Numerical Algorithms
https://doi.org/10.1007/s11075-023-01511-2

ORIGINAL PAPER

An algebraically stabilized method for convection–diffusion–reaction problems with optimal experimental convergence rates on general meshes

Petr Knobloch[1]

Received: 12 August 2022 / Accepted: 31 January 2023

Abstract
Algebraically stabilized finite element discretizations of scalar steady-state convection–diffusion–reaction equations often provide accurate approximate solutions satisfying the discrete maximum principle (DMP). However, it was observed that a deterioration of the accuracy and convergence rates may occur for some problems if meshes without local symmetries are used. The paper investigates these phenomena both numerically and analytically and the findings are used to design a new algebraic stabilization called Symmetrized Monotone Upwind-type Algebraically Stabilized (SMUAS) method. It is proved that the SMUAS method is linearity preserving and satisfies the DMP on arbitrary simplicial meshes. Moreover, numerical results indicate that the SMUAS method leads to optimal convergence rates on general simplicial meshes.

Keywords Convection–diffusion–reaction problems · Algebraic stabilization · Discrete maximum principle · Optimal convergence

Mathematics Subject Classification (2010) 65N12 · 65N30

1 Introduction

Convection, diffusion, and reaction are basic physical mechanisms which play an important role in many mathematical models used in science and technology. A

✉ Petr Knobloch
knobloch@karlin.mff.cuni.cz

[1] Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, Praha 8, 18675, Czech Republic

Published online: 04 April 2023

frequently used model problem for studying numerical techniques for the mentioned class of models is the scalar steady-state convection–diffusion–reaction problem

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

where $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is a bounded domain, $\varepsilon > 0$ is a constant diffusion coefficient, $\boldsymbol{b}$ is the convection field, c is the reaction field, and the right-hand side g is a source of the unknown quantity u. Note that the model problem (1) itself has also a clear physical meaning since it may describe, e.g., the distribution of temperature or concentration. For our mathematical considerations, we will assume that the boundary $\partial\Omega$ of Ω is polyhedral and Lipschitz-continuous (if $d \geq 2$) and that $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, and $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$. Moreover, it will be assumed that the data satisfy the conditions

$$\nabla\cdot\boldsymbol{b} = 0\,, \qquad c \geq \sigma_0 \geq 0 \qquad \text{in } \Omega\,, \tag{2}$$

where σ_0 is a constant.

In most applications, the convective transport strongly dominates the diffusion which causes that the solution u comprises so-called layers, which are narrow regions where u changes abruptly. The presence of layers makes the numerical solution of (1) very challenging since standard approaches provide solutions polluted by spurious oscillations unless the layers are resolved by the mesh. A well-known remedy is a stabilization of the standard discretization, e.g., by adding additional stabilization terms (see, e.g., [36]). To obtain accurate approximations, the stabilization has to be adopted to the character of the approximated solution which inevitably leads to nonlinear methods. However, many of such stabilization techniques still do not remove the spurious oscillations completely since the stabilization effect is influenced by many factors, like the used mesh or the considered data (cf. [18–20]). Although the remaining spurious oscillations are often quite small, they may be not acceptable in some applications, e.g., if the oscillating solution should serve as input data for other equations. A possible remedy is to apply methods satisfying the discrete maximum principle (DMP) (see, e.g., the recent review paper [8]). The DMP excludes many types of oscillating solutions that otherwise frequently appear when solving convection-dominated problems. A further reason for requiring the validity of the DMP is that a maximum principle holds for the continuous problem (1) if $c \geq 0$ (cf. [12, 14]) and it is important that this physical property is preserved by the discrete problem.

An interesting class of methods satisfying the DMP (often under some assumptions on the mesh) are algebraically stabilized finite element schemes, e.g., algebraic flux correction (AFC) schemes. These methods have been developed intensively in recent years (see, e.g., [2, 7, 15, 28–35]). The origins of this approach can be tracked back to [10, 37]. In these schemes, the stabilization is performed on the basis of the algebraic system of equations corresponding to the Galerkin finite element method. It involves so-called limiters, which restrict the stabilized discretization mainly to a vicinity of layers to ensure the satisfaction of the DMP without compromising the accuracy. There are several limiters proposed in the literature, like the so-called

Kuzmin [29], BJK [7], or BBK [4] limiters. Both, the Kuzmin and the BBK limiters were utilized in [3] for defining a scheme that blends a standard linear stabilized scheme in smooth regions and a nonlinear stabilized method in a vicinity of layers.

An important feature of algebraically stabilized schemes is that they not only satisfy the DMP but also usually provide sharp approximations of layers (cf. the numerical results in, e.g., [1, 16, 23, 31]). In this paper, we concentrate on schemes based on the idea of algebraic flux correction. Many properties of the AFC schemes are already well understood since these schemes were investigated in a number of papers (see, e.g., [5–7, 9, 24–26]) where one can find results on the existence of solutions, local and global DMPs, error estimates, and further properties. However, it was observed already in [6] that convergence rates of these schemes may be suboptimal on some meshes, even if problems without layers are considered. The aim of the present paper is to explain this behavior in some model cases and, using the results of this analysis, to propose modifications of the considered methods leading to optimal convergence rates. This will lead to a new algebraic stabilization called Symmetrized Monotone Upwind-type Algebraically Stabilized (SMUAS) method for which the solvability, linearity preservation and DMP will be proved on arbitrary simplicial meshes. Moreover, various numerical results will be reported that show that, in many cases, the SMUAS method leads to more accurate results than other algebraic stabilizations. In addition, the numerical results indicate that the SMUAS method converges with optimal rates on general simplicial meshes. Let us mention that the analysis of AFC schemes also demonstrates the interesting fact that certain types of spurious oscillations may be still present in the approximate solutions despite the validity of the DMP. This contradicts the frequently made claim that the DMP guarantees that no spurious oscillations appear.

The plan of the paper is as follows. In the next section, we define a Galerkin finite element discretization of (1) and the corresponding linear algebraic problem. Then, in Section 3, we introduce a general algebraic stabilization and summarize its main properties. Section 4 provides three examples of algebraic stabilizations. The first one is the AFC scheme with the Kuzmin limiter, the deficiencies of which are then analyzed in Section 5. The other two examples in Section 4 are the AFC scheme with the BJK limiter, for which also some results are reported in Section 5, and the MUAS method. The MUAS method is used as the basis for defining the new algebraic stabilization in Section 6. After analyzing the new method, various numerical results will be presented.

2 Galerkin finite element discretization

A finite element discretization of the convection–diffusion–reaction problem (1) is based on its weak formulation, which reads:

Find $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u, v) = (g, v) \qquad \forall\ v \in H_0^1(\Omega)\,,$$

where

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

As usual, $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. It is well known that this weak formulation has a unique solution (cf. [12]).

To define a finite element discretization of problem (1), we consider a simplicial triangulation $\mathscr{T}_h$ of $\overline{\Omega}$ which is assumed to belong to a regular family of triangulations in the sense of [11]. Furthermore, we introduce finite element spaces

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in P_1(T)\ \forall\, T \in \mathscr{T}_h\}\,, \qquad V_h = W_h \cap H_0^1(\Omega)\,,$$

consisting of continuous piecewise linear functions. The vertices of the triangulation $\mathscr{T}_h$ will be denoted by $x_1, \ldots, x_N$ and we assume that $x_1, \ldots, x_M \in \Omega$ and $x_{M+1}, \ldots, x_N \in \partial\Omega$. Then, the usual basis functions $\varphi_1, \ldots, \varphi_N$ of W_h are defined by the conditions $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol. Obviously, the functions $\varphi_1, \ldots, \varphi_M$ form a basis of V_h. Any function $u_h \in W_h$ can be written in a unique way in the form

$$u_h = \sum_{i=1}^{N} u_i\, \varphi_i \tag{3}$$

and hence it can be identified with the coefficient vector $\mathrm{U} = (u_1, \ldots, u_N)$.

Now, an approximate solution of problem (1) 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

$$a(u_h, v_h) = (g, v_h) \qquad \forall\ v_h \in V_h\,. \tag{4}$$

It is easy to show that the discrete problem (4) has a unique solution.

We denote

$$a_{ij} = a(\varphi_j, \varphi_i)\,, \qquad i, j = 1, \ldots, N\,, \tag{5}$$

$$g_i = (g, \varphi_i)\,, \qquad i = 1, \ldots, M\,, \tag{6}$$

$$u_i^b = u_b(x_i)\,, \qquad i = M+1, \ldots, N\,. \tag{7}$$

Then, u_h is a solution of the finite-dimensional problem (4) if and only if the coefficient vector $(u_1, \ldots, u_N)$ corresponding to u_h satisfies the algebraic problem

$$\sum_{j=1}^{N} a_{ij}\, u_j = g_i\,, \qquad i = 1, \ldots, M\,, \tag{8}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{9}$$

As discussed in the introduction, the above discretization is not appropriate in the convection-dominated regime and a stabilization has to be applied. The most common way is to introduce additional stabilization terms in the discrete problem (4) (see, e.g., [36]). However, another attractive possibility is to modify the algebraic problem (8), (9), which will be pursued in this paper.

3 A general algebraic stabilization

The stabilizing effect of various approaches used to suppress the spurious oscillations present in the solutions of the Galerkin discretization is due to the fact that these methods add a certain amount of artificial diffusion to the Galerkin FEM. However, if this amount is too large, the approximate solution becomes inaccurate due to an excessive smearing of the layers. It turns out that accurate solutions can be obtained only if the amount of the artificial diffusion respects the local behavior of the solution (see, e.g., [8]). This motivates us to stabilize the algebraic problem (8), (9) by adding an artificial diffusion matrix $\mathbb{B}(\mathrm{U}) = (b_{ij}(\mathrm{U}))_{i,j=1}^{N}$ which depends on the unknown approximate solution $\mathrm{U} = (u_1, \ldots, u_N)$. Here, we shall describe this approach only briefly and refer to the recent paper [21] for a more detailed presentation.

Based on the above discussion, we will consider the nonlinear algebraic problem

$$\sum_{j=1}^{N} (a_{ij} + b_{ij}(\mathrm{U}))\, u_j = g_i\,, \qquad i = 1, \ldots, M\,, \tag{10}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{11}$$

We assume that, for any $\mathrm{U} \in \mathbb{R}^N$, the matrix $\mathbb{B}(\mathrm{U})$ satisfies

$$b_{ij}(\mathrm{U}) = b_{ji}(\mathrm{U})\,, \qquad i, j = 1, \ldots, N\,, \tag{12}$$

$$b_{ij}(\mathrm{U}) \le 0, \qquad i, j = 1, \ldots, N\,,\ i \ne j\,, \tag{13}$$

$$\sum_{j=1}^{N} b_{ij}(\mathrm{U}) = 0\,, \qquad i = 1, \ldots, N\,. \tag{14}$$

Moreover, we assume that $\mathbb{B}(\mathrm{U})$ has the typical sparsity pattern of finite element matrices, i.e.,

$$b_{ij}(\mathrm{U}) = 0 \qquad \forall\ j \notin S_i \cup \{i\},\ i = 1, \ldots, M\,, \tag{15}$$

where

$$S_i = \{j \in \{1, \ldots, N\} \setminus \{i\}\,;\ x_i \text{ and } x_j \text{ are end points of the same edge}\}\,.$$

These assumptions are motivated by the fact that the properties (12)–(15) are satisfied for the diffusion matrix $(\varepsilon\,(\nabla\varphi_j, \nabla\varphi_i))_{i,j=1}^{N}$ if the triangulation $\mathscr{T}_h$ is weakly acute, i.e., if the angles between facets of $\mathscr{T}_h$ do not exceed $\pi/2$. It is also important that the properties (12)–(14) assure that the matrix $\mathbb{B}(\mathrm{U})$ is positive semidefinite for any $\mathrm{U} \in \mathbb{R}^N$ (see [21]).

To prove the solvability of the system (10), (11), we make the following assumption, which is motivated by the definitions of the matrix $\mathbb{B}(\mathrm{U})$ considered in this paper.

Assumption (A1) For any $i \in \{1, \ldots, M\}$ and any $j \in \{1, \ldots, N\}$, the function $b_{ij}(\mathrm{U})(u_j - u_i)$ is a continuous function of $\mathrm{U} = (u_1, \ldots, u_N) \in \mathbb{R}^N$ and, for any $i \in \{1, \ldots, M\}$ and any $j \in \{M+1, \ldots, N\}$, the function $b_{ij}(\mathrm{U})$ is a bounded function of $\mathrm{U} \in \mathbb{R}^N$.

Theorem 1 *Let* (12)–(14) *hold and let Assumption (A1) be satisfied. Then, there exists a solution of the nonlinear problem* (10), (11).

Proof See [21]. □

The construction of the matrix $\mathbb{B}(\mathrm{U})$ is usually based on the requirement that the problem (10), (11) satisfies the DMP. One can formulate various conditions that guarantee that a nonlinear discrete problem satisfies the DMP or at least preserves the positivity (cf. [8]). For our purposes, the following assumption is useful.

Assumption (A2) Consider any $\mathrm{U} = (u_1, \ldots, u_N) \in \mathbb{R}^N$ and any $i \in \{1, \ldots, M\}$. If u_i is a strict local extremum of U with respect to S_i, i.e.,

$$u_i > u_j \quad \forall\ j \in S_i \qquad \text{or} \qquad u_i < u_j \quad \forall\ j \in S_i\,,$$

then

$$a_{ij} + b_{ij}(\mathrm{U}) \le 0 \qquad \forall\ j \in S_i\,.$$

Under the above assumptions, it is possible to prove that the approximate solution obtained using the nonlinear problem (10), (11) satisfies a direct analogue of the maximum principles which hold for the problem (1) (see, e.g., [12] for the classical solutions and [14] for the weak solutions).

Theorem 2 *Let the assumptions stated in* Section 1 *be satisfied and let the matrix* $\mathbb{B}(\mathrm{U})$ *satisfy* (12)–(15) *and Assumptions (A1) and (A2). Consider any nonempty set* $\mathscr{G}_h \subset \mathscr{T}_h$ *and define*

$$G_h = \bigcup_{T \in \mathscr{G}_h} T\,.$$

Let $\mathrm{U} \in \mathbb{R}^N$ *be a solution of* (10) *and let* $u_h \in W_h$ *be the corresponding finite element function given by* (3)*. Then, one has the DMP*

$$g \le 0 \quad \text{in } G_h \quad \Rightarrow \quad \max_{G_h} u_h \le \max_{\partial G_h} u_h^+\,,$$
$$g \ge 0 \quad \text{in } G_h \quad \Rightarrow \quad \min_{G_h} u_h \ge \min_{\partial G_h} u_h^-\,,$$

where $u_h^+ = \max\{u_h, 0\}$ *and* $u_h^- = \min\{u_h, 0\}$*. If, in addition,* $c = 0$ *in* G_h*, then*

$$g \le 0 \quad \text{in } G_h \quad \Rightarrow \quad \max_{G_h} u_h = \max_{\partial G_h} u_h\,,$$
$$g \ge 0 \quad \text{in } G_h \quad \Rightarrow \quad \min_{G_h} u_h = \min_{\partial G_h} u_h\,.$$

Proof See [21]. □

We will close this section with a brief discussion of a priori error estimates available for the nonlinear problem (10), (11). To derive an error estimate, it is convenient

to write (10), (11) as a variational problem where the algebraic stabilization term is represented using the form

$$b_h(w; z, v) = \sum_{i,j=1}^{N} b_{ij}(w)\, z(x_j)\, v(x_i)\,, \qquad w, z, v \in C(\overline{\Omega})\,,$$

with $b_{ij}(w) := b_{ij}(\{w(x_i)\}_{i=1}^N)$ (see [21] for details). This variational problem is stable with respect to the solution-dependent norm on W_h defined by

$$\|v\|_h := \Big(\varepsilon\, |v|_{1,\Omega}^2 + \sigma_0\, \|v\|_{0,\Omega}^2 + b_h(u_h; v, v)\Big)^{1/2}, \qquad v \in H^1(\Omega) \cap C(\overline{\Omega})\,,$$

assuming that $\sigma_0 > 0$ in (2). This shows that the problem (10), (11) really provides a stronger stability than the original problem (8), (9).

The algebraic stabilization term leads to a consistency error whose behavior with respect to h depends on how the artificial diffusion matrix is constructed. Often, one has

$$|b_{ij}(u_h)| \le \max\{|a_{ij}|, |a_{ji}|\} \qquad \forall\, i \neq j\,,$$

which will be also the case in this paper. Under this assumption and assuming further that the weak solution of (1) satisfies $u \in H^2(\Omega)$ and that $\sigma_0 > 0$, one can prove (cf. [21]) that the finite element function $u_h \in W_h$, corresponding via (3) to the solution $\mathrm{U} \in \mathbb{R}^N$ of the nonlinear algebraic problem (10), (11), satisfies the estimate

$$\begin{aligned} \|u - u_h\|_h \;\le\; & C\,(\varepsilon + \sigma_0^{-1}\,\{\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\} + \sigma_0 h^2)^{1/2}\, h\, \|u\|_{2,\Omega} \\ & + C\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h + \|c\|_{0,\infty,\Omega}\, h^2)^{1/2}\, |i_h u|_{1,\Omega}\,, \end{aligned} \tag{16}$$

where the constant C is independent of h and the data of problem (1). If $\sigma_0 = 0$, then the estimate is deteriorated by a negative power of ε (see [6] for details). We also refer to [6] and [9] for slightly improved error estimates under various additional assumptions.

The estimate (16) does not imply any convergence in the diffusion-dominated case (when $\varepsilon > \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$) and it guarantees only the convergence order $1/2$ in the convection-dominated case. Numerical results presented in [6] show that this result is sharp under the general assumptions made up to now. It is of course desirable to design the artificial diffusion matrix $\mathbb{B}(\mathrm{U})$ in such a way that optimal convergence rates with respect to various norms are obtained. For some algebraic stabilizations, optimal convergence rates were indeed observed; however, more detailed convergence studies revealed that the convergence rates often depend on the considered meshes and data (cf. [6, 7]). The aim of this paper is to analyze some of these observations and to propose an algebraic stabilization for which optimal convergence rates can be observed in a wide range of situations, in particular, for various types of simplicial meshes.

4 Examples of algebraic stabilizations

In this section, we present three examples of algebraic stabilizations based on the papers [29], [7], and [21], respectively. All these stabilizations fit into the framework of the previous section.

4.1 Algebraic flux correction with the Kuzmin limiter

To derive an algebraic flux correction (AFC) scheme for the problem (8), (9), one first introduces the artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ by

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

Note that this matrix possesses the properties (12)–(15). If $(\mathbb{D}\,\mathrm{U})_i$ is added to the left-hand side of (8), one obtains a problem satisfying the DMP. However, this stabilized problem is too diffusive. Therefore, one first adds the term $(\mathbb{D}\,\mathrm{U})_i$ to both sides of (8), uses the identity

$$(\mathbb{D}\,\mathrm{U})_i = \sum_{j=1}^N f_{ij} \qquad \text{with} \qquad f_{ij} = d_{ij}\,(u_j - u_i)$$

and then, on the right-hand side, one limits those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. The limiting is achieved by multiplying the fluxes by solution dependent limiters $\alpha_{ij} \in [0, 1]$ satisfying

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i, j = 1, \ldots, N\,. \tag{17}$$

This leads to the algebraic problem (10), (11) with

$$b_{ij}(\mathrm{U}) = (1 - \alpha_{ij}(\mathrm{U}))\, d_{ij} \qquad \forall\, i \neq j\,, \qquad\qquad b_{ii}(\mathrm{U}) = -\sum_{j\neq i} b_{ij}(\mathrm{U})\,. \tag{18}$$

This matrix $(b_{ij}(\mathrm{U}))_{i,j=1}^N$ satisfies the assumptions (12)–(15). A theoretical analysis of this AFC scheme concerning the solvability, local DMP, and error estimation can be found in [6] where also a detailed derivation of the scheme is presented.

The properties of the above-described AFC scheme significantly depend on the choice of the limiters α_{ij}. Here, we present the Kuzmin limiter proposed in [29] which was thoroughly investigated in [6] and can be considered as a standard limiter for algebraic stabilizations of steady-state convection–diffusion–reaction equations.

To define the limiter of [29], one first computes, for $i = 1, \ldots, M$,

$$P_i^+ = \sum_{\substack{j\in S_i \\ a_{ji}\le a_{ij}}} f_{ij}^+\,, \qquad P_i^- = \sum_{\substack{j\in S_i \\ a_{ji}\le a_{ij}}} f_{ij}^-\,, \qquad Q_i^+ = -\sum_{j\in S_i} f_{ij}^-\,, \qquad Q_i^- = -\sum_{j\in S_i} f_{ij}^+\,, \tag{19}$$

where $f_{ij} = d_{ij}(u_j - u_i)$, $f_{ij}^+ = \max\{0, f_{ij}\}$, and $f_{ij}^- = \min\{0, f_{ij}\}$. Then, one defines

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \qquad i = 1, \ldots, M. \tag{20}$$

If P_i^+ or P_i^- vanishes, one sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. At Dirichlet nodes, these quantities are also set to be 1, i.e.,

$$R_i^+ = 1, \quad R_i^- = 1, \qquad i = M+1, \ldots, N. \tag{21}$$

Furthermore, one sets

$$\widetilde{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0, \end{cases} \qquad i, j = 1, \ldots, N. \tag{22}$$

Finally, one defines

$$\alpha_{ij} = \alpha_{ji} = \widetilde{\alpha}_{ij} \qquad \text{if} \quad a_{ji} \le a_{ij}, \qquad i, j = 1, \ldots, N. \tag{23}$$

It was proved in [6] that $\alpha_{ij}(\mathrm{U})(u_j - u_i)$ are continuous functions of $\mathrm{U} \in \mathbb{R}^N$ so that the assumption (A1) is satisfied for $b_{ij}(\mathrm{U})$ defined by (18) with the Kuzmin limiter. The validity of (A2) was proved in [25] under the assumption

$$\min\{a_{ij}, a_{ji}\} \le 0 \qquad \forall\, i = 1, \ldots, M, \;\; j = 1, \ldots, N, \;\; i \ne j. \tag{24}$$

On the other hand, it was shown in [25] that the DMP generally does not hold if the condition (24) is not satisfied. Since the convection matrix is skew-symmetric, the condition (24) can be violated if the diffusion matrix has large positive entries (which may occur if the angles between facets of $\mathscr{T}_h$ exceed $\pi/2$) or if the reaction coefficient c is large. As a remedy for the latter case, a lumping of the reaction term was considered in [6]. This, however, may increase the smearing of layers as demonstrated in [21]. Let us mention that, in the two-dimensional case and for $c = 0$ or a lumped reaction term, the validity of (24) is guaranteed for Delaunay meshes (i.e., meshes where the sum of any pair of angles opposite a common edge is smaller than, or equal to, π).

Since it is desirable that the DMP holds on arbitrary simplicial meshes and without a lumping of the reaction term, it is necessary to apply other limiters or different algebraic stabilizations. This will be the subject of the following two sections.

4.2 Algebraic flux correction with the BJK limiter

In this section, we again consider an AFC scheme, i.e., the matrix $\mathbb{B}(\mathrm{U})$ in (10), (11) is defined by (18). A small difference to the previous section is that the matrix $(a_{ij})_{i,j=1}^N$ is modified by

$$a_{ji} := 0 \quad \text{if} \quad a_{ij} < 0, \qquad i = 1, \ldots, M, \;\; j = M+1, \ldots, N.$$

This modification affects only the definition of the matrix $\mathbb{D}$ and reduces the amount of artificial diffusion introduced by the algebraic stabilization. We shall describe the

so-called BJK limiter proposed in [7] using some ideas of [31]. The definition of this limiter is inspired by the Zalesak algorithm [37] for the time-dependent case.

The definition of the limiter again relies on local quantities P_i^+, P_i^-, Q_i^+, Q_i^- which are now computed for $i = 1, \ldots, M$ by

$$P_i^+ = \sum_{j \in S_i} f_{ij}^+ , \qquad P_i^- = \sum_{j \in S_i} f_{ij}^- , \tag{25}$$

$$Q_i^+ = q_i \, (u_i - u_i^{\max}) , \qquad Q_i^- = q_i \, (u_i - u_i^{\min}) , \tag{26}$$

where again $f_{ij} = d_{ij} \, (u_j - u_i)$ and

$$u_i^{\max} = \max_{j \in S_i \cup \{i\}} u_j , \qquad u_i^{\min} = \min_{j \in S_i \cup \{i\}} u_j , \qquad q_i = \sum_{j \in S_i} d_{ij} .$$

Then, one defines

$$R_i^+ = \min \left\{ 1, \frac{\mu_i \, Q_i^+}{P_i^+} \right\} , \quad R_i^- = \min \left\{ 1, \frac{\mu_i \, Q_i^-}{P_i^-} \right\} , \qquad i = 1, \ldots, M , \tag{27}$$

with fixed constants $\mu_i > 0$. If P_i^+ or P_i^- vanishes, one again sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. The definition (21) of $R_i^\pm$ at Dirichlet nodes is applied, too, and one again defines the factors $\widetilde{\alpha}_{ij}$ by (22). Finally, the limiter functions are defined by

$$\alpha_{ij} = \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\} , \qquad i, j = 1, \ldots, N . \tag{28}$$

The validity of the Assumptions (A1) and (A2) was proved in [7] without any additional assumptions on the matrix $(a_{ij})_{i,j=1}^N$. Thus, in particular, the DMP holds for arbitrary simplicial meshes and any nonnegative reaction coefficient c. Moreover, it was shown in [7] that the constants μ_i can be defined in such a way that the AFC scheme with the BJK limiter is linearity preserving, i.e., $\mathbb{B}(u) = 0$ for $u \in P_1(\mathbb{R}^d)$. This property may lead to improved convergence results (see, e.g., [4, 9]).

To formulate a sufficient condition for the linearity preservation, we introduce the patches

$$\Delta_i = \cup\{T \in \mathscr{T}_h \, : \, x_i \in T\} , \qquad i = 1, \ldots, M , \tag{29}$$

consisting of simplices from $\mathscr{T}_h$ sharing the vertex x_i. Then, the AFC scheme with the BJK limiter is linearity preserving if

$$\mu_i \geq \frac{\max\limits_{x_j \in \partial \Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial \Delta_i^{\mathrm{conv}})} , \qquad i = 1, \ldots, M , \tag{30}$$

where Δ_i^{conv} is the convex hull of Δ_i. It was also proved in [7] that it suffices to set $\mu_i = 1$ if the patch Δ_i is symmetric with respect to the vertex x_i. Note that large values of the constants μ_i cause that more limiters α_{ij} are equal to 1 and hence less artificial diffusion is added, which makes it possible to obtain sharp approximations of layers. On the other hand, however, large values of μ_i's also cause that the numerical solution of the nonlinear algebraic problem becomes more involved.

4.3 Monotone upwind-type algebraically stabilized method

Although the BJK limiter presented in the previous section has nice theoretical properties, numerical experiments revealed that it has also some drawbacks in comparison with the Kuzmin limiter. In particular, the nonlinear algebraic problems are much more difficult to solve and the approximate solutions are sometimes less accurate away from layers (see [17, 22]). Therefore, another approach based on the Kuzmin limiter was developed in [21, 27] that will be presented in this section.

As we mentioned in Section 4.1, the DMP generally does not hold for the AFC scheme with the Kuzmin limiter if the condition (24) is not satisfied. The need of (24) for proving the Assumption (A2) is a consequence of the condition $a_{ji} \le a_{ij}$ used in (23) to symmetrize the factors $\widetilde{\alpha}_{ij}$. A possible remedy is to replace (23) by (28) and to define $P_i^{\pm}$ by (25). Then, the DMP is satisfied without any additional condition on the matrix $(a_{ij})_{i,j=1}^N$ but the method is more diffusive then the scheme from Section 4.1 (see [25]).

The inequality $a_{ji} < a_{ij}$ often means that the vertex x_i lies in the upwind direction with respect to the vertex x_j (see [25] for a discussion on this topic). Consequently, the use of the inequality $a_{ji} \le a_{ij}$ in (23) causes that $\alpha_{ij} = \alpha_{ji}$ is defined using quantities computed at the upwind vertex of the edge with end points x_i, x_j. It turns out that this feature has a positive influence on the quality of the approximate solutions and on the convergence of the iterative process for solving the nonlinear problem (10), (11).

In order to obtain a method possessing the mentioned upwind feature and satisfying the DMP on arbitrary simplicial meshes, the definition of the matrix $\mathbb{B}(\mathrm{U})$ was changed in [21] to

$$b_{ij}(\mathrm{U}) = -\max\{\beta_{ij}(\mathrm{U})\, a_{ij}, 0, \beta_{ji}(\mathrm{U})\, a_{ji}\}\,, \qquad i, j = 1, \dots, N\,,\ i \neq j\,, \tag{31}$$

$$b_{ii}(\mathrm{U}) = -\sum_{j \neq i} b_{ij}(\mathrm{U})\,, \qquad i = 1, \dots, N\,, \tag{32}$$

with some solution-dependent factors $\beta_{ij}(\mathrm{U}) \in [0, 1]$. This matrix again satisfies the assumptions (12)–(15) but, in contrast to (18), the formula (31) leads to a symmetric matrix $\mathbb{B}(\mathrm{U})$ also if the factors β_{ij} are not symmetric. This makes it possible to get rid of the symmetry condition (17).

If the condition (24) is satisfied, then

$$b_{ij}(\mathrm{U}) = \begin{cases} \beta_{ij}(\mathrm{U})\, d_{ij} & \text{if } a_{ji} \le a_{ij}\,, \\ \beta_{ji}(\mathrm{U})\, d_{ij} & \text{otherwise}\,, \end{cases}$$

for $i = 1, \dots, M$ and $j = 1, \dots, N$ with $i \neq j$. Thus, in this case, the definition (31) implicitly comprises the favorable upwind feature discussed above and the method (10), (11) can be again written in the form of an AFC scheme. Moreover, if the functions β_{ij} form a symmetric matrix and $\alpha_{ij} = 1 - \beta_{ij}$, then the definitions (18) and (31), (32) are equivalent.

Thus, let us consider the algebraic problem (10), (11) with the artificial diffusion matrix given by (31) and (32) and with any functions $\beta_{ij} : \mathbb{R}^N \to [0, 1]$ satisfying, for any $i, j \in \{1, \ldots, N\}$,

$$\text{if } a_{ij} > 0, \text{ then } \beta_{ij}(\mathrm{U})(u_j - u_i) \text{ is a continuous function of } \mathrm{U} \in \mathbb{R}^N . \tag{33}$$

Then, one has the following existence result.

Theorem 3 *Let the matrix $(b_{ij}(\mathrm{U}))_{i,j=1}^N$ be defined by (31) and (32) with functions $\beta_{ij} : \mathbb{R}^N \to [0, 1]$ satisfying (33) for any $i, j \in \{1, \ldots, N\}$. Then, Assumption (A1) is satisfied and the nonlinear algebraic problem (10), (11) has a solution.*

Proof See [21]. □

Rewriting the definition of the Kuzmin limiter under the condition (24), the following definition of β_{ij} was introduced in [21]. First, for any $i \in \{1, \ldots, M\}$, one computes

$$P_i^+ = \sum_{\substack{j \in S_i \\ a_{ij} > 0}} a_{ij} (u_i - u_j)^+ , \qquad P_i^- = \sum_{\substack{j \in S_i \\ a_{ij} > 0}} a_{ij} (u_i - u_j)^- , \tag{34}$$

$$Q_i^+ = \sum_{j \in S_i} s_{ij} (u_j - u_i)^+ , \qquad Q_i^- = \sum_{j \in S_i} s_{ij} (u_j - u_i)^- , \tag{35}$$

with

$$s_{ij} = \max\{|a_{ij}|, a_{ji}\} .$$

Then, one defines $R_i^\pm$ by (20) and (21), and sets

$$\beta_{ij} = \begin{cases} 1 - R_i^+ & \text{if } u_i > u_j , \\ 0 & \text{if } u_i = u_j , \\ 1 - R_i^- & \text{if } u_i < u_j , \end{cases} \qquad i, j = 1, \ldots, N . \tag{36}$$

It was proved in [21] that the resulting method satisfies the Assumptions (A1) and (A2) without any additional assumptions on the matrix $(a_{ij})_{i,j=1}^N$. Thus, the DMP holds on arbitrary simplicial meshes and for any nonnegative reaction coefficient c. Due to the above-discussed upwind feature, the name Monotone Upwind-type Algebraically Stabilized (MUAS) method was introduced in [21].

If the condition (24) holds, then the only difference between the MUAS method and the AFC scheme with the Kuzmin limiter is the definition of $Q_i^\pm$ since the relations (19) give (35) with $s_{ij} = |d_{ij}|$. In the convection-dominated regime, the difference is negligible and both methods lead to almost the same results. Therefore, the MUAS method preserves the advantages of the AFC scheme from Section 4.1 which are available under the condition (24). Note that, without the assumption (24), the application of the AFC scheme with the Kuzmin limiter does not make much sense since the main goal of the AFC, i.e., the validity of the DMP, is not achieved in general. In the diffusion-dominated case, the use of s_{ij} instead of $|d_{ij}|$ may improve the accuracy and convergence behavior when non-Delaunay meshes are used (see [21]).

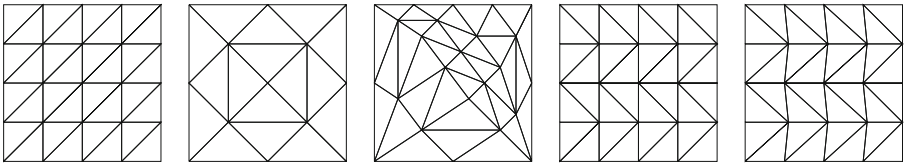


Fig. 1 Grids 1–5 (left to right)

5 Numerical and analytical studies of AFC schemes

The convergence properties of the AFC scheme with the Kuzmin limiter from Section 4.1 were thoroughly tested in [6] for various grids and the following example.

Example 1 Problem (1) 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 so that

$$u(x, y) = 100\, x^2\, (1 - x)^2\, y\, (1 - y)\, (1 - 2y)$$

is the solution of (1).

The coarsest levels of the grids considered in [6] are shown in Fig. 1. Grids 1, 2, and 3 were refined uniformly whereas Grid 4 was always 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 the tenth of the horizontal mesh width on each even horizontal mesh line. Note that Grids 3 and 5 are not of Delaunay type.

Errors of the approximate solutions of Example 1 with $\varepsilon = 10^{-8}$ computed using the AFC scheme with the Kuzmin limiter for Grid 1 can be seen in Table 1. The results slightly differ from those in [6] since, in contrast to the present paper, a lumping of the reaction term was applied in [6]. The value of ne represents the number of edges along one horizontal mesh line (thus, $ne = 4$ for Grid 1 in Fig. 1). One observes the usual optimal orders of convergence with respect to the L^2 norm and the H^1 seminorm. Moreover, the convergence order with respect to the norm $\|\cdot\|_h$ is much higher than predicted by (16). However, if the computation is repeated on

Table 1 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 1 computed using the AFC scheme with the Kuzmin limiter

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	1.934e−2	1.60	4.937e−1	0.98	5.007e−2	1.87
32	5.359e−3	1.85	2.305e−1	1.10	1.149e−2	2.12
64	1.385e−3	1.95	1.082e−1	1.09	2.649e−3	2.12
128	3.442e−4	2.01	5.154e−2	1.07	6.152e−4	2.11
256	8.536e−5	2.01	2.566e−2	1.01	1.586e−4	1.96
512	2.126e−5	2.01	1.342e−2	0.93	3.876e−5	2.03

Table 2 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 4 computed using the AFC scheme with the Kuzmin limiter

ne	$\|u-u_h\|_{0,\Omega}$	Order	$\|u-u_h\|_{1,\Omega}$	Order	$\|u-u_h\|_h$	Order
16	2.019e−2	1.65	6.005e−1	0.68	5.663e−2	1.74
32	6.285e−3	1.68	4.832e−1	0.31	2.138e−2	1.41
64	2.308e−3	1.45	4.549e−1	0.09	9.485e−3	1.17
128	1.092e−3	1.08	4.442e−1	0.03	4.490e−3	1.08
256	5.543e−4	0.98	4.368e−1	0.02	2.187e−3	1.04
512	2.823e−4	0.97	4.327e−1	0.01	1.083e−3	1.01

Grid 4, one observes in Table 2 that the convergence orders with respect to all three norms deteriorate by 1 and, in particular, one has no error reduction with respect to the H^1 seminorm. A similar behavior can be observed for Grids 3 and 5. For Grid 2, the deterioration of the convergence is less pronounced but the convergence orders are also far from being optimal (see [6] for the case with a lumped reaction term leading to similar results). Let us mention that, in all these computations, the matrix $(a_{ij})_{i,j=1}^N$ satisfies the condition (24), which guarantees the validity of the DMP.

On the other hand, there are various grids for which optimal convergence orders can be observed. Examples of such grids are given in Fig. 2. The finer variants of Grid 6 are obtained by uniform refinement like for Grid 1 whereas Grid 7 is obtained from Grid 6 by changing the directions of some of the diagonals. Grid 8 is obtained from Grid 6 by adding the second diagonal in each small square. Finer variants of Grid 9 are also not constructed by refining the coarse level but each level is constructed separately (cf. the rightmost grid in Fig. 2). Obviously, the basic difference between Grids 2–5 and Grids 1 and 6–9 is that, in the latter case, (most of) the patches Δ_i defined by (29) are symmetric, i.e., for any vertex $x_j \in \partial\Delta_i$ there exists a vertex $x_k \in \partial\Delta_i$ such that $(x_j + x_k)/2 = x_i$. Thus, it seems that the local symmetry of the grids is important for optimal convergence rates.

To understand why the approximate solutions on Grid 4 do not converge in the H^1 seminorm, let us have a look at the graphs of some of these solutions. Figure 3 (left) shows the solution computed for $ne = 32$ and it can be seen that the solution is polluted by an oscillating component (for the sake of clarity, the solution is drawn only along grid lines of Grid 4 which are parallel to the coordinate axes). This is also clearly seen from Fig. 3 (right) which shows the wildly oscillating error $u_h - i_h u$, where i_h is the usual Lagrange interpolation operator. The observed structure of the

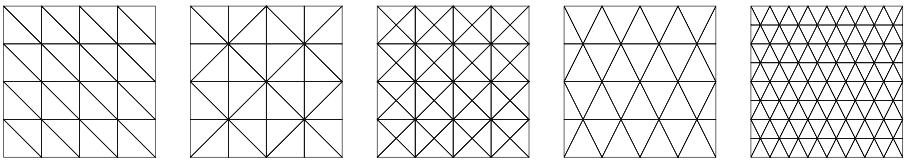


Fig. 2 Grids 6–9 (left to right) and a finer variant of Grid 9 (rightmost)

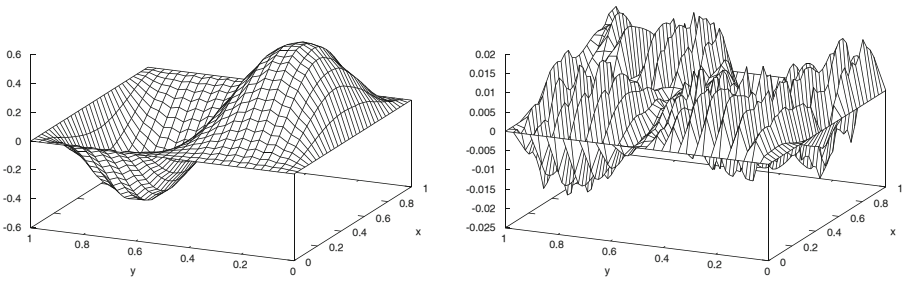


Fig. 3 Example 1: $\varepsilon = 10^{-8}$, approximate solution computed using the AFC scheme with the Kuzmin limiter on Grid 4 with $ne = 32$ (left) and the corresponding error $u_h - i_h u$ (right)

solution remains preserved also on finer meshes. Figure 4 shows the errors $u_h - i_h u$ along the line $x = 0.25$ on three successive meshes and indicates that the H^1 seminorm of the error will not change significantly when switching to finer meshes (notice the different scales on the vertical axes). It should be mentioned that this type of oscillations does not represent a violation of the DMP. The oscillatory behavior of the approximate solutions suggests that the accuracy might be improved by a local averaging. This is indeed possible but the convergence rates generally still remain suboptimal.

Figures 3 and 4 explain why the H^1 seminorm of the error of the approximate solution does not tend to zero on Grid 4 and now the main question is why the observed oscillations are not suppressed by the algebraic stabilization. To answer this question, we shall consider simpler examples than Example 1. We start with the following almost trivial case.

Example 2 Problem (1) is considered with $\Omega = (0, 1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1, 0)^T$, $c = 0$, $g = 1$, and $u_b(x, y) = x$.

Of course, the exact solution of this example is $u(x, y) = x$ and hence the Galerkin FEM gives the exact solution on any mesh. However, if one applies the AFC scheme with the Kuzmin limiter on Grid 4, one obtains the oscillating solution shown in Fig. 5 (left). Again, the structure of the solution is preserved also on finer meshes. Moreover, numerical tests show that the size of the oscillations is proportional to h

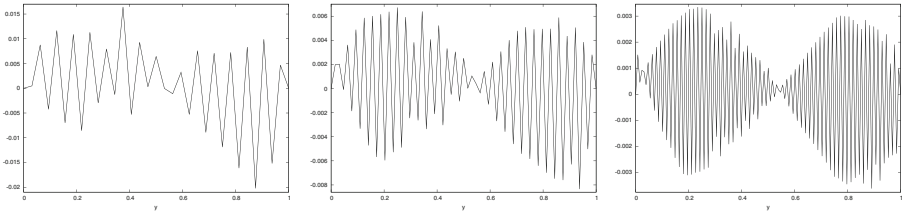


Fig. 4 Example 1: $\varepsilon = 10^{-8}$, errors $u_h - i_h u$ along the line $x = 0.25$ for approximate solutions computed using the AFC scheme with the Kuzmin limiter on Grid 4 with $ne = 32$, $ne = 64$ and $ne = 128$ (left to right)

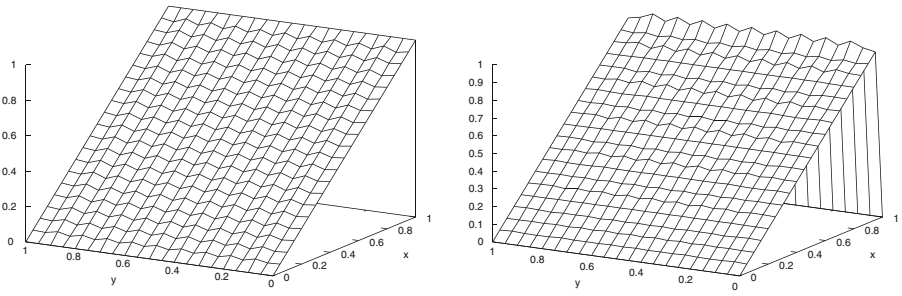


Fig. 5 Approximate solutions computed using the AFC scheme with the Kuzmin limiter on Grid 4 with $ne = 20$: Example 2 (left) and Example 3 computed with the modification (47), (48) (right)

so that one can again expect that the H^1 seminorm of the error will not tend to zero. This is confirmed by the results shown in Table 3.

Before we start our analytical investigations of this surprising observation, let us have a closer look at Grid 4 and the matrix entries corresponding to Example 2. First, note that all the patches Δ_i defined in (29) have the same geometry for Grid 4 but they possess two types of orientation with respect to the constant convection vector $\boldsymbol{b}$. This can be seen in Fig. 6 where a part of Grid 4 is shown. One type of orientation of the patches is represented by the patch around the node A and the other one by the patch around the node B. Note that A lies on an even horizontal grid line and B on an odd horizontal grid line.

In the previous sections, we referred to the nodes x_i of the triangulation through their indices i. In the following, it will be more convenient to use directly the notation for nodes. Thus, for example, the matrix entry a_{ij} will be denoted by a_{AB} if $x_i = A$ and $x_j = B$. Then, the linear system (8), (9), can be written in the form

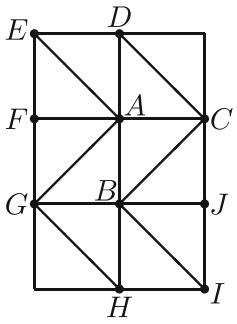
$$\sum_{Q \in \mathscr{I}(P)} a_{PQ}\, u_Q = g_P \quad \forall\ P \in \mathscr{N}_h^i\,, \qquad u_P = u_P^b \quad \forall\ P \in \mathscr{N}_h^b\,, \tag{37}$$

where $\mathscr{N}_h^i$ is the set of interior nodes of $\mathscr{T}_h$, $\mathscr{N}_h^b$ is the set of boundary nodes of $\mathscr{T}_h$, and $\mathscr{I}(P) \subset \mathscr{N}_h^i \cup \mathscr{N}_h^b$ consists of the node P and all nodes connected to P by edges

Table 3 Example 2: numerical results for Grid 4 computed using the AFC scheme with the Kuzmin limiter

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	8.104e−3	0.82	4.401e−1	−0.20	1.179e−2	0.78
32	4.291e−3	0.92	4.700e−1	−0.09	6.227e−3	0.92
64	2.204e−3	0.96	4.851e−1	−0.05	3.157e−3	0.98
128	1.117e−3	0.98	4.926e−1	−0.02	1.580e−3	1.00
256	5.618e−4	0.99	4.963e−1	−0.01	7.893e−4	1.00
512	2.817e−4	1.00	4.982e−1	−0.01	3.974e−4	0.99

Fig. 6 A part of Grid 4



of $\mathcal{T}_h$. Note that $\mathcal{I}(A) = \{A, B, C, D, E, F, G\}$ and $\mathcal{I}(B) = \{B, A, G, H, I, J, C\}$ in the case depicted in Fig. 6.

Considering the notation introduced in Fig. 6 and the data of Example 2, the entries of the Galerkin matrix $(a_{ij})_{i,j=1}^{N}$ defined in (5) are given by

$$
\begin{aligned}
a_{AB} &= a_{AD} = a_{HB} = -\varepsilon + \frac{h}{6}, \qquad & a_{AC} &= a_{FA} = a_{BJ} = a_{GB} = -\varepsilon + \frac{h}{3},\\
a_{BA} &= a_{DA} = a_{BH} = -\varepsilon - \frac{h}{6}, \qquad & a_{CA} &= a_{AF} = a_{JB} = a_{BG} = -\varepsilon - \frac{h}{3},\\
a_{EA} &= a_{GA} = a_{BC} = a_{BI} = \frac{h}{6}, \qquad & a_{AA} &= 4\,\varepsilon,\\
a_{AE} &= a_{AG} = a_{CB} = a_{IB} = -\frac{h}{6}, \qquad & a_{BB} &= 4\,\varepsilon,
\end{aligned}
$$

where h is the mesh width in the directions of the coordinate axes. In our analytical considerations, it will be always assumed that

$$\varepsilon < \frac{h}{9}. \tag{38}$$

Since the data of Example 2 are constant and the triangulation is uniform, the matrix entries do not depend on the actual position of the nodes A and B. The above values of the entries of the Galerkin matrix imply that the relations for $P_i^{\pm}$ from (19) can be written in the form

$$P_A^{\pm} = f_{AB}^{\pm} + f_{AC}^{\pm} + f_{AD}^{\pm}, \qquad P_B^{\pm} = f_{BI}^{\pm} + f_{BJ}^{\pm} + f_{BC}^{\pm}. \tag{39}$$

Finally, note that, under the assumption (38), one has

$$
\begin{aligned}
d_{AB} &= d_{AD} = \varepsilon - \frac{h}{6}, \qquad & d_{AC} &= d_{AF} = \varepsilon - \frac{h}{3}, \qquad & d_{AE} &= d_{AG} = -\frac{h}{6},\\
d_{BA} &= d_{BH} = \varepsilon - \frac{h}{6}, \qquad & d_{BG} &= d_{BJ} = \varepsilon - \frac{h}{3}, \qquad & d_{BI} &= d_{BC} = -\frac{h}{6}.
\end{aligned}
$$

A closer look at Fig. 5 (left) reveals that, in a large part of the computational domain, the discrete solution u_h is approximately given by

$$u_h(x, y) = x + \alpha \qquad \text{along odd horizontal grid lines}\,, \tag{40}$$

$$u_h(x, y) = x - \beta \qquad \text{along even horizontal grid lines}\,, \tag{41}$$

where α and β are positive constants. A direct computation shows that the nodal values of this function satisfy

$$\sum_{Q \in \mathscr{I}(A)} a_{AQ}\, u_Q = h^2 - 2\,\delta\,\varepsilon\,, \qquad \sum_{Q \in \mathscr{I}(B)} a_{BQ}\, u_Q = h^2 + 2\,\delta\,\varepsilon\,, \tag{42}$$

where $\delta = \alpha + \beta$. For the data of Example 2, one has $g_P = h^2$ in (37) and hence one observes that the function u_h given by (40) and (41) satisfies the Galerkin discretization up to the perturbation $2\,\delta\,\varepsilon$. This leads us to the surprising conclusion that, for the oscillating solution shown in Fig. 5 (left), the AFC stabilization term should be nearly zero.

Thus, let us investigate the AFC stabilization term when it is applied to a function satisfying (40) and (41). If we consider $\delta \le h$, then

$$f_{AB} \le 0\,, \quad f_{AC} \le 0\,, \quad f_{AD} \le 0\,, \quad f_{AE} \ge 0\,, \quad f_{AF} \ge 0\,, \quad f_{AG} \ge 0\,, \tag{43}$$

$$f_{BI} \le 0\,, \quad f_{BJ} \le 0\,, \quad f_{BC} \le 0\,, \quad f_{BA} \ge 0\,, \quad f_{BG} \ge 0\,, \quad f_{BH} \ge 0\,, \tag{44}$$

which together with (39) implies that $P_A^+ = P_B^+ = 0$ and hence $R_A^+ = R_B^+ = 1$. Furthermore, it follows from (19), (43), and (44) that

$$Q_A^- = -f_{AE}^+ - f_{AF}^+ - f_{AG}^+\,, \qquad Q_B^- = -f_{BA}^+ - f_{BG}^+ - f_{BH}^+\,. \tag{45}$$

Then, a direct computation gives

$$P_A^- = Q_B^- = \varepsilon\,(h + 2\,\delta) - \frac{h}{3}\,(h + \delta)\,,$$

$$Q_A^- = P_B^- = \varepsilon\,h - \frac{h}{3}\,(2\,h - \delta)\,.$$

Moreover, setting

$$\delta = \frac{h}{2}\,\frac{h}{h - 3\,\varepsilon}\,, \tag{46}$$

one obtains $P_A^- = Q_A^-$ and $P_B^- = Q_B^-$, which implies that $R_A^- = R_B^- = 1$. Since the nodes A and B were chosen arbitrarily, one observes that, for any function u_h given by (40) and (41) with $\alpha + \beta$ equal to δ from (46), the AFC stabilization term vanishes. This shows that our conjecture was correct.

The fact that the AFC scheme with the Kuzmin limiter does not reproduce the exact solution $u(x, y) = x$ implies that the method is not linearity preserving or not uniquely solvable. The following lemma shows that the former possibility holds true.

Lemma 1 *Let $u \in P_1(\mathbb{R}^2)$ be an arbitrary first degree polynomial and let us consider the arrangement from* Fig. 6 *and the above matrix entries corresponding to Example 2. Then, the quantities from* (19) *computed using the nodal values of u satisfy*

$$P_A^+ \le Q_A^+\,, \qquad P_A^- \ge Q_A^-$$

and

$$P_B^+ \le \frac{2h - 3\varepsilon}{h - 3\varepsilon} Q_B^+, \qquad P_B^- \ge \frac{2h - 3\varepsilon}{h - 3\varepsilon} Q_B^-,$$

where the latter inequalities are sharp. Consequently, the AFC scheme with the Kuzmin limiter is not linearity preserving on Grid 4 when applied to Example 2.

Proof First consider the inequalities at the node A. Since the values of $P_A^\pm$ and $Q_A^\pm$ do not change if a constant function is added to u, one can consider $u_A = 0$. Moreover, the ratios Q_A^+/P_A^+ and Q_A^-/P_A^- do not change if u is multiplied by a positive constant. Thus, it suffices to consider three types of functions u: with $u_C = -u_F = 1$, $u_C = -u_F = -1$, and $u_C = u_F = 0$. These functions are then determined by the value u_B and it is sufficient to consider $u_B \ge 0$ in view of the axisymmetry of the patch Δ_A. Then, it is straightforward to verify that the inequalities at the node A hold.

The inequalities at the node B can be verified analogously. Equalities hold for u given by $u_A = u_B = 0$, $u_G = 1$ and $u_A = u_B = 0$, $u_G = -1$, respectively. For these functions, one gets $(\mathbb{B}(\mathrm{U})\mathrm{U})_B = h^2/(6h - 9\varepsilon)$ and $(\mathbb{B}(\mathrm{U})\mathrm{U})_B = -h^2/(6h - 9\varepsilon)$, respectively, which means that the considered method is not linearity preserving. □

In view of the previous lemma, it is not surprising that the exact solution of Example 2 is not recovered by the AFC scheme with the Kuzmin limiter on Grid 4. Nevertheless, it is rather disappointing that, for this very simple example, the H^1 seminorm of the error cannot be reduced by considering finer meshes.

On the other hand, we also see from Lemma 1 that it is easy to modify the AFC scheme with the Kuzmin limiter in such a way that the method becomes linearity preserving for the considered case. In fact, similarly as in (27), it suffices to replace $R_B^\pm$ by

$$R_B^+ = \min\left\{1, \frac{\mu\, Q_B^+}{P_B^+}\right\}, \quad R_B^- = \min\left\{1, \frac{\mu\, Q_B^-}{P_B^-}\right\}, \tag{47}$$

with an appropriate positive constant μ. It can be easily verified that this does not change other properties of the method formulated so far. To simplify our analytical considerations, we shall use

$$\mu = \frac{2h - 3\varepsilon}{h - 4\varepsilon}, \tag{48}$$

which is a slightly larger value than suggested by Lemma 1. Nevertheless, for values of h and ε considered in our numerical computations, this modification is negligible.

If one now repeats the computation leading to the result in Fig. 5 (left) with $R_i^\pm$ given by (47), (48) at nodes lying on odd horizontal grid lines, one obtains the exact solution $u_h(x, y) = x$. This is not surprising since this exact solution solves the Galerkin discretization and the AFC stabilization term now vanishes for first degree polynomials. Thus, let us consider the following slightly more difficult example.

Example 3 Problem (1) is considered with $\Omega = (0, 1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1, 0)^T$, $c = 0$, $g = 1$, and

$$u_b(x, y) = x - \frac{\mathrm{e}^{\frac{x}{\varepsilon}} - 1}{\mathrm{e}^{\frac{1}{\varepsilon}} - 1}. \tag{49}$$

The formula in (49) not only defines the boundary condition but it also represents the solution $u = u(x, y)$ of Example 3. In most of Ω, $u(x, y)$ is very close to x, only in the vicinity of the outflow boundary $x = 1$ it abruptly falls to 0 and exhibits an exponential boundary layer. The approximate solution obtained on Grid 4 using the AFC scheme with the Kuzmin limiter modified by (47), (48) is depicted in Fig. 5 (right) and one can observe that it is again rather poor. The character of the solution remains the same also on finer meshes where one can observe that, in a large part of the computational domain, the discrete solution u_h is approximately given by three parameters. For example, for $ne = 80$, one can deduce the following form of the discrete solution:

$$u_h(x, y) = \begin{cases} x + \alpha & \text{if } x = (3k - 1)\, h\,, \\ x + \beta & \text{otherwise}\,, \end{cases} \qquad \text{along odd horizontal grid lines}\,, \tag{50}$$

$$u_h(x, y) = \begin{cases} x + \beta & \text{if } x = (3k + 1)\, h\,, \\ x - \gamma & \text{otherwise}\,, \end{cases} \qquad \text{along even horizontal grid lines}\,, \tag{51}$$

where k is an arbitrary integer and α, β, γ are positive constants.

Let us now again investigate when a function u_h given by (50) and (51) satisfies the Galerkin discretization or the AFC scheme. Due to the definition of u_h, one has to distinguish six cases: whether the node under consideration lies on an odd or an even horizontal grid line and whether the respective vertical grid line is expressed by $x = (3k - 1)\, h$, $x = 3k\, h$, or $x = (3k + 1)\, h$. Like in (42), one derives in all six cases that, under the condition

$$\alpha = 2\,\beta + \gamma\,, \tag{52}$$

u_h satisfies the Galerkin discretization up to a perturbation $\kappa\,(\beta + \gamma)\,\varepsilon$ with $\kappa \in \{-5, -3, -1, 1, 2, 6\}$. Since the computed discrete solutions approximately satisfy (52), we again expect that the AFC stabilization term is nearly zero for them.

To compute the AFC stabilization term for u_h given by (50) and (51), we shall assume that, apart from (38) and (52), one also has

$$\beta + \gamma \le \frac{h}{4}\,. \tag{53}$$

Then, it is easy to verify that, in all cases, the fluxes again have the signs given in (43) and (44). Thus, one again immediately obtains that $R_A^+ = R_B^+ = 1$ and the relations (45) hold. Using (39), (45), (47), and (48), a lengthy but straightforward computation reveals that, in all three cases, $R_A^- = R_B^- = 1$, which means that the AFC stabilization term again vanishes. Note that the condition (53) allows more flexibility in defining the function u_h than (46) which determines the respective u_h uniquely up to an additive constant.

There are two important conclusions of the above discussion. The first one, a more general, is that approximate solutions may be polluted by spurious oscillations despite the validity of the DMP. This may happen also if the right-hand side g vanishes (in contrast to the above examples) (see Example 5 below). The second conclusion is that there are oscillating functions (which may solve, e.g., a Galerkin

discretization) for which the algebraic stabilization term vanishes. This is a surprising observation that does not correspond to the usual experience that, in case of an oscillating solution, a stabilization introduces an artificial diffusion in the discrete problem to suppress the oscillations. However, it is worth noting that if a discretization of Example 2 on some of the Grids 1, 4–7, and 9 leads to an oscillating approximate solution of the type (40) and (41), then also residual-based stabilizations (see, e.g., [36]) are not able to suppress the oscillations since the residual vanishes on any element of the triangulation.

Let us mention that if Example 3 is solved on Grid 1 using the AFC scheme with the Kuzmin limiter, one obtains the nodally exact solution, except for the rightmost vertical interior grid line (see Fig. 7 (left)). However, if $R_i^{\pm}$ are defined by (27) with $\mu_i = 2$, then oscillations again appear (see Fig. 7 (second from left)). Since the AFC scheme with the Kuzmin limiter is linearity preserving on Grid 1 for constant data, this shows that the symmetry of the patches and the linearity preservation are not sufficient for obtaining an accurate approximate solution. The error at the rightmost vertical interior grid line appears independently of the choice of the limiter as it was proved in [24] so that a refinement of the mesh along the outflow boundary is needed for enhancing the accuracy.

Let us now change the boundary condition of Example 3 to the homogeneous one, i.e., consider

Example 4 Problem (1) is considered with $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (1,0)^T$, $c = 0$, $g = 1$, and $u_b = 0$.

The solution of this example possesses not only an exponential boundary layer at the outflow boundary but also two parabolic boundary layers. The AFC scheme with the Kuzmin limiter on Grid 1 provides the approximate solution shown in Fig. 7 (second from right). One can observe that, in the region of the numerical parabolic boundary layers, the approximate solution is not monotone in the crosswind direction. This can be improved by defining $P_i^{\pm}$ by (25) instead of (19) (see Fig. 7 (right)). In general, this modification decreases $R_i^{\pm}$ so that more artificial diffusion is introduced, which may lead to a more pronounced smearing of layers. Then, the accuracy can be again enhanced by using a finer mesh in the boundary layer region.

Let us mention that, for the finite element functions given by (40), (41) or (50), (51) and for the matrix entries corresponding to Grid 4 and the data of Example 2, the values of the Kuzmin limiter are determined only by the quantities R_i^-. Since it follows from (43), (44) that the quantities P_i^- attain the same values for both definitions

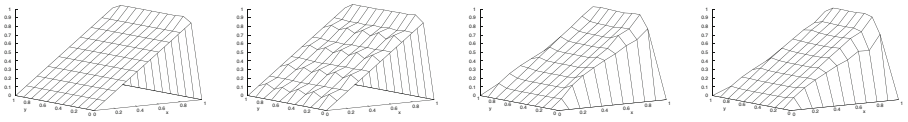


Fig. 7 Approximate solutions computed using the AFC scheme with the Kuzmin limiter on Grid 1 with $ne = 10$: Example 3 (left), Example 3, $R_i^{\pm}$ defined by (27) with $\mu_i = 2$ (second from left), Example 4 (second from right), Example 4, $P_i^{\pm}$ defined by (25) (right)

(19) and (25), the above analytical results remain the same also if $P_i^{\pm}$ are defined by (25). Also the result in Fig. 7 (left) is not affected by computing $P_i^{\pm}$ using (25).

Figure 8 shows results for Example 4 computed using the AFC scheme with the BJK limiter on Grid 1. As we know from Section 4.2, one can consider $\mu_i = 1$ in (27) for Grid 1 to guarantee the linearity preservation, which leads to the oscillatory solution in Fig. 8 (left). If one uses $\mu_i = 2$ as suggested by the formula (30), the oscillations become even larger (see Fig. 8 (right)). This again demonstrates that the symmetry of the patches and the linearity preservation are not sufficient for obtaining an accurate approximate solution. Moreover, the results presented in Figs. 5, 7, and 8 show that using the modification (27) (with $\mu_i > 1$) of (20) (e.g., to enforce the linearity preservation or to reduce the amount of artificial diffusion) is not a good idea since it allows more oscillatory solutions. In fact, this is not surprising since, for any finite element function, for which the quantities $R_i^{\pm}$ do not vanish, one can find μ_i such that the AFC stabilization term vanishes.

In particular, one should avoid such constructions of limiters for which $R_i^+ = R_i^- = 1$ may occur for oscillating functions. If $P_i^{\pm}$ are defined by (25) and $Q_i^{\pm}$ by (19), as suggested in the discussion to Fig. 7, then $P_i^+ + Q_i^- = P_i^- + Q_i^+ = 0$ and it is easy to verify that $R_i^+ = R_i^- = 1$ is equivalent to $P_i^+ = Q_i^+$ and $P_i^- = Q_i^-$, i.e., to

$$0 = \sum_{j \in S_i} (f_{ij}^+ + f_{ij}^-) = \sum_{j \in S_i} f_{ij} = \sum_{j \in S_i} d_{ij} (u_j - u_i) .$$

Thus, $R_i^+ = R_i^- = 1$ holds if and only if $u_i = \bar{u}_i$ where $\bar{u}_i$ is a local average defined by

$$\bar{u}_i = \frac{\sum_{j \in S_i} |d_{ij}| \, u_j}{\sum_{j \in S_i} |d_{ij}|} . \tag{54}$$

To avoid oscillating approximate solutions, the local averages $\bar{u}_i$ should be good approximations of the values u_i for smoothly varying functions, which is the case for locally symmetric meshes like Grid 1 but not meshes with unsymmetric patches like Grid 4. This probably contributes to the better performance of the AFC scheme with the Kuzmin limiter on locally symmetric meshes.

The above examples have all non-vanishing right-hand sides g so that the DMP provides only one-sided local bounds on approximate solutions. To demonstrate that the above-discussed phenomena are not restricted to this case, let us consider the following example with a vanishing right-hand side.

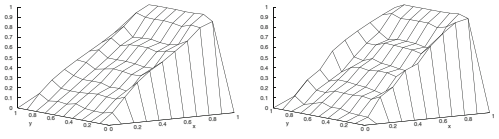


Fig. 8 Example 4: approximate solutions computed using the AFC scheme with the BJK limiter on Grid 1 with $ne = 10$ for $\mu_i = 1$ (left) and $\mu_i = 2$ (right)

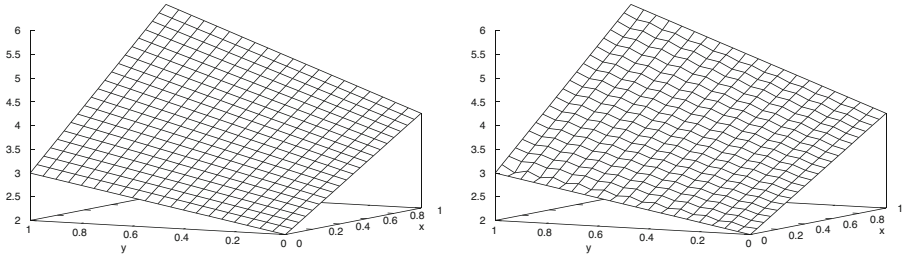


Fig. 9 Example 5: approximate solutions computed using the AFC scheme with the Kuzmin limiter on Grid 1 (left) and on Grid 4 (right), in both cases with $ne = 20$

Example 5 Problem (1) is considered with $\Omega = (0, 1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (-2, -3)^T$, $g = 0$, and

$$c(x, y) = \frac{3\,x + 2\,y + 7}{(x + 1)(y + 2)}\,, \qquad u_b(x, y) = (x + 1)(y + 2)\,.$$

Note that the solution of Example 5 is $u(x, y) = (x + 1)(y + 2)$. Whereas, on Grid 1, the AFC scheme with the Kuzmin limiter leads to an accurate approximation, the results on Grid 4 are again polluted by spurious oscillations (see Fig. 9). Moreover, on Grid 1, one can observe the same optimal convergence rates as in Table 1 whereas an analogous reduction of the convergence rates as in Table 2 is observed on Grid 4.

6 Symmetrized monotone upwind-type algebraically stabilized method

The aim of this section is to design an algebraic stabilization which will not suffer from the deficiencies discussed and analyzed in the previous section. The starting point will be the MUAS method of Section 4.3 since this method has the favorable property being of upwind type and satisfies the DMP on arbitrary simplicial meshes.

It was argued below Example 4 in the previous section that $P_i^{\pm}$ should be defined by (25) instead of (19). Consequently, the relations (34) in the MUAS method should be changed to

$$P_i^+ = \sum_{j \in S_i} |d_{ij}|\,(u_i - u_j)^+\,, \qquad P_i^- = \sum_{j \in S_i} |d_{ij}|\,(u_i - u_j)^-\,. \tag{55}$$

Moreover, it was observed in the previous section that two properties seem to be important for obtaining accurate results using algebraic stabilizations: local symmetries of triangulations and the linearity preservation. As it was demonstrated that the linearity preservation should not be enforced using (27), our goal will be to get this property by symmetrizing the definitions of $P_i^{\pm}$ and $Q_i^{\pm}$ in a suitable way.

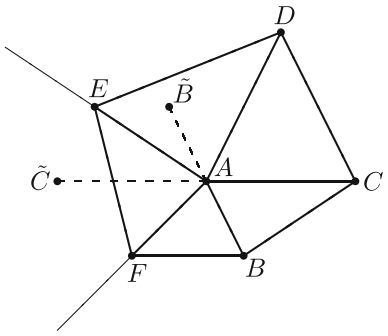


Fig. 10 Construction of symmetrically placed points

To introduce the mentioned symmetry, we will extend the definitions of $P_i^{\pm}$ and $Q_i^{\pm}$ by considering values at symmetrically placed points. The construction is illustrated by Fig. 10 where a patch Δ_A around a node A is shown. Each node P connected to A by an edge is mapped to a point $\tilde{P}$ in a symmetric way with respect to A and then the idea is to compute the value u_{AP} at $\tilde{P}$ of the finite element function u_h corresponding to U via (3). This is easy in case of the node B from Fig. 10 since $\tilde{B} \in \Delta_A$. However, in case of the node C, the symmetrically placed point $\tilde{C}$ lies outside Δ_A. In this case, we extend the linear function u_h from the triangle AEF to the convex set surrounded by the half lines AE and AF and define u_{AC} as the value of this extended function at $\tilde{C}$. This makes more sense than considering the actual value $u_h(\tilde{C})$. The value u_{AC} can be easily computed using the gradient of u_h on AEF since

$$u_{AC} = u_A + \nabla u_h|_{AEF} \cdot (\tilde{C} - A) = u_A + \nabla u_h|_{AEF} \cdot (A - C)\,.$$

Of course, an analogous relation holds for u_{AB}, too.

Using the above-defined values u_{AP}, we can now symmetrize the definitions of the quantities $P_i^{\pm}$ and $Q_i^{\pm}$ in (55) and (35), respectively, for any $i \in \{1, \ldots, M\}$ by setting

$$P_i^+ = \sum_{j \in S_i} |d_{ij}| \left\{(u_i - u_j)^+ + (u_i - u_{ij})^+\right\}, \tag{56}$$

$$P_i^- = \sum_{j \in S_i} |d_{ij}| \left\{(u_i - u_j)^- + (u_i - u_{ij})^-\right\}, \tag{57}$$

$$Q_i^+ = \sum_{j \in S_i} s_{ij} \left\{(u_j - u_i)^+ + (u_{ij} - u_i)^+\right\}, \tag{58}$$

$$Q_i^- = \sum_{j \in S_i} s_{ij} \left\{(u_j - u_i)^- + (u_{ij} - u_i)^-\right\}, \tag{59}$$

where

$$u_{ij} = u_i + \nabla u_h|_{T_{ij}} \cdot (x_i - x_j) \qquad \forall\ j \in S_i\,, \tag{60}$$

and $T_{ij} \subset \Delta_i$ is a simplex intersected by the half line $\{x_i + \alpha\,(x_i - x_j)\,;\ \alpha > 0\}$ (like the triangle AEF in Fig. 10 for $x_i = A$ and $x_j = C$). As we will see below, this

modification of the MUAS method leads to optimal convergence rates in cases where the algebraic stabilizations of Section 4 provide suboptimal convergence results.

The above definitions of $P_i^{\pm}$ and $Q_i^{\pm}$ can be generalized to

$$P_i^+ = \sum_{j\in S_i,\, a_{ij}>0 \vee a_{ji}>0} p_{ij}\,\{(u_i-u_j)^+ + (u_i-u_{ij})^+\}\,, \tag{61}$$

$$P_i^- = \sum_{j\in S_i,\, a_{ij}>0 \vee a_{ji}>0} p_{ij}\,\{(u_i-u_j)^- + (u_i-u_{ij})^-\}\,, \tag{62}$$

$$Q_i^+ = \sum_{j\in S_i} q_{ij}\,\{(u_j-u_i)^+ + (u_{ij}-u_i)^+\}\,, \tag{63}$$

$$Q_i^- = \sum_{j\in S_i} q_{ij}\,\{(u_j-u_i)^- + (u_{ij}-u_i)^-\}\,, \tag{64}$$

with some weighting factors satisfying, for any $j \in S_i$, $i = 1,\dots,M$,

$$0 \le p_{ij} \le q_{ij}\,, \tag{65}$$

$$p_{ij} > 0 \quad \text{if} \quad a_{ij} > 0\,. \tag{66}$$

We name the resulting scheme Symmetrized Monotone Upwind-type Algebraically Stabilized (SMUAS) method. Let us recall that the stabilization matrix of the SMUAS method is given by (31) and (32) with β_{ij} determined by (36), (20), (21), and (61)–(64) where u_{ij} are defined by (60) and p_{ij}, q_{ij} satisfy (65) and (66).

Remark 1 If $P_i^+ = 0$, then R_i^+ can be defined arbitrarily (and the same holds for P_i^- and R_i^-). Indeed, P_i^+ is used only for defining β_{ij} with j such that $u_i > u_j$. Then, if $P_i^+ = 0$, one has $a_{ij} \le 0$ due to (66) and hence the matrix $\mathbb{B}(\mathrm{U})$ defined by (31), (32) does not depend on these β_{ij}.

Remark 2 The condition (65) assures that the SMUAS method is linearity preserving. Indeed, if $u_h \in P_1(\mathbb{R}^d)$, then $u_i - u_{ij} = u_j - u_i$ for any $i \in \{1,\dots,M\}$ and $j \in S_i$ and hence one gets

$$P_i^+ = -P_i^- \le \sum_{j\in S_i} p_{ij}\,|u_i-u_j| \le \sum_{j\in S_i} q_{ij}\,|u_i-u_j| = Q_i^+ = -Q_i^-\,,$$

so that $R_i^+ = R_i^- = 1$ for $i = 1,\dots,N$ and the stabilization term vanishes.

Remark 3 The condition $(a_{ij} > 0 \vee a_{ji} > 0)$ in (61) and (62) restricts the summation to those indices $j \in S_i$ for which $d_{ij} \ne 0$ (cf. (56) and (57)). This is important to obtain optimal convergence rates in the diffusion-dominated regime.

Of course, the properties of the SMUAS method depend on the choice of the weighting factors p_{ij}, q_{ij}. The relations (56)–(59) correspond to

$$p_{ij} = \max\{a_{ij}, 0, a_{ji}\}\,, \quad q_{ij} = \max\{|a_{ij}|, a_{ji}\}\,, \quad i = 1,\dots,M\,,\ j \in S_i\,. \tag{67}$$

Another possibility is to simply set

$$p_{ij} = q_{ij} = 1\,, \quad i = 1, \ldots, M\,, \;\; j \in S_i\,. \tag{68}$$

More generally, let us consider weighting factors satisfying $p_{ij} = q_{ij}$ for $i = 1, \ldots, M$ and $j \in S_i$ but not necessarily equal to 1. In the convection-dominated regime, the condition $(a_{ij} > 0 \vee a_{ji} > 0)$ usually holds for any $j \in S_i$ due to the skew-symmetry of the convection matrix and hence one obtains that $P_i^+ + Q_i^- = P_i^- + Q_i^+ = 0$. Thus, if $R_i^+ = R_i^- = 1$ for some $i \in \{1, \ldots, M\}$ (so that $\beta_{ij} = 0$ for all $j \in S_i$), one finds out that $u_i = \bar{u}_i$ where $\bar{u}_i$ is a local average defined by

$$\bar{u}_i = \frac{\sum_{j \in S_i} p_{ij}\,(u_j + u_{ij})}{2\sum_{j \in S_i} p_{ij}}\,,$$

see the derivation leading to (54). Therefore, the choice of the weights p_{ij} may be also guided by the requirement that the local averages $\bar{u}_i$ are good approximations of the values u_i for smoothly varying functions. Then, the weights p_{ij} should depend on the distribution of the nodes x_j, $j \in S_i$, and on their distances to x_i.

Remark 4 It is not always necessary to use all the additional terms in (61)–(64). For example, let us consider the patch around the node A in Fig. 6. Then, the nodes B, D and C, F are symmetric with respect to A. Therefore, using (68) and assuming that the condition $(a_{ij} > 0 \vee a_{ji} > 0)$ holds for any $j \in S_i$ and $x_i = A$, it is sufficient to introduce only symmetrically placed points to the nodes E and G. However, for simplicity of the presentation (and also of implementation), we do not consider such variants of the above formulas in this paper.

Now, let us prove that the SMUAS method satisfies Assumptions (A1) and (A2) from Section 3.

Theorem 4 *The stabilization matrix of the SMUAS method satisfies Assumption (A1).*

Proof In view of Theorem 3, it suffices to prove (33). Since $\beta_{ij} \equiv 0$ for any $j \in \{1, \ldots, N\}$ if $i \in \{M+1, \ldots, N\}$, consider any $i \in \{1, \ldots, M\}$. Let $j \in S_i$ be such that $a_{ij} > 0$. We want to show that $\Phi(\mathrm{U}) := \beta_{ij}(\mathrm{U})(u_j - u_i)$ is continuous at a fixed but arbitrary point $\bar{\mathrm{U}} = (\bar{u}_1, \ldots, \bar{u}_N) \in \mathbb{R}^N$. If $\bar{u}_i = \bar{u}_j$, then $\Phi(\bar{\mathrm{U}}) = 0$ and the continuity at $\bar{\mathrm{U}}$ follows from the estimates

$$|\Phi(\mathrm{U}) - \Phi(\bar{\mathrm{U}})| = |\Phi(\mathrm{U})| \le |u_i - u_j| \le \sqrt{2}\,\|\mathrm{U} - \bar{\mathrm{U}}\|\,,$$

where $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^N$. Thus, let $\bar{u}_i > \bar{u}_j$ and denote

$$B = \left\{ \mathrm{U} \in \mathbb{R}^N;\; \|\mathrm{U} - \bar{\mathrm{U}}\| \le \frac{1}{2}|\bar{u}_i - \bar{u}_j| \right\}.$$

Then, $u_i > u_j$ for $\mathrm{U} \in B$ and hence

$$\Phi(\mathrm{U}) = (1 - R_i^+(\mathrm{U}))\,(u_j - u_i) \qquad \forall\, \mathrm{U} \in B\,.$$

Since both P_i^+ and Q_i^+ are continuous and P_i^+ is positive in B due to (66), the function Φ is continuous in B and hence also at $\bar{\mathrm{U}}$. If $\bar{u}_i < \bar{u}_j$, one proceeds analogously. □

Theorem 5 *The stabilization matrix of the SMUAS method satisfies Assumption (A2).*

Proof Consider any $\mathrm{U} = (u_1, \ldots, u_N) \in \mathbb{R}^N$, $i \in \{1, \ldots, M\}$, and $j \in S_i$. Let u_i be a strict local extremum of U with respect to S_i. We want to prove that

$$a_{ij} + b_{ij}(\mathrm{U}) \le 0\,. \tag{69}$$

If $a_{ij} \le 0$, then (69) holds since $b_{ij}(\mathrm{U}) \le 0$. Thus, let $a_{ij} > 0$. First, assume that $u_i > u_k$ for any $k \in S_i$. Then, for any simplex $T \subset \Delta_i$ and any vector $\boldsymbol{a} \in \mathbb{R}^d$ pointing from x_i into T, one has $\boldsymbol{a} \cdot \nabla u_h|_T < 0$ with u_h defined by (3). Thus, $u_i > u_{ik}$ for any $k \in S_i$ according to (60), which implies that $Q_i^+ = 0$. Moreover, $P_i^+ \ge p_{ij}\,(u_i - u_j)^+ > 0$ in view of (66) and hence $\beta_{ij} = 1 - R_i^+ = 1$. Similarly, if $u_i < u_k$ for any $k \in S_i$, then also $u_i < u_{ik}$ for any $k \in S_i$ and hence $Q_i^- = 0$. Since $P_i^- \le p_{ij}\,(u_i - u_j)^- < 0$, one obtains $\beta_{ij} = 1 - R_i^- = 1$. Therefore, in both cases, $b_{ij}(\mathrm{U}) \le -a_{ij}$, which proves (69). □

The above theorems imply that the SMUAS method is solvable (cf. Theorem 1) and satisfies the DMP formulated in Theorem 2. Moreover, as shown in Remark 2, the SMUAS method is linearity preserving. It is important that all these properties hold for arbitrary simplicial meshes. For regular families of triangulations, one also has the error estimate (16).

However, as we have seen in Section 5, such theoretical properties do not guarantee that an algebraically stabilized method will provide an accurate approximate solution and that the approximate solutions will converge to the exact solution in usual norms. Thus, let us investigate the properties of the SMUAS method numerically. We start with Example 1 for $\varepsilon = 10^{-8}$ and Grid 4, for which suboptimal convergence results are presented in Table 2 for the AFC scheme with the Kuzmin limiter (that is equivalent to the MUAS method from Section 4.3 in this case). The results for the SMUAS method with p_{ij}, q_{ij} defined by (67) are shown in Table 4.

Table 4 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 4 computed using the SMUAS method with p_{ij}, q_{ij} defined by (67)

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	2.147e−2	1.61	4.734e−1	0.98	5.530e−2	1.92
32	6.353e−3	1.76	2.529e−1	0.90	1.479e−2	1.90
64	1.783e−3	1.83	1.363e−1	0.89	3.922e−3	1.92
128	4.706e−4	1.92	7.220e−2	0.92	1.054e−3	1.90
256	1.221e−4	1.95	3.807e−2	0.92	2.940e−4	1.84
512	3.135e−5	1.96	2.002e−2	0.93	7.896e−5	1.90

Table 5 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 4 computed using the SMUAS method with p_{ij}, q_{ij} defined by (68)

ne	$\|u-u_h\|_{0,\Omega}$	Order	$\|u-u_h\|_{1,\Omega}$	Order	$\|u-u_h\|_h$	Order
16	2.208e−2	1.60	4.748e−1	0.99	5.702e−2	1.91
32	6.605e−3	1.74	2.515e−1	0.92	1.530e−2	1.90
64	1.860e−3	1.83	1.336e−1	0.91	4.008e−3	1.93
128	4.924e−4	1.92	6.959e−2	0.94	1.046e−3	1.94
256	1.279e−4	1.95	3.635e−2	0.94	2.823e−4	1.89
512	3.291e−5	1.96	1.917e−2	0.92	7.358e−5	1.94

One observes a higher accuracy of the results than in Table 2 and the experimental convergence rates tend to the optimal values. Using the SMUAS method with p_{ij}, q_{ij} defined by (68) leads to similar results (see Table 5). Also for other test examples, the results obtained using (67) and (68) were similar and hence we will not present any other comparisons of results for these two choices of the weighting factors here. Similar convergence rates as in Tables 4 and 5 can be observed for all other grids from Figs. 1 and 2 (see Tables 6 and 7 for illustration). In Tables 6 and 7, ne represents the number of edges along the part of $\partial\Omega$ lying on the line $y = 1$, i.e., in Fig. 1, $ne = 2$ for Grid 2 and $ne = 4$ for Grid 3 so that the numbers of triangles in grids with the same value of ne are similar. The higher accuracy of the SMUAS method can be also seen from Fig. 11 if one compares it with Fig. 3.

It was reported in [6] for Example 1 that, in the diffusion-dominated case $\varepsilon = 10$, the solutions of the AFC scheme with the Kuzmin limiter do not converge for the non-Delaunay Grid 5 in any of the three norms considered in the above tables. Recall that Grid 5 was obtained from Grid 4 by shifting some of the nodes by $h/10$ to the right, where h is the horizontal mesh width in Grid 4. If the shift is $h/2$, then the experimental convergence rates tend to zero already on relatively coarse meshes (cf. [21]). The MUAS method from Section 4.3 shows an improved behavior. In particular, for the shift $h/2$, it leads to a convergence in all three norms and the convergence rates in the L^2 norm and the H^1 seminorm are near to the optimal values. However, if the shift is $0.8h$, then the accuracy deteriorates and the convergence rates

Table 6 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 2 computed using the SMUAS method with p_{ij}, q_{ij} defined by (67)

ne	$\|u-u_h\|_{0,\Omega}$	Order	$\|u-u_h\|_{1,\Omega}$	Order	$\|u-u_h\|_h$	Order
16	9.950e−3	1.75	2.432e−1	1.12	1.956e−2	2.10
32	2.764e−3	1.85	1.125e−1	1.11	4.638e−3	2.08
64	7.357e−4	1.91	5.210e−2	1.11	1.139e−3	2.03
128	1.889e−4	1.96	2.461e−2	1.08	2.812e−4	2.02
256	4.778e−5	1.98	1.189e−2	1.05	6.888e−5	2.03
512	1.202e−5	1.99	5.837e−3	1.03	1.709e−5	2.01

Table 7 Example 1: $\varepsilon = 10^{-8}$, numerical results for Grid 3 computed using the SMUAS method with p_{ij}, q_{ij} defined by (68)

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	2.218e−2	1.45	4.338e−1	1.00	5.994e−2	1.86
32	7.257e−3	1.61	2.175e−1	1.00	1.637e−2	1.87
64	2.093e−3	1.79	1.009e−1	1.11	4.066e−3	2.01
128	5.703e−4	1.88	4.635e−2	1.12	1.000e−3	2.02
256	1.520e−4	1.91	2.184e−2	1.09	2.524e−4	1.99
512	4.046e−5	1.91	1.046e−2	1.06	6.448e−5	1.97

tend to zero also for the MUAS method (cf. [21]). It is conjectured in [21] that this behavior is connected with the fact that the MUAS method is linearity preserving for the shift $h/2$ but not for $0.8h$. This conjecture is supported by the results obtained for the SMUAS method which is always linearity preserving and, indeed, leads to optimal convergence rates even on the distorted mesh corresponding to the shift $0.8h$ (see Table 8 and Fig. 12 (left)). Moreover, also a further deformation of the mesh leads to similar results (cf. Table 9 and Fig. 12 (middle) corresponding to the shift $1.5h$ (rightmost interior nodes on even horizontal mesh lines are shifted by $0.75h$)).

The SMUAS method seems to work well also on pathological meshes with very enlongated triangles. Table 10 shows results obtained on a triangular version of the mesh4_1 family from the FVCA5 benchmark [13]. This family consists of six meshes and the coarsest one is depicted in Fig. 12 (right). In contrast to the meshes of the type of Grid 5, the orientation of the edges in the 17 vertical strips of the coarsest mesh of the mesh4_1 family is preserved also in finer meshes. The value ne again represents the number of edges along the part of $\partial\Omega$ lying on the line $y = 1$ but now ne increases linearly ($ne = 17\,i$ in the ith row of Table 10) whereas in all the tables considered before the increase was exponential ($ne = 8 \cdot 2^i$ in the ith row). Consequently, the convergence orders in Table 10 tend to the optimal values slower.

It is not surprising that the SMUAS method provides the exact solution on any mesh if it is applied to Example 2. For Example 3 and Grid 4, the solution of the SMUAS method is nodally exact except for the rightmost vertical interior grid line, similarly as for the AFC scheme with the Kuzmin limiter and Grid 1 in Fig. 7 (left).

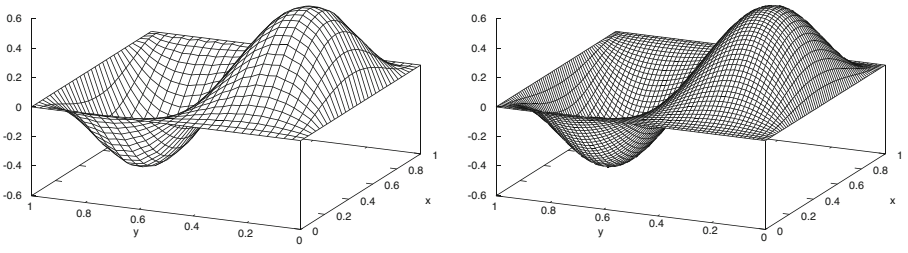


Fig. 11 Example 1: $\varepsilon = 10^{-8}$, approximate solutions computed using the SMUAS method with p_{ij}, q_{ij} defined by (68) on Grid 4 with $ne = 32$ (left) and $ne = 64$ (right)

Table 8 Example 1: $\varepsilon = 10$, numerical results computed using the SMUAS method with p_{ij}, q_{ij} defined by (67) on triangulations of the type of Grid 5 obtained by shifting the respective interior nodes by eight tenths of the horizontal mesh width

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	3.155e−2	1.65	5.855e−1	0.83	1.976e+0	0.90
32	7.267e−3	2.12	3.002e−1	0.96	9.676e−1	1.03
64	1.665e−3	2.13	1.518e−1	0.98	4.826e−1	1.00
128	4.111e−4	2.02	7.642e−2	0.99	2.420e−1	1.00
256	1.048e−4	1.97	3.837e−2	0.99	1.214e−1	1.00
512	2.659e−5	1.98	1.922e−2	1.00	6.080e−2	1.00

Also for Example 4, the SMUAS method on Grid 4 provides an approximate solution which is nodally exact in most of the computational domain (see Fig. 13 (left)). The approximation of the boundary layers should be improved by local mesh refinement. Finally, also in case of Example 5, an application of the SMUAS method on Grid 4 leads to a much more accurate approximate solution than the AFC scheme with the Kuzmin limiter (see Figs. 13 (right) and 9 (right)). Moreover, the SMUAS method again shows optimal convergence rates.

Summarizing our numerical results, one can state that the SMUAS method led to optimal convergence rates in all our numerical tests involving various types of simplicial meshes whereas, in many cases, the algebraic stabilizations from Section 4 lead to suboptimal convergence rates or do not converge at all. A theoretical explanation of the observed optimal convergence behavior of the SMUAS method is left to future work.

The properties of the SMUAS method proved in this paper together with the observed optimal experimental convergence rates make the SMUAS method superior to the remaining three algebraically stabilized methods discussed in this paper. This is also illustrated by Table 11 where the basic properties of all four methods are compared.

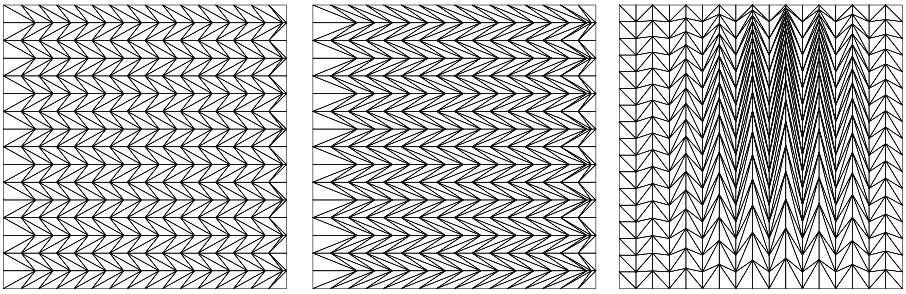


Fig. 12 Grids used for computing the results in Table 8 (left, $ne = 16$), Table 9 (middle, $ne = 16$), and Table 10 (right, $ne = 17$)

Table 9 Example 1: $\varepsilon = 10$, numerical results computed using the SMUAS method with p_{ij}, q_{ij} defined by (67) on triangulations of the type of Grid 5 obtained by shifting the respective interior nodes by one and half of the horizontal mesh width

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
16	7.338e−2	1.10	8.767e−1	0.63	3.620e+0	0.74
32	2.287e−2	1.68	4.854e−1	0.85	1.732e+0	1.06
64	5.395e−3	2.08	2.506e−1	0.95	8.234e−1	1.07
128	1.319e−3	2.03	1.281e−1	0.97	4.091e−1	1.01
256	3.541e−4	1.90	6.501e−2	0.98	2.062e−1	0.99
512	9.619e−5	1.88	3.275e−2	0.99	1.037e−1	0.99

Table 10 Example 1: $\varepsilon = 10$, numerical results computed using the SMUAS method with p_{ij}, q_{ij} defined by (67) on a triangular version of the mesh4_1 family from the FVCA5 benchmark

ne	$\|u - u_h\|_{0,\Omega}$	Order	$\|u - u_h\|_{1,\Omega}$	Order	$\|u - u_h\|_h$	Order
17	1.847e−1		1.532e+0		1.021e+1	
34	1.295e−1	0.52	1.210e+0	0.35	7.641e+0	0.43
51	9.021e−2	0.90	9.790e−1	0.53	5.663e+0	0.75
68	6.288e−2	1.26	8.018e−1	0.70	4.329e+0	0.94
85	4.504e−2	1.50	6.684e−1	0.82	3.459e+0	1.01
102	3.344e−2	1.64	5.681e−1	0.90	2.881e+0	1.01

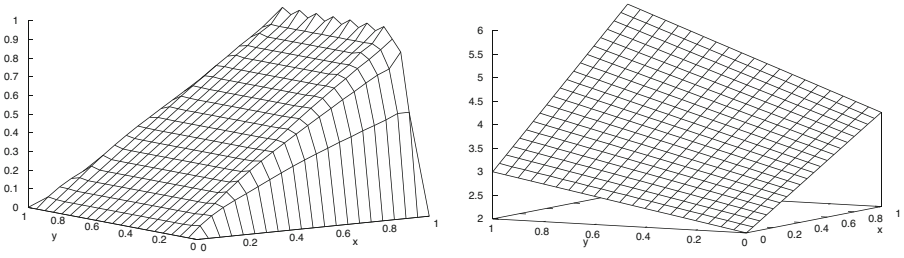


Fig. 13 Approximate solutions computed using the SMUAS method with p_{ij}, q_{ij} defined by (67) on Grid 4 with $ne = 20$ for Example 4 (left) and Example 5 (right)

Table 11 Comparison of the properties of the AFC scheme with the Kuzmin limiter (Section 4.1), the AFC scheme with the BJK limiter (Section 4.2), the MUAS method (Section 4.3), and the SMUAS method (Section 6)

	Kuzmin	BJK	MUAS	SMUAS
DMP on arbitrary meshes	✗	✓	✓	✓
Linearity preservation	✗	✓	✗	✓
Method of upwind type	✓	✗	✓	✓
Optimal convergence rates	✗	✗	✗	✓

Remark 5 To compute solutions of the algebraic stabilizations considered in this paper, systems of nonlinear algebraic equations have to be solved. For the AFC schemes with the Kuzmin and BJK limiters, various nonlinear solvers have been studied in [17] and it turned out that a simple fixed point iteration called *fixed point rhs* was the most efficient method. This approach was also used to compute the results presented in this paper. We observed that the convergence of this nonlinear solver for the SMUAS method was very similar as for the AFC scheme with the Kuzmin limiter and we refer to [17, 22] where detailed convergence studies for *fixed point rhs* applied to the AFC scheme with the Kuzmin limiter can be found for various test problems. These convergence studies investigate the influence of the mesh width on the numbers of iterations and computing times and include test problems with interior layers and a non-constant convection field.

Funding Open access publishing supported by the National Technical Library in Prague. This work has been supported through the grant no. 22-01591S of the Czech Science Foundation.

Declarations

Competing interests The author declares no competing interests.

References

1. Augustin, M., Caiazzo, A., Fiebach, A., Fuhrmann, J., John, V., Linke, A., Umla, R.: An assessment of discretizations for convection-dominated convection–diffusion equations. Comput. Methods Appl. Mech. Engrg. **200**(47-48), 3395–3409 (2011)
2. Badia, S., Bonilla, J.: Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. Comput. Methods Appl. Mech. Engrg. **313**, 133–158 (2017)

3. Barrenechea, G.R., Burman, E., Karakatsani, F.: Blending low-order stabilised finite element methods: A positivity-preserving local projection method for the convection–diffusion equation. Comput. Methods Appl. Mech. Engrg. **317**, 1169–1193 (2017)
4. Barrenechea, G.R., Burman, E., Karakatsani, F.: Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. Numer. Math. **135**(2), 521–545 (2017)
5. Barrenechea, G.R., John, V., Knobloch, P.: 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)
6. Barrenechea, G.R., John, V., Knobloch, P.: Analysis of algebraic flux correction schemes. SIAM J. Numer. Anal. **54**(4), 2427–2451 (2016)
7. Barrenechea, G.R., John, V., Knobloch, P.: An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. Math. Models Methods Appl. Sci. **27**(3), 525–548 (2017)
8. Barrenechea, G.R., John, V., Knobloch, P.: Finite element methods respecting the discrete maximum principle for convection-diffusion equations. arXiv:2204.07480 (2022)
9. Barrenechea, G.R., John, V., Knobloch, P., Rankin, R.: A unified analysis of algebraic flux correction schemes for convection–diffusion equations. SeMA J. **75**(4), 655–685 (2018)
10. Boris, J.P., Book, D.L.: Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. J. Comput. Phys. **11**(1), 38–69 (1973)
11. Ciarlet, P.G.: The finite element method for elliptic problems. North-Holland, Amsterdam (1978)
12. Evans, L.C.: Partial differential equations, 2nd edn. American Mathematical Society, Providence, RI (2010)
13. FVCA5 benchmark, 5th International Symposium on Finite Volumes for Complex Applications, June 8–13, 2008, Aussois, France. https://www.i2m.univ-amu.fr/fvca5/benchmark/Meshes/
14. Gilbarg, D., Trudinger, N.S.: Elliptic partial differential equations of second order. Springer, Berlin (2001)
15. Guermond, J.-L., Nazarov, M., Popov, B., Yang, Y.: A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations. SIAM J. Numer. Anal. **52**(4), 2163–2182 (2014)
16. Gurris, M., Kuzmin, D., Turek, S.: Implicit finite element schemes for the stationary compressible Euler equations. Int. J. Numer. Methods Fluids **69**(1), 1–28 (2012)
17. Jha, A., John, V.: A study of solvers for nonlinear AFC discretizations of convection–diffusion equations. Comput. Math. Appl. **78**(9), 3117–3138 (2019)
18. 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**(17-20), 2197–2215 (2007)
19. John, V., Knobloch, P.: On the performance of SOLD methods for convection-diffusion problems with interior layers. Int. J. Comput. Sci. Math. **1**(2-4), 245–258 (2007)
20. 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. Engrg. **197**(21-24), 1997–2014 (2008)
21. John, V., Knobloch, P.: On algebraically stabilized schemes for convection–diffusion–reaction problems. Numer. Math. **152**(3), 553–585 (2022)
22. John, V., Knobloch, P., Pártl, O.: A numerical assessment of finite element discretizations for convection-diffusion-reaction equations satisfying discrete maximum principles. Comput. Methods Appl. Math. https://doi.org/10.1515/cmam-2022-0125 (2022)
23. John, V., Schmeyer, E.: Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. Comput. Methods Appl. Mech. Engrg. **198**(3-4), 475–494 (2008)
24. Knobloch, P.: On the application of algebraic flux correction schemes to problems with non-vanishing right-hand side. In: Knobloch, P. (ed.) Boundary and Interior Layers, Computational and Asymptotic Methods – BAIL 2014. Lect. Notes Comput. Sci. Eng., vol. 108, pp. 99–109. Springer, Cham (2015)
25. Knobloch, P.: On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. In: Huang, Z., Stynes, M., Zhang, Z. (eds.) Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016. Lect. Notes Comput. Sci. Eng., vol. 120, pp. 129–139. Springer, Cham (2017)

26. Knobloch, P.: A linearity preserving algebraic flux correction scheme of upwind type satisfying the discrete maximum principle on arbitrary meshes. In: Radu, F.A., Kumar, K., Berre, I., Nordbotten, J.M., Pop, I.S. (eds.) Numerical Mathematics and Advanced Applications ENUMATH 2017. Lect. Notes Comput. Sci. Eng., vol. 126, pp. 909–918. Springer, Cham (2019)
27. Knobloch, P.: A new algebraically stabilized method for convection–diffusion–reaction equations. In: Vermolen, F.J., Vuik, C. (eds.) Numerical Mathematics and Advanced Applications ENUMATH 2019. Lect. Notes Comput. Sci. Eng., vol. 139, pp. 605–613. Springer, Cham (2021)
28. Kuzmin, D.: On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. J. Comput. Phys. **219**(2), 513–531 (2006)
29. Kuzmin, D.: Algebraic flux correction for finite element discretizations of coupled systems. In: Papadrakakis, M., Oñate, E., Schrefler, B. (eds.) Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, pp. 653–656. CIMNE, Barcelona (2007)
30. Kuzmin, D.: Explicit and implicit FEM-FCT algorithms with flux linearization. J. Comput. Phys. **228**(7), 2517–2534 (2009)
31. Kuzmin, D.: Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. J. Comput. Appl. Math. **236**(9), 2317–2337 (2012)
32. Kuzmin, D.: Algebraic flux correction I. Scalar conservation laws. In: Kuzmin, D., Löhner, R., Turek, S. (eds.) Flux-Corrected Transport. Principles, Algorithms, and Applications. 2nd edn., pp. 145–192. Springer, Dordrecht (2012)
33. Kuzmin, D., Shadid, J.N.: Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations. Int. J. Numer. Methods Fluids **84**(11), 675–695 (2017)
34. Kuzmin, D., Turek, S.: High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. J. Comput. Phys. **198**(1), 131–158 (2004)
35. Lohmann, C., Kuzmin, D., Shadid, J.N., Mabuza, S.: Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. J. Comput. Phys. **344**, 151–186 (2017)
36. Roos, H.-G., Stynes, M., Tobiska, L.: Robust numerical methods for singularly perturbed differential equations. Convection–diffusion–reaction and flow problems. Springer, Berlin (2008)
37. Zalesak, S.T.: Fully multidimensional flux-corrected transport algorithms for fluids. J. Comput. Phys. **31**(3), 335–362 (1979)