

# A unified analysis of algebraic flux correction schemes for convection–diffusion equations

Gabriel R. Barrenechea<sup>1</sup> · Volker John<sup>2,3</sup> ·  
Petr Knobloch<sup>4</sup> · Richard Rankin<sup>5</sup>

Received: 1 March 2018 / Accepted: 26 April 2018 / Published online: 9 May 2018  
© The Author(s) 2018

**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 schemes. Then, specific versions of the method, that is, different definitions for the flux limiters, are reviewed and their main results stated. Numerical studies compare the different versions of the scheme.

**Keywords** Steady-state convection-diffusion equation · Algebraic flux correction · Edge diffusion · Discrete maximum principle · Comparison of limiters

**Mathematics Subject Classification** 65N30 · 65N12 · 65N15

---

✉ Gabriel R. Barrenechea  
gabriel.barrenechea@strath.ac.uk

Volker John  
john@wias-berlin.de

Petr Knobloch  
knobloch@karlin.mff.cuni.cz

Richard Rankin  
Richard.Rankin@nottingham.edu.cn

<sup>1</sup> Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XH, Scotland, UK

<sup>2</sup> Weierstrass Institute for Applied Analysis and Stochastics, Leibniz Institute in Forschungsverbund Berlin e. V. (WIAS), Mohrenstr. 39, 10117 Berlin, Germany

<sup>3</sup> Department of Mathematics and Computer Science, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany

<sup>4</sup> Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University, Sokolovská 83, 18675 Praha 8, Czech Republic

<sup>5</sup> School of Mathematical Sciences, University of Nottingham Ningbo China, 199 Taikang East Road, Ningbo 315100, China

## 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  $u : \Omega \rightarrow \mathbb{R}$  such that

$$-\varepsilon \Delta u + \mathbf{b} \cdot \nabla u + cu = g \quad \text{in } \Omega, \quad u = u_D \quad \text{on } \partial\Omega, \quad (1)$$

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,  $\mathbf{b} \in W^{1,\infty}(\Omega)^d$  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 [15, 22] was the first milestone in this direction. Solutions computed with this method usually have sharp layers at the correct position, but there are still non-negligible spurious oscillations in a vicinity of layers. Since the publication of [15, 22] 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. [16]) and the local projection stabilization (LPS) method (see [12] 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 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<sup>1</sup> 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 [42, 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 [25, 26], and the

<sup>1</sup> A linear discretization of (1) is a discretization that leads to a linear system of equations.

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, 25, 28, 29]. 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 [14, 43]. Later, the idea was extended to the finite element context, e.g., in [4, 39]. In the last fifteen years, there has been an intensive development of these methods, e.g., see [33–37].

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 [8, 9, 11]. The first paper [8] focuses on the solvability of the nonlinear scheme, while [9] presents the first error analysis of the AFC schemes. Interestingly, the paper [9] also presented negative results, in the sense that it was shown that unless some restrictions are imposed on the mesh, the numerical scheme may not converge. Finally, in the recent paper [11] the role of the linearity preservation was studied. This study is also complemented by the work [10], 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 those 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 Sect. 2, a rewriting as an edge diffusion scheme is presented. A unified analysis is given in Sect. 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 Sect. 4. Strategies for the solution of the nonlinear problems are discussed in Sect. 5. Numerical studies for different limiters used in AFC schemes and the edge diffusion scheme proposed in [10] are presented in Sect. 6. Finally, Sect. 7 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

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

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

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

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

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

- $\{\mathcal{T}_h\}_{h>0}$  denotes a family of shape regular simplicial triangulations of  $\overline{\Omega}$ .
- For a given triangulation  $\mathcal{T}_h$ ,  $\mathcal{E}_h$  denotes the set of its internal edges.

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

$$\mathcal{E}_i := \{E \in \mathcal{E}_h : \mathbf{x}_i \text{ is an endpoint of } E\}.$$

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

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

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

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

- These spaces have standard nodal basis functions denoted by  $\{\varphi_1, \dots, \varphi_N\}$ , uniquely determined by the conditions  $\varphi_i(\mathbf{x}_j) = \delta_{ij}$  for all  $i, j = 1, \dots, N$ . We further notice that  $\text{supp } \varphi_i = \Delta_i$ .
- The Lagrange interpolation operator  $i_h : C^0(\overline{\Omega}) \rightarrow V_h$  is given by

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

In addition, we set

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

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

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

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  $\mathbf{U} \in \mathbb{R}^N$  of its coefficients with respect to the basis  $\{\varphi_1, \dots, \varphi_N\}$  of  $V_h$ . Then  $\mathbf{U} \equiv (u_1, \dots, u_N)$  satisfies the following system of linear equations:

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

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

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.,

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

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

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

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

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

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

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  $\mathbb{D} \mathbf{U}$ . Since the row sums of the matrix  $\mathbb{D}$  vanish, it follows that

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

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:

$$(\mathbb{A} \mathbf{U})_i + \sum_{j=1}^N f_{ij} = g_i, \quad i = 1, \dots, M,$$

$$u_i = u_D(\mathbf{x}_i), \quad i = M + 1, \dots, N.$$

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

$$\sum_{j=1}^N a_{ij} u_j + \sum_{j=1}^N \beta_{ij}(\mathbf{U}) d_{ij} (u_j - u_i) = g_i, \quad i = 1, \dots, M, \tag{11}$$

$$u_i = u_D(\mathbf{x}_i), \quad i = M + 1, \dots, N. \tag{12}$$

For  $\beta_{ij} = 0$ , the original system (5) is recovered. Hence, intuitively, the coefficients  $\beta_{ij}$  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  $\beta_{ij}$  are symmetric, i.e.,

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

This requirement also has a mathematical justification. As a matter of fact, in [8], the possible non-existence of solutions has been shown if this restriction is ignored. Note that (11) does not involve  $\beta_{ij}$  with  $i \in \{M + 1, \dots, N\}$  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 [9] for problem (11), (12): Find  $u_h \in V_h$  such that  $u_h|_{\partial\Omega} = i_h u_D$ , and

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

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

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

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

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

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

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

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

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

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

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

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

From now on we will suppose that, for every edge  $E \in \mathcal{E}_h$ ,  $d_E$  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  $\beta_E$  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  $E \in \mathcal{E}_h$ , the function  $\beta_E(u_h)(\nabla u_h)|_E \cdot \mathbf{t}_E$  is a continuous function of  $u_h \in V_h$ .

It will be shown in Sect. 4 that the limiters defined in [9–11] 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 [38], 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  $\beta_E$ .

#### 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  $T : X \rightarrow X$  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 1 can be found in [40, 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  $T : V_{h,0} \rightarrow [V_{h,0}]'$  by

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

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  $v_h \in V_{h,0}$ ,

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

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

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

Then, the definition of the operator  $T$  yields

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

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

$$\begin{aligned} |a(u_{h,D}, v_h)| &= |\varepsilon(\nabla u_{h,D}, \nabla v_h)_\Omega + (\mathbf{b} \cdot \nabla u_{h,D}, v_h)_\Omega + (c u_{h,D}, v_h)_\Omega| \\ &\leq \varepsilon |u_{h,D}|_{1,\Omega} |v_h|_{1,\Omega} + \sqrt{d} \|\mathbf{b}\|_{\infty,\Omega} |u_{h,D}|_{1,\Omega} \|v_h\|_{0,\Omega} \\ &\quad + \|c\|_{\infty,\Omega} \|u_{h,D}\|_{0,\Omega} \|v_h\|_{0,\Omega} \leq C |u_{h,D}|_{1,\Omega} |v_h|_{1,\Omega}. \end{aligned}$$

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

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

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

### 3.2 The discrete maximum principle

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

To prove the DMP, we make the following general assumption, which is a reformulation of an analogous assumption introduced in [32].

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

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

then

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

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

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

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

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

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

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

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

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

where

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

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

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

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

These properties follow from the fact that  $(\mathbf{b} \cdot \nabla v, v)_{\Omega} = 0$  for any  $v \in H_0^1(\Omega)$ ,  $D_h(u_h; \varphi_i, \varphi_i) \geq 0$ , and  $\sum_{j=1}^N \varphi_j = 1$  in  $\Omega$ .

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

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

To prove (19), let  $c = 0$  in  $\Delta_i$  and assume that  $\max_{\Delta_i} u_h > \max_{\partial\Delta_i} u_h$ . Since the maximum of  $u_h$  on  $\Delta_i$  is attained at vertices of the elements of  $\mathcal{T}_h$  making up  $\Delta_i$ , this means that  $u_h(x_i)$  is the strict maximum of  $u_h$  on  $\Delta_i$ . Then Assumption (A2) implies that the sum in (24) is non-negative. Since  $A_i = 0$  (see (23)) and  $\tilde{a}_{ii} > 0$  (see (22)), there is  $j \in S_i$  such that  $\tilde{a}_{ij} < 0$  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  $\beta_E$  satisfying Assumptions (A1) and (A2). Then*

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

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

If, in addition,  $c = 0$  in  $\Omega$ , then

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

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

*Proof* The proof is based on the technique used in [31, Theorems 5.1 and 5.2]. Let  $u_h \in V_h$  satisfy (14) and let  $g \leq 0$  in  $\Omega$ . Then the nodal values of  $u_h$  satisfy (21) and, due to (7), one has

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

Note that

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

Let

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

First, let us show that

$$\tilde{a}_{ij} \leq 0 \quad \forall i \in J \cap \{1, \dots, M\}, j \notin J. \tag{30}$$

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

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

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

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

As  $u_i - u_j > 0$ , it follows that  $\tilde{a}_{ij} \leq 0$ . For  $j \notin S_i \cup \{i\}$ , one has  $\tilde{a}_{ij} = 0$ , 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  $\Omega$  and hence  $\sum_{j=1}^N \tilde{a}_{ij} = 0$  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  $\sum_{j=1}^N \tilde{a}_{ij} > 0$ , 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

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

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

$$0 \geq \sum_{j \in J} \tilde{a}_{ij} \geq - \sum_{j \notin J} \tilde{a}_{ij} \geq 0,$$

which gives

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

Thus, the matrix  $(\tilde{a}_{ij})_{i,j \in J}$  is singular and hence there exist real numbers  $\{v_i\}_{i \in J}$ , not all equal to zero, such that  $\sum_{i \in J} \tilde{a}_{ij} v_i = 0$  for  $j = 1, \dots, M$ . Consequently, the matrix  $(\tilde{a}_{ij})_{i,j=1}^M$  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)

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

(note that the first inequality implies that  $r > 0$ ). Hence,  $s \leq r$ , 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. □

*Remark 2* The global DMP stated in (27) and (28) assures that the global maximum (resp. minimum) is attained on the boundary of  $\Omega$  but does not exclude that it is also attained at internal nodes. Therefore, it is called the weak DMP. The strong DMP states that the global maximum (resp. minimum) is attained *only* on the boundary of  $\Omega$  and easily follows if  $c = 0$  in  $\Omega$  and  $g$  is negative (resp. positive). Then one has

$$g < 0 \quad \text{in } \Omega \quad \Rightarrow \quad u_h(\mathbf{x}_i) < \max_{\partial\Omega} u_h, \quad i = 1, \dots, M, \tag{32}$$

$$g > 0 \quad \text{in } \Omega \quad \Rightarrow \quad u_h(\mathbf{x}_i) > \min_{\partial\Omega} u_h, \quad i = 1, \dots, M. \tag{33}$$

Indeed, if  $u_h(\mathbf{x}_i) = \max_{\partial\Omega} u_h$  for some  $i \in \{1, \dots, M\}$ , then  $i \in J$  and hence it follows from (21), (23), and (30) that

$$g_i = \sum_{j=1}^N \tilde{a}_{ij} u_j = \sum_{j \notin J} \tilde{a}_{ij} (u_j - u_i) \geq 0,$$

which is a contradiction. The statement (33) follows analogously.

### 3.3 An a priori error estimate

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

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

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

**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*

$$\|u - u_h\|_h \leq C (\varepsilon + c_0^{-1} \{\|\mathbf{b}\|_{\infty,\Omega}^2 + \|c\|_{\infty,\Omega}^2 h^2\})^{1/2} h |u|_{2,\Omega} + D_h(u_h; i_h u, i_h u)^{1/2}.$$

*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

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

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

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

Next, the continuity of  $a$  gives

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

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

$$D_h(u_h; i_h u, e_h) \leq D_h(u_h; i_h u, i_h u)^{\frac{1}{2}} D_h(u_h; e_h, e_h)^{\frac{1}{2}} \leq D_h(u_h; i_h u, i_h u)^{\frac{1}{2}} \|e_h\|_h.$$

Combining the above relations proves the result. □

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

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

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

*one has*

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

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

$$D_h(u_h; i_h u, i_h u) \leq C (\varepsilon + \|\mathbf{b}\|_{\infty,\Omega} h + \|c\|_{\infty,\Omega} h^2) |i_h u|_{1,\Omega}^2.$$

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

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

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

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

which finishes the proof. □

Lemma 2 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 [9] that these results are sharp. On the other hand, the results of [10, 11] 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 AFC vanishes in regions where the approximate solution is a polynomial of degree 1. This property can be formulated in terms of the limiters  $\beta_E$  in the following way.

**Assumption (A3)** The limiters  $\beta_E$  possess the linearity-preservation property, i.e.,

$$\beta_E(u_h) = 0 \quad \text{if } u_h|_{\omega_E} \in \mathbb{P}_1(\omega_E) \quad \forall E \in \mathcal{E}_h.$$

The linearity preservation leads to an improved bound of the consistency error provided that the limiters satisfy the following Lipschitz-continuity assumption.

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

$$\left| \beta_E(u_h)(\nabla u_h)|_E \cdot \mathbf{t}_E - \beta_E(v_h)(\nabla v_h)|_E \cdot \mathbf{t}_E \right| \leq C \sum_{E' \in \mathcal{E}_i \cup \mathcal{E}_j} \left| (\nabla(u_h - v_h))|_{E'} \cdot \mathbf{t}_{E'} \right|,$$

where  $C > 0$  is independent of  $u_h, v_h, i$ , and  $j$ .

**Lemma 3** (Improved estimate of the consistency error) *Let the limiters  $\beta_E$  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 refinement of the technique used in [10, Theorem 4]. Let us write  $D_h = \sum_{E \in \mathcal{E}_h} D_E$  with

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

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

$$\begin{aligned} & |D_E(u_h; u_h, w_h) - D_E(v_h; v_h, w_h)| \\ & \leq C |d_E| h_E^2 \sum_{E' \in \mathcal{E}_i \cup \mathcal{E}_j} \left| (\nabla(u_h - v_h))|_{E'} \cdot \mathbf{t}_{E'} \right| \left| (\nabla w_h)|_E \cdot \mathbf{t}_E \right| \\ & \leq \tilde{C} |d_E| h_E^{2-d} |u_h - v_h|_{1,\omega_E} |w_h|_{1,\omega_E}. \end{aligned} \tag{34}$$

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 2, 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

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

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

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

Using standard finite element approximation results (see [18]),  $i_E u$  satisfies

$$|u - i_E u|_{1, \omega_E} \leq C h_E |u|_{2, \omega_E}. \tag{35}$$

Outside  $\omega_E$ , 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 (34), (35), and the shape regularity of  $\mathcal{T}_h$ , one obtains

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

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 Kuzmin limiter

In this section we review the results obtained when implementing the method with the definition of the limiters proposed in [34]. In that work, the algorithm, originally proposed by Zalesak in [43] was adapted to the steady-state case, and exploited further. We refer then to this limiter as the Kuzmin limiter. This limiter has been used in numerous works, for example [5,9], where a detailed study of its performance for the convection–diffusion equation is carried out. The numbers  $d_E$  in the definition of  $D_h$  are given by (8) in this case.

After having presented the Kuzmin 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 [11, Ex. 7.2] with the help of a numerical example that the AFC scheme with the Kuzmin limiter is not linearity preserving in general.

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

$$P_i^+ := \sum_{\substack{j \in S_i \\ a_{ji} \leq a_{ij}}} f_{ij}^+, \quad P_i^- := \sum_{\substack{j \in S_i \\ a_{ji} \leq a_{ij}}} f_{ij}^-, \quad Q_i^+ := - \sum_{j \in S_i} f_{ij}^-, \quad Q_i^- := - \sum_{j \in S_i} f_{ij}^+, \tag{36}$$

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

$$R_i^+ := \min \left\{ 1, \frac{Q_i^+}{P_i^+} \right\}, \quad R_i^- := \min \left\{ 1, \frac{Q_i^-}{P_i^-} \right\}, \quad i = 1, \dots, M. \tag{37}$$

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.,

$$R_i^+ := 1, \quad R_i^- := 1, \quad i = M + 1, \dots, N. \tag{38}$$

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

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0, \end{cases} \quad \alpha_{ji} := \alpha_{ij}. \tag{39}$$

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

There is an obvious ambiguity in the definition of  $\beta_E$  if  $a_{ij} = a_{ji}$ . This ambiguity does not influence the resulting method if  $\min\{a_{ij}, a_{ji}\} \leq 0$  since then  $d_E = 0$  and the respective term with  $\beta_E$  does not occur in (15). To fulfill the condition  $\min\{a_{ij}, a_{ji}\} \leq 0$ , which also assures the DMP (cf. Lemma 5), it may help to replace the matrix corresponding to the reaction term by a lumped diagonal matrix, see [9].

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

*Proof* Let  $E$  be an internal edge that connects the nodes  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . 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  $\beta_E(u) = \beta_{ij} = \beta_{ji}$ ,  $\alpha_{ij} = \alpha_{ji}$ , we can restrict these considerations to the situation that  $a_{ji} \leq a_{ij}$ . 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 (39), (37), and (36), we obtain

$$\alpha_{ij}(u_h) = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{f_{ij} + \tilde{P}_i^+} \quad \text{with} \quad \tilde{P}_i^+ = \sum_{\substack{k \in S_i \\ a_{ki} \leq a_{ik}, k \neq j}} f_{ik}^+. \tag{40}$$

The numerator and the denominator are continuous functions and by assumption, the denominator is positive in a neighborhood of  $\bar{u}_h$ . Hence  $\alpha_{ij}$  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

$$\alpha_{ij}(u_h) = R_i^- = \frac{\min\{-P_i^-, -Q_i^-\}}{|f_{ij}| - \tilde{P}_i^-} \quad \text{with} \quad \tilde{P}_i^- = \sum_{\substack{k \in S_i \\ a_{ki} \leq a_{ik}, k \neq j}} f_{ik}^-. \tag{41}$$

Using exactly the same reasoning as above, we conclude that  $\alpha_{ij}$  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  $\alpha_{ij}$  is bounded by definition,  $\alpha_{ij}(u_h)(u_j - u_i) \rightarrow 0$  as  $u_j \rightarrow u_i$ . Consequently,  $\alpha_{ij}(u_h)(u_j - u_i)$  is continuous at  $\bar{u}_h$ . □

*Remark 3* In [9] 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 (40) and (41) of the coefficients  $\alpha_{ij}$ . 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*

$$\min\{a_{ij}, a_{ji}\} \leq 0 \quad \forall i = 1, \dots, M, \quad j = 1, \dots, N, \quad i \neq j. \tag{42}$$

*Then the Kuzmin limiter satisfies Assumption (A2).*

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

$$a_{ij} + (1 - \alpha_{ij}(u_h)) d_{ij} \leq 0. \tag{43}$$

If  $a_{ij} \leq 0$ , then (43) holds since  $(1 - \alpha_{ij}(u_h)) d_{ij} \leq 0$ . If  $a_{ij} > 0$ , then  $a_{ji} \leq 0$  due to (42) and hence  $a_{ji} \leq a_{ij}$  and  $d_{ij} = -a_{ij} < 0$ . Thus, if  $u_i > u_k$  for any  $k \in S_i$ , then  $f_{ij} > 0$  and  $f_{ik} \geq 0$  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  $f_{ik} \leq 0$  for  $k \in S_i$ , so that  $\alpha_{ij} = R_i^- = 0$ . Since  $a_{ij} + d_{ij} \leq 0$ , one concludes that (43) holds.  $\square$

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

Here we present a limiter recently proposed in [11] using some ideas of [35]. 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 Kuzmin 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

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

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

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

with fixed constants  $\gamma_i > 0$ . Then one defines the quantities  $R_i^+$  and  $R_i^-$  again by (37) and (38) and one sets

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

Finally, the limiters are defined by  $\beta_E := 1 - \min\{\tilde{\alpha}_{ij}, \tilde{\alpha}_{ji}\}$  for any internal edge  $E \in \mathcal{E}_h$  having the endpoints  $x_i, x_j$ .

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

*Proof* The validity of Assumption (A1) follows analogously as in the proof of Lemma 4. Let us prove Assumption (A2). Consider any  $u_h \in V_h, i \in \{1, \dots, M\}$ , and  $j \in S_i$  and assume that  $u_i := u_h(x_i)$  is a strict local extremum of  $u_h$  in  $\Delta_i$ . Then we want to prove that

$$a_{ij} + (1 - \min\{\tilde{\alpha}_{ij}(u_h), \tilde{\alpha}_{ji}(u_h)\}) d_{ij} \leq 0. \tag{44}$$

If  $d_{ij} = 0$ , then  $a_{ij} \leq 0$  and hence (44) 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  $\tilde{\alpha}_{ij} = R_i^+ = 0$ . Since  $a_{ij} + d_{ij} \leq 0$ , one obtains (44). If  $u_i < u_k$  for any  $k \in S_i$ , (44) follows analogously.  $\square$

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

$$u_i - u_i^{\min} \leq \gamma_i (u_i^{\max} - u_i) \quad \forall u \in \mathbb{P}_1(\mathbb{R}^d). \tag{45}$$

It was proved in [11] that (45) holds with  $\gamma_i = 1$  if the patch  $\Delta_i$  is symmetric with respect to the vertex  $\mathbf{x}_i$ , and with

$$\gamma_i = \frac{\max_{\mathbf{x}_j \in \partial \Delta_i} |\mathbf{x}_i - \mathbf{x}_j|}{\text{dist}(\mathbf{x}_i, \partial \Delta_i^{\text{conv}})}$$

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

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

*Proof* Consider any  $i \in \{1, \dots, M\}$ . 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 (45)

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

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

*Remark 4* Note that large values of the constants  $\gamma_i$  cause that more limiters  $\alpha_{ij}$  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  $\gamma_i$ '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 [10] and its main results. This limiter, also referred to as *smoothness-based viscosity*, has its origin in the finite volume literature (see, e.g., [24] and [23]), and has also been used (although in a slightly modified way) in the recent work [19].

The numbers  $d_E$  in the definition of  $D_h$  are given by  $d_E = \gamma_0 h_E^{d-1}$ , where  $\gamma_0$  is a fixed parameter, dependent on the data of (1). The limiters  $\beta_E$ ,  $E \in \mathcal{E}_h$ , are given by the following algorithm: for  $w_h \in V_h$ , one defines  $\xi_{w_h}$  as the unique element in  $V_h$  whose nodal values are given by

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

Then, on each  $E \in \mathcal{E}_h$ ,  $\beta_E$  is defined by

$$\beta_E(w_h) := \max_{\mathbf{x} \in E} [\xi_{w_h}(\mathbf{x})]^p, \quad p \in [1, +\infty). \tag{46}$$

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 [10] 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 [10] for details).

Two remarks can be rapidly made about this definition of the limiter. First, if a function  $v_h$  has an extremum at an internal node  $x_i$ , and  $E \in \mathcal{E}_i$ , 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  $x_i$ , and every  $j \in S_i$ , there exists  $k \in S_i$  such that  $x_k - x_i = -(x_j - x_i)$ . So, if the mesh is symmetric with respect to its internal nodes, if  $E \in \mathcal{E}_h$  has endpoints  $x_i$  and  $x_j$ , and  $v_h \in \mathbb{P}_1(\omega_E)$ , then

$$\sum_{l \in S_i} (v_h(x_i) - v_h(x_l)) = 0 \quad \text{and} \quad \sum_{l \in S_j} (v_h(x_j) - v_h(x_l)) = 0,$$

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 5* In [10, 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 (46) satisfies Assumptions (A1), (A2), and (A4).

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

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

and  $\gamma_0 \geq C_0 \|b\|_{\infty, \Omega} + C_1 \|c\|_{\infty, \Omega} h$  (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 [10, Lemma 1] the following result is proven: for any internal node  $x_i$  the following holds

$$|\xi_{u_h}(x_i) - \xi_{v_h}(x_i)| \leq 4 \frac{\sum_{E' \in \mathcal{E}_i} h_{E'} |\nabla(u_h - v_h) \cdot \mathbf{t}_{E'}|}{\sum_{E' \in \mathcal{E}_i} h_{E'} (|\nabla u_h \cdot \mathbf{t}_{E'}| + |\nabla v_h \cdot \mathbf{t}_{E'}|)}.$$

Let us now suppose that, for  $E \in \mathcal{E}_h$  having endpoints  $x_i, x_j$ ,  $\beta_E(u_h) = \xi_{u_h}(x_i)$  and  $\beta_E(v_h) = \xi_{v_h}(x_j)$ . Then, using that  $0 \leq \xi_{v_h}(x_i) \leq 1$  one obtains

$$\begin{aligned} \beta_E(u_h) \nabla u_h \cdot \mathbf{t}_E - \beta_E(v_h) \nabla v_h \cdot \mathbf{t}_E &\leq (\xi_{u_h}(x_i) - \xi_{u_h}(x_j)) \nabla u_h \cdot \mathbf{t}_E \\ &\quad + |\xi_{u_h}(x_j)| |\nabla(u_h - v_h) \cdot \mathbf{t}_E| + |\xi_{u_h}(x_j) - \xi_{v_h}(x_j)| |\nabla v_h \cdot \mathbf{t}_E| \\ &\leq (\xi_{u_h}(x_i) - \xi_{u_h}(x_j)) \nabla u_h \cdot \mathbf{t}_E + 5 \sum_{E' \in \mathcal{E}_j} |\nabla(u_h - v_h) \cdot \mathbf{t}_{E'}|. \end{aligned}$$

In a completely analogous way one obtains

$$\begin{aligned} & \beta_E(u_h)\nabla u_h \cdot \mathbf{t}_E - \beta_E(v_h)\nabla v_h \cdot \mathbf{t}_E \\ & \leq (\xi_{v_h}(\mathbf{x}_i) - \xi_{v_h}(\mathbf{x}_j))\nabla u_h \cdot \mathbf{t}_E + 5 \sum_{E' \in \mathcal{E}_i} |\nabla(u_h - v_h) \cdot \mathbf{t}_{E'}|. \end{aligned}$$

Thus, since  $\xi_{u_h}(\mathbf{x}_i) - \xi_{u_h}(\mathbf{x}_j) \geq 0$  and  $\xi_{v_h}(\mathbf{x}_i) - \xi_{v_h}(\mathbf{x}_j) \leq 0$ ,

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

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  $\mathbf{x}_i$ . Let  $j \in S_i$ , and let  $E \in \mathcal{E}_h$  be the edge with endpoints  $\mathbf{x}_i$  and  $\mathbf{x}_j$ . Then, as was mentioned earlier,  $\beta_E(u_h) = 1$ . Thus, using the shape regularity of the mesh, one obtains

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

and Assumption (A2) follows. □

It follows from Lemma 2 that the consistency error  $D_h(u_h; i_h u, i_h u)$  can be bounded as follows:

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

Moreover, if the mesh is symmetric with respect to its internal nodes, then Lemma 3 implies that the following bound holds for the consistency error

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

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  $O(h)$ .

### 4.4 Related recent work

We finish this section by mentioning that, to the research reviewed in this paper, work has been done in parallel, e.g., in [6, 7, 19, 20]. In those references, a stabilizing term similar to the one defined in (15) is added to the formulation, and referred to as the *Graph Laplacian*. The stabilizing mechanism of the methods presented in those works and the ones reviewed in this manuscript are very similar. There are, nevertheless, significant differences in the limiters  $\beta_E$ , some of the results obtained in terms of the satisfaction of the discrete maximum principle, and the linearity preservation of the final schemes.

For example, in [6] the emphasis is in the regularization of the limiter proposed there (related to the BBK one) in order to make the limiter differentiable, to allow the use of Newton’s method to solve the nonlinear system. The regularization proposed in there used regularization parameters that had an impact on the performance of the method. In addition, although the results concerning the discrete maximum principle were not too different from the ones reviewed in this work, the linearity preservation was not guaranteed for the regularized limiters. In [19] the definition of the limiter (non-differentiable this time) is modified using generalized barycentric coordinates in order to make it differentiable on meshes for which the support of basis functions is convex.

A final important difference between the works reviewed in this paper and the above-mentioned references consists in the emphasis. While the papers just quoted deal with first order hyperbolic systems, the results reviewed in this paper deal with the convection–diffusion equation. The presence of the Laplacian in the partial differential equation makes the method satisfy very different properties, especially on non-Delaunay meshes, as it can be seen in [9] where an example of non-convergence was given for the Kuzmin limiter.

### 5 Iterative schemes for solving the nonlinear problem

Consider the weak formulation (13) and the equivalent formulation (11), (12) in matrix-vector notation. For simplicity, we will restrict the discussion to the case of homogeneous boundary conditions. These formulations represent a nonlinear problem since the coefficients  $\beta_{ij}$  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

$$a \left( \tilde{u}_h^{(m+1)}, v_h \right) + D_h \left( u_h^{(m)}; \tilde{u}_h^{(m+1)}, v_h \right) = (g, v_h)_{\Omega} \quad \forall v_h \in V_{h,0}. \tag{47}$$

The matrix-vector form of (47) is

$$\sum_{j=1}^N a_{ij} \tilde{u}_j^{(m+1)} + \sum_{j=1}^N \beta_{ij}^{(m)} d_{ij} \left( \tilde{u}_j^{(m+1)} - \tilde{u}_i^{(m+1)} \right) = g_i, \quad i = 1, \dots, M, \tag{48}$$

$$\tilde{u}_i^{(m+1)} = 0, \quad i = M + 1, \dots, N,$$

where  $\beta_{ij}^{(m)} = \beta_{ij} \left( u^{(m)} \right)$ . In the iterations (47) and (48), 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 (48) the iteration

$$\begin{aligned} \sum_{j=1}^N (a_{ij} + d_{ij}) \tilde{u}_j^{(m+1)} &= g_i + \sum_{j=1}^N (1 - \beta_{ij}^{(m)}) d_{ij} (u_j^{(m)} - u_i^{(m)}) \\ &= g_i + \sum_{j=1}^N (1 - \beta_{ij}^{(m)}) f_{ij}^{(m)}, \quad i = 1, \dots, M, \\ \tilde{u}_i^{(m+1)} &= 0, \quad i = M + 1, \dots, N. \end{aligned} \tag{49}$$

Using a sparse direct solver, then the matrix of (49) 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

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

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

$$DF(u^{(m)}) \tilde{u}_h^{(m+1)} = DF(u^{(m)}) u_h^{(m)} - F(u^{(m)}), \tag{50}$$

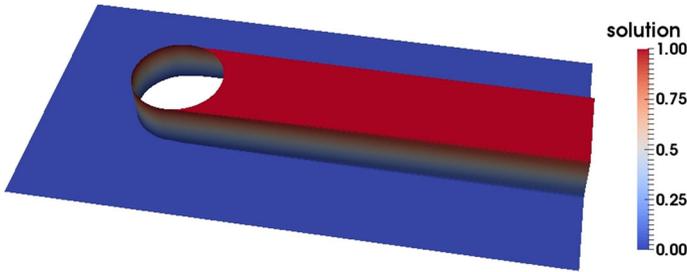
where  $DF(u^{(m)})$  is the Jacobian, which can be computed by applying the product rule and the chain rule, and observing that the derivative of the limiter with respect to the Dirichlet nodes is not needed since these values are fixed

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

Hence, the entries of the matrix that has to be inverted in (50) are given by

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

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



**Fig. 1** Hemker example: reference solution for  $\varepsilon = 10^{-4}$  on a fine grid

the particular limiter that is used in the simulations. Note that the derivation presented above requires smoothness of the limiters. If this property is not given, a possible approach consists in modifying the limiters such that they become sufficiently smooth, as is proposed in [6] for a transport equation. The exploration of strategies for applying Newton-type methods to (modified) AFC schemes for convection–diffusion–reaction equations is currently ongoing and will be reported elsewhere.

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 [26], an automatic strategy for adapting the parameter during the iteration is described. In [5], the use of the so-called Anderson acceleration, proposed in [3,41], is advocated. The Anderson acceleration stores vectors from previous iterations and builds with them second order information.

## 6 Numerical studies

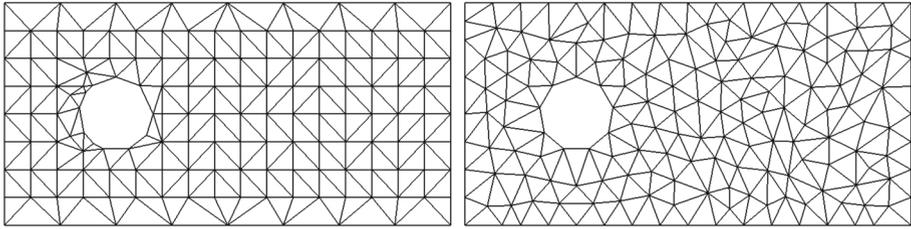
### 6.1 The Hemker example

We will consider the so-called Hemker example, which was proposed in [21]. It models the convection of temperature from a hot circle (2d cylinder) in a channel. The convection field is constant. There are exponential layers at the circle and interior layers downstream from 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.

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

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

In [5], a reference solution on a fine grid containing 48,252,416 degrees of freedom on a  $\mathbb{Q}_1$  mesh (for details, see [5]) was computed for  $\varepsilon = 10^{-4}$ , see Fig. 1. Quantities of interest defined in [5] are the magnitude of the over- and under-shoots, the difference to the reference solution on selected cut lines, and the smearing of the interior layer at a certain cutline downstream from the cylinder. For the concrete definition of these quantities see [5].



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

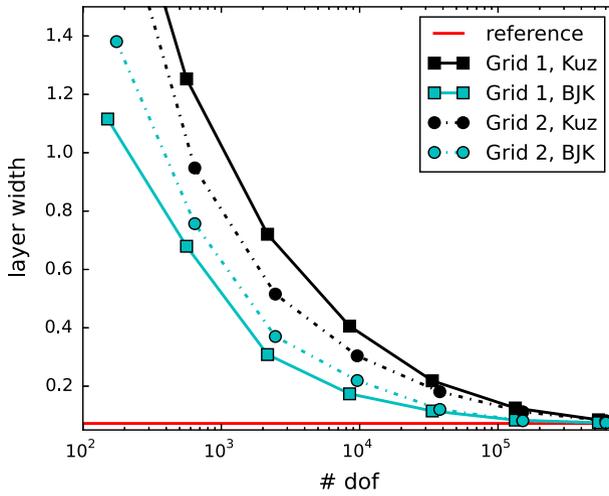
The simulations were performed with  $\mathbb{P}_1$  finite elements on two types of grids, see Fig. 2. Grid 1 is aligned downstream to the convection and it has edges at the position where the interior layer is expected. The stopping criterion for the nonlinear iteration was based on the Euclidean norm of the residual vector, which should be smaller or equal than  $10^{-13} (\#\text{dof})^{\frac{1}{2}}$ , where  $\#\text{dof}$  is the number of degrees of freedom (including Dirichlet nodes) on the respective 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, [17]. The simulations were performed with the code MooNMD [27].

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.

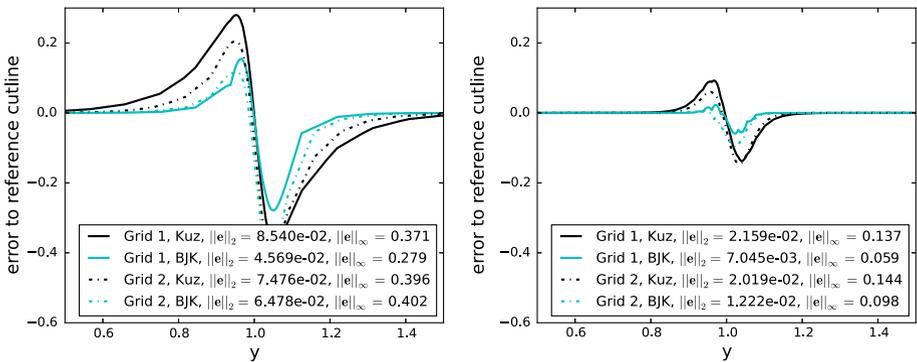
Another quantity of interest studies the smearing of the interior layer at  $x = 4$ , see Fig. 3. It can be seen that the smearing introduced by the BJK limiter is always smaller than with the Kuzmin limiter. In particular, on the aligned Grid 1, the results with the BJK limiter are much better. This 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  $\mathbf{e}$  contains the differences of the reference solution and the numerical solution in these points. The errors in the Euclidean norm  $\|\mathbf{e}\|_2$  and the maximum norm  $\|\mathbf{e}\|_\infty$  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.

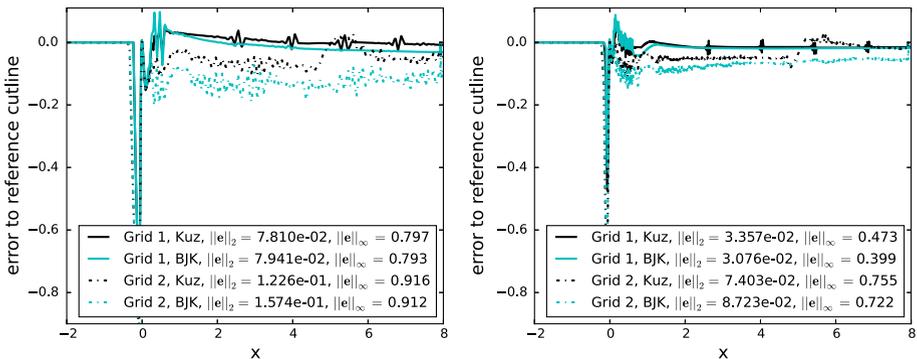
Finally, the costs for solving the nonlinear problems is studied. In the used code, only the fixed point iteration (49) is implemented. Either, the selection of the damping parameter as described in [26] can be used or the Anderson acceleration with  $l > 0$  vectors and a fixed damping parameter  $\omega$ . Results are presented for  $\omega = 0.5$  and  $l = 5, 10, 25$  vectors in the Anderson acceleration. 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 Kuzmin 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. For example, results have been obtained for other values of the damping parameter  $\omega$  (not reported here due to space restrictions). More precisely, the use of  $\omega = 0.25$  gave similar results to the ones shown



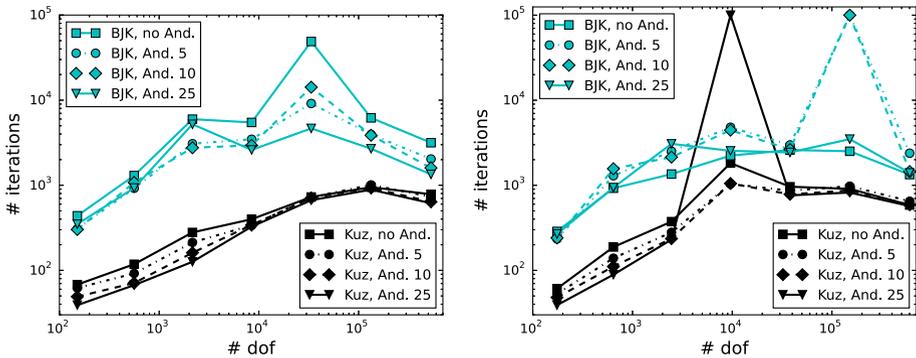
**Fig. 3** Hemker example: width of the interior layer at  $x = 4$



**Fig. 4** Hemker example: errors to the outline at  $x = 4$ , left level 3 (nearly 10,000 d.o.f.s), right level 5 (nearly 150,000 d.o.f.s)



**Fig. 5** Hemker example: errors to the outline at  $y = 1$ , left level 3 (nearly 10,000 d.o.f.s), right level 5 (nearly 150,000 d.o.f.s)



**Fig. 6** Hemker example: number of iterations for solving the nonlinear problems, for  $\omega = 0.5, l = 5, 10, 25$  vectors in the Anderson acceleration, on Grid 1 (left), Grid 2 (right); 100,000 iterations means that the stopping criterion was not reached

in Fig. 6, needing a few more iterations in most cases. Using  $\omega = 0.7$ , we observed that the stopping criterion was not reached in many more cases than for  $\omega = 0.5$ .

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

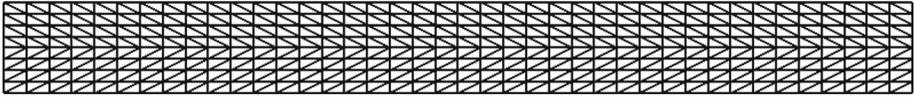
A motivation for studying convection–diffusion equations in channel geometries comes from the simulation of population balance systems in chemical engineering. For experiments, chemical engineers often use long and thin pipes. That means, the diameter of the pipes is of the order of a few millimeters 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 [13,28].

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  $\varepsilon = 10^{-10}$ ,  $\mathbf{b} = (1, 0)^T, c = g = 0$  be the coefficients of the problem. The boundary condition is an impulse in a center strip at the inlet of the domain and there is a homogeneous Neumann boundary condition at the outlet:

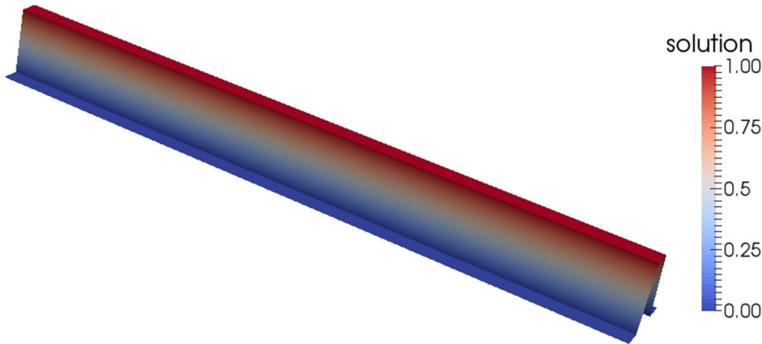
$$u = \begin{cases} 1 & x = 0, y \in [0.375, 0.625], \\ 0 & x = 0, y \notin [0.375, 0.625], \\ 0 & y = 0 \text{ or } y = 1, \end{cases} \quad \varepsilon \nabla u \cdot \mathbf{n} = 0 \quad \text{on } x = 10, y \in (0, 1).$$

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

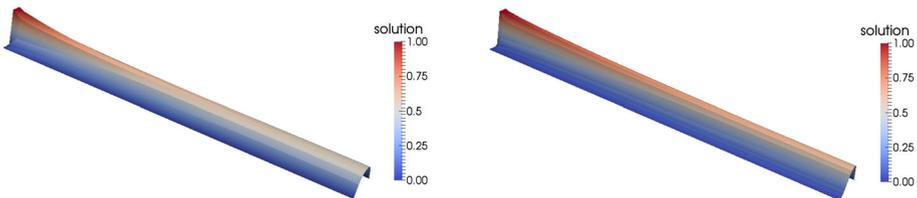
The coarsest grid is presented in Fig. 7. There are horizontal lines at both positions where the inlet condition has its jumps. The same stopping criterion for the solution of the nonlinear problem as in the Hemker example was used.



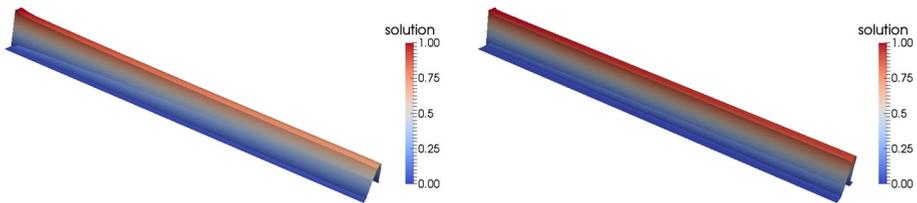
**Fig. 7** Transport of an impulse: initial grid, level 0



**Fig. 8** Transport of an impulse: solution obtained with the SUPG method on level 0



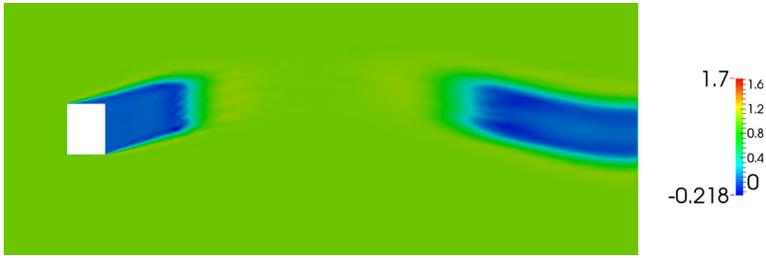
**Fig. 9** Transport of an impulse: solutions obtained with the Kuzmin limiter on levels 0 (left) and 1 (right)



**Fig. 10** Transport of an impulse: solutions obtained with the BJK limiter on levels 0 (left) and 1 (right)

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 Kuzmin 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 Kuzmin limiter, compare Figs. 9 and 10. We could observe that the solution for the Kuzmin 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 [28]. Algebraic stabilizations are by construction multi-dimensional schemes, i.e., there is no dimensional



**Fig. 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

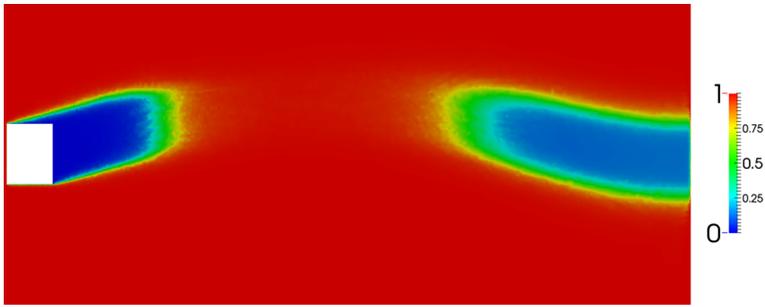
splitting in the construction of the limiters. Such a splitting would be of advantage in this example since it is basically one-dimensional. However, the limiters see the layers of the solution that are vertical to the convection and they do not recognize that it is not necessary to introduce notable diffusion for preventing spurious oscillations.

### 6.3 A three-dimensional example

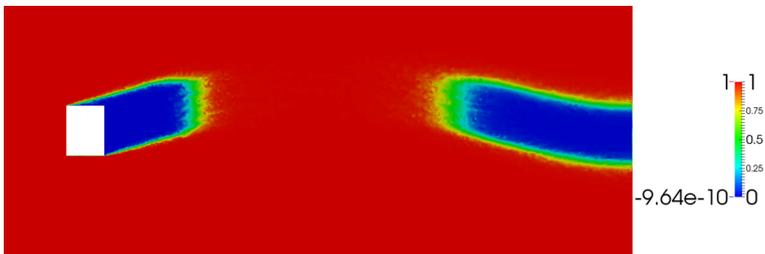
Let  $\Omega = \Omega_1 \setminus \overline{\Omega_2}$  with  $\Omega_1 = (0, 5) \times (0, 2) \times (0, 2)$  and  $\Omega_2 = (0.5, 0.8) \times (0.8, 1.2) \times (0.8, 1.2)$ . We consider problem (1) with  $\varepsilon = 10^{-5}$ ,  $\mathbf{b} = (1, \ell(x), \ell(x))^T$  where  $\ell(x) = (0.19x^3 - 1.42x^2 + 2.38x)/4$ ,  $u_D = 1$  on  $\partial\Omega_1$ ,  $c = g = 0$ , 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 [26, Figure 12], and the initial guess was obtained using a standard unstabilized Galerkin approximation.

Slices along the plane  $z = 1$  of the solution obtained with the different methods are shown in Figs. 11, 12, 13, 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 a highly refined mesh, containing 135,408,953 elements, was built. A highly accurate (although not fully resolved) SUPG solution was computed in this mesh, and a slice of this solution along the plane  $z = 1$  is presented in Fig. 15. Finally, in Fig. 16, we compare all these approximations and depict the cross section of them on the line  $y = z = 1$ .

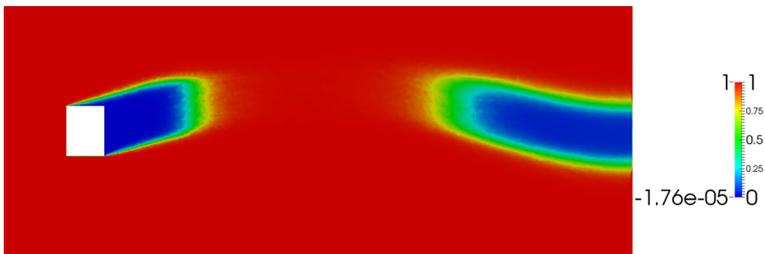
There is a slight violation of the DMP for the method with the Kuzmin limiter. This violation is due to the fact that the mesh does not respect the hypotheses under which the DMP can be shown, cf. Lemma 5. For this mesh we have found violations of this condition, which explains the numerical results 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. In fact, for this example the method using the Kuzmin limiter took 70 iterations to reach convergence, while the method using the BBK limiter took 166 iterations, and the use of the BJK limiter took 1117 iterations to reach convergence. As was mentioned earlier, the BJK and Kuzmin limiters provide sharper profiles than the BBK one. This has been observed not only in this example. This behavior seems to be related to the equal weight given to all fluxes by the construction of the BBK limiter, different from the Kuzmin one (essentially, an upwind limiter), and the BJK limiter, which has the flux associated to



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



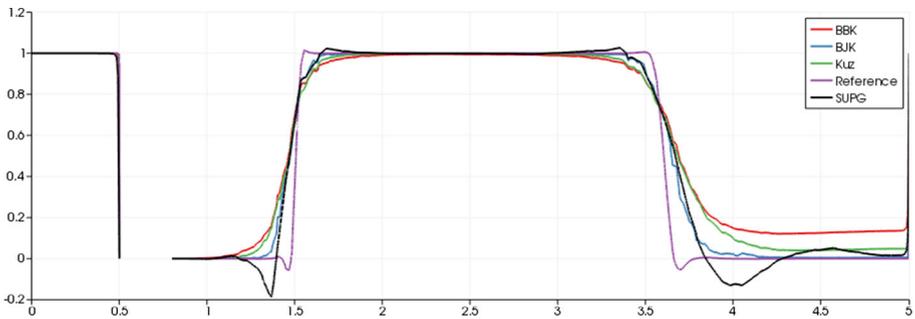
**Fig. 13** The 3d example: the slice at  $z = 1$  of the approximation obtained using the BJK method. It took 1117 iterations for the Euclidean norm of the residual in the damped fixed point algorithm to be less than the tolerance of  $10^{-6}$



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



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



**Fig. 16** The 3d example: the approximations shown in Figs. 11 (SUPG), 12 (BBK), 13 (BJK), 14 (Kuz) and 15 (reference) on the line  $y = z = 1$

the local extremum as main ingredient, and also includes some explicit mesh information to make it linearity preserving.

## 7 Open problems

The improvement of AFC schemes and the further development of their analysis have been listed in [30] 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 [9] is sharp. However, one can observe for the Kuzmin limiter and the BJK limiter higher orders of convergence than proved in [9], 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.

**Open Access** This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

## References

1. Ainsworth, M., Allendes, A., Barrenechea, G.R., Rankin, R.: Fully computable a posteriori error bounds for stabilised FEM approximations of convection–reaction–diffusion problems in three dimensions. *Int. J. Numer. Methods Fluids* **73**(9), 765–790 (2013)
2. Allendes, A., Barrenechea, G.R., Rankin, R.: 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. Anderson, D.G.: Iterative procedures for nonlinear integral equations. *J. Assoc. Comput. Mach.* **12**, 547–560 (1965)
4. Arminjon, P., Dervieux, A.: Construction of TVD-like artificial viscosities on two-dimensional arbitrary FEM grids. *J. Comput. Phys.* **106**(1), 176–198 (1993)

5. 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. Eng.* **200**(47–48), 3395–3409 (2011)
6. Badia, S., Bonilla, J.: Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Comput. Methods Appl. Mech. Eng.* **313**, 133–158 (2017)
7. Badia, S., Hierro, A.: On monotonicity-preserving stabilized finite element approximations of transport problems. *SIAM J. Sci. Comput.* **36**(6), A2673–A2697 (2014)
8. 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)
9. Barrenechea, G.R., John, V., Knobloch, P.: Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.* **54**(4), 2427–2451 (2016)
10. 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)
11. 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)
12. Becker, R., Braack, M.: A two-level stabilization scheme for the Navier–Stokes equations. In: Feistauer, M., Dolejší, V., Knobloch, P., Najzar, K. (eds.) *Numerical Mathematics and Advanced Applications*, pp. 123–130. Springer, Berlin (2004)
13. Bordás, R., John, V., Schmeier, E., Thévenin, D.: Numerical methods for the simulation of a coalescence-driven droplet size distribution. *Theor. Comput. Fluid Dyn.* **27**(3–4), 253–271 (2013)
14. 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)
15. Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov–Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations. *Comput. Methods Appl. Mech. Eng.* **32**(1–3), 199–259 (1982). FENOMECH '81, Part I (Stuttgart, 1981)
16. Burman, E., Hansbo, P.: Edge stabilization for Galerkin approximations of convection–diffusion–reaction problems. *Comput. Methods Appl. Mech. Eng.* **193**(15–16), 1437–1453 (2004)
17. Davis, T.A.: Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.* **30**(2), 196–199 (2004)
18. Ern, A., Guermond, J.-L.: *Theory and Practice of Finite Elements*, Applied Mathematical Sciences, vol. 159. Springer, New York (2004)
19. Guermond, J.-L., Popov, B.: Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal.* **55**(6), 3120–3146 (2017)
20. 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)
21. Hemker, P.W.: A singularly perturbed model problem for numerical computation. *J. Comput. Appl. Math.* **76**(1–2), 277–285 (1996)
22. Hughes, T.J.R., Brooks, A.: 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), AMD, vol. 34, pp. 19–35. Amer. Soc. Mech. Engrs. (ASME), New York (1979)
23. Jameson, A.: Origins and further development of the Jameson–Schmidt–Turkel scheme. *AIAA J.* **55**, 1487–1510 (2017)
24. Jameson, A., Schmidt, W., Turkel, E.: Numerical solution of the Euler equations by finite volume methods using Runge–Kutta time-stepping schemes. *AIAA J.* In: *14th AIAA Fluid and Plasma Dynamics Conference*, AIAA paper 1981-1259 (1981)
25. John, V., Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. I. A review. *Comput. Methods Appl. Mech. Eng.* **196**(17–20), 2197–2215 (2007)
26. John, V., Knobloch, P.: On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations. II. Analysis for  $P_1$  and  $Q_1$  finite elements. *Comput. Methods Appl. Mech. Eng.* **197**(21–24), 1997–2014 (2008)
27. John, V., Matthies, G.: MoonMMD—a program package based on mapped finite element methods. *Comput. Vis. Sci.* **6**(2–3), 163–169 (2004)
28. John, V., Novo, J.: On (essentially) non-oscillatory discretizations of evolutionary convection–diffusion equations. *J. Comput. Phys.* **231**(4), 1570–1586 (2012)
29. John, V., Schmeier, E.: Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion. *Comput. Methods Appl. Mech. Eng.* **198**(3–4), 475–494 (2008)

30. John, V., Knobloch, P., Novo, J.: Finite elements for scalar convection-dominated equations and incompressible flow problems—a never ending story? *Comput. Vis. Sci.* (2018). <https://doi.org/10.1007/s00791-018-0290-5>
31. Knobloch, P.: Numerical solution of convection–diffusion equations using a nonlinear method of upwind type. *J. Sci. Comput.* **43**(3), 454–470 (2010)
32. Knobloch, P.: On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. In: Zhang, Z., Huang, Z., Stynes, M. (eds.) *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016, Lecture Notes in Computational Science and Engineering*, vol. 120, pp. 129–139. Springer, New York (2017)
33. 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)
34. Kuzmin, D.: Algebraic flux correction for finite element discretizations of coupled systems. In: Papadrakakis, M., Oñate, E., Schrefler, B. (eds.) *Proceedings of the International Conference on Computational Methods for Coupled Problems in Science and Engineering*, pp. 1–5. CIMNE, Barcelona (2007)
35. Kuzmin, D.: Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *J. Comput. Appl. Math.* **236**(9), 2317–2337 (2012)
36. Kuzmin, D., Möller, M.: Algebraic flux correction I. Scalar conservation laws. In: Kuzmin, D., Löhner, R., Turek, S. (eds.) *Flux-Corrected Transport. Principles, Algorithms, and Applications*, pp. 155–206. Springer, Berlin (2005)
37. Kuzmin, D., Turek, S.: High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter. *J. Comput. Phys.* **198**(1), 131–158 (2004)
38. 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)
39. Löhner, R., Morgan, K., Peraire, J., Vahdati, M.: Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier–Stokes equations. *Int. J. Numer. Methods Fluids* **7**(10), 1093–1109 (1987)
40. Temam, R.: *Navier–Stokes Equations. Theory and Numerical Analysis. Studies in Mathematics and its Applications*, vol. 2. North-Holland, Amsterdam, (1977)
41. Walker, H.F., Ni, P.: Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.* **49**(4), 1715–1735 (2011)
42. Wesseling, P.: *Principles of Computational Fluid Dynamics*, Springer Series in Computational Mathematics, vol. 29. Springer, Berlin (2001)
43. Zalesak, S.T.: Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.* **31**(3), 335–362 (1979)