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

Preprint ISSN 2198-5855

Analysis of algebraic flux correction schemes

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

submitted: May 12, 2015

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

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

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

No. 2107
Berlin 2015

2010 *Mathematics Subject Classification.* 65N12 , 65N30.

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

This work has been funded by the Leverhulme Trust under grant RPG-2012-483. The work of P. Knobloch has been partially supported through the grant No.13-00522S of the Czech Science Foundation.

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

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

Abstract. A family of algebraic flux correction 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.

1. Introduction. Many processes from nature and industry can be modelled using (systems of) partial differential equations. 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 in modifying the Galerkin FEM in such a way that the effect of small scales is taken into account already 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 [17, 15, 16], is performed for transport equations and they are called flux-corrected transport (FCT) schemes (see also [6] 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 [3, 9], where the convection–diffusion equations are part of population balance systems. In [9] it is explicitly emphasized that the FCT scheme was preferred to 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 [8]). 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 methods only acts 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 partial differential equation 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, up to our best knowledge, the work [13]).

This paper analyzes algebraic stabilizations for linear steady-state boundary value problems. These methods are called algebraic flux correction (AFC) schemes. Apart from obvious properties of these methods, which are the basis of their construction, there has not been any numerical analysis until very recently. The first contribution in this field is [2], where some preliminary results on the analysis of an AFC scheme (cf. [12]) 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 into 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 having introduced 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. Also, even if the AFC family of methods is built to preserve the discrete maximum principle, we have not been able to find a general proof of this fact in the steady-state case. Then, we give a general proof for this property in Section 5. 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 finite element methods so far. 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 equations 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 in the convection-dominated case. The orders of convergence for standard norms depend on the concrete grid and they are sometimes suboptimal. Finally, 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 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 $\mathrm{U} \equiv (u_1, \dots, u_N)$ satisfies a system of linear equations of the form

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

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

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

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

It is natural to require that the maximum principle also holds for the discrete problem (2.1), (2.2). Due to (2.3), 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

$$\forall\ i \in \{1, \dots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \neq i,\, a_{ij} \neq 0} u_j\,, \tag{2.4}$$

or, at least,

$$\forall\ i \in \{1, \dots, M\}: \qquad \sum_{j=1}^{N} a_{ij}\, u_j \le 0 \quad \Rightarrow \quad u_i \le \max_{j \neq i,\, a_{ij} \neq 0} u_j^+\,, \tag{2.5}$$

where $u_j^+ = \max\{0, u_j\}$. It can be shown (cf. Appendix A below), that (2.4) holds if and only if

$$a_{ij} \le 0 \qquad \forall\ i \neq j,\ i = 1, \dots, M,\ j = 1, \dots, N\,, \tag{2.6}$$

and

$$\sum_{j=1}^{N} a_{ij} = 0\,, \qquad i = 1, \dots, M\,. \tag{2.7}$$

The discrete maximum principle (2.5) holds if and only if (2.6) is satisfied and

$$\sum_{j=1}^{N} a_{ij} \geq 0\,, \qquad i = 1, \dots, M\,. \tag{2.8}$$

While the conditions (2.7) or (2.8) are often satisfied, the property (2.6) does not hold for many discretizations, in particular, of convection-dominated problems. The aim of the algebraic flux correction method is to modify the algebraic system (2.1) 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 algebraic flux correction 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

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall\, i \neq j\,, \qquad\qquad d_{ii} = -\sum_{j\neq i} d_{ij}\,. \tag{2.9}$$

Then the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ satisfies the necessary conditions for the discrete maximum principle provided that (2.7) or (2.8) holds for the matrix $\mathbb{A}$.

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

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + (\mathbb{D}\,\mathrm{U})_i\,, \qquad i = 1, \dots, M\,. \tag{2.10}$$

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

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

where $f_{ij} = d_{ij}\,(u_j - u_i)$. Clearly, $f_{ij} = -f_{ji}$ for all $i, j = 1, \dots, N$. Now the idea of the algebraic flux correction schemes is to limit those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (2.1) (or, equivalently, (2.10)) is replaced by

$$(\tilde{\mathbb{A}}\,\mathrm{U})_i = g_i + \sum_{j\neq i} \alpha_{ij}\, f_{ij}\,, \qquad i = 1, \dots, M\,, \tag{2.11}$$

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. They can be chosen in various ways but their definition is always based on the above fluxes f_{ij}, see [11, 12, 14, 15, 16] for examples. To guarantee that the resulting scheme is conservative, one should require that the coefficients α_{ij} are symmetric, i.e.,

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i, j = 1, \dots, N\,. \tag{2.12}$$

Rewriting the equation (2.11) using the definition of the matrix $\tilde{\mathbb{A}}$, one obtains the final form of the algebraic flux correction scheme to be investigated in this paper. It is the following system of nonlinear equations:

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1 - \alpha_{ij})\, d_{ij}\, (u_j - u_i) = g_i\,, \qquad i = 1, \dots, M\,, \tag{2.13}$$

$$u_i = u_i^b\,, \qquad i = M + 1, \dots, N\,, \tag{2.14}$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \dots, u_N) \in [0,1]$, $i, j = 1, \dots, N$, satisfy (2.12).

3. Solvability of the algebraic flux correction scheme and of its linearized variant. In this section we prove that the nonlinear problem (2.13), (2.14) is solvable under a continuity assumption on α_{ij}. As a consequence, we obtain the unique solvability of the linearized problem (2.13), (2.14) (with α_{ij} independent of the solution), which is useful for computing the solution of (2.13), (2.14) numerically using a fixed point iteration. The following result will be of great use in the proof of existence of solutions below.

LEMMA 3.1. *Consider any $\mu_{ij} = \mu_{ji} \le 0$, $i, j = 1, \dots, N$. Then*

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

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 shall use the following consequence of Brower's fixed-point Theorem, whose proof can be found in [18, p. 164, Lemma 1.4].

LEMMA 3.2. *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 $K > 0$ a real number such that $(Tx, x)_X > 0$ for any $x \in X$ with $\|x\|_X = K$. Then there exists $x \in X$ such that $\|x\|_X < K$ and $Tx = 0$.*

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

THEOREM 3.3. *Let* (2.3) *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* (2.12)*. Then there exists a solution of the nonlinear problem* (2.13)*,* (2.14)*.*

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 shall 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 (2.3),

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

Thus, one has

$$\sum_{i,j=1}^{M} v_i\, a_{ij}\, v_j \geq C_M\, \|\widetilde{\mathrm{V}}\|^2 \qquad \forall\ \widetilde{\mathrm{V}} \in \mathbb{R}^M\,. \tag{3.1}$$

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 (2.13), (2.14) if and only if $T\,\widetilde{\mathrm{U}} = 0$. The operator T is continuous and, in view of (3.1), Lemma 3.1, and Hölder's and Young's inequalities, one derives

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

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 3.2, there exists $\widetilde{\mathrm{U}} \in \mathbb{R}^M$ such that $T\,\widetilde{\mathrm{U}} = 0$. □

Corollary 3.4. *Let* (2.3) *hold. Consider any* $\alpha_{ij} \in [0,1]$, $i,j = 1,\ldots,N$, *satisfying* (2.12). *Then the system* (2.13), (2.14) *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 3.3, 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 3.1. *The statement of Corollary 3.4 can be proved directly (without using Theorem 3.3) by showing that the homogeneous system*

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j=1}^{N} (1-\alpha_{ij})\, d_{ij}\,(u_j - u_i) = 0\,, \qquad i = 1,\ldots,M\,, \tag{3.2}$$

$$u_i = 0\,, \qquad i = M+1,\ldots,N\,, \tag{3.3}$$

has only the trivial solution. Indeed, if $U = (u_1,\ldots,u_N)$ *solves* (3.2), (3.3), *then, according to Lemma 3.1, 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* (2.3).

Finally, let us formulate sufficient conditions on the functions α_{ij} assuring the validity of the continuity assumption in Theorem 3.3 for many particular examples of the functions α_{ij} used in practice (cf., e.g., [11, 15, 16]).

LEMMA 3.5. *Consider any $i, j \in \{1, \dots, N\}$ and let $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ satisfy*

$$\alpha_{ij}(U) = \frac{A_{ij}(U)}{|u_j - u_i| + B_{ij}(U)} \qquad \forall\, U \equiv (u_1, \dots, u_N) \in \mathbb{R}^N,\ u_i \neq u_j\,, \tag{3.4}$$

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 $u_1, \dots, u_N$ 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 $\bar{U} \equiv (\bar{u}_1, \dots, \bar{u}_N) \in \mathbb{R}^N$. If $\bar{u}_i \neq \bar{u}_j$, then there is a neighbourhood of $\bar{U}$, where the denominator from (3.4) 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 $U \equiv (u_1, \dots, u_N) \in \mathbb{R}^N$. 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 $U = (u_1, \dots, u_N)$ and $\bar{U} = (\bar{u}_1, \dots, \bar{u}_N)$. 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

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

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 shall show that it satisfies the assumptions of Lemma 3.5 and hence leads to a solvable nonlinear problem (2.13), (2.14).

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 $i = 1, \dots, N$ in the following way. First, one initializes all these quantities by zero. Then one goes through all pairs of indices $i, j \in \{1, \dots, N\}$ and one performs the updates

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

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^-, $i = 1, \dots, N$, one defines

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

If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Furthermore, according to [10], these quantities are set to 1 at Dirichlet nodes, i.e.,

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

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

$$\alpha_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0\,, \\ 1 & \text{if } f_{ij} = 0\,, \\ R_i^- & \text{if } f_{ij} < 0\,, \end{cases} \qquad \alpha_{ji} := \alpha_{ij}\,. \tag{4.1}$$

It is worth mentioning that this algorithm is the one presented in [12] (that originates from the ideas of [20]), to which, following [10], 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

$$P_i^+ = \sum_{\substack{j=1\\ a_{ji} \le a_{ij}}}^N f_{ij}^+\,, \qquad P_i^- = \sum_{\substack{j=1\\ a_{ji} \le a_{ij}}}^N f_{ij}^-\,, \qquad Q_i^+ = -\sum_{j=1}^N f_{ij}^-\,, \qquad Q_i^- = -\sum_{j=1}^N f_{ij}^+\,, \tag{4.2}$$

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 3.3, and then, that they lead to a solvable nonlinear problem (2.13), (2.14).

Lemma 4.1. *The above coefficients α_{ij} are such that $\alpha_{ij}(u_1,\dots,u_N)(u_j - u_i)$ are Lipschitz-continuous functions of $u_1,\dots,u_N$ on $\mathbb{R}^N$.*

Proof. Consider any $i,j \in \{1,\dots,N\}$. It suffices to consider the case $\alpha_{ij} \not\equiv 1$ (and hence $d_{ij} \ne 0$). Furthermore, due to (2.12), one may assume that $a_{ji} \le a_{ij}$. If $u_i > u_j$, then $f_{ij} > 0$ and hence

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

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

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

Thus, α_{ij} is of the form (3.4) 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 3.5 are satisfied and the result immediately follows from Lemma 3.5. ∎

Remark 4.1. *There is an apparent ambiguity in the definition of the coefficients α_{ij} if $a_{ij} = a_{ji}$. However, often $a_{ij} + a_{ji} \le 0$ (cf. assumption* (5.2) *in the next section), and then $a_{ij} = a_{ji} \le 0$. Thus, if the artificial diffusion matrix is defined by* (2.9)*, one obtains $d_{ij} = 0$ so that the respective α_{ij} does not occur in the nonlinear problem* (2.13), (2.14)*, 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,

$$a_{ii} > 0\,, \qquad \sum_{j=1}^{N} a_{ij} \geq 0 \qquad \forall\, i = 1,\dots,M\,, \tag{5.1}$$

$$a_{kl} + a_{lk} \leq 0 \qquad \forall\, k,l = 1,\dots,N\,,\ k \neq l\,,\ k \leq M \text{ or } l \leq M\,, \tag{5.2}$$

and we recall that $d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\}$ for all $i,j = 1,\dots,N$, $i \neq j$ (cf. (2.9)). The first condition in (5.1) is a consequence of (2.3), the second one is a necessary condition for the validity of the discrete maximum principle in case of the linear problem (2.1), (2.2). Note that the row sums are not affected by adding the nonlinear term in (2.13). Condition (5.2) is weaker than (2.6). In Section 7, we present a discrete problem for which all the assumptions in (5.1) and (5.2) are satisfied.

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

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

the sets of upwind nodes, and by

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

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

Thanks to (5.2), for any $i \in \{1,\dots,M\}$ and $j \in \{1,\dots,N\}$ 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 (4.2) defining P_i^+ and P_i^- can be written in the form

$$P_i^+ = \sum_{j\in \mathrm{Do}_i} f_{ij}^+\,, \qquad P_i^- = \sum_{j\in \mathrm{Do}_i} f_{ij}^-\,, \qquad i = 1,\dots,M\,. \tag{5.3}$$

Moreover, the second term on the left-hand side of (2.13) 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 $i \in \{1,\dots,M\}$ and $j \in \mathrm{Do}_i$, and consequently, $\alpha_{ji}\, f_{ji} = R_j^+\, f_{ji}^+ + R_j^-\, f_{ji}^-$ if $i \in \{1,\dots,M\}$ 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 algebraic flux correction scheme (2.13), (2.14) can be written in the form

$$\sum_{j=1}^{N} a_{ij}\, u_j + \sum_{j\in \mathrm{Do}_i} (Z_i^+\, f_{ij}^+ + Z_i^-\, f_{ij}^-) + \sum_{j\in \mathrm{Up}_i} (Z_j^+\, f_{ij}^- + Z_j^-\, f_{ij}^+) = g_i\,, \qquad i = 1,\dots,M\,, \tag{5.4}$$

$$u_i = u_i^b\,, \qquad i = M+1,\dots,N\,. \tag{5.5}$$

Next, defining

$$A_i = u_i \sum_{j=1}^{N} a_{ij}\,, \tag{5.6}$$

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 (5.2), 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. (5.3)), (5.4) is equivalent to

$$A_i - P_i^+\, R_i^+ - P_i^-\, R_i^- + \sum_{j\in \mathrm{Up}_i} (Z_j^+\, f_{ij}^- + Z_j^-\, f_{ij}^+ + a_{ij}\,(u_j-u_i)) = g_i\,. \tag{5.7}$$

The following is a preliminary technical result.

LEMMA 5.1. *Consider any $i \in \{1,\dots,M\}$ and let $u_i \le u_j$ for all $j \in \mathrm{Up}_i$. Then*

$$A_i - P_i^-\, R_i^- + R_i^+ \sum_{j\in \mathrm{Do}_i} a_{ij}\,(u_j-u_i)^- + \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^+\, d_{ij})\,|u_j-u_i| = g_i\,. \tag{5.8}$$

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

$$A_i - P_i^+\, R_i^+ + R_i^- \sum_{j\in \mathrm{Do}_i} a_{ij}\,(u_j-u_i)^+ - \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^-\, d_{ij})\,|u_j-u_i| = g_i\,. \tag{5.9}$$

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 (5.7). □

The following result is a quick consequence of the above lemma, whose implications will become apparent in Corollary 5.3 below.

COROLLARY 5.2. *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*

$$A_i + \sum_{j\in \mathrm{Up}_i} (a_{ij} + Z_j^+\, d_{ij})\,|u_j-u_i| = g_i\,. \tag{5.10}$$

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

$$A_i - \sum_{j \in \mathrm{Up}_i} (a_{ij} + Z_j^- \, d_{ij}) \, |u_j - u_i| = g_i \,. \tag{5.11}$$

Proof. One has $f_{ij}^+ = 0$ for $j = 1, \ldots, N$ and hence $Q_i^- = 0$, which gives $P_i^- \, R_i^- = 0$. Then, (5.10) follows from (5.8). To prove (5.11) it is enough to note that $f_{ij}^- = 0$ for $j = 1, \ldots, N$, which leads to $Q_i^+ = 0$ and $P_i^+ \, R_i^+ = 0$, and then apply (5.9). □

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

COROLLARY 5.3. *Consider any* $i \in \{1, \ldots, M\}$*. Then*

$$g_i \leq 0 \quad \Rightarrow \quad u_i \leq \max_{j \neq i, \, a_{ij} \neq 0} u_j \quad \textit{for } u_i \geq 0 \quad \Rightarrow \quad u_i \leq \max_{j \neq i, \, a_{ij} \neq 0} u_j^+ \,, \tag{5.12}$$

$$g_i \geq 0 \quad \Rightarrow \quad u_i \geq \min_{j \neq i, \, a_{ij} \neq 0} u_j \quad \textit{for } u_i \leq 0 \quad \Rightarrow \quad u_i \geq \min_{j \neq i, \, a_{ij} \neq 0} u_j^- \,. \tag{5.13}$$

Proof. Let $u_i \geq 0$. Then thanks to (5.1), $A_i \geq 0$ (where A_i is defined in (5.6)). If $u_i > u_j$ for all $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, then (5.11) holds with a positive left-hand side. Thus, if $g_i \leq 0$, then $u_i \leq u_j$ for some $j \in \mathrm{Up}_i \cup \mathrm{Do}_i$, which implies (5.12). The second statement is proved in an analogous way. □

REMARK 5.1. *It is worth remarking that, if* $\sum_{j=1}^N a_{ij} = 0$*, then the previous results can be strengthen since Lemma 5.1 and Corollary 5.2 hold with* $A_i = 0$*. Then Corollary 5.3 is valid without the restriction on the sign of* u_i*, i.e., for any* $i \in \{1, \ldots, M\}$*, one has*

$$g_i \leq 0 \qquad \Rightarrow \quad u_i \leq \max_{j \neq i, \, a_{ij} \neq 0} u_j \,,$$

$$g_i \geq 0 \qquad \Rightarrow \quad u_i \geq \min_{j \neq i, \, a_{ij} \neq 0} u_j \,.$$

This is in accordance with the corresponding results for partial differential equations (see, e.g., [5]).

6. Variational form of the algebraic flux correction scheme and error estimation. In this section we show how the linear system (2.1), (2.2) originates from a variational problem representing a finite element discretization and how, in turn, the nonlinear algebraic problem (2.13), (2.14) can be put into a variational form. Then the derivation of an error estimate is discussed. It is important to notice that all 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 $\Omega \subset \mathbb{R}^d$, $d \geq 1$, be a bounded domain and let the boundary $\partial\Omega$ of Ω be Lipschitz-continuous and polyhedral (if $d \geq 2$). Let $a : H^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be a bilinear form, $u_b \in H^{1/2}(\partial\Omega) \cap C(\partial\Omega)$, and $g \in H^{-1}(\Omega)$ and let us consider the variational problem:

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

$$a(u, v) = \langle g, v \rangle \qquad \forall \; v \in H_0^1(\Omega) \,. \tag{6.1}$$

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

To solve (6.1) 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 $\varphi_1, \dots, \varphi_N$ and assume that the functions $\varphi_1, \dots, \varphi_M$ (with $0 < M < N$) form a basis in V_h. In addition, we assume that there are points $x_1, \dots, x_N \in \overline{\Omega}$ such that $\varphi_i(x_j) = \delta_{ij}$, $i, j = 1, \dots, N$, where δ_{ij} is the Kronecker symbol, and that $x_{M+1}, \dots, x_N \in \partial\Omega$ (note that $x_1, \dots, x_M \in \Omega$). 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.,

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

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

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

$$a_h(v_h, v_h) \geq C_a\,\|v_h\|_a^2 \qquad \forall\ v_h \in V_h\,, \tag{6.2}$$

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 (6.1) can be introduced as the solution of the following finite-dimensional problem:

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

$$a_h(u_h, v_h) = \langle g, v_h \rangle \qquad \forall\ v_h \in V_h\,. \tag{6.3}$$

We denote

$$a_{ij} = a_h(\varphi_j, \varphi_i)\,, \qquad i, j = 1, \dots, N\,, \tag{6.4}$$

$$g_i = \langle g, \varphi_i \rangle\,, \qquad i = 1, \dots, M\,, \tag{6.5}$$

$$u_i^b = u_b(x_i)\,, \qquad i = M+1, \dots, N\,. \tag{6.6}$$

Then u_h is a solution of the finite-dimensional problem (6.3) if and only if it satisfies the relations (2.1) and (2.2). Moreover, the matrix $(a_{ij})_{i,j=1}^M$ satisfies (2.3). We denote

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

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

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

and hence we realize that the corresponding flux correction scheme (2.13), (2.14) is equivalent to the following variational problem:

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

$$a_h(u_h, v_h) + d_h(u_h; u_h, v_h) = \langle g, v_h \rangle \qquad \forall\ v_h \in V_h\,. \tag{6.7}$$

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 3.1) and hence it satisfies Schwarz's inequality

$$|d_h(w;z,v)|^2 \le d_h(w;z,z)\, d_h(w;v,v) \qquad \forall\ w,z,v \in C(\overline{\Omega})\,. \tag{6.8}$$

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 (6.7) 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 (6.7) 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 (6.1) and (6.7), 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 $u_h - z_h \in V_h$, one derives using (6.2) and (6.8) 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

$$\|u - u_h\|_h \le \inf_{z_h \in W_h^b} \left\{ \|u - z_h\|_h + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(z_h,v_h)}{\|v_h\|_h} + (d_h(u_h;z_h,z_h))^{1/2} \right\}. \tag{6.9}$$

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 (6.9) one gets the estimate

$$\|u - u_h\|_h \le C_a^{1/2}\,\|u - i_h u\|_a + \sup_{v_h \in V_h} \frac{a(u,v_h) - a_h(i_h u,v_h)}{\|v_h\|_h} + (d_h(u_h;i_h u,i_h u))^{1/2}\,. \tag{6.10}$$

Thus, as usual, the error of the discrete solution is estimated by an interpolation error and a consistency error. In the following section we shall 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

$$-\varepsilon\,\Delta u + \boldsymbol{b}\cdot\nabla u + c\,u = g \quad \text{in } \Omega\,, \qquad\qquad u = u_b \quad \text{on } \partial\Omega\,, \tag{7.1}$$

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

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

where σ_0 is a constant. The weak solution of (7.1) satisfies (6.1) 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 (7.1) exists, is unique, and satisfies the maximum principle (cf. [5]).

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 (7.1) has only nonnegative entries, which may cause a violation of the condition (2.6). In order to overcome this, we replace the matrix corresponding to the reaction term by a simple diagonal approximation:

$$(c\,u_h, v_h) = \sum_{i=1}^{M} (c\,u_h, \varphi_i)\,v_i \approx \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i \qquad \forall\ u_h \in W_h,\ v_h \in V_h\,. \tag{7.2}$$

This has the extra impact that it makes the matrix $\mathbb{D}$ to be independent of c (see below). The error introduced by this approximation is estimated in the following lemma.

Lemma 7.1. *There is a constant C independent of h such that*

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

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

$$\begin{aligned}(c\,u_h, v_h) - \sum_{i=1}^{M} (c, \varphi_i)\,u_i\,v_i &= \sum_{i=1}^{M} (c\,(u_h - u_i), \varphi_i)\,v_i = \sum_{T\in\mathscr{T}_h} \sum_{\substack{i\,=\,1\\ x_i\,\in\,T}}^{M} (c\,(u_h - u_i), \varphi_i)_T\,v_i\\ &\leq \|c\|_{0,\infty,\Omega} \sum_{T\in\mathscr{T}_h} \sum_{\substack{i\,=\,1\\ x_i\,\in\,T}}^{M} \|u_h - u_i\|_{0,1,T}\,|v_i|\,.\end{aligned}$$

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

Using the approximation (7.2), the bilinear form a_h in (6.3) is given by

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

and satisfies (6.2) 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 (7.1). The bilinear form a_h defines the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$ whose entries are given by (6.4). The artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is defined using (2.9), and thus, it is independent of c.

REMARK 7.1. *It is easy to verify that the matrix $\mathbb{A}$ satisfies* (5.1)*. The assumption* (5.2) *holds if and only if*

$$(\nabla\varphi_k,\nabla\varphi_l) \le 0 \qquad \forall\; k,l = 1,\dots,N\,,\;\; k\neq l\,,\;\; k\le M \;\text{ or }\; l\le M\,. \tag{7.3}$$

The validity of (7.3) *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* (7.3) *that $\mathscr{T}_h$ is a Delaunay triangulation, i.e., that the sum of any pair of angles opposite a common edge is smaller than, or equal to, π.*

Now we can discuss the estimation of the terms on the right-hand side of the error estimate (6.10). To this end, we assume that $u\in H^2(\Omega)$. Then, standard interpolation estimates (cf. [4]) give

$$\|u - i_h u\|_a \le C\,(\varepsilon + \sigma_0\,h^2)^{1/2}\,h\,|u|_{2,\Omega}\,. \tag{7.4}$$

The remaining two terms on the right-hand side of (6.10) will be estimated in the following two lemmas.

LEMMA 7.2. *Let $\sigma_0 > 0$. Then there is a constant C independent of h and the data of problem* (7.1) *such that, for any $u\in H^2(\Omega)$,*

$$\sup_{v_h\in V_h}\frac{a(u,v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \le C\,(\varepsilon + \sigma_0^{-1}\,\{\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2\})^{1/2}\,h\,\|u\|_{2,\Omega}\,. \tag{7.5}$$

If $c\equiv 0$, then

$$\sup_{v_h\in V_h}\frac{a(u,v_h) - a_h(i_h u, v_h)}{\|v_h\|_h} \le C\,(\varepsilon + \varepsilon^{-1}\,\|\boldsymbol{b}\|_{0,\infty,\Omega}^2\,h^2)^{1/2}\,h\,|u|_{2,\Omega}\,. \tag{7.6}$$

Proof. Consider any $u \in H^2(\Omega)$ and $v_h \in V_h$. Then, in view of Lemma 7.1,

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

Therefore, if $\sigma_0 > 0$, one obtains (7.5). If $c \equiv 0$, one can employ the fact that

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

which leads to (7.6). ◻

Lemma 7.2 shows that, if $\sigma_0 > 0$, one obtains from (6.10)

$$\|u - u_h\|_h \le C\,h\,\|u\|_{2,\Omega} + (d_h(u_h; i_h u, i_h u))^{1/2}\,, \tag{7.7}$$

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 (7.6) 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 (7.6) 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 (6.10).

Lemma 7.3. *Let the matrix $\mathbb{D}$ be defined by* (2.9). *Then there is a constant C independent of h and the data of problem* (7.1) *such that*

$$d_h(w_h; i_h u, i_h u) \le C\,(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h)\,|i_h u|^2_{1,\Omega} \qquad \forall\ w_h \in W_h,\ u \in C(\overline{\Omega})\,. \tag{7.8}$$

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

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

Therefore, using Lemma 3.1, one derives for any $w_h \in W_h$ and $u \in C(\overline{\Omega})$

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

Since

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

one obtains the statement of the lemma. □

One observes that if $d_h(u_h; i_h u, i_h u)$ in (7.7) is estimated using Lemma 7.3, the convergence order is reduced. As a matter of fact, (7.4), (7.5) and (7.8) lead to the following global error estimate.

COROLLARY 7.4. *Let $u \in H^2(\Omega)$ be the solution of (7.1), and u_h be a solution of the discrete problem (6.7). Then, if $\sigma_0 > 0$, there exists a constant $C > 0$, independent of h and the data of (7.1) such that*

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

REMARK 7.2. *A careful inspection of the proof of Lemma 7.3 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 (7.8) reduces to*

$$d_h(w_h; i_h u, i_h u) \le C\,\|\boldsymbol{b}\|_{0,\infty,\Omega}\,h\,|i_h u|^2_{1,\Omega} \qquad \forall\, w_h \in W_h,\ u \in C(\overline{\Omega})\,. \tag{7.9}$$

This estimate implies an $\mathcal{O}(\sqrt{h})$ error estimate in (7.7), 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 (7.8) reduces to*

$$d_h(w_h; i_h u, i_h u) \le C\,\varepsilon\,|i_h u|^2_{1,\Omega} \qquad \forall\, w_h \in W_h,\ u \in C(\overline{\Omega})\,, \tag{7.10}$$

which does not imply any convergence of $\|u - u_h\|_h$. However, this result can be improved for suitable types of the meshes that are used. 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 non-positive and hence $|d_{ij}| \le \|\boldsymbol{b}\|_{0,\infty,\Omega}\,h/3$ for $i \ne j$. Thus, the estimate (7.9) is again valid and leads to an $\mathcal{O}(\sqrt{h})$ in estimate (7.7), which will be confirmed by numerical experiments in Section 8.*

  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 the off-diagonal entries of the matrix $\mathbb{A}$ are non-positive 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 (6.7) reduces to the original linear method (6.3). 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* (6.7) *does not reduce to the original linear method* (6.3) *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* (7.10) *cannot be improved in general. Indeed, if* $\theta_{ij} > \pi/2$ *and* $\varepsilon \geq \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h\, |\tan\theta_{ij}|$*, then* $|d_{ij}| \geq \varepsilon\, |\cot\theta_{ij}|/3$*. 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.*

As we mentioned, 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 7.3 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 the estimate* (7.9) *predicts. 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.*

8. Numerical results. This section presents numerical results obtained with the algebraic flux correction scheme applied to the convection–diffusion–reaction equation (7.1). For the sake of brevity, the presentation will be restricted to studies of the convergence of the method for the below example with smooth solution. Results for an example with layers can be found, e.g., in [1].

EXAMPLE 8.1. *Problem* (7.1) *is considered with* $\Omega = (0,1)^2$*, with different values of* ε*, and with* $\boldsymbol{b} = (3,2)^T$*,* $c = 1$*,* $u_b = 0$*, and the right-hand side* g *chosen such that*

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

is the solution of (7.1)*.*

In the numerical simulations, $\mathbb{P}_1$ finite elements were used on triangular grids. Mass lumping (cf. (7.2)) was performed for the reactive term, but only very small differences could be observed 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 $a_{ji} \leq a_{ij}$, 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 results. The nonlinear discrete equations were solved with a fixed point iteration with Anderson acceleration [19]. The iterations were stopped if 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 them was MooNMD [7].

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

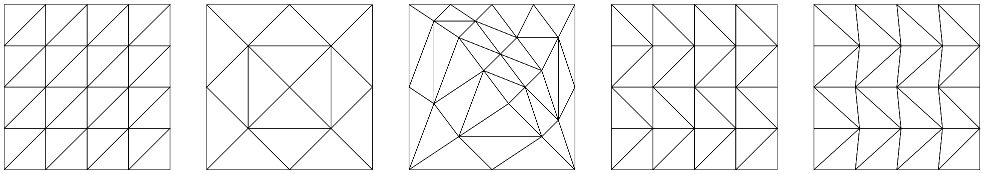

FIG. 1. *Grids 1 – 5 (left to right), level 0. The differences between Grid 4 and Grid 5 are described in the text.*

nodes to the right by the 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 7.2 satisfies $\theta_{ij} > \pi$.

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, the presentation of the results for $c = 0$ is omitted.

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 (2.13), (2.14) by a linear problem, which is, essentially, a stabilized method adding first order artificial diffusion to the original problem (2.1), (2.2). 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 preceeding column. The results in Table 1 indicate that the estimate (7.9) of $d_h(w_h; i_h u, i_h u)$, and hence also the estimate for $\|u - u_h\|_h$ given in Corollary 7.4, are sharp.

TABLE 1
Example 8.1, $\varepsilon = 10^{-8}$, numerical results for Grid 1 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	1.951e−2	0.79	4.408e−1	0.47	2.528e−1	0.43	2.535e−1	0.43
4	1.087e−2	0.84	3.228e−1	0.45	1.833e−1	0.46	1.836e−1	0.47
5	5.769e−3	0.91	2.334e−1	0.47	1.313e−1	0.48	1.314e−1	0.48
6	2.974e−3	0.96	1.670e−1	0.48	9.348e−2	0.49	9.353e−2	0.49
7	1.510e−3	0.98	1.188e−1	0.49	6.632e−2	0.50	6.634e−2	0.50
8	7.613e−4	0.99	8.429e−2	0.50	4.698e−2	0.50	4.698e−2	0.50

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 for all simulations around one. In all cases, this order is dominated by the order of convergence of $d_h(u_h, i_h u, i_h u)^{1/2}$. 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 8.1, $\varepsilon = 10^{-8}$, numerical results for Grid 1 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\vert e_h\vert_{1,\Omega}$	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	1.114e−1	0.97
4	1.408e−3	1.95	1.074e−1	1.09	5.317e−2	1.06	5.319e−2	1.07
5	3.493e−4	2.01	5.113e−2	1.07	2.472e−2	1.11	2.472e−2	1.11
6	8.652e−5	2.01	2.546e−2	1.01	1.158e−2	1.09	1.158e−2	1.09
7	2.152e−5	2.01	1.321e−2	0.95	5.533e−3	1.07	5.533e−3	1.07
8	5.357e−6	2.01	6.822e−3	0.95	2.685e−3	1.04	2.685e−3	1.04

Table 3

Example 8.1, $\varepsilon = 10^{-8}$, numerical results for Grid 2 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\vert e_h\vert_{1,\Omega}$	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	1.239e−1	1.04
4	2.516e−3	1.76	1.954e−1	0.57	5.884e−2	1.07	5.889e−2	1.07
5	8.369e−4	1.59	1.380e−1	0.50	2.801e−2	1.07	2.802e−2	1.07
6	2.891e−4	1.53	1.031e−1	0.42	1.356e−2	1.05	1.357e−2	1.05
7	1.103e−4	1.39	7.865e−2	0.39	6.638e−3	1.03	6.639e−3	1.03
8	4.136e−5	1.42	6.524e−2	0.27	3.263e−3	1.02	3.263e−3	1.02
9	1.539e−5	1.43	5.768e−2	0.18	1.618e−3	1.01	1.618e−3	1.01

Table 4

Example 8.1, $\varepsilon = 10^{-8}$, numerical results for Grid 3 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\vert e_h\vert_{1,\Omega}$	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	9.209e−2	1.06
4	2.216e−3	1.47	2.244e−1	0.51	4.488e−2	1.03	4.493e−2	1.04
5	9.946e−4	1.16	1.821e−1	0.30	2.224e−2	1.01	2.226e−2	1.01
6	4.993e−4	0.99	1.559e−1	0.22	1.124e−2	0.98	1.125e−2	0.98
7	2.519e−4	0.99	1.375e−1	0.18	5.676e−3	0.98	5.682e−3	0.98
8	1.277e−4	0.98	1.231e−1	0.16	2.871e−3	0.98	2.874e−3	0.98

Table 5

Example 8.1, $\varepsilon = 10^{-8}$, numerical results for Grid 4 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\vert e_h\vert_{1,\Omega}$	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	9.835e−2	1.06
4	2.313e−3	1.46	4.543e−1	0.09	4.341e−2	1.18	4.347e−2	1.18
5	1.089e−3	1.09	4.434e−1	0.03	1.830e−2	1.25	1.833e−2	1.25
6	5.527e−4	0.98	4.361e−1	0.02	8.276e−3	1.14	8.295e−3	1.14
7	2.817e−4	0.97	4.320e−1	0.01	3.926e−3	1.08	3.936e−3	1.08
8	1.425e−4	0.98	4.297e−1	0.01	1.915e−3	1.04	1.921e−3	1.03

Table 6
Example 8.1, $\varepsilon = 10^{-8}$, 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	1.002e−1	1.07
4	2.687e−3	1.37	5.395e−1	0.06	4.405e−2	1.18	4.413e−2	1.18
5	1.304e−3	1.04	5.294e−1	0.03	1.896e−2	1.22	1.901e−2	1.22
6	6.645e−4	0.97	5.225e−1	0.02	8.792e−3	1.11	8.817e−3	1.11
7	3.382e−4	0.97	5.186e−1	0.01	4.235e−3	1.05	4.249e−3	1.05
8	1.708e−4	0.99	5.164e−1	0.01	2.083e−3	1.02	2.091e−3	1.02

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 to the numerical studies presented in Section 8.1 is the behavior of the weights. They do not stay constant but they 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 7.3 is too rough in some cases.

Table 7
Example 8.1, $\varepsilon = 10^{-8}$, 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

For the diffusion-dominated regime, numerical results are presented in Tables 8–10. For Grid 1, all convergence orders 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 of 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 7.2.

Table 8
Example 8.1, $\varepsilon = 10$, 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	5.674e−1	0.99
4	5.379e−4	2.00	8.799e−2	1.00	5.643e−2	1.02	2.839e−1	1.00
5	1.345e−4	2.00	4.401e−2	1.00	2.792e−2	1.02	1.420e−1	1.00
6	3.360e−5	2.00	2.201e−2	1.00	1.387e−2	1.01	7.097e−2	1.00
7	8.398e−6	2.00	1.100e−2	1.00	6.912e−3	1.00	3.548e−2	1.00

9. Summary and Outlook. An algebraic flux correction 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, it is the first time that a convergence analysis of an algebraic flux correction scheme was performed. For a practical choice of the limiters, a local discrete

TABLE 9
Example 8.1, $\varepsilon = 10$, numerical results for Grid 4 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\lvert e_h\rvert_{1,\Omega}$	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	5.898e−1	0.94
4	6.209e−4	1.82	8.800e−2	1.00	1.473e−1	0.43	3.148e−1	0.91
5	1.940e−4	1.68	4.402e−2	1.00	1.069e−1	0.46	1.755e−1	0.84
6	6.899e−5	1.49	2.201e−2	1.00	7.657e−2	0.48	1.035e−1	0.76
7	2.789e−5	1.31	1.101e−2	1.00	5.450e−2	0.49	6.467e−2	0.68
8	1.239e−5	1.17	5.503e−3	1.00	3.867e−2	0.50	4.240e−2	0.61

TABLE 10
Example 8.1, $\varepsilon = 10$, numerical results for Grid 5 and α_{ij} from Section 4.

l	$\|e_h\|_{0,\Omega}$	ord.	$\lvert e_h\rvert_{1,\Omega}$	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	1.494e+0	0.25
4	1.123e−2	0.15	1.558e−1	0.52	1.316e+0	0.00	1.406e+0	0.09
5	1.090e−2	0.04	1.333e−1	0.22	1.313e+0	0.00	1.380e+0	0.03
6	1.080e−2	0.01	1.269e−1	0.07	1.312e+0	0.00	1.372e+0	0.01
7	1.077e−2	0.00	1.252e−1	0.02	1.311e+0	0.00	1.369e+0	0.00
8	1.076e−2	0.00	1.248e−1	0.00	1.310e+0	0.00	1.369e+0	0.00

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 for the convection-dominated case.

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

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

LEMMA 9.1. *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 (2.5) holds for any $u_1,\dots,u_N \in \mathbb{R}$ if and only if the conditions (2.6) and (2.8) are satisfied.*

Proof. Let us assume that at least one of the conditions (2.6) and (2.8) is not valid. We shall construct a counterexample to the validity of (2.5). If (2.6) does not hold, i.e., if $a_{ik} > 0$ for some $i \in \{1,\dots,M\}$ and $k \in \{1,\dots,N\}$, $k \neq i$, the 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 (2.5) does not hold. If (2.8) is not valid, i.e., if $\sum_{j=1}^N a_{ij} < 0$

for some $i \in \{1,\ldots,M\}$, then we set

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

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 (2.5) does not hold. This proves that the validity of (2.5) for any $u_1,\ldots,u_N \in \mathbb{R}$ implies (2.6) and (2.8).

Now let us assume that the conditions (2.6) and (2.8) are satisfied. Consider any $i \in \{1,\ldots,M\}$ and any $u_1,\ldots,u_N \in \mathbb{R}$ 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

$$\begin{aligned} \text{(9.1)}\qquad a_{ii}\,u_i &\leq \sum_{\substack{j\,=\,1\\ j\,\neq\, i}}^{N} (-a_{ij})\,u_j = \sum_{\substack{j\,=\,1\\ j\,\neq\, i}}^{N} (-a_{ij})\,(u_j - c) + \sum_{\substack{j\,=\,1\\ j\,\neq\, i}}^{N} (-a_{ij})\,c \\ &\leq c \sum_{\substack{j\,=\,1\\ j\,\neq\, i}}^{N} (-a_{ij}) \leq c\,a_{ii}\,, \end{aligned}$$

which implies that $u_i \leq c$. □

Lemma 9.2. *Let us consider a matrix* $(a_{ij})_{j=1,\ldots,N}^{i=1,\ldots,M}$ *with* $0 < M < N$ *and let* $a_{ii} > 0$ *for* $i = 1,\ldots,M$*. Then* (2.4) *holds for any* $u_1,\ldots,u_N \in \mathbb{R}$ *if and only if the conditions* (2.6) *and* (2.7) *are satisfied.*

Proof. Let us assume that at least one of the conditions (2.6) and (2.7) is not valid. Since the counterexamples from the proof of Lemma 9.1 can be used also here, it suffices to consider the case when $\sum_{j=1}^N a_{ij} > 0$ for some $i \in \{1,\ldots,M\}$. We set

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

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 (2.4) does not hold. This proves that the validity of (2.4) for any $u_1,\ldots,u_N \in \mathbb{R}$ implies (2.6) and (2.7).

Now let us assume that the conditions (2.6) and (2.7) are satisfied. Consider any $i \in \{1,\ldots,M\}$ and any $u_1,\ldots,u_N \in \mathbb{R}$ such that $\sum_{j=1}^N a_{ij}\,u_j \leq 0$. Setting

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

the statement (9.1) remains valid (the last '$\leq$' can be changed to '$=$') and hence $u_i \leq c$. □

REFERENCES

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

[2] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Some analytical results for an algebraic flux correction scheme for a steady convection-diffusion equation in one dimension. IMA J. Numer. Anal., 2015. in press.

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

[4] Alexandre Ern and Jean-Luc Guermond. Theory and Practice of Finite Elements. Springer-Verlag, New York, 2004.

[5] David Gilbarg and Neil S. Trudinger. Elliptic partial differential equations of second order, volume 224 of Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, second edition, 1983.

[6] Marcel Gurris, Dmitri Kuzmin, and Stefan Turek. Implicit finite element schemes for the stationary compressible Euler equations. Internat. J. Numer. Methods Fluids, 69(1):1–28, 2012.

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

[8] Volker John, Teodora Mitkova, Michael Roland, Kai Sundmacher, Lutz Tobiska, and Andreas Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. Chemical Engineering Science, 64:733–741, 2009.

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

[10] Dmitri Kuzmin. Personal communication.

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

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

[13] Dmitri Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. Journal of Computational and Applied Mathematics, 218:79–87, 2008.

[14] Dmitri Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. J. Comput. Appl. Math., 218:79–87, 2008.

[15] Dmitri Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. J. Comput. Phys., 228:2517–2534, 2009.

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

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

[18] Roger Temam. Navier-Stokes equations. Theory and numerical analysis. North-Holland, Amsterdam, 1977.

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

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