Weierstraß-Institut
für Angewandte Analysis und Stochastik
Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

A unified analysis of Algebraic Flux Correction schemes for convection-diffusion equations

Gabriel R. Barrenechea[1], Volker John[2], Petr Knobloch[3], Richard Rankin[4]

submitted: January 25, 2018

[1] University of Strathclyde
Department of Mathematics and Statistics
26 Richmond Street
Glasgow G1 1XH
Scotland
E-Mail: gabriel.barrenechea@strath.ac.uk

[2] Weierstrass Institute
Mohrenstr. 39, 10117 Berlin, Germany
and Free University of Berlin
Dep. of Mathematics and Computer Science
Arnimallee 6, 14195 Berlin, Germany
E-Mail: volker.john@wias-berlin.de

[3] Charles University in Prague
Faculty of Mathematics and Physics
Department of Numerical Mathematics
Sokolovská 83, 18675 Praha 8, Czech Republic
E-Mail: knobloch@karlin.mff.cuni.cz

[4] The University of Nottingham Ningbo China
School of Mathematical Sciences
199 Taikang East Road, Ningbo, 315100
China
E-Mail: Richard.Rankin@nottingham.edu.cn

No. 2475

Berlin 2018

2010 *Mathematics Subject Classification.* 65N30.

Key words and phrases. scalar convection-diffusion equations, algebraic stabilizations, edge-based diffusion scheme, discrete maximum principle, error analysis.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

A unified analysis of Algebraic Flux Correction schemes for convection-diffusion equations

Gabriel R. Barrenechea, Volker John, Petr Knobloch, Richard Rankin

Abstract

Recent results on the numerical analysis of Algebraic Flux Correction (AFC) finite element schemes for scalar convection-diffusion equations are reviewed and presented in a unified way. A general form of the method is presented using a link between AFC schemes and nonlinear edge-based diffusion scheme. Then, specific versions of the method, this is, different definitions for the flux limiters, are reviewed and their main results stated. Numerical studies compare the different versions of the scheme.

1 Introduction

Scalar convection-diffusion equations model the convective and molecular transport of a quantity like temperature or concentration. In applications, the convective transport is usually dominant, which is the case of interest in this paper.

Here, we consider the steady-state situation, where the mathematical problem is formulated as follows: Find $u : \overline{\Omega} \to \mathbb{R}$ such that

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

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded polygonal or polyhedral domain with a Lipschitz continuous boundary $\partial\Omega$, $\varepsilon > 0$ is a constant diffusion coefficient, $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$ is a solenoidal convection field, $c \in L^\infty(\Omega)$ is a non-negative reaction coefficient, $g \in L^2(\Omega)$ is an outer source of the quantity u, and $u_D \in H^{\frac{1}{2}}(\partial\Omega) \cap C^0(\partial\Omega)$ is a boundary datum.

A characteristic feature of solutions of (1) is the appearance of layers, i.e., of narrow regions where the solution has a large gradient. These regions are usually so narrow that the layers cannot be resolved by affordable grids. It is well known that standard discretizations cannot cope with this situation and they lead to meaningless numerical solutions that are globally polluted with huge spurious oscillations. The remedy consists in using stabilized discretizations. In the context of finite element methods, the proposal of the streamline-upwind Petrov–Galerkin (SUPG) method in [13, 18] was the first milestone in this direction. Solutions computed with this method have usually sharp layers at the correct position, but there are still non-negligible spurious oscillations in a vicinity of layers. Since the publication of [13, 18] the development and analysis of stabilized discretizations for convection-dominated equations has been an active field of research.

In this research, one can distinguish two directions. The first one is the development of stabilized methods with a provable order of convergence in appropriate norms. Examples of this direction are the continuous interior penalty (CIP) method (see, e.g. [14]) and the local projection stabilization (LPS) method (see [10] for the first application of this method to a convection-dominated equation). The second direction consists in finding stabilized methods that compute solutions without spurious oscillations

DOI 10.20347/WIAS.PREPRINT.2475

and still with sharp layers. The property of being free of spurious oscillations can be expressed mathematically with the satisfaction of the discrete maximum principle (DMP). Usually, the satisfaction of the DMP is proved by the sufficient condition that the matrix of a linear discretization[1] is an M-matrix. However, it is well known that, in the limit case $\varepsilon = 0$, there is a barrier of the order of the local discretization error for linear discretizations with M-matrices: these discretizations are at most of first order, e.g., see [37, Thm. 4.2.2].

Since the property of being free of spurious oscillations might be of utmost importance for a method to be applicable in practice, a significant amount of work has been devoted to the development of such methods. Due to the order barrier for linear discretizations, nonlinear discretizations became of interest. One further argument in favor of using nonlinear discretizations for a convection-dominated problem stems from the fact that most of the applications in which convection dominates are modeled by nonlinear partial differential equations. Then, the use of a nonlinear discretization does not constitute a significant overhead. Since the late 1980s, there have been a number of proposals to remove the spurious oscillations of the SUPG method by adding appropriate nonlinear terms. This class of methods is called spurious oscillations at layers diminishing (SOLD) methods, or shock capturing methods. A comprehensive review was carried out in the companion papers [19, 20], and the main conclusion of it was that none of the proposed SOLD methods reduced the spurious oscillations sufficiently well.

Algebraic stabilizations, so-called Algebraic Flux Correction (AFC) schemes, became of interest to us as a result of numerical assessments of stabilized discretizations in [5, 19, 24, 23]. The main motivation for the design of AFC methods is the satisfaction of the DMP. In addition, they provide reasonably sharp approximations of the layers. In contrast to SOLD methods, which are based on variational formulations, the main idea of AFC schemes consists in modifying the algebraic system corresponding to a discrete problem, typically the Galerkin discretization, by means of solution-dependent flux corrections. Consequently, AFC schemes are nonlinear. The basic philosophy of flux correction schemes was formulated already in [12, 38]. Later, the idea was extended to the finite element context, e.g., in [4, 34]. In the last fifteen years, there has been an intensive development of these methods, e.g., see [27, 28, 30, 31, 32].

None of the above references deals with the mathematical analysis of the AFC methods. In fact, the first contributions to the numerical analysis of AFC schemes were presented only recently in [7, 8, 9]. The first paper [7] focuses on the solvability of the nonlinear scheme, while [8] presents the first error analysis of the AFC schemes. Interestingly, the paper [8] also presented negative results, in the sense that it was shown that unless some restrictions are imposed in the mesh, the numerical scheme may not converge. Finally, in the recent paper [9] the role of the linearity preservation was studied. This study is also complemented by the work [6], where a link between the AFC schemes and a nonlinear edge-based diffusion scheme is presented, and the linearity preservation of the scheme is also studied in detail. This latter reformulation offers the applicability of different tools than used so far for the analysis of AFC schemes. In particular, it facilitated the a posteriori error analysis of the AFC method, presented in [2]. Thus, the present paper aims at providing a review of these works, and performing the analysis in a unified framework.

The rest of the manuscript is organized as follows. After having introduced AFC methods in Section 2, a rewriting as an edge diffusion scheme is presented. A unified analysis is given in Section 3, covering the existence of a solution, minimal conditions for the validity of the DMP, and finite element error estimates. Three definitions of limiters are provided in Section 4. Strategies for the solution of the nonlinear problems are discussed in Section 5. Numerical studies for different limiters used in AFC schemes and the edge diffusion scheme proposed in [6] are presented in Section 6. Finally, Section 7

[1] A linear discretization of (1) is a discretization that leads to a linear system of equations.

states the most important open problems in the field of AFC schemes.

2 The model problem and a unified presentation of AFC schemes

The weak formulation of (1) reads: Find $u \in H^1(\Omega)$ such that $u|_{\partial\Omega} = u_D$, and

$$a(u,v) = (g,v)_\Omega \qquad \forall\, v \in H_0^1(\Omega)\,, \tag{2}$$

where $(\cdot,\cdot)_\Omega$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$ and the bilinear form $a(\cdot,\cdot)$ is given by

$$a(u,v) = \varepsilon(\nabla u, \nabla v)_\Omega + (\boldsymbol{b}\cdot\nabla u, v)_\Omega + (c\,u, v)_\Omega\,. \tag{3}$$

Thanks to the Poincaré inequality, and the fact that $\boldsymbol{b}$ is solenoidal and c is non-negative, this problem has a unique solution.

To discretize the problem (1), we introduce the following notation:

- $\{\mathscr{T}_h\}_{h>0}$ denotes a family of shape regular simplicial triangulations of $\overline{\Omega}$.
- For a given triangulation $\mathscr{T}_h$, $\mathscr{E}_h$ denotes the set of its internal edges.
- For every edge $E \in \mathscr{E}_h$, we denote by h_E the length of E and by $\boldsymbol{x}_{E,1}$, $\boldsymbol{x}_{E,2}$ the endpoints of E. Furthermore, for every $E \in \mathscr{E}_h$, we choose one unit tangent vector $\boldsymbol{t}_E$. Its orientation is of no importance.
- For every edge $E \in \mathscr{E}_h$, we define the neighborhood $\omega_E := \cup\{T \in \mathscr{T}_h : T \cap E \neq \emptyset\}$.
- For a given triangulation $\mathscr{T}_h$, $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}$ is the set of its nodes. We will assume that the nodes $\boldsymbol{x}_1,\dots,\boldsymbol{x}_M$ are the internal nodes, and $\boldsymbol{x}_{M+1},\dots,\boldsymbol{x}_N$ are boundary nodes, i.e., the nodes where the Dirichlet boundary condition is imposed.
- For a node $\boldsymbol{x}_i$, $i = 1,\dots,N$, we define

  $$\mathscr{E}_i := \{E \in \mathscr{E}_h : \boldsymbol{x}_i \text{ is an endpoint of } E\}\,.$$

- For a node $\boldsymbol{x}_i$, $i = 1,\dots,N$, we define $\Delta_i := \{T \in \mathscr{T}_h : \boldsymbol{x}_i \in T\}$.
- For an interior node $\boldsymbol{x}_i$, $i = 1,\dots,M$, we define the index set of its neighbors

  $$S_i := \{j \in \{1,\dots,N\}\setminus\{i\} \;:\; \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are endpoints of the same internal edge } E \in \mathscr{E}_h\}\,.$$

- The finite element spaces used in this work are given by

  $$V_h := \{v_h \in C^0(\overline{\Omega}) : v_h|_T \in \mathbb{P}_1(T)\ \ \forall T \in \mathscr{T}_h\}\,, \qquad V_{h,0} := V_h \cap H_0^1(\Omega)\,.$$

- These spaces have standard nodal basis functions denoted by $\{\varphi_1,\dots,\varphi_N\}$, uniquely determined by the conditions $\varphi_i(\boldsymbol{x}_j) = \delta_{ij}$ for all $i,j = 1,\dots,N$. We further notice that $\operatorname{supp}\varphi_i = \Delta_i$.

- The Lagrange interpolation operator $i_h : C^0(\overline{\Omega}) \to V_h$ is given by

$$i_h v = \sum_{i=1}^{N} v(\boldsymbol{x}_i)\,\varphi_i\,.$$

In addition, we set

$$i_h u_D = \sum_{i=M+1}^{N} u_D(\boldsymbol{x}_i)\,\varphi_i|_{\partial\Omega}\,.$$

The Galerkin scheme associated to (2) is given as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h u_D$, and

$$a(u_h, v_h) = (g, v_h)_\Omega \qquad \forall\, v_h \in V_{h,0}\,. \tag{4}$$

This scheme is well known to lead to inaccurate results on affordable grids.

The first step towards the building of an AFC scheme is the writing of the Galerkin method (4) in matrix form. For this, we introduce the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, where $a_{ij} = a(\varphi_j, \varphi_i)$. Then, we represent the discrete solution by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to the basis $\{\varphi_1, \dots, \varphi_N\}$ of V_h. Then $\mathrm{U} \equiv (u_1, \dots, u_N)$ satisfies the following system of linear equations:

$$\sum_{j=1}^{N} a_{ij}\, u_j = g_i\,, \qquad i = 1, \dots, M\,, \tag{5}$$

$$u_i = u_D(\boldsymbol{x}_i)\,, \qquad i = M+1, \dots, N\,, \tag{6}$$

where $g_i = (g, \varphi_i)_\Omega$ for $i = 1, \dots, M$. Thanks to the ellipticity of $a(\cdot,\cdot)$ on $V_{h,0}$, the matrix $(a_{ij})_{i,j=1}^M$ is positive definite, i.e.,

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j > 0 \qquad \forall\ (u_1, \dots, u_M) \in \mathbb{R}^M \setminus \{0\}\,. \tag{7}$$

Using the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, we introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ with entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\ i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j\neq i} d_{ij}\,. \tag{8}$$

The first step of defining an AFC scheme is then to add artificial diffusion to the algebraic system. More precisely, the problem (4) is replaced by

$$(\mathbb{A}\,\mathrm{U})_i + (\mathbb{D}\,\mathrm{U})_i = g_i\,, \qquad\qquad i = 1, \dots, M\,, \tag{9}$$

$$u_i = u_D(\boldsymbol{x}_i)\,, \qquad i = M+1, \dots, N\,. \tag{10}$$

In practice, the solution of such a perturbed scheme, which corresponds to simple upwinding, is too diffusive to be of interest. Then, the aim of AFC schemes is to localize this added diffusion in such a way that the DMP is respected, while the internal and boundary layers are not too smeared. This requires a finer analysis of the structure of the product $\mathbb{D}\,\mathrm{U}$. Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\,\mathrm{U})_i = \sum_{j\neq i} f_{ij}\,, \qquad i = 1, \dots, N\,,$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \dots, N$. Then, a further rewriting of (9) reads as follows:

$$(\mathbb{A}\,\mathrm{U})_i + \sum_{j=1}^{N} f_{ij} = g_i\,, \qquad i = 1, \dots, M\,,$$
$$u_i = u_D(\boldsymbol{x}_i)\,, \qquad i = M+1, \dots, N\,.$$

The next fundamental step in the building of an AFC scheme is to limit the fluxes f_{ij}. In other words, the idea is to localize the diffusion to the areas surrounding extrema and layers. To this end, we introduce solution-dependent correction factors (or flux limiters) $\beta_{ij} \in [0,1]$, and replace system (9) by

$$\sum_{j=1}^{N} a_{ij}\,u_j + \sum_{j=1}^{N} \beta_{ij}(\mathrm{U})\,d_{ij}\,(u_j - u_i) = g_i\,, \qquad i = 1, \dots, M\,, \tag{11}$$
$$u_i = u_D(\boldsymbol{x}_i)\,, \qquad i = M+1, \dots, N\,. \tag{12}$$

For $\beta_{ij} = 0$, the original system (5) is recovered. Hence, intuitively, the coefficients β_{ij} should be as close to 0 as possible to limit the modifications of the original problem. So far, these coefficients have been chosen in various ways, and their definition is always based on the fluxes f_{ij}. To guarantee that the resulting scheme is conservative, and to be able to show existence of solutions, one should require that the coefficients β_{ij} are symmetric, i.e.,

$$\beta_{ij} = \beta_{ji}\,, \qquad i, j = 1, \dots, N\,.$$

This requirement also has a mathematical justification. As a matter of fact, in [7], the possible non-existence of solutions has been shown if this restriction is ignored. Note that (11) does not involve β_{ij} with $j \in \{M+1, \dots, N\}$ and hence these values can be chosen arbitrarily. We define them by the above symmetry condition and by the requirement that $\beta_{ij} = 0$ if $i, j \in \{M+1, \dots, N\}$.

2.1 A variational formulation and a rewriting as an edge diffusion scheme

Our starting point is the following variational formulation presented in [8] for problem (11), (12): Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h u_D$, and

$$a(u_h, v_h) + D_h(u_h; u_h, v_h) = (g, v_h)_\Omega \qquad \forall\, v_h \in V_{h,0}\,. \tag{13}$$

Here, the nonlinear form $D_h(\cdot;\cdot,\cdot)$ is given by

$$D_h(z; v, w) = \sum_{i,j=1}^{N} \beta_{ij}(z)\,d_{ij}\,(v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i))\,w(\boldsymbol{x}_i)\,.$$

We now rewrite this nonlinear form using the symmetry of d_{ij} and β_{ij}:

$$\begin{aligned} & D_h(z;v,w) \\ &= \sum_{i>j} \beta_{ij}(z)\, d_{ij}\, (v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i)) w(\boldsymbol{x}_i) + \sum_{i<j} \beta_{ij}(z)\, d_{ij}\, (v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i)) w(\boldsymbol{x}_i) \\ &= \sum_{i>j} \beta_{ij}(z)\, d_{ij}\, (v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i)) w(\boldsymbol{x}_i) + \sum_{i>j} \beta_{ji}(z)\, d_{ji}\, (v(\boldsymbol{x}_i) - v(\boldsymbol{x}_j)) w(\boldsymbol{x}_j) \\ &= \sum_{i>j} \beta_{ij}(z)\, d_{ij}\, (v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i))(w(\boldsymbol{x}_i) - w(\boldsymbol{x}_j)) \\ &= \sum_{E \in \mathscr{E}_h} \beta_E(z)\, |d_E|\, (v(\boldsymbol{x}_{E,1}) - v(\boldsymbol{x}_{E,2}))(w(\boldsymbol{x}_{E,1}) - w(\boldsymbol{x}_{E,2}))\,, \end{aligned}$$

where we have denoted $\beta_E = \beta_{ij} = \beta_{ji}$ and $d_E = d_{ij} = d_{ji}$ for any edge $E \in \mathscr{E}_h$ that has the endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$.

Hence, with this rewriting of D_h, we can state the following general form of an AFC scheme: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h u_D$, and

$$a_h(u_h; u_h, v_h) = (g, v_h)_\Omega \qquad \forall\, v_h \in V_{h,0}\,, \tag{14}$$

where $a_h(z; v, w) = a(v, w) + D_h(z; v, w)$ with $a(\cdot,\cdot)$ being defined in (3) and $D_h(\cdot;\cdot,\cdot)$ given by

$$D_h(z;v,w) = \sum_{E \in \mathscr{E}_h} \beta_E(z)\, |d_E|\, (v(\boldsymbol{x}_{E,1}) - v(\boldsymbol{x}_{E,2}))(w(\boldsymbol{x}_{E,1}) - w(\boldsymbol{x}_{E,2}))\,. \tag{15}$$

Since, for any function from V_h, the restriction to an any edge E of $\mathscr{E}_h$ is a linear function, one has

$$D_h(z;v,w) = \sum_{E \in \mathscr{E}_h} \beta_E(z)\, |d_E|\, h_E\, (\nabla v \cdot \boldsymbol{t}_E, \nabla w \cdot \boldsymbol{t}_E)_E \qquad \forall\; v, w \in V_h\,. \tag{16}$$

From now on we will suppose that, for every edge $E \in \mathscr{E}_h$, d_E is a real number, not necessarily linked to the matrix $\mathbb{A}$. This flexibility will allow us to include a wider class of methods in our presentation.

The solution-dependent limiters β_E are still assumed to satisfy $\beta_E \in [0,1]$ and to assure the solvability of (14) (see the next section), we further make the following continuity assumption:

Assumption (A1): For any $E \in \mathscr{E}_h$, the function $\beta_E(u_h)(\nabla u_h)|_E \cdot \boldsymbol{t}_E$ is a continuous function of $u_h \in V_h$.

It will be shown in Section 4 that the limiters defined in [6, 8, 9] satisfy Assumption (A1).

Remark 1 The fact that the restriction of the functions v and w to the internal edges is a linear function is what makes it possible to obtain the expression (16) for D_h. This property also holds for unmapped $\mathbb{Q}_1$ finite elements, and for mapped $\mathbb{Q}_1$ finite elements on parallelepipeds, although in that case this methodology would lead to a completely different method, as the cross-terms would not be included in the method. The implications of this remark are the topic of current investigations and are to be reported elsewhere. On a related note to the previous point, AFC-related schemes using higher order elements combined with Bernstein basis functions have been developed recently in the work [33], but the full stability and error analysis of the methods is lacking.

3 General properties of the nonlinear scheme

In this section we present the main results associated to the nonlinear scheme (14). More precisely, we present results on its solvability, minimal conditions for the validity of the discrete maximum principle, and a first error estimate for the method. In the following section the conditions imposed herein will be checked for different definitions of the limiters β_E.

3.1 Existence of solutions

Lemma 1 (Consequence of Brouwer's fixed-point theorem) *Let X be a finite-dimensional Hilbert space with inner product $(\cdot,\cdot)_X$ and norm $\|\cdot\|_X$. Let $T : X \to X$ be a continuous mapping and $K > 0$ a real number such that $(Tx, x)_X > 0$ for any $x \in X$ with $\|x\|_X = K$. Then there exists $x \in X$ such that $\|x\|_X < K$ and $Tx = 0$.*

A proof of Lemma 3.1 can be found in [35, p. 164, Lemma 1.4]. Now, the existence of solutions for the nonlinear scheme (14) can be proved.

Theorem 1 (Existence of a solution of (14)**)** *If Assumption (A1) holds, then there exists a solution u_h of* (14).

Proof For this proof only, we will consider constants $C > 0$ that may depend on the data of (1) and h. In addition, we will make use of a function $u_{h,D} \in V_h$, which is an extension of the boundary datum $i_h u_D$. Let us first define the nonlinear mapping $T : V_{h,0} \to [V_{h,0}]'$ by

$$\langle Tv_h, w_h\rangle := a(v_h + u_{h,D}, w_h) + D_h(v_h + u_{h,D}; v_h + u_{h,D}, w_h) - (g, w_h)_\Omega \,.$$

Since $a(\cdot,\cdot)$ is a continuous bilinear form, Assumption (A1) implies that T is a continuous mapping. Next, from the definition of $a(\cdot,\cdot)$, it follows that, for any $v_h \in V_{h,0}$,

$$a(v_h, v_h) = \varepsilon\,|v_h|^2_{1,\Omega} + (c\,v_h, v_h) \ge \varepsilon\,|v_h|^2_{1,\Omega}\,.$$

Moreover, (16) and the fact that $\beta_E(v_h + u_{h,D}) \ge 0$ give

$$D_h(v_h + u_{h,D}; v_h, v_h) = \sum_{E\in\mathscr{E}_h} \beta_E(v_h + u_{h,D})\,|d_E|\,h_E\,\|\nabla v_h\cdot \boldsymbol{t}_E\|^2_{0,E} \ge 0.$$

Then, the definition of the operator T yields

$$\langle Tv_h, v_h\rangle \ge \varepsilon|v_h|^2_{1,\Omega} + a(u_{h,D}, v_h) + D_h(v_h + u_{h,D}; u_{h,D}, v_h) - (g, v_h)_\Omega\,.$$

The terms involving $u_{h,D}$ are bounded next. The Cauchy–Schwarz and Poincaré inequalities lead to

$$\begin{aligned}|a(u_{h,D}, v_h)| &= \left|\varepsilon(\nabla u_{h,D}, \nabla v_h)_\Omega + (\boldsymbol{b}\cdot\nabla u_{h,D}, v_h)_\Omega + (c\,u_{h,D}, v_h)_\Omega\right|\\ &\le \varepsilon\,|u_{h,D}|_{1,\Omega}\,|v_h|_{1,\Omega} + \|\boldsymbol{b}\|_{\infty,\Omega}\,|u_{h,D}|_{1,\Omega}\,\|v_h\|_{0,\Omega} + \|c\|_{\infty,\Omega}\,\|u_{h,D}\|_{0,\Omega}\,\|v_h\|_{0,\Omega}\\ &\le C\|u_{h,D}\|_{1,\Omega}\,|v_h|_{1,\Omega}\,.\end{aligned}$$

In addition, using the shape regularity of the mesh sequence, $\beta_E(\cdot) \le 1$, and the local trace inequality, one arrives at

$$\begin{aligned}&|D_h(v_h + u_{h,D}; u_{h,D}, v_h)|\\ &\quad= \left|\sum_{E\in\mathscr{E}_h} \beta_E(v_h + u_{h,D})\,|d_E|\,h_E\,(\nabla u_{h,D}\cdot\boldsymbol{t}_E, \nabla v_h\cdot\boldsymbol{t}_E)_E\right|\\ &\quad\le \sum_{E\in\mathscr{E}_h} |d_E|\,h_E\,\|\nabla u_{h,D}\cdot\boldsymbol{t}_E\|_{0,E}\,\|\nabla v_h\cdot\boldsymbol{t}_E\|_{0,E} \le C\,|u_{h,D}|_{1,\Omega}\,|v_h|_{1,\Omega}\,.\end{aligned}$$

Finally, the application of the Poincaré and Young inequalities gives

$$\langle T v_h, v_h\rangle \geq \varepsilon |v_h|_{1,\Omega}^2 - C\,\|u_{h,D}\|_{1,\Omega} |v_h|_{1,\Omega} - \|g\|_{0,\Omega}\|v_h\|_{0,\Omega} \geq \frac{\varepsilon}{2}|v_h|_{1,\Omega}^2 - C_0\,.$$

Thus, for $v_h \in V_{h,0}$ such that $|v_h|_{1,\Omega} > (2\,C_0/\varepsilon)^{\frac12}$ there holds $\langle T v_h, v_h\rangle > 0$. Lemma 3.1 implies that there exists $v_h \in V_{h,0}$ such that $|v_h|_{1,\Omega} < 2\,(C_0/\varepsilon)^{\frac12}$ and $T v_h = 0$. In other words, $u_h := v_h + u_{h,D}$ solves (14).

3.2 The DMP

In this section we shall formulate general properties of the limiters β_E under which the AFC scheme (14) satisfies the local and global DMP. The local DMP will be formulated on the patches Δ_i defined in Section 2.

To prove the DMP, we make the following general assumption, which is a reformulation of an analogous assumption introduced in [26].

Assumption (A2): Consider any $u_h \in V_h$ and any $i \in \{1,\dots,M\}$. If $u_h(\boldsymbol{x}_i)$ is a strict local extremum of u_h on Δ_i, i.e.,

$$u_h(\boldsymbol{x}_i) > u_h(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Delta_i \setminus \{\boldsymbol{x}_i\} \qquad \text{or} \qquad u_h(\boldsymbol{x}_i) < u_h(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \Delta_i \setminus \{\boldsymbol{x}_i\}\,,$$

then

$$a_h(u_h;\varphi_j,\varphi_i) \leq 0 \qquad \forall\, j \in S_i\,.$$

Theorem 2 (Local DMP) *Let $u_h \in V_h$ be a solution of* (14) *with limiters β_E satisfying Assumption (A2). Consider any $i \in \{1,\dots,M\}$. Then*

$$g \leq 0 \ \text{ in } \Delta_i \quad \Rightarrow \quad \max_{\Delta_i} u_h \leq \max_{\partial\Delta_i} u_h^+\,, \tag{17}$$

$$g \geq 0 \ \text{ in } \Delta_i \quad \Rightarrow \quad \min_{\Delta_i} u_h \geq \min_{\partial\Delta_i} u_h^-\,, \tag{18}$$

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

$$g \leq 0 \ \text{ in } \Delta_i \quad \Rightarrow \quad \max_{\Delta_i} u_h = \max_{\partial\Delta_i} u_h\,, \tag{19}$$

$$g \geq 0 \ \text{ in } \Delta_i \quad \Rightarrow \quad \min_{\Delta_i} u_h = \min_{\partial\Delta_i} u_h\,. \tag{20}$$

Proof Let $u_h \in V_h$ satisfy (14) and let us denote $u_i = u_h(\boldsymbol{x}_i)$, $i = 1,\dots,N$. Then $u_h = \sum_{j=1}^N u_j\,\varphi_j$ and one has

$$\sum_{j=1}^N \widetilde{a}_{ij}\,u_j = g_i\,, \qquad i = 1,\dots,M\,, \tag{21}$$

where

$$\begin{aligned} \widetilde{a}_{ij} &= a_h(u_h;\varphi_j,\varphi_i)\,, \qquad i = 1,\dots,M\,,\ j = 1,\dots,N\,,\\ g_i &= (g,\varphi_i)_{\Delta_i}\,, \qquad i = 1,\dots,M\,. \end{aligned}$$

Moreover, for $i = 1, \dots, M$, one derives

$$\widetilde{a}_{ii} \ge a(\varphi_i, \varphi_i) \ge \varepsilon\, |\varphi_i|^2_{1,\Omega} > 0\,, \tag{22}$$

$$\sum_{j=1}^{N} \widetilde{a}_{ij} = a_h(u_h; 1, \varphi_i) = (c, \varphi_i)_{\Delta_i} \ge 0\,. \tag{23}$$

These properties follow from the fact that $(\boldsymbol{b}\cdot\nabla v, v)_\Omega = 0$ for any $v \in H^1_0(\Omega)$, $D_h(u_h; \varphi_i, \varphi_i) \ge 0$, and $\sum_{j=1}^N \varphi_j = 1$ in Ω.

Consider any $i \in \{1, \dots, M\}$ and let $g \le 0$ in Δ_i so that $g_i \le 0$. Let us denote $A_i = \sum_{j=1}^N \widetilde{a}_{ij}$. Since $\widetilde{a}_{ij} = 0$ for any $j \notin S_i \cup \{i\}$, it follows from (21) that

$$A_i\, u_i + \sum_{j\in S_i} \widetilde{a}_{ij}\,(u_j - u_i) = g_i\,. \tag{24}$$

To prove (19), let $c = 0$ in Δ_i and assume that $\max_{\Delta_i} u_h > \max_{\partial\Delta_i} u_h$. Since the maximum of u_h on Δ_i is attained at vertices of the elements of $\mathscr{T}_h$ making up Δ_i, this means that $u_h(\boldsymbol{x}_i)$ is the strict maximum of u_h on Δ_i. Then Assumption (A2) implies that the sum in (24) is non-negative. Since $A_i = 0$ (see (23)) and $\widetilde{a}_{ii} > 0$ (see (22)), there is $j \in S_i$ such that $\widetilde{a}_{ij} < 0$ and hence the left-hand side of (24) is positive, which is a contradiction.

For proving (17), it suffices to consider the case $A_i > 0$. Let us assume that $\max_{\Delta_i} u_h > \max_{\partial\Delta_i} u_h^+$. Then again $\max_{\Delta_i} u_h > \max_{\partial\Delta_i} u_h$ and also $u_i > 0$. Like before, the sum in (24) is non-negative and since $A_i\, u_i > 0$, the left-hand side of (24) is positive, which is again a contradiction proving the assertion.

The implications (18) and (20) follow in an analogous way.

Theorem 3 (Global DMP) *Let $u_h \in V_h$ be a solution of* (14) *with limiters β_E satisfying Assumptions (A2) and (A1). Then*

$$g \le 0 \quad \text{in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u_h \le \max_{\partial\Omega} u_h^+\,, \tag{25}$$

$$g \ge 0 \quad \text{in } \Omega \quad \Rightarrow \quad \min_{\overline{\Omega}} u_h \ge \min_{\partial\Omega} u_h^-\,. \tag{26}$$

If, in addition, $c = 0$ in Ω, then

$$g \le 0 \quad \text{in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u_h = \max_{\partial\Omega} u_h\,, \tag{27}$$

$$g \ge 0 \quad \text{in } \Omega \quad \Rightarrow \quad \min_{\overline{\Omega}} u_h = \min_{\partial\Omega} u_h\,. \tag{28}$$

Proof The proof is based on the technique used in [25, Theorems 5.1 and 5.2]. Let $u_h \in V_h$ satisfy (14) and let $g \le 0$ in Ω. Then the nodal values of u_h satisfy (21) and, due to (7), one has

$$\sum_{i,j=1}^{M} v_i\, \widetilde{a}_{ij}\, v_j \ge \sum_{i,j=1}^{M} v_i\, a_{ij}\, v_j > 0 \qquad \forall\ (v_1, \dots, v_M) \in \mathbb{R}^M \setminus \{0\}\,. \tag{29}$$

Note that

$$\max_{\overline{\Omega}} u_h = \max\{u_i\ :\ i = 1, \dots, N\}\,,$$
$$\max_{\partial\Omega} u_h = \max\{u_i\ :\ i = M+1, \dots, N\}\,.$$

Let

$$s = \max\{u_i\ :\ i = 1,\dots,N\}\,, \qquad J = \{i \in \{1,\dots,N\}\ :\ u_i = s\}\,.$$

First, let us show that

$$\widetilde{a}_{ij} \le 0 \qquad \forall\ i \in J \cap \{1,\dots,M\},\ j \notin J\,. \tag{30}$$

Let $i \in J \cap \{1,\dots,M\}$ and $j \in S_i \setminus J$. Then $\widetilde{a}_{ij} = a_{ij} - \beta_E(u_h)\,|d_E|$, where E is the edge with endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. For any $k \in \mathbb{N}$, define the function $u_h^k = u_h + \varphi_i/k$. Then $u_h^k(\boldsymbol{x}_i)$ is the strict maximum of u_h^k on $\overline{\Omega}$ and hence, in view of Assumption (A2),

$$(a_{ij} - \beta_E(u_h^k)\,|d_E|)\,(u_i^k - u_j^k) = a_h(u_h^k;\varphi_j,\varphi_i)\,(u_i^k - u_j^k) \le 0\,,$$

where $u_i^k = u_h^k(\boldsymbol{x}_i)$ and $u_j^k = u_h^k(\boldsymbol{x}_j)$. Since $u_h^k \to u_h$ for $k \to \infty$, Assumption (A1) implies that

$$(a_{ij} - \beta_E(u_h)\,|d_E|)\,(u_i - u_j) \le 0\,.$$

As $u_i - u_j > 0$, it follows that $\widetilde{a}_{ij} \le 0$. For $j \notin S_i \cup \{i\}$, one has $\widetilde{a}_{ij} = 0$, which completes the proof of (30).

Now we want to prove that the relations (21), (23), (29), and (30) imply (25) and (27). If $c = 0$ in Ω and hence $\sum_{j=1}^N \widetilde{a}_{ij} = 0$ for $i = 1,\dots,M$ (see (23)), then (21) still holds if one adds a constant to all components of the vector $(u_1,\dots,u_N)$ so that one can assume that $s > 0$. If $\sum_{j=1}^N \widetilde{a}_{ij} > 0$, then $s > 0$ can be also assumed since otherwise (25) trivially holds.

Thus, let $s > 0$ and let us assume that (27) does not hold, which implies that $J \subset \{1,\dots,M\}$. We shall prove that then

$$\exists\ k \in J: \quad \mu_k := \sum_{j\in J} \widetilde{a}_{kj} > 0\,. \tag{31}$$

Let (31) do not hold. Then, applying (23) and (30), one derives for any $i \in J$

$$0 \ge \sum_{j\in J} \widetilde{a}_{ij} \ge -\sum_{j\notin J} \widetilde{a}_{ij} \ge 0\,,$$

which gives

$$\sum_{j\in J} \widetilde{a}_{ij} = 0 \quad \forall\ i \in J\,, \qquad \widetilde{a}_{ij} = 0 \quad \forall\ i \in J,\ j \notin J\,.$$

Thus, the matrix $(\widetilde{a}_{ij})_{i,j\in J}$ is singular and hence there exist real numbers $\{v_i\}_{i\in J}$, not all equal to zero, such that $\sum_{i\in J} \widetilde{a}_{ij}\, v_i = 0$ for $j = 1,\dots,M$. Consequently, the matrix $(\widetilde{a}_{ij})_{i,j=1}^M$ is singular, which contradicts (29). Therefore, (31) holds and hence, denoting $r = \max\{u_i\ :\ i = 1,\dots,N,\ i \notin J\}$, one obtains using (21), (30), and (23)

$$s\,\mu_k = \sum_{j\in J} \widetilde{a}_{kj}\, u_j = g_k - \sum_{j\notin J} \widetilde{a}_{kj}\, u_j \le g_k + r \sum_{j\notin J} (-\widetilde{a}_{kj}) \le r\,\mu_k$$

(note that the first inequality implies that $r > 0$). Hence, $s \le r$, which is a contradiction to the definition of J. Therefore (27) and hence also (25) holds.

The relations (26) and (28) can be proved analogously.

3.3 An a priori error estimate

The error estimate will be proven using the following mesh-dependent norm.

$$\|v\|_h := \left(\varepsilon\,|v|_{1,\Omega}^2 + c_0\,\|v\|_{0,\Omega}^2 + D_h(u_h;v,v)\right)^{\frac{1}{2}} \ ,$$

where D_h is defined in (15) and $c_0 := \inf \operatorname{ess}_\Omega\ c$.

Theorem 4 (Error estimate) *Let us suppose that the solution of* (2) *belongs to $H^2(\Omega)$ and that $c_0 > 0$. Then, there exists $C > 0$, independent of h and the data of* (1)*, such that*

$$\|u - u_h\|_h \le C\big(\varepsilon + c_0^{-1}\{\|\boldsymbol{b}\|_{\infty,\Omega}^2 + \|c\|_{\infty,\Omega}^2\, h^2\}\big)^{\frac{1}{2}} h\,|u|_{2,\Omega} + D_h(u_h;i_hu,i_hu)^{\frac{1}{2}}\,.$$

Proof We decompose the error in the usual way $u - u_h = (u - i_h u) + (i_h u - u_h) =: \rho_h + e_h$. First, we notice that $D_h(u_h;\rho_h,\rho_h) = 0$, and then, standard interpolation estimates lead to

$$\|\rho_h\|_h \le C(\varepsilon + c_0\, h^2)^{\frac{1}{2}}\, h\,|u|_{2,\Omega}\,.$$

To bound the discrete error e_h we use the ellipticity of $a(\cdot,\cdot)$, the properties of $D_h(\cdot;\cdot,\cdot)$, and the relations (14) and (2) to get

$$\begin{aligned}\|e_h\|_h^2 &\le a(e_h,e_h) + D_h(u_h;e_h,e_h)\\ &= a(i_hu,e_h) - \big\{a(u_h,e_h) + D_h(u_h;u_h,e_h)\big\} + D_h(u_h;i_hu,e_h)\\ &= -\,a(\rho_h,e_h) + D_h(u_h;i_hu,e_h)\,.\end{aligned}$$

Next, the continuity of a gives

$$\begin{aligned}a(\rho_h,e_h) &\le \left(\left[\varepsilon^{\frac{1}{2}} + c_0^{-\frac{1}{2}}\,\|\boldsymbol{b}\|_{\infty,\Omega}\right]|\rho_h|_{1,\Omega} + c_0^{-\frac{1}{2}}\,\|c\|_{\infty,\Omega}\,\|\rho_h\|_{0,\Omega}\right)\|e_h\|_h\\ &\le C\left(\varepsilon^{\frac{1}{2}} + c_0^{-\frac{1}{2}}\,\|\boldsymbol{b}\|_{\infty,\Omega} + c_0^{-\frac{1}{2}}\,\|c\|_{\infty,\Omega}\, h\right) h\,|u|_{2,\Omega}\,\|e_h\|_h\,.\end{aligned}$$

Moreover, since $D_h(u_h;\cdot,\cdot)$ is a symmetric positive semi-definite bilinear form, it satisfies Cauchy–Schwarz inequality, which gives

$$D_h(u_h;i_hu,e_h) \le D_h(u_h;i_hu,i_hu)^{\frac{1}{2}} D_h(u_h;e_h,e_h)^{\frac{1}{2}} \le D_h(u_h;i_hu,i_hu)^{\frac{1}{2}}\|e_h\|_h\,.$$

Combining the above relations proves the result.

A simple estimate of the consistency error $D_h(u_h;i_hu,i_hu)^{\frac{1}{2}}$ is given in the following lemma.

Lemma 2 (Basic estimate of the consistency error) *Denoting*

$$A_h = \max_{E\in\mathscr{E}_h}\left(|d_E|\,h_E^{2-d}\right),$$

one has

$$D_h(u_h;i_hu,i_hu) \le C\,A_h\,|i_hu|_{1,\Omega}^2 \qquad \forall\ u_h \in V_h,\ u \in C^0(\overline{\Omega})\,.$$

If, in particular, d_E are defined by (8)*, then*

$$D_h(u_h;i_hu,i_hu) \le C\,(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega}\, h + \|c\|_{\infty,\Omega}\, h^2)\,|i_hu|_{1,\Omega}^2\,.$$

Proof Using $\beta_E \leq 1$ and the shape regularity of $\mathscr{T}_h$ implies that

$$D_h(u_h; i_h u, i_h u) \leq A_h \sum_{E \in \mathscr{E}_h} h_E^{d-1} \, \|\nabla i_h u \cdot \boldsymbol{t}_E\|_{0,E}^2 \leq C \, A_h \, |i_h u|_{1,\Omega}^2 \, .$$

If d_E is defined by (8) for an internal edge E with endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, then

$$\begin{aligned} |d_E| &\leq \sum_{T \in \mathscr{T}_h,\, \boldsymbol{x}_i, \boldsymbol{x}_j \in T} \Big(\varepsilon \, |\varphi_i|_{1,T} \, |\varphi_j|_{1,T} + \|c\|_{\infty,T} \, \|\varphi_i\|_{0,T} \, \|\varphi_j\|_{0,T} \\ &\qquad\qquad + \|\boldsymbol{b}\|_{\infty,T} \left\{ |\varphi_i|_{1,T} \, \|\varphi_j\|_{0,T} + |\varphi_j|_{1,T} \, \|\varphi_i\|_{0,T} \right\} \Big) \\ &\leq C \, h_E^{d-2} \, \left(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} \, h + \|c\|_{\infty,\Omega} \, h^2 \right) \, , \end{aligned}$$

which finishes the proof.

Lemma 3.3 shows that if d_E is defined by (8), then the convergence order of $\|u - u_h\|_h$ is reduced to 1/2 in the convection-dominated case and no convergence follows in the diffusion-dominated case. It was demonstrated in [8] that these results are sharp. On the other hand, the results of [6, 9] indicate that a better convergence behaviour in the diffusion-dominated case may be expected if the AFC scheme is linearity preserving, i.e., if the stabilization originating from the algebraic flux correction vanishes in regions where the approximate solution is a polynomial of degree 1. This property can be formulated in terms of the limiters β_E in the following way.

Assumption (A3): The limiters β_E possess the linearity-preservation property, i.e.,

$$\beta_E(u_h) = 0 \qquad \text{if } \; u_h|_{\omega_E} \in \mathbb{P}_1(\omega_E) \qquad \forall \; E \in \mathscr{E}_h \, .$$

This property leads to an improved bound of the consistency error provided that the limiters satisfy the following Lipschitz-continuity assumption.

Assumption (A4): For any $E \in \mathscr{E}_h$ with endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, the function $\beta_E(u_h)(\nabla u_h)|_E \cdot \boldsymbol{t}_E$ is Lipschitz continuous in the sense that

$$\Big| \beta_E(u_h)(\nabla u_h)|_E \cdot \boldsymbol{t}_E - \beta_E(v_h)(\nabla v_h)|_E \cdot \boldsymbol{t}_E \Big| \leq C \sum_{E' \in \mathscr{E}_i \cup \mathscr{E}_j} \Big| (\nabla(u_h - v_h))|_{E'} \cdot \boldsymbol{t}_{E'} \Big| .$$

Lemma 3 (Improved estimate of the consistency error) *Let the limiters β_E satisfy Assumptions (A3) and (A4). Then*

$$D_h(u_h; i_h u, i_h u) \leq \frac{\varepsilon}{2} \, |u_h - i_h u|_{1,\Omega}^2 + C \, \frac{A_h^2}{\varepsilon} \, |i_h u|_{1,\Omega}^2 + \varepsilon \, h^2 \, |u|_{2,\Omega}^2 \, .$$

Proof The proof is a refinment of the technique used in [6, Theorem 4]. Let us write $D_h = \sum_{E \in \mathscr{E}_h} D_E$ with

$$D_E(z; v, w) = \beta_E(z) \, |d_E| \, h_E \, (\nabla v \cdot \boldsymbol{t}_E, \nabla w \cdot \boldsymbol{t}_E)_E \, .$$

Then it follows from Assumption (A4) and the shape regularity of $\mathscr{T}_h$ that, for any $u_h, v_h, w_h \in V_h$,

$$\begin{aligned} &|D_E(u_h; u_h, w_h) - D_E(v_h; v_h, w_h)| \\ &\quad \leq C \, |d_E| \, h_E^2 \sum_{E' \in \mathscr{E}_i \cup \mathscr{E}_j} \Big| (\nabla(u_h - v_h))|_{E'} \cdot \boldsymbol{t}_{E'} \Big| \, \Big| (\nabla w_h)|_E \cdot \boldsymbol{t}_E \Big| \\ &\quad \leq \widetilde{C} \, |d_E| \, h_E^{2-d} \, |u_h - v_h|_{1,\omega_E} \, |w_h|_{1,\omega_E} \, . \end{aligned} \tag{32}$$

Consequently,

$$|D_h(u_h; u_h, w_h) - D_h(v_h; v_h, w_h)| \leq C\, A_h\, |u_h - v_h|_{1,\Omega}\, |w_h|_{1,\Omega}\,.$$

Like in Lemma 3.3, one also obtains

$$|D_h(u_h; v_h, w_h)| \leq C\, A_h\, |v_h|_{1,\Omega}\, |w_h|_{1,\Omega}\,.$$

Using the last two estimates and applying Young's inequality, one obtains

$$\begin{aligned} D_h(u_h; i_h u, i_h u) &= D_h(u_h; i_h u - u_h, i_h u) \\ &\quad + \{D_h(u_h; u_h, i_h u) - D_h(i_h u; i_h u, i_h u)\} + D_h(i_h u; i_h u, i_h u) \\ &\leq \frac{\varepsilon}{2}\, |u_h - i_h u|_{1,\Omega}^2 + C\, \frac{A_h^2}{\varepsilon}\, |i_h u|_{1,\Omega}^2 + D_h(i_h u; i_h u, i_h u)\,. \end{aligned}$$

To bound the last term, we use the linearity preservation and the Lipschitz continuity of D_E. More precisely, for a given $E \in \mathscr{E}_h$, we introduce the function $i_E u \in \mathbb{P}_1(\omega_E)$ as the unique solution of the problem

$$(\nabla i_E u, \nabla \psi)_{\omega_E} = (\nabla u, \nabla \psi)_{\omega_E} \quad \forall\, \psi \in \mathbb{P}_1(\omega_E)\,, \qquad (i_E u, 1)_{\omega_E} = (u, 1)_{\omega_E}\,.$$

Using standard finite element approximation results (see [16]), $i_E u$ satisfies

$$|u - i_E u|_{1,\omega_E} \leq C\, h_E |u|_{2,\omega_E}\,. \tag{33}$$

Outside ω_E, the function $i_E u$ can be arbitrarily extended to a function from V_h. In view of Assumption (A3), one has $D_E(i_E u; i_E u, i_h u) = 0$ and hence, using (32), (33), and the shape regularity of $\mathscr{T}_h$, one obtains

$$\begin{aligned} D_E(i_h u; i_h u, i_h u) &= D_E(i_h u; i_h u, i_h u) - D_E(i_E u; i_E u, i_h u) \\ &\leq C\, |d_E|\, h_E^{2-d}\, |i_h u - i_E u|_{1,\omega_E}\, |i_h u|_{1,\omega_E} \leq \widetilde{C}\, |d_E|\, h_E^{3-d}\, |u|_{2,\omega_E}\, |i_h u|_{1,\omega_E}\,. \end{aligned}$$

This implies that

$$D_h(i_h u; i_h u, i_h u) \leq C\, A_h\, h\, |u|_{2,\Omega}\, |i_h u|_{1,\Omega} \leq \frac{C^2 A_h^2}{4\,\varepsilon}\, |i_h u|_{1,\Omega}^2 + \varepsilon\, h^2\, |u|_{2,\Omega}^2\,,$$

which completes the proof.

4 Various definitions of the limiters

4.1 The Zalesak limiter

The Zalesak limiter is a classical limiter that originates from ideas of [38] and it was proposed to be applied in AFC finite element schemes in [29]. Numerical simulations with this limiter can be found also in [5, 8]. The numbers d_E in the definition of D_h are given by (8) in this case.

After having presented the Zalesak limiter, we will show that it satisfies Assumption (A1) and, under an additional assumption on the matrix $\mathbb{A}$, also Assumption (A2). Consequently, the nonlinear problem (14) possesses a solution and satisfies the discrete maximum principle. It was demonstrated in [9, Ex. 7.2] with the help of a numerical example that the AFC scheme with the Zalesak limiter is not linearity preserving in general.

The definition of the coefficients for the Zalesak limiter relies on the values P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i = 1, \ldots, M$ by

$$P_i^+ := \sum_{\substack{j \in S_i \\ a_{ji} \le a_{ij}}} f_{ij}^+ \,, \quad P_i^- := \sum_{\substack{j \in S_i \\ a_{ji} \le a_{ij}}} f_{ij}^- \,, \quad Q_i^+ := -\sum_{j \in S_i} f_{ij}^- \,, \quad Q_i^- := -\sum_{j \in S_i} f_{ij}^+ \,, \tag{34}$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$, $f_{ij}^+ = \max\{0, f_{ij}\}$, and $f_{ij}^- = \min\{0, f_{ij}\}$. These values can be computed by performing a loop over all internal edges. After this loop, one defines

$$R_i^+ := \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- := \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \ldots, M \,. \tag{35}$$

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

$$R_i^+ := 1 \,, \quad R_i^- := 1 \,, \quad i = M+1, \ldots, N \,. \tag{36}$$

Then, for any $i, j \in \{1, \ldots, N\}$ such that $a_{ji} \le a_{ij}$, we set

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \,, \\ 1 & \text{if } f_{ij} = 0 \,, \\ R_i^- & \text{if } f_{ij} < 0 \,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij} \,. \tag{37}$$

The final step consists in defining $\beta_E := 1 - \alpha_{ij}$ for any internal edge $E \in \mathscr{E}_h$ having the endpoints $\boldsymbol{x}_i$, $\boldsymbol{x}_j$.

There is an obvious ambiguity in the definition of β_E if $a_{ij} = a_{ji}$. This ambiguity does not influence the resulting method if $\min\{a_{ij}, a_{ji}\} \le 0$ since then $d_E = 0$ and the respective term with β_E does not occur in (15). To fulfill the condition $\min\{a_{ij}, a_{ji}\} \le 0$, which also assures the DMP (cf. Lemma 4.1), it may help to replace the matrix corresponding to the reaction term by a lumped diagonal matrix, see [8].

Lemma 4 *The Zalesak limiter satisfies Assumption (A1).*

Proof Let E be an internal edge that connects the nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Then it suffices to show that $\alpha_{ij}(u_h)(u_j - u_i)$ is a continuous function of $u_h \in V_h$. Because of $\beta_E(u) = \beta_{ij} = \beta_{ji}$, $\alpha_{ij} = \alpha_{ij}$, we can restrict these considerations to the situation that $a_{ji} \le a_{ij}$. Moreover, it suffices to consider $d_{ij} < 0$ since otherwise $\alpha_{ij} \equiv 1$.

As first case, $\bar{u}_h \in V_h$ such that $f_{ij}(\bar{u}_h) > 0$ will be considered. Then $\bar{u}_i > \bar{u}_j$ and hence $f_{ij}(u_h) > 0$ in a neighborhood of $\bar{u}_h$. Using (37), (35), and (34), we obtain

$$\alpha_{ij}(u_h) = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{f_{ij} + \widetilde{P}_i^+} \quad \text{with} \quad \widetilde{P}_i^+ = \sum_{\substack{k \in S_i \\ a_{ki} \le a_{ik},\, k \ne j}} f_{ik}^+ \,. \tag{38}$$

The numerator and the denominator are continuous functions and by assumption, the denominator is positive in a neighborhood of $\bar{u}_h$. Hence α_{ij} is a continuous function at $\bar{u}_h$. In the same way, we get for the case $f_{ij}(\bar{u}_h) < 0$ first that $\bar{u}_i < \bar{u}_j$ and second the representation formula

$$\alpha_{ij}(u_h) = R_i^- = \frac{\min\{-P_i^-, -Q_i^-\}}{|f_{ij}| - \widetilde{P}_i^-} \quad \text{with} \quad \widetilde{P}_i^- = \sum_{\substack{k \in S_i \\ a_{ki} \le a_{ik},\, k \ne j}} f_{ik}^- \,. \tag{39}$$

Using exactly the same reasoning as above, we conclude that α_{ij} is continuous at $\bar{u}_h$ in this case.

The last case is $f_{ij}(\bar{u}_h) = 0$ which leads to $\alpha_{ij}(\bar{u}_h)(\bar{u}_j - \bar{u}_i) = 0$. Since α_{ij} is bounded by definition, $\alpha_{ij}(u_h)(u_j - u_i) \to 0$ as $u_j \to u_i$. Consequently, $\alpha_{ij}(u_h)(u_j - u_i)$ is continuous at $\bar{u}_h$.

Remark 2 In [8] it was shown that the terms $\alpha_{ij}(u_h)(u_j - u_i)$ are even Lipschitz-continuous. The proof of this property is based on the representations (38) and (39) of the coefficients α_{ij}. The sums in these representations are Lipschitz-continuous and then one can show that the function which is obtained by multiplying these representations with $(u_j - u_i)$ is Lipschitz-continuous, too.

Lemma 5 *Let the matrix of the system* (5) *satisfy*

$$\min\{a_{ij}, a_{ji}\} \le 0 \qquad \forall\, i = 1, \dots, M\,,\ j = 1, \dots, N\,,\ i \neq j\,. \tag{40}$$

Then the Zalesak limiter satisfies Assumption (A2).

Proof Consider any $u_h \in V_h$, $i \in \{1, \dots, M\}$, and $j \in S_i$. Let $u_i := u_h(\boldsymbol{x}_i)$ be a strict local extremum of u_h in Δ_i. We want to prove that

$$a_{ij} + (1 - \alpha_{ij}(u_h))\, d_{ij} \le 0\,. \tag{41}$$

If $a_{ij} \le 0$, then (41) holds since $(1 - \alpha_{ij}(u_h))\, d_{ij} \le 0$. If $a_{ij} > 0$, then $a_{ji} \le 0$ due to (40) and hence $a_{ji} \le a_{ij}$ and $d_{ij} = -a_{ij} < 0$. Thus, if $u_i > u_k$ for any $k \in S_i$, then $f_{ij} > 0$ and $f_{ik} \ge 0$ for $k \in S_i$, so that $\alpha_{ij} = R_i^+ = 0$. Similarly, if $u_i < u_k$ for any $k \in S_i$, then $f_{ij} < 0$ and $f_{ik} \le 0$ for $k \in S_i$, so that $\alpha_{ij} = R_i^- = 0$. Since $a_{ij} + d_{ij} \le 0$, one concludes that (41) holds.

4.2 A limiter leading to linearity preservation and DMP on general meshes (BJK limiter)

Here we present a limiter recently proposed in [9] using some ideas of [30]. This limiter is designed in such a way that the AFC scheme satisfies the discrete maximum principle and linearity-preservation property on arbitrary meshes, which is a substantial improvement in comparison with the Zalesak limiter. Like in the previous section, the numbers d_E used in (15) are given by (8).

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

$$P_i^+ := \sum_{j \in S_i} f_{ij}^+\,, \qquad P_i^- := \sum_{j \in S_i} f_{ij}^-\,,$$
$$Q_i^+ := q_i\,(u_i - u_i^{\max})\,, \qquad Q_i^- := q_i\,(u_i - u_i^{\min})\,,$$

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

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

with fixed constants $\gamma_i > 0$. Then one defines the quantities R_i^+ and R_i^- again by (35) and (36) and one sets

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

Finally, the limiters are defined by $\beta_E := 1 - \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}$ for any internal edge $E \in \mathscr{E}_h$ having the endpoints $\boldsymbol{x}_i$, $\boldsymbol{x}_j$.

Lemma 6 *The above limiter satisfies Assumptions (A2) and (A1).*

Proof The validity of Assumption (A1) follows analogously as in the proof of Lemma 4.1. Let us prove Assumption (A2). Consider any $u_h \in V_h$, $i \in \{1, \ldots, M\}$, and $j \in S_i$ and assume that $u_i := u_h(\boldsymbol{x}_i)$ is a strict local extremum of u_h in Δ_i. Then we want to prove that

$$a_{ij} + (1 - \min\{\widetilde{\alpha}_{ij}(u_h), \widetilde{\alpha}_{ji}(u_h)\})\, d_{ij} \leq 0\,. \tag{42}$$

If $d_{ij} = 0$, then $a_{ij} \leq 0$ and hence (42) holds. Thus, let us assume that $d_{ij} < 0$. If $u_i > u_k$ for any $k \in S_i$, then $f_{ij} > 0$ and $u_i^{\max} = u_i$ so that $P_i^+ > 0$, $Q_i^+ = 0$ and $\widetilde{\alpha}_{ij} = R_i^+ = 0$. Since $a_{ij} + d_{ij} \leq 0$, one obtains (42). If $u_i < u_k$ for any $k \in S_i$, (42) follows analogously.

The constants γ_i can be adjusted in such a way that the linearity-preservation assumption (A3) is satisfied. In fact, it suffices to use such constants that

$$u_i - u_i^{\min} \leq \gamma_i\,(u_i^{\max} - u_i) \qquad \forall\; u \in \mathbb{P}_1(\mathbb{R}^d)\,. \tag{43}$$

It was proved in [9] that (43) holds with $\gamma_i = 1$ if the patch Δ_i is symmetric with respect to the vertex $\boldsymbol{x}_i$, and with

$$\gamma_i = \frac{\max\limits_{\boldsymbol{x}_j \in \partial\Delta_i} |\boldsymbol{x}_i - \boldsymbol{x}_j|}{\operatorname{dist}(\boldsymbol{x}_i, \partial\Delta_i^{\rm conv})}$$

in general, where $\Delta_i^{\rm conv}$ is the convex hull of Δ_i.

Lemma 7 *The above limiter satisfies Assumption (A3).*

Proof Consider any $i \in \{1, \ldots, M\}$. Since $R_i^+(u_h)$ and $R_i^-(u_h)$ depend on u_h only through $u_h|_{\Delta_i}$, it suffices to verify that, for any $u_h \in \mathbb{P}_1(\mathbb{R}^d)$, one has $R_i^+(u_h) = R_i^-(u_h) = 1$. One obtains using (43)

$$P_i^+ = \sum_{\substack{j\,\in\,S_i\\ u_j\,<\,u_i}} d_{ij}\,(u_j - u_i) \leq \sum_{j \in S_i} d_{ij}\,(u_i^{\min} - u_i) \leq \sum_{j \in S_i} d_{ij}\,\gamma_i\,(u_i - u_i^{\max}) = Q_i^+$$

and hence $R_i^+ = 1$. Similarly, one obtains $R_i^- = 1$.

Remark 3 Note that large values of the constants γ_i cause that more limiters α_{ij} will be 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 γ_i's also cause that the numerical solution of the nonlinear algebraic problem becomes more involved.

4.3 A limiter based on the variation of the discrete solution (BBK limiter)

In this section we review briefly the limiter presented in [6] and its main results. In this case, the numbers d_E in the definition of D_h are given by $d_E = \gamma_0\, h_E^{d-1}$, where γ_0 is a fixed parameter, dependent on the data of (1).

The limiters β_E, $E \in \mathscr{E}_h$, are given by the following algorithm: for $w_h \in V_h$, one defines ξ_{w_h} as the unique element in V_h whose nodal values are given by

$$\xi_{w_h}(\boldsymbol{x}_i) := \begin{cases} \dfrac{\left|\sum_{j\in S_i}\big(w_h(\boldsymbol{x}_i)-w_h(\boldsymbol{x}_j)\big)\right|}{\sum_{j\in S_i}|w_h(\boldsymbol{x}_i)-w_h(\boldsymbol{x}_j)|}, & \text{if } \displaystyle\sum_{j\in S_i}|w_h(\boldsymbol{x}_i)-w_h(\boldsymbol{x}_j)| \neq 0\,,\\ 0, & \text{otherwise}\,. \end{cases}$$

Then, on each $E \in \mathscr{E}_h$, β_E is defined by

$$\beta_E(w_h) := \max_{\boldsymbol{x}\in E}\big[\xi_{w_h}(\boldsymbol{x})\big]^p \quad , \quad p \in\ [1,+\infty)\,. \tag{44}$$

The value for p determines the rate of decay of the numerical diffusion with the distance to the critical points. A value closer to 1 adds more diffusion in the far field, while a larger value makes the diffusion vanish faster, but on the other hand, increasing p may make the nonlinear system more difficult to solve. In our experience, values up to $p = 20$ are considered safe to use (see [6] for a detailed discussion). In this section we will detail the proof of the results using $p = 1$, but these extend to $p > 1$ without difficulty (see [6] for details).

Two remarks can be rapidly made about this definition of the limiter. First, if a function v_h has en extremum at an internal node $\boldsymbol{x}_i$, and $E \in \mathscr{E}_i$, then $\beta_E(v_h) = 1$. This will be of paramount importance for the satisfaction of the DMP. Moreover, for meshes which have a certain structure, the method is linearity preserving, i.e., Assumption (A3) holds. More precisely, we will say that a mesh is symmetric with respect to its inner nodes if, for every node $\boldsymbol{x}_i$, and every $j \in S_i$, there exists $k \in S_i$ such that $\boldsymbol{x}_k - \boldsymbol{x}_i = -(\boldsymbol{x}_j - \boldsymbol{x}_i)$. So, if the mesh is symmetric with respect to its interior nodes, if $E \in \mathscr{E}_h$ has endpoints $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, and $v_h \in \mathbb{P}_1(\omega_E)$, then

$$\sum_{l\in S_i}\big(v_h(\boldsymbol{x}_i)-v_h(\boldsymbol{x}_l)\big) = 0 \quad\text{and}\quad \sum_{l\in S_j}\big(v_h(\boldsymbol{x}_j)-v_h(\boldsymbol{x}_l)\big) = 0\,,$$

which gives $\beta_E(v_h) = 0$. So, the method does not add extra diffusion in smooth regions, whenever the mesh is sufficiently structured.

Remark 4 In [6, Remark 1] a process to generate a method which is linearity preserving on general meshes is described. It involves a minimization process per node to determine a set of weights. The same results that hold for the method presented in this work hold for that variant.

The next result states that the limiter defined in (44) satisfies Assumptions (A1), (A2), and (A4).

Lemma 8 *The limiter defined in this section satisfies Assumptions (A1) and (A4). Moreover, if the triangulation $\mathscr{T}_h$ is such that*

$$(\nabla\varphi_j, \nabla\varphi_i)_\Omega \le 0\,, \qquad i = 1,\dots,M\,, \quad j = 1,\dots,N\,,$$

and $\gamma_0 \ge C_0\|\boldsymbol{b}\|_{\infty,\Omega} + C_1\,\|c\|_{\infty,\Omega}\,h$ (where C_0 and C_1 are two constants independent of h, but large enough), then Assumption (A2) is fulfilled, too.

Proof To prove (A1) and (A4), let $u_h, v_h \in V_h$. First, in [6, Lemma 1] the following result is proven: for any interior node $\boldsymbol{x}_i$ the following holds

$$|\xi_{u_h}(\boldsymbol{x}_i) - \xi_{v_h}(\boldsymbol{x}_i)| \le 4\frac{\sum_{E'\in\mathscr{E}_i} h_{E'}|\nabla(u_h - v_h)\cdot \boldsymbol{t}_{E'}|}{\sum_{E'\in\mathscr{E}_i} h_{E'}\big(|\nabla u_h\cdot \boldsymbol{t}_{E'}| + |\nabla v_h \cdot \boldsymbol{t}_{E'}|\big)}\,.$$

Let us now suppose that, for $E \in \mathscr{E}_h$ having end points $\boldsymbol{x}_i, \boldsymbol{x}_j$, $\beta_E(u_h) = \xi_{u_h}(\boldsymbol{x}_i)$ and $\beta_E(v_h) = \xi_{v_h}(\boldsymbol{x}_j)$. Then, using that $0 \le \xi_{v_h}(\boldsymbol{x}_i) \le 1$ one obtains

$$\begin{aligned}
\beta_E(u_h)\nabla u_h \cdot \boldsymbol{t}_E - \beta_E(v_h)\nabla v_h \cdot \boldsymbol{t}_E &\le \big(\xi_{u_h}(\boldsymbol{x}_i) - \xi_{u_h}(\boldsymbol{x}_j)\big)\nabla u_h \cdot \boldsymbol{t}_E \\
&\quad + |\xi_{u_h}(\boldsymbol{x}_j)|\,|\nabla(u_h - v_h)\cdot \boldsymbol{t}_E| + |\xi_{u_h}(\boldsymbol{x}_j) - \xi_{v_h}(\boldsymbol{x}_j)|\,|\nabla v_h \cdot \boldsymbol{t}_E| \\
&\le \big(\xi_{u_h}(\boldsymbol{x}_i) - \xi_{u_h}(\boldsymbol{x}_j)\big)\nabla u_h \cdot \boldsymbol{t}_E + 5\sum_{E' \in \mathscr{E}_j} |\nabla(u_h - v_h)\cdot \boldsymbol{t}_{E'}|\,.
\end{aligned}$$

In a completely analogous way one obtains

$$\begin{aligned}
&\beta_E(u_h)\nabla u_h \cdot \boldsymbol{t}_E - \beta_E(v_h)\nabla v_h \cdot \boldsymbol{t}_E \\
&\le \big(\xi_{v_h}(\boldsymbol{x}_i) - \xi_{v_h}(\boldsymbol{x}_j)\big)\nabla u_h \cdot \boldsymbol{t}_E + 5\sum_{E' \in \mathscr{E}_i} |\nabla(u_h - v_h)\cdot \boldsymbol{t}_{E'}|\,.
\end{aligned}$$

Thus, since $\xi_{u_h}(\boldsymbol{x}_i) - \xi_{u_h}(\boldsymbol{x}_j) \ge 0$ and $\xi_{v_h}(\boldsymbol{x}_i) - \xi_{v_h}(\boldsymbol{x}_j) \le 0$,

$$\begin{aligned}
&\beta_E(u_h)\nabla u_h \cdot \boldsymbol{t}_E - \beta_E(v_h)\nabla v_h \cdot \boldsymbol{t}_E \\
&\le \min\big\{\big(\xi_{u_h}(\boldsymbol{x}_i) - \xi_{u_h}(\boldsymbol{x}_j)\big)\nabla u_h \cdot \boldsymbol{t}_E, \big(\xi_{v_h}(\boldsymbol{x}_i) - \xi_{v_h}(\boldsymbol{x}_j)\big)\nabla u_h \cdot \boldsymbol{t}_E\big\} \\
&\quad + 5\sum_{E' \in \mathscr{E}_i \cup \mathscr{E}_j} |\nabla(u_h - v_h)\cdot \boldsymbol{t}_{E'}| \\
&\le 5\sum_{E' \in \mathscr{E}_i \cup \mathscr{E}_j} |\nabla(u_h - v_h)\cdot \boldsymbol{t}_{E'}|\,,
\end{aligned}$$

which proves Assumption (A4) and hence also (A1).

To prove (A2) let us suppose that u_h, solution of (14), has an extremum at the internal node $\boldsymbol{x}_i$. Let $j \in S_i$, and let $E \in \mathscr{E}_h$ be the edge with end points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Then, as it was mentioned earlier, $\beta_E(u_h) = 1$. Thus, using the shape regularity of the mesh, one obtains

$$\begin{aligned}
a_h(u_h;\varphi_j,\varphi_i) &= \varepsilon(\nabla\varphi_j,\nabla\varphi_i)_\Omega + (\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i)_\Omega \\
&\quad + (c\,\varphi_j,\varphi_i)_\Omega + \gamma_0\,\beta_E(u_h)\,h_E^d\,(\nabla\varphi_j\cdot\boldsymbol{t}_E,\nabla\varphi_i\cdot\boldsymbol{t}_E)_E \\
&\le C_0\,\|\boldsymbol{b}\|_{\infty,\Omega}\,h_E^{d-1} + C_1\,\|c\|_{\infty,\Omega}\,h_E^d - \gamma_0 h_E^{d-1} \\
&= (C_0\,\|\boldsymbol{b}\|_{\infty,\Omega} + C_1\,\|c\|_{\infty,\Omega}\,h - \gamma_0)\,h_E^{d-1}\,,
\end{aligned}$$

and Assumption (A2) follows.

It follows from Lemma 3.3 that the consistency error $D_h(u_h; i_h u, i_h u)$ can be bounded as follows:

$$D_h(u_h; i_h u, i_h u) \le C\,\gamma_0\,h\,|i_h u|_{1,\Omega}^2\,.$$

Moreover, if the mesh is symmetric with respect to its internal nodes, then Lemma 3.3 implies that the following bound holds for the consistency error

$$D_h(u_h; i_h u, i_h u) \le \frac{\varepsilon}{2}\,|u_h - i_h u|_{1,\Omega}^2 + C\,\gamma_0^2\,\frac{h^2}{\varepsilon}\,|i_h u|_{1,\Omega}^2 + \varepsilon\,h^2\,|u|_{2,\Omega}^2\,.$$

Thus, the method with the definition of the limiters from this section converges for every regular mesh, and, in addition, in the case in which the limiters are linearity preserving, the convergence order increases from $O(h^{\frac{1}{2}})$ to $O(h)$.

5 Iterative schemes for solving the nonlinear problem

Consider the weak formulation (13) and the equivalent formulation (11), (12) in matrix-vector notation. These formulations represent a nonlinear problem since the coefficients β_{ij} depend on the finite element solution u_h. Applying an iterative scheme for solving the nonlinear problem, our experience is that usually damping is necessary to achieve convergence. Let $u_h^{(m)}$, $m \geq 0$, be a given approximation of u_h.

A fixed point iteration can be defined as follows. In a first step, a finite element function $\tilde{u}_h^{(m+1)}$ is computed by solving: Find $\tilde{u}_h^{(m+1)} \in V_{h,0}$ such that

$$a\left(\tilde{u}_h^{(m+1)}, v_h\right) + D_h\left(u_h^{(m)}; \tilde{u}_h^{(m+1)}, v_h\right) = (g, v_h)_\Omega \qquad \forall\, v_h \in V_{h,0}\,. \tag{45}$$

The matrix-vector form of (45) is

$$\begin{aligned} &\sum_{j=1}^N a_{ij}\,\tilde{u}_j^{(m+1)} + \sum_{j=1}^N \beta_{ij}^{(m)}\, d_{ij}\,\left(\tilde{u}_j^{(m+1)} - \tilde{u}_i^{(m+1)}\right) = g_i\,, \quad i = 1,\ldots,M\,, \\ &\tilde{u}_i^{(m+1)} = 0\,, \quad i = M+1,\ldots,N\,, \end{aligned} \tag{46}$$

where $\beta_{ij}^{(m)} = \beta_{ij}\left(u^{(m)}\right)$. In the iterations (45) and (46), the matrix of the problem changes in each iteration.

It is also possible to perform a fixed point iteration in such a way that only the right-hand side changes. Using the relation

$$\sum_{j=1}^N \beta_{ij} d_{ij}(u_j - u_i) = \sum_{j=1}^N d_{ij}u_j - u_i \underbrace{\sum_{j=1}^N d_{ij}}_{=0} - \sum_{j=1}^N (1 - \beta_{ij}d_{ij})(u_j - u_i),$$

one can consider instead of (46) the iteration

$$\begin{aligned} \sum_{j=1}^N (a_{ij} + d_{ij})\,\tilde{u}_j^{(m+1)} &= g_i + \sum_{j=1}^N \left(1 - \beta_{ij}^{(m)}\right) d_{ij}\left(u_j^{(m)} - u_i^{(m)}\right) \\ &= g_i + \sum_{j=1}^N \left(1 - \beta_{ij}^{(m)}\right) f_{ij}^{(m)}\,, \quad i = 1,\ldots,M\,, \\ \tilde{u}_i^{(m+1)} &= 0\,, \quad i = M+1,\ldots,N\,. \end{aligned} \tag{47}$$

Using a sparse direct solver, then the matrix of (47) has to be factorized only once and in all subsequent iterations, only the solutions of the triangular systems have to be computed.

Another approach for solving the nonlinear problem is a (damped) Newton method. Let us consider as starting point for deriving this method the matrix-vector formulation (11), (12). Let the i-th equation be written in the form

$$F_i(u) = \sum_{j=1}^N a_{ij}u_j + \sum_{j=1}^N \beta_{ij}(u) d_{ij}(u_j - u_i) - g_i = 0, \quad i = 1,\ldots,M,$$

then the intermediate solution in Newton's method is computed by solving

$$DF\left(u^{(m)}\right)\tilde{u}_h^{(m+1)} = DF\left(u^{(m)}\right)u_h^{(m)} - F\left(u^{(m)}\right)\,, \tag{48}$$

where $DF\left(u^{(m)}\right)$ is the Jacobian, which can be computed by applying the product rule and the chain rule, and observing that the derivative of the limiter with respect to the Dirichlet nodes is not needed since these values are fixed

$$\begin{aligned} & DF_i(u)[v] \\ &= \sum_{j=1}^{N} a_{ij}v_j + \sum_{j=1}^{N} \beta_{ij}(u)d_{ij}(v_j - v_i) + \sum_{j=1}^{N}\left(\sum_{k=1}^{M}\frac{\partial \beta_{ij}}{\partial u_k}(u)v_k\right)d_{ij}(u_j-u_i) \\ &= \sum_{j=1}^{N} a_{ij}v_j + \sum_{j=1}^{N} \beta_{ij}(u)d_{ij}v_j - \left(\sum_{j=1}^{N}\beta_{ij}(u)d_{ij}\right)v_i \\ &\quad + \sum_{j=1}^{M}\left(\sum_{k=1}^{N}\frac{\partial \beta_{ik}}{\partial u_j}(u)d_{ik}(u_k-u_i)\right)v_j. \end{aligned}$$

Hence, the entries of the matrix that has to be inverted in (48) are given by

$$\begin{aligned} & DF\left(u^{(m)}\right)_{ij} \\ &= \begin{cases} a_{ij} + \beta_{ij}^{(m)}d_{ij} + \displaystyle\sum_{k=1}^{N}\frac{\partial \beta_{ik}^{(m)}}{\partial u_j}d_{ik}\left(u_k^{(m)} - u_i^{(m)}\right) & \text{if } i \neq j, \\ a_{ii} - \displaystyle\sum_{\substack{j=1\\ j\neq i}}^{N}\beta_{ij}^{(m)}d_{ij} + \sum_{k=1}^{N}\frac{\partial \beta_{ik}^{(m)}}{\partial u_i}d_{ik}\left(u_k^{(m)} - u_i^{(m)}\right) & \text{if } i = j, \end{cases} \end{aligned}$$

for $i = 1, \dots M$, $j = 1, \dots N$. The last $N - M$ rows have just the diagonal entry 1. The derivatives of the limiter with respect to the solution are needed. These derivatives depend on the particular limiter that is used in the simulations.

Let $\omega^{(m+1)} \in [\omega_0, 1]$, $\omega_0 > 0$, be a damping factor. The next iterate is given by

$$u_h^{(m+1)} = u_h^{(m)} + \omega^{(m+1)}\left(\tilde{u}_h^{(m+1)} - u_h^{(m)}\right).$$

The choice of appropriate damping parameters is essential for the efficiency of the iteration. In [20], an automatic strategy for adapting the parameter during the iteration is described. In [5], the use of the so-called Anderson acceleration, proposed in [3, 36], is advocated. The Anderson acceleration stores vectors from previous iterations and builds with them second order information.

6 Numerical studies

6.1 The Hemker example

We will consider the so-called Hemker example, which was proposed in [17]. It models the convection of temperature from a hot circle (2d cylinder) in a channel. The convection field is constant. There are exponential layers a the circle and interior layers downstream the circle. The Hemker problem can be considered as a standard benchmark problem for convection-diffusion equations. It was used in [5] for comparing a number of stabilized discretizations. Here, the same setup as in this paper will be considered.

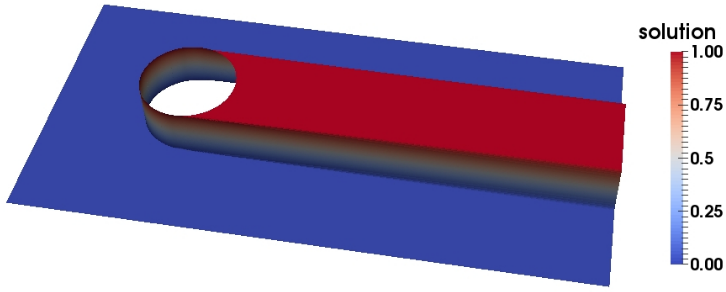

Figure 1: Hemker example: Solution for $\varepsilon = 10^{-4}$.

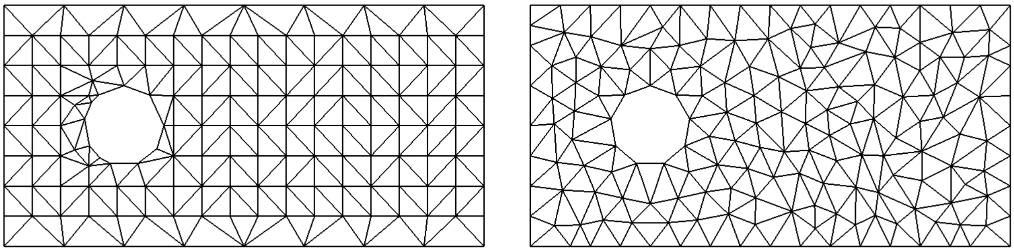

Figure 2: Hemker example: Grid 1 (left) and Grid 2 (right), both level 0.

This problem is defined in $\Omega = \{[-3,9] \times [-3,3]\} \setminus \{(x,y) \ : \ x^2 + y^2 < 1\}$, the coefficients are $\boldsymbol{b} = (1,0)^T$, $c = 0$, $f = 0$, and the boundary conditions are given by

$$u(x,y) = \begin{cases} 0, & \text{for } x = -3, \\ 1, & \text{for } x^2 + y^2 = 1, \\ \varepsilon \nabla u \cdot \mathbf{n} = 0, & \text{else.} \end{cases}$$

In [5], a reference solution on a fine grid was computed for $\varepsilon = 10^{-4}$, see Fig. 1. Quantities of interest defined in [5] are the magnitude of the over- and undershoots, the difference to the reference solution on selected cut lines, and the smearing of the interior layer at a certain cutline downstream the cylinder. For the concrete definition of these quantities, it is referred to [5].

The simulations were performed with P_1 finite elements on two types of grids, see Fig. 2. Grid 1 is aligned downstream to the convection and it has edges at the position where the interior layer is expected. The stopping criterion for the nonlinear iteration was based on the Euclidean norm of the residual vector, which should be smaller or equal than $10^{-13}\,(\#\mathrm{dof})^{\frac{1}{2}}$, where $\#\mathrm{dof}$ is the number of degrees of freedom (including Dirichlet nodes) on the respected grid. As initial iterate, a function that vanishes on all degrees of freedom was used. The linear systems were solved with the sparse direct solver UMFPACK, [15]. The simulations were performed with the code MooNMD [22].

By construction of the Hemker example, the solution takes values in $[0,1]$. The first quantity of interest from [5] considers the violation of this range by the numerical solutions. Since the AFC methods satisfy the DMP, it is expected that there are no violations if the nonlinear problems are solved exactly. In fact, we could observe in the numerical results only negligible violations of the order of the stopping criterion for the iteration of the nonlinear problem.

Another quantity of interest studies the smearing of the interior layer at $x = 4$, see Fig. 3. It can be seen that the smearing introduced by the BJK limiter is always smaller than with the Zalesak limiter. In particular, on the aligned Grid 1, the results with the BJK limiter are much better. This statement is supported by considering the error to the reference solution at the cutline $x = 4$, see Fig. 4. To compute the errors, 10001 equidistant points were taken on the cutline and the vector $\boldsymbol{e}$ contains

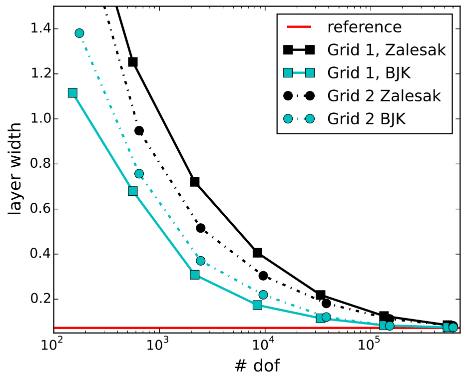

Figure 3: Hemker example: Width of the interior layer at $x = 4$.

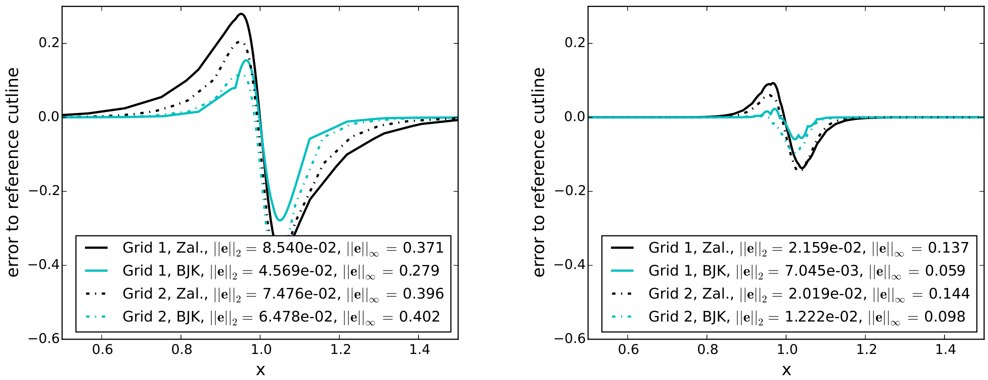

Figure 4: Hemker example: Errors to the cutline at $x = 4$, left level 3 (nearly 10000 d.o.f.s), right level 5 (nearly 150000 d.o.f.s).

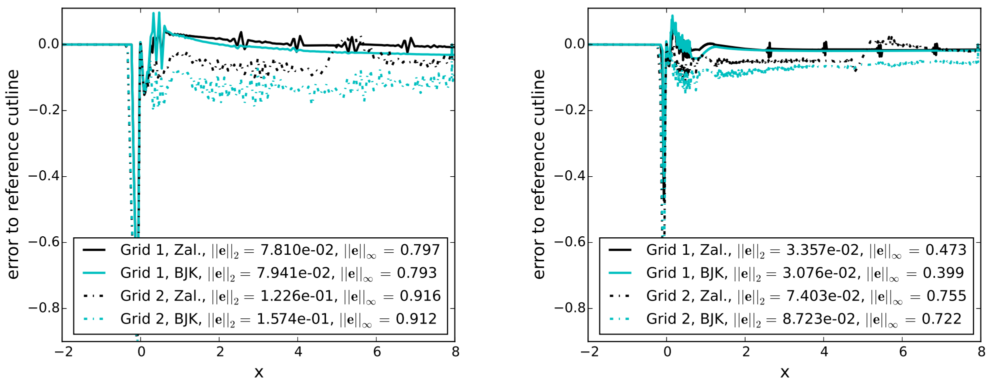

Figure 5: Hemker example: Errors to the cutline at $y = 1$, left level 3 (nearly 10000 d.o.f.s), right level 5 (nearly 150000 d.o.f.s).

the differences of the reference solution and the numerical solution in these points. The errors in the Euclidean norm $\|\boldsymbol{e}\|_2$ and the maximum norm $\|\boldsymbol{e}\|_\infty$ are given.

At the cutline $y = 1$, the results obtained with both limiters are similar, compare Fig. 5. The negative peak of the error is at the circle in a neighborhood of the point $(0, 1)$. In this neighborhood, there is the transition from the exponential layer to the interior layer.

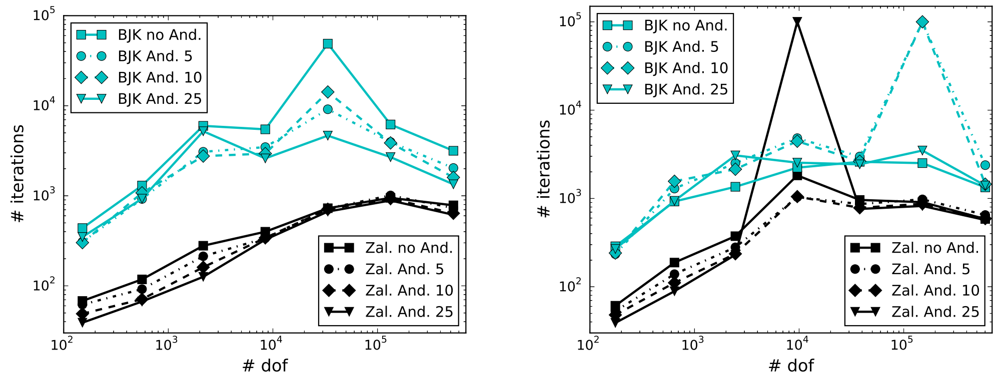

Figure 6: Hemker example: Number of iterations for solving the nonlinear problems, Grid 1 (left), Grid 2 (right); 100 000 iterations means that the stopping criterion was not reached.

Finally, the costs for solving the nonlinear problems is studied. In the used code, only the fixed point iteration (47) is implemented. Either, the selection of the damping parameter as described in [20] can be used or the Anderson acceleration with $l > 0$ vectors and a fixed damping parameter ω. Results are presented for $\omega = 0.5$. The numbers of iterations that were necessary for solving the nonlinear problems are illustrated in Fig. 6. It can be seen that generally fewer iterations were needed for the Zalesak limiter. On the structured grid, the variant with 25 vectors in the Anderson acceleration needed often the smallest number of iterations and the fixed point iteration with an adaptive selection of the damping parameter needed most iterations. But on the unstructured grid, there is no clear picture. Using many vectors in the Anderson iteration did even result in failing to reach the stopping criterion on certain levels. Altogether, these results show a bottleneck of AFC schemes that can hopefully be reduced or cured by using more advanced methods.

6.2 Illustration of the smearing of layers

A motivation for studying convection-diffusion equations in channel geometries comes from the simulation of population balance systems in chemical engineering. For experiments, chemical engineers often use long and thin pipes. That means, the diameter of the pipes is of the order of few millimeters or centimeters and the length of the order of several meters. There are several specific properties when considering convection-diffusion equations in pipes or channels. First, a preferred flow direction exists. Second, the grids are eventually aligned with the flow direction and third, the mesh cells might be anisotropic. For convection-dominated problems there is the experience that it is of advantage to align the grid with the convection. In the literature, one finds already observations that report notable smearing of layers for algebraic stabilizations in examples where the grid is aligned to the convection, e.g., in [23, 11].

This example considers a straight 2d channel, where the convection is a constant vector pointing into the direction of the channel. Let $\Omega = (0,10) \times (0,1)$ and let $\varepsilon = 10^{-10}$, $\boldsymbol{b} = (1,0)^T, c = f = 0$ be the coefficients of the problem. The boundary condition is of impulse form at the inlet of the domain and there is a homogeneous Neumann boundary condition at the outlet:

$$u = \begin{cases} 1 & x = 0, y \in [0.375, 0.625], \\ 0 & x = 0, y \notin [0.375, 0.625], \\ 0 & y = 0 \text{ or } y = 1, \end{cases} \qquad \varepsilon \nabla u \cdot \boldsymbol{n} = 0 \quad \text{on } x = 10, y \in (0,1).$$

DOI 10.20347/WIAS.PREPRINT.2475

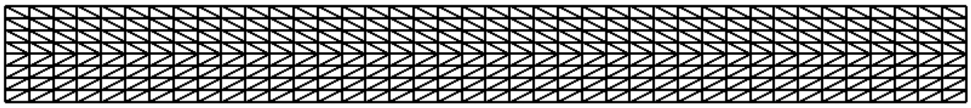

Figure 7: Transport of an impulse: initial grid, level 0.

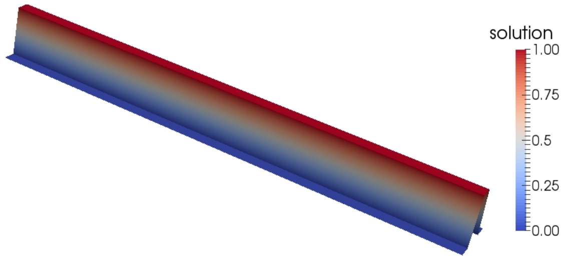

Figure 8: Transport of an impulse: Solution obtained with the SUPG method on level 0.

Because of the very small diffusion coefficient, one expects that the initial condition is transported from the inlet to the outlet.

The coarsest grid is presented in Fig. 7. There are horizontal lines at both position where the inlet condition has its jumps. The same stopping criterion for the solution of the nonlinear problem as in the Hemker example was used.

Applying the SUPG method, one obtains a solution with sharp layers in the whole channel and with basically no spurious oscillations already on level 0, compare Fig. 8. In contrast, the solutions computed with the AFC schemes showed a notable smearing of the layers, in particular the solutions obtained with the Zalesak limiter, see Fig. 9. One can see that the layers become sharper when refining the grid. The solutions computed with the BJK limiter are considerably more accurate than those obtained with the Zalesak limiter, compare Figs. 9 and 10. We could observe that the solution for the Zalesak limiter on level 3 looks similarly accurate as the solution of the BJK limiter on level 1.

The deeper understanding of the reasons for the smearing effect and the finding of remedies are open problems. So far, the probably best explanation is given in [23]. Algebraic stabilizations are by construction multi-dimensional schemes, i.e., there is no dimensional splitting in the construction of the limiters. Such a splitting would be of advantage in this example since it is basically one-dimensional. However, the limiters see the layers of the solution that are vertical to the convection and they do not recognize that it is not necessary to introduce notable diffusion for preventing spurious oscillations.

6.3 A three-dimensional example

Let $\Omega = \Omega_1 \setminus \overline{\Omega_2}$ with $\Omega_1 = (0,5) \times (0,2) \times (0,2)$ and $\Omega_2 = (0.5, 0.8) \times (0.8, 1.2) \times (0.8, 1.2)$. We consider problem (1) with $\varepsilon = 10^{-5}$, $\boldsymbol{b} = (1, g(x), g(x))^T$ where $g(x) = (0.19x^3 - 1.42x^2 + 2.38x)/4$, $u_D = 1$ on $\partial\Omega_1$, $c = 0$, and $u_D = 0$ on $\partial\Omega_2$. An initial mesh containing 842 elements was generated using gmsh and adaptively refined to a mesh containing 1,308,237 elements by using an SUPG method combined with the a posteriori error estimator from [1]. This adaptively refined mesh was then used to obtain approximations using various AFC methods. The nonlinear problems were solved using the damped fixed point algorithm from [20, Figure 12], and the initial guess was obtained using a standard unstabilized Galerkin approximation.

Slides along the plane $z = 1$ of the solution obtained with the different methods are shown in Figs. 11

DOI 10.20347/WIAS.PREPRINT.2475

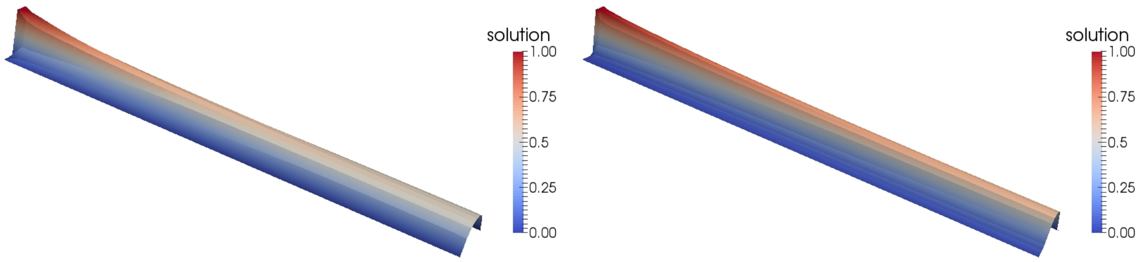

Figure 9: Transport of an impulse: Solutions obtained with the Zalesak limiter on levels 0 and 1.

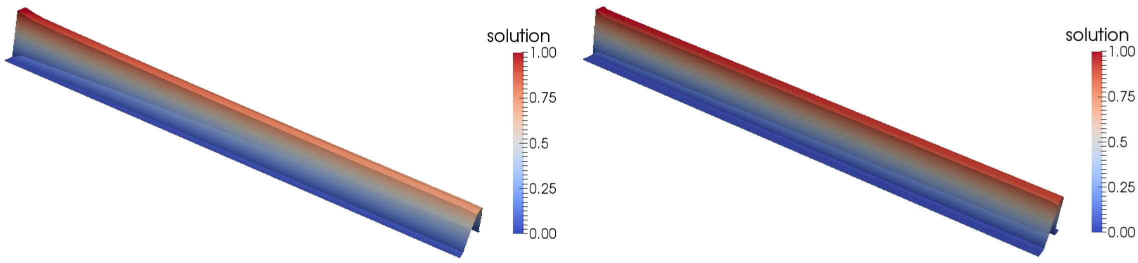

Figure 10: Transport of an impulse: Solutions obtained with the BJK limiter on levels 0 and 1.

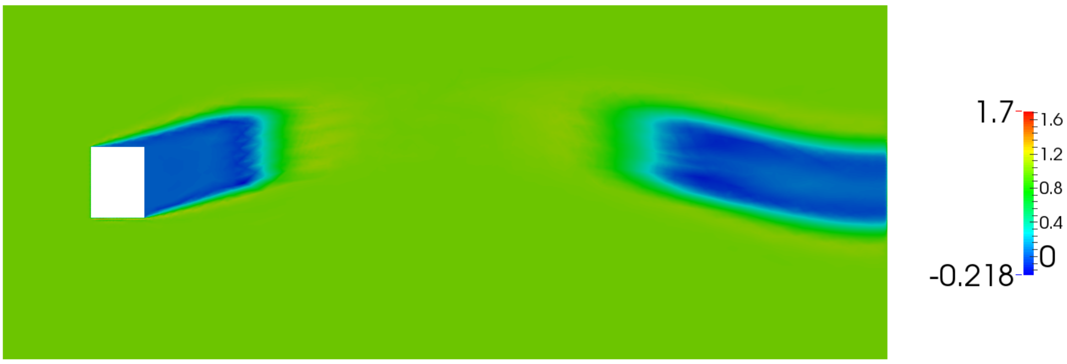

Figure 11: The 3d example: The slice at $z = 1$ of the approximation obtained using the SUPG method on an adaptively refined mesh containing 1,308,237 elements.

– 14. To obtain a reference solution, we pursued this approach further and obtained a sequence of adaptively refined meshes using the same error estimator until we built a highly refined mesh containing 135,408,953 elements. A highly accurate (although not fullly resolved) SUPG solution was computed in this mesh, and a slice of this solution along the plane $z = 1$ this is presented in Fig. 15. Finally, in Figure 16, we compare all these approximations and depict the cross section of them on the line $y = z = 1$.

There is a slight violation of the DMP for the method with the Zalesak limiter. This violation is due to the fact that the mesh does not respect the hypotheses under which the DMP can be shown, cf. Lemma 4.1. For this mesh we have founds violations of this condition, which explains the numerical results and confirms the sharpness of the analytical results. The boundary and inner layers are significantly sharper for the method with the BJK limiter, although this comes at the price of having to perform significantly more fixed point iterations than with the other methods.

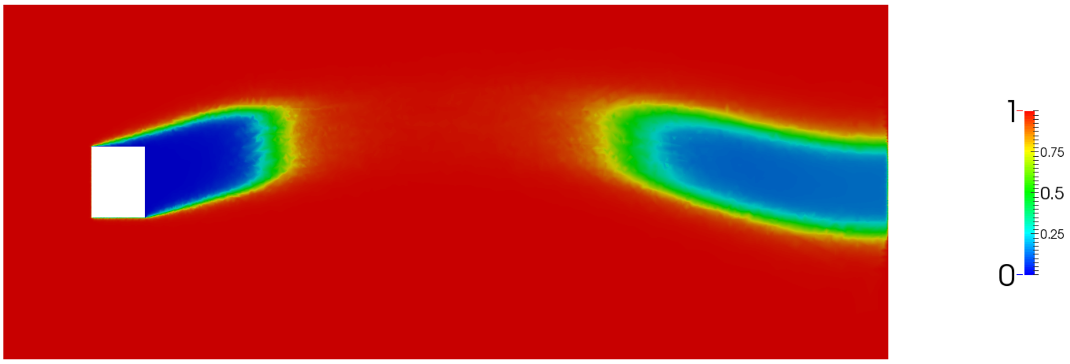

Figure 12: The 3d example: The slice at $z = 1$ of the approximation obtained using the (BBK) method. It took 166 iterations for the Euclidian norm of the residual in the damped fixed point algorithm to be less than the tolerance of 10^{-6}.

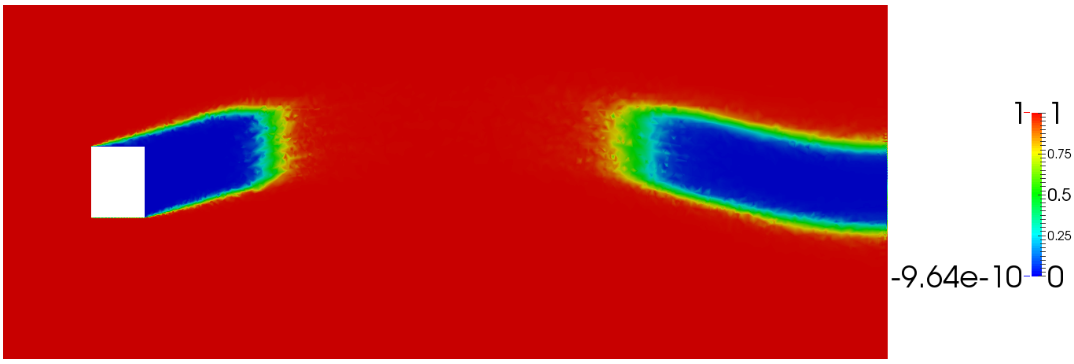

Figure 13: The 3d example: The slice at $z = 1$ of the approximation obtained using the (BJK) method. It took 1117 iterations for the Euclidian norm of the residual in the damped fixed point algorithm to be less than the tolerance of 10^{-6}.

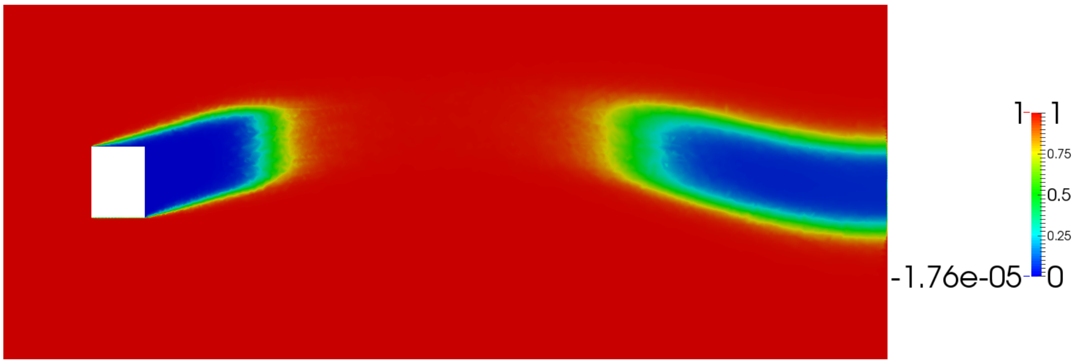

Figure 14: The 3d example: The slice at $z = 1$ of the approximation obtained by the Zalesak limiters. It took 70 iterations for the Euclidean norm of the residual in the damped fixed point algorithm to be less than the tolerance of 10^{-6}.

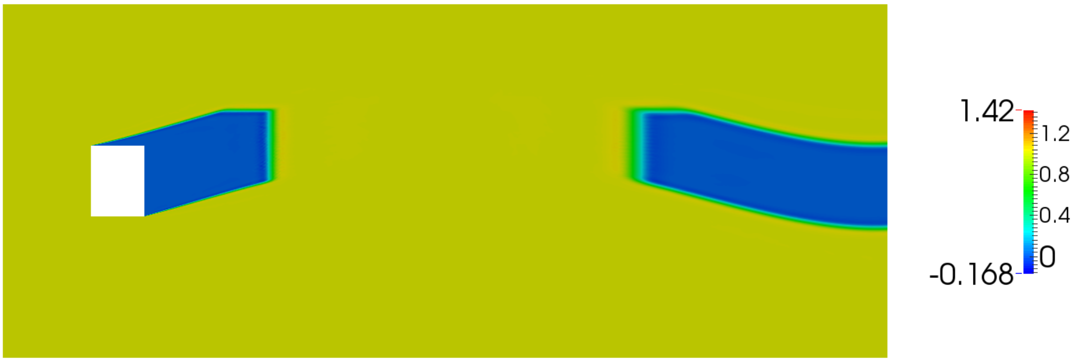

Figure 15: The 3d example: The approximations obtained using the SUPG method on an adaptively refined mesh containing 135,408,953 elements.

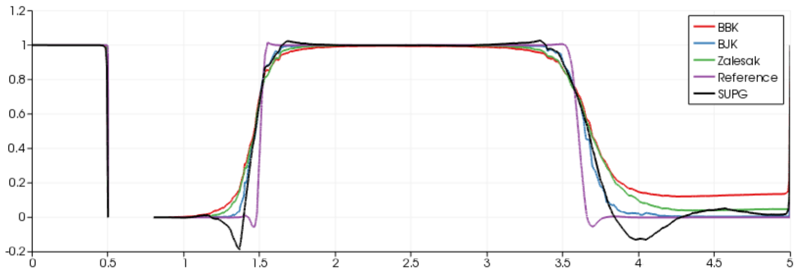

Figure 16: The 3d example: The approximations shown in Figures 11 (SUPG), 12 (BBK), 13 (BJK), 14 (Zalesak) and 15 (Reference) on the line $y = z = 1$.

7 Open problems

The improvement of AFC schemes and the further development of their analysis have been listed in [21] among the most important open problems for H^1-conforming finite elements for convection-diffusion equations. Some concrete issues are the following. It was shown by means of a numerical example that the general a priori estimate given in [8] is sharp. However, one can observe for the Zalesak limiter and the BJK limiter higher orders of convergence than proved in [8], at least on special grids. So far, there is no concrete characterization of the necessary properties of such grids and no corresponding analysis. A priori analysis of AFC schemes for anisotropic grids remains an open problem. In addition, numerical analysis of AFC schemes for time-dependent equations is not available. Last but not least, efficient numerical methods for solving the nonlinear problems have to be developed.

References

[1] Mark Ainsworth, Alejandro Allendes, Gabriel R. Barrenechea, and Richard Rankin. Fully computable a posteriori error bounds for stabilised FEM approximations of convection-reaction-diffusion problems in three dimensions. *Internat. J. Numer. Methods Fluids*, 73(9):765–790, 2013.

[2] Alejandro Allendes, Gabriel R. Barrenechea, and Richard Rankin. Fully computable error estimation of a nonlinear, positivity-preserving discretization of the convection-diffusion-reaction

equation. *SIAM J. Sci. Comput.*, 39(5):A1903–A1927, 2017.

[3] Donald G. Anderson. Iterative procedures for nonlinear integral equations. *J. Assoc. Comput. Mach.*, 12:547–560, 1965.

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

[5] Matthias Augustin, Alfonso Caiazzo, André Fiebach, Jürgen Fuhrmann, Volker John, Alexander Linke, and Rudolf Umla. An assessment of discretizations for convection-dominated convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 200(47-48):3395–3409, 2011.

[6] Gabriel R. Barrenechea, Erik Burman, and Fotini Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math.*, 135(2):521–545, 2017.

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

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

[9] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. 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.

[10] Roland Becker and Malte Braack. A two-level stabilization scheme for the Navier-Stokes equations. In M. Feistauer, V. Dolejší, P. Knobloch, and K. Najzar, editors, *Numerical mathematics and advanced applications*, pages 123–130. Springer, Berlin, 2004.

[11] Róbert Bordás, Volker John, Ellen Schmeyer, and Dominique Thévenin. Numerical methods for the simulation of a coalescence-driven droplet size distribution. *Theoretical and Computational Fluid Dynamics*, 27(3-4):253–271, 2013.

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

[13] Alexander N. Brooks and Thomas J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982. FENOMECH '81, Part I (Stuttgart, 1981).

[14] Erik Burman and Peter Hansbo. Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems. *Comput. Methods Appl. Mech. Engrg.*, 193(15-16):1437–1453, 2004.

[15] Timothy A. Davis. Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Software*, 30(2):196–199, 2004.

[16] Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[17] P. W. Hemker. A singularly perturbed model problem for numerical computation. *J. Comput. Appl. Math.*, 76(1-2):277–285, 1996.

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

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

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

[21] Volker John, Petr Knobloch, and Julia Novo. Finite elements for scalar convection-dominated equations and incompressible flow problems – a never ending story? *Comput. Vis. Sci.*, 2018. in press.

[22] Volker John and Gunar Matthies. MooNMD—a program package based on mapped finite element methods. *Comput. Vis. Sci.*, 6(2-3):163–169, 2004.

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

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

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

[26] Petr Knobloch. On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. In Z. Huang et al., editor, *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016*, volume 120 of *Lecture Notes in Computational Science and Engineering*. Springer, 2017. To appear.

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

[28] Dmitri Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. In Manolis Papadrakakis, Eugenio Oñate, and Bernard Schrefler, editors, *Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering*, pages 1–5. CIMNE, Barcelona, 2007.

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

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

[31] Dmitri Kuzmin and Matthias Möller. Algebraic flux correction I. Scalar conservation laws. In Dmitri Kuzmin, Rainald Löhner, and Stefan Turek, editors, *Flux-Corrected Transport. Principles, Algorithms, and Applications*, pages 155–206. Springer-Verlag, Berlin, 2005.

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

[33] Christoph Lohmann, Dmitri Kuzmin, John N. Shadid, and Sibusiso Mabuza. Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *J. Comput. Phys.*, 344:151–186, 2017.

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

[35] Roger Temam. *Navier-Stokes equations. Theory and numerical analysis*. North-Holland Publishing Co., Amsterdam, 1977. Studies in Mathematics and its Applications, Vol. 2.

[36] Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49(4):1715–1735, 2011.

[37] Pieter Wesseling. *Principles of computational fluid dynamics*, volume 29 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2001.

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