FINITE ELEMENT METHODS RESPECTING THE DISCRETE MAXIMUM PRINCIPLE FOR CONVECTION-DIFFUSION EQUATIONS*

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

Abstract. Convection-diffusion-reaction equations model the conservation of scalar quantities. From the analytic point of view, solution of these equations satisfy under certain conditions maximum principles, which represent physical bounds of the solution. That the same bounds are respected by numerical approximations of the solution is often of utmost importance in practice. The mathematical formulation of this property, which contributes to the physical consistency of a method, is called Discrete Maximum Principle (DMP). In many applications, convection dominates diffusion by several orders of magnitude. It is well known that standard discretizations typically do not satisfy the DMP in this convection-dominated regime. In fact, in this case, it turns out to be a challenging problem to construct discretizations that, on the one hand, respect the DMP and, on the other hand, compute accurate solutions. This paper presents a survey on finite element methods, with a main focus on the convection-dominated regime, that satisfy a local or a global DMP. The concepts of the underlying numerical analysis are discussed. The survey reveals that for the steady-state problem there are only a few discretizations, all of them nonlinear, that at the same time satisfy the DMP and compute reasonably accurate solutions, e.g., algebraically stabilized schemes. Moreover, most of these discretizations have been developed in recent years, showing the enormous progress that has been achieved lately. Methods based on algebraic stabilization, nonlinear and linear ones, are currently as well the only finite element methods that combine the satisfaction of the global DMP and accurate numerical results for the evolutionary equations in the convection-dominated situation.

Key words. convection-diffusion-reaction equations; convection-dominated regime; stabilized finite element methods; discrete maximum principle (DMP); matrices of non-negative type; algebraically stabilized schemes

AMS subject classifications. 65N30; 65M60

CONTENTS

*Submitted to the editors DATE.

Funding: The research of all authors was supported by the programme *Research in pairs* of the Mathematisches Forschungsinstitut Oberwolfach (MFO), grant No. 1937p. The work of Gabriel R. Barrenechea has been partially funded by the Leverhume Trust via the Research Fellowship No. RF-2019-510. The work of Volker John has been supported by the Research Training 'Group Differential Equation- and Data-driven Models in Life Sciences and Fluid Dynamics' (DAEDALUS), RTG 2433, funded by the German Research Foundation (DFG). The work of Petr Knobloch has been supported by the grant No. 22-01591S of the Czech Science Foundation.

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

‡Weierstrass Institute for Applied Analysis and Stochastics, Leibniz Institute im Forschungsverbund Berlin e. V. (WIAS), Mohrenstr. 29, 10117 Berlin, and Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany. john@wias-berlin.de.

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

This is a peer reviewed, accepted author manuscript of the following research article: Barrenechea, G. R., John, V., & Knobloch, P. (Accepted/In press - 17 January 2023). Finite element methods respecting the discrete maximum principle for convection-diffusion equations. *SIAM Review*.

1. Introduction.

Partial differential equations (PDEs) or systems of them are widely used for modeling processes from nature and industry. Usually, an analytic solution cannot be obtained. In practice, numerical methods are utilized for computing approximations of the solution. Such numerical methods consist of several components, like discretizations with respect to different variables, approaches for solving nonlinear problems, and solvers for systems of linear algebraic equations. The actual choice of these components might be dictated by different goals, like efficiency, or accuracy with respect to quantities of interest. A particular aspect of the second goal is the so-called physical consistency of a method, i.e., certain fundamental physical properties of the solution of the PDE should be inherited by the numerical solution. For many practitioners, the physical consistency is an essential criterion for utilizing a numerical method.

Classes of PDEs that can be found in many models from applications are elliptic linear second order equations

$$-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u = f \quad \text{in } \Omega, \tag{1.1}$$

and their parabolic counterparts

$$\partial_t u - \varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u = f \quad \text{in } (0,T]\times\Omega. \tag{1.2}$$

In these equations $\Omega \subset \mathbb{R}^d$, $d \geq 1$, is a spatial domain, $(0, T]$ a time interval, and u is some scalar quantity like the temperature or a concentration. This scalar quantity is transported by molecular diffusion with the diffusion coefficient ε [m^2/s] and by convective transport with the velocity field $\boldsymbol{b}$ [m/s]. The zeroth order term in (1.1) and (1.2) is called reactive term with the reaction coefficient σ [$1/\mathrm{s}$] and the term on the right-hand side describes sinks and sources of the scalar quantity. Both equations (1.1) and (1.2) have to be equipped with suitable boundary conditions at the boundary $\partial\Omega$ of Ω and (1.2) also with an initial condition at $t = 0$ in order to define well-posed problems. Then, the analysis of (1.1) and (1.2) is very well understood. In particular, it can be shown that under appropriate assumptions on the data of the problems, so-called Maximum Principles (MP) are satisfied. That means, loosely speaking, that the solution at some point or in some subdomain can be bounded a priori, e.g., for a global MP by the values on $\partial\Omega$ and, for the evolutionary problem, also on $\{0\} \times \Omega$. In case that the assumptions for the satisfaction of the MP are satisfied, it represents a fundamental physical property of solutions of (1.1) and (1.2).

A physically consistent discretization of (1.1) and (1.2) should satisfy discrete counterparts of the MP, the so-called Discrete Maximum Principle (DMP). Discretizations that do not fulfill the DMP are prone to numerical solutions with unphysical values, so-called spurious oscillations. Usually, equations of type (1.1) and (1.2) are part of coupled problems and their numerical solution serves as input data for other equations. With spurious oscillations in this input, there is a high probability that also the numerical solutions of the remaining equations possess unphysical values and finally the numerical simulation of the coupled problem might blow up, as it is our own experience reported in [73]. Consequently, the satisfaction of the DMP is essential for discretizations of (1.1) and (1.2) to be useful for simulations in applications. If this property is satisfied, then efficiency or the satisfaction of other physical properties, like conservation properties, or the accuracy with respect to quantities of interest, like norms in Sobolev spaces, are further criteria for selecting a method.

The first proof of a maximum principle for a discretization of a PDE was presented by Gershgorin [48] already in 1930. A generalization of this result is given in the monograph by Collatz [34] from 1955, whose English translation is [35]. The consideration of discrete analogs of maximum principles can be found in papers by Bramble and Hubbard [19, 20] published in the early 1960s. In 1970, Ciarlet presented in [31] necessary and sufficient conditions for a discretization to satisfy a DMP. In all these works, finite difference methods are considered. However, all arguments from linear algebra that were utilized in these papers can be applied analogously to linear systems of equations arising from other discretizations. The first work that studies the DMP explicitly for finite element methods was published in 1973 by Ciarlet and Raviart [32]. Since then, numerous papers appeared studying the DMP for different discretizations of elliptic and parabolic boundary value problems.

Convection-diffusion-reaction equations (1.1) and (1.2) possess a feature that makes the computation of a numerical solution challenging. In most applications, the convective transport by the velocity field strongly dominates the diffusive transport. Hence, the first order term in (1.1) and (1.2) is dominant. Under appropriate conditions on the smoothness of the data, it can be shown that (weak) solutions of (1.1) and (1.2) do not possess jumps, but they exhibit so-called layers. Layers are very thin regions where the norm of the gradient of the solution is very large. In the convection-dominated regime, the width of layer regions is much smaller than the affordable mesh width, apart from special cases when anisotropic layer-adapted meshes can be constructed. Hence, in general, layers cannot be resolved. Standard discretizations, like the Galerkin finite element method or central finite differences, cannot cope with this situation. In general, numerical solutions computed with such discretizations are globally polluted with spurious oscillations. A well-known remedy consists in using so-called stabilized discretizations.

Finite element methods are a popular approach for discretizing spatial derivatives. Major reasons include, but are not limited to, that unstructured meshes can be used easily, such that domains with complicated boundaries can be coped with, and that for many problems they allow an error analysis. In a nutshell, finite element methods start with a weak formulation of the PDE, replace the infinite-dimensional function spaces with finite-dimensional ones, usually consisting of piecewise polynomial functions, and they might approximate, modify or extend the forms (functionals, bilinear forms etc.) of the weak formulation. This procedure does not pay attention to physical consistency. The situation is different for other approaches, like finite volume methods, where a goal of the discretization process is to transfer conservation properties from the continuous to the discrete equation. However, in view of the attractive features of finite element methods, there has been a great interest in studying to which extent they lead to physically consistent discretizations and, in case of unsatisfactory findings, in developing modifications that possess the desired physical consistency.

The goal of the present paper consists in providing a survey on finite element methods that satisfy local or global DMPs for linear elliptic or parabolic problems. To keep the presentation focussed on the DMPs, other properties of the respective methods, like results from the finite element convergence theory, will be discussed only in the form of brief comments. On the one hand, many proofs concerning the DMPs use just basic tools from linear algebra and they will be presented such that main ideas of the numerical analysis become clear. But on the other hand, since this survey is intended also for an audience without special knowledge in the mathematical analysis of the finite element method, it is referred to the literature for some other proofs, in particular for those which require many technical steps. Although the considered problems (1.1) and (1.2) are linear, both linear as well as nonlinear finite element methods for their discretization have been proposed. A nonlinear method contains stabilization terms whose parameters depend on the numerical solution. That such methods can be suitable becomes clear from the above described form of the solution: there are layers and gently varying parts in the solution and an adequate discretization should treat both parts differently.

After formulating the steady-state problem and general notations in Section 2, the following Section 3 will introduce general results concerning the DMP for both linear and nonlinear discretizations. Then, several sections follow that consider discretizations of the steady-state problem. First, problems without convection, in particular the Poisson problem, will be discussed in Section 4. Then, linear discretizations and finally nonlinear discretizations of convection-diffusion-reaction problems will be reviewed in Sections 5 and 6, respectively. The theoretical considerations are illustrated by numerical results in Section 7. In all these sections, only discretizations with conforming piecewise linear ($\mathbb{P}_1$) finite elements are considered, since most of the literature is for this case. Methods for parabolic problems, and $\mathbb{P}_1$ finite elements in space, will be reviewed in Section 8. The survey reveals that many finite element methods that satisfy the DMP for $\mathbb{P}_1$ finite elements transferred ideas from finite volume methods, like upwind techniques or the consideration of fluxes. Finite elements different than $\mathbb{P}_1$ are the topic of Section 9. The available results for the satisfaction of the DMP for other $H^1(\Omega)$-conforming finite elements, often even only for the Poisson problem, pose usually very restrictive requirements on the shape of the mesh cells, or they are even negative. Thus, it turns out that the restriction to the $\mathbb{P}_1$ finite element in the literature (and the previous sections) has mathematical reasons. In addition, non-conforming finite elements are discussed. Then, Section 10 provides brief comments on methods that satisfy the DMP for hyperbolic conservation laws. Finally, a summary and an outlook are presented in Section 11.

2. The steady-state model problem, general notations.

Let $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$, be a bounded domain with polygonal resp. polyhedral and Lipschitz continuous boundary $\partial\Omega$. For a domain $D \subset \Omega$ we denote by $W^{m,p}(D)$ the space of functions in $L^p(D)$ with weak derivatives up to order m belonging to $L^p(D)$, with the usual convention $W^{0,p}(D) = L^p(D)$. The notation $W_0^{m,p}(D)$ denotes the closure of $C_0^\infty(D)$ in $W^{m,p}(D)$. If $p = 2$ and $m > 0$, the usual notations $H^m(D)$ and $H_0^m(D)$ are used instead of $W^{m,p}(D)$ and $W_0^{m,p}(D)$, respectively. The norm (seminorm) in $W^{m,p}(D)$ is denoted by $\|\cdot\|_{m,p,D}$ ($|\cdot|_{m,p,D}$), and whenever $p = 2$, the index p will be dropped from the notation, this is, $\|\cdot\|_{m,D} = \|\cdot\|_{m,2,D}$. The inner product in $L^2(D)$ or $L^2(D)^d$ is denoted by $(\cdot,\cdot)_D$, and the subindex will be dropped if $D = \Omega$. The Euclidean norm of a vector is denoted by $|\cdot|$. Finally, for a number $a \in \mathbb{R}$, we define its positive and negative parts as follows:

$$a^+ := \max\{a, 0\} \geq 0 \qquad \text{and} \qquad a^- := \min\{a, 0\} \leq 0\,,$$

and the same notation is used to define the positive and negative parts of a real-valued function.

2.1. The steady-state model problem.

Defining a characteristic length scale and a characteristic scale of the sought quantity, the steady-state equation (1.1) can be transformed to a dimensionless problem, where we use for simplicity the same notations: Find $u : \overline{\Omega} \to \mathbb{R}$ such that

$$\begin{array}{rcll} -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u &=& f & \text{in } \Omega\,, \\ u &=& g & \text{on } \partial\Omega\,. \end{array} \tag{2.1}$$

For simplifying the following presentation, we will suppose that $\varepsilon > 0$ and $\sigma \geq 0$ are constants and that $\boldsymbol{b}$ is solenoidal.

Let $\boldsymbol{b} \in W^{1,\infty}(\Omega)^d$, $f \in L^2(\Omega)$, and $g \in H^{1/2}(\partial\Omega)$, then the weak formulation of (2.1) reads as follows: Find $u \in H^1(\Omega)$ such that $u|_{\partial\Omega} = g$ and

$$a(u,v) = (f,v) \qquad \forall\, v \in H_0^1(\Omega)\,, \tag{2.2}$$

where $a(\cdot,\cdot)$ is the bilinear form given by

$$a(u,v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u + \sigma u, v)\,. \tag{2.3}$$

Under the stated assumptions on the smoothness of the data, the existence and uniqueness of a solution of (2.2) can be concluded from the Lax–Milgram theorem. The weak maximum principle for a sufficiently regular solution reads as follows, e.g., see [49, Chapter 3.1] or [42, Chapter 6.4.1].

THEOREM 2.1 (Weak maximum principle). *Let $u \in C^2(\Omega) \cap C(\overline{\Omega})$. Then*

$$\begin{array}{lll} -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u \leq 0 \quad \text{in } \Omega & \Longrightarrow & \max\limits_{\boldsymbol{x}\in\overline{\Omega}} u(\boldsymbol{x}) \leq \max\limits_{\boldsymbol{x}\in\partial\Omega} u^+(\boldsymbol{x}), \\ -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u \geq 0 \quad \text{in } \Omega & \Longrightarrow & \min\limits_{\boldsymbol{x}\in\overline{\Omega}} u(\boldsymbol{x}) \geq \min\limits_{\boldsymbol{x}\in\partial\Omega} u^-(\boldsymbol{x}). \end{array}$$

If $\sigma = 0$*, then*

$$
\begin{aligned}
-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u \le 0 \quad \text{in } \Omega \quad &\Longrightarrow \quad \max_{\boldsymbol{x}\in\overline{\Omega}} u(\boldsymbol{x}) = \max_{\boldsymbol{x}\in\partial\Omega} u(\boldsymbol{x}),\\
-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u \ge 0 \quad \text{in } \Omega \quad &\Longrightarrow \quad \min_{\boldsymbol{x}\in\overline{\Omega}} u(\boldsymbol{x}) = \min_{\boldsymbol{x}\in\partial\Omega} u(\boldsymbol{x}).
\end{aligned}
$$

2.2. Triangulations and finite element spaces. We denote by $\{\mathscr{T}_h\}_{h>0}$ a family of conforming and regular simplicial triangulations of Ω consisting of mesh cells K. Note that each mesh cell is the image of a fixed reference cell $\hat{K}$ via an affine map. We use the notion of facet to denote an edge in 2d or a face in 3d. Let $h_G = \operatorname{diam}(G)$ be the diameter of a set G and $h = \max\{h_K : K \in \mathscr{T}_h\}$. For a mesh $\mathscr{T}_h$, the following notations are used:

- internal vertices: $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_M\}$, vertices on the boundary: $\{\boldsymbol{x}_{M+1}, \dots, \boldsymbol{x}_N\}$,
- set of internal facets: $\mathscr{F}_I$, set of boundary facets: $\mathscr{F}_\partial$, set of all facets: $\mathscr{F}_h = \mathscr{F}_I \cup \mathscr{F}_\partial$,
- set of internal edges: $\mathscr{E}_I$, set of boundary edges: $\mathscr{E}_\partial$, set of all edges: $\mathscr{E}_h = \mathscr{E}_I \cup \mathscr{E}_\partial$,
- for $K \in \mathscr{T}_h, F \in \mathscr{F}_h$, and a vertex $\boldsymbol{x}_i$, we define the sets

$$
\begin{aligned}
\mathscr{F}_K &= \{F \in \mathscr{F}_h : F \subset K\}, & \mathscr{F}_i &= \{F \in \mathscr{F}_h : \boldsymbol{x}_i \in F\},\\
\mathscr{E}_K &= \{E \in \mathscr{E}_h : E \subset K\}, & \mathscr{E}_F &= \{E \in \mathscr{E}_h : E \subset F\},
\end{aligned}
$$

- for $K \in \mathscr{T}_h, F \in \mathscr{F}_h$, $E \in \mathscr{E}_h$, and a vertex $\boldsymbol{x}_i$, we define the following subsets of $\overline{\Omega}$

$$
\begin{aligned}
\omega_K &= \cup\{K' \in \mathscr{T}_h : K \cap K' \neq \emptyset\}, & \omega_F &= \cup\{K \in \mathscr{T}_h : F \subset K\},\\
\tilde{\omega}_F &= \cup\{K \in \mathscr{T}_h : K \cap F \neq \emptyset\}, & \omega_E &= \cup\{K \in \mathscr{T}_h : E \subset K\},\\
\omega_i &= \cup\{K \in \mathscr{T}_h : \boldsymbol{x}_i \in K\},
\end{aligned}
$$

- for a vertex $\boldsymbol{x}_i$, we define the set of indices corresponding to neighbor vertices by

$$S_i = \{j \in \{1, \dots, N\} \setminus \{i\} \,:\, \boldsymbol{x}_i \text{ and } \boldsymbol{x}_j \text{ are endpoints of } E \in \mathscr{E}_h\}, \tag{2.4}$$

- for a facet $F \in \mathscr{F}_I$, we denote the jump of a function across F by $[\![\cdot]\!]_F$. The orientation of the jump is irrelevant, but fixed.

Note that from the regularity of the triangulations a minimal angle condition follows, e.g., see [21, Section 4.3]. In particular, the number of mesh cells in ω_K, ω_E, and ω_i is bounded uniformly for all K, E, i, and h. In addition, the mesh regularity implies that there exists a positive constant ρ such that

$$h_K \le \rho\, h_F \qquad \forall\ K \subset \tilde{\omega}_F\,. \tag{2.5}$$

Let $\boldsymbol{x}_i, \boldsymbol{x}_j$ be two vertices that are connected by an edge $E_{ij} \in \mathscr{E}_h$ (or, simply E when there is no possible confusion) and $K \subset \omega_{E_{ij}}$, then, compare Figure 1 for the two-dimensional situation,

- F_i^K and F_j^K are the facets of K opposite $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively, with outer unit normals $\boldsymbol{n}_i^K$ and $\boldsymbol{n}_j^K$, respectively,
- θ_E^K is the angle formed by F_i^K and F_j^K, or, more precisely, θ_E^K is the dihedral angle given by (cf. [22]))

$$\cos\theta_E^K = -\boldsymbol{n}_i^K \cdot \boldsymbol{n}_j^K\,, \tag{2.6}$$

- $\kappa_E^K = F_i^K \cap F_j^K$; when $d = 2$, we will adopt the convention $|\kappa_E^K| = 1$,

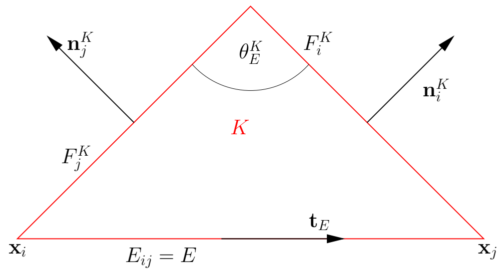


FIG. 1. *Notations for a triangle.*

- $\boldsymbol{t}_E = (\boldsymbol{x}_j - \boldsymbol{x}_i)/|\boldsymbol{x}_j - \boldsymbol{x}_i|$, where the orientation of this tangent vector is irrelevant, but fixed,
- $\delta_E v := v(\boldsymbol{x}_j) - v(\boldsymbol{x}_i)$ for any function $v \in C^0(\overline{\Omega})$ if the tangent vector $\boldsymbol{t}_E$ points from $\boldsymbol{x}_i$ to $\boldsymbol{x}_j$, and $\delta_E v := v(\boldsymbol{x}_i) - v(\boldsymbol{x}_j)$ in the other situation.

Whether or not a discretization satisfies a DMP might depend on properties of the underlying mesh or family of meshes. Some relevant properties in two and three dimensions are defined next.

DEFINITION 2.2 (Properties of meshes). *A mesh $\mathscr{T}_h$ will be said to be connected if, for any two vertices $\boldsymbol{x}_i, \boldsymbol{x}_j$, there exists a path $j_0, \ldots, j_s$ such that $E_{ij_0}, E_{j_0j_1}, \ldots, E_{j_sj}$ are all edges in $\mathscr{E}_h$. In addition, the mesh $\mathscr{T}_h$ will be said to be:*
- *weakly acute: if every internal dihedral angle θ of the mesh satisfies $\theta \le \frac{\pi}{2}$,*
- *of Xu–Zikatanov (XZ) type (cf. [135]): if, for every $E \in \mathscr{E}_I$, the following holds*

$$\sum_{K \subset \omega_E} |\kappa_E^K| \cot \theta_E^K \ge 0\,, \tag{2.7}$$

- *of Delaunay type: if the interior of the circumscribed sphere of any simplex from the mesh $\mathscr{T}_h$ does not contain any vertex of $\mathscr{T}_h$.*

For $d = 2$, the definition of a Delaunay mesh can be equivalently stated as follows: for every $E = K \cap K' \in \mathscr{E}_I$ there holds

$$\theta_E^K + \theta_E^{K'} \le \pi\,.$$

In two dimensions, the XZ-criterion and the Delaunay property are equivalent.

DEFINITION 2.3 (Strictly acute and average acute families of meshes). *A mesh family $\{\mathscr{T}_h\}_{h>0}$ will be said to be strictly acute if there is a constant $\delta > 0$ independent of h such that every internal dihedral angle θ of any of the meshes satisfies*

$$\theta \le \frac{\pi}{2} - \delta\,. \tag{2.8}$$

In two dimensions, a family $\{\mathscr{T}_h\}_{h>0}$ will be said to be average acute if, for every $h > 0$ and every edge $E = K \cap K' \in \mathscr{E}_I$, the following holds:

$$\theta_E^K + \theta_E^{K'} \le \pi - \delta\,, \tag{2.9}$$

where $\delta > 0$ is independent of h.

As already mentioned, most discretizations discussed in this survey are based on continuous piecewise linear finite elements. The corresponding finite element spaces and interpolation operators for this case will be defined next. Associated with the vertices $\{\boldsymbol{x}_1,\dots,\boldsymbol{x}_N\}$, the standard continuous piecewise linear basis functions $\phi_1,\dots,\phi_N$ are given by the property $\phi_i(\boldsymbol{x}_j)=\delta_{ij}$ for $i,j\in\{1,\dots,N\}$. Then, the corresponding conforming finite element spaces are

$$V_h := \operatorname{span}\{\phi_1,\dots,\phi_N\} \quad \text{and} \quad V_{h,0} := \operatorname{span}\{\phi_1,\dots,\phi_M\}\,. \tag{2.10}$$

Associated with V_h, the Lagrange interpolation operator is defined by

$$i_h : C^0(\overline{\Omega}) \to V_h\,, \quad v \mapsto i_h v = \sum_{i=1}^N v(\boldsymbol{x}_i)\phi_i\,.$$

We will also use the symbol i_h to interpolate functions with domain in the boundary of Ω, this is, $i_h g = \sum_{i=M+1}^N g(\boldsymbol{x}_i)\phi_i$.

2.3. Finite element matrices. In this section, the main finite element matrices are introduced. The diffusion matrix $\mathbb{A}_{\rm d}$, the convection matrix $\mathbb{A}_{\rm c}$, and the reaction matrix $\mathbb{M}_{\rm c}$, which is also called consistent mass matrix, are defined by

$$\mathbb{A}_{\rm d} = (\ell_{ij})_{i,j=1}^N \quad \text{where } \ell_{ij} = (\nabla\phi_j,\nabla\phi_i) \quad \text{for } i,j=1,\dots,N\,, \tag{2.11}$$

$$\mathbb{A}_{\rm c} = (c_{ij})_{i,j=1}^N \quad \text{where } c_{ij} = (\boldsymbol{b}\cdot\nabla\phi_j,\phi_i) \quad \text{for } i,j=1,\dots,N\,, \tag{2.12}$$

$$\mathbb{M}_{\rm c} = (m_{ij})_{i,j=1}^N \quad \text{where } m_{ij} = (\phi_j,\phi_i) \quad \text{for } i,j=1,\dots,N\,. \tag{2.13}$$

The entries of the matrices can be written as a sum of local entries, e.g.,

$$\ell_{ij} = \sum_{K\subset\omega_i\cap\omega_j} \ell_{ij}^K \quad \text{with } \ell_{ij}^K = (\nabla\phi_j,\nabla\phi_i)_K,$$

and analogously for c_{ij} and m_{ij}.

In the derivations made in the coming sections, having exact formulae for the diffusion and consistent mass matrices will be of much use. A basic tool in the derivations below is a formula relating the gradient of the barycentric coordinates and the normal outward vector to K. Since the basis function $\phi_i|_K$ vanishes on F_i^K, its derivative in any direction tangent to F_i^K vanishes. So, $\nabla\phi_i|_K$ is proportional to the unit normal $\boldsymbol{n}_i^K$. Consider the height vector $\boldsymbol{h}_i$ from F_i^K to $\boldsymbol{x}_i$. This vector is parallel to $\boldsymbol{n}_i^K$, pointing in the opposite direction, and the derivative of $\phi_i|_K$ in the direction of $\boldsymbol{h}_i$ is the constant $1/|\boldsymbol{h}_i|$. Hence, using the formula for the volume of the simplex K leads to

$$\nabla\phi_i|_K = -\frac{1}{|\boldsymbol{h}_i|}\boldsymbol{n}_i^K = -\frac{|F_i^K|}{d|K|}\boldsymbol{n}_i^K\,. \tag{2.14}$$

So, in view of (2.6), the local diffusion matrix is given by

$$\ell_{ij}^K = (\nabla\phi_j,\nabla\phi_i)_K = |K|\frac{|F_j^K|\,|F_i^K|}{d^2|K|^2}\boldsymbol{n}_j^K\cdot\boldsymbol{n}_i^K = -\frac{|F_j^K|\,|F_i^K|}{d^2|K|}\cos\theta_E^K. \tag{2.15}$$

Concerning the mass matrix and using the formula for the integral of a product of barycentric coordinates, see, e.g., [131] where this is proven in any space dimension, one gets

$$m_{ij}^K = \begin{cases} \dfrac{2|K|}{(d+1)(d+2)} & i=j,\\[2mm] \dfrac{|K|}{(d+1)(d+2)} & \text{else}\,. \end{cases} \tag{2.16}$$

Both in the steady-state and time-dependent situations, mass lumping is a widely used technique to discretize terms without spatial derivatives. The derivation of mass lumping starts with the construction of a dual mesh, which is a technique from finite volume methods. For each node $\boldsymbol{x}_i$, all mesh cells $K \subset \omega_i$ are considered. In each mesh cell, a polyhedral subset with volume $|K|/(d+1)$ assigned to $\boldsymbol{x}_i$ is constructed. The vertices of this subset are $\boldsymbol{x}_i$, the barycenter of K, midpoints of edges of K containing $\boldsymbol{x}_i$, and, if $d = 3$, also the barycenters of faces of K containing $\boldsymbol{x}_i$. Now, the dual mesh cell D_i is defined by the union of these subsets from all $K \subset \omega_i$. Consequently, one has

$$|D_i| = \frac{|\omega_i|}{d+1}\,.$$

Piecewise constant basis functions, given by

$$\psi_i(\boldsymbol{x}) = \begin{cases} 1 & \text{if } \boldsymbol{x} \in D_i\,, \\ 0 & \text{else}\,, \end{cases} \qquad i = 1, \ldots, N, \tag{2.17}$$

are associated with this dual mesh. With the help of these functions, the following lumping operator is defined

$$\mathscr{L} \;:\; C(\overline{\Omega}) \to L^2(\Omega)\,, \quad v \mapsto \mathscr{L}v = \sum_{i=1}^{N} v(\boldsymbol{x}_i)\psi_i\,. \tag{2.18}$$

In addition, the lumped $L^2(\Omega)$ inner product $(\cdot,\cdot)_h : C(\overline{\Omega}) \times C(\overline{\Omega}) \to \mathbb{R}$ is given by

$$(f, g)_h = (\mathscr{L}f, \mathscr{L}g)\,. \tag{2.19}$$

Since $\{\psi_i\}_{i=1}^N$ is an orthogonal set in $L^2(\Omega)$ and $(\psi_i, \psi_i) = |D_i|$, one obtains

$$(f, g)_h = \sum_{i,j=1}^{N} f(\boldsymbol{x}_j)g(\boldsymbol{x}_i)(\psi_j, \psi_i) = \sum_{i=1}^{N} |D_i| f(\boldsymbol{x}_i)g(\boldsymbol{x}_i)\,.$$

Using the lumped inner product, the following seminorm is induced in $C(\overline{\Omega})$, which is a norm in V_h,

$$|f|_h := (f, f)_h^{1/2} = \left(\sum_{i=1}^{N} |D_i|\,|f(\boldsymbol{x}_i)|^2\right)^{1/2}\,.$$

Finally, the lumped mass matrix, which is a diagonal matrix, is defined as follows

$$\mathbb{M}_l = (\tilde{m}_{ij})_{i,j=1}^N \quad \text{where} \quad \tilde{m}_{ij} = (\phi_j, \phi_i)_h = (\mathscr{L}\phi_j, \mathscr{L}\phi_i) = |D_i|\delta_{ij}\,. \tag{2.20}$$

Utilizing an exact quadrature rule for linears and the fact that the basis functions of V_h form a partition of unity yields

$$\tilde{m}_{ii} = |D_i| = \sum_{K \subset \omega_i} \frac{|K|}{d+1} = \sum_{K \subset \omega_i} (1, \phi_i)_K = (1, \phi_i) = \sum_{j=1}^{N} (\phi_j, \phi_i) = \sum_{j=1}^{N} m_{ij}\,. \tag{2.21}$$

So, the lumped mass matrix can be computed directly from the consistent mass matrix, without the need to build the dual mesh.

3. General results on DMP satisfying discretizations. This section provides conditions for the satisfaction of local and global DMPs that are based on special properties of matrices for general linear discrete problems, and of nonlinear forms for general nonlinear discretizations. The presentation of the theory for linear discretizations is based on the concept of matrices of non-negative type, instead on the traditional approach with monotone matrices or, more special, M-matrices. This concept enables also the consideration of local DMPs.

3.1. Linear discretizations. Let a matrix $(a_{ij})_{j=1,\dots,N}^{i=1,\dots,M} \in \mathbb{R}^{M\times N}$ and real numbers $f_1,\dots,f_M,g_1,\dots,g_{N-M}$ with $M<N$ be given. A linear discretization leads to a system of linear algebraic equations of the following form: Find $\boldsymbol{u}=(u_1,\dots,u_N)^T \in \mathbb{R}^N$ such that

$$\sum_{j=1}^{N} a_{ij}u_j = f_i \qquad \text{for } i=1,\dots,M\,, \tag{3.1}$$

$$u_i = g_{i-M} \qquad \text{for } i=M+1,\dots,N\,. \tag{3.2}$$

Remark 3.1. The system matrix of the system (3.1)-(3.2) is of the form

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_{\mathrm{I}} & \mathbb{A}_{\mathrm{B}} \\ \mathbb{O} & \mathbb{I} \end{pmatrix}\,, \tag{3.3}$$

where $\mathbb{A}_{\mathrm{I}} \in \mathbb{R}^{M\times M}$ is the matrix associated with the internal (or non-Dirichlet) degrees of freedom, $\mathbb{A}_{\mathrm{B}} \in \mathbb{R}^{M\times(N-M)}$ is the matrix that couples the boundary values to the values in the interior of the domain, $\mathbb{I}\in\mathbb{R}^{(N-M)\times(N-M)}$ is the identity matrix and $\mathbb{O}\in\mathbb{R}^{(N-M)\times M}$ a matrix consisting of zeros. In what follows, $\mathbb{A}$ will always denote the matrix given by (3.3). □

Definition 3.2 (Matrix of non-negative type). *A matrix* $(a_{ij})_{j=1,\dots,n}^{i=1,\dots,m} \in \mathbb{R}^{m\times n}$ *($m,n\in\mathbb{N}$) will be said to be of non-negative type if*

$$a_{ij} \le 0 \qquad \forall\ i\ne j,\ 1\le i\le m,\ 1\le j\le n\,, \tag{3.4}$$

$$\sum_{j=1}^{n} a_{ij} \ge 0 \qquad \forall\ 1\le i\le m\,. \tag{3.5}$$

One should notice that the notion of a matrix of non-negative type must not be confused with the notion of a non-negative matrix as it is studied, e.g., in [126, Chapter 2].

Remark 3.3. In some cases, e.g., when $\sigma=0$ in (2.1), the matrix $\mathbb{A}$ will satisfy a stronger property than (3.5), namely

$$\sum_{j=1}^{N} a_{ij} = 0 \qquad \forall\ 1\le i\le M\,. \tag{3.6}$$

With this property, it will be possible to derive stronger statements for the DMP than with (3.5). □

The next result is a local version of the results given in [31, 32].

Theorem 3.4 (Local DMP in the case of matrices of non-negative type). *Let* $a_{ii}>0$ *for* $i=1,\dots,M$. *Then, any possible solution of* (3.1)-(3.2) *satisfies*

$$f_i \le 0 \implies u_i \le \max_{j\ne i, a_{ij}\ne 0} u_j^+\,, \qquad f_i \ge 0 \implies u_i \ge \min_{j\ne i, a_{ij}\ne 0} u_j^- \tag{3.7}$$

for all $i = 1, \ldots, M$ if and only if $\mathbb{A}$ is of non-negative type. The implications

$$f_i \le 0 \quad \Longrightarrow \quad u_i \le \max_{j\ne i, a_{ij}\ne 0} u_j\,, \qquad\qquad f_i \ge 0 \quad \Longrightarrow \quad u_i \ge \max_{j\ne i, a_{ij}\ne 0} u_j \tag{3.8}$$

hold true for all $i = 1, \ldots, M$ if and only if $\mathbb{A}$ is of non-negative type and satisfies in addition (3.6).

Proof. Consider any $i \in \{1, \ldots, M\}$ and let $f_i \le 0$. If $\mathbb{A}$ is of non-negative type, then it follows from (3.1), (3.4), and (3.5) that

$$a_{ii}\, u_i = f_i - \sum_{j\ne i} a_{ij}\, u_j \le \sum_{j\ne i} (-a_{ij}) \max_{j\ne i, a_{ij}\ne 0} u_j^+ \le a_{ii} \max_{j\ne i, a_{ij}\ne 0} u_j^+\,,$$

which implies (3.7). If, in addition, (3.6) holds, then (3.8) follows from

$$a_{ii}\, u_i = f_i - \sum_{j\ne i} a_{ij}\, u_j \le \sum_{j\ne i} (-a_{ij}) \max_{j\ne i, a_{ij}\ne 0} u_j = a_{ii} \max_{j\ne i, a_{ij}\ne 0} u_j\,.$$

The statements for $f_i \ge 0$ follow analogously. The necessity of the conditions on $\mathbb{A}$ can be proved by constructing appropriate counterexamples, see [12, Appendix]. ∎

In the context of numerical approximation of PDEs, Theorem 3.4 implies a local DMP. It should be emphasized that for the local DMP the invertibility of $\mathbb{A}$ is not a necessary condition. In particular, it holds also for convection-diffusion equations (2.1), without reactive term, and with pure Neumann boundary conditions as long as their discretization leads to a system matrix of non-negative type and there is a solution.

Next, the global version of the DMP is shown. Its proof is based on a technique developed in [81] and can be considered as a generalization of [31, Theorem 3].

THEOREM 3.5 (Global DMP in the case of matrices of non-negative type). *Let us suppose that $\mathbb{A}$ is of non-negative type and that the matrix $\mathbb{A}_{\mathrm{I}} = (a_{ij})_{i,j=1}^M$ is invertible. Then, system (3.1)-(3.2) possesses a unique solution. This solution satisfies*

$$\begin{aligned} f_i \le 0 \quad \forall\ i = 1, \ldots, M \quad &\Longrightarrow \quad \max_{i=1,\ldots,N} u_i \le \max_{j=M+1,\ldots,N} u_j^+\,,\\ f_i \ge 0 \quad \forall\ i = 1, \ldots, M \quad &\Longrightarrow \quad \min_{i=1,\ldots,N} u_i \ge \min_{j=M+1,\ldots,N} u_j^-\,. \end{aligned} \tag{3.9}$$

In addition, if $\mathbb{A}$ satisfies (3.6), the following holds

$$\begin{aligned} f_i \le 0 \quad \forall\ i = 1, \ldots, M \quad &\Longrightarrow \quad \max_{i=1,\ldots,N} u_i = \max_{j=M+1,\ldots,N} u_j\,,\\ f_i \ge 0 \quad \forall\ i = 1, \ldots, M \quad &\Longrightarrow \quad \min_{i=1,\ldots,N} u_i = \min_{j=M+1,\ldots,N} u_j\,. \end{aligned} \tag{3.10}$$

Proof. Inserting the values from (3.2) in (3.1) leads to a linear system of equations for $u_1, \ldots, u_M$ with the matrix $\mathbb{A}_{\mathrm{I}}$. From the assumed invertibility of this matrix, the existence of a unique solution of (3.1)-(3.2) follows.

Next, the first statement of (3.9) will be shown. The second statement of (3.9) follows by changing the signs of $\boldsymbol{u}$ and of the right-hand side of (3.1)-(3.2). Let

$$s = \max_{i=1,\ldots,N} u_i \quad \text{and} \quad J = \{i \in \{1, \ldots, N\} : u_i = s\}\,.$$

If $s \le 0$, then (3.9) holds trivially. So, consider $s > 0$ and assume that $J \subset \{1, \ldots, M\}$. It will be shown that

$$\exists k \in J \text{ such that } \ \mu_k := \sum_{j\in J} a_{kj} > 0\,. \tag{3.11}$$

Let us suppose that (3.11) does not hold. Then, one concludes by combining (3.4) and (3.5) that

$$\sum_{j\in J} a_{ij} = 0 \qquad \forall\, i \in J\,.$$

Hence, the matrix $(a_{ij})_{i,j\in J}$ is singular because the sum of its columns is zero. With $(a_{ij})_{i,j\in J}$, also its transposed $(a_{ji})_{i,j\in J}$ is singular. Hence, there exist numbers $v_i, i \in J$, not all zero, such that

$$\sum_{i\in J} a_{ij} v_i = 0 \qquad \forall\, j \in J\,. \tag{3.12}$$

In addition, applying that $\mathbb{A}$ is of non-negative type one finds that $a_{ij} = 0$ for all $i \in J$ and all $j \notin J$. Using this property, (3.12), and defining the vector $\tilde{\boldsymbol{v}} = (\tilde{v}_i)_{i=1}^M$, where $\tilde{v}_i = v_i$ if $i \in J$, and $\tilde{v}_i = 0$ otherwise, yields

$$\sum_{i=1}^{M} a_{ij}\tilde{v}_i = \sum_{i\in J} a_{ij} v_i = 0\,,$$

for all $j \in \{1,\dots,M\}$. This implies that the matrix $\mathbb{A}_{\mathrm{I}}$ is singular, which contradicts the hypothesis. So, (3.11) holds.

Denoting now

$$r = \max_{i\notin J} u_i^+\,,$$

one obtains with $f_i \le 0$ for all i, (3.4), and (3.5)

$$\begin{aligned} s\mu_k &= \sum_{j\in J} a_{kj}u_j = f_k - \sum_{j\notin J} a_{kj}u_j \le -\sum_{j\notin J} a_{kj}u_j = \sum_{j\notin J}(-a_{kj})u_j \le r\sum_{j\notin J}(-a_{kj}) \\ &= r\left(\sum_{j=1}^{N}(-a_{kj}) + \sum_{j\in J} a_{kj}\right) \le r\mu_k\,. \end{aligned}$$

This implies that $s \le r$, which is a contradiction to the definition of s. Hence, $J \cap \{M+1,\dots,N\} \ne \emptyset$ and (3.9) follows.

The validity of (3.10) easily follows from (3.9). Since (3.6) holds, one can add a sufficiently large positive constant $q > 0$ to every u_i in such a way that all components of this new vector $\tilde{\boldsymbol{u}}$ are positive. Then, the first statement of (3.9) holds for $\tilde{\boldsymbol{u}}$ without the positive parts, which implies the first statement of (3.10). ∎

Remark 3.6. If the global DMP (3.9) holds and $\boldsymbol{u} \in \mathbb{R}^N$ is such that $u_{M+1} = \ldots = u_N = 0$ and $\boldsymbol{u}_{\mathrm{I}} := (u_1,\dots,u_M)^T$ satisfies $\mathbb{A}_{\mathrm{I}}\boldsymbol{u}_{\mathrm{I}} = 0$, then $\max_{i=1,\dots,N} u_i \le 0$ and $\min_{i=1,\dots,N} u_i \ge 0$ so that $\boldsymbol{u} = 0$. Consequently, the validity of the global DMP (3.9) implies that the matrix $\mathbb{A}_{\mathrm{I}}$ is invertible. Thus, this additional assumption (in comparison to the assumptions of Theorem 3.4 for the local DMP) is necessary. □

Remark 3.7. It is easy to construct a matrix $\mathbb{A}$ of non-negative type and a vector $\boldsymbol{u} = (u_1,\dots,u_N)^T$ such that the right-hand side of some of the implications in Theorem 3.4 holds for all $i = 1,\dots,M$ but the corresponding right-hand side in Theorem 3.5 is not satisfied. Thus, a global DMP cannot be obtained as a consequence of the validity of the corresponding local DMPs. On the other hand, it can also happen that the global DMP holds but the local one not since the assumption that $\mathbb{A}$ is of non-negative type is not necessary for the validity of the global DMP. □

Remark 3.8. A situation considered sometimes in the literature is the case of homogeneous Dirichlet boundary values. In this case, the proof of Theorem 3.5 does not require any assumptions on the submatrix $\mathbb{A}_{\mathrm{B}} = (a_{ij})_{j=M+1,\ldots,N}^{i=1,\ldots,M}$. However, such assumptions are needed in the general case, and consequently considering homogeneous Dirichlet boundary conditions is only a particular situation. □

Remark 3.9. From the previous theorems, it follows that both the local and global DMPs are satisfied if $\mathbb{A}$ is of non-negative type and $\mathbb{A}_{\mathrm{I}}$ is invertible. Since $\det \mathbb{A} = \det \mathbb{A}_{\mathrm{I}}$, one observes that $\mathbb{A}_{\mathrm{I}}$ is invertible if and only if $\mathbb{A}$ is invertible. Moreover, a direct calculation shows that

$$\mathbb{A} = \begin{pmatrix} \mathbb{A}_{\mathrm{I}} & \mathbb{A}_{\mathrm{B}} \\ \mathbb{O} & \mathbb{I} \end{pmatrix} \quad \Longleftrightarrow \quad \mathbb{A}^{-1} = \begin{pmatrix} \mathbb{A}_{\mathrm{I}}^{-1} & -\mathbb{A}_{\mathrm{I}}^{-1}\mathbb{A}_{\mathrm{B}} \\ \mathbb{O} & \mathbb{I} \end{pmatrix}. \tag{3.13}$$

In addition, an interesting observation is that the proof of (3.9) allows that $\mathbb{A}_{\mathrm{B}} = \mathbb{O}$. Hence, there is no connection between the degrees of freedom and the prescribed values on the boundary. In contrast, (3.6) in combination with the invertibility of $\mathbb{A}_{\mathrm{I}}$ requires that $\mathbb{A}_{\mathrm{B}} \neq \mathbb{O}$. □

As discussed in the previous remark, the invertibility of $\mathbb{A}_{\mathrm{I}}$ is a necessary and sufficient condition for the well-posedness of the discrete problem and is also necessary for proving that a method satisfies a global DMP (cf. Remark 3.6). Then, under the assumptions of the previous theorems, the matrix $\mathbb{A}_{\mathrm{I}}$ is of non-negative type (since $\mathbb{A}$ is) and invertible. It will be shown in Corollary 3.13 that these properties imply that the matrix $\mathbb{A}_{\mathrm{I}}$ belongs to the class of M-matrices defined next.

Definition 3.10 (M-matrix, monotone matrix). *A matrix $\mathbb{Q} = (q_{ij})_{i,j=1}^{n}$ is an M-matrix if:*

i) The off-diagonal entries are non-positive, i.e., $q_{ij} \le 0$, $i, j = 1, \ldots, n$, $i \neq j$;
ii) $\mathbb{Q}$ is non-singular; and
iii) $\mathbb{Q}^{-1} \ge 0$.

A matrix that satisfies conditions ii) and iii) is called monotone matrix.

In the above definition, the condition $\mathbb{Q}^{-1} \ge 0$ means that all entries of the matrix $\mathbb{Q}^{-1}$ are non-negative. In the following, an analogous notation will be used also for vectors, e.g., $\boldsymbol{v} \ge 0$ means that all entries of the vector $\boldsymbol{v}$ are non-negative.

Remark 3.11. A monotone matrix $\mathbb{Q}$ can be equivalently characterized by the property that, for any $\boldsymbol{v} \in \mathbb{R}^n$, the validity of $\mathbb{Q}\boldsymbol{v} \ge 0$ implies $\boldsymbol{v} \ge 0$. Indeed, if this implication holds, then $\mathbb{Q}$ is non-singular (since $\mathbb{Q}\boldsymbol{v} = 0$ implies both $\boldsymbol{v} \ge 0$ and $-\boldsymbol{v} \ge 0$) and if $\boldsymbol{v}$ is any column of $\mathbb{Q}^{-1}$, one has $\mathbb{Q}\boldsymbol{v} \ge 0$ and hence $\boldsymbol{v} \ge 0$ so that $\mathbb{Q}^{-1} \ge 0$. On the other hand, if $\mathbb{Q}$ is monotone, then $\mathbb{Q}\boldsymbol{v} \ge 0$ implies that $\boldsymbol{v} = \mathbb{Q}^{-1}\mathbb{Q}\boldsymbol{v} \ge 0$. □

Theorem 3.12 (Equivalence of the monotonicity and the global DMP). *Let the row sums of the matrix $\mathbb{A}$ be non-negative. Then the global DMP (3.9) is satisfied if and only if $\mathbb{A}$ is monotone.*

Proof. If the global DMP holds, then, for any $\boldsymbol{v} \in \mathbb{R}^N$ satisfying $\mathbb{A}\boldsymbol{v} \ge 0$, one has $v_i \ge \min_{j=M+1,\ldots,N} v_j^- = 0$ for all $i = 1, \ldots, N$ so that $\mathbb{A}$ is monotone due to Remark 3.11. Reciprocally, let $\mathbb{A}$ be monotone and let $\boldsymbol{u} \in \mathbb{R}^N$ be the solution of

(3.1)-(3.2) with $f_i \geq 0$, $i = 1, \ldots, M$. Set $c := \min_{j=M+1,\ldots,N} u_j^-$ and define $\boldsymbol{v} \in \mathbb{R}^N$ by $v_i = u_i - c$. Since $c \leq 0$ and the row sums of $\mathbb{A}$ are non-negative, one has $\mathbb{A}\boldsymbol{v} \geq 0$. Then the monotonicity of $\mathbb{A}$ implies that $\boldsymbol{v} \geq 0$ and hence $u_i \geq c$ for $i = 1, \ldots, N$. Thus the global DMP holds. ∎

COROLLARY 3.13 (M-matrix property of $\mathbb{A}$). *If the matrix $\mathbb{A}$ is invertible and of non-negative type, then both $\mathbb{A}$ and $\mathbb{A}_{\mathrm{I}}$ are M-matrices.*

Proof. If $\mathbb{A}$ is invertible and of non-negative type, then, according to Theorem 3.5, the global DMP (3.9) is satisfied and $\mathbb{A}$ is monotone in view of Theorem 3.12. Consequently, $\mathbb{A}$ is an M-matrix. In view of (3.13), $\mathbb{A}_{\mathrm{I}}$ is an M-matrix as well. ∎

Remark 3.14. Using (3.13), it follows immediately that if $\mathbb{A}$ is an M-matrix (monotone matrix) also $\mathbb{A}_{\mathrm{I}}$ is an M-matrix (monotone matrix). Conversely, if $\mathbb{A}_{\mathrm{I}}$ is an M-matrix (monotone matrix) and $\mathbb{A}_{\mathrm{B}} \leq 0$ (in particular, if $\mathbb{A}$ is of non-negative type), then $\mathbb{A}$ is an M-matrix (monotone matrix). □

Remark 3.15. The analysis for linear discretizations was performed purely on the algebraic level. We like to emphasize that the results concerning the vector $\boldsymbol{u}$ with respect to the DMP can be transferred to the corresponding finite element function only in special cases, like for the $\mathbb{P}_1$ finite element. Finite element spaces where such a transfer is not possible are discussed in Section 9. □

3.2. Nonlinear discretizations. In this section we will deal with two types of nonlinear discretizations of (2.1) which will be considered in variational forms with the $\mathbb{P}_1$ finite element spaces (2.10):
Type I: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$a(u_h, v_h) + j_h(u_h; v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,, \tag{3.14}$$

where $a(\cdot,\cdot)$ is the bilinear form given by (2.3), and $j_h(\cdot;\cdot)$ is a nonlinear stabilizing term, linear in the second argument.
Type II: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$a(u_h, v_h) + d_h(u_h; u_h, v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,, \tag{3.15}$$

where $a(\cdot,\cdot)$ is the bilinear form given by (2.3), and $d_h(\cdot;\cdot,\cdot)$ is nonlinear in the first argument and linear in the remaining two arguments. We assume that $d_h(\cdot;\cdot,\cdot)$ vanishes if the second argument is constant, i.e.,

$$d_h(w_h; 1, v_h) = 0 \qquad \forall\, w_h, v_h \in V_h \tag{3.16}$$

and that, for all $w_h \in V_h$, the bilinear form $d_h(w_h;\cdot,\cdot)$ is positive semidefinite, i.e.,

$$d_h(w_h; v_h, v_h) \geq 0 \qquad \forall\, w_h, v_h \in V_h\,. \tag{3.17}$$

Due to the nonlinear character of (3.14) and (3.15) the results presented in the last section cannot be applied. We present below two criteria for the satisfaction of the DMP. In both cases the criteria are related to the following remark: in order to prove the DMP, the only argument used concerns the entries of the row that corresponds to a node where an extremum of a discrete solution is encountered. So, to prove the DMP, it is not necessary to modify every equation, but only those associated with local extrema of a solution u_h. Based on this idea, in [28] a criterion was proposed in order to prove the DMP for a nonlinear discretization of Type I. Here, we present the following two variants of this criterion.

Definition 3.16 (Strong and weak DMP properties). *The nonlinear form $j_h(\cdot;\cdot)$ is said to satisfy the strong DMP property if the following condition holds: If u_h attains a strict local minimum (maximum) at an interior node $\boldsymbol{x}_i$, then there exist constants $\alpha_F > 0$, $F \in \mathscr{F}_i$, such that*

$$a(u_h, \phi_i) + j_h(u_h; \phi_i) \leq - \sum_{F \in \mathscr{F}_i} \alpha_F \, |[\![\nabla u_h]\!]_F| \,,$$

(resp. $\geq \sum_{F\in\mathscr{F}_i} \alpha_F |[\![\nabla u_h]\!]_F|$). The form $j_h(\cdot;\cdot)$ is said to satisfy the weak DMP property if the same conclusion holds under the extra assumption that the local minimum (maximum) satisfies $u_h(\boldsymbol{x}_i) < 0$ (resp. $u_h(\boldsymbol{x}_i) > 0$).

Definition 3.17 (Strong and weak DMP properties for non-strict extrema). *The nonlinear form $j_h(\cdot;\cdot)$ is said to satisfy the strong or weak DMP property for non-strict extrema if the conditions from Definition 3.16 hold not only in case of a strict local minimum (maximum) but also in case of a non-strict local minimum (maximum) of u_h at the node $\boldsymbol{x}_i$.*

Theorem 3.18 (Local and global DMPs for nonlinear discretizations of Type I). *Let us suppose that $j_h(\cdot;\cdot)$ satisfies the weak DMP property. Then, method (3.14) satisfies the local DMP in the following sense:*

$$(3.18) \quad (f, \phi_i) \leq 0 \implies \max_{\omega_i} u_h \leq \max_{\partial\omega_i} u_h^+ \,, \qquad (f, \phi_i) \geq 0 \implies \min_{\omega_i} u_h \geq \min_{\partial\omega_i} u_h^- \,,$$

for all $i = 1, \dots, M$. If $j_h(\cdot;\cdot)$ satisfies the strong DMP property, (3.14) satisfies the local DMP in the following sense:

$$(3.19) \quad (f, \phi_i) \leq 0 \implies \max_{\omega_i} u_h = \max_{\partial\omega_i} u_h \,, \qquad (f, \phi_i) \geq 0 \implies \min_{\omega_i} u_h = \min_{\partial\omega_i} u_h \,,$$

for all $i = 1, \dots, M$. In addition, the global DMP is also satisfied in the following form

$$(3.20) \quad f \leq 0 \text{ in } \Omega \implies \max_{\overline{\Omega}} u_h \leq \max_{\partial\Omega} u_h^+ \,, \qquad f \geq 0 \text{ in } \Omega \implies \min_{\overline{\Omega}} u_h \geq \min_{\partial\Omega} u_h^- \,,$$

if $j_h(\cdot;\cdot)$ satisfies the weak DMP property for non-strict extrema and in the form

$$(3.21) \quad f \leq 0 \text{ in } \Omega \implies \max_{\overline{\Omega}} u_h = \max_{\partial\Omega} u_h \,, \qquad f \geq 0 \text{ in } \Omega \implies \min_{\overline{\Omega}} u_h = \min_{\partial\Omega} u_h \,,$$

if $j_h(\cdot;\cdot)$ satisfies the strong DMP property for non-strict extrema.

Proof. The idea of the proof originates from [28]. Consider any $i \in \{1, \dots, M\}$ and let $(f, \phi_i) \leq 0$. Since $\max_{\omega_i} u_h$ is attained at a node, one has $\max_{\omega_i} u_h = \max\{u_h(\boldsymbol{x}_i), \max_{\partial\omega_i} u_h\} \leq \max\{u_h(\boldsymbol{x}_i), \max_{\partial\omega_i} u_h^+\}$. Thus, (3.18) trivially holds if $u_h(\boldsymbol{x}_i) \leq 0$ and hence it suffices to assume that $u_h(\boldsymbol{x}_i) > 0$ or that the strong DMP property holds. Let us assume that $u_h(\boldsymbol{x}_i) > \max_{\partial\omega_i} u_h$. Then u_h attains a strict local maximum at $\boldsymbol{x}_i$ and hence the strong (weak) DMP property implies that

$$0 \geq (f, \phi_i) = a(u_h, \phi_i) + j_h(u_h; \phi_i) \geq \sum_{F \in \mathscr{F}_i} \alpha_F |[\![\nabla u_h]\!]_F| \,.$$

Thus, ∇u_h is a constant in ω_i and hence u_h is a $\mathbb{P}_1$ function in ω_i, which is a contradiction since u_h was assumed to attain a strict local extremum in $\boldsymbol{x}_i$. Consequently,

$u_h(\boldsymbol{x}_i) \le \max_{\partial\omega_i} u_h$, which proves (3.19) and also (3.18). If $(f, \phi_i) \ge 0$, one can proceed analogously.

For the global results (3.20), (3.21), let us suppose that $f \le 0$ in Ω and that the solution attains a global maximum at $\boldsymbol{x}_i$ with some $i \in \{1, \ldots, M\}$. If only the weak DMP property holds, it is again sufficient to assume that $u_h(\boldsymbol{x}_i) > 0$. Then, analogously as for the local result, one deduces that u_h is a $\mathbb{P}_1$ function in ω_i. Since u_h attains an extremum at $\boldsymbol{x}_i$, it has to be constant in ω_i, and thus the global maximum is attained at a node $\boldsymbol{x}_j \in \partial\omega_i$. If $\boldsymbol{x}_j \in \partial\Omega$, there is nothing more to prove. Otherwise, we proceed as above and conclude that u_h is constant in ω_j as well. Continuing in the same fashion, and using that the mesh is connected, one can conclude that the global maximum is reached at a point on the boundary $\partial\Omega$. ∎

To treat problems of Type II, we introduce the following condition, reminiscent of [82] (see also [14]).

Definition 3.19 (Algebraic DMP property). *We will say that $d_h(\cdot;\cdot,\cdot)$ satisfies the algebraic DMP property if the following condition holds: Consider any $u_h \in V_h$ and any $i \in \{1, \ldots, M\}$. If $u_h(\boldsymbol{x}_i)$ is a strict local extremum of u_h on ω_i, i.e.,*

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

then

$$a(\phi_j, \phi_i) + d_h(u_h; \phi_j, \phi_i) \le 0 \qquad \forall\, j \in S_i \tag{3.22}$$

and

$$d_h(u_h; \phi_j, \phi_i) = 0 \qquad \forall\, j \notin S_i \cup \{i\}\,. \tag{3.23}$$

One can notice that, in essence, what (3.22) states is that only the i^{th} row in the nonlinear system (3.15) behaves like a matrix of non-negative type, and not all the rows, in contrast to the case of linear discretizations. The algebraic DMP property is sufficient for proving the local DMP. The proof of the global DMP requires a sign condition also in case of non-strict extrema.

Definition 3.20 (Algebraic DMP property for non-strict extrema). *We will say that $d_h(\cdot;\cdot,\cdot)$ satisfies the algebraic DMP property for non-strict extrema if the following condition holds: Consider any $u_h \in V_h$ and any $i \in \{1, \ldots, M\}$. If $u_h(\boldsymbol{x}_i)$ is a local extremum of u_h on ω_i, i.e.,*

$$u_h(\boldsymbol{x}_i) \ge u_h(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \omega_i \qquad \text{or} \qquad u_h(\boldsymbol{x}_i) \le u_h(\boldsymbol{x}) \quad \forall\, \boldsymbol{x} \in \omega_i\,,$$

then

$$a(\phi_j, \phi_i) + d_h(u_h; \phi_j, \phi_i) \le 0 \qquad \forall\, j \in S_i \text{ with } u_h(\boldsymbol{x}_j) \ne u_h(\boldsymbol{x}_i) \tag{3.24}$$

and (3.23) holds.

Theorem 3.21 (Local and global DMPs for nonlinear discretizations of Type II). *Let $u_h \in V_h$ be a solution of (3.15) and let us suppose that $d_h(\cdot;\cdot,\cdot)$ satisfies the algebraic DMP property. Then the local DMP (3.18) holds for all $i = 1, \ldots, M$. If, in addition, $\sigma = 0$, then also the stronger form (3.19) of the local DMP holds for all $i = 1, \ldots, M$.*

If $d_h(\cdot;\cdot,\cdot)$ satisfies the algebraic DMP property for non-strict extrema, then the global DMP (3.20) is satisfied. If, in addition, $\sigma = 0$, then also the stronger form (3.21) of the global DMP holds.

Proof. Denote $u_i = u_h(\boldsymbol{x}_i)$ and $\tilde{a}_{ij} = a(\phi_j,\phi_i) + d_h(u_h;\phi_j,\phi_i)$ for $i,j = 1,\dots,N$, and let us prove the local versions of the DMP. Consider any $i \in \{1,\dots,M\}$ and let $(f,\phi_i) \le 0$. If $\sigma > 0$, it suffices to consider $u_i > 0$ since otherwise (3.18) trivially holds (cf. the beginning of the proof of Theorem 3.18). Let us assume that $u_i > u_j$ for all $j \in S_i$. If $d_h(\cdot;\cdot,\cdot)$ satisfies the algebraic DMP property, then it follows from (3.15) and (3.23) that

$$A_i\, u_i + \sum_{j\in S_i} \tilde{a}_{ij}\,(u_j - u_i) = (f,\phi_i)\,, \tag{3.25}$$

where $A_i := \sum_{j=1}^N \tilde{a}_{ij} = (\sigma,\phi_i)$ due to (3.16). Moreover, (3.22) implies that the sum in (3.25) is non-negative. If $\sigma = 0$, then $A_i = 0$ and hence there is $j \in S_i$ such that $\tilde{a}_{ij} < 0$ since $\tilde{a}_{ii} \ge \varepsilon\,|\phi_i|^2_{1,\Omega} > 0$ (see (3.17)). This implies that the sum in (3.25) is positive. If $\sigma > 0$, then $A_i\,u_i > 0$. Thus, in both cases, the left-hand side of (3.25) is positive, which is a contradiction. Therefore, there is $j \in S_i$ such that $u_i \le u_j$, which proves (3.18) and (3.19). If $(f,\phi_i) \ge 0$, one can proceed analogously.

The proof of the global DMP can be carried out analogously as for Theorem 3.5, see also the proof of Theorem 3 in [14]. ∎

4. Linear discretizations of steady-state problems without convection.

This first section on linear discretizations is devoted to the special case of (2.1) where $\boldsymbol{b} = \boldsymbol{0}$. For all linear discretizations, the proofs of the DMP will consist of checking the hypotheses of Theorem 3.4. It turns out that the DMP is satisfied only under appropriate requirements on the mesh.

A careful inspection of the statements of the results from Section 3.1 reveals that one only needs to show properties for the first M rows of the coefficient matrix of system (3.1)-(3.2), that is, one only needs to worry about the equations associated with nodes interior to Ω. This observation motivates to define, for $\mathbb{A} \in \mathbb{R}^{N\times N}$, the matrix $(\mathbb{A})^M \in \mathbb{R}^{M\times N}$ as the matrix containing only the first M rows of $\mathbb{A}$. In fact, showing that $(\mathbb{A})^M$ is of non-negative type is what is needed to use Theorems 3.4 and 3.5 due to the expression (3.3) for the matrix associated with the system (3.1)-(3.2).

4.1. The Poisson problem.

In this section we will discuss necessary and sufficient conditions for the satisfaction of the DMP for the Poisson problem. The argument relies on proving that the diffusion matrix $(\mathbb{A}_{\rm d})^M$, defined in (2.11), is of non-negative type. For the finite element method the first result in this direction is given in [32]. Since in that paper the partial differential equation is a reaction-diffusion equation, the mesh is supposed to be acute and fine enough (see Section 4.2 below). Later, for the Poisson problem in 2d, it was noted that it is only needed for the mesh to satisfy the Delaunay criterion, see [121, p. 78]. Extensions to three space dimensions can be found in [21].

We start noticing that using (2.15) leads to the first proof of the satisfaction of the DMP for the Poisson problem. In fact, if the mesh $\mathscr{T}_h$ is weakly acute, then, using (2.15), one has $\ell_{ij} = \sum_{K\subset\omega_i\cap\omega_j} \ell^K_{ij} \le 0$ for $i \ne j$. This observation has been widely used in the literature and provides a sufficient condition for the satisfaction of the DMP for the Poisson equation. The proof we present next was first given in [135, Lemma 2.1] and has the advantage that it presents a necessary and sufficient condition on the mesh to guarantee the satisfaction of the local DMP.

THEOREM 4.1 (Sufficient and necessary condition for $(\mathbb{A}_{\rm d})^M$ to be of non-negative type, [135]). *A sufficient condition for the matrix $(\mathbb{A}_{\rm d})^M$ to be of non-negative type is* *that the mesh $\mathscr{T}_h$ satisfies the XZ-criterion (2.7). If any internal edge of $\mathscr{T}_h$ has at least one endpoint in Ω, then this condition is necessary. In addition, $(\mathbb{A}_{\rm d})^M$ satisfies (3.6).*

Proof. Let $\boldsymbol{x}_i, \boldsymbol{x}_j$ be two different nodes contained in the same mesh cell $K \in \mathscr{T}_h$. We recall the following formulas for the volume of a simplex

$$|K| = \frac{|F_i^K||F_j^K|}{2}\sin\theta^K_{E_{ij}} \quad \text{if } d=2, \qquad |K| = \frac{2|F_i^K||F_j^K|}{3|\kappa^K_{E_{ij}}|}\sin\theta^K_{E_{ij}} \quad \text{if } d=3\,.$$

Inserting them in (2.15), and using the convention that $|\kappa^K_{E_{ij}}| = 1$ if $d = 2$ gives

$$\ell^K_{ij} = -\frac{1}{d(d-1)}|\kappa^K_{E_{ij}}|\cot\theta^K_{E_{ij}}\,. \tag{4.1}$$

Thus, for $i \in \{1,\dots,M\}$ and $j \in S_i$,

$$\ell_{ij} = \sum_{K\subset\omega_{E_{ij}}} \ell^K_{ij} = -\sum_{K\subset\omega_{E_{ij}}} \frac{|\kappa^K_{E_{ij}}|\cot\theta^K_{E_{ij}}}{d(d-1)}\,, \tag{4.2}$$

and then (3.4) is satisfied if (2.7) holds. If the set $\mathscr{E}_I$ consists only of edges E_{ij} with $i \in \{1,\dots,M\}$ and $j \in S_i$, then (2.7) is necessary for the validity of (3.4). Finally, since the basis functions form a partition of unity, one has

$$\sum_{j=1}^{N} \ell_{ij} = \sum_{j=1}^{N} (\nabla\phi_j, \nabla\phi_i) = (\nabla 1, \nabla\phi_i) = 0\,. \tag{4.3}$$

So, (3.6) is satisfied, and in particular (3.5). ∎

Remark 4.2. The statement of Theorem 4.1 implies, in connection with Theorem 3.4, that the local DMP is satisfied if and only if the mesh is of XZ-type, with the slight exception concerning edges whose endpoints are both on $\partial\Omega$. In addition, Theorems 4.1 and 3.5 show that the validity of the XZ-criterion implies the global DMP. However, in this case, the XZ-criterion is not necessary. Indeed, in [39] a two-dimensional example is constructed where the global DMP is satisfied although the mesh is not of XZ-type. Nevertheless, in general, if the mesh is not of XZ-type, then the global DMP might be violated as an example in [22] demonstrates. □

Remark 4.3. Let $\mathbb{A}_{\rm d,I} \in \mathbb{R}^{M\times M}$ denote the $M \times M$ submatrix of the diffusion matrix only considering the non-Dirichlet nodes, i.e., the analog of $\mathbb{A}_{\rm I}$ in (3.3). Then, $\mathbb{A}_{\rm d,I}$ is non-singular, since the corresponding bilinear form is elliptic on $H^1_0(\Omega)$. □

Remark 4.4. A Poisson problem with heterogeneous anisotropic diffusion is given by

$$\begin{array}{rcll} -\nabla\cdot(\mathbb{E}(\boldsymbol{x})\nabla u) &=& f & \text{in } \Omega\,,\\ u &=& g & \text{on } \partial\Omega\,, \end{array} \tag{4.4}$$

with the symmetric diffusion tensor $\mathbb{E}(\boldsymbol{x})$. The tensor $\mathbb{E}$ depends on the spatial variable $\boldsymbol{x}$, which makes it heterogeneous, and in addition it is allowed to have different eigenvalues at a given $\boldsymbol{x}$, making it anisotropic. In any case, it will be assumed that $\mathbb{E}$ is symmetric and strictly positive-definite in Ω. Numerous applications lead to heterogeneous anisotropic diffusion, such as image processing [124] and atmospheric modelling [120], just to name a few.

Problem (4.4) was considered in [101] for $\mathbb{P}_1$ finite elements in two and three dimensions. The main condition on the mesh is the following: for every element K it is assumed that

$$\left(\boldsymbol{n}_i^K\right)^T \mathbb{E}_K \boldsymbol{n}_j^K \leq 0 \quad \forall\, \boldsymbol{x}_i, \boldsymbol{x}_j \in K,\ \boldsymbol{x}_i \neq \boldsymbol{x}_j, \quad \forall\, K \in \mathscr{T}_h\,, \tag{4.5}$$

where $\mathbb{E}_K$ stands for an approximation of the integral of $\mathbb{E}$ in K using quadrature. By writing the global matrix as sum of local contributions it is proven that under this assumption the system matrix is of non-negative type, from which the validity of the DMP can be concluded using the results presented in Section 3.1. It can be readily seen that in the special case $\mathbb{E}_K = \mathbb{I}$, (4.5) reduces to the weakly acute angle condition from Definition 2.2. A comprehensive interpretation of (4.5) is provided in [59]. It turns out that (4.5) is equivalent to the requirement that the angles are weakly acute with respect to an inner product induced by $\mathbb{E}_K^{-1}$. Condition (4.5) can be expressed in terms of the map from the reference cell to K. This formulation was utilized in [101] for the construction of appropriate meshes on which the numerical solution satisfies the global DMP.

Later, in [59], the analysis from [101] was refined for the two-dimensional situation in order to obtain a condition weaker than (4.5). The numerical analysis studies the global stiffness matrix, in contrast to the analysis from [101], and in the isotropic case $\mathbb{E}_K = \mathbb{I}$ the resulting condition becomes that the mesh has to be Delaunay. □

4.2. The reaction-diffusion equation and mass lumping. So far the reaction was set to be zero to show the intrinsic link between the geometry of the mesh and the properties of the matrix $\mathbb{A}_\mathrm{d}$. If reaction is added, the satisfaction of the DMP is in fact harder than for the plain diffusion equation, as the next result shows.

LEMMA 4.5 (Sufficient condition for $(\varepsilon\mathbb{A}_\mathrm{d} + \sigma\mathbb{M}_\mathrm{c})^M$ to be of non-negative type). *Let $\mathbb{M}_\mathrm{c}$ be the consistent mass matrix defined in (2.13). Then, $(\varepsilon\mathbb{A}_\mathrm{d} + \sigma\mathbb{M}_\mathrm{c})^M$ is of non-negative type if the mesh family $\{\mathscr{T}_h\}_{h>0}$ is strictly acute and h satisfies*

$$h^2 \leq C\frac{\varepsilon}{\sigma}\,\cot\left(\frac{\pi}{2} - \delta\right) = C\frac{\varepsilon}{\sigma}\,\tan\delta\,, \tag{4.6}$$

where δ is the angle from (2.8), $C = 12$ in 2d, and C depends only on the shape regularity of the mesh family $\{\mathscr{T}_h\}_{h>0}$ in 3d.

Proof. The satisfaction of (3.5) follows from (4.3) and the fact that the row sum of the consistent mass matrix is positive, compare (2.21).

Consider two nodes $\boldsymbol{x}_i \neq \boldsymbol{x}_j$ of a mesh cell $K \in \mathscr{T}_h$. The shape regularity of the mesh implies that there is a constant C_0 such that $|\kappa_{E_{ij}}^K| \geq C_0 h_K^{d-2}$ (note that one can set $C_0 = 1$ if $d = 2$). Since $|K| \leq h_K^d/(d(d-1))$, one obtains using (4.1), the exact formula for the local mass matrix (2.16), and the fact that the cotangent is monotonically decreasing

$$\begin{aligned}\varepsilon\ell_{ij}^K + \sigma m_{ij}^K &= -\varepsilon\frac{|\kappa_{E_{ij}}^K|\cot\theta_{E_{ij}}^K}{d(d-1)} + \sigma\frac{|K|}{(d+1)(d+2)}\\ &\leq h_K^{d-2}\,\frac{(d-2)!}{(d+2)!}\left(-\varepsilon\, C_0\,(d+1)(d+2)\cot(\frac{\pi}{2} - \delta) + \sigma h_K^2\right).\end{aligned}$$

Hence, (4.6) with $C = C_0(d+1)(d+2)$ leads to $\varepsilon\ell_{ij}+\sigma m_{ij} \le 0$ for $i \ne j$, thus proving (3.4). ◻

The last result shows that the presence of a positive reaction term makes the satisfaction of the DMP more difficult than for the Poisson problem. In fact, the presence of the reaction imposes a restriction on the size of the mesh (cf. (4.6)) as well as a stronger restriction on the geometry. While the need for a strictly acute mesh family is clear from the proof, the restriction on the mesh size has been slightly relaxed in, e.g., [23], although some size restriction is always present as long as the consistent mass matrix is used (see [23] for examples of non-satisfaction of the DMP if the mesh is not refined enough). So, we now move onto the presentation of a mass-lumping strategy that allows one to remove the size restriction without affecting accuracy. The mass-lumped discretization of the reaction-diffusion equation reads as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$\varepsilon(\nabla u_h, \nabla v_h) + \sigma(u_h, v_h)_h = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,,$$

with $(\cdot,\cdot)_h$ defined in (2.19). The following result shows that the stiffness matrix of this modified Galerkin discretization is of non-negative type under the same conditions as the stiffness matrix of the pure diffusion problem. Thus, the modification removes the restriction on the mesh size from Lemma 4.5.

Corollary 4.6 (Sufficient and necessary condition for $(\varepsilon\mathbb{A}_{\rm d} + \sigma\mathbb{M}_{\rm l})^M$ to be of non-negative type). *Let $\mathbb{M}_{\rm l}$ be the lumped mass matrix defined in (2.20). Then, a sufficient condition for the matrix $(\varepsilon\mathbb{A}_{\rm d} + \sigma\mathbb{M}_{\rm l})^M$ to be of non-negative type is that the mesh $\mathscr{T}_h$ is of XZ-type. If any internal edge of $\mathscr{T}_h$ has at least one endpoint in Ω, then this condition is necessary.*

Proof. The proof follows by realizing that the lumping process removes the positive off-diagonal entries of $\mathbb{M}_{\rm c}$, and then it becomes a direct application of Theorem 4.1. ◻

Remark 4.7. This section is finished with a brief discussion concerning the fact that an appropriate stabilized method for the reaction-diffusion equation also helps relaxing the mesh conditions for the satisfaction of the DMP, even if it uses the consistent mass matrix. This method, known as Unusual Stabilized finite element method (USFEM), was introduced in [46] and reads as follows: find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$\begin{aligned}(4.7)\qquad \varepsilon\,(\nabla u_h, \nabla v_h) + \sigma\,(u_h, v_h) - \sum_{K\in\mathscr{T}_h} \frac{h_K^2}{\sigma h_K^2+\varepsilon}(\sigma u_h, \sigma v_h)_K \\ = \quad (f, v_h) - \sum_{K\in\mathscr{T}_h} \frac{h_K^2}{\sigma h_K^2+\varepsilon}(f, \sigma v_h)_K \quad \forall\ v_h \in V_{h,0}.\end{aligned}$$

The USFEM improves stability by subtracting a term of reaction type from both sides of the finite element equation. As a consequence, the corresponding matrix $(\mathbb{A})^M$ has the entries

$$a_{ij} = \varepsilon\,(\nabla\phi_j, \nabla\phi_i) + \sum_{K\in\mathscr{T}_h} \frac{\sigma\varepsilon}{\sigma h_K^2+\varepsilon}(\phi_j, \phi_i)_K\,.$$

Following the same steps as in the proof of Lemma 4.5, one can see that $a_{ij} \le 0$

requires the mesh family to be strictly acute and

$$\frac{\varepsilon}{\sigma h_K^2 + \varepsilon} h_K^2 \le C \frac{\varepsilon}{\sigma} \tan\delta \qquad \forall\, K \in \mathscr{T}_h\,, \tag{4.8}$$

where δ is the angle from (2.8) and C is the same as in (4.6). In the interesting case $\varepsilon \ll \sigma$, (4.8) is a much milder condition than (4.6). Moreover, (4.8) does not restrict h_K at all if $C \tan\delta \ge 1$. Likewise important, the sign of the right-hand side of (4.7) is not affected, since it can be written for every basis function ϕ_i as

$$\sum_{K\in\mathscr{T}_h} \frac{\varepsilon}{\sigma h_K^2 + \varepsilon} (f, \phi_i)_K\,.$$

Thus, for a uniform mesh with $h_K = h$ for any $K \in \mathscr{T}_h$, the USFEM is equivalent to replacing ε by $\varepsilon + \sigma h^2$ in the standard Galerkin discretization so that it just adds isotropic linear artificial diffusion of amount σh^2, cf. Section 5.2.

In summary, the USFEM (4.7) preserves the DMP whenever (4.8) is satisfied. □

5. Linear discretizations of the steady-state problem. In this section the main ideas for a linear discretization of the convection-diffusion equation (2.1) are given. It should be kept in mind that the presentation of this and the following sections focuses on the convection-dominated regime, even if this is not always explicitly stated, i.e., ε has to be thought of being (very) small. First, to justify the need for stabilization we describe the standard Galerkin method and make it explicit that, unless the mesh is acute, and prohibitively refined, the DMP cannot hold. So, we then consider stabilized discretizations, where we review linear artificial diffusion, upwind methods, and the edge-averaged finite element method.

5.1. The Galerkin finite element method. The Galerkin finite element method reads as follows: Find $u_h \in V_h$ such that $u_h = i_h g$ on $\partial\Omega$ and

$$a(u_h, v_h) = (f, v_h) \quad \forall v_h \in V_{h,0}\,, \tag{5.1}$$

where $a(\cdot,\cdot)$ is defined in (2.3). Following classical arguments (see, e.g. [41]) one can derive optimal order error estimates, but with a constant that behaves like $\|\boldsymbol{b}\|_{0,\infty,\Omega} h/\varepsilon$, thus making these estimates not useful in practice, and somehow explaining why non-localized spurious oscillations appear in the simulations. This feature is shared by all central discretizations of the convective term (see, e.g., [116] for extensive discussions on this issue).

To illustrate the restrictions of the Galerkin method with respect to the satisfaction of the DMP we focus on the special case where $d = 2$ and $\sigma = 0$. Here, the matrix associated with (5.1) is $(\mathbb{A})^M = (\varepsilon\mathbb{A}_{\mathrm{d}} + \mathbb{A}_{\mathrm{c}})^M$, compare (2.11) and (2.12). Since $\boldsymbol{b}$ is solenoidal, $\mathbb{A}_{\mathrm{c}}$ satisfies

$$c_{ij} = -c_{ji} \qquad \text{for all } i, j = 1, \ldots, M\,, \tag{5.2}$$

i.e., there is a partial antisymmetry.

The next result states that the Galerkin method satisfies the DMP if the mesh family $\{\mathscr{T}_h\}_{h>0}$ is average acute and h is sufficiently small.

THEOREM 5.1 (Conditions on the Galerkin method in 2d to satisfy the DMP). *Suppose that $d = 2$, $\sigma = 0$, the mesh family $\{\mathscr{T}_h\}_{h>0}$ is average acute, and the data and the mesh satisfy: for all $E = K \cap K' \in \mathscr{E}_I$, it holds*

$$\frac{(h_K + h_{K'})\|\boldsymbol{b}\|_{0,\infty,\omega_E}}{3\tan\frac{\delta}{2}} \le \varepsilon\,, \tag{5.3}$$

where δ is the angle from (2.9). Then, the matrix $(\varepsilon\mathbb{A}_{\rm d}+\mathbb{A}_{\rm c})^M$ is of non-negative type and satisfies (3.6).

Proof. Since the basis functions $\phi_1,\dots,\phi_N$ form a partition of unity, $(\varepsilon\mathbb{A}_{\rm d}+\mathbb{A}_{\rm c})^M$ satisfies

$$\sum_{j=1}^N a_{ij} = \varepsilon\,(\nabla 1,\nabla\phi_i) + (\boldsymbol{b}\cdot\nabla 1,\phi_i) = 0\,,\quad i=1,\dots,M\,, \tag{5.4}$$

which proves (3.6). It remains to show (3.4). Let $E=K\cap K'\in\mathscr{E}_I$ with endpoints $\boldsymbol{x}_i,\boldsymbol{x}_j$, $i\in\{1,\dots,M\}$, $j\in\{1,\dots,N\}$. Using (4.2) and $|\kappa_{E_{ij}}|=1$ yields

$$\begin{aligned}\ell_{ij} &= (\nabla\phi_j,\nabla\phi_i)_K + (\nabla\phi_j,\nabla\phi_i)_{K'}\\ &= -\frac{1}{2}\cot\theta_E^K - \frac{1}{2}\cot\theta_E^{K'} = -\frac{\sin(\theta_E^K+\theta_E^{K'})}{2\sin\theta_E^K\sin\theta_E^{K'}}\,.\end{aligned} \tag{5.5}$$

In addition, since $\theta_E^K,\theta_E^{K'}\in(0,\pi)$, one has

$$\begin{aligned}\sin^2\left(\frac{\theta_E^K+\theta_E^{K'}}{2}\right) &= \frac{1-\cos(\theta_E^K+\theta_E^{K'})}{2}\\ &= \frac{1-\cos\theta_E^K\cos\theta_E^{K'}}{2} + \frac{\sin\theta_E^K\sin\theta_E^{K'}}{2} > \frac{\sin\theta_E^K\sin\theta_E^{K'}}{2} > 0\,.\end{aligned} \tag{5.6}$$

Observing that the right-hand side of (5.5) is negative, since the mesh family is average acute and $\theta_E^K,\theta_E^{K'}\in(0,\pi)$, inserting (5.6) in (5.5), and using the monotonicity of the cotangent leads to

$$\begin{aligned}\ell_{ij} &< -\frac{\sin(\theta_E^K+\theta_E^{K'})}{4\sin^2\left(\frac{\theta_E^K+\theta_E^{K'}}{2}\right)} = -\frac{1}{2}\cot\frac{\theta_E^K+\theta_E^{K'}}{2}\\ &\le -\frac{1}{2}\cot\left(\frac{\pi}{2}-\frac{\delta}{2}\right) = -\frac{1}{2}\tan\frac{\delta}{2} < 0\,.\end{aligned} \tag{5.7}$$

Concerning the convective term, a direct calculation using (2.14), Hölder's inequality, and that the diameter of any facet of K is bounded by h_K, gives

$$(\boldsymbol{b}\cdot\nabla\phi_j,\phi_i)_K = -\frac{|F_j^K|}{2|K|}(\boldsymbol{b}\cdot\boldsymbol{n}_j^K,\phi_i)_K \le \frac{h_K\|\boldsymbol{b}\|_{0,\infty,K}}{2|K|}\frac{|K|}{3} \le \frac{h_K\|\boldsymbol{b}\|_{0,\infty,K}}{6}\,. \tag{5.8}$$

From (5.7) and (5.8), one obtains the following upper bound for the off-diagonal matrix entries

$$a_{ij} = \varepsilon\ell_{ij} + c_{ij} \le -\frac{\varepsilon}{2}\tan\frac{\delta}{2} + \frac{(h_K+h_{K'})\|\boldsymbol{b}\|_{0,\infty,\omega_E}}{6} \tag{5.9}$$

and hence $a_{ij}\le 0$ if (5.3) holds. □

Remark 5.2. The geometrical hypothesis on $\mathscr{T}_h$ cannot be relaxed. Indeed, suppose that $\{\mathscr{T}_h\}_{h>0}$ is not average acute and choose an internal edge $E=K\cap K'\in\mathscr{E}_I$ with endpoints $\boldsymbol{x}_i,\boldsymbol{x}_j$, $i,j\in\{1,\dots,M\}$, such that $\theta_E^K+\theta_E^{K'}=\pi$. Then, thanks to (5.5), it follows that $\ell_{ij}=\ell_{ji}=0$. So, since $\mathbb{A}_{\rm c}$ satisfies (5.2), then for any $\boldsymbol{b}$ such that $c_{ij}\neq 0$, one has $c_{ij}>0$ or $c_{ji}>0$, which implies that $\mathbb{A}$ does not satisfy (3.4). □

The discussion in this section shows that the Galerkin method will not satisfy the DMP in any practical situation. These observations were made as early as [77]. On the other hand, supposing the mesh family $\{\mathscr{T}_h\}_{h>0}$ is average acute relaxes the hypotheses made by [77, 33, 26], since in those works the results were proven for strictly acute mesh families.

Remark 5.3. The analysis of [101] and [59] for heterogeneous anisotropic diffusion problems (cf. Remark 4.4) was extended to convection-diffusion-reaction problems in [107]. Since a Galerkin discretization without mass lumping was considered, a condition on the fineness of the mesh appears for the satisfaction of the DMP, cf. Lemma 4.5 and Theorem 5.1. □

Concentrating for a brief discussion of an error estimate on the impact of diffusion and convection, i.e., considering $\sigma = 0$ and homogeneous Dirichlet boundary conditions, one finds under the assumption that $u \in H^2(\Omega)$ that

$$|u - u_h|_{1,\Omega} \le Ch \left(1 + \frac{\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h}{\varepsilon}\right) |u|_{2,\Omega}, \tag{5.10}$$

where C comes from interpolation error estimates in the $L^2(\Omega)$ norm and in the $H^1(\Omega)$ seminorm. The term in the parentheses is very large in the convection-dominated case so that, although (5.10) predicts first order error reduction, the error bound is not useful as long as h is not very small. In fact, large errors can be observed for the Galerkin method on coarse grids if the solution of (2.1) possesses layers.

5.2. Isotropic linear artificial diffusion. Restriction (5.3) can be circumvented by either refining the mesh or making the diffusion of the discrete problem larger. This section will analyze a method that takes the latter approach and adds artificial diffusion to the problem. It will turn out that the diffusion added needs to be of a size proportional to the mesh size. This method will also be supplemented with a mass lumping strategy in order to avoid technical complications due to the presence of reaction.

The following finite element method with added artificial diffusion will be studied: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$a_h(u_h, v_h) + s_h(u_h, v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,, \tag{5.11}$$

where the bilinear form $a_h(\cdot,\cdot)$ is given by

$$a_h(u, v) = \varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u, v) + \sigma\,(u, v)_h\,, \tag{5.12}$$

with $(\cdot,\cdot)_h$ being the mass-lumped inner product defined in (2.19), and the added linear artificial diffusion term is given by

$$s_h(u_h, v_h) = \sum_{K\in\mathscr{T}_h} \tilde{\varepsilon}_K (\nabla u_h, \nabla v_h)_K\,, \quad \tilde{\varepsilon}_K \ge 0\,.$$

In this section we consider the following expression for the added diffusion [77]:

$$\tilde{\varepsilon}_K := \max\left\{ c_0 \frac{h_K \|\boldsymbol{b}\|_{0,\infty,\Omega}}{\tan\frac{\delta}{2}} - \varepsilon, 0 \right\}\,, \tag{5.13}$$

where δ is the constant from (2.9) and $c_0 > 0$ is a constant that is only linked to the shape regularity of the triangulation, see (5.15) below. One notices the close relation between (5.13) and (5.3). In fact, the added diffusion is built in such a way that once the mesh is sufficiently fine, (5.11) reduces to the standard Galerkin method (up to the lumping of the reaction term). Later works proposed slightly different versions of $\tilde{\varepsilon}_K$, e.g., see [33, 26].

The analysis of (5.11) was carried out originally in [77] under the assumption that the mesh families are strictly acute. The analysis presented below is detailed for $d = 2$, and relaxes this hypothesis and requires only average acute mesh families (the case $d = 3$ is discussed in Remark 5.6).

THEOREM 5.4 (DMP for isotropic linear artificial diffusion in 2d). *Let us suppose $d = 2$, that the mesh family is average acute, $\tilde{\varepsilon}_K$ are defined by (5.13), and c_0 is large enough (see (5.15)). Then, (5.11) satisfies the DMP.*

Proof. The proof consists in rewriting method (5.11) as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$\sum_{K\in\mathscr{T}_h} (\varepsilon + \tilde{\varepsilon}_K)(\nabla u_h, \nabla v_h)_K + (\boldsymbol{b}\cdot\nabla u_h, v_h) + \sigma\,(u_h, v_h)_h = (f, v_h) \quad \forall\, v_h \in V_{h,0}\,.$$

Let $i \in \{1, \ldots, M\}$ and $i \neq j \in \{1, \ldots, N\}$. Since the off-diagonal elements of the lumped mass matrix vanish, one gets

$$a_{ij} = \sum_{K\in\mathscr{T}_h} (\varepsilon + \tilde{\varepsilon}_K)(\nabla\phi_j, \nabla\phi_i)_K + c_{ij}\,.$$

Using the notation from the proof of Theorem 5.1 and assuming that $(\nabla\phi_j, \nabla\phi_i)_K \le 0$ and $(\nabla\phi_j, \nabla\phi_i)_{K'} \le 0$, one can use the fact that

$$\varepsilon + \tilde{\varepsilon}_K \ge c_0 \frac{\|\boldsymbol{b}\|_{0,\infty,\Omega} h_K}{\tan\frac{\delta}{2}} \ge \frac{\|\boldsymbol{b}\|_{0,\infty,\Omega}(h_K + h_{K'})}{3\tan\frac{\delta}{2}}\,, \tag{5.14}$$

if

$$c_0 \ge \max_{K,K'\in\mathscr{T}_h : K\cap K'\in\mathscr{E}_I} \frac{h_K + h_{K'}}{3\min\{h_K, h_{K'}\}}\,, \tag{5.15}$$

which is a constant uniformly bounded thanks to the mesh regularity. Then an application of the techniques used to prove Theorem 5.1 shows that the system matrix of method (5.11) is of non-negative type. If, e.g., $(\nabla\phi_j, \nabla\phi_i)_{K'} > 0$, then $(\nabla\phi_j, \nabla\phi_i)_K \le 0$ since the mesh family is average acute. Moreover, since $\theta_E^{K'} \ge \frac{\pi}{2}$, one has $h_{K'} = h_E \le h_K$. Therefore, $\varepsilon + \tilde{\varepsilon}_{K'} \le \varepsilon + \tilde{\varepsilon}_K$ and hence

$$(\varepsilon + \tilde{\varepsilon}_K)(\nabla\phi_j, \nabla\phi_i)_K + (\varepsilon + \tilde{\varepsilon}_{K'})(\nabla\phi_j, \nabla\phi_i)_{K'} \le (\varepsilon + \tilde{\varepsilon}_K)\,\ell_{ij}\,.$$

Now one can apply (5.14) and conclude that $a_{ij} \le 0$ analogously as before. For $\sigma = 0$, the method satisfies (3.6). Finally, the theorem follows from the results of Section 3.1. ∎

Remark 5.5. Once again, the hypothesis on the mesh family being average acute is sharp. In fact, analogous considerations as made in Remark 5.2 hold in this case. □

Remark 5.6. We now briefly discuss the case $d = 3$. For this case one needs to assume that the mesh family $\{\mathscr{T}_h\}_{h>0}$ is strictly acute. Let $\delta > 0$ be the angle from

(2.8), and let the added diffusion be given by

$$\tilde{\varepsilon}_K = \max\left\{ c_0 \frac{h_K \|\boldsymbol{b}\|_{0,\infty,K}}{\tan\delta} - \varepsilon, 0 \right\}.$$

Then, following the same steps as to reach (5.14) and using that $|\kappa^K_{E_{ij}}| \geq C h_K$ (thanks to the mesh regularity) one gets

$$\begin{aligned}
a_{ij} &= \sum_{K\in\mathscr{T}_h} (\varepsilon + \tilde{\varepsilon}_K)(\nabla\phi_j, \nabla\phi_i)_K + c_{ij} \\
&\leq \sum_{K\subset\omega_i\cap\omega_j} \left\{ -\frac{\varepsilon+\tilde{\varepsilon}_K}{6} |\kappa^K_{E_{ij}}| \cot\theta^K_{E_{ij}} + \frac{h_K^2 \|\boldsymbol{b}\|_{0,\infty,K}}{24} \right\} \\
&\leq \sum_{K\subset\omega_i\cap\omega_j} \left\{ -C\, c_0 \frac{h_K \|\boldsymbol{b}\|_{0,\infty,K}}{6\tan\delta} h_K \tan\delta + \frac{h_K^2 \|\boldsymbol{b}\|_{0,\infty,K}}{24} \right\} \\
&= \sum_{K\subset\omega_i\cap\omega_j} h_K^2 \|\boldsymbol{b}\|_{0,\infty,K} \left\{ -\frac{C\, c_0}{6} + \frac{1}{24} \right\}.
\end{aligned}$$

By supposing c_0 is large enough one concludes that $a_{ij} \leq 0$. Thus, in three space dimensions the same result holds as in 2d under the assumption of a strictly acute mesh family. □

The last theorem shows that method (5.11) satisfies the DMP under much milder assumptions than the Galerkin method.

We finish this section with a short comment on an error estimate for method (5.11). We place ourselves in the same situation as in Section 5.1, i.e., $\sigma = 0$, $g = 0$, and $u \in H^2(\Omega)$, and assuming $\tilde{\varepsilon}_K = \tilde{\varepsilon}$ for any $K \in \mathscr{T}_h$, gives the estimate

$$\begin{aligned}
|u - u_h|_{1,\Omega} &\leq Ch\left(1 + \frac{\|\boldsymbol{b}\|_{0,\infty,\Omega}\, h}{\varepsilon+\tilde{\varepsilon}}\right)|u|_{2,\Omega} + \frac{\tilde{\varepsilon}}{\varepsilon+\tilde{\varepsilon}}|u|_{1,\Omega} \\
&\leq Ch\left(1 + \frac{\tan\delta}{c_0}\right)|u|_{2,\Omega} + \frac{\tilde{\varepsilon}}{\varepsilon+\tilde{\varepsilon}}|u|_{1,\Omega},
\end{aligned}$$

where C is again only linked to interpolation error estimates. In contrast to the error estimate (5.10) for the Galerkin method, the factor in front of $|u|_{2,\Omega}$ is of order $\mathcal{O}(1)$. However, due to the consistency error estimated by the term including $|u|_{1,\Omega}$, there is no reduction of the bound proportional to some power of the mesh size as long as the Péclet number $\|\boldsymbol{b}\|_{0,\infty,\Omega} h/\varepsilon$ is large. Note that this second term is strictly monotonically decreasing as $\tilde{\varepsilon}$ tends to zero and eventually it vanishes.

An extension of the linear isotropic diffusion method has recently been proposed in [9]. The interest in this extension by itself is limited, but it opens the door for a LPS-based nonlinear discretization, to be presented in Section 6.4.

5.3. Upwind finite element methods. In this section, one of the earliest proposals for satisfying the DMP in the framework of finite element methods for convection-diffusion equations is reviewed. The basic idea of this method consists in discretizing the convective term in a finite volume manner and utilizing an upwind technique. The first method of this type was developed in [122]. An improved method is presented in [3] and an extension to non-conforming finite elements in [110], see Section 9.3 for more details. Although the methods from [122, 3] were originally proposed for transient problems, compare Section 8.3, we present here their steady-state versions as they contain the main ideas. From the numerical experience reported in the literature, it is known that linear upwind methods lead to solutions with smeared layers, see also Section 7. This situation might explain that, to the best of our knowledge, the methods from [122, 3] are rarely used nowadays. So, their presentation will be kept brief, with an emphasis on the earlier method from [122].

In [122], a two-dimensional problem without reactive term is considered. These assumptions will be relaxed below. In the first step of this method, one defines for an internal node $\boldsymbol{x}_i$ a so-called upwind simplex $K_i^{\rm up}$: $\boldsymbol{x}_i$ is a vertex of $K_i^{\rm up}$ and the straight half-line starting at $\boldsymbol{x}_i$ with direction $-\boldsymbol{b}(\boldsymbol{x}_i)$ intersects $K_i^{\rm up}$. If this line is parallel to a face (edge) F, then one chooses one element of ω_F at random. For nodes at the boundary, the construction is performed analogously. If $-\boldsymbol{b}(\boldsymbol{x}_i)$ points outside the domain, then $\boldsymbol{x}_i$ belongs to the inlet boundary, which means that a Dirichlet condition is imposed at it, and, in turn, the test functions vanish at $\boldsymbol{x}_i$. This means that the upwind simplex can be chosen at random, as this choice will not affect the result. To simplify the presentation, we define the upwind simplex as the empty set in this case. If $\boldsymbol{b}(\boldsymbol{x}_i) = \boldsymbol{0}$, one uses an arbitrary element of ω_i as $K_i^{\rm up}$. The choice of the upwind element is motivated by the following observation. Let $\boldsymbol{x}_j, j \neq i$, be the other vertices of the simplex $K_i^{\rm up}$. By construction, it holds that $|\sphericalangle(-\boldsymbol{b}(\boldsymbol{x}_i), \boldsymbol{n}_i)| < \pi/2$ and $\pi/2 \le |\sphericalangle(-\boldsymbol{b}(\boldsymbol{x}_i), \boldsymbol{n}_j)| < 3\pi/2$ for $j \neq i$, where $\boldsymbol{n}_i$ and $\boldsymbol{n}_j$ are the outer unit normals to the facets of $K_i^{\rm up}$ opposite $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, respectively. From (2.14), it follows that

$$(5.16) \qquad \boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla \phi_i|_{K_i^{\rm up}} \ge 0 \quad \text{and} \quad \boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla \phi_j|_{K_i^{\rm up}} \le 0 \quad \text{for } j \neq i\,,$$

which will be of major importance later. With these definitions, the upwind method reads as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$(5.17) \qquad \varepsilon(\nabla u_h, \nabla v_h) + \sum_{j=1}^{N} \Big(\boldsymbol{b}(\boldsymbol{x}_j) \cdot \nabla u_h|_{K_j^{\rm up}} \psi_j, \mathscr{L} v_h\Big) + \sigma\,(u_h, v_h)_h = (f, v_h)_h\,,$$

for all $v_h \in V_{h,0}$, where ψ_j is the dual basis function defined in (2.17), $\mathscr{L}$ the lumping operator from (2.18), and $(\cdot,\cdot)_h$ the lumped inner product defined in (2.19). The term $\nabla u_h|_{K_j^{\rm up}}$ is defined to be the zero vector if the upwind simplex is the empty set, otherwise it is a constant vector on $K_j^{\rm up}$.

The analysis of the method simplifies greatly if one rewrites the convective term. Noticing that the dual basis functions $\psi_1, \dots, \psi_N$ are orthogonal in $L^2(\Omega)$ and using (2.21), one can see that for every $v_h \in V_h$ the following holds

$$\begin{aligned}
&\sum_{j=1}^{N} \Big(\boldsymbol{b}(\boldsymbol{x}_j) \cdot \nabla u_h|_{K_j^{\rm up}} \psi_j, \mathscr{L} v_h\Big) \\
&\quad = \sum_{i,j=1}^{N} \boldsymbol{b}(\boldsymbol{x}_j) \cdot \nabla u_h|_{K_j^{\rm up}} v_h(\boldsymbol{x}_i)(\psi_j, \psi_i) = \sum_{i=1}^{N} \boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla u_h|_{K_i^{\rm up}} v_h(\boldsymbol{x}_i)|D_i| \\
&\quad = \sum_{i=1}^{N} \boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla u_h|_{K_i^{\rm up}} v_h(\boldsymbol{x}_i)(1, \phi_i) = \sum_{i=1}^{N} (\boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla u_h|_{K_i^{\rm up}}, \phi_i)\, v_h(\boldsymbol{x}_i)\,.
\end{aligned}$$

Thus, method (5.17) can be rewritten as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$\varepsilon(\nabla u_h, \nabla v_h) + \sum_{i=1}^{N} (\boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla u_h|_{K_i^{\mathrm{up}}}, \phi_i)\, v_h(\boldsymbol{x}_i) + \sigma\, (u_h, v_h)_h = (f, v_h)_h\,,$$

for all $v_h \in V_{h,0}$.

The result below establishes well-posedness and the satisfaction of the DMP. In addition, this result also relaxes the hypotheses made on the mesh family from strictly acute to the XZ-criterion.

THEOREM 5.7 (DMP for the upwind finite element method). *Let us suppose that the mesh satisfies the XZ-criterion. Then, the matrix corresponding to the discrete problem* (5.17) *is of non-negative type and hence the solution satisfies the local DMP. In addition, the discrete problem* (5.17) *is well posed and then also the global DMP follows.*

Proof. We will show that $(\varepsilon\, \mathbb{A}_{\mathrm{d}} + \hat{\mathbb{A}}_{\mathrm{c}} + \sigma \mathbb{M}_{\mathrm{l}})^M$, where

$$\hat{\mathbb{A}}_{\mathrm{c}} = (\hat{c}_{ij}) \quad \text{with} \quad \hat{c}_{ij} := (\boldsymbol{b}(\boldsymbol{x}_i) \cdot \nabla \phi_j|_{K_i^{\mathrm{up}}}, \phi_i)\,,$$

is of non-negative type. From Corollary 4.6 it is known that $(\varepsilon\, \mathbb{A}_{\mathrm{d}} + \sigma \mathbb{M}_{\mathrm{l}})^M$ is of non-negative type if the mesh satisfies the XZ-criterion. Moreover, thanks to (5.16) and to the fact that the basis functions form a partition of unity on K_i^{up}, one has for $i, j = 1, \ldots, N$

$$\hat{c}_{ii} \ge 0\,, \qquad \hat{c}_{ij} \le 0 \quad \text{for } i \ne j\,, \quad \text{and} \quad \sum_{j=1}^{N} \hat{c}_{ij} = 0\,.$$

Hence, $\hat{\mathbb{A}}_{\mathrm{c}}$ is also of non-negative type. It follows that $(\varepsilon\, \mathbb{A}_{\mathrm{d}} + \hat{\mathbb{A}}_{\mathrm{c}} + \sigma \mathbb{M}_{\mathrm{l}})^M$ is of non-negative type and since the diagonal entries of this matrix are positive, the method satisfies the local DMP thanks to Theorem 3.4.

Since $\varepsilon(\ell_{ij})_{i,j=1}^M$ is of non-negative type and it is invertible (thanks to Remark 4.3), and $(\hat{c}_{ij})_{i,j=1}^M, (\sigma \tilde{m}_{ij})_{i,j=1}^M$ are of non-negative type, an application of [81, Theorem 5.1] shows that $(\varepsilon \ell_{ij} + \hat{c}_{ij} + \sigma \tilde{m}_{ij})_{i,j=1}^M$ is invertible, which, in turn, implies that (5.17) has a unique solution. Finally, an application of Theorem 3.5 leads to the satisfaction of the global DMP. ∎

Alternative versions of the upwind method for $\mathbb{P}_1$ finite elements have been proposed over the years. For example, in [3], also for time-dependent convection-diffusion equations, a method was proposed motivated by the fact that the exact solution satisfies a discrete analog of a mass conservation property if a special boundary condition is applied, see Section 8.3 for some details. This is an additional feature compared with the method from [122]. Domains $\Omega \subset \mathbb{R}^d$ and triangulations of weakly acute type are considered in [3]. Again, the barycentric cell D_i around a vertex $\boldsymbol{x}_i$ is constructed. Then, appropriate discrete fluxes β_{ij} across the individual parts of ∂D_i are defined, which is a technique from finite volume methods. The discrete convective term has the form

$$\sum_{i=1}^{N} \sum_{j \in S_i} \left(\beta_{ij}^{+} u_h(\boldsymbol{x}_i) + \beta_{ij}^{-} u_h(\boldsymbol{x}_j) \right) v_h(\boldsymbol{x}_i),$$

with S_i defined in (2.4). The coefficients β_{ij} should satisfy several conditions and concrete choices are given in [3]. The off-diagonal entries of the convection matrix are always non-positive and, for a particular choice of the coefficients β_{ij} specified in [3], the row sums of this matrix vanish and thus the convection matrix is of non-negative type. Under these assumptions, the statements of Theorem 5.7 can be transferred literally to the method from [3].

One further upwind method, based on a slightly different choice of the domains for the dual basis, was presented in [75]. A proposal for partial upwinding can be found in [60]. For a unified presentation of upwind finite element methods and some numerical results we refer to [79].

The numerical analysis of several linear finite element upwind schemes can be found in [60], in particular in [60, Section 4.7] for the steady-state convection-diffusion equation ($\sigma = 0$) in two dimensions. The error analysis for one of the methods is presented in detail. For weakly acute triangulations, sufficiently small mesh width, and u being regular enough, the estimate

$$\|u - u_h\|_{0,\infty,\Omega} \le Ch$$

is proved, with C being independent of ε. It is remarked that the same result holds true for the methods from [3, 122]. The d-dimensional convection-diffusion-reaction equation is studied in [3], where the reaction coefficient is assumed to be constant and mass lumping is used for the reactive term. It is proved that there exists a positive constant C, which does not depend on ε, such that

$$\|i_h u - u_h\|_{0,\infty,\Omega} \le Ch\|u\|_{2,p,\Omega}, \quad p > d,$$

if the reaction constant is sufficiently large.

5.4. The edge-averaged finite element method. This section describes the method proposed in [135] and its main properties.

A part of the analysis will be performed under the assumption that the matrix $\mathbb{A}_{\mathrm{d,I}}$ is irreducible. Let us mention that if the mesh is connected (see Definition 2.2), then the diffusion matrix $\mathbb{A}_{\mathrm{d}}$ (including all boundary nodes) is irreducible, compare [39, Rem. 2.3]. As shown in the same paper, this property does not necessarily imply the irreducibility of $\mathbb{A}_{\mathrm{d,I}}$. Despite this, it needs to be considered that the example provided in [39] is rather pathological. In fact, in the same paper it is already noted that refining the mesh once removes the reducibility of $\mathbb{A}_{\mathrm{d,I}}$. Thus, from the available experience, one might state that the reducibility of $\mathbb{A}_{\mathrm{d,I}}$ is an exceptional situation that can be cured by mesh refinements (with the resulting mesh being still very coarse). For this reason, assuming that the matrix $\mathbb{A}_{\mathrm{d,I}}$ is irreducible does not seem to be a big loss of generality.

The following rewriting of the discrete Laplacian matrix $\mathbb{A}_{\mathrm{d}}$, which was at the heart of the proof of Theorem 4.1, will be fundamental for the derivation of the method. Consider any $u_h, v_h \in V_h$ and any $K \in \mathscr{T}_h$, and denote by $\mathscr{I}_K$ the index set of nodes contained in K. Since the local diffusion matrices are symmetric and have zero row sums, a direct calculation using δ_E defined in Section 2.2 yields

$$\begin{aligned}(\nabla u_h, \nabla v_h)_K &= \sum_{i,j\in\mathscr{I}_K} \ell_{ij}^K u_i v_j = \sum_{i,j\in\mathscr{I}_K} \ell_{ij}^K u_i (v_j - v_i) \\ &= \sum_{i,j\in\mathscr{I}_K, i<j} \ell_{ij}^K (u_i - u_j)(v_j - v_i) = -\sum_{i,j\in\mathscr{I}_K, i<j} \ell_{ij}^K \delta_{E_{ij}} u_h \, \delta_{E_{ij}} v_h \,,\end{aligned}$$

where we use the notation $u_i = u_h(\boldsymbol{x}_i)$, $v_i = v_h(\boldsymbol{x}_i)$, $i = 1, \dots, N$. This formula is a sum over the edges of K, where every edge appears exactly once. Hence, denoting

$$\lambda_E^K = \frac{|\kappa_E^K| \cot \theta_E^K}{d(d-1)} \,,$$

it follows from (4.1) that

$$(\nabla u_h, \nabla v_h)_K = \sum_{E \in \mathscr{E}_K} \lambda_E^K \, \delta_E u_h \, \delta_E v_h \,.$$

Consider any $\boldsymbol{a} \in \mathbb{R}^d$ and set $u_h(\boldsymbol{x}) = \boldsymbol{a} \cdot \boldsymbol{x}$. Then $u_h \in V_h$, $\nabla u_h = \boldsymbol{a}$, and $\delta_E u_h = h_E \, \boldsymbol{a} \cdot \boldsymbol{t}_E$ for any $E \in \mathscr{E}_h$. Thus, the previous identity implies that

$$(\boldsymbol{a}, \nabla v_h)_K = \sum_{E \in \mathscr{E}_K} h_E \, \lambda_E^K \, \boldsymbol{a} \cdot \boldsymbol{t}_E \, \delta_E v_h \qquad \forall \, \boldsymbol{a} \in \mathbb{R}^d, \, v_h \in V_h, \, K \in \mathscr{T}_h \,. \tag{5.18}$$

Another fundamental ingredient in the derivation of the method is the consideration of a conservative form of the convective term. We will present, just for simplicity, the case $\sigma = 0$, although the case $\sigma > 0$ is also treated in [135] using a mass-lumping strategy. Then, applying integration by parts, the bilinear form $a(\cdot, \cdot)$ defined in (2.3) satisfies

$$a(u, v) = (\varepsilon \nabla u - \boldsymbol{b} \, u, \nabla v) \qquad \forall \, u \in H^1(\Omega), \, v \in H_0^1(\Omega) \,. \tag{5.19}$$

The quantity $\boldsymbol{J}(u) = \varepsilon \nabla u - \boldsymbol{b} \, u$ is called total flux.

A further ingredient is a function χ_E defined, for each edge $E \in \mathscr{E}_h$, by

$$\frac{\partial \chi_E}{\partial \boldsymbol{t}_E} = - \frac{\boldsymbol{b} \cdot \boldsymbol{t}_E}{\varepsilon} \,,$$

which determines χ_E uniquely up to an additive constant. This definition implies that, for $u \in C^1(\overline{\Omega})$, one has

$$\frac{\partial (e^{\chi_E} u)}{\partial \boldsymbol{t}_E} = \frac{1}{\varepsilon} \, e^{\chi_E} \, \boldsymbol{J}(u) \cdot \boldsymbol{t}_E \,,$$

which leads to

$$\delta_E \left(e^{\chi_E} u \right) = \frac{1}{\varepsilon} \int_E e^{\chi_E} \boldsymbol{J}(u) \cdot \boldsymbol{t}_E \; ds \,.$$

Thus, approximating $\boldsymbol{J}(u)$ on $K \subset \omega_E$ by a constant vector $\boldsymbol{J}_K(u)$ leads to the relation

$$\boldsymbol{J}_K(u) \cdot \boldsymbol{t}_E \approx \varepsilon \, \frac{\delta_E (e^{\chi_E} u)}{\int_E e^{\chi_E} ds} \,. \tag{5.20}$$

Now, using the approximations $\boldsymbol{J}_K(u)$ in (5.19) with $v = v_h \in V_{h,0}$ and applying (5.18) and (5.20) leads to

$$\begin{aligned} a(u, v_h) &\approx \sum_{K \in \mathscr{T}_h} (\boldsymbol{J}_K(u), \nabla v_h)_K = \sum_{K \in \mathscr{T}_h} \sum_{E \in \mathscr{E}_K} h_E \, \lambda_E^K \, \boldsymbol{J}_K(u) \cdot \boldsymbol{t}_E \, \delta_E v_h \\ &\approx \sum_{K \in \mathscr{T}_h} \sum_{E \in \mathscr{E}_K} \lambda_E^K \, \tilde{\varepsilon}_E(\boldsymbol{b}) \, \delta_E (e^{\chi_E} u) \, \delta_E v_h \,, \end{aligned}$$

where

$$\tilde{\varepsilon}_E(\boldsymbol{b}) = \frac{\varepsilon\, h_E}{\int_E e^{\chi_E}\, ds}$$

is the harmonic average of $\varepsilon\, e^{-\chi_E}$ on the edge E. This suggests to introduce the bilinear form

$$a_h(u_h, v_h) = \sum_{E\in\mathscr{E}_h} \left(\sum_{K\subset\omega_E} \lambda_E^K \right) \tilde{\varepsilon}_E(\boldsymbol{b})\, \delta_E(e^{\chi_E} u_h)\, \delta_E v_h\,,$$

which leads to the following Xu–Zikatanov, or edge-averaged, finite element method: Find $u_h \in V_h$, such that $u_h|_{\partial\Omega} = i_h g$, and

$$a_h(u_h, v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,. \tag{5.21}$$

It is worth stressing that if one replaces χ_E by $\chi_E + c$, $c \in \mathbb{R}$, then, in exact arithmetic, the bilinear form $a_h(\cdot,\cdot)$ is not affected. Thus, the fact that χ_E is defined up to an additive constant has no effect in method (5.21). It is observed in [8] that in two dimensions the edge-averaged finite element method is equivalent to the Scharfetter–Gummel finite volume scheme.

For analyzing (5.21), first two properties of its system matrix will be proven. More precisely, we define the matrix $(\mathbb{A})^M = (a_{ij})_{j=1,\dots,N}^{i=1,\dots,M}$ given by $a_{ij} = a_h(\phi_j, \phi_i)$. Then, the following results hold.

Lemma 5.8 (Properties of the system matrix of (5.21)). *If the matrix $\mathbb{A}_{\mathrm{d,I}}$ is irreducible, then the matrix $\mathbb{A}_{\mathrm{I}} = (a_{ij})_{i,j=1}^M$ is irreducible, too. In addition, if the XZ-condition (2.7) is satisfied, the diagonal entries of $\mathbb{A}_{\mathrm{I}} = (a_{ij})_{i,j=1}^M$ are positive.*

Proof. Consider any $i,j \in \{1,\dots,M\}$, $i \neq j$. If $\boldsymbol{x}_i$, $\boldsymbol{x}_j$ are not endpoints of the same edge, then $a_{ij} = 0 = \ell_{ij}$. Otherwise, in view of (4.2),

$$a_{ij} = -\left(\sum_{K\subset\omega_{E_{ij}}} \lambda_{E_{ij}}^K \right) \tilde{\varepsilon}_{E_{ij}}(\boldsymbol{b})\, e^{\chi_{E_{ij}}(\boldsymbol{x}_j)} = \ell_{ij}\, \tilde{\varepsilon}_{E_{ij}}(\boldsymbol{b})\, e^{\chi_{E_{ij}}(\boldsymbol{x}_j)}\,. \tag{5.22}$$

The positivity of the last two factors implies that $a_{ij} = 0$ if and only if $\ell_{ij} = 0$, which proves the first part of the lemma. Furthermore, again in view of (4.2),

$$a_{ii} = \sum_{E\in\mathscr{E}_h:\, \boldsymbol{x}_i\in E} \left(\sum_{K\subset\omega_E} \lambda_E^K \right) \tilde{\varepsilon}_E(\boldsymbol{b})\, e^{\chi_E(\boldsymbol{x}_i)} = -\sum_{j\in S_i} \ell_{ij}\, \tilde{\varepsilon}_{E_{ij}}(\boldsymbol{b})\, e^{\chi_{E_{ij}}(\boldsymbol{x}_i)}$$

for any $i \in \{1,\dots,M\}$. If (2.7) holds, then (4.2) implies that $\ell_{ij} \le 0$ for all $j \neq i$ and since $\ell_{ii} = |\phi_i|_{1,\Omega}^2 > 0$, it follows from (4.3) that $\ell_{ij} < 0$ for at least one index $j \neq i$. Therefore, $a_{ii} > 0$, which finishes the proof. ∎

Theorem 5.9 (M-matrix property of the system matrix of the edge-averaged FEM). *Let the mesh be of XZ-type and let the matrix $\mathbb{A}_{\mathrm{d,I}}$ be irreducible. Then the system matrix of the discretization (5.21) is an M-matrix.*

Proof. First, note that the matrix $\mathbb{A}_{\mathrm{I}}$ is irreducible by Lemma 5.8. We extend the matrix $(\mathbb{A})^M$ to an $N \times N$ matrix by setting $a_{ij} = a_h(\phi_j, \phi_i)$ for all $i,j = 1,\dots,N$. Then the representation (5.22) holds if $j \in S_i$, and $a_{ij} = 0$ if $j \notin S_i \cup \{i\}$. Since $\mathscr{T}_h$

satisfies the XZ-condition (2.7), one observes immediately that $a_{ij} \le 0$ if $j \neq i$ and $i \le M$ or $j \le M$. Moreover, from the definition of δ_E, it follows directly that

$$\sum_{i=1}^{N} a_{ij} = a_h(\phi_j, 1) = 0\,, \qquad j = 1, \ldots, N\,.$$

Since the matrix $\mathbb{A}_{\mathrm{d}}$ is irreducible, there is $\tilde{i} \in \{M+1, \ldots, N\}$ and $\tilde{j} \in \{1, \ldots, M\}$ such that $a_{\tilde{i}\tilde{j}} < 0$, which implies that at least one column sum of $\mathbb{A}_{\mathrm{I}}$ is strictly positive (while the remaining ones are at least non-negative). Hence, $\mathbb{A}_{\mathrm{I}}^T$ is irreducibly diagonally dominant and then, according to [126, Theorem 3.27], $\mathbb{A}_{\mathrm{I}}^T$ is an M-matrix. Consequently, also $\mathbb{A}_{\mathrm{I}}$ is an M-matrix and the theorem follows from Remark 3.14. ∎

The last result generalizes the result presented in [135, Lemma 6.2] where it is shown that the bilinear form $a_h(\cdot, \cdot)$ from (5.21) satisfies an inf-sup condition for sufficiently small h, and thus showing well-posedness of (5.21) for sufficiently refined meshes (although we should mention that this generalization is already hinted in [135, Remark 6.1]).

Remark 5.10. The M-matrix property proved in Theorem 5.9 immediately implies the positivity preservation of the discrete problem (5.21), i.e., if the right-hand side f and the boundary condition g are non-negative, then also the discrete solution u_h is non-negative. However, the M-matrix property does not imply the local or global DMP. The validity of the DMPs follows from Theorems 3.4 and 3.5 if the convection field $\boldsymbol{b}$ is constant since then the validity of (3.6) can be shown. However, in general, the validity of the local and global DMPs is open. □

The discrete problem (5.21) is well-posed under the assumptions of Theorem 5.9 since the system matrix is an M-matrix. In more general situations the well-posedness for sufficiently small mesh sizes is shown in [135]. That paper presents also an error estimate of the form

$$\|i_h u - u_h\|_{1,\Omega} \le C h \left(\sum_{K \in \mathscr{T}_h} |\boldsymbol{J}(u)|^2_{1,p,K} + \sum_{K \in \mathscr{T}_h} |\sigma u|^2_{1,r,K} \right)^{1/2},$$

assuming that the terms on the right-hand side are well defined for sufficiently large values of p and r, where the concrete values depend on the dimension.

6. Nonlinear stabilized discretizations of the steady-state problem. One common feature of all the discretizations presented in the previous section is that they add global stabilizing terms, that is, the methods modify the formulation in the whole domain (equivalently, they modify every row in the system matrix). As a consequence, linear stabilized methods that respect the DMP provide, in general, very diffused solutions. Now, as it was mentioned earlier, in order to prove the DMP, one only needs to analyze the rows of the matrix associated with nodes where an extremum is attained. So, ideally, a method should modify only these rows of the matrix in order to have a good performance. The selection of these rows depends on the solution itself, thus such a method is necessarily nonlinear. This is why in this section we present several nonlinear finite element methods for the convection-diffusion equation that respect the DMP. In contrast to linear methods, some of the nonlinear approaches even satisfy the DMP on general meshes, i.e., without any assumptions on the angles in the meshes.

6.1. The Mizukami–Hughes method. The Mizukami–Hughes method is a nonlinear Petrov–Galerkin method proposed in [109] and improved and further developed in [78, 80, 81]. The idea of the method is to create an upwind effect by means of solution-dependent weighting functions which guarantee that the approximate solution satisfies a linear system with a matrix of non-negative type. Up to the best of our knowledge, this is the first nonlinear DMP-satisfying method proposed for the numerical solution of (2.1). We shall confine ourselves to the two-dimensional case and to $\sigma = 0$. Extensions to $\sigma > 0$ and to three space dimensions can be found in [78].

For any interior node $\boldsymbol{x}_i$, $i \in \{1, \ldots, M\}$, we introduce the weighting function

$$\widetilde{\phi}_i = \phi_i + \sum_{K \subset \omega_i} C_i^K \chi_K \,.$$

Here χ_K denotes the characteristic functions of mesh cells K (i.e., $\chi_K = 1$ in K and $\chi_K = 0$ elsewhere) and C_i^K are constants which will be determined later. The discretization of the convection-diffusion equation reads as follows: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$\varepsilon\,(\nabla u_h, \nabla \phi_i) + (\boldsymbol{b}_h \cdot \nabla u_h, \widetilde{\phi}_i) = (f, \widetilde{\phi}_i)\,, \qquad i = 1, \ldots, M\,, \tag{6.1}$$

where $\boldsymbol{b}_h$ is a piecewise constant approximation of $\boldsymbol{b}$. We shall also use the notation $\boldsymbol{b}_K := \boldsymbol{b}_h|_K$ for $K \in \mathscr{T}_h$. The simplest choice is to set $\boldsymbol{b}_K$ equal to the value of $\boldsymbol{b}$ at the barycenter of K.

The definition of the constants C_i^K is based on the requirement that the local convection matrix $\hat{\mathbb{A}}_c^K$ with entries

$$\hat{c}_{ij}^K = (\boldsymbol{b}_K \cdot \nabla \phi_j, \widetilde{\phi}_i)_K\,, \qquad i = 1, \ldots, M\,,\ \ j = 1, \ldots, N\,,\ \ \boldsymbol{x}_i, \boldsymbol{x}_j \in K\,, \tag{6.2}$$

is of non-negative type. In [109], it was further required that

$$C_i^K \geq -\tfrac{1}{3} \qquad \forall\, i \in \{1, \ldots, N\}\,,\ \boldsymbol{x}_i \in K\,, \qquad\qquad \sum_{\substack{i=1\\ \boldsymbol{x}_i \in K}}^{N} C_i^K = 0\,. \tag{6.3}$$

As we will see, the choice of the constants C_i^K significantly depends on the direction of the convection vector $\boldsymbol{b}_K$ with respect to the edges of K. To characterize the direction of $\boldsymbol{b}_K$, we decompose any triangle K into vertex zones and edge zones by drawing lines parallel to the edges of K which all intersect at the barycenter of K, see Fig. 2. Denoting the vertices of K by $\boldsymbol{x}_1$, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$, the set containing the vertex $\boldsymbol{x}_i$, $i = 1, 2, 3$, will be called vertex zone VZ_i. The remaining three sets are called edge zones and the edge zone opposite the vertex $\boldsymbol{x}_i$ will be denoted by EZ_i. The common part of the boundaries of two adjacent zones is included in the respective vertex zone. The fact that the vector $\boldsymbol{b}_K$ points from the barycenter of K into VZ_i or EZ_i will be shortly expressed by $\boldsymbol{b}_K \in \mathrm{VZ}_i$ or $\boldsymbol{b}_K \in \mathrm{EZ}_i$, respectively. Without loss of generality, one may assume that the vertices of K are numbered in such a way that $\boldsymbol{b}_K \in \mathrm{VZ}_1$ or $\boldsymbol{b}_K \in \mathrm{EZ}_1$ as depicted in Fig. 2.

Using (2.14), it is easy to see that

$$\begin{aligned}
\boldsymbol{b}_K \in \mathrm{VZ}_1 \quad &\Longleftrightarrow \quad \boldsymbol{b}_K \cdot \nabla\phi_1 > 0\,, \quad \boldsymbol{b}_K \cdot \nabla\phi_2 \leq 0\,, \quad \boldsymbol{b}_K \cdot \nabla\phi_3 \leq 0\,,\\
\boldsymbol{b}_K \in \mathrm{EZ}_1 \quad &\Longleftrightarrow \quad \boldsymbol{b}_K \cdot \nabla\phi_1 < 0\,, \quad \boldsymbol{b}_K \cdot \nabla\phi_2 > 0\,, \quad \boldsymbol{b}_K \cdot \nabla\phi_3 > 0\,,
\end{aligned}$$

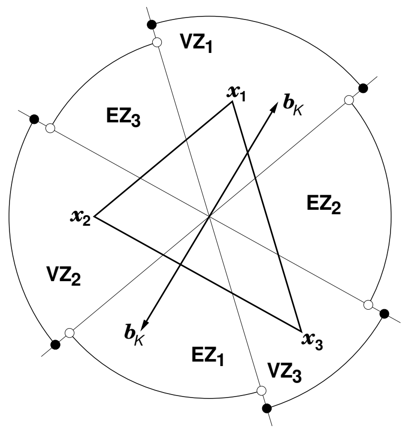


FIG. 2. *Definition of edge zones and vertex zones.*

where we write $\nabla\phi_i$ instead of $\nabla\phi_i|_K$ for simplicity. Note that $\hat{\mathbb{A}}_{\rm c}^K$ has always zero row sums, so that one has to assure only that $\hat{c}_{ij}^K \le 0$ for $i \neq j$. Since

$$\hat{c}_{ij}^K = \boldsymbol{b}_K \cdot \nabla\phi_j|_K \,|K|\,(\tfrac{1}{3} + C_i^K)\,,$$

one observes that, if $\boldsymbol{b}_K \in {\rm VZ}_1$, this condition on $\hat{\mathbb{A}}_{\rm c}^K$ can be easily satisfied by setting

$$C_1^K = \tfrac{2}{3}\,, \qquad C_2^K = C_3^K = -\tfrac{1}{3}\,. \tag{6.4}$$

However, if $\boldsymbol{b}_K \in {\rm EZ}_1$, it is generally not possible to choose the constants C_1^K, C_2^K, C_3^K in such a way that (6.3) holds and $\hat{\mathbb{A}}_{\rm c}^K$ is of non-negative type.

Nevertheless, Mizukami and Hughes [109] made the important observation that u still solves the equation (2.1) if $\boldsymbol{b}$ is replaced by any function $\tilde{\boldsymbol{b}}$ such that $\tilde{\boldsymbol{b}} - \boldsymbol{b}$ is orthogonal to ∇u. This suggests to define the constants C_i^K in such a way that the matrix $\hat{\mathbb{A}}_{\rm c}^K$ is of non-negative type for $\boldsymbol{b}_K$ replaced by a function $\tilde{\boldsymbol{b}}_K$ pointing into a vertex zone and preserving the product $\boldsymbol{b}_K \cdot \nabla u_h|_K$. Note that the local convection matrix $\hat{\mathbb{A}}_{\rm c}^K$ will be still defined using $\boldsymbol{b}_K$ and the vector $\tilde{\boldsymbol{b}}_K$ is used only for defining the constants C_i^K. Since the constants C_i^K depend through $\tilde{\boldsymbol{b}}_K$ on the unknown discrete solution u_h, the resulting discrete problem is nonlinear.

Let us assume that $\boldsymbol{b}_K \in {\rm EZ}_1$ and $\boldsymbol{b}_K \cdot \nabla u_h|_K \neq 0$ and let $\boldsymbol{w} \neq \boldsymbol{0}$ be a vector orthogonal to $\nabla u_h|_K$. We introduce the sets

$$V_k = \{\alpha \in \mathbb{R};\ \boldsymbol{b}_K + \alpha\,\boldsymbol{w} \in {\rm VZ}_k\}\,, \qquad k = 2, 3\,.$$

The vectors $\boldsymbol{b}_K + \alpha\,\boldsymbol{w}$ play the role of $\tilde{\boldsymbol{b}}_K$ mentioned above. Is is easy to see that $V_2 \cup V_3 \neq \emptyset$. Mizukami and Hughes showed that, depending on V_2 and V_3, the following values of the constants C_i^K should be used:

$$V_2 \neq \emptyset \ \ \& \ \ V_3 = \emptyset \quad \Longrightarrow \quad C_2^K = \tfrac{2}{3}\,, \quad C_1^K = C_3^K = -\tfrac{1}{3}\,, \tag{6.5}$$

$$V_2 = \emptyset \ \ \& \ \ V_3 \neq \emptyset \quad \Longrightarrow \quad C_3^K = \tfrac{2}{3}\,, \quad C_1^K = C_2^K = -\tfrac{1}{3}\,, \tag{6.6}$$

$$V_2 \neq \emptyset \ \ \& \ \ V_3 \neq \emptyset \quad \Longrightarrow \quad C_1^K = -\tfrac{1}{3}\,, \quad C_2^K + C_3^K = \tfrac{1}{3}\,, \tag{6.7}$$

$$C_2^K > -\tfrac{1}{3}\,, \quad C_3^K > -\tfrac{1}{3}\,.$$

It was observed in [78] that the definition of C_i^K's proposed in [109] for the case (6.7) depends on the orientation of $\boldsymbol{b}_K$ and $\nabla u_h|_K$ in a discontinuous way. This may deteriorate the quality of the discrete solution and prevent the nonlinear iterative process from converging. Therefore, another definition of these constants was introduced in [78] for which the dependence on the orientation of $\boldsymbol{b}_K$ and $\nabla u_h|_K$ is continuous. To avoid technical digressions, we refer to [78] for details.

It was also demonstrated in [78] that, in some cases, the solutions of the original Mizukami–Hughes method do not approximate boundary layers in a correct way. Therefore, if $\boldsymbol{b}_K$ points into an edge zone, it was proposed to set

$$C_1^K = C_2^K = C_3^K = -\tfrac{1}{3} \tag{6.8}$$

for any mesh cell $K \in \mathscr{T}_h$ having a node on $\partial\Omega$. Except for cases where these mesh cells form a strip along the boundary of an approximately constant width, the definition (6.8) is used also for mesh cells whose all nodes are connected by edges to nodes on $\partial\Omega$. The choice (6.8) suppresses the influence of the Dirichlet boundary condition on the approximate solution inside Ω, which may be important if K lies in the numerical boundary layer.

If $\boldsymbol{b}_K \in \mathrm{EZ}_1$, $\boldsymbol{b}_K \cdot \nabla u_h|_K = 0$ and (6.8) is not used, then one sets

$$C_1^K = -\tfrac{1}{3}\,, \qquad C_2^K = C_3^K = \tfrac{1}{6}\,. \tag{6.9}$$

Finally, one sets $C_1^K = C_2^K = C_3^K = 0$ if $\boldsymbol{b}_K = \mathbf{0}$.

Although the system matrix of (6.1) is in general not of non-negative type, one can prove that, for meshes of XZ-type, the solution vector solves a linear system of the form (3.1)–(3.2) with a non-singular matrix of non-negative type, which implies that the solution of the Mizukami–Hughes method satisfies local and global DMPs.

THEOREM 6.1 (Matrix of non-negative type for the Mizukami–Hughes method). *Let the mesh $\mathscr{T}_h$ be of XZ-type. Then the solution of the Mizukami–Hughes method (6.1) satisfies a linear system of the type (3.1)–(3.2) with $f_i = (f, \widetilde{\phi}_i)$, $i = 1, \dots, M$, and $g_{i-M} = g(\boldsymbol{x}_i)$, $i = M+1, \dots, N$, such that the corresponding system matrix $\mathbb{A}$ given in (3.3) is of non-negative type and its block $\mathbb{A}_{\mathrm{I}}$ is invertible.*

Proof. Let $\boldsymbol{u}$ be the coefficient vector corresponding to the solution of (6.1). We shall show that, for any $K \in \mathscr{T}_h$, there is a matrix $\tilde{\mathbb{A}}_{\mathrm{c}}^K$ of non-negative type such that

$$\tilde{\mathbb{A}}_{\mathrm{c}}^K\, \boldsymbol{u}^K = \hat{\mathbb{A}}_{\mathrm{c}}^K\, \boldsymbol{u}^K\,, \tag{6.10}$$

where $\hat{\mathbb{A}}_{\mathrm{c}}^K$ is defined by (6.2) and $\boldsymbol{u}^K$ consists of the components of $\boldsymbol{u}$ corresponding to nodes of K. If $\boldsymbol{b}_K = \mathbf{0}$ or C_i^K's are defined in (6.4) or (6.8), one can take $\tilde{\mathbb{A}}_{\mathrm{c}}^K = \hat{\mathbb{A}}_{\mathrm{c}}^K$. In case of (6.9) which is used if $\boldsymbol{b}_K \cdot \nabla u_h|_K = 0$, one can set $\tilde{\mathbb{A}}_{\mathrm{c}}^K = 0$. It remains to define $\tilde{\mathbb{A}}_{\mathrm{c}}^K$ in cases when the constants C_i^K are defined by (6.5)–(6.7), which assumes that $\boldsymbol{b}_K \in \mathrm{EZ}_1$ and $\boldsymbol{b}_K \cdot \nabla u_h|_K \neq 0$. First, we introduce some auxiliary notation. If, for some $k \in \{2, 3\}$, the set V_k is non-empty, we choose $\alpha_k \in V_k$ and define the matrix $\tilde{\mathbb{A}}_{\mathrm{c}}^{K,k}$ with entries

$$\tilde{c}_{ij}^{K,k} = (\boldsymbol{b}_K + \alpha_k\, \boldsymbol{w}) \cdot \nabla\phi_j|_K\, |K|\, (\tfrac{1}{3} + C_i^{K,k})\,, \qquad i, j = 1, 2, 3 \;\; (\boldsymbol{x}_i \in \Omega)\,,$$

where $C_i^{K,2}$ are defined as in (6.5) and $C_i^{K,3}$ as in (6.6). If $V_k = \emptyset$, we set $\tilde{\mathbb{A}}_{\mathrm{c}}^{K,k} = 0$. Then the matrices $\tilde{\mathbb{A}}_{\mathrm{c}}^{K,2}$ and $\tilde{\mathbb{A}}_{\mathrm{c}}^{K,3}$ are of non-negative type and hence also

$$\tilde{\mathbb{A}}_{\mathrm{c}}^K := (\tfrac{1}{3} + C_2^K)\, \tilde{\mathbb{A}}_{\mathrm{c}}^{K,2} + (\tfrac{1}{3} + C_3^K)\, \tilde{\mathbb{A}}_{\mathrm{c}}^{K,3}$$

is of non-negative type. Since $\boldsymbol{w} \cdot \nabla u_h|_K = 0$ and

$$(\tfrac{1}{3} + C_2^K)(\tfrac{1}{3} + C_i^{K,2}) + (\tfrac{1}{3} + C_3^K)(\tfrac{1}{3} + C_i^{K,3}) = \tfrac{1}{3} + C_i^K\,, \quad i = 1, 2, 3\,,$$

one obtains (6.10).

The matrices $\hat{\mathbb{A}}_{\mathrm{c}}^K$ and $\tilde{\mathbb{A}}_{\mathrm{c}}^K$ are assembled to $M \times N$ matrices $\hat{\mathbb{A}}_{\mathrm{c,MH}}$ and $\tilde{\mathbb{A}}_{\mathrm{c,MH}}$ for which $\hat{\mathbb{A}}_{\mathrm{c,MH}}\,\boldsymbol{u} = \tilde{\mathbb{A}}_{\mathrm{c,MH}}\,\boldsymbol{u}$ and $\tilde{\mathbb{A}}_{\mathrm{c,MH}}$ is of non-negative type. Since $\boldsymbol{u}$ corresponds to the solution of (6.1), one has $(\varepsilon\,(\mathbb{A}_{\mathrm{d}})^M + \hat{\mathbb{A}}_{\mathrm{c,MH}})\,\boldsymbol{u} = \boldsymbol{f}$ with $\boldsymbol{f} = (f_1, \ldots, f_M)$ introduced in the formulation of the theorem. As $\mathscr{T}_h$ is of XZ-type, the matrix $(\mathbb{A}_{\mathrm{d}})^M$ is of non-negative type according to Theorem 4.1. Thus $\boldsymbol{u}$ also satisfies the linear system $(\varepsilon\,(\mathbb{A}_{\mathrm{d}})^M + \tilde{\mathbb{A}}_{\mathrm{c,MH}})\,\boldsymbol{u} = \boldsymbol{f}$ and the matrix $\mathbb{A}^M := \varepsilon\,(\mathbb{A}_{\mathrm{d}})^M + \tilde{\mathbb{A}}_{\mathrm{c,MH}}$ is of non-negative type. Since the block $\mathbb{A}_{\mathrm{d,I}}$ of $\mathbb{A}_{\mathrm{d}}$ is invertible (cf. Remark 4.3), it follows that also $\mathbb{A}_{\mathrm{I}}$ is invertible (see [81, Theorem 5.1]). This finishes the proof. ∎

As discussed in [81], the Mizukami–Hughes method corresponds to the discretization of the convective term by standard upwind differencing. This is appropriate if the diffusion ε is small in comparison to $\boldsymbol{b}$. However, if this is not the case, such a discretization leads to a low accuracy since too much artificial diffusion is introduced. Therefore, in [81], the constants C_i^K were defined in such a way that the matrix $\widetilde{\varepsilon}\,\mathbb{A}_{\mathrm{d}}^K + \hat{\mathbb{A}}_{\mathrm{c}}^K$ is of non-negative type, where $\mathbb{A}_{\mathrm{d}}^K$ is the local diffusion matrix and $\widetilde{\varepsilon} \in (0, \varepsilon)$ is close to ε. This does not change the method much in the convection-dominated case but improves the accuracy if ε is not small.

To the best of our knowledge, there are no error estimates available for the Mizukami–Hughes method. Also, the solvability of the nonlinear problem seems to be still an open problem.

6.2. Burman–Ern Methods. In this section we will present the finite element method, based on a continuous interior penalty idea, presented in [28]. The analysis of this method requires the mesh to be of XZ-type, so we will assume that throughout this section. In the work [28] the method is presented with two stabilizations, namely, a linear one (e.g., SUPG or CIP), and the nonlinear stabilizing term responsible for the DMP. To keep the discussion brief, we will start discussing the case of the reduced method, that is, the method only adds the nonlinear stabilization to the Galerkin formulation. The proof of the local DMP (cf. Theorem 3.18) is achieved by proving that the nonlinear problem satisfies the weak DMP property (cf. Definition 3.16). So, as a motivation for the definition of the method we will now suppose that $u_h(\boldsymbol{x}_i) < 0$, $i \in \{1, \ldots, M\}$, is a local minimum in ω_i and will bound $a(u_h, \phi_i)$. Thanks to the fact that the mesh is of XZ-type one has $\ell_{ij} \le 0$ for all $j \ne i$ (cf. Theorem 4.1), and consequently

$$(\nabla u_h, \nabla\phi_i) = \sum_{j\in S_i} \ell_{ij}(u_h(\boldsymbol{x}_j) - u_h(\boldsymbol{x}_i)) \le 0\,. \qquad (6.11)$$

In addition, if the function u_h changes sign inside $K \subset \omega_i$, using a Taylor expansion at a zero of u_h, one gets

$$(u_h, \phi_i)_K \le \frac{|K|}{d+1}\, h_K\, \left|\nabla u_h|_K\right|.$$

If $u_h \le 0$ in K then one just bounds $(u_h, \phi_i)_K \le 0$. The convective term is bounded in a similar way leading to

$$(\boldsymbol{b} \cdot \nabla u_h + \sigma u_h, \phi_i) \le \frac{1}{d+1} \sum_{K \subset \omega_i} \big(\|\boldsymbol{b}\|_{0,\infty,K} + \sigma\, h_K \big) |K| \, \big|\nabla u_h|_K\big| .$$

Next, to bound the gradient of u_h in the last inequality one uses that $u_h(\boldsymbol{x}_i)$ is a local minimum and then the following bound holds (see [28, Lemma 2.7] for the proof):

$$|\nabla u_h|_K| \le \sum_{F \in \mathscr{F}_i} |[\![\nabla u_h]\!]_F| \qquad \forall\ K \subset \omega_i ,$$

which leads to

$$\begin{aligned} a(u_h, \phi_i) &\le \frac{1}{d+1} \sum_{F \in \mathscr{F}_i} \sum_{K \subset \omega_i} \big(\|\boldsymbol{b}\|_{0,\infty,K} + \sigma\, h_K \big)\, |K| \, |[\![\nabla u_h]\!]_F| \qquad (6.12) \\ &\le \frac{1}{d+1} \sum_{F \in \mathscr{F}_i} \big(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\, \sigma\, h_F \big) |\omega_i| \, |[\![\nabla u_h]\!]_F| , \end{aligned}$$

where we used the fact that, in view of (2.5), one has $h_K \le \rho\, h_F$ for any $K \subset \omega_i$ and $F \in \mathscr{F}_i$. Since $|\omega_i| \le \Omega_d \, (\max_{K \subset \omega_i} h_K)^d$, where Ω_d is the measure of the unit ball in $\mathbb{R}^d$, one has $|\omega_i| \le \Omega_d\, \rho^d\, h_F^d$ for any $F \in \mathscr{F}_i$. Using the mesh regularity, one gets $|\omega_i| \le C\, \rho^d\, h_F\, |F|$, which gives

$$a(u_h, \phi_i) \le \frac{C \rho^d}{d+1} \sum_{F \in \mathscr{F}_i} \big(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\, \sigma\, h_F \big) h_F |F| \, |[\![\nabla u_h]\!]_F| . \qquad (6.13)$$

From the discussion above, one sees that in order to prove the DMP, one needs to control a term related to the jumps of the gradients of the discrete solution across the facets containing the local extrema. Motivated by this observation, in [28] the following method is proposed: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$a(u_h, v_h) + j_h(u_h; v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0} . \qquad (6.14)$$

Here, $j_h(\cdot;\cdot)$ is a stabilizing form given by

$$j_h(u_h; v_h) = c_\rho \sum_{F \in \mathscr{F}_I} \big(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\, \sigma\, h_F \big)\, h_F \, \big(|[\![\nabla u_h]\!]_F| \, , b_F(u_h; v_h) \big)_F , \qquad (6.15)$$

$$b_F(u_h; v_h) = \sum_{E \in \mathscr{E}_F} h_E \, \mathrm{sign}(\nabla u_h \cdot \boldsymbol{t}_E) \nabla v_h \cdot \boldsymbol{t}_E . \qquad (6.16)$$

The parameter $c_\rho > 0$ depends on the mesh regularity through the quantity ρ. Using a regularized problem and Brouwer's fixed-point theorem in [28] it is proven that (6.14) admits at least one solution. Under the hypothesis that the mesh is of XZ-type, the following result regarding the local DMP can be shown.

THEOREM 6.2 (Local DMP for the Burman–Ern method). *Let us suppose that the mesh is of XZ-type. Then, if c_ρ is sufficiently large, the nonlinear form $j_h(\cdot;\cdot)$ satisfies the weak DMP property if $\sigma > 0$ and the strong DMP property if $\sigma = 0$. Consequently, method (6.14) satisfies the local DMP from Theorem 3.18.*

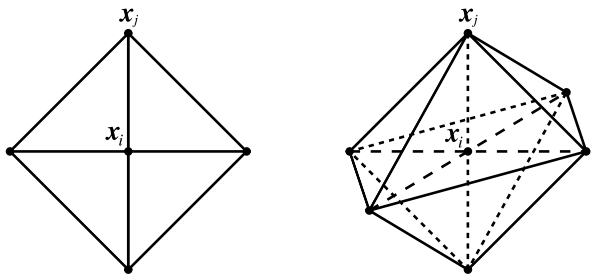


FIG. 3. *Examples of patches ω_i in 2d and 3d.*

Proof. Let us suppose that $u_h \in V_h$ has a strict local minimum at the interior node $\boldsymbol{x}_i$. Then, for any $F \in \mathscr{F}_i$, one has

$$b_F(u_h;\phi_i) = \sum_{j\in S_i:\, E_{ij}\subset F} \operatorname{sign}\big(u_h(\boldsymbol{x}_j) - u_h(\boldsymbol{x}_i)\big)\,\big(\phi_i(\boldsymbol{x}_j) - \phi_i(\boldsymbol{x}_i)\big) = -(d-1)\,, \tag{6.17}$$

since $\operatorname{card}\{j \in S_i : E_{ij} \subset F\} = d-1$. This implies that

$$j_h(u_h;\phi_i) \le -c_\rho \sum_{F\in\mathscr{F}_i} \big(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\,\sigma\,h_F\big)\,h_F\,|F|\,|[\![\nabla u_h]\!]_F|\,.$$

Thus, combining this last bound with (6.13) (which was derived for $u_h(\boldsymbol{x}_i) < 0$ but holds also for $u_h(\boldsymbol{x}_i) \ge 0$ if $\sigma = 0$) gives

$$a(u_h,\phi_i) + j_h(u_h;\phi_i) \le \left(\frac{C\rho^d}{d+1} - c_\rho\right) \sum_{F\in\mathscr{F}_i} \big(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\,\sigma\,h_F\big)\,h_F\,|F|\,|[\![\nabla u_h]\!]_F|\,,$$

and the proof follows choosing c_ρ large enough provided that $\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\,\sigma\,h_F > 0$ for all $F \in \mathscr{F}_i$. If this is not the case, one can employ the fact that the previous inequality holds with the term $\varepsilon\,(\nabla u_h, \nabla\phi_i)$ on the right-hand side. When deriving (6.12), this term was estimated by (6.11). However, since now $u_h(\boldsymbol{x}_i)$ is a strict local minimum, it follows from (6.11) that $(\nabla u_h, \nabla\phi_i)$ is negative and hence can be estimated by $-\sum_{F\in\mathscr{F}_i} \alpha_F\,|[\![\nabla u_h]\!]_F|$ with suitable positive constants α_F. This finishes the proof. ∎

Remark 6.3. The validity of the global DMP seems to be open for method (6.14) since, in general, the stabilizing form $j_h(\cdot;\cdot)$ defined in (6.15) does not allow to prove the strong and weak DMP properties for non-strict extrema formulated in Definition 3.17. To see this, let us consider the patches ω_i depicted in Fig. 3. Let us decompose ω_i into the sets

$$\omega_i^1 = \cup\{K \subset \omega_i : \boldsymbol{x}_j \in K\}\,, \qquad \omega_i^2 = \cup\{K \subset \omega_i : \boldsymbol{x}_j \notin K\}\,.$$

Let $u_h \in V_h$ be such that $u_h(\boldsymbol{x}_j) \ne u_h(\boldsymbol{x}_i)$ and $u_h(\boldsymbol{x}_k) = u_h(\boldsymbol{x}_i)$ for any vertex $\boldsymbol{x}_k \in \omega_i^2$. Then $u_h \in \mathbb{P}_1(\omega_i^1)$, u_h is constant in ω_i^2, it is not constant in ω_i, and attains a local extremum at $\boldsymbol{x}_i$. Consequently, $b_F(u_h;v_h) = 0$ for any $F \subset \omega_i^2$ and any $v_h \in V_h$. On the other hand, for any $F \in \mathscr{F}_i$ such that $\boldsymbol{x}_j \in F$, one has $[\![\nabla u_h]\!]_F = \boldsymbol{0}$ since ∇u_h is constant in ω_i^1. Thus, $j_h(u_h,\phi_i) = 0$ which means that the term j_h cannot be used to enforce the strong or weak DMP property for non-strict extrema at the node $\boldsymbol{x}_i$. □

An alternative definition, hinted in [28, Theorem 3.5], and developed further in [29, Section 2.4], can be obtained by replacing $|[\![\nabla u_h]\!]_F|$ in (6.15) by

$$m_F(u_h) = \max_{F' \in \mathscr{F}_I :\, F' \subset \omega_F} |[\![\nabla u_h]\!]_{F'}| \,.$$

Then

$$j_h(u_h; v_h) = c_\rho \sum_{F \in \mathscr{F}_I} \left(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F}\, h_F + \rho\, \sigma\, h_F^2\right) (m_F(u_h)\,, b_F(u_h; v_h)\,)_F \,. \tag{6.18}$$

For this stabilizing term, one can prove also the DMP properties for non-strict extrema formulated in Definition 3.17.

THEOREM 6.4 (DMP for (6.18)). *Let us suppose that the mesh is of XZ-type. Then, if c_ρ is sufficiently large, the nonlinear form $j_h(\cdot;\cdot)$ defined in (6.18) satisfies the weak DMP property for non-strict extrema if $\sigma > 0$ and the strong DMP property for non-strict extrema if $\sigma = 0$. Consequently, method (6.14) with $j_h(\cdot;\cdot)$ from (6.18) satisfies both the local and the global DMPs from Theorem 3.18.*

Proof. Let us suppose that $u_h \in V_h$ has a local minimum at the interior node $\boldsymbol{x}_i$. Then, for any $F \in \mathscr{F}_i$, one has

$$b_F(u_h; \phi_i) = \sum_{j \in S_i :\, E_{ij} \subset F} \operatorname{sign}\left(u_h(\boldsymbol{x}_j) - u_h(\boldsymbol{x}_i)\right) \left(\phi_i(\boldsymbol{x}_j) - \phi_i(\boldsymbol{x}_i)\right) \le 0 \,.$$

Consider any $F \in \mathscr{F}_i$. If $[\![\nabla u_h]\!]_F \neq \boldsymbol{0}$, then there exists a vertex $\boldsymbol{x}_j \in \omega_F$ such that $u_h(\boldsymbol{x}_j) \neq u_h(\boldsymbol{x}_i)$. Let $F'' \subset \omega_F$ be a facet such that $\boldsymbol{x}_i, \boldsymbol{x}_j \in F''$. Then

$$b_{F''}(u_h; \phi_i) \le -\operatorname{sign}\left(u_h(\boldsymbol{x}_j) - u_h(\boldsymbol{x}_i)\right) = -1 \,.$$

Since $F \subset \omega_{F''}$, one gets

$$|[\![\nabla u_h]\!]_F| \le -m_{F''}(u_h)\, b_{F''}(u_h; \phi_i) \,.$$

If $[\![\nabla u_h]\!]_F = \boldsymbol{0}$, then this inequality holds with any $F'' \in \mathscr{F}_i$ satisfying $F'' \subset \omega_F$ since the right-hand side is nonnegative. Hence one finds that

$$\begin{aligned} \sum_{F \in \mathscr{F}_i} |[\![\nabla u_h]\!]_F| &\le - \sum_{F \in \mathscr{F}_i} m_{F''}(u_h)\, b_{F''}(u_h; \phi_i) \\ &\le -(2\,d-1) \sum_{F \in \mathscr{F}_i} m_F(u_h)\, b_F(u_h; \phi_i) \end{aligned} \tag{6.19}$$

as the number of facets $F \in \mathscr{F}_i$ satisfying $F \subset \omega_{F''}$ for a given $F'' \in \mathscr{F}_i$ is $2\,d-1$. Using this estimate in the first inequality of (6.12) (which was derived for $u_h(\boldsymbol{x}_i) < 0$ but holds also for $u_h(\boldsymbol{x}_i) \ge 0$ if $\sigma = 0$) and performing the same manipulations as used to derive (6.13), one obtains

$$a(u_h, \phi_i) \le -C\, \rho^d\, \frac{2\,d-1}{d+1} \sum_{F \in \mathscr{F}_i} \left(\|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F} + \rho\, \sigma\, h_F\right) h_F\, (m_F(u_h)\,, b_F(u_h; \phi_i)\,)_F \,,$$

where C is the same constant as in (6.13). Thus, $a(u_h, \phi_i) + j_h(u_h; \phi_i) \le \frac{1}{2} j_h(u_h; \phi_i)$ if $c_\rho \ge 2\, C\, \rho^d\, (2\,d-1)/(d+1)$. According to (6.19), one has

$$j_h(u_h; \phi_i) \le -\frac{c_\rho}{2\,d-1} \min_{F \in \mathscr{F}_i} \left\{ \left(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\tilde{\omega}_F}\, h_F + \rho\, \sigma\, h_F^2\right) |F| \right\} \sum_{F \in \mathscr{F}_i} |[\![\nabla u_h]\!]_F| \,,$$

which completes the proof. ∎

Remark 6.5. The methods just analyzed need the mesh to be of XZ-type. To avoid this restriction, in [27] the following method was proposed for the Poisson problem: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$(6.20)\qquad (\nabla u_h, \nabla v_h) + \delta \sum_{F\in\mathscr{F}_I} (|[\![\nabla u_h]\!]_F|\,, b_F(u_h; v_h)\,)_F = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,,$$

where b_F is defined as in (6.16) and $\delta > 0$. Then, for $\delta > \frac{1}{d(d-1)}$, method (6.20) satisfies the strong DMP property for any mesh. In fact, the main argument of the proof is the following observation from [27]: regardless of the mesh,

$$(6.21)\qquad (\nabla u_h, \nabla \phi_i) = \sum_{F\in\mathscr{F}_i} \big([\![\nabla u_h]\!]_F \cdot \boldsymbol{n}_F, \phi_i)_F = \sum_{F\in\mathscr{F}_i} \frac{|F|}{d} [\![\nabla u_h]\!]_F \cdot \boldsymbol{n}_F\,,$$

where $\boldsymbol{n}_F$ is the unit normal vector to F in the direction corresponding to the orientation of the jump $[\![\cdot]\!]_F$. So, if u_h has a strict local minimum at an interior node $\boldsymbol{x}_i$, it follows from (6.17) that

$$(\nabla u_h, \nabla \phi_i) + \delta \sum_{F\in\mathscr{F}_I} (|[\![\nabla u_h]\!]_F|\,, b_F(u_h; \phi_i)\,)_F \le \sum_{F\in\mathscr{F}_i} \left(\frac{1}{d} - \delta(d-1)\right) |F|\,|[\![\nabla u_h]\!]_F|\,.$$

Thus, for $\delta > \frac{1}{d(d-1)}$ (6.20) satisfies the strong DMP criterion.

The main difference between (6.20) and (6.14) resides on the size of the stabilization term. In fact, only considering the powers of h involved, the stabilization given in (6.20) is one size larger than the one from (6.14), as (6.20) is designed to match the behavior of the diffusion matrix given by (6.21). So, even if this term is positive (as it would happen if a mesh that is not of XZ-type is used), then the stabilization is large enough to compensate for that. Even if in [27] an extension to the convection-diffusion equation has been studied, this variant does not seem to have been applied to convection-dominated problems in later years. □

Method (6.14) is the simplest form of a Burman–Ern method that respects the local DMP. In the presence of dominating convection, sometimes it is recommended to first add a linear stabilization term to stabilize the convection, and only then to add a nonlinear stabilization to ensure the satisfaction of the DMP. With this objective in mind, this approach was pursued in [28] by using a linear stabilization which can be given by the SUPG or CIP stabilization. We now summarize briefly the results proven for the latter option. The CIP stabilizing term is defined as follows (see, e.g., [29])

$$s_h(u_h, v_h) = \sum_{F\in\mathscr{F}_I} \gamma_{\text{cip}}\, \|\boldsymbol{b}\|_{0,\infty,\Omega} h_F^2\, ([\![\nabla u_h]\!]_F, [\![\nabla v_h]\!]_F)_F\,,$$

where $\gamma_{\text{cip}} > 0$. Using this stabilizing term, the following stabilized method is proposed in [28]: Find $u_h \in V_h$ such that $u_h|_{\partial\Omega} = i_h g$, and

$$(6.22)\qquad a(u_h, v_h) + s_h(u_h, v_h) + j_h(u_h; v_h) = (f, v_h) \qquad \forall\, v_h \in V_{h,0}\,,$$

with $j_h(\cdot;\cdot)$ being a combination of (6.15) and (6.18). The corresponding analogue of Theorem 6.4 was proven for (6.22) in [28, Theorem 3.5]. For the diffusion-dominated regime, i.e., with the assumption $ch \le \varepsilon$ for some appropriate constant c, the following error estimate appears as a corollary of [28, Theorem 3.10]:

$$\begin{aligned}(6.23)\qquad \varepsilon^{\frac{1}{2}}|u-u_h|_{1,\Omega} + \sigma^{\frac{1}{2}}\|u-u_h\|_{0,\Omega} + \|h^{\frac{1}{2}}\boldsymbol{b}\cdot\nabla(u-u_h)\|_{0,\Omega} \\ + s_h(u-u_h,u-u_h)^{\frac{1}{2}} \le C\left(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega}\, h + \sigma\, h^2\right)^{\frac{1}{2}} h\,\|u\|_{2,\Omega}\,,\end{aligned}$$

where $C > 0$ is independent of h and all the physical parameters, provided that the exact solution u belongs to $H^2(\Omega)$.

The combination of linear and nonlinear stabilizations has two main effects in this context. First, the addition of the linear stabilization term $s_h(\cdot,\cdot)$ allows for the extra control on the convective term appearing in (6.14), which is responsible for the estimate (6.23). This control is not possible to achieve if only the nonlinear stabilization $j_h(\cdot,\cdot)$ is used. The second main effect is computational. It can be observed that, while the nonlinear stabilization $j_h(\cdot,\cdot)$ is local (in the sense that it is active mostly in the vicinity of extrema and layers), the linear stabilization term $s_h(\cdot,\cdot)$ is global, and thus it helps dampening oscillations that appear away from the layers.

Remark 6.6. Finally, it is worth mentioning that the works reviewed in this section were not the first effort that was made in this direction by the authors. In fact, in their previous paper [26] the authors proposed a nonlinear diffusion method that, under the assumption of acute meshes, satisfies the global DMP. To improve the convergence of the nonlinear solver, absolute values in the nonlinear terms were regularized, which however leads to a violation of the DMP. Comprehensive numerical tests of three variants of the methods from [26] can be also found in [66, 67]. In particular, in [67], the authors did not succeed to solve the respective nonlinear problems in a number of cases. □

6.3. Algebraic Flux Correction methods. Algebraic flux correction (AFC) methods belong to the class of algebraically stabilized schemes which have been intensively developed in recent years, see, e.g., [4, 13, 52, 83, 86, 87, 89, 90, 91, 96, 98, 104]. In contrast to the methods discussed in the previous sections, the stabilization is not introduced in a variational form but the starting point is the system of linear algebraic equations corresponding to the Galerkin FEM discretization. Then, a nonlinear algebraic term is added to the linear system in order to enforce a DMP without an excessive smearing of the layers.

Let $\mathbb{A}_{\rm N}$ be the matrix corresponding to the standard Galerkin FEM (5.1) with Neumann boundary conditions, i.e.,

$$(6.24)\qquad \mathbb{A}_{\rm N} = \varepsilon\mathbb{A}_{\rm d} + \mathbb{A}_{\rm c} + \sigma\mathbb{M}_{\rm c}\,.$$

We will also consider a lumping of the reaction term in (5.1), which leads to a matrix given by

$$(6.25)\qquad \mathbb{A}_{\rm N} = \varepsilon\mathbb{A}_{\rm d} + \mathbb{A}_{\rm c} + \sigma\mathbb{M}_{\rm l}\,.$$

The discrete problem is then equivalent to the system (3.1), (3.2), where $f_i = (f,\phi_i)$ for $i = 1,\dots,M$ and $g_{i-M} = g(\boldsymbol{x}_i)$ for $i = M+1,\dots,N$. To derive an AFC scheme, first a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ is introduced by

$$(6.26)\qquad d_{ij} = -\max\{0, a_{ij}, a_{ji}\} \quad \text{for } i \ne j, \qquad d_{ii} = -\sum_{j=1, j\ne i}^{N} d_{ij}\,.$$

Hence $\mathbb{D}$ has zero row and column sums and the matrix $\mathbb{A}_{\mathrm{N}} + \mathbb{D}$ is of non-negative type. Thus, replacing $\mathbb{A}_{\mathrm{N}}$ by $\mathbb{A}_{\mathrm{N}} + \mathbb{D}$ in (3.1), one obtains the stabilized problem

$$(\mathbb{A}_{\mathrm{N}} + \mathbb{D})^M \boldsymbol{u} = \boldsymbol{f}$$

satisfying the DMP (with $\boldsymbol{f} = (f_1, \ldots, f_M)^T$). However, like for the similar linear artificial diffusion method of Section 5.2, the added artificial diffusion is usually too large and leads to an excessive smearing of layers. Therefore, it is necessary to restrict the artificial diffusion to regions where the solution changes abruptly. Since these regions are not known a priori, this will again lead to a nonlinear method.

The original derivation of the AFC method, e.g., in [87], is performed in such a way that first the term $(\mathbb{D}\boldsymbol{u})_i$ is added to both sides of (3.1) leading to

$$(\mathbb{A}_{\mathrm{N}} + \mathbb{D})^M \boldsymbol{u} = \boldsymbol{f} + \mathbb{D}^M \boldsymbol{u}\,, \tag{6.27}$$

and then the identity

$$(\mathbb{D}\boldsymbol{u})_i = \sum_{j=1}^N f_{ij} \qquad \text{with} \qquad f_{ij} = d_{ij}\,(u_j - u_i)$$

is used. The quantities f_{ij} are called fluxes since they can be interpreted as quantities which correspond to the intensity of the flow of u between the nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, see also the explanation of the concept of fluxes at the beginning of Section 8.4. It turns out that spurious oscillations in the approximate solution can be suppressed by damping the above-introduced fluxes f_{ij} appearing on the right-hand side of (6.27). This damping is often called limiting and it is achieved by multiplying the fluxes by solution-dependent correction factors $\alpha_{ij} \in [0,1]$ called limiters. This leads to the nonlinear algebraic problem

$$\sum_{j=1}^N a_{ij}\,u_j + \sum_{j=1}^N (1 - \alpha_{ij}(\boldsymbol{u}))\,d_{ij}\,(u_j - u_i) = f_i \qquad \text{for } i = 1, \ldots, M\,, \tag{6.28}$$

$$u_i = g_{i-M} \qquad \text{for } i = M+1, \ldots, N\,. \tag{6.29}$$

It is assumed that

$$\alpha_{ij} = \alpha_{ji}\,, \qquad i,j = 1, \ldots, N\,, \tag{6.30}$$

and that, for any $i,j \in \{1, \ldots, N\}$, the function $\alpha_{ij}(\boldsymbol{u})(u_j - u_i)$ is a continuous function of $\boldsymbol{u} \in \mathbb{R}^N$. A theoretical analysis of the AFC scheme (6.28), (6.29) concerning the solvability, local DMP and error estimation can be found in [12]; see also [2, 63] for a posteriori error estimators.

The symmetry condition (6.30) is particularly important for several reasons. First, it guarantees that the resulting method is conservative. Second, it implies that the matrix corresponding to the term arising from the AFC is positive semidefinite. This shows that this term really enhances the stability of the method and enables to estimate the error of the approximate solution, see [12]. Finally, it was demonstrated in [11] that, without the symmetry condition (6.30), the nonlinear algebraic problem (6.28), (6.29) is not solvable in general.

Recently, motivated by [4], a generalization of (6.28) was proposed in [83] by introducing the matrix $\mathbb{B}(\boldsymbol{u}) = (b_{ij}(\boldsymbol{u}))_{i,j=1}^N$ given by

$$b_{ij}(\boldsymbol{u}) = -\max\{0, (1-\alpha_{ij}(\boldsymbol{u}))\, a_{ij}, (1-\alpha_{ji}(\boldsymbol{u}))\, a_{ji}\} \quad \text{for } i \neq j, \tag{6.31}$$

$$b_{ii}(\boldsymbol{u}) = -\sum_{j=1, j\neq i}^N b_{ij}(\boldsymbol{u}). \tag{6.32}$$

Then, instead of (6.28), (6.29), the following algebraically stabilized problem is considered

$$\sum_{j=1}^N a_{ij}\, u_j + \sum_{j=1}^N b_{ij}(\boldsymbol{u})\,(u_j - u_i) = f_i \qquad \text{for } i = 1,\dots,M\,, \tag{6.33}$$

$$u_i = g_{i-M} \qquad \text{for } i = M+1,\dots,N\,. \tag{6.34}$$

Under condition (6.30), both algebraic problems, (6.28), (6.29) and (6.33), (6.34), are equivalent. However, the advantage of (6.33), (6.34) is that the symmetry condition (6.30) is no longer necessary. Note that the matrix $\mathbb{B}(\boldsymbol{u})$ is symmetric, has nonpositive off-diagonal entries and has zero row and column sums. These properties imply that

$$\sum_{i,j=1}^N v_i\, b_{ij}(\boldsymbol{u})\,(v_j - v_i) = -\frac{1}{2}\sum_{i,j=1}^N b_{ij}(\boldsymbol{u})\,(v_j - v_i)^2 \geq 0 \quad \forall\ \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^N\,.$$

Thus, the matrix $\mathbb{B}(\boldsymbol{u})$ is positive semidefinite for any $\boldsymbol{u} \in \mathbb{R}^N$.

To write the above algebraic problem in a variational form, we denote

$$d_h(w; z, v) = \sum_{i,j=1}^N b_{ij}(w)\, z(\boldsymbol{x}_j)\, v(\boldsymbol{x}_i) \qquad \forall\ w, z, v \in C(\overline{\Omega})\,,$$

with $b_{ij}(w) := b_{ij}(\{w(\boldsymbol{x}_i)\}_{i=1}^N)$. Then

$$d_h(w; \phi_j, \phi_i) = b_{ij}(w) \qquad \forall\ w \in C(\overline{\Omega}),\ i, j = 1,\dots,N\,, \tag{6.35}$$

and (6.33), (6.34) is equivalent to problem (3.15), where $a(\cdot,\cdot)$ is defined by (2.3) in case of $\mathbb{A}_{\rm N}$ given by (6.24) and by (5.12) if $\mathbb{A}_{\rm N}$ given by (6.25) is considered. The property (6.32) immediately implies the validity of (3.16). Since the matrix $\mathbb{B}(\boldsymbol{u})$ is positive semidefinite, the form d_h also satisfies (3.17). Finally, since $a_{ij} = a_{ji} = 0$ if $j \notin S_i \cup \{i\}$, one has

$$d_h(w; \phi_j, \phi_i) = 0 \qquad \forall\ w \in C(\overline{\Omega}),\ j \notin S_i \cup \{i\},\ i = 1,\dots,N\,, \tag{6.36}$$

so that (3.23) always holds.

Of course, the properties of an algebraically stabilized scheme significantly depend on the choice of the limiters α_{ij}. Their design principles often originate from the time-dependent case where they should guarantee the positivity preservation, see Section 8.4. In the steady case, a standard limiter is the Kuzmin limiter proposed in [87] which was thoroughly investigated in [12]. To define the limiter of [87], one first computes, for $i = 1, \ldots, M$,

$$P_i^+ = \sum_{\substack{j \in S_i \\ a_{ji} \leq a_{ij}}} f_{ij}^+ \,, \qquad P_i^- = \sum_{\substack{j \in S_i \\ a_{ji} \leq a_{ij}}} f_{ij}^- \,, \tag{6.37}$$

$$Q_i^+ = -\sum_{j \in S_i} f_{ij}^- \,, \qquad Q_i^- = -\sum_{j \in S_i} f_{ij}^+ \,, \tag{6.38}$$

where $f_{ij} = d_{ij}\,(u_j - u_i)$, $f_{ij}^+ = \max\{0, f_{ij}\}$, and $f_{ij}^- = \min\{0, f_{ij}\}$. We recall that d_{ij} is defined in (6.26) using the matrix $\mathbb{A}_{\mathrm{N}}$ from (6.24) or (6.25). Also the matrix entries appearing in (6.37) are taken from this matrix. Then, one defines

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \qquad i = 1, \ldots, M \,. \tag{6.39}$$

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

$$R_i^+ = 1 \,, \quad R_i^- = 1 \,, \qquad i = M+1, \ldots, N \,. \tag{6.40}$$

Furthermore, one sets

$$\widetilde{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \,, \\ 1 & \text{if } f_{ij} = 0 \,, \\ R_i^- & \text{if } f_{ij} < 0 \,, \end{cases} \qquad i, j = 1, \ldots, N \,. \tag{6.41}$$

Finally, one defines

$$\alpha_{ij} = \alpha_{ji} = \widetilde{\alpha}_{ij} \qquad \text{if} \quad a_{ji} \leq a_{ij} \,, \qquad i, j = 1, \ldots, N \,. \tag{6.42}$$

THEOREM 6.7 (DMP for the AFC scheme with Kuzmin limiter). *Let*

$$\min\{a_{ij}, a_{ji}\} \leq 0 \qquad \forall\; i = 1, \ldots, M \,,\; j = 1, \ldots, N \,,\; i \neq j \,. \tag{6.43}$$

Then the AFC scheme (6.28), (6.29) with the Kuzmin limiter defined by (6.37)–(6.42) satisfies the algebraic DMP property formulated in Definition 3.19 and also the algebraic DMP property for non-strict extrema from Definition 3.20.

Proof. Consider any $u_h \in V_h$, $i \in \{1, \ldots, M\}$, and $j \in S_i$. Let $\boldsymbol{u}$ be the vector of nodal values of u_h and assume that u_i is a local extremum of u_h on ω_i and that $u_i \neq u_j$. We want to prove that

$$a_{ij} + (1 - \alpha_{ij}(\boldsymbol{u}))\, d_{ij} \leq 0 \,. \tag{6.44}$$

If $a_{ij} \leq 0$, then (6.44) holds since $(1 - \alpha_{ij}(\boldsymbol{u}))\, d_{ij} \leq 0$. If $a_{ij} > 0$, then $a_{ji} \leq 0$ due to (6.43) and hence $a_{ji} < a_{ij}$ and $d_{ij} = -a_{ij} < 0$. Thus, if $u_i \geq u_k$ for all $k \in S_i$, then $f_{ij} > 0$ and $f_{ik} \geq 0$ for $k \in S_i$, so that $\alpha_{ij} = R_i^+ = 0$. Similarly, if $u_i \leq u_k$ for all $k \in S_i$, then $f_{ij} < 0$ and $f_{ik} \leq 0$ for $k \in S_i$, so that $\alpha_{ij} = R_i^- = 0$. Since $a_{ij} + d_{ij} = 0$, one concludes that (6.44) holds. ∎

If the matrix (6.25) with lumped reaction term is considered, then the validity of (6.43) is guaranteed if the triangulation $\mathscr{T}_h$ satisfies the XZ-criterion (2.7). The condition (6.43) may be satisfied also if the XZ-criterion is violated, particularly, in the convection-dominated case, since the convection matrix is skew-symmetric. However, in general, the validity of a DMP cannot be guaranteed without the XZ-criterion. Moreover, if the matrix (6.25) is replaced by (6.24), then the validity of (6.43) may be lost since some off-diagonal entries of the matrix $\mathbb{M}_\mathrm{c}$ are positive.

It was shown in [82] that the DMP generally does not hold if condition (6.43) is not satisfied. This is due to the condition $a_{ji} \le a_{ij}$ used in (6.42) to symmetrize the factors $\widetilde{\alpha}_{ij}$. Therefore, in [83], it was proposed to use the above limiter in the formulation (6.33), (6.34) without the symmetry condition (6.42). To obtain a well defined problem satisfying a continuity assumption on $\alpha_{ij}(\boldsymbol{u})(u_j - u_i)$, the definition of $P_i^\pm$ was replaced by

$$(6.45) \qquad P_i^+ = \sum_{\substack{j \,\in\, S_i \\ a_{ij} \,>\, 0}} a_{ij}\,(u_i - u_j)^+ \,, \qquad P_i^- = \sum_{\substack{j \,\in\, S_i \\ a_{ij} \,>\, 0}} a_{ij}\,(u_i - u_j)^- \,.$$

Then the DMP is satisfied without any additional condition on the matrix $\mathbb{A}_\mathrm{N}$, which means that it holds for any triangulation $\mathscr{T}_h$ and also without the lumping of the matrix $\mathbb{M}_\mathrm{c}$ in the Galerkin FEM. Note, however, that if the reaction term is dominant, some lumping may be performed by the algebraic flux correction scheme.

THEOREM 6.8 (DMP for the algebraically stabilized scheme with modified Kuzmin limiter). *Let us consider the algebraically stabilized scheme* (6.33)*,* (6.34) *with* $\alpha_{ij} = \widetilde{\alpha}_{ij}$ *for* $i, j = 1, \ldots, N$*, where* $\widetilde{\alpha}_{ij}$ *is defined by* (6.45) *and* (6.38)*–*(6.41)*. Then the algebraic DMP property and the algebraic DMP property for non-strict extrema are satisfied.*

Proof. The proof is similar as for Theorem 6.7. Under the assumptions made before (6.44) we now want to prove that

$$(6.46) \qquad a_{ij} - \max\{0, (1 - \widetilde{\alpha}_{ij}(\boldsymbol{u}))\,a_{ij}, (1 - \widetilde{\alpha}_{ji}(\boldsymbol{u}))\,a_{ji}\} \le 0\,.$$

Since this clearly holds if $a_{ij} \le 0$, it suffices to investigate the case $a_{ij} > 0$. If $u_i \ge u_k$ for all $k \in S_i$, then $P_i^+ \ge a_{ij}\,(u_i - u_j)^+ > 0$, $f_{ij} > 0$ and $f_{ik} \ge 0$ for $k \in S_i$, so that $\widetilde{\alpha}_{ij} = R_i^+ = 0$. If $u_i \le u_k$ for all $k \in S_i$, then $P_i^- \le a_{ij}\,(u_i - u_j)^- < 0$, $f_{ij} < 0$ and $f_{ik} \le 0$ for $k \in S_i$, so that $\widetilde{\alpha}_{ij} = R_i^- = 0$. This implies (6.46). ∎

If condition (6.43) holds, then (6.37) and (6.45) are equivalent, and $b_{ij}(\boldsymbol{u})$ defined using the modified Kuzmin limiter from Theorem 6.8 satisfies $b_{ij}(\boldsymbol{u}) = (1 - \alpha_{ij}(\boldsymbol{u}))d_{ij}$ with the Kuzmin limiter α_{ij} from (6.42). Thus, under condition (6.43), both approaches described above are equivalent. The modified Kuzmin limiter was further improved and reformulated in [69] leading to the Monotone Upwind-type Algebraically Stabilized (MUAS) method. The paper [69] also contains a detailed analysis of algebraically stabilized methods of the type (6.33), (6.34). Further analytical and numerical studies of these approaches recently inspired the design of the Symmetrized Monotone Upwind-type Algebraically Stabilized (SMUAS) method in [84].

Another way how to construct a limiter leading to the DMP on arbitrary meshes and without an explicit lumping of the matrix $\mathbb{M}_\mathrm{c}$ was proposed in [13], using some ideas of [91]. The definition of this limiter, which we call BJK limiter, is inspired by the Zalesak algorithm that will be derived in Section 8.4 for the time-dependent case. It again relies on local quantities P_i^+, P_i^-, Q_i^+, Q_i^- which are now computed for $i = 1, \ldots, M$ by

$$P_i^+ = \sum_{j \in S_i} f_{ij}^+ , \qquad P_i^- = \sum_{j \in S_i} f_{ij}^- , \tag{6.47}$$

$$Q_i^+ = q_i \,(u_i - u_i^{\max}) , \qquad Q_i^- = q_i \,(u_i - u_i^{\min}) , \tag{6.48}$$

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

$$u_i^{\max} = \max_{j \in S_i \cup \{i\}} u_j , \qquad u_i^{\min} = \min_{j \in S_i \cup \{i\}} u_j , \qquad q_i = \gamma_i \sum_{j \in S_i} d_{ij} , \tag{6.49}$$

with fixed constants $\gamma_i > 0$. Then one defines the factors $\widetilde{\alpha}_{ij}$ by (6.39)–(6.41). Finally, the limiters are defined by

$$\alpha_{ij} = \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\} , \qquad i, j = 1, \ldots, N . \tag{6.50}$$

THEOREM 6.9 (DMP for the AFC scheme with BJK limiter). *The AFC scheme* (6.28)*,* (6.29) *with the BJK limiter defined by* (6.47)*–*(6.49)*,* (6.39)*–*(6.41)*, and* (6.50) *satisfies the algebraic DMP property and also the algebraic DMP property for non-strict extrema.*

Proof. The proof is similar as for Theorem 6.7. Under the assumptions made before (6.44) we now want to prove that

$$a_{ij} + (1 - \min\{\widetilde{\alpha}_{ij}(\boldsymbol{u}), \widetilde{\alpha}_{ji}(\boldsymbol{u})\})\, d_{ij} \le 0 . \tag{6.51}$$

If $d_{ij} = 0$, then $a_{ij} \le 0$ and hence (6.51) holds. Thus, let us assume that $d_{ij} < 0$. If $u_i \ge u_k$ for all $k \in S_i$, then $f_{ij} > 0$ and $u_i^{\max} = u_i$ so that $P_i^+ > 0$, $Q_i^+ = 0$ and $\widetilde{\alpha}_{ij} = R_i^+ = 0$. Since $a_{ij} + d_{ij} \le 0$, one obtains (6.51). If $u_i \le u_k$ for all $k \in S_i$, (6.51) follows analogously. ∎

It was proved in [13] that, for

$$\gamma_i \ge \frac{\max_{\boldsymbol{x}_j \in \partial\omega_i} |\boldsymbol{x}_i - \boldsymbol{x}_j|}{\operatorname{dist}(\boldsymbol{x}_i, \partial\omega_i^{\text{conv}})} ,$$

where ω_i^{conv} is the convex hull of ω_i, the AFC scheme with the BJK limiter is linearity preserving, i.e., $\mathbb{B}(u) = 0$ for $u \in \mathbb{P}_1(\mathbb{R}^d)$. This property may lead to improved convergence results, see, e.g., [10, 14]. Note that large values of the constants γ_i cause that more limiters α_{ij} will be equal to 1 and hence less artificial diffusion is added, which makes it possible to obtain sharp approximations of layers. On the other hand, however, large values of γ_i's also cause that the numerical solution of the nonlinear algebraic problem becomes more involved.

Remark 6.10. The various limiters discussed above are inspired by techniques used in the time-dependent case, where a classical approach is the above-mentioned Zalesak algorithm (cf. Section 8.4). This algorithm cannot be simply applied to the steady-state case since the quantities $Q_i^\pm$ are defined using the mass matrix from the discretization of the time-derivative, and a provisional solution of an explicit low-order scheme. The Kuzmin limiter formulated in (6.37)–(6.42) circumvents this problem by defining $Q_i^\pm$ analogously as $P_i^\pm$ in the Zalesak algorithm. The design of the BJK limiter is formally closer to the Zalesak limiter and relies on a carefully selected multiplicative factor in the definition of $Q_i^\pm$. The remaining approaches mentioned above use various modifications of the Kuzmin limiter. As discussed above, the original Kuzmin limiter satisfies the DMP only under the condition (6.43) whereas the other approaches satisfy the DMP without any condition on the stiffness matrix. In addition the BJK limiter and the SMUAS limiter [84] are linearity preserving on arbitrary simplicial meshes. Nevertheless, it is difficult to assess the quality of the resulting schemes from these theoretical properties. Indeed, recent numerical results [64, 65, 71, 84] reveal that depending on considered data and the used criterion (e.g., accuracy, efficiency or experimental convergence rate), one can come to various conclusions concerning the quality of the methods. For example, the BJK limiter often leads to sharp approximations of layers but the nonlinear algebraic problems are difficult to solve and the approximate solutions may be less accurate away from layers than for the Kuzmin limiter. □

Finally, let us present another way how to define the matrix $\mathbb{B}(\boldsymbol{u})$ in the algebraically stabilized problem (6.33), (6.34), the so-called BBK method proposed in [10]. It is also referred to as smoothness-based viscosity and has its origin in the finite volume literature (see, e.g., [62] and [61]).

Given $\boldsymbol{u} \in \mathbb{R}^N$, one first defines the function $\xi_{\boldsymbol{u}} \in V_h$ whose nodal values are given by

$$\xi_{\boldsymbol{u}}(\boldsymbol{x}_i) = \begin{cases} \dfrac{\left|\sum_{j\in S_i}(u_i - u_j)\right|}{\sum_{j\in S_i}|u_i - u_j|} & \text{if } \displaystyle\sum_{j\in S_i}|u_i - u_j| \neq 0\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1,\dots,N\,. \tag{6.52}$$

Then, for any $i, j \in \{1,\dots,N\}$ such that there is an edge $E \in \mathscr{E}_h$ with endpoints $\boldsymbol{x}_i, \boldsymbol{x}_j$, one sets

$$b_{ij}(\boldsymbol{u}) = -\gamma_0\, h_E^{d-1} \max_{\boldsymbol{x}\in E} \left[\xi_{\boldsymbol{u}}(\boldsymbol{x})\right]^p, \qquad p \in [1, +\infty)\,, \tag{6.53}$$

where γ_0 is a fixed parameter, dependent on the data of (2.1). For other pairs of $i \neq j$, one sets $b_{ij}(\boldsymbol{u}) = 0$. Finally, the diagonal entries of the matrix $\mathbb{B}(\boldsymbol{u})$ are again defined by (6.32). Then the corresponding form d_h again satisfies (3.16), (3.17), and (6.36).

The value of p determines the rate of decay of the numerical diffusion with the distance to the critical points. A value closer to 1 adds more diffusion far away from layers and extrema, while a larger value makes the diffusion vanish faster, but on the other hand, increasing p may make the nonlinear system more difficult to solve. In our experience, values up to $p = 20$ are considered safe to use (see [10] for a detailed discussion). Note also that, on symmetric meshes, the method is linearity preserving.

THEOREM 6.11 (DMP for the BBK method). *Let the triangulation $\mathscr{T}_h$ satisfy the XZ-criterion (2.7). Then there exist constants C_0 and C_1 depending only on the shape regularity of $\mathscr{T}_h$ such that if $\gamma_0 \geq C_0\|\boldsymbol{b}\|_{0,\infty,\Omega} + C_1\,\sigma\, h$, then the algebraically stabilized scheme (6.33), (6.34) with $\mathbb{B}(\boldsymbol{u})$ defined by (6.52), (6.53) satisfies the algebraic DMP property and also the algebraic DMP property for non-strict extrema.*

Proof. We again start with the assumptions made in the proof of Theorem 6.7 before (6.44). Then $\xi_{\boldsymbol{u}}(\boldsymbol{x}_i) = 1$ and hence $b_{ij}(\boldsymbol{u}) = -\gamma_0\, h_E^{d-1}$. In view of (6.35), Theorem 4.1, and the shape regularity of the mesh, one obtains

$$\begin{aligned} a(\phi_j,\phi_i) + d_h(u_h;\phi_j,\phi_i) &= \varepsilon\,(\nabla\phi_j,\nabla\phi_i) + (\boldsymbol{b}\cdot\nabla\phi_j,\phi_i) + \sigma\,(\phi_j,\phi_i) - \gamma_0\,h_E^d \\ &\le (C_0\,\|\boldsymbol{b}\|_{0,\infty,\Omega} + C_1\,\sigma\,h - \gamma_0)\,h_E^{d-1} \end{aligned}$$

and the result follows. ∎

Let us now briefly discuss different approaches to make the BBK method linearity preserving on general meshes. The common point to all those alternatives is to introduce positive constants β_{ij} for $j \in S_i$ and modify slightly the definition (6.52) of $\xi_{\boldsymbol{u}}(\boldsymbol{x}_i)$ as follows

$$\xi_{\boldsymbol{u}}(\boldsymbol{x}_i) = \begin{cases} \dfrac{\left|\sum_{j\in S_i}\beta_{ij}(u_i-u_j)\right|}{\sum_{j\in S_i}\beta_{ij}|u_i-u_j|} & \text{if } \displaystyle\sum_{j\in S_i}|u_i-u_j| \neq 0\,, \\ 0 & \text{otherwise}\,, \end{cases} \qquad i = 1,\dots,N\,.$$

In [10, Remark 1] a process to generate a linearity preserving method is described. It involves solving local minimization problems in each node to determine the value of β_{ij}. An alternative approach is presented in [53, Section 4.3]. If the support of the basis functions ϕ_i is convex, then there exists a set of generalized barycentric coordinates $(\omega_{ij})_{j\in S_i}$ such that its elements are non-negative functions, form a partition of unity, and $\boldsymbol{x} = \sum_{j\in S_i}\omega_{ij}(\boldsymbol{x})\boldsymbol{x}_j$ for all $\boldsymbol{x} \in \omega_i$. A process to build these coordinates in higher dimensions is proposed in [132] (see also [45] for a comprehensive review on the topic of generalized barycentric coordinates). Then, taking $\beta_{ij} = \omega_{ij}(\boldsymbol{x}_i)$, it can be proven that the resulting method is linearity preserving.

We end this section again by discussing the solvability and error estimates. It can be proven by means of Brouwer's fixed-point theorem that the nonlinear algebraic problem (6.33), (6.34) is solvable provided that the entries of the matrix $\mathbb{B}(\boldsymbol{u})$ are bounded functions of $\boldsymbol{u} \in \mathbb{R}^N$ and, for any $i, j \in \{1,\dots,N\}$, the functions $b_{ij}(\boldsymbol{u})(u_j - u_i)$ are continuous, see, e.g., [69]. This is the case for all the methods discussed in this section, cf. [14, 69, 84]. A natural norm for estimating the errors of the solutions to the nonlinear problems considered in this section is the solution-dependent norm proposed in [12] given by

$$\|v\|_h := \left(\varepsilon\,|v|_{1,\Omega}^2 + \sigma\,\|v\|_{0,\Omega}^2 + d_h(u_h;v,v)\right)^{1/2}.$$

Then, if $u \in H^2(\Omega)$ and $\sigma > 0$, one has (cf., e.g., [12])

$$\|u - u_h\|_h \le C\,(\varepsilon + \sigma^{-1}\,\|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \sigma)\,h\,\|u\|_{2,\Omega} + (d_h(u_h;i_hu,i_hu))^{1/2}\,,$$

where C is independent of h and the data of the problem (2.1). The term $(d_h(u_h;i_hu,i_hu))^{1/2}$ represents an estimate of the consistency error induced by the algebraic stabilizations. As its precise definition varies according to the choice of limiters, it is to be expected that different convergence orders may be proven for the different choices of limiters. A common feature of the analyses presented in [12, 10] is the following: an $\mathcal{O}(h^{1/2})$ convergence can be proven for meshes of XZ-type. For non-XZ meshes, this convergence order can be proven only in the convection-dominated case in general since certain entries of the diffusion matrix may be positive. Indeed, examples of non-convergence in the diffusion-dominated case are shown for the

Kuzmin limiter in [12]. Moreover, it was proven in [10, 14] that the combination of Lipschitz continuity and linearity preservation leads to an (ε-dependent) improved error estimate of order $\mathcal{O}(h)$.

6.4. A monotone Local Projection Stabilized (LPS) method. In this section we will review a LPS method that respects the DMP proposed in [9]. Its motivation, already hinted in [16], is to start with an optimal order stabilized method based on facets (e.g. CIP), and to introduce a nonlinear switch that makes the method become a first order linear artificial diffusion method in the vicinity of layers and extrema.

The monotone LPS method is given by (3.15) with

$$d_h(w_h;u_h,v_h) = \sum_{F\in\mathscr{F}_I}\Big[\tau_F\alpha_F(w_h)(\nabla u_h,\nabla v_h)_{\omega_F} + \gamma_F\big(1-\alpha_F\,(w_h)\big)(\nabla u_h - G_F\nabla u_h, \nabla v_h - G_F\nabla v_h)_{\omega_F}\Big]. \tag{6.54}$$

Here, for each $F\in\mathscr{F}_I$, the operator G_F provides a local mean value defined by

$$G_F q = \frac{(q,1)_{\omega_F}}{|\omega_F|}\,, \qquad q\in L^1(\omega_F)\,,$$

which is computed component-wise in the case of vector-valued functions, and τ_F, γ_F are stabilization parameters given by

$$\tau_F = c_0\|\boldsymbol{b}\|_{0,\infty,\omega_F}h_F \qquad \text{and} \qquad \gamma_F = \gamma_0\min\left\{\|\boldsymbol{b}\|_{0,\infty,\omega_F}h_F, \frac{h_F^2}{\varepsilon}\right\}, \tag{6.55}$$

with positive constants c_0 and γ_0. The nonlinear switches α_F need to be designed in such a way that they detect regions of extrema and large variations in the gradients, the latter indicating the possible presence of layers. For now, we will just assume that they satisfy the following two basic assumptions:

i) $\alpha_F : V_h \to [0,1]$ are continuous functions; and
ii) $\alpha_F(u_h) = 1$ whenever u_h attains a local extremum at a node of a mesh cell containing F.

In [9] it was proposed to define α_F using regularized versions of the Kuzmin limiter (6.42) or the smoothness-based indicator (6.52).

The form $d_h(\cdot;\cdot,\cdot)$ obviously satisfies the assumptions (3.16) and (3.17). In addition, since $(q - G_F q, 1)_{\omega_F} = 0$ for any $q\in L^1(\omega_F)$ and $F\in\mathscr{F}_I$, it can be also written as

$$d_h(w_h;u_h,v_h) = \sum_{F\in\mathscr{F}_I}\Big[\tau_F\alpha_F(w_h)(\nabla u_h,\nabla v_h)_{\omega_F} + \gamma_F\big(1-\alpha_F\,(w_h)\big)(\nabla u_h - G_F\nabla u_h, \nabla v_h)_{\omega_F}\Big]. \tag{6.56}$$

Remark 6.12. A more natural way of writing (6.56) would be to express the stabilizing term as follows

$$\sum_{F\in\mathscr{F}_I}\tilde\tau_F(\nabla u_h - \beta_F(u_h)G_F\nabla u_h, \nabla v_h)_{\omega_F}\,,$$

where $\tilde{\tau}_F$ is a stabilization parameter, and $\beta_F(u_h) = 1 - \alpha_F(u_h)$. This writing does represent the idea of a method that includes transitions between low-order artificial diffusion and higher order local projection, while at the same time stressing the character of combining linear and nonlinear stabilization terms, as it was made in Section 6.2 for the method given by (6.22). Unfortunately, numerical experimentation has shown that to obtain accurate results the stabilization parameters for the linear diffusion and local projection parts need to be of significantly different sizes. This has led to the (less natural) writing (6.54) for the stabilization term.

It is also worth mentioning that a similar strategy to the above monotone LPS method, although using a local projection related to the Scott–Zhang interpolation operator, was used in [6] to approximate the transport problem. □

In [9] it was proven that, under the assumptions i) and ii) on the limiters, the discrete problem has at least one solution. Concerning the satisfaction of the DMP, we now report a proof slightly more specific than the one provided in [9, § 2.3]. To avoid technical complications, we will present this result in two space dimensions and will suppose that $\sigma = 0$.

THEOREM 6.13 (DMP for the monotone LPS method). *Let us suppose that $d = 2$, the mesh family $\{\mathscr{T}_h\}_{h>0}$ is weakly acute and average acute, $\sigma = 0$ and the nonlinear switches α_F satisfy ii). Then, there exists a constant $C > 0$ depending only on the shape regularity of the mesh family $\{\mathscr{T}_h\}_{h>0}$ such that, if c_0 from (6.55) satisfies*

$$c_0 \geq C \cot \frac{\delta}{2}\,, \tag{6.57}$$

where δ is the angle appearing in (2.9), then the form $d_h(\cdot;\cdot,\cdot)$ defined in (6.54) satisfies the algebraic DMP property and also the algebraic DMP property for non-strict extrema.

Proof. Consider any $u_h \in V_h$ and let us suppose that u_h attains a local extremum at an interior node $\boldsymbol{x}_i \in \Omega$. Consider any $j \in \{1, \ldots, N\}$. Since $\alpha_F(u_h) = 1$ for any $F \subset \omega_i$ and $\nabla\phi_i|_{\omega_F} = 0$ for any $F \not\subset \omega_i$, it follows from (6.56) that

$$d_h(u_h; \phi_j, \phi_i) = \sum_{F \in \mathscr{F}_I,\, F \subset \omega_i} \tau_F\, (\nabla\phi_j, \nabla\phi_i)_{\omega_F}\,,$$

which implies (3.23). Now consider any $j \in S_i$ and let us denote by $E = K \cap K'$ the edge connecting $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. Since the mesh is weakly acute, one has $(\nabla\phi_j, \nabla\phi_i)_K \leq 0$ for all $K \in \mathscr{T}_h$, which leads to $d_h(u_h; \phi_j, \phi_i) \leq \tau_E\, (\nabla\phi_j, \nabla\phi_i)_{\omega_E} = \tau_E\, \ell_{ij}$. Thus, applying (5.9), one arrives at

$$\begin{aligned} a(\phi_j, \phi_i) + d_h(u_h; \phi_j, \phi_i) &\leq \tau_E\, \ell_{ij} + c_{ij} = c_0\, h_E\, \|\boldsymbol{b}\|_{0,\infty,\omega_E}\, \ell_{ij} + c_{ij} \\ &\leq -\frac{c_0\, h_E\, \|\boldsymbol{b}\|_{0,\infty,\omega_E}}{2}\, \tan\frac{\delta}{2} + \frac{(h_K + h_{K'})\|\boldsymbol{b}\|_{0,\infty,\omega_E}}{6}\,. \end{aligned}$$

Thanks to the mesh regularity, one has $h_K + h_{K'} \leq \tilde{C}\, h_E$, where $\tilde{C}$ does not depend on the mesh size h. Hence, if (6.57) holds with $C = \tilde{C}/3$, one obtains (3.22) and (3.24). □

We finish this section by summarizing the error estimates available for the method discussed in this section. Following standard estimates involving stability and asymptotic consistency, an $\mathcal{O}(h^{1/2})$ error estimate can be proven. In [9, Section 2.4] a more refined analysis is carried out assuming that the functions α_F decay with an appropriate rate away from the layers, in other words, assuming that the nonlinear switch is active in only a small region of the computational domain. More precisely, one starts defining the region

$$S_\alpha := \bigcup \left\{ K \in \mathscr{T}_h : \max_{F \in \mathscr{F}_K} \alpha_F(u_h) > h^2 \right\} ,$$

and assumes that $|S_\alpha| = C\, h^s$ with $s > 0$. In addition, for $r > 0$ one defines the set

$$S_{h,\mathrm{ext}} := \left\{ \boldsymbol{x} \in \Omega : |\nabla u(\boldsymbol{x})| \le C\, h^r\, |u|_{2,\infty,\Omega} \right\} ,$$

and requires that

$$\sup_{\boldsymbol{x} \in S_\alpha} \inf_{\boldsymbol{y} \in S_{h,\mathrm{ext}}} |\boldsymbol{x} - \boldsymbol{y}| \le C\, h^r .$$

Under these assumptions the following error estimate is proven in [9, Lemma 2.6]

$$\begin{aligned} &\varepsilon^{\frac{1}{2}} |u - u_h|_{1,\Omega} + \sigma^{\frac{1}{2}} \|u - u_h\|_{0,\Omega} + d_h(u_h; u - u_h, u - u_h)^{\frac{1}{2}} \\ &\quad \le \quad C \left(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + (\sigma + \sigma^{-1} |\boldsymbol{b}|^2_{1,\infty,\Omega}) h^2 \right)^{\frac{1}{2}} h\, |u|_{2,\Omega} + C\, h^{\frac{1+s}{2}} \left(h + h^r \right) |u|_{2,\infty,\Omega} . \end{aligned}$$

Supposing in addition that $r + s/2 \ge 1$ the improved estimate

$$\begin{aligned} &\varepsilon^{\frac{1}{2}} |u - u_h|_{1,\Omega} + \sigma^{\frac{1}{2}} \|u - u_h\|_{0,\Omega} + d_h(u_h; u - u_h, u - u_h)^{\frac{1}{2}} \\ &\quad \le \quad C \left(\varepsilon + \|\boldsymbol{b}\|_{\infty,\Omega} h + (\sigma + \sigma^{-1} |\boldsymbol{b}|^2_{1,\infty,\Omega}) h^2 \right)^{\frac{1}{2}} h\, |u|_{2,\infty,\Omega} \end{aligned}$$

is obtained.

7. A numerical illustration.

This section presents a brief numerical study that illustrates the behavior of several methods discussed in the previous chapters.

In the considered example, a profile defined on the inlet boundary is transported through the domain $\Omega = (0,1)^2$. The data of (2.1) are given by $\varepsilon = 10^{-5}$, $\boldsymbol{b} = (-y, x)^T$, and $\sigma = f = 0$. Hence, the problem satisfies the conditions for the weak maximum principle from Theorem 2.1 for $\sigma = 0$. The Dirichlet boundary condition at the inlet boundary $y = 0$ is prescribed by

$$u(x,0) = \begin{cases} \dfrac{x - 0.375}{\xi} + 1 & \text{if } x \in [0.375 - \xi, 0.375), \\[2mm] -0.75 \dfrac{x - 0.5}{0.125} + 0.25 & \text{if } x \in [0.375, 0.5), \\[2mm] 0.25 \dfrac{x - 0.625}{0.125} + 0.5 & \text{if } x \in [0.5, 0.625), \\[2mm] -0.5 \dfrac{x - 0.625}{\xi} + 0.5 & \text{if } x \in [0.625, 0.625 + \xi), \\[2mm] 32(x - 0.75)(1 - x) & \text{if } x \in [0.75, 1], \\[2mm] 0 & \text{else,} \end{cases}$$

with $\xi = 10^{-3}$. A homogeneous Dirichlet boundary condition is prescribed at the boundary $x = 1$ and homogeneous Neumann conditions on the remaining part of

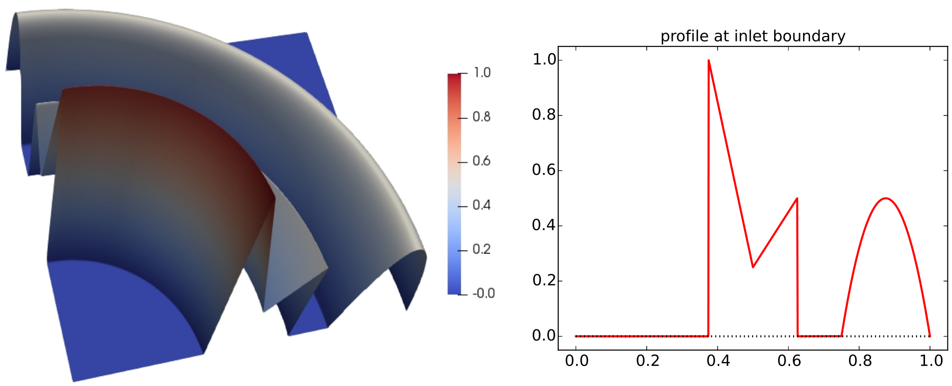
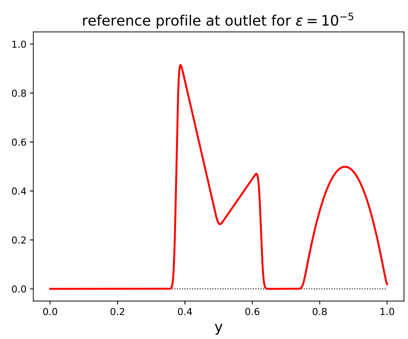


FIG. 4. *Numerical approximation of the solution (left) and profile at the inlet boundary (right).*



quantity of interest	reference value
first maximum value	9.148468e-01
minimum value	2.642484e-01
second maximum value	4.699239e-01
width of the left profile	2.628492e-01
maximum of the bump	4.989947e-01
width of the bump	2.367020e-01
$u(0,1)$	1.914778e-02

FIG. 5. *Reference solution at the outlet boundary $x = 0$ and corresponding reference values.*

the boundary. Figure 4 presents a numerical approximation of the solution and an illustration of the inlet condition.

For assessing the different methods, certain characteristic values of the solution at the outlet boundary $x = 0$ are monitored. A reference solution was computed with the $\mathbb{Q}_2$ Galerkin FEM on a grid consisting of 4096×4096 squares (67 125 249 degrees of freedom, including Dirichlet nodes). Figure 5 depicts the reference solution at the outlet boundary. For defining the reference values, the outlet boundary was decomposed into 100 000 intervals and the corresponding nodal values were used for computing the maximal and minimal values. The width of the left profile was defined by the condition $u(0, y) \geq 0.1$ for $y \leq 0.7$. For the width of the bump, also the condition $u(0, y) \geq 0.1$ was used for computing the left point. Then, the width is defined by subtracting the y-coordinate of this point from 1. In all simulations, a linear interpolation was used for computing the widths. For the reference values, the above mentioned decomposition of the outlet boundary was used and for the other simulations, an interpolation of the nodal values was applied. The reference values are provided in Figure 5.

Simulations were performed for $\mathbb{P}_1$ finite elements. Initially, the domain was decomposed into two triangles by using the diagonal from $(0, 1)$ to $(1, 0)$. Then, this decomposition was refined uniformly using red refinements. Linear systems of equations were solved with the sparse direct solver UMFPACK [36] and nonlinear problems were solved with a simple fixed point iteration, e.g., see [67] or the method *fixed point rhs* from [64], which has been proven to be the most efficient solver for AFC methods in the numerical studies of those papers. The iterations were stopped if the Euclidean norm of the residual vector was smaller than 10^{-10}. Most of the computational results have been double checked with two codes, one of them ParMooN, cf. [47, 134].

From our numerical studies, only results will be presented where the numerical solution does not exhibit spurious oscillations, or more precisely, where the spurious oscillations are at most of the order of round-off errors from floating point arithmetics or the stopping criterion for the iteration of a nonlinear discrete problem. There are many methods that compute solutions with small but still notable spurious oscillations, like some of the spurious oscillations at layers diminishing (SOLD) methods that can be found in the survey [66]. However, such methods are not the topic of this review.

The goal of computing oscillation-free numerical solutions could not be achieved for all methods presented in Section 6. The proof of the DMP property for the edge stabilization method of Burman and Ern from [28] requires that the parameter c_ρ from (6.15) is sufficiently large, compare Theorem 6.2. In the numerical studies in [28], this parameter was set probably to $c_\rho = 5$ (this information is provided for an example with smooth solution but not for an example with layers). But even with this parameter, notable spurious oscillations of the method are reported in [28, Table 3] for the case of a comparatively large diffusion coefficient. For the example studied here, we were able to solve the nonlinear problems (with two different codes) for method (6.14)–(6.16) for parameters $c_\rho \precsim 0.005$. If a standard SUPG term is included, a numerical solution of the nonlinear problem was possible for $c_\rho \precsim 0.05$, which is the parameter choice for this method from [66]. But in both cases and on all grids there are notable undershoots of the computed solutions. This is the reason why we have not reported the results from that method in this survey.

The precise definition of the constants C_i^K used in the implementation of the Mizukami–Hughes method can be found in [78, Fig. 8] or [81, Fig. 5]. The algebraically stabilized method with BBK limiter was used with the parameters $\gamma_0 = 0.75$ and $p = 10$.

Figure 6 presents the differences of the reference value and the values computed with the different methods for all quantities of interest. It can be seen that all nonlinear methods are much more accurate than the used linear method. The accuracy that is reached for the linear upwind method with about 1 000 000 degrees of freedom is usually achieved with the nonlinear methods already for about 4 000 or 16 000 degrees of freedom. One can also observe that there are some differences in the accuracy of the results computed with the different nonlinear discretizations, in particular on coarser grids. However, a comprehensive comparison of the different nonlinear methods, e.g., at other examples or with respect to the computational costs for solving the nonlinear problem, is outside the scope of this review. Some numerical comparisons of algebraically stabilized schemes can be found already in [14, 64].

In summary, the main messages that should be conveyed with this numerical study are that many nonlinear discretizations which satisfy the DMP are much more accurate than linear discretizations with this property and that linear discretizations require prohibitively fine grids for computing accurate results if the solution possesses layers. This message is also supported by the recent paper [71] that contains results of comprehensive numerical studies not only for the methods considered in this section but also for the edge-averaged method from Section 5.4, the MUAS method [68] (see also Section 6.3), and the monolithic convex limiting approach [92].

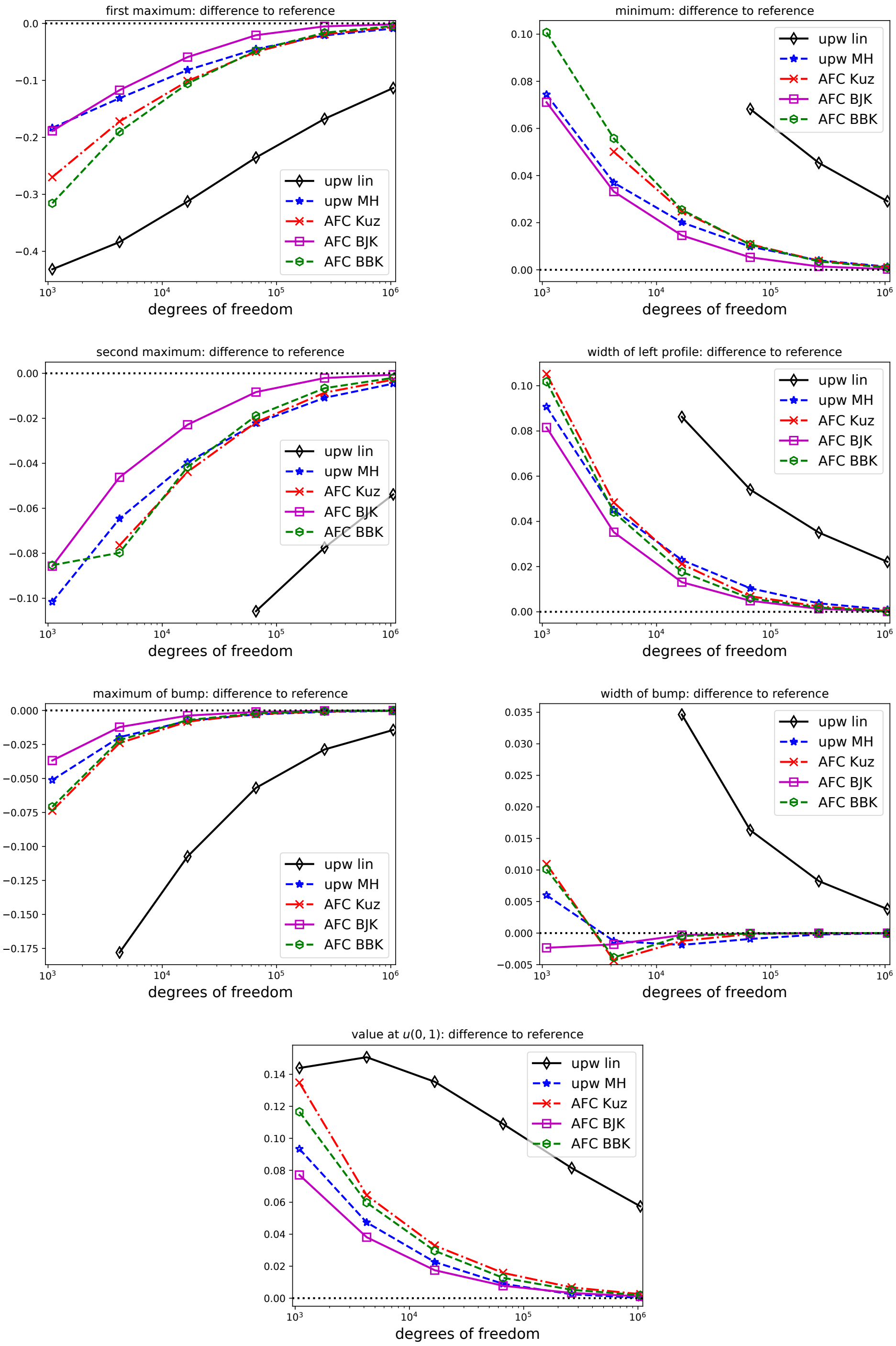


FIG. 6. *Differences of reference value and computed values for the quantities of interest.*

8. Time-dependent problem. This section considers discretizations of time-dependent convection-diffusion-reaction equations, which use one-step θ-schemes in time and finite element methods in space, and which satisfy a DMP. A few linear discretizations in space will be presented briefly and the class of FEM Flux-Corrected-Transport (FCT) schemes, which are usually nonlinear in space, will be discussed in detail.

8.1. The continuous problem. A time-dependent or evolutionary convection-diffusion-reaction initial-boundary value problem is given by

$$\begin{array}{rcll} \partial_t u - \varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + \sigma u &=& f & \text{in } (0,T]\times\Omega\,,\\ u &=& g & \text{on } (0,T]\times\partial\Omega\,,\\ u(0,\cdot) &=& u_0 & \text{in } \Omega\,, \end{array} \tag{8.1}$$

where for the data of the problem, the same notations are used as in the steady-state case. For simplicity, we will again suppose that $\varepsilon > 0$ and $\sigma \ge 0$ are constants and that $\boldsymbol{b}$ is solenoidal. In (8.1), T is the final time and $u_0 = u_0(\boldsymbol{x})$ is a given initial condition. The velocity field $\boldsymbol{b}$, the right-hand side f, and the boundary condition g might depend on time and space. For brevity, the notation $\Omega_T = (0,T]\times\Omega$ is introduced and the parabolic boundary is denoted by $\Gamma_T = \overline{\Omega}_T \setminus \Omega_T$. Note that if $\sigma < 0$, then a change of variable $\check{u}(t,\boldsymbol{x}) = u(t,\boldsymbol{x})\exp(-\kappa t)$ leads to an evolutionary convection-diffusion-reaction equation for $\check{u}$ with the same terms for diffusion and convection, but the coefficient of the reactive term becomes $\sigma+\kappa$, such that $\sigma+\kappa\ge 0$ holds for sufficiently large κ. In this way, many results obtained for $\sigma \ge 0$ can be extended to $\sigma < 0$.

Consider for the moment a problem with $g = 0$ on $(0,T]\times\partial\Omega$. Then, the definition and the analysis of a weak solution of (8.1) can be found, e.g., in [42, Chapter 7.1]. For $\boldsymbol{b}\in L^\infty(0,T;L^\infty(\Omega))$, $f\in L^2(\Omega_T)$, and $u_0\in L^2(\Omega)$, a function $u\in L^2(0,T;H_0^1(\Omega))$ with $\partial_t u \in L^2(0,T;H^{-1}(\Omega))$ is a weak solution of the convection-diffusion-reaction initial-boundary value problem if $u(0) = u_0$ and

$$\langle \partial_t u, v\rangle + \varepsilon(\nabla u,\nabla v) + (\boldsymbol{b}\cdot\nabla u + \sigma u, v) = (f,v) \quad \forall\ v\in H_0^1(\Omega)$$

almost everywhere in $[0,T]$, where $\langle\cdot,\cdot\rangle$ denotes the duality pairing between $H^{-1}(\Omega)$ and $H_0^1(\Omega)$. The existence of a weak solution of (8.1) can be proven with the Galerkin method, see also [42]. For proving uniqueness, it suffices to show that the fully homogeneous problem ($f = 0$, $g = 0$, $u_0 = 0$) possesses only the trivial solution, because the problem is linear. This statement can be proven using the Gronwall lemma. Note that the condition $\sigma \ge 0$ is not needed for these results. If g does not vanish and it is sufficiently smooth, which will be assumed from now on, a problem with homogeneous boundary conditions can be constructed in the usual way by using a lifting of g into Ω for each time and considering a problem for the difference of u and the lifting.

If $\sigma = 0$, problem (8.1) can be equivalently written in the form

$$\begin{array}{rcll} \partial_t u + \nabla\cdot(-\varepsilon\nabla u + \boldsymbol{b}u) &=& f & \text{in } (0,T]\times\Omega\,,\\ u &=& g & \text{on } (0,T]\times\partial\Omega\,,\\ u(0,\cdot) &=& u_0 & \text{in } \Omega\,, \end{array} \tag{8.2}$$

which is called conservative form and results from modeling the conservation of physical quantities. In (8.2), $-\varepsilon\nabla u$ is called diffusive flux and $\boldsymbol{b}u$ convective flux.

8.2. Maximum principle, DMP, and positivity preservation. It will be assumed in this section that $\boldsymbol{b} \in C(\overline{\Omega}_T)$, such that this function is in particular bounded. From the practical point of view, the following weak maximum principle is of importance. Its proof can be found in [42, Chapter 7.1.4], where also a strong maximum principle is proven.

THEOREM 8.1 (Weak maximum principle). *Let* $u \in C^2(\Omega_T) \cap C(\overline{\Omega}_T)$. *Then*

$$\partial_t u - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u + \sigma u \le 0 \quad \text{in } \Omega_T \quad \Longrightarrow \quad \max_{(t,\boldsymbol{x})\in\overline{\Omega}_T} u(t,\boldsymbol{x}) \le \max_{(t,\boldsymbol{x})\in\Gamma_T} u^+(t,\boldsymbol{x}). \tag{8.3}$$

$$\partial_t u - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u + \sigma u \ge 0 \quad \text{in } \Omega_T \quad \Longrightarrow \quad \min_{(t,\boldsymbol{x})\in\overline{\Omega}_T} u(t,\boldsymbol{x}) \ge \min_{(t,\boldsymbol{x})\in\Gamma_T} u^-(t,\boldsymbol{x}). \tag{8.4}$$

If $\sigma = 0$, *then*

$$\partial_t u - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u \le 0 \quad \text{in } \Omega_T \quad \Longrightarrow \quad \max_{(t,\boldsymbol{x})\in\overline{\Omega}_T} u(t,\boldsymbol{x}) = \max_{(t,\boldsymbol{x})\in\Gamma_T} u(t,\boldsymbol{x}). \tag{8.5}$$

$$\partial_t u - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u \ge 0 \quad \text{in } \Omega_T \quad \Longrightarrow \quad \min_{(t,\boldsymbol{x})\in\overline{\Omega}_T} u(t,\boldsymbol{x}) = \min_{(t,\boldsymbol{x})\in\Gamma_T} u(t,\boldsymbol{x}). \tag{8.6}$$

Consider problem (8.1) with $\sigma = 0$ and $f = 0$. For a sufficiently smooth solution, it follows from (8.5) and (8.6) that

$$\min_{(t,\boldsymbol{x})\in\Gamma_T} u(t,\boldsymbol{x}) \le u(t,\boldsymbol{x}) \le \max_{(t,\boldsymbol{x})\in\Gamma_T} u(t,\boldsymbol{x}) \quad \forall\ (t,\boldsymbol{x}) \in \Omega_T. \tag{8.7}$$

Physical quantities whose behavior is modeled with convection-diffusion-reaction equations are often by definition non-negative, like concentrations or the temperature (in Kelvin). The mathematical formulation of this property is the so-called positivity preservation. Let the data of (8.1) be non-negative, i.e., $f \ge 0$ in Ω_T (no sinks), $g \ge 0$ on $(0,T] \times \partial\Omega$, and $u_0 \ge 0$ in Ω. Then it follows from (8.4) that $u \ge 0$ in Ω_T. If $\sigma < 0$, then as already explained in Section 8.1, one can transform problem (8.1) to an equivalent problem for $\breve{u}(t,\boldsymbol{x}) = u(t,\boldsymbol{x}) \exp(-\kappa t)$ with non-negative reaction coefficient and non-negative data on the right-hand sides. Then (8.4) implies that $\breve{u} \ge 0$ in Ω_T whence also $u \ge 0$ in Ω_T. Thus, independently of the sign of σ, the non-negativity of the data f, g, u_0 is sufficient for obtaining a non-negative solution. Therefore, besides the local and global DMP, also the positivity preservation of discretizations of the time-dependent problem is often studied in the literature.

Consider from now on the case that the right-hand side of (8.1) is identically zero. Moreover, for simplicity, we assume that the boundary condition g is independent of time. Let the time interval be decomposed by $0 = t^0 < t^1 < \ldots < t^J = T$. After having applied a one-step θ scheme in time and a linear discretization in space to (8.1), one arrives at time instant t^{n+1} at an algebraic problem of the form

$$\mathbb{B}\boldsymbol{u}^{n+1} = \mathbb{K}\boldsymbol{u}^n, \tag{8.8}$$

where $\boldsymbol{u}^{n+1}$ is the sought solution vector at t^{n+1} and $\boldsymbol{u}^n$ is the solution at time t^n. The matrices $\mathbb{B}$ and $\mathbb{K}$ have the form (3.3) so that the last $N - M$ equations of (8.8) set the Dirichlet boundary conditions for $\boldsymbol{u}^{n+1}$; we recall that the last $N - M$ entries of $\boldsymbol{u}^n$ and $\boldsymbol{u}^{n+1}$ contain the boundary values. We assume that the matrices $\mathbb{B}$ and $\mathbb{K}$ possess the typical sparsity pattern corresponding to discretizations with $\mathbb{P}_1$ finite elements, i.e.,

$$b_{ij} = k_{ij} = 0 \qquad \forall\ j \notin S_i \cup \{i\}\,,\ 1 \le i \le M\,, \tag{8.9}$$

where S_i is defined by (2.4).

Since the right-hand side of (8.1) is identically zero, all cases of the maximum principle from Theorem 8.1 apply. Now, conditions on the matrices $\mathbb{B}$ and $\mathbb{K}$ will be derived such that a discrete version of (8.7) holds.

LEMMA 8.2 (Local DMP). *Consider any* $n \in \{0, \dots, J-1\}$ *and denote*

$$u_i^{\min} = \min\left\{\min_{j\in S_i\cup\{i\}} u_j^n, \min_{j\in S_i} u_j^{n+1}\right\}, \quad u_i^{\max} = \max\left\{\max_{j\in S_i\cup\{i\}} u_j^n, \max_{j\in S_i} u_j^{n+1}\right\}$$

for $i = 1, \dots, M$. *Assume that* (8.8) *holds with* (8.9) *and*

$$b_{ii} > 0,\ \ k_{ii} \ge 0,\ \ b_{ij} \le 0,\ \ k_{ij} \ge 0 \qquad \forall\ j \in S_i\,,\ 1 \le i \le M\,. \tag{8.10}$$

If

$$\sum_{j\in S_i\cup\{i\}} b_{ij} = \sum_{j\in S_i\cup\{i\}} k_{ij}\,, \qquad 1 \le i \le M\,,$$

then it follows that

$$u_i^{\min} \le u_i^{n+1} \le u_i^{\max}\,, \qquad 1 \le i \le M\,.$$

Proof. The proof will be given for the upper bound, the statement for the lower bound can be derived analogously. Consider any $i \in \{1, \dots, M\}$. Let $w_j = u_j^{n+1} - u_i^{\max}$ and $v_j = u_j^n - u_i^{\max}$ for $j = 1, \dots, N$. Then $w_j \le 0$ for all $j \in S_i$ and $v_j \le 0$ for all $j \in S_i \cup \{i\}$. A direct calculation, utilizing the assumption on the row sums, reveals that

$$b_{ii} w_i = k_{ii} v_i + \sum_{j\in S_i} \left(k_{ij} v_j - b_{ij} w_j\right).$$

By construction and assumption (8.10), the coefficient on the left-hand side is positive and the right-hand side is non-positive. Hence, one obtains $w_i \le 0$, which is equivalent to $u_i^{n+1} \le u_i^{\max}$. ∎

For studying global properties, it is convenient to write (8.8) without the (trivial) equations for the values on the Dirichlet boundary:

$$(\mathbb{B}_{\rm I}|\mathbb{B}_{\rm B}) \begin{pmatrix} \boldsymbol{u}_{\rm I}^{n+1} \\ \boldsymbol{u}_{\rm B}^{n+1} \end{pmatrix} = (\mathbb{K}_{\rm I}|\mathbb{K}_{\rm B}) \begin{pmatrix} \boldsymbol{u}_{\rm I}^{n} \\ \boldsymbol{u}_{\rm B}^{n} \end{pmatrix}, \tag{8.11}$$

with $\mathbb{B}_{\rm I}, \mathbb{K}_{\rm I} \in \mathbb{R}^{M\times M}$, $\mathbb{B}_{\rm B}, \mathbb{K}_{\rm B} \in \mathbb{R}^{M\times(N-M)}$, $\boldsymbol{u}_{\rm I}^{n+1}, \boldsymbol{u}_{\rm I}^{n} \in \mathbb{R}^M$, and $\boldsymbol{u}_{\rm B}^{n+1}, \boldsymbol{u}_{\rm B}^{n} \in \mathbb{R}^{N-M}$. It will be assumed that $\mathbb{B}_{\rm I}$ is invertible. Note that from setting the Dirichlet boundary conditions, $\boldsymbol{u}_{\rm B}^{n+1} = \boldsymbol{u}_{\rm B}^{n}$, but for the following considerations, these vectors might be even different.

DEFINITION 8.3 (Positivity preservation). *Method* (8.11) *is said to be positivity preserving if the inequality* $\boldsymbol{u}_{\rm I}^{n+1} \ge 0$ *is valid for all non-negative vectors* $\boldsymbol{u}_{\rm B}^{n+1}$, $\boldsymbol{u}_{\rm I}^{n}$, $\boldsymbol{u}_{\rm B}^{n}$.

THEOREM 8.4 (Necessary and sufficient conditions for positivity preservation). *Method* (8.11) *is positivity preserving if and only if the two conditions*

$$\mathbb{B}_{\rm I}^{-1}(\mathbb{K}_{\rm I}|\mathbb{K}_{\rm B}) \ \ge\ 0, \tag{8.12}$$

$$-\mathbb{B}_{\rm I}^{-1}\mathbb{B}_{\rm B} \ \ge\ 0, \tag{8.13}$$

hold.

Proof. The statement of the theorem follows immediately from the following representation

$$\boldsymbol{u}_{\mathrm{I}}^{n+1} = \mathbb{B}_{\mathrm{I}}^{-1}(\mathbb{K}_{\mathrm{I}}|\mathbb{K}_{\mathrm{B}}) \begin{pmatrix} \boldsymbol{u}_{\mathrm{I}}^{n} \\ \boldsymbol{u}_{\mathrm{B}}^{n} \end{pmatrix} - \mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}}^{n+1},$$

which is obtained from (8.11). ∎

Definition 8.5 (Global DMP). *Method (8.11) is said to satisfy the (global) DMP if*

$$\min\left\{\boldsymbol{u}_{\mathrm{B}}^{n+1}, \boldsymbol{u}_{\mathrm{I}}^{n}, \boldsymbol{u}_{\mathrm{B}}^{n}\right\} \le u_i^{n+1} \le \max\left\{\boldsymbol{u}_{\mathrm{B}}^{n+1}, \boldsymbol{u}_{\mathrm{I}}^{n}, \boldsymbol{u}_{\mathrm{B}}^{n}\right\}, \quad 1 \le i \le M, \tag{8.14}$$

for each choice $\boldsymbol{u}_{\mathrm{B}}^{n+1}, \boldsymbol{u}_{\mathrm{I}}^{n}, \boldsymbol{u}_{\mathrm{B}}^{n}$, *where* $(u_i^{n+1})_{i=1}^M = \boldsymbol{u}_{\mathrm{I}}^{n+1}$.

In the following, a vector of length $k \in \mathbb{N}$ where all entries are 1 is denoted by $\mathbf{1}_k$.

Theorem 8.6 (Necessary and sufficient conditions for the global DMP). *Method (8.11) satisfies the global DMP if and only if (8.12), (8.13), and*

$$(\mathbb{B}_{\mathrm{I}}|\mathbb{B}_{\mathrm{B}})\mathbf{1}_N = (\mathbb{K}_{\mathrm{I}}|\mathbb{K}_{\mathrm{B}})\mathbf{1}_N \tag{8.15}$$

hold, i.e., the i*th row sums of* $(\mathbb{B}_{\mathrm{I}}|\mathbb{B}_{\mathrm{B}})$ *and* $(\mathbb{K}_{\mathrm{I}}|\mathbb{K}_{\mathrm{B}})$ *are identical,* $i = 1, \ldots, M$.

Proof. The proof follows [43].

i) DMP $\Longrightarrow$ *(8.12), (8.13), (8.15).* If $\boldsymbol{u}_{\mathrm{B}}^{n+1}$, $\boldsymbol{u}_{\mathrm{I}}^{n}$, and $\boldsymbol{u}_{\mathrm{B}}^{n}$ are arbitrary non-negative vectors, then the left-hand inequality of (8.14) states that $\boldsymbol{u}_{\mathrm{I}}^{n+1}$ is also non-negative. Hence, the method is positivity preserving and it follows from Theorem 8.4 that (8.12) and (8.13) are satisfied.

Choosing in (8.14) $\boldsymbol{u}_{\mathrm{B}}^{n+1} = \mathbf{1}_{N-M}$, $\boldsymbol{u}_{\mathrm{I}}^{n} = \mathbf{1}_M$, and $\boldsymbol{u}_{\mathrm{B}}^{n} = \mathbf{1}_{N-M}$ yields $\boldsymbol{u}_{\mathrm{I}}^{n+1} = \mathbf{1}_M$. Inserting these vectors in (8.11) shows that (8.15) is satisfied.

ii) (8.12), (8.13), (8.15) $\Longrightarrow$ *DMP.* Denoting $u_{\max}^n = \max\{\boldsymbol{u}_{\mathrm{B}}^{n+1}, \boldsymbol{u}_{\mathrm{I}}^{n}, \boldsymbol{u}_{\mathrm{B}}^{n}\}$ and using (8.12), (8.15), and (8.13), gives

$$\begin{aligned}
\boldsymbol{u}_{\mathrm{I}}^{n+1} &= -\mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}}^{n+1} + \mathbb{B}_{\mathrm{I}}^{-1}(\mathbb{K}_{\mathrm{I}}|\mathbb{K}_{\mathrm{B}}) \begin{pmatrix} \boldsymbol{u}_{\mathrm{I}}^{n} \\ \boldsymbol{u}_{\mathrm{B}}^{n} \end{pmatrix} \\
&\le -\mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}}^{n+1} + u_{\max}^n\mathbb{B}_{\mathrm{I}}^{-1}(\mathbb{K}_{\mathrm{I}}|\mathbb{K}_{\mathrm{B}})\mathbf{1}_N \\
&= -\mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}}\boldsymbol{u}_{\mathrm{B}}^{n+1} + u_{\max}^n\mathbb{B}_{\mathrm{I}}^{-1}(\mathbb{B}_{\mathrm{I}}|\mathbb{B}_{\mathrm{B}})\mathbf{1}_N \\
&= -\mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}}(\boldsymbol{u}_{\mathrm{B}}^{n+1} - u_{\max}^n\mathbf{1}_{N-M}) + u_{\max}^n\mathbf{1}_M \le u_{\max}^n\mathbf{1}_M\,,
\end{aligned}$$

which is equivalent to the right-hand inequality in (8.14). The left-hand inequality is proven similarly. ∎

The concepts of positivity preservation and of the global DMP can be extended to non-vanishing right-hand sides, see [43]. The necessary and sufficient requirements on the matrices for the satisfaction of these properties are the same as given in Theorems 8.4 and 8.6.

Corollary 8.7 (Positivity preservation and global DMP for monotone matrices). *Let the matrix*

$$\mathbb{B} = \begin{pmatrix} \mathbb{B}_{\mathrm{I}} & \mathbb{B}_{\mathrm{B}} \\ \mathbb{O} & \mathbb{I} \end{pmatrix}$$

be monotone and let $\mathbb{K} \ge 0$. *Then method (8.11) is positivity preserving. If, in addition, the* i*th row sums of* $\mathbb{B}$ *and* $\mathbb{K}$ *are identical,* $i = 1, \ldots, M$, *then method (8.11) satisfies the global DMP.*

Proof. From computing the inverse of $\mathbb{B}$, compare (3.13), it follows that $\mathbb{B}_{\mathrm{I}}^{-1} \geq 0$ and $-\mathbb{B}_{\mathrm{I}}^{-1}\mathbb{B}_{\mathrm{B}} \geq 0$. Since $\mathbb{K} \geq 0$, the conditions (8.12) and (8.13) are satisfied. Thus, the corollary follows from Theorems 8.4 and 8.6. ∎

Remark 8.8. Note that if $\mathbb{B}$ is a monotone matrix, $\mathbb{K} \geq 0$, and $\boldsymbol{u}^n \geq 0$, then it immediately follows that the solution of (8.8) satisfies $\boldsymbol{u}^{n+1} \geq 0$. □

Another property that is often studied for discretizations of scalar evolutionary transport problems is the local extremum diminishing (LED) property. Considering a method that is only semi-discrete in space, the LED condition is as follows: if u_i is a local maximum in space, then $du_i/dt \leq 0$ and if u_i is a local minimum in space, then $du_i/dt \geq 0$, i.e., a local maximum does not increase and a local minimum does not decrease. For a fully discrete method, discretized with a one-step θ-scheme, the LED property states that if $u_i^{n+\theta} = \theta u_i^{n+1} + (1-\theta)u_i^n$ is a local maximum in space, then $u_i^{n+1} \leq u_i^n$ and similarly for a local minimum, e.g., see [5].

Section 8.4 will discuss a class of nonlinear discretizations in some detail. A motivation for considering such discretizations for the convection-dominated regime is provided by a study of the limit case of (8.1) with respect to small diffusion, i.e., the transport equation where $\varepsilon = 0$. Consider this case with constant convection $b \neq 0$ and $\sigma = f = 0$ in one dimension on the infinite domain $\Omega = (-\infty, \infty)$. The domain is decomposed using an equidistant grid with mesh width h and the nodes x_i, $i \in \mathbb{Z}$. Then, the application of an explicit one-step θ-scheme leads to a problem of the form

$$u_j^{n+1} = \sum_{i=-S}^{S} \gamma_i u_{j+i}^n, \qquad j \in \mathbb{Z}\,, \tag{8.16}$$

where S is determined by the width of the stencil. For this kind of problem there exists the notion of a monotonicity preserving scheme: for all monotone discrete initial conditions u^0, the solution u^n possesses the same monotonicity for all $n \geq 1$. It can be shown that the scheme is monotonicity preserving if and only if $\gamma_i \geq 0$ for all $i \in \{-S, \ldots, S\}$. Then, Godunov's order barrier theorem [50] states that if $C_{\mathrm{CFL}} = |b|\tau/h \notin \mathbb{N}$, a linear monotonicity preserving method of form (8.16) cannot compute solutions exactly that are polynomials of degree 2. Hence, a linear monotonicity preserving method has to be of low order. For a more recent presentation of this topic see [133]. Using an implicit one-step scheme or a linear multi-step scheme instead of an explicit one-step scheme does not solve this issue, see [133, Thm. 9.2.4].

The condition on the non-negativity of γ_i resembles condition (8.12), which is necessary for the positivity preservation and the satisfaction of the DMP. Thus, one can expect that for (8.1), in the convection-dominated regime, a linear discretization that possesses these properties will be only of low-order. There is no mathematical proof of this expectation but computational evidence. This issue motivates the construction of nonlinear discretizations to obtain accurate schemes for (8.1) that are positivity preserving and satisfy the DMP.

8.3. Linear methods. Utilizing a one-step θ-scheme in combination with the Galerkin or some stabilized finite element method for the discretization of (8.1) with $f = 0$ leads to an algebraic system of the form

$$(\mathbb{M}_{\mathrm{c}} + \theta\tau\mathbb{A}_1)^M\, \boldsymbol{u}^{n+1} = (\mathbb{M}_{\mathrm{c}} - (1-\theta)\tau\mathbb{A}_2)^M\, \boldsymbol{u}^n, \quad u_i^{n+1} = g_{i-M}^{n+1}, \tag{8.17}$$

$i = M+1, \ldots, N$, where $\mathbb{M}_{\mathrm{c}}$ is the consistent mass matrix defined in (2.13), $\mathbb{A}_1$, $\mathbb{A}_2$ are stiffness matrices, and $\tau = t^{n+1} - t^n$ is the current time step. Consider a uniform spatial grid with mesh width h. Then, for standard Lagrangian finite element spaces, $\mathbb{M}_c$ possesses positive off-diagonal entries of order $\mathcal{O}(h^d)$, compare (2.16) for $\mathbb{P}_1$ finite elements. Consequently, $\mathbb{M}_c$ is not an M-matrix and as can be checked easily, e.g., for a one-dimensional problem, $\mathbb{M}_c$ is not a monotone matrix. The off-diagonal entries of $\tau\mathbb{A}_1$ are of order $\mathcal{O}(\tau h^{d-2})$ for the diffusive term and $\mathcal{O}(\tau h^{d-1})$ for the convective term. Hence, if τ is sufficiently small, the system matrix of (8.17) cannot be an M-matrix. In particular, any finite element analysis that considers the so-called continuous-in-time situation, i.e., only a semi-discretization in space, cannot apply the concept of M-matrices. It is shown in [123] that a standard continuous-in-time finite element discretization of the heat equation cannot be positivity preserving and it cannot satisfy the global DMP. One can only hope for non-positive off-diagonal entries of the system matrix of (8.17) if τ is of order $\max\{h, h^2\}$. In fact, for the heat equation, discretized with a one-step θ-scheme and the Galerkin FEM, sufficient conditions for the satisfaction of the DMP were derived in [43] that include a lower and an upper bound for the length of the time step, which are both of order $\mathcal{O}(h^2)$.

Note that this issue does not appear for finite volume and finite difference methods, where the temporal discretization leads to a diagonal matrix with positive diagonal entries. Studying positivity preservation and the DMP with the concept of M-matrices for finite element methods, the common way consists in applying mass lumping, which is presented in Section 2.3. Utilizing a lumped mass matrix, the positivity preservation can be proven for the heat equation in two dimensions, $\mathbb{P}_1$ finite elements, and under certain additional assumptions, see [118]. An extension of this result to three dimensions is also possible.

In [44] a class of problems was studied which includes the linear convection-diffusion-reaction equation as a special case. The considered discretization was a one-step θ-scheme combined with the Galerkin FEM. The DMP is proven under a number of assumptions. Because of using the Galerkin FEM, the mesh width has to be sufficiently small, compare [44, Thm. 5.2 (ii)], in particular the bound for the mesh width tends to zero as $\varepsilon \to 0$. For a sufficiently small mesh width, there is a lower bound for the time step of order $\mathcal{O}(h^2)$.

As already mentioned in Section 5.3, the upwind finite element method proposed in [122] was formulated and studied for a two-dimensional time-dependent equation. The analysis is performed for the forward Euler scheme, where a lumped mass matrix is utilized, so that the discretization of the time derivative corresponds to a finite difference or finite volume one. The key ingredient of this method is the discretization of the convective term, which is described in Section 5.3. From the proof presented in [122], it can be seen that the assumptions of Corollary 8.7 are satisfied under an appropriate CFL condition, hence the method satisfies the DMP. In the final part of [122], it is mentioned that the analysis can be extended to the (mass lumped) backward Euler scheme and to time-dependent convection fields.

The upwind method proposed and analyzed in [3] was also already presented in Section 5.3. In [3], it was studied for the conservative form (8.2) of the convection-diffusion equation. In contrast with the method from [122], it satisfies a discrete analog of a mass conservation property if (8.2) is equipped with so-called free boundary condition

$$\varepsilon \frac{\partial u}{\partial \boldsymbol{n}} - \boldsymbol{b} \cdot \boldsymbol{n}\, u = 0 \quad \text{on } (0, T] \times \partial\Omega.$$

The upwind method is analyzed for this boundary condition, steady-state convection fields, and the mass lumped forward Euler scheme so that an appropriate CFL condition becomes necessary throughout the analysis. A brief description of the discretization of the convective term, leading to a convection matrix $\tilde{\mathbb{A}}_{\mathrm{c}}$, is already provided in Section 5.3. Thus, the discretization of (8.2) with $f = 0$ and the free boundary condition is of the form

$$(\mathbb{M}_{\mathrm{l}})^M \boldsymbol{u}^{n+1} = (\mathbb{M}_{\mathrm{l}})^M \boldsymbol{u}^n - \tau(\varepsilon \mathbb{A}_{\mathrm{d}} + \tilde{\mathbb{A}}_{\mathrm{c}})^M \boldsymbol{u}^n.$$

The construction of $\tilde{\mathbb{A}}_{\mathrm{c}}$ assures that its row sums vanish. The row sums of $\mathbb{A}_{\mathrm{d}}$ also vanish, see (4.3), and hence the positivity preservation and the satisfaction of the global DMP for this upwind method can be inferred from Corollary 8.7.

Remark 8.9. The techniques of [101, 59] developed for problems with heterogeneous anisotropic diffusion, see Remark 4.4, were applied to study also the DMP for the heat equation in [102]. The $\mathbb{P}_1$ finite element in space is combined with a one-step θ-method in time. Concerning the spatial mesh, the same conditions apply as for the steady-state diffusion problem. Using a lumped mass matrix, one obtains a restriction for the length of the time step, which is of the form

$$\tau \le C \min_{K \in \mathscr{T}_h} \min_{F \in \mathscr{F}_K} \frac{h_{K,F}^2}{\lambda_{\max}(\bar{\mathbb{E}}_K)},$$

where $h_{K,F}$ is the height from the facet $F \subset K$ to the vertex of K opposite F and $\bar{\mathbb{E}}_K$ is defined to be the integral mean of the diffusion tensor $\mathbb{E}$ on K. □

8.4. FEM Flux-Corrected-Transport (FCT) schemes. A physical quantity is called extensive if it scales with the size of the physical problem. Examples are mass, momentum, or energy. Fluxes are quantities of an extensive variable that moves from one location in space to another one. That means, the amount of the variable that is removed from the first location is added at the second location. If numerical methods are formulated in terms of fluxes, they are called conservative if the same principle is applied as mentioned above: what is removed from one degree of freedom is added to another one. The conservation of physical quantities in a numerical method contributes to the physical consistency of this method and thus, it helps that the method becomes accepted by practitioners.

The usual starting point for the construction of numerical methods based on fluxes is the conservative form (8.2) of the convection-diffusion equation. Natural discretizations for this form are finite difference and finite volume methods.

For illustrating the concept of numerical fluxes, consider a finite difference method for the one-dimensional analog of (8.2)

$$\begin{array}{rcll} \partial_t u + \partial_x\left(-\varepsilon \partial_x u + bu\right) & = & 0 & \text{in } (0,T] \times \Omega\,, \\ u & = & 0 & \text{on } (0,T] \times \partial\Omega\,, \\ u(\cdot,0) & = & u_0 & \text{in } \Omega\,, \end{array} \tag{8.18}$$

with $\Omega = (\xi_{\mathrm{l}}, \xi_{\mathrm{r}})$, $\xi_{\mathrm{l}} < \xi_{\mathrm{r}}$. Let $\overline{\Omega}$ be triangulated using an equidistant grid with mesh width h and nodes $\{x_i\}_{i=1}^N$, $x_1 = \xi_{\mathrm{l}}$, $x_N = \xi_{\mathrm{r}}$, $x_i < x_{i+1}$. Consider the step from time instant t^n to t^{n+1}. A finite difference approximation of (8.18) is said to be of conservative form, if it can be written for inner nodes in the form

$$u_i^{n+1} = u_i^n + \frac{\tau}{\frac{1}{2}(x_{i+1} - x_{i-1})}\left(f_{i-1/2} - f_{i+1/2}\right),$$

where $f_{i+1/2}$ and $f_{i-1/2}$ are numerical fluxes depending on diffusion and convection at one or several time levels. Utilizing the explicit Euler scheme for discretizing (8.18) in time, the standard 3 point stencil for the discretization of the second derivative and a central finite difference defined on the points $x_{i+1/2} = (x_{i+1} + x_i)/2$ and $x_{i-1/2} = (x_i + x_{i-1})/2$ for the convective term yields

$$\begin{aligned} u_i^{n+1} &= u_i^n + \tau \left[\varepsilon \frac{u_{i+1}^n - 2u_i^n + u_{i-1}^n}{h^2} - \frac{b_{i+1/2}^n u_{i+1/2}^n - b_{i-1/2}^n u_{i-1/2}^n}{h} \right] \\ &= u_i^n + \frac{\tau}{h} \left[-\varepsilon \frac{u_i^n - u_{i-1}^n}{h} + b_{i-1/2}^n u_{i-1/2}^n - \left(-\varepsilon \frac{u_{i+1}^n - u_i^n}{h} + b_{i+1/2}^n u_{i+1/2}^n \right) \right]. \end{aligned}$$

Hence, the numerical analog of the fluxes of the continuous problem, see the end of Section 8.1, is given by

$$f_{i+1/2} = -\varepsilon \frac{u_{i+1}^n - u_i^n}{h} + b_{i+1/2}^n u_{i+1/2}^n \,,$$

where the first term on the right-hand side is the numerical diffusive flux and the second term the numerical convective flux. Usually, the values $u_{i\pm1/2}^n$ at $x_{i\pm1/2}$ are approximated using the values at the neighboring nodes with the aim to obtain a stable discretization. A classical example is the one-sided upwind approximation.

The first development and implementation of a FCT scheme was performed for a finite difference method in one dimension in [18]. Consider the step from one discrete time level to the next one, then the basic approach is as follows:

1. A (linear) scheme is needed that guarantees that no nonphysical values are computed. Such a scheme has to utilize low-order fluxes, which possess a large amount of numerical diffusion.
2. A second (linear) scheme with high-order fluxes is used, which is highly accurate for smooth regions of the solution. This scheme has only a small amount of numerical diffusion and its solution has spurious oscillations in a vicinity of layers or shocks.
3. So-called antidiffusive fluxes are defined by the difference of the high and low-order fluxes from the two schemes.
4. The solution at the new time level is obtained by adding appropriately weighted (limited) antidiffusive fluxes to the solution of the low-order scheme. The limiting process has to ensure that no unphysical values are created in this step. For smooth parts of the solution, the high-order scheme should be recovered.

FCT schemes were then transferred to one-dimensional finite volume methods. It turned out that the limiter for one-dimensional problems proposed in [18] does not work properly in multiple dimensions. Thus, the next milestone in the development of FCT schemes was the proposal of a new limiter that works in multiple dimensions in [136], the nowadays so-called Zalesak limiter. This limiter will be described within the presentation of the FEM-FCT methods. A good survey of the motivations for deriving FCT schemes and their main design principles can be found in the paper [137], which concentrates on finite volume schemes on structured grids.

The development of FCT schemes for finite element methods was driven by the goal to apply the FCT methodology on unstructured grids. To this end, a concept that resembles fluxes was introduced in finite element methods, the so-called algebraic fluxes. Algebraic fluxes are quantities f_{ij} between adjacent degrees of freedom i and j that are derived from algebraic quantities like matrices and vectors and for which $f_{ij} = -f_{ji}$ (the flux property) holds. The vast majority of FEM-FCT methods have been developed for $\mathbb{P}_1$ and $\mathbb{Q}_1$ finite elements, where the degrees of freedom are function values at the vertices of the mesh cells. The first FEM-FCT schemes were proposed in [105, 111]. Since then, FEM-FCT schemes have been improved and further developed, e.g., in [97, 86, 89, 91, 104], see also the surveys in [90] and [95, Chapters 6.3, 7.5, 7.6]. Nevertheless, theoretical results on FEM-FCT schemes for time-dependent convection-diffusion-reaction problems are far less developed than for the related algebraically stabilized methods proposed for the steady-state problem and discussed in Section 6.3. In particular, we are not aware of any error estimates.

Whereas the FCT methodology is used in finite difference and finite volume schemes directly to define a discretization of the convection and diffusion operators with the goal to satisfy the DMP locally, its application in the FEM is more indirect. There, the Galerkin FEM discretization is reformulated equivalently such that the system matrix becomes an M-matrix and then the FCT methodology is utilized to modify the right-hand side such that the M-matrix property of the system matrix allows to satisfy the global DMP and the positivity preservation.

In the following, a FEM-FCT scheme will be presented in detail, thereby explaining the derivation and application of the Zalesak limiter. The starting point is now problem (8.1) and it is again assumed that the right-hand side vanishes. Moreover, for simplicity, we assume that the velocity field $\boldsymbol{b}$ does not depend on time.

The high-order method from Step 2 of the basic FCT approach is the standard Galerkin FEM. Using a one-step θ-scheme as temporal discretization, $\theta \in (0,1]$, leads to the linear algebraic system

$$(8.19) \qquad (\mathbb{M}_{\rm c} + \theta\tau\mathbb{A}_{\rm N})^M \boldsymbol{u}^{n+1} = (\mathbb{M}_{\rm c} - (1-\theta)\tau\mathbb{A}_{\rm N})^M \boldsymbol{u}^n,$$

where the matrix $\mathbb{A}_{\rm N}$ is defined by (6.24). The system (8.19) has to be supplemented by Dirichlet boundary conditions for $\boldsymbol{u}^{n+1}$. Like for the algebraic flux correction in the steady case, we define the matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$ by (6.26) using the entries of $\mathbb{A}_{\rm N}$. In addition, we introduce the matrix $\mathbb{L} = (l_{ij})_{i,j=1}^N$ defined by

$$\mathbb{L} = \mathbb{A}_{\rm N} + \mathbb{D}.$$

As discussed in Section 6.3, the matrix $\mathbb{L}$ is of non-negative type and $\mathbb{D}$ is positive semidefinite.

Next, the low-order scheme from Step 1 of the basic FCT algorithm is given by

$$(8.20) \quad (\mathbb{M}_{\rm l} + \theta\tau\mathbb{L})^M \tilde{\boldsymbol{u}} = (\mathbb{M}_{\rm l} - (1-\theta)\tau\mathbb{L})^M \boldsymbol{u}^n, \quad \tilde{u}_i = g_{i-M}^{n+1}, \; i = M+1,\dots,N,$$

where the lumped mass matrix $\mathbb{M}_{\rm l}$ is defined in (2.20). Due to the assumptions on the data of (8.1), the matrix $(\mathbb{A}_{\rm N})_{\rm I}$ is positive definite and hence also the matrix $(\mathbb{M}_{\rm l} + \theta\tau\mathbb{L})_{\rm I}$ is positive definite. Consequently, the system matrix of (8.20), defined by extending the matrix $(\mathbb{M}_{\rm l} + \theta\tau\mathbb{L})^M$ by the lower blocks of (3.3), is invertible. Since it is of non-negative type, Corollary 3.13 implies that the system matrix of (8.20) is an M-matrix. Thus, in view of Corollary 8.7, method (8.20) is positivity preserving if

$$(8.21) \qquad (\mathbb{M}_{\rm l} - (1-\theta)\tau\mathbb{L})^M \geq 0.$$

To simplify the presentation, we denote the diagonal entries of $\mathbb{M}_{\rm l}$ by m_i instead of $\tilde{m}_{ii}$ considered in (2.20). Since $\mathbb{L}$ is of non-negative type and $\mathbb{L}_{\rm I}$ is positive definite, one has $l_{ii} > 0$ and $l_{ij} \leq 0$ for $j \neq i$, $i = 1,\dots,M$. Hence (8.21) holds if and only if $(1-\theta)\tau l_{ii} \leq m_i$ for all $i = 1,\dots,M$, which is satisfied if $\theta = 1$ or if

$$(8.22) \qquad \tau \leq \frac{m_i}{(1-\theta) l_{ii}}, \qquad i = 1,\dots,M\,.$$

This is a CFL condition which can be checked easily in simulations.

Although the solution of (8.20) does not possess unphysical values under the CFL condition (8.22), it is usually very inaccurate. In the FEM-FCT methodology, a correction term $\tau\overline{\boldsymbol{f}}$ is added, which leads to a method of the form

$$(8.23)\qquad \left(\mathbb{M}_{\mathrm{l}} + \theta\tau\mathbb{L}\right)^M \boldsymbol{u}^{n+1} = \left(\mathbb{M}_{\mathrm{l}} - (1-\theta)\tau\mathbb{L}\right)^M \boldsymbol{u}^n + \tau\overline{\boldsymbol{f}}.$$

If the solution is smooth in the whole domain, then (8.23) should recover the high-order method. A direct calculation, subtracting (8.19) from (8.23), shows that in this case

$$\tau\overline{\boldsymbol{f}} = \left(\mathbb{M}_{\mathrm{l}} - \mathbb{M}_{\mathrm{c}}\right)^M \left(\boldsymbol{u}^{n+1} - \boldsymbol{u}^n\right) + \tau\left(\mathbb{D}\right)^M \left(\theta\boldsymbol{u}^{n+1} + (1-\theta)\boldsymbol{u}^n\right)$$

is the appropriate correction. The expression on the right-hand side can be written in terms of algebraic fluxes. Using the definition (2.20) of the lumped mass matrix and that the row sums of $\mathbb{D}$ are zero, one obtains by a straightforward calculation

$$\begin{aligned}\tau\left(\overline{\boldsymbol{f}}\right)_i &= \sum_{j=1}^N \left[-m_{ij}\left(u_j^{n+1} - u_i^{n+1}\right) + m_{ij}\left(u_j^n - u_i^n\right)\right] \\ &\quad + \tau\sum_{j=1}^N \left[\theta d_{ij}\left(u_j^{n+1} - u_i^{n+1}\right) + (1-\theta)d_{ij}\left(u_j^n - u_i^n\right)\right].\end{aligned}$$

For computing the right-hand side, again the matrices without having imposed Dirichlet boundary conditions are used. Thus, the antidiffusive fluxes from Step 3 of the basic FCT algorithm are given by

$$(8.24)\qquad \begin{aligned} f_{ij} &= \frac{1}{\tau}\left[-m_{ij}\left(u_j^{n+1} - u_i^{n+1}\right) + m_{ij}\left(u_j^n - u_i^n\right)\right] \\ &\quad + \left[\theta d_{ij}\left(u_j^{n+1} - u_i^{n+1}\right) + (1-\theta)d_{ij}\left(u_j^n - u_i^n\right)\right], \quad i,j = 1,\ldots,N. \end{aligned}$$

Because $\mathbb{M}_{\mathrm{c}}$ and $\mathbb{D}$ are symmetric matrices, one has $f_{ij} = -f_{ji}$. Note that the fluxes depend on (unknown) values of the numerical solution at time level t^{n+1}.

Now, following Step 4 of the basic FCT algorithm, the solution for the inner nodes at the next time level is defined by

$$(8.25)\qquad \left(\mathbb{M}_{\mathrm{l}} + \theta\tau\mathbb{L}\right)^M \boldsymbol{u}^{n+1} = \left(\mathbb{M}_{\mathrm{l}} - (1-\theta)\tau\mathbb{L}\right)^M \boldsymbol{u}^n + \tau\left(\sum_{j=1}^N \alpha_{ij} f_{ij}\right)_{i=1}^M,$$

where the limiters $\alpha_{ij} = \alpha_{ji} \in [0,1]$ have to be chosen appropriately.

In order to apply the framework presented in Section 8.2, the nonlinear problem (8.25) is written in the following way:

$$(8.26)\qquad \left(\mathbb{M}_{\mathrm{l}}\right)^M \overline{\boldsymbol{u}} = \left(\mathbb{M}_{\mathrm{l}} - (1-\theta)\tau\mathbb{L}\right)^M \boldsymbol{u}^n,$$

$$(8.27)\qquad \left(\mathbb{M}_{\mathrm{l}}\right)^M \tilde{\boldsymbol{u}} = \left(\mathbb{M}_{\mathrm{l}}\right)^M \overline{\boldsymbol{u}} + \tau\left(\sum_{j=1}^N \left(\alpha_{ij} f_{ij}\right)^{[n+1]}\right)_{i=1}^M,$$

$$(8.28)\qquad \left(\mathbb{M}_{\mathrm{l}} + \theta\tau\mathbb{L}\right)^M \boldsymbol{u}^{n+1} = \left(\mathbb{M}_{\mathrm{l}}\right)^M \tilde{\boldsymbol{u}},$$

where the superscript $[n+1]$ indicates that the fluxes and limiters depend on the solution at time instant t^{n+1}. The function $\overline{\boldsymbol{u}}$, which is equipped with the boundary conditions at $t^{n+1-\theta}$, has to be computed only in the first step. This function is needed because it enters the definition of a lower and an upper bound in the limiting process, see (8.29) below. Then, solving (8.27)–(8.28) has to be performed with an iterative process, where the boundary conditions at t^{n+1} are utilized in (8.28).

First, positivity preservation will be discussed. Let $\boldsymbol{u}^n \geq 0$. Assuming the validity of the CFL condition (8.22), one has (8.21) and hence $\overline{\boldsymbol{u}} \geq 0$ since $\mathbb{M}_\mathrm{l}$ is a diagonal matrix with positive diagonal entries. In the next step, $u_i^{\min} \geq 0$, $i = 1, \ldots, M$, are chosen and the limiters are determined such that $\tilde{u}_i \geq u_i^{\min}$, $i = 1, \ldots, M$, in (8.27). Finally, since $\mathbb{M}_\mathrm{l} \geq 0$ and the system matrix of (8.28) equipped with Dirichlet boundary conditions (which are assumed to be non-negative) is an M-matrix, it follows from Corollary 8.7 that $\boldsymbol{u}^{n+1} \geq 0$.

For studying the satisfaction of the global DMP (cf. Definition 8.5), the computation of the limiters has to be explained in detail. Let $\boldsymbol{u}^{(m)}$ be an approximation of $\boldsymbol{u}^{n+1}$ after the mth iteration for solving (8.27)–(8.28). Then, the algebraic fluxes defined in (8.24) are approximated using $\boldsymbol{u}^{(m)}$ instead of $\boldsymbol{u}^{n+1}$, leading to fluxes $f_{ij}^{(m)}$. Consider any $i \in \{1, \ldots, M\}$ and define

$$\overline{u}_i^{\min} = \min_{j \in S_i \cup \{i\}} \overline{u}_j \,, \qquad \overline{u}_i^{\max} = \max_{j \in S_i \cup \{i\}} \overline{u}_j \,, \tag{8.29}$$

with S_i given by (2.4). Then the limiters $\alpha_{ij}^{(m)}$, where the superscript indicates that they depend on $f_{ij}^{(m)}$, are computed such that

$$\overline{u}_i^{\min} \leq \tilde{u}_i \leq \overline{u}_i^{\max}, \tag{8.30}$$

where $\tilde{\boldsymbol{u}}$ is the solution of (8.27) with the fluxes $f_{ij}^{(m)}$ and the limiters $\alpha_{ij}^{(m)}$. Consider the upper bound and introduce non-negative numbers R_i^+ such that $\alpha_{ij}^{(m)} \leq R_i^+$ if $f_{ij}^{(m)} > 0$. Then

$$\begin{aligned} \tilde{u}_i &= \overline{u}_i + \frac{\tau}{m_i} \sum_{j=1}^N \alpha_{ij}^{(m)} f_{ij}^{(m)} \leq \overline{u}_i + \frac{\tau}{m_i} \sum_{j=1}^N \alpha_{ij}^{(m)} \left(f_{ij}^{(m)} \right)^+ \\ &\leq \overline{u}_i + \frac{\tau}{m_i} R_i^+ \sum_{j=1}^N \left(f_{ij}^{(m)} \right)^+ . \end{aligned}$$

Thus, to satisfy the upper bound in (8.30), it suffices to require that

$$R_i^+ \leq \frac{m_i}{\tau} \left(\overline{u}_i^{\max} - \overline{u}_i \right) \left(\sum_{j=1}^N \left(f_{ij}^{(m)} \right)^+ \right)^{-1}, \tag{8.31}$$

where the right-hand side is non-negative thanks to the definition (8.29) of $\overline{u}_i^{\max}$. Note that if $\left(f_{ij}^{(m)}\right)^+ = 0$ for all $j = 1, \ldots, N$, then the upper bound in (8.30) always holds and R_i^+ can be defined arbitrarily. Similarly, to satisfy the lower bound in (8.30), it suffices to require that $\alpha_{ij}^{(m)} \leq R_i^-$ if $f_{ij}^{(m)} < 0$ with

$$R_i^- \leq \frac{m_i}{\tau} \left(\overline{u}_i^{\min} - \overline{u}_i \right) \left(\sum_{j=1}^N \left(f_{ij}^{(m)} \right)^- \right)^{-1} . \tag{8.32}$$

Like in the previous case, if $\left(f_{ij}^{(m)}\right)^- = 0$ for all $j = 1, \ldots, N$, then the lower bound in (8.30) always holds and R_i^- can be defined arbitrarily. Since the limiters need to belong to $[0, 1]$ by definition, one has to require $R_i^+ \le 1$ and $R_i^- \le 1$ besides the conditions (8.31) and (8.32). In addition, one has to take into account that the flux property is maintained after having applied the limiters, i.e., $\alpha_{ij}^{(m)} f_{ij}^{(m)} = -\alpha_{ji}^{(m)} f_{ji}^{(m)}$, which requires $\alpha_{ij}^{(m)} = \alpha_{ji}^{(m)}$ since $f_{ij}^{(m)} = -f_{ji}^{(m)}$. Thus, one has to take the smaller value of the above-derived bounds for $\alpha_{ij}^{(m)}$ and $\alpha_{ji}^{(m)}$. Summarizing all these considerations leads to the algorithm for the Zalesak limiter from [136], where for the sake of clarity the iteration index is neglected in its presentation:

1. Compute
$$P_i^+ = \sum_{j=1, j\ne i}^{N} f_{ij}^+, \qquad P_i^- = \sum_{j=1, j\ne i}^{N} f_{ij}^-.$$
2. Compute
$$Q_i^+ = \frac{m_i}{\tau}\left(\overline{u}_i^{\max} - \overline{u}_i\right), \qquad Q_i^- = \frac{m_i}{\tau}\left(\overline{u}_i^{\min} - \overline{u}_i\right).$$
3. Compute
$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \qquad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}.$$
   If the denominator is zero, set the value equal to 1. In addition, both values are set to be 1 at Dirichlet nodes.
4. Compute
$$\alpha_{ij} = \begin{cases} \min\{R_i^+, R_j^-\} & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ \min\{R_i^-, R_j^+\} & \text{if } f_{ij} < 0. \end{cases}$$
   Note that the value for $f_{ij} = 0$ does not possess any impact.

It should be emphasized that, like in the steady-state case, the fluxes and limiters are computed on the basis of the matrices for Neumann boundary conditions.

The nonlinear discretization (8.25), or equivalently (8.26)–(8.28), together with a limiter of the form of Zalesak's limiter and fluxes depending on $\boldsymbol{u}^{n+1}$ is called nonlinear FEM-FCT scheme. The standard approach for computing an approximation to the solution, which is already sketched above, is summarized in Algorithm 8.1. The following theorem shows that, under appropriate conditions, all iterates satisfy the global DMP.

THEOREM 8.10 (Global DMP for the iterates of Algorithm 8.1). *Denote*

$$u^{\min} = \min\left\{u_1^n, \ldots, u_N^n, g_1^{n+1-\theta}, \ldots, g_{N-M}^{n+1-\theta}, g_1^{n+1}, \ldots, g_{N-M}^{n+1}\right\}, \tag{8.33}$$
$$u^{\max} = \max\left\{u_1^n, \ldots, u_N^n, g_1^{n+1-\theta}, \ldots, g_{N-M}^{n+1-\theta}, g_1^{n+1}, \ldots, g_{N-M}^{n+1}\right\}. \tag{8.34}$$

Let $\theta = 1$ or the CFL condition (8.22) be satisfied and let $\boldsymbol{u}^{(0)} = \boldsymbol{u}^n$ in Algorithm 8.1. Let all row sums of $(\mathbb{L})^M$ vanish and let the Zalesak algorithm be applied to compute the flux limiters. Then all iterates $\boldsymbol{u}^{(m)}$, $m = 0, 1, \ldots$, satisfy $u^{\min} \le u_i^{(m)} \le u^{\max}$, $i = 1, \ldots, N$.

Proof. Note that the boundary values of $\overline{\boldsymbol{u}}$ are $g_1^{n+1-\theta}, \ldots, g_{N-M}^{n+1-\theta}$. The CFL condition implies that (8.21) holds. Thus, if all row sums of $(\mathbb{L})^M$ vanish, then the

Algorithm 8.1 Iterative scheme for computing an approximation of the solution of the nonlinear FEM-FCT problem. Let $\boldsymbol{u}^{(0)} = \boldsymbol{u}^n$ and let tol > 0 and a damping factor $\rho \in (0,1]$ be given.

1: Solve (8.26).
2: **for** $m = 0,1,\ldots$ **do**
3: Compute the algebraic fluxes $f_{ij}^{(m)}$ as in (8.24) with $\boldsymbol{u}^{n+1}$ replaced by $\boldsymbol{u}^{(m)}$ and the corresponding limiters $\alpha_{ij}^{(m)}$ by Zalesak's algorithm, such that $(\mathbb{M}_{\mathrm{l}})^M \tilde{\boldsymbol{u}}$ can be computed from (8.27).
4: **if** $\left|(\mathbb{M}_{\mathrm{l}} + \theta\tau\mathbb{L})^M \boldsymbol{u}^{(m)} - (\mathbb{M}_{\mathrm{l}})^M \tilde{\boldsymbol{u}}\right| \le \mathrm{tol}$ **then**
5: Set $\boldsymbol{u}^{n+1} := \boldsymbol{u}^{(m)}$, break.
6: **end if**
7: Solve (8.28) with the right-hand side $(\mathbb{M}_{\mathrm{l}})^M \tilde{\boldsymbol{u}}$ and Dirichlet boundary conditions at t^{n+1}. Denote the solution $\hat{\boldsymbol{u}}$ and set $\boldsymbol{u}^{(m+1)} = \boldsymbol{u}^{(m)} + \rho\big(\hat{\boldsymbol{u}} - \boldsymbol{u}^{(m)}\big)$ for the inner nodes and $\boldsymbol{u}^{(m+1)} = \hat{\boldsymbol{u}}$ for the boundary nodes.
8: **end for**

matrices of equation (8.26) satisfy the assumptions of Corollary 8.7. Hence it follows that $u^{\min} \le \overline{u}_i \le u^{\max}$, $i = 1,\ldots,N$. Since the Zalesak limiter is constructed in such a way that the solution of (8.27) satisfies (8.30), one also has

$$u^{\min} \le \tilde{u}_i \le u^{\max}, \qquad i = 1,\ldots,M.$$

As already mentioned above, the matrix on the left-hand side of (8.28), extended by the rows for the Dirichlet conditions, is an M-matrix. Since the row sums of $(\mathbb{L})^M$ vanish, the matrices in (8.28) satisfy the assumptions of Corollary 8.7 and hence

$$u^{\min} \le \min\left\{\tilde{u}_1,\ldots,\tilde{u}_M, g_1^{n+1},\ldots,g_{N-M}^{n+1}\right\} \le \hat{u}_i\,,$$
$$\hat{u}_i \le \max\left\{\tilde{u}_1,\ldots,\tilde{u}_M, g_1^{n+1},\ldots,g_{N-M}^{n+1}\right\} \le u^{\max}\,,$$

for $i = 1,\ldots,N$. Finally, from $\boldsymbol{u}^{(m+1)} = (1-\rho)\boldsymbol{u}^{(m)} + \rho\hat{\boldsymbol{u}}$ for the inner nodes, it can be inferred that $u^{\min} \le u_i^{(m+1)} \le u^{\max}$, $i = 1,\ldots,N$. ∎

Note that the statement of Theorem 8.10 does not depend on the form of the algebraic fluxes.

Now, one has to study under which conditions the row sums of $(\mathbb{L})^M$ vanish. Since the row sums of $\mathbb{D}$ are zero by construction, the row sums of $(\mathbb{L})^M$ vanish if and only if the row sums of the matrix $(\mathbb{A}_{\mathrm{N}})^M$ vanish. In view of (5.4), this is the case if and only if $\sigma = 0$. The assumption that $\sigma = 0$ has to be expected since it appears already for the continuous version (8.7) of the maximum principle.

Remark 8.11. The group finite element method is an alternative assembling routine of the convective term for $\mathbb{P}_1$ and $\mathbb{Q}_1$ finite elements that is based on matrix-vector multiplications instead on numerical quadrature. It introduces a consistency error, see [15] for a numerical analysis of the method, but it is usually considerably more efficient than the standard discretization, see [74]. The ith row sum of the matrix for the convective term reads as follows [15, 74] for $i = 1, \ldots, M$

$$\sum_{j=1}^{N} \left(\sum_{k=1}^{d} \left(\partial_k \phi_j, \phi_i \right) b_k(\boldsymbol{x}_j) \right),$$

where $b_k(\boldsymbol{x}_j)$ is the value of the kth component of $\boldsymbol{b}$ at the node $\boldsymbol{x}_j$. With the same argument as for the standard discretization, one finds that this row sum vanishes if $\boldsymbol{b}$ is constant with respect to space, i.e., $b_k(\boldsymbol{x}_j) = b_k$. But for general convection fields, the row sums do not vanish and hence, for the group finite element method, the satisfaction of the global DMP can be inferred from Theorem 8.10 only for very special (academic) convection fields. □

LEMMA 8.12 (Local DMP for both substeps of the FEM-FCT scheme). *Let the assumptions of Theorem 8.10 be satisfied, then the substeps of the FEM-FCT scheme satisfy the following local DMPs:*

i) The solution $\overline{\boldsymbol{u}}$ of (8.26) satisfies

$$\min_{j \in S_i \cup \{i\}} u_j^n \le \overline{u}_i \le \max_{j \in S_i \cup \{i\}} u_j^n\,, \qquad 1 \le i \le M\,. \tag{8.35}$$

ii) The solution $\boldsymbol{u}^{n+1}$ of (8.28) satisfies

$$\min\left\{ \overline{u}_i^{\min}, \min_{j \in S_i} u_j^{n+1} \right\} \le u_i^{n+1} \le \max\left\{ \overline{u}_i^{\max}, \max_{j \in S_i} u_j^{n+1} \right\}, \qquad 1 \le i \le M\,. \tag{8.36}$$

Proof. Consider any $i \in \{1, \ldots, M\}$. We will prove only the upper bounds in (8.35) and (8.36) since the proofs of the lower bounds proceed along the same lines.

Denote by $u_i^{\max}$ the right-hand side of (8.35) and set $\mathbb{K} = \mathbb{M}_l - (1-\theta)\tau\mathbb{L}$. Then $(\mathbb{K})^M \ge 0$ due to (8.21). Using the notation $\mathbb{K} = (k_{ij})_{i,j=1}^N$ and the row sum property of $(\mathbb{L})^M$, the solution of (8.26) satisfies

$$m_i \overline{u}_i = \sum_{j \in S_i \cup \{i\}} k_{ij} \left(u_j^n - u_i^{\max} \right) + m_i\, u_i^{\max} \le m_i\, u_i^{\max}\,,$$

which implies the upper bound in (8.35).

Now denote by $u_i^{\max}$ the right-hand side of (8.36). Then the ith row of (8.28) can be written in the form

$$(m_i + \theta\tau l_{ii})\big(u_i^{n+1} - u_i^{\max}\big) = m_i \left(\tilde{u}_i - u_i^{\max} \right) - \theta\tau \sum_{j \in S_i} l_{ij} \left(u_j^{n+1} - u_i^{\max} \right). \tag{8.37}$$

Since $l_{ij} \le 0$ for $j \in S_i$ and the Zalesak limiter is constructed in such a way that $\tilde{\boldsymbol{u}}$ satisfies (8.30), the right-hand side of (8.37) is non-positive. As discussed above, the matrix $\mathbb{L}_{\mathrm{I}}$ is positive definite and hence $l_{ii} > 0$. Thus, (8.37) implies the upper bound in (8.36). □

Summarizing the statements of Lemma 8.12, one finds that the solution of the nonlinear problem (8.25) satisfies

$$u_i^{n+1} \le \max\left\{ \overline{u}_i, \max_{j \in S_i} \overline{u}_j, \max_{j \in S_i} u_j^{n+1} \right\} \le \max\left\{ \max_{j \in S_i \cup \{i\}} u_j^n, \max_{j \in S_i} \overline{u}_j, \max_{j \in S_i} u_j^{n+1} \right\}.$$

Consequently, one cannot conclude that a local DMP of the form formulated in Lemma 8.2 is satisfied for (8.25) since the values of the intermediate solution $\overline{\boldsymbol{u}}$ might determine the maximum on the right-hand side of the above estimate. Likewise, one cannot prove the LED property for the fully discrete problem, but only for both substeps individually. For instance, if u_i^n is a local maximum, it cannot be excluded that $\overline{u}_j > \overline{u}_i$ for some $j \in S_i$. In this case, it is $\overline{u}_i^{\max} \neq \overline{u}_i$ and the LED property of the second substep does not provide information on the value of u_i^{n+1}. That the local DMP and the LED property, which are usually stated in the literature for the semi-discrete problem, cannot be transferred to the fully discrete problem is already mentioned in [103, Ex. 4.56].

In [68], the existence of a solution of (8.26)–(8.28) is proven for arbitrary time steps. The existence and uniqueness of a solution for sufficiently small time steps is shown in [70].

We like to mention that there are in practice a couple of algorithmic issues and variations of the FEM-FCT scheme, like prelimiting. Since this topic is outside the scope of this survey, we refer to [90] or [95, Chapters 7.5, 7.6] for detailed presentations. Note that the global DMP is still satisfied as long as the fluxes are modified before the application of the Zalesak limiter.

Method (8.26)–(8.28) with the fluxes (8.24) and the bounds for the limiter (8.29) is a nonlinear scheme. As shown in Theorem 8.10, an accurate solution of the nonlinear problem is not necessary in order to satisfy the global DMP, since it is satisfied for each iterate, but the accuracy of the numerical solution depends on how accurately the nonlinear problems are solved. However, in practice, it might be of advantage to use a linear version of a FEM-FCT scheme for the sake of high efficiency, thereby accepting some loss of accuracy. Note that already the first FEM-FCT scheme proposed in [111] is a linear scheme. Linear FEM-FCT schemes are systematically derived in [89].

The source of nonlinearity of a nonlinear FEM-FCT scheme is the definition (8.24) of the algebraic fluxes. A linear FEM-FCT scheme can be also considered in the form (8.25), however, the fluxes f_{ij} are independent of the solution $\boldsymbol{u}^{n+1}$ at the new time level. To define these fluxes, the values of $\boldsymbol{u}^{n+1}$ in the formula (8.24) are approximated by the solution of an appropriate problem, e.g., the high-order method (8.19) or the low-order method (8.20), or by extrapolating the solution $\overline{\boldsymbol{u}}$ of the explicit scheme (8.26) to the time level t^{n+1}. For $\theta = 1/2$, such extrapolation was considered in [74], leading to the approximation of $\boldsymbol{u}^{n+1}$ by $2\,\overline{\boldsymbol{u}} - \boldsymbol{u}^n$. Then the fluxes are given by

$$f_{ij} = -m_{ij}\left(\hat{u}_j - \hat{u}_i\right) + d_{ij}\left(\overline{u}_j - \overline{u}_i\right) \tag{8.38}$$

with $\hat{\boldsymbol{u}} = 2(\overline{\boldsymbol{u}} - \boldsymbol{u}^n)/\tau$. Note that

$$(\mathbb{M}_{\mathrm{l}})^M \hat{\boldsymbol{u}} = -(\mathbb{L})^M \boldsymbol{u}^n\,, \tag{8.39}$$

i.e., $\hat{\boldsymbol{u}}$ is an approximation of the time derivative of u corresponding to the low-order scheme (8.20) with $\theta = 0$. Independently of how the algebraic fluxes are defined, the limiting procedure remains the same as for the nonlinear FEM-FCT scheme. In particular, the bounds (8.29) for the limiter are defined using the solution of (8.26). Thus, one obtains the following analog of Theorem 8.10.

COROLLARY 8.13 (Global DMP for the linear FEM-FCT scheme with Zalesak limiter). *Let the algebraic fluxes be defined by* (8.24) *with* $\boldsymbol{u}^{n+1}$ *approximated using the solution of a problem depending on* $\boldsymbol{u}^n$ *such that the fluxes are independent of* $\boldsymbol{u}^{n+1}$. *Let* $\theta = 1$ *or the CFL condition* (8.22) *be satisfied, and let the bounds of the limiter be defined by* (8.29) *with* $\overline{\boldsymbol{u}}$ *from* (8.26). *Let all row sums of* $(\mathbb{L})^M$ *vanish and let the Zalesak algorithm be applied for computing the flux limiters. Then the solution* *of the linear scheme* (8.25) *satisfies* $u^{\min} \le u_i^{n+1} \le u^{\max}$*,* $i = 1,\ldots,N$*, where* $u^{\min}$ *and* $u^{\max}$ *are defined by* (8.33) *and* (8.34)*, respectively.*

Proof. The proof proceeds along the lines of the corresponding proof for the nonlinear FEM-FCT scheme. It was already noted that the concrete form of the fluxes does not play any role. ∎

Another linearization strategy proposed in [89] is a predictor-corrector approach directly based on the basic FCT algorithm. In the first step, an intermediate solution $\overline{\boldsymbol{u}}$ at time level t^{n+1} is computed, e.g., by solving a problem of form (8.20). In this step, one has to ensure that $\overline{\boldsymbol{u}}$ satisfies a global DMP, which will give rise to a CFL condition, like (8.22). The solution $\overline{\boldsymbol{u}}$ is used for computing the algebraic fluxes and the bounds (8.29) for the limiter. Then the flux limiters are computed in the same way as for the nonlinear FEM-FCT method and a corrected solution is defined by

$$(\mathbb{M}_l)^M \boldsymbol{u}^{n+1} = (\mathbb{M}_l)^M \overline{\boldsymbol{u}} + \tau \left(\sum_{j=1}^{N} \alpha_{ij} f_{ij} \right)_{i=1}^{M} \tag{8.40}$$

and Dirichlet boundary conditions at t^{n+1}. The algebraic fluxes can be defined by the formula (8.24) with $\boldsymbol{u}^{n+1}$ replaced by $\overline{\boldsymbol{u}}$, as considered in [103]. In [89], the formula (8.24) is considered with $\theta = 1$, leading to (8.38), where $\hat{\boldsymbol{u}}$ is again an approximation of the discrete time derivative $(\boldsymbol{u}^{n+1} - \boldsymbol{u}^n)/\tau$ which can be defined by (8.39), see [89, 90] for alternative proposals.

THEOREM 8.14 (Global DMP for the predictor-corrector FEM-FCT scheme with Zalesak limiter). *Let* $\overline{\boldsymbol{u}}$ *be the solution of* (8.20) *and let the bounds of the limiter be defined by* (8.29) *using this* $\overline{\boldsymbol{u}}$*. Let the algebraic fluxes be defined by an approximation of* (8.24) *such that they are independent of* $\boldsymbol{u}^{n+1}$ *and let the Zalesak algorithm be applied for computing the flux limiters. Let* $\theta = 1$ *or the CFL condition* (8.22) *be satisfied, and let all row sums of* $(\mathbb{L})^M$ *vanish. Then the corrected solution defined by* (8.40) *satisfies* $u^{\min} \le u_i^{n+1} \le u^{\max}$*,* $i = 1,\ldots,N$*, where* $u^{\min} = \min\{u_1^n,\ldots,u_N^n, g_1^{n+1},\ldots,g_{N-M}^{n+1}\}$ *and* $u^{\max} = \max\{u_1^n,\ldots,u_N^n, g_1^{n+1},\ldots,g_{N-M}^{n+1}\}$*.*

Proof. Since the matrices in (8.20) satisfy all the assumptions of Corollary 8.7, the solution $\overline{\boldsymbol{u}}$ of (8.20) satisfies $u^{\min} \le \overline{u}_i \le u^{\max}$, $i = 1,\ldots,N$. The Zalesak limiter is constructed in such a way that the corrected solution satisfies $\overline{u}_i^{\min} \le u_i^{n+1} \le \overline{u}_i^{\max}$, $i = 1,\ldots,M$, which implies the theorem. ∎

For a comprehensive evaluation of the gain of efficiency and loss of accuracy in using a linear scheme for several academic problems, we refer to the numerical studies in [74]. In that paper, one can find also comparisons with a linear upwind finite element method and an example where some shortcomings of the FEM-FCT method are presented.

9. Other types of finite elements.

This section discusses results concerning the DMP and corresponding methods for finite elements other than continuous piecewise linears. It turns out that the results are often negative, at least in dimensions higher than one, and that there are only few methods for which a DMP can be proven. This situation justifies the concentration on the $\mathbb{P}_1$ finite element in the previous sections.

9.1. $\mathbb{Q}_1$ finite element.

Triangulations made of quadrilaterals in two dimensions or hexahedra in three dimensions are widely used for problems from fluid dynamics. The lowest order continuous finite element space on such triangulations is the space $\mathbb{Q}_1$ consisting of piecewise d-linear functions. Strictly speaking, one has to distinguish between two types of such spaces, namely mapped and unmapped $\mathbb{Q}_1$ finite elements. For the mapped version the local space is defined on a reference cell $\hat{K}$, e.g., $\hat{K} = [-1,1]^d$. Then, the finite element space on a physical mesh cell K is given by the reference map from $\hat{K}$ to K. For the unmapped version the local functions are defined directly on the physical mesh cells. Both definitions coincide if the reference map is affine, i.e., if K is a parallelepiped. If this is not the case, the image of a d-linear function defined on $\hat{K}$ will not be a d-linear function on K.

Concerning $\mathbb{Q}_1$ finite elements, investigations of the DMP have been concentrated so far on meshes whose cells are Cartesian products of intervals, sometimes called blocks in the literature. For the Poisson equation in two dimensions, it had been observed already in [30] that the DMP is violated if the aspect ratio, i.e., the ratio of the lengths of the longest edge and the shortest edge of the cell, becomes too large. Based on the tensor-product representation of the basis functions by one-dimensional basis functions, one can derive with a straightforward calculation a formula for the local entries ℓ_{ij}^K of the diffusion matrix, compare [128, Sec. 4.6]. If the corresponding nodes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ share a common edge E_1, then one finds in particular that

$$\ell_{ij}^K = -\frac{|K|}{3^{d-1}}\left(\frac{1}{h_{E_1}^2} - \sum_{k=2}^{d}\frac{1}{2h_{E_k}^2}\right),$$

where $E_1,\ldots,E_d$ are mutually orthogonal edges of K. Thus, for $d=2$, one obtains a non-positive entry, which is condition (3.4) for a matrix of non-negative type, if the aspect ratio is lower than or equal to $\sqrt{2}$. For $d=3$, the mentioned entries are non-negative only for cubes, namely $\ell_{ij}^K = 0$, see also [76]. Considering the relaxed requirement that the diffusion matrix should be monotone, then numerical studies in [85] reveal that the aspect ratios might be larger, at least on sufficiently fine grids, about 2.16 for $d=2$ and 1.05 for $d=3$. An extension of the analysis to reaction-diffusion equations can be found in [128, Sec. 4.6].

9.2. Higher order H^1-conforming finite elements. Concerning the investigation of the DMP, a major difference between higher order H^1-conforming finite element functions and $\mathbb{P}_1$ functions is as follows. Whereas local extrema are attained for $\mathbb{P}_1$ functions only in the degrees of freedom, i.e., geometrically at the vertices of the mesh cells, this is not the case for higher order finite element functions. As simple example, a one-dimensional standard $\mathbb{P}_2$ basis function is depicted in Figure 7, which takes its minimum between the locations of the degrees of freedom.

A local DMP whose definition is restricted to the degrees of freedom has been studied for the Poisson equation in two dimensions already in [106, 57]. It is shown in [57] that such a DMP is satisfied for $\mathbb{P}_2$ finite elements only in special situations: on triangulations with equilateral triangles and on meshes consisting of squares in which the squares are divided by arbitrary diagonals. Note that these special triangulations impose severe restrictions on admissible forms of the domain. A more recent numerical study in [127] shows that for $\mathbb{P}_2$ elements also triangulations with 'nearly' equilateral triangles lead to a satisfaction of the DMP with respect to the degrees of freedom and that such a DMP is not satisfied for finite elements of degree three and higher. In addition, it is discussed in [57] that even on special grids a DMP for the degrees of freedom is not valid for $\mathbb{P}_3$ finite elements.

A proposal for extending an algebraically stabilized scheme to $\mathbb{P}_2$ finite elements

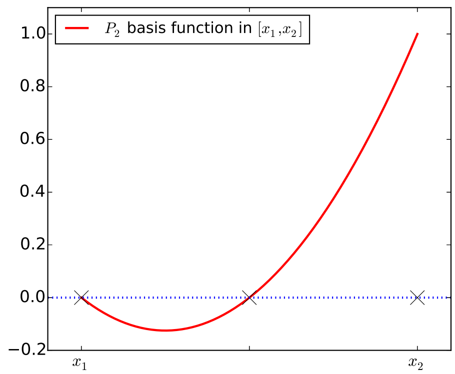


FIG. 7. *Basis function for* $\mathbb{P}_2$ *in the interval* $[x_1, x_2]$*. The degrees of freedom are indicated with black crosses. The function is non-negative at the degrees of freedom, but takes negative values in the interval.*

such that the DMP for the nodal values is satisfied can be found in [88].

Already in [57], an example is given that the DMP for the degrees of freedom does not imply a DMP for the finite element function. This issue might be crucial in coupled problems, when the $\mathbb{P}_2$ finite element solution is a coefficient in other equations and sufficiently accurate quadrature rules have to be utilized for assembling the finite element terms of the other equations. Usually, the nodes of such quadrature rules do not coincide with the geometric positions of the degrees of freedom of the $\mathbb{P}_2$ finite element function.

In [106], the special case of a triangulation consisting of squares that are divided by diagonals which have all the same direction is studied. The proof of the DMP relies on a sufficient condition for the system matrix to be monotone. This condition is based, interestingly, on an additive decomposition of the system matrix, in its diagonal, a term that contains all positive off-diagonal entries, and a term that contains all negative off-diagonal entries. Then, it is assumed that the last term admits another additive decomposition that satisfies appropriate properties. A way that might be successful for deriving such a decomposition is provided. For details, it is referred to [106, 100].

At least for one-dimensional problems, some progress concerning the validation of the DMP has been achieved, e.g., in [129, 130]. These results will not be discussed here since they do not generalize to higher dimensions. Another direction of research, inspired by [117], consists in proving a so-called weak DMP, i.e., in showing that $\|u_h\|_{\infty,\Omega} \le C\|u_h\|_{\infty,\partial\Omega}$, where C is independent of the mesh width, e.g., see [99] for a recent contribution. Although mathematically certainly of interest, the weak DMP does not ensure the physical consistency of the numerical solution, even for $C = 1$, e.g., if the solution is a concentration that should take values in $[0, 1]$ in Ω and equals 1 at some part of $\partial\Omega$, then negative values can still appear in a corresponding numerical solution. A further direction of research consists in applying finite difference techniques for deriving a discrete problem for $\mathbb{Q}_2$ finite elements, e.g., see [100] for a recent paper, which studies reaction-diffusion equations in two dimensions. Such methods possess the usual restriction of finite difference methods to simple domains. Results presented in [100] include the satisfaction of the global DMP on uniform meshes for the Poisson equation. If the uniform mesh is sufficiently fine, then the global DMP is also satisfied for the reaction-diffusion equation.

Remark 9.1. *Bernstein finite element methods.* The presentation of the FCT schemes in Section 8.4 is completely algebraic, it did not exploit any special property of $\mathbb{P}_1$ finite elements. Only some general properties were used, like that the finite element basis forms a partition of unity and that the off-diagonal entries of $\mathbb{M}_c$ are non-negative in order to obtain a well-defined lumped mass matrix. These two properties are also satisfied if the finite element basis consists of local Bernstein polynomials of some degree. The finite element solution can be represented as a linear combination of these basis functions, which are non-negative, with so-called Bernstein coefficients. However, even in points that are degrees of freedom, the value of the solution usually does not coincide with one of the Bernstein coefficients, in contrast to Lagrangian basis functions. All statements proved in Section 8.4 can be transferred to a FEM-FCT scheme with Bernstein polynomials, where everywhere the solution u has to be replaced by the Bernstein coefficients, because they appear in the algebraic problems. Such a scheme for scalar transport equations is studied in [104]. □

9.3. Non-conforming finite elements of Crouzeix–Raviart type. Consider a simplicial triangulation $\mathscr{T}_h$ of Ω. Then, the lowest order non-conforming finite element space of Crouzeix–Raviart-type, proposed in [32], is defined by

$$\mathbb{P}_1^{\mathrm{nc}} = \big\{v_h \in L^2(\Omega) \ : \ v_h|_K \in \mathbb{P}_1(K) \ \forall\, K \in \mathscr{T}_h, \ v_h \text{ is continuous at the barycenters of all facets}\big\}.$$

Functions from $\mathbb{P}_1^{\mathrm{nc}}$ are usually discontinuous across facets, so $\mathbb{P}_1^{\mathrm{nc}} \not\subset H^1(\Omega)$. The degrees of freedom are assigned to the facets. Consequently, the support of each nodal basis function consists of not more than two mesh cells. This property results in a small communication overhead in simulations on parallel computers. Furthermore, the localized support leads to quite sparse matrices for many discretizations.

An upwind method for $\mathbb{P}_1^{\mathrm{nc}}$ was proposed in [110]. To this end, a dual domain or lumping domain for each degree of freedom is considered. Since the degrees of freedom are assigned to the facets, the construction of the dual domain is much easier than for $\mathbb{P}_1$. For each degree of freedom, it is the polytope whose vertices are the vertices of the corresponding facet and the barycenter(s) of the mesh cell(s) where the facet belongs to. Integration by parts on the dual grid is applied to the convective term and then the fluxes across the facets of the dual mesh cells are approximated by an upwind technique. The construction of the upwind fluxes leads on triangulations of acute type to a convection matrix that is of non-negative type. Also the diffusion matrix for $\mathbb{P}_1^{\mathrm{nc}}$ is of non-negative type on acute grids. Its restriction to the degrees of freedom that are not on the Dirichlet boundary is invertible, since the corresponding bilinear form is coercive with respect to a piecewise defined $H^1(\Omega)$ seminorm, which is a norm in the subspace of $\mathbb{P}_1^{\mathrm{nc}}$ consisting of functions vanishing at barycenters of facets contained in the Dirichlet boundary. Thus, from [81, Theorem 5.1] one can conclude the existence of a unique solution of the discrete problem and from Theorems 3.4 and 3.5 the satisfaction of the local and global DMP for the degrees of freedom, respectively, on acute triangulations.

To the best of our knowledge, this upwind method is nowadays rarely used for the numerical solution of convection-diffusion-reaction equations. However, it gained some usefulness in the construction of multigrid methods for incompressible flow problems. For such problems, the pair $\mathbb{P}_1^{\mathrm{nc}}/\mathbb{P}_0$ satisfies a discrete inf-sup condition and applying the upwind technique from [110] leads to a convection-stabilized discretization of the incompressible Navier–Stokes equations. It was proposed in [72] to utilize this discretization on lower levels of a multigrid method, leading to the so-called multiple discretization multilevel (MDML) method. A more recent comparison of solvers for the incompressible Navier–Stokes equations that includes the MDML method can be found in [1].

The upwind technique from [110] can be extended in a straightforward way to non-conforming rotated bilinear finite elements of lowest order for quadrilaterals and hexahedra proposed in [113], see [125].

9.4. Discontinuous Galerkin finite element methods. Discontinuous Galerkin (DG) methods were already proposed in [114] for first order hyperbolic problems. They started to become also popular for discretizing second order elliptic equations in the 1990s. Meanwhile, a number of monographs are available, e.g., [115, 37, 38].

In DG methods, the finite element space consists of piecewise polynomials that are completely discontinuous across facets of the mesh cells. Thus, a DG finite element function is usually not contained in $H^1(\Omega)$.

For DG methods, the notion of 'satisfying a DMP' has to be revisited. In several papers on time-dependent transport and convection-diffusion equations, e.g., [138, 140], the fact that DG allows to use the cell averages in natural way has been used to restrict the DMP to these quantities, and then the following criterion has been proposed: let the cell-wise averages of the DG solution u^n at time instant t^n be in $[u^{\min}, u^{\max}]$, then the DG method satisfies a DMP if the averages of u^{n+1} at t^{n+1} are also contained in this interval. For a detailed discussion of such methods, it is referred to the respective literature, e.g., [119]. An alternative approach, more algebraic and based in the concept of invariant sets and domains for hyperbolic problems, is followed in [54, 56, 112].

In here, we will detail an approach proposed for the convection-diffusion equation in [7]. We start by defining the first order discontinuous space on a simplicial grid, that is[1]

$$\mathbb{P}_1^{\text{disc}} = \left\{ v_h \in L^2(\Omega) \ : \ v_h|_K \in \mathbb{P}_1(K) \ \forall\, K \in \mathscr{T}_h \right\}.$$

This space is equipped with the basis $\{\phi_i^K\}$, where for a mesh cell K and a node i such that $\boldsymbol{x}_i$ is a vertex of K, the function ϕ_i^K is defined as follows: ϕ_i^K is linear in K, $\phi_i^K|_K(\boldsymbol{x}_i) = 1$, $\phi_i^K|_K = 0$ at all other vertices of K, and ϕ_i^K vanishes outside of K. The restriction of $v_h \in \mathbb{P}_1^{\text{disc}}$ to a mesh cell K is denoted by v_h^K.

The first observation is that even the notion of a local extremum is not clear for functions from $\mathbb{P}_1^{\text{disc}}$, compare Fig. 8, where the values at $\boldsymbol{x}_i$ are both a strict local minimum and a strict local maximum. To this end, the following definition was introduced in [7].

Definition 9.2 (Local discrete extremum for $\mathbb{P}_1^{\text{disc}}$). *The function $u_h \in \mathbb{P}_1^{\text{disc}}$ has a local discrete minimum (resp. maximum) at the vertex $\boldsymbol{x}_i$ in K if $u_h^K(\boldsymbol{x}_i) \le u_h(\boldsymbol{x})$ (resp. $u_h^K(\boldsymbol{x}_i) \ge u_h(\boldsymbol{x})$) for all $\boldsymbol{x} \in \omega_i$.*

Then, a definition of a DMP for methods using $\mathbb{P}_1^{\text{disc}}$ is given in [7], which is inspired by Definition 3.17 for nonlinear forms with $\mathbb{P}_1$ functions.

[1] Strictly speaking, the functions of $\mathbb{P}_1^{\text{disc}}$ are well-defined only on the interiors of the mesh cells, since the limits to the same point at the boundaries of mesh cells, approached from different mesh cells, are usually different. To simplify the presentation, we will nevertheless speak of values on facets or at vertices and mean always the limit from the corresponding mesh cell.

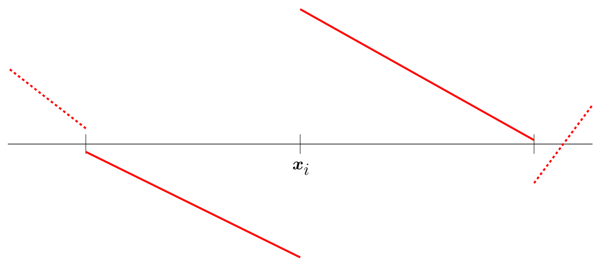


FIG. 8. $\mathbb{P}_1^{\mathrm{disc}}$ *function (in red) with local minimum and local maximum at* $\boldsymbol{x}_i$.

DEFINITION 9.3 (DMP for $\mathbb{P}_1^{\mathrm{disc}}$). *Let* $a_h : \mathbb{P}_1^{\mathrm{disc}} \times \mathbb{P}_1^{\mathrm{disc}} \to \mathbb{R}$ *be a bilinear form. This bilinear form is said to possess the DMP property if for all* $u_h \in \mathbb{P}_1^{\mathrm{disc}}$ *and for all interior vertices* $\boldsymbol{x}_i$ *where* u_h *is locally minimal (resp. maximal) at* $\boldsymbol{x}_i$ *in* K, *there exist constants* $\alpha_F > 0$ *and* $\zeta_K > 0$ *such that*

$$a_h\left(u_h, \phi_i^K\right) \le -\sum_{F\in\mathscr{F}_i\cap\mathscr{F}_K} \frac{\alpha_F}{h_F}\int_F |[\![u_h]\!]_F| \, d\boldsymbol{s} - \frac{\zeta_K}{h_K}\int_K \left|\nabla u_h^K\right| \, d\boldsymbol{x}, \tag{9.1}$$

(resp. $a_h\left(u_h, \phi_i^K\right) \ge \sum_{F\in\mathscr{F}_i\cap\mathscr{F}_K} \frac{\alpha_F}{h_F}\int_F |[\![u_h]\!]_F| \, d\boldsymbol{s} + \frac{\zeta_K}{h_K}\int_K \left|\nabla u_h^K\right| \, d\boldsymbol{x}$*).*

Next, the consistency of the preceding definitions will be shown.

LEMMA 9.4 (Consequences of the satisfaction of the DMP). *Let* $a_h : \mathbb{P}_1^{\mathrm{disc}} \times \mathbb{P}_1^{\mathrm{disc}} \to \mathbb{R}$ *be a bilinear form that satisfies the DMP property from Definition 9.3 and consider the problem to find* $u_h \in \mathbb{P}_1^{\mathrm{disc}}$ *such that* $a_h(u_h, v_h) = (f, v_h)$ *for all* $v_h \in \mathbb{P}_1^{\mathrm{disc}}$.

i) If $f \ge 0$ *(resp.* $f \le 0$*), then* u_h *does not possess a strict local discrete minimum (resp. maximum), see Definition 9.2, at any interior point.*

ii) If $f \ge 0$ *(resp.* $f \le 0$*), then* u_h *attains its global minimum (resp. maximum) at the boundary* $\partial\Omega$.

Proof. i) Assume that u_h has a strict local discrete minimum at the interior node $\boldsymbol{x}_i$ in the mesh cell K. Since $a_h(\cdot,\cdot)$ satisfies the DMP property, it follows from (9.1) that $a_h(u_h, \phi_i^K) \le 0$. On the other hand, one has $(f, \phi_i^K) \ge 0$ and then $a_h(u_h, \phi_i^K) = 0$ holds. From (9.1), one infers that then $\nabla u_h^K = \boldsymbol{0}$ and hence u_h^K is constant so that the minimum is not strict.

ii) If $u_h^K(\boldsymbol{x}_i)$ is a global minimum for some mesh cell K and some interior node $\boldsymbol{x}_i \in K$, then it is also a local minimum and from the proof of i), one gets that u_h^K is constant. Moreover, it follows from the DMP property that $[\![u_h]\!]_F = 0$ for all $F \in \mathscr{F}_i \cap \mathscr{F}_K$. Let $K' \subset \omega_i$ be a mesh cell that shares a common facet F with K. As the jump $[\![u_h]\!]_F$ vanishes, then $u_h^K(\boldsymbol{x}) = u_h^{K'}(\boldsymbol{x})$ for all $\boldsymbol{x} \in F$, and in particular $u_h^K(\boldsymbol{x}_i) = u_h^{K'}(\boldsymbol{x}_i)$. Thus, $u_h^{K'}(\boldsymbol{x}_i)$ also is a global minimum and it follows that $u_h^{K'}$ is constant. By induction, one finds that $u_h|_{\omega_i} = u_h^K(\boldsymbol{x}_i)$ is a constant. Then, again by induction, it follows that u_h is constant in Ω and in particular that $u_h|_{\partial\Omega} = u_h^K(\boldsymbol{x}_i)$. Hence, the global minimum is attained at the boundary of Ω. ∎

One type of equations studied in [7] is a steady-state convection-diffusion equation with conservative form of the convective term and solenoidal convection field. For the DG discretization of the diffusive term, the standard incomplete interior penalty (IIP) method is used. This choice is motivated by the analysis of one-dimensional diffusion problems that are discretized with DG methods, see [58]. The convective term is integrated by parts and then an upwind discretization at interior facets is utilized. In addition, and this is the major algorithmic proposal of [7], a nonlinear, locally defined artificial diffusion term built with the help of a shock detector is added. For a one-dimensional problem, the DMP, according to Definition 9.3, is proven. There are no analytic results for multiple dimensions. The main obstacle for such results is that a DMP is not available already for the usual interior penalty discretizations of the diffusion term. In the numerical studies presented in [7], small violations of the DMP can be observed for a simulation performed on an acute mesh in two dimensions.

A method that addresses the above mentioned issue of the DMP for interior penalty discretizations of the diffusive term is proposed in [5]. This method augments the symmetric interior penalty method with a nonlinear discrete diffusion operator related to the AFC/FCT schemes described in previous sections. Then, it is shown in [5] that the proposed scheme for the steady-state convection-diffusion problem satisfies a local DMP if the right-hand side of the equation vanishes identically. This statement holds for arbitrary admissible grids and $\mathbb{P}_1^{\mathrm{disc}}$ finite elements on simplices and discontinuous piecewise d-linear elements on quadrilaterals or hexahedra. For the time-dependent case, a semi-discrete problem in space is considered and it is shown that the discrete scheme is LED, again in case that the right-hand side of the problem is identically zero.

High-order DG schemes based on algebraic flux correction were recently developed in [56, 112] for hyperbolic conservation laws. While in [56] monolithic convex limiting with subcell flux limiters is used, in [112] an FCT-type predictor-corrector algorithm is advocated. The bound preserving DG scheme of [56] employs Bernstein polynomials to facilitate the use of very high order spatial approximations. The limiting strategy of [112] is tailor-made for Legendre-Gauss-Lobatto DG bases, and makes use of a novel sparse low-order invariant domain preserving method whose stencil does not grow with the polynomial degree of the corresponding high-order method. The invariant domain preservation is proved under a CFL condition.

10. Brief comments on hyperbolic conservation laws. The aim of this section is to discuss briefly results on the satisfaction of the DMP for transport equations and nonlinear hyperbolic conservation laws. Presenting in detail the amount and variety of works devoted to hyperbolic problems requires a review on its own and it is clearly outside the scope of the present survey. In particular, in this section we will only focus on continuous finite element methods, since for discontinuous Galerkin approaches there exist several well documented reviews (e.g., [139, 119]). In addition, in recent years there has been an increasing interest in seeking suitable conforming approximations for hyperbolic problems, since conforming approximations do not have a built-in stability, and hence the challenge of finding structure-preserving stabilizing terms is different from the discontinuous counterparts.

The model problem considered in this section is the extreme case $\varepsilon = 0$, this is, the transport equation, or, more generally, conservation equations of the form

$$\partial_t u + \operatorname{div} \boldsymbol{f}(u) = 0 \qquad \text{in } \Omega\,, \tag{10.1}$$

where $\boldsymbol{f}(u)$ is the flux function, provided with appropriate (inlet) boundary and initial conditions. If $\boldsymbol{f}(u) = \boldsymbol{b}u$, then (10.1) reduces to the linear transport equation.

Remark 10.1. It is worth mentioning that the case $\varepsilon = 0$ allows to propose methods that respect the DMP on general meshes in a more natural way. In fact, the added viscosity methods only need to deal with compensating for the wrong signed terms in the convection matrix, and not with the possibly positive terms in the diffusion matrix, which are of a different order in terms of the mesh size. For example, in [25] an appropriate combination of upwinding and FCT-related techniques is used to propose a nonlinear stabilized scheme that preserves the DMP for the linear transport equation. In addition, the time discretization is based on an explicit method, so the overhead of using a nonlinear discretization is minimal. On the other hand, it is important to mention that the discrete maximum principle is not, in general, enough to prove the convergence of a numerical scheme to the entropy solution of (10.1), as it has been mentioned in, e.g., [53], where the authors show that, in order to converge to the entropy solution, the scheme needs also to control the maximum wave speed. More precisely, in Lemma 4.6 in that reference, it is shown that the FCT algorithm, equipped with a limiter related to the Zalesak one, might not converge for certain nonlinear fluxes, which is then confirmed in the numerical experiments for Burgers' equation. □

We start by mentioning that most of the references quoted in Section 8.4 were, in fact, works developed for the transport, or Euler, equations. So, this section will be devoted to describing some of the more recent developments of DMP-preserving schemes for this problem. In [92], using the framework of algebraic flux correction and invariant domain preserving schemes, a monolithic approach to convex limiting is introduced for hyperbolic conservation laws. The convex limiting is thoroughly discussed for both scalar conservation laws (including the transport equation) and hyperbolic systems. In the context of the enriched finite element method (proposed originally in [17]), a FCT scheme for the transport equation is proposed in [94] where the DMP is proven (under appropriate CFL conditions) for both the continuous and discontinuous parts of the solution.

In [51] a first order added diffusion/viscosity method with an explicit time discretization is proposed for (10.1). The DMP for the resulting scheme is proven under a CFL condition. On uniform meshes, the bilinear form of the first order diffusion used in [51] corresponds to the matrix $M_{\rm C} - M_{\rm L}$ used in [105]. Later, in [55] the authors show that it is impossible to propose an explicit continuous finite element method that is stabilized with artificial viscosity and satisfies the DMP if the time derivative is approximated using the consistent mass matrix. In the paper [52] the authors propose a different technique: first, a higher order added viscosity (defined as the minimum between the first order viscosity and the entropy residual) is added. The DMP cannot be proven for the resulting scheme, so they use a technique related to the FCT method (linked to the graph-Laplacian writing of the added viscosity), supplied with flux limiters related to those described in Section 8.4 (based on the Zalesak algorithm), and an approximation of the inverse of the consistent mass matrix to correct the scheme. The combination of these techniques allows for the proof of the DMP. Later, in [53] a method, again related to the FCT family, is proposed, equipped with three different limiters, namely the Zalesak limiter, the smoothness-based indicator, and a greedy viscosity algorithm. In addition, the satisfaction of the DMP and the convergence to the entropy solution are shown. Some comparisons in terms of robustness and reliability are also carried out in [53]. Another work devoted to stabilization by the nonlinear diffusion operator (also referred to as graph Laplacian in some papers) is the work [4], where a regularization of the definition of the limiters is proposed in order to obtain twice differentiable limiters and to make the discretization amenable to the use of Newton's method to solve the algebraic system.

In the context of the Burgers equation, in [24] numerical viscosity is added to satisfy the DMP and prove convergence to the entropy solution of the hyperbolic equation. In one space dimension the method consists of adding a numerical diffusion of the form $(\varepsilon(u_h)\partial_x u_h, \partial_x v_h)$ where $\varepsilon(u_h)$ is designed to satisfy several hypotheses. These conditions imply the Lipschitz continuity of the stabilization and the fact that the problem satisfies the strong DMP property (similar to those in Section 3.2). Under these assumptions, the finite element method is proven to converge to the entropy solution of Burgers' equation. Later, in [6], essentially the same assumptions are imposed on the coefficient of the added diffusion, with the difference that in this case the diffusion is of the form of a local projection stabilization method. The method is proven to be LED and to converge to the entropy solution.

We next comment on the possibility of using both linear and nonlinear stabilizing terms in conservation laws. In fact, as it was mentioned in previous sections, it has been observed in several works that the use of a nonlinear stabilization (e.g., FCT) alone does not suffice to build a convergent method. For example, in [53] it is shown that using nonlinear stabilization alone leads, in certain cases, to failure in convergence of the scheme. So, the authors take a different approach by first adding an entropy viscosity to a method by using the consistent mass matrix, thus violating the DMP, and then applying a FCT technique as a post-processing to produce a DMP-preserving approximate solution. In addition, in the work [40] a combined use of linear (edge-based) stabilization and a nonlinear entropy viscosity is advocated. It is shown in that reference that the addition of linear stabilization, if not weighted properly, can actually hinder the satisfaction of the DMP and increase the entropy violations, and even in some extreme cases, make a convergent method converge to the wrong weak solution. So, a nonlinear weight is introduced to balance the influence of the stabilizing terms and secure convergence to the entropy solution. We should, nevertheless, mention that even if the entropy viscosity method is claimed to satisfy a *weakened* maximum principle, there is no proof of DMP-satisfaction (or weakened DMP) available, although the authors show numerical evidence supporting the claim that the weighted method does satisfy a weakened DMP.

We finish this short section by mentioning two relatively recent works where DMP-preserving methods are introduced and that use LPS-related methods as linear stabilization. In [93] a linear stabilizing term is first introduced. This term penalizes the fluctuations between the discrete solution and its local average (thus inspired by the LPS idea, but departing from the classical LPS approaches). This method preserves the DMP but provides inaccurate results, so the target function, that is, the function with respect to which the fluctuation is computed, is modified by adding to it an approximation of its gradient. This approximation is then limited using limiters that guarantee the LED property and linearity preservation (on general meshes) of the resulting scheme. The authors claim that the linearity preserving limiter introduced in [93, Section 7] can also be applied in different contexts, e.g., the AFC and FCT schemes. The resulting method is tested in steady-state and time-dependent schemes showing that the combination of the gradient approximation as high order stabilization with the LED limiter localizes the stabilization enough as to reduce the oscillations around the shocks without smearing the profiles in excess. Finally, in [108] the authors present a nonlinear stabilization through discrete artificial diffusion supplemented by a monotone local projection operator based on limiting at the semi-discrete level. The resulting method respects the DMP and is linearity preserving. The impact of the local projection operator is studied in the numerical experiments where it is shown that its addition (that acts as a high order background dissipation) helps to reduce the terracing (and even eliminates it in some cases).

11. Summary. For convection-dominated convection-diffusion problems it is a challenging task to construct discretizations that at the same time satisfy the DMP and compute accurate solutions. Enormous efforts have been spent since the 1980s in the development of schemes that enrich traditional stabilized finite element methods with extra terms to reduce the size of spurious oscillations, leading to the class of SOLD methods. However, this development turned out to be only little successful with respect to designing methods for which the DMP can be proven rigorously, since only the Mizukami–Hughes method satisfies this property. In the 2000s, a different class of methods was started to be developed, namely algebraically stabilized finite element methods. In that decade, FEM-FCT schemes for the time-dependent problem were proposed and at the end of that decade, the first AFC method for the steady-state problem. Then, in recent years, the analysis for AFC methods have been developed and further methods for the steady-state problem, like modifications and extensions of algebraic stabilizations, have been developed. For all of these schemes, the DMP can be proven, sometimes under conditions on the data or the mesh. In summary, there are meanwhile several, but still surprisingly few, finite element methods available that satisfy the DMP and compute simultaneously quite accurate results.

For the steady-state problem, all DMP-respecting finite element schemes with accurate solutions are nonlinear. It can be seen in the numerical example from Section 7 that, on the one hand, there are differences concerning the accuracy of the computed solutions, but on the other hand, the differences are not large. For the practical use of these methods, also aspects like the efficiency for solving the nonlinear problems and the efforts for implementing the methods in three dimensions are important. Concerning the first issue, whose investigation is outside the scope of this survey, a comprehensive comparison of two algebraically stabilized schemes can be found in [64]. Simulations of three-dimensional problems with various algebraic stabilizations can be found in [14, 64]. Note the many algebraic stabilizations do work only with the matrices and vectors such that their implementation can be carried out independently of the dimension of the problem.

There is a similar situation for the time-dependent problem: algebraically stabilized schemes are the currently best available finite element methods that satisfy the global DMP and compute accurate solutions. Here, also a linear variant is available which showed in several applications a good balance of accuracy and efficiency.

Acknowledgment. This work was initiated at a stay of the three authors at the Mathematisches Forschungsinstitut Oberwolfach (MFO) within the *Research in pairs* programme, grant No. 1937p.

REFERENCES

[1] N. Ahmed, C. Bartsch, V. John, and U. Wilbrandt, *An assessment of some solvers for saddle point problems emerging from the incompressible Navier-Stokes equations*, Comput. Methods Appl. Mech. Engrg., 331 (2018), pp. 492–513, https://doi.org/10.1016/j.cma.2017.12.004.

[2] A. Allendes, G. R. Barrenechea, and R. Rankin, *Fully computable error estimation of a nonlinear, positivity-preserving discretization of the convection-diffusion-reaction equation*, SIAM J. Sci. Comput., 39 (2017), pp. A1903–A1927, https://doi.org/10.1137/16M1092763.

[3] K. Baba and M. Tabata, *On a conservative upwind finite element scheme for convective diffusion equations*, RAIRO Anal. Numér., 15 (1981), pp. 3–25, https://doi.org/10.1051/m2an/1981150100031.

[4] S. Badia and J. Bonilla, *Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization*, Comput. Methods Appl. Mech. Engrg., 313 (2017), pp. 133–158, https://doi.org/10.1016/j.cma.2016.09.035.
[5] S. Badia, J. Bonilla, and A. Hierro, *Differentiable monotonicity-preserving schemes for discontinuous Galerkin methods on arbitrary meshes*, Comput. Methods Appl. Mech. Engrg., 320 (2017), pp. 582–605, https://doi.org/10.1016/j.cma.2017.03.032.
[6] S. Badia and A. Hierro, *On monotonicity-preserving stabilized finite element approximations of transport problems*, SIAM J. Sci. Comput., 36 (2014), pp. A2673–A2697, https://doi.org/10.1137/130927206.
[7] S. Badia and A. Hierro, *On discrete maximum principles for discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., 286 (2015), pp. 107–122, https://doi.org/10.1016/j.cma.2014.12.006.
[8] R. E. Bank, W. M. Coughran jr., and L. C. Cowsar, *The finite volume Scharfetter-Gummel method for steady convection diffusion equations*, Comput. Vis. Sci., 1 (1998), pp. 123–136, https://doi.org/10.1007/s007910050012.
[9] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Blending low-order stabilised finite element methods: a positivity-preserving local projection method for the convection-diffusion equation*, Comput. Methods Appl. Mech. Engrg., 317 (2017), pp. 1169–1193, https://doi.org/10.1016/j.cma.2017.01.016.
[10] G. R. Barrenechea, E. Burman, and F. Karakatsani, *Edge-based nonlinear diffusion for finite element approximations of convection-diffusion equations and its relation to algebraic flux-correction schemes*, Numer. Math., 135 (2017), pp. 521–545, https://doi.org/10.1007/s00211-016-0808-z.
[11] 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, https://doi.org/10.1093/imanum/dru041.
[12] G. R. Barrenechea, V. John, and P. Knobloch, *Analysis of algebraic flux correction schemes*, SIAM J. Numer. Anal., 54 (2016), pp. 2427–2451, https://doi.org/10.1137/15M1018216.
[13] G. R. Barrenechea, V. John, and P. Knobloch, *An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes*, Math. Models Methods Appl. Sci., 27 (2017), pp. 525–548, https://doi.org/10.1142/S0218202517500087.
[14] G. R. Barrenechea, V. John, P. Knobloch, and R. Rankin, *A unified analysis of algebraic flux correction schemes for convection-diffusion equations*, SeMA J., 75 (2018), pp. 655–685, https://doi.org/10.1007/s40324-018-0160-6.
[15] G. R. Barrenechea and P. Knobloch, *Analysis of a group finite element formulation*, Appl. Numer. Math., 118 (2017), pp. 238–248, https://doi.org/10.1016/j.apnum.2017.03.008.
[16] R. Becker, E. Burman, and P. Hansbo, *A finite element time relaxation method*, C. R. Math. Acad. Sci. Paris, 349 (2011), pp. 353–356, https://doi.org/10.1016/j.crma.2010.12.010.
[17] R. Becker, E. Burman, P. Hansbo, and M. G. Larson, *A reduced $\mathbb{P}^1$-discontinuous Galerkin method.*, Chalmers Finite Element Center Preprint 2003-13, Chalmers University of Technology, Göteborg, Sweden, 2003.
[18] J. P. Boris and D. L. Book, *Flux-corrected transport. I: SHASTA, a fluid transport algorithm that works.*, J. Comput. Phys., 11 (1973), pp. 38–69, https://doi.org/10.1016/0021-9991(73)90147-2.
[19] J. H. Bramble and B. E. Hubbard, *On the formulation of finite difference analogues of the Dirichlet problem for Poisson's equation*, Numer. Math., 4 (1962), pp. 313–327, https://doi.org/10.1007/BF01386325.
[20] J. H. Bramble and B. E. Hubbard, *New monotone type approximations for elliptic problems*, Math. Comp., 18 (1964), pp. 349–367, https://doi.org/10.1090/S0025-5718-1964-0165702-X.
[21] J. Brandts, S. Korotov, and M. Křížek, *Simplicial Partitions with Applications to the Finite Element Method*, Springer-Verlag, Cham, 2020, https://doi.org/10.1007/978-3-030-55677-8.
[22] J. Brandts, S. Korotov, M. Křížek, and J. Šolc, *On nonobtuse simplicial partitions*, SIAM Rev., 51 (2009), pp. 317–335, https://doi.org/10.1137/060669073.
[23] J. H. Brandts, S. Korotov, and M. Křížek, *The discrete maximum principle for linear simplicial finite element approximations of a reaction-diffusion problem*, Linear Algebra Appl., 429 (2008), pp. 2344–2357, https://doi.org/10.1016/j.laa.2008.06.011.
[24] E. Burman, *On nonlinear artificial viscosity, discrete maximum principle and hyperbolic con-*

servation laws, BIT, 47 (2007), pp. 715–733, https://doi.org/10.1007/s10543-007-0147-7.

[25] E. Burman, *A monotonicity preserving, nonlinear, finite element upwind method for the transport equation*, Applied Mathematics Letters, 49 (2015), pp. 141–146, https://doi.org/https://doi.org/10.1016/j.aml.2015.05.005.

[26] E. Burman and A. Ern, *Nonlinear diffusion and discrete maximum principle for stabilized Galerkin approximations of the convection–diffusion-reaction equation*, Comput. Methods Appl. Mech. Engrg., 191 (2002), pp. 3833–3855, https://doi.org/10.1016/S0045-7825(02)00318-3.

[27] E. Burman and A. Ern, *Discrete maximum principle for Galerkin approximations of the Laplace operator on arbitrary meshes*, C. R. Math. Acad. Sci. Paris, 338 (2004), pp. 641–646, https://doi.org/10.1016/j.crma.2004.02.010.

[28] E. Burman and A. Ern, *Stabilized Galerkin approximation of convection-diffusion-reaction equations: discrete maximum principle and convergence*, Math. Comp., 74 (2005), pp. 1637–1652, https://doi.org/10.1090/S0025-5718-05-01761-8.

[29] E. Burman and P. Hansbo, *Edge stabilization for Galerkin approximations of convection-diffusion-reaction problems*, Comput. Methods Appl. Mech. Engrg., 193 (2004), pp. 1437–1453, https://doi.org/10.1016/j.cma.2003.12.032.

[30] I. Christie and C. Hall, *The maximum principle for bilinear elements*, Internat. J. Numer. Methods Engrg., 20 (1984), pp. 549–553, https://doi.org/10.1002/nme.1620200312.

[31] P. G. Ciarlet, *Discrete maximum principle for finite-difference operators*, Aequationes Math., 4 (1970), pp. 338–352, https://doi.org/10.1007/BF01844166.

[32] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and uniform convergence for the finite element method*, Comput. Methods Appl. Mech. Engrg., 2 (1973), pp. 17–31, https://doi.org/10.1016/0045-7825(73)90019-4.

[33] R. Codina, *A discontinuity-capturing crosswind-dissipation for the finite element solution of the convection-diffusion equation*, Comput. Methods Appl. Mech. Engrg., 110 (1993), pp. 325–342, https://doi.org/10.1016/0045-7825(93)90213-H.

[34] L. Collatz, *Numerische Behandlung von Differentialgleichungen*, Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen mit besonderer Berücksichtigung der Anwendungsgebiete, Bd. LX, Springer-Verlag, Berlin, 1955. 2te Aufl.

[35] L. Collatz, *The numerical treatment of differential equations. 3d ed*, Die Grundlehren der mathematischen Wissenschaften, Bd. 60, Springer-Verlag, Berlin, 1960. Translated from a supplemented version of the 2d German edition by P. G. Williams.

[36] T. A. Davis, *Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method*, ACM Trans. Math. Software, 30 (2004), pp. 196–199, https://doi.org/10.1145/992200.992206.

[37] D. A. Di Pietro and A. Ern, *Mathematical aspects of discontinuous Galerkin methods*, Springer, Heidelberg, 2012, https://doi.org/10.1007/978-3-642-22980-0.

[38] V. Dolejší and M. Feistauer, *Discontinuous Galerkin method. Analysis and applications to compressible flow*, Springer, Cham, 2015, https://doi.org/10.1007/978-3-319-19267-3.

[39] A. Drăgănescu, T. F. Dupont, and L. R. Scott, *Failure of the discrete maximum principle for an elliptic finite element problem*, Math. Comp., 74 (2005), pp. 1–23, https://doi.org/10.1090/S0025-5718-04-01651-5.

[40] A. Ern and J.-L. Guermond, *Weighting the edge stabilization*, SIAM J. Numer. Anal., 51 (2013), pp. 1655–1677, https://doi.org/10.1137/120867482.

[41] A. Ern and J.-L. Guermond, *Finite elements II—Galerkin approximation, elliptic and mixed PDEs*, Springer, Cham, 2021, https://doi.org/10.1007/978-3-030-56923-5.

[42] L. C. Evans, *Partial differential equations*, vol. 19 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, second ed., 2010.

[43] I. Faragó and R. Horváth, *Discrete maximum principle and adequate discretizations of linear parabolic problems*, SIAM J. Sci. Comput., 28 (2006), pp. 2313–2336, https://doi.org/10.1137/050627241.

[44] I. Faragó, J. Karátson, and S. Korotov, *Discrete maximum principles for nonlinear parabolic PDE systems*, IMA J. Numer. Anal., 32 (2012), pp. 1541–1573, https://doi.org/10.1093/imanum/drr050.

[45] M. S. Floater, *Generalized barycentric coordinates and applications*, Acta Numerica, 24 (2015), pp. 161–214, https://doi.org/10.1017/S0962492914000129.

[46] L. P. Franca and C. Farhat, *Bubble functions prompt unusual stabilized finite element methods*, Comput. Methods Appl. Mech. Engrg., 123 (1995), pp. 299–308, https://doi.org/10.1016/0045-7825(94)00721-X.

[47] S. Ganesan, V. John, G. Matthies, R. Meesala, S. Abdus, and U. Wilbrandt, *An object oriented parallel finite element scheme for computing PDEs: Design and implementa-*

tion, in IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW) Hyderabad, IEEE, 2016, pp. 106–115, https://doi.org/10.1109/HiPCW.2016.023.

[48] S. A. Gershgorin, *Fehlerabschätzung für das Differenzenverfahren zur Lösung partieller Differentialgleichungen*, Z. Angew. Math. Mech., 10 (1930), pp. 373–382, https://doi.org/10.1002/zamm.19300100409.

[49] D. Gilbarg and N. S. Trudinger, *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin, second ed., 2001, https://doi.org/10.1007/978-3-642-61798-0.

[50] S. K. Godunov, *A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics*, Mat. Sb. (N.S.), 47 (89) (1959), pp. 271–306.

[51] J.-L. Guermond and M. Nazarov, *A maximum-principle preserving C^0 finite element method for scalar conservation equations*, Comput. Methods Appl. Mech. Engrg., 272 (2014), pp. 198–213, https://doi.org/10.1016/j.cma.2013.12.015.

[52] J.-L. Guermond, M. Nazarov, B. Popov, and Y. Yang, *A second-order maximum principle preserving Lagrange finite element technique for nonlinear scalar conservation equations*, SIAM J. Numer. Anal., 52 (2014), pp. 2163–2182, https://doi.org/10.1137/130950240.

[53] J.-L. Guermond and B. Popov, *Invariant domains and second-order continuous finite element approximation for scalar conservation equations*, SIAM J. Numer. Anal., 55 (2017), pp. 3120–3146, https://doi.org/10.1137/16M1106560.

[54] J.-L. Guermond, B. Popov, and I. Tomas, *Invariant domain preserving discretization-independent schemes and convex limiting for hyperbolic systems*, Comput. Methods Appl. Mech. Engrg., 347 (2019), pp. 143–175, https://doi.org/10.1016/j.cma.2018.11.036.

[55] J.-L. Guermond, B. Popov, and Y. Yang, *The effect of the consistent mass matrix on the maximum-principle for scalar conservation equations*, J. Sci. Comput., 70 (2017), pp. 1358–1366, https://doi.org/10.1007/s10915-016-0285-7.

[56] H. Hajduk, *Monolithic convex limiting in discontinuous Galerkin discretizations of hyperbolic conservation laws*, Comput. Math. Appl., 87 (2021), pp. 120–138, https://doi.org/10.1016/j.camwa.2021.02.012.

[57] W. Höhn and H.-D. Mittelmann, *Some remarks on the discrete maximum-principle for finite elements of higher order*, Computing, 27 (1981), pp. 145–154, https://doi.org/10.1007/BF02243548.

[58] T. L. Horváth and M. E. Mincsovics, *Discrete maximum principle for interior penalty discontinuous Galerkin methods*, Cent. Eur. J. Math., 11 (2013), pp. 664–679, https://doi.org/10.2478/s11533-012-0154-z.

[59] W. Huang, *Discrete maximum principle and a Delaunay-type mesh condition for linear finite element approximations of two-dimensional anisotropic diffusion problems*, Numer. Math. Theory Methods Appl., 4 (2011), pp. 319–334, https://doi.org/10.4208/nmtma.2011.m1024.

[60] T. Ikeda, *Maximum principle in finite element models for convection-diffusion phenomena*, North-Holland, Amsterdam, 1983.

[61] A. Jameson, *Origins and further development of the Jameson-Schmidt-Turkel scheme*, AIAA J., 55 (2017), pp. 1487–1510, https://doi.org/10.2514/1.J055493.

[62] A. Jameson, W. Schmidt, and E. Turkel, *Numerical solution of the Euler equations by finite volume methods using Runge-Kutta time-stepping schemes*, in 14th AIAA Fluid and Plasma Dynamics Conference, Palo Alto, CA (USA), 23-25 Jun 1981, AIAA meeting paper 1981-1259, 1981, https://doi.org/10.2514/6.1981-1259.

[63] A. Jha, *A residual based a posteriori error estimators for AFC schemes for convection-diffusion equations*, Comput. Math. Appl., 97 (2021), pp. 86–99, https://doi.org/10.1016/j.camwa.2021.05.031.

[64] A. Jha and V. John, *A study of solvers for nonlinear AFC discretizations of convection-diffusion equations*, Comput. Math. Appl., 78 (2019), pp. 3117–3138, https://doi.org/10.1016/j.camwa.2019.04.020.

[65] A. Jha, V. John, and P. Knobloch, *Adaptive grids in the context of algebraic stabilizations for convection–diffusion–reaction equations*, 2022, https://arxiv.org/abs/2007.08405.

[66] V. John and P. Knobloch, *On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review*, Comput. Methods Appl. Mech. Engrg., 196 (2007), pp. 2197–2215, https://doi.org/10.1016/j.cma.2006.11.013.

[67] V. John and P. Knobloch, *On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part II – Analysis for P_1 and Q_1 finite elements*, Comput. Methods Appl. Mech. Engrg., 197 (2008), pp. 1997–2014, https://doi.org/10.1016/j.cma.2007.12.019.

[68] V. John and P. Knobloch, *Existence of solutions of a finite element flux-corrected-transport*

scheme, Appl. Math. Lett., 115 (2021), p. Paper No. 106932, https://doi.org/10.1016/j.aml.2020.106932.

[69] V. John and P. Knobloch, *On algebraically stabilized schemes for convection–diffusion–reaction problems*, Numer. Math., 152 (2022), pp. 553–585, https://doi.org/10.1007/s00211-022-01325-9.

[70] V. John, P. Knobloch, and P. Korsmeier, *On the solvability of the nonlinear problems in an algebraically stabilized finite element method for evolutionary transport-dominated equations*, Math. Comp., 90 (2021), pp. 595–611, https://doi.org/10.1090/mcom/3576.

[71] V. John, P. Knobloch, and O. Pártl, *A numerical assessment of finite element discretizations for convection-diffusion-reaction equations satisfying discrete maximum principles*, Comput. Methods Appl. Math., (2022), https://doi.org/10.1515/cmam-2022-0125.

[72] V. John and G. Matthies, *Higher-order finite element discretizations in a benchmark problem for incompressible flows.*, Int. J. Numer. Methods Fluids, 37 (2001), pp. 885–903, https://doi.org/10.1002/fld.195.

[73] 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*, Chemical Engineering Science, 64 (2009), pp. 733–741, https://doi.org/10.1016/j.ces.2008.05.004.

[74] V. John and J. Novo, *On (essentially) non-oscillatory discretizations of evolutionary convection-diffusion equations*, J. Comput. Phys., 231 (2012), pp. 1570–1586, https://doi.org/10.1016/j.jcp.2011.10.025.

[75] H. Kanayama, *Discrete maximum principles for salinity distribution in a bay: conservation law and maximum principle*, Theoretical Appl. Mech., 28 (1978), pp. 559–579.

[76] J. Karátson, S. Korotov, and M. Křížek, *On discrete maximum principles for nonlinear elliptic problems*, Math. Comput. Simulation, 76 (2007), pp. 99–108, https://doi.org/10.1016/j.matcom.2007.01.011.

[77] F. Kikuchi, *Discrete maximum principle and artificial viscosity in finite element approximations to convective diffusion equations*, Institute of Space and Aeronautical Science, University of Tokyo, 550 (1977).

[78] P. Knobloch, *Improvements of the Mizukami–Hughes method for convection–diffusion equations*, Comput. Methods Appl. Mech. Engrg., 196 (2006), pp. 579–594, https://doi.org/10.1016/j.cma.2006.06.004.

[79] P. Knobloch, *Numerical solution of convection–diffusion equations using upwinding techniques satisfying the discrete maximum principle*, in Proceedings of Czech-Japanese Seminar in Applied Mathematics 2005, M. Beneš, M. Kimura, and T. Nakaki, eds., vol. 3 of COE Lect. Note, Kyushu Univ., Fukuoka, 2006, pp. 69–76.

[80] P. Knobloch, *Application of the Mizukami–Hughes method to bilinear finite elements*, in Proceedings of Czech-Japanese Seminar in Applied Mathematics 2006, M. Beneš, M. Kimura, and T. Nakaki, eds., vol. 6 of COE Lect. Note, Kyushu Univ., Fukuoka, 2007, pp. 137–147.

[81] P. Knobloch, *Numerical solution of convection-diffusion equations using a nonlinear method of upwind type*, J. Sci. Comput., 43 (2010), pp. 454–470, https://doi.org/10.1007/s10915-008-9260-2.

[82] P. Knobloch, *On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type*, in Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016, Z. Huang, M. Stynes, and Z. Zhang, eds., vol. 120 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Cham, 2017, pp. 129–139, https://doi.org/10.1007/978-3-319-67202-1_10.

[83] P. Knobloch, *A new algebraically stabilized method for convection–diffusion–reaction equations*, in Numerical mathematics and advanced applications ENUMATH 2019, F. Vermolen and C. Vuik, eds., vol. 139 of Lect. Notes Comput. Sci. Eng., Springer-Verlag, Cham, 2021, pp. 605–613, https://doi.org/10.1007/978-3-030-55874-1_59.

[84] P. Knobloch, *An algebraically stabilized method for convection–diffusion–reaction problems with optimal experimental convergence rates on general meshes*, 2022, https://arxiv.org/abs/2208.07705.

[85] S. Korotov and T. Vejchodský, *A comparison of simplicial and block finite elements*, in Numerical mathematics and advanced applications 2009. Proceedings of ENUMATH 2009, G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva, eds., Springer-Verlag, Berlin, 2010, pp. 533–541, https://doi.org/10.1007/978-3-642-11795-4_57.

[86] 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, https://doi.org/10.1016/j.jcp.2006.03.034.

[87] D. Kuzmin, *Algebraic flux correction for finite element discretizations of coupled systems*, in

Proceedings of the Int. Conf. on Computational Methods for Coupled Problems in Science and Engineering, M. Papadrakakis, E. Oñate, and B. Schrefler, eds., CIMNE, Barcelona, 2007, pp. 1–5.

[88] D. Kuzmin, *On the design of algebraic flux correction schemes for quadratic finite elements*, J. Comput. Appl. Math., 218 (2008), pp. 79–87, https://doi.org/10.1016/j.cam.2007.04.045.

[89] D. Kuzmin, *Explicit and implicit FEM-FCT algorithms with flux linearization*, J. Comput. Phys., 228 (2009), pp. 2517–2534, https://doi.org/10.1016/j.jcp.2008.12.011.

[90] D. Kuzmin, *Algebraic flux correction I. Scalar conservation laws*, in Flux-corrected transport. Principles, algorithms, and applications, D. Kuzmin, R. Löhner, and S. Turek, eds., Springer, Dordrecht, second ed., 2012, pp. 145–192, https://doi.org/10.1007/978-94-007-4038-9_6.

[91] D. Kuzmin, *Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes*, J. Comput. Appl. Math., 236 (2012), pp. 2317–2337, https://doi.org/10.1016/j.cam.2011.11.019.

[92] D. Kuzmin, *Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws*, Comput. Methods Appl. Mech. Engrg., 361 (2020), p. Paper No. 112804, https://doi.org/10.1016/j.cma.2019.112804.

[93] D. Kuzmin, S. Basting, and J. N. Shadid, *Linearity-preserving monotone local projection stabilization schemes for continuous finite elements*, Comput. Methods Appl. Mech. Engrg., 322 (2017), pp. 23–41, https://doi.org/10.1016/j.cma.2017.04.030.

[94] D. Kuzmin, H. Hajduk, and A. Rupp, *Locally bound-preserving enriched Galerkin methods for the linear advection equation*, Comput. & Fluids, 205 (2020), p. Paper No. 104525, https://doi.org/10.1016/j.compfluid.2020.104525.

[95] D. Kuzmin and J. Hämäläinen, *Finite element methods for computational fluid dynamics: A practical guide*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2015, https://doi.org/10.1137/1.9781611973617.

[96] D. Kuzmin and J. N. Shadid, *Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations*, Internat. J. Numer. Methods Fluids, 84 (2017), pp. 675–695, https://doi.org/10.1002/fld.4365.

[97] D. Kuzmin and S. Turek, *Flux correction tools for finite elements*, J. Comput. Phys., 175 (2002), pp. 525–558, https://doi.org/10.1006/jcph.2001.6955.

[98] D. Kuzmin and S. Turek, *High-resolution FEM-TVD schemes based on a fully multidimensional flux limiter*, J. Comput. Phys., 198 (2004), pp. 131–158, https://doi.org/10.1016/j.jcp.2004.01.015.

[99] D. Leykekhman and B. Li, *Weak discrete maximum principle of finite element methods in convex polyhedra*, Math. Comp., 90 (2021), pp. 1–18, https://doi.org/10.1090/mcom/3560.

[100] H. Li and X. Zhang, *On the monotonicity and discrete maximum principle of the finite difference implementation of C^0-Q^2 finite element method*, Numer. Math., 145 (2020), pp. 437–472, https://doi.org/10.1007/s00211-020-01110-6.

[101] X. Li and W. Huang, *An anisotropic mesh adaptation method for the finite element solution of heterogeneous anisotropic diffusion problems*, J. Comput. Phys., 229 (2010), pp. 8072–8094, https://doi.org/10.1016/j.jcp.2010.07.009.

[102] X. Li and W. Huang, *Maximum principle for the finite element solution of time-dependent anisotropic diffusion problems*, Numer. Methods Partial Differential Equations, 29 (2013), pp. 1963–1985, https://doi.org/10.1002/num.21784.

[103] C. Lohmann, *Physics-compatible finite element methods for scalar and tensorial advection problems*, Springer Spektrum, Wiesbaden, 2019, https://doi.org/10.1007/978-3-658-27737-6.

[104] C. Lohmann, D. Kuzmin, J. N. Shadid, and S. Mabuza, *Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements*, J. Comput. Phys., 344 (2017), pp. 151–186, https://doi.org/10.1016/j.jcp.2017.04.059.

[105] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati, *Finite element flux-corrected transport (FEM-FCT) for the Euler and Navier-Stokes equations.*, Int. J. Numer. Methods Fluids, 7 (1987), pp. 1093–1109, https://doi.org/10.1002/fld.1650071007.

[106] J. Lorenz, *Zur Inversmonotonie diskreter Probleme*, Numer. Math., 27 (1976/77), pp. 227–238, https://doi.org/10.1007/BF01396643.

[107] C. Lu, W. Huang, and J. Qiu, *Maximum principle in linear finite element approximations of anisotropic diffusion-convection-reaction problems*, Numer. Math., 127 (2014), pp. 515–537, https://doi.org/10.1007/s00211-013-0595-8.

[108] S. Mabuza, J. N. Shadid, and D. Kuzmin, *Local bounds preserving stabilization for continuous Galerkin discretization of hyperbolic systems*, J. Comput. Phys., 361 (2018), pp. 82–

110, https://doi.org/10.1016/j.jcp.2018.01.048.
[109] A. Mizukami and T. J. R. Hughes, *A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle*, Comput. Methods Appl. Mech. Engrg., 50 (1985), pp. 181–193, https://doi.org/10.1016/0045-7825(85)90089-1.
[110] K. Ohmori and T. Ushijima, *A technique of upstream type applied to a linear nonconforming finite element approximation of convective diffusion equations*, RAIRO Anal. Numér., 18 (1984), pp. 309–332, https://doi.org/10.1051/m2an/1984180303091.
[111] A. K. Parrott and M. A. Christie, *FCT applied to the 2-D finite element solution of tracer transport by single phase flow in a porous medium*, in Numerical methods for fluid dynamics II, Proc. Conf., Reading/UK 1985, vol. 7 of Inst. Math. Appl. Conf. Ser., New Ser., 1986, pp. 609–619.
[112] W. Pazner, *Sparse invariant domain preserving discontinuous Galerkin methods with subcell convex limiting*, Comput. Methods Appl. Mech. Engrg., 382 (2021), p. Paper No. 113876, https://doi.org/10.1016/j.cma.2021.113876.
[113] R. Rannacher and S. Turek, *Simple nonconforming quadrilateral Stokes element*, Numer. Methods Partial Differential Equations, 8 (1992), pp. 97–111, https://doi.org/10.1002/num.1690080202.
[114] W. Reed and T. Hill, *Triangular mesh methods for the neutron transport equation*, Technical Report LA-UR-73-479, Los Alamos Scientific Laboratory, Los Alamos, NM, 1973.
[115] B. Rivière, *Discontinuous Galerkin methods for solving elliptic and parabolic equations. Theory and implementation*, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008, https://doi.org/10.1137/1.9780898717440.
[116] H.-G. Roos, M. Stynes, and L. Tobiska, *Robust numerical methods for singularly perturbed differential equations. Convection-diffusion-reaction and flow problems*, Springer-Verlag, Berlin, second ed., 2008, https://doi.org/10.1007/978-3-540-34467-4.
[117] A. H. Schatz, *A weak discrete maximum principle and stability of the finite element method in L_∞ on plane polygonal domains. I*, Math. Comp., 34 (1980), pp. 77–91, https://doi.org/10.2307/2006221.
[118] A. H. Schatz, V. Thomée, and L. B. Wahlbin, *On positivity and maximum-norm contractivity in time stepping methods for parabolic equations*, Comput. Methods Appl. Math., 10 (2010), pp. 421–443, https://doi.org/10.2478/cmam-2010-0025.
[119] C.-W. Shu, *Discontinuous Galerkin methods for time-dependent convection dominated problems: basics, recent developments and comparison with other methods*, in Building bridges: connections and challenges in modern approaches to numerical partial differential equations, G. R. Barrenechea, F. Brezzi, A. Cangiani, and E. H. Georgoulis, eds., vol. 114 of Lect. Notes Comput. Sci. Eng., Springer, Cham, 2016, pp. 369–397, https://doi.org/10.1007/978-3-319-41640-3_12.
[120] J. M. Stockie, *The mathematics of atmospheric dispersion modeling*, SIAM Rev., 53 (2011), pp. 349–372, https://doi.org/10.1137/10080991X.
[121] G. Strang and G. J. Fix, *An analysis of the finite element method*, Prentice-Hall, Inc., Englewood Cliffs, N. J., 1973.
[122] M. Tabata, *A finite element approximation corresponding to the upwind finite differencing*, Mem. Numer. Math., 4 (1977), pp. 47–63.
[123] V. Thomée and L. B. Wahlbin, *On the existence of maximum principles in parabolic finite element equations*, Math. Comp., 77 (2008), pp. 11–19, https://doi.org/10.1090/S0025-5718-07-02021-2.
[124] C. Tsiotsios and M. Petrou, *On the choice of the parameters for anisotropic diffusion in image processing*, Pattern Recognition, 46 (2013), pp. 1369–1381, https://doi.org/https://doi.org/10.1016/j.patcog.2012.11.012.
[125] S. Turek, *Tools for simulating nonstationary incompressible flow via discretely divergence-free finite element models*, Internat. J. Numer. Methods Fluids, 18 (1994), pp. 71–105, https://doi.org/10.1002/fld.1650180105.
[126] R. S. Varga, *Matrix iterative analysis*, Springer-Verlag, Berlin, 2000, https://doi.org/10.1007/978-3-642-05156-2.
[127] T. Vejchodský, *Angle conditions for discrete maximum principles in higher-order FEM*, in Numerical mathematics and advanced applications 2009. Proceedings of ENUMATH 2009, G. Kreiss, P. Lötstedt, A. Målqvist, and M. Neytcheva, eds., Springer-Verlag, Berlin, 2010, pp. 901–909, https://doi.org/10.1007/978-3-642-11795-4_97.
[128] T. Vejchodský, *Discrete Maximum Principles*, habilitation, Charles University Prague, Faculty of Mathematics and Physics, 2011.
[129] T. Vejchodský and P. Šolín, *Discrete maximum principle for a 1D problem with piecewise-*

constant coefficients solved by hp-FEM, J. Numer. Math., 15 (2007), pp. 233–243, https://doi.org/10.1515/jnma.2007.011.

[130] T. Vejchodský and P. Šolín, *Discrete maximum principle for higher-order finite elements in 1D*, Math. Comp., 76 (2007), pp. 1833–1846, https://doi.org/10.1090/S0025-5718-07-02022-4.

[131] F. J. Vermolen and A. Segal, *On an integration rule for products of barycentric coordinates over simplexes in* $\mathbb{R}^n$, J. Comput. Appl. Math., 330 (2018), pp. 289–294, https://doi.org/10.1016/j.cam.2017.09.013.

[132] J. Warren, S. Schaefer, A. N. Hirani, and M. Desbrun, *Barycentric coordinates for convex sets*, Adv.Comput. Math., 39 (2007), pp. 319–338, https://doi.org/10.1007/s10444-005-9008-6.

[133] P. Wesseling, *Principles of computational fluid dynamics*, Springer-Verlag, Berlin, 2001, https://doi.org/10.1007/978-3-642-05146-3.

[134] U. Wilbrandt, C. Bartsch, N. Ahmed, N. Alia, F. Anker, L. Blank, A. Caiazzo, S. Ganesan, S. Giere, G. Matthies, R. Meesala, A. Shamim, J. Venkatesan, and V. John, *ParMooN—A modernized program package based on mapped finite elements*, Comput. Math. Appl., 74 (2017), pp. 74–88, https://doi.org/10.1016/j.camwa.2016.12.020.

[135] J. Xu and L. Zikatanov, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp., 68 (1999), pp. 1429–1446, https://doi.org/10.1090/S0025-5718-99-01148-5.

[136] S. T. Zalesak, *Fully multidimensional flux-corrected transport algorithms for fluids*, J. Comput. Phys., 31 (1979), pp. 335–362, https://doi.org/10.1016/0021-9991(79)90051-2.

[137] S. T. Zalesak, *The design of flux-corrected transport (FCT) algorithms for structured grids*, in Flux-corrected transport. Principles, algorithms, and applications, D. Kuzmin, R. Löhner, and S. Turek, eds., Springer, Dordrecht, second ed., 2012, pp. 23–65, https://doi.org/10.1007/978-94-007-4038-9_2.

[138] X. Zhang and C.-W. Shu, *On maximum-principle-satisfying high order schemes for scalar conservation laws*, J. Comput. Phys., 229 (2010), pp. 3091–3120, https://doi.org/10.1016/j.jcp.2009.12.030.

[139] X. Zhang and C.-W. Shu, *Maximum-principle-satisfying and positivity-preserving high-order schemes for conservation laws: Survey and new developments*, Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci., 467 (2011), pp. 2752–2776, https://doi.org/10.1098/rspa.2011.0153.

[140] Y. Zhang, X. Zhang, and C.-W. Shu, *Maximum-principle-satisfying second order discontinuous Galerkin schemes for convection-diffusion equations on triangular meshes*, J. Comput. Phys., 234 (2013), pp. 295–316, https://doi.org/10.1016/j.jcp.2012.09.032.