4	0.98	1.231e−1	0.16	2.871e−3	0.98	5.099e−4	0.92

TABLE 5
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 4 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.383e−3	1.70	4.826e−1	0.31	9.814e−2	1.06	2.143e−2	1.45
4	2.313e−3	1.46	4.543e−1	0.09	4.341e−2	1.18	9.455e−3	1.18
5	1.089e−3	1.09	4.434e−1	0.03	1.830e−2	1.25	4.469e−3	1.08
6	5.527e−4	0.98	4.361e−1	0.02	8.276e−3	1.14	2.176e−3	1.04
7	2.817e−4	0.97	4.320e−1	0.01	3.926e−3	1.08	1.077e−3	1.01
8	1.425e−4	0.98	4.297e−1	0.01	1.915e−3	1.04	5.381e−4	1.00

TABLE 6
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 5 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.925e−3	1.66	5.638e−1	0.25	9.992e−2	1.06	2.486e−2	1.37
4	2.687e−3	1.37	5.395e−1	0.06	4.405e−2	1.18	1.140e−2	1.12
5	1.304e−3	1.04	5.294e−1	0.03	1.896e−2	1.22	5.491e−3	1.05
6	6.645e−4	0.97	5.225e−1	0.02	8.792e−3	1.11	2.711e−3	1.02
7	3.382e−4	0.97	5.186e−1	0.01	4.235e−3	1.05	1.349e−3	1.01
8	1.708e−4	0.99	5.164e−1	0.01	2.083e−3	1.02	6.755e−4	1.00

TABLE 7
Example 20, $\varepsilon = 10^{-8}$, *grid* 1, *arithmetic mean of* $\{1 - \alpha_{ij}(u_h)\}$ *with* α_{ij} *from section* 4.

Level	3	4	5	6	7	8
$1 - \overline{\alpha}(u_h)$	1.09e−1	5.94e−2	3.16e−2	1.73e−2	9.60e−3	5.27e−3
Order	0.83	0.87	0.91	0.87	0.85	0.87

9. Summary and outlook. An algebraic flux correction (AFC) scheme applied to linear boundary value problems was analyzed. The existence of a solution, existence and uniqueness of a solution of a linearized problem, and an a priori error estimate were proved under rather general assumptions on the limiters α_{ij}. To the best of our knowledge, this is the first time that convergence analysis of an AFC scheme was performed. For a practical choice of the limiters, a local discrete maximum principle was proved. The theory for the abstract problem was applied to steady-state convection-diffusion-reaction equations, where in particular an error estimate was derived. Numerical studies showed that this estimate is sharp for the general assumptions on the limiters used in the analysis. Using the standard limiters, a higher order of convergence was observed than predicted.

As a next step we intend to specialize the convergence results to the standard limiters. This step requires an analysis of the algorithm presented in section 4, which seems to be intricate due to the dependency of the limiters on the solution of the discrete problem. From the numerical aspect, the observed dependency of errors in

TABLE 8
Example 20, $\varepsilon = 10$, *numerical results for grid* 1 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.148e−3	1.98	1.757e−1	0.99	1.144e−1	1.00	5.557e−1	0.99
4	5.379e−4	2.00	8.799e−2	1.00	5.643e−2	1.02	2.783e−1	1.00
5	1.345e−4	2.00	4.401e−2	1.00	2.792e−2	1.02	1.392e−1	1.00
6	3.360e−5	2.00	2.201e−2	1.00	1.387e−2	1.01	6.960e−2	1.00
7	8.398e−6	2.00	1.100e−2	1.00	6.912e−3	1.00	3.480e−2	1.00

TABLE 9
Example 20, $\varepsilon = 10$, *numerical results for grid* 4 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.187e−3	1.89	1.756e−1	0.99	1.983e−1	0.37	5.554e−1	0.99
4	6.209e−4	1.82	8.800e−2	1.00	1.473e−1	0.43	2.783e−1	1.00
5	1.940e−4	1.68	4.402e−2	1.00	1.069e−1	0.46	1.392e−1	1.00
6	6.899e−5	1.49	2.201e−2	1.00	7.657e−2	0.48	6.961e−2	1.00
7	2.789e−5	1.31	1.101e−2	1.00	5.450e−2	0.49	3.481e−2	1.00
8	1.239e−5	1.17	5.503e−3	1.00	3.867e−2	0.50	1.740e−2	1.00

TABLE 10
Example 20, $\varepsilon = 10$, *numerical results for grid* 5 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	1.248e−2	0.48	2.229e−1	0.79	1.317e+0	-0.03	7.211e−1	0.77
4	1.123e−2	0.15	1.558e−1	0.52	1.316e+0	0.00	5.135e−1	0.49
5	1.090e−2	0.04	1.333e−1	0.22	1.313e+0	0.00	4.452e−1	0.21
6	1.080e−2	0.01	1.269e−1	0.07	1.312e+0	0.00	4.259e−1	0.06
7	1.077e−2	0.00	1.252e−1	0.02	1.311e+0	0.00	4.207e−1	0.02
8	1.076e−2	0.00	1.248e−1	0.00	1.310e+0	0.00	4.193e−1	0.00

TABLE 11
Example 20, $\varepsilon = 10$, *numerical results for grid* 5, α_{ij} *from section* 4, *and* d_{ij} *defined by* (54) *instead of* (9).

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.319e−3	1.94	1.849e−1	0.98	1.581e−1	0.74	5.846e−1	0.98
4	6.098e−4	1.93	9.275e−2	1.00	1.040e−1	0.60	2.933e−1	1.00
5	1.676e−4	1.86	4.642e−2	1.00	7.244e−2	0.52	1.468e−1	1.00
6	4.979e−5	1.75	2.322e−2	1.00	5.105e−2	0.50	7.343e−2	1.00
7	1.659e−5	1.59	1.161e−2	1.00	3.607e−2	0.50	3.672e−2	1.00
8	6.302e−6	1.40	5.806e−3	1.00	2.550e−2	0.50	1.836e−2	1.00

standard norms on the concrete grid is remarkable. Comprehensive numerical studies that clarify which types of grids should be used and which types should be avoided are necessary, and this will be the subject of future research.

Appendix. For completeness, we report the proofs of some classical results on the relation between M-matrices and discrete maximum principles.

LEMMA 21. *Let us consider a matrix* $(a_{ij})_{j=1,\dots,N}^{i=1,\dots,M}$ *with* $0 < M < N$, *and let* $a_{ii} > 0$ *for* $i = 1, \dots, M$. *Then* (5) *holds for any* $u_1, \dots, u_N \in \mathbb{R}$ *if and only if the conditions* (6) *and* (8) *are satisfied.*

Proof. Let us assume that at least one of the conditions (6) and (8) is not valid. We will construct a counterexample to the validity of (5). If (6) does not hold, i.e., if

$a_{ik} > 0$ for some $i \in \{1, \ldots, M\}$ and $k \in \{1, \ldots, N\}$, $k \neq i$, then we set

$$u_i = 1, \qquad u_k = -\frac{a_{ii}}{a_{ik}}, \qquad u_j = 0 \quad \forall\, j \in \{1, \ldots, N\},\, j \neq i, k\,.$$

Then $u_k < 0$, and hence $\max\{u_j^+ \,;\, j \neq i,\, a_{ij} \neq 0\} = 0 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = a_{ii}\, u_i + a_{ik}\, u_k = 0$ so that (5) does not hold. If (8) is not valid, i.e., if $\sum_{j=1}^N a_{ij} < 0$ for some $i \in \{1, \ldots, M\}$, then we set

$$u_i = 1 - \frac{1}{a_{ii}} \sum_{j=1}^N a_{ij}, \qquad u_j = 1 \quad \forall\, j \in \{1, \ldots, N\},\, j \neq i\,.$$

Then $\max\{u_j^+ \,;\, j \neq i,\, a_{ij} \neq 0\} = 1 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = \sum_{j=1}^N a_{ij} + a_{ii}\,(u_i - 1) = 0$ so that again (5) does not hold. This proves that the validity of (5) for any $u_1, \ldots, u_N \in \mathbb{R}$ implies (6) and (8).

Now let us assume that the conditions (6) and (8) are satisfied. Consider any $i \in \{1, \ldots, M\}$ and any $u_1, \ldots, u_N \in \mathbb{R}$ such that $\sum_{j=1}^N a_{ij}\, u_j \leq 0$. Setting

$$c := \max_{j \neq i,\, a_{ij} \neq 0} u_j^+\,,$$

one has

$$\begin{aligned} a_{ii}\, u_i &\leq \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\, u_j = \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\,(u_j - c) + \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\, c \\ &\leq c \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij}) \leq c\, a_{ii}\,, \end{aligned} \tag{57}$$

which implies that $u_i \leq c$. ∎

LEMMA 22. *Let us consider a matrix $(a_{ij})_{j=1,\ldots,N}^{i=1,\ldots,M}$ with $0 < M < N$, and let $a_{ii} > 0$ for $i = 1, \ldots, M$. Then* (4) *holds for any $u_1, \ldots, u_N \in \mathbb{R}$ if and only if the conditions* (6) *and* (7) *are satisfied.*

Proof. Let us assume that at least one of the conditions (6) and (7) is not valid. Since the counterexamples from the proof of Lemma 21 can be used also here, it suffices to consider the case when $\sum_{j=1}^N a_{ij} > 0$ for some $i \in \{1, \ldots, M\}$. We set

$$u_i = -1 + \frac{1}{a_{ii}} \sum_{j=1}^N a_{ij}, \qquad u_j = -1 \quad \forall\, j \in \{1, \ldots, N\},\, j \neq i\,.$$

Then $\max\{u_j \,;\, j \neq i,\, a_{ij} \neq 0\} = -1 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = -\sum_{j=1}^N a_{ij} + a_{ii}\,(u_i + 1) = 0$ so that (4) does not hold. This proves that the validity of (4) for any $u_1, \ldots, u_N \in \mathbb{R}$ implies (6) and (7).

Now let us assume that the conditions (6) and (7) are satisfied. Consider any $i \in \{1, \ldots, M\}$ and any $u_1, \ldots, u_N \in \mathbb{R}$ such that $\sum_{j=1}^N a_{ij}\, u_j \leq 0$. Setting

$$c := \max_{j \neq i,\, a_{ij} \neq 0} u_j\,,$$

statement (57) remains valid (the last $\leq$ can be changed to $=$), and hence $u_i \leq c$. ∎

REFERENCES

[1] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla, *An assessment of discretizations for convection-dominated convection-diffusion equations*, Comput. Methods Appl. Mech. Engrg., 200 (2011), pp. 3395–3409.

[2] G. R. Barrenechea, V. John, and P. Knobloch, *Some analytical results for an algebraic flux correction scheme for a steady convection-diffusion equation in one dimension*, IMA J. Numer. Anal., 35 (2015), pp. 1729–1756.

[3] P. Bochev, M. Perego, and K. Peterson, *Formulation and analysis of a parameter-free stabilized finite element method*, SIAM J. Numer. Anal., 53 (2015), pp. 2363–2388, doi:10.1137/14096284X.

[4] R. Bordás, V. John, E. Schmeyer, and D. Thévenin, *Numerical methods for the simulation of a coalescence-driven droplet size distribution*, Theor. Comput. Fluid Dyn., 27 (2013), pp. 253–271.

[5] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, New York, 2004.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Sciences] 224, Springer-Verlag, Berlin, 1983.

[7] M. Gurris, D. Kuzmin, and S. Turek, *Implicit finite element schemes for the stationary compressible Euler equations*, Internat. J. Numer. Methods Fluids, 69 (2012), pp. 1–28.

[8] W. Hundsdorfer and C. Montijn, *A note on flux limiting for diffusion discretizations*, IMA J. Numer. Anal., 24 (2004), pp. 635–642.

[9] V. John and G. Matthies, *MooNMD—a program package based on mapped finite element methods*, Comput. Vis. Sci., 6 (2004), pp. 163–169.

[10] V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska, and A. Voigt, *Simulations of population balance systems with one internal coordinate using finite element methods*, Chem. Engrg. Sci., 64 (2009), pp. 733–741.

[11] V. John and M. Roland, *On the impact of the scheme for solving the higher dimensional equation in coupled population balance systems*, Internat. J. Numer. Methods Engrg., 82 (2010), pp. 1450–1474.

[12] D. Kuzmin, *Private communication.*

[13] D. Kuzmin, *On the design of general-purpose flux limiters for finite element schemes.* I. *Scalar convection*, J. Comput. Phys., 219 (2006), pp. 513–531.

[14] D. Kuzmin, *Algebraic flux correction for finite element discretizations of coupled systems*, in Proceedings of the International Conference on Computational Methods for Coupled Problems in Science and Engineering, M. Papadrakakis, E. Oñate, and B. Schrefler, eds., CIMNE, Barcelona, 2007, pp. 1–5.

[15] D. Kuzmin, *On the design of algebraic flux correction schemes for quadratic finite elements*, J. Comput. Appl. Math., 218 (2008), pp. 79–87.

[16] D. Kuzmin, *Explicit and implicit FEM-FCT algorithms with flux linearization*, J. Comput. Phys., 228 (2009), pp. 2517–2534.

[17] D. Kuzmin, *Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes*, J. Comput. Appl. Math., 236 (2012), pp. 2317–2337.

[18] D. Kuzmin and M. Möller, *Algebraic flux correction* I. *Scalar conservation laws*, in Flux-Corrected Transport. Principles, Algorithms, and Applications, D. Kuzmin, R. Löhner, and S. Turek, eds., Springer-Verlag, Berlin, 2005, pp. 155–206.

[19] D. Kuzmin, M. J. Shashkov, and D. Svyatskiy, *A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems*, J. Comput. Phys., 228 (2009), pp. 3448–3463.

[20] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North–Holland, Amsterdam, 1977.

[21] H. F. Walker and P. Ni, *Anderson acceleration for fixed-point iterations*, SIAM J. Numer. Anal., 49 (2011), pp. 1715–1735, doi:10.1137/10078356X.

[22] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362.

Mathematical Models and Methods in Applied Sciences

AN ALGEBRAIC FLUX CORRECTION SCHEME SATISFYING THE DISCRETE MAXIMUM PRINCIPLE AND LINEARITY PRESERVATION ON GENERAL MESHES

GABRIEL R. BARRENECHEA

Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland
gabriel.barrenechea@strath.ac.uk

VOLKER JOHN

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany, and
Free University of Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany
john@wias-berlin.de

PETR KNOBLOCH*

Charles University, Faculty of Mathematics and Physics, Department of Numerical Mathematics, Sokolovská 83, 18675 Praha 8, Czech Republic
knobloch@karlin.mff.cuni.cz

This work is devoted to the proposal of a new flux limiter that makes the algebraic flux correction finite element scheme linearity and positivity preserving on general simplicial meshes. Minimal assumptions on the limiter are given in order to guarantee the validity of the discrete maximum principle, and then a precise definition of it is proposed and analyzed. Numerical results for convection-diffusion problems confirm the theory.

Keywords: finite element method; convection-diffusion equation; algebraic flux correction; discrete maximum principle; linearity preservation.

AMS Subject Classification: 65N30, 65N12, 65N15

1. Introduction

The numerical stability of a convection-diffusion equation is, for the most part, due to the presence of the diffusion term. Then, when convection dominates diffusion, it is natural to expect that instabilities appear in the numerical solution. These

*Corresponding author

instabilities result in the presence of large over and undershoots, which are a sign of a violation of the discrete maximum principle (DMP). To correct the violation of the DMP, many methods have been proposed and analyzed over the years. The first attempt was to add enough numerical diffusion to make the problem diffusion-dominated, and then the DMP follows under appropriate assumptions (see, e.g., Ref. 23). This crude strategy leads to numerical results which are extremely diffusive, and then not usable in practice. This fact motivated the introduction of the so-called shock-capturing methods, which are characterized by adding an extra term to the discrete formulation. This extra term contains a viscosity coefficient which is solution-dependent, hence making the method nonlinear (see Ref. 21 for a review). Nonlinear discretizations are not necessarily guaranteed to preserve the DMP, and, to the best of our knowledge, the first one was the work of Ref. 31. Later approaches include Refs. 9, 11, 3, 4, 14, 5.

All the above-mentioned references share two main hypotheses, namely, the need to use first-order polynomials, and certain assumptions on the mesh. More precisely, in the two-dimensional case the mesh is supposed to be a Delaunay one. This restriction can be tracked back to the first work concerning the validity of the DMP, even for a Laplace equation, i.e., the work of Ref. 13. Since then, several generalizations and attempts to overcome that restriction have been done. For example, in Ref. 10 an anisotropic Laplacian was added to the formulation, and the DMP can be proved for more general cases. More recently, in the context of hyperbolic equations, the works of Refs. 18, 17 propose methods that can overcome this restriction, while at the same time providing approximations that converge to the entropy solution. It is important to remark that these last references' possible extension to the case in which diffusion is present in the equations does not seem to be an easy task.

One particular nonlinear discretization, designed to satisfy the DMP by construction, is the one known as Algebraic Flux Correction (AFC) method. The origins of this method can be tracked back to Refs. 8, 33, and it has enjoyed active development in the last decade thanks to the work of D. Kuzmin and co-workers (see Refs. 24, 25, 26, 27, 28, and Ref. 29 for a recent review). This class of methods, unlike previous discretizations, is not based on a variational formulation of the problem, but rather on a restatement of the resulting linear system in which the right-hand side is written as the sum of antidiffusive fluxes. This restatement shows that these fluxes are responsible for the violation of the DMP, and then AFC schemes limit them using solution-dependent limiters. Despite the fact of providing good numerical results (apart from the above-cited references, see also the review works of Refs. 22, 1 for some further numerical results), until very recently, no mathematical analysis had been carried out for the AFC schemes. The first works in this direction are, to the best of our knowledge, Refs. 6, 7. Surprisingly, the proof of the DMP given in Ref. 7 also requires the use of a Delaunay mesh. Then, despite the fact that the geometry of the mesh does not enter explicitly in the definition of the AFC methods, some results on them still depend on the geometry of the mesh. This fact motivates the search for modifications of the limiters that generate

methods satisfying the DMP on general meshes.

Another important property that is often required for numerical discretizations is the so-called linearity preservation. This property demands that the modification added to the formulation vanishes if the solution is a polynomial of degree 1 (at least locally). This restriction, which can be interpreted as a weak consistency requirement, is believed to lead to improved accuracy in regions where the solution is smooth. In fact, in previous works, linearity preservation was linked to good convergence properties for diffusion problems (see, e.g., Refs. 20, 30). Even if this is a requirement that may seem natural, this condition was proposed in a very heuristic manner. As a matter of fact, in many works the proposed method has been claimed to be linearity preserving, but a proof of this fact is just hinted, or even lacking. In addition, although this property, so far, has not been proved mathematically to be a sufficient, or even a necessary, condition for good numerical behavior, it has been observed in different works (see, e.g., Ref. 12, and, especially, the introduction in Ref. 15 for a discussion), that linearity preservation improves the quality of the numerical solution on distorted meshes.

Based on the above considerations, our main objective in this work is to propose a definition of the limiters in an AFC method for a convection-diffusion-reaction equation that achieves two main goals: satisfaction of the DMP and linearity preservation, both on general simplicial meshes. To achieve this, we write down the main requirements to be satisfied by the limiters, and proceed to modify the algorithm proposed in Ref. 28 in such a way that these two properties are valid on general meshes. More precisely, the limiters from Ref. 28 are modified with factors that depend on the geometry of the elements that share a given node of the triangulation. Hence, this approach introduces explicit geometric information about the mesh into the algorithm.

Numerical studies will support the analytical results. In addition they show that the numerical solutions obtained with the new limiter possess further desirable properties compared with the solutions computed with the limiter from Ref. 25, which is considered to be a method of choice: it exhibits optimal convergence on distorted meshes in the diffusion-dominated regime and a sharper layer is obtained in a standard test problem for the convection-dominated case.

It is worth mentioning that methods of AFC type we have found in the literature do not satisfy the objectives of our paper in the required generality. For example, the techniques of Ref. 28, used as a basis for our method, are proved to be linearity preserving only on symmetric meshes as we discuss in Remark 6.3 below. The method recently presented in Ref. 5 has been proved to preserve the DMP only for meshes that satisfy the condition of Xu & Zikatanov[32], and this condition is sharp when the diffusion dominates. The linearity preservation of this method is again restricted to symmetric meshes. An alternative making the method linearity preserving for more general meshes requires solving an optimization problem for each interior node of the mesh, thus rendering the method more involved. Very recently, another monotone and linearity preserving method was proposed in Ref. 2

for conservation laws. However, it is not clear whether the DMP still holds when this method is applied to a convection-diffusion-reaction equation, which is our problem of interest. Moreover, the authors of Ref. 2 propose to use a regularization strategy to make the method twice differentiable and hence suitable for applying Newton's method but then the linearity preservation property is lost. Thus, to the best of our knowledge, the method presented in this paper is the first method that satisfies both the DMP and linearity preservation on general simplicial meshes, when the equation under consideration is a convection-diffusion-reaction equation. In particular, as a special result, a monotone and linearity preserving discretization of the Poisson equation on general simplicial meshes is obtained.

The rest of the paper is organized as follows. In Sec. 2, AFC schemes are presented in their most general form. Then, the minimal requirements on the limiter in order to satisfy the DMP are laid down in Sec. 3. Our concrete proposal for the limiter is given in Sec. 4. Sec. 5 is devoted to the application of the AFC scheme to the convection-diffusion-reaction equation and its analysis. The final ingredient in the definition of the limiter, namely, the computation of the multiplicative factor introduced in order to make the method linearity preserving, is presented in Sec. 6. Finally, some numerical results supporting our claims are given in Sec. 7.

2. An Algebraic Flux Correction Scheme

Consider a linear boundary value problem for which the maximum principle holds. Let us discretize this problem by the finite element method. Then, the discrete solution can be represented by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of U ($0 < M < N$) correspond to nodes where Dirichlet boundary conditions are prescribed whereas the first M components of U are computed using the finite element discretization of the underlying partial differential equation. Then $\mathrm{U} \equiv (u_1, \ldots, u_N)$ satisfies a system of linear equations of the form

$$\sum_{j=1}^{N} a_{ij}\, u_j = g_i\,, \qquad i = 1, \ldots, M\,, \tag{2.1}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{2.2}$$

We assume that the matrix $(a_{ij})_{i,j=1}^{M}$ is positive definite, i.e.,

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j > 0 \qquad \forall\, (u_1, \ldots, u_M) \in \mathbb{R}^M \setminus \{0\}\,. \tag{2.3}$$

To introduce an algebraic flux correction scheme, we first extend the matrix of (2.1) to a matrix $\mathbb{A} = (a_{ij})_{i,j=1}^{N}$. For example, one can simply use the finite element matrix corresponding to the above-mentioned finite element discretization in the case when homogeneous natural boundary conditions are used instead of the

Dirichlet ones. We shall consider this matrix with the following modification:

$$a_{ji} := 0 \quad \text{if} \quad a_{ij} < 0\,, \qquad i = 1, \ldots, M\,,\ j = M+1, \ldots, N\,. \tag{2.4}$$

This reduces the amount of artificial diffusion introduced by the matrix $\mathbb{D}$ defined next.

Using the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, we introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ with entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\ i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j \neq i} d_{ij}\,. \tag{2.5}$$

This definition guarantees that the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ has positive diagonal entries and non-positive off-diagonal entries. If, in addition,

$$\sum_{j=1}^N a_{ij} \geq 0\,, \qquad\qquad i = 1, \ldots, M\,, \tag{2.6}$$

then the matrix $\tilde{\mathbb{A}}$ satisfies sufficient conditions to preserve the discrete maximum principle. Note that the property (2.6) is usually satisfied by finite element discretizations of elliptic equations arising in applications.

Going back to the solution of (2.1), this system is equivalent to

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + (\mathbb{D}\,\mathrm{U})_i\,, \qquad i = 1, \ldots, M\,. \tag{2.7}$$

Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\,\mathrm{U})_i = \sum_{j \neq i} f_{ij}\,, \qquad i = 1, \ldots, N\,,$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. The idea of the algebraic flux correction scheme is to limit those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (2.1) (or, equivalently, (2.7)) is replaced by

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij}\, f_{ij}\,, \qquad i = 1, \ldots, M\,, \tag{2.8}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (2.1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem. So far, these coefficients have been chosen in various ways, and their definition is always based on the above fluxes f_{ij}, see Refs. 24, 25, 26, 27, 28 for examples. To guarantee that the resulting scheme is conservative, and to be able to show existence of solutions, one should require that the coefficients α_{ij} are symmetric, i.e.,

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i, j = 1, \ldots, M\,. \tag{2.9}$$

Rewriting the equation (2.8) using the definition of the matrix $\tilde{\mathbb{A}}$, one obtains the following expression for the algebraic flux correction scheme:

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1-\alpha_{ij})\, d_{ij}\, (u_j - u_i) = g_i\,, \qquad i = 1, \ldots, M\,, \tag{2.10}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,, \tag{2.11}$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \ldots, u_N) \in [0,1]$, $i = 1, \ldots, M$, $j = 1, \ldots, N$, satisfy (2.9).

The following theorem states sufficient conditions on the limiters α_{ij} assuring the solvability of the nonlinear discrete problem (2.10), (2.11). Our proposal for such limiters will be given in Sec. 4.

Theorem 2.1. *Let* (2.3) *hold. For any* $i \in \{1, \ldots, M\}$, $j \in \{1, \ldots, N\}$, *let* $\alpha_{ij} : \mathbb{R}^N \to [0,1]$ *be such that* $\alpha_{ij}(u_1, \ldots, u_N)(u_j - u_i)$ *is a continuous function of* $u_1, \ldots, u_N$. *Finally, let the functions* α_{ij} *satisfy* (2.9). *Then there exists a solution of the nonlinear problem* (2.10), (2.11).

Proof. See Theorem 3.3 in Ref. 7. □

It is worth mentioning that the symmetry property (2.9) is necessary for the validity of Theorem 2.1, see Ref. 6.

3. The Discrete Maximum Principle

As it was mentioned in the introduction, the main motivation of AFC schemes is to respect the DMP. In this section, we state some minimal assumptions on the limiters α_{ij} in order to satisfy this property.

Given $i \in \{1, \ldots, M\}$, the discrete maximum principle will be formulated locally, with respect to an index set $S_i \subset \{1, \ldots, N\}$. We assume that

$$S_i \supset \{j \in \{1, \ldots, N\} \setminus \{i\} \,:\, a_{ij} \neq 0 \ \text{ or } \ a_{ji} > 0\}\,, \qquad i = 1, \ldots, M\,. \tag{3.1}$$

The proof of the discrete maximum principle requires only that $\{\alpha_{ij} d_{ij}\}_{j \in S_i}$ vanish if u_i is a strict local extremum. More precisely, we assume that, for any $i \in \{1, \ldots, M\}$ and any $U = (u_1, \ldots, u_N) \in \mathbb{R}^N$, the limiters α_{ij} satisfy

$$u_i > u_j \quad \forall\, j \in S_i \quad \text{or} \quad u_i < u_j \quad \forall\, j \in S_i \qquad \Rightarrow \qquad \alpha_{ij}(U) d_{ij} = 0 \quad \forall\, j \in S_i\,. \tag{3.2}$$

The matrix $\mathbb{A}$ will be supposed to satisfy (2.6). Then the only assumption on $\mathbb{A}$ for proving the local discrete maximum principle at $i \in \{1, \ldots, M\}$ will be that

$$\text{there exists } j \in \{1, \ldots, N\},\, j \neq i \,: \qquad a_{ij} < 0 \quad \text{or} \quad a_{ij} < a_{ji}\,. \tag{3.3}$$

Note that the diagonal entry a_{ii} can be arbitrary. The condition (3.3) is typically satisfied, in particular, by the matrix associated to a finite element discretization of

the convection-diffusion equation (see Lemma 5.1 and Remark 5.2 below for details). If (3.3) does not hold but

$$A_i := \sum_{j=1}^{N} a_{ij} > 0\,, \tag{3.4}$$

then still a slightly weaker statement on the DMP can be proved. If $A_i = 0$ and $a_{ii} > 0$ (as implied by (2.3)), then (3.3) is always satisfied.

With the above hypotheses, we prove the main result of this section.

Theorem 3.1. *Let the matrix $\mathbb{A}$ satisfy* (2.6) *and let the limiters α_{ij} satisfy* (3.2). *Let $(u_1, \ldots, u_N) \in \mathbb{R}^N$ satisfy* (2.10). *Consider any $i \in \{1, \ldots, M\}$. If* (3.3) *holds, one has*

$$g_i \le 0 \quad \Rightarrow \quad \big(\textit{if } u_i \ge 0,\ \textit{then } u_i \le \max_{j\in S_i} u_j\big)\,, \tag{3.5}$$

$$g_i \ge 0 \quad \Rightarrow \quad \big(\textit{if } u_i \le 0,\ \textit{then } u_i \ge \min_{j\in S_i} u_j\big)\,. \tag{3.6}$$

If $A_i > 0$, one has

$$g_i \le 0 \quad \Rightarrow \quad \big(\textit{if } u_i > 0,\ \textit{then } u_i \le \max_{j\in S_i} u_j\big)\,, \tag{3.7}$$

$$g_i \ge 0 \quad \Rightarrow \quad \big(\textit{if } u_i < 0,\ \textit{then } u_i \ge \min_{j\in S_i} u_j\big)\,. \tag{3.8}$$

Consequently, if (3.3) *holds or $A_i > 0$, one has*

$$g_i \le 0 \quad \Rightarrow \quad u_i \le \max_{j\in S_i} u_j^+\,, \tag{3.9}$$

$$g_i \ge 0 \quad \Rightarrow \quad u_i \ge \min_{j\in S_i} u_j^-\,, \tag{3.10}$$

where $u_j^+ := \max\{0, u_j\}$ and $u_j^- := \min\{0, u_j\}$.

Proof. Since $d_{ij} = 0$ for any $i \in \{1, \ldots, M\}$ and $j \notin S_i \cup \{i\}$, the equation (2.10) can be written in the form

$$A_i\, u_i + \sum_{j\in S_i} \left[a_{ij} + (1 - \alpha_{ij}(U))\, d_{ij}\right](u_j - u_i) = g_i\,, \qquad i = 1, \ldots, M\,. \tag{3.11}$$

Consider any $i \in \{1, \ldots, M\}$ and let $g_i \le 0$ and $u_i \ge 0$. Let us assume that $u_i > u_j$ for all $j \in S_i$. Then (3.11) and (3.2) imply that

$$A_i\, u_i + \sum_{j\in S_i} (a_{ij} + d_{ij})\,(u_j - u_i) = g_i\,. \tag{3.12}$$

Due to the definition of d_{ij} (cf. (2.5)), one has $a_{ij} + d_{ij} \le 0$ for $j \ne i$. Moreover, if (3.3) holds, there is a $j \in S_i$ such that $a_{ij} + d_{ij} < 0$. Hence the left-hand side of (3.12) is strictly positive, which is a contradiction. If $A_i > 0$ and $u_i > 0$, then (3.12) implies that $g_i \ge A_i\, u_i > 0$. This is, again, a contradiction. Therefore, there is a $j \in S_i$ such that $u_i \le u_j$, which proves (3.5) and (3.7). The statements (3.6) and (3.8) follow

in an analogous way. Finally, (3.9) and (3.10) are immediate consequences of the preceding statements. □

Assuming equality instead of inequality in (2.6), the following stronger result can be proved.

Theorem 3.2. *Let the limiters α_{ij} satisfy* (3.2) *and let $(u_1, \ldots, u_N) \in \mathbb{R}^N$ satisfy* (2.10). *Consider any $i \in \{1, \ldots, M\}$. If $A_i = 0$ and* (3.3) *holds, then one has*

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j \in S_i} u_j \,,$$
$$g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j \in S_i} u_j \,.$$

Proof. The proof from the previous result can be applied, with the minor difference that, since $A_i = 0$, the restriction on the sign of u_i is not needed. □

4. Definition of α_{ij}

The last section imposed minimal conditions that the limiter α_{ij} used in (2.10) should satisfy in order to guarantee the discrete maximum principle. In this section we design a limiter that fulfills those hypotheses. Additionally, we are interested in proposing a limiter that makes the method linearity preserving on general simplicial meshes. Our proposal is related to the one from Ref. 28 which is, however, not proved to be linearity preserving on general meshes, see Remark 6.3. The main difference between our proposal and the one from Ref. 28 is the definition of the constant γ_i below, which will be later derived to impose linearity preservation on general simplicial meshes. We shall show that it provides limiters that guarantee the solvability of (2.10), (2.11), and the validity of the discrete maximum principle.

First, for any $i \in \{1, \ldots, M\}$, we set

$$u_i^{\max} := \max_{j \in S_i \cup \{i\}} u_j \,, \qquad u_i^{\min} := \min_{j \in S_i \cup \{i\}} u_j \,, \qquad q_i := \gamma_i \sum_{j \in S_i} d_{ij} \,, \tag{4.1}$$

where S_i is an index set satisfying (3.1) and $\gamma_i > 0$ is a fixed constant, whose value will be defined later (see (6.5) in Theorem 6.1). Furthermore, for any $i \in \{1, \ldots, M\}$, we set

$$P_i^+ := \sum_{j \in S_i} f_{ij}^+ \,, \quad P_i^- := \sum_{j \in S_i} f_{ij}^- \,, \quad Q_i^+ := q_i \,(u_i - u_i^{\max}) \,, \quad Q_i^- := q_i \,(u_i - u_i^{\min}) \,,$$

and we define

$$R_i^+ := \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\} , \qquad R_i^- := \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\} .$$

If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Finally, we set

$$\widetilde{\alpha}_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \,, \\ 1 & \text{if } f_{ij} = 0 \,, \\ R_i^- & \text{if } f_{ij} < 0 \,, \end{cases} \qquad i = 1, \ldots, M, \; j = 1, \ldots, N \,,$$

and define

$$\alpha_{ij} := \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}\,, \qquad i,j = 1,\dots,M\,,$$
$$\alpha_{ij} := \widetilde{\alpha}_{ij}\,, \qquad i = 1,\dots,M,\ j = M+1,\dots,N\,.$$

The symmetry condition (2.9) is guaranteed by the last step of this algorithm.

The following result shows that the above limiter satisfies (3.2). Then, the resulting method respects the discrete maximum principle, independently of the geometry of the mesh, provided $\mathbb{A}$ satisfies (2.6) and at least one of the conditions (3.3) and (3.4) for any $i \in \{1,\dots,M\}$.

Lemma 4.1. *The limiter α_{ij} defined in this section satisfies* (3.2).

Proof. Consider any $i \in \{1,\dots,M\}$ and $U = (u_1,\dots,u_N) \in \mathbb{R}^N$ such that $u_i > u_j$ for all $j \in S_i$. Then, $u_i^{\max} = u_i$ and hence $Q_i^+ = 0$. Choose any $j \in S_i$ and let us show that $\alpha_{ij}(U)d_{ij} = 0$. It suffices to consider $d_{ij} \neq 0$. But then $f_{ij} > 0$ and hence $P_i^+ > 0$, leading to $R_i^+ = 0$. Consequently $\widetilde{\alpha}_{ij}(U) = 0$, thus giving $\alpha_{ij}(U) = 0$. If $u_i < u_j$ for all $j \in S_i$, then the proof is analogous. □

In addition to the last lemma, the following result states that the limiter α_{ij} satisfies the continuity conditions from Theorem 2.1, and hence problem (2.10), (2.11) has a solution. Its proof is very similar to Lemma 4.1 in Ref. 7, and then we give an abridged form of it for completeness.

Lemma 4.2. *The coefficients α_{ij} are such that $\phi_{ij}(U) := \alpha_{ij}(u_1,\dots,u_N)(u_j - u_i)$ are continuous functions of $u_1,\dots,u_N$ on $\mathbb{R}^N$.*

Proof. Consider any $i \in \{1,\dots,M\}$, $j \in \{1,\dots,N\}$. Let us first investigate the continuity of $\widetilde{\alpha}_{ij}$. It suffices to consider the case $\widetilde{\alpha}_{ij} \not\equiv 1$ (and hence $d_{ij} \neq 0$ and $j \in S_i$). Let $U = \{u_i\}_{i=1}^N \in \mathbb{R}^N$. We first consider $u_i > u_j$. Then, $f_{ij} > 0$ and one obtains

$$\widetilde{\alpha}_{ij}(U) = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{|f_{ij}| + \widetilde{P}_i^+} \qquad \text{with} \qquad \widetilde{P}_i^+ = \sum_{k \in S_i \setminus \{j\}} f_{ik}^+\,.$$

Since $u_i > u_j$, there is a neighborhood of U where the denominator of the above expression does not vanish, and then the function $\widetilde{\alpha}_{ij}$ is continuous in U. Now, if $u_j > u_i$, by the same arguments one can deduce that $\widetilde{\alpha}_{ij}$ is continuous in U. Thus, if $u_i \neq u_j$, then $\widetilde{\alpha}_{ij}$, and therefore ϕ_{ij}, is continuous in U. Finally, if $u_i = u_j$, then $\phi_{ij}(U) = 0$. Let $V = \{v_i\}_{i=1}^N \in \mathbb{R}^N$. Then, since $\alpha_{ij}(U) \in [0,1]$, one obtains

$$|\phi_{ij}(V) - \phi_{ij}(U)| = |\phi_{ij}(V)| = |\alpha_{ij}(V)|\,|v_j - v_i| \leq |v_j - u_j - (v_i - u_i)| \leq \sqrt{2}\|V - U\|_{\mathbb{R}^N}\,.$$

Then, $\phi_{ij}(V) \to \phi_{ij}(U)$ if $V \to U$ and ϕ_{ij} is continuous in U. This finishes the proof. □

We finish this section by making some comments on the choice of the factors γ_i used in (4.1). First, the proof of the discrete maximum principle is independent of

their values, and then, it can be applied for choices other than the one introduced in this paper, e.g., the ones from Ref. 28. Once this is said, the actual value of γ_i has two main impacts in the performance of the AFC scheme. First, if chosen appropriately (as it will be done in Sec. 6 below), then it can be proved that the resulting scheme is linearity preserving on general simplicial meshes. Second, it influences the amount of artificial diffusion added by the AFC term to the original system (2.1). If γ_i's are increased, then more limiters α_{ij} will be equal to 1 and hence less artificial diffusion will be added. If γ_i's are decreased, then more limiters α_{ij} will be smaller than 1 and hence more artificial diffusion will be added. Thus, to reduce smearing of approximate solutions represented by the values $u_1, \ldots, u_N$, large values of γ_i's are convenient. The downside of this is that, for large values of γ_i's, the limiters $\alpha_{ij}(u_1, \ldots, u_N)$ change very rapidly near local extrema in u_i and hence the numerical solution of the nonlinear algebraic problem becomes more involved.

5. The AFC Scheme for Convection-Diffusion-Reaction Equations

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded polyhedral domain with Lipschitz boundary. Let us consider the steady-state convection-diffusion-reaction equation

$$-\varepsilon \, \Delta u + \boldsymbol{b} \cdot \nabla u + c \, u = g \quad \text{in } \Omega \,, \qquad\qquad u = u_b \quad \text{on } \partial\Omega \,, \tag{5.1}$$

where $\varepsilon \in (0, \varepsilon_0)$ with $\varepsilon_0 < +\infty$ is a constant, and $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, and $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$ are given functions satisfying

$$\nabla \cdot \boldsymbol{b} = 0 \,, \qquad c \geq \sigma_0 \geq 0 \qquad \text{in } \Omega \,,$$

where σ_0 is a constant. The weak solution of (5.1) is a function $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u, v) = (g, v) \qquad \forall \; v \in H^1_0(\Omega) \,, \tag{5.2}$$

with

$$a(u, v) = \varepsilon \, (\nabla u, \nabla v) + (\boldsymbol{b} \cdot \nabla u, v) + (c \, u, v) \,.$$

Here we adopt the usual notation for Sobolev spaces. In particular, $(\cdot, \cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. Since $c \geq \sigma_0$ in Ω and $\boldsymbol{b}$ is solenoidal, then

$$a(v, v) \geq \; \|v\|_a^2 \qquad \forall \; v \in H^1_0(\Omega) \,, \tag{5.3}$$

with

$$\|v\|_a^2 = \varepsilon \, |v|_{1,\Omega}^2 + \sigma_0 \, \|v\|_{0,\Omega}^2 \,.$$

It is well known that the weak solution of (5.1) exists, is unique, and satisfies the maximum principle (cf. Ref. 16).

Let $\mathscr{T}_h$ belong to a regular family of triangulations of $\overline{\Omega}$ consisting of simplices. We introduce the finite element spaces

$$W_h = \{ v_h \in C(\overline{\Omega}) \, : \, v_h|_T \in \mathbb{P}_1(T) \; \forall \, T \in \mathscr{T}_h \} \,, \qquad V_h = W_h \cap H^1_0(\Omega) \,,$$

consisting of continuous piecewise linear functions. From now on, we denote by $x_1, \ldots, x_N$ the vertices of the triangulation $\mathscr{T}_h$ and assume that $x_1, \ldots, x_M \in \Omega$ and $x_{M+1}, \ldots, x_N \in \partial\Omega$. Furthermore, we denote by $\varphi_1, \ldots, \varphi_N$ the usual basis functions of W_h, i.e., we assume that $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol. Then the functions $\varphi_1, \ldots, \varphi_M$ form a basis in V_h.

Now, an approximate solution of the variational problem (5.2) can be introduced as the solution of the following finite-dimensional problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a(u_h, v_h) = (g, v_h) \qquad \forall\ v_h \in V_h\,. \tag{5.4}$$

We denote

$$a_{ij} = a(\varphi_j, \varphi_i)\,, \qquad i, j = 1, \ldots, N\,, \tag{5.5}$$

$$g_i = (g, \varphi_i)\,, \qquad i = 1, \ldots, M\,, \tag{5.6}$$

$$u_i^b = u_b(x_i)\,, \qquad i = M+1, \ldots, N\,. \tag{5.7}$$

Then u_h solves (5.4) if and only if its coefficient vector with respect to the basis of W_h satisfies the relations (2.1) and (2.2). The bilinear form a defines the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$ whose entries are given by (5.5) and (2.4). Finally, thanks to (5.3) the matrix $(a_{ij})_{i,j=1}^M$ satisfies (2.3), and it follows that the problem (5.4) has a unique solution.

The artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is defined using (2.5). We introduce the nonlinear form

$$d_h(w; z, v) := \sum_{i,j=1}^{N} (1 - \alpha_{ij}(w))\, d_{ij}\, (z(x_j) - z(x_i))\, v(x_i) \qquad \forall\ w, z, v \in C(\overline{\Omega})\,,$$

with $\alpha_{ij}(w) := \alpha_{ij}(\{w(x_i)\}_{i=1}^N)$. Then the corresponding flux correction scheme (2.10), (2.11) can be rewritten as the following variational problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a(u_h, v_h) + d_h(u_h; u_h, v_h) = (g, v_h) \qquad \forall\ v_h \in V_h\,. \tag{5.8}$$

Since the limiters α_{ij} defined in the last section satisfy the assumptions of Theorem 2.1, and the bilinear form a is elliptic, then the problem (5.8) has a solution. A natural (solution dependent) norm on V_h corresponding to the left-hand side of (5.8) is defined by

$$\|v_h\|_h := \Big(\|v_h\|_a^2 + d_h(u_h; v_h, v_h)\Big)^{1/2}\,, \qquad v_h \in V_h\,.$$

Assuming that $u \in H^2(\Omega)$ and following completely analogous steps as the ones from Sec. 7 in Ref. 7 it follows that, if $\sigma_0 > 0$, the following error bound holds

$$\|u - u_h\|_h \le C\, h\, \|u\|_{2,\Omega} + (d_h(u_h; i_h u, i_h u))^{1/2}\,, \tag{5.9}$$

where $C > 0$ is independent of u, h, and ε, and $i_h u$ stands for the Lagrange interpolant of u. For the last term in (5.9), using the proof of Lemma 7.3 from Ref. 7, it follows that

$$d_h(w_h; i_h u, i_h u) \leq C \max_{i,j=1,\ldots,N} (|d_{ij}|\,|x_i - x_j|^{2-d})\,|i_h u|^2_{1,\Omega} \quad \forall\, w_h \in W_h,\, u \in C(\overline{\Omega})\,, \tag{5.10}$$

where C is independent of h and the data of problem (5.1). This result shows that the error $\|u - u_h\|_h$ will tend to zero as long as the product $|d_{ij}|\,|x_i - x_j|^{2-d}$ tends to zero. This implies that the method will converge as long as the matrix $\mathbb{A}$ tends to be an M-matrix, and this speed of convergence is fast enough to compensate for the negative power of h arising from $|x_i - x_j|^{2-d}$ in the three-dimensional case. Hence, it is natural to expect that the convergence properties of the method will vary according to the geometry of the mesh. In particular, for the convection-dominated regime, an $O(h^{1/2})$ estimate of $\|u-u_h\|_h$ can be shown irrespectively of the geometry of the mesh. On the contrary, for the diffusion-dominated regime, the convergence rates will vary dramatically depending on the geometrical properties of the mesh (see Ref. 7 for details). This was illustrated numerically in Ref. 7 for the limiter defined in Ref. 25. In some particular cases a better than expected convergence was observed, but the theoretical justification of this fact, which requires a more refined estimation of $d_h(u_h; i_h u, i_h u)$ for particular limiters, does not seem to be an easy task, and it will be the subject of our future research.

The above results are valid for any limiters α_{ij} satisfying the assumptions of Sec. 2 (resp. of Theorem 2.1) and hence, in particular, for the limiter from Sec. 4. To apply this limiter, we have to specify the sets S_i satisfying (3.1). The simplest possibility is to use

$$S_i = \{j \in \{1,\ldots,N\} \setminus \{i\} \,:\, x_i \text{ and } x_j \text{ are end points of the same edge}\}\,, \tag{5.11}$$

where $i = 1,\ldots,M$. This definition of S_i was used in the computations reported in Sec. 7. To finish the definition of α_{ij}, we have to define the factors γ_i used in (4.1). This will be done in the following section.

Remark 5.1. Usually, results on the discrete maximum principle like in Theorems 3.1 and 3.2 are proved for Delaunay meshes with respect to sets $S_i = \{j \in \{1,\ldots,N\} \setminus \{i\} \,:\, a_{ij} \neq 0\}$. For $c = 0$, this definition and the set used in (3.1) coincide in Delaunay meshes. Indeed, for such a mesh, the validity of $a_{ji} > 0$ in (3.1) implies that $a_{ij} \neq 0$ since $a_{ij} + a_{ji} = 2\,\varepsilon\,(\nabla\varphi_i, \nabla\varphi_j) \leq 0$. Whenever $c > 0$, then the two definitions no longer coincide, the set induced by (3.1) can be larger, and hence the final result is slightly weaker. The stronger assumption (3.1) is made in order to guarantee our results to be valid on arbitrary meshes.

We close this section by showing that the matrix $\mathbb{A}$ defined above satisfies the assumptions made on it to prove the discrete maximum principle.

Lemma 5.1. *The matrix* $\mathbb{A}$ *defined in* (5.5) *and* (2.4) *satisfies the assumption*

(2.6). *Moreover, for any* $i \in \{1,\dots,M\}$, *the assumption* (3.3) *holds if* $A_i = 0$ *or*

$$\text{there exists } j \in \{1,\dots,N\} : \quad (\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) \neq 0\,. \tag{5.12}$$

Proof. The validity of (2.6) follows immediately from the property $\sum_{j=1}^N \varphi_j = 1$ and the nonnegativity of c. Consider any $i \in \{1,\dots,M\}$. If $A_i = 0$, then there is $j \in \{1,\dots,N\}$, $j \neq i$, with $a_{ij} < 0$ since $a_{ii} \geq \varepsilon\,|\varphi_i|^2_{1,\Omega} > 0$. Hence (3.3) holds. Let us assume (5.12) and let (3.3) does not hold, i.e.,

$$a_{ij} \geq 0 \quad \text{and} \quad a_{ij} \geq a_{ji} \qquad \forall\, j \in \{1,\dots,N\}\,,\ j \neq i\,. \tag{5.13}$$

Under this assumption, then the modification (2.4) is not used for the matrix entries in (5.13), and the original matrix remains unchanged. Hence, in view of the second inequality in (5.13), one has

$$(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) \geq (\boldsymbol{b}\cdot\nabla\varphi_i,\varphi_j) = -(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) \qquad \forall\, j \in \{1,\dots,N\}\,,\ j \neq i\,,$$

so that

$$(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) \geq 0 \qquad \forall\, j \in \{1,\dots,N\}\,,\ j \neq i\,.$$

Since $(\boldsymbol{b}\cdot\nabla\varphi_i,\varphi_i) = 0$ and $\sum_{j=1}^N (\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) = 0$, one deduces that

$$(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i) = 0 \qquad \forall\, j \in \{1,\dots,N\}\,,$$

which is in contradiction with (5.12). □

Remark 5.2. According to the previous lemma, the validity of (3.3) is not guaranteed if the convection term does not contribute to the i-th row of the matrix $\mathbb{A}$. Although this cannot be excluded, it is a rather exceptional situation and hence (3.3) will typically hold if $\boldsymbol{b}$ does not vanish identically in $\operatorname{supp}\varphi_i$. Lemma 5.1 also shows that (3.3) holds if $c \equiv 0$ since then $A_i = 0$ for any $i \in \{1,\dots,M\}$. Thus, if the reaction term for $c > 0$ is discretized using a lumping like in Ref. 7, the off-diagonal entries of $\mathbb{A}$ are the same as for $c \equiv 0$ and hence (3.3) again holds although $A_i > 0$.

6. Linearity Preservation

Let us consider the limiter from Sec. 4 with the sets S_i defined in (5.11). In this section we finish the definition of this limiter by specifying the parameters γ_i that make it possible to prove that the resulting scheme is linearity preserving on general simplicial meshes. We recall that $x_1,\dots,x_N$ stand for the vertices of $\mathscr{T}_h$, and that $x_1,\dots,x_M \in \Omega$. We shall show that the factors γ_i in (4.1) can be defined in such a way that

$$\widetilde{\alpha}_{ij}(u) = 1 \qquad \forall\, u \in \mathbb{P}_1(\mathbb{R}^d)\,, \qquad i = 1,\dots,M,\ j = 1,\dots,N\,. \tag{6.1}$$

Then the AFC scheme (2.10), (2.11) will be linearity preserving. Let us consider any function $u \in \mathbb{P}_1(\mathbb{R}^d)$ and set $u_i = u(x_i)$, $i = 1,\dots,N$. Then, if one wants to satisfy (6.1), one needs

$$Q_i^+ \geq P_i^+ \quad \text{if } f_{ij} > 0\,, \qquad Q_i^- \leq P_i^- \quad \text{if } f_{ij} < 0\,. \tag{6.2}$$

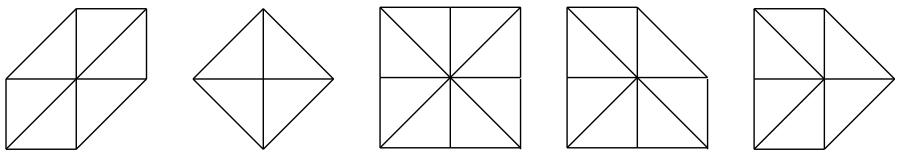


Fig. 1. Examples of patches Δ_i for $d = 2$.

Sufficient conditions for (6.2) are the inequalities

$$u_i - u_i^{\min} \leq \gamma_i \, (u_i^{\max} - u_i) \,, \qquad u_i^{\max} - u_i \leq \gamma_i \, (u_i - u_i^{\min}) \,. \tag{6.3}$$

Note that it suffices to find γ_i such that

$$u_i - u_i^{\min} \leq \gamma_i \, (u_i^{\max} - u_i) \qquad \forall \; u \in \mathbb{P}_1(\mathbb{R}^d) \,, \tag{6.4}$$

since then the second inequality in (6.3) follows from (6.4) by changing the sign of u. Thus, the validity of (6.4) assures that the AFC scheme (2.10), (2.11) based on the limiter from Sec. 4 is linearity preserving.

To discuss the validity of (6.4), it is convenient to introduce the patch $\Delta_i = \operatorname{supp} \varphi_i$ for any interior vertex x_i of the triangulation $\mathscr{T}_h$. Thus, Δ_i is a patch consisting of simplices $T \in \mathscr{T}_h$ sharing the vertex x_i, see Fig. 1. Then the sets S_i defined in (5.11) satisfy

$$S_i = \{ j \in \{1, \ldots, N\} \, : \, x_j \in \partial \Delta_i \} \,,$$

and one has

$$u_i^{\min} = \min_{\Delta_i} u \,, \qquad u_i^{\max} = \max_{\Delta_i} u \,.$$

Note that, for $u \in \mathbb{P}_1(\mathbb{R}^d)$, $u_i^{\min}$ and $u_i^{\max}$ are attained at vertices lying on $\partial \Delta_i$.

If the patch Δ_i is symmetric with respect to the vertex x_i (like the first three patches from the left in Fig. 1), then the inequality (6.4) holds with $\gamma_i = 1$ as the following lemma shows.

Lemma 6.1. *Let Δ_i be symmetric with respect to x_i. Then*

$$u_i - u_i^{\min} = u_i^{\max} - u_i \qquad \forall \; u \in \mathbb{P}_1(\mathbb{R}^d) \,.$$

Proof. Let us assume that $u_i - u_i^{\min} < u_i^{\max} - u_i$. There exists a vertex $x_j \in \partial \Delta_i$ such that $u_i^{\max} = u_j$. Furthermore, due to the symmetry of Δ_i, there is a vertex $x_k \in \partial \Delta_i$ such that $(x_j + x_k)/2 = x_i$. Then $u_j + u_k = 2 \, u_i$ and hence

$$u_i - u_i^{\min} < u_i^{\max} - u_i = u_j - u_i = u_i - u_k \,.$$

Consequently, $u_k < u_i^{\min}$, which is a contradiction. Analogously, it can be shown that $u_i - u_i^{\min} > u_i^{\max} - u_i$ leads to a contradiction. □

For general patches Δ_i, a possible factor γ_i is computed in the following theorem.

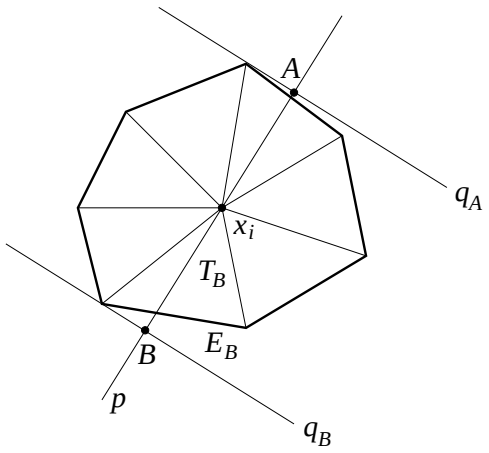


Fig. 2. Patch Δ_i with notation from the proof of Theorem 6.1.

Theorem 6.1. *Let $x_1, \dots, x_M \in \Omega$. For any $i \in \{1, \dots, M\}$, let Δ_i be the above-defined patch corresponding to the vertex x_i and let Δ_i^{conv} be its convex hull. Let*

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{dist(x_i, \partial\Delta_i^{\mathrm{conv}})}, \qquad i = 1, \dots, M. \tag{6.5}$$

Then the inequalities (6.4) *hold and hence the AFC scheme* (2.10), (2.11) *with the limiter from Sec. 4 is linearity preserving.*

Proof. For simplicity, we shall present the proof for $d = 2$. For $d = 3$ one can proceed analogously. Consider a patch Δ_i and let $u \in \mathbb{P}_1(\mathbb{R}^2)$ be any nonconstant linear function. Let p be the line in the direction of ∇u containing the vertex x_i. Then there are uniquely determined points $A, B \in p$ such that $u(A) = u_i^{\min}$, $u(B) = u_i^{\max}$. Let q_A and q_B be lines orthogonal to p intersecting the line p at the points A and B, respectively, see Fig. 2. Since u is constant along lines perpendicular to p, the patch Δ_i is contained in the strip between the lines q_A and q_B. Consequently, each of these lines intersects Δ_i only at points on $\partial\Delta_i$ comprising at least one vertex. Moreover, any such vertex lies on the boundary of the convex hull Δ_i^{conv}. To find a constant γ_i for which the inequality (6.4) holds, we have to estimate the ratio

$$\frac{u_i - u_i^{\min}}{u_i^{\max} - u_i} = \frac{u(x_i) - u(A)}{u(B) - u(x_i)} = \frac{|x_i - A|}{|B - x_i|}.$$

Since q_A contains a vertex x_k lying on $\partial\Delta_i^{\mathrm{conv}}$, one has

$$|x_i - A| \le |x_i - x_k| \le \max_{x_j \in \partial\Delta_i^{\mathrm{conv}}} |x_i - x_j| = \max_{x_j \in \partial\Delta_i} |x_i - x_j|.$$

On the other hand, if T_B is a triangle whose vertices are x_i and two consecutive vertices on $\partial\Delta_i^{\mathrm{conv}}$ such that the half-line x_iB intersects T_B (see Fig. 2), then

$$|B - x_i| \ge \mathrm{dist}(x_i, E_B),$$

where E_B is the edge of T_B opposite x_i. Consequently,

$$|B - x_i| \geq \operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}}) \,,$$

which gives (6.5). □

Remark 6.1. For the patches in Fig. 1, the formula (6.5) gives the values 2, $\sqrt{2}$, $\sqrt{2}$, 2, and 2, respectively (from the left to the right). Since the first three patches from the left are symmetric, Lemma 6.1 shows that the formula (6.5) is not optimal in general. The last two patches in Fig. 1 are nonsymmetric and, for the linear function $u(x, y) = x + y$, one obtains $u_i - u_i^{\min} = 2\,(u_i^{\max} - u_i)$. Thus, for these two patches, the formula (6.5) gives the optimal values.

This possible lack of optimality arises from the fact that we have used the worst case scenario, this is, when the extrema of the function u are attained at the vertices closest to, and furthest away from, x_i, to derive the formula (6.5). This reasoning about the worst case scenario is adapted to three space dimensions in a straightforward way.

Remark 6.2. Let us briefly mention the computation of the denominator in (6.5). First, any vertex $x_j \in \partial\Delta_i$ is shifted in the direction of the edge $x_i x_j$ on the boundary of the convex hull Δ_i^{conv}. Then one goes through all simplices T forming Δ_i^{conv} and, denoting by E the edge (or face) of T opposite x_i, one computes $\operatorname{dist}(x_i, E)$. This is particularly easy in the two-dimensional case: If T possesses an obtuse angle at an end point of E, say P, then $\operatorname{dist}(x_i, E) = |x_i - P|$. If both angles of T at the end points of E are non-obtuse, then $\operatorname{dist}(x_i, E) = 2\,|T|/|E|$. In the three-dimensional case, the computation of $\operatorname{dist}(x_i, E)$ is more involved. Nevertheless, one can replace it by $3\,|T|/|E| \leq \operatorname{dist}(x_i, E)$ (and possibly increase the value of γ_i). Another possibility is to replace $\operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}})$ by the smallest diameter of inscribed balls of simplices forming Δ_i^{conv}.

Remark 6.3. As already mentioned, the limiter proposed in this paper is related to a method presented in Ref. 28. Although the methods of Ref. 28 are claimed to be linearity preserving, it turns out that the respective proofs are not valid for general meshes. The reason is that they rely on the validity of the inequality

$$u_i - u_j \leq \gamma_{ij}\,(u_i^{\max} - u_i) \tag{6.6}$$

for any $u \in \mathbb{P}_1(\mathbb{R}^d)$ and $j \in S_i$ (with S_i defined in (5.11)), where

$$\gamma_{ij} = \frac{2}{m_i} \sum_{k \neq i} |\boldsymbol{c}_{ik} \cdot (x_i - x_j)| \,, \quad m_i = \int_\Omega \varphi_i \,\mathrm{d}x \,, \quad \boldsymbol{c}_{ik} = \int_\Omega \varphi_i \,\nabla\varphi_k \,\mathrm{d}x \,.$$

To prove (6.6), one uses the fact that $m_i \nabla u = \sum_k \boldsymbol{c}_{ik}\, u_k = \sum_k \boldsymbol{c}_{ik}(u_k - u_i)$ and $u_i - u_j = \nabla u \cdot (x_i - x_j)$, which leads to

$$u_i - u_j = \frac{1}{m_i} \sum_{k \neq i} \boldsymbol{c}_{ik} \cdot (x_i - x_j)\,(u_k - u_i) \,. \tag{6.7}$$

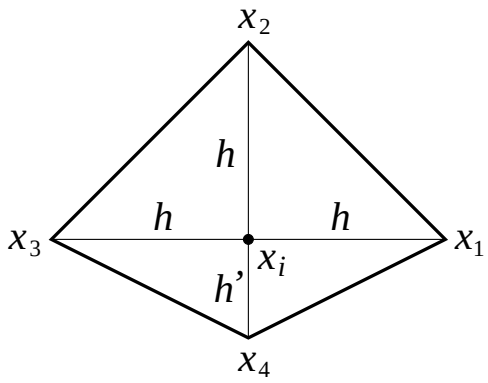


Fig. 3. Patch Δ_i for constructing a counterexample in Remark 6.3.

If the patch Δ_i is symmetric with respect to x_i, then $|u_k - u_i| \le u_i^{\max} - u_i$ for any $k \in S_i$ due to Lemma 6.1 and hence (6.7) implies (6.6). On the other hand, for non-symmetric patches, the inequality $|u_k - u_i| \le u_i^{\max} - u_i$ may be violated. Therefore, in general, (6.6) does not hold, as one can see from the following counterexample. Let us consider the patch Δ_i depicted in Fig. 3 consisting of four right-angled triangles such that the vertices x_1, x_2, x_3 have the same distance h from x_i whereas the distance of x_4 from x_i is h'. Then $\gamma_{i2} = 4\,h/(h + h')$. If $u \in \mathbb{P}_1(\mathbb{R}^2)$ satisfies $u_4 = u_i^{\max}$, then $u_i - u_2 = (u_i^{\max} - u_i)\,h/h'$ and hence (6.6) may hold with $j = 2$ only if $h \le 3\,h'$.

We finish this section by stating that the definition of the limiter presented in this work introduces explicit geometric information about the mesh into the method. This is not the standard way of defining the limiters (as the usual definitions use only the matrix entries and the solution values), and is different from the one used in Ref. 28, but it has been proved to be of fundamental importance to ensure linearity preservation on general meshes.

7. Numerical Studies

The numerical studies will illustrate the properties of the AFC scheme (2.10), (2.11) with the limiter proposed in Sec. 4 for the convection-diffusion-reaction equation from Sec. 5. If not specified otherwise, the parameters γ_i from (4.1) are defined by the formula (6.5). In addition, the results will be compared with those obtained with the limiter from Ref. 25. The limiter from Ref. 25 can be considered as a standard limiter for algebraic stabilizations of steady-state convection-diffusion-reaction equations.

For the sake of brevity, only results computed on a distorted mesh, see Fig. 4 (left), will be presented in detail. The mesh was constructed starting from the Delaunay mesh depicted in Fig. 4 (right) by shifting interior nodes to the right by half of the horizontal mesh width on each even horizontal mesh line. Therefore, for

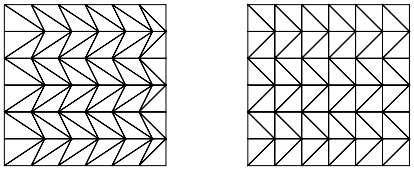


Fig. 4. Distorted mesh used in the simulations (left) and starting point for its construction (right).

most of the diagonal edges, the sum of the two angles opposite the edge is greater than $5\pi/4$ and hence the mesh is not of Delaunay type. We shall characterize the meshes by the number of edges ne along one horizontal (or equally vertical) mesh line (thus, $ne = 6$ for both meshes in Fig. 4).

Results for three examples will be presented. In the first example, the order of convergence is studied, in both the convection-dominated and diffusion-dominated regime. The second example investigates the linearity preservation property. Finally, a standard test problem with boundary layers and an interior layer is considered.

The nonlinear discrete problems were solved with a damped Newton's method.

Example 7.1. *Polynomial solution.* Problem (5.1) is considered with $\Omega = (0,1)^2$, $\boldsymbol{b} = (3,2)^T$, $c = 1$, $u_b = 0$, and the right-hand side g is chosen so that, for a given value of ε,

$$u(x,y) = 100\,x^2\,(1-x)^2\,y\,(1-y)\,(1-2y)$$

is the solution of (5.1).

The order of convergence of the error $e_h := u - u_h$ measured in various norms for the limiter proposed in Sec. 4 is presented in Table 1 for the convection-dominated case and in Table 2 for the diffusion-dominated regime. In addition, the tables show the consistency error $d_h^{1/2}(u_h) := d_h(u_h; i_h u, i_h u)^{1/2}$, cf. the estimate (5.9).

Concerning the convection-dominated case, results for the limiter from Ref. 25 on a mesh of the same type can be found in Table 6 from Ref. 7. Comparing the results, it can be seen that for both limiters the convergence orders of e_h are similar in all three norms. We could observe that this statement holds also for other meshes, in particular for more regular ones.

The situation is much different in the diffusion-dominated regime. Whereas the limiter from Sec. 4 leads to errors that decay with an optimal rate, see Table 2, the method with the limiter from Ref. 25 does not converge at all, compare Table 10 from Ref. 7. This favorable behavior of the new limiter seems to be important in situations where the convection field is a flow field. In this case there might be subregions of the domain in which the problem is diffusion-dominated.

We believe that the optimal convergence of the limiter proposed in Sec. 4 is connected with its linearity preservation property on general simplicial meshes. A similar behavior has been observed in Ref. 30, where linearity preserving limiters

Table 1. Example 7.1, $\varepsilon = 10^{-8}$, numerical results for α_{ij} from Sec. 4.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	2.722e−2	1.15	1.401e+0	0.02	9.086e−2	1.76	7.428e−2	1.21
32	1.035e−2	1.40	1.041e+0	0.43	2.287e−2	1.99	2.563e−2	1.54
64	5.099e−3	1.02	8.907e−1	0.23	6.219e−3	1.88	1.113e−2	1.20
128	2.555e−3	1.00	8.952e−1	-0.01	2.308e−3	1.43	5.240e−3	1.09
256	1.299e−3	0.98	8.991e−1	-0.01	8.409e−4	1.46	2.538e−3	1.05

Table 2. Example 7.1, $\varepsilon = 10$, numerical results for α_{ij} from Sec. 4.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	1.786e−2	1.74	4.726e−1	0.87	9.284e−1	1.13	1.522e+0	0.88
32	4.218e−3	2.08	2.404e−1	0.98	3.035e−1	1.61	7.633e−1	1.00
64	1.016e−3	2.05	1.213e−1	0.99	1.077e−1	1.49	3.841e−1	0.99
128	2.545e−4	2.00	6.082e−2	1.00	3.816e−2	1.50	1.924e−1	1.00
256	6.439e−5	1.98	3.045e−2	1.00	1.361e−2	1.49	9.632e−2	1.00
512	1.628e−5	1.98	1.524e−2	1.00	4.896e−3	1.47	4.819e−2	1.00

Table 3. Example 7.1, $\varepsilon = 10$, numerical results for α_{ij} from Sec. 4 and γ_i replaced by $\gamma_i/4$.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	4.543e−2	0.91	5.801e−1	0.68	2.753e+0	0.32	2.051e+0	0.65
32	3.095e−2	0.55	3.939e−1	0.56	2.362e+0	0.22	1.404e+0	0.55
64	2.622e−2	0.24	3.138e−1	0.33	2.199e+0	0.10	1.127e+0	0.32
128	2.428e−2	0.11	2.826e−1	0.15	2.118e+0	0.05	1.018e+0	0.15
256	2.341e−2	0.05	2.707e−1	0.06	2.078e+0	0.03	9.756e−1	0.06
512	2.301e−2	0.03	2.660e−1	0.03	2.059e+0	0.01	9.582e−1	0.03

are used to approximate a diffusion problem. The theoretical justification of this statement is not yet available, and will be the topic of our future research.

Further evidence in support of the above claim is given in Table 3. Here we present results obtained with the limiter from Sec. 4 for parameters γ_i defined as a quarter of the value provided by the formula (6.5). Then the method is not linearity preserving and we observe that the errors of the approximate solutions do

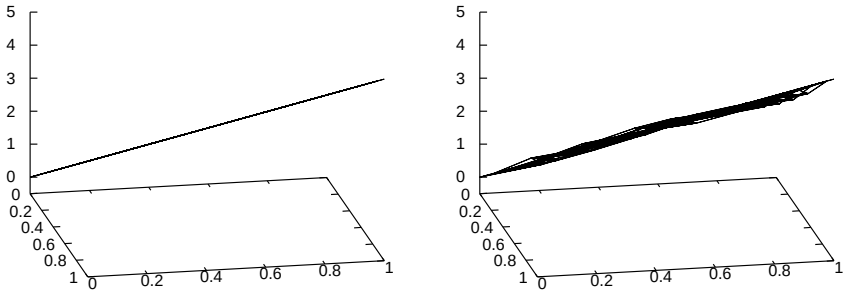


Fig. 5. Example 7.2, solution with the limiter from Sec. 4 (left) and that from Ref. 25 (right).

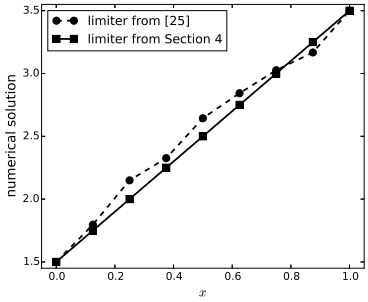


Fig. 6. Example 7.2, cross section of the solutions at $y = 0.5$.

not converge to zero.

Example 7.2. *Linear solution.* The data for this example were chosen to be $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (2\,y - x, -3\,x + y)^T$, $c = 0$, and the boundary condition u_b and the right-hand side g were set so that

$$u(x,y) = 2\,x + 3\,y$$

is the solution of (5.1).

This example serves for showing on the one hand the linearity preservation of the limiter from Sec. 4 on the considered distorted mesh. On the other hand, it also demonstrates that the limiter from Ref. 25 does not possess this property. Results for simulations with $ne = 8$ are presented in Fig. 5 and for a closer inspection also a cross-section of the two solutions is shown in Fig. 6. The limiter proposed in Sec. 4 provides a solution which is virtually the analytical solution (the maximum error is of the order of 10^{-10}, which is in accordance with the stopping criterion for the nonlinear iteration). For the limiter from Ref. 25, the violation of the linearity preservation is clearly visible.

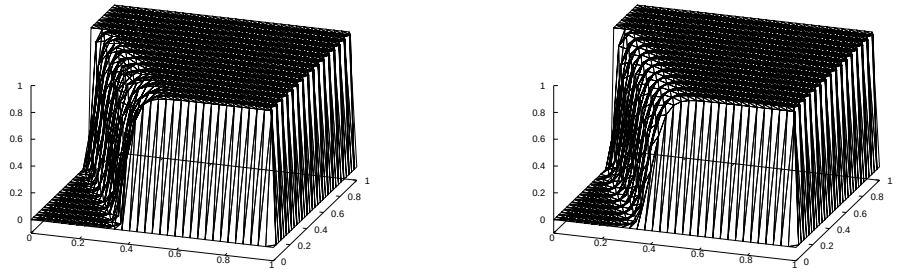


Fig. 7. Example 7.3, solutions obtained with the limiter defined in Sec. 4 (left) and the limiter from Ref. 25 (right).

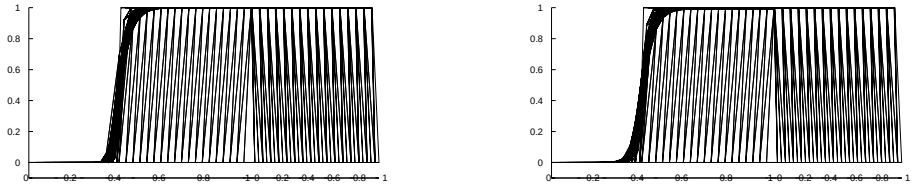


Fig. 8. Example 7.3, solutions obtained with the limiter defined in Sec. 4 (left) and the limiter from Ref. 25 (right). Both solutions respect the discrete maximum principle. The solution with the proposed limiter shows a sharper interior layer, especially at the bottom. A slight smearing can be observed along the boundary layer at $y = 0$ for the limiter from Ref. 25.

Example 7.3. *Solution with layers.* The final example considers a standard test problem defined in Ref. 19. This problem is given by $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = 0$, $g = 0$, and the boundary condition

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

Note that the boundary condition from Example 7.3 can be easily changed to an infinitely smooth function that coincides with u_b from Example 7.3 at all boundary vertices of the mesh used for the computations presented in this section. Then Example 7.3 also formally fits into the framework considered in Sec. 5.

The solutions computed with both limiters are presented in Figs. 7 and 8. It can be observed that both definitions of the limiters provide an acceptable solution. They obey the DMP and all boundary layers are sharp. A close look at the interior layer, in particular at the bottom, shows that the layer of the solution computed

with the limiter from Sec. 4 is a little bit sharper. Also, a slight smearing of the boundary layer at $y = 0$ is visible for the limiter from Ref. 25.

8. Conclusions and Outlook

This paper proposed a new limiter for algebraic stabilizations of steady-state convection-diffusion-reaction equations within the framework of finite element methods. The main goal of the construction of the new limiter was that the resulting scheme should obey the DMP and it should possess the linearity preservation property on general simplicial meshes. Both properties could be achieved and proved. The definition of the new limiter does not only rely on algebraic data but also requires some geometric information (on the local mesh), like the limiter of Ref. 2. We think that the enrichment of algebraic stabilizations with geometric information is in general a promising approach for designing stabilized methods. In contrast to the limiters of Refs. 2 and 5, the new limiter does not depend on any user-chosen parameter (like the exponent p in case of Refs. 2, 5) controlling the amount of numerical diffusion added to the method, which makes the present approach more practical.

The numerical studies showed an optimal order of convergence in the diffusion-dominated regime, which is not present for the limiter from Ref. 25. As already mentioned, we believe that this behavior of the new limiter is somehow connected to the linearity preservation, but the proof is open. A further topic of our future work will be the analysis, and possibly improvement, of algebraic stabilizations for time-dependent problems.

Acknowledgment

The work of G.R. Barrenechea has been partially funded by the Leverhulme Trust via the Research Project Grant No. RPG-2012-483. The work of V. John has been partially supported via the grant Jo329/10-2 within the DFG priority programme 1679: Dynamic simulation of interconnected solids processes. The work of P. Knobloch has been partially supported through the grant No. 16-03230S of the Czech Science Foundation.

References

1. Matthias Augustin, Alfonso Caiazzo, André Fiebach, Jürgen Fuhrmann, Volker John, Alexander Linke, and Rudolf Umla. An assessment of discretizations for convection-dominated convection-diffusion equations. Comput. Methods Appl. Mech. Engrg., 200(47-48):3395–3409, 2011.
2. Santiago Badia and Jesús Bonilla. Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. Comput. Methods Appl. Mech. Engrg., 313:133–158, 2017.
3. Santiago Badia and Alba Hierro. On monotonicity-preserving stabilized finite element approximations of transport problems. SIAM J. Sci. Comput., 36(6):A2673–A2697, 2014.

4. Santiago Badia and Alba Hierro. On discrete maximum principles for discontinuous Galerkin methods. Comput. Methods Appl. Mech. Engrg., 286:107–122, 2015.
5. Gabriel R. Barrenechea, Erik Burman, and Fotini Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes. Numer. Math., 2017. to appear.
6. Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Some analytical results for an algebraic flux correction scheme for a steady convection-diffusion equation in one dimension. IMA J. Numer. Anal., 35(4):1729–1756, 2015.
7. Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal., 54(4):2427–2451, 2016.
8. Jay P. Boris and David L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. J. Comput. Phys., 11:38–69, 1973.
9. Erik Burman and Alexandre Ern. Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion-reaction equation. Comput. Methods Appl. Mech. Engrg., 191(35):3833–3855, 2002.
10. Erik Burman and Alexandre Ern. Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes. C. R. Math. Acad. Sci. Paris, 338(8):641–646, 2004.
11. Erik Burman and Alexandre Ern. Stabilized Galerkin approximation of convection–diffusion–reaction equations: discrete maximum principle and convergence. Math. Comp., 74:1637–1652, 2005.
12. L. A. Catalano, P. De Palma, M. Napolitano, and G. Pascazio. A critical analysis of multi-dimensional upwinding for the Euler equations. Comput. & Fluids, 25(1):29–38, 1996.
13. Philippe G. Ciarlet and Pierre-Arnaud Raviart. Maximum principle and uniform convergence for the finite element method. Comput. Methods Appl. Mech. Engrg., 2:17–31, 1973.
14. Alexandre Ern and Jean-Luc Guermond. Weighting the edge stabilization. SIAM J. Numer. Anal., 51(3):1655–1677, 2013.
15. Zhiming Gao and Jiming Wu. A linearity-preserving cell-centered scheme for the heterogeneous and anisotropic diffusion equations on general meshes. Internat. J. Numer. Methods Fluids, 67(12):2157–2183, 2011.
16. David Gilbarg and Neil S. Trudinger. Elliptic partial differential equations of second order, volume 224 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, second edition, 1983.
17. Jean-Luc Guermond and Murtazo Nazarov. A maximum-principle preserving C^0 finite element method for scalar conservation equations. Comput. Methods Appl. Mech. Engrg., 272:198–213, 2014.
18. Jean-Luc Guermond, Murtazo Nazarov, Bojan Popov, and Yong Yang. A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. SIAM J. Numer. Anal., 52(4):2163–2182, 2014.
19. Thomas J. R. Hughes, Michel Mallet, and Akira Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. Comput. Methods Appl. Mech. Engrg., 54(3):341–355, 1986.
20. Willem Hundsdorfer and Carolynne Montijn. A note on flux limiting for diffusion discretizations. IMA J. Numer. Anal., 24(4):635–642, 2004.
21. Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review. Comput. Methods Appl. Mech. Engrg., 196:2197–2215, 2007.
22. Volker John and Ellen Schmeyer. Finite element methods for time-dependent

convection–diffusion–reaction equations with small diffusion. Comput. Methods Appl. Mech. Engrg., 198:475–494, 2008.
23. Claes Johnson. Numerical solution of partial differential equations by the finite element method. Dover Publications, Inc., Mineola, NY, 2009. Reprint of the 1987 edition.
24. Dmitri Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. J. Comput. Phys., 219:513–531, 2006.
25. Dmitri Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. In M. Papadrakakis, E. Oñate, and B. Schrefler, editors, Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, pages 1–5. CIMNE, Barcelona, 2007.
26. Dmitri Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. J. Comput. Appl. Math., 218:79–87, 2008.
27. Dmitri Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. J. Comput. Phys., 228:2517–2534, 2009.
28. Dmitri Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. J. Comput. Appl. Math., 236:2317–2337, 2012.
29. Dmitri Kuzmin and Jari Hämäläinen. Finite element methods for computational fluid dynamics, volume 14 of Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015. A practical guide.
30. Dmitri Kuzmin, Mikkhail J. Shashkov, and Daniil Svyatskiy. A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems. J. Comput. Phys., 228(9):3448–3463, 2009.
31. Akira Mizukami and Thomas 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(2):181–193, 1985.
32. Jinchao Xu and Ludmil Zikatanov. A monotone finite element scheme for convection-diffusion equations. Math. Comp., 68(228):1429–1446, 1999.
33. Steven T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys., 31:335–362, 1979.