Mathematical Models and Methods in Applied Sciences
Vol. 27, No. 3 (2017) 525–548

DOI: 10.1142/S0218202517500087

World Scientific
www.worldscientific.com

An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes

Gabriel R. Barrenechea

Department of Mathematics and Statistics,
University of Strathclyde,
26 Richmond Street, Glasgow G1 1XH, UK
gabriel.barrenechea@strath.ac.uk

Volker John

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin, Germany
and
Department of Mathematics and Computer Science,
Free University of Berlin,
Arnimallee 6, 14195 Berlin, Germany
john@wias-berlin.de

Petr Knobloch*

Department of Numerical Mathematics,
Charles University, Faculty of Mathematics and Physics,
Sokolovská 83, 18675 Praha 8, Czech Republic
knobloch@karlin.mff.cuni.cz

Received 8 June 2016
Revised 21 December 2016
Accepted 23 December 2016
Published 3 March 2017
Communicated by E. Süli

This work is devoted to the proposal of a new flux limiter that makes the algebraic flux correction finite element scheme linearity and positivity preserving on general simplicial meshes. Minimal assumptions on the limiter are given in order to guarantee the validity of the discrete maximum principle, and then a precise definition of it is proposed and analyzed. Numerical results for convection–diffusion problems confirm the theory.

Keywords: Finite element method; convection–diffusion equation; algebraic flux correction; discrete maximum principle; linearity preservation.

AMS Subject Classification: 65N30, 65N12, 65N15

*Corresponding author

1. Introduction

The numerical stability of a convection–diffusion equation is, for the most part, due to the presence of the diffusion term. Then, when convection dominates diffusion, it is natural to expect that instabilities appear in the numerical solution. These instabilities result in the presence of large over- and undershoots, which are a sign of a violation of the discrete maximum principle (DMP). To correct the violation of the DMP, many methods have been proposed and analyzed over the years. The first attempt was to add enough numerical diffusion to make the problem diffusion-dominated, and then the DMP follows under appropriate assumptions (see, e.g. Ref. 23). This crude strategy leads to numerical results which are extremely diffusive, and then not usable in practice. This fact motivated the introduction of the so-called shock-capturing methods, which are characterized by adding an extra term to the discrete formulation. This extra term contains a viscosity coefficient which is solution-dependent, hence making the method nonlinear (see Ref. 21 for a review). Nonlinear discretizations are not necessarily guaranteed to preserve the DMP, and, to the best of our knowledge, the first one was the work of Ref. 31. Later approaches include Refs. 9, 11, 3, 4, 14 and 5.

All the above-mentioned references share two main hypotheses, namely, the need to use first-order polynomials, and certain assumptions on the mesh. More precisely, in the two-dimensional case the mesh is supposed to be a Delaunay one. This restriction can be tracked back to the first work concerning the validity of the DMP, even for a Laplace equation, i.e. the work of Ref. 13. Since then, several generalizations and attempts to overcome that restriction have been done. For example, in Ref. 10 an anisotropic Laplacian was added to the formulation, and the DMP can be proved for more general cases. More recently, in the context of hyperbolic equations, the works of Refs. 18 and 17 propose methods that can overcome this restriction, while at the same time providing approximations that converge to the entropy solution. It is important to remark that these last references' possible extension to the case in which diffusion is present in the equations does not seem to be an easy task.

One particular nonlinear discretization, designed to satisfy the DMP by construction, is the one known as algebraic flux correction (AFC) method. The origins of this method can be tracked back to Refs. 8 and 33, and it has enjoyed active development in the last decade thanks to the work of Kuzmin and co-workers (see Refs. 24–28, and Ref. 29 for a recent review). This class of methods, unlike previous discretizations, is not based on a variational formulation of the problem, but rather on a restatement of the resulting linear system in which the right-hand side is written as the sum of antidiffusive fluxes. This restatement shows that these fluxes are responsible for the violation of the DMP, and then AFC schemes limit them using solution-dependent limiters. Despite the fact of providing good numerical results (apart from the above-cited references, see also the review works of Refs. 22 and 1 for some further numerical results), until very recently, no mathematical analysis

had been carried out for the AFC schemes. The first works in this direction are, to the best of our knowledge, Refs. 6 and 7. Surprisingly, the proof of the DMP given in Ref. 7 also requires the use of a Delaunay mesh. Then, despite the fact that the geometry of the mesh does not enter explicitly in the definition of the AFC methods, some results on them still depend on the geometry of the mesh. This fact motivates the search for modifications of the limiters that generate methods satisfying the DMP on general meshes.

Another important property that is often required for numerical discretizations is the so-called linearity preservation. This property demands that the modification added to the formulation vanishes if the solution is a polynomial of degree 1 (at least locally). This restriction, which can be interpreted as a weak consistency requirement, is believed to lead to improved accuracy in regions where the solution is smooth. In fact, in previous works, linearity preservation was linked to good convergence properties for diffusion problems (see, e.g. Refs. 20 and 30). Even if this is a requirement that may seem natural, this condition was proposed in a very heuristic manner. As a matter of fact, in many works the proposed method has been claimed to be linearity preserving, but a proof of this fact is just hinted, or even lacking. In addition, although this property, so far, has not been proved mathematically to be a sufficient, or even a necessary, condition for good numerical behavior, it has been observed in different works (see, e.g. Ref. 12, and, especially, the introduction in Ref. 15 for a discussion) that linearity preservation improves the quality of the numerical solution on distorted meshes.

Based on the above considerations, our main objective in this work is to propose a definition of the limiters in an AFC method for a convection–diffusion–reaction equation that achieves two main goals: satisfaction of the DMP and linearity preservation, both on general simplicial meshes. To achieve this, we write down the main requirements to be satisfied by the limiters, and proceed to modify the algorithm proposed in Ref. 28 in such a way that these two properties are valid on general meshes. More precisely, the limiters from Ref. 28 are modified with factors that depend on the geometry of the elements that share a given node of the triangulation. Hence, this approach introduces explicit geometric information about the mesh into the algorithm.

Numerical studies will support the analytical results. In addition they show that the numerical solutions obtained with the new limiter possess further desirable properties compared with the solutions computed with the limiter from Ref. 25, which is considered to be a method of choice: it exhibits optimal convergence on distorted meshes in the diffusion-dominated regime and a sharper layer is obtained in a standard test problem for the convection-dominated case.

It is worth mentioning that methods of AFC type we have found in the literature do not satisfy the objectives of our paper in the required generality. For example, the techniques of Ref. 28, used as a basis for our method, are proved to be linearity preserving only on symmetric meshes as we discuss in Remark 6.3 below. The method recently presented in Ref. 5 has been proved to preserve the DMP

only for meshes that satisfy the condition of Xu and Zikatanov from Ref. 32, and this condition is sharp when the diffusion dominates. The linearity preservation of this method is again restricted to symmetric meshes. An alternative making the method linearity preserving for more general meshes requires solving an optimization problem for each interior node of the mesh, thus rendering the method more involved. Very recently, another monotone and linearity preserving method was proposed in Ref. 2 for conservation laws. However, it is not clear whether the DMP still holds when this method is applied to a convection–diffusion–reaction equation, which is our problem of interest. Moreover, the authors of Ref. 2 propose to use a regularization strategy to make the method twice differentiable and hence suitable for applying Newton's method but then the linearity preservation property is lost. Thus, to the best of our knowledge, the method presented in this paper is the first method that satisfies both the DMP and linearity preservation on general simplicial meshes, when the equation under consideration is a convection–diffusion–reaction equation. In particular, as a special result, a monotone and linearity preserving discretization of the Poisson equation on general simplicial meshes is obtained.

The rest of the paper is organized as follows. In Sec. 2, AFC schemes are presented in their most general form. Then, the minimal requirements on the limiter in order to satisfy the DMP are laid down in Sec. 3. Our concrete proposal for the limiter is given in Sec. 4. Section 5 is devoted to the application of the AFC scheme to the convection–diffusion–reaction equation and its analysis. The final ingredient in the definition of the limiter, namely, the computation of the multiplicative factor introduced in order to make the method linearity preserving, is presented in Sec. 6. Finally, some numerical results supporting our claims are given in Sec. 7.

2. An Algebraic Flux Correction Scheme

Consider a linear boundary value problem for which the maximum principle holds. Let us discretize this problem by the finite element method. Then, the discrete solution can be represented by a vector $\mathrm{U} \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of U ($0 < M < N$) correspond to nodes where Dirichlet boundary conditions are prescribed whereas the first M components of U are computed using the finite element discretization of the underlying partial differential equation. Then $U \equiv (u_1, \ldots, u_N)$ satisfies a system of linear equations of the form:

$$\sum_{j=1}^{N} a_{ij} u_j = g_i, \quad i = 1, \ldots, M, \tag{2.1}$$

$$u_i = u_i^b, \quad i = M+1, \ldots, N. \tag{2.2}$$

We assume that the matrix $(a_{ij})_{i,j=1}^{M}$ is positive definite, i.e.

$$\sum_{i,j=1}^{M} u_i a_{ij} u_j > 0 \quad \forall (u_1, \ldots, u_M) \in \mathbb{R}^M \backslash \{0\}. \tag{2.3}$$

To introduce an algebraic flux correction scheme, we first extend the matrix of (2.1) to a matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$. For example, one can simply use the finite element matrix corresponding to the above-mentioned finite element discretization in the case when homogeneous natural boundary conditions are used instead of the Dirichlet ones. We shall consider this matrix with the following modification:

$$a_{ji} := 0 \quad \text{if } a_{ij} < 0, \quad i = 1, \ldots, M, \;\; j = M+1, \ldots, N. \tag{2.4}$$

This reduces the amount of artificial diffusion introduced by the matrix $\mathbb{D}$ defined next.

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

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \quad \forall\, i \neq j, \quad d_{ii} = -\sum_{j \neq i} d_{ij}. \tag{2.5}$$

This definition guarantees that the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ has positive diagonal entries and nonpositive off-diagonal entries. If, in addition,

$$\sum_{j=1}^N a_{ij} \geq 0, \quad i = 1, \ldots, M, \tag{2.6}$$

then the matrix $\tilde{\mathbb{A}}$ satisfies sufficient conditions to preserve the discrete maximum principle. Note that the property (2.6) is usually satisfied by finite element discretizations of elliptic equations arising in applications.

Going back to the solution of (2.1), this system is equivalent to

$$(\tilde{\mathbb{A}}\mathrm{U})_i = g_i + (\mathbb{D}\mathrm{U})_i, \quad i = 1, \ldots, M. \tag{2.7}$$

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

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

where $f_{ij} = d_{ij}(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. The idea of the algebraic flux correction scheme is to limit those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (2.1) (or, equivalently, (2.7)) is replaced by

$$(\tilde{\mathbb{A}}\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad i = 1, \ldots, M, \tag{2.8}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (2.1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem. So far, these coefficients have been chosen in various ways, and their definition is always based on the above fluxes f_{ij}, see Refs. 24–28 for examples. To guarantee that the resulting scheme is conservative, and to be able to show existence of solutions, one should

require that the coefficients α_{ij} are symmetric, i.e.

$$\alpha_{ij} = \alpha_{ji}, \quad i,j = 1,\ldots,M. \tag{2.9}$$

Rewriting Eq. (2.8) using the definition of the matrix $\tilde{\mathbb{A}}$, one obtains the following expression for the algebraic flux correction scheme:

$$\sum_{j=1}^{N} a_{ij}u_j + \sum_{j=1}^{N}(1-\alpha_{ij})d_{ij}(u_j - u_i) = g_i, \quad i = 1,\ldots,M, \tag{2.10}$$

$$u_i = u_i^b, \quad i = M+1,\ldots,N, \tag{2.11}$$

where $\alpha_{ij} = \alpha_{ij}(u_1,\ldots,u_N) \in [0,1]$, $i = 1,\ldots,M$, $j = 1,\ldots,N$, satisfy (2.9).

The following theorem states sufficient conditions on the limiters α_{ij} assuring the solvability of the nonlinear discrete problem (2.10), (2.11). Our proposal for such limiters will be given in Sec. 4.

Theorem 2.1. *Let* (2.3) *hold. For any* $i \in \{1,\ldots,M\}$, $j \in \{1,\ldots,N\}$, *let* $\alpha_{ij} : \mathbb{R}^N \to [0,1]$ *be such that* $\alpha_{ij}(u_1,\ldots,u_N)(u_j - u_i)$ *is a continuous function of* $u_1,\ldots,u_N$. *Finally, let the functions* α_{ij} *satisfy* (2.9). *Then there exists a solution of the nonlinear problem* (2.10), (2.11).

Proof. See Theorem 3.3 in Ref. 7. □

It is worth mentioning that the symmetry property (2.9) is necessary for the validity of Theorem 2.1, see Ref. 6.

3. The Discrete Maximum Principle

As it was mentioned in the Introduction, the main motivation of AFC schemes is to respect the DMP. In this section, we state some minimal assumptions on the limiters α_{ij} in order to satisfy this property.

Given $i \in \{1,\ldots,M\}$, the discrete maximum principle will be formulated locally, with respect to an index set $S_i \subset \{1,\ldots,N\}$. We assume that

$$S_i \supset \{j \in \{1,\ldots,N\}\backslash\{i\} : a_{ij} \neq 0 \text{ or } a_{ji} > 0\}, \quad i = 1,\ldots,M. \tag{3.1}$$

The proof of the discrete maximum principle requires only that $\{\alpha_{ij}d_{ij}\}_{j\in S_i}$ vanish if u_i is a strict local extremum. More precisely, we assume that, for any $i \in \{1,\ldots,M\}$ and any $U = (u_1,\ldots,u_N) \in \mathbb{R}^N$, the limiters α_{ij} satisfy

$$u_i > u_j \quad \forall\, j \in S_i \quad \text{or} \quad u_i < u_j \quad \forall\, j \in S_i \;\Rightarrow\; \alpha_{ij}(U)d_{ij} = 0 \quad \forall\, j \in S_i. \tag{3.2}$$

The matrix $\mathbb{A}$ will be supposed to satisfy (2.6). Then the only assumption on $\mathbb{A}$ for proving the local discrete maximum principle at $i \in \{1,\ldots,M\}$ will be that

$$\text{there exists } j \in \{1,\ldots,N\},\ j \neq i : \quad a_{ij} < 0 \quad \text{or} \quad a_{ij} < a_{ji}. \tag{3.3}$$

Note that the diagonal entry a_{ii} can be arbitrary. The condition (3.3) is typically satisfied, in particular, by the matrix associated to a finite element discretization

of the convection–diffusion equation (see Lemma 5.1 and Remark 5.2 below for details). If (3.3) does not hold but

$$A_i := \sum_{j=1}^{N} a_{ij} > 0, \tag{3.4}$$

then still a slightly weaker statement on the DMP can be proved. If $A_i = 0$ and $a_{ii} > 0$ (as implied by (2.3)), then (3.3) is always satisfied.

With the above hypotheses, we prove the main result of this section.

Theorem 3.1. *Let the matrix* $\mathbb{A}$ *satisfy* (2.6) *and let the limiters* α_{ij} *satisfy* (3.2). *Let* $(u_1, \ldots, u_N) \in \mathbb{R}^N$ *satisfy* (2.10). *Consider any* $i \in \{1, \ldots, M\}$. *If* (3.3) *holds, one has*:

$$g_i \le 0 \Rightarrow \left(\text{if } u_i \ge 0, \text{then } u_i \le \max_{j \in S_i} u_j \right), \tag{3.5}$$

$$g_i \ge 0 \Rightarrow \left(\text{if } u_i \le 0, \text{then } u_i \ge \min_{j \in S_i} u_j \right). \tag{3.6}$$

If $A_i > 0$, *one has*:

$$g_i \le 0 \Rightarrow \left(\text{if } u_i > 0, \text{then } u_i \le \max_{j \in S_i} u_j \right), \tag{3.7}$$

$$g_i \ge 0 \Rightarrow \left(\text{if } u_i < 0, \text{then } u_i \ge \min_{j \in S_i} u_j \right). \tag{3.8}$$

Consequently, if (3.3) *holds or* $A_i > 0$, *one has*:

$$g_i \le 0 \Rightarrow u_i \le \max_{j \in S_i} u_j^+, \tag{3.9}$$

$$g_i \ge 0 \Rightarrow u_i \ge \min_{j \in S_i} u_j^-, \tag{3.10}$$

where $u_j^+ := \max\{0, u_j\}$ *and* $u_j^- := \min\{0, u_j\}$.

Proof. Since $d_{ij} = 0$ for any $i \in \{1, \ldots, M\}$ and $j \notin S_i \cup \{i\}$, Eq. (2.10) can be written in the form

$$A_i u_i + \sum_{j \in S_i} [a_{ij} + (1 - \alpha_{ij}(U)) d_{ij}](u_j - u_i) = g_i, \quad i = 1, \ldots, M. \tag{3.11}$$

Consider any $i \in \{1, \ldots, M\}$ and let $g_i \le 0$ and $u_i \ge 0$. Let us assume that $u_i > u_j$ for all $j \in S_i$. Then (3.11) and (3.2) imply that

$$A_i u_i + \sum_{j \in S_i} (a_{ij} + d_{ij})(u_j - u_i) = g_i. \tag{3.12}$$

Due to the definition of d_{ij} (cf. (2.5)), one has $a_{ij} + d_{ij} \le 0$ for $j \ne i$. Moreover, if (3.3) holds, there is a $j \in S_i$ such that $a_{ij} + d_{ij} < 0$. Hence the left-hand side of (3.12) is strictly positive, which is a contradiction. If $A_i > 0$ and $u_i > 0$,

then (3.12) implies that $g_i \geq A_i u_i > 0$. This is, again, a contradiction. Therefore, there is a $j \in S_i$ such that $u_i \leq u_j$, which proves (3.5) and (3.7). The statements (3.6) and (3.8) follow in an analogous way. Finally, (3.9) and (3.10) are immediate consequences of the preceding statements. □

Assuming equality instead of inequality in (2.6), the following stronger result can be proved.

Theorem 3.2. *Let the limiters α_{ij} satisfy* (3.2) *and let $(u_1, \ldots, u_N) \in \mathbb{R}^N$ satisfy* (2.10). *Consider any $i \in \{1, \ldots, M\}$. If $A_i = 0$ and* (3.3) *holds, then one has*:

$$g_i \leq 0 \Rightarrow u_i \leq \max_{j \in S_i} u_j,$$

$$g_i \geq 0 \Rightarrow u_i \geq \min_{j \in S_i} u_j.$$

Proof. The proof from the previous result can be applied with the minor difference that, since $A_i = 0$, the restriction on the sign of u_i is not needed. □

4. Definition of α_{ij}

The last section imposed minimal conditions that the limiter α_{ij} used in (2.10) should satisfy in order to guarantee the discrete maximum principle. In this section, we design a limiter that fulfills those hypotheses. Additionally, we are interested in proposing a limiter that makes the method linearity preserving on general simplicial meshes. Our proposal is related to the one from Ref. 28 which is, however, not proved to be linearity preserving on general meshes, see Remark 6.3. The main difference between our proposal and the one from Ref. 28 is the definition of the constant γ_i below, which will be later derived to impose linearity preservation on general simplicial meshes. We shall show that it provides limiters that guarantee the solvability of (2.10), (2.11), and the validity of the discrete maximum principle.

First, for any $i \in \{1, \ldots, M\}$, we set

$$u_i^{\max} := \max_{j \in S_i \cup \{i\}} u_j, \quad u_i^{\min} := \min_{j \in S_i \cup \{i\}} u_j, \quad q_i := \gamma_i \sum_{j \in S_i} d_{ij}, \tag{4.1}$$

where S_i is an index set satisfying (3.1) and $\gamma_i > 0$ is a fixed constant, whose value will be defined later (see (6.5) in Theorem 6.1). Furthermore, for any $i \in \{1, \ldots, M\}$, we set

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

and we define

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

If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Finally, we set

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

and define

$$\begin{aligned} \alpha_{ij} &:= \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}, \quad && i, j = 1, \dots, M, \\ \alpha_{ij} &:= \widetilde{\alpha}_{ij}, && i = 1, \dots, M, \;\; j = M+1, \dots, N. \end{aligned}$$

The symmetry condition (2.9) is guaranteed by the last step of this algorithm.

The following result shows that the above limiter satisfies (3.2). Then, the resulting method respects the discrete maximum principle, independently of the geometry of the mesh, provided $\mathbb{A}$ satisfies (2.6) and at least one of the conditions (3.3) and (3.4) for any $i \in \{1, \dots, M\}$.

Lemma 4.1. *The limiter α_{ij} defined in this section satisfies* (3.2).

Proof. Consider any $i \in \{1, \dots, M\}$ and $U = (u_1, \dots, u_N) \in \mathbb{R}^N$ such that $u_i > u_j$ for all $j \in S_i$. Then, $u_i^{\max} = u_i$ and hence $Q_i^+ = 0$. Choose any $j \in S_i$ and let us show that $\alpha_{ij}(U) d_{ij} = 0$. It suffices to consider $d_{ij} \neq 0$. But then $f_{ij} > 0$ and hence $P_i^+ > 0$, leading to $R_i^+ = 0$. Consequently $\widetilde{\alpha}_{ij}(U) = 0$, thus giving $\alpha_{ij}(U) = 0$. If $u_i < u_j$ for all $j \in S_i$, then the proof is analogous. □

In addition to the last lemma, the following result states that the limiter α_{ij} satisfies the continuity conditions from Theorem 2.1, and hence problem (2.10), (2.11) has a solution. Its proof is very similar to Lemma 4.1 in Ref. 7, and then we give an abridged form of it for completeness.

Lemma 4.2. *The coefficients α_{ij} are such that $\phi_{ij}(U) := \alpha_{ij}(u_1, \dots, u_N)(u_j - u_i)$ are continuous functions of $u_1, \dots, u_N$ on $\mathbb{R}^N$.*

Proof. Consider any $i \in \{1, \dots, M\}$, $j \in \{1, \dots, N\}$. Let us first investigate the continuity of $\widetilde{\alpha}_{ij}$. It suffices to consider the case $\widetilde{\alpha}_{ij} \not\equiv 1$ (and hence $d_{ij} \neq 0$ and $j \in S_i$). Let $U = \{u_i\}_{i=1}^N \in \mathbb{R}^N$. We first consider $u_i > u_j$. Then, $f_{ij} > 0$ and one obtains

$$\widetilde{\alpha}_{ij}(U) = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{|f_{ij}| + \widetilde{P}_i^+} \quad \text{with} \quad \widetilde{P}_i^+ = \sum_{k \in S_i \setminus \{j\}} f_{ik}^+.$$

Since $u_i > u_j$, there is a neighborhood of U where the denominator of the above expression does not vanish, and then the function $\widetilde{\alpha}_{ij}$ is continuous in U. Now, if $u_j > u_i$, by the same arguments one can deduce that $\widetilde{\alpha}_{ij}$ is continuous in U. Thus,

if $u_i \neq u_j$, then $\widetilde{\alpha}_{ij}$, and therefore ϕ_{ij}, is continuous in U. Finally, if $u_i = u_j$, then $\phi_{ij}(U) = 0$. Let $V = \{v_i\}_{i=1}^N \in \mathbb{R}^N$. Then, since $\alpha_{ij}(U) \in [0,1]$, one obtains

$$\begin{aligned}&|\phi_{ij}(V) - \phi_{ij}(U)| \\ &\quad = |\phi_{ij}(V)| = |\alpha_{ij}(V)||v_j - v_i| \leq |v_j - u_j - (v_i - u_i)| \leq \sqrt{2}\|V - U\|_{\mathbb{R}^N}.\end{aligned}$$

Then, $\phi_{ij}(V) \to \phi_{ij}(U)$ if $V \to U$ and ϕ_{ij} is continuous in U. This finishes the proof. □

We finish this section by making some comments on the choice of the factors γ_i used in (4.1). First, the proof of the discrete maximum principle is independent of their values, and then, it can be applied for choices other than the one introduced in this paper, e.g. the ones from Ref. 28. Once this is said, the actual value of γ_i has two main impacts on the performance of the AFC scheme. First, if chosen appropriately (as it will be done in Sec. 6 below), then it can be proved that the resulting scheme is linearity preserving on general simplicial meshes. Second, it influences the amount of artificial diffusion added by the AFC term to the original system (2.1). If γ_i's are increased, then more limiters α_{ij} will be equal to 1 and hence less artificial diffusion will be added. If γ_i's are decreased, then more limiters α_{ij} will be smaller than 1 and hence more artificial diffusion will be added. Thus, to reduce smearing of approximate solutions represented by the values $u_1, \ldots, u_N$, large values of γ_i's are convenient. The downside of this is that, for large values of γ_i's, the limiters $\alpha_{ij}(u_1, \ldots, u_N)$ change very rapidly near local extrema in u_i and hence the numerical solution of the nonlinear algebraic problem becomes more involved.

5. The AFC Scheme for Convection–Diffusion–Reaction Equations

Let $\Omega \subset \mathbb{R}^d$, $d = 2, 3$, be a bounded polyhedral domain with Lipschitz boundary. Let us consider the steady-state convection–diffusion–reaction equation

$$-\varepsilon\Delta u + \boldsymbol{b} \cdot \nabla u + cu = g \quad \text{in } \Omega, \quad u = u_b \quad \text{on } \partial\Omega, \tag{5.1}$$

where $\varepsilon \in (0, \varepsilon_0)$ with $\varepsilon_0 < +\infty$ is a constant, and $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $c \in L^\infty(\Omega)$, $g \in L^2(\Omega)$, and $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap C(\partial\Omega)$ are given functions satisfying

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

where σ_0 is a constant. The weak solution of (5.1) is a function $u \in H^1(\Omega)$ such that $u = u_b$ on $\partial\Omega$ and

$$a(u, v) = (g, v) \quad \forall\, v \in H_0^1(\Omega), \tag{5.2}$$

with

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

Here we adopt the usual notation for Sobolev spaces. In particular, $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. Since $c \geq \sigma_0$ in Ω and $\boldsymbol{b}$ is solenoidal, then

$$a(v,v) \geq \|v\|_a^2 \quad \forall\, v \in H_0^1(\Omega), \tag{5.3}$$

with

$$\|v\|_a^2 = \varepsilon |v|_{1,\Omega}^2 + \sigma_0 \|v\|_{0,\Omega}^2.$$

It is well known that the weak solution of (5.1) exists, is unique, and satisfies the maximum principle (cf. Ref. 16).

Let $\mathscr{T}_h$ belong to a regular family of triangulations of $\overline{\Omega}$ consisting of simplices. We introduce the finite element spaces:

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

consisting of continuous piecewise linear functions. From now on, we denote by $x_1, \ldots, x_N$ the vertices of the triangulation $\mathscr{T}_h$ and assume that $x_1, \ldots, x_M \in \Omega$ and $x_{M+1}, \ldots, x_N \in \partial\Omega$. Furthermore, we denote by $\varphi_1, \ldots, \varphi_N$ the usual basis functions of W_h, i.e. we assume that $\varphi_i(x_j) = \delta_{ij}$, $i,j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol. Then the functions $\varphi_1, \ldots, \varphi_M$ form a basis in V_h.

Now, an approximate solution of the variational problem (5.2) can be introduced as the solution of the following finite-dimensional problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a(u_h, v_h) = (g, v_h) \quad \forall\, v_h \in V_h. \tag{5.4}$$

We denote

$$a_{ij} = a(\varphi_j, \varphi_i), \quad i,j = 1, \ldots, N, \tag{5.5}$$

$$g_i = (g, \varphi_i), \qquad i = 1, \ldots, M, \tag{5.6}$$

$$u_i^b = u_b(x_i), \qquad i = M+1, \ldots, N. \tag{5.7}$$

Then u_h solves (5.4) if and only if its coefficient vector with respect to the basis of W_h satisfies the relations (2.1) and (2.2). The bilinear form a defines the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$ whose entries are given by (5.5) and (2.4). Finally, thanks to (5.3) the matrix $(a_{ij})_{i,j=1}^M$ satisfies (2.3), and it follows that the problem (5.4) has a unique solution.

The artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is defined using (2.5). We introduce the nonlinear form

$$d_h(w; z, v) := \sum_{i,j=1}^N (1 - \alpha_{ij}(w)) d_{ij} (z(x_j) - z(x_i)) v(x_i) \quad \forall\, w, z, v \in C(\overline{\Omega}),$$

with $\alpha_{ij}(w) := \alpha_{ij}(\{w(x_k)\}_{k=1}^N)$. Then the corresponding flux correction scheme (2.10), (2.11) can be rewritten as the following variational problem:

Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \ldots, N$, and

$$a(u_h, v_h) + d_h(u_h; u_h, v_h) = (g, v_h) \quad \forall\, v_h \in V_h. \tag{5.8}$$

Since the limiters α_{ij} defined in the last section satisfy the assumptions of Theorem 2.1, and the bilinear form a is elliptic, then the problem (5.8) has a solution. A natural (solution-dependent) norm on V_h corresponding to the left-hand side of (5.8) is defined by

$$\|v_h\|_h := (\|v_h\|_a^2 + d_h(u_h; v_h, v_h))^{1/2}, \quad v_h \in V_h.$$

Assuming that $u \in H^2(\Omega)$ and following completely analogous steps as the ones from Sec. 7 in Ref. 7 it follows that, if $\sigma_0 > 0$, the following error bound holds

$$\|u - u_h\|_h \leq Ch\|u\|_{2,\Omega} + (d_h(u_h; i_h u, i_h u))^{1/2}, \tag{5.9}$$

where $C > 0$ is independent of u, h and ε, and $i_h u$ stands for the Lagrange interpolant of u. For the last term in (5.9), using the proof of Lemma 7.3 from Ref. 7, it follows that

$$d_h(w_h; i_h u, i_h u) \leq C \max_{i,j=1,\ldots,N} (|d_{ij}||x_i - x_j|^{2-d}) |i_h u|_{1,\Omega}^2 \quad \forall\, w_h \in W_h,\ u \in C(\overline{\Omega}), \tag{5.10}$$

where C is independent of h and the data of problem (5.1). This result shows that the error $\|u - u_h\|_h$ will tend to zero as long as the product $|d_{ij}|\,|x_i - x_j|^{2-d}$ tends to zero. This implies that the method will converge as long as the matrix $\mathbb{A}$ tends to be an M-matrix, and this speed of convergence is fast enough to compensate for the negative power of h arising from $|x_i - x_j|^{2-d}$ in the three-dimensional case. Hence, it is natural to expect that the convergence properties of the method will vary according to the geometry of the mesh. In particular, for the convection-dominated regime, an $O(h^{1/2})$ estimate of $\|u - u_h\|_h$ can be shown irrespectively of the geometry of the mesh. On the contrary, for the diffusion-dominated regime, the convergence rates will vary dramatically depending on the geometrical properties of the mesh (see Ref. 7 for details). This was illustrated numerically in Ref. 7 for the limiter defined in Ref. 25. In some particular cases a better than expected convergence was observed, but the theoretical justification of this fact, which requires a more refined estimation of $d_h(u_h; i_h u, i_h u)$ for particular limiters, does not seem to be an easy task, and it will be the subject of our future research.

The above results are valid for any limiters α_{ij} satisfying the assumptions of Sec. 2 (respectively of Theorem 2.1) and hence, in particular, for the limiter from Sec. 4. To apply this limiter, we have to specify the sets S_i satisfying (3.1). The simplest possibility is to use

$$S_i = \{j \in \{1, \ldots, N\} \backslash \{i\} : x_i \text{ and } x_j \text{ are end points of the same edge}\}, \tag{5.11}$$

where $i = 1, \ldots, M$. This definition of S_i was used in the computations reported in Sec. 7. To finish the definition of α_{ij}, we have to define the factors γ_i used in (4.1). This will be done in the following section.

Remark 5.1. Usually, results on the discrete maximum principle like in Theorems 3.1 and 3.2 are proved for Delaunay meshes with respect to sets $S_i = \{j \in \{1, \ldots, N\} \backslash \{i\} : a_{ij} \neq 0\}$. For $c = 0$, this definition and the set used in (3.1) coincide for Delaunay meshes. Indeed, for such a mesh, the validity of $a_{ji} > 0$ in (3.1) implies that $a_{ij} \neq 0$ since $a_{ij} + a_{ji} = 2\varepsilon(\nabla\varphi_i, \nabla\varphi_j) \leq 0$. Whenever $c > 0$, then the two definitions no longer coincide, the set induced by (3.1) can be larger, and hence the final result is slightly weaker. The stronger assumption (3.1) is made in order to guarantee our results to be valid on arbitrary meshes.

We close this section by showing that the matrix $\mathbb{A}$ defined above satisfies the assumptions made on it to prove the discrete maximum principle.

Lemma 5.1. *The matrix* $\mathbb{A}$ *defined in* (5.5) *and* (2.4) *satisfies the assumption* (2.6). *Moreover, for any* $i \in \{1, \ldots, M\}$, *the assumption* (3.3) *holds if* $A_i = 0$ *or*

$$\text{there exists } j \in \{1, \ldots, N\} : (\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) \neq 0. \tag{5.12}$$

Proof. The validity of (2.6) follows immediately from the property $\sum_{j=1}^{N} \varphi_j = 1$ and the non-negativity of c. Consider any $i \in \{1, \ldots, M\}$. If $A_i = 0$, then there is $j \in \{1, \ldots, N\}$, $j \neq i$, with $a_{ij} < 0$ since $a_{ii} \geq \varepsilon|\varphi_i|^2_{1,\Omega} > 0$. Hence (3.3) holds. Let us assume (5.12) and let (3.3) does not hold, i.e.

$$a_{ij} \geq 0 \quad \text{and} \quad a_{ij} \geq a_{ji} \quad \forall\, j \in \{1, \ldots, N\}, \quad j \neq i. \tag{5.13}$$

Under this assumption, then the modification (2.4) is not used for the matrix entries in (5.13), and the original matrix remains unchanged. Hence, in view of the second inequality in (5.13), one has

$$(\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) \geq (\boldsymbol{b} \cdot \nabla\varphi_i, \varphi_j) = -(\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) \quad \forall\, j \in \{1, \ldots, N\}, \;\; j \neq i,$$

so that

$$(\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) \geq 0 \quad \forall\, j \in \{1, \ldots, N\}, \quad j \neq i.$$

Since $(\boldsymbol{b} \cdot \nabla\varphi_i, \varphi_i) = 0$ and $\sum_{j=1}^{N} (\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) = 0$, one deduces that

$$(\boldsymbol{b} \cdot \nabla\varphi_j, \varphi_i) = 0 \quad \forall\, j \in \{1, \ldots, N\},$$

which is in contradiction with (5.12). □

Remark 5.2. According to the previous lemma, the validity of (3.3) is not guaranteed if the convection term does not contribute to the ith row of the matrix $\mathbb{A}$. Although this cannot be excluded, it is a rather exceptional situation and hence (3.3) will typically hold if $\boldsymbol{b}$ does not vanish identically in supp φ_i. Lemma 5.1 also shows that (3.3) holds if $c \equiv 0$ since then $A_i = 0$ for any $i \in \{1, \ldots, M\}$. Thus, if the

reaction term for $c > 0$ is discretized using a lumping like in Ref. 7, the off-diagonal entries of $\mathbb{A}$ are the same as for $c \equiv 0$ and hence (3.3) again holds although $A_i > 0$.

6. Linearity Preservation

Let us consider the limiter from Sec. 4 with the sets S_i defined in (5.11). In this section, we finish the definition of this limiter by specifying the parameters γ_i that make it possible to prove that the resulting scheme is linearity preserving on general simplicial meshes. We recall that $x_1, \ldots, x_N$ stand for the vertices of $\mathscr{T}_h$, and that $x_1, \ldots, x_M \in \Omega$. We shall show that the factors γ_i in (4.1) can be defined in such a way that

$$\widetilde{\alpha}_{ij}(u) = 1 \quad \forall\, u \in \mathbb{P}_1(\mathbb{R}^d), \quad i = 1, \ldots, M, \quad j = 1, \ldots, N. \tag{6.1}$$

Then the AFC scheme (2.10), (2.11) will be linearity preserving. Let us consider any function $u \in \mathbb{P}_1(\mathbb{R}^d)$ and set $u_i = u(x_i)$, $i = 1, \ldots, N$. Then, if one wants to satisfy (6.1), one needs

$$Q_i^+ \geq P_i^+ \quad \text{if } f_{ij} > 0, \quad Q_i^- \leq P_i^- \quad \text{if } f_{ij} < 0. \tag{6.2}$$

Sufficient conditions for (6.2) are the inequalities

$$u_i - u_i^{\min} \leq \gamma_i (u_i^{\max} - u_i), \quad u_i^{\max} - u_i \leq \gamma_i (u_i - u_i^{\min}). \tag{6.3}$$

Note that it suffices to find γ_i such that

$$u_i - u_i^{\min} \leq \gamma_i (u_i^{\max} - u_i) \quad \forall\, u \in \mathbb{P}_1(\mathbb{R}^d), \tag{6.4}$$

since then the second inequality in (6.3) follows from (6.4) by changing the sign of u. Thus, the validity of (6.4) assures that the AFC scheme (2.10), (2.11) based on the limiter from Sec. 4 is linearity preserving.

To discuss the validity of (6.4), it is convenient to introduce the patch $\Delta_i = \operatorname{supp} \varphi_i$ for any interior vertex x_i of the triangulation $\mathscr{T}_h$. Thus, Δ_i is a patch consisting of simplices $T \in \mathscr{T}_h$ sharing the vertex x_i, see Fig. 1. Then the sets S_i defined in (5.11) satisfy

$$S_i = \{j \in \{1, \ldots, N\} : x_j \in \partial \Delta_i\},$$

and one has

$$u_i^{\min} = \min_{\Delta_i} u, \quad u_i^{\max} = \max_{\Delta_i} u.$$

Note that, for $u \in \mathbb{P}_1(\mathbb{R}^d)$, $u_i^{\min}$ and $u_i^{\max}$ are attained at vertices lying on $\partial \Delta_i$.

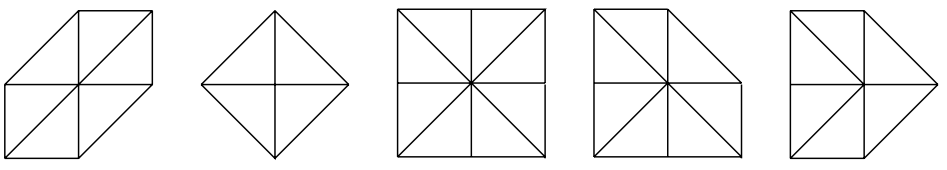

Fig. 1. Examples of patches Δ_i for $d = 2$.

If the patch Δ_i is symmetric with respect to the vertex x_i (like the first three patches from the left in Fig. 1), then the inequality (6.4) holds with $\gamma_i = 1$ as the following lemma shows.

Lemma 6.1. *Let Δ_i be symmetric with respect to x_i. Then*

$$u_i - u_i^{\min} = u_i^{\max} - u_i \quad \forall\, u \in \mathbb{P}_1(\mathbb{R}^d).$$

Proof. Let us assume that $u_i - u_i^{\min} < u_i^{\max} - u_i$. There exists a vertex $x_j \in \partial\Delta_i$ such that $u_i^{\max} = u_j$. Furthermore, due to the symmetry of Δ_i, there is a vertex $x_k \in \partial\Delta_i$ such that $(x_j + x_k)/2 = x_i$. Then $u_j + u_k = 2\,u_i$ and hence

$$u_i - u_i^{\min} < u_i^{\max} - u_i = u_j - u_i = u_i - u_k.$$

Consequently, $u_k < u_i^{\min}$, which is a contradiction. Analogously, it can be shown that $u_i - u_i^{\min} > u_i^{\max} - u_i$ leads to a contradiction. □

For general patches Δ_i, a possible factor γ_i is computed in the following theorem.

Theorem 6.1. *Let $x_1, \ldots, x_M \in \Omega$. For any $i \in \{1, \ldots, M\}$, let Δ_i be the above-defined patch corresponding to the vertex x_i and let Δ_i^{conv} be its convex hull. Let*

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}})}, \quad i = 1, \ldots, M. \tag{6.5}$$

Then the inequalities (6.4) *hold and hence the AFC scheme* (2.10), (2.11) *with the limiter from Sec.* 4 *is linearity preserving.*

Proof. For simplicity, we shall present the proof for $d = 2$. For $d = 3$ one can proceed analogously. Consider a patch Δ_i and let $u \in \mathbb{P}_1(\mathbb{R}^2)$ be any nonconstant linear function. Let p be the line in the direction of ∇u containing the vertex x_i. Then there are uniquely determined points $A, B \in p$ such that $u(A) = u_i^{\min}$, $u(B) = u_i^{\max}$. Let q_A and q_B be lines orthogonal to p intersecting the line p at the points A and B, respectively, see Fig. 2. Since u is constant along lines perpendicular to p, the patch Δ_i is contained in the strip between the lines q_A and q_B. Consequently, each of these lines intersects Δ_i only at points on $\partial\Delta_i$ comprising at least one vertex. Moreover, any such vertex lies on the boundary of the convex hull Δ_i^{conv}. To find a constant γ_i for which the inequality (6.4) holds, we have to estimate the ratio

$$\frac{u_i - u_i^{\min}}{u_i^{\max} - u_i} = \frac{u(x_i) - u(A)}{u(B) - u(x_i)} = \frac{|x_i - A|}{|B - x_i|}.$$

Since q_A contains a vertex x_k lying on $\partial\Delta_i^{\mathrm{conv}}$, one has

$$|x_i - A| \le |x_i - x_k| \le \max_{x_j \in \partial\Delta_i^{\mathrm{conv}}} |x_i - x_j| = \max_{x_j \in \partial\Delta_i} |x_i - x_j|.$$

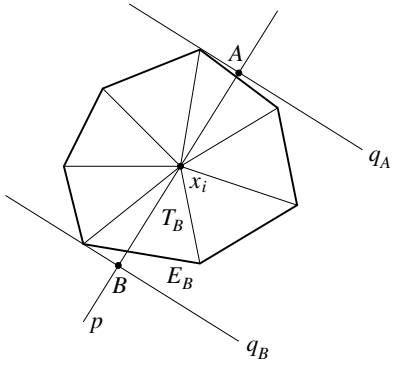

Fig. 2. Patch Δ_i with notation from the proof of Theorem 6.1.

On the other hand, if T_B is a triangle whose vertices are x_i and two consecutive vertices on $\partial\Delta_i^{\mathrm{conv}}$ such that the half-line x_iB intersects T_B (see Fig. 2), then

$$|B - x_i| \geq \operatorname{dist}(x_i, E_B),$$

where E_B is the edge of T_B opposite x_i. Consequently,

$$|B - x_i| \geq \operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}}),$$

which gives (6.5). □

Remark 6.1. For the patches in Fig. 1, the formula (6.5) gives the values 2, $\sqrt{2}$, $\sqrt{2}$, 2 and 2, respectively (from the left to the right). Since the first three patches from the left are symmetric, Lemma 6.1 shows that the formula (6.5) is not optimal in general. The last two patches in Fig. 1 are nonsymmetric and, for the linear function $u(x, y) = x + y$, one obtains $u_i - u_i^{\min} = 2(u_i^{\max} - u_i)$. Thus, for these two patches, the formula (6.5) gives the optimal values.

This possible lack of optimality arises from the fact that we have used the worst case scenario, that is, when the extrema of the function u are attained at the vertices closest to, and furthest away from, x_i, to derive the formula (6.5). This reasoning about the worst case scenario is adapted to three-space dimensions in a straightforward way.

Remark 6.2. Let us briefly mention the computation of the denominator in (6.5). First, any vertex $x_j \in \partial\Delta_i$ is shifted in the direction of the edge x_ix_j on the boundary of the convex hull Δ_i^{conv}. Then one goes through all simplices T forming Δ_i^{conv} and, denoting by E the edge (or face) of T opposite x_i, one computes $\operatorname{dist}(x_i, E)$. This is particularly easy in the two-dimensional case: if T possesses an obtuse angle at an end point of E, say P, then $\operatorname{dist}(x_i, E) = |x_i - P|$. If both angles of T at the end points of E are nonobtuse, then $\operatorname{dist}(x_i, E) = 2\,|T|/|E|$. In the three-dimensional case, the computation of $\operatorname{dist}(x_i, E)$ is more involved. Nevertheless, one can replace

it by $3\,|T|/|E| \le \operatorname{dist}(x_i, E)$ (and possibly increase the value of γ_i). Another possibility is to replace $\operatorname{dist}(x_i, \partial\Delta_i^{\mathrm{conv}})$ by the smallest diameter of inscribed balls of simplices forming Δ_i^{conv}.

Remark 6.3. As already mentioned, the limiter proposed in this paper is related to a method presented in Ref. 28. Although the methods of Ref. 28 are claimed to be linearity preserving, it turns out that the respective proofs are not valid for general meshes. The reason is that they rely on the validity of the inequality

$$u_i - u_j \le \gamma_{ij}(u_i^{\max} - u_i), \tag{6.6}$$

for any $u \in \mathbb{P}_1(\mathbb{R}^d)$ and $j \in S_i$ (with S_i defined in (5.11)), where

$$\gamma_{ij} = \frac{2}{m_i} \sum_{k \ne i} |\boldsymbol{c}_{ik} \cdot (x_i - x_j)|, \quad m_i = \int_\Omega \varphi_i \,\mathrm{d}x\,, \quad \boldsymbol{c}_{ik} = \int_\Omega \varphi_i \,\nabla\varphi_k \,\mathrm{d}x.$$

To prove (6.6), one uses the fact that $m_i \nabla u = \sum_k \boldsymbol{c}_{ik} u_k = \sum_k \boldsymbol{c}_{ik}(u_k - u_i)$ and $u_i - u_j = \nabla u \cdot (x_i - x_j)$, which lead to

$$u_i - u_j = \frac{1}{m_i} \sum_{k \ne i} \boldsymbol{c}_{ik} \cdot (x_i - x_j)(u_k - u_i). \tag{6.7}$$

If the patch Δ_i is symmetric with respect to x_i, then $|u_k - u_i| \le u_i^{\max} - u_i$ for any $k \in S_i$ due to Lemma 6.1 and hence (6.7) implies (6.6). On the other hand, for non-symmetric patches, the inequality $|u_k - u_i| \le u_i^{\max} - u_i$ may be violated. Therefore, in general, (6.6) does not hold, as one can see from the following counterexample. Let us consider the patch Δ_i depicted in Fig. 3 consisting of four right-angled triangles such that the vertices x_1, x_2, x_3 have the same distance h from x_i whereas the distance of x_4 from x_i is h'. Then $\gamma_{i2} = 4\,h/(h + h')$. If $u \in \mathbb{P}_1(\mathbb{R}^2)$ satisfies $u_4 = u_i^{\max}$, then $u_i - u_2 = (u_i^{\max} - u_i)h/h'$ and hence (6.6) may hold with $j = 2$ only if $h \le 3h'$.

We finish this section by stating that the definition of the limiter presented in this work introduces explicit geometric information about the mesh into the method. This is not the standard way of defining the limiters (as the usual definitions use only the matrix entries and the solution values), and is different from the one used in Ref. 28, but it has been proved to be of fundamental importance to ensure linearity preservation on general meshes.

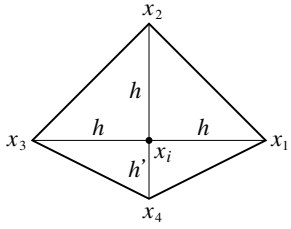

Fig. 3. Patch Δ_i for constructing a counterexample in Remark 6.3.

7. Numerical Studies

The numerical studies will illustrate the properties of the AFC scheme (2.10), (2.11) with the limiter proposed in Sec. 4 for the convection–diffusion–reaction equation from Sec. 5. If not specified otherwise, the parameters γ_i from (4.1) are defined by the formula (6.5). In addition, the results will be compared with those obtained with the limiter from Ref. 25. The limiter from Ref. 25 can be considered as a standard limiter for algebraic stabilizations of steady-state convection–diffusion–reaction equations.

For the sake of brevity, only results computed on a distorted mesh, see Fig. 4 (left), will be presented in detail. The mesh was constructed starting from the Delaunay mesh depicted in Fig. 4 (right) by shifting interior nodes to the right by half of the horizontal mesh width on each even horizontal mesh line. Therefore, for most of the diagonal edges, the sum of the two angles opposite the edge is greater than $5\pi/4$ and hence the mesh is not of Delaunay type. We shall characterize the meshes by the number of edges ne along one horizontal (or equally vertical) mesh line (thus, $ne = 6$ for both meshes in Fig. 4).

Results for three examples will be presented. In the first example, the order of convergence is studied, in both the convection-dominated and diffusion-dominated regime. The second example investigates the linearity preservation property. Finally, a standard test problem with boundary layers and an interior layer is considered.

The nonlinear discrete problems were solved with a damped Newton's method.

Example 7.1. *Polynomial solution.* Problem (5.1) is considered with $\Omega = (0,1)^2$, $\boldsymbol{b} = (3,2)^T$, $c = 1$, $u_b = 0$, and the right-hand side g is chosen so that, for a given value of ε,

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

is the solution of (5.1).

The order of convergence of the error $e_h := u - u_h$ measured in various norms for the limiter proposed in Sec. 4 is presented in Table 1 for the convection-dominated case and in Table 2 for the diffusion-dominated regime. In addition, the tables show the consistency error $d_h^{1/2}(u_h) := d_h(u_h; i_h u, i_h u)^{1/2}$, cf. estimate (5.9).

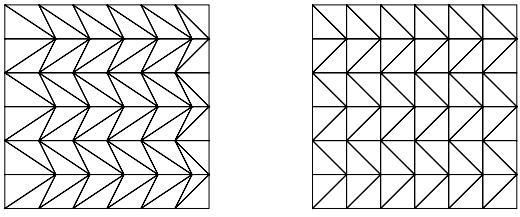

Fig. 4. Distorted mesh used in the simulations (left) and starting point for its construction (right).

Table 1. Example 7.1, $\varepsilon = 10^{-8}$, numerical results for α_{ij} from Sec. 4.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	2.722e−2	1.15	1.401e+0	0.02	9.086e−2	1.76	7.428e−2	1.21
32	1.035e−2	1.40	1.041e+0	0.43	2.287e−2	1.99	2.563e−2	1.54
64	5.099e−3	1.02	8.907e−1	0.23	6.219e−3	1.88	1.113e−2	1.20
128	2.555e−3	1.00	8.952e−1	−0.01	2.308e−3	1.43	5.240e−3	1.09
256	1.299e−3	0.98	8.991e−1	−0.01	8.409e−4	1.46	2.538e−3	1.05

Table 2. Example 7.1, $\varepsilon = 10$, numerical results for α_{ij} from Sec. 4.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	1.786e−2	1.74	4.726e−1	0.87	9.284e−1	1.13	1.522e+0	0.88
32	4.218e−3	2.08	2.404e−1	0.98	3.035e−1	1.61	7.633e−1	1.00
64	1.016e−3	2.05	1.213e−1	0.99	1.077e−1	1.49	3.841e−1	0.99
128	2.545e−4	2.00	6.082e−2	1.00	3.816e−2	1.50	1.924e−1	1.00
256	6.439e−5	1.98	3.045e−2	1.00	1.361e−2	1.49	9.632e−2	1.00
512	1.628e−5	1.98	1.524e−2	1.00	4.896e−3	1.47	4.819e−2	1.00

Concerning the convection-dominated case, results for the limiter from Ref. 25 on a mesh of the same type can be found in Table 6 from Ref. 7. Comparing the results, it can be seen that for both limiters the convergence orders of e_h are similar in all three norms. We could observe that this statement holds also for other meshes, in particular for more regular ones.

The situation is much different in the diffusion-dominated regime. Whereas the limiter from Sec. 4 leads to errors that decay with an optimal rate, see Table 2, the method with the limiter from Ref. 25 does not converge at all, cf. Table 10 from Ref. 7. This favorable behavior of the new limiter seems to be important in situations where the convection field is a flow field. In this case, there might be subregions of the domain in which the problem is diffusion-dominated.

We believe that the optimal convergence of the limiter proposed in Sec. 4 is connected with its linearity preservation property on general simplicial meshes. A similar behavior has been observed in Ref. 30, where linearity preserving limiters

Table 3. Example 7.1, $\varepsilon = 10$, numerical results for α_{ij} from Sec. 4 and γ_i replaced by $\gamma_i/4$.

ne	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
16	4.543e−2	0.91	5.801e−1	0.68	2.753e+0	0.32	2.051e+0	0.65
32	3.095e−2	0.55	3.939e−1	0.56	2.362e+0	0.22	1.404e+0	0.55
64	2.622e−2	0.24	3.138e−1	0.33	2.199e+0	0.10	1.127e+0	0.32
128	2.428e−2	0.11	2.826e−1	0.15	2.118e+0	0.05	1.018e+0	0.15
256	2.341e−2	0.05	2.707e−1	0.06	2.078e+0	0.03	9.756e−1	0.06
512	2.301e−2	0.03	2.660e−1	0.03	2.059e+0	0.01	9.582e−1	0.03

are used to approximate a diffusion problem. The theoretical justification of this statement is not yet available, and will be the topic of our future research.

Further evidence in support of the above claim is given in Table 3. Here we present results obtained with the limiter from Sec. 4 for parameters γ_i defined as a quarter of the value provided by the formula (6.5). Then the method is not linearity preserving and we observe that the errors of the approximate solutions do not converge to zero.

Example 7.2. *Linear solution.* The data for this example were chosen to be $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (2y - x, -3x + y)^T$, $c = 0$, and the boundary condition u_b and the right-hand side g were set so that

$$u(x,y) = 2x + 3y$$

is the solution of (5.1).

This example serves for showing on the one hand the linearity preservation of the limiter from Sec. 4 on the considered distorted mesh. On the other hand, it also demonstrates that the limiter from Ref. 25 does not possess this property. Results for simulations with $ne = 8$ are presented in Fig. 5 and for a closer inspection also a cross-section of the two solutions is shown in Fig. 6. The limiter proposed in Sec. 4 provides a solution which is virtually the analytical solution (the maximum error is of the order of 10^{-10}, which is in accordance with the stopping criterion for the nonlinear iteration). For the limiter from Ref. 25, the violation of the linearity preservation is clearly visible.

Example 7.3. *Solution with layers.* The final example considers a standard test problem defined in Ref. 19. This problem is given by $\Omega = (0,1)^2$, $\varepsilon = 10^{-8}$, $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))^T$, $c = 0$, $g = 0$, and the boundary condition

$$u_b(x,y) = \begin{cases} 0 & \text{for } x = 1 \text{ or } y \leq 0.7, \\ 1 & \text{else.} \end{cases}$$

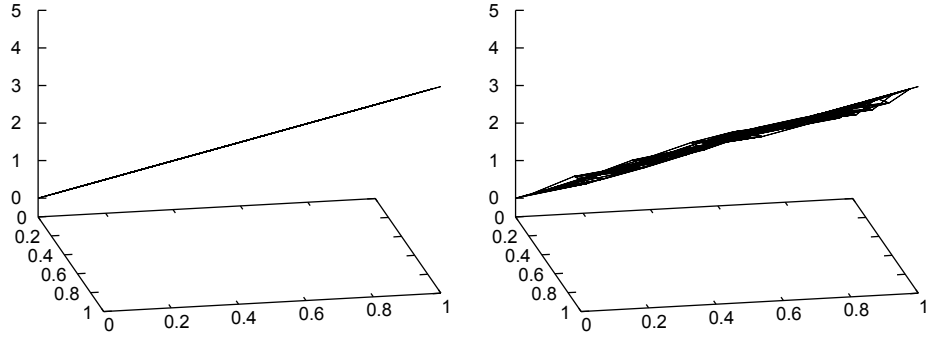

Fig. 5. Example 7.2, solution with the limiter from Sec. 4 (left) and that from Ref. 25 (right).

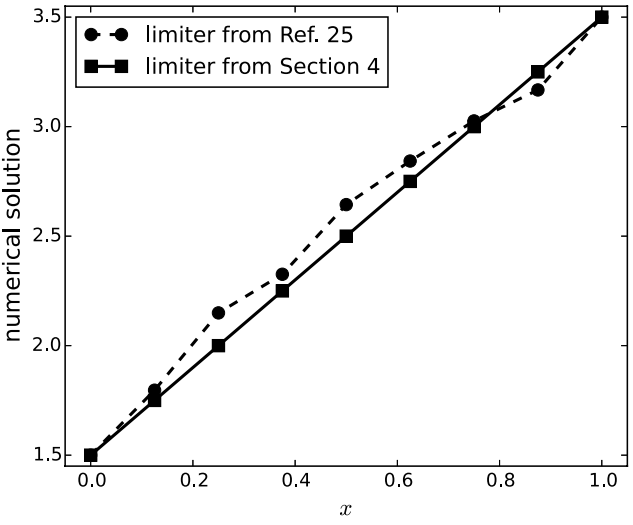

Fig. 6. Example 7.2, cross-section of the solutions at $y = 0.5$.

Note that the boundary condition from Example 7.3 can be easily changed to an infinitely smooth function that coincides with u_b from Example 7.3 at all boundary vertices of the mesh used for the computations presented in this section. Then Example 7.3 also formally fits into the framework considered in Sec. 5.

The solutions computed with both limiters are presented in Figs. 7 and 8. It can be observed that both definitions of the limiters provide an acceptable solution. They obey the DMP and all boundary layers are sharp. A close look at the interior layer, in particular at the bottom, shows that the layer of the solution computed with the limiter from Sec. 4 is a little bit sharper. Also, a slight smearing of the boundary layer at $y = 0$ is visible for the limiter from Ref. 25.

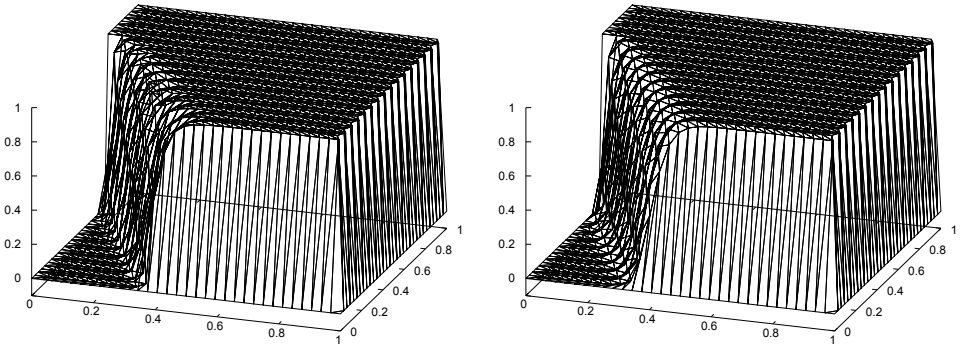

Fig. 7. Example 7.3, solutions obtained with the limiter defined in Sec. 4 (left) and the limiter from Ref. 25 (right).

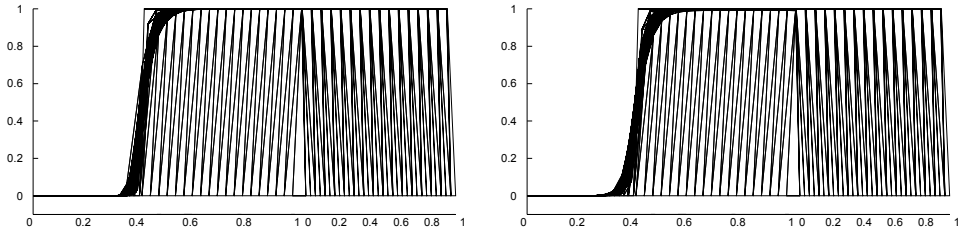

Fig. 8. Example 7.3, solutions obtained with the limiter defined in Sec. 4 (left) and the limiter from Ref. 25 (right). Both solutions respect the discrete maximum principle. The solution with the proposed limiter shows a sharper interior layer, especially at the bottom. A slight smearing can be observed along the boundary layer at $y = 0$ for the limiter from Ref. 25.

8. Conclusions and Outlook

This paper proposed a new limiter for algebraic stabilizations of steady-state convection–diffusion–reaction equations within the framework of finite element methods. The main goal of the construction of the new limiter was that the resulting scheme should obey the DMP and it should possess the linearity preservation property on general simplicial meshes. Both properties could be achieved and proved. The definition of the new limiter does not only rely on algebraic data but also requires some geometric information (on the local mesh), like the limiter of Ref. 2. We think that the enrichment of algebraic stabilizations with geometric information is in general a promising approach for designing stabilized methods. In contrast to the limiters of Refs. 2 and 5, the new limiter does not depend on any user-chosen parameter (like the exponent p in case of Refs. 2 and 5) controlling the amount of numerical diffusion added to the method, which makes the present approach more practical.

The numerical studies showed an optimal order of convergence in the diffusion-dominated regime, which is not present for the limiter from Ref. 25. As already mentioned, we believe that this behavior of the new limiter is somehow connected to the linearity preservation, but the proof is open. A further topic of our future work will be the analysis, and possibly improvement, of algebraic stabilizations for time-dependent problems.

Acknowledgments

The work of G.R.B. has been partially funded by the Leverhulme Trust via the Research Project Grant No. RPG-2012-483. The work of V.J. has been partially supported via the Grant Jo329/10-2 within the DFG priority programme 1679: Dynamic Simulation of Interconnected Solids Processes. The work of P.K. has been partially supported through the Grant No. 16-03230S of the Czech Science Foundation.

References

1. M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke and R. Umla, An assessment of discretizations for convection-dominated convection–diffusion equations, *Comput. Methods Appl. Mech. Engrg.* **200** (2011) 3395–3409.
2. S. Badia and J. Bonilla, Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization, *Comput. Methods Appl. Mech. Engrg.* **313** (2017) 133–158.
3. S. Badia and A. Hierro, On monotonicity-preserving stabilized finite element approximations of transport problems, *SIAM J. Sci. Comput.* **36** (2014) A2673–A2697.
4. S. Badia and A. Hierro, On discrete maximum principles for discontinuous Galerkin methods, *Comput. Methods Appl. Mech. Engrg.* **286** (2015) 107–122.
5. G. R. Barrenechea, E. Burman and F. Karakatsani, Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes, *Numer. Math.* **135** (2017) 521–545.
6. G. R. Barrenechea, V. John and P. Knobloch, Some analytical results for an algebraic flux correction scheme for a steady convection–diffusion equation in one dimension, *IMA J. Numer. Anal.* **35** (2015) 1729–1756.
7. G. R. Barrenechea, V. John and P. Knobloch, Analysis of algebraic flux correction schemes, *SIAM J. Numer. Anal.* **54** (2016) 2427–2451.
8. J. P. Boris and D. L. Book, Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works, *J. Comput. Phys.* **11** (1973) 38–69.
9. E. Burman and A. Ern, Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion–reaction equation, *Comput. Methods Appl. Mech. Engrg.* **191** (2002) 3833–3855.
10. E. Burman and A. Ern, Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes, *C. R. Math. Acad. Sci. Paris* **338** (2004) 641–646.
11. E. Burman and A. Ern, Stabilized Galerkin approximation of convection–diffusion–reaction equations: Discrete maximum principle and convergence, *Math. Comput.* **74** (2005) 1637–1652.
12. L. A. Catalano, P. De Palma, M. Napolitano and G. Pascazio, A critical analysis of multi-dimensional upwinding for the Euler equations, *Comput. Fluids* **25** (1996) 29–38.
13. P. G. Ciarlet and P.-A. Raviart, Maximum principle and uniform convergence for the finite element method, *Comput. Methods Appl. Mech. Engrg.* **2** (1973) 17–31.
14. A. Ern and J.-L. Guermond, Weighting the edge stabilization, *SIAM J. Numer. Anal.* **51** (2013) 1655–1677.
15. Z. Gao and J. Wu, A linearity-preserving cell-centered scheme for the heterogeneous and anisotropic diffusion equations on general meshes, *Int. J. Numer. Methods Fluids* **67** (2011) 2157–2183.
16. D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd edn., Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], Vol. 224 (Springer, 1983).
17. J.-L. Guermond and M. Nazarov, A maximum-principle preserving C^0 finite element method for scalar conservation equations, *Comput. Methods Appl. Mech. Engrg.* **272** (2014) 198–213.
18. J.-L. Guermond, M. Nazarov, B. Popov and Y. Yang, A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations, *SIAM J. Numer. Anal.* **52** (2014) 2163–2182.

19. T. J. R. Hughes, M. Mallet and A. Mizukami, A new finite element formulation for computational fluid dynamics. II. Beyond SUPG, *Comput. Methods Appl. Mech. Engrg.* **54** (1986) 341–355.
20. W. Hundsdorfer and C. Montijn, A note on flux limiting for diffusion discretizations, *IMA J. Numer. Anal.* **24** (2004) 635–642.
21. V. John and P. Knobloch, On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I — A review, *Comput. Methods Appl. Mech. Engrg.* **196** (2007) 2197–2215.
22. V. John and E. Schmeyer, Finite element methods for time-dependent convection–diffusion–reaction equations with small diffusion, *Comput. Methods Appl. Mech. Engrg.* **198** (2008) 475–494.
23. C. Johnson, *Numerical Solution of Partial Differential Equations by the Finite Element Method* (Dover Publications, Inc., Mineola, NY, 2009). Reprint of the 1987 edition.
24. D. Kuzmin, On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection, *J. Comput. Phys.* **219** (2006) 513–531.
25. D. Kuzmin, Algebraic flux correction for finite element discretizations of coupled systems, in *Proc. Int. Conf. Computational Methods for Coupled Problems in Science and Engineering*, eds. M. Papadrakakis, E. Oñate and B. Schrefler, CIMNE, Barcelona (2007), pp. 1–5.
26. D. Kuzmin, On the design of algebraic flux correction schemes for quadratic finite elements, *J. Comput. Appl. Math.* **218** (2008) 79–87.
27. D. Kuzmin, Explicit and implicit FEM–FCT algorithms with flux linearization, *J. Comput. Phys.* **228** (2009) 2517–2534.
28. D. Kuzmin, Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes, *J. Comput. Appl. Math.* **236** (2012) 2317–2337.
29. D. Kuzmin and J. Hämäläinen, *Finite Element Methods for Computational Fluid Dynamics*: *A Practical Guide*, Computational Science and Engineering, Vol. 14 (SIAM, 2015).
30. D. Kuzmin, M. J. Shashkov and D. Svyatskiy, A constrained finite element method satisfying the discrete maximum principle for anisotropic diffusion problems, *J. Comput. Phys.* **228** (2009) 3448–3463.
31. A. Mizukami and T. J. R. Hughes, A Petrov–Galerkin finite element method for convection-dominated flows: An accurate upwinding technique for satisfying the maximum principle, *Comput. Methods Appl. Mech. Engrg.* **50** (1985) 181–193.
32. J. Xu and L. Zikatanov, A monotone finite element scheme for convection–diffusion equations, *Math. Comput.* **68** (1999) 1429–1446.
33. S. T. Zalesak, Fully multidimensional flux-corrected transport algorithms for fluids, *J. Comput. Phys.* **31** (1979) 335–362.