No. 564 **April 2017**

Flux-corrected transport algorithms preserving the eigenvalue range of symmetric tensor quantities

C. Lohmann

ISSN: 2190-1767

Flux-corrected transport algorithms preserving the eigenvalue range of symmetric tensor quantities

Christoph Lohmann[a]

[a]*Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany*

Abstract

This paper presents a new approach to constraining the eigenvalue range of symmetric tensors in numerical advection schemes based on the flux-corrected transport (FCT) algorithm and a continuous finite element discretization. In the context of element-based FEM-FCT schemes for scalar conservation laws, the numerical solution is evolved using local extremum diminishing (LED) antidiffusive corrections of a low order approximation which is assumed to satisfy the relevant inequality constraints. The application of a limiter to antidiffusive element contributions guarantees that the corrected solution remains bounded by the local maxima and minima of the low order predictor.

The FCT algorithm to be presented in this paper guarantees the LED property for the largest and smallest eigenvalues of the transported tensor at the low order evolution step. At the antidiffusive correction step, this property is preserved by limiting the antidiffusive element contributions to all components of the tensor in a synchronized manner. The definition of the element-based correction factors for FCT is based on perturbation bounds for auxiliary tensors which are constrained to be positive semidefinite to enforce the generalized LED condition. The derivation of sharp bounds involves calculating the roots of polynomials of degree up to 3. As inexpensive and numerically stable alternatives, limiting techniques based on appropriate approximations are considered. The ability of the new limiters to enforce local bounds for the eigenvalue range is confirmed by numerical results for 2D advection problems.

Keywords: tensor quantity, continuous Galerkin method, flux-corrected transport, artificial diffusion, local discrete maximum principles

1. Introduction

During the last decades, advanced flux-corrected transport (FCT) algorithms have been developed for the numerical solution of scalar hyperbolic partial differential equations. They distinguish themselves from other stabilization techniques like streamline upwind Petrov-Galerkin (SUPG) by algebraically preserving local maximum principles while obtaining high order of accuracy in regions where the solution is smooth. This ensures monotonicity of the solution and robust algorithms.

While the FCT methodology has been successfully extended to various CFD problems like the Euler equations, one current issue of research is the limiting of (symmetric) tensor quantities, which occur, e.g., in context of orientation and stress tensors and will be discussed in this paper. In contrast to scalar variables, it is not entirely clarified which quantity corresponding to tensors should be observed and constrained to satisfy a relevant local maximum principle for the algorithm. The use of algorithms that limit each tensor component separately is not recommended since such limiting techniques are frame dependent and may fail to preserve physical properties like the definiteness. Therefore, Luttwak and Falcovitz [14] proposed a tensor image polyhedron (TIP) approach based on the convex hull idea initially developed for vectors [13]: The scaled/modified quantity of interest has to be located in the convex hull of neighboring vectors/tensors. This guarantees that tensor components are constrained in a frame invariant manner and, hence, preserves the symmetry of numerical solutions. Related and more efficient extensions are proposed in [12], where bounding boxes (BB) enclosing the convex hull are exploited. Unfortunately, TIP and BB approaches are designed

Email address: christoph.lohmann@math.tu-dortmund.de (Christoph Lohmann)

to limit tensor entries without taking other properties of tensors into consideration. In context of bounds preserving reconstruction (remapping), slope limiters for stress tensors were recently developed in [6, 16] treating principal invariants as quantities of interest: After separating the trace, a scalar approach can be applied to the trace while limiting the resulting tensor with vanishing trace by observing the second (and third) principal invariant separately. This allows a conservative restriction of the second invariant of deviatoric stress tensors, which is proportional to the elastic energy density [6]. In addition, eigenvalues are constrained implicitly because of their relation to principal invariants.

In this paper, an approach is proposed to limit the eigenvalues of symmetric tensors in a range preserving manner: The smallest/largest eigenvalue is bounded below/above by the local minimum/maximum of eigenvalues corresponding to a low order approximation. For this reason, physical properties should be preserved (especially in context of orientation tensors) and the definiteness is maintained. After calculating corresponding bounds, a simplified limiting procedure can be applied by exploiting polynomials of degree up to 3 without calculating further eigenvalues.

The paper is written as follows: After motivating the limiting of (the range of) eigenvalues (Sec. 2), the (element-based) low order method originally developed for scalar transport equations is extended to tensorial variables and the corresponding local extremum diminishing (LED) property is proved for the semidiscrete and fully discrete problem (Sec. 3). In Sec. 5, criteria are developed to preserve this property when adding scaled antidiffusive element contributions (defined in Sec. 4). These criteria are based on auxiliary tensors which are desired to be positive semidefinite. To reach this, eigenvalue calculations are necessary, when using worst case estimates or a theorem by Caron et al. [2]. Instead, principal invariants can be considered to ensure the positive semidefiniteness of the auxiliary tensors (Sec. 5.3). Further restrictions of tensor quantities can be defined by including local maximum principles for the trace (Sec. 6). Finally, the proposed limiting algorithms are validated and assessed in Sec. 7 using tensorial extensions of familiar FCT benchmarks.

1.1. Index convention for tensors

Without loss of generality, the (real-valued) eigenvalues of a symmetric tensor $A \in \mathbb{R}^{d\times d}$, $d = 2, 3$, are given in a sorted manner by $\lambda_1(A) \leq \ldots \leq \lambda_d(A)$. The notation can be shortened by using the abbreviations $a_1 \coloneqq \lambda_1(A), \ldots, a_d \coloneqq \lambda_d(A)$ while a_{kl}, $1 \leq k, l \leq d$, are the tensor entries corresponding to A. The eigenvalues and entries of a tensor A_i are referred to as $a_{i,k}$ and $a_{i,kl}$, respectively. Furthermore, if A is positive semidefinite, i.e., $0 \leq a_1 \leq \ldots \leq a_d$, the notation $A \geq 0$ is used (similar for $A \leq 0$, $A > 0$, and $A < 0$).

2. Properties of interest

The problem to be considered is given by the linear transport equation

$$\begin{cases} \partial_t U + \operatorname{div}(\mathbf{v}U) = 0 & \text{in } \Omega, \quad (1a)\\ U(\cdot, t) = U_{\text{in}} & \text{on } \Gamma_{\text{in}} = \{\mathbf{x} \in \partial\Omega : \mathbf{n}(\mathbf{x}) \cdot \mathbf{v} < 0\}, \quad (1b)\\ U(\cdot, 0) = U_0 & \text{in } \Omega, \quad (1c) \end{cases}$$

where $\Omega \subset \mathbb{R}^d$, $d = 2, 3$ is a bounded domain, Γ_{in} is the inflow part of the boundary $\partial\Omega$, $\mathbf{n} : \partial\Omega \to \mathbb{R}^d$ is the unit outward normal vector, $\mathbf{v} : \Omega\times\mathbb{R}_0^+ \to \mathbb{R}^d$ is the velocity field, and $U : \Omega\times\mathbb{R}_0^+ \to \mathbb{R}^{d\times d}$ is the unknown symmetric tensor and variable of interest. The initial and boundary conditions are given by the (symmetric) tensor fields $U_0 : \Omega \to \mathbb{R}^{d\times d}$ and $U_{\text{in}} : \Gamma_{\text{in}} \times \mathbb{R}_0^+ \to \mathbb{R}^{d\times d}$.

At the continuous level, a scalar solution of the transport equation (1) is positivity preserving and satisfies the maximum principle if $\operatorname{div}(\mathbf{v}) = 0$. When convecting tensorial unknowns with a divergence-free velocity field, each scalar quantity $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ differentially depending on the tensor entries evolves in the same manner as the solution of the scalar transport equation

$$\begin{aligned} \partial_t f(U) + \operatorname{div}(\mathbf{v} f(U)) &= (\nabla_U f(U)) : \partial_t U + \operatorname{div}(\mathbf{v}) f(U) + \mathbf{v} \cdot \operatorname{grad}(f(U))\\ &= (\nabla_U f(U)) : \partial_t U + \operatorname{div}(\mathbf{v}) f(U) + (\nabla_U f(U)) : (\mathbf{v} \cdot \operatorname{grad}(U))\\ &= (\nabla_U f(U)) : \partial_t U + (\nabla_U f(U)) : (\operatorname{div}(\mathbf{v}U)) + \operatorname{div}(\mathbf{v}) (f(U) - (\nabla_U f(U)) : U)\\ &= (\nabla_U f(U)) : (\partial_t U + \operatorname{div}(\mathbf{v}U)) + \operatorname{div}(\mathbf{v}) (f(U) - (\nabla_U f(U)) : U)\\ &= \operatorname{div}(\mathbf{v}) (f(U) - (\nabla_U f(U)) : U). \end{aligned}$$

In particular, entries, eigenvalues, and principal invariants of tensorial solutions of (1) satisfy local maximum principles if $\operatorname{div}(\mathbf{v}) = 0$. However, enforcing all these properties at the discrete level results in fairly restrictive conditions. The following Riemann problem with U_L for $x \leq 0$ and U_R for $x > 0$ defined by

$$U_L = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \qquad U_R = \begin{pmatrix} 2 & 0 \\ 0 & 1 \end{pmatrix}$$

illustrates, why limiting eigenvalues in case of continuous FE approximations can be more appropriate than restricting principal invariants and more likely to produce physically reasonable results: First of all, a numerical algorithm (preserving certain local maximum principles) for a tensor quantity must be frame invariant to reproduce existing symmetries of test problems correctly. Furthermore, in case of continuous FE approximations, discontinuities cannot be represented exactly. Therefore, artificial diffusion has to be added, which leads to a smooth blending $U(x)$ between U_L and U_R. Due to the frame invariance, $U(x)$ remains diagonal. Desired local maximum principles for the range of eigenvalues lead to the restriction

$$1 \leq u_1(x) \leq u_2(x) \leq 2 \qquad \Longrightarrow \qquad 1 \leq u_{11}(x), u_{22}(x) \leq 2,$$

where the index convention as described in Sec. 1.1 is used for eigenvalues and tensor components. In particular, a smooth blending $U(x)$ given by a convex combination of U_L and U_R is admissible while unphysical overshoots/undershoots of eigenvalues are prohibited. If (additionally) a local extremum diminishing (LED) property for the principal invariants should be satisfied, the inequality constraints simplify to the conditions

$$\mathrm{I}_1(U(x)) = u_1(x) + u_2(x) = 3, \qquad \mathrm{I}_2(U(x)) = u_1(x) \cdot u_2(x) = 2. \tag{2}$$

Using the frame invariance of the algorithm, (2) implies

$$U(x) \in \{U_L, U_R\}$$

and no smooth blending satisfying local maximum principles for the principal invariants is possible between U_L and U_R. If the frame invariance of the solution is neglected, a possible (not unique) blending is given by rotations of U_L/U_R, i.e.

$$U_\theta = \begin{pmatrix} 1 + \sin^2\theta & -\sin\theta\cos\theta \\ -\sin\theta\cos\theta & 1 + \cos^2\theta \end{pmatrix} \qquad 0 \leq \theta < 2\pi,$$

where $U_0 = U_\pi = U_L$ and $U_{\pi/2} = U_{3\pi/2} = U_R$.

For this reason, further on the local maximum principle for principal invariants, which is satisfied for the exact solution, is not adopted to the numerical method and the treatment of eigenvalues will be discussed.

3. Low order method

The standard Galerkin discretization of (1) is given by

$$\begin{gathered} \sum_{j=1}^{N_{\mathrm{dof}}} m_{ij} \frac{\mathrm{d}u_{i,kl}}{\mathrm{d}t} = \sum_{j=1}^{N_{\mathrm{dof}}} k_{ij} u_{j,kl} + b_{i,kl} \qquad \text{for all } 1 \leq i, j \leq N_{\mathrm{dof}},\ j \neq i \text{ and } 1 \leq k, l \leq d, \\ \text{where} \qquad m_{ij} \coloneqq \sum_{e=1}^{N_{\mathrm{elem}}} m_{ij}^e, \qquad k_{ij} \coloneqq \sum_{e=1}^{N_{\mathrm{elem}}} k_{ij}^e, \end{gathered} \tag{3}$$

where, $u_{i,kl}$, $1 \leq k, l \leq d$, are the entries of the tensor $U_i \in \mathbb{R}^{d \times d}$ corresponding to the degree of freedom i (compare to Sec. 1.1). In addition, B_i denotes (weak) Dirichlet boundary conditions and m_{ij}^e and k_{ij}^e are the entries of the element-based mass and convection matrix, respectively. This method does not preserve the range of eigenvalues and, hence, is used to derive a possible low order method, which satisfies local maximum principles for the eigenvalue range of the tensor quantity.

Since the tensorial transport equation (and the Galerkin discretization (3)) can be treated as one scalar transport equation for each entry of the tensorial solution, it is reasonable to use the same low order method for each component as in the scalar case. Therefore, the semidiscrete low order method given by [7]

$$m_i \frac{\mathrm{d}u_{i,kl}}{\mathrm{d}t} = \sum_{j=1}^{N_{\mathrm{dof}}} l_{ij} u_{j,kl} + b_{i,kl} \qquad \text{with} \quad m_i > 0, \quad l_{ij} \geq 0 \quad \text{for all } 1 \leq i, j \leq N_{\mathrm{dof}},\ j \neq i \text{ and } 1 \leq k, l \leq d \tag{4}$$

is considered, where m_i and l_{ij} are low order approximations to the entries of the mass and convection matrix, respectively. For example, a low order scheme with desired properties can be constructed using mass lumping and element-based discrete upwinding [9, 11]

$$m_i \coloneqq \sum_{e=1}^{N_{\mathrm{elem}}} m_i^e \coloneqq \sum_{e=1}^{N_{\mathrm{elem}}} \sum_{j=1}^{N_{\mathrm{dof}}} m_{ij}^e, \qquad l_{ij} \coloneqq k_{ij} + d_{ij} \coloneqq \sum_{e=1}^{N_{\mathrm{elem}}} k_{ij}^e + d_{ij}^e \qquad \text{where} \quad d_{ij}^e \coloneqq \begin{cases} \max\{-k_{ij}^e, 0, -k_{ji}^e\} & : j \neq i, \\ -\sum_{k \neq i} d_{ik}^e & : j = i. \end{cases}$$

For detailed derivations (for scalar transport equations), readers are referred to [7] and references therein.

In case of scalar variables, the low order method ensures the nonnegativity preservation and eliminates spurious oscillations at discontinuities. Furthermore, it can be proved that local extrema diminish if the velocity field is divergence-free. These properties are also satisfied for each entry of the tensor quantity due to missing couplings between different components. However, single entries have no physical meaning and, hence, their LED property is just a bonus. As explained above, LED properties for the eigenvalues are more important for tensor quantities. Hence, the ability of the low order method to preserve the (local) range of eigenvalues is discussed below.

For each symmetric tensor $U_i \in \mathbb{R}^{d \times d}$ there exists an orthogonal tensor $Q_i \in \mathbb{R}^{d \times d}$ such that $U_i = Q_i^\top \tilde{U}_i Q_i$ and $\tilde{U}_i = Q_i U_i Q_i^\top$, where $\tilde{U}_i \in \mathbb{R}^{d \times d}$ is a diagonal tensor with the eigenvalues $u_{i,1} \leq \ldots \leq u_{i,d}$ of U_i on its diagonal, i.e. $\tilde{u}_{i,kk} = u_{i,k}$, $1 \leq k \leq d$. Differentiation with respect to time leads to

$$\begin{aligned} \frac{\mathrm{d}\tilde{U}_i}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}(Q_i U_i Q_i^\top) &= (\mathrm{d}_t Q_i) U_i Q_i^\top + Q_i (\mathrm{d}_t U_i) Q_i^\top + Q_i U_i (\mathrm{d}_t Q_i^\top) \\ &= (\mathrm{d}_t Q_i) Q_i^\top \tilde{U}_i + Q_i (\mathrm{d}_t U_i) Q_i^\top + \tilde{U}_i Q_i (\mathrm{d}_t Q_i)^\top \\ &= H_i \tilde{U}_i + Q_i (\mathrm{d}_t U_i) Q_i^\top - \tilde{U}_i H_i, \end{aligned}$$

where

$$H_i \coloneqq (\mathrm{d}_t Q_i) Q_i^\top = \mathrm{d}_t(Q_i Q_i^\top) - Q_i (\mathrm{d}_t Q_i^\top) = \mathrm{d}_t \mathbb{1} - Q_i (\mathrm{d}_t Q_i^\top) = -Q_i (\mathrm{d}_t Q_i)^\top = -H_i^\top.$$

Substituting (4) for $\mathrm{d}_t U_i$, neglecting boundary conditions B_i, and assuming that $u_{i,1}$ is a local minimum, i.e., $u_{i,1} \leq u_{j,1} \leq u_{j,k}$ for all neighboring degrees of freedom j and $1 \leq k \leq d$, results in

$$\begin{aligned} \frac{\mathrm{d}u_{i,1}}{\mathrm{d}t} = \frac{\mathrm{d}\tilde{u}_{i,11}}{\mathrm{d}t} &= h_{i,11} \tilde{u}_{i,11} + (Q_i (\mathrm{d}_t U_i) Q_i^\top)_{11} - \tilde{u}_{i,11} h_{i,11} = (Q_i (\mathrm{d}_t U_i) Q_i^\top)_{11} \\ &= \Big(Q_i m_i^{-1} \sum_{j=1}^{N_{\mathrm{dof}}} l_{ij} U_j Q_i^\top \Big)_{11} = m_i^{-1} \sum_{j=1,\, j \neq i}^{N_{\mathrm{dof}}} l_{ij} (Q_i U_j Q_i^\top)_{11} + m_i^{-1} l_{ii} (Q_i U_i Q_i^\top)_{11} \\ &= m_i^{-1} \sum_{j=1,\, j \neq i}^{N_{\mathrm{dof}}} l_{ij} (Q_i U_j Q_i^\top)_{11} - m_i^{-1} \Big(\sum_{j=1,\, j \neq i}^{N_{\mathrm{dof}}} l_{ij} \Big) u_{i,1} (Q_i Q_i^\top)_{11} = m_i^{-1} \sum_{j=1,\, j \neq i}^{N_{\mathrm{dof}}} l_{ij} (Q_i (U_j - u_{i,1} \mathbb{1}) Q_i^\top)_{11} \geq 0 \end{aligned}$$

if $\sum_{j=1}^{N_{\mathrm{dof}}} l_{ij} = 0$ for all $1 \leq j \leq N_{\mathrm{dof}}$, because the diagonal entries of positive semidefinite tensors are nonnegative, $U_j - u_{i,1} \mathbb{1}$ is positive semidefinite due to $u_{i,1} \leq u_{j,1}$ ($\mathbb{1} \in \mathbb{R}^{d \times d}$ is the identity matrix), and $l_{ij} \geq 0$, $i \neq j$. Hence, a local minimum of (the smallest) eigenvalues cannot decrease and, similarly, a local maximum of (the largest) eigenvalues cannot increase. This can be interpreted as a local extremum diminishing property for (the eigenvalue range of) tensor quantities of the semidiscrete problem and justifies the choice of the same low order method as in the case of scalar transport equations.

This property carries over to the fully discrete problem after applying a strong stability preserving time integrator (eventually time step restrictions are required) [3, 4]: Assuming that the system after a two-level time discretization is given by (boundary conditions are neglected)

$$\sum_{j=1}^{N_{\text{dof}}} a_{ij} U_j^{n+1} = \sum_{j=1}^{N_{\text{dof}}} b_{ij} U_j^n \qquad \text{for all } 0 \le i \le N_{\text{dof}},$$

the eigenvalue range is preserved if

$$\sum_{j=1}^{N_{\text{dof}}} (a_{ij} - b_{ij}) = 0, \quad a_{ii} > 0, \quad b_{ii} \ge 0, \quad a_{ij} \le 0, \quad b_{ij} \ge 0 \qquad \text{for all } 0 \le i \ne j \le N_{\text{dof}}. \tag{5}$$

This criterion is satisfied for the low order method (4) discretized with the standard θ-scheme under corresponding CFL-like conditions [9]. Here, U_j^{n+1} and U_j^n are the tensors at node j and t^{n+1} and t^n, respectively. If $u_i^{\min}$ and $u_i^{\max}$ are lower and upper bounds at node i defined by (compare with [9])

$$u_i^{\min} := \min \left\{ \min_{\substack{1 \le j \le N_{\text{dof}}, \\ l_{ij} \ne 0}} u_{j,1}^n, \min_{\substack{1 \le j \le N_{\text{dof}}, \\ j \ne i,\, l_{ij} \ne 0}} u_{j,1}^{n+1} \right\}, \quad u_i^{\max} := \max \left\{ \max_{\substack{1 \le j \le N_{\text{dof}}, \\ l_{ij} \ne 0}} u_{j,d}^n, \max_{\substack{1 \le j \le N_{\text{dof}}, \\ j \ne i,\, l_{ij} \ne 0}} u_{j,d}^{n+1} \right\} \qquad \text{for all } 1 \le i \le N_{\text{dof}},$$

then $u_{i,1}^{n+1} \ge u_i^{\min}$ is satisfied for all $1 \le i \le N_{\text{dof}}$

$$\begin{aligned}
& a_{ii} U_i^{n+1} = b_{ii} U_i^n - \sum_{j=1, j\ne i}^{N_{\text{dof}}} a_{ij} U_j^{n+1} + \sum_{j=1, j\ne i}^{N_{\text{dof}}} b_{ij} U_j^n \\
\iff \quad & a_{ii} U_i^{n+1} = b_{ii} U_i^n - \sum_{j=1, j\ne i}^{N_{\text{dof}}} a_{ij} U_j^{n+1} + \sum_{j=1, j\ne i}^{N_{\text{dof}}} b_{ij} U_j^n + u_i^{\min} \mathbb{1} \sum_{j=1}^{N_{\text{dof}}} (a_{ij} - b_{ij}) \\
\iff \quad & a_{ii} (U_i^{n+1} - u_i^{\min} \mathbb{1}) = \underbrace{b_{ii} (U_i^n - u_i^{\min} \mathbb{1})}_{\ge 0} + \sum_{j=1, j\ne i}^{N_{\text{dof}}} \underbrace{(-a_{ij}) (U_j^{n+1} - u_i^{\min} \mathbb{1})}_{\ge 0} + \sum_{j=1, j\ne i}^{N_{\text{dof}}} \underbrace{b_{ij} (U_j^n - u_i^{\min} \mathbb{1})}_{\ge 0} \ge 0.
\end{aligned} \tag{6}$$

Furthermore, no new minimum of minimal eigenvalues can be generated at t^{n+1} (except for the boundary): Otherwise, there exists a degree of freedom i, where the new local minimum is attained, which means

$$u_{i,1}^{n+1} \le \min_{\substack{1 \le j \le N_{\text{dof}}, \\ l_{ij} \ne 0}} u_{j,1}^n, \qquad u_{i,1}^{n+1} \le \min_{\substack{1 \le j \le N_{\text{dof}}, \\ j \ne i,\, l_{ij} \ne 0}} u_{j,1}^{n+1} \qquad \Longrightarrow \qquad u_{i,1}^{n+1} = u_i^{\min}.$$

Then, (6) leads to

$$0 = a_{ii} (u_{i,1}^{n+1} - u_i^{\min}) \ge b_{ii} (u_{i,1}^n - u_i^{\min}) + \sum_{j=1, j\ne i}^{N_{\text{dof}}} (-a_{ij}) (u_{j,1}^{n+1} - u_i^{\min}) + \sum_{j=1, j\ne i}^{N_{\text{dof}}} b_{ij} (u_{j,1}^n - u_i^{\min}) \ge 0 \tag{7}$$

using (5) and $a_1 + b_1 \le \lambda_1(A + B) \le \lambda_d(A + B) \le a_d + b_d$

$$\begin{aligned}
& \mathbf{x}^\top (A + B - (a_1 + b_1)\mathbb{1})\, \mathbf{x} = \mathbf{x}^\top (A - a_1 \mathbb{1}) \mathbf{x} + \mathbf{x}^\top (B - b_1 \mathbb{1}) \mathbf{x} \ge 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \quad \Longrightarrow \quad \lambda_1(A + B) \ge a_1 + b_1, \\
& \mathbf{x}^\top (A + B - (a_d + b_d)\mathbb{1})\, \mathbf{x} = \mathbf{x}^\top (A - a_d \mathbb{1}) \mathbf{x} + \mathbf{x}^\top (B - b_d \mathbb{1}) \mathbf{x} \le 0 \quad \text{for all } \mathbf{x} \in \mathbb{R}^d \quad \Longrightarrow \quad \lambda_d(A + B) \le a_d + b_d.
\end{aligned} \tag{8}$$

In particular, (7) results in

$$u_{j,1}^n = u_i^{\min} \quad \text{for all } 1 \le j \le N_{\text{dof}} \text{ with } b_{ij} \ne 0 \qquad \text{and} \qquad u_{j,1}^{n+1} = u_i^{\min} \quad \text{for all } 1 \le j \le N_{\text{dof}} \text{ with } a_{ij} \ne 0$$

due to (5). Hence, the minimum is also attained at t^n and no new minimum is generated at t^{n+1}. Similar results can be obtained for the upper bound $u_i^{\max}$ and the LED property is valid at the discrete level, too.

Furthermore, each function $f : \mathbb{R}^{d\times d} \to \mathbb{R}$ linearly depending on U_i, i.e., satisfying $f(\alpha A + B) = \alpha f(A) + f(B)$ for all $A, B \in \mathbb{R}^{d\times d}$, $\alpha \in \mathbb{R}$, is local extremum diminishing: Assuming $f(U_i)$ is a local minimum, i.e., $f(U_i) \le f(U_j)$ for all neighboring nodes j, the time derivative can be written as

$$\begin{aligned}\frac{\mathrm{d}f(U_i)}{\mathrm{d}t} &= \sum_{k,l=1}^{d} f(\mathbf{e}_k\mathbf{e}_l^\top)\frac{\mathrm{d}u_{i,kl}}{\mathrm{d}t} = \sum_{k,l=1}^{d} m_i^{-1} f(\mathbf{e}_k\mathbf{e}_l^\top)\Big(\sum_{j=1}^{N_{\mathrm{dof}}} l_{ij}u_{j,kl}\Big)\\ &= m_i^{-1}\sum_{j=1}^{N_{\mathrm{dof}}} l_{ij}\sum_{k,l=1}^{d} u_{j,kl} f(\mathbf{e}_k\mathbf{e}_l^\top) = m_i^{-1}\sum_{j=1,\, j\neq i}^{N_{\mathrm{dof}}} l_{ij}\Big(f(U_j) - f(U_i)\Big) \ge 0\end{aligned}$$

and similarly for a local maximum. Therefore, the LED property is satisfied for the trace in particular.

Hence, the standard low order method for FCT algorithms yields the LED property for eigenvalues and the trace without additional limitations or modifications. This property is not valid for principal invariants due to the nonlinear dependence on the tensor entries. A low order method preserving the LED property for the determinant must be nonlinear when the tensor itself is the convected variable.

4. Antidiffusive correction

The low order method (4) is derived from the Galerkin method by adding artificial diffusion to guarantee discrete maximum principles for the local range of eigenvalues. This leads to a stable algorithm, which is fairly diffusive and, hence, not recommended for a numerical approximation per se. Nevertheless, it provides useful bounds for a correction procedure, which is based on the addition of antidiffusive components. These components can be designed to convert the low order method into a given arbitrary high order (stabilized) scheme. In the simplest case, the Galerkin method is reconstructed by eliminating artificial diffusion using the low order solution: Let the low order method (4) be discretized in time by the θ-scheme ($0 \le \theta \le 1$)

$$m_i\frac{U_i^L - U_i^n}{\Delta t} = \sum_{j=1}^{N_{\mathrm{dof}}} l_{ij}\Big(\theta U_j^L + (1-\theta)U_j^n\Big) + \theta B_i^{n+1} + (1-\theta)B_i^n,$$

where U_i^L and U_i^n are the low order solution and solution at t^n, respectively. Then the linearized high order Galerkin solution U^H is given by

$$m_i U_i^H = m_i U_i^L + \sum_{e=1}^{N_{\mathrm{elem}}} F_i^e,$$

where F_i^e are element-based antidiffusive components defined by

$$F_i^e = m_i^e(U_i^L - U_i^n) - \sum_{j=1}^{N_{\mathrm{dof}}} m_{ij}^e(U_j^L - U_j^n) + \Delta t\sum_{j=1}^{N_{\mathrm{dof}}} d_{ij}^e U_j^n = \sum_{j=1}^{N_{\mathrm{dof}}}\Big[m_{ij}^e\Big((U_i^L - U_i^n) - (U_j^L - U_j^n)\Big) + \Delta t d_{ij}^e U_j^n\Big], \qquad (9)$$

which do not contain any mass, i.e., $\sum_{i=1}^{N_{\mathrm{dof}}} F_i^e = 0 \cdot \mathbb{1}$. The reconstructed solution U^H has no guarantee that local maximum principles are satisfied. Therefore, the antidiffusive components are limited by element-based correction factors $0 \le \alpha^e \le 1$, which enforce specific constraints. Then, the monotonicity preserving solution at t^{n+1} is given by

$$m_i U_i^{n+1} = m_i U_i^L + \sum_{e=1}^{N_{\mathrm{elem}}} \alpha^e F_i^e.$$

Clearly, the crucial part of each FCT algorithm is the way in which the correction factors are defined to prevent the antidiffusive correction from producing artificial oscillations.

5. Eigenvalue range limiting

In case of tensor quantities, eigenvalues should not increase or decrease arbitrarily to preserve physically reasonable solutions. Especially orientation tensors have to stay positive semi definite to guarantee a stable numerical simulation. While the low order method guarantees the LED property for the eigenvalues, the antidiffusive components must be scaled by a correction factor for that reason. In general, these correction factors are calculated by estimating the quantity of interest depending on α^e (e.g., related derivations for the Euler equations can be found in [8]). In the two- and three-dimensional space, there exists an explicit formula for eigenvalue computations just depending on the principal invariants of the symmetric tensor.

2D

$$u_1 = \frac{1}{2}\Big(\mathrm{I}_1 - \sqrt{\mathrm{I}_1^2 - 4\mathrm{I}_2}\Big), \qquad u_2 = \frac{1}{2}\Big(\mathrm{I}_1 + \sqrt{\mathrm{I}_1^2 - 4\mathrm{I}_2}\Big), \tag{10a}$$

$$\mathrm{I}_1(U) = u_1 + u_2 = \mathrm{tr}(U) = u_{11} + u_{22}, \qquad \mathrm{I}_2(U) = u_1 u_2 = \det(U) = u_{11}u_{22} - u_{12}^2, \tag{10b}$$

3D [5, 17]

$$\begin{cases} u_1 = \frac{\mathrm{I}_1}{3} - 2\sqrt{v}\cos(\frac{\pi}{3} - \phi), & v = (\frac{\mathrm{I}_1}{3})^2 - \frac{\mathrm{I}_2}{3}, \\ u_2 = \frac{\mathrm{I}_1}{3} - 2\sqrt{v}\cos(\frac{\pi}{3} + \phi), & s = (\frac{\mathrm{I}_1}{3})^3 - \frac{\mathrm{I}_1\cdot\mathrm{I}_2}{6} + \frac{\mathrm{I}_3}{2}, \\ u_3 = \frac{\mathrm{I}_1}{3} + 2\sqrt{v}\cos(\phi), & \phi = \frac{1}{3}\arccos\left(s\sqrt{v^{-3}}\right), \end{cases} \tag{10c}$$

$$\begin{cases} \mathrm{I}_1(U) = u_1 + u_2 + u_3 & = \mathrm{tr}(U) & = u_{11} + u_{22} + u_{33}, \\ \mathrm{I}_2(U) = u_1u_2 + u_1u_3 + u_2u_3 = \frac{1}{2}\left((\mathrm{tr}U)^2 - \mathrm{tr}(U^2)\right) & & = u_{11}u_{22} + u_{11}u_{33} + u_{22}u_{33} - u_{12}^2 - u_{13}^2 - u_{23}^2, \\ \mathrm{I}_3(U) = u_1u_2u_3 & = \det(U) & = u_{11}u_{22}u_{33} + 2u_{12}u_{13}u_{23} - u_{22}u_{13}^2 - u_{33}u_{12}^2 - u_{11}u_{23}^2. \end{cases} \tag{10d}$$

However, due to their complexity, these relations are not recommended for estimation of correction factors. Instead of that, auxiliary tensors are defined by

$$U_i^{\min} := U_i^L - u_i^{\min}\mathbb{1} + \alpha_i^- m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}(F_i^e)_-, \qquad U_i^{\max} := u_i^{\max}\mathbb{1} - U_i^L - \alpha_i^+ m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}(F_i^e)_+, \tag{11}$$

where $u_i^{\min}, u_i^{\max} \in \mathbb{R}$ are lower and upper bounds for the eigenvalues [1, 7], $(F_i^e)_-$ and $(F_i^e)_+$ are negative and positive parts of the eigenvalue decomposition of $F_i^e = (F_i^e)_- + (F_i^e)_+$, and $0 \le \alpha_i^-, \alpha_i^+ \le 1$ are nodal-based correction factors. If $U_i^{\min}$ and $U_i^{\max}$ are positive semidefinite and $\alpha^e \le \min\{\alpha_i^-, \alpha_i^+\}$ for all nodes i corresponding to element e, then $U_i^{n+1} - u_i^{\min}\mathbb{1}$ and $u_i^{\max}\mathbb{1} - U_i^{n+1}$ are also positive semidefinite

$$U_i^{\min} \le U_i^L - u_i^{\min}\mathbb{1} + m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e(F_i^e)_- \le U_i^L - u_i^{\min}\mathbb{1} + m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e F_i^e = U_i^{n+1} - u_i^{\min}\mathbb{1},$$

$$U_i^{\max} \le u_i^{\max}\mathbb{1} - U_i^L - m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e(F_i^e)_+ \le u_i^{\max}\mathbb{1} - U_i^L - m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e F_i^e = u_i^{\max}\mathbb{1} - U_i^{n+1}$$

and all eigenvalues of U_i^{n+1} are bounded by $u_i^{\min}$ and $u_i^{\max}$, i.e., $u_i^{\min} \le u_{i,k}^{n+1} \le u_i^{\max}$ for all $1 \le k \le d$. A limiter based on these estimates involves the computation of (global) sums of negative and positive fluxes $\sum_{e=1}^{N_{\mathrm{dof}}}(F_i^e)_-$ and $\sum_{e=1}^{N_{\mathrm{dof}}}(F_i^e)_+$ as in Zalesak's FCT algorithm [19] and, hence, the eigenvalue decomposition of F_i^e is required. To reduce the computational effort and localize the auxiliary tensors, the following inequality for the eigenvalues can be considered

$$\underbrace{m_i^{-1}\Big(\sum_{e=1}^{N_{\mathrm{elem}}} m_i^e\Big)}_{=1}u_i^{\min} = u_i^{\min} \overset{!}{\le} \lambda_k(U_i^{n+1}) = \lambda_k(U_i^L + m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e F_i^e) = \lambda_k\Big((m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}} m_i^e)U_i^L + m_i^{-1}\sum_{e=1}^{N_{\mathrm{elem}}}\alpha^e F_i^e\Big)$$

$$= m_i^{-1}\lambda_k\Big(\sum_{e=1}^{N_{\mathrm{elem}}}\left(m_i^e U_i^L + \alpha^e F_i^e\right)\Big) \overset{!}{\le} u_i^{\max} = m_i^{-1}\Big(\sum_{e=1}^{N_{\mathrm{elem}}} m_i^e\Big)u_i^{\max}.$$

Due to (8), this condition is satisfied if the following relation holds

$$m_i^e u_i^{\min} \le \lambda_k \left(m_i^e U_i^L + \alpha^e F_i^e \right) \le m_i^e u_i^{\max} \qquad \text{for all } 1 \le e \le N_{\text{elem}} \text{ and } 1 \le k \le d.$$

Therefore, the element-based auxiliary tensors

$$U_i^{\min,e} := m_i^e (U_i^L - u_i^{\min} \mathbb{1}) + \alpha_i^{e,-} F_i^e, \qquad U_i^{\max,e} := m_i^e (u_i^{\max} \mathbb{1} - U_i^L) - \alpha_i^{e,+} F_i^e \tag{12}$$

must be positive semidefinite. Then, $\alpha^e \le \min_i\{\alpha_i^{e,-}, \alpha_i^{e,+}\}$ guarantees the boundedness of eigenvalues for all nodes i corresponding to element e.

In what follows, algorithms are presented that ensure the positive semidefiniteness of $A + \alpha B$, $A, B \in \mathbb{R}^{d\times d}$, depending on a correction factor $0 \le \alpha \le 1$. More precisely, we are interested in finding the largest $\alpha \in [0, 1]$, which satisfies the condition

$$A + \beta B \text{ is positive semidefinite for all } 0 \le \beta \le \alpha, \text{ where } A \text{ is positive semidefinite.} \tag{13}$$

The proposed methods can be used to enforce local maximum principles for the eigenvalues of tensor quantities based on the global or element-based approach (11) or (12), respectively.

5.1. min-min *criterion*

The most obvious way for defining $0 \le \alpha \le 1$ is by considering the smallest eigenvalues of A and B independently and exploiting (8), which yields

$$0 \overset{!}{\le} a_1 + \beta b_1 \le \lambda_1(A + \beta B) \quad \text{for all } 0 \le \beta \le \alpha \qquad \Longrightarrow \qquad \alpha := \begin{cases} -a_1 b_1^{-1} & : b_1 < -a_1, \\ 1 & : b_1 \ge -a_1. \end{cases} \tag{14}$$

This approach can be implemented quite easily but needs the computation of the eigenvalues of two tensors. Due to the separated treatment of A and B, the resulting α is not optimal when the system is ill conditioned. For example,

$$\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix} + \beta \begin{pmatrix} -3 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} 1 - 3\beta & 0 \\ 0 & 0 \end{pmatrix} \tag{15}$$

is positive semidefinite for all $0 \le \beta \le \frac{1}{3}$. However, algorithm (14) results in $\alpha = 0$.

5.2. Regularized criterion

Assuming A is positive definite, Caron et al. [2] have shown that $A + \beta B$ is positive semidefinite if and only if $\beta \in [\underline{\beta}, \bar{\beta}]$, where $\underline{\beta} < \bar{\beta} \in \bar{\mathbb{R}}$ are defined by

$$\underline{\beta} := \begin{cases} -\left(\lambda_d(A^{-1}B)\right)^{-1} & : \lambda_d(A^{-1}B) > 0, \\ -\infty & : \lambda_d(A^{-1}B) \le 0, \end{cases} \qquad \bar{\beta} := \begin{cases} -\left(\lambda_1(A^{-1}B)\right)^{-1} & : \lambda_1(A^{-1}B) < 0, \\ +\infty & : \lambda_1(A^{-1}B) \ge 0. \end{cases}$$

This leads to the following definition of the correction factor $0 \le \alpha \le 1$

$$\alpha := \begin{cases} -\left(\lambda_1(A^{-1}B)\right)^{-1} & : \lambda_1(A^{-1}B) < -1, \\ 1 & : \lambda_1(A^{-1}B) \ge -1. \end{cases} \tag{16}$$

In the context of FCT algorithms as described above, A is often nearly singular and the computation of $A^{-1}B$ is ill conditioned, due to the definition of local bounds. Especially at a local extremum, A is singular and (16) is not suitable. To rectify this, a stabilized version can be constructed by introducing a small regularization parameter $\varepsilon > 0$ and adding $\varepsilon\mathbb{1}$ to A. Then the smallest eigenvalue of $(A + \varepsilon\mathbb{1})^{-1}B$ can be expressed by

$$\lambda_1((A + \varepsilon\mathbb{1})^{-1}B) = \lambda_1\left((Q^\top \tilde{A} Q)^{-1}B\right) = \lambda_1\left(\tilde{A}^{-1}(QBQ^\top)\right) = \lambda_1\left(\tilde{A}^{-1/2}(QBQ^\top)\tilde{A}^{-1/2}\right) \tag{17}$$

using the eigenvalue decomposition $A + \varepsilon\mathbb{1} = Q^\top \tilde{A} Q$, where Q and $\tilde{A}$ are orthogonal and diagonal matrices. The resulting α guarantees the positive semidefiniteness of $A + \varepsilon\mathbb{1} + \alpha B$ and, hence, eigenvalues can violate local bounds by ε. Therefore, local maximum principles of the FCT solution are not preserved exactly.

To see this, applying (16) with regularization parameter ε to example (15) yields

$$(A + \varepsilon\mathbb{1})^{-1} B = \begin{pmatrix} 1+\varepsilon & 0 \\ 0 & \varepsilon \end{pmatrix}^{-1} \begin{pmatrix} -3 & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} -3(1+\varepsilon)^{-1} & 0 \\ 0 & 0 \end{pmatrix} \qquad \Longrightarrow \qquad \alpha := \tfrac{1}{3}(1+\varepsilon).$$

Of course, it is sufficient to add the regularization parameter ε just to the vanishing eigenvalues, which results in the exact solution $\alpha = \frac{1}{3}$.

This more accurate, but not strictly eigenvalue range preserving, algorithm needs the eigenvalue decomposition of A (if (17) is used) and one additional computation of the smallest eigenvalue of $A^{-1}B$. Therefore, it is significantly more expensive than the min-min criterion.

5.3. Invariant criterion

While the first algorithm proposed so far seems to be too restrictive for general tensors A and B, FCT solutions depending on the second one tend to violate the bounds for the eigenvalue range. A third approach depends on principal invariants and exploits the fact that a matrix is positive semidefinite if and only if all principal invariants $\mathrm{I}_1, \ldots, \mathrm{I}_d$ are nonnegative. Therefore, α has to be defined as

$$\alpha := \min\{\alpha_1, \ldots, \alpha_d\}, \qquad \text{where} \qquad \alpha_k = \inf\left(\{\beta \geq 0 : \mathrm{I}_k(A + \beta B) < 0\} \cup \{1\}\right) \quad \text{for all } 1 \leq k \leq d. \tag{18}$$

Here, α_k corresponds to the first root of the polynomial $\mathrm{i}_k(\beta) := \mathrm{I}_k(A + \beta B)$ in $[0, 1]$, where i_k changes its sign from positive to negative. For regular/positive definite matrices A, the minimum of α is reached at α_d, the smallest positive value, for which $A + \beta B$ becomes singular (by definition of the determinant I_d). Furthermore, this value coincides with the correction factor of the Regularized criterion when using $\varepsilon = 0$. In contrast, no additional regularization is necessary if A is singular.

While there exist stable and direct algorithms for calculating α_1 and α_2 [15], roots of a cubic polynomial can be determined accurately and efficiently using polynomial fitting [18]. However, to avoid the computation of all roots of the cubic polynomial $\mathrm{i}_3(\beta)$, the following pseudo algorithm can be used:

0. At steps 1, 2, and 3 below, we are searching for an interval $[\underline{\beta}, \bar{\beta}]$, which contains the desired root α_3 of $\mathrm{i}_3(\beta)$, where the sign changes from positive to negative (if existing). This interval is initialized by $[\underline{\beta}, \bar{\beta}] := [0, 1]$.
1. To estimate α_3, we calculate critical points of $\mathrm{i}_3(\beta)$, i.e. the roots of $\mathrm{i}_3{}'(\beta)$. Due to the above sign change assumption, there is at most one root of $\mathrm{i}_3(\beta)$ between two neighboring extrema.
2. Starting with the smallest root of $\mathrm{i}_3{}'(\beta)$, iterate over all roots $\tilde{\beta} \in [0, 1]$.
   - If $\mathrm{i}_3(\tilde{\beta}) \geq 0$, update $\underline{\beta} := \tilde{\beta}$ (the desired root of i_3 cannot be in $[\underline{\beta}, \tilde{\beta}]$).
   - Otherwise, we have $\alpha_3 \in [\underline{\beta}, \bar{\beta}]$, where $\bar{\beta} := \tilde{\beta}$. → Go to 4.
3. If no interval $[\underline{\beta}, \bar{\beta}]$ so far contains α_3, check if $\mathrm{i}_3(1) \geq 0$. If this is the case, there is no desired root in $[0, 1]$ and $\alpha_3 := 1$. Otherwise, we have $\alpha_3 \in [\underline{\beta}, \bar{\beta}]$ with $\bar{\beta} = 1$. → Go to 4.
4. $\alpha_3 \in [\underline{\beta}, \bar{\beta}]$ can be calculated iteratively by an algorithm, which yields guaranteed bounds after each iteration (e.g., bisection method).

Of course, this algorithm can also be used to find the desired root α_2 of the quadratic polynomial $\mathrm{i}_2(\beta)$.

Due to involved methods for calculating the required roots $\alpha_1, \ldots, \alpha_d$, this criterion can be most expensive (depending on the used root-finding algorithm). On the other hand, it calculates the most accurate correction factors while preserving the eigenvalue range.

The approach of using the principal invariants to define α such that (13) is satisfied can also be exploited to define an *approximate Invariant criterion* based on the inequality

$$0 \leq \beta \leq 1 \qquad \Longrightarrow \qquad C\beta^{k+1} \leq C\beta^k \quad \text{for all } C \geq 0 \text{ and } k \in \mathbb{N}. \tag{19}$$

To describe the corresponding method, let the polynomial $\mathrm{i}_k(\beta)$ be given by

$$\mathrm{i}_k(\beta) = \mathrm{I}_k(A + \beta B) = c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + c_1^{(k)}\beta + c_0^{(k)},$$

which should be nonnegative for all $0 \le \beta \le \alpha_k$ due to (18). The easiest way is to estimate $\mathrm{i}_k(\beta)$ by a linear polynomial

$$\begin{aligned}
\mathrm{i}_k(\beta) &= c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots c_1^{(k)}\beta + c_0^{(k)} \\
&\ge (\min\{0, c_k^{(k)}\} + c_{k-1}^{(k)})\beta^{k-1} + \ldots c_1^{(k)}\beta + c_0^{(k)} \\
&\ge \ldots \\
&\ge \Big(\min\Big\{0, \ldots, \min\{0, \min\{0, c_k^{(k)}\} + c_{k-1}^{(k)}\} + \ldots + c_2^{(k)}\Big\} + c_1^{(k)}\Big)\beta + c_0^{(k)} \overset{!}{\ge} 0
\end{aligned} \tag{20}$$

due to (19). If $c_1^{(k)} \ge 0$, this estimate can be improved by using a quadratic approximation

$$\begin{aligned}
\mathrm{i}_k(\beta) &= c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + c_1^{(k)}\beta + c_0^{(k)} \\
&\ge c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + c_3^{(k)}\beta^3 + (c_2^{(k)} + c_1^{(k)})\beta^2 + c_0^{(k)} \\
&\ge (\min\{0, c_k^{(k)}\} + c_{k-1}^{(k)})\beta^{k-1} + \ldots + c_3^{(k)}\beta^3 + (c_2^{(k)} + c_1^{(k)})\beta^2 + c_0^{(k)} \\
&\ge \ldots \\
&\ge \Big(\min\Big\{0, \ldots \min\{0, \min\{0, c_k^{(k)}\} + c_{k-1}^{(k)}\} + \ldots + c_3^{(k)}\Big\} + (c_2^{(k)} + c_1^{(k)})\Big)\beta^2 + c_0^{(k)} \overset{!}{\ge} 0.
\end{aligned}$$

In general, the above approach leads to estimates of the form

$$\begin{aligned}
\mathrm{i}_k(\beta) &= c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + c_1^{(k)}\beta + c_0^{(k)} \\
&\ge c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + \Big(\sum_{s=1}^{r} c_s^{(k)}\Big)\beta^r + c_0^{(k)} \\
&\ge \Big(\min\Big\{\{0, \ldots \min\{0, \min\{0, c_k^{(k)}\} + c_{k-1}^{(k)}\} + \ldots + c_{r+1}^{(k)}\Big\} + \sum_{s=1}^{r} c_s^{(k)}\Big)\beta^r + c_0^{(k)} =: c^{(k)}\beta^r + c_0^{(k)} \overset{!}{\ge} 0,
\end{aligned} \tag{21}$$

where $r = \min\{1 \le s \le d : \sum_{i=1}^{s} c_i^{(k)} < 0\}$. Then, the correction factor α_k is defined by

$$\alpha_k := \begin{cases} \Big(-c_0^{(k)}(c^{(k)})^{-1}\Big)^{1/r} & : c^{(k)} < -c_0^{(k)}, \\ 1 & : c^{(k)} \ge -c_0^{(k)}. \end{cases}$$

Furthermore, this approach can be improved by exploiting (18) and already known quantities $\alpha_1, \ldots, \alpha_{k-1}$ when calculating α_k: If the fluxes are prelimited by defining $F_i^e := \alpha^e F_i^e$, i.e., $B := \alpha^e B$, then the estimates of (20) and (21) are less restrictive and more accurate results can be expected.

In case of example (15), both approaches based on the Invariant criterion (exact and approximate) yield the optimal correction factor $\alpha = \frac{1}{3}$ (note that $\mathrm{I}_2(A + \beta B) = 0$ for all $0 \le \beta \le 1$). However, the modified tensor of example (15) given by

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + \beta \begin{pmatrix} -3 & 0 \\ 0 & -3 \end{pmatrix} = \begin{pmatrix} 1 - 3\beta & 0 \\ 0 & 1 - 3\beta \end{pmatrix}$$

is still positive semidefinite for all $0 \le \beta \le \frac{1}{3}$. While the min-min criterion leads to the optimal value $\alpha = \frac{1}{3}$, the approximate Invariant criterion using (20) or (21) yields

$$\mathrm{i}_3(\beta) = (1 - 2\beta)^2 = 1 - 6\beta + 9\beta^2 \ge 1 - 6\beta \overset{!}{\ge} 0 \qquad \Longrightarrow \qquad \alpha_3 := \tfrac{1}{6},$$

which is more restrictive. On the other hand, the approximate Invariant criterion does not involve any computation of eigenvalues and is therefore the most efficient algorithm for calculating the correction factor α.

All proposed approaches are frame invariant, i.e., they do not yield different correction factors α if a similarity transformation is applied to $A + \beta B$. In case of the min-min criterion or Regularized criterion this can be seen directly because of the frame invariance of eigenvalues. In case of the Invariant criterion, consider an arbitrary orthogonal tensor $Q \in \mathbb{R}^{d\times d}$. Then the principal invariant I_k satisfies

$$c_k^{(k)}\beta^k + c_{k-1}^{(k)}\beta^{k-1} + \ldots + c_0^{(k)} = \mathrm{I}_k(A+\beta B) = \mathrm{I}_k\left(Q^\top(A+\beta B)Q\right) =: \tilde{c}_k^{(k)}\beta^k + \tilde{c}_{k-1}^{(k)}\beta^{k-1} + \ldots + \tilde{c}_0^{(k)}$$

for all $0 \le \beta \le 1$ and $1 \le k \le d$. This can be rewritten as

$$0 = (c_k^{(k)} - \tilde{c}_k^{(k)})\beta^k + (c_{k-1}^{(k)} - \tilde{c}_{k-1}^{(k)})\beta^{k-1} + \ldots + (c_0^{(k)} - \tilde{c}_0^{(k)}) \qquad \text{for all } 0 \le \beta \le 1 \text{ and } 1 \le k \le d.$$

Therefore, $c_k^{(k)} = \tilde{c}_k^{(k)}$ for all $1 \le k \le d$ must be valid and limiters just depending on $c_0^{(0)}, \ldots, c_d^{(d)}$ are frame invariant, which is particularly the case for the (approximate) Invariant criterion.

6. Trace limiter

In the last section, algorithms are proposed for limiting antidiffusive fluxes such that local minima/maxima of the smallest/largest eigenvalue cannot decrease/increase. The admissible eigenvalue range is determined using the low order method presented in Sec. 3, which was shown to preserve the eigenvalue range. In particular, the so-defined LED property for tensor quantities guarantees that the eigenvalues satisfy global bounds. Furthermore, the FCT correction step implicitly constrains the magnitude of principal invariants due to

$$|\mathrm{I}_k(U)| = \Big|\sum_{1\le i_1<\ldots<i_k\le d} u_{i_1}\cdot\ldots\cdot u_{i_k}\Big| \le (d-k+1)\Big(\max_{1\le l\le d}|u_l|\Big)^k \qquad \text{for all } 1 \le k \le d.$$

However, the principal invariants of the low order solution are not local extremum diminishing and possibly violate local bounds. Clearly the correction step of the FCT algorithm cannot rectify this drawback of the low order method. Therefore, constraining principal invariants in context of continuous FEM-FCT algorithms is not recommended and just results in stronger limiting.

Nevertheless, as described on page 6, the low order method possesses the LED property for the trace. So, it is worthwhile to preserve this benefit when it comes to adding antidiffusive fluxes. This can be done in the same Zalesak-like manner as for scalar quantities, due to the linear dependence of the trace on tensor entries. In the localized version of element-based FCT, this method is given by

$$\alpha^{\mathrm{tr},e} := \min_{\substack{1\le j\le N_{\mathrm{dof}},\\ l_{ij}\neq 0}} R_i^{\mathrm{tr},e}, \qquad R_i^{\mathrm{tr},e} = \begin{cases} \min\left\{1, \frac{m_i^e(\mathrm{tr}_i^{\max}-\mathrm{tr}_i^L)}{\mathrm{tr}(F_i^e)}\right\} & : \mathrm{tr}(F_i^e) > 0, \\ 1 & : \mathrm{tr}(F_i^e) = 0, \\ \min\left\{1, \frac{m_i^e(\mathrm{tr}_i^{\min}-\mathrm{tr}_i^L)}{\mathrm{tr}(F_i^e)}\right\} & : \mathrm{tr}(F_i^e) < 0, \end{cases}$$

where $\mathrm{tr}_i^{\min}$ and $\mathrm{tr}_i^{\max}$ are local bounds for the trace [1, 7] and tr_i^L is the trace of the low order solution in node i. Then, the element-based correction factor can be defined by $\alpha^e := \min\{\alpha^e, \alpha^{\mathrm{tr},e}\}$. If the approximate Invariant criterion is used or global sums of positive and negative antidiffusive element contributions are calculated, it is recommended to limit the trace first and scale the element contributions F^e with $\alpha^{\mathrm{tr},e}$ before applying the eigenvalue range limiter. This results in more accurate solutions than the reverse approach, because estimates for correction factors of eigenvalues are based on smaller absolute fluxes.

6.1. Special case: Orientation tensors

The trace of general tensors is the only principal invariant which depends linearly on tensor entries and, hence, is worthwhile to be limited. While bounding eigenvalues is quite complicated due to nonlinear dependencies, methods for the trace can be adopted directly from scalar FCT algorithms. However, in case of orientation tensors, which possess a unit trace property, this application is redundant as shown below.

Orientation tensors are defined by their positive semidefiniteness combined with a unit trace property. While the positive semidefiniteness is preserved by the eigenvalue range limiter of Sec. 5, an arbitrary space discretization of the form $\sum_{j=1}^{N_{\text{dof}}}(m_{ij}\mathrm{d}_t U_j - k_{ij}U_j) = 0$ leads to

$$\sum_{j=1}^{N_{\text{dof}}} m_{ij}\frac{\mathrm{d}\text{tr}(U_j)}{\mathrm{d}t} = \sum_{j=1}^{N_{\text{dof}}} m_{ij}\sum_{k=1}^{d}\frac{\mathrm{d}u_{j,kk}}{\mathrm{d}t} = \sum_{k=1}^{d}\sum_{j=1}^{N_{\text{dof}}} k_{ij}u_{j,kk} = \sum_{j=1}^{N_{\text{dof}}} k_{ij}\text{tr}(U_j) = \sum_{j=1}^{N_{\text{dof}}} k_{ij} \qquad \text{for all } 1 \le i \le N_{\text{dof}}.$$

Therefore, the first principal invariant stays constant if the mass matrix is regular and the convection matrix has vanishing row sums, i.e., $\sum_{j=1}^{N_{\text{dof}}} k_{ij} = 0$. Clearly, this is particularly true for the high order Galerkin method (3) and the corresponding low order counterpart (4) if $\text{div}(\mathbf{v}) = 0$ and, therefore, the zero-sum condition holds at the discrete level. Finally, high order corrections depending on antidiffusive element contributions (9) do not change the trace due to the definition of d_{ij}^e and

$$\text{tr}(F_i^e) = \sum_{k=1}^{d} f_{i,kk}^e = \sum_{j=1}^{N_{\text{dof}}}\Big[m_{ij}^e\Big(\sum_{k=1}^{d} u_{i,kk}^L - \sum_{k=1}^{d} u_{i,kk}^n - \sum_{k=1}^{d} u_{j,kk}^L + \sum_{k=1}^{d} u_{j,kk}^n\Big) + \Delta t d_{ij}^e \sum_{k=1}^{d} u_{j,kk}^n\Big] = \Delta t \sum_{j=1}^{N_{\text{dof}}} d_{ij}^e = 0$$

$$\Longrightarrow \qquad \text{tr}(U_i^{n+1}) = \text{tr}(U_i^L) + m_i^{-1}\sum_{e=1}^{N_{\text{elem}}} \alpha^e \text{tr}(F_i^e) = \text{tr}(U_i^L) \equiv \text{const.}$$

Therefore, convecting all diagonal entries of orientation tensors is redundant in numerical applications and an arbitrary diagonal entry can be neglected and determined when required using the definition of the trace.

7. Numerical experiments

In Secs. 5 and 6, FCT algorithms for symmetric tensors are proposed to preserve the monotonicity of eigenvalues and the trace. While standard approaches for scalar quantities can be easily adopted to restrict antidiffusive element contributions with respect to the trace, methods for constraining eigenvalues require more advanced techniques. Therefore, different criteria were developed exploiting the desired positive semidefiniteness of auxiliary tensors of the form $A + \alpha B$.

In this section, the different criteria are compared with each other using tensorial extensions of generally accepted benchmarks in the context of monotonicity preserving methods. Solutions of componentwise limiters are observed to justify the necessity of eigenvalue limiting. Furthermore, the influence of first bounding the trace is discussed briefly.

The results are presented by plotting minimal and maximal eigenvalues (color bar is restricted to interval $[0, 1]$; overshoots/undershoots are plotted in magenta; contour lines exist for values 0.0, 0.1, ..., 1.0) and measure the L^2- and L^1-errors of the Frobenius and spectral norm $\|\cdot\|_F$ and $\|\cdot\|_2$, respectively. To identify different FCT algorithms, limiters for the eigenvalue range are abbreviated in the following manner

$$\texttt{ER}_{\texttt{cr}}\texttt{-FCT} \qquad \text{where} \qquad \texttt{cr} \in \begin{cases} \texttt{min} & \text{: min-min criterion,} \\ \texttt{reg} & \text{: Regularized criterion,} \\ \texttt{inv} & \text{: Invariant criterion,} \\ \texttt{app} & \text{: approximate Invariant criterion.} \end{cases}$$

Furthermore, the following synonyms

$$\texttt{tr-FCT}, \qquad \texttt{TE}_{\texttt{sep}}\texttt{-FCT}, \qquad \texttt{TE}_{\texttt{syn}}\texttt{-FCT}$$

are used for the trace and separated/synchronized tensor entry limiters, respectively. If the trace is handled before applying another limiter like the 'invariant based eigenvalue limiter' $\texttt{ER}_{\texttt{inv}}\texttt{-FCT}$, the shortened notation $\texttt{tr-ER}_{\texttt{inv}}\texttt{-FCT}$ is used.

In all numerical examples, the Crank-Nicolson method is used as the time integrator of the low order method ($\theta = \frac{1}{2}$). If not mentioned otherwise, the spatial domain $\Omega = (0, 1)^2$ is discretized uniformly using $128 \cdot 128 = 16384$

quadrilaterals with $(128+1)^2 = 16641$ degrees of freedom (bilinear finite elements). By choosing $\Delta t = 10^{-3}$, the CFL condition

$$\mathrm{CFL} = \frac{\|\mathbf{v}\|\Delta t}{\Delta x} = \frac{\|\mathbf{v}\|10^{-3}}{(128)^{-1}} \leq 0.2$$

is satisfied in each benchmark.

7.1. Solid body rotation

The first test problem is given by a customization of the solid body rotation benchmark first introduced by Zalesak [19] and extended by LeVeque [10]. In the scalar case three solid bodies, a slotted cylinder, a sharp cone, and a smooth hump, are rotated around the center of the two-dimensional domain $\Omega = (0,1)^2$ using the time-independent and divergence-free velocity field

$$\mathbf{v}(x,y) \coloneqq \left(\frac{1}{2} - y, x - \frac{1}{2}\right)^\top.$$

At $T = 2\pi$ the exact solution coincides with the initial condition and corresponding errors are calculated.

In this case, the unknown solution is given by a three-dimensional tensor still convected in the two-dimensional domain. Similarly to the scalar benchmark, the initial condition is defined piecewise by four solid bodies

$$U_0(x,y) \coloneqq \begin{cases} U^{(1)}\left(\frac{x-0.25}{0.15}, \frac{y-0.5}{0.15}\right) & : \sqrt{(x-0.25)^2 + (y-0.5)^2} \leq 0.15, \quad \text{'hump'} \\ U^{(2)}\left(\frac{x-0.5}{0.15}, \frac{y-0.25}{0.15}\right) & : \sqrt{(x-0.5)^2 + (y-0.25)^2} \leq 0.15, \quad \text{'cone'} \\ U^{(3)}\left(\frac{x-0.75}{0.15}, \frac{y-0.5}{0.15}\right) & : \sqrt{(x-0.75)^2 + (y-0.5)^2} \leq 0.15, \quad \text{'semi-ellipse'} \\ U^{(4)}\left(\frac{x-0.5}{0.15}, \frac{y-0.75}{0.15}\right) & : \sqrt{(x-0.5)^2 + (y-0.75)^2} \leq 0.15, \quad \text{'slotted cylinder'} \\ 0 \cdot \mathbb{1} & : \text{otherwise,} \end{cases} \tag{22}$$

where the positive semidefinite tensors $U^{(1)}$, $U^{(2)}$, $U^{(3)}$, and $U^{(4)}$ are described by their eigenvalue decomposition as follows

$$U^{(1)}(x,y) \coloneqq \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & \sin\phi & -\cos\phi \end{pmatrix} \frac{1}{r} \begin{pmatrix} x & y & 0 \\ y & -x & 0 \\ 0 & 0 & r \end{pmatrix} \begin{pmatrix} u_1^{(1)} & 0 & 0 \\ 0 & u_2^{(1)} & 0 \\ 0 & 0 & u_3^{(1)} \end{pmatrix} \frac{1}{r} \begin{pmatrix} x & y & 0 \\ y & -x & 0 \\ 0 & 0 & r \end{pmatrix} \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos\phi & \sin\phi \\ 0 & \sin\phi & -\cos\phi \end{pmatrix},$$

$$\text{where} \quad u_1^{(1)} = \left(\tfrac{1}{2}\left(1 + \cos(\pi r)\right)\right)^3, \quad u_2^{(1)} = \left(\tfrac{1}{2}\left(1 + \cos(\pi r)\right)\right)^2, \quad u_3^{(1)} = \tfrac{1}{2}\left(1 + \cos(\pi r)\right), \quad \phi = \tfrac{1}{2}\operatorname{arctan2}(x,y),$$

$$U^{(2)}(x,y) \coloneqq \frac{1}{10}\begin{pmatrix} 10 & 0 & 0 \\ 0 & 8 & 6 \\ 0 & 6 & -8 \end{pmatrix} \frac{1}{r} \begin{pmatrix} x & y & 0 \\ y & -x & 0 \\ 0 & 0 & r \end{pmatrix} \begin{pmatrix} u_1^{(2)} & 0 & 0 \\ 0 & u_a^{(2)} & 0 \\ 0 & 0 & u_b^{(2)} \end{pmatrix} \frac{1}{r} \begin{pmatrix} x & y & 0 \\ y & -x & 0 \\ 0 & 0 & r \end{pmatrix} \frac{1}{10}\begin{pmatrix} 10 & 0 & 0 \\ 0 & 8 & 6 \\ 0 & 6 & -8 \end{pmatrix},$$

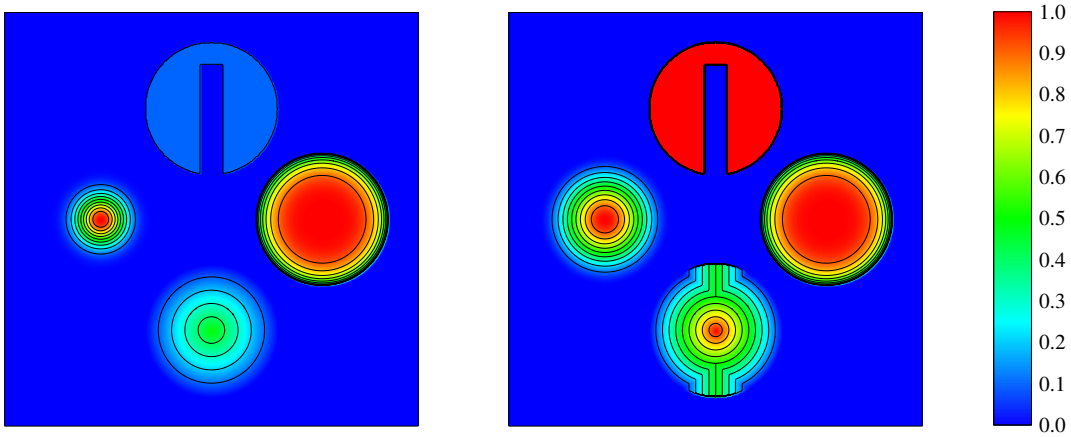
$$\text{where} \quad u_1^{(2)} = \tfrac{1}{2} - \tfrac{1}{2}r, \quad u_a^{(2)} = \tfrac{1}{2} - \tfrac{1}{2}|x|, \quad u_b^{(2)} = 1 - r,$$

$$U^{(3)}(x,y) \coloneqq \begin{pmatrix} u_1^{(3)} & 0 & 0 \\ 0 & u_1^{(3)} & 0 \\ 0 & 0 & u_1^{(3)} \end{pmatrix},$$

$$\text{where} \quad u_1^{(3)} = u_2^{(3)} = u_3^{(3)} = \sqrt{1 - r^2},$$

$$U^{(4)}(x,y) \coloneqq \begin{cases} \frac{1}{10}\begin{pmatrix} -8 & 6 & 0 \\ 6 & 8 & 0 \\ 0 & 0 & 10 \end{pmatrix} \begin{pmatrix} u_3^{(4)} & 0 & 0 \\ 0 & u_1^{(4)} & 0 \\ 0 & 0 & u_2^{(4)} \end{pmatrix} \frac{1}{10}\begin{pmatrix} -8 & 6 & 0 \\ 6 & 8 & 0 \\ 0 & 0 & 10 \end{pmatrix} & : \left(|x| \geq \tfrac{1}{6} \vee y > \tfrac{2}{3}\right) \wedge (x > 0), \\ \frac{1}{10}\begin{pmatrix} -8 & 6 & 0 \\ 6 & 8 & 0 \\ 0 & 0 & 10 \end{pmatrix} \begin{pmatrix} u_1^{(4)} & 0 & 0 \\ 0 & u_3^{(4)} & 0 \\ 0 & 0 & u_2^{(4)} \end{pmatrix} \frac{1}{10}\begin{pmatrix} -8 & 6 & 0 \\ 6 & 8 & 0 \\ 0 & 0 & 10 \end{pmatrix} & : \left(|x| \geq \tfrac{1}{6} \vee y > \tfrac{2}{3}\right) \wedge (x < 0), \\ 0 \cdot \mathbb{1} & : \text{otherwhere,} \end{cases}$$

$$\text{where} \quad u_1^{(4)} = 0.1, \quad u_2^{(4)} = 0.45, \quad u_3^{(4)} = 1,$$



(a) Minimal eigenvalue $u_{h,1}$. (b) Maximal eigenvalue $u_{h,3}$.

Figure 1: Solid body rotation and swirling flow: Range of eigenvalues of exact solution at initial and final time.

using the radius $r := \sqrt{x^2 + y^2}$. The intermediate and largest eigenvalue of $U^{(2)}$ are given by

$$u_2^{(2)} = \begin{cases} u_a^{(2)} & : |x| \geq 2r - 1, \\ u_b^{(2)} & : |x| < 2r - 1, \end{cases} \qquad u_3^{(2)} = \begin{cases} u_b^{(2)} & : |x| \geq 2r - 1, \\ u_a^{(2)} & : |x| < 2r - 1 \end{cases}$$

while 0 and 1 are the (globally) smallest and largest eigenvalue of U_0.

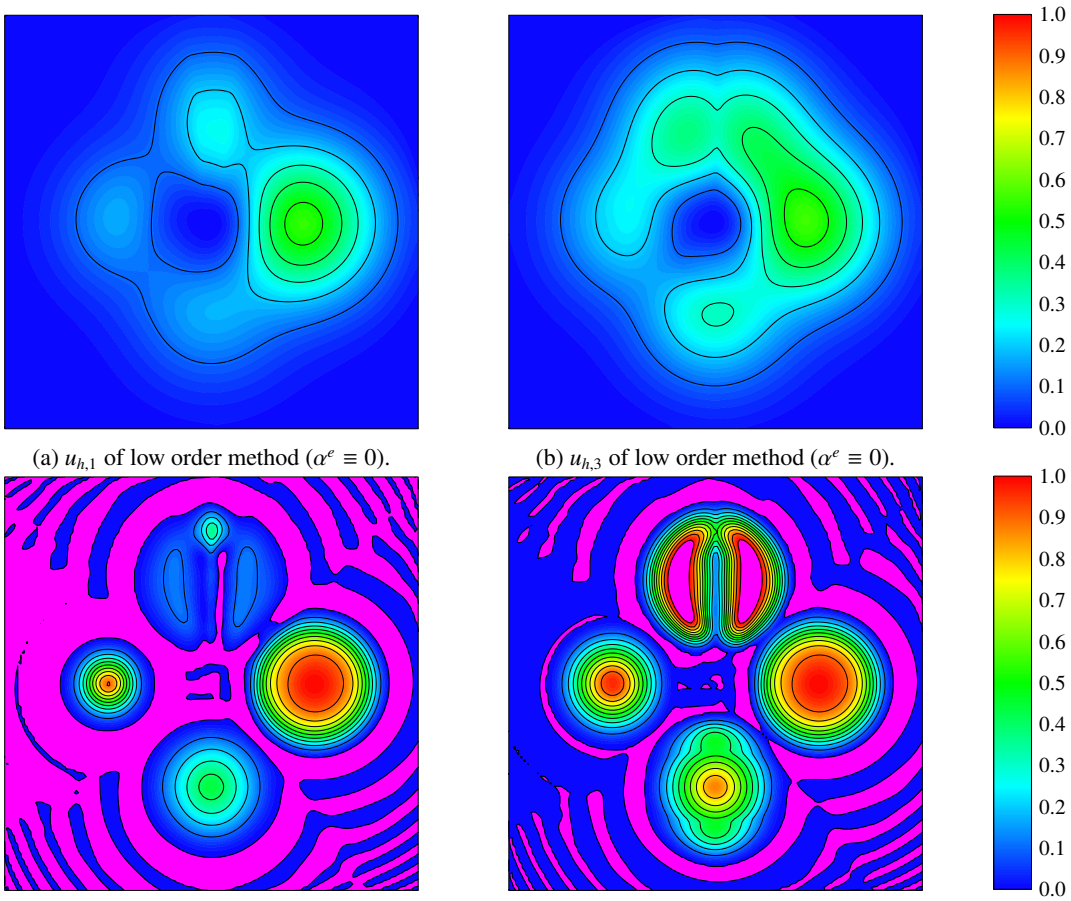
The descriptions 'hump', 'cone', 'semiellipse', and 'slotted cylinder' of (22) are based on the design of corresponding minimal/maximal eigenvalues (Fig. 1) and their similarities to counterparts of the scalar benchmark. For this reason, the structure of the exact solution is well suited for studying the behavior of the monotonicity preserving methods near 'discontinuities' (slotted cylinder and semiellipse) and in regions of smooth and sharp peaks (hump and cone).

Using the low order method (Figs. 2a and 2b), details of the initial condition are completely smoothed out by adding artificial diffusion in such a way that local maximum principles are preserved algebraically. Initially separated bodies merge and produce a single monotone and smooth body satisfying global bounds for the range of eigenvalues as well as tensor entries as proved in Sec. 3.

If linearized antidiffusive element contributions are added without corrections (Sec. 4), the accuracy of the numerical solution improves and contours of each body become visible (Figs. 2c and 2d). However, local maximum principles for the eigenvalues and tensor components do not hold and spurious oscillations with overshoots/undershoots occur. As shown in Table 1, global maximum principles are still violated if FCT algorithms for the tensor entries are applied, i.e., $\texttt{TE}_{\texttt{sep}}\texttt{-FCT}$ and $\texttt{TE}_{\texttt{syn}}\texttt{-FCT}$. Even though $\texttt{TE}_{\texttt{sep}}\texttt{-FCT}$ produces the most accurate result of all considered methods, the LED property for the eigenvalue range is violated and, in particular, tensors do not stay positive semidefinite (Figs. 3a and 3b).

In contrast to this method, the synchronized counterpart $\texttt{TE}_{\texttt{syn}}\texttt{-FCT}$ produces diffusive results with excessive distortions (Figs. 3c and 3d). This behavior can be explained by the drawback of any synchronized FCT algorithm: If a quantity of interest is almost constant (but not exactly), the upper and lower bounds are close to each other. Then, small changes in antidiffusive element contributions lead to highly varying correction factors. Moreover, the absolute value of corresponding antidiffusive element contributions is small (due to almost constant function values) and scaling with an arbitrary correction factor has no meaningful influence. If correction factors are synchronized instead, large element contributions have to be scaled with such oscillating correction factors and, hence, produce inaccurate results with remarkable ripples even if local maximum principles are preserved.

So far, only the low order method was found to preserve the definiteness of the tensorial solution. In particular, local maximum principles for the range of eigenvalues are not satisfied if limited antidiffusive element contributions are added by taking the LED property for each tensor entry into account. These violations can be avoided by applying



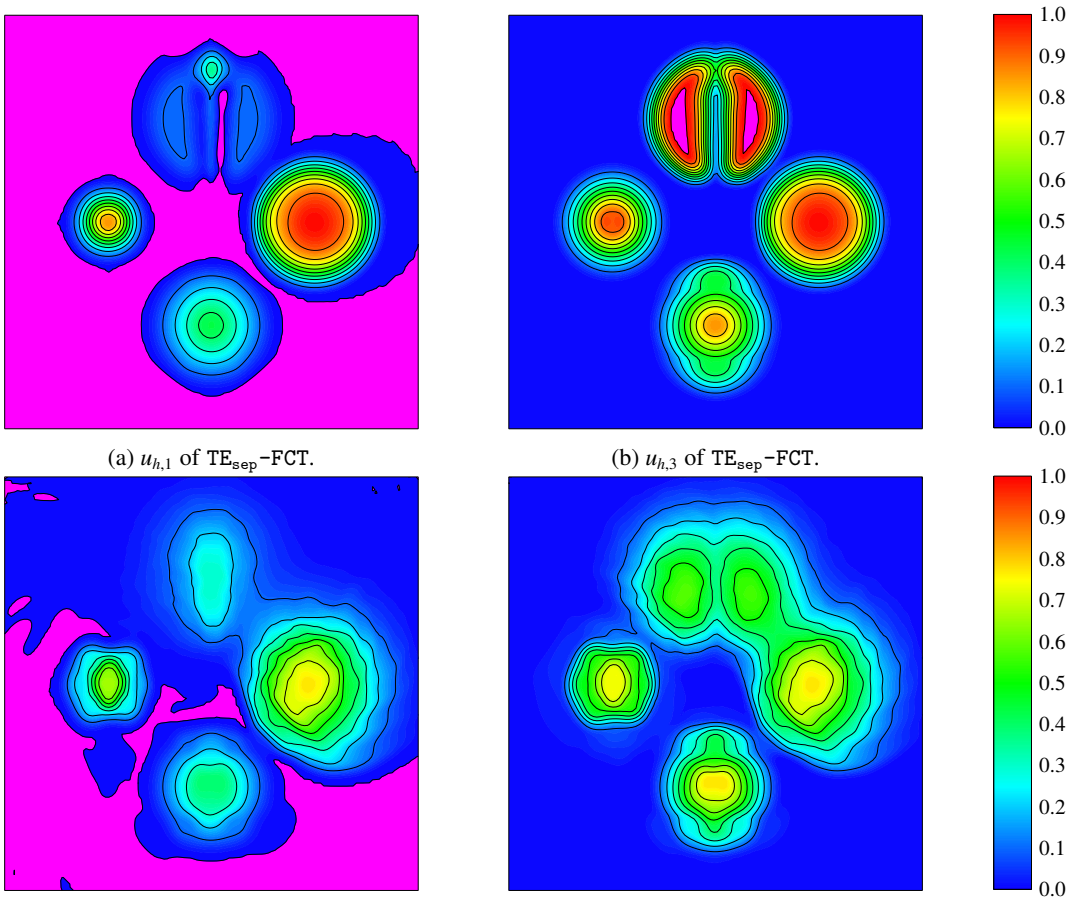
(a) $u_{h,1}$ of low order method ($\alpha^e \equiv 0$).

(b) $u_{h,3}$ of low order method ($\alpha^e \equiv 0$).

(c) $u_{h,1}$ of uncorrected method ($\alpha^e \equiv 1$).

(d) $u_{h,3}$ of uncorrected method ($\alpha^e \equiv 1$).

Figure 2: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$, range of eigenvalues of low order method and FCT algorithm with $\alpha^e \equiv 1$.



(a) $u_{h,1}$ of $\texttt{TE}_{\texttt{sep}}\texttt{-FCT}$.

(b) $u_{h,3}$ of $\texttt{TE}_{\texttt{sep}}\texttt{-FCT}$.

(c) $u_{h,1}$ of $\texttt{TE}_{\texttt{syn}}\texttt{-FCT}$.

(d) $u_{h,3}$ of $\texttt{TE}_{\texttt{syn}}\texttt{-FCT}$.

Figure 3: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128 + 1)^2$, range of eigenvalues of component-based FCT algorithms.

method	$L^1 - \|\cdot\|_F$	$L^2 - \|\cdot\|_F$	$L^1 - \|\cdot\|_2$	$L^2 - \|\cdot\|_2$	$u_{h,1}$	$u_{h,3}$
$\alpha^e \equiv 0$	1.984E−1	3.080E−1	1.487E−1	2.385E−1	−7.088E−13	0.558
$\alpha^e \equiv 1$	4.058E−2	1.015E−1	3.382E−2	8.860E−2	−6.907E−2	1.118
tr-FCT	3.674E−2	1.016E−1	3.057E−2	8.861E−2	−7.418E−3	1.006
ER_{min}-FCT	4.867E−2	1.177E−1	4.128E−2	1.020E−1	−3.088E−17	0.998
ER_{app}-FCT	4.386E−2	1.104E−1	3.713E−2	9.616E−2	−1.636E−11	1.000
ER_{inv}-FCT	4.378E−2	1.104E−1	3.708E−2	9.622E−2	−2.624E−10	1.000
ER_{reg}-FCT	4.378E−2	1.104E−1	3.708E−2	9.622E−2	−4.108E−10	1.000
TE_{sep}-FCT	3.646E−2	1.011E−1	3.036E−2	8.821E−2	−2.760E−2	1.003
TE_{syn}-FCT	1.208E−1	2.209E−1	9.294E−2	1.775E−1	−3.821E−3	0.776
tr-ER_{min}-FCT	4.941E−2	1.186E−1	4.185E−2	1.025E−1	−3.678E−17	0.994
tr-ER_{app}-FCT	4.494E−2	1.116E−1	3.798E−2	9.691E−2	−2.572E−11	1.000
tr-ER_{inv}-FCT	4.483E−2	1.115E−1	3.789E−2	9.682E−2	−2.017E−9	1.000
tr-ER_{reg}-FCT	4.483E−2	1.115E−1	3.789E−2	9.682E−2	−2.370E−9	1.000
tr-TE_{sep}-FCT	3.668E−2	1.017E−1	3.053E−2	8.866E−2	−1.074E−2	1.000
tr-TE_{syn}-FCT	1.206E−1	2.206E−1	9.280E−2	1.772E−1	−3.761E−3	0.787

Table 1: Solid body rotation: $T = 2\pi$, $\Delta t \approx 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, errors and range of eigenvalues of different numerical methods.

limiters for the range of eigenvalues as proposed in Sec. 5. As can be seen from Table 1, global maximum principles for the eigenvalues are satisfied even if L^1- and L^2-errors increase compared to TE_{sep}-FCT. However, this error behavior is not surprising, because TE_{sep}-FCT is the only algorithm that allows using individually chosen correction factors for each tensor entry and, hence, is most flexible.

In the following study, the different limiters for the range of eigenvalues are examined in detail.

7.2. Swirling flow

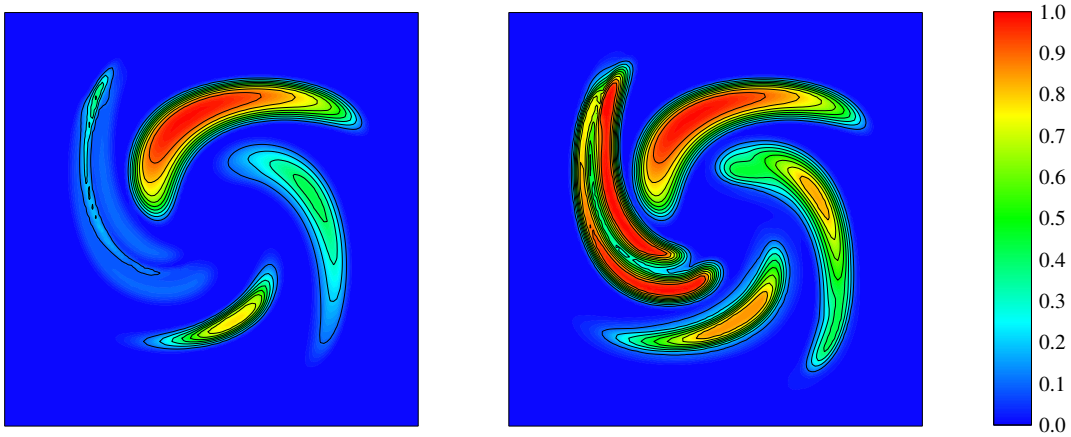
A more complex benchmark is given by the 'swirling deformation flow' as proposed by LeVeque [10]: The time-independent velocity field of the solid body rotation is replaced by

$$\mathbf{v}(x,y,t) \coloneqq \left(\sin^2(\pi x)\sin(2\pi y)g(t), -\sin^2(\pi y)\sin(2\pi x)g(t)\right)^\top, \qquad \text{div}\,(\mathbf{v}(x,y)) = 0,$$

where $g(t) = \cos(\pi t/T)$ describes the time dependency on the interval $0 \le t \le T \coloneqq \frac{3}{2}$. In this case, the velocity increases in a smooth manner and deforms the initial solution. After slowing down and changing its sign at $\frac{T}{2}$, when the maximal deformation is reached, the velocity field reverses such that the initial solution is recovered at the final time T. As seen in Fig. 4, the choice of the final time $T = \frac{3}{2}$ guarantees a reasonable amount of deformation taking the mesh size into account. Due to the complexity of the velocity field, which still yields the exact solution analytically, this benchmark is recommended to evaluate different limiting techniques. Furthermore, no (inflow) boundary condition has to be applied, because $\mathbf{v}$ vanishes on the boundary of the spatial domain $\partial\Omega$.

Fig. (5) shows the smallest and largest eigenvalue of the final solution using different limiting techniques for the range of eigenvalues. Results of ER_{reg}-FCT nearly coincide with the ones produced by ER_{inv}-FCT (see Tab. 1 and 2) and, hence, are omitted. The FCT algorithm ER_{min}-FCT calculates the correction factors by treating each term of the sum in (14) separately. This produces the most diffusive approximation, including peak clipping at the slotted cylinder for $u_{h,3}$. Nevertheless, the minimal eigenvalue $u_{h,1}$ (Fig. 5a) is comparable to the one of the other eigenvalue range limiters (Figs. 5c and 5e).

Optically, there is hardly any difference between ER_{app}-FCT and ER_{inv}-FCT for $u_{h,1}$ and $u_{h,3}$. Table 2 indicates that errors of ER_{app}-FCT grow and the global maximum of $u_{h,3}$ decreases. However, even in this benchmark, the differences are marginal and the increased computational cost of ER_{inv}-FCT seems to be not worthwhile. Additional numerical experiments have shown that even the FCT algorithm using linearization (20) produces no remarkable



(a) $u_{h,1}$ of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$. (b) $u_{h,3}$ of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$.

Figure 4: Swirling flow: $T = \frac{3}{4}$, $\Delta t = 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, range of eigenvalues of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$.

differences compared to $\mathtt{ER}_{\mathtt{app}}\mathtt{-FCT}$. However, eigenvalue range limiters using the Invariant criterion are less diffusive than adding antidiffusive element contributions after applying the min-min criterion.

If the trace limiter is applied before handling the range of eigenvalues, additional restrictions have to be taken into account and less antidiffusive element contributions can be added. Fig. 6 shows the solution using $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$ and the counterpart $\mathtt{tr\text{-}ER}_{\mathtt{inv}}\mathtt{-FCT}$. Once again, changes can be seen at the peaks of $u_{h,3}$ at the slotted cylinder and cone. Table 1 and 2 confirm the statement that each FCT algorithm gets more diffusive if the trace is constrained, too.

The proposed FCT algorithms for tensor quantities are at most first order accurate in the $L^1 - \|\cdot\|_F$ norm (see Table 3). Here, the experimental order of convergence (EOC) is computed using the formula [10]

$$\text{EOC} = \log\left(\frac{E(h_2)}{E(h_1)}\right)\log\left(\frac{h_2}{h_1}\right)^{-1},$$

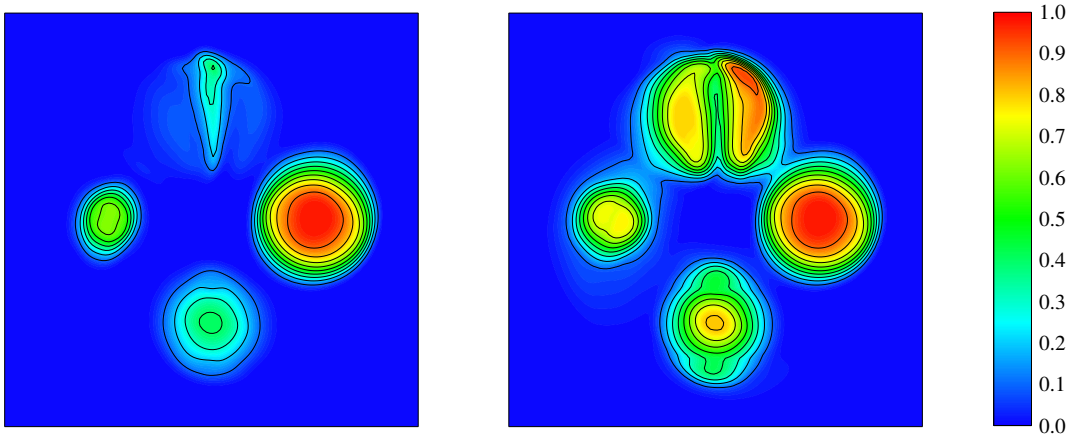
where $E(h_2)$ and $E(h_1)$ are numerical errors corresponding to the mesh sizes h_2 and h_1. If the tensorial solution is diagonal, the Invariant criterion coincides with the synchronized FCT algorithm for each tensor entry, i.e., $\mathtt{TE}_{\mathtt{syn}}\mathtt{-FCT}$, and the same effective order of convergence is attained.

7.3. Drawback of eigenvalue range limiters

As mentioned in Sec. 7.1, synchronized FCT algorithms produce inaccurate solutions if one quantity of interest is nearly constant. The proposed eigenvalue range limiters restrict the minimal and maximal eigenvalue separately and, hence, synchronize correction factors, too (to enforce positive semidefiniteness of $U_i^{\min,e}$ and $U_i^{\max,e}$). Therefore, if for instance the minimal eigenvalue is almost constant, eigenvalue range limiters scale antidiffusive element contributions in an unnatural manner and the maximal eigenvalue of the solution exhibits artificial ripples. Fig. 7 shows the eigenvalues of the swirling flow solution using $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$ with a modified initial solution: One eigenvalue of the slotted cylinder is set to zero, i.e., $u_2^{(4)} = 0$, which becomes the new (constant) minimal eigenvalue. Small variations occur numerically due to larger minimal eigenvalues at the other bodies of the profile. This produces artifacts such that the maximal eigenvalue of the slotted cylinder at the final time is comparable to the one of $\mathtt{TE}_{\mathtt{syn}}\mathtt{-FCT}$ while $u_{h,1}$ remains constant (compare Figs. 3d and 7b).

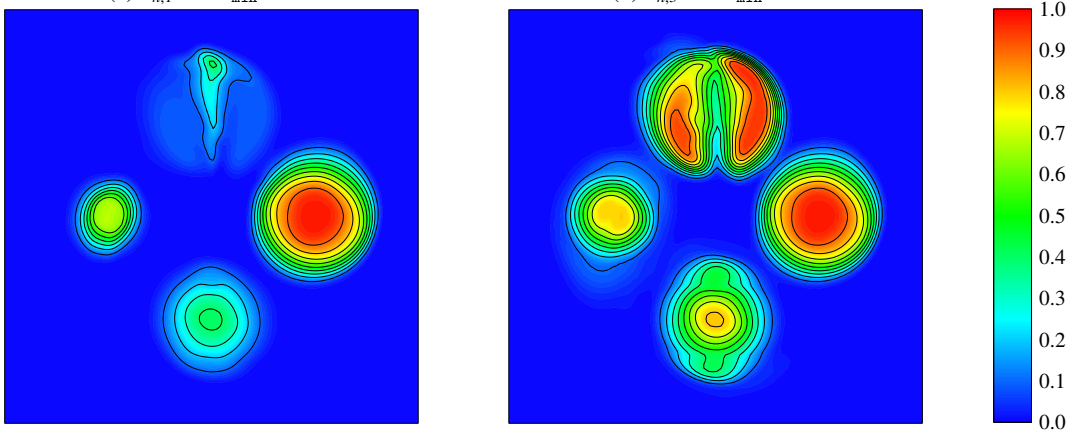
8. Conclusions

Preserving the definiteness of tensor quantities is mandatory in various applications of computational fluid dynamics. A natural way of guaranteeing this property is given by constraining the eigenvalues. As shown in Sec. 3, the low order method originally developed for scalar transport equations can be extended to tensorial variables in such a way



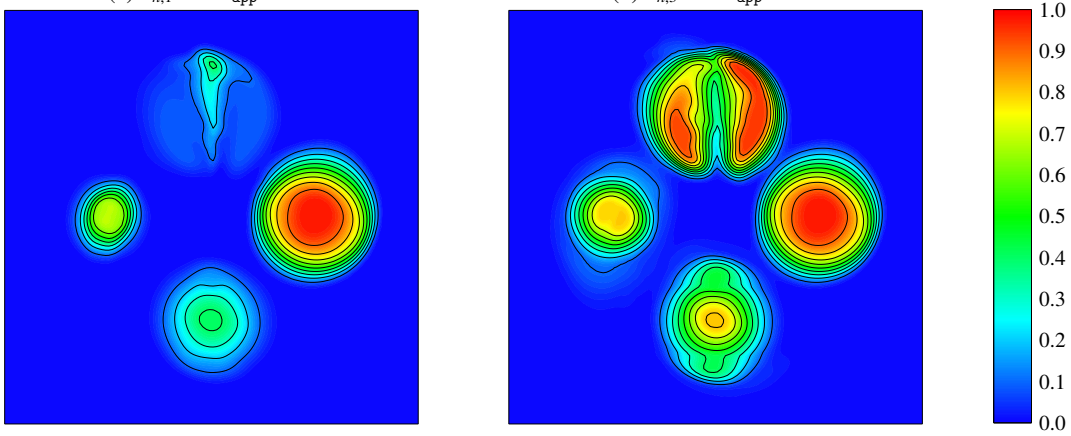
(a) $u_{h,1}$ of $\mathtt{ER}_{\mathtt{min}}$-$\mathtt{FCT}$.

(b) $u_{h,3}$ of $\mathtt{ER}_{\mathtt{min}}$-$\mathtt{FCT}$.



(c) $u_{h,1}$ of $\mathtt{ER}_{\mathtt{app}}$-$\mathtt{FCT}$.

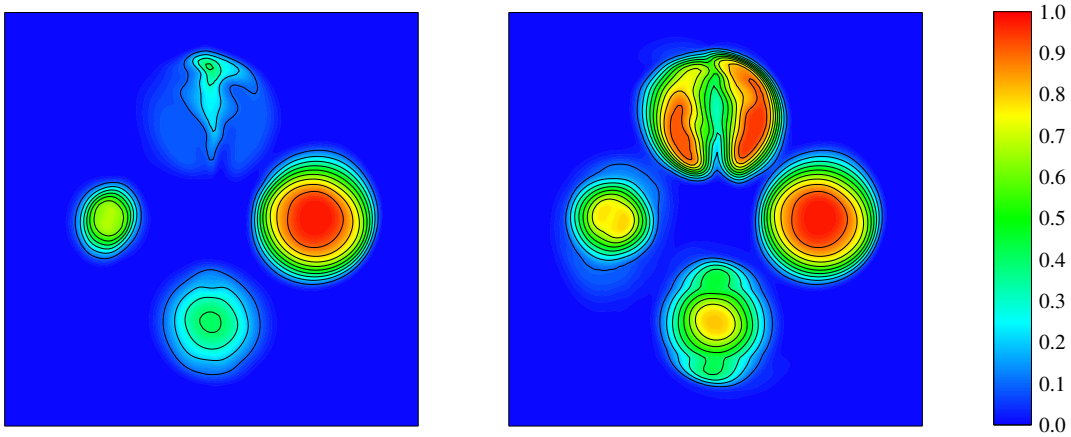
(d) $u_{h,3}$ of $\mathtt{ER}_{\mathtt{app}}$-$\mathtt{FCT}$.



(e) $u_{h,1}$ of $\mathtt{ER}_{\mathtt{inv}}$-$\mathtt{FCT}$.

(f) $u_{h,3}$ of $\mathtt{ER}_{\mathtt{inv}}$-$\mathtt{FCT}$.

Figure 5: Swirling flow: $T = \frac{3}{2}$, $\Delta t = 10^{-3}$, $N_{\mathrm{dof}} = (128 + 1)^2$, range of eigenvalues of different eigenvalue range limiters.



(a) $u_{h,1}$ of $\texttt{tr-ER}_{\texttt{inv}}\texttt{-FCT}$. (b) $u_{h,3}$ of $\texttt{tr-ER}_{\texttt{inv}}\texttt{-FCT}$.

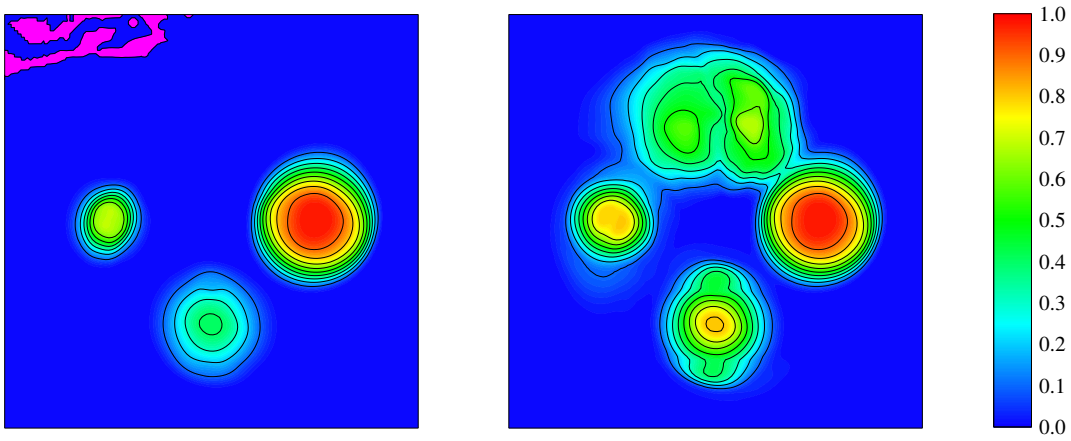
Figure 6: Swirling flow: $T = \frac{3}{2}$, $\Delta t = 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, range of eigenvalues of $\texttt{ER}_{\texttt{inv}}\texttt{-FCT}$ with preceding trace limiting, i.e., $\texttt{tr-ER}_{\texttt{inv}}\texttt{-FCT}$.

method	$L^1 - \|\cdot\|_F$	$L^2 - \|\cdot\|_F$	$L^1 - \|\cdot\|_2$	$L^2 - \|\cdot\|_2$	$u_{h,1}$	$u_{h,3}$
$\alpha^e \equiv 0$	1.742E−1	2.816E−1	1.322E−1	2.196E−1	3.497E−46	0.627
$\alpha^e \equiv 1$	4.363E−2	1.063E−1	3.585E−2	9.120E−2	−6.919E−2	1.108
tr-FCT	4.324E−2	1.116E−1	3.563E−2	9.551E−2	−1.622E−2	0.997
$\texttt{ER}_{\texttt{min}}\texttt{-FCT}$	5.896E−2	1.351E−1	4.972E−2	1.156E−1	1.027E−53	0.983
$\texttt{ER}_{\texttt{app}}\texttt{-FCT}$	4.882E−2	1.173E−1	4.073E−2	1.002E−1	2.234E−52	0.974
$\texttt{ER}_{\texttt{inv}}\texttt{-FCT}$	4.828E−2	1.167E−1	4.031E−2	9.981E−2	4.938E−54	0.983
$\texttt{ER}_{\texttt{reg}}\texttt{-FCT}$	4.828E−2	1.167E−1	4.031E−2	9.981E−2	−1.047E−13	0.983
$\texttt{TE}_{\texttt{sep}}\texttt{-FCT}$	4.181E−2	1.090E−1	3.435E−2	9.324E−2	−3.413E−2	0.999
$\texttt{TE}_{\texttt{syn}}\texttt{-FCT}$	1.155E−1	2.162E−1	8.945E−2	1.738E−1	−4.620E−3	0.798
$\texttt{tr-ER}_{\texttt{min}}\texttt{-FCT}$	5.977E−2	1.365E−1	5.032E−2	1.166E−1	1.837E−53	0.983
$\texttt{tr-ER}_{\texttt{app}}\texttt{-FCT}$	4.988E−2	1.194E−1	4.156E−2	1.018E−1	2.957E−52	0.974
$\texttt{tr-ER}_{\texttt{inv}}\texttt{-FCT}$	4.912E−2	1.183E−1	4.099E−2	1.010E−1	9.848E−54	0.983
$\texttt{tr-ER}_{\texttt{reg}}\texttt{-FCT}$	4.912E−2	1.183E−1	4.099E−2	1.010E−1	−7.173E−14	0.983
$\texttt{tr-TE}_{\texttt{sep}}\texttt{-FCT}$	4.305E−2	1.116E−1	3.551E−2	9.552E−2	−2.033E−2	0.983
$\texttt{tr-TE}_{\texttt{syn}}\texttt{-FCT}$	1.155E−1	2.162E−1	8.955E−2	1.741E−1	−4.517E−3	0.800

Table 2: Swirling flow: $T = 1.5$, $\Delta t = 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, errors and range of eigenvalues of different numerical methods.

N_{dof}	$L^1 - \|\cdot\|_F$		$L^2 - \|\cdot\|_F$		$L^1 - \|\cdot\|_2$		$L^2 - \|\cdot\|_2$	
	error	EOC	error	EOC	error	EOC	error	EOC
$(2^5+1)^2$	1.553E−1		2.607E−1		1.214E−1		2.093E−1	
$(2^6+1)^2$	8.843E−2	0.812	1.741E−1	0.582	7.221E−2	0.749	1.449E−1	0.531
$(2^7+1)^2$	4.828E−2	0.873	1.167E−1	0.577	4.031E−2	0.841	9.981E−2	0.537
$(2^8+1)^2$	2.550E−2	0.921	8.346E−2	0.484	2.185E−2	0.884	7.347E−2	0.442
$(2^9+1)^2$	1.294E−2	0.978	5.922E−2	0.495	1.125E−2	0.958	5.287E−2	0.475
$(2^{10}+1)^2$	7.236E−3	0.839	4.524E−2	0.389	6.336E−3	0.828	4.054E−2	0.383

Table 3: Swirling flow: $T = \frac{3}{2}$, $\Delta t = 10^{-3}2^{l-7}$, $N_{\text{dof}} = (2^l+1)^2$, errors and range of eigenvalues of $\texttt{ER}_{\texttt{inv}}\texttt{-FCT}$ on different mesh levels l.



(a) $u_{h,1}$ of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$. (b) $u_{h,3}$ of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$.

Figure 7: Swirling flow: $T = \frac{3}{2}$, $\Delta t = 10^{-3}$, $N_{\text{dof}} = (128+1)^2$, range of eigenvalues of $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$ with modified initial condition $u_2^{(4)} = 0$.

that LED properties for the range of eigenvalues are satisfied. This makes it possible to limit antidiffusive element contributions such that local maximum principles are valid for the entire FCT algorithm.

Corresponding limiting criteria are defined in Sec. 5 taking advantage of auxiliary tensors: If $U_i^{\min,e}$ and $U_i^{\max,e}$ are positive semidefinite, inequality constraints for the minimal and maximal eigenvalue are satisfied and the solution stays bounded. For this purpose, different frame invariant approaches are presented using simple estimates, regularization techniques, or by observing principal invariants.

The proposed treatments are validated by considering tensorial extensions of standard benchmarks for the linear transport equation (Sec. 7). In this study, frame dependent scalar FCT algorithms ($\mathtt{TE}_{\mathtt{syn}}\mathtt{-FCT}$ and $\mathtt{TE}_{\mathtt{sep}}\mathtt{-FCT}$) produce overshoots/undershoots and, especially, fail to preserve the definiteness of the exact solution. Eigenvalue range limiters enforce local maximum principles for the (range of) eigenvalues, but tend to produce rather diffusive results in specific benchmarks due to synchronizing correction factors (Sec. 7.3). However, reasonable accuracy is achieved by using $\mathtt{ER}_{\mathtt{app}}\mathtt{-FCT}$, while more expensive algorithms like $\mathtt{ER}_{\mathtt{inv}}\mathtt{-FCT}$ and $\mathtt{ER}_{\mathtt{reg}}\mathtt{-FCT}$ produce the most accurate results. The peak clipping effect is more pronounced if $\mathtt{ER}_{\mathtt{min}}\mathtt{-FCT}$ is used instead or if additional limiting is performed to enforce the maximum principle for the trace.

In summary, the eigenvalue range criterion represents a useful tool for constraining tensor quantities in FCT algorithms. The accuracy of limiting techniques can be improved by using sharper estimates or less restrictive bounds.

Acknowledgements

The author would like to thank Dmitri Kuzmin and Steffen Basting (both TU Dortmund University) for fruitful discussions on calculating reasonable correction factors for tensors.

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

References

[1] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works. *Journal of computational physics*, 11(1):38–69, 1973.

[2] R. J. Caron, H. Song, and T. Traynor. Positive Semidefinite Intervals for Matrix Pencils. Technical report, Internal Report, Department of Mathematics and Statistics, University of Windsor, Ontario, Canada, 2005.

[3] S. Gottlieb, C.-W. Shu, and E. Tadmor. Strong stability-preserving high-order time discretization methods. *SIAM Review*, 43:89–112, 2001.

[4] S. Gottlieb, D. Ketcheson, and C.-W. Shu. *Strong Stability Preserving Runge-Kutta and Multistep Time Discretizations*. World Scientific, 2011.

[5] K. M. Hasan, P. J. Basser, D. L. Parker, and A. L. Alexander. Analytical Computation of the Eigenvalues and Eigenvectors in DT-MRI. *Journal of Magnetic Resonance*, 152(1):41 – 47, 2001. ISSN 1090-7807. doi: http://dx.doi.org/10.1006/jmre.2001.2400. URL `http://www.sciencedirect.com/science/article/pii/S1090780701924000`.

[6] 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.

[7] 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`.

[8] D. Kuzmin and C. Lohmann. Synchronized slope limiting in discontinuous Galerkin methods for the equations of gas dynamics. Technical report, Fakultät für Mathematik, TU Dortmund, May 2016. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 541.

[9] D. Kuzmin and S. Turek. Flux correction tools for finite elements. *Journal of Computational Physics*, 175(2):525–558, 2002.

[10] R. J. LeVeque. High-resolution conservative algorithms for advection in incompressible flow. *SIAM Journal on Numerical Analysis*, 33(2): 627–665, 1996.

[11] R. Löhner, K. Morgan, J. Peraire, and M. Vahdati. Finite element flux-corrected transport (FEM–FCT) for the Euler and Navier–Stokes equations. *International Journal for Numerical Methods in Fluids*, 7(10):1093–1109, 1987.

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

[13] 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.

[14] 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.

[15] W. H. Press, B. P. Flannery, S. A. Teukolsky, and W. T. Vetterling. Numerical recipes in C: the art of scientific programming. *Cambridge University Press*, 10:408–412, 1992.

[16] 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.

[17] O. K. Smith. Eigenvalues of a Symmetric 3 ×3 Matrix. *Commun. ACM*, 4(4):168–, Apr. 1961. ISSN 0001-0782. doi: 10.1145/355578.366316. URL `http://doi.acm.org/10.1145/355578.366316`.

[18] P. Strobach. Solving cubics by polynomial fitting. *Journal of computational and applied mathematics*, 235(9):3033–3052, 2011.

[19] S. T. Zalesak. Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids. *Journal of computational physics*, 31(3):335–362, 1979.