No. 629 **April 2020**

An algebraic flux correction scheme facilitating the use of Newton-like solution strategies

C. Lohmann

ISSN: 2190-1767

An algebraic flux correction scheme facilitating the use of Newton-like solution strategies*

Christoph Lohmann†

April 01, 2020

Abstract

Building on recent advances in the analysis and design of algebraic flux correction (AFC) schemes, new differentiable limiter functions are constructed for efficient nonlinear solution strategies. The proposed scaling parameters are used to limit artificial diffusion operators incorporated into the residual of a high order target scheme to produce accurate and bound-preserving finite element approximations to hyperbolic problems. Due to this stabilization procedure, the occurring system becomes highly nonlinear and the efficient computation of corresponding solutions is a challenging task. The presented regularization approach makes the AFC residual twice continuously differentiable so that Newton's method converges quadratically for sufficiently good initial guesses. Furthermore, the performance of each nonlinear iteration is improved by expressing the Jacobian as the sum and product of matrices having the same sparsity pattern as the Galerkin system matrix. Eventually, the AFC methodology constructed for stationary problems is extended to transient test cases and validated numerically by applying it to several numerical benchmarks.

Keywords. Algebraic flux correction, Newton-like methods, Jacobian, linearity preservation, regularization

1 Introduction

The algebraic flux correction (AFC) methodology introduced by Kuzmin [Kuz07] is a well-established and promising approach for calculating bound-preserving finite element (FE) solutions to (steady) hyperbolic problems. The technique is based on an algebraically defined artificial diffusion operator which is incorporated into a high order target discretization to provably guarantee the validity of discrete maximum principles (DMPs). The so defined low order scheme is very diffusive and calls for a nonlinear correction to remove redundant diffusivity and to improve the accuracy of the resulting AFC solution. For this purpose, antidiffusive fluxes are multiplied by solution dependent scaling parameters and added to the residual of the

*This research was supported by the German Research Association (DFG) under grant KU 1530/23-1.

†Institute of Applied Mathematics (LS III), TU Dortmund University, Vogelpothsweg 87, D-44227 Dortmund, Germany (christoph.lohmann@math.tu-dortmund.de)

low order scheme. These so called correction factors are defined in such a way that the high order target discretization is recovered in regions where the solution is smooth. Whenever the discrete solution attains a local extremum, the method is designed to fall back to the low order scheme which prevents the occurrence of unphysical overshoots and undershoots. Furthermore, the validity of local and global discrete maximum principles is preserved [Bar+18; Loh19b] and guarantees bound-preserving numerical solutions.

In recent years, the AFC approach has received an increased interest and its theoretical foundations have been laid. For example, Barrenechea et al. [Bar+16] proved the existence of a solution and some a priori error estimates in the context of the convection-diffusion-reaction equation. An improved convergence result was shown in [Bar+18] for linearity preserving limiters which are designed to recover the high order target whenever the exact solution is (globally) linear. Representatives of this class of correction factors were constructed and successfully applied, e.g., in [BB17; Bar+17b; Bar+17a; Kuz+17]. Extensions of the AFC methodology to nonlinear systems of hyperbolic problems can found, e.g., in [Kuz07; Bad+19; Mab+18; Kuz20].

Solution dependent correction factors are mandatory for the construction of accurate and bound-preserving finite element schemes. However, they also make the corresponding system of equations highly nonlinear and extend the local domain of dependence. Therefore, robust iterative solution techniques as well as good preconditioners are desirable to prevent inordinately high computational costs for solving the resulting problems. First investigations in this direction were accomplished by Möller [Möl07; Möl08] who used finite differences to approximate the 'exact' Jacobian of the non-differentiable AFC residual. In [BB17], second order of convergence was locally achieved by exploiting the exact Jacobian of a regularized residual. Recently, Jha and John [JJ18; JJ19] observed that a low order approximation of the Jacobian was most efficient with respect to computing times, at least for the preconditioners considered in their numerical studies.

As pointed out above, the theoretical analysis of algebraic flux correction schemes and its application in different areas of computational fluid dynamics have imposed several new requirements on the design of correction factors. In particular, convergence of nonlinear solvers is a prerequisite for a numerical scheme to produce physics-compatible solutions. In what follows, we present a new way of defining the correction factors and discuss existing relationships to other AFC schemes. The proposed limiter combines several benefits and is well suited to calculate accurate and bound-preserving FE solutions to steady problems or when using implicit time integrators in the context of transient evolution equations. In particular, we are focusing on the following design criteria:

- *Linearity preservation* property on general meshes for accurate solutions;
- *Unique regularization* for efficient solution strategies;
- *Lipschitz continuity* of the residual to guarantee well-posedness;
- Validity of *discrete maximum principles* even for regularized problem;
- *Efficient calculation* of correction factors and corresponding Jacobian;
- Straightforward *extension to time dependent problems.*

This work is organized as follows: Section 2 presents the main idea of the AFC methodology for the steady advection-reaction equation and summarizes some notational conventions. In Section 3, the new approach to defining the nodal correction factors is presented. Furthermore, several properties are discussed to reduce the number of (optional) parameters. Then an efficient way to determine the exact Jacobian of the AFC residual is summarized in Section 4. A brief outlook into the calculation of bound-preserving finite element solutions to the transient advection equation is given in Section 5. Finally, some numerical results are presented to illustrate the potential of the proposed technique in Section 6.

2 Preliminaries

In this section, we outline the main idea of the algebraic flux correction methodology and introduce some symbols to be used throughout this work. For this purpose, we consider the steady hyperbolic model problem in which the solution $u : \Omega \to \mathbb{R}$ satisfies the advection-reaction equation in non-conservative form

$$\mathbf{v} \cdot \operatorname{grad}(u) + cu = s \qquad \text{in } \Omega, \tag{1a}$$

$$u = u_{\text{in}} \qquad \text{on } \Gamma_{\text{in}} \coloneqq \big\{\mathbf{s} \in \partial\Omega \,\big|\, \mathbf{n}(\mathbf{s}) \cdot \mathbf{v}(\mathbf{s}) < 0\big\}, \tag{1b}$$

where $u_{\text{in}} : \Gamma_{\text{in}} \to \mathbb{R}$ is the inflow boundary data and $\Omega \subset \mathbb{R}^d$, $d \in \{1, 2, 3\}$, denotes a polyhedral domain with outward normal vector $\mathbf{n} : \partial\Omega \to \mathbb{R}^d$. Furthermore, the velocity field is given by $\mathbf{v} : \bar{\Omega} \to \mathbb{R}^d$ while $c : \Omega \to \mathbb{R}$ and $s : \Omega \to \mathbb{R}$ are the reactivity parameter and external source, respectively. In the weak formulation of problem (1), a solution is sought so that

$$\int_\Omega \varphi \mathbf{v} \cdot \operatorname{grad}(u) + \int_\Omega c\varphi u + \int_{\Gamma_{\text{in}}} |\mathbf{v} \cdot \mathbf{n}| \varphi u = \int_\Omega s\varphi + \int_{\Gamma_{\text{in}}} |\mathbf{v} \cdot \mathbf{n}| \varphi u_{\text{in}} \quad \forall \varphi \tag{2}$$

is valid, where the inflow boundary condition is prescribed in a weak manner. Then the Galerkin discretization of (2) using continuous and linear finite elements reads: Find a solution $u_h = \sum_{j=1}^N u_j \varphi_j$ s.t.

$$\sum_{j=1}^N a_{ij} u_j = g_i \quad \forall i \in \{1, \dots, N\}, \tag{3}$$

where $u = (u_j)_{j=1}^N$ is the vector of degrees of freedom and $\varphi_1, \dots, \varphi_N : \Omega \to \mathbb{R}$ denote the linear Lagrange basis functions corresponding to the nodes $\mathbf{x}_1, \dots, \mathbf{x}_N$ of the triangulation $\mathcal{T}_h$. Furthermore, the entries of the system matrix $\mathcal{A} = (a_{ij})_{i,j=1}^N$ and right hand side vector $g = (g_i)_{i=1}^N$ are given by

$$a_{ij} \coloneqq \int_\Omega \varphi_i \mathbf{v} \cdot \operatorname{grad}(\varphi_j) + \int_\Omega c\varphi_i\varphi_j + \int_{\Gamma_{\text{in}}} |\mathbf{v} \cdot \mathbf{n}| \varphi_i\varphi_j \quad \forall i, j \in \{1, \dots, N\},$$

$$g_i \coloneqq \int_\Omega s\varphi_i + \int_{\Gamma_{\text{in}}} |\mathbf{v} \cdot \mathbf{n}| \varphi_i u_{\text{in}} \qquad \forall i \in \{1, \dots, N\}.$$

It is well known that the finite element function u_h corresponding to the solution of (3) is possibly polluted by spurious oscillations even if the exact solution to (1) may be smooth. In particular, overshoots and undershoots can violate physical bounds and result in unrealistic, or

even unstable, simulations. To stabilize the solution of (3), we introduce an artificial diffusion operator $\mathcal{D} = (d_{ij})_{i,j=1}^N$ so that $\mathcal{A} - \mathcal{D}$ is a so called M-matrix under some coercivity condition [Loh19b]. The entries of $\mathcal{D}$ are given by

$$d_{ij} \coloneqq \begin{cases} \max\{a_{ij}, 0, a_{ji}\} & : j \neq i, \\ -\sum_{k \neq i} d_{ik} & : j = i \end{cases} \quad \forall i, j \in \{1, \dots, N\} \tag{4}$$

and guarantee that the solution to

$$(\mathcal{A} - \mathcal{D})u = g \tag{5}$$

satisfies some discrete maximum principles. However, the low order method (5) is very diffusive by construction and calls for some nonlinear antidiffusive correction. For this purpose, the AFC problem based on some solution dependent correction factors $\alpha_{ij} = \alpha_{ji} : \mathbb{R}^N \to [0, 1]$ reads: Find $u \in \mathbb{R}^N$ s.t.

$$\mathscr{R}(u) \coloneqq \mathcal{A}u - \hat{f}(u) - g = 0, \tag{6}$$

where the entries of $\hat{f} = (\hat{f}_i)_{i=1}^N$ are given by

$$\hat{f}_i = \hat{f}_i(u) = \sum_{j=1, j \neq i}^N \big(1 - \alpha_{ij}(u)\big) \underbrace{d_{ij}(u_j - u_i)}_{=: f_{ij}} = \sum_{j=1, j \neq i}^N (1 - \alpha_{ij}) f_{ij} \quad \forall i \in \{1, \dots, N\}. \tag{7}$$

In what follows, we restrict ourselves to edge-based correction factors α_{ij} which are given by the product of two auxiliary quantities β_i and β_j, i.e.,

$$\alpha_{ij} \coloneqq \beta_i \beta_j \quad \forall i, j,\, i \neq j. \tag{8}$$

These nodal correction factors $\beta_i : \mathbb{R}^N \to [0, 1]$ are responsible for the validity of discrete maximum principles in node i and, for this, should vanish whenever u_i is a local extremum. Otherwise, they are designed to be as close as possible to one to produce accurate results. In the literature, one can find different improvements of (8) which still suffice to guarantee that the AFC scheme (6) is bound-preserving. For example, the correction factor α_{ij} might be chosen as the minimum of the nodal quantities or depend on the sign of $u_i - u_j$ and/or a_{ij}. However, we only focus on the definition of the nodal correction factors β_i and do not consider further specializations.

For the analysis of the algebraic flux correction scheme (6), we introduce the index sets

$$\begin{aligned} \mathcal{N}_i &\coloneqq \big\{ j \in \{1, \dots, N\} \,\big|\, \operatorname{supp}(\varphi_i) \cap \operatorname{supp}(\varphi_j) \neq \emptyset \big\} && \forall i \in \{1, \dots, N\}, \\ \mathcal{E}_i &\coloneqq \{ e \in \{1, \dots, E\} \,|\, \operatorname{supp}(\varphi_i) \cap K^e \neq \emptyset \} && \forall i \in \{1, \dots, N\}, \\ \mathcal{N}^e &\coloneqq \{ j \in \{1, \dots, N\} \,|\, \operatorname{supp}(\varphi_j) \cap K^e \neq \emptyset \} && \forall e \in \{1, \dots, E\}, \end{aligned}$$

where $K^e \subset \bar{\Omega}$, $e \in \{1, \dots, E\}$, are the elements of the triangulation $\mathcal{T}_h$ under consideration. Thus, the support of the Lagrange basis function φ_i is given by the patch of elements containing node i, i.e., $\varpi_i \coloneqq \bigcup_{e \in \mathcal{E}_i} K^e$, and $a_{ij} = d_{ij} = 0$ for all $j \notin \mathcal{N}_i$.

3 Nodal correction factors

In this section, we introduce a new way of defining nodal correction factors and show that for certain values of the involved free parameters the result is bounded above and below by the limiting functions of two approaches known from the literature. After explaining the ramifications of this relationship, we present an efficient way to calculate the weights which are responsible for the linearity preservation property. Finally, a suitable regularization is considered which preserves the validity of discrete maximum principles and guarantees the Lipschitz continuity of the residual no matter how the regularization parameter is chosen.

3.1 Definition

The AFC schemes considered in [Bar+17a; KS17; Kuz+17] use nodal correction factors of the form

$$\beta_i^{(1)} = \begin{cases} 1 - \left(\frac{\left|\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}(u_i-u_k)\right|}{\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|} \right)^{2(p+1)} & : \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k| > 0, \\ 1 & : \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k| = 0, \end{cases} \tag{9}$$

where $p \in \{-\frac{1}{2}, 0, \frac{1}{2}, 1, \ldots\}$ is a parameter that controls the accuracy of the AFC solution. The weights $(w_{ik})_{k\in\mathcal{N}_i\backslash\{i\}}$ are chosen to be not smaller than one and can be determined so that the limiter is linearity preserving [Bar+17a; KS17; Kuz+17]. The special treatment of a vanishing denominator in the definition of $\beta_i^{(1)}$ is only required to prevent division by zero. For all $j \in \mathcal{N}_i \setminus \{i\}$ s.t. $u_i = u_j$, we have $\alpha_{ij} f_{ij} = \alpha_{ij} d_{ij}(u_j - u_i) = 0$ no matter how α_{ij} is chosen. Therefore, we omit this special case in what follows and, without loss of generality, consider

$$\beta_i^{(1)} = 1 - \left(\frac{\left|\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}(u_i-u_k)\right|}{\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|} \right)^{2(p+1)} \tag{10}$$

instead of (9). Practically, the same behavior can be achieved by adding a small positive constant to the denominator of the ratio.

Denoting the nonnegative part of $x \in \mathbb{R}$ by $|x|_+ = \max(0, x)$ and using

$$s_i^+ = \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|_+, \quad s_i^- = \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_k-u_i|_+,$$

the correction factor $\beta_i^{(1)}$ can be equivalently expressed by

$$\begin{aligned} \beta_i^{(1)} &= 1 - \left(\frac{\left(\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}(u_i-u_k)\right)^2}{\left(\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|\right)^2} \right)^{p+1} \\ &= 1 - \left(1 - \frac{4\left(\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|_+\right)\left(\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_k-u_i|_+\right)}{\left(\sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|u_i-u_k|\right)^2} \right)^{p+1} \\ &= 1 - \left(1 - \frac{4 s_i^+ s_i^-}{(s_i^+ + s_i^-)^2} \right)^{p+1}. \end{aligned} \tag{11}$$

We can easily see that the ratio in (11) vanishes whenever

$$u_i = u_i^{\max} := \max_{j\in\mathcal{N}_i\setminus\{i\}} u_j \quad \text{or} \quad u_i = u_i^{\min} := \min_{j\in\mathcal{N}_i\setminus\{i\}} u_j$$

and, hence, we have $\beta_i^{(1)} = 0$ if u_i is a local extremum. Therefore, the validity of discrete maximum principles is satisfied as proved in [Loh19b, Section 4.5.1]. In contrast to (10), this property holds true if a different quantity $\tilde{u}$ is used to evaluate the denominator of the fraction. We will exploit this property in Section 5 for the construction of a bound-preserving AFC scheme for time dependent problems. Additionally, quantity $\beta_i^{(1)}$ can be augmented by another parameter $q \geqslant 1$ to control the accuracy of the AFC scheme:

$$\beta_i^{(2)} = 1 - \max\left(0, 1 - q\frac{4s_i^+ s_i^-}{(s_i^+ + s_i^-)^2}\right)^{p+1} \tag{12}$$

Actually, we have $\beta_i^{(2)} \geqslant \beta_i^{(1)}$ and the nodal correction factor increases monotonically with q if $s_i^+ s_i^- > 0$. Furthermore, the correction factors coincide with the non-regularized quantities presented in [Loh19a, Section 7.2] if the weights w_{ik} are set to d_{ik} for all $k \in \mathcal{N}_i \setminus \{i\}$.

Next, we establish a relation between $\beta_i^{(2)}$ and the correction factors considered in [Bar+17b, Section 4] which are defined by

$$\begin{aligned}
Q_i^+ &:= q_i'|d_{ii}|(u_i^{\max} - u_i), \quad Q_i^- := q_i'|d_{ii}|(u_i - u_i^{\min}) && \forall i \in \{1,\dots,N\},\\
P_i^+ &:= \sum_{j\in\mathcal{N}_i} d_{ij}\max(0, u_i - u_j), \quad P_i^- := \sum_{j\in\mathcal{N}_i} d_{ij}\max(0, u_j - u_i) && \forall i \in \{1,\dots,N\},\\
\beta_i^{(3),+} &:= \min\Big(1, \frac{Q_i^+}{P_i^+}\Big), \quad \beta_i^{(3),-} := \min\Big(1, \frac{Q_i^-}{P_i^-}\Big) && \forall i \in \{1,\dots,N\},\\
\beta_{ij}^{(3)} &:= \begin{cases} \beta_i^{(3),+} & : u_i > u_j,\\ 1 & : u_i = u_j,\\ \beta_i^{(3),-} & : u_i < u_j \end{cases} \quad \alpha_{ij}^{(3)} := \min(\beta_{ij}^{(3)}, \beta_{ji}^{(3)}) && \forall i,j \in \{1,\dots,N\}, j \neq i
\end{aligned}$$

for some $q_i' > 0$. Without loss of generality, let us assume that $u_i > u_j$ is valid. Then the auxiliary quantity $\beta_{ij}^{(3)}$ satisfies

$$\begin{aligned}
\beta_{ij}^{(3)} = \beta_i^{(3),+} &= \min\left(1, \frac{q_i'|d_{ii}|(u_i^{\max} - u_i)}{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_i - u_k|_+}\right)\\
&\geqslant \min\left(1, \frac{q_i'\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_k - u_i|_+}{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_i - u_k|}\right)\\
&\geqslant \min\left(1, q_i'\frac{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_k - u_i|_+}{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_i - u_k|}\frac{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_i - u_k|_+}{\sum_{k\in\mathcal{N}_i\setminus\{i\}} d_{ik}|u_i - u_k|}\right)
\end{aligned}$$

and, hence, is bounded below by $\beta_i^{(2)}$ for $p = 0$, $q_i' = 4q$, and $w_{ik} = d_{ik}$ for all $k \in \mathcal{N}_i \setminus \{i\}$. Therefore, the nodal correction factor $\beta_i^{(2)}$ is bounded above and below by two previously published limiters and the resulting AFC scheme can be analyzed in the same way. In what follows, it turns out that definition $\beta_i := \beta_i^{(2)}$ complies with all design criteria written down in the introduction.

3.2 Linearity preservation

In this section, we define the weights of the correction factors so that the limiter under investigation is linearity preserving no matter how the parameters $p \in \mathbb{N}_0$ and $q \geqslant 1$ are chosen. The values of $(w_{ik})_{k\in\mathcal{N}_i\setminus\{i\}}$ are defined in an a priori manner by geometric interpretations similar to the ones considered in [BB17, Section 3]. For the special case of a two dimensional domain, we also present a formula which only depends on the location of the nodes and does not require any further information regarding the triangulation. We will see that the construction of positive weights guaranteeing the linearity preservation property is only possible for nodal correction factors which are not associated to nodes located on the boundary. Therefore, all other weights are set to one, the value which is attained at interior nodes of a symmetric mesh.

To derive conditions for the linearity preservation property, let us assume that the finite element function u_h is (locally) linear, but not constant to avoid divisions by zero. Then the AFC scheme (6) coincides with its high order counterpart (3) if all nodal correction factors β_i are equal to one. This is the case if

$$\begin{aligned} 1 \leqslant q\frac{4s_i^+s_i^-}{(s_i^+ + s_i^-)^2} &= 1 + \frac{4(q-1)s_i^+s_i^-}{(s_i^+ + s_i^-)^2} - \frac{(s_i^+ - s_i^-)^2}{(s_i^+ + s_i^-)^2} \\ &= 1 + \frac{4(q-1)s_i^+s_i^-}{(s_i^+ + s_i^-)^2} - \frac{\left(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}(u_i - u_k)\right)^2}{\left(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i - u_k|\right)^2} \\ &= 1 + \frac{4(q-1)s_i^+s_i^-}{(s_i^+ + s_i^-)^2} - \frac{\left(\nabla u_h \cdot \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}(\mathbf{x}_i - \mathbf{x}_k)\right)^2}{\left(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i - u_k|\right)^2} \end{aligned} \tag{13}$$

because $u_i - u_k = \nabla u_h \cdot (\mathbf{x}_i - \mathbf{x}_k)$ for all $k \in \mathcal{N}_i \setminus \{i\}$ due to the fact that u_h is linear. Thus, we have $\beta_i = 1$ if $(w_{ik})_{k\in\mathcal{N}_i\setminus\{i\}}$ solves [Bar+17a, (2.14)]

$$\mathbf{0} = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}(\mathbf{x}_i - \mathbf{x}_k). \tag{14}$$

It is easy to verify that this problem possesses at least one solution with positive weights if

$$\rho_i \coloneqq \min_{\mathbf{x}\in\partial\varpi_i^{\mathrm{conv}}} \|\mathbf{x} - \mathbf{x}_i\|_2 > 0, \quad \varpi_i^{\mathrm{conv}} \coloneqq \Big\{ \sum_{j\in\mathcal{N}_i} \lambda_j\mathbf{x}_j \;\Big|\; \sum_{j\in\mathcal{N}_i} \lambda_j = 1,\ \lambda_j \geqslant 0\ \forall j \in \mathcal{N}_i \Big\}. \tag{15}$$

In particular, condition (14) is satisfied for $w_{ik} = 1$, $k \in \mathcal{N}_i \setminus \{i\}$, if the considered mesh is symmetric, i.e., $\forall j \in \mathcal{N}_i \setminus \{i\}\ \exists k_j \in \mathcal{N}_i \setminus \{i\}$ s.t. $\mathbf{x}_{k_j} = 2\mathbf{x}_i - \mathbf{x}_j$. However, if node $\mathbf{x}_i$ is located on the boundary, condition (15) might be violated and there must not exist a solution to (14) with positive entries. Therefore, we set $w_{ik} = 1$ for all $i \in \{1, \dots, N\}$ s.t. $\mathbf{x}_i \in \partial\Omega$ and all $k \in \mathcal{N}_i \setminus \{i\}$.

In the literature, one can find different approaches to define some weights satisfying (14). For example, the use of an optimization problem was mentioned in [Bar+17a, (2.16)] while Kuzmin et al. [Kuz+17, Section 7] defined w_{ik} using reconstructed gradients that depend on the unknown degrees of freedom. Inspired by the reconstruction technique considered in [BB17, Section 3], we determine the values of $w_{ik} \geqslant 1$ for all $k \in \mathcal{N}_i \setminus \{i\}$ by

$$w_{ik} = 1 + \lambda_i^{-1}\varphi_k(\mathbf{x}_i + \lambda_i\mathbf{y}_i) \quad \forall k \in \mathcal{N}_i \setminus \{i\}, \qquad \mathbf{y}_i \coloneqq \sum_{k\in\mathcal{N}_i\setminus\{i\}} (\mathbf{x}_i - \mathbf{x}_k), \tag{16a}$$

where the parameter $\lambda_i > 0$ is defined as follows (see Fig. 1):

$$\lambda_i = \begin{cases} 1 & : \mathbf{y}_i = \mathbf{0}, \\ \max\{\lambda_i > 0 \,|\, \mathbf{x}_i + \lambda_i \mathbf{y}_i \in \partial\varpi_i\} & : \mathbf{y}_i \neq \mathbf{0} \end{cases} \tag{16b}$$

Indeed, the weights given by (16) solve (14) because

$$\begin{aligned} \sum_{k \in \mathcal{N}_i \setminus \{i\}} w_{ik}(\mathbf{x}_i - \mathbf{x}_k) &= \mathbf{y}_i + \lambda_i^{-1} \sum_{k \in \mathcal{N}_i \setminus \{i\}} \varphi_k(\mathbf{x}_i + \lambda_i \mathbf{y}_i)(\mathbf{x}_i - \mathbf{x}_k) \\ &= \mathbf{y}_i + \lambda_i^{-1} \mathbf{x}_i \underbrace{\sum_{k \in \mathcal{N}_i} \varphi_k(\mathbf{x}_i + \lambda_i \mathbf{y}_i)}_{=1} - \lambda_i^{-1} \underbrace{\sum_{k \in \mathcal{N}_i} \varphi_k(\mathbf{x}_i + \lambda_i \mathbf{y}_i)\mathbf{x}_k}_{=\mathbf{x}_i + \lambda_i \mathbf{y}_i} = \mathbf{0} \end{aligned}$$

due to the fact that $\mathbf{x}_i + \lambda_i \mathbf{y}_i \in \varpi_i$ and

$$\sum_{k \in \mathcal{N}_i} \varphi_k(\mathbf{x}) = 1, \quad \sum_{k \in \mathcal{N}_i} \varphi_k(\mathbf{x}) \mathbf{x}_k = \mathbf{x} \qquad \forall \mathbf{x} \in \varpi_i$$

holds for the (multi-)linear Lagrange basis functions $\varphi_1, \ldots, \varphi_N$.

Note that $\varphi_k(\mathbf{x}) = 0$ whenever $\mathbf{x} \notin \varpi_k$ and, hence, the weight w_{ik} can only be greater than one if $k \in \mathcal{N}^{e_i}$ for some $e_i \in \mathcal{E}_i$ s.t. $\mathbf{x}_i + \lambda_i \mathbf{y}_i \in K^{e_i}$. In two dimensions, the quantity $\mathbf{x}_i + \lambda_i \mathbf{y}_i$ is located on an edge with endpoints $\mathbf{x}_{k_i^+}$ and $\mathbf{x}_{k_i^-}$, $k_i^+, k_i^- \in \mathcal{N}_i \setminus \{i\}$, whenever $\mathbf{y}_i \neq \mathbf{0}$. Therefore, we have $\varphi_k(\mathbf{x} + \lambda_i \mathbf{y}_i) = 0$ for all $k \in \mathcal{N}_i \setminus \{i, k_i^+, k_i^-\}$ and, hence, at most two weights can be greater than one (see Fig. 1). If ϖ_i is star-shaped with respect to $\mathbf{x}_i$, the corresponding indices $k_i^+, k_i^- \in \mathcal{N}_i \setminus \{i\}$ can be determined by

$$k_i^{\pm} := \operatorname*{arg\,max}_{j \in \mathcal{N}_i \setminus \{i\}} (\pm z_{ij}), \qquad z_{ij} = \begin{cases} 3 + \frac{\mathbf{y}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{\|\mathbf{y}_i\| \|\mathbf{x}_j - \mathbf{x}_i\|} & : \mathbf{y}_{i,2}(\mathbf{x}_{j,1} - \mathbf{x}_{i,1}) \geqslant \mathbf{y}_{i,1}(\mathbf{x}_{j,2} - \mathbf{x}_{i,2}), \\ 1 - \frac{\mathbf{y}_i \cdot (\mathbf{x}_j - \mathbf{x}_i)}{\|\mathbf{y}_i\| \|\mathbf{x}_j - \mathbf{x}_i\|} & : \mathbf{y}_{i,2}(\mathbf{x}_{j,1} - \mathbf{x}_{i,1}) < \mathbf{y}_{i,1}(\mathbf{x}_{j,2} - \mathbf{x}_{i,2}) \end{cases} \tag{17}$$

and maximize/minimize the angle between the edges given by $\mathbf{x}_i$ and $\mathbf{x}_k$ as well as $\mathbf{x}_i$ and $\mathbf{x}_i + \mathbf{y}_i$ (measured counterclockwise). According to this observation, we have

$$\begin{aligned} \mathbf{y}_i = \sum_{k \in \mathcal{N}_i \setminus \{i\}} (\mathbf{x}_i - \mathbf{x}_k) &= \sum_{k \in \mathcal{N}_i \setminus \{i\}} w_{ik}(\mathbf{x}_i - \mathbf{x}_k) - (1 - w_{ik_i^+})(\mathbf{x}_i - \mathbf{x}_{k_i^+}) - (1 - w_{ik_i^-})(\mathbf{x}_i - \mathbf{x}_{k_i^-}) \\ &= (w_{ik_i^+} - 1)(\mathbf{x}_i - \mathbf{x}_{k_i^+}) + (w_{ik_i^-} - 1)(\mathbf{x}_i - \mathbf{x}_{k_i^-}) \end{aligned}$$

and the weights given by (16) satisfy

$$w_{ik} = 1 \quad \forall k \notin \mathcal{N}_i \setminus \{i, k_i^+, k_i^-\}, \qquad (w_{ik_i^+} - 1)(\mathbf{x}_i - \mathbf{x}_{k_i^+}) + (w_{ik_i^-} - 1)(\mathbf{x}_i - \mathbf{x}_{k_i^-}) = \mathbf{y}_i. \tag{18}$$

Therefore, the values of $(w_{ik})_{k \in \mathcal{N}_i \setminus \{i\}}$ can be determined by (17) and (18) without using the definition of the basis functions $\varphi_1, \ldots, \varphi_N$ or the connectivity graph of the triangulation. In particular, there is no need to compute λ_i and an associated element index $e_i \in \mathcal{E}_i$ which makes the computation very efficient.

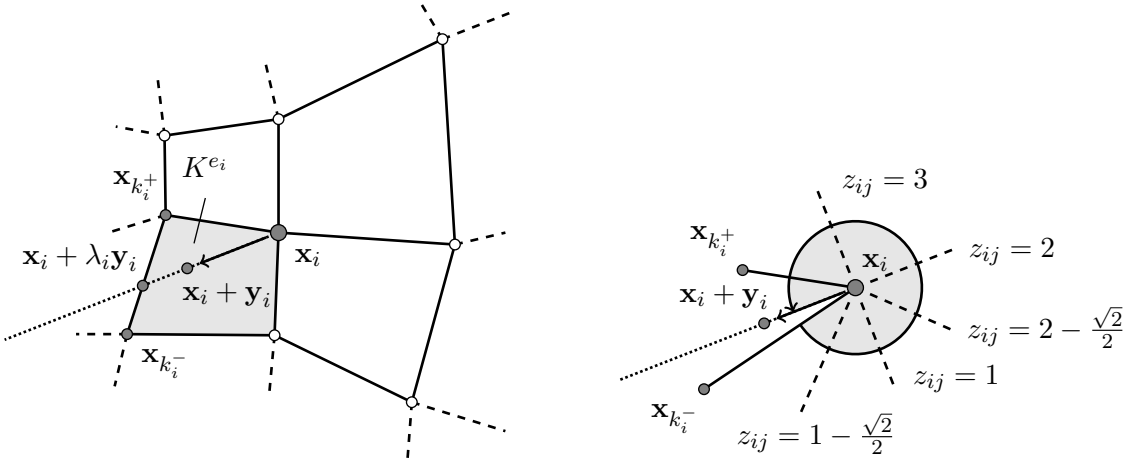


Figure 1: Calculation of weights $(w_{ik})_{k\in\mathcal{N}_i\setminus\{i\}}$ illustrated in two dimensional domain.

3.3 Regularization

The limiter defined above exploits the nonnegative part of scalar differences to make the scheme bound-preserving. However, this cut-off function is globally continuous but not in C^1 and prevents the AFC residual from being differentiable. Therefore, iterative solvers may not converge with the optimal order and the use of regularization techniques is desirable. For this purpose, smoothed correction factors should be designed to guarantee that the resulting AFC scheme is still bound-preserving and at least as stable as the original method.

In what follows, we consider correction factors $\beta_{i,\varepsilon}$ which can be written in the form

$$\beta_{i,\varepsilon} \coloneqq 1 - \max\left(0, 1 - q\frac{4s_i^{+,\delta_i}s_i^{-,\delta_i}}{(s_i^{+,\delta_i} + s_i^{-,\delta_i} + \gamma_i)^2}\right)^{p+1}, \tag{19a}$$

$$s_i^{+,\delta_i} = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i - u_k|_{+,\delta_i}, \quad s_i^{-,\delta_i} = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_k - u_i|_{+,\delta_i}, \tag{19b}$$

where $\delta_i, \gamma_i > 0$ will be defined in terms of the regularization parameter $\varepsilon > 0$ in such a way that $\lim_{\varepsilon\searrow 0}\beta_{i,\varepsilon} = \beta_i =: \beta_{i,0}$. In (19b), the nonnegative part of a scalar quantity $|\cdot|_+$ is approximated by [Loh19a]

$$|x|_{+,\delta_i} \coloneqq \frac{|x|_+^{\tilde{p}+1}}{|x|^{\tilde{p}} + \delta_i} \quad \forall x \in \mathbb{R}, \tag{20}$$

for some $\tilde{p} \in \mathbb{N}$. It is easy to verify that $|\cdot|_{+,\delta_i}$ uniformly converges to $|\cdot|_+$ because

$$|x - \delta_i'|_+ \leqslant |x|_{+,\delta_i} \leqslant |x|_+ \quad \forall x \in \mathbb{R}, \qquad \delta_i' \coloneqq \sqrt[\tilde{p}]{\delta_i(\tilde{p}-1)^{\tilde{p}-1}\tilde{p}^{-\tilde{p}}} \leqslant \sqrt[\tilde{p}]{\delta_i}. \tag{21}$$

Property (21) also guarantees that the fraction in (19a) vanishes at a local extremum and the AFC scheme (6) stays bound-preserving when β_i is replaced by $\beta_{i,\varepsilon}$ for $\varepsilon > 0$.

Furthermore, the approximation $|\cdot|_{+,\delta_i}$ is $\tilde{p}$-times continuously differentiable due to the fact that

$$\frac{\partial}{\partial x}|x|_+^{p'+1} = (p'+1)|x|_+^{p'} \quad \forall x \quad \forall p' \in \mathbb{N} \tag{22}$$

and the first derivative is bounded above and below by

$$0 \leqslant \frac{\partial |x|_{+,\delta_i}}{\partial x} = \frac{|x|_+^{\tilde{p}}}{|x|^{\tilde{p}} + \delta_i}\Big(1 + \frac{\delta_i \tilde{p}}{|x|^{\tilde{p}} + \delta_i}\Big) \leqslant \frac{1}{4}(\tilde{p}+1)^2 \tilde{p}^{-1} \quad \forall x. \tag{23}$$

If $\gamma_i \geqslant \delta_i'$ is satisfied, property (23) can be used to show that the AFC residual is Lipschitz continuous with a Lipschitz constant bounded above independently of δ_i and γ_i. For the sake of simplicity, we only show this result for $q = 1$ and $p = 0$ (cf. [Bar+16]). The general statement can be similarly verified by following the ideas presented in [Loh19b, Lemma 4.76]. The crucial part of the proof is the use of the lower bound for γ_i to show the nonnegativity of

$$\begin{aligned}
B_{ij} &\coloneqq s_i^{+,\delta_i} + s_i^{-,\delta_i} + \gamma_i - |u_i - u_j| \\
&= \sum_{k \in \mathcal{N}_i \setminus \{i\}} w_{ik}\big(|u_k - u_i|_{+,\delta_i} + |u_i - u_k|_{+,\delta_i}\big) + \gamma_i - |u_i - u_j| \\
&= \sum_{k \in \mathcal{N}_i \setminus \{i,j\}} w_{ik}\big(|u_k - u_i|_{+,\delta_i} + |u_i - u_k|_{+,\delta_i}\big) \\
&\quad + w_{ij}\big(|u_j - u_i|_{+,\delta_i} + |u_i - u_j|_{+,\delta_i}\big) + \gamma_i - |u_i - u_j| \\
&\geqslant \big(|u_j - u_i|_{+,\delta_i} + |u_i - u_j|_{+,\delta_i}\big) + \delta_i' - |u_i - u_j| \geqslant 0 \quad \forall j \in \mathcal{N}_i \setminus \{i\}
\end{aligned}$$

by virtue of (21). Then $u \mapsto A_i^{\pm}(u) \coloneqq 2 s_i^{\pm,\delta_i}$ and $u \mapsto B_{ij}(u)$ are nonnegative functions which satisfy the Lipschitz condition for constants that do not depend on δ_i and γ_i because the first derivative of $|\cdot|_{x,\delta_i}$ is uniformly bounded according to (23). Thus, for all $u, \bar{u} \in \mathbb{R}^N$ and $\zeta_i^{\pm} \coloneqq 2 s_i^{\pm,\delta_i}(s_i^{+,\delta_i} + s_i^{-,\delta_i} + \gamma_i)^{-1} \leqslant 2$, we have

$$\begin{aligned}
|\zeta_i^+ - \bar{\zeta}_i^+|\,|u_j - u_i| &= \Big|\frac{A_i^+}{B_{ij} + |u_j - u_i|} - \frac{\bar{A}_i^+}{\bar{B}_{ij} + |\bar{u}_j - \bar{u}_i|}\Big||u_j - u_i| \\
&\leqslant \frac{|A_i^+ - \bar{A}_i^+|}{B_{ij} + |u_j - u_i|}|u_j - u_i| + \frac{\bar{A}_i}{\bar{B}_{ij} + |\bar{u}_j - \bar{u}_i|}\frac{|B_{ij} - \bar{B}_{ij}| + \big||u_j - u_i| - |\bar{u}_j - \bar{u}_i|\big|}{B_{ij} + |u_j - u_i|}|u_j - u_i| \\
&\leqslant |A_i^+ - \bar{A}_i^+| + 2\big(|B_{ij} - \bar{B}_{ij}| + |u_j - \bar{u}_j| + |u_i - \bar{u}_i|\big),
\end{aligned}$$

as well as

$$\begin{aligned}
&\Big|\alpha_{ij}(u)(u_j - u_i) - \alpha_{ij}(\bar{u})(\bar{u}_j - \bar{u}_i)\Big| \\
&\quad = \big|\zeta_i^+\zeta_i^-\zeta_j^+\zeta_j^-(u_j - u_i) - \bar{\zeta}_i^+\bar{\zeta}_i^-\bar{\zeta}_j^+\bar{\zeta}_j^-(\bar{u}_j - \bar{u}_i)\big| \\
&\quad \leqslant \Big(|\zeta_i^+ - \bar{\zeta}_i^+|\zeta_i^-\zeta_j^+\zeta_j^- + \bar{\zeta}_i^+|\zeta_i^- - \bar{\zeta}_i^-|\zeta_j^+\zeta_j^- + \bar{\zeta}_i^+\bar{\zeta}_i^-|\zeta_j^+ - \bar{\zeta}_j^+|\zeta_j^- \\
&\qquad\qquad + \bar{\zeta}_i^+\bar{\zeta}_i^-\bar{\zeta}_j^+|\zeta_j^- - \bar{\zeta}_j^-|\Big)|u_j - u_i| + \bar{\zeta}_i^+\bar{\zeta}_i^-\bar{\zeta}_j^+\bar{\zeta}_j^-\big|(u_j - u_i) - (\bar{u}_j - \bar{u}_i)\big| \\
&\quad \leqslant 2^3\big(|\zeta_i^+ - \bar{\zeta}_i^+| + |\zeta_i^- - \bar{\zeta}_i^-| + |\zeta_j^+ - \bar{\zeta}_j^+| + |\zeta_j^- - \bar{\zeta}_j^-|\big)|u_j - u_i| + 2^4\big|(u_j - u_i) - (\bar{u}_j - \bar{u}_i)\big|,
\end{aligned}$$

where quantities equipped with a bar are evaluated at $\bar{u}$. Combining these two estimates, it is easy to verify that the Lipschitz constant of the AFC residual $\mathscr{R}$ using $p = 0$ and $q = 1$ is bounded above by a value that does not depend on δ_i and γ_i under the condition $\gamma_i \geqslant \delta_i'$. On the hand, the parameter γ_i should be chosen as small as possible for an accurate approximation $\beta_{i,\varepsilon} \approx \beta_i$. For that reason, we fix the optional parameter γ_i in the regularization procedure by setting $\gamma_i = \delta_i'$.

Clearly, the AFC scheme (6) using the regularized correction factors $\beta_{i,\varepsilon}$ is not linearity preserving any longer because the results of Section 3.2 do not hold for $\varepsilon > 0$. However, the values of q and δ_i can be determined to guarantee that $\beta_{i,\varepsilon} = 1$ is still satisfied at least for all (globally) linear functions u_h with $\|\nabla u_h\|_2 \geqslant \eta$ for some $\eta > 0$. For this purpose, let us assume that the condition

$$\delta_i' \leqslant \frac{2\big(\sqrt{q}-1\big)}{1+2w_i\sqrt{q}}\rho_i\eta, \qquad w_i \coloneqq \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}, \quad \rho_i \coloneqq \min_{\mathbf{x}\in\partial\varpi_i^{\text{conv}}} \|\mathbf{x}-\mathbf{x}_i\|_2 \tag{24}$$

is valid. Then we have

$$(1+2w_i\sqrt{q})\delta_i' \leqslant 2(\sqrt{q}-1)\rho_i\eta \leqslant 2(\sqrt{q}-1)s_i^+ \quad \Longrightarrow \quad 2s_i^+ + \delta_i' \leqslant 2\sqrt{q}(s_i^+ - \delta_i' w_i)$$

and, by virtue of the fact that $w_{ik} \geqslant 1$ for all $k \in \mathcal{N}_i \setminus \{i\}$,

$$s_i^+ = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i-u_k|_+ \geqslant \sum_{k\in\mathcal{N}_i\setminus\{i\}} |u_i-u_k|_+ \geqslant \min_{\mathbf{x}\in\partial\varpi_i^{\text{conv}}} \|\mathbf{x}-\mathbf{x}_i\|_2\|\nabla u_h\|_2 \geqslant \rho_i\eta.$$

Thus, the validity of condition (24) implies that all correction factors are equal to one by

$$\begin{aligned}
(s_i^{+,\delta_i} + s_i^{-,\delta_i} + \gamma_i)^2 &\leqslant (2s_i^+ + \delta_i')^2 \\
&\leqslant 4q(s_i^+ - \delta_i' w_i)^2 \\
&= 4q\Big(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}\big(|u_i-u_k|_+ - \delta_i'\big)\Big)\Big(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}\big(|u_k-u_i|_+ - \delta_i'\big)\Big) \\
&\leqslant 4q\Big(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i-u_k|_{+,\delta_i}\Big)\Big(\sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_k-u_i|_{+,\delta_i}\Big) \leqslant 4q s_i^{+,\delta_i} s_i^{-,\delta_i}
\end{aligned}$$

because the inequality $s_i^{\pm,\delta_i} \leqslant s_i^{\pm}$ holds due to (21) and $s_i^+ = s_i^-$ is satisfied for linear functions by definition of the weights $(w_{ik})_{k\in\mathcal{N}_i\setminus\{i\}}$.

According to (24), the quantity q seems to have only a minor influence on the accuracy of the solution while the high order target scheme is recovered for more and more linear functions if the parameter δ_i' is decreased. However, the right hand side of inequality (24) is proportional to the local mesh size ρ_i and the value of δ_i' should be adapted accordingly to guarantee that $\beta_{i,\varepsilon} = 1$ for linear functions u_h with $\|\nabla u_h\|_2 \geqslant \eta$ no matter how fine the mesh is chosen. Therefore, we set

$$\gamma_i = \delta_i' = \varepsilon h_i^r, \quad h_i \coloneqq \max_{k\in\mathcal{N}_i\setminus\{i\}} \|\mathbf{x}_i-\mathbf{x}_k\|_2$$

for $r \in \mathbb{N}$. Actually, the numerical examples presented below illustrate that $r \geqslant 2$ seems to be necessary to achieve the optimal rate of convergence.

According to the above analysis, the free parameters of the regularized AFC scheme are given by $q \geqslant 1$, $p \in \mathbb{N}_0$, $\tilde{p} \in \mathbb{N}$, $r \in \mathbb{N}$, and $\varepsilon \geqslant 0$. These values can be adapted (i) to improve the accuracy of the corresponding FE solution, (ii) to increase the regularity of the AFC residual, and (iii) to minimize the numerical effort for solving the nonlinear system of equations. Unfortunately, the optimal balance between accuracy and numerical effort is not well defined and difficult to achieve in practice. However, the choice of $p = \tilde{p} = 2$ suffices to guarantee that Newton's method converges quadratically for $\varepsilon > 0$ because then the AFC residual is twice

continuously differentiable. On the other hand, the parameters q and ε can still be adapted to improve the accuracy of the scheme. Furthermore, we set $r = 2$ for an optimal convergence behavior in terms of accuracy while still keeping γ_i and δ_i as large as possible for a smooth AFC residual. The final regularized nodal correction factors are given by

$$\beta_{i,\varepsilon} \coloneqq 1 - \max\left(0, 1 - q\frac{4s_i^{+,\delta_i}s_i^{-,\delta_i}}{(s_i^{+,\delta_i} + s_i^{-,\delta_i} + \delta_i')^2}\right)^{p+1},$$

$$s_i^{+,\delta_i} = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_i - u_k|_{+,\delta_i}, \quad s_i^{-,\delta_i} = \sum_{k\in\mathcal{N}_i\setminus\{i\}} w_{ik}|u_k - u_i|_{+,\delta_i},$$

$$\delta_i' = \varepsilon h_i^r, \quad \delta_i = (p-1)^{1-p}(p\varepsilon h_i^r)^p, \quad h_i \coloneqq \max_{k\in\mathcal{N}_i\setminus\{i\}} \|\mathbf{x}_i - \mathbf{x}_k\|_2, \quad r = 2, \quad p = \tilde{p} = 2$$

and depend on the free parameters $q \geqslant 1$ and $\varepsilon > 0$. In the numerical examples presented below, all other parameters are defined as above if not mentioned otherwise.

4 Jacobian of AFC residual

After regularization of the correction factors as above, the residual of the resulting AFC scheme becomes twice continuously differentiable and calls for the use of more sophisticated solvers than simple fixed point iterations. In particular, Newton's method can be applied and converges quadratically for sufficiently good initial guesses. However, the exact Jacobian of the residual must be efficiently computable to benefit from this theoretical advantage. Unfortunately, the local domain of dependence is extended by definition of the (nodal) correction factors and the Jacobian becomes more dense than the high order system matrix. Therefore, custom-made techniques for calculating the Jacobian are required to reduce the total effort of the numerical solution procedure. In [Loh19a, Section 5.1], the author showed that the Jacobian admits a decomposition into a sum and a product of matrices having the same sparsity pattern as the system matrix, at least when the correction factors provide a special nodal representation. Based on this algebraic manipulation, the Jacobian can be explicitly computed using a matrix-matrix product or its application can be implicitly performed by means of a few matrix-vector multiplications.

In what follows, we briefly generalize the decomposition of the Jacobian to AFC residuals using some edge-based correction factors and present corresponding results for element-based AFC approaches.

4.1 Edge-based correction factors

In contrast to the assumptions considered in [Loh19a, Section 5.1], we no longer require that the correction factors α_{ij} be the product of two nodal correction factors. Instead, let us assume that $\alpha_{ij} = \alpha_{ji}$ is twice continuously differentiable and only depends on some auxiliary quantities β_{ij} and β_{ji} with

$$\frac{\partial\beta_{ij}}{\partial u_k} = \begin{cases} r_{ij}q_{ik} + s_{ij,i} & : k = i, \\ r_{ij}q_{ik} + s_{ij,j} & : k = j, \\ r_{ij}q_{ik} & : k \in \mathcal{N}_i \setminus \{i,j\}, \\ 0 & : k \notin \mathcal{N}_i \end{cases} \quad \forall k \in \{1,\dots,N\},$$

where, for the sake of simplicity, all other entries of $\mathcal{R} = (r_{ij})_{i,j=1}^N$, $\mathcal{Q} = (q_{ij})_{i,j=1}^N$, and $\mathcal{S}^{(k)} = (s_{ij,k})_{i,j=1}^N$, $k \in \{1,\dots,N\}$, are set to zero. Then the Jacobian $\mathcal{J} \in \mathbb{R}^{N\times N}$ of the residual $\mathscr{R}$ defined by (6) satisfies

$$
\begin{aligned}
&(\mathcal{J}v)_i - \sum_{j\in\mathcal{N}_i} a_{ij}v_j + \sum_{j\in\mathcal{N}_i\setminus\{i\}} d_{ij}(1-\alpha_{ij})(v_j - v_i)\\
&\quad= \sum_k \sum_{j\in\mathcal{N}_i\setminus\{i\}} \big((\partial_{\beta_{ij}}\alpha_{ij})(v_k\partial_{u_k}\beta_{ij}) + (\partial_{\beta_{ji}}\alpha_{ij})(v_k\partial_{u_k}\beta_{ji})\big) f_{ij}\\
&\quad= \sum_{j\in\mathcal{N}_i\setminus\{i\}} \Big((\partial_{\beta_{ij}}\alpha_{ij})\big(\sum_k v_k\partial_{u_k}\beta_{ij}\big) + (\partial_{\beta_{ji}}\alpha_{ij})\big(\sum_k v_k\partial_{u_k}\beta_{ji}\big)\Big) f_{ij}\\
&\quad= \sum_{j\in\mathcal{N}_i\setminus\{i\}} (\partial_{\beta_{ij}}\alpha_{ij})\big(v_i s_{ij,i} + v_j s_{ij,j} + \sum_k v_k r_{ij} q_{ik}\big) f_{ij}\\
&\qquad+ \sum_{j\in\mathcal{N}_i\setminus\{i\}} (\partial_{\beta_{ji}}\alpha_{ij})\big(v_j s_{ji,j} + v_i s_{ji,i} + \sum_k v_k r_{ji} q_{jk}\big) f_{ij}\\
&\quad= \Big(\sum_{j\in\mathcal{N}_i\setminus\{i\}} (\partial_{\beta_{ij}}\alpha_{ij}) r_{ij} f_{ij}\Big)\big(\sum_k v_k q_{ik}\big) + \sum_{j\in\mathcal{N}_i\setminus\{i\}} (\partial_{\beta_{ji}}\alpha_{ij}) r_{ji} f_{ij}\big(\sum_k v_k q_{jk}\big)\\
&\qquad+ \Big(\sum_{j\in\mathcal{N}_i\setminus\{i\}} \big((\partial_{\beta_{ij}}\alpha_{ij}) s_{ij,i} + (\partial_{\beta_{ji}}\alpha_{ij}) s_{ji,i}\big) f_{ij}\Big) v_i\\
&\qquad+ \sum_{j\in\mathcal{N}_i\setminus\{i\}} \big((\partial_{\beta_{ij}}\alpha_{ij}) s_{ij,j} + (\partial_{\beta_{ji}}\alpha_{ij}) s_{ji,j}\big) f_{ij} v_j \qquad \forall i \in \{1,\dots,N\} \quad \forall v \in \mathbb{R}^N
\end{aligned}
$$

and, hence, can be written as

$$\mathcal{J} = \mathcal{A} - \hat{\mathcal{D}} + \mathcal{T} + \mathcal{P}\mathcal{Q}, \tag{25}$$

where the entries of $\hat{\mathcal{D}} = (\hat{d}_{ij})_{i,j=1}^N$, $\mathcal{T} = (t_{ij})_{i,j=1}^N$, and $\mathcal{P} = (p_{ij})_{i,j=1}^N$ are given by

$$
\begin{aligned}
\hat{d}_{ij} &:= \begin{cases} (1-\alpha_{ij}) d_{ij} & : i \neq j, \\ -\sum_{k\in\mathcal{N}_i\setminus\{i\}} \hat{d}_{ik} & : i = j \end{cases} && \forall i,j \in \{1,\dots,N\},\\
t_{ij} &:= \begin{cases} \big((\partial_{\beta_{ij}}\alpha_{ij}) s_{ij,j} + (\partial_{\beta_{ji}}\alpha_{ij}) s_{ji,j}\big) f_{ij} & : i \neq j, \\ -\sum_{k\in\mathcal{N}_j\setminus\{j\}} t_{kj} & : i = j \end{cases} && \forall i,j \in \{1,\dots,N\},\\
p_{ij} &:= \begin{cases} (\partial_{\beta_{ji}}\alpha_{ij}) r_{ji} f_{ij} & : i \neq j, \\ -\sum_{k\in\mathcal{N}_j\setminus\{j\}} p_{kj} & : i = j \end{cases} && \forall i,j \in \{1,\dots,N\}
\end{aligned}
$$

due to the fact that $f_{ij} = -f_{ji}$ and $\alpha_{ij} = \alpha_{ji}$. By definition of the artificial diffusion operator $\mathcal{D}$, all auxiliary matrices have the same sparsity pattern as the system matrix $\mathcal{A}$ and, hence, can be efficiently generated. If the auxiliary correction factors β_{ij} are nodal quantities, i.e., $\beta_{ij} = \beta_i$, and only depend on $(u_k - u_i)_{k\in\mathcal{N}_i}$, we can even set $r_{ij} = 1$ as well as $s_{ij,i} = s_{ij,j} = 0$ for all $j \in \mathcal{N}_i \setminus \{i\}$. Then $\mathcal{T} = 0$ and the row sums of $\mathcal{Q}$ vanish because

$$q_{ii} = r_{ij} q_{ii} + s_{ij,i} = \frac{\partial\beta_{ij}}{\partial u_i} = -\sum_{k\in\mathcal{N}_i\setminus\{i\}} \frac{\partial\beta_{ij}}{\partial u_k} = -s_{ij,j} - r_{ij}\sum_{k\in\mathcal{N}_i\setminus\{i\}} q_{ik} = -\sum_{k\in\mathcal{N}_i\setminus\{i\}} q_{ik}.$$

For example, the regularized correction factors defined in Section 3 fit into this category and formula (25) can be employed to determine the exact Jacobian of the nonlinear AFC residual in an simple and efficient manner.

4.2 Element-based approach

In Section 1, the AFC methodology was introduced by incorporating nonlinear (diffusive) fluxes into the high order Galerkin discretization to compute highly accurate and bound-preserving finite element solutions. As considered, e.g., in [Loh+17; Kuz+17], similar techniques can also be used to stabilize the high order target method in an element-based manner. For this purpose, some element contributions $f^e \in \mathbb{R}^N$, $e \in \{1, \dots, E\}$, are scaled by means of solution-dependent correction factors $\alpha^e : \mathbb{R}^N \to [0, 1]$ and result in diffusive terms like (cf. (7))

$$\hat{f}_i = \sum_{e \in \mathcal{E}_i} \big(1 - \alpha^e(u)\big) f_i^e \quad \forall i \in \{1, \dots, N\}. \tag{26}$$

In what follows, we extend the idea of decomposing the Jacobian to the element-based AFC approach for the sake of completeness. To this end, we again assume that the correction factor α^e is twice continuously differentiable and satisfies

$$\alpha^e = \alpha^e\big((\beta_i)_{i \in \mathcal{N}^e}\big) \quad \forall e \in \{1, \dots, E\}$$

for some nodal correction factors β_i with

$$\frac{\partial \beta_i}{\partial u_k} = \begin{cases} q_{ik} & : k \in \mathcal{N}_i, \\ 0 & : k \notin \mathcal{N}_i \end{cases} \quad \forall k \in \{1, \dots, N\}.$$

As above, we set $q_{ik} = 0$ if $k \notin \mathcal{N}_i$. Then the Jacobian $\mathcal{J} \in \mathbb{R}^{N \times N}$ of the residual $\mathscr{R}$ satisfies

$$\begin{aligned}
&(\mathcal{J}v)_i - \sum_{j \in \mathcal{N}_i} a_{ij} v_j + \sum_k \sum_{e \in \mathcal{E}_i} (1 - \alpha^e) v_k (\partial_{u_k} f_i^e) \\
&\quad = \sum_k \sum_{e \in \mathcal{E}_i} f_i^e (v_k \partial_{u_k} \alpha^e) = \sum_k \sum_{e \in \mathcal{E}_i} f_i^e \sum_{l \in \mathcal{N}^e} (\partial_{\beta_l} \alpha^e)(v_k \partial_{u_k} \beta_l) \\
&\quad = \sum_{e \in \mathcal{E}_i} f_i^e \sum_{l \in \mathcal{N}^e} (\partial_{\beta_l} \alpha^e) \sum_{k \in \mathcal{N}_l} (v_k q_{lk}) = \sum_l \Big(\sum_{e \in \mathcal{E}_i} f_i^e (\partial_{\beta_l} \alpha^e) \Big) \Big(\sum_{k \in \mathcal{N}_l} (v_k q_{lk}) \Big) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall i \in \{1, \dots, N\} \quad \forall v \in \mathbb{R}^N
\end{aligned}$$

and can be written as

$$\mathcal{J} = \mathcal{A} - \hat{\mathcal{D}} + \mathcal{P}\mathcal{Q},$$

where the entries of the auxiliary matrices $\hat{\mathcal{D}} = (\hat{d}_{ij})_{i,j=1}^N$ and $\mathcal{P} = (p_{ij})_{i,j=1}^N$ are given by

$$\begin{aligned}
\hat{d}_{ij} &\coloneqq \sum_{e \in \mathcal{E}_i} (1 - \alpha^e)(\partial_{u_j} f_i^e) \quad \forall i, j \in \{1, \dots, N\}, \\
p_{ij} &\coloneqq \sum_{e \in \mathcal{E}_i} f_i^e (\partial_{\beta_j} \alpha^e) \qquad\quad \forall i, j \in \{1, \dots, N\}.
\end{aligned}$$

5 Time dependent problem

After focusing on the AFC methodology for the steady hyperbolic model problem (1), we now present the main idea of extending the monolithic limiting technique to its transient counterpart.

In this case, many beneficial properties like the preservation of discrete maximum principles should be maintained, at least under some CFL-like time step restriction. Furthermore, we design the involved correction factors so that the right hand side of the semi-discrete problem is Lipschitz continuous guaranteeing unique steady state solutions and well-posed problems when using implicit time integrators.

In this section, the continuous problem under consideration is given by the transient advection-reaction equation in non-conservative form

$$\frac{\partial u}{\partial t} + \mathbf{v} \cdot \operatorname{grad}(u) + cu = s \qquad \text{in } \Omega \times (0,T), \tag{27a}$$

$$u = u_{\text{in}} \qquad \text{on } \Gamma_{\text{in}} \coloneqq \big\{(\mathbf{s},t) \in \partial\Omega \times (0,T) \,\big|\, \mathbf{n}(\mathbf{s}) \cdot \mathbf{v}(\mathbf{s},t) < 0\big\}, \tag{27b}$$

$$u(\cdot,0) = u_0 \qquad \text{in } \Omega, \tag{27c}$$

where $T > 0$ is the final time and $u_0 : \Omega \to \mathbb{R}$ denotes the initial data for the solution $u : \Omega \times [0,T] \to \mathbb{R}$. Furthermore, the functions $\mathbf{v}$, c, s, and u_{in} are time dependent counterparts of the quantities introduced in Section 2. Then the degrees of freedom of the Galerkin approximation $u_h : \Omega \times [0,T] \to \mathbb{R}$ solve the semi-discrete problem

$$\sum_{j \in \mathcal{N}_i} m_{ij} \frac{\mathrm{d}u_j}{\mathrm{d}t} + \sum_{j \in \mathcal{N}_i} a_{ij} u_j = g_i \quad \forall i \in \{1, \dots, N\},$$

where the consistent mass matrix $\mathcal{M} = (m_{ij})_{i,j=1}^N$ is defined by

$$m_{ij} = \int_\Omega \varphi_i \varphi_j \quad \forall i,j \in \{1, \dots, N\}.$$

As in the stationary case, the high order finite element solution is possibly polluted by spurious oscillations and may violate some discrete maximum principles. Therefore, we consider the bound-preserving AFC scheme given by [Kuz12]

$$m_i \frac{\mathrm{d}u_i}{\mathrm{d}t} + \big(\mathscr{F}(u,t)\big)_i = 0 \quad \forall i \in \{1, \dots, N\}, \tag{28a}$$

$$\big(\mathscr{F}(u,t)\big)_i \coloneqq \sum_{j \in \mathcal{N}_i \setminus \{i\}} \dot{\alpha}_{ij} m_{ij} (\dot{u}_j - \dot{u}_i) + \sum_{j \in \mathcal{N}_i} a_{ij} u_j - \sum_{j \in \mathcal{N}_i \setminus \{i\}} (1 - \alpha_{ij}) d_{ij} (u_j - u_i) - g_i, \tag{28b}$$

where the entries of the lumped mass matrix $\mathcal{M}_{\text{L}} = (m_i \delta_{ij})_{i,j=1}^N$ are given by $m_i = \sum_{j \in \mathcal{N}_i} m_{ij}$ and $\dot{u}_i$ approximates $\mathrm{d}_t u_i$ in terms of the dissipation parameter $\omega \in [0,1]$

$$\dot{u}_i = m_i^{-1} \Big(g_i - \sum_{j \in \mathcal{N}_i} a_{ij} u_j + \sum_{j \in \mathcal{N}_i \setminus \{i\}} (1 - \omega \alpha_{ij}) d_{ij} (u_j - u_i) \Big) \quad \forall i \in \{1, \dots, N\}.$$

For the sake of simplicity, the correction factors α_{ij} and $\dot{\alpha}_{ij}$ are given as above by the product of regularized nodal quantities, i.e., $\alpha_{ij} = \beta_{i,\varepsilon} \beta_{j,\varepsilon}$ and $\dot{\alpha}_{ij} = \dot{\beta}_{i,\varepsilon} \dot{\beta}_{j,\varepsilon}$, and should guarantee the Lipschitz continuity of $\mathscr{F} : \mathbb{R}^N \to \mathbb{R}^N$. For that reason, the nodal correction factors $\beta_{i,\varepsilon}$ are

defined as above while $\dot{\beta}_{i,\varepsilon}$ is similarly introduced by

$$\dot{\beta}_{i,\varepsilon} \coloneqq \dot{\beta}^{+}_{i,\varepsilon}\dot{\beta}^{-}_{i,\varepsilon}, \quad \dot{\beta}^{\pm}_{i,\varepsilon} \coloneqq 1 - \max\left(0, 1 - \sqrt{\dot{q}}\frac{2r_i^{\pm,\delta_i}}{t_i^{+,\delta_i} + t_i^{-,\delta_i} + \gamma_i}\right)^p,$$

$$r_i^{+,\delta_i} \coloneqq \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}\left|\frac{d_{ik}}{m_{ik}}(u_k - u_i)\right|_{+,\delta_i}, \quad r_i^{-,\delta_i} \coloneqq \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}\left|\frac{d_{ik}}{m_{ik}}(u_i - u_k)\right|_{+,\delta_i},$$

$$t_i^{+,\delta_i} \coloneqq \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|\dot{u}_k - \dot{u}_i|_{+,\delta_i}, \quad t_i^{-,\delta_i} \coloneqq \sum_{k\in\mathcal{N}_i\backslash\{i\}} w_{ik}|\dot{u}_i - \dot{u}_k|_{+,\delta_i},$$

where the terms of the sum in the numerator $r_i^{\pm,\delta_i}$ are scaled by $\frac{d_{ik}}{m_{ik}}$ to obtain a fraction without units. Note that $\dot{\beta}_{i,\varepsilon}$ is given as the product of two auxiliary quantities $\dot{\beta}^{+}_{i,\varepsilon}$ and $\dot{\beta}^{-}_{i,\varepsilon}$ which vanish whenever u_i is a local maximum or minimum, respectively. This slight modification is required because the denominator of the correction factors $\dot{\beta}^{\pm}_{i,\varepsilon}$ is evaluated by $\dot{u}$ instead of u and, hence, $r_i^{\pm,\delta_i}(t_i^{+,\delta_i} + t_i^{-,\delta_i} + \gamma_i)^{-1}$ is not bounded above by some $C \neq C(u, \dot{u})$. By virtue of this adjustment, the Lipschitz continuity of $\mathscr{F}$ can be proved as illustrated in Section 3.3.

Adapting the results of Section 4.1, the exact Jacobian $\mathcal{J}$ of $\mathscr{F}$ can be written as

$$\mathcal{J} = (\mathcal{A} - \hat{\mathcal{D}} + \mathcal{P}\mathcal{Q}) + \dot{\mathcal{P}}\dot{\mathcal{Q}}^{\mathrm{nu}} - (\hat{\mathcal{M}} + \dot{\mathcal{P}}\dot{\mathcal{Q}}^{\mathrm{de}})\mathcal{M}_{\mathrm{L}}^{-1}(\mathcal{A} - (1-\omega)\mathcal{D} - \omega\hat{\mathcal{D}} + \omega\mathcal{P}\mathcal{Q}), \tag{29}$$

where the rows of $\dot{\mathcal{Q}}^{\mathrm{nu}} = (\dot{q}_{ij}^{\mathrm{nu}})_{i,j=1}^{N}$ and $\dot{\mathcal{Q}}^{\mathrm{de}} = (\dot{q}_{ij}^{\mathrm{de}})_{i,j=1}^{N}$ contain the derivatives of $\dot{\beta}_{i,\varepsilon}$ with respect to the components of u and $\dot{u}$, respectively, i.e.,

$$\dot{q}_{ij}^{\mathrm{nu}} \coloneqq \partial_{u_j}\dot{\beta}_{i,\varepsilon} \quad \dot{q}_{ij}^{\mathrm{de}} \coloneqq \partial_{\dot{u}_j}\dot{\beta}_{i,\varepsilon} \qquad \forall i,j \in \{1,\dots,N\}.$$

Furthermore, the auxiliary matrices $\hat{\mathcal{M}} = (\hat{m}_{ij})_{i,j=1}^{N}$ and $\dot{\mathcal{P}} = (\dot{p}_{ij})_{i,j=1}^{N}$ are defined by

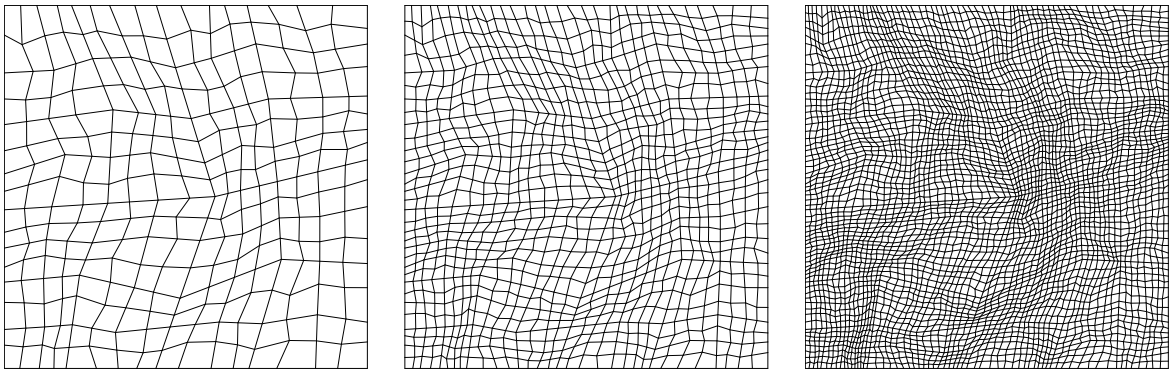
$$\hat{m}_{ij} \coloneqq \begin{cases} \dot{\alpha}_{ij}m_{ij} & : i \neq j, \\ -\sum_{k\in\mathcal{N}_i\backslash\{i\}} \hat{m}_{ik} & : i = j \end{cases} \quad \forall i,j \in \{1,\dots,N\},$$

$$\dot{p}_{ij} \coloneqq \begin{cases} \dot{\beta}_{i,\varepsilon}m_{ij}(\dot{u}_j - \dot{u}_i) & : i \neq j, \\ -\sum_{i\in\mathcal{N}_j\backslash\{j\}} \dot{p}_{ij} & : i = j \end{cases} \quad \forall i,j \in \{1,\dots,N\}$$

while $\hat{\mathcal{D}}$, $\mathcal{P}$, and $\mathcal{Q}$ are determined as above. Representation (29) involves multiplication of up to four matrices having the same sparsity pattern as the system matrix and solution of corresponding linear systems is expensive. Therefore, it might not be worthwhile to employ Newton's method in the context of implicit time integrators and simple fixed-point iterations can be more efficient due to reasonably good initial guesses. However, detailed numerical studies are required to evaluate different solution strategies and are beyond the scope of this work.

6 Numerical examples

In this section, we apply the AFC methodology using the proposed nodal correction factors to some steady and transient two dimensional test problems. The domain of all considered benchmarks is given by the unit square, i.e., $\Omega = (0,1)^2$, and is decomposed into a structured



(a) Level 4 with $(2^4)^2 = 16^2$ cells. (b) Level 5 with $(2^5)^2 = 32^2$ cells. (c) Level 6 with $(2^6)^2 = 64^2$ cells.

Figure 2: Illustration of distorted triangulations which are given by transformed Shestakov meshes [She+90] using $a = 0.35$.

mesh containing 2^l quadrilateral elements in each coordinate direction on level $l \in \mathbb{N}$. The nodes are either uniformly distributed or provide a disturbed mesh as constructed in [She+90] (cf. Fig. 2).

In the stationary case, all nonlinear systems under consideration are solved using a damped Newton's method where the solution update is relaxed in terms of an Armijo condition [Arm66]. More precisely, the employed algorithm reads:

0. Set $u^{(0)} = (\mathcal{A} - \mathcal{D})^{-1} g$ and $m = 0$.
1. If $\left\| \mathcal{M}_{\mathrm{L}}^{-1} \mathscr{R}(u^{(m)}) \right\|_{\tilde{2}} \leqslant 10^{-8}$, the approximate solution is given by $u^{(m)}$.
2. Compute $\Delta u^{(m+1)} = -\mathcal{J}(u^{(m)})^{-1} \mathscr{R}(u^{(m)})$.
3. Determine $s \in \mathbb{N}_0$ as the smallest integer for which
$$\left\| \mathcal{M}_{\mathrm{L}}^{-1} \mathscr{R}(u^{(m)} + 2^{-s} \Delta u^{(m+1)}) \right\|_{\tilde{2}} \leqslant (1 + 10^{-2} 2^{-s}) \left\| \mathcal{M}_{\mathrm{L}}^{-1} \mathscr{R}(u^{(m)}) \right\|_{\tilde{2}}. \tag{30}$$
4. Set $u^{(m+1)} = u^{(m)} + 2^{-s} \Delta u^{(m+1)}$ as well as $m = m + 1$ and go to step 1.

In this solution procedure, the weighted ℓ_2-norm $\| \cdot \|_{\tilde{2}} : \mathbb{R}^N \to [0, \infty)$ is given by

$$\|v\|_{\tilde{2}} \coloneqq \Big(\sum_{i=1}^{N} m_i v_i^2 \Big)^{\frac{1}{2}} \quad \forall v \in \mathbb{R}^N$$

and approximates the L^2-norm of the finite element function v_h associated with the degrees of freedom $v \in \mathbb{R}^N$. Although the AFC residual is only differentiable for $\varepsilon > 0$, the same algorithm is also used if the regularization parameter vanishes. In this case, the derivative of $|\cdot|_+$ is set to

$$\frac{\partial |x|_+}{\partial x} \coloneqq \begin{cases} 0 & : x \leqslant 0, \\ 1 & : x > 0 \end{cases} \quad \forall x \in \mathbb{R}.$$

For some steady state results presented below, the Streamline upwind/Petrov-Galerkin (SUPG) method [BH82] is employed as the high order target of the AFC scheme. In these cases, we use the weak formulation

$$\int_\Omega \big(\varphi + \tau \mathbf{v}\cdot\operatorname{grad}(\varphi)\big)\mathbf{v}\cdot\operatorname{grad}(u) + \int_\Omega c\big(\varphi + \tau\mathbf{v}\cdot\operatorname{grad}(\varphi)\big)u + \int_{\Gamma_{\text{in}}} |\mathbf{v}\cdot\mathbf{n}|\varphi u$$
$$= \int_\Omega s\big(\varphi + \tau\mathbf{v}\cdot\operatorname{grad}(\varphi)\big) + \int_{\Gamma_{\text{in}}} |\mathbf{v}\cdot\mathbf{n}|\varphi u_{\text{in}} \quad \forall\varphi$$

to compute the entries of the system matrix $\mathcal{A}$ and right hand side g, where the stabilization parameter $\tau = \tau(\mathbf{x})$ is determined by (cf. [JK07])

$$\tau(\mathbf{x}) \coloneqq \frac{h_K(\mathbf{x})}{2\big\|\mathbf{v}(\mathbf{x})\big\|_2}, \qquad h_K(\mathbf{x}) = \max_{K^e \text{ s.t. } \mathbf{x}\in K^e} \operatorname{diam}(K^e)$$

and $\operatorname{diam}(K^e) \coloneqq \sup\big\{\|\mathbf{y}_1 - \mathbf{y}_2\| \;\big|\; \mathbf{y}_1, \mathbf{y}_2 \in K^e\big\}$. If not mentioned otherwise, the Galerkin discretization is considered as introduced in Section 2.

6.1 Smooth circular convection

In the first example, the influence of the parameters ε, r, and q on (i) the accuracy of the AFC scheme and (ii) the required number of damped Newton iterations for computing the corresponding solution is demonstrated. For this purpose, we focus on the circular convection benchmark [Hub07]

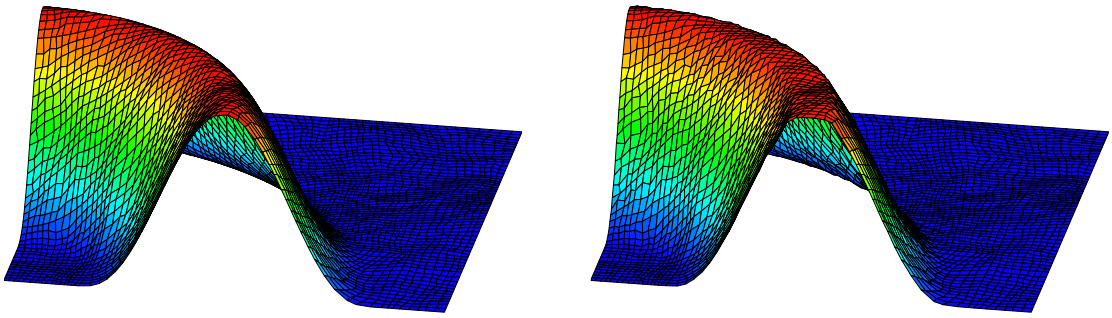
$$\begin{aligned} \mathbf{v}\cdot\operatorname{grad}(u) &= 0 && \text{in } \Omega, \\ u &= u_{\text{in}} && \text{on } \Gamma_{\text{in}} \coloneqq [0,1]\times\{0\} \cup \{1\}\times[0,1] \end{aligned}$$

using the velocity field $\mathbf{v}(\mathbf{x}) = (-x_2, x_1)^\top$ and choose the prescribed inflow boundary data u_{in} so that the exact solution $u \in C^2(\Omega;\mathbb{R})$ reads (cf. Fig. 3)

$$u(\mathbf{x}) = \max\big(0, 3\|\mathbf{x}\|_2 - \tfrac{1}{2}\big)^3 \max\big(0, \tfrac{5}{2} - 3\|\mathbf{x}\|_2\big)^3 \quad \forall \mathbf{x}\in\Omega. \tag{31}$$

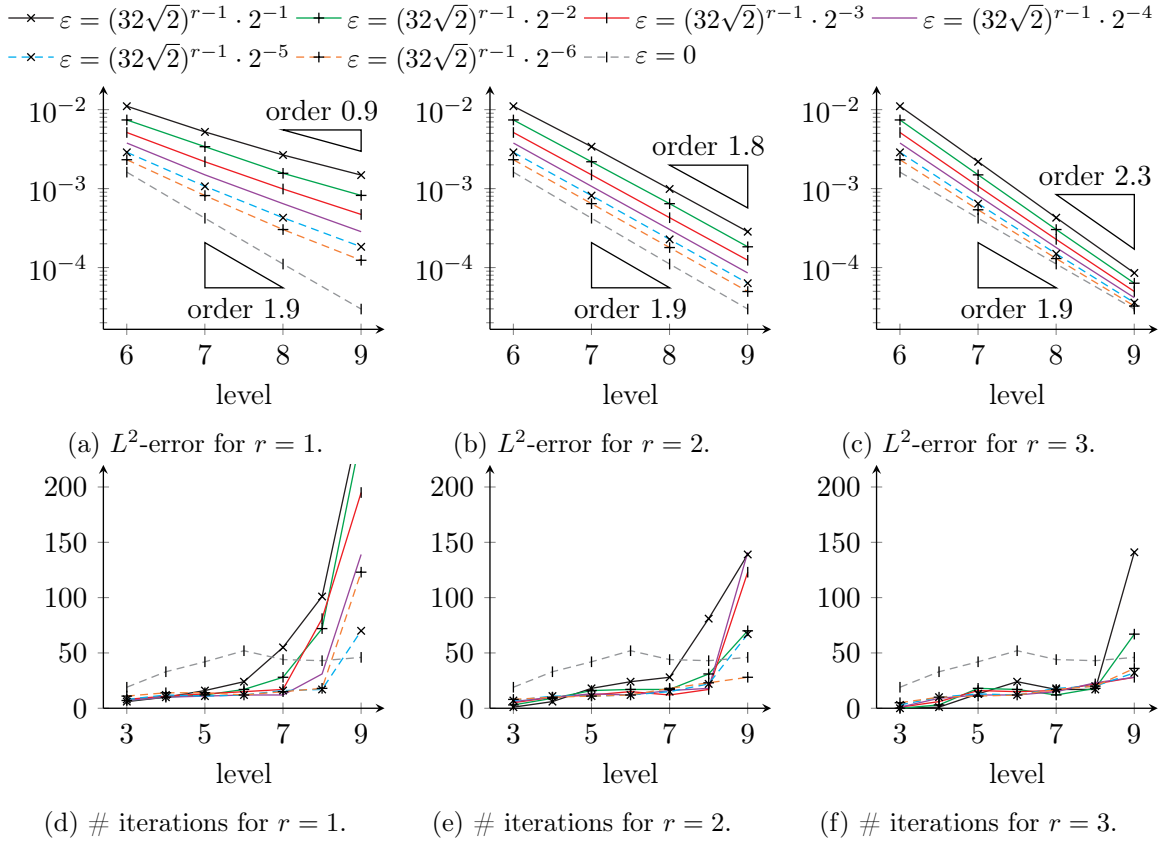
As expected, the AFC solution for $q = 2$ and $\varepsilon = 0$ converges on a uniform mesh with order 2 to the exact solution in L^2 because the employed scheme is linearity preserving (cf. Fig. 4). If the correction factors are regularized, this property is violated and the order of convergence is affected by the parameter r: While the AFC solution only converges with order 1 for $r = 1$, second order of convergence can be observed for $r = 2$ no matter how $\varepsilon > 0$ is chosen. If the exponent r is set to 3, the experimental order of convergence can 'locally' even be higher than 2. However, this effect vanishes on finer meshes because the AFC solution for $\varepsilon > 0$ cannot be (significantly) more accurate than the solution of the unregularized AFC scheme.

In this numerical study, the AFC solution for $q = 2$ and $\varepsilon = 0$ can be computed with less than 52 damped Newton iterations no matter how fine the mesh is. In contrast to this, the total number of nonlinear iterations required to solve the regularized AFC problem seems to increase with the mesh level l. Consequently, the use of regularized correction factors only pays off on coarse meshes for the considered values of ε and $r = 1, 2$. However, the total number of nonlinear iterations can also be halved on finer meshes if the parameter δ_i' is sufficiently small.



(a) Interpolation of exact solution. (b) AFC solution using $q = 2$, $\varepsilon = 0$.

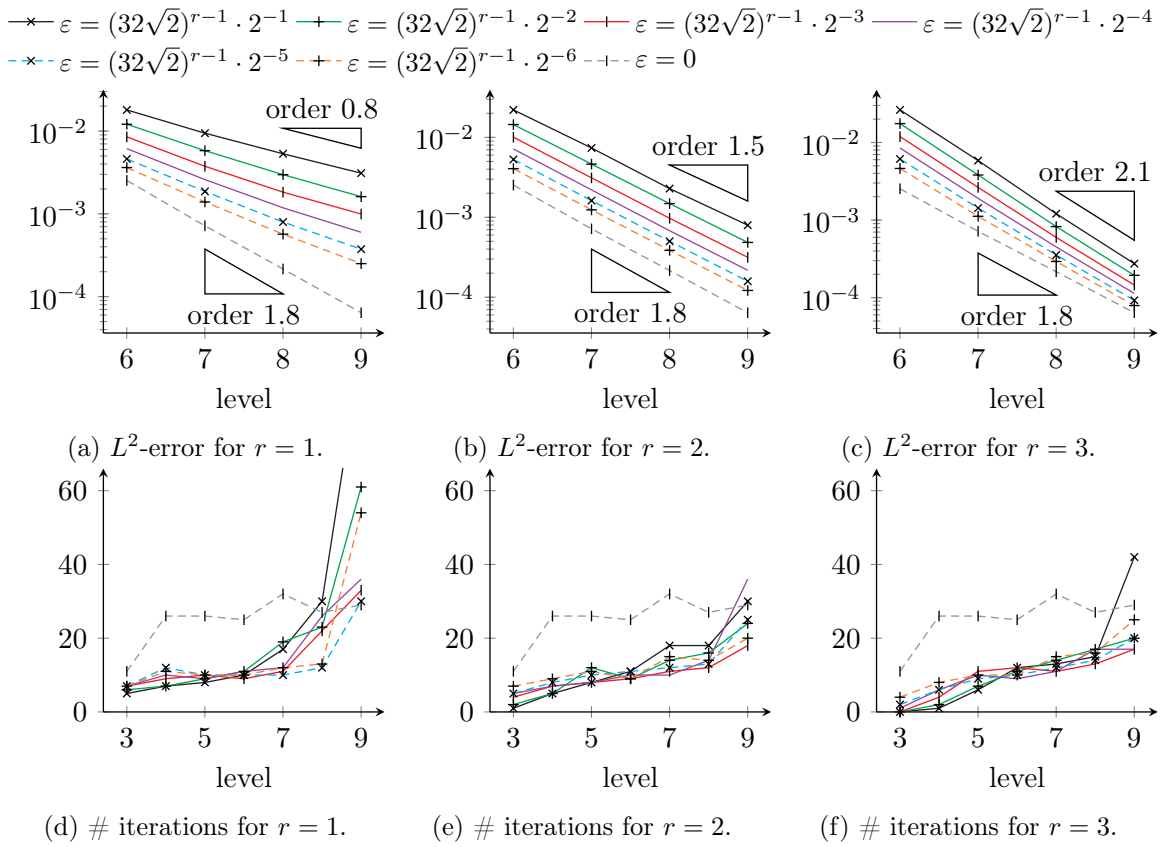
Figure 3: Smooth circular convection: Different discrete solutions on level 6 of distorted mesh.



(a) L^2-error for $r = 1$. (b) L^2-error for $r = 2$. (c) L^2-error for $r = 3$.

(d) # iterations for $r = 1$. (e) # iterations for $r = 2$. (f) # iterations for $r = 3$.

Figure 4: Smooth circular convection: Grid convergence study for L^2-error as well as number of iterations on uniform mesh using Galerkin method as high order target and $q = 2$.



(a) L^2-error for $r = 1$. (b) L^2-error for $r = 2$. (c) L^2-error for $r = 3$.

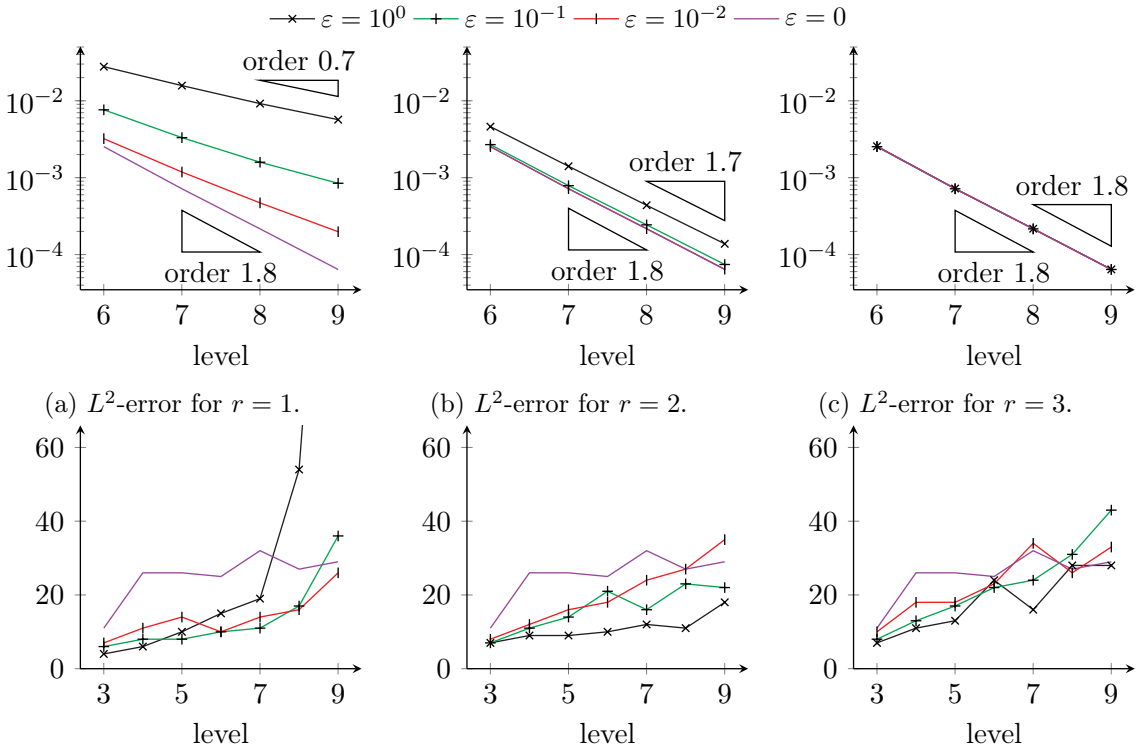
(d) # iterations for $r = 1$. (e) # iterations for $r = 2$. (f) # iterations for $r = 3$.

Figure 5: Smooth circular convection: Grid convergence study for L^2-error as well as number of iterations on distorted mesh using SUPG method as high order target and $q = 2$.

The results presented in Fig. 5 are computed analogously to the previous ones on distorted meshes by employing the SUPG method as the high order target. In this case, the optimal order of convergence of the unregularized AFC scheme is given by 1.8 and can be again approximately recovered by using the smoothed correction factors for $r \geqslant 2$. Furthermore, the required number of nonlinear iterations decreases if the AFC residual is regularized using a parameter δ_i' which is sufficiently small.

So far, the AFC solutions were computed using relatively large regularization parameters to clearly illustrate the influence of the exponent r. By using smaller values of ε, the accuracy of the solution obviously improves (cf. Fig. 6 for $r = 2, 3$) but at the same time the total number of nonlinear iterations increases. Therefore, the benefit of regularization seems to be less significant. Especially for $r = 3$, there is hardly any difference between the L^2-error of the solutions produced by the smoothed and unregularized AFC schemes while the required number of iterations also behaves similarly. As a consequence, the choice of $r = 2$ might be reasonable for further numerical studies and the parameter ε has to be selected appropriately to achieve the optimal ratio between accuracy and solvability.

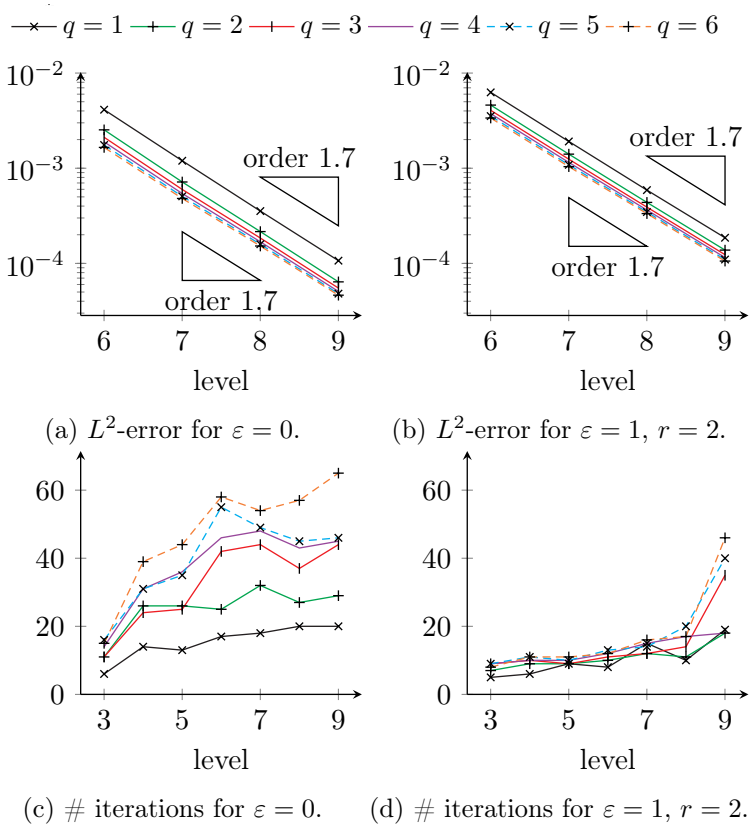
In Fig. 7, the influence of the parameter q is illustrated for the unregularized AFC scheme and the smoothed one using $\varepsilon = 1$. In both cases, the accuracy of the solution improves by



(a) L^2-error for $r = 1$. (b) L^2-error for $r = 2$. (c) L^2-error for $r = 3$.

(d) # iterations for $r = 1$. (e) # iterations for $r = 2$. (f) # iterations for $r = 3$.

Figure 6: Smooth circular convection: Grid convergence study for L^2-error as well as number of iterations on distorted mesh using SUPG method as high order target and $q = 2$.



(a) L^2-error for $\varepsilon = 0$. (b) L^2-error for $\varepsilon = 1$, $r = 2$.

(c) # iterations for $\varepsilon = 0$. (d) # iterations for $\varepsilon = 1$, $r = 2$.

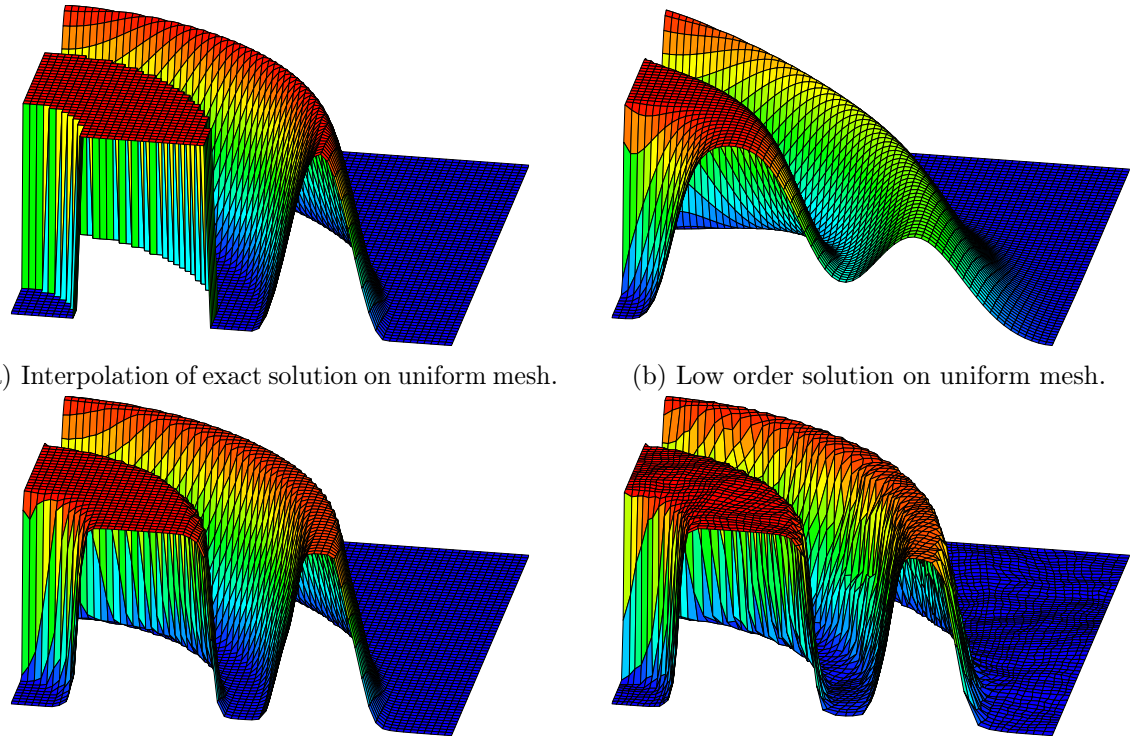
Figure 7: Smooth circular convection: Grid convergence study for L^2-error as well as total number of iterations on distorted mesh using SUPG method as high order target and different values of q.

increasing q and seems to converge to a mesh-dependent value. On the other hand, the total number of nonlinear iterations grows linearly for $\varepsilon = 0$ so that $q = 2$ might be the optimal choice between accuracy and numerical effort. If the AFC residual is regularized using $\varepsilon = 1$, nearly the same number of damped Newton steps are necessary for all values of q, at least for level $l \leqslant 8$. This behavior is caused by the fact that the large regularization parameter $\varepsilon = 1$ is mainly responsible for the stiffness of the nonlinear system and deteriorates the accuracy of the solution.

6.2 Discontinuous circular convection

The exact solution of the above test problem is twice continuously differentiable and, particularly, does not possess any discontinuity. To illustrate the full potential of the AFC scheme for steady hyperbolic problems, we now replace the exact solution (31) by the composition (cf. [Hub07])

$$u(\mathbf{x}) = \begin{cases} 1 & : 0.15 \leqslant \|\mathbf{x}\|_2 \leqslant 0.45, \\ \cos^2\left(10\pi \frac{\|\mathbf{x}\|_2 - 0.7}{3}\right) & : 0.55 \leqslant \|\mathbf{x}\|_2 \leqslant 0.85, \\ 0 & : \text{otherwise} \end{cases} \qquad \forall \mathbf{x} \in \Omega. \tag{32}$$



(a) Interpolation of exact solution on uniform mesh.

(b) Low order solution on uniform mesh.

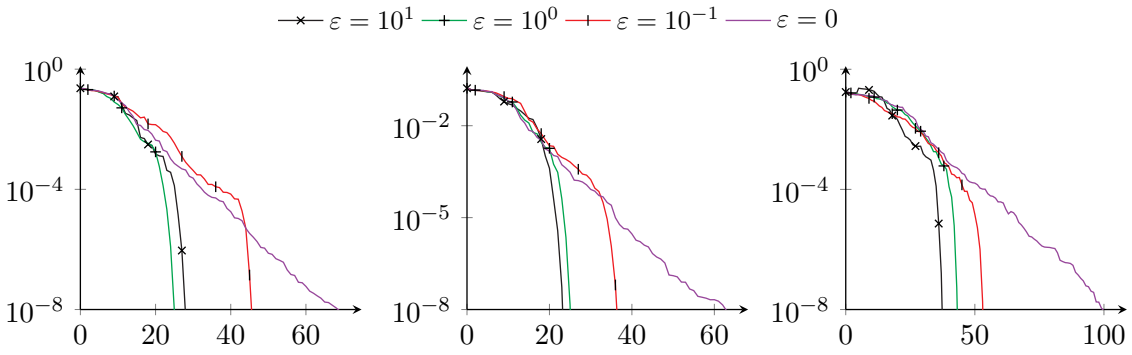
(c) AFC solution on uniform mesh.

(d) AFC solution on distorted mesh.

Figure 8: Discontinuous circular convection: Different discrete solutions on level 6 of uniform and distorted mesh. AFC solutions are computed using $q = 2$ and $\varepsilon = 0$.

In Fig. 8, the pointwise interpolation of u and different bound-preserving finite element approximations are presented. The low order solution solving (6) for $q = 0$ is very diffusive and the prescribed boundary data u_{in} is completely smeared out as the solution profile is convected away from the inflow boundary. If $q = 2$ and $\varepsilon = 0$ are used in the definition of the nodal correction factors, the accuracy improves while the validity of discrete maximum principles is still preserved. In particular, the discontinuities occurring in the exact solution are approximated by steep gradients without creating overshoots and undershoots. As the mesh is distorted, the AFC solution becomes slightly more diffusive. However, this drawback is due to the deteriorated quality of the mesh and might not be caused by a disadvantage of the AFC scheme.

To compute these bound-preserving finite element functions, more than 60 nonlinear iterations are required (cf. Fig. 9). However, AFC solutions with the same accuracy can also be computed if the correction factors are regularized using $\varepsilon = 10^{-1}$ (solutions not shown here). Actually, there is hardly any difference between the AFC solutions, but the number of nonlinear iterations is drastically reduced. The reason for this behavior is the fact that the smoothed AFC residual is twice continuously differentiable so that Newton's method converges quadratically after some initial iterates. The use of larger regularization parameters further reduces the numerical effort but results in slightly more diffusive solutions.



(a) Distorted mesh, $q = 2$. (b) Uniform mesh, $q = 2$. (c) Uniform mesh, $q = 4$.

Figure 9: Discontinuous circular convection: History of norm of residual during nonlinear iterations for different values of ε and q on level 7 of uniform and distorted mesh.

6.3 Convection of discontinuity

In the last steady state test case, we focus on the convection of a discontinuity using the constant velocity field $\mathbf{v} = (\frac{1}{2}, -\sin\frac{\pi}{3})^\top$ where the exact solution is given by [BB17, Section 9.1]

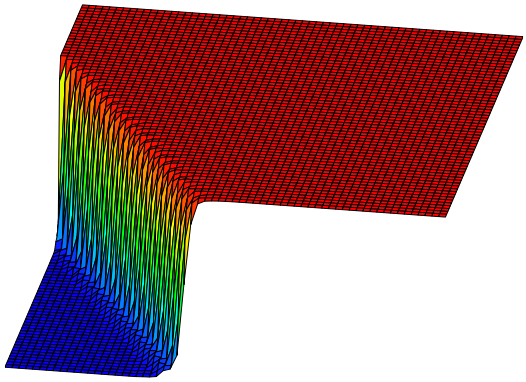
$$u(\mathbf{x}) = \begin{cases} 1 & : x_2 > 0.7 - 2x_1 \sin\frac{\pi}{3}, \\ 0 & : \text{otherwise.} \end{cases} \tag{33}$$

As observed in [Loh19a] for another definition of the correction factors, the AFC solution based on the Galerkin discretization is polluted by some terracing effect close to the discontinuity although discrete maximum principles are satisfied (cf. Fig. 10). The author assumed that this artifact might disappear by incorporating some stabilization term into the high order target scheme. For this purpose, we replace the Galerkin discretization by the SUPG method which corresponds to the addition of some high order diffusivity in streamline direction. By doing so, the solution becomes slightly more diffusive but the artificial terraces disappear no matter if the mesh is distorted or not.

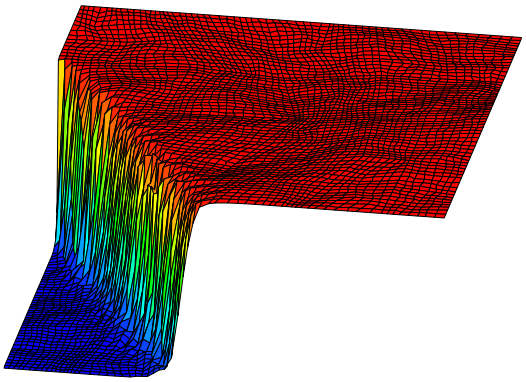
6.4 Solid body rotation

Finally, the AFC scheme of Section 5 is applied to the solid body rotation benchmark which goes back to [Zal79; LeV96]. Three solid bodies with radius $r = 0.15$ initially centered at $\mathbf{x}^{(1)} = (0.25, 0.5)^\top$, $\mathbf{x}^{(2)} = (0.5, 0.25)^\top$, and $\mathbf{x}^{(3)} = (0.5, 0.75)^\top$ are rotated once around the center of the domain $\Omega = (0,1)^2$ using the velocity field $\mathbf{v}(\mathbf{x}) = (\frac{1}{2} - x_2, x_1 - \frac{1}{2})^\top$ and the inflow boundary condition $u_{\text{in}} = 0$. At the final time $T = 2\pi$, the exact solution coincides with the initial profile u_0 which provides the decomposition (cf. Fig. 11)

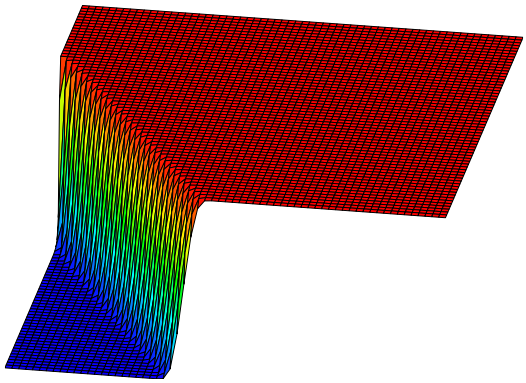
$$u_0(\mathbf{x}) \coloneqq \begin{cases} u^{(1)}\big((\mathbf{x} - \mathbf{x}^{(1)})r^{-1}\big) & : \|\mathbf{x} - \mathbf{x}^{(1)}\|_2 \leqslant r, \\ u^{(2)}\big((\mathbf{x} - \mathbf{x}^{(2)})r^{-1}\big) & : \|\mathbf{x} - \mathbf{x}^{(2)}\|_2 \leqslant r, \\ u^{(3)}\big((\mathbf{x} - \mathbf{x}^{(3)})r^{-1}\big) & : \|\mathbf{x} - \mathbf{x}^{(3)}\|_2 \leqslant r, \\ 0 & : \text{otherwise} \end{cases} \qquad \forall \mathbf{x} \in \Omega, \tag{34}$$



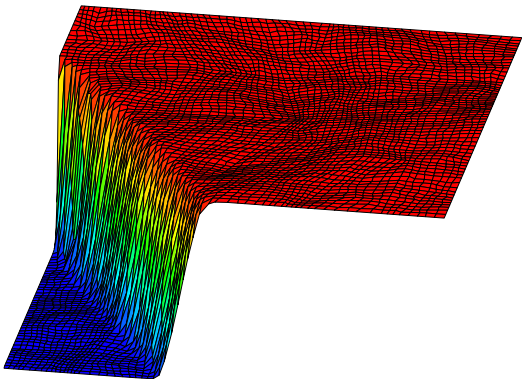
(a) Stabilized Galerkin solution on uniform mesh.



(b) Stabilized Galerkin solution on distorted mesh.



(c) Stabilized SUPG solution on uniform mesh.



(d) Stabilized SUPG solution on distorted mesh.

Figure 10: Convection of discontinuity: Different AFC solutions on level 6 of uniform and distorted mesh using $q = 4$ and $\varepsilon = 0$.

where the shapes of the 'smooth hump', 'sharp cone', and 'slotted cylinder' are determined by

$$\begin{aligned} u^{(1)}(\mathbf{x}) &= \tfrac{1}{4}\big(1+\cos(\pi\|\mathbf{x}\|_2)\big), \\ u^{(2)}(\mathbf{x}) &= 1-\|\mathbf{x}\|_2, \\ u^{(3)}(\mathbf{x}) &= \begin{cases} 1 & : |x_1| \geqslant \frac{1}{6} \vee x_2 \geqslant \frac{2}{3}, \\ 0 & : \text{otherwise} \end{cases} \end{aligned} \qquad \forall \mathbf{x} \in \mathbb{R}^2,\ \|\mathbf{x}\|_2 \leqslant 1. \tag{35}$$

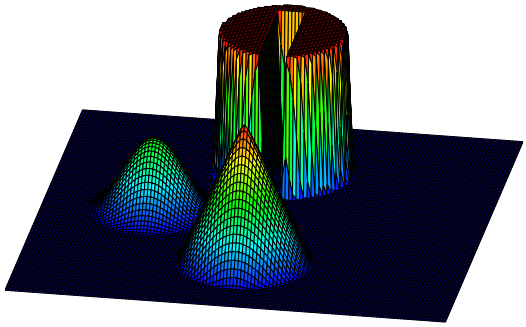
The bound-preserving discrete initial profile $u_{h,0}$ is computed using a lumped L^2-projection and is then evolved in time using the Crank-Nicolson scheme and $\Delta t \approx 10^{-3}$. This implicit time integrator requires the solution of a nonlinear system of equations in each time step and, hence, might not outperform an explicit time discretization. However, the θ-scheme with $\theta = \frac{1}{2}$ is used to illustrate the possibility of using implicit time integration techniques and the stability of the semi-discrete problem. In this case, the nonlinear solution update is performed using the following undamped fixed-point iteration based on a low order preconditioner:

0. Set $k^{(1)} = -\mathcal{M}_{\mathrm{L}}^{-1}\mathscr{F}(u^n, n\Delta t)$, $k^{(2,0)} = 0 \in \mathbb{R}^N$, and $m = 0$
1. Set $r^{(2,m)} = \mathcal{M}_{\mathrm{L}} k^{(2,m)} + \mathscr{F}(u^n + \frac{\Delta t}{2}k^{(1)} + \frac{\Delta t}{2}k^{(2,m)}, (n+1)\Delta t)$.
2. If $\|\mathcal{M}_{\mathrm{L}}^{-1} r^{(2,m)}\|_{\tilde{2}} \leqslant 10^{-8}\Delta t^{-1}$, set $k^{(2)} \coloneqq k^{(2,m)}$ and go to step 5.
3. Compute $\Delta k^{(2,m+1)} = -\big(\mathcal{M}_{\mathrm{L}} + \frac{\Delta t}{2}(\mathcal{A}-\mathcal{D})\big)^{-1} r^{(2,m)}$.
4. Set $k^{(2,m+1)} = k^{(2,m)} + \Delta k^{(2,m+1)}$ as well as $m = m+1$ and go to step 1.
5. Set $u^{n+1} = u^n + \frac{\Delta t}{2}k^{(1)} + \frac{\Delta t}{2}k^{(2)}$.

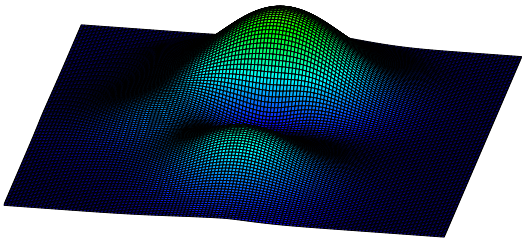
For the numerical results presented in Fig. 11, the dissipation parameter $\omega = 0.6$ is chosen to achieve an accurate high order target scheme which only creates some local ripples close to discontinuities. However, the corresponding finite element solution violates the global bounds of the exact solution and calls for some nonlinear stabilization. The AFC methodology using this high order target scheme, $q = 2$, and $\varepsilon = 0$ prevents the occurrence of local artifacts and produces a bound-preserving finite element solution which is very accurate for this test problem although some peak clipping effects can be observed at the summit of the smooth hump and the sharp cone.

7 Summary

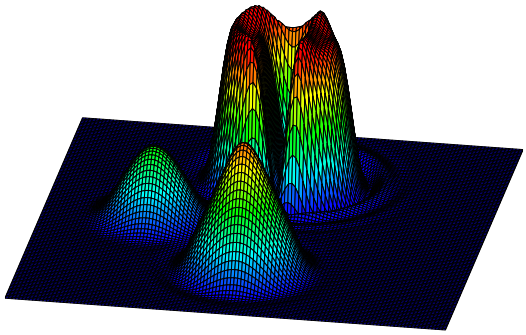
In this work, a new way of defining the correction factors of AFC schemes is proposed to reduce the numerical effort for solving the resulting nonlinear systems. The regularization technique introduced for this purpose makes the AFC residual twice continuously differentiable so that Newton's method converges quadratically for sufficiently good initial guesses. Although the advantage of this approach could be observed in numerical examples, the optimal choice of the regularization parameter balancing accuracy and efficient solvability is a challenging task. On the other hand, preconditioning the fixed-point iteration using the 'exact' Jacobian seems to work quite well even for the unregularized AFC scheme. A decomposition of the Jacobian is exploited to efficiently compute its entries and further reduce the numerical effort.



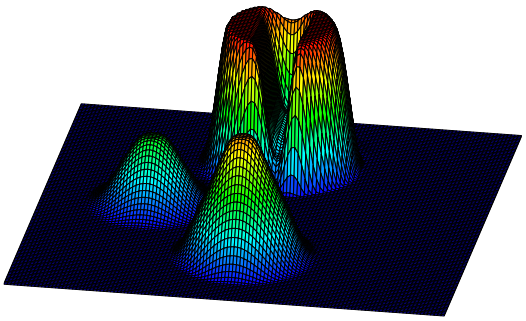
(a) Initial data using lumped L^2-projection.



(b) Low order solution using $\alpha_{ij} = \dot{\alpha}_{ij} = 0$.



(c) High order solution using $\alpha_{ij} = \dot{\alpha}_{ij} = 1$ and $\omega = 0.6$.



(d) AFC solution using $\omega = 0.6$, $q = 2$, and $\varepsilon = 0$.

Figure 11: Solid body rotation. Different discrete solutions on level 7 of uniform mesh.

The weights of the limiting function formula are defined so that the unregularized correction factors vanish if the solution is (locally) linear. Numerical experiments indicate that this so called linearity preservation property makes the solution second order accurate on uniform meshes. This order of convergence is preserved if the correction factors are smoothed using an appropriately chosen regularization technique.

Furthermore, the limiting strategy is extended to the treatment of transient problems. In this case, the proposed correction factors are designed so that the nonlinear AFC residual is Lipschitz continuous. This facilitates the use of implicit time integrators and guarantees well-posedness of steady state problems. If the backward Euler scheme is employed, bound-preserving solutions can be obtained without any CFL-like time step restriction. However, this scheme is only first order accurate in time and the use of inordinately large time steps may cause convergence problems. To achieve higher overall efficiency with implicit AFC schemes, further effort is worth expending on the development of customized limiting techniques that lead to nonlinear systems with desirable properties.

References

[Arm66] L. Armijo. "Minimization of functions having Lipschitz continuous first partial derivatives." In: *Pacific J. Math.* 16.1 (1966), pp. 1–3.

[BB17] S. Badia and J. Bonilla. "Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization". In: *Computer Methods in Applied Mechanics and Engineering* 313 (2017), pp. 133–158.

[Bad+19] S. Badia, J. Bonilla, S. Mabuza, and J. N. Shadid. *On differentiable local bounds preserving stabilization for Euler equations.* 2019.

[Bar+17a] 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". In: *Numerische Mathematik* 135.2 (02/2017), pp. 521–545.

[Bar+16] G. R. Barrenechea, V. John, and P. Knobloch. "Analysis of Algebraic Flux Correction Schemes". In: *SIAM Journal on Numerical Analysis* 54.4 (2016), pp. 2427–2451.

[Bar+17b] G. R. Barrenechea, V. John, and P. Knobloch. "An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes". In: *Mathematical Models and Methods in Applied Sciences* 27.03 (2017), pp. 525–548.

[Bar+18] G. R. Barrenechea, V. John, P. Knobloch, and R. Rankin. "A unified analysis of algebraic flux correction schemes for convection–diffusion equations". In: *SeMA Journal* 75 (05/2018), pp. 655–685.

[BH82] A. N. Brooks and T. J. Hughes. "Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations". In: *Computer Methods in Applied Mechanics and Engineering* 32.1 (1982), pp. 199–259.

[Hub07] M. Hubbard. “Non-oscillatory third order fluctuation splitting schemes for steady scalar conservation laws”. In: *Journal of Computational Physics* 222.2 (2007), pp. 740–768.

[JJ18] A. Jha and V. John. “On basic iteration schemes for nonlinear AFC discretizations”. In: *WIAS Preprint 2533* (2018). Weierstrass Institute for Applied Analysis and Stochastics.

[JJ19] A. Jha and V. John. “A study of solvers for nonlinear AFC discretizations of convection–diffusion equations”. In: *Computers & Mathematics with Applications* (2019).

[JK07] V. John and P. Knobloch. “On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I – A review”. In: *Computer Methods in Applied Mechanics and Engineering* 196.17 (2007), pp. 2197–2215.

[Kuz07] D. Kuzmin. “Algebraic flux correction for finite element discretizations of coupled systems”. In: *Computational Methods for Coupled Problems in Science and Engineering II, CIMNE, Barcelona* (2007), pp. 653–656.

[Kuz12] D. Kuzmin. “Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes”. In: *Journal of Computational and Applied Mathematics* 236.9 (2012), pp. 2317–2337.

[Kuz20] D. Kuzmin. “Monolithic convex limiting for continuous finite element discretizations of hyperbolic conservation laws”. In: *Computer Methods in Applied Mechanics and Engineering* 361 (2020), p. 112804.

[Kuz+17] D. Kuzmin, S. Basting, and J. N. Shadid. “Linearity-preserving monotone local projection stabilization schemes for continuous finite elements”. In: *Computer Methods in Applied Mechanics and Engineering* 322 (2017), pp. 23–41.

[KS17] D. Kuzmin and J. N. Shadid. “Gradient-based nodal limiters for artificial diffusion operators in finite element schemes for transport equations”. In: *International Journal for Numerical Methods in Fluids* 84.11 (2017), pp. 675–695.

[LeV96] R. J. LeVeque. “High-resolution conservative algorithms for advection in incompressible flow”. In: *SIAM Journal on Numerical Analysis* 33.2 (1996), pp. 627–665.

[Loh19a] C. Lohmann. *On the solvability and iterative solution of algebraic flux correction problems for convection–reaction equations.* Tech. rep. Ergebnisberichte des Instituts für Angewandte Mathematik, Nummer 612. Fakultät für Mathematik, TU Dortmund, 2019.

[Loh19b] C. Lohmann. *Physics-Compatible Finite Element Methods for Scalar and Tensorial Advection Problems.* Springer Fachmedien Wiesbaden, 2019, pp. 1–283.

[Loh+17] C. Lohmann, D. Kuzmin, J. N. Shadid, and S. Mabuza. “Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements”. In: *Journal of Computational Physics* 344 (2017), pp. 151–186.

[Mab+18] S. Mabuza, J. N. Shadid, and D. Kuzmin. “Local bounds preserving stabilization for continuous Galerkin discretization of hyperbolic systems”. In: *Journal of Computational Physics* 361 (2018), pp. 82–110.

[Möl07] M. Möller. “Efficient solution techniques for implicit finite element schemes with flux limiters”. In: *International Journal for Numerical Methods in Fluids* 55.7 (2007), pp. 611–635.

[Möl08] M. Möller. “Adaptive high-resolution finite element schemes”. PhD thesis. PhD thesis, Dortmund University of Technology, 2008.

[She+90] A. Shestakov, D. Kershaw, and G. Zimmerman. “Test problems in radiative transfer calculations”. In: *Nuclear Science and Engineering* 105.1 (1990), pp. 88–104.

[Zal79] S. T. Zalesak. “Fully Multidimensional Flux-Corrected Transport Algorithms for Fluids”. In: *Journal of Computational Physics* 31.3 (1979), pp. 335–362.