SIAM J. NUMER. ANAL.
Vol. 54, No. 4, pp. 2427–2451

ANALYSIS OF ALGEBRAIC FLUX CORRECTION SCHEMES*

GABRIEL R. BARRENECHEA†, VOLKER JOHN‡, AND PETR KNOBLOCH§

Abstract. A family of algebraic flux correction (AFC) schemes for linear boundary value problems in any space dimension is studied. These methods' main feature is that they limit the fluxes along each one of the edges of the triangulation, and we suppose that the limiters used are symmetric. For an abstract problem, the existence of a solution, existence and uniqueness of the solution of a linearized problem, and an a priori error estimate are proved under rather general assumptions on the limiters. For a particular (but standard in practice) choice of the limiters, it is shown that a local discrete maximum principle holds. The theory developed for the abstract problem is applied to convection-diffusion-reaction equations, where in particular an error estimate is derived. Numerical studies show its sharpness.

Key words. algebraic flux correction method, linear boundary value problem, well-posedness, discrete maximum principle, convergence analysis, convection-diffusion-reaction equations

AMS subject classifications. 65N12, 65N30

DOI. 10.1137/15M1018216

1. Introduction.

Many processes from nature and industry can be modeled using (systems of) partial differential equations (PDEs). Usually, these equations cannot be solved analytically. Instead, only numerical approximations can be computed, e.g., by using a finite element method (FEM). The Galerkin FEM replaces just the infinite-dimensional spaces from the variational form of the differential equation with finite-dimensional counterparts. However, if the considered problem contains a wide range of important scales, the Galerkin FEM does not give useful numerical results unless all scales are resolved. For many problems, the resolution of all scales is not affordable because of the huge computational costs (memory, computing time). The remedy consists of modifying the Galerkin FEM in such a way that the effect of small scales is taken into account on grids which do not resolve all scales. This methodology is usually called stabilization. The most common strategy modifies or enriches the Galerkin FEM, e.g., such that the new discrete problem provides additional control of the error in appropriate norms. An alternative approach acts on the algebraic level; i.e., algebraic representations of discrete operators and vectors are modified before computing a numerical solution. This paper studies a method of the latter type.

Applications of algebraically stabilized FEMs can be found in particular for convection-dominated problems. Their construction, e.g., in [18, 16, 17], is performed for transport equations, and they are called flux-corrected transport (FCT) schemes

*Received by the editors April 23, 2015; accepted for publication (in revised form) March 30, 2016; published electronically August 16, 2016. The research of the authors was funded by the Leverhulme Trust under grant RPG-2012-483.

http://www.siam.org/journals/sinum/54-4/M101821.html

†Department of Mathematics and Statistics, University of Strathclyde, Glasgow G1 1XH, Scotland (gabriel.barrenechea@strath.ac.uk).

‡Weierstrass Institute for Applied Analysis and Stochastics (WIAS), 10117 Berlin, Germany, and Department of Mathematics and Computer Science, Free University of Berlin, 14195 Berlin, Germany (john@wias-berlin.de). The research of this author was partially supported by grant Jo329/10-2 within the DFG priority programme 1679: Dynamic simulation of interconnected solids processes.

§Department of Numerical Mathematics, Faculty of Mathematics and Physics, Charles University in Prague, 18675 Praha 8, Czech Republic (knobloch@karlin.mff.cuni.cz). The research of this author was partially supported through grant 13-00522S of the Czech Science Foundation.

(see also [7] for their application to compressible flows). These schemes can be used also for the discretization of time-dependent convection-diffusion equations, e.g., as in [4, 11], where the convection-diffusion equations are part of population balance systems. In [11] it is explicitly emphasized that the FCT scheme was preferred over the popular streamline-upwind Petrov–Galerkin (SUPG) stabilization, which adds an additional term to the Galerkin FEM, because of a former bad experience with this stabilization. More precisely, the lack of positivity of the solution provided by SUPG caused blow ups in finite time for some nonlinear coupled problems in chemical engineering (for details, see [10]). Altogether, the advantages of the FCT methods, compared with the majority of other stabilized methods, are as follows. First, their construction relies on the goal of conservation and of satisfying a discrete maximum principle. Second, since this sort of method acts only at the algebraic level, without taking into consideration the weak formulation, their implementation is independent of the space dimension. The importance of these two points for many applications does not need to be emphasized. However, there are also drawbacks. First, for most methods, one has to solve a nonlinear discrete problem, even when the PDE to be solved is linear. This issue is, in our opinion, of minor importance, since in applications one encounters generally nonlinear problems. Second, the FCT methodology has, so far, been applied successfully only for lowest order finite elements, which limits the accuracy of the computed solutions to the best approximation in these spaces (the only exception of this fact being, to the best of our knowledge, the work [15]).

This paper analyzes algebraic stabilizations for linear steady-state boundary value problems. These methods are called algebraic flux correction (AFC) schemes. Apart from the obvious properties of these methods, which are the basis of their construction, there has been no numerical analysis of them until very recently. The first contribution in this field is [2], where some preliminary results on the analysis of an AFC scheme (cf. [14]) for a linear steady-state convection-diffusion-reaction equation in one space dimension were reported. The discretization studied in [2] is in some sense more general than the AFC methodology used in practice. In the methodology of [2], one has to compute limiters $\alpha_{ij} \in [0,1]$ (see below), and in contrast to the common application of AFC schemes, it was not assumed that $\alpha_{ij} = \alpha_{ji}$, which may cause a lack of conservation. Besides other properties, it was proved in [2] that the nonlinear discrete problem might not even possess a solution. Thus, there is an important physical as well as a strong mathematical reason for including the symmetry condition in the scheme, which will be done in this paper.

The first part of the paper (sections 2–6) considers a general linear boundary value problem in several space dimensions. After introducing a nonlinear AFC scheme in section 2, the existence of a solution is proved, and then the existence of a unique solution of the linearized scheme is shown, both in section 3. The symmetry of the limiters, i.e., $\alpha_{ij} = \alpha_{ji}$, the requirement that $\alpha_{ij} \in [0,1]$, and a continuity assumption are the minimal assumptions used in this section. Section 4 considers a concrete choice of the limiters, which is a standard definition found in the literature. It is shown that these limiters satisfy the assumptions made in the preceding analysis, so they lead to discrete problems that have a solution. In section 5 we give a general proof of the discrete maximum principle, since we have not been able to find it in the literature, although the AFC family of methods is built to preserve this property. In section 6, the AFC scheme is formulated in a variational form and an abstract error estimate is derived, with only the same minimal assumptions on the limiters as used in section 3. As usual for stabilized methods, the norm for which the error estimate is given contains a contribution from the stabilization. To the best of our

knowledge, this is the first error estimate for algebraically stabilized FEMs. In the second part of the paper (sections 7–8), the abstract theory is applied to steady-state linear convection-diffusion-reaction equations. In section 7 an error estimate for this kind of equation is derived. Numerical studies are presented in section 8. It is shown that within the minimal assumptions on the limiters used in the analysis, the derived error estimate is sharp. However, applying the definition of the limiters as discussed in section 5, one can observe a higher order of convergence. The orders of convergence for standard norms depend on the concrete grid and are sometimes suboptimal. Finally, in an appendix at the end of the paper a few supplementary results are proved.

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

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

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,. \tag{2}$$

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

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j > 0 \quad \forall\, (u_1, \ldots, u_M) \in \mathbb{R}^M \setminus \{0\}\,. \tag{3}$$

It is natural to require that the maximum principle also hold for the discrete problem (1), (2). Due to (3), the diagonal entries of the matrix $(a_{ij})_{i,j=1}^{M}$ are positive, and hence, locally, the discrete maximum principle corresponds to the statement

$$\forall\, i \in \{1, \ldots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i,\, a_{ij} \ne 0} u_j \tag{4}$$

or, at least,

$$\forall\, i \in \{1, \ldots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i,\, a_{ij} \ne 0} u_j^+\,, \tag{5}$$

where $u_j^+ = \max\{0, u_j\}$. It can be shown (cf. the appendix), that (4) holds if and only if

$$a_{ij} \le 0 \qquad \forall\, i \ne j,\, i = 1, \ldots, M,\, j = 1, \ldots, N\,, \tag{6}$$

and

$$\sum_{j=1}^{N} a_{ij} = 0\,, \qquad i = 1, \ldots, M\,. \tag{7}$$

The discrete maximum principle (5) holds if and only if (6) is satisfied and

$$\sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1, \ldots, M\,. \tag{8}$$

While conditions (7) or (8) are often satisfied, the property (6) does not hold for many discretizations, in particular, of convection-dominated problems. The aim of the AFC method is to modify the algebraic system (1) in such a way that the necessary conditions for the validity of the discrete maximum principle are satisfied and layers are not excessively smeared.

The starting point of the AFC algorithm is the finite element matrix $\mathbb{A} = (a_{ij})_{i,j=1}^{N}$ corresponding to the above-mentioned finite element discretization in the case where homogeneous natural boundary conditions are used instead of the Dirichlet ones. We introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^{N}$ possessing the entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \quad \forall\, i \neq j\,, \qquad d_{ii} = -\sum_{j \neq i} d_{ij}\,. \tag{9}$$

Then the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ satisfies the necessary conditions for the discrete maximum principle provided that (7) or (8) holds for the matrix $\mathbb{A}$.

Going back to the solution of (1), this system is equivalent to

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + (\mathbb{D}\,\mathrm{U})_i\,, \qquad i = 1, \ldots, M\,. \tag{10}$$

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

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

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$. Now the idea of the AFC schemes is to limit those antidiffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (1) (or, equivalently, (10)) is replaced by

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j \neq i} \alpha_{ij}\, f_{ij}\,, \qquad i = 1, \ldots, M\,, \tag{11}$$

with solution-dependent correction factors $\alpha_{ij} \in [0, 1]$. For $\alpha_{ij} = 1$, the original system (1) is recovered. Hence, intuitively, the coefficients α_{ij} should be as close to 1 as possible to limit the modifications of the original problem. They can be chosen in various ways, but their definition is always based on the above fluxes f_{ij}; see [13, 14, 15, 16, 17] for examples. To guarantee that the resulting scheme is conservative, one should require that the coefficients α_{ij} be symmetric, i.e.,

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i, j = 1, \ldots, N\,. \tag{12}$$

Rewriting (11) using the definition of the matrix $\tilde{\mathbb{A}}$, one obtains the final form of the AFC scheme to be investigated in this paper. It is the following system of nonlinear equations:

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) = g_i\,, \qquad i = 1, \ldots, M\,, \tag{13}$$

$$u_i = u_i^b\,, \qquad i = M+1, \ldots, N\,, \tag{14}$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \ldots, u_N) \in [0, 1]$, $i, j = 1, \ldots, N$, satisfy (12).

3. Solvability of the algebraic flux correction scheme and of its linearized variant. In this section we prove that the nonlinear problem (13), (14) is solvable under a continuity assumption on α_{ij}. As a consequence, we obtain the unique solvability of the linearized problem (13), (14) (with α_{ij} independent of the solution), which is useful for computing the solution of (13), (14) numerically using a fixed-point iteration. The following result will be of great use in the proof of existence of solutions below.

LEMMA 1. *Consider any $\mu_{ij} = \mu_{ji} \le 0$, $i, j = 1, \dots, N$. Then*

$$\sum_{i,j=1}^{N} v_i \, \mu_{ij} \, (v_j - v_i) = - \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} \mu_{ij} \, (v_i - v_j)^2 \ge 0 \quad \forall \, v_1, \dots, v_N \in \mathbb{R} \, .$$

Proof. A quick calculation shows that

$$\begin{aligned} \sum_{i,j=1}^{N} v_i \, \mu_{ij} \, (v_j - v_i) &= \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} v_i \, \mu_{ij} \, (v_j - v_i) + \sum_{\substack{j,\, i \,=\, 1 \\ j \,>\, i}}^{N} v_j \, \mu_{ji} \, (v_i - v_j) \\ &= - \sum_{\substack{i,\, j \,=\, 1 \\ i \,<\, j}}^{N} \mu_{ij} \, (v_i - v_j)^2 \ge 0 \, , \end{aligned}$$

and the proof is finished. ∎

For proving the solvability of the nonlinear problem, we use the following consequence of Brower's fixed-point theorem, whose proof can be found in [20, Lemma 1.4, p. 164].

LEMMA 2. *Let X be a finite-dimensional Hilbert space with inner product $(\cdot, \cdot)_X$ and norm $\| \cdot \|_X$. Let $T : X \to X$ be a continuous mapping, and let $K > 0$ be a real number such that $(Tx, x)_X > 0$ for any $x \in X$ with $\|x\|_X = K$. Then there exists $x \in X$ such that $\|x\|_X < K$ and $Tx = 0$.*

The following is our main result on existence of solutions for the AFC scheme.

THEOREM 3. *Let* (3) *hold. For any $i, j \in \{1, \dots, N\}$, let $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ be such that $\alpha_{ij}(u_1, \dots, u_N)(u_j - u_i)$ is a continuous function of $u_1, \dots, u_N$. Finally, let the functions α_{ij} satisfy* (12)*. Then there exists a solution of the nonlinear problem* (13)*,* (14)*.*

Proof. Throughout this proof, we denote by $\widetilde{\mathrm{V}} \equiv (v_1, \dots, v_M)$ the elements of the space $\mathbb{R}^M$ and, if v_i with $i \in \{M+1, \dots, N\}$ occurs, we always assume that $v_i = u_i^b$. To any $\widetilde{\mathrm{V}} \in \mathbb{R}^M$ we assign $\mathrm{V} := (v_1, \dots, v_N)$. Furthermore, we set $\mathrm{G} := (g_1, \dots, g_M)$. We denote by $(\cdot, \cdot)$ the usual inner product in $\mathbb{R}^M$ and by $\| \cdot \|$ the corresponding (Euclidean) norm.

It is easy to show by contradiction that, in view of (3),

$$C_M := \inf_{\|\widetilde{\mathrm{V}}\| = 1} \sum_{i,j=1}^{M} v_i \, a_{ij} \, v_j > 0 \, .$$

Thus, one has

$$\sum_{i,j=1}^{M} v_i \, a_{ij} \, v_j \ge C_M \, \|\widetilde{\mathrm{V}}\|^2 \quad \forall \, \widetilde{\mathrm{V}} \in \mathbb{R}^M \, . \tag{15}$$

Let us define the operator $T : \mathbb{R}^M \to \mathbb{R}^M$ by

$$(T\,\widetilde{\mathrm{V}})_i = \sum_{j=1}^{N} a_{ij}\, v_j + \sum_{j=1}^{N} [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - v_i) - g_i\,, \qquad i = 1, \ldots, M\,.$$

Then U is a solution of the nonlinear problem (13), (14) if and only if $T\,\widetilde{\mathrm{U}} = 0$. The operator T is continuous and, in view of (15), Lemma 1, and Hölder's and Young's inequalities, one derives

$$\begin{aligned}(T\,\widetilde{\mathrm{V}}, \widetilde{\mathrm{V}}) &= \sum_{i,j=1}^{M} v_i\, a_{ij}\, v_j + \sum_{i,j=1}^{N} v_i\, [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - v_i) \\ &\quad + \sum_{i=1}^{M} v_i \sum_{j=M+1}^{N} a_{ij}\, u_j^b - \sum_{i=M+1}^{N} u_i^b \sum_{j=1}^{N} [1 - \alpha_{ij}(\mathrm{V})]\, d_{ij}\, (v_j - u_i^b) - (\mathrm{G}, \widetilde{\mathrm{V}}) \\ &\geq C_M\, \|\widetilde{\mathrm{V}}\|^2 - C_0 - C_1\, \|\widetilde{\mathrm{V}}\| \geq \frac{C_M}{2}\, \|\widetilde{\mathrm{V}}\|^2 - C_2\,,\end{aligned}$$

where C_0, C_1, and C_2 are positive constants that do not depend on $\widetilde{\mathrm{V}}$. Then for any $\widetilde{\mathrm{V}} \in \mathbb{R}^M$ satisfying $\|\widetilde{\mathrm{V}}\| = \sqrt{3\, C_2 / C_M}$, one has $(T\,\widetilde{\mathrm{V}}, \widetilde{\mathrm{V}}) > 0$, and hence, according to Lemma 2, there exists $\widetilde{\mathrm{U}} \in \mathbb{R}^M$ such that $T\,\widetilde{\mathrm{U}} = 0$. ∎

COROLLARY 4. *Let* (3) *hold. Consider any* $\alpha_{ij} \in [0, 1]$, $i, j = 1, \ldots, N$, *satisfying* (12). *Then the system* (13), (14) *has a unique solution for any* $g_1, \ldots, g_M \in \mathbb{R}$ *and* $u_{M+1}^b, \ldots, u_N^b \in \mathbb{R}$.

Proof. According to Theorem 3, for any values of $g_1, \ldots, g_M$ and $u_{M+1}^b, \ldots, u_N^b$, there exists a solution of the considered linear system. Consequently, the solutions have to be unique. ∎

Remark 5. The statement of Corollary 4 can be proved directly (without using Theorem 3) by showing that the homogeneous system

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) = 0\,, \qquad i = 1, \ldots, M\,, \tag{16}$$

$$u_i = 0\,, \qquad i = M+1, \ldots, N\,, \tag{17}$$

has only the trivial solution. Indeed, if $U = (u_1, \ldots, u_N)$ solves (16), (17), then according to Lemma 1, one has

$$\sum_{i,j=1}^{M} u_i\, a_{ij}\, u_j = -\sum_{i,j=1}^{N} u_i\, (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) \leq 0\,.$$

Therefore, $u_i = 0$, $i = 1, \ldots, M$, in view of (3).

Finally, let us formulate sufficient conditions on the functions α_{ij}, ensuring the validity of the continuity assumption in Theorem 3 for many particular examples of the functions α_{ij} used in practice (cf., e.g., [13, 16, 17]).

LEMMA 6. *Consider any* $i, j \in \{1, \ldots, N\}$, *and let* $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ *satisfy*

$$\alpha_{ij}(U) = \frac{A_{ij}(U)}{|u_j - u_i| + B_{ij}(U)} \quad \forall\, U \equiv (u_1, \ldots, u_N) \in \mathbb{R}^N,\ u_i \neq u_j\,, \tag{18}$$

where $A_{ij}, B_{ij} : \mathbb{R}^N \to [0,\infty)$ are nonnegative functions that are continuous at any point $U \in \mathbb{R}^N$ with $u_i \neq u_j$. Then $\Phi_{ij}(U) := \alpha_{ij}(U)(u_j - u_i)$ is a continuous function of $u_1, \ldots, u_N$ on $\mathbb{R}^N$. Moreover, if the functions A_{ij}, B_{ij} are Lipschitz-continuous with the constant L in the sets $\{U \in \mathbb{R}^N;\ u_i < u_j\}$ and $\{U \in \mathbb{R}^N;\ u_i > u_j\}$, then the function Φ_{ij} is Lipschitz-continuous on $\mathbb{R}^N$, with the constant $2\,L + \sqrt{2}$.

Proof. Consider any $\bar{U} \equiv (\bar{u}_1, \ldots, \bar{u}_N) \in \mathbb{R}^N$. If $\bar{u}_i \neq \bar{u}_j$, then there is a neighbourhood of $\bar{U}$, where the denominator from (18) does not vanish and the functions A_{ij}, B_{ij} are continuous so that α_{ij} is continuous at $\bar{U}$. If $\bar{u}_i = \bar{u}_j$, we employ the fact that $\alpha_{ij} \in [0,1]$, which implies that $|\alpha_{ij}(U)(u_j - u_i)| \leq |u_j - u_i| \leq \sqrt{2}\,\|U - \bar{U}\|$ for any $U \equiv (u_1, \ldots, u_N) \in \mathbb{R}^N$. Thus, $\alpha_{ij}(U)(u_j - u_i)$ is continuous at $\bar{U}$.

To prove the Lipschitz-continuity of Φ_{ij}, consider any U, $\bar{U} \in \mathbb{R}^N$ with $U = (u_1, \ldots, u_N)$ and $\bar{U} = (\bar{u}_1, \ldots, \bar{u}_N)$. Set $v = u_j - u_i$, $\bar{v} = \bar{u}_j - \bar{u}_i$. If $v\,\bar{v} \leq 0$, then

$$|\Phi_{ij}(U) - \Phi_{ij}(\bar{U})| \leq |v| + |\bar{v}| = |v - \bar{v}| \leq \sqrt{2}\,\|U - \bar{U}\|\,.$$

If $v\,\bar{v} > 0$, then

$$\Phi_{ij}(U) - \Phi_{ij}(\bar{U}) = (A_{ij}(U) - A_{ij}(\bar{U}))\,\frac{\bar{v}}{|\bar{v}| + B_{ij}(\bar{U})} + \alpha_{ij}(U)\,\frac{(B_{ij}(\bar{U}) - B_{ij}(U))\,\bar{v} + (v - \bar{v})\,B_{ij}(\bar{U})}{|\bar{v}| + B_{ij}(\bar{U})},$$

and hence

$$|\Phi_{ij}(U) - \Phi_{ij}(\bar{U})| \leq |A_{ij}(U) - A_{ij}(\bar{U})| + |B_{ij}(U) - B_{ij}(\bar{U})| + |v - \bar{v}|\,.$$

This proves the lemma. ∎

4. An example of the choice of α_{ij}. In this section we present a concrete choice of the limiters α_{ij}. This choice is often used in computations, and we show that it satisfies the assumptions of Lemma 6 and hence leads to a solvable nonlinear problem (13), (14).

The definition of the coefficients α_{ij} considered in this section relies on the values P_i^+, P_i^-, Q_i^+, Q_i^- computed for $i = 1, \ldots, N$ in the following way. First, one initializes all these quantities by zero. Then one goes through all pairs of indices $i, j \in \{1, \ldots, N\}$ and performs the updates

$$\begin{aligned}
&P_i^+ := P_i^+ + \max\{0, f_{ij}\}\,,\ P_i^- := P_i^- - \max\{0, f_{ji}\} \quad \text{if } a_{ji} \leq a_{ij}\,,\\
&Q_i^+ := Q_i^+ + \max\{0, f_{ji}\}\,,\ Q_i^- := Q_i^- - \max\{0, f_{ij}\} \quad \text{if } i < j\,,\\
&Q_j^+ := Q_j^+ + \max\{0, f_{ij}\}\,,\ Q_j^- := Q_j^- - \max\{0, f_{ji}\} \quad \text{if } i < j\,,
\end{aligned}$$

where we again use the notation $f_{ij} = d_{ij}\,(u_j - u_i)$. After having computed the values P_i^+, P_i^-, Q_i^+, Q_i^-, $i = 1, \ldots, N$, one defines

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

If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Furthermore, according to [12], these quantities are set to 1 at Dirichlet nodes, i.e.,

$$R_i^+ := 1\,, \quad R_i^- := 1\,, \qquad i = M+1, \ldots, N\,.$$

Finally, for any $i, j \in \{1, \ldots, N\}$ such that $a_{ji} \leq a_{ij}$, one sets

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}\,. \tag{19}$$

It is worth mentioning that this algorithm is the one presented in [14] (that originates from the ideas of [22]) to which, following [12], the symmetry condition $\alpha_{ij} = \alpha_{ji}$ has been added.

Note that the quantities P_i^+, P_i^-, Q_i^+, Q_i^- can be expressed in the form

$$P_i^+ = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} f_{ij}^+\,, \qquad P_i^- = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} f_{ij}^-\,, \qquad Q_i^+ = -\sum_{j=1}^{N} f_{ij}^-\,, \qquad Q_i^- = -\sum_{j=1}^{N} f_{ij}^+\,, \tag{20}$$

where $f_{ij}^+ = \max\{0, f_{ij}\}$ and $f_{ij}^- = \min\{0, f_{ij}\}$.

The following result shows that the above coefficients α_{ij} satisfy the hypotheses of Theorem 3, and then that they lead to a solvable nonlinear problem (13), (14).

LEMMA 7. *The above coefficients α_{ij} are such that $\alpha_{ij}(u_1, \ldots, u_N)(u_j - u_i)$ are Lipschitz-continuous functions of $u_1, \ldots, u_N$ on $\mathbb{R}^N$.*

Proof. Consider any $i, j \in \{1, \ldots, N\}$. It suffices to consider the case $\alpha_{ij} \not\equiv 1$ (and hence $d_{ij} \neq 0$). Furthermore, due to (12), one may assume that $a_{ji} \leq a_{ij}$. If $u_i > u_j$, then $f_{ij} > 0$, and hence

$$\alpha_{ij} = R_i^+ = \frac{\min\{P_i^+, Q_i^+\}}{|f_{ij}| + \widetilde{P}_i^+} \qquad \text{with} \qquad \widetilde{P}_i^+ = \sum_{\substack{k=1 \\ a_{ki} \leq a_{ik},\, k \neq j}}^{N} f_{ik}^+\,.$$

If $u_i < u_j$, then $f_{ij} < 0$ so that

$$\alpha_{ij} = R_i^- = \frac{\min\{-P_i^-, -Q_i^-\}}{|f_{ij}| - \widetilde{P}_i^-} \qquad \text{with} \qquad \widetilde{P}_i^- = \sum_{\substack{k=1 \\ a_{ki} \leq a_{ik},\, k \neq j}}^{N} f_{ik}^-\,.$$

Thus, α_{ij} is of the form (18), with functions A_{ij} and B_{ij} satisfying

$$A_{ij} = \frac{1}{|d_{ij}|} \begin{cases} \min\{-P_i^-, -Q_i^-\} & \text{if } u_i < u_j\,, \\ \min\{P_i^+, Q_i^+\} & \text{if } u_i > u_j\,, \end{cases} \qquad B_{ij} = \frac{1}{|d_{ij}|} \begin{cases} -\widetilde{P}_i^- & \text{if } u_i < u_j\,, \\ \widetilde{P}_i^+ & \text{if } u_i > u_j\,. \end{cases}$$

Since the maximum or minimum of two Lipschitz-continuous functions with constant L is again a Lipschitz-continuous function with constant L, the functions A_{ij} and B_{ij} are Lipschitz-continuous with constant $\sqrt{2}\,(\sum_{k=1}^{N} |d_{ik}|)/|d_{ij}|$ in the sets $\{u_i < u_j\}$ and $\{u_i > u_j\}$. Then the hypotheses of Lemma 6 are satisfied, and the result immediately follows from Lemma 6. ∎

Remark 8. There is an apparent ambiguity in the definition of the coefficients α_{ij} if $a_{ij} = a_{ji}$. However, often $a_{ij} + a_{ji} \leq 0$ (cf. assumption (22) in the next section), and then $a_{ij} = a_{ji} \leq 0$. Thus, if the artificial diffusion matrix is defined by (9), one obtains $d_{ij} = 0$ so that the respective α_{ij} does not occur in the nonlinear problem (13), (14) and can be defined arbitrarily.

5. The discrete maximum principle. In this section we prove several versions of the discrete maximum principle for the case when the coefficients α_{ij} are defined as in the previous section. We start with the main assumptions needed for the proofs, namely,

$$a_{ii} > 0\,, \qquad \sum_{j=1}^{N} a_{ij} \geq 0 \quad \forall\, i = 1, \ldots, M\,, \tag{21}$$

$$a_{kl} + a_{lk} \leq 0 \quad \forall\, k, l = 1, \ldots, N\,,\ k \neq l\,,\ k \leq M,\ \text{or}\ l \leq M\,, \tag{22}$$

and we recall that $d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\}$ for all $i, j = 1, \ldots, N$, $i \neq j$ (cf. (9)). The first condition in (21) is a consequence of (3), and the second is a necessary condition for the validity of the discrete maximum principle in the case of linear problem (1), (2). Note that the row sums are not affected by adding the nonlinear term in (13). Condition (22) is weaker than (6). In section 7, we present a discrete problem for which all the assumptions in (21) and (22) are satisfied.

Also, we present some notation that will be useful in what follows. We denote by

$$\mathrm{Up}_i = \{j \in \{1, \ldots, N\}\,;\ j \neq i,\ a_{ij} < 0\}\,, \qquad i = 1, \ldots, M\,,$$

the sets of upwind nodes, and by

$$\mathrm{Do}_i = \{j \in \{1, \ldots, N\}\,;\ j \neq i,\ a_{ij} > 0\}\,, \qquad i = 1, \ldots, M\,,$$

the sets of downwind nodes. In what follows, we shall tacitly assume that these sets are not empty.

Thanks to (22), for any $i \in \{1, \ldots, M\}$ and $j \in \{1, \ldots, N\}$ such that $i \neq j$ and $d_{ij} \neq 0$, one derives

$$a_{ij} < a_{ji} \quad \Leftrightarrow \quad j \in \mathrm{Up}_i\,, \qquad\qquad a_{ji} \leq a_{ij} \quad \Leftrightarrow \quad j \in \mathrm{Do}_i\,.$$

Therefore, the sums in (20) defining P_i^+ and P_i^- can be written in the form

$$P_i^+ = \sum_{j \in \mathrm{Do}_i} f_{ij}^+\,, \qquad P_i^- = \sum_{j \in \mathrm{Do}_i} f_{ij}^-\,, \qquad i = 1, \ldots, M\,. \tag{23}$$

Moreover, the second term on the left-hand side of (13) can be written as

$$\begin{aligned}\sum_{j=1}^{N} (1 - \alpha_{ij})\, f_{ij} &= \sum_{j=1}^{N} f_{ij} - \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{N} \alpha_{ij}\, f_{ij} + \sum_{\substack{j=1 \\ a_{ij} < a_{ji}}}^{N} \alpha_{ji}\, f_{ji} \\ &= \sum_{j=1}^{N} f_{ij} - \sum_{j \in \mathrm{Do}_i} \alpha_{ij}\, f_{ij} + \sum_{j \in \mathrm{Up}_i} \alpha_{ji}\, f_{ji}\,.\end{aligned}$$

Furthermore, $\alpha_{ij}\, f_{ij} = R_i^+\, f_{ij}^+ + R_i^-\, f_{ij}^-$ for $i \in \{1, \ldots, M\}$ and $j \in \mathrm{Do}_i$, and consequently, $\alpha_{ji}\, f_{ji} = R_j^+\, f_{ji}^+ + R_j^-\, f_{ji}^-$ if $i \in \{1, \ldots, M\}$ and $j \in \mathrm{Up}_i$. Then since $f_{ji}^+ = -f_{ij}^-$ and $f_{ji}^- = -f_{ij}^+$, one obtains

$$\sum_{j=1}^{N} (1 - \alpha_{ij})\, f_{ij} = \sum_{j=1}^{N} f_{ij} - \sum_{j \in \mathrm{Do}_i} (R_i^+\, f_{ij}^+ + R_i^-\, f_{ij}^-) - \sum_{j \in \mathrm{Up}_i} (R_j^+\, f_{ij}^- + R_j^-\, f_{ij}^+)\,.$$

Finally, denoting $Z_i^+ := 1 - R_i^+$ and $Z_i^- := 1 - R_i^-$, it follows that

$$\sum_{j=1}^{N} (1-\alpha_{ij})\, f_{ij} = \sum_{j\in \mathrm{Do}_i} (Z_i^+ f_{ij}^+ + Z_i^- f_{ij}^-) + \sum_{j\in \mathrm{Up}_i} (Z_j^+ f_{ij}^- + Z_j^- f_{ij}^+)\,.$$

Thus, the AFC scheme (13), (14) can be written in the form

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j\in \mathrm{Do}_i} (Z_i^+ f_{ij}^+ + Z_i^- f_{ij}^-) + \sum_{j\in \mathrm{Up}_i} (Z_j^+ f_{ij}^- + Z_j^- f_{ij}^+) = g_i\,, \qquad i = 1,\dots,M\,, \tag{24}$$

$$u_i = u_i^b\,, \qquad i = M+1,\dots,N\,. \tag{25}$$

Next, defining

$$A_i = u_i \sum_{j=1}^{N} a_{ij}\,, \tag{26}$$

one derives, for any $i \in \{1,\dots,M\}$,

$$\sum_{j=1}^{N} a_{ij}\, u_j = \sum_{j=1}^{N} a_{ij}\,(u_j - u_i) + A_i = \sum_{j\in \mathrm{Up}_i} a_{ij}\,(u_j - u_i) + \sum_{j\in \mathrm{Do}_i} a_{ij}\,(u_j - u_i) + A_i\,.$$

In view of (22), one has $a_{ij} = -d_{ij}$ for $j \in \mathrm{Do}_i$, and then

$$\sum_{j=1}^{N} a_{ij}\, u_j = \sum_{j\in \mathrm{Up}_i} a_{ij}\,(u_j - u_i) - \sum_{j\in \mathrm{Do}_i} f_{ij} + A_i\,.$$

Therefore, using that $\sum_{j\in \mathrm{Do}_i} f_{ij} = P_i^+ + P_i^-$ (cf. (23)), (24) is equivalent to

$$A_i - P_i^+\, R_i^+ - P_i^-\, R_i^- + \sum_{j\in \mathrm{Up}_i} (Z_j^+ f_{ij}^- + Z_j^- f_{ij}^+ + a_{ij}\,(u_j - u_i)) = g_i\,. \tag{27}$$

The following is a preliminary technical result.

LEMMA 9. *Consider any $i \in \{1,\dots,M\}$, and let $u_i \le u_j$ for all $j \in \mathrm{Up}_i$. Then*

$$A_i - P_i^-\, R_i^- + R_i^+ \sum_{j\in \mathrm{Do}_i} a_{ij}\,(u_j - u_i)^- + \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^+\, d_{ij})\,|u_j - u_i| = g_i\,. \tag{28}$$

On the other hand, if $u_i \ge u_j$ for all $j \in \mathrm{Up}_i$, then

$$A_i - P_i^+\, R_i^+ + R_i^- \sum_{j\in \mathrm{Do}_i} a_{ij}\,(u_j - u_i)^+ - \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^-\, d_{ij})\,|u_j - u_i| = g_i\,. \tag{29}$$

Proof. Since $f_{ij}^+ = d_{ij}\,(u_j - u_i)^-$, $f_{ij}^- = d_{ij}\,(u_j - u_i)^+$, and $d_{ij} = -a_{ij}$ if $j \in \mathrm{Do}_i$, the lemma follows immediately from (27). ∎

The following result is a quick consequence of the above lemma, whose implications will become apparent in Corollary 11.

COROLLARY 10. *Consider any* $i \in \{1, \dots, M\}$, *and let* $u_i \le u_j$ *for all* $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$. *Then*

$$A_i + \sum_{j \in \mathrm{Up}_i} (a_{ij} + Z_j^+ \, d_{ij}) \, |u_j - u_i| = g_i \,. \tag{30}$$

On the other hand, if $u_i \ge u_j$ *for all* $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, *then*

$$A_i - \sum_{j \in \mathrm{Up}_i} (a_{ij} + Z_j^- \, d_{ij}) \, |u_j - u_i| = g_i \,. \tag{31}$$

Proof. One has $f_{ij}^+ = 0$ for $j = 1, \dots, N$, and hence $Q_i^- = 0$, which gives $P_i^- \, R_i^- = 0$. Then (30) follows from (28). To prove (31) it is enough to note that $f_{ij}^- = 0$ for $j = 1, \dots, N$, which leads to $Q_i^+ = 0$ and $P_i^+ \, R_i^+ = 0$, and then apply (29). ∎

Finally, the following corollary states that if $g_i \le 0$ (≥ 0), then u_i cannot be a strict *positive (negative)* local maximum (minimum).

COROLLARY 11. *Consider any* $i \in \{1, \dots, M\}$. *Then*

$$g_i \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j \quad \textit{for } u_i \ge 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j^+ \,, \tag{32}$$

$$g_i \ge 0 \quad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j \quad \textit{for } u_i \le 0 \quad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j^- \,. \tag{33}$$

Proof. Let $u_i \ge 0$. Then thanks to (21), $A_i \ge 0$ (where A_i is defined in (26)). If $u_i > u_j$ for all $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, then (31) holds with a positive left-hand side. Thus, if $g_i \le 0$, then $u_i \le u_j$ for some $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, which implies (32). The second statement is proved in an analogous way. ∎

Remark 12. It is worth remarking that, if $\sum_{j=1}^N a_{ij} = 0$, then the previous results can be strengthened since Lemma 9 and Corollary 10 hold with $A_i = 0$. Then Corollary 11 is valid without the restriction on the sign of u_i; i.e., for any $i \in \{1, \dots, M\}$, one has

$$g_i \le 0 \quad \Rightarrow \quad u_i \le \max_{j \ne i, \, a_{ij} \ne 0} u_j \,,$$

$$g_i \ge 0 \quad \Rightarrow \quad u_i \ge \min_{j \ne i, \, a_{ij} \ne 0} u_j \,.$$

This is in accordance with the corresponding results for PDEs (see, e.g., [6]).

6. Variational form of the algebraic flux correction scheme and error estimation. In this section we show how the linear system (1), (2) originates from a variational problem representing a finite element discretization and how, in turn, the nonlinear algebraic problem (13), (14) can be put into a variational form. Then the derivation of an error estimate is discussed. It is important to notice that all of the results of this section, and the following one, are valid for limiters α_{ij} that are only required to belong to $[0, 1]$.

Let $\Omega \subset \mathbb{R}^d$, $d \ge 1$, be a bounded domain and let the boundary $\partial\Omega$ of Ω be Lipschitz-continuous and polyhedral (if $d \ge 2$). Let $a : H^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be a bilinear form, let $u_b \in H^{1/2}(\partial\Omega) \cap C(\partial\Omega)$, let $g \in H^{-1}(\Omega)$, and consider the following variational problem:

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

$$a(u, v) = \langle g, v \rangle \quad \forall \; v \in H_0^1(\Omega) \,. \tag{34}$$

An example of such a variational problem will be presented in the next section.

To solve (34) numerically, let us introduce a finite element space $W_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$ approximating the space $H^1(\Omega)$, and set $V_h := W_h \cap H_0^1(\Omega)$. We denote the basis functions of W_h by $\varphi_1, \ldots, \varphi_N$ and assume that the functions $\varphi_1, \ldots, \varphi_M$ (with $0 < M < N$) form a basis in V_h. In addition, we assume that there are points $x_1, \ldots, x_N \in \overline{\Omega}$ such that $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \ldots, N$, where δ_{ij} is the Kronecker symbol, and that $x_{M+1}, \ldots, x_N \in \partial\Omega$ (note that $x_1, \ldots, x_M \in \Omega$). Since constant functions are always required to be contained in W_h, one has $\sum_{i=1}^N \varphi_i = 1$ in Ω. In what follows, for any $u_h \in W_h$ (or v_h, z_h, etc.), we shall denote by $\{u_i\}_{i=1}^N$ (or $\{v_i\}_{i=1}^N$, $\{z_i\}_{i=1}^N$, etc.) the uniquely determined coefficients with respect to the above basis of W_h, i.e.,

$$u_h = \sum_{i=1}^N u_i \, \varphi_i \qquad \left(\text{or} \quad v_h = \sum_{i=1}^N v_i \, \varphi_i \,, \quad z_h = \sum_{i=1}^N z_i \, \varphi_i \,, \quad \text{etc.} \right) .$$

Of course, $u_i = u_h(x_i)$ (or $v_i = v_h(x_i)$, $z_i = z_h(x_i)$, etc.) for any $i \in \{1, \ldots, N\}$.

It is sometimes convenient (cf. section 7) to approximate the bilinear form a by a bilinear form $a_h : W_h \times V_h \to \mathbb{R}$. We assume that a_h is elliptic on the space V_h; i.e., there is a constant $C_a > 0$ such that

$$a_h(v_h, v_h) \geq C_a \, \|v_h\|_a^2 \quad \forall \; v_h \in V_h \,, \tag{35}$$

where $\|\cdot\|_a$ is a norm on the space $H_0^1(\Omega)$ but generally only a seminorm on the space $H^1(\Omega)$.

Now an approximate solution of the variational problem (34) can be introduced as the solution of the following finite-dimensional problem:

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

$$a_h(u_h, v_h) = \langle g, v_h \rangle \quad \forall \; v_h \in V_h \,. \tag{36}$$

We denote

$$a_{ij} = a_h(\varphi_j, \varphi_i) \,, \qquad i, j = 1, \ldots, N \,, \tag{37}$$

$$g_i = \langle g, \varphi_i \rangle \,, \qquad i = 1, \ldots, M \,, \tag{38}$$

$$u_i^b = u_b(x_i) \,, \qquad i = M+1, \ldots, N \,. \tag{39}$$

Then u_h is a solution of the finite-dimensional problem (36) if and only if it satisfies the relations (1) and (2). Moreover, the matrix $(a_{ij})_{i,j=1}^M$ satisfies (3). We denote

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

with $\alpha_{ij}(w) := \alpha_{ij}(\{w(x_i)\}_{i=1}^N)$. This implies that

$$d_h(w_h; z_h, v_h) = \sum_{i,j=1}^N (1 - \alpha_{ij}(w_h)) \, d_{ij} \, (z_j - z_i) \, v_i \quad \forall \; w_h, z_h, v_h \in W_h \,,$$

and hence we realize that the corresponding flux correction scheme (13), (14) is equivalent to the following variational problem:

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

$$a_h(u_h, v_h) + d_h(u_h; u_h, v_h) = \langle g, v_h \rangle \quad \forall \; v_h \in V_h \,. \tag{40}$$

For any $w \in C(\overline{\Omega})$, the mapping $d_h(w;\cdot,\cdot) : C(\overline{\Omega}) \times C(\overline{\Omega}) \to \mathbb{R}$ is a nonnegative symmetric bilinear form (cf. Lemma 1), and hence it satisfies Schwarz's inequality

$$|d_h(w;z,v)|^2 \le d_h(w;z,z)\, d_h(w;v,v) \quad \forall\ w,z,v \in C(\overline{\Omega})\,. \tag{41}$$

Thus, for any $w \in C(\overline{\Omega})$, the functional $(d_h(w;\cdot,\cdot))^{1/2}$ is a seminorm on $C(\overline{\Omega})$.

Now let $u_h \in W_h$ be a solution of (40), and let us derive an estimate of the error $u - u_h$. A natural norm on V_h corresponding to the left-hand side of (40) is defined by

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

Note that $\|\cdot\|_h$ may be only a seminorm on W_h and that it is not defined on the space $H^1(\Omega)$. We introduce the set

$$W_h^b = \{z_h \in W_h\,;\ z_h(x_i) = u_b(x_i),\ i = M+1,\dots,N\}$$

and consider any $v_h \in V_h$ and $z_h \in W_h^b$. Then, according to (34) and (40), one obtains

$$a_h(u_h - z_h, v_h) + d_h(u_h; u_h - z_h, v_h) = a(u, v_h) - a_h(z_h, v_h) - d_h(u_h; z_h, v_h)\,.$$

Since $u_h - z_h \in V_h$, using (35) and (41) one derives that

$$\|u_h - z_h\|_h \le \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(z_h,v_h)}{\|v_h\|_h} + (d_h(u_h; z_h, z_h))^{1/2}\,.$$

Assuming that $u \in C(\overline{\Omega})$, adding $\|u - z_h\|_h$ to both sides of this estimate and using the triangle inequality, one obtains

$$\|u - u_h\|_h \le \inf_{z_h \in W_h^b} \left\{\|u - z_h\|_h + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(z_h,v_h)}{\|v_h\|_h} + (d_h(u_h; z_h, z_h))^{1/2}\right\}. \tag{42}$$

Let us introduce the Lagrange interpolation operator $i_h : C(\overline{\Omega}) \to W_h$ by

$$i_h v = \sum_{i=1}^{N} v(x_i)\, \varphi_i\,, \qquad v \in C(\overline{\Omega})\,.$$

Then $i_h u \in W_h^b$, and hence using (42) one gets the estimate

$$\|u - u_h\|_h \le C_a^{1/2} \|u - i_h u\|_a + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} + (d_h(u_h; i_h u, i_h u))^{1/2}\,. \tag{43}$$

Thus, as usual, the error of the discrete solution is estimated by an interpolation error and a consistency error. In the following section we estimate these terms for a discretization of a convection-diffusion-reaction equation.

7. Application to a convection-diffusion-reaction equation. Let Ω be as in section 6, and let us consider the steady-state convection-diffusion-reaction equation

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = g \quad \text{in } \Omega\,, \qquad u = u_b \quad \text{on } \partial\Omega\,, \tag{44}$$

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

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

where σ_0 is a constant. The weak solution of (44) satisfies (34) with

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + (c\,u, v) \quad \text{and} \quad \langle g, v\rangle = (g, v)\,,$$

where $(\cdot,\cdot)$ denotes the inner product in $L^2(\Omega)$ or $L^2(\Omega)^d$. It is well known that the weak solution of (44) exists, is unique, and satisfies the maximum principle (cf. [6]).

Let $\mathscr{T}_h$ belong to a regular family of triangulations of Ω consisting of simplices. We consider a space $W_h \subset H^1(\Omega)$ consisting of continuous piecewise linear functions, i.e.,

$$W_h = \{v_h \in C(\overline{\Omega})\,;\ v_h|_T \in \mathbb{P}_1(T)\ \forall\, T \in \mathscr{T}_h\}\,.$$

The points x_i assigned to the basis functions φ_i introduced in the previous section are vertices of the triangulation $\mathscr{T}_h$.

The matrix corresponding to the reaction term $(c\,u_h, v_h)$ in the Galerkin finite element discretization of (44) has only nonnegative entries, which may cause a violation of the condition (6). In order to overcome this, we replace the matrix corresponding to the reaction term by a simple diagonal approximation:

$$(c\,u_h, v_h) = \sum_{i=1}^{M} (c\,u_h, \varphi_i)\,v_i \approx \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \quad \forall\ u_h \in W_h,\ v_h \in V_h\,. \tag{45}$$

This has the extra impact of making the matrix $\mathbb{D}$ independent of c (see below). An alternative diagonal approximation of the reaction matrix can be defined using a low-order nodal quadrature for the reaction term, in which case the estimation of the associated error follows standard approaches (provided that c has a higher regularity than the one assumed so far). The error incurred by the use of (45) is estimated in the next lemma.

LEMMA 13. *There is a constant C independent of h such that*

$$\left|(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i\right| \le C\,h\,\|c\|_{0,\infty,\Omega}\,|u_h|_{1,\Omega}\,\|v_h\|_{0,\Omega}$$

for all $c \in L^\infty(\Omega)$, $u_h \in W_h$, and $v_h \in V_h$.

Proof. Consider any $c \in L^\infty(\Omega)$, $u_h \in W_h$, and $v_h \in V_h$. Then

$$(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i = \sum_{i=1}^{M} (c\,(u_h - u_i), \varphi_i)\,v_i = \sum_{T\in\mathscr{T}_h} \sum_{\substack{i=1\\ x_i\in T}}^{M} (c\,(u_h - u_i), \varphi_i)_T\,v_i$$

$$\le \|c\|_{0,\infty,\Omega} \sum_{T\in\mathscr{T}_h} \sum_{\substack{i=1\\ x_i\in T}}^{M} \|u_h - u_i\|_{0,1,T}\,|v_i|\,.$$

Next, using the Cauchy–Schwarz inequality one obtains

$$\|u_h - u_i\|_{0,1,T} \le |T|^{1/2}\|u_h - u_i\|_{0,T} \le h_T^{d/2}\|\nabla u_h\cdot(x - x_i)\|_{0,T} \le h_T^{1+d/2}|u_h|_{1,T}\,,$$

where $h_T = \operatorname{diam}(T)$. Consequently,

$$(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \le\ h\,\|c\|_{0,\infty,\Omega} \sum_{T\in\mathscr{T}_h} |u_h|_{1,T}\,h_T^{d/2} \sum_{x_i\in T} |v_h(x_i)|\,.$$

Since $h_T^{d/2} \sum_{x_i \in T} |v_h(x_i)| \leq C \|v_h\|_{0,T}$, the lemma follows by applying Hölder's inequality. ∎

Using the approximation (45), the bilinear form a_h in (36) is given by

$$a_h(u_h, v_h) = \varepsilon (\nabla u_h, \nabla v_h) + (\boldsymbol{b} \cdot \nabla u_h, v_h) + \sum_{i=1}^{M} (c, \varphi_i)\, u_i\, v_i \quad \forall\ u_h \in W_h,\ v_h \in V_h$$

and satisfies (35), with

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

and $C_a > 0$ independent of h and the data of (44). The bilinear form a_h defines the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, whose entries are given by (37). The artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is defined using (9), and thus it is independent of c.

Remark 14. It is easy to verify that the matrix $\mathbb{A}$ satisfies (21). The assumption (22) holds if and only if

$$(\nabla \varphi_k, \nabla \varphi_l) \leq 0 \quad \forall\ k, l = 1, \ldots, N,\ k \neq l\,,\ k \leq M,\ \text{or}\ l \leq M\,. \tag{46}$$

The validity of (46) is guaranteed if the triangulation $\mathscr{T}_h$ is weakly acute, i.e., if the angles between faces in $\mathscr{T}_h$ do not exceed $\pi/2$. In the two-dimensional case, it is sufficient for (46) that $\mathscr{T}_h$ is a Delaunay triangulation, i.e., that the sum of any pair of angles opposite a common edge is less than or equal to π.

Now we can discuss the estimation of the terms on the right-hand side of the error estimate (43). To this end, we assume that $u \in H^2(\Omega)$. Then, standard interpolation estimates (cf. [5]) give

$$\|u - i_h u\|_a \leq C\, (\varepsilon + \sigma_0\, h^2)^{1/2}\, h\, |u|_{2,\Omega}\,. \tag{47}$$

The remaining two terms on the right-hand side of (43) will be estimated in the following two lemmas.

LEMMA 15. *Let $\sigma_0 > 0$. Then there is a constant C independent of h and the data of problem* (44) *such that for any $u \in H^2(\Omega)$,*

$$\sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \leq C\, (\varepsilon + \sigma_0^{-1}\, \{\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\})^{1/2}\, h\, \|u\|_{2,\Omega}\,. \tag{48}$$

If $c \equiv 0$, then

$$\sup_{v_h \in V_h} \frac{a(u, v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \leq C\, (\varepsilon + \varepsilon^{-1}\, \|\boldsymbol{b}\|_{0,\infty,\Omega}^2\, h^2)^{1/2}\, h\, |u|_{2,\Omega}\,. \tag{49}$$

Proof. Consider any $u \in H^2(\Omega)$ and $v_h \in V_h$. Then, in view of Lemma 13,

$$\begin{aligned} a(u, v_h) - a_h(i_h u, v_h) &= \varepsilon\, (\nabla(u - i_h u), \nabla v_h) + (\boldsymbol{b} \cdot \nabla(u - i_h u), v_h) \\ &\quad + (c\, (u - i_h u), v_h) + (c\, i_h u, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\, (i_h u)(x_i)\, v_i \\ &\leq C\, (\varepsilon\, |v_h|_{1,\Omega} + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega} + \|c\|_{0,\infty,\Omega}\, \|v_h\|_{0,\Omega})\, h\, \|u\|_{2,\Omega}\,. \end{aligned}$$

Therefore, if $\sigma_0 > 0$, one obtains (48). If $c \equiv 0$, one can employ the fact that

$$(\boldsymbol{b} \cdot \nabla(u - i_h u), v_h) = -(u - i_h u, \boldsymbol{b} \cdot \nabla v_h) \leq C\, h^2\, |u|_{2,\Omega}\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, |v_h|_{1,\Omega}\,,$$

which leads to (49). ∎

Lemma 15 shows that, if $\sigma_0 > 0$, one obtains from (43)

$$\|u - u_h\|_h \leq C\, h\, \|u\|_{2,\Omega} + (d_h(u_h; i_h u, i_h u))^{1/2}\,, \tag{50}$$

where C is independent of u, h, and ε. However, if $c \equiv 0$ (hence $\sigma_0 = 0$), one cannot avoid an explicit negative power of ε in the estimate (49) since the seminorm $(d_h(u_h; v_h, v_h))^{1/2}$ cannot be used for estimating v_h due to the possibly vanishing factors $(1 - \alpha_{ij}(u_h))$. The negative power of ε in (49) is somewhat compensated by the presence of h in the numerator. Still, this estimate can be considered fully satisfactory only if $h \lesssim \varepsilon^{1/2}$.

Finally, let us estimate the last term on the right-hand side of (43).

LEMMA 16. *Let the matrix $\mathbb{D}$ be defined by* (9)*. Then there is a constant C independent of h and the data of problem* (44) *such that*

$$d_h(w_h; i_h u, i_h u) \leq C\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)\, |i_h u|^2_{1,\Omega} \quad \forall\ w_h \in W_h,\ u \in C(\overline{\Omega})\,. \tag{51}$$

Proof. Consider any $i, j \in \{1, \ldots, N\}$ such that $i \neq j$ and $d_{ij} \neq 0$. Then

$$\begin{aligned} |d_{ij}| &\leq \sum_{T \in \mathscr{T}_h,\, x_i, x_j \in T} \left(\varepsilon\, |\varphi_i|_{1,T}\, |\varphi_j|_{1,T} + \|\boldsymbol{b}\|_{0,\infty,T} \left\{ |\varphi_i|_{1,T}\, \|\varphi_j\|_{0,T} + |\varphi_j|_{1,T}\, \|\varphi_i\|_{0,T} \right\}\right) \\ &\leq C \sum_{T \in \mathscr{T}_h,\, x_i, x_j \in T} \left(\varepsilon\, h_T^{d-2} + \|\boldsymbol{b}\|_{0,\infty,T}\, h_T^{d-1}\right) \leq \widetilde{C}\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)\, |x_i - x_j|^{d-2}\,. \end{aligned}$$

Therefore, using Lemma 1, one derives for any $w_h \in W_h$ and $u \in C(\overline{\Omega})$

$$\begin{aligned} d_h(w_h; i_h u, i_h u) &= \sum_{\substack{i,\, j = 1 \\ i < j}}^{N} (1 - \alpha_{ij}(w_h))\, |d_{ij}|\, [u(x_i) - u(x_j)]^2 \\ &\leq \sum_{T \in \mathscr{T}_h} \sum_{x_i, x_j \in T} |d_{ij}|\, [u(x_i) - u(x_j)]^2 \\ &\leq \widetilde{C}\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h) \sum_{T \in \mathscr{T}_h} h_T^{d-2} \sum_{x_i, x_j \in T} [u(x_i) - u(x_j)]^2\,. \end{aligned}$$

Since

$$h_T^{d-2} \sum_{x_i, x_j \in T} [u(x_i) - u(x_j)]^2 \leq C\, |i_h u|^2_{1,T}\,,$$

one obtains the statement of the lemma. ∎

One observes that if $d_h(u_h; i_h u, i_h u)$ in (50) is estimated using Lemma 16, the convergence order is reduced. As a matter of fact, (47), (48), and (51) lead to the following global error estimate.

COROLLARY 17. *Let $u \in H^2(\Omega)$ be the solution of* (44)*, and let u_h be a solution of the discrete problem* (40)*. Then if $\sigma_0 > 0$, there exists a constant $C > 0$ independent of h and the data of* (44) *such that*

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

Remark 18. A careful inspection of the proof of Lemma 16 reveals that the convergence order of the term $d_h(u_h; i_h u, i_h u)$ depends on the relation between ε and $\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$ and on properties of the triangulations $\mathscr{T}_h$. For simplicity, the discussion will be restricted to the two-dimensional case, but the same arguments are valid (with minor modifications) in the higher-dimensional case. We distinguish the following cases:

- *convection-dominated regime* ($\varepsilon < \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$): the estimate (51) reduces to

$$d_h(w_h; i_h u, i_h u) \le C\, \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |i_h u|^2_{1,\Omega} \quad \forall\ w_h \in W_h,\ u \in C(\overline{\Omega})\,. \tag{52}$$

This estimate implies an $\mathcal{O}(\sqrt{h})$ error estimate in (50), which will be confirmed by numerical experiments in section 8 for a particular choice of the coefficients α_{ij}.

- *diffusion-dominated regime* ($\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h$). In this case, the estimate (51) reduces to

$$d_h(w_h; i_h u, i_h u) \le C\, \varepsilon\, |i_h u|^2_{1,\Omega} \quad \forall\ w_h \in W_h,\ u \in C(\overline{\Omega})\,, \tag{53}$$

which does not imply any convergence of $\|u - u_h\|_h$. However, this result can be improved for suitable types of meshes. To characterize the geometry of a triangulation $\mathscr{T}_h$, we introduce a quantity θ_{ij} for any edge E_{ij} with end points x_i, x_j. If $E_{ij} \subset \partial\Omega$, then θ_{ij} is the angle opposite E_{ij}. If $E_{ij} \not\subset \partial\Omega$, then θ_{ij} is the average of the pair of angles opposite E_{ij}. Finally, we denote by θ_h the maximum of all θ_{ij}. Then we consider the following values of θ_h:

(a) $\theta_h \le \pi/2$, i.e., $\mathscr{T}_h$ is a *Delaunay triangulation* (in particular, $\mathscr{T}_h$ may consist of *weakly acute triangles*, i.e., with all angles $\le \pi/2$). Then the off-diagonal entries of the diffusion matrix are all nonpositive, and hence $|d_{ij}| \le \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h/3$ for $i \ne j$. Thus, the estimate (52) is again valid and leads to an $\mathcal{O}(\sqrt{h})$ in estimate (50).

(b) $\theta_h < \pi/2$, a particular case of (a), satisfied, e.g., for $\mathscr{T}_h$ consisting of *acute triangles* (all angles $< \pi/2$). Then all off-diagonal entries of the diffusion matrix are negative, and hence all off-diagonal entries of the matrix $\mathbb{A}$ are nonpositive in the strongly diffusion-dominated case (precisely, if $\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, (\tan\theta_h)/3$). In this case, all entries of the artificial diffusion matrix $\mathbb{D}$ vanish, and hence the AFC method (40) reduces to the original linear method (36). Consequently, the standard $\mathcal{O}(h)$ error estimate of $\|u - u_h\|_h$ is valid.

(c) $\theta_h = \pi/2$, again a particular case of (a) which may happen, e.g., if $\mathscr{T}_h$ consists of right-angled triangles. Then some off-diagonal entries of the diffusion matrix vanish, and hence the corresponding entries d_{ij} do not vanish in general. Thus, if $\theta_h = \pi/2$ for all $\mathscr{T}_h$ in the family of triangulations, then, in contrast to the previous case, the AFC method (40) does not reduce to the original linear method (36) for $h \to 0$.

(d) $\theta_h > \pi/2$, i.e., $\mathscr{T}_h$ is *not of Delaunay type*, which implies that $\mathscr{T}_h$ contains *obtuse triangles* (with an angle $> \pi/2$). In this case, some off-diagonal entries of the diffusion matrix are positive, and hence the estimate (53) cannot be improved in general. Indeed, if $\theta_{ij} > \pi/2$ and $\varepsilon \ge \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |\tan\theta_{ij}|$, then $|d_{ij}| \ge \varepsilon\, |\cot\theta_{ij}|/3$. Thus, if the mesh is not of Delaunay type, the results presented in this work do not prove convergence of the method, which will be also confirmed by numerical experiments presented in section 8. Note also that, in this case, the results of section 5 are not valid for the AFC scheme considered in this section.

It is worth remarking that these last results are the best that can be obtained using the general approach described in the previous sections, combined with the choice for limiters α_{ij} from section 4. As a matter of fact, the algebraic construction of the method has been carried out using a rather general splitting of the stiffness matrix. Now, for the convection-diffusion equation, the lack of convergence of the method for non-Delaunay meshes can be overcome by changing the way the matrix $\mathbb{D}$ is built. In fact, if instead of using the whole stiffness matrix to build $\mathbb{D}$, we use only the convection matrix to build it, that is,

$$d_{ij} = -\max\{(\boldsymbol{b}\cdot\nabla\varphi_j,\varphi_i), 0, (\boldsymbol{b}\cdot\nabla\varphi_i,\varphi_j)\} \quad \forall\, i\neq j\,; \tag{54}$$

then the estimate (51) in Lemma 16 becomes

$$d_h(w_h; i_h u, i_h u) \leq C\,\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |i_h u|^2_{1,\Omega} \quad \forall\, w_h\in W_h,\, u\in C(\overline{\Omega})\,.$$

This leads to an $\mathcal{O}(\sqrt{h})$ estimate of $\|u-u_h\|_h$, even on non-Delaunay meshes in the diffusion-dominated regime. An alternative way to solve this would be to change the definition of the limiters α_{ij} to make them more suitable for diffusion problems. Examples of limiters suitable for diffusion problems can be found, e.g., in [8, 19], but their applicability to convection-dominated problems has yet to be explored.

We finally mention that numerical results in section 8 indicate that the estimates of $d_h(w_h; i_h u, i_h u)$ discussed above are sharp. Note, however, that the only properties of the coefficients α_{ij} used in the proof of Lemma 16 were the fact that their values are from the interval $[0,1]$ and that $\alpha_{ij}=\alpha_{ji}$. If the coefficients α_{ij} are defined as in section 4, then in the convection-dominated regime, better convergence rates are observed than those predicted by estimate (52). Some deeper analysis of this choice of α_{ij} might lead to an improved estimate of $d_h(w_h; i_h u, i_h u)$ in the convection-dominated case.

Remark 19. We finish this section by making some comments on the stability of the nonlinear discretization (40) with W_h defined in section 7. Our objective is to show that this formulation can be viewed as a way of adding numerical diffusion to the Galerkin discretization. We restrict our discussion to the two-dimensional case, but the results can be extended to three space dimensions. First, given $u_h\in W_h$, we divide the triangulation $\mathscr{T}_h$ as $\mathscr{T}_h=\mathscr{T}_1\cup\mathscr{T}_2$, where $\mathscr{T}_1$ and $\mathscr{T}_2$ are disjoint and $T\in\mathscr{T}_1$ if and only if for at least two edges of T we have $(1-\alpha_{ij}(u_h))|d_{ij}|>0$. We will denote by α_T the minimum value of these nonzero quantities. Typically, T will belong to $\mathscr{T}_1$ if there is an extremum of u_h in a vertex of T or if u_h has a layer through T. Then from the proof of Lemma 1, and using a scaling argument, it is not difficult to realize that for any $v_h\in W_h$,

$$\begin{aligned} d_h(u_h; v_h, v_h) &= \frac{1}{2}\sum_{i,j=1}^{N} (1-\alpha_{ij}(u_h))\,|d_{ij}|\,(v_i-v_j)^2 \\ &\geq \frac{1}{12}\sum_{T\in\mathscr{T}_1}\sum_{x_i,x_j\in T}\alpha_T\,(v_i-v_j)^2 \geq C\sum_{T\in\mathscr{T}_1}\alpha_T\,|v_h|^2_{1,T}\,. \end{aligned}$$

Note that for simplicity, we used the inequality

$$(v_i-v_j)^2+(v_j-v_k)^2 \geq \frac{1}{3}\left((v_i-v_j)^2+(v_j-v_k)^2+(v_k-v_i)^2\right) \quad \forall\, i,j,k\,.$$

Then, AFC methods add numerical diffusion on certain elements of the triangulation, namely, the elements which contain extrema of the discrete solution or lie in its layer regions.

In addition, we can also compare this last result with a parameter-free stabilized method proposed in [3]. That method is based on rewriting the gradient of the $\mathbb{P}_1$ basis functions in terms of the Nédélec edge FEM. More precisely, the stabilization term added to the Galerkin formulation in [3] reads as follows:

$$Q(u_h, v_h) = (\Theta_h(u_h), \Theta_h(v_h))\,, \tag{55}$$

with

$$\Theta_h(u_h) = \sum_{E\in\mathcal{E}_h} \tilde{\theta}_E\,(u_h(x_{E1}) - u_h(x_{E2}))\,\boldsymbol{N}_E\,, \tag{56}$$

where $\mathcal{E}_h$ stands for the set of edges of the triangulation $\mathscr{T}_h$, x_{E1}, x_{E2} are the end points of an edge E, and $\boldsymbol{N}_E$ stands for the basis function of the Nédélec space associated to E. In (56), $\tilde{\theta}_E$ is a positive parameter depending on the edge Péclet number (for details, see [3, eqs. (2.14) and (2.10)]). With these definitions, the term defined in (55) satisfies

$$\begin{aligned} Q(u_h, u_h) &= \sum_{E,E'\in\mathcal{E}_h} \tilde{\theta}_E\,\tilde{\theta}_{E'}\,(u_h(x_{E1}) - u_h(x_{E2}))\,(u_h(x_{E'1}) - u_h(x_{E'2}))\,(\boldsymbol{N}_E, \boldsymbol{N}_{E'}) \\ &\approx \sum_{E\in\mathcal{E}_h} |E|^{d-2}\big(\tilde{\theta}_E\,(u_h(x_{E1}) - u_h(x_{E2}))\big)^2\,, \end{aligned}$$

where by $\approx$ we mean that both terms bound each other with constants that do not depend on h. Then we see that the method from [3] can be seen as well as a "linearized" version of (40) (where we choose α_{ij} in such a way that $(1-\alpha_{ij}(u_h))|d_{ij}| = \tilde{\theta}_E^2\,|E|^{d-2}$ for every edge E). This also explains the fact that only $\mathcal{O}(\sqrt{h})$ convergence has been obtained in Table 1 (where we choose $\alpha_{ij}(u_h) = 0.5$ for every edge). As a matter of fact, that was the order of convergence proven in [3].

8. Numerical results. This section presents numerical results obtained with the AFC scheme applied to the convection-diffusion-reaction equation (44). For the sake of brevity, the presentation is restricted to studies of the convergence of the method for the following example with smooth solution. Results for an example with layers can be found, e.g., in [1].

Example 20. Problem (44) is considered with $\Omega = (0,1)^2$, with different values of ε, and with $\boldsymbol{b} = (3,2)^T$, $c = 1$, $u_b = 0$, and the right-hand side g chosen such that

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

is the solution of (44).

In the numerical simulations, $\mathbb{P}_1$ finite elements were used on triangular grids. Mass lumping (cf. (45)) was performed for the reactive term, but only very small differences could be observed compared to results obtained without mass lumping. If x_i is a Dirichlet node, we set $R_i^+ := 1$, $R_i^- := 1$, leading to $\alpha_{ij} = 1$ if $a_{ji} \le a_{ij}$; see section 4. Concerning the errors in $\|\cdot\|_h$, qualitatively the same results were obtained with and without this definition. However, the errors in other norms of interest were sometimes clearly smaller with this definition, and we decided to present these better

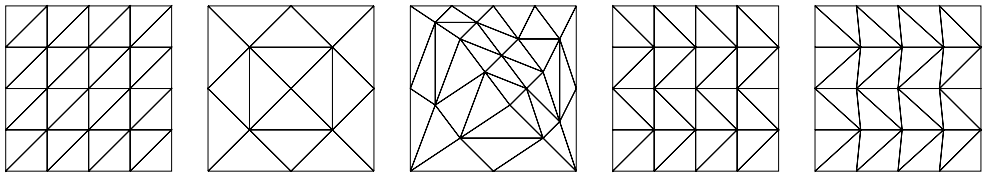

FIG. 1. *Grids* 1–5 *(left to right), level* 0. *The differences between grid* 4 *and grid* 5 *are described in the text.*

results. The nonlinear discrete equations were solved with a fixed-point iteration with Anderson acceleration [21]. The iterations were stopped when the Euclidean norm of the residual vector was smaller than 10^{-9}. All simulations were double-checked by computing them with two different codes, one of which was MooNMD [9].

Simulations were performed on several structured and unstructured grids; see Figure 1 for the coarsest grids (level 0). Grids 1, 2, and 3 were refined uniformly. Grid 4 was obtained from grid 1 by changing the directions of the diagonals in even rows of squares (from below). Grid 5 was obtained from grid 4 by shifting interior nodes to the right by a tenth of the horizontal mesh width on each even horizontal mesh line. Therefore, for any diagonal edge E_{ij} of grid 5, the value θ_{ij} introduced in Remark 18 satisfies $\theta_{ij} > \pi/2$.

Considering a problem without reaction, i.e., with $c = 0$ instead of $c = 1$, and otherwise the same setup, one obtains qualitatively the same results as below. For the sake of brevity, we omit the results for $c = 0$.

8.1. Constant weights α_{ij}. The case of constant weights $\alpha_{ij} = 0.5$ (with the modification at Dirichlet nodes mentioned above) fits into the presented error analysis. Fixing the weights independently of the approximate solution u_h replaces the nonlinear problem (13), (14) by a linear problem, which is essentially a stabilized method adding first-order artificial diffusion to the original problem (1), (2). Then, some suboptimal convergence results are to be expected. Table 1 shows numerical results obtained in the convection-dominated regime for grid 1. In the first row of the table, we use the following notation: l is the grid level, $e_h = u - u_h$, $d_h^{1/2}(u_h) = d_h(u_h; i_h u, i_h u)^{1/2}$, and "ord." denotes experimental convergence orders computed from values in the preceding column. The results in Table 1 indicate that the estimate (52) of $d_h(w_h; i_h u, i_h u)$ and also the estimate for $\|u - u_h\|_h$ given in Corollary 17 are sharp.

TABLE 1
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 5 *and constant weights* α_{ij}.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.622e−2	0.66	7.668e−1	0.11	2.722e−1	0.43	9.666e−2	0.57
4	1.527e−2	0.78	7.021e−1	0.13	1.975e−1	0.46	6.397e−2	0.60
5	8.260e−3	0.89	6.489e−1	0.11	1.415e−1	0.48	4.274e−2	0.58
6	4.295e−3	0.94	6.149e−1	0.08	1.008e−1	0.49	2.912e−2	0.55
7	2.189e−3	0.97	5.956e−1	0.05	7.150e−2	0.50	2.015e−2	0.53
8	1.105e−3	0.99	5.854e−1	0.02	5.065e−2	0.50	1.408e−2	0.52

8.2. Weights computed with the algorithm from section 4. As already mentioned, the computation of the weights as presented in section 4 is a standard choice in practice. For the convection-dominated regime, numerical results are

presented in Tables 2–6. It can be observed that the order of convergence of $\|u-u_h\|_h$ is around two on grid 1 and around one for all other simulations. The errors $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ behave differently on different grids. For grid 1, which is of Friedrichs–Keller type (it consists of three sets of parallel lines), one can see the optimal order of convergence for $\|u-u_h\|_{0,\Omega}$ and also the convergence of $|u-u_h|_{1,\Omega}$ is almost optimal. For grids 2–5, the orders of convergence of $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ are clearly smaller than the optimal order. Moreover, for grids 4 and 5, the convergence order of $|u-u_h|_{1,\Omega}$ tends to zero for $h \to 0$.

TABLE 2
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 1 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	5.457e−3	1.85	2.287e−1	1.10	1.112e−1	0.97	1.163e−2	2.11
4	1.408e−3	1.95	1.074e−1	1.09	5.317e−2	1.06	2.683e−3	2.12
5	3.493e−4	2.01	5.113e−2	1.07	2.472e−2	1.11	6.410e−4	2.07
6	8.652e−5	2.01	2.546e−2	1.01	1.158e−2	1.09	1.633e−4	1.97
7	2.152e−5	2.01	1.321e−2	0.95	5.533e−3	1.07	4.099e−5	1.99
8	5.357e−6	2.01	6.822e−3	0.95	2.685e−3	1.04	1.018e−5	2.01

TABLE 3
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 2 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	8.533e−3	1.86	2.901e−1	1.00	1.236e−1	1.03	1.855e−2	1.91
4	2.516e−3	1.76	1.954e−1	0.57	5.884e−2	1.07	6.065e−3	1.61
5	8.369e−4	1.59	1.380e−1	0.50	2.801e−2	1.07	2.640e−3	1.20
6	2.891e−4	1.53	1.031e−1	0.42	1.356e−2	1.05	1.254e−3	1.07
7	1.103e−4	1.39	7.865e−2	0.39	6.638e−3	1.03	5.938e−4	1.08
8	4.136e−5	1.42	6.524e−2	0.27	3.263e−3	1.02	2.924e−4	1.02
9	1.539e−5	1.43	5.768e−2	0.18	1.618e−3	1.01	1.436e−4	1.03

In summary, in the convection-dominated regime, the numerical studies for the choice of the weights as presented in section 4 show a higher order of error reduction than in the worst case which was considered in the analysis. The difference with respect to the numerical studies of section 8.1 is the behavior of the weights. They do not stay constant but converge in the mean to 1; see Table 7 which shows a representative result for the arithmetic mean value of $\{1-\alpha_{ij}(u_h)\}$. This indicates that the estimate $1-\alpha_{ij}(u_h) \le 1$ used in the proof of Lemma 16 is too rough in some cases.

For the diffusion-dominated regime, numerical results are presented in Tables 8–10. For grid 1, the convergence orders of $\|u-u_h\|_{0,\Omega}$ and $|u-u_h|_{1,\Omega}$ are again optimal, but for grid 4 only $|u-u_h|_{1,\Omega}$ is still optimal, whereas $d_h(u_h; i_h u, i_h u)^{1/2}$ converges with the order 1/2. For grid 5, no convergence is observed. The observations with respect to convergence orders of $d_h(u_h; i_h u, i_h u)^{1/2}$ on grids 4 and 5 are in accordance with the discussion in Remark 18. If the matrix $\mathbb{D}$ is defined using the convection matrix only (i.e., by (54)), then on grids 1 and 4 the results qualitatively do not change, whereas on grid 5, we observe an analogous behavior as on grid 4; see Table 11.

TABLE 4
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 3 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.125e−3	1.61	3.202e−1	0.71	9.189e−2	1.05	1.569e−2	1.81
4	2.216e−3	1.47	2.244e−1	0.51	4.488e−2	1.03	6.502e−3	1.27
5	9.946e−4	1.16	1.821e−1	0.30	2.224e−2	1.01	3.376e−3	0.95
6	4.993e−4	0.99	1.559e−1	0.22	1.124e−2	0.98	1.802e−3	0.91
7	2.519e−4	0.99	1.375e−1	0.18	5.676e−3	0.98	9.649e−4	0.90
8	1.277e−4	0.98	1.231e−1	0.16	2.871e−3	0.98	5.099e−4	0.92

TABLE 5
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 4 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.383e−3	1.70	4.826e−1	0.31	9.814e−2	1.06	2.143e−2	1.45
4	2.313e−3	1.46	4.543e−1	0.09	4.341e−2	1.18	9.455e−3	1.18
5	1.089e−3	1.09	4.434e−1	0.03	1.830e−2	1.25	4.469e−3	1.08
6	5.527e−4	0.98	4.361e−1	0.02	8.276e−3	1.14	2.176e−3	1.04
7	2.817e−4	0.97	4.320e−1	0.01	3.926e−3	1.08	1.077e−3	1.01
8	1.425e−4	0.98	4.297e−1	0.01	1.915e−3	1.04	5.381e−4	1.00

TABLE 6
Example 20, $\varepsilon = 10^{-8}$, *numerical results for grid* 5 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	6.925e−3	1.66	5.638e−1	0.25	9.992e−2	1.06	2.486e−2	1.37
4	2.687e−3	1.37	5.395e−1	0.06	4.405e−2	1.18	1.140e−2	1.12
5	1.304e−3	1.04	5.294e−1	0.03	1.896e−2	1.22	5.491e−3	1.05
6	6.645e−4	0.97	5.225e−1	0.02	8.792e−3	1.11	2.711e−3	1.02
7	3.382e−4	0.97	5.186e−1	0.01	4.235e−3	1.05	1.349e−3	1.01
8	1.708e−4	0.99	5.164e−1	0.01	2.083e−3	1.02	6.755e−4	1.00

TABLE 7
Example 20, $\varepsilon = 10^{-8}$, *grid* 1, *arithmetic mean of* $\{1 - \alpha_{ij}(u_h)\}$ *with* α_{ij} *from section* 4.

Level	3	4	5	6	7	8
$1 - \overline{\alpha}(u_h)$	1.09e−1	5.94e−2	3.16e−2	1.73e−2	9.60e−3	5.27e−3
Order	0.83	0.87	0.91	0.87	0.85	0.87

9. Summary and outlook. An algebraic flux correction (AFC) scheme applied to linear boundary value problems was analyzed. The existence of a solution, existence and uniqueness of a solution of a linearized problem, and an a priori error estimate were proved under rather general assumptions on the limiters α_{ij}. To the best of our knowledge, this is the first time that convergence analysis of an AFC scheme was performed. For a practical choice of the limiters, a local discrete maximum principle was proved. The theory for the abstract problem was applied to steady-state convection-diffusion-reaction equations, where in particular an error estimate was derived. Numerical studies showed that this estimate is sharp for the general assumptions on the limiters used in the analysis. Using the standard limiters, a higher order of convergence was observed than predicted.

As a next step we intend to specialize the convergence results to the standard limiters. This step requires an analysis of the algorithm presented in section 4, which seems to be intricate due to the dependency of the limiters on the solution of the discrete problem. From the numerical aspect, the observed dependency of errors in

TABLE 8
Example 20, $\varepsilon = 10$, *numerical results for grid* 1 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.148e−3	1.98	1.757e−1	0.99	1.144e−1	1.00	5.557e−1	0.99
4	5.379e−4	2.00	8.799e−2	1.00	5.643e−2	1.02	2.783e−1	1.00
5	1.345e−4	2.00	4.401e−2	1.00	2.792e−2	1.02	1.392e−1	1.00
6	3.360e−5	2.00	2.201e−2	1.00	1.387e−2	1.01	6.960e−2	1.00
7	8.398e−6	2.00	1.100e−2	1.00	6.912e−3	1.00	3.480e−2	1.00

TABLE 9
Example 20, $\varepsilon = 10$, *numerical results for grid* 4 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.187e−3	1.89	1.756e−1	0.99	1.983e−1	0.37	5.554e−1	0.99
4	6.209e−4	1.82	8.800e−2	1.00	1.473e−1	0.43	2.783e−1	1.00
5	1.940e−4	1.68	4.402e−2	1.00	1.069e−1	0.46	1.392e−1	1.00
6	6.899e−5	1.49	2.201e−2	1.00	7.657e−2	0.48	6.961e−2	1.00
7	2.789e−5	1.31	1.101e−2	1.00	5.450e−2	0.49	3.481e−2	1.00
8	1.239e−5	1.17	5.503e−3	1.00	3.867e−2	0.50	1.740e−2	1.00

TABLE 10
Example 20, $\varepsilon = 10$, *numerical results for grid* 5 *and* α_{ij} *from section* 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	1.248e−2	0.48	2.229e−1	0.79	1.317e+0	-0.03	7.211e−1	0.77
4	1.123e−2	0.15	1.558e−1	0.52	1.316e+0	0.00	5.135e−1	0.49
5	1.090e−2	0.04	1.333e−1	0.22	1.313e+0	0.00	4.452e−1	0.21
6	1.080e−2	0.01	1.269e−1	0.07	1.312e+0	0.00	4.259e−1	0.06
7	1.077e−2	0.00	1.252e−1	0.02	1.311e+0	0.00	4.207e−1	0.02
8	1.076e−2	0.00	1.248e−1	0.00	1.310e+0	0.00	4.193e−1	0.00

TABLE 11
Example 20, $\varepsilon = 10$, *numerical results for grid* 5, α_{ij} *from section* 4, *and* d_{ij} *defined by* (54) *instead of* (9).

l	$\|e_h\|_{0,\Omega}$	ord.	$\|e_h\|_{1,\Omega}$	ord.	$d_h^{1/2}(u_h)$	ord.	$\|e_h\|_h$	ord.
3	2.319e−3	1.94	1.849e−1	0.98	1.581e−1	0.74	5.846e−1	0.98
4	6.098e−4	1.93	9.275e−2	1.00	1.040e−1	0.60	2.933e−1	1.00
5	1.676e−4	1.86	4.642e−2	1.00	7.244e−2	0.52	1.468e−1	1.00
6	4.979e−5	1.75	2.322e−2	1.00	5.105e−2	0.50	7.343e−2	1.00
7	1.659e−5	1.59	1.161e−2	1.00	3.607e−2	0.50	3.672e−2	1.00
8	6.302e−6	1.40	5.806e−3	1.00	2.550e−2	0.50	1.836e−2	1.00

standard norms on the concrete grid is remarkable. Comprehensive numerical studies that clarify which types of grids should be used and which types should be avoided are necessary, and this will be the subject of future research.

Appendix. For completeness, we report the proofs of some classical results on the relation between M-matrices and discrete maximum principles.

LEMMA 21. *Let us consider a matrix* $(a_{ij})_{j=1,\dots,N}^{i=1,\dots,M}$ *with* $0 < M < N$, *and let* $a_{ii} > 0$ *for* $i = 1, \dots, M$. *Then* (5) *holds for any* $u_1, \dots, u_N \in \mathbb{R}$ *if and only if the conditions* (6) *and* (8) *are satisfied.*

Proof. Let us assume that at least one of the conditions (6) and (8) is not valid. We will construct a counterexample to the validity of (5). If (6) does not hold, i.e., if

$a_{ik} > 0$ for some $i \in \{1, \dots, M\}$ and $k \in \{1, \dots, N\}$, $k \neq i$, then we set

$$u_i = 1, \qquad u_k = -\frac{a_{ii}}{a_{ik}}, \qquad u_j = 0 \quad \forall\, j \in \{1, \dots, N\},\, j \neq i, k\,.$$

Then $u_k < 0$, and hence $\max\{u_j^+\,;\, j \neq i,\, a_{ij} \neq 0\} = 0 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = a_{ii}\, u_i + a_{ik}\, u_k = 0$ so that (5) does not hold. If (8) is not valid, i.e., if $\sum_{j=1}^N a_{ij} < 0$ for some $i \in \{1, \dots, M\}$, then we set

$$u_i = 1 - \frac{1}{a_{ii}} \sum_{j=1}^N a_{ij}, \qquad u_j = 1 \quad \forall\, j \in \{1, \dots, N\},\, j \neq i\,.$$

Then $\max\{u_j^+\,;\, j \neq i,\, a_{ij} \neq 0\} = 1 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = \sum_{j=1}^N a_{ij} + a_{ii}\,(u_i - 1) = 0$ so that again (5) does not hold. This proves that the validity of (5) for any $u_1, \dots, u_N \in \mathbb{R}$ implies (6) and (8).

Now let us assume that the conditions (6) and (8) are satisfied. Consider any $i \in \{1, \dots, M\}$ and any $u_1, \dots, u_N \in \mathbb{R}$ such that $\sum_{j=1}^N a_{ij}\, u_j \leq 0$. Setting

$$c := \max_{j \neq i,\, a_{ij} \neq 0} u_j^+\,,$$

one has

$$\begin{aligned} a_{ii}\, u_i \leq \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\, u_j &= \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\,(u_j - c) + \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij})\, c \\ &\leq c \sum_{\substack{j\,=\,1\\ j\,\neq\,i}}^N (-a_{ij}) \leq c\, a_{ii}\,, \end{aligned} \tag{57}$$

which implies that $u_i \leq c$. ∎

LEMMA 22. *Let us consider a matrix $(a_{ij})_{j=1,\dots,N}^{i=1,\dots,M}$ with $0 < M < N$, and let $a_{ii} > 0$ for $i = 1, \dots, M$. Then* (4) *holds for any $u_1, \dots, u_N \in \mathbb{R}$ if and only if the conditions* (6) *and* (7) *are satisfied.*

Proof. Let us assume that at least one of the conditions (6) and (7) is not valid. Since the counterexamples from the proof of Lemma 21 can be used also here, it suffices to consider the case when $\sum_{j=1}^N a_{ij} > 0$ for some $i \in \{1, \dots, M\}$. We set

$$u_i = -1 + \frac{1}{a_{ii}} \sum_{j=1}^N a_{ij}, \qquad u_j = -1 \quad \forall\, j \in \{1, \dots, N\},\, j \neq i\,.$$

Then $\max\{u_j\,;\, j \neq i,\, a_{ij} \neq 0\} = -1 < u_i$, whereas $\sum_{j=1}^N a_{ij}\, u_j = -\sum_{j=1}^N a_{ij} + a_{ii}\,(u_i + 1) = 0$ so that (4) does not hold. This proves that the validity of (4) for any $u_1, \dots, u_N \in \mathbb{R}$ implies (6) and (7).

Now let us assume that the conditions (6) and (7) are satisfied. Consider any $i \in \{1, \dots, M\}$ and any $u_1, \dots, u_N \in \mathbb{R}$ such that $\sum_{j=1}^N a_{ij}\, u_j \leq 0$. Setting

$$c := \max_{j \neq i,\, a_{ij} \neq 0} u_j\,,$$

statement (57) remains valid (the last $\leq$ can be changed to $=$), and hence $u_i \leq c$. ∎

REFERENCES

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

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

[3] P. Bochev, M. Perego, and K. Peterson, *Formulation and analysis of a parameter-free stabilized finite element method*, SIAM J. Numer. Anal., 53 (2015), pp. 2363–2388, doi:10.1137/14096284X.

[4] R. Bordás, V. John, E. Schmeyer, and D. Thévenin, *Numerical methods for the simulation of a coalescence-driven droplet size distribution*, Theor. Comput. Fluid Dyn., 27 (2013), pp. 253–271.

[5] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, Springer-Verlag, New York, 2004.

[6] D. Gilbarg and N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order*, 2nd ed., Grundlehren Math. Wiss. [Fundamental Principles of Mathematical Sciences] 224, Springer-Verlag, Berlin, 1983.

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

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

[9] V. John and G. Matthies, *MooNMD—a program package based on mapped finite element methods*, Comput. Vis. Sci., 6 (2004), pp. 163–169.

[10] V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska, and A. Voigt, *Simulations of population balance systems with one internal coordinate using finite element methods*, Chem. Engrg. Sci., 64 (2009), pp. 733–741.

[11] V. John and M. Roland, *On the impact of the scheme for solving the higher dimensional equation in coupled population balance systems*, Internat. J. Numer. Methods Engrg., 82 (2010), pp. 1450–1474.

[12] D. Kuzmin, *Private communication.*

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

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

[15] D. Kuzmin, *On the design of algebraic flux correction schemes for quadratic finite elements*, J. Comput. Appl. Math., 218 (2008), pp. 79–87.

[16] D. Kuzmin, *Explicit and implicit FEM-FCT algorithms with flux linearization*, J. Comput. Phys., 228 (2009), pp. 2517–2534.

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

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

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

[20] R. Temam, *Navier-Stokes Equations. Theory and Numerical Analysis*, North–Holland, Amsterdam, 1977.

[21] H. F. Walker and P. Ni, *Anderson acceleration for fixed-point iterations*, SIAM J. Numer. Anal., 49 (2011), pp. 1715–1735, doi:10.1137/10078356X.

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