No. 584 **April 2018**

Algebraic flux correction schemes preserving the eigenvalue range of symmetric tensor fields

C. Lohmann

ISSN: 2190-1767

Mathematical Modelling and Numerical Analysis Will be set by the publisher
Modélisation Mathématique et Analyse Numérique

ALGEBRAIC FLUX CORRECTION SCHEMES PRESERVING THE EIGENVALUE RANGE OF SYMMETRIC TENSOR FIELDS

Christoph Lohmann[1]

Abstract. This work extends the algebraic flux correction (AFC) paradigm to finite element discretizations of conservation laws for symmetric tensor fields. The proposed algorithms are designed to enforce discrete maximum principles and preserve the eigenvalue range of evolving tensors. To that end, a continuous Galerkin approximation is modified by adding a linear artificial diffusion operator and a nonlinear antidiffusive correction. The latter is decomposed into edge-based fluxes and constrained to prevent violations of local bounds for the minimal and maximal eigenvalues. In contrast to the flux-corrected transport (FCT) algorithm developed previously by the author and existing slope limiting techniques for stress tensors , the admissible eigenvalue range is defined implicitly and the limited antidiffusive terms are incorporated into the residual of the nonlinear system. In addition to scalar limiters that use a common correction factor for all components of a tensor-valued antidiffusive flux, tensor limiters are designed using spectral decompositions. The new limiter functions are analyzed using tensorial extensions of the existing AFC theory for scalar convection-diffusion equations. The proposed methodology is backed by rigorous proofs of eigenvalue range preservation and Lipschitz continuity. Convergence of pseudo time-stepping methods to stationary solutions is demonstrated in numerical studies.

1991 Mathematics Subject Classification. 65N12, 65N30.

The dates will be set by the publisher.

1. Introduction

The need to impose physically motivated bounds on the eigenvalues or principal invariants of symmetric tensor quantities arises in many applications of numerical solution methods to real life problems. A variety of nonlinear high resolution schemes based on the use of flux or slope limiting techniques have been developed for scalar conserved quantities and hyperbolic systems but the design and analysis of bounds preserving limiters for symmetric tensor quantities poses additional challenges which we address in this work. The proposed methodology extends the algebraic flux correction (AFC) paradigm and the underlying theory to tensor fields.

In the context of scalar conservation laws, algebraic flux correction (AFC) constrains the coefficients of a finite element approximation so as to satisfy (local) maximum principles while preserving the total mass. The AFC methodology traces its origins to the flux corrected transport (FCT) algorithm which applies antidiffusive fluxes to a low order solution (predictor) and limits them in a way which guarantees preservation of local bounds for the final solution (corrector). In contrast to FCT algorithms of this kind, the AFC methodology proposed

Keywords and phrases: advection of tensor fields, continuous Galerkin method, algebraic flux correction, artificial diffusion, discrete maximum principles

[1] Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany (e-mail: christoph.lohmann@math.tu-dortmund.de)

by Kuzmin [16] inserts limited antidiffusive fluxes into the residual of the nonlinear system which blends a high order target discretization and its monotone low order counterpart so that the resulting scheme becomes local extremum diminishing (LED). Algebraic limiting techniques of this kind were proposed and analyzed, e.g., in [4, 5, 6, 7, 18]. In recent contributions to the field, the principle of linearity preservation was recognized to be an important prerequisite for achieving the optimal order of convergence for sufficiently smooth data on general meshes [6, 18]. Moreover, differentiable versions of nodal limiters were developed to facilitate iterative solution of nonlinear systems [4].

The development of meaningful limiting techniques for tensor quantities is not straightforward and has become an important issue in numerous research areas: image analysis (structure tensors), diffusion tensor magnetic resonance imaging (DT-MRI; diffusion tensors), fluid and solid dynamics (Cauchy stress tensors), civil engineering and solid mechanics (inertia, diffusion and permittivity tensors), fiber suspensions (orientation tensors) – to name just a few [2, 3, 10, 14, 25]. While for scalar fields it is almost clarified which properties are important for calculating physics-compatible numerical solutions (e.g., local maximum principles and conservation of mass), this issue is ambiguous for (symmetric) tensor quantities and the best choice of a limiting strategy may depend on the underlying problem. In the context of bounds preserving reconstruction (remapping), tailor-made slope limiters for stress tensors were proposed in the recent work of Shashkov et al. [14, 27]. Their limiter prevents oscillations of the second invariant which represents a conserved quantity proportional to the elastic energy density. The general limiting framework proposed by Luttwak et al. [22, 23, 24] constrains tensors or vectors to lie in the convex hull of tensors/vectors that are known to be physically admissible. Other general-purpose limiting approaches proposed to date include the generalized Osher-Sethian scheme by Burgeth et al. [10] and a limiter constraining the tensor components along flow-related directions [25].

The present paper is based on the idea introduced by the author in the context of an eigenvalue range preserving FCT algorithm [20] designed to satisfy (local) maximum principles and preserve definiteness when it comes to antidiffusive corrections of the underlying low order approximation. This FCT limiter is frame invariant and preserves a constant trace, which makes it a useful tool for numerical treatment of orientation tensors in fiber suspension flow models. However, predictor-corrector methods of FCT type are not well suited for solving stationary transport problems and theoretical analysis. Therefore, we develop and analyze alternative limiting approaches in this paper building on recent advances in the field of AFC schemes for scalar convection-diffusion equations.

The article is structured as follows: In the next section, we introduce some basic notation, summarize important properties of (symmetric) tensor quantities and explain the main idea behind the AFC methodology for scalar fields. In Section 3, a first basic limiter for symmetric tensor quantities is developed to enforce sufficient conditions of eigenvalue range preservation using a scalar correction factor which scales each component of the antidiffusive flux in the same manner and exploits a special form of the system matrix. Theoretical proofs of its Lipschitz continuity and local maximum principles justify the definition. This limiter is generalized to arbitrary system matrices in Section 4. However, it turns out that the resulting AFC method can be quite diffusive due to the use of a scalar-valued correction factor. This drawback is cured in Secs. 5 and 6 by introducing tensorial correction factors based on spectral decompositions. In the last theoretical section (Sec. 7) of this paper, the concept of dual limiting is introduced for antidiffusive fluxes that may violate eigenvalue-based maximum principles at both nodes. The segregated treatment of such fluxes improves the accuracy of the AFC method at the expense of just a minor increase in computational costs. Finally, the proposed algorithms are compared to each other and the implications of varying the fine-tuning parameters are explored in applications to newly defined test problems (Sec. 8).

2. Notation and preliminaries

2.1. Notation

In this work, we are interested in monotonicity preserving algorithms for symmetric tensors with the dimension $d \times d$, $1 \leqslant d \leqslant 3$. This space is denoted by $\mathbb{S}_d \subset \mathbb{R}^{d\times d}$ while $\mathbb{S}_{d,+} \subset \mathbb{S}_d$ stands for the subset of positive

semidefinite tensors, where $\mathbf{V} \in \mathbb{S}_{d,+}$ is equivalent to $\mathbf{V} \succcurlyeq \mathbf{0}$ and $\mathbf{0} \preccurlyeq \mathbf{V}$. Here, the tensor inequality $\mathbf{V} \succcurlyeq \mathbf{U}$ should be understood in the sense that all eigenvalues of $\mathbf{V} - \mathbf{U}$ are nonnegative (cf. Löwner ordering [21]), i.e., $\lambda_k(\mathbf{V} - \mathbf{U}) \geqslant 0$, $1 \leqslant k \leqslant d$, while $\mathbf{0} \in \mathbb{S}_d$ and $\mathbf{I} \in \mathbb{S}_d$ denote the zero and identity tensor, respectively. Furthermore, we adopt the notation that tensor quantities in the $d \times d$-dimensional space are written in capital and boldface letters, while vectors $\mathbf{v} \in \mathbb{R}^d$ are labeled in lowercase and boldface letters. If we introduce a vector $\mathbf{v}_k$ corresponding to a tensor quantity $\mathbf{V}$, $\mathbf{v}_k$ corresponds to the k-th column vector of $\mathbf{V}$. The boldface notation should distinguish tensors and vectors in the $d \times d$- and d-dimensional space from matrices and vectors corresponding to the finite element space with dimension N.

A tensor quantity furnished with a tilde like $\tilde{\mathbf{V}}$ denotes the diagonal tensor whose diagonal entries are the sorted eigenvalues $v_k \coloneqq \lambda_k(\mathbf{V})$ of $\mathbf{V}$, i.e.,

$$v_1 \leqslant \ldots \leqslant v_d \quad \text{and} \quad \tilde{\mathbf{V}} \coloneqq \operatorname{diag}(v_1, \ldots, v_d).$$

Then, we have

$$\mathbf{V} = \mathbf{Q}\tilde{\mathbf{V}}\mathbf{Q}^\top,$$

where $\mathbf{Q} \in \mathbb{R}^{d \times d}$ is the orthogonal tensor having the normalized eigenvector $\mathbf{q}_k$ corresponding to the eigenvalue v_k as its k-th column vector, i.e., $\mathbf{V}\mathbf{q}_k = v_k\mathbf{q}_k$. In addition to the notation with one subscript, lowercase letters with two subscripts $v_{k\ell}$ denote the tensor entries of $\mathbf{V}$, i.e., $\mathbf{V} = (v_{k\ell})_{k,\ell=1}^d$. If not mentioned otherwise, k and ℓ are reserved for exclusive use in sums over the space dimension, i.e., sums over k and ℓ are restricted to the range $1 \leqslant k, \ell \leqslant d$.

2.2. **Basic results for tensor norms**

To prove the existence of a solution and the Lipschitz continuity of the proposed flux limiters, we will make use of the spectral and Frobenius norm $\|\cdot\|_2$ and $\|\cdot\|_\mathrm{F}$, respectively. Therefore, we summarize some of the most important properties.

The spectral norm $\|\cdot\|_2$ is the induced matrix norm of the Euclidean vector norm and equal to the largest absolute eigenvalue for symmetric tensors $\mathbf{V} \in \mathbb{S}_d$, i.e.,

$$\|\mathbf{V}\mathbf{x}\|_2 \leqslant \|\mathbf{V}\|_2\|\mathbf{x}\|_2, \qquad \text{for all } \mathbf{V} \in \mathbb{R}^{d\times d} \text{ and } \mathbf{x} \in \mathbb{R}^d, \qquad \|\mathbf{V}\|_2 = \max\big(|v_1|, |v_d|\big) \qquad \text{for all } \mathbf{V} \in \mathbb{S}_d.$$

In contrast to that, the Frobenius norm $\|\cdot\|_\mathrm{F}$ is not an induced norm. According to the invariance of the trace $\operatorname{tr}(\cdot)$ under cyclic permutations

$$\operatorname{tr}(\mathbf{V}\mathbf{W}) = \operatorname{tr}(\mathbf{W}\mathbf{V}) \qquad \text{for all } \mathbf{V}, \mathbf{W} \in \mathbb{R}^{d\times d}$$

and the definition of the Frobenius inner product $(\cdot,\cdot)_\mathrm{F}$

$$(\mathbf{V}, \mathbf{W})_\mathrm{F} \coloneqq \mathbf{V} : \mathbf{W} \coloneqq \sum_{k,\ell=1}^{d} v_{k\ell}w_{k\ell} = \operatorname{tr}(\mathbf{V}^\top\mathbf{W}) = \operatorname{tr}(\mathbf{W}^\top\mathbf{V}),$$

the Frobenius norm $\|\cdot\|_\mathrm{F}$ satisfies the identity

$$\|\mathbf{V}\|_\mathrm{F}^2 \coloneqq (\mathbf{V}, \mathbf{V})_\mathrm{F} = \operatorname{tr}(\mathbf{V}^2) = \operatorname{tr}(\mathbf{Q}\tilde{\mathbf{V}}^2\mathbf{Q}^\top) = \operatorname{tr}(\tilde{\mathbf{V}}^2\mathbf{Q}^\top\mathbf{Q}) = \operatorname{tr}(\tilde{\mathbf{V}}^2) = \sum_{k=1}^{d} v_k^2 \qquad \text{for all } \mathbf{V} \in \mathbb{S}_d. \tag{1}$$

Both norms can be calculated just by considering the eigenvalues. Therefore, they are frame invariant, i.e.,

$$\|\mathbf{Q}\mathbf{V}\mathbf{Q}^\top\|_2 = \|\mathbf{V}\|_2, \qquad \|\mathbf{Q}\mathbf{V}\mathbf{Q}^\top\|_\mathrm{F} = \|\mathbf{V}\|_\mathrm{F} \qquad \text{for all } \mathbf{V}, \mathbf{Q} \in \mathbb{R}^{d\times d} \text{ with } \mathbf{Q}\mathbf{Q}^\top = \mathbf{I}$$

and equivalent to each other

$$\|\mathbf{V}\|_2^2 = \lambda_d(\mathbf{V}^\top\mathbf{V}) \leqslant \sum_{k=1}^{d} \lambda_k(\mathbf{V}^\top\mathbf{V}) = \|\mathbf{V}\|_{\mathrm{F}}^2 \leqslant d\|\mathbf{V}\|_2^2$$
$$\implies \qquad \|\mathbf{V}\|_2 \leqslant \|\mathbf{V}\|_{\mathrm{F}} \leqslant \sqrt{d}\|\mathbf{V}\|_2 \qquad \text{for all } \mathbf{V} \in \mathbb{R}^{d\times d}.$$

Furthermore, the following inequalities hold

$$\begin{aligned} \|\mathbf{V}\mathbf{W}\|_{\mathrm{F}}^2 &= \sum_{k=1}^{d} \|\mathbf{V}\mathbf{w}_k\|_2^2 \leqslant \|\mathbf{V}\|_2^2 \sum_{k=1}^{d} \|\mathbf{w}_k\|_2^2 = \|\mathbf{V}\|_2^2\|\mathbf{W}\|_{\mathrm{F}}^2, \\ \|\mathbf{W}\mathbf{V}\|_{\mathrm{F}}^2 &= \|\mathbf{V}^\top\mathbf{W}^\top\|_{\mathrm{F}}^2 \leqslant \|\mathbf{V}^\top\|_2^2\|\mathbf{W}^\top\|_{\mathrm{F}}^2 = \|\mathbf{V}\|_2^2\|\mathbf{W}\|_{\mathrm{F}}^2 \end{aligned} \qquad \text{for all } \mathbf{V}, \mathbf{W} \in \mathbb{R}^{d\times d}, \tag{2}$$

where we used the notation that $\mathbf{w}_k \in \mathbb{R}^d$ is the k-th column vector of $\mathbf{W} \in \mathbb{R}^{d\times d}$.

2.3. **Lipschitz continuity of selected scalar and tensorial operations**

In what follows, we prove the Lipschitz continuity of the absolute value, maximum, and minimum. For scalars $u, v, \bar{u}, \bar{v} \in \mathbb{R}$, we have due to the reverse triangle inequality

$$\big||v| - |\bar{v}|\big| \leqslant |v - \bar{v}|, \tag{3a}$$

$$\begin{aligned} \big|\max(u,v) - \max(\bar{u},\bar{v})\big| &= \tfrac{1}{2}\big|(u+v) + |u-v| - (\bar{u}+\bar{v}) - |\bar{u}-\bar{v}|\big| \\ &\leqslant \tfrac{1}{2}|u+v-\bar{u}-\bar{v}| + \tfrac{1}{2}\big||u-v| - |\bar{u}-\bar{v}|\big| \leqslant |u-\bar{u}| + |v-\bar{v}|, \end{aligned} \tag{3b}$$

$$\begin{aligned} \big|\min(u,v) - \min(\bar{u},\bar{v})\big| &= \tfrac{1}{2}\big|(u+v) - |u-v| - (\bar{u}+\bar{v}) + |\bar{u}-\bar{v}|\big| \\ &\leqslant \tfrac{1}{2}|u+v-\bar{u}-\bar{v}| + \tfrac{1}{2}\big||u-v| - |\bar{u}-\bar{v}|\big| \leqslant |u-\bar{u}| + |v-\bar{v}|. \end{aligned} \tag{3c}$$

In the case of tensor quantities, the operations are defined as follows for all $\mathbf{U}, \mathbf{V} \in \mathbb{S}_d$

$$|\mathbf{V}| \coloneqq \mathbf{Q}|\tilde{\mathbf{V}}|\mathbf{Q}^\top, \qquad \max(\mathbf{U},\mathbf{V}) \coloneqq \tfrac{1}{2}(\mathbf{U}+\mathbf{V}) + \tfrac{1}{2}|\mathbf{U}-\mathbf{V}|, \qquad \min(\mathbf{U},\mathbf{V}) \coloneqq \tfrac{1}{2}(\mathbf{U}+\mathbf{V}) - \tfrac{1}{2}|\mathbf{U}-\mathbf{V}|, \tag{4}$$

where $|\tilde{\mathbf{V}}|$ is defined as the tensor with the absolute values of the diagonal tensor $\tilde{\mathbf{V}}$ and $\mathbf{Q} \in \mathbb{R}^{d\times d}$ is the orthogonal tensor such that $\mathbf{V} = \mathbf{Q}\tilde{\mathbf{V}}\mathbf{Q}^\top$. In other words, $|\mathbf{V}|$ is the tensor with the absolute eigenvalues of $\mathbf{V}$ corresponding to the same eigenvectors. As expected, the minimum of two tensors is bounded above in the sense of Löwner ordering by one of its arguments

$$\min(\mathbf{U},\mathbf{V}) = \tfrac{1}{2}\big((\mathbf{U}+\mathbf{V}) - \underbrace{|\mathbf{U}-\mathbf{V}|}_{\succcurlyeq \mathbf{V}-\mathbf{U}}\big) \preccurlyeq \tfrac{1}{2}\big((\mathbf{U}+\mathbf{V}) + (\mathbf{U}-\mathbf{V})\big) = \mathbf{U}. \tag{5}$$

However, if both arguments are bounded below, the minimum can violate this bound in contrast to the scalar definition of the minimum

$$\mathbf{U}, \mathbf{V} \succcurlyeq \mathbf{W} \qquad \not\Longrightarrow \qquad \min(\mathbf{U},\mathbf{V}) \succcurlyeq \mathbf{W}.$$

A counterexample in $\mathbb{R}^{2\times 2}$ is given by

$$\mathbf{U} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix} \succcurlyeq \mathbf{0}, \quad \mathbf{V} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} \succcurlyeq \mathbf{0} \qquad \implies \qquad \min(\mathbf{U},\mathbf{V}) = \tfrac{1}{2}\begin{pmatrix} 5 & 4 \\ 4 & 5 \end{pmatrix} - \tfrac{1}{2}\begin{pmatrix} 3 & 0 \\ 0 & 3 \end{pmatrix} = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \not\succcurlyeq \mathbf{0}.$$

The operations defined in (4) are Lipschitz continuous for symmetric tensors: We will prove this for the Frobenius norm $\|\cdot\|_{\mathrm{F}}$ with Lipschitz constant $L = 1$. Without loss of generality, let us assume that $\mathbf{V}$ is diagonal, i.e.,

$\mathbf{V} = \tilde{\mathbf{V}}$. Then, we have

$$\big\||\mathbf{V}| - |\bar{\mathbf{V}}|\big\|_{\mathrm{F}}^2 = \|\mathbf{V}\|_{\mathrm{F}}^2 + \|\bar{\mathbf{V}}\|_{\mathrm{F}}^2 - 2\big(|\mathbf{V}|, |\bar{\mathbf{V}}|\big)_{\mathrm{F}} \leqslant \|\mathbf{V}\|_{\mathrm{F}}^2 + \|\bar{\mathbf{V}}\|_{\mathrm{F}}^2 - 2(\mathbf{V}, \bar{\mathbf{V}})_{\mathrm{F}} = \|\mathbf{V} - \bar{\mathbf{V}}\|_{\mathrm{F}}^2, \tag{6}$$

where the inequality holds due to

$$(\mathbf{V}, \bar{\mathbf{V}})_{\mathrm{F}} = \sum_{k=1}^{d} v_{kk}\bar{v}_{kk} = \sum_{k=1}^{d} v_{kk} \sum_{\ell=1}^{d} \bar{q}_{k\ell}\bar{v}_{\ell}\bar{q}_{k\ell} \leqslant \sum_{k=1}^{d} |v_{kk}| \sum_{\ell=1}^{d} |\bar{v}_{\ell}|\bar{q}_{k\ell}^2 \leqslant \big(|\mathbf{V}|, |\bar{\mathbf{V}}|\big)_{\mathrm{F}},$$

because $\mathbf{V}$ is diagonal and $\bar{\mathbf{V}} = \bar{\mathbf{Q}}\tilde{\bar{\mathbf{V}}}\bar{\mathbf{Q}}^{\top}$. This allows us to prove the Lipschitz continuity of $\max(\cdot,\cdot)$ and $\min(\cdot,\cdot)$ in the same way as in the scalar case:

$$\begin{aligned}
&\big\|\max(\mathbf{U}, \mathbf{V}) - \max(\bar{\mathbf{U}}, \bar{\mathbf{V}})\big\|_{\mathrm{F}} \\
&\quad = \tfrac{1}{2}\big\|(\mathbf{U} + \mathbf{V}) + |\mathbf{U} - \mathbf{V}| - (\bar{\mathbf{U}} + \bar{\mathbf{V}}) - |\bar{\mathbf{U}} - \bar{\mathbf{V}}|\big\|_{\mathrm{F}} \\
&\quad \leqslant \tfrac{1}{2}\|\mathbf{U} + \mathbf{V} - \bar{\mathbf{U}} - \bar{\mathbf{V}}\|_{\mathrm{F}} + \tfrac{1}{2}\big\||\mathbf{U} - \mathbf{V}| - |\bar{\mathbf{U}} - \bar{\mathbf{V}}|\big\|_{\mathrm{F}} \\
&\quad \leqslant \tfrac{1}{2}\|\mathbf{U} + \mathbf{V} - \bar{\mathbf{U}} - \bar{\mathbf{V}}\|_{\mathrm{F}} + \tfrac{1}{2}\|\mathbf{U} - \mathbf{V} - \bar{\mathbf{U}} + \bar{\mathbf{V}}\|_{\mathrm{F}} \leqslant \|\mathbf{U} - \bar{\mathbf{U}}\|_{\mathrm{F}} + \|\mathbf{V} - \bar{\mathbf{V}}\|_{\mathrm{F}},
\end{aligned} \tag{7a}$$

$$\begin{aligned}
&\big\|\min(\mathbf{U}, \mathbf{V}) - \min(\bar{\mathbf{U}}, \bar{\mathbf{V}})\big\|_{\mathrm{F}} \\
&\quad = \tfrac{1}{2}\big\|(\mathbf{U} + \mathbf{V}) - |\mathbf{U} - \mathbf{V}| - (\bar{\mathbf{U}} + \bar{\mathbf{V}}) + |\bar{\mathbf{U}} - \bar{\mathbf{V}}|\big\|_{\mathrm{F}} \\
&\quad \leqslant \tfrac{1}{2}\|\mathbf{U} + \mathbf{V} - \bar{\mathbf{U}} - \bar{\mathbf{V}}\|_{\mathrm{F}} + \tfrac{1}{2}\big\||\mathbf{U} - \mathbf{V}| - |\bar{\mathbf{U}} - \bar{\mathbf{V}}|\big\|_{\mathrm{F}} \\
&\quad \leqslant \tfrac{1}{2}\|\mathbf{U} + \mathbf{V} - \bar{\mathbf{U}} - \bar{\mathbf{V}}\|_{\mathrm{F}} + \tfrac{1}{2}\|\mathbf{U} - \mathbf{V} - \bar{\mathbf{U}} + \bar{\mathbf{V}}\|_{\mathrm{F}} \leqslant \|\mathbf{U} - \bar{\mathbf{U}}\|_{\mathrm{F}} + \|\mathbf{V} - \bar{\mathbf{V}}\|_{\mathrm{F}}.
\end{aligned} \tag{7b}$$

All three inequalities are sharp.

2.4. **Idea of algebraic flux correction in the scalar case**

In this section we briefly describe the basic concept of algebraic flux correction methods. The limiting techniques to be considered are based on the one developed in [16] to stabilize a numerical method corresponding to a scalar boundary value problem. It is closely related to the flux-corrected transport (FCT) methodology originally introduced in [9, 32]. Both methods have the capability to continuously blend a high order approximation of the discretized problem, typically given by the Galerkin method, and its low order counterpart, which is defined by adding artificial diffusion such that local maximum principles are satisfied algebraically. Because the high order method is not able to preserve (local) maximum principles, the algorithm should fall back to the low order method at local extrema, while being as close as possible to the high order solution everywhere else. This guarantees the boundedness of the function values by the extrema of the neighboring degrees of freedom and, hence, respects the local extremum diminishing (LED) property.

FCT algorithms utilize this idea in an predictor-corrector like manner: After calculating the low order solution, antidiffusive fluxes are added, which are limited to recover a (linearized) high order method in smooth regions without producing overshoots and undershoots elsewhere [17]. In contrast to this, the AFC method proposed by Kuzmin [16] leads to a nonlinear system, which blends both methods only depending on the solution itself. Due to this monolithic limiting strategy, the method can be used for stationary problems and time-dependent solutions are less sensitive to the choice of the time increment. Barrenechea et al. [5] were the first ones to investigate theoretical aspects of the nonlinear method and prove the existence of a solution.

To present the basic aspects of the AFC methodology, let us start with an arbitrary linear system of equations, which stems from the finite element discretization of a scalar boundary value problem and is given by

$$\sum_{j=1}^{N} a_{ij}u_j = g_i \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{8}$$

where $u = (u_i)_{i=1}^N, g = (g_i)_{i=1}^N$, $u_i, g_i \in \mathbb{R}$ for all $1 \leqslant i \leqslant N$, denote the degrees of freedom of the solution and right hand side, respectively, and $A = (a_{ij})_{i,j=1}^N$ is a positive definite matrix, i.e.,

$$C_M \coloneqq \inf_{v \in \mathbb{R}^N, \|v\|_2 = 1} \sum_{i,j=1}^N v_i a_{ij} v_j > 0. \tag{9}$$

The right hand side g_i of (8) depends on known data (boundary conditions, initial data, solution values from the previous time step). While the discrete problem (8) corresponds to the use of weakly imposed Dirichlet boundary conditions, strongly enforced function values also lead to a system with positive definite system matrix if the rows corresponding to boundary values are scaled in a reasonable manner. For the definition and detailed analysis of a stationary scalar problem, where strongly enforced Dirichlet boundary conditions are treated separately, we refer the reader to [5].

As presented in [5], if $a_{ij} \leqslant 0$ for all $j \neq i$, problem (8) satisfies the following discrete maximum principles

$$\sum_{j=1}^N a_{ij} = 0 \quad \text{and} \quad g_i \leqslant 0 \qquad \Longrightarrow \qquad u_i \leqslant \max_{j \neq i, a_{ij} \neq 0} u_j, \tag{10a}$$

$$\sum_{j=1}^N a_{ij} = 0 \quad \text{and} \quad g_i \geqslant 0 \qquad \Longrightarrow \qquad u_i \geqslant \min_{j \neq i, a_{ij} \neq 0} u_j, \tag{10b}$$

$$\sum_{j=1}^N a_{ij} \geqslant 0 \quad \text{and} \quad g_i \leqslant 0 \qquad \Longrightarrow \qquad u_i \leqslant \max\big(0, \max_{j \neq i, a_{ij} \neq 0} u_j\big), \tag{10c}$$

$$\sum_{j=1}^N a_{ij} \geqslant 0 \quad \text{and} \quad g_i \geqslant 0 \qquad \Longrightarrow \qquad u_i \geqslant \min\big(0, \min_{j \neq i, a_{ij} \neq 0} u_j\big). \tag{10d}$$

For example, (10c) can be shown as follows [5]

$$\begin{aligned} a_{ii} u_i \leqslant -\sum_{j \neq i} a_{ij} u_j &= \sum_{j \neq i, a_{ij} \neq 0} (-a_{ij}) \big(u_j - \max_{j \neq i, a_{ij} \neq 0} u_j\big) + \sum_{j \neq i, a_{ij} \neq 0} (-a_{ij}) \big(\max_{j \neq i, a_{ij} \neq 0} u_j\big) \\ &\leqslant \sum_{j \neq i, a_{ij} \neq 0} (-a_{ij}) \big(\max_{j \neq i, a_{ij} \neq 0} u_j\big) \leqslant a_{ii} \max\big(0, \max_{j \neq i, a_{ij} \neq 0} u_j\big). \end{aligned}$$

The proof for (10a) is the same, except that the last inequality is an identity

$$a_{ii} u_i \leqslant -\sum_{j \neq i} a_{ij} u_j \leqslant \sum_{j \neq i, a_{ij} \neq 0} (-a_{ij}) \big(\max_{j \neq i, a_{ij} \neq 0} u_j\big) = a_{ii} \max_{j \neq i, a_{ij} \neq 0} u_j.$$

These maximum principles are the discrete analogues of continuously defined maximum principles that, e.g., scalar convection equations satisfy [11, 26]. However, a general discretization of such a boundary value problem mostly does not satisfy the above sufficient conditions. For that reason, the basic idea is to introduce artificial diffusion, which corrects troublesome matrix entries a_{ij} in a consistent manner: If the artificial diffusion matrix $D = (d_{ij})_{i,j=1}^N \in \mathbb{R}^{N \times N}$ is defined by (similarly to the definition in [17] for $a_{ij} = -k_{ij}$)

$$d_{ij} \coloneqq \max\{a_{ij}, 0, a_{ji}\} \geqslant 0 \quad \text{for all } j \neq i, \qquad d_{ii} \coloneqq -\sum_{j \neq i} d_{ij} \leqslant 0, \tag{11}$$

then $B \coloneqq A - D$ satisfies the requirements of the sign for the discrete maximum principles. Furthermore, D has vanishing row and column sums, i.e.,

$$\sum_j d_{ij} = \sum_i d_{ij} = 0 \qquad \Longrightarrow \qquad \sum_j b_{ij} = \sum_j a_{ij}, \qquad \sum_i b_{ij} = \sum_i a_{ij}.$$

It follows that the total (and local) mass does not change (sum over i) and the properties of the original system corresponding to the discrete maximum principle are preserved (sum over j). Additionally, D is negative semidefinite and, hence, B is positive definite, due to the positive definiteness of A.

In the FCT methodology, matrix B is used to calculate a low order solution u^L, which satisfies the local extremum diminishing (LED) property, but is highly diffusive. Then, local bounds $u_i^{\min}$ and $u_i^{\max}$ depending on u^L and correction factors $0 \leqslant \alpha_{ij} = \alpha_{ji} \leqslant 1$ are defined so that the FCT solution is guaranteed to be located between $u_i^{\min}$ and $u_i^{\max}$, i.e.

$$u_i^{\min} \quad \leqslant \quad u_i = u_i^L + \frac{1}{m_i} \sum_j \alpha_{ij} f_{ij} \quad \leqslant \quad u_i^{\max}, \tag{12}$$

where $f_{ij} = -f_{ji}$ are antidiffusive fluxes. Here, m_i is the i-th diagonal entry of the lumped mass matrix and $m_i u_i$ is the mass corresponding to the degree of freedom i. The interested reader is referred to [17] for a more detailed description of this approach.

In the AFC framework defined by Kuzmin in [16], artificial diffusion and limited antidiffusive fluxes are incorporated into the nonlinear system

$$\sum_{j=1}^N b_{ij} u_j + \sum_{j \neq i} \alpha_{ij} d_{ij} u_j = \sum_{j=1}^N a_{ij} u_j + \sum_{j \neq i} (1 - \alpha_{ij}) d_{ij} (u_i - u_j) = g_i \qquad \text{for all } 1 \leqslant i \leqslant N,$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \ldots, u_N)$ depends on the solution and is responsible for the nonlinearity. As in the case of FCT, the correction factors blend the stabilized low order approximation ($\alpha_{ij} = 0$) and the high order target method, which is recovered by $\alpha_{ij} = 1$. If $\alpha_{ij} = 0$ whenever u_i is a local extremum, we can show that u_i is bounded above/below by the maximum/minimum of the neighboring function values and the LED property is satisfied [5].

To solve the nonlinear system, the Newton method can be used if α_{ij} is differentiable. Otherwise, pseudo time stepping approaches are commonly used to calculate the solution. In each pseudo time step, the solution of the nonlinear problem can be calculated using a fixed point iteration, which converges if there exists a unique solution, $\alpha_{ij}(u_i - u_j)$ is Lipschitz continuous, and the pseudo time increment is chosen small enough [1, Proposition 4.3]. Barrenechea et al. have shown that the algorithm proposed in [16] yields a solution [5] and that the solution is unique [6] in the special case of a steady-state convection-diffusion-reaction equation with a solenoidal velocity field.

2.5. **Extension of algebraic flux correction to the tensor case**

In this paper, we extend the concept of algebraic flux correction proposed in [16] to tensor quantities and prove the existence of a solution while following the proof techniques of [5]. A general linear system of equations for a symmetric tensor field is given by

$$\sum_{j=1}^N a_{ij} \mathbf{U}_j = \mathbf{G}_i \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{13}$$

where $\mathbf{U}_i \in \mathbb{S}_d$ and $\mathbf{G}_i \in \mathbb{S}_d$. In (13), each tensor entry $(u_{j,k\ell})_{j=1}^N$ satisfies a linear system with the same system matrix $A = (a_{ij})_{i,j=1}^N$ and different right-hand sides $(g_{i,k\ell})_{i=1}^N$. Then, the AFC system reads

$$\sum_{j=1}^{N} a_{ij}\mathbf{U}_j + \sum_{j\neq i} d_{ij}(\mathcal{I} - \mathcal{A}_{ij})[\mathbf{U}_i - \mathbf{U}_j] = \mathbf{G}_i \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{14}$$

where the artificial diffusion coefficient d_{ij} is defined as in the scalar case and $\mathcal{I}, \mathcal{A}_{ij} : \mathbb{S}_d \to \mathbb{S}_d$ are the identity and limiting operators. The definition of d_{ij} as in the scalar case has already proved itself to be viable in the FCT framework [20], where additionally to the LED property for each tensor entry, discrete maximum principles for the range of eigenvalues and the trace are fulfilled. The limiting operator $\mathcal{A}_{ij}$ should ideally return $\mathbf{U}_i - \mathbf{U}_j$ to recover the high order approximation and produce $\mathbf{0}$ in the worst case.

In what follows, we introduce two different concepts of defining $\mathcal{A}_{ij}$, which are able to preserve the LED property for the range of eigenvalues. This means that the maximal eigenvalue $u_{i,d}$ of $\mathbf{U}_i$ is bounded above by the maximum of maximal eigenvalues in the neighborhood of node i and the minimal eigenvalue $u_{i,1}$ is bounded below in a similar vein.

3. One-node scalar limiting

To start with the introduction of a simple tensor limiter, we assume that the system matrix A satisfies

$$\min(a_{ij}, a_{ji}) \leqslant 0 \qquad \text{for all } 1 \leqslant i, j \leqslant N \text{ s.t. } i \neq j. \tag{15}$$

This allows us to define the limiter in an upwind based manner (similarly to [16]), because for each edge there exists at least one harmless node, where the satisfaction of maximum principles does not depend on the edge contribution (see the proof of the LED property below). In Sec. 4, we discuss a more general case, where restriction (15) does not have to be satisfied.

The simplest approach to define the limiting operator $\mathcal{A}_{ij}$ is to use a scalar correction factor, which multiplies each entry of $\mathbf{U}_i - \mathbf{U}_j$ by the same value, i.e.,

$$\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] \coloneqq \begin{cases} \alpha_{ij}(\mathbf{U}_i - \mathbf{U}_j) & : a_{ij} > 0, a_{ji} \leqslant 0, \\ \alpha_{ji}(\mathbf{U}_i - \mathbf{U}_j) & : a_{ij} \leqslant 0, a_{ji} > 0, \\ \mathbf{U}_i - \mathbf{U}_j & : a_{ij} \leqslant 0, a_{ji} \leqslant 0, \end{cases} \tag{16}$$

where $0 \leqslant \alpha_{ij} \leqslant 1$ is a scalar correction factor depending on the solution $U = (\mathbf{U}_i)_{i=1}^N$ (at the moment arbitrary) and responsible for the validity of maximum principles at node i. The restriction to symmetric limiting operators $\mathcal{A}_{ij} = \mathcal{A}_{ji}$ is required for conservation reasons. In this context, the notation $\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]$ means the application of $\mathcal{A}_{ij}$ to $\mathbf{U}_i - \mathbf{U}_j$, where the correction factors depend on U. Most often the dependency on U is neglected and we just write $\mathcal{A}_{ij}[\cdot]$ instead of $\mathcal{A}_{ij}(U)[\cdot]$.

In what follows, we prove the existence of a solution of (14) following the techniques used in [5]: First of all, using the definition (16) of $\mathcal{A}_{ij}$, for an arbitrary tensor $\mathbf{V} \in \mathbb{S}_d$, we have

$$\big(\mathbf{V}, (\mathcal{I} - \mathcal{A}_{ij})[\mathbf{V}]\big)_{\mathrm{F}} = (1 - \alpha_{ij})(\mathbf{V}, \mathbf{V})_{\mathrm{F}} \geqslant 0 \qquad \text{if } a_{ij} > 0, a_{ji} \leqslant 0.$$

In the cases $a_{ij} \leqslant 0, a_{ji} > 0$ and $a_{ij} \leqslant 0, a_{ji} \leqslant 0$, the correction factor α_{ij} has to be replaced by α_{ji} and 1, respectively, and the same property is satisfied. This implies the nonnegativity of

$$\big(\mathbf{U}, (\mathcal{I} - \mathcal{A}_{ij})[\mathbf{U} - \mathbf{V}]\big)_{\mathrm{F}} + \big(\mathbf{V}, (\mathcal{I} - \mathcal{A}_{ji})[\mathbf{V} - \mathbf{U}]\big)_{\mathrm{F}} = \big(\mathbf{U} - \mathbf{V}, (\mathcal{I} - \mathcal{A}_{ij})[\mathbf{U} - \mathbf{V}]\big)_{\mathrm{F}} \geqslant 0$$

and, hence,

$$\sum_{i,j=1}^{N} d_{ij}\big(\mathbf{V}_i, (\mathcal{I}-\mathcal{A}_{ij})[\mathbf{V}_i-\mathbf{V}_j]\big)_{\mathrm{F}} = \sum_{i,j=1,i<j}^{N} d_{ij}\big(\mathbf{V}_i, (\mathcal{I}-\mathcal{A}_{ij})[\mathbf{V}_i-\mathbf{V}_j]\big)_{\mathrm{F}} + d_{ij}\big(\mathbf{V}_j, (\mathcal{I}-\mathcal{A}_{ji})[\mathbf{V}_j-\mathbf{V}_i]\big)_{\mathrm{F}} \geqslant 0. \quad (17)$$

Furthermore, we define the scalar product $(\cdot,\cdot)_{2,\mathrm{F}}$ and corresponding norm $\|\cdot\|_{2,\mathrm{F}}$ of a tensorial solution by

$$(V,U)_{2,\mathrm{F}} \coloneqq \sum_{i=1}^{N} (\mathbf{V}_i, \mathbf{U}_i)_{\mathrm{F}}, \qquad \|V\|_{2,\mathrm{F}}^2 \coloneqq (V,V)_{2,\mathrm{F}} = \sum_{i=1}^{N} \|\mathbf{V}_i\|_{\mathrm{F}}^2 \qquad \text{for all } V,U \in \mathbb{S}_d^N$$

so that, using the definition of $C_M > 0$ (cf. (9)), the inequality

$$\sum_{i,j=1}^{N} a_{ij}(\mathbf{V}_i,\mathbf{V}_j)_{\mathrm{F}} = \sum_{k,\ell=1}^{d}\sum_{i,j=1}^{N} a_{ij} v_{i,k\ell} v_{j,k\ell} \geqslant \sum_{k,\ell=1}^{d} C_M \sum_{i=1}^{N} v_{i,k\ell}^2 = C_M \sum_{i=1}^{N} \|\mathbf{V}_i\|_{\mathrm{F}}^2 = C_M \|V\|_{2,\mathrm{F}}^2 \quad (18)$$

holds. If we now define the operator $\mathcal{T}_i : \mathbb{S}_d^N \to \mathbb{S}_d$ depending on $V = (\mathbf{V}_i)_{i=1}^N$ by

$$\mathcal{T}_i[V] = \sum_{j=1}^{N} a_{ij}\mathbf{V}_j + \sum_{j=1}^{N} d_{ij}(\mathcal{I}-\mathcal{A}_{ij})[\mathbf{V}_i-\mathbf{V}_j] - \mathbf{G}_i \qquad \text{for all } 1 \leqslant i \leqslant N,$$

then using (17), (18), and the Cauchy-Schwarz and Young's inequalities, we obtain (cf. [5])

$$\begin{aligned}\big((\mathcal{T}_i)_{i=1}^N[V], V\big)_{2,\mathrm{F}} &= \sum_{i,j=1}^{N} a_{ij}(\mathbf{V}_i,\mathbf{V}_j)_{\mathrm{F}} + \sum_{i,j=1}^{N} d_{ij}\big(\mathbf{V}_i, (\mathcal{I}-\mathcal{A}_{ij})[\mathbf{V}_i-\mathbf{V}_j]\big)_{\mathrm{F}} - \sum_{i=1}^{N}(\mathbf{G}_i,\mathbf{V}_i)_{\mathrm{F}} \\ &\geqslant C_M\|V\|_{2,\mathrm{F}}^2 - C_0 - C_1\|V\|_{2,\mathrm{F}} \geqslant \frac{C_M}{2}\|V\|_{2,\mathrm{F}}^2 - C_2,\end{aligned}$$

where the constants $C_0, C_1, C_2 > 0$ do not depend on the solution and we have

$$\big((\mathcal{T}_i)_{i=1}^N[\tilde{V}], \tilde{V}\big)_{2,\mathrm{F}} > 0 \qquad \text{for all } \tilde{V} \in \mathbb{S}_d^N \text{ s.t. } \|\tilde{V}\|_{2,\mathrm{F}}^2 = \frac{3C_2}{C_M}. \quad (19)$$

Thus, according to the following Lemma, there exists a solution $V \in \mathbb{S}_d^N$ such that (14) is satisfied if $\mathcal{A}_{ij}[\mathbf{V}_i-\mathbf{V}_j]$ is a continuous function of $\mathbf{V}_1,\dots,\mathbf{V}_N$ for all $1 \leqslant i \leqslant N$.

Lemma ([30, p. 164, Lemma 1.4])**.** *Let X be a finite-dimensional Hilbert space with inner product $(\cdot,\cdot)_X$ and norm $\|\cdot\|_X$. Let $T : X \to X$ be a continuous mapping and $K > 0$ a real number such that*

$$(Tx,x)_X > 0 \qquad \textit{for all } x \in X \textit{ with } \|x\|_X = K. \quad (20)$$

Then there exists $x \in X$ such that $\|x\|_X \leqslant K$ and $Tx = 0$.

Due to the nonlinearity of the considered system (14), the solution is not necessarily unique. However, Barrenechea et al. [6] have shown the uniqueness in the special case of a scalar steady-state convection-diffusion-reaction equation with a solenoidal velocity field. In general, only the linearized problem, where $\mathcal{A}_{ij}$ is explicitly given and independent of the solution V, possesses a unique solution.

Before focusing on a specific definition of the limiting operator and showing its continuity, let us consider a more general form:

If the application of $\mathcal{A}_{ij}$ can be expressed by (cf. [5])

$$\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j] = \alpha_{ij}(\mathbf{U}_i - \mathbf{U}_j), \quad 0 \leqslant \alpha_{ij} \coloneqq \frac{A_{ij}}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + B_{ij}} \leqslant 1 \quad \text{for all } U \in \mathbb{S}_d^N \quad \text{if } a_{ij} > 0, a_{ji} \leqslant 0, \tag{21}$$

where A_{ij} and B_{ij} are nonnegative and (Lipschitz) continuous functions depending on U (similarly in other cases), then the application of the limiter is (Lipschitz) continuous, too.

To show the continuity of $\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j]$ in $\bar{U} \in \mathbb{S}_d^N$ chosen arbitrarily, we only have to take a look at $\bar{U}$ with $\bar{\mathbf{U}}_i = \bar{\mathbf{U}}_j$, otherwise the denominator is not vanishing and both parts of the fraction are continuous. If $\bar{\mathbf{U}}_i = \bar{\mathbf{U}}_j$, we obtain

$$\begin{aligned}\big\|\mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j] - \mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]\big\|_{\mathrm{F}} &= \big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]\big\|_{\mathrm{F}} \\ &\leqslant \|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} = \big\|(\mathbf{U}_i - \mathbf{U}_j) - (\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j)\big\|_{\mathrm{F}} \leqslant \sqrt{2}\|U - \bar{U}\|_{2,\mathrm{F}} \quad \text{for all } U \in \mathbb{S}_d^N\end{aligned}$$

because $|\alpha_{ij}| \leqslant 1$.

Let us now show the Lipschitz continuity of $\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]$. The cases in which $\bar{\mathbf{U}}_i = \bar{\mathbf{U}}_j$ or $\mathbf{U}_i = \mathbf{U}_j$ can be treated in the same way as before. Therefore, without loss of generality we can assume that the denominators are nonvanishing. For the sake of simplicity, we use the following abbreviations

$$\begin{aligned} A_{ij} &\coloneqq A_{ij}(U), \qquad B_{ij} \coloneqq B_{ij}(U), \qquad \alpha_{ij} \coloneqq \alpha_{ij}(U), \qquad \mathbf{W}_{ij} \coloneqq \mathbf{U}_i - \mathbf{U}_j, \\ \bar{A}_{ij} &\coloneqq A_{ij}(\bar{U}), \qquad \bar{B}_{ij} \coloneqq B_{ij}(\bar{U}), \qquad \bar{\alpha}_{ij} \coloneqq \alpha_{ij}(\bar{U}), \qquad \bar{\mathbf{W}}_{ij} \coloneqq \bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j. \end{aligned} \tag{22}$$

Then, we have

$$\begin{aligned}\alpha_{ij} - \bar{\alpha}_{ij} &= \frac{A_{ij} - \bar{A}_{ij}}{\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}} + \frac{\bar{A}_{ij}}{\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}} - \frac{\bar{A}_{ij}}{\|\bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \bar{B}_{ij}} \\ &= \frac{A_{ij} - \bar{A}_{ij}}{\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}} + \bar{A}_{ij}\frac{\big(\|\bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \bar{B}_{ij}\big) - \big(\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}\big)}{\big(\|\bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \bar{B}_{ij}\big)\big(\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}\big)} \\ &= \frac{(A_{ij} - \bar{A}_{ij}) + \bar{\alpha}_{ij}\big(\|\bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \bar{B}_{ij}\big) - \bar{\alpha}_{ij}\big(\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}\big)}{\|\mathbf{W}_{ij}\|_{\mathrm{F}} + B_{ij}}\end{aligned} \tag{23}$$

and, hence, for $a_{ij} > 0$, $\mathbf{U}_i \neq \mathbf{U}_j$, and $\bar{\mathbf{U}}_i \neq \bar{\mathbf{U}}_j$

$$\begin{aligned}\big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] - \mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j]\big\|_{\mathrm{F}} &= \big\|\alpha_{ij}\mathbf{W}_{ij} - \bar{\alpha}_{ij}\bar{\mathbf{W}}_{ij}\big\|_{\mathrm{F}} \\ &\leqslant |\alpha_{ij} - \bar{\alpha}_{ij}|\,\|\mathbf{W}_{ij}\|_{\mathrm{F}} + \bar{\alpha}_{ij}\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\ &\leqslant 2\bar{\alpha}_{ij}\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \bar{\alpha}_{ij}|B_{ij} - \bar{B}_{ij}| + |A_{ij} - \bar{A}_{ij}| \\ &\leqslant 2\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}} + 2\|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}} + \big|B_{ij} - \bar{B}_{ij}\big| + \big|A_{ij} - \bar{A}_{ij}\big|\end{aligned} \tag{24}$$

due to the nonnegativity of B_{ij} and the *reverse triangle inequality*

$$\big|\|\mathbf{A}\|_{\mathrm{F}} - \|\mathbf{B}\|_{\mathrm{F}}\big| \leqslant \|\mathbf{A} - \mathbf{B}\|_{\mathrm{F}} \qquad \text{for all } \mathbf{A}, \mathbf{B} \in \mathbb{S}_d. \tag{25}$$

Therefore, $\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j]$ is (Lipschitz) continuous if A_{ij} and B_{ij} are (Lipschitz) continuous.

3.1. **Example of a scalar eigenvalue range limiter**

A scalar eigenvalue range limiter that admits representation (21) uses the correction factor

$$\alpha_{ij} \coloneqq \begin{cases} 1 & : a_{ij} \leqslant 0, \\ \min\Big\{1, \frac{q(u_i^{\max} - u_{i,d})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}, \frac{q(u_{i,1} - u_i^{\min})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}\Big\} & : a_{ij} > 0, \end{cases} \tag{26}$$

where $u_i^{\min}$ and $u_i^{\max}$ are local bounds for the eigenvalues $u_{i,k}$ of the tensor $\mathbf{U}_i$ such that

$$u_i^{\min} \leqslant u_{i,1} \leqslant \ldots \leqslant u_{i,d} \leqslant u_i^{\max} \tag{27}$$

and $q > 0$ is a positive adjustable constant (see below). For example, $u_i^{\min}$ and $u_i^{\max}$ can be defined as the local extrema in terms of the eigenvalues of the neighboring degrees of freedom, i.e.,

$$u_i^{\min} \coloneqq \min_{j,a_{ij}\neq 0} u_{j,1}, \qquad u_i^{\max} \coloneqq \max_{j,a_{ij}\neq 0} u_{j,d}.$$

Limiter (26) is designed such that α_{ij} vanishes if node i is a local extremum, i.e., $u_{i,d} = u_i^{\max}$ or $u_{i,1} = u_i^{\min}$. In what follows, this allows us to prove local maximum principles for the limiter.

By increasing q, the solution becomes less diffusive, but the Lipschitz constant increases and the nonlinear system becomes increasingly ill-conditioned. Due to (26), the functions $A_{ij}, B_{ij} : \mathbb{S}_d^N \to \mathbb{R}_0^+$ are given by

$$\begin{aligned} A_{ij}(U) &\coloneqq \min\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon, A_{ij}^{\max}(U), A_{ij}^{\min}(U)\big), \qquad B_{ij}(U) \coloneqq \varepsilon \geqslant 0, \\ A_{ij}^{\max}(U) &\coloneqq q(u_i^{\max} - u_{i,d}) \geqslant 0, \qquad A_{ij}^{\min}(U) \coloneqq q(u_{i,1} - u_i^{\min}) \geqslant 0. \end{aligned}$$

The Lipschitz continuity of B_{ij} can be shown trivially. Next, let us prove the Lipschitz continuity of $A_{ij}^{\max}$. First, for an arbitrary finite index set J, we have

$$\begin{aligned} &\max_{j\in J} v_j - \max_{j\in J} \bar{v}_j = \max_{j\in J}\big(v_j - \max_{l\in J} \bar{v}_l\big) \leqslant \max_{j\in J}(v_j - \bar{v}_j) \leqslant \max_{j\in J} |v_j - \bar{v}_j| \\ &\qquad\Longrightarrow \qquad \Big|\max_{j\in J} v_j - \max_{j} \bar{v}_j\Big| \leqslant \max_{j\in J} |v_j - \bar{v}_j| \end{aligned} \qquad \text{for all } v_j, \bar{v}_j \in \mathbb{R}. \tag{28}$$

Then, due to the *Wielandt-Hoffman theorem* [31, p. 104]

$$|v_k - \bar{v}_k|^2 \leqslant \sum_{\ell=1}^{d} |v_\ell - \bar{v}_\ell|^2 \leqslant \|\mathbf{V} - \bar{\mathbf{V}}\|_{\mathrm{F}}^2 \qquad \text{for all } \mathbf{V}, \bar{\mathbf{V}} \in \mathbb{S}_d \text{ and } 1 \leqslant k \leqslant d, \tag{29}$$

the Lipschitz continuity of $A_{ij}^{\max}$ follows from

$$\begin{aligned} \big|A_{ij}^{\max}(U) - A_{ij}^{\max}(\bar{U})\big| &= q\Big|\big(\max_{l,a_{il}\neq 0} u_{l,d} - u_{i,d}\big) - \big(\max_{l,a_{il}\neq 0} \bar{u}_{l,d} - \bar{u}_{i,d}\big)\Big| \\ &= q\Big|\max_{l,a_{il}\neq 0}(u_{l,d} - u_{i,d}) - \max_{l,a_{il}\neq 0}(\bar{u}_{l,d} - \bar{u}_{i,d})\Big| \\ &\leqslant q \max_{l\neq i,a_{il}\neq 0}\big|(u_{l,d} - u_{i,d}) - (\bar{u}_{l,d} - \bar{u}_{i,d})\big| \\ &\leqslant q\big(\max_{l\neq i,a_{il}\neq 0} |u_{l,d} - \bar{u}_{l,d}| + |u_{i,d} - \bar{u}_{i,d}|\big) \\ &\leqslant q\big(\max_{l\neq i,a_{il}\neq 0} \|\mathbf{U}_l - \bar{\mathbf{U}}_l\|_{\mathrm{F}} + \|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}}\big) \leqslant \sqrt{2}q\|U - \bar{U}\|_{2,\mathrm{F}}. \end{aligned}$$

For the proof of the Lipschitz continuity of $A_{ij}^{\min}$, consider $\mathbf{U}_i' \coloneqq -\mathbf{U}_i$ and $\bar{\mathbf{U}}_i' \coloneqq -\bar{\mathbf{U}}_i$, $1 \leqslant i \leqslant N$, and make use of the Lipschitz continuity of $A_{ij}^{\max}$. Thus, the proposed limiter (26) is Lipschitz continuous, too, because the minimum of two Lipschitz continuous functions is Lipschitz continuous due to (3c), and it follows that the AFC system (14) possesses a solution.

A desirable property of an AFC limiter is linearity preservation, which means that the method is consistent with (locally) linear functions: If the solution of the high order method is a linear function, then the limiter should produce $\alpha_{ij} = 1$, which is equivalent to

$$q(u_i^{\max} - u_{i,d}) \geqslant \|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon, \qquad \text{and} \qquad q(u_{i,1} - u_i^{\min}) \geqslant \|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon.$$

This will be the case if

$$u_i^{\min} + \tfrac{1}{q}\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon\big) =: \hat{u}_i^{\min} \leqslant u_{i,1}, \qquad u_{i,d} \leqslant \hat{u}_i^{\max} := u_i^{\max} - \tfrac{1}{q}\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon\big).$$

As long as $u_i^{\min} < u_{i,1}$ and $u_{i,d} < u_i^{\max}$, the limiter is linearity preserving for sufficiently large values of the coefficient q (see Fig. 1a). In the limit $q \to \infty$ the method preserves all linear functions that do not have constant minimal or maximal eigenvalues. Constant functions are preserved, because

$$\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j] = \mathcal{A}_{ij}[\mathbf{0}] = \mathbf{0} = \mathbf{U}_i - \mathbf{U}_j.$$

In [7, Theorem 6.1], the authors establish a lower bound for q only depending on the triangulation such that their scalar AFC method is linearity preserving. However, this approach does not seem to be suitable for enforcing linearity preservation in tensorial extensions intended to prove the LED property for the range of eigenvalues. For example, consider the two-dimensional tensors

$$\mathbf{U}_{i-1} = \begin{pmatrix} 1 & 0 \\ 0 & -1+\delta \end{pmatrix}, \qquad \mathbf{U}_i = \begin{pmatrix} 1 & 0 \\ 0 & \delta \end{pmatrix}, \qquad \mathbf{U}_{i+1} = \begin{pmatrix} 1 & 0 \\ 0 & 1+\delta \end{pmatrix},$$

where $0 \leqslant \delta \ll 1$ is a small nonnegative perturbation. They correspond to a linear tensor function $\mathbf{U} : \mathbb{R} \to \mathbb{S}_2$, if the corresponding consecutive nodes are arranged equidistantly. Then, we have

$$\alpha_{i+1,i} \overset{!}{=} 1 \qquad \Longleftrightarrow \qquad \frac{q(u_i^{\max} - u_{i,d})}{\|\mathbf{U}_i - \mathbf{U}_{i+1}\|_{\mathrm{F}} + \varepsilon} = q\delta(1+\varepsilon)^{-1} \overset{!}{\geqslant} 1$$

and q would have to depend nonlinearly on the perturbation δ. Especially in the case $\delta = 0$, there exists no $q > 0$ such that $\alpha_{ij} = 1$ due to $u_i^{\max} = u_{i,d}$ even though $\mathbf{U}$ is not constant. In the numerical examples, this issue will be discussed again.

In definition (26), the correction factors α_{ij} depend piecewise linearly on the eigenvalues $u_{i,1}$ and $u_{i,d}$. This leads to an abrupt derivative change in the transition between the linearity preserving region, where $\alpha_{ij} = 1$, and the region, where $\alpha_{ij} < 1$ (see Fig. 1b). To avoid this, a more general scalar limiter can be constructed, which smooths this blending by introducing a new parameter $p \in \mathbb{N}$. In this version, the correction factor is defined by

$$\alpha_{ij} := \begin{cases} 1 & : a_{ij} \leqslant 0, \\ 1 - \Big(1 - \min\Big\{1, \frac{q(u_i^{\max} - u_{i,d})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}, \frac{q(u_{i,1} - u_i^{\min})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}\Big\}\Big)^p & : a_{ij} > 0 \end{cases} \tag{30}$$

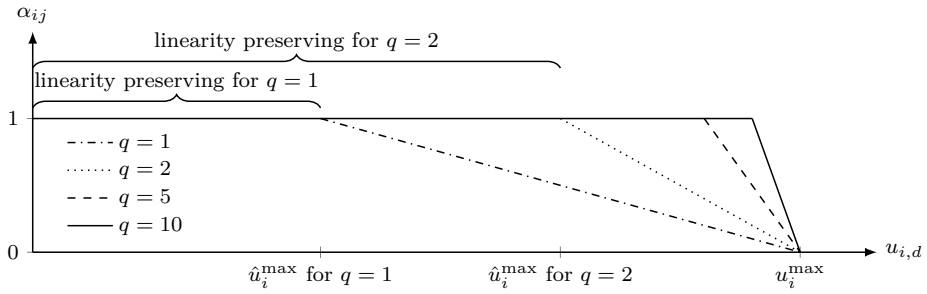
and coincides with the definition in (26) if $p = 1$. This results in a less diffusive limiter (due to larger correction factors; see Fig. 1b) and a greater Lipschitz constant. For the proof of Lipschitz continuity, we refer to Sec. 5, where a more general approach is considered and the limiter (30) can be treated similarly.

In what follows, we prove the local extremum diminishing (LED) property of the proposed limiters (26) and (30). For this purpose, we assume that the system matrix $A = (a_{ij})_{i,j=1}^N$ satisfies conditions (15) and

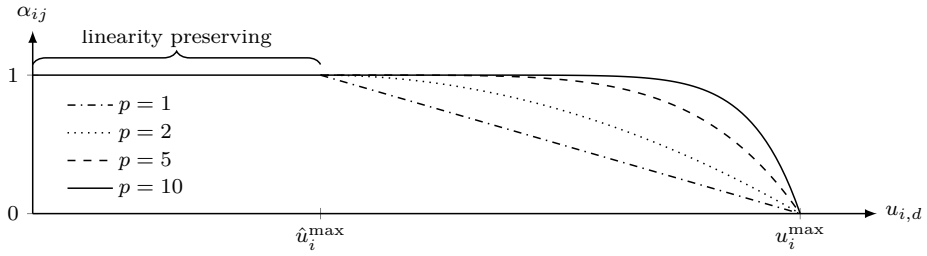
$$\sum_{j=1}^N a_{ij} \geqslant 0, \qquad a_{ii} > 0 \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{31}$$

while the artificial diffusion is defined as in (11). If A is positive definite, which is a requirement for the existence of an AFC solution, the condition $a_{ii} > 0$ is fulfilled automatically. The below proof of the maximum principle remains valid under the weaker assumption that

$$a_{ii} + \sum_{j \neq i, a_{ij} > 0} d_{ij} > 0 \text{ if } u_{i,1} \text{ or } u_{i,d} \text{ is a local extremum.} \tag{32}$$



(A) Dependence on $q > 0$ for fixed $p = 1$.



(B) Dependence on $p \in \mathbb{N}$ for fixed $q > 0$.

FIGURE 1. Influence of parameters on correction factors for the maximal eigenvalue.

This condition is automatically satisfied if $a_{ii} \geqslant 0$ and there exists a j such that $a_{ij} > 0$.

As already discussed, the condition $\min(a_{ij}, a_{ji}) \leqslant 0$ allows the definition of the limiter in an upwind based manner and can be used instead of the requirement $a_{ij} + a_{ji} \leqslant 0$ in [5]. If this is satisfied, either $a_{ij} \leqslant 0$ or $a_{ji} \leqslant 0$ for each edge ij and there is at most one troubled node (node j or i, respectively), where maximum principles could be violated by the antidiffusive flux corresponding to the edge ij.

To prove this, first of all, let us transform the AFC system (14) by exploiting definition (16) and separating $\mathbf{U}_i$

$$\begin{aligned}&\Big(a_{ii} + \sum_{j\neq i, a_{ij}>0} d_{ij}(1-\alpha_{ij}) + \sum_{j\neq i, a_{ij}\leqslant 0} d_{ij}(1-\alpha_{ji})\Big)\mathbf{U}_i \\ &+ \sum_{j\neq i, a_{ij}>0} \big(a_{ij} - d_{ij}(1-\alpha_{ij})\big)\mathbf{U}_j + \sum_{j\neq i, a_{ij}\leqslant 0} \big(a_{ij} - d_{ij}(1-\alpha_{ji})\big)\mathbf{U}_j = \mathbf{G}_i.\end{aligned}$$

Here, we do not need to treat the special case $a_{ij}, a_{ji} \leqslant 0$ separately, because $\alpha_{ji} = 1$ if $a_{ji} \leqslant 0$ by definition. Now, if $u_{i,d}$ is a local maximum of maximal eigenvalues, i.e.,

$$u_{i,d} = u_i^{\max} := \max_{j, a_{ij}\neq 0} u_{j,d} \quad \geqslant \quad \tilde{u}_i^{\max} := \max_{j\neq i, a_{ij}\neq 0} u_{j,d},$$

and $\mathbf{G}_i \preccurlyeq \mathbf{0}$, then we have $\alpha_{ij} = 0$ for all $1 \leqslant j \leqslant N$ due to the definition of the correction factors and

$$
\begin{aligned}
&\overbrace{\Big(a_{ii} + \sum_{j\neq i, a_{ij}>0} d_{ij} + \sum_{j\neq i, a_{ij}\leqslant 0} d_{ij}(1-\alpha_{ji})\Big)}^{>0}\mathbf{U}_i \\
&\quad\preccurlyeq -\sum_{j\neq i, a_{ij}>0}(a_{ij}-d_{ij})\mathbf{U}_j - \sum_{j\neq i, a_{ij}\leqslant 0}\big(a_{ij}-d_{ij}(1-\alpha_{ji})\big)\mathbf{U}_j \\
&\quad= \sum_{j\neq i, a_{ij}>0} \underbrace{(a_{ij}-d_{ij})}_{\leqslant 0 \text{ due to (11)}} \underbrace{(\tilde{u}_i^{\max}\mathbf{I}-\mathbf{U}_j)}_{\succcurlyeq \mathbf{0}} + \sum_{j\neq i, a_{ij}\leqslant 0} \underbrace{\big(a_{ij}-d_{ij}(1-\alpha_{ji})\big)}_{\leqslant 0} \underbrace{(\tilde{u}_i^{\max}\mathbf{I}-\mathbf{U}_j)}_{\succcurlyeq \mathbf{0}} \\
&\qquad - \sum_{j\neq i, a_{ij}>0}(a_{ij}-d_{ij})\tilde{u}_i^{\max}\mathbf{I} - \sum_{j\neq i, a_{ij}\leqslant 0}\big(a_{ij}-d_{ij}(1-\alpha_{ji})\big)\tilde{u}_i^{\max}\mathbf{I} \\
&\quad\preccurlyeq -\sum_{j\neq i} a_{ij}\tilde{u}_i^{\max}\mathbf{I} + \sum_{j\neq i, a_{ij}>0} d_{ij}\tilde{u}_i^{\max}\mathbf{I} + \sum_{j\neq i, a_{ij}\leqslant 0} d_{ij}(1-\alpha_{ji})\tilde{u}_i^{\max}\mathbf{I} \\
&\quad\preccurlyeq \Big(a_{ii} + \sum_{j\neq i, a_{ij}>0} d_{ij} + \sum_{j\neq i, a_{ij}\geqslant 0} d_{ij}(1-\alpha_{ji})\Big)\max(0,\tilde{u}_i^{\max})\mathbf{I},
\end{aligned}
\tag{33}
$$

where we used $\sum_j a_{ij} \geqslant 0$ in the last inequality. Therefore,

$$
\mathbf{G}_i \preccurlyeq \mathbf{0} \qquad \Longrightarrow \qquad u_{i,d} \leqslant \max\big(0, \max_{j\neq i, a_{ij}\neq 0} u_{j,d}\big) \qquad \text{for all } 1 \leqslant i \leqslant N \tag{34a}
$$

and the maximal eigenvalue $u_{i,d}$ is bounded above by the neighboring maximal eigenvalues or zero. Here, the upwind based design of the limiter is exploited, because $a_{ij} - d_{ij}(1-\alpha_{ji}) \leqslant 0$ for any $0 \leqslant \alpha_{ji} \leqslant 1$ if $a_{ij} \leqslant 0$. Similarly to (33), the minimal eigenvalue $u_{i,1}$ is bounded below as follows

$$
\mathbf{G}_i \succcurlyeq \mathbf{0} \qquad \Longrightarrow \qquad u_{i,1} \geqslant \min\big(0, \min_{j\neq i, a_{ij}\neq 0} u_{j,1}\big) \qquad \text{for all } 1 \leqslant i \leqslant N. \tag{34b}
$$

In the case $\sum_j a_{ij} = 0$, the last inequality in (33) becomes an identity and we can prove that

$$
\sum_j a_{ij} = 0 \quad \text{and} \quad \mathbf{G}_i \preccurlyeq \mathbf{0} \quad \Longrightarrow \quad u_{i,d} \leqslant \max_{j\neq i, a_{ij}\neq 0} u_{j,d} \quad \text{for all } 1 \leqslant i \leqslant N, \tag{34c}
$$

$$
\sum_j a_{ij} = 0 \quad \text{and} \quad \mathbf{G}_i \succcurlyeq \mathbf{0} \quad \Longrightarrow \quad u_{i,1} \geqslant \min_{j\neq i, a_{ij}\neq 0} u_{j,1} \quad \text{for all } 1 \leqslant i \leqslant N. \tag{34d}
$$

These maximum principles guarantee not only preservation of definiteness but also eigenvalue range preservation.

4. Two-node scalar limiting

For the proposed limiter to be eigenvalue range preserving, each edge must possess at least one harmless node, such that the antidiffusive flux does not have to be scaled to satisfy maximum principles at this node. This requirement restricts the set of admissible system matrices and excludes certain discretizations of a boundary value problem (see below).

To avoid this disadvantage, we introduce a two-node version of the scalar limiter, in which the antidiffusive flux is scaled so that both nodes are allowed to be troubled and maximum principles are still satisfied. If the assumption (15) is waived, the limiting operator can be defined by

$$
\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j] \coloneqq \min(\alpha_{ij}, \alpha_{ji})(\mathbf{U}_i - \mathbf{U}_j), \tag{35}
$$

where $0 \leqslant \alpha_{ij} \leqslant 1$ guarantees the satisfaction of maximum principles at node i while $0 \leqslant \alpha_{ji} \leqslant 1$ is responsible for preventing local overshoots and undershoots at node j.

If α_{ij} and α_{ji} can be expressed as before in the form

$$0 \leqslant \alpha_{ij} \coloneqq \frac{A_{ij}}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + B_{ij}} \leqslant 1 \qquad \text{and} \qquad 0 \leqslant \alpha_{ji} \coloneqq \frac{A_{ji}}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + B_{ji}} \leqslant 1, \tag{36}$$

where $A_{ij}, A_{ji}, B_{ij}, B_{ji} : \mathbb{S}_d^N \to \mathbb{R}_0^+$ are nonnegative and Lipschitz continuous functions, then the synchronized version is also Lipschitz continuous due to (3c), (23), and (25)

$$\begin{aligned}
&\big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] - \mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j]\big\|_{\mathrm{F}} \\
&\quad = \big\|\min(\alpha_{ij}, \alpha_{ji})\mathbf{W}_{ij} - \min(\bar{\alpha}_{ij}, \bar{\alpha}_{ji})\bar{\mathbf{W}}_{ij}\big\|_{\mathrm{F}} \\
&\quad \leqslant \big|\min(\alpha_{ij}, \alpha_{ji}) - \min(\bar{\alpha}_{ij}, \bar{\alpha}_{ji})\big|\|\mathbf{W}_{ij}\|_{\mathrm{F}} + \min(\bar{\alpha}_{ij}, \bar{\alpha}_{ji})\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\
&\quad \leqslant \big(|\alpha_{ij} - \bar{\alpha}_{ij}| + |\alpha_{ji} - \bar{\alpha}_{ji}|\big)\|\mathbf{W}_{ij}\|_{\mathrm{F}} + \min(\bar{\alpha}_{ij}, \bar{\alpha}_{ji})\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\
&\quad \leqslant 3\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + |A_{ij} - \bar{A}_{ij}| + |A_{ji} - \bar{A}_{ji}| + |B_{ij} - \bar{B}_{ij}| + |B_{ji} - \bar{B}_{ji}|
\end{aligned}$$

and there exists a solution of the AFC system (14). The correction factors α_{ij} defined as before, e.g., by (26) or (30), are admissible and guarantee local maximum principles without distinguishing between good and bad edge contributions (the proof is very similar to (33)). Therefore, each contribution can be potentially troubled and (35) does not impose as many restrictions on the system matrix A as the one-node limiter (16).

If the system matrix possesses the property $\min(a_{ij}, a_{ji}) \leqslant 0$ for all $i \neq j$, then the new limiter (35) coincides with (16), because the definition of the correction factor α_{ij} already takes the sign of a_{ij} into consideration.

5. One-node tensorial limiting

In Sec. 3 (and Sec. 4), we focused on a simple scalar limiting approach, whereby each tensor entry of the antidiffusive flux is scaled by the same correction factor. Thus, all eigenvalues of the tensor quantity are coupled with each other and the resulting solution seems to be unnecessarily diffusive. In what follows, we replace the scalar limiting coefficient by a tensorial correction factor, which is able to treat each eigenvalue at the troubled node in a segregated manner. Limiting strategies of this kind produce less diffusive results, but give rise to increasing computational costs and tend to degrade the convergence behavior of iterative solvers.

Assuming (15) is valid, we define the one-node tensor limiting operator $\mathcal{A}_{ij}$ as follows:

$$\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j] \coloneqq \begin{cases} \frac{1}{2}\big((\mathbf{U}_i - \mathbf{U}_j)\mathbf{S}_{ij} + \mathbf{S}_{ij}(\mathbf{U}_i - \mathbf{U}_j)\big) & : a_{ij} > 0, a_{ji} \leqslant 0, \\ \frac{1}{2}\big((\mathbf{U}_i - \mathbf{U}_j)\mathbf{S}_{ji} + \mathbf{S}_{ji}(\mathbf{U}_i - \mathbf{U}_j)\big) & : a_{ij} \leqslant 0, a_{ji} > 0, \\ \mathbf{U}_i - \mathbf{U}_j & : a_{ij} \leqslant 0, a_{ji} \leqslant 0, \end{cases} \tag{37}$$

where $\mathbf{S}_{ij} \coloneqq \mathbf{Q}_i \tilde{\mathbf{S}}_{ij} \mathbf{Q}_i^\top$ is the tensorial correction factor, $\mathbf{Q}_i$ is the orthogonal tensor of the spectral decomposition $\mathbf{U}_i = \mathbf{Q}_i \tilde{\mathbf{U}}_i \mathbf{Q}_i^\top$, and the diagonal tensor $\tilde{\mathbf{S}}_{ij}$ contains the correction factors $s_{ij,k} \coloneqq \alpha_{ij,k}$ with $0 \leqslant \alpha_{ij,k} \leqslant 1$ defined below. As in the scalar case, node i is the troubled node of the edge ij if $a_{ij} > 0$ and $\alpha_{ij,k}$ are defined so that maximum principles are preserved at node i.

By definition (37), we have due to the Cauchy-Schwarz inequality and (2)

$$\begin{aligned}
\big(\mathbf{V}, (\mathcal{I} - \mathcal{A}_{ij})[\mathbf{V}]\big)_{\mathrm{F}} &= (\mathbf{V}, \mathbf{V})_{\mathrm{F}} - \tfrac{1}{2}(\mathbf{V}, \mathbf{V}\mathbf{S}_{ij} + \mathbf{S}_{ij}\mathbf{V})_{\mathrm{F}} = \|\mathbf{V}\|_{\mathrm{F}}^2 - (\mathbf{V}, \mathbf{S}_{ij}\mathbf{V})_{\mathrm{F}} \\
&\geqslant \|\mathbf{V}\|_{\mathrm{F}}^2 - \|\mathbf{V}\|_{\mathrm{F}}\|\mathbf{S}_{ij}\mathbf{V}\|_{\mathrm{F}} \geqslant \|\mathbf{V}\|_{\mathrm{F}}^2 - \|\mathbf{S}_{ij}\|_2\|\mathbf{V}\|_{\mathrm{F}}^2 \geqslant 0 \qquad \text{if } a_{ij} > 0, a_{ji} \leqslant 0,
\end{aligned} \tag{38}$$

because the trace is invariant under cyclic permutations and $\|\mathbf{S}_{ij}\|_2^2 = \max_k s_{ij,k}^2 \leqslant 1$ holds. Further on, we can show that

$$\big\|\mathcal{A}_{ij}[\mathbf{V}]\big\|_{\mathrm{F}}^2 \leqslant \|\mathbf{S}_{ij}\mathbf{V}\|_{\mathrm{F}}^2 \leqslant \|\mathbf{S}_{ij}\|_2^2\|\mathbf{V}\|_{\mathrm{F}}^2 \leqslant \|\mathbf{V}\|_{\mathrm{F}}^2 \qquad \text{if } a_{ij} > 0, a_{ji} \leqslant 0. \tag{39}$$

In the case $a_{ij} \leqslant 0, a_{ji} > 0$, the estimates are the same except that $\mathbf{S}_{ij}$ has to be replaced by $\mathbf{S}_{ji}$.

To show the Lipschitz continuity of the new limiting operator, we start with the general form

$$\mathbf{0} \preccurlyeq \mathbf{S}_{ij} \coloneqq \frac{\mathbf{A}_{ij}}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + B_{ij}} \preccurlyeq \mathbf{I},$$

where $\mathbf{A}_{ij} : \mathbb{S}_d^N \to \mathbb{S}_{d,+}$ and $B_{ij} : \mathbb{S}_d^N \to \mathbb{R}_0^+$ are Lipschitz continuous functions depending on the solution U. In contrast to the scalar approach (21), the numerator is now a positive semidefinite tensor quantity. Let $\bar{U} \in \mathbb{S}_d^N$ be arbitrary. In the special case $\bar{\mathbf{U}}_i = \bar{\mathbf{U}}_j$, continuity follows directly from (39)

$$\begin{aligned}
&\big\|\mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j] - \mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]\big\|_{\mathrm{F}} = \big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j]\big\|_{\mathrm{F}} \\
&\quad \leqslant \|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} = \big\|(\mathbf{U}_i - \mathbf{U}_j) - (\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j)\big\|_{\mathrm{F}} \leqslant \sqrt{2}\|U - \bar{U}\|_{2,\mathrm{F}} \quad \text{for all } U \in \mathbb{S}_d^N.
\end{aligned}$$

Let us now consider the case $\mathbf{U}_i \neq \mathbf{U}_j$, $\bar{\mathbf{U}}_i \neq \bar{\mathbf{U}}_j$. Using the same notation as in the scalar case (see (22)), the tensorial extension of (23) (replace α_{ij} and A_{ij} by $\mathbf{S}_{ij}$ and $\mathbf{A}_{ij}$, respectively), and (2), we obtain

$$\begin{aligned}
&\big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] - \mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j]\big\|_{\mathrm{F}} \\
&\quad = \tfrac{1}{2}\big\|(\mathbf{W}_{ij}\mathbf{S}_{ij} + \mathbf{S}_{ij}\mathbf{W}_{ij}) - (\bar{\mathbf{W}}_{ij}\bar{\mathbf{S}}_{ij} + \bar{\mathbf{S}}_{ij}\bar{\mathbf{W}}_{ij})\big\|_{\mathrm{F}} \\
&\quad \leqslant \tfrac{1}{2}\big\|\mathbf{W}_{ij}(\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij})\big\|_{\mathrm{F}} + \tfrac{1}{2}\big\|(\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij})\mathbf{W}_{ij}\big\|_{\mathrm{F}} + \tfrac{1}{2}\big\|(\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij})\bar{\mathbf{S}}_{ij}\big\|_{\mathrm{F}} + \tfrac{1}{2}\big\|\bar{\mathbf{S}}_{ij}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij})\big\|_{\mathrm{F}} \\
&\quad \leqslant \|\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij}\|_2\|\mathbf{W}_{ij}\|_{\mathrm{F}} + \|\bar{\mathbf{S}}_{ij}\|_2\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\
&\quad \leqslant \|\mathbf{A}_{ij} - \bar{\mathbf{A}}_{ij}\|_2 + \|\bar{\mathbf{S}}_{ij}\|_2|B_{ij} - \bar{B}_{ij}| + 2\|\bar{\mathbf{S}}_{ij}\|_2\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\
&\quad \leqslant 2\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \|\mathbf{A}_{ij} - \bar{\mathbf{A}}_{ij}\|_2 + |B_{ij} - \bar{B}_{ij}|
\end{aligned} \tag{40}$$

and $\mathcal{A}_{ij}[\mathbf{U}_i - \mathbf{U}_j]$ is Lipschitz continuous if $\mathbf{A}_{ij}$ and B_{ij} are Lipschitz continuous.

5.1. **Example of a tensorial eigenvalue range limiter**

Similarly to the scalar one-node eigenvalue range limiter, the correction factors $\alpha_{ij,k}$ can be defined by

$$\alpha_{ij,k} \coloneqq \begin{cases} 1 & : a_{ij} \leqslant 0, \\ 1 - \Big(1 - \min\Big\{1, \frac{q(u_i^{\max} - u_{i,k})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}, \frac{q(u_{i,k} - u_i^{\min})}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}\Big\}\Big)^p & : a_{ij} > 0 \end{cases} \tag{41}$$

for all $1 \leqslant k \leqslant d$, where $q > 0$ and $p \in \mathbb{N}$ are adjustable parameters of the limiter. As in the scalar case, increasing q or p results in a less diffusive limiter with a larger Lipschitz constant. Additionally, the limiter should be less restrictive than its scalar counterpart due to the estimates

$$u_i^{\max} - u_{i,k} \geqslant u_i^{\max} - u_{i,d}, \qquad u_{i,k} - u_i^{\min} \geqslant u_{i,1} - u_i^{\min} \qquad \text{for all } 1 \leqslant k \leqslant d.$$

If $a_{ij} > 0$, the correction factor $0 \leqslant \alpha_{ij,k} \leqslant 1$ defined by (41) can be written conveniently as

$$\begin{aligned}
\alpha_{ij,k} &= \frac{a_{ij,k}}{\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon}, \qquad a_{ij,k} \coloneqq \frac{\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon\big)^p - c_{ij,k}^p}{\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon\big)^{p-1}}, \\
c_{ij,k} &\coloneqq \max\big(0, \|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon - q\min(u_i^{\max} - u_{i,k}, u_{i,k} - u_i^{\min})\big),
\end{aligned} \tag{42}$$

where $0 \leqslant c_{ij,k} \leqslant \|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + \varepsilon$. Hence, the tensor $\mathbf{0} \preccurlyeq \mathbf{S}_{ij} \preccurlyeq \mathbf{I}$, which scales the antidiffusive flux, is given by

$$\mathbf{S}_{ij} \coloneqq \frac{\mathbf{A}_{ij}}{\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}}, \qquad \mathbf{A}_{ij} \coloneqq \frac{\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big)^p \mathbf{I} - \mathbf{C}_{ij}^p}{\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big)^{p-1}}, \qquad B_{ij} \coloneqq \varepsilon,$$
$$\mathbf{C}_{ij} \coloneqq \mathbf{Q}_i \tilde{\mathbf{C}}_{ij} \mathbf{Q}_i^\top = \max\big(\mathbf{0}, \big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big)\mathbf{I} - q \min(u_i^{\max}\mathbf{I} - \mathbf{U}_i, \mathbf{U}_i - u_i^{\min}\mathbf{I})\big),$$

where we used the definitions (4) for $\max(\cdot,\cdot)$ and $\min(\cdot,\cdot)$ of tensors, e.g.,

$$\begin{aligned}
&\mathbf{q}_{i,k}^\top \min(u_i^{\max}\mathbf{I} - \mathbf{U}_i, \mathbf{U}_i - u_i^{\min}\mathbf{I})\mathbf{q}_{i,k} \\
&\quad = \tfrac{1}{2}\mathbf{q}_{i,k}^\top \big((u_i^{\max}\mathbf{I} - \mathbf{U}_i) + (\mathbf{U}_i - u_i^{\min}\mathbf{I}) - \big|(u_i^{\max}\mathbf{I} - \mathbf{U}_i) - (\mathbf{U}_i - u_i^{\min}\mathbf{I})\big|\big)\mathbf{q}_{i,k} \\
&\quad = \tfrac{1}{2}\big((u_i^{\max} - u_{i,k}) + (u_{i,k} - u_i^{\min}) - \big|(u_i^{\max} - u_{i,k}) - (u_{i,k} - u_i^{\min})\big|\big) \\
&\quad = \min(u_i^{\max} - u_{i,k}, u_{i,k} - u_i^{\min}).
\end{aligned}$$

If $a_{ij} \leqslant 0$, we have

$$a_{ij,k} \coloneqq \|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + \varepsilon \quad \text{for all } 1 \leqslant k \leqslant d \qquad \Longrightarrow \qquad \mathbf{C}_{ij} \coloneqq \mathbf{0}, \qquad \mathbf{S}_{ij} = \mathbf{I}.$$

Note that the definitions of $\alpha_{ij,k}$, $a_{ij,k}$, and $c_{ij,k}$ in (42) do not follow the convention that eigenvalues are sorted, because they define auxiliary quantities corresponding to an eigenvalue $u_{i,k}$, which is responsible for the order of the eigenvectors $\mathbf{q}_k$.

While the Lipschitz continuity of $B_{ij} = \varepsilon$ is trivial, we need to show the Lipschitz continuity of $\mathbf{A}_{ij}$. First of all, using the telescope sum, $\mathbf{D}_{ij} \coloneqq \mathbf{C}_{ij}(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij})^{-1}$, and $\bar{\mathbf{D}}_{ij} \coloneqq \bar{\mathbf{C}}_{ij}(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} + \bar{B}_{ij})^{-1}$, we obtain

$$\begin{aligned}
&\mathbf{A}_{ij} - \bar{\mathbf{A}}_{ij} - \big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij} - \|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} - \bar{B}_{ij}\big)\mathbf{I} \\
&\quad = \bar{\mathbf{D}}_{ij}^p\big(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} + \bar{B}_{ij}\big) - \mathbf{D}_{ij}^p\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big) \\
&\quad = \sum_{l=0}^{p-1} \mathbf{D}_{ij}^l \bar{\mathbf{D}}_{ij}^{p-l}\big(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} + \bar{B}_{ij}\big) - \sum_{l=1}^{p-1} \mathbf{D}_{ij}^l \bar{\mathbf{D}}_{ij}^{p-l}\big(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} + \bar{B}_{ij}\big) \\
&\qquad + \sum_{l=1}^{p-1} \mathbf{D}_{ij}^l \bar{\mathbf{D}}_{ij}^{p-l}\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big) - \sum_{l=0}^{p-1} \mathbf{D}_{ij}^{l+1} \bar{\mathbf{D}}_{ij}^{p-1-l}\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij}\big) \\
&\quad = \sum_{l=0}^{p-1} \mathbf{D}_{ij}^l (\bar{\mathbf{C}}_{ij} - \mathbf{C}_{ij}) \bar{\mathbf{D}}_{ij}^{p-1-l} + \sum_{l=1}^{p-1} \mathbf{D}_{ij}^l \bar{\mathbf{D}}_{ij}^{p-l}\big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + B_{ij} - \|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F} - \bar{B}_{ij}\big).
\end{aligned}$$

According to (2), this yields the following estimate for the Frobenius norm

$$\begin{aligned}
\big\|\mathbf{A}_{ij} - \bar{\mathbf{A}}_{ij}\big\|_\mathrm{F} &\leqslant \Big(\big|\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} - \|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F}\big| + |B_{ij} - \bar{B}_{ij}|\Big)\|\mathbf{I}\|_\mathrm{F} + \|\mathbf{C}_{ij} - \bar{\mathbf{C}}_{ij}\|_\mathrm{F}\Big(\sum_{l=0}^{p-1} \|\mathbf{D}_{ij}\|_2^l \|\bar{\mathbf{D}}_{ij}\|_2^{p-1-l}\Big) \\
&\quad + \Big(\big|\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} - \|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_\mathrm{F}\big| + |B_{ij} - \bar{B}_{ij}|\Big)\|\mathbf{I}\|_\mathrm{F}\Big(\sum_{l=1}^{p-1} \|\mathbf{D}_{ij}\|_2^l \|\bar{\mathbf{D}}_{ij}\|_2^{p-l}\Big) \\
&\leqslant p\sqrt{d}\big(\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_\mathrm{F} + \|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_\mathrm{F} + |B_{ij} - \bar{B}_{ij}|\big) + p\|\mathbf{C}_{ij} - \bar{\mathbf{C}}_{ij}\|_\mathrm{F},
\end{aligned}$$

because $\|\mathbf{D}_{ij}\|_2, \|\bar{\mathbf{D}}_{ij}\|_2 \leqslant 1$. Therefore, $\mathbf{A}_{ij}$ is Lipschitz continuous if additionally $\mathbf{C}_{ij}$ is Lipschitz continuous, which is fulfilled due to (7), (25), (28), (29), and

$$\|\mathbf{C}_{ij} - \bar{\mathbf{C}}_{ij}\|_\mathrm{F} = \Big\|\max\big(\mathbf{0}, \big(\|\mathbf{U}_i - \mathbf{U}_j\|_\mathrm{F} + \varepsilon\big)\mathbf{I} - q \min(u_i^{\max}\mathbf{I} - \mathbf{U}_i, \mathbf{U}_i - u_i^{\min}\mathbf{I})\big)$$

$$
\begin{aligned}
&\quad - \max\big(\mathbf{0}, \big(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_{\mathrm{F}} + \varepsilon\big)\mathbf{I} - q\min(\bar{u}_i^{\max}\mathbf{I} - \bar{\mathbf{U}}_i, \bar{\mathbf{U}}_i - \bar{u}_i^{\min}\mathbf{I})\big)\Big\|_{\mathrm{F}} \\
&\leqslant \Big\|\big(\|\mathbf{U}_i - \mathbf{U}_j\|_{\mathrm{F}} + \varepsilon\big)\mathbf{I} - q\min(u_i^{\max}\mathbf{I} - \mathbf{U}_i, \mathbf{U}_i - u_i^{\min}\mathbf{I}) \\
&\quad - \big(\|\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j\|_{\mathrm{F}} + \varepsilon\big)\mathbf{I} + q\min(\bar{u}_i^{\max}\mathbf{I} - \bar{\mathbf{U}}_i, \bar{\mathbf{U}}_i - \bar{u}_i^{\min}\mathbf{I})\Big\|_{\mathrm{F}} \\
&\leqslant \|\mathbf{I}\|_{\mathrm{F}}\big(\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}} + \|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}}\big) \\
&\quad + q\|u_i^{\max}\mathbf{I} - \mathbf{U}_i - \bar{u}_i^{\max}\mathbf{I} + \bar{\mathbf{U}}_i\|_{\mathrm{F}} + q\|\mathbf{U}_i - u_i^{\min}\mathbf{I} - \bar{\mathbf{U}}_i + \bar{u}_i^{\min}\mathbf{I}\|_{\mathrm{F}} \\
&\leqslant \big(\|\mathbf{I}\|_{\mathrm{F}} + q\big)\big(\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}} + \|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}}\big) + q\|\mathbf{I}\|_{\mathrm{F}}\big(|u_i^{\max} - \bar{u}_i^{\max}| + |u_i^{\min} - \bar{u}_i^{\min}|\big) \\
&\leqslant \big(\sqrt{d} + q\big)\big(\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}} + \|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}}\big) + q\sqrt{d}\max_{j,a_{ij}\neq 0}|u_{j,d} - \bar{u}_{j,d}| + q\sqrt{d}\max_{j,a_{ij}\neq 0}|u_{j,1} - \bar{u}_{j,1}| \\
&\leqslant \big(\sqrt{d} + q\big)\big(\|\mathbf{U}_i - \bar{\mathbf{U}}_i\|_{\mathrm{F}} + \|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}}\big) + 2q\sqrt{d}\max_{j,a_{ij}\neq 0}\|\mathbf{U}_j - \bar{\mathbf{U}}_j\|_{\mathrm{F}}.
\end{aligned}
$$

In summary, the limiter defined by (41) is Lipschitz continuous, which proves the existence of a solution to the AFC problem (14). In a similar vein, the Lipschitz continuity of the generalized scalar tensor limiter (30) can be proved, where $\mathbf{A}_{ij}$ and $\mathbf{C}_{ij}$ are replaced by scalar quantities.

Furthermore, we can prove maximum principles as in the case of the scalar eigenvalue range limiter (26). Before doing this particularly for the proposed limiter (41), we define sufficient requirements for a linear limiting operator $\mathcal{A}_{ij}$ to satisfy local maximum principles:

$$
\begin{aligned}
&\text{If } u_{i,d} = u_i^{\max}\text{, then for all } j \neq i \text{ s.t. } a_{ij} \neq 0 \text{ and } u_{j,d} \leqslant v \leqslant \max(0, u_{i,d}) \\
&\quad \exists\, \underline{\alpha}_{ij}^{\max} \leqslant 1: \qquad \mathbf{q}_{i,d}^{\top}\mathcal{A}_{ij}[\mathbf{U}_i - v\mathbf{I}]\mathbf{q}_{i,d} \leqslant \underline{\alpha}_{ij}^{\max}(u_{i,d} - v), \\
&\qquad a_{ij} > 0: \qquad \mathbf{q}_{i,d}^{\top}\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,d} \geqslant 0, \\
&\qquad a_{ij} \leqslant 0: \qquad \mathbf{q}_{i,d}^{\top}\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,d} \geqslant \mathbf{q}_{i,d}^{\top}\mathbf{U}_j\mathbf{q}_{i,d} - v =: \hat{u}_{j,d} - v,
\end{aligned}
\tag{43a}
$$

$$
\begin{aligned}
&\text{If } u_{i,1} = u_i^{\min}\text{, then for all } j \neq i \text{ s.t. } a_{ij} \neq 0 \text{ and } \min(0, u_{i,1}) \leqslant v \leqslant u_{j,1} \\
&\quad \exists\, \underline{\alpha}_{ij}^{\min} \leqslant 1: \qquad \mathbf{q}_{i,1}^{\top}\mathcal{A}_{ij}[\mathbf{U}_i - v\mathbf{I}]\mathbf{q}_{i,1} \geqslant \underline{\alpha}_{ij}^{\min}(u_{i,1} - v), \\
&\qquad a_{ij} > 0: \qquad \mathbf{q}_{i,1}^{\top}\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,1} \leqslant 0, \\
&\qquad a_{ij} \leqslant 0: \qquad \mathbf{q}_{i,1}^{\top}\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,1} \leqslant \mathbf{q}_{i,1}^{\top}\mathbf{U}_j\mathbf{q}_{i,1} - v =: \hat{u}_{j,1} - v.
\end{aligned}
\tag{43b}
$$

If these assumptions and (31) hold, the solution of the AFC system (14) satisfies

$$
\hat{g}_{i,d} \leqslant 0 \quad \Longrightarrow \quad u_{i,d} \leqslant \max\big(0, \max_{j\neq i, a_{ij}\neq 0} u_{j,d}\big) \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{44a}
$$

$$
\hat{g}_{i,1} \geqslant 0 \quad \Longrightarrow \quad u_{i,1} \geqslant \min\big(0, \min_{j\neq i, a_{ij}\neq 0} u_{j,1}\big) \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{44b}
$$

$$
\sum_j a_{ij} = 0 \quad \text{and} \quad \hat{g}_{i,d} \leqslant 0 \quad \Longrightarrow \quad u_{i,d} \leqslant \max_{j\neq i, a_{ij}\neq 0} u_{j,d} \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{44c}
$$

$$
\sum_j a_{ij} = 0 \quad \text{and} \quad \hat{g}_{i,1} \geqslant 0 \quad \Longrightarrow \quad u_{i,1} \geqslant \min_{j\neq i, a_{ij}\neq 0} u_{j,1} \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{44d}
$$

where $\hat{g}_{i,1} \coloneqq \mathbf{q}_{i,1}^{\top}\mathbf{G}_i\mathbf{q}_{i,1}$ and $\hat{g}_{i,d} \coloneqq \mathbf{q}_{i,d}^{\top}\mathbf{G}_i\mathbf{q}_{i,d}$. To prove this, first of all, we note that for any $v \in \mathbb{R}$ the AFC system (14) can be written as

$$
\big(\sum_j a_{ij}\big)v\mathbf{I} + \big(a_{ii}\mathcal{I} + \sum_{j\neq i} d_{ij}(\mathcal{I} - \mathcal{A}_{ij})\big)\big[\mathbf{U}_i - v\mathbf{I}\big] + \sum_{j\neq i}\big(a_{ij}\mathcal{I} - d_{ij}(\mathcal{I} - \mathcal{A}_{ij})\big)\big[\mathbf{U}_j - v\mathbf{I}\big] = \mathbf{G}_i. \tag{45}
$$

To prove (44a), let us assume that $u_{i,d}$ is a local maximum of maximal eigenvalues, i.e.,

$$u_{i,d} = u_i^{\max} \coloneqq \max_{j,a_{ij}\neq 0} u_{j,d} \quad \geqslant \quad \tilde{u}_i^{\max} \coloneqq \max_{j\neq i,a_{ij}\neq 0} u_{j,d}.$$

Otherwise, (44a) holds trivially. Then, on the one hand, according to (43a), we have

$$\mathbf{q}_{i,d}^\top\big(a_{ij}\mathcal{I} - d_{ij}(\mathcal{I}-\mathcal{A}_{ij})\big)\big[\mathbf{U}_j - v\mathbf{I}\big]\mathbf{q}_{i,d} \geqslant 0 \qquad \text{for all } 1\leqslant j\leqslant N \text{ and all } u_{j,d}\leqslant v\leqslant \max(0,u_{i,d})$$

because, if $a_{ij}>0$

$$\mathbf{q}_{i,d}^\top\big(a_{ij}\mathcal{I} - d_{ij}(\mathcal{I}-\mathcal{A}_{ij})\big)\big[\mathbf{U}_j - v\mathbf{I}\big]\mathbf{q}_{i,d} = \underbrace{(a_{ij}-d_{ij})}_{\leqslant 0}\underbrace{(\hat{u}_{j,d}-v)}_{\leqslant 0} + d_{ij}\underbrace{\mathbf{q}_{i,d}^\top\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,d}}_{\geqslant 0} \geqslant 0$$

due to the min-max theorem, i.e., $\hat{u}_{j,d}\leqslant u_{j,d}\leqslant v$, and otherwise

$$\mathbf{q}_{i,d}^\top\big(a_{ij}\mathcal{I} - d_{ij}(\mathcal{I}-\mathcal{A}_{ij})\big)[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,d} = (a_{ij}-d_{ij})(\hat{u}_{j,d}-v) + d_{ij}\mathbf{q}_{i,d}^\top\mathcal{A}_{ij}[\mathbf{U}_j - v\mathbf{I}]\mathbf{q}_{i,d} \geqslant \underbrace{a_{ij}}_{\leqslant 0}\underbrace{(\hat{u}_{j,d}-v)}_{\leqslant 0} \geqslant 0.$$

On the other hand, the first condition of (43a) leads to

$$\begin{aligned}\mathbf{q}_{i,d}^\top\big(a_{ii}\mathcal{I} + \sum_{j\neq i} d_{ij}(\mathcal{I}-\mathcal{A}_{ij})\big)\big[\mathbf{U}_i - v\mathbf{I}\big]\mathbf{q}_{i,d} &= \big(a_{ii}+\sum_{j\neq i} d_{ij}\big)(u_{i,d}-v) - \sum_{j\neq i} d_{ij}\mathbf{q}_{i,d}^\top\mathcal{A}_{ij}\big[\mathbf{U}_i - v\mathbf{I}\big]\mathbf{q}_{i,d}\\ &\geqslant \big(a_{ii}+\sum_{j\neq i} d_{ij}(1-\underline{\alpha}_{ij}^{\max})\big)(u_{i,d}-v).\end{aligned}$$

Therefore, using $v=\max(0,\tilde{u}_i^{\max})$, multiplication of (45) from left and right by $\mathbf{q}_{i,d}^\top$ and $\mathbf{q}_{i,d}$ yields

$$\hat{g}_{i,d} \geqslant \underbrace{\big(\sum_j a_{ij}\big)\max(0,\tilde{u}_i^{\max})}_{\geqslant 0} + \underbrace{\big(a_{ii}+\sum_{j\neq i} d_{ij}(1-\underline{\alpha}_{ij}^{\max})\big)}_{>0}\big(u_{i,d}-\max(0,\tilde{u}_i^{\max})\big). \tag{46}$$

Thus, the maximal eigenvalue is bounded above by $\max(0,\tilde{u}_i^{\max})$ if $\hat{g}_{i,d}\leqslant 0$. The other maximum principles of (44) can be proved similarly.

Obviously, the proofs are also valid in the case of the scalar limiting operators defined in Secs. 3 and 4. Thus, the requirements $\mathbf{G}_i \succcurlyeq \mathbf{0}$ and $\mathbf{G}_i \preccurlyeq \mathbf{0}$ of (34) are sufficient, but not necessary, and can be replaced by $\hat{g}_{i,d}\geqslant 0$ and $\hat{g}_{i,1}\leqslant 0$.

Using (43), it is also possible to derive the global maximum principles

$$\hat{g}_{i,d}\leqslant 0 \quad\Longrightarrow\quad u_{i,d}\leqslant \max\Big(0,\max_{1\leqslant j\leqslant N}\big\{u_{j,d}:\hat{g}_{j,d}>0\big\}\Big) \qquad \text{for all } 1\leqslant i\leqslant N, \tag{47a}$$

$$\hat{g}_{i,1}\geqslant 0 \quad\Longrightarrow\quad u_{i,1}\geqslant \min\Big(0,\min_{1\leqslant j\leqslant N}\big\{u_{j,1}:\hat{g}_{j,1}<0\big\}\Big) \qquad \text{for all } 1\leqslant i\leqslant N, \tag{47b}$$

$$\sum_j a_{ij}=0 \text{ and } \hat{g}_{i,d}\leqslant 0 \quad\Longrightarrow\quad u_{i,d}\leqslant \max_{1\leqslant j\leqslant N}\big\{u_{j,d}:\hat{g}_{j,d}>0 \text{ or } \textstyle\sum_l a_{jl}>0\big\} \quad \text{for all } 1\leqslant i\leqslant N, \tag{47c}$$

$$\sum_j a_{ij}=0 \text{ and } \hat{g}_{i,1}\geqslant 0 \quad\Longrightarrow\quad u_{i,1}\geqslant \min_{1\leqslant j\leqslant N}\big\{u_{j,1}:\hat{g}_{j,1}<0 \text{ or } \textstyle\sum_l a_{jl}>0\big\} \quad \text{for all } 1\leqslant i\leqslant N. \tag{47d}$$

If the set on the right-hand side is empty, there exists no meaningful maximum principle. However, corresponding estimates regarding strongly enforced (Dirichlet) boundary conditions as considered in the scalar case in [8] remain valid. The corresponding proof is very similar to the one presented below, which is inspired by [8, 15].

To prove (47a), let us focus on a fixed but arbitrary node i, such that $u_{i,d}$ is the global maximum of maximal eigenvalues and $\hat{g}_{i,d} \leqslant 0$. If there is no such node, we are done. Otherwise, we define

$$J \coloneqq \{1 \leqslant l \leqslant N : \hat{u}_{l,d} = u_{i,d} \text{ and } \hat{g}_{l,d} \leqslant 0\}.$$

In particular, we have $u_{l,d} = \hat{u}_{l,d} = u_{i,d}$ and $u_{l,d}$ also coincides with the global maximum for all $l \in J$ because $\mathbf{q}^\top \mathbf{V}\mathbf{q} \leqslant v_d$ for all normalized vectors $\mathbf{q} \in \mathbb{R}^d$ and $\mathbf{V} \in \mathbb{S}_d$. Hence, conditions (43a) are valid for all $l \in J$. Similarly to the derivation of (46), we choose $v = u_{i,d} = u_{l,d}$ (replace $\max(0, \tilde{u}_i^{\max})$ by $u_{l,d}$) and obtain

$$\begin{aligned}
0 \geqslant \hat{g}_{l,d} &\geqslant \big(\sum_j a_{lj}\big) u_{l,d} + \sum_{j\neq l, a_{lj}>0} \underbrace{(a_{lj} - d_{lj})}_{=:\hat{a}_{lj}} (\hat{u}_{j,d} - u_{l,d}) + \sum_{j\neq l, a_{lj}\leqslant 0} \underbrace{a_{lj}}_{=:\hat{a}_{lj}} (\hat{u}_{j,d} - u_{l,d}) \\
&= \big(\sum_j a_{lj} - \sum_{j\neq l} a_{lj} + \sum_{j\neq l, a_{lj}>0} d_{lj}\big) u_{l,d} + \sum_{j\neq l} \hat{a}_{lj}\hat{u}_{j,d} \\
&= \underbrace{\big(a_{ll} + \sum_{j\neq l, a_{lj}>0} d_{lj}\big)}_{=:\hat{a}_{ll}>0} u_{l,d} + \sum_{j\neq l} \hat{a}_{lj}\hat{u}_{j,d} = \sum_j \hat{a}_{lj}\hat{u}_{j,d} =: \hat{f}_{l,d} \qquad \text{for all } l \in J,
\end{aligned}$$

where (i) $\hat{a}_{lj} \leqslant \min(0, a_{lj})$ if $j \neq l$ and (ii) $\sum_j \hat{a}_{lj} = \sum_j a_{lj} \geqslant 0$ for all $l \in J$. Then, similarly to [15, proof of Theorem 5.2]

$$\text{there exists } \iota \in J \text{ such that } \quad \sum_{j\in J} \hat{a}_{\iota j} > 0 \tag{48}$$

due to the positive definiteness of A, (i), (ii), and

$$0 < C_M\Big(\sum_{l\in J} 1\Big) \leqslant \sum_{l,j\in J} a_{lj} = \sum_{l,j\in J} \hat{a}_{lj} + \sum_{l\in J, j\notin J} \underbrace{(\hat{a}_{lj} - a_{lj})}_{\leqslant 0} \leqslant \sum_{l,j\in J} \hat{a}_{lj}.$$

Finally, introducing $w \coloneqq \max\{\hat{u}_{j,d} : j \notin J\}$, we find that

$$u_{i,d} \sum_{j\in J} \hat{a}_{\iota j} = \sum_{j\in J} \hat{a}_{\iota j}\hat{u}_{j,d} = \hat{f}_{\iota,d} - \sum_{j\notin J} \hat{a}_{\iota j}\hat{u}_{j,d} \leqslant \hat{g}_{\iota,d} - w \sum_{j\notin J} \hat{a}_{\iota j} \leqslant \max(0, w) \sum_{j\in J} \hat{a}_{\iota j}.$$

Hence, $u_{i,d} \leqslant \max(0, w)$ and (47a) holds. The other inequalities of (47) can be shown in a similar vein.

Let us now show that requirements (43) are valid for the tensorial limiter defined by (37) and (41): If $v \geqslant u_{j,d}$ and $u_{i,d}$ is a local maximum (implies $\alpha_{ij,d} = 0$ due to (41)), conditions (43a) hold due to

$$\mathbf{q}_{i,d}^\top \mathbf{S}_{ij}(\mathbf{U}_i - v\mathbf{I})\mathbf{q}_{i,d} = \alpha_{ij,d}(u_{i,d} - v) = 0, \tag{49a}$$

$$\mathbf{q}_{i,d}^\top \mathbf{S}_{ji}(\mathbf{U}_i - v\mathbf{I})\mathbf{q}_{i,d} = \mathbf{q}_{i,d}^\top \mathbf{S}_{ji}\mathbf{q}_{i,d}(u_{i,d} - v) = \hat{\alpha}_{ji,d}(u_{i,d} - v), \tag{49b}$$

$$\mathbf{q}_{i,d}^\top \mathbf{S}_{ij}(\mathbf{U}_j - v\mathbf{I})\mathbf{q}_{i,d} = \alpha_{ij,d}\mathbf{q}_{i,d}^\top(\mathbf{U}_j - v\mathbf{I})\mathbf{q}_{i,d} = 0, \tag{49c}$$

$$\begin{aligned}
\mathbf{q}_{i,d}^\top \mathbf{S}_{ji}(\mathbf{U}_j - v\mathbf{I})\mathbf{q}_{i,d} &= \mathbf{q}_{i,d}^\top \mathbf{Q}_j \tilde{\mathbf{S}}_{ji} \mathbf{Q}_j^\top \mathbf{Q}_j (\tilde{\mathbf{U}}_j - v\mathbf{I}) \mathbf{Q}_j^\top \mathbf{q}_{i,d} = \sum_{k=1}^{d} \underbrace{(\mathbf{q}_{i,d}^\top \mathbf{q}_{j,k})^2}_{\geqslant 0} \underbrace{\alpha_{ji,k}}_{\in[0,1]} \underbrace{(u_{j,k} - v)}_{\leqslant 0} \\
&\geqslant \sum_{k=1}^{d} (\mathbf{q}_{i,d}^\top \mathbf{q}_{j,k})(u_{j,k} - v)(\mathbf{q}_{j,k}^\top \mathbf{q}_{i,d}) = \mathbf{q}_{i,d}^\top \mathbf{Q}_j (\tilde{\mathbf{U}}_j - v\mathbf{I}) \mathbf{Q}_j^\top \mathbf{q}_{i,d} = \hat{u}_{j,d} - v,
\end{aligned} \tag{49d}$$

where $\hat{\alpha}_{ji,d} \coloneqq \mathbf{q}_{i,d}^\top \mathbf{S}_{ji}\mathbf{q}_{i,d} \leqslant 1$ due to $\mathbf{S}_{ji} \preccurlyeq \mathbf{I}$. Similarly, conditions (43b) can be shown and therefore the tensorial limiting operator satisfies the LED properties.

6. Two-node tensorial limiting

After considering the two-node extension of the scalar limiter, an obvious way of extending the one-node tensor limiter is to set

$$\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] \coloneqq \tfrac{1}{2}\big((\mathbf{U}_i - \mathbf{U}_j)\min(\mathbf{S}_{ij}, \mathbf{S}_{ji}) + \min(\mathbf{S}_{ij}, \mathbf{S}_{ji})(\mathbf{U}_i - \mathbf{U}_j)\big)$$

if $a_{ij}, a_{ji} > 0$. While the Lipschitz continuity of this limiter can be shown by exploiting the basic ideas used before, the preservation of local maximum principles is not provable.

As an alternative, we consider a two-node limiting strategy which uses the product of tensorial correction factors. Since $\mathbf{U}_i$ and $\mathbf{S}_{ij}$, as well as $\mathbf{U}_j$ and $\mathbf{S}_{ji}$ are simultaneously diagonalizable, the matrix multiplications $\mathbf{U}_i\mathbf{S}_{ij}$ and $\mathbf{U}_j\mathbf{S}_{ji}$ are commutative and we can define

$$\begin{aligned}\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] &= \tfrac{1}{2}(\mathbf{S}_{ji}\mathbf{S}_{ij}\mathbf{U}_i - \mathbf{S}_{ij}\mathbf{S}_{ji}\mathbf{U}_j) + \tfrac{1}{2}(\mathbf{U}_i\mathbf{S}_{ij}\mathbf{S}_{ji} - \mathbf{U}_j\mathbf{S}_{ji}\mathbf{S}_{ij}) \\ &= \tfrac{1}{2}\mathbf{S}_{ij}(\mathbf{U}_i - \mathbf{U}_j)\mathbf{S}_{ji} + \tfrac{1}{2}\mathbf{S}_{ji}(\mathbf{U}_i - \mathbf{U}_j)\mathbf{S}_{ij}.\end{aligned} \tag{50}$$

Note that this definition coincides with (37) if $a_{ij} \leqslant 0$ or $a_{ji} \leqslant 0$ as in the scalar case. For the Lipschitz continuity (especially if $a_{ij}, a_{ji} > 0$), similarly to (40), we have

$$\begin{aligned}&\big\|\mathcal{A}_{ij}(U)[\mathbf{U}_i - \mathbf{U}_j] - \mathcal{A}_{ij}(\bar{U})[\bar{\mathbf{U}}_i - \bar{\mathbf{U}}_j]\big\|_{\mathrm{F}} \\ &\quad= \tfrac{1}{2}\big\|(\mathbf{S}_{ij}\mathbf{W}_{ij}\mathbf{S}_{ji} + \mathbf{S}_{ji}\mathbf{W}_{ij}\mathbf{S}_{ij}) - (\bar{\mathbf{S}}_{ij}\bar{\mathbf{W}}_{ij}\bar{\mathbf{S}}_{ji} + \bar{\mathbf{S}}_{ji}\bar{\mathbf{W}}_{ij}\bar{\mathbf{S}}_{ij})\big\|_{\mathrm{F}} \\ &\quad\leqslant \tfrac{1}{2}\big\|(\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij})\mathbf{W}_{ij}\mathbf{S}_{ji} + \mathbf{S}_{ji}\mathbf{W}_{ij}(\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij})\big\|_{\mathrm{F}} + \tfrac{1}{2}\big\|\bar{\mathbf{S}}_{ij}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij})\mathbf{S}_{ji} + \mathbf{S}_{ji}(\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij})\bar{\mathbf{S}}_{ij}\big\|_{\mathrm{F}} \\ &\qquad+ \tfrac{1}{2}\big\|\bar{\mathbf{S}}_{ij}\bar{\mathbf{W}}_{ij}(\mathbf{S}_{ji} - \bar{\mathbf{S}}_{ji}) + (\mathbf{S}_{ji} - \bar{\mathbf{S}}_{ji})\bar{\mathbf{W}}_{ij}\bar{\mathbf{S}}_{ij}\big\|_{\mathrm{F}} \\ &\quad\leqslant \|\mathbf{S}_{ij} - \bar{\mathbf{S}}_{ij}\|_2\|\mathbf{S}_{ji}\|_2\|\mathbf{W}_{ij}\|_{\mathrm{F}} + \|\bar{\mathbf{S}}_{ij}\|_2\|\mathbf{S}_{ji}\|_2\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \|\mathbf{S}_{ji} - \bar{\mathbf{S}}_{ji}\|_2\|\bar{\mathbf{S}}_{ij}\|_2\|\bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} \\ &\quad\leqslant 3\|\mathbf{W}_{ij} - \bar{\mathbf{W}}_{ij}\|_{\mathrm{F}} + \|\mathbf{A}_{ij} - \bar{\mathbf{A}}_{ij}\|_2 + \|\mathbf{A}_{ji} - \bar{\mathbf{A}}_{ji}\|_2 + |B_{ij} - \bar{B}_{ij}| + |B_{ji} - \bar{B}_{ji}|\end{aligned} \tag{51}$$

if $\mathbf{U}_i \neq \mathbf{U}_j$ and $\bar{\mathbf{U}}_i \neq \bar{\mathbf{U}}_j$. The other case can be handled as before. Therefore, the two-node tensorial limiter defined by (50) is Lipschitz continuous if $\mathbf{A}_{ij}$ and B_{ij} are Lipschitz continuous. Furthermore, the coercivity of the AFC problem can be shown as in (38)

$$\begin{aligned}\big(\mathbf{V}, (\mathcal{I} - \mathcal{A}_{ij})[\mathbf{V}]\big)_{\mathrm{F}} &= (\mathbf{V}, \mathbf{V})_{\mathrm{F}} - \tfrac{1}{2}(\mathbf{V}, \mathbf{S}_{ij}\mathbf{V}\mathbf{S}_{ji})_{\mathrm{F}} + \tfrac{1}{2}(\mathbf{V}, \mathbf{S}_{ji}\mathbf{V}\mathbf{S}_{ij})_{\mathrm{F}} = (\mathbf{V}, \mathbf{V})_{\mathrm{F}} - (\mathbf{V}, \mathbf{S}_{ij}\mathbf{V}\mathbf{S}_{ji})_{\mathrm{F}} \\ &\geqslant \|\mathbf{V}\|_{\mathrm{F}}^2 - \|\mathbf{V}\|_{\mathrm{F}}\|\mathbf{S}_{ij}\mathbf{V}\mathbf{S}_{ji}\|_{\mathrm{F}} \geqslant \|\mathbf{V}\|_{\mathrm{F}}^2 - \|\mathbf{S}_{ij}\|_2\|\mathbf{S}_{ji}\|_2\|\mathbf{V}\|_{\mathrm{F}}^2 \geqslant 0\end{aligned} \tag{52}$$

if $a_{ij}, a_{ji} > 0$ and a solution of the nonlinear discrete problem exists. The solution satisfies the relevant maximum principles for the range of eigenvalues, because the sufficient conditions (43) are fulfilled: For example, if $u_{i,d}$ is a local maximum of maximal eigenvalues and $a_{ij}, a_{ji} > 0$, we have $\alpha_{ij,d} = 0$ and

$$\begin{aligned}\mathbf{q}_{i,d}^\top\mathbf{S}_{ij}(\mathbf{U}_i - v\mathbf{I})\mathbf{S}_{ji}\mathbf{q}_{i,d} &= \alpha_{ij,d}(u_{i,d} - v)\mathbf{q}_{i,d}^\top\mathbf{S}_{ji}\mathbf{q}_{i,d} = 0, \\ \mathbf{q}_{i,d}^\top\mathbf{S}_{ij}(\mathbf{U}_j - v\mathbf{I})\mathbf{S}_{ji}\mathbf{q}_{i,d} &= \alpha_{ij,d}\mathbf{q}_{i,d}^\top(\mathbf{U}_j - v\mathbf{I})\mathbf{S}_{ji}\mathbf{q}_{i,d} = 0\end{aligned} \qquad \text{for all } v \in \mathbb{R}.$$

Therefore, (50) defines a reasonable extension of the one-node tensor limiter, which is Lipschitz continuous and satisfies maximum principles for the range of eigenvalues.

7. Dual limiting

So far, we have added artificial diffusion to the system matrix A to enforce sufficient conditions for local maximum principles and applied limited antidiffusive fluxes to reduce discretization errors while preserving the eigenvalue range. As we have seen, the optimal choice of the limiting strategy depends on the sign of the matrix

entries a_{ij} and a_{ji}. For instance, the limiting operator presented in the previous section identifies node i as an untroubled node and applies a one-sided limiter if $a_{ij} \leqslant 0$. In the case $a_{ij} > 0$ and $a_{ji} > 0$, a two-node limiter is invoked regardless of the magnitudes of a_{ij} and a_{ji}. While two-node limiting is certainly appropriate for symmetric operators ($a_{ij} = a_{ji} > 0$), it is hardly optimal in the case $0 < \min(a_{ij}, a_{ji}) \ll \max(a_{ij}, a_{ji})$ in which a large percentage of the antidiffusive flux could be handled using a one-node limiter without posing any hazard to the other node. Moreover, a different limiting strategy is selected whenever the sign of a_{ij} or a_{ji} changes. Hence, the outcome of composite limiting does not depend continuously on the coefficients of A (small variations of a_{ij} may lead to significant changes in the magnitude of the limited antidiffusive flux).

To avoid this drawback, we construct and limit the artificial diffusion operator using a decomposition of the system matrix A into two parts. The first one is a symmetric matrix with off-diagonal entries $a'_{ij} = \max(0, \min(a_{ij}, a_{ji}))$ and calls for the use of a two-node limiting strategy. The second one contains the remainder and satisfies condition (15) that is used in the proof of (local) maximum principles for one-node limiters. This splitting is based on the idea of 'prelimiting' in the context of upwind-biased AFC schemes for scalar transport equations [17, Section 7.4].

The dual limiting approach that we propose in this section is a generalization of prelimiting. The diffusion matrix $D' = (d'_{ij})_{i,j=1}^N$ that transforms A into a form suitable for one-node limiting is defined by

$$d'_{ij} \coloneqq \max\big(0, \min(a_{ij}, a_{ji})\big) \geqslant 0 \quad \text{for all } i \neq j, \qquad d'_{ii} \coloneqq -\sum_{j=1}^{N} d'_{ij} \leqslant 0. \tag{53}$$

This definition corresponds to the algebraic splitting $A = A' + A''$, where $A' = D'$ and $A'' = A - D'$ satisfies the one-node LED condition (15). To transform A'' into an M-matrix, we apply the artificial diffusion matrix $D'' = (d''_{ij})_{i,j=1}^N$ defined as in (11) in terms of the coefficients a''_{ij} and a''_{ji}. The off-diagonal entries of $A - D' - D''$ satisfy

$$a_{ij} - d'_{ij} - d''_{ij} \leqslant 0 \quad \text{and} \quad a_{ij} - d'_{ij} - d''_{ij} \leqslant 0 \qquad \text{for all } 1 \leqslant i, j \leqslant N,\ i \neq j.$$

In fact, the artificial diffusion $D = D' + D''$ is the same as the one that we used before. However, now we have the option of constraining the antidiffusive fluxes corresponding to D'' using a one-node limiter, while using a two-node limiter for fluxes corresponding to D'. Clearly, the so-defined dual limiting algorithm is less restrictive than the previously considered two-node and composite limiters. If A is symmetric or skew symmetric (leading to $D'' = 0$ or $D' = 0$) then the dual limiting approach will automatically select the optimal limiting strategy (one-node or two-node) for antidiffusive fluxes associated with the nontrivial part of D. In general, the prelimited AFC system reads

$$\sum_{j=1}^{N} a_{ij}\mathbf{U}_j + d'_{ij}(\mathcal{I} - \mathcal{A}'_{ij})[\mathbf{U}_i - \mathbf{U}_j] + d''_{ij}(\mathcal{I} - \mathcal{A}''_{ij})[\mathbf{U}_i - \mathbf{U}_j] = \mathbf{G}_i \qquad \text{for all } 1 \leqslant i \leqslant N, \tag{54}$$

where the different branches of the limiting operators $\mathcal{A}'_{ij}$ and $\mathcal{A}''_{ij}$ are selected depending on the entries of A and $A' = A - D'$, respectively. The LED property can be shown by extending the above proofs in a straightforward manner. The unsplit version of a given limiter for $D = D' + D''$ corresponds to the case $\mathcal{A}'_{ij} = \mathcal{A}''_{ij}$.

8. Numerical examples

In what follows, we analyze the proposed limiting techniques and illustrate their benefits and drawbacks using numerical studies for two stationary benchmarks: The first one deals with pure advection, which is optionally stabilized using the streamline upwind Petrov-Galerkin (SUPG); the second one represents an elliptic problem dominated by anisotropic diffusion. In both cases, the spatial domains are two-dimensional, while the dimension of the tensor quantity is 3×3.

The numerical solution to system (54) is marched to the steady state using the implicit Euler method with the lumped mass matrix. The resulting nonlinear system of each pseudo time step is solved numerically using

a fixed point iteration method in which the limited antidiffusive terms are calculated using the data from the previous iteration, i.e., for every $1 \leqslant s \leqslant S$ and for all $1 \leqslant i \leqslant N$ we solve

$$\begin{aligned} m_i \mathbf{U}_i^{n+1,s+1} + \Delta t \sum_{j=1}^{N} (a_{ij} - d_{ij}) \mathbf{U}_j^{n+1,s+1} \\ = m_i \mathbf{U}_i^{n,s} + \Delta t \mathbf{G}_i + \Delta t \sum_{j=1}^{N} \big(d'_{ij} \mathcal{A}'_{ij}(U^{n+1,s}) + d''_{ij} \mathcal{A}''_{ij}(U^{n+1,s}) \big) [\mathbf{U}_i^{n+1,s} - \mathbf{U}_j^{n+1,s}]. \end{aligned} \tag{55}$$

Here, $S \in \mathbb{N}$ is the total number of fixed point iterations, $\Delta t > 0$ is the pseudo time increment and m_i is the i-th diagonal entry of the lumped mass matrix. The default initial condition for the fixed point iteration is $U^{n+1,1} \coloneqq U^n$, where U^n is the solution from the last time step. Since we are just interested in the steady state solution, it is worthwhile to terminate the fixed point loop after one cycle, i.e., we set $S = 1$, and update the solution just once per time step without checking the convergence criteria. Thus, the number of pseudo time steps (or final time T) is an indicator for the degree of nonlinearity and Lipschitz constant of the method under investigation.

In the description of numerical experiments, we use the following abbreviations for the proposed methods/limiters

low order	$\widehat{=}$	Galerkin method with artificial diffusion (11)
Galerkin	$\widehat{=}$	Galerkin method (without stabilization)
SUPG	$\widehat{=}$	Galerkin method with SUPG stabilization
scalar limiter	$\widehat{=}$	AFC method (54) using limiter (35)
tensor limiter	$\widehat{=}$	AFC method (54) using limiter (50)

where the correction factors are calculated using (30) and (41) for the scalar and tensor limiter, respectively. If not mentioned otherwise, the parameters are chosen to be $p = q = 2$. In the case of an AFC method, the target scheme can be either the standard Galerkin method or SUPG (specified each time).

8.1. **Circular convection**

Let the tensorial extension of the stationary circular convection problem (cf. [12]) be given by

$$\begin{cases} \operatorname{div}(\mathbf{v}\mathbf{U}) = 0 & \text{in } \Omega = (0,1)^2, \\ \mathbf{U} = \mathbf{U}_{\text{in}} & \text{on } \Gamma_{\text{in}} = [0,1] \times \{0\} \cup \{1\} \times [0,1], \end{cases} \quad \text{where } \mathbf{v} = (-x_2, x_1)^\top \tag{56}$$

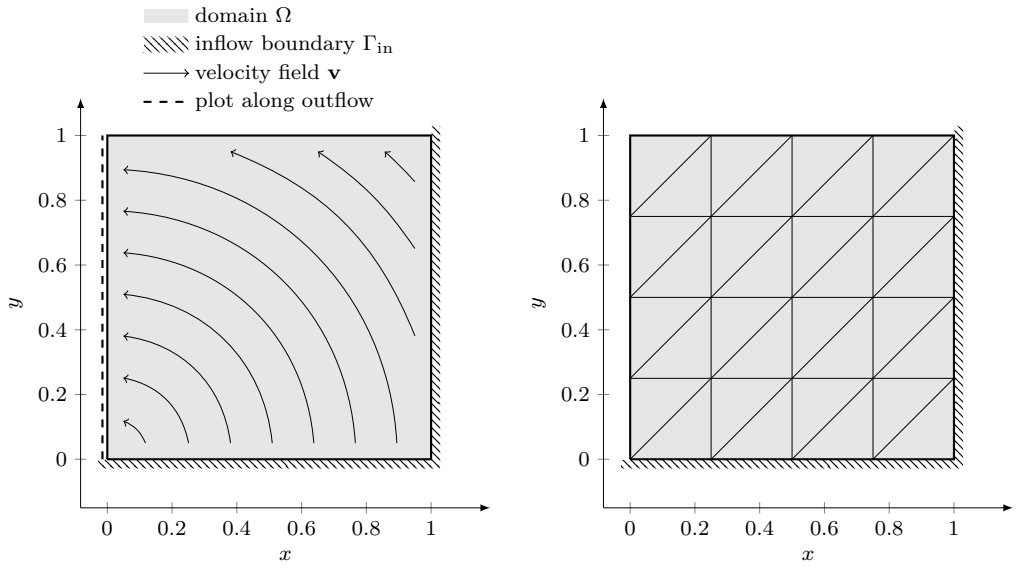
and the inflow boundary condition $\mathbf{U}_{\text{in}} : \Gamma_{\text{in}} \to \mathbb{S}_d$ be convected around the center of the vortex which is located at the lower left corner $\mathbf{x}^* = (0,0)^\top$ of the domain Ω (see Fig. 2a). Then, the exact solution $\mathbf{U} : \Omega \to \mathbb{S}_d$ depends only on the distance $r = \|\mathbf{x} - \mathbf{x}^*\|_2 = \|\mathbf{x}\|_2$ from the origin and is uniquely defined by the inflow boundary condition $\mathbf{U}_{\text{in}}$.

The Galerkin discretization including weakly enforced Dirichlet boundary conditions leads to the following system of equations for the nodal values $u_{j,k\ell}$ of tensor entries (cf. [20])

$$\sum_{j=1}^{N} a_{ij} u_{j,k\ell} = g_{i,k\ell} \qquad \text{for all } 1 \leqslant i \leqslant N,$$

$$a_{ij} \coloneqq -\int_\Omega \operatorname{grad}(\varphi_i) \cdot \mathbf{v} \varphi_j \, \mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} \mathbf{v} \cdot \mathbf{n} \varphi_i \varphi_j \, \mathrm{d}\mathbf{s}, \qquad g_{i,k\ell} \coloneqq -\int_{\Gamma_{\text{in}}} u_{\text{in},k\ell} \varphi_i \mathbf{v} \cdot \mathbf{n} \, \mathrm{d}\mathbf{s}, \tag{57}$$

where φ_j and φ_i are scalar trial and test functions of the continuous and piecewise linear finite element space $V_h = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\}$ with dimension N and $\Gamma_{\text{out}} \coloneqq \Gamma \setminus \Gamma_{\text{in}}$ is the outflow part of the boundary $\Gamma \coloneqq \partial\Omega$.



(A) Domain Ω with velocity field $\mathbf{v} = (-y, x)^\top$ and inflow boundary Γ_{in}.

(B) Mesh on level 2 (32 cells and 25 nodes).

FIGURE 2. Circular convection: Geometry of the domain with velocity field and illustration of uniform triangulation.

Thanks to the solenoidal velocity field, i.e., $\operatorname{div}(\mathbf{v}) = 0$, the system matrix $A \in \mathbb{R}^{N\times N}$ is positive semidefinite

$$\begin{aligned}\sum_{i,j=1}^{N} u_i a_{ij} u_j &= -\int_\Omega \operatorname{grad}(u_h)\cdot\mathbf{v}u_h \,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} u_h^2 \mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s}\\ &= -\tfrac{1}{2}\int_\Omega \operatorname{div}(\mathbf{v}u_h^2)\,\mathrm{d}\mathbf{x} + \int_{\Gamma_{\text{out}}} u_h^2\mathbf{v}\cdot\mathbf{n}\,\mathrm{d}\mathbf{s}\\ &= -\tfrac{1}{2}\int_{\Gamma_{\text{in}}} u_h^2 \underbrace{\mathbf{v}\cdot\mathbf{n}}_{\leqslant 0}\,\mathrm{d}\mathbf{s} + \tfrac{1}{2}\int_{\Gamma_{\text{out}}} u_h^2 \underbrace{\mathbf{v}\cdot\mathbf{n}}_{\geqslant 0}\,\mathrm{d}\mathbf{s} = \tfrac{1}{2}\int_\Gamma u_h^2|\mathbf{v}\cdot\mathbf{n}|\,\mathrm{d}\mathbf{s} \geqslant 0\end{aligned}$$

for all scalar finite element functions $u_h \in V_h$ with degrees of freedom u_i. Equality holds for all functions vanishing on the boundary.

The resulting discretization is unstable so that the solution can exhibit high-frequency oscillations inside the domain. This can be avoided, e.g., by adding streamline upwind Petrov Galerkin (SUPG) stabilization to the system matrix. The corresponding stabilization matrix $S = (s_{ij})_{i,j=1}^N$ is defined by [13, 28, 29]

$$s_{ij} \coloneqq \int_\Omega \tau(\mathbf{v}\cdot\operatorname{grad}\varphi_i)\operatorname{div}(\mathbf{v}\varphi_j)\,\mathrm{d}\mathbf{x}, \tag{58}$$

where the SUPG parameter $\tau \coloneqq h(2\|\mathbf{v}\|_2 + \varepsilon)^{-1} > 0$ depends on the diameter $h = \operatorname{diam}(K)$ of the element K and a small number $\varepsilon > 0$ to avoid divisions by zero. If $\operatorname{div}(\mathbf{v}) = 0$, the stabilization matrix defined in this way is symmetric, has vanishing row and column sums, and is positive semidefinite, because for any scalar finite

element function $u_h \in V_h$ we have

$$\sum_{i,j=1}^{N} u_i s_{ij} u_j = \int_\Omega \tau(\mathbf{v}\cdot\operatorname{grad} u_h)\operatorname{div}(\mathbf{v}u_h)\,\mathrm{d}\mathbf{x} = \int_\Omega \tau(\mathbf{v}\cdot\operatorname{grad} u_h)^2\,\mathrm{d}\mathbf{x} \geqslant 0.$$

Since the system matrix A is not positive definite in the case of stationary divergence-free advection (with and without SUPG stabilization), the above analysis is not directly applicable to our hyperbolic model problem. However, due to the use of a pseudo time stepping method, the contribution of the lumped mass matrix associated with the pseudo-transient makes A positive definite and, hence, guarantees the existence of a solution in each pseudo time step.

In this benchmark, we choose $\Delta t = 10^{-4}$ and abort the simulation if the pseudo time reaches 100 or the relative error between two consecutive approximations becomes smaller than 10^{-14}. The initial condition for the pseudo time stepping approach is given by a scaled identity tensor such that the trace corresponds to that of the inflow boundary condition, i.e., $\mathbf{U}_{\text{init}} = \frac{1}{3}\mathbf{I}$. If not mentioned otherwise, the computational domain is given by a uniform triangular mesh on level 7 (consisting of $2\cdot(2^7)^2 = 32\,768$ cells and $(2^8+1)^2 = 16\,641$ nodes; see Fig. 2b), which is optionally distorted after the refinement. In contrast to formulation (57), we use strongly enforced Dirichlet boundary conditions in the practical implementation.

8.1.1. *Discontinuous solution*

To begin with, the inflow boundary condition is chosen as a piecewise constant tensor quantity defined by

$$\mathbf{U} = \begin{cases} \mathbf{U}_1 & : 0 \leqslant r < \frac{1}{5}, \\ \mathbf{U}_2 & : \frac{1}{5} \leqslant r < \frac{2}{5}, \\ \mathbf{U}_3 & : \frac{2}{5} \leqslant r < \frac{3}{5}, \\ \mathbf{U}_4 & : \frac{3}{5} \leqslant r < \frac{4}{5}, \\ \mathbf{U}_5 & : \frac{4}{5} \leqslant r < 1, \end{cases}$$

where the constant parts and their eigenvalue decompositions are given by

$$\mathbf{U}_1 = \tfrac{1}{3}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

$$\mathbf{U}_2 = \tfrac{1}{75}\begin{pmatrix} 32 & 24 & 0 \\ 24 & 18 & 0 \\ 0 & 0 & 25 \end{pmatrix} = \tfrac{1}{5}\begin{pmatrix} 4 & 3 & 0 \\ 3 & -4 & 0 \\ 0 & 0 & 5 \end{pmatrix}\begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}\tfrac{1}{5}\begin{pmatrix} 4 & 3 & 0 \\ 3 & -4 & 0 \\ 0 & 0 & 5 \end{pmatrix},$$

$$\mathbf{U}_3 = \tfrac{1}{3}\begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} & 0 \\ 0 & 0 & 2 \end{pmatrix}\begin{pmatrix} 0 & 0 & 0 \\ 0 & \frac{2}{3} & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}\tfrac{1}{2}\begin{pmatrix} \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

$$\mathbf{U}_4 = \tfrac{1}{3}\begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \tfrac{1}{2}\begin{pmatrix} \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} & 0 \\ 0 & 0 & 2 \end{pmatrix}\begin{pmatrix} \frac{2}{3} & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & \frac{1}{3} \end{pmatrix}\tfrac{1}{2}\begin{pmatrix} \sqrt{2} & \sqrt{2} & 0 \\ \sqrt{2} & -\sqrt{2} & 0 \\ 0 & 0 & 2 \end{pmatrix},$$

$$\mathbf{U}_5 = \tfrac{1}{3}\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix} = \tfrac{\sqrt{6}}{6}\begin{pmatrix} \sqrt{2} & \sqrt{3} & 1 \\ \sqrt{2} & -\sqrt{3} & 1 \\ \sqrt{2} & 0 & -2 \end{pmatrix}\begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}\tfrac{\sqrt{6}}{6}\begin{pmatrix} \sqrt{2} & \sqrt{2} & \sqrt{2} \\ \sqrt{3} & -\sqrt{3} & 0 \\ 1 & 1 & -2 \end{pmatrix}.$$

These data are defined in such a way that the trace is identically 1 on the inflow boundary Γ_{in} and the eigenvalues are bounded by 0 and 1. Due to the solenoidal velocity field, these properties are preserved inside the domain

Ω. The discontinuities are chosen so as to demonstrate the ability of numerical methods to handle different scenarios: $\mathbf{U}_1$ is characterized by the fact that its minimal and maximal eigenvalues coincide. The discontinuous transition to $\mathbf{U}_2$ produces a tensor with distinct eigenvalues. At the discontinuity separating $\mathbf{U}_2$ and $\mathbf{U}_3$ the eigenvectors corresponding to the minimal and maximal eigenvalue are interchanged and slightly distorted, whereas the eigenvectors of the swapped eigenvalues of $\mathbf{U}_3$ and $\mathbf{U}_4$ are exactly the same. In fact, the structure of the jumps between $\mathbf{U}_2$ and $\mathbf{U}_4$ corresponds to a two-dimensional tensor problem because $u_{1,ij} = u_{2,ij} = u_{3,ij}$ if $i = 3$ or $j = 3$ and the eigenvalue corresponding to the eigenvector $(0, 0, 1)^\top$ is located between the other eigenvalues (this eigenvalue does not affect the upper and lower bounds for correction factors). Finally, $\mathbf{U}_4$ and $\mathbf{U}_5$ possess the same eigenvector $(1, -1, 0)^\top$ corresponding to the minimal eigenvalue 0 while the remainder differs.

Figure 3 shows the minimal and maximal eigenvalues produced by different methods on the uniform mesh without using any limiter. As already mentioned, the Galerkin method is highly unstable and the convergence to the exact solution cannot be guaranteed. In this case, the solution exhibits high-frequency oscillations, which violate local maximum principles for the range of eigenvalues. These oscillations are damped and clustered around the discontinuities if SUPG stabilization is applied (see Fig. 4a). The low order counterpart of the Galerkin method (that is the scheme derived by adding an artificial diffusion operator D to the stiffness matrix A of the Galerkin discretization) generates enormous amounts of artificial diffusion. Local maximum principles for the range of eigenvalues are satisfied and no oscillations appear in the solution. Along the streamline, the minimal and maximal eigenvalues move towards the intermediate eigenvalue while the trace is preserved. For $r < 0.8$, there is no discontinuity in the intermediate eigenvalue and the corresponding eigenvector. Thus, the artificial diffusion does not affect this part of the tensor field and the maximal eigenvalue decreases in the same way as the minimal eigenvalue increases due to the constant trace (see Fig. 4a).

In Fig. 4b, the behavior of the proposed limiting techniques is illustrated. If the Galerkin method is used as the AFC target, the system matrix A is skew symmetric in the interior of the domain (due to divergence free velocity field $\mathbf{v}$), $D' = 0$, and there is no need for using a two-node limiter. Both limiters produce nearly the same results at the first three discontinuities ($r < 0.7$), while the tensorial approach tends to be less restrictive. At the last discontinuity, where the eigenvector of the minimal eigenvalue does not change, scalar limiting is less accurate and the solution is comparable to the low order approximation (see Fig. 4a). This loss of accuracy occurs, because there is no discontinuity in the minimal eigenvalue (and the corresponding eigenvector) and artificial diffusion only affects the other two eigenvalues. Thus, the minimal eigenvalue stays constant, the corresponding correction factor is zero, and the synchronization of the correction factors corresponding to the minimal and maximal eigenvalue (cf. (26) and (30)) results in a vanishing antidiffusive flux.

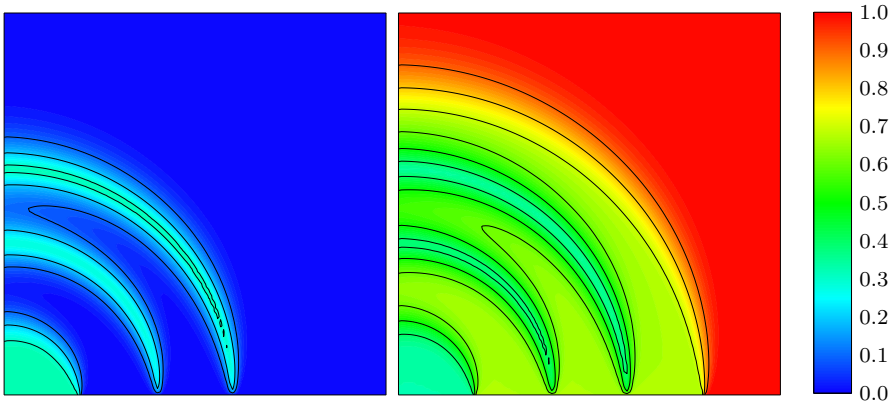
On the other hand, scalar limiting has the benefit that the intermediate eigenvalue is controlled implicitly, too, due to the preservation of the constant trace and maximum principles for the other eigenvalues (three conditions for three eigenvalues). Therefore, the intermediate eigenvalue does not oscillate and stays stable. In contrast to this, the tensor limiter ignores the trace so that high-frequency oscillations can occur in the intermediate eigenvalue without violating maximum principles for the range of eigenvalues.

If the target method is stabilized using SUPG, a symmetric, positive semidefinite operator is added to the system matrix and a two-node limiter is mandatory for an LED method. As a result, spurious oscillations of the intermediate eigenvalue vanish if tensorial limiting is applied. Additionally, variations of the trace are reduced, but not removed completely (see Tab. 1) and the method is still incapable for enforcing the trace preservation.

Table 1 summarizes the most important data of this example on the uniform and distorted meshes (disturbed on final level 7). Here, e.g., $L^1-\|\cdot\|_{\mathrm{F}}$ denotes the L^1 norm of the Frobenius error defined as a function depending on $\mathbf{x}$. As expected, the tensorial limiting technique (50) is not able to keep the trace constant. However, due to fewer constraints, it produces the most accurate results. The behavior of the methods under investigation remains unchanged if the mesh is distorted.

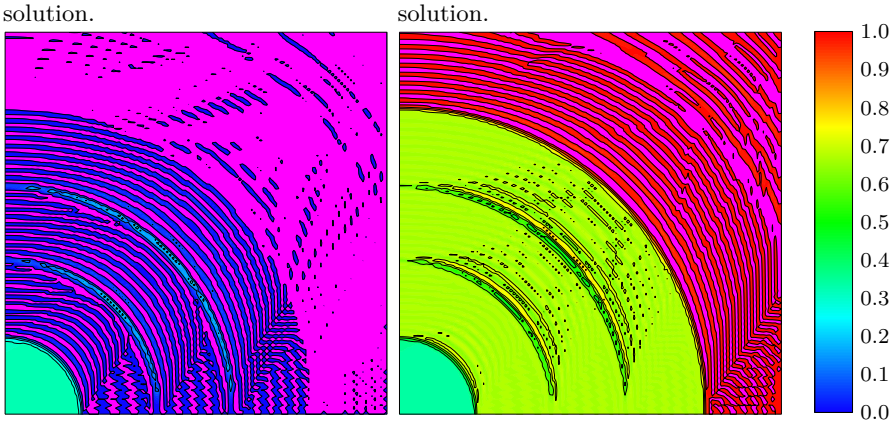
8.1.2. *Smooth solution*

To analyze the accuracy and convergence behavior of the proposed AFC methods with respect to variations of parameters and mesh refinement, we replace the discontinuous inflow boundary condition by the smooth



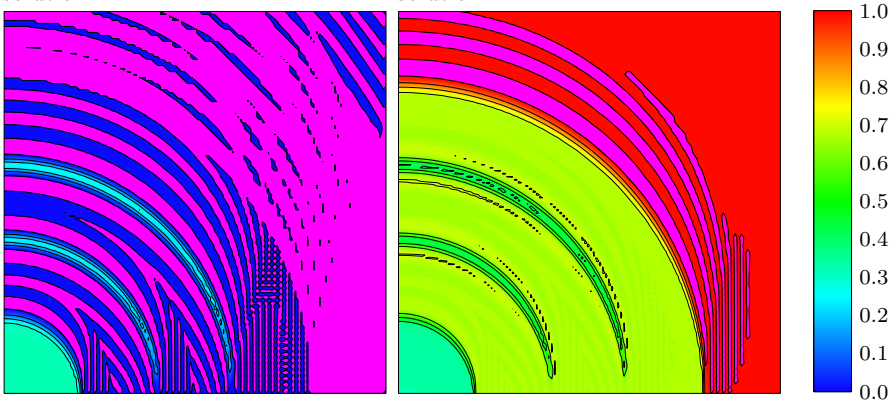
(A) Minimal eigenvalue of low order solution.

(B) Maximal eigenvalue of low order solution.



(C) Minimal eigenvalue of Galerkin solution.

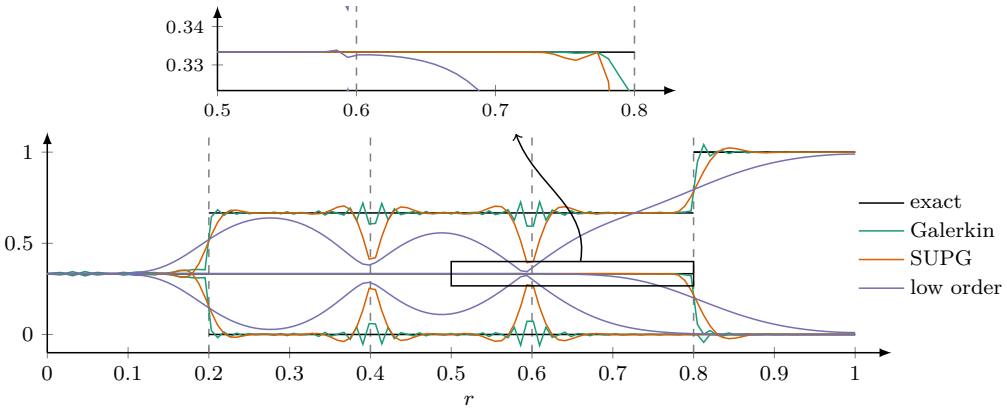
(D) Maximal eigenvalue of Galerkin solution.



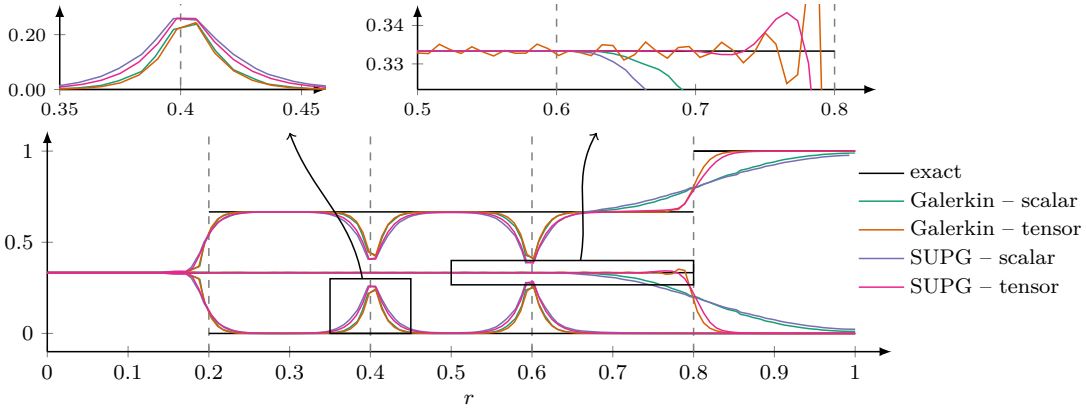
(E) Minimal eigenvalue of SUPG solution.

(F) Maximal eigenvalue of SUPG solution.

FIGURE 3. Circular convection (discontinuous test): Eigenvalue range of the low order and Galerkin solutions on the uniform level 7 mesh. Overshoots and undershoots are plotted in magenta.



(A) Solutions without limiting.



(B) Limited Galerkin and SUPG solutions.

FIGURE 4. Circular convection (discontinuous test): Cutline $x_1 = 0$ profiles (cf. Fig. 2a) and zooms of eigenvalues corresponding to different numerical solutions on the uniform level 7 mesh.

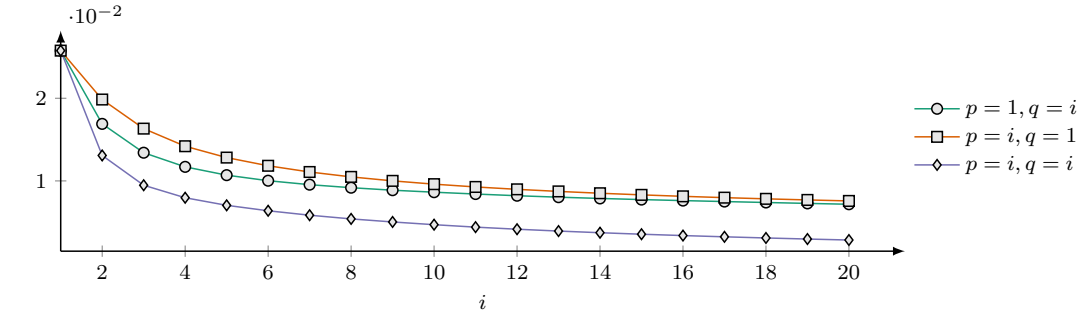
function

$$\mathbf{U} = \begin{pmatrix} \sin\tilde{r} & \cos\tilde{r} & 0 \\ \cos\tilde{r} & -\sin\tilde{r} & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sin\tilde{r} & 0 & 0 \\ 0 & 1-\sin\tilde{r} & 0 \\ 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} \sin\tilde{r} & \cos\tilde{r} & 0 \\ \cos\tilde{r} & -\sin\tilde{r} & 0 \\ 0 & 0 & 1 \end{pmatrix} \qquad \tilde{r} := \tfrac{3}{4}\pi r.$$

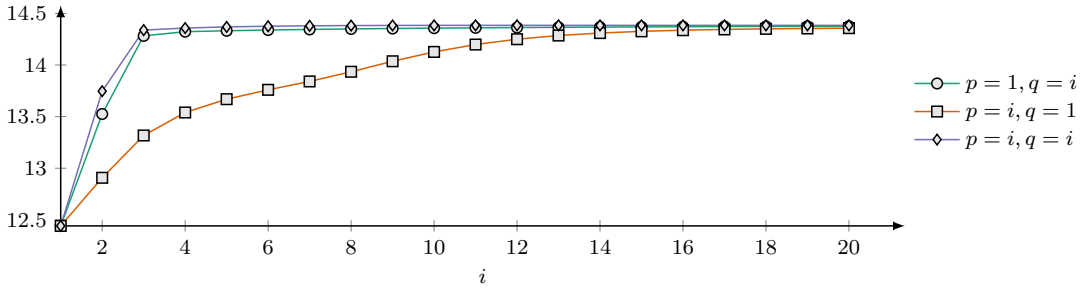
As the mesh is refined, the SUPG approximation exhibits second-order convergence to the exact solution on uniform and distorted meshes alike (see Tab. 2). While the unstabilized Galerkin method achieves the same behavior on the uniform mesh, the order of convergence decreases to approximately 1.3 on the distorted mesh for the considered L^1 integral of the Frobenius error norm. Note that this is not obvious, because there is no guarantee that the Galerkin approximation converges at all. To discuss the convergence behavior of the proposed AFC methods, the tensor limiter (50) using different parameters p and q is considered as an example. This limiting technique is not able to achieve the same convergence behavior as the high order methods and falls back to first order of accuracy if the parameters are too small with respect to the mesh size. As we increase p and q, the order of convergence of the AFC solution approaches that of the Galerkin/SUPG solution. However, mesh refinement requires simultaneous adjustment of the parameters. In contrast to [7], relatively large values must be used to achieve the optimal convergence behavior, because there exists no lower bound for q which guarantees that the method is linearity preserving.

TABLE 1. Circular convection (discontinuous test): Errors of different numerical solutions on the uniform and distorted level 7 mesh.

	method	$L^1 - \|\cdot\|_F$	$L^2 - \|\cdot\|_F$	$L^1 - \|\cdot\|_2$	$L^2 - \|\cdot\|_2$	$\mathrm{tr} - \|\cdot\|_\infty$
Galerkin uniform	low order	9.79e−2	1.56e−1	6.97e−2	1.10e−1	0.00
	high order	3.40e−2	6.81e−2	2.48e−2	4.88e−2	0.00
	scalar	5.68e−2	1.11e−1	4.02e−2	7.82e−2	0.00
	tensor	2.47e−2	7.62e−2	1.79e−2	5.46e−2	5.76e−2
Galerkin distorted	low order	9.86e−2	1.56e−1	7.02e−2	1.11e−1	0.00
	high order	6.50e−2	9.28e−2	4.90e−2	6.85e−2	0.00
	scalar	5.80e−2	1.12e−1	4.10e−2	7.90e−2	0.00
	tensor	2.68e−2	7.79e−2	1.96e−2	5.60e−2	5.08e−2
SUPG uniform	low order	1.11e−1	1.68e−1	7.94e−2	1.20e−1	0.00
	high order	2.69e−2	7.72e−2	1.90e−2	5.46e−2	0.00
	scalar	6.80e−2	1.22e−1	4.82e−2	8.62e−2	0.00
	tensor	3.40e−2	8.84e−2	2.47e−2	6.33e−2	3.08e−2
SUPG distorted	low order	1.15e−1	1.72e−1	8.26e−2	1.23e−1	0.00
	high order	2.82e−2	7.99e−2	1.99e−2	5.65e−2	0.00
	scalar	7.23e−2	1.26e−1	5.13e−2	8.91e−2	0.00
	tensor	3.82e−2	9.37e−2	2.81e−2	6.76e−2	4.99e−2



(A) $L^1 - \|\cdot\|_F$ error



(B) final pseudo time T

FIGURE 5. Circular convection (smooth test): Influence of the parameters p and q on (A) $L^1 - \|\cdot\|_F$ error and (B) final pseudo time for the SUPG method limited by the tensor limiter on the distorted level 7 mesh.

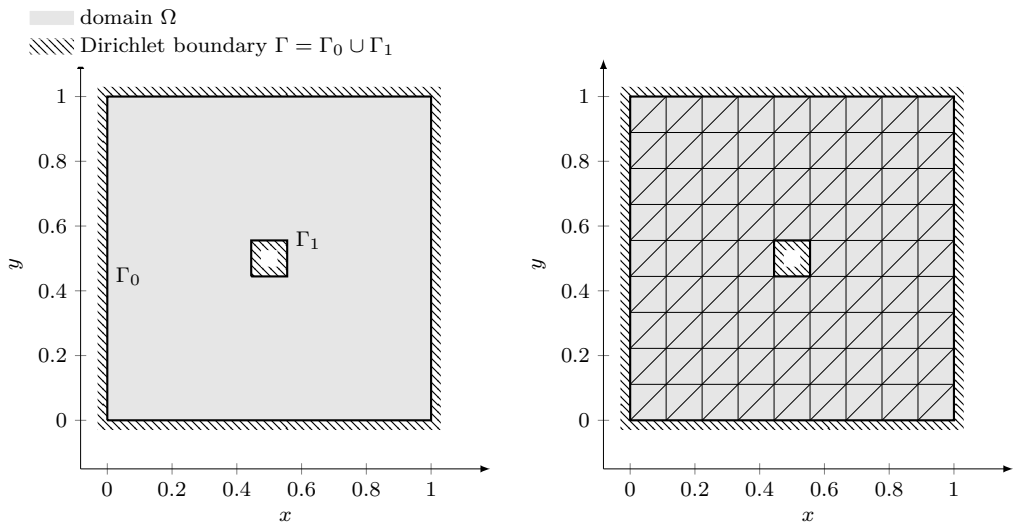
TABLE 2. Circular convection (smooth test): Convergence of the L^1 integral of the Frobenius error norm ($L^1 - \|\cdot\|_F$) for different numerical solutions with Galerkin and SUPG as AFC target on the uniform and distorted mesh.

	mesh level	uniform mesh				distorted mesh			
		Galerkin		SUPG		Galerkin		SUPG	
		error	EOC	error	EOC	error	EOC	error	EOC
high order	3	4.69e−2		1.06e−1		7.41e−2		1.11e−1	
	4	1.14e−2	2.04	2.01e−2	2.40	2.09e−2	1.83	2.26e−2	2.30
	5	2.88e−3	1.98	3.67e−3	2.45	8.76e−3	1.25	4.64e−3	2.28
	6	7.31e−4	1.98	7.48e−4	2.29	2.65e−3	1.73	9.80e−4	2.24
	7	1.85e−4	1.99	1.66e−4	2.17	1.18e−3	1.17	2.25e−4	2.13
low order	3	2.72e−1		2.90e−1		2.78e−1		2.97e−1	
	4	1.89e−1	0.52	2.22e−1	0.39	1.95e−1	0.51	2.32e−1	0.36
	5	1.18e−1	0.68	1.50e−1	0.56	1.21e−1	0.69	1.60e−1	0.54
	6	6.81e−2	0.80	9.10e−2	0.73	6.91e−2	0.81	9.73e−2	0.72
	7	3.69e−2	0.88	5.08e−2	0.84	3.75e−2	0.88	5.48e−2	0.83
tensor $p = q = 2$	3	1.40e−1		1.93e−1		1.42e−1		2.00e−1	
	4	6.50e−2	1.11	9.76e−2	0.98	7.02e−2	1.02	1.07e−1	0.90
	5	3.17e−2	1.04	4.86e−2	1.00	3.45e−2	1.03	5.48e−2	0.96
	6	1.53e−2	1.05	2.38e−2	1.03	1.68e−2	1.04	2.67e−2	1.04
	7	7.53e−3	1.03	1.16e−2	1.04	8.10e−3	1.05	1.31e−2	1.03
tensor $p = q = 5$	3	8.95e−2		1.41e−1		1.04e−1		1.49e−1	
	4	3.56e−2	1.33	5.23e−2	1.43	4.22e−2	1.30	5.97e−2	1.31
	5	1.69e−2	1.08	2.43e−2	1.11	2.13e−2	0.99	2.77e−2	1.11
	6	8.66e−3	0.96	1.25e−2	0.96	1.02e−2	1.06	1.38e−2	1.01
	7	4.44e−3	0.96	6.37e−3	0.97	5.37e−3	0.92	7.05e−3	0.97
tensor $p = q = 10$	3	6.99e−2		1.22e−1		9.28e−2		1.28e−1	
	4	2.71e−2	1.37	3.45e−2	1.82	3.12e−2	1.57	4.01e−2	1.67
	5	1.09e−2	1.31	1.43e−2	1.27	1.46e−2	1.10	1.66e−2	1.27
	6	5.64e−3	0.95	7.92e−3	0.85	7.26e−3	1.00	8.56e−3	0.96
	7	3.00e−3	0.91	4.29e−3	0.89	3.92e−3	0.89	4.70e−3	0.86
tensor $p = q = 50$	3	6.81e−2		1.17e−1		9.12e−2		1.23e−1	
	4	2.32e−2	1.56	2.38e−2	2.30	2.75e−2	1.73	2.86e−2	2.11
	5	7.03e−3	1.72	5.55e−3	2.10	1.19e−2	1.20	7.43e−3	1.94
	6	2.65e−3	1.41	2.33e−3	1.25	4.89e−3	1.29	2.97e−3	1.32
	7	1.20e−3	1.14	1.19e−3	0.97	2.49e−3	0.97	1.40e−3	1.09

An increase in q produces larger correction factors and has a stronger positive impact on the accuracy of numerical solutions than an increase in p (see Fig. 5), but requires a larger number of pseudo time steps to obtain a converged solution. Even in the scalar case, there is a direct relationship between the amount of numerical diffusion and numbers of outer iterations/pseudo time steps.

8.2. **Anisotropic diffusion**

The second benchmark is a tensorial extension of the anisotropic diffusion problem considered by Lipnikov et al. [19]. In this example, a two-node limiting technique must be used to preserve maximum principles. The computational domain is the unit square with a rectangular hole in the middle (cf. Fig. 6a). Two different constant boundary conditions are imposed on the outer and inner boundary Γ_0 and Γ_1, respectively. The exact solution varies between both values in a smooth and monotone way. The proposed tensor version of this



(A) Domain Ω with Dirichlet boundary Γ. (B) Mesh on level 0 (160 cells and 100 nodes).

FIGURE 6. Anisotropic diffusion: Geometry of the domain and illustration of uniform triangulation.

benchmark reads

$$\begin{cases} -\operatorname{div}(\mathbf{D}\operatorname{grad}\mathbf{U}) = 0 & \text{in } \Omega = (0,1)^2 \setminus [\frac{4}{9},\frac{5}{9}], \\ \mathbf{U} = \mathbf{U}_5 & \text{on } \Gamma_0 = \{0,1\}\times[0,1]\cup[0,1]\times\{0,1\}, \\ \mathbf{U} = \mathbf{U}_3 & \text{on } \Gamma_1 = \{\frac{4}{9},\frac{5}{9}\}\times[\frac{4}{9},\frac{5}{9}]\cup[\frac{4}{9},\frac{5}{9}]\times\{\frac{4}{9},\frac{5}{9}\}, \end{cases} \tag{59}$$

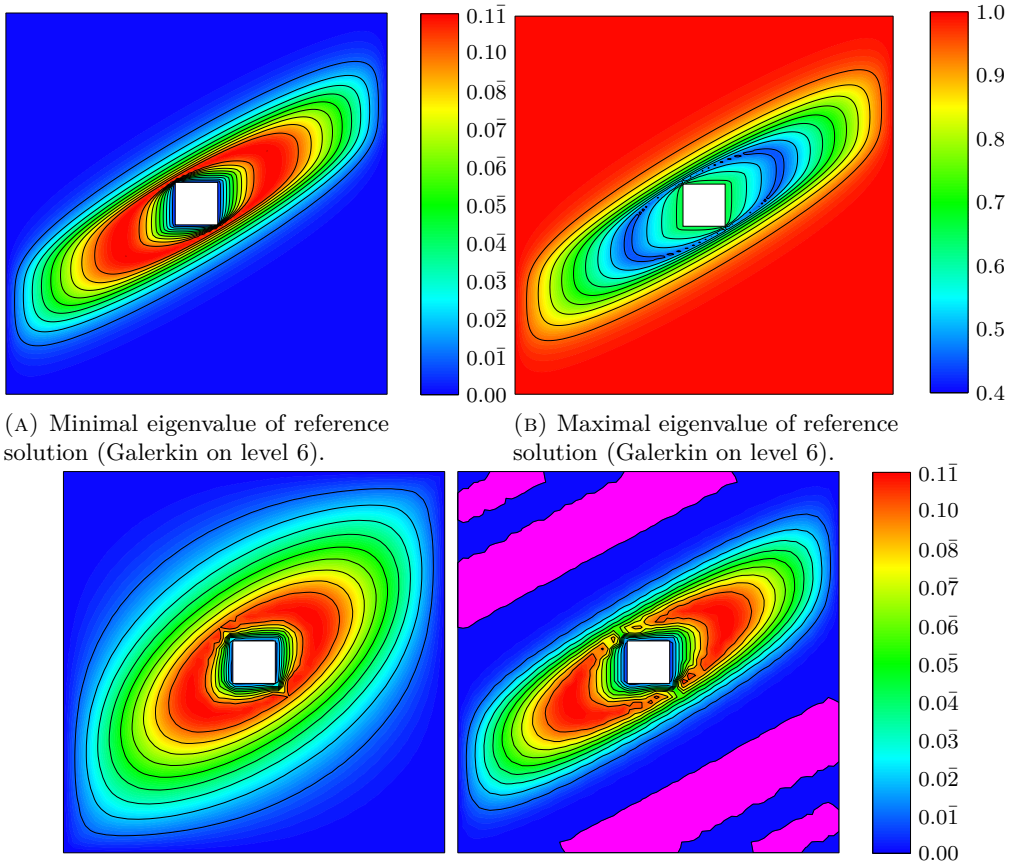
where the anisotropic diffusion operator $\mathbf{D}$ is given by

$$\mathbf{D} \coloneqq \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix} \begin{pmatrix} 100 & 0 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{pmatrix}, \qquad \theta = \frac{\pi}{6}$$

and $\mathbf{U}_5$ and $\mathbf{U}_3$ are defined as before (see Fig. 7a and 7b). The domain is discretized using a uniform mesh of $2\cdot(9\cdot 2^2)^2 - 2\cdot(2^2)^2 = 2560$ triangles and $(9\cdot 2^2+1)^2 - (2^2-1)^2 = 1360$ nodes (level 2; see Fig. 6b). Using linear finite elements (and a continuous Galerkin method as before), the eigenvalues of the unconstrained Galerkin solution violate the analytic bounds (see Fig. 7d). Optical irregularities are due to visualization effects on the coarse mesh with mesh size $h = \frac{1}{32}$. While the corresponding low order method reproduces the solution very accurately along the diagonal where $x_1 = x_2$, the solution is smoothed along the orthogonal direction.

In this case, the scalar and tensorial limiting techniques produce comparable results with respect to their accuracy (see Fig. 8). While the AFC solution using the tensor limiter with parameters $p = q = 2$ exhibits some artifacts on the diagonal $x_1 = x_2$, the minimal eigenvalue of the scalar limiting approximation has a pronounced local extremum. As we increase the values of the limiting parameters p and q, both approximations approach the Galerkin solution without generating undershoots on any mesh. While the peak in the minimal eigenvalue of the scalar limiting solution for $p = q = 5$ stays more pronounced than the one of the Galerkin approximation, the opposite is the case when the tensor limiter is applied.

The AFC approach using the scalar limiting technique produces very accurate results for problem (59), but it falls back to the low order method if the inner boundary condition is replaced by $\mathbf{U}_4$. In this case, the inner and outer boundary condition possess the same minimal eigenvalue and the same corresponding eigenvector. Hence, the scalar correction factors are set to zero (cf. Sec. 8.1.1).



(A) Minimal eigenvalue of reference solution (Galerkin on level 6).

(B) Maximal eigenvalue of reference solution (Galerkin on level 6).

(C) Minimal eigenvalue of low order level 2 solution.

(D) Minimal eigenvalue of Galerkin level 2 solution; minimum value is −3.83e−5.

FIGURE 7. Anisotropic diffusion: Eigenvalues of a reference solution and minimal eigenvalue of the low order and Galerkin solution on the uniform level 2 mesh. Overshoots and undershoots are plotted in magenta.

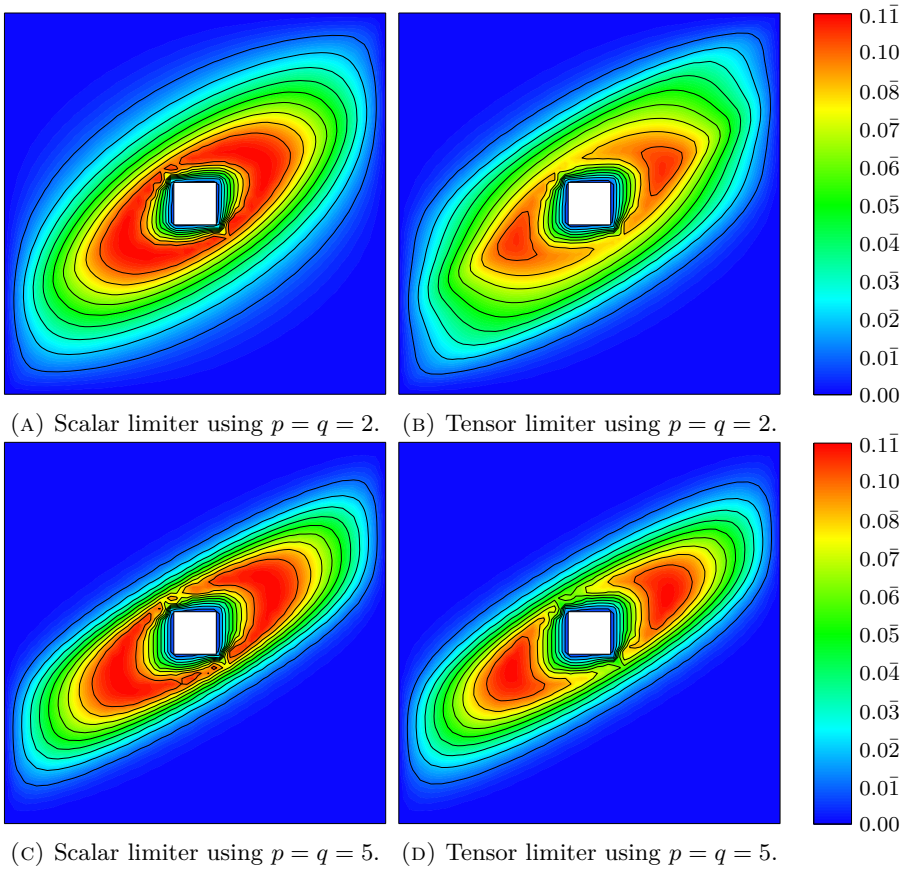
9. Conclusion

In this paper, we proposed nonlinear algebraic flux correction schemes for constraining finite element approximations to symmetric tensor quantities in a manner which guarantees that the range of eigenvalues satisfies customized maximum principles. Starting with a simple scalar approach, a more robust extension based on tensorial correction factors was derived.

The proposed algorithms are applicable to general positive definite system matrices and theoretically supported by proofs of local and global maximum principles and their Lipschitz continuity.

Numerical experiments show that scalar correction factors can be too restrictive due to risky synchronizations of the correction factors corresponding to the minimal and maximal eigenvalue. Tensorial limiters overcome this drawback by introducing individually chosen correction factors for each eigenvalue and, thus, tend to produce better results. However, currently the scalar limiting procedure is the only approach, which also preserves a constant trace by definition.

Unfortunately, there seems to be no way to define the involved parameters of the limiters a priori so that the AFC method is linearity preserving and a high order of convergence can be guaranteed on general meshes.



(A) Scalar limiter using $p = q = 2$. (B) Tensor limiter using $p = q = 2$.

(C) Scalar limiter using $p = q = 5$. (D) Tensor limiter using $p = q = 5$.

FIGURE 8. Anisotropic diffusion: Minimal eigenvalues corresponding to different AFC solutions on the uniform level 2 mesh.

Further work is required to resolve this issue. Additionally, appropriate solvers should be designed to reduce the number of costly fixed point iterations in each pseudo time step.

Acknowledgments

The author would like to thank Dmitri Kuzmin (TU Dortmund University) for fruitful discussions on the definition of promising limiting techniques.

The research was sponsored by the German Research Association (DFG) under grant KU 1530/13-1.

REFERENCES

[1] R. Abgrall. High order schemes for hyperbolic problems using globally continuous approximation and avoiding mass matrices. *Journal of Scientific Computing*, 73(2-3):461–494, 2017.

[2] S. G. Advani and C. L. Tucker III. The use of tensors to describe and predict fiber orientation in short fiber composites. *Journal of Rheology*, 31(8):751–784, 1987.

[3] M. C. Altan and L. Tang. Orientation tensors in simple flows of dilute suspensions of non-Brownian rigid ellipsoids, comparison of analytical and approximate solutions. *Rheologica Acta*, 32(3):227–244, 1993.

[4] S. Badia and J. Bonilla. Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Computer Methods in Applied Mechanics and Engineering*, 313:133 – 158, 2017. ISSN 0045-7825. doi: https://doi.org/10.1016/j.cma.2016.09.035. URL http://www.sciencedirect.com/science/article/pii/S0045782516306405.

[5] G. R. Barrenechea, V. John, and P. Knobloch. Analysis of algebraic flux correction schemes. *SIAM Journal on Numerical Analysis*, 54(4):2427–2451, 2016. doi: 10.1137/15M1018216. URL https://doi.org/10.1137/15M1018216.

[6] 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. *Numerische Mathematik*, 135(2):521–545, Feb 2017. ISSN 0945-3245. doi: 10.1007/s00211-016-0808-z. URL https://doi.org/10.1007/s00211-016-0808-z.

[7] G. R. Barrenechea, V. John, and P. Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Mathematical Models and Methods in Applied Sciences*, 27(03):525–548, 2017. doi: 10.1142/S0218202517500087. URL http://www.worldscientific.com/doi/abs/10.1142/S0218202517500087.

[8] G. R. Barrenechea, V. John, P. Knobloch, and R. Rankin. A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *WIAS Preprint No. 2475*, 2018. doi: https://doi.org/10.20347/WIAS.PREPRINT.2475. URL http://www.wias-berlin.de/publications/wias-publ/run.jsp?template=abstract&type=Preprint&year=2018&number=2475.

[9] J. Boris and D. Book. I. SHASTA, a fluid transport algorithm that works. *J. Comput. Phys.*, 11:38–69, 1973.

[10] B. Burgeth, A. Bruhn, S. Didas, J. Weickert, and M. Welk. Morphology for matrix data: Ordering versus pde-based approach. *Image and Vision Computing*, 25(4):496–511, 2007.

[11] D. Gilbarg and N. S. Trudinger. *Elliptic partial differential equations of second order*. springer, 2015.

[12] M. Hubbard. Non-oscillatory third order fluctuation splitting schemes for steady scalar conservation laws. *Journal of Computational Physics*, 222(2):740 – 768, 2007. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2006.08.007. URL http://www.sciencedirect.com/science/article/pii/S0021999106003937.

[13] V. John and P. Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I–A review. *Computer Methods in Applied Mechanics and Engineering*, 196(17):2197–2215, 2007.

[14] M. Klíma, M. Kuchařík, M. J. Shashkov, and J. Velechovský. Bound-Preserving Reconstruction of Tensor Quantities for Remap in ALE Fluid Dynamics. Technical report, Los Alamos National Laboratory (LANL), 2017. LA-UR-17-20068, Proceedings of XVI International Conference on Hyperbolic Problems Theory, Numerics and Applications, Aachen (Germany), Aug. 1-5, 2016.

[15] P. Knobloch. Numerical solution of convection–diffusion equations using a nonlinear method of upwind type. *Journal of Scientific Computing*, 43(3):454–470, Jun 2010. ISSN 1573-7691. doi: 10.1007/s10915-008-9260-2. URL https://doi.org/10.1007/s10915-008-9260-2.

[16] D. Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. *Computational Methods for Coupled Problems in Science and Engineering II, CIMNE, Barcelona*, pages 653–656, 2007.

[17] D. Kuzmin. Algebraic flux correction I. Scalar conservation laws. In D. Kuzmin, R. Löhner, and S. Turek, editors, *Flux-Corrected Transport*, Scientific Computation, pages 145–192. Springer Netherlands, 2012. ISBN 978-94-007-4037-2. URL http://link.springer.com/chapter/10.1007/978-94-007-4038-9_6.

[18] D. Kuzmin, S. Basting, and J. N. Shadid. Linearity-preserving monotone local projection stabilization schemes for continuous finite elements. *Computer Methods in Applied Mechanics and Engineering*, 322:23–41, 2017.

[19] K. Lipnikov, M. Shashkov, D. Svyatskiy, and Y. Vassilevski. Monotone finite volume schemes for diffusion equations on unstructured triangular and shape-regular polygonal meshes. *Journal of Computational Physics*, 227(1):492–512, 2007.

[20] C. Lohmann. Flux-corrected transport algorithms preserving the eigenvalue range of symmetric tensor quantities. *Journal of Computational Physics*, 350(Supplement C):907 – 926, 2017. ISSN 0021-9991. doi: https://doi.org/10.1016/j.jcp.2017.09.009. URL http://www.sciencedirect.com/science/article/pii/S002199911730668X.

[21] K. Löwner. Über monotone Matrixfunktionen. *Mathematische Zeitschrift*, 38(1):177–216, 1934. URL http://eudml.org/doc/168495.

[22] G. Luttwak. On the Extension of Monotonicity to Multi-Dimensional Flows. 2016.

[23] G. Luttwak and J. Falcovitz. Vector Image Polygon (VIP) limiters in ALE Hydrodynamics. In *EPJ Web of Conferences*, volume 10, page 00020. EDP Sciences, 2010.

[24] G. Luttwak and J. Falcovitz. Slope limiting for vectors: A novel vector limiting algorithm. *International Journal for Numerical Methods in Fluids*, 65(11-12):1365–1375, 2011.

[25] P.-H. Maire, R. Abgrall, J. Breil, R. Loubère, and B. Rebourcet. A nominally second-order cell-centered Lagrangian scheme for simulating elastic–plastic flows on two-dimensional unstructured grids. *Journal of Computational Physics*, 235:626–665, 2013.

[26] M. H. Protter and H. F. Weinberger. *Maximum principles in differential equations*. Springer Science & Business Media, 2012.

[27] S. K. Sambasivan, M. J. Shashkov, and D. E. Burton. Exploration of new limiter schemes for stress tensors in lagrangian and ALE hydrocodes. *Computers & Fluids*, 83:98–114, 2013.

[28] Y.-T. Shih and H. C. Elman. Modified streamline diffusion schemes for convection-diffusion problems. *Computer Methods in Applied Mechanics and Engineering*, 174(1):137 – 151, 1999. ISSN 0045-7825. doi: https://doi.org/10.1016/S0045-7825(98)00283-7. URL http://www.sciencedirect.com/science/article/pii/S0045782598002837.

[29] M. Stynes and L. Tobiska. Necessary L2-uniform convergence conditions for difference schemes for two-dimensional convection-diffusion problems. *Computers & Mathematics with Applications*, 29(4): 45 – 53, 1995. ISSN 0898-1221. doi: https://doi.org/10.1016/0898-1221(94)00237-F. URL http://www.sciencedirect.com/science/article/pii/089812219400237F.

[30] R. Temam. *Navier-Stokes equations*, volume 2. North-Holland Amsterdam, 1984.

[31] J. H. Wilkinson, editor. *The Algebraic Eigenvalue Problem*. Oxford University Press, Inc., New York, NY, USA, 1988. ISBN 0-198-53418-3.

[32] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *Journal of Computational Physics*, 31:335–362, June 1979. doi: 10.1016/0021-9991(79)90051-2.