A Study of Solvers for Nonlinear AFC Discretizations of Convection-Diffusion Equations

Abhinav Jha[b], Volker John[a,b,*]

[a] *Weierstrass Institute for Applied Analysis and Stochastics (WIAS), Mohrenstr. 39, 10117 Berlin, Germany*
[b] *Freie Universität Berlin, Department of Mathematics and Computer Science, Arnimallee 6, 14195 Berlin, Germany*

Abstract

Nonlinear discretizations are necessary for convection-diffusion equations for obtaining accurate solutions that satisfy the discrete maximum principle. The numerical solution of the arising nonlinear problems is often difficult. This paper presents several approaches for solving the nonlinear problems of algebraic flux correction (AFC) schemes for the Kuzmin limiter and the BJK limiter. Comprehensive numerical studies are performed at examples that model the transport of energy from a body in a flow field in two and three dimensions. It turns out that the most efficient approach, from the point of view of computing times, is a simple fixed point iteration, because the iteration matrix possesses properties that can be exploited by the solvers of the arising linear systems of equations.

Keywords: steady-state convection-diffusion equations, algebraic flux correction (AFC) schemes, Kuzmin and BJK limiter, mixed fixed point iteration, formal Newton methods

1. Introduction

Partial differential equations (PDEs) are used to model many processes in nature and industry. The solutions of these equations should reflect important features of the modeled process, like mass conservation or the restriction of the values to an admissible interval. But usually, these solutions cannot be computed analytically and some numerical method has to be applied to calculate approximations. From the practical point of view, it is often essential that also the numerical solutions possess those features which are of importance for the solutions of the PDE. This property is often of utmost significance for a numerical solution to be accepted by practitioners. However, many discretizations do not lead to numerical solutions that respect the important features of the solution of the PDE. As already mentioned, such solutions might be rejected in practice and therefore it is worthwhile to study in much detail discretizations that do respect these features.

The PDE considered in this paper is the steady-state convection-diffusion equation

$$\begin{aligned} -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u &= 0 \quad \text{in } \Omega, \\ u &= u^b \quad \text{on } \Gamma_D, \\ -\varepsilon\nabla u\cdot\boldsymbol{n} &= 0 \quad \text{on } \Gamma_N, \end{aligned} \tag{1}$$

where $\Omega \subset \mathbb{R}^d$, $d \in \{2,3\}$ is a bounded domain with boundary $\Gamma = \Gamma_D \cup \Gamma_N$, $\Gamma_D \cap \Gamma_N = \emptyset$, $\text{meas}_{d-1}(\Gamma_D) > 0$, and $\boldsymbol{n}$ being the outward pointing unit normal

*Corresponding author
Email addresses: jha@wias-berlin.de (Abhinav Jha), john@wias-berlin.de (Volker John)

on Γ. Problem (1) models the transport of the scalar quantity u (temperature, energy). The transport consists of a molecular transport with the constant diffusion coefficient $\varepsilon > 0$ and a convective transport with a flow field $\boldsymbol{b}$ with $\nabla \cdot \boldsymbol{b} = 0$. On Γ_D, Dirichlet boundary conditions are set and on Γ_N, Neumann boundary conditions are prescribed.

An important feature of the solution of (1) is that u satisfies the maximum principle (MP), i.e., u takes its minimal and maximal value at the Dirichlet boundary Γ_D. Since there are no sources and sinks in (1), there is no physical mechanism for obtaining lower or larger values than the extremal values at the Dirichlet boundary. Another feature is that in practice the convective transport is usually much stronger than the molecular transport, i.e., it is $\|\boldsymbol{b}\|_{L^\infty(\Omega)} \gg \varepsilon$, which is called the convection-dominated regime.

Using standard discretizations for (1) in the convection-dominated regime, like the central finite difference method or the Galerkin finite element method, on affordable grids leads to numerical solutions that are globally polluted with spurious oscillations and which are useless in practice. The development of discretizations for this regime focused on introducing terms for stabilization and often also on achieving a high order of error reduction in certain norms. The starting point of this development was the SUPG method (Streamline-Upwind Petrov–Galerkin) from [11, 6]. Meanwhile, many other approaches have been proposed, see [23] for a comprehensive overview. However, it turned out that most discretizations do not satisfy the discrete maximum principle (DMP), which is the analog of the MP. Other discretizations, like certain upwind schemes, satisfy the DMP, but the numerical solutions are quite inaccurate, in particular the layers of the solution are heavily smeared. This situation might be inevitable for linear discretizations of (1), since, in the limit case $\varepsilon = 0$, it is known that a linear discretization leading to an M-matrix, which is a usual criterion for the satisfaction of the DMP, cannot possess a local discretization error of second order, see [28, Chap. 4.4]. A similar mathematical result for small positive ε is not known, but the experience is that the same situation holds also in this case. At any rate, already in the 1980's nonlinear discretizations of (1) were proposed, e.g., in [22, 12], to reduce or remove the spurious oscillations of linear stabilized schemes. The nonlinearity is introduced by a stabilization parameter that depends on the numerical solution. However, in comprehensive studies [14, 15], it was shown that only very few of the proposed nonlinear schemes remove or significantly reduce spurious oscillations. These studies did not yet include algebraic stabilizations. However, in the subsequent study [1], the algebraic flux correction (AFC) scheme with Kuzmin limiter, which is also included in the studies of the present paper, proved to compute numerical solutions without over- and undershoots in a two-dimensional benchmark problem.

With the nonlinearity, a new issue arises: the efficient solution of the nonlinear problems. As noted in [16], an ideal discretization of (1) should satisfy the following properties:

- the numerical solution should be accurate, in particular it should exhibit sharp layers,
- the numerical solution must not have spurious oscillations,
- the numerical solution has to be computed efficiently.

And as also noted in [16], currently no scheme is known that satisfies all these properties.

AFC schemes, the topic of this paper, satisfy the first two requirements. However, as it can be seen in the literature, e.g., in [1, 5], an efficient solver for the nonlinear problems is not yet available. This paper addresses this issue: several approaches for solving the nonlinear problems will be presented and studied comprehensively. The first steps of this study were already performed in [13]. In this

paper, two basic fixed point iterations and a basic formal Newton method were investigated at simple academic examples in two dimensions. From the algorithmic point of view, the current paper proposes a mixed fixed point iteration, a refinement of the formal Newton method, and it considers a regularized formal Newton method. Additional algorithmic components included in the studies are Anderson acceleration [27] and the projection to admissible values [2]. The numerical studies were performed for examples in two and three dimensions that model flows around a body and the transport of u from this body in downstream direction. We are aware of only one rather short presentation of a 3d example in the literature so far in [5]. In 3d, an iterative solver for the linear problems of equations is used, which seems to be also a new aspect within the framework of AFC schemes.

Another scheme that satisfies the first two properties, and thus is an alternative to AFC schemes, is the Mizukami–Hughes method [22] with the improvements from [19]. However, it is reported [14], where this method is applied to steady-state convection-diffusion equations, that the solution of the nonlinear problems arising in this scheme might be also difficult. Thus, the efficient solution of the equations arising in nonlinear discretizations of (1) seems to be a more general difficulty.

2. Algebraic Flux Correction Schemes

Presentations of AFC schemes can be found already elsewhere in the literature. However, for this paper to be self-contained, and above all, for the description of the algorithms for solving the nonlinear problems, we think it inevitable to repeat a presentation of these schemes here. To be more general, a convection-diffusion equation is considered where the right-hand side is some function f, instead of the homogeneous right-hand side in (1).

2.1. General Approach

Applying a conforming Galerkin discretization with P_1 or Q_1 finite elements for discretizing the convection-diffusion problem with homogeneous Neumann boundary conditions leads to a linear system of equations $A\underline{u} = \underline{f}$, $A = (a_{ij})_{i,j=1}^n \in \mathbb{R}^{n\times n}, \underline{u}, \underline{f} \in \mathbb{R}^n$, or in detail

$$\sum_{j=1}^{n} a_{ij}u_j = f_i, \quad i = 1, \ldots, n. \tag{2}$$

Now, one defines a symmetric artificial diffusion matrix $D = (d_{ij})_{i,j=1}^n$ by

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \quad \text{for } i \neq j, \quad d_{ii} = -\sum_{j=1, j\neq i}^{n} d_{ij}. \tag{3}$$

With this matrix, (2) can be written as

$$(\hat{A}\underline{u})_i = f_i + (D\underline{u})_i, \quad i = 1, \ldots, n, \quad \text{with } \hat{A} = A + D. \tag{4}$$

By construction, the row sums of D vanish. Hence, it holds

$$(D\underline{u})_i = \sum_{j=1}^{n} d_{ij}u_j - u_i \sum_{j=1}^{n} d_{ij} = \sum_{j=1, j\neq i}^{n} d_{ij}(u_j - u_i) = \sum_{j=1, j\neq i}^{n} f_{ij}, \quad i = 1, \ldots, n,$$

where $f_{ij} = d_{ij}(u_j - u_i) = -f_{ji}$ are the so-called fluxes. The goal of AFC schemes consists in limiting those fluxes that cause spurious oscillations by introducing

solution-dependent weights $\alpha_{ij} = \alpha_{ij}(\underline{u}) = \alpha_{ji}(\underline{u}) \in [0,1]$ and considering instead of (4) the nonlinear system of equations

$$(\hat{A}\underline{u})_i = f_i + \sum_{j=1, j\neq i}^{n} \alpha_{ij} f_{ij}, \quad i = 1, \ldots, n.$$

At this point, the Dirichlet boundary conditions are incorporated in the AFC scheme. Let the entries are ordered such that the $(n-m)$ Dirichlet values, $m < n$, are at the end of $\underline{u}$, then the system of equations takes the form

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} u_j + \sum_{j=1}^{n} (1-\alpha_{ij}) d_{ij} (u_j - u_i) &= f_i, \qquad i = 1, \ldots, m, \\ u_i &= u_i^b, \qquad i = m+1, \ldots, n. \end{aligned} \tag{5}$$

The symmetry of the limiter is important for the AFC method to be conservative, see [21], and for proving the existence of a solution of (5), see [3].

By construction, the diagonal entries of $\hat{A}$ are positive and the off-diagonal entries non-positive. After having incorporated the Dirichlet boundary conditions, the resulting modified matrix $\hat{A}$ can be expected to be an irreducibly diagonally dominant matrix in usual situations. From all these properties, it follows that $\hat{A}$ is an M-matrix, [26].

In the literature concerning algebraic stabilizations for steady-state convection-diffusion equations, essentially two types of limiters are proposed.

2.2. The Kuzmin Limiter

This limiter, proposed in [20], starts by computing

$$P_i^+ = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{n} f_{ij}^+, \; P_i^- = \sum_{\substack{j=1 \\ a_{ji} \leq a_{ij}}}^{n} f_{ij}^-, \; Q_i^+ = -\sum_{j=1}^{n} f_{ij}^-, \; Q_i^- = -\sum_{j=1}^{n} f_{ij}^+, \tag{6}$$

$i = 1, \ldots, n$, where $f_{ij}^+ = \max\{0, f_{ij}\}$ and $f_{ij}^- = \min\{0, f_{ij}\}$. Next, one calculates

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \ldots, m. \tag{7}$$

If P_i^+ or P_i^- is zero, one sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. At Dirichlet nodes, one sets

$$R_i^+ = 1, \quad R_i^- = 1, \quad i = m+1, \ldots, n. \tag{8}$$

Finally, for any $i, j \in \{1, \ldots, n\}$ such that $a_{ji} \leq a_{ij}$, the limiter is defined by

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \\ 1 & \text{if } f_{ij} = 0 \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}, \quad \alpha_{ji} = \alpha_{ij}. \tag{9}$$

The Kuzmin limiter can be applied to P_1 and Q_1 finite elements, see [3] for some details of its implementation. In [3], the Kuzmin limiter was analyzed for steady-state convection-diffusion-reaction equations with Dirichlet boundary conditions and P_1 finite elements. Existence of a solution of the nonlinear problem and the satisfaction of the DMP, under the restriction that the mesh is weakly acute, are proved. The uniqueness of the solution as well as the extension of the analysis to mixed boundary conditions are open problems. We like to note that for the Galerkin method with P_1 finite elements and diffusion-reaction equations, one can find an analysis of the DMP in the case of mixed boundary conditions in [18].

2.3. The BJK Limiter.

This limiter was developed in [4] for P_1 finite elements. As first step, one defines for $i = 1, \ldots, n$

$$u_i^{\max} = \max_{j \in S_i \cup \{i\}} u_j, \quad u_i^{\min} = \min_{j \in S_i \cup \{i\}} u_j, \quad q_i = \gamma_i \sum_{j \in S_i} d_{ij}, \tag{10}$$

where γ_i is a positive constant that was computed for interior nodes as in [4, Rem. 6.2] and the index set S_i was to be chosen as the set of all degrees of freedom $j \neq i$ for which there is an entry in the sparsity pattern of A, i.e., S_i is the set of all direct neighbor degrees of freedom of i. For nodes on the Neumann boundary, γ_i was computed using the same formulas as for interior nodes with the natural restriction of the patch of mesh cells corresponding to the Neumann node. As next step, one computes for $i = 1, \ldots, m$

$$P_i^+ = \sum_{j \in S_i} f_{ij}^+, \; P_i^- = \sum_{j \in S_i} f_{ij}^-, \; Q_i^+ = q_i(u_i - u_i^{\max}), \; Q_i^- = q_i(u_i - u_i^{\min}), \tag{11}$$

and then, one sets

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

If P_i^+ or P_i^- vanishes, one sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. Then, (8) is applied for the Dirichlet nodes and the quantities

$$\bar{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \\ 1 & \text{if } f_{ij} = 0 \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}, \quad i = 1, \ldots, m, \; j = 1, \ldots, n, \tag{12}$$

are calculated. Finally, one sets

$$\alpha_{ij} = \min\{\bar{\alpha}_{ij}, \bar{\alpha}_{ji}\}, \quad i, j = 1, \ldots, m, \tag{13}$$

$$\alpha_{ij} = \bar{\alpha}_{ij}, \quad i = 1, \ldots, m, \; j = m+1, \ldots, n. \tag{14}$$

It is proved in [4], in the case of Dirichlet boundary conditions, that a solution of the AFC method (5) exists, that it satisfies the DMP and it is linearity preserving, all on arbitrary simplicial grids. The uniqueness of the solution and the study of mixed boundary conditions are open questions.

3. Nonlinear Iteration Schemes

Consider the nonlinear problem (5) in the form

$$\begin{aligned} F(\underline{u}) &= \underline{0} \quad \text{with} \qquad (15) \\ F_i(\underline{u}) &= \sum_{j=1}^{n} a_{ij} u_j + \sum_{j=1}^{n} (1 - \alpha_{ij}(\underline{u})) d_{ij}(u_j - u_i) - f_i = 0, \quad i = 1, \ldots, m, \\ F_i(\underline{u}) &= u_i - u_i^b = 0, \quad i = m+1, \ldots, n. \end{aligned}$$

Then, a damped iteration for solving (15) is given by

$$\underline{u}^{(\nu+1)} = \underline{u}^{(\nu)} - \omega^{(\nu)} B^{-1} F\left(\underline{u}^{(\nu)}\right), \quad \nu = 0, 1, \ldots, \tag{16}$$

where $B \in \mathbb{R}^{n \times n}$ is a non-singular matrix. A vector $\underline{u}$ is a solution of the nonlinear problem (5) if and only if it is a fixed point of (16). The choice of the damping parameter $\omega^{(\nu)}$ is briefly discussed in Section 4.1.

3.1. The Mixed Fixed Point Iteration

Utilizing some kind of simple fixed point iteration is a natural starting point for the construction of solvers for the nonlinear problem (15). A straightforward idea consists in using for the construction of the left-hand side of (15) the currently available values for the limiter, leading in the iteration step $(\nu+1)$ to a linear system of equations of the form

$$\begin{aligned} \sum_{j=1}^{n} a_{ij} u_j^{(\nu+1)} + \sum_{j=1}^{n} \left(1 - \alpha_{ij}^{(\nu)}\right) d_{ij} \left(u_j^{(\nu+1)} - u_i^{(\nu+1)}\right) &= f_i, \quad i = 1, \ldots, m, \\ u_i^{(\nu+1)} &= u_i^b, \quad i = m+1, \ldots, n, \end{aligned} \tag{17}$$

with $\alpha_{ij}^{(\nu)} = \alpha_{ij}\left(u^{(\nu)}\right)$. This method is called *fixed point matrix* in [13]. It is shown in [3, 4] that, in the case of Dirichlet boundary conditions, the linear system (17) has a unique solution for both the Kuzmin and the BJK limiter.

Another simple fixed point iteration can be derived by using that the row sums of the matrix D vanish, such that

$$\sum_{j=1}^{n} \left(1 - \alpha_{ij}^{(\nu)}\right) d_{ij} \left(u_j^{(\nu+1)} - u_i^{(\nu+1)}\right) = \sum_{j=1}^{n} d_{ij} u_j^{(\nu+1)} - \sum_{j=1}^{n} \alpha_{ij}^{(\nu)} d_{ij} \left(u_j^{(\nu+1)} - u_i^{(\nu+1)}\right).$$

Then, a fixed point iteration is given by

$$\begin{aligned} \sum_{j=1}^{n} (a_{ij} + d_{ij}) u_j^{(\nu+1)} &= f_i + \sum_{j=1}^{n} \alpha_{ij}^{(\nu)} f_{ij}^{(\nu)}, \quad i = 1, \ldots, m, \\ u_i^{(\nu+1)} &= u_i^b, \qquad\qquad\qquad\quad i = m+1, \ldots, n, \end{aligned} \tag{18}$$

where $f_{ij}^{(\nu)}$ is the flux computed with the limiter $\alpha_{ij}^{(\nu)}$. In [13], this method is called *fixed point rhs*. A distinct feature of *fixed point rhs* is that the matrix $A + D = \hat{A}$ does not depend on the iterate and thus, in each iteration step, the matrix of the linear system of equations to be solved is the same. Hence, applying a sparse direct solver, the whole iteration requires just one matrix factorization in the first iteration step and in all subsequent iterations, only two triangular systems have to be solved.

Remark 1. The methods *fixed point matrix* and *fixed point rhs* were already studied in [13] at academic examples in two dimensions. In these studies, it could be observed that both methods behaved often rather differently. The method *fixed point matrix* often failed to converge on fine grids. The studies in [13] applied a sparse direct solver. It turned out that if *fixed point rhs* did converge, it was by far more efficient than *fixed point matrix*.

The numerical studies in Section 5 will consider also examples in three dimensions. In this situation, the sparse factorization of a sparse matrix is much more involved than in two dimensions, such that the use of iterative solvers for the arising linear systems of equations becomes necessary. For iterative solvers, it is a priori not of advantage for *fixed point rhs* that there is the same matrix in each iteration step. However, the matrices of *fixed point rhs* and *fixed point matrix* are different and iterative methods might behave differently.

Our expectation before performing the numerical studies of [13] was that the method *fixed point matrix* might need generally fewer iterations than *fixed point rhs*, because *fixed point matrix* is a less explicit method since it uses the current iterate for assembling the matrix and not only for assembling the right-hand side. However, as indicated in Remark 1, the expectation was not met. But we think

that a less explicit fixed point iteration than *fixed point rhs* is worth to be studied. To this end, we define the mixed fixed point iteration

$$\begin{aligned} \sum_{j=1}^{n} (a_{ij} + d_{ij})\, u_j^{(\nu+1)} - \omega_{\mathrm{fp}} \sum_{j=1}^{n} \alpha_{ij}^{(\nu)} d_{ij} \left(u_j^{(\nu+1)} - u_i^{(\nu+1)} \right) & \\ = \quad f_i + (1 - \omega_{\mathrm{fp}}) \sum_{j=1}^{n} \alpha_{ij}^{(\nu)} f_{ij}^{(\nu)}, \quad i = 1, \ldots, m, & \\ u_i^{(\nu+1)} \quad = \quad u_i^b, \qquad\qquad\qquad\qquad i = m+1, \ldots, n, & \end{aligned} \tag{19}$$

with the mixing parameter $\omega_{\mathrm{fp}} \in [0, 1]$. For $\omega_{\mathrm{fp}} = 0$, one gets *fixed point rhs* and for $\omega_{\mathrm{fp}} = 1$, the method *fixed point matrix* is obtained. With respect to the fixed point iteration (16), method (19) uses the matrix B with

$$B\left(\underline{u}^{(\nu)}\right)_{ij} = \begin{cases} a_{ij} + d_{ij} - \omega_{\mathrm{fp}} \alpha_{ij}^{(\nu)} d_{ij} & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \omega_{\mathrm{fp}} \displaystyle\sum_{j=1, j\neq i}^{n} \alpha_{ij}^{(\nu)} d_{ij} & \text{if } i = j, \end{cases}$$

for $i = 1, \ldots, m, j = 1, \ldots, n$. The last $n - m$ rows have just the diagonal entry 1. Comprehensive numerical studies with the method *mixed fixed point*(ω_{fp}) from (19) are presented in Section 5.

3.2. A Formal Newton Method

This section presents a formal Newton method for solving (15). We call this method formal because, as it will be discussed below, there are situations where the differentiability requirements for Newton's method are not satisfied.

3.2.1. Derivation

For Newton's method, the matrix B in (16) is the Jacobian of F. Considering (15) for $i = 1, \ldots, m$, one can compute the Jacobian formally, using standard calculus, as

$$\begin{aligned} DF_i(\underline{u})[\underline{v}] \quad = \quad & \sum_{j=1}^{n} a_{ij} v_j + \sum_{j=1}^{n} (1 - \alpha_{ij}(\underline{u})) d_{ij} (v_j - v_i) \\ & - \sum_{j=1}^{n} \left(\sum_{k=1}^{m} \frac{\partial \alpha_{ij}}{\partial u_k}(\underline{u}) v_k \right) d_{ij} (u_j - u_i) \\ = \quad & \sum_{j=1}^{n} a_{ij} v_j + \sum_{j=1}^{n} (1 - \alpha_{ij}(\underline{u})) d_{ij} v_j - \left(\sum_{j=1}^{n} (1 - \alpha_{ij}(\underline{u})) d_{ij} \right) v_i \\ & - \sum_{j=1}^{n} \left(\sum_{k=1}^{m} \frac{\partial \alpha_{ij}}{\partial u_k}(\underline{u}) v_k \right) d_{ij} (u_j - u_i). \end{aligned}$$

Hence, the entries of the matrix that has to be inverted in (16) are given by

$$\begin{aligned} B\left(\underline{u}^{(\nu)}\right)_{ij} &= DF\left(\underline{u}^{(\nu)}\right)_{ij} \\ &= \begin{cases} a_{ij} + d_{ij} - \alpha_{ij}^{(\nu)} d_{ij} - \displaystyle\sum_{k=1}^{n} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_j} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \displaystyle\sum_{k=1, k\neq i}^{n} \alpha_{ik}^{(\nu)} d_{ik} - \sum_{k=1}^{n} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_i} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) & \text{if } i = j, \end{cases} \end{aligned} \tag{20}$$

for $i = 1, \ldots, m, j = 1, \ldots, n$. The last $n - m$ rows have just the diagonal entry 1.

One can see that in the Jacobian the partial derivatives of the limiter with respect to the solution vector are contained. The application of Newton's method requires smoothness of the limiter such that all terms in (20) are well defined. This property is not given, neither for the Kuzmin limiter nor for the BJK limiter.

For the presentation of one approach below, it is of advantage to start with a different representation of the Jacobian. Let $\beta_{ik}^{(\nu)} = \alpha_{ik}^{(\nu)} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right)$. Then, it is

$$\begin{aligned} \frac{\partial \beta_{ik}^{(\nu)}}{\partial u_j} &= \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_j} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) + \alpha_{ik}^{(\nu)} \frac{\partial \left(d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) \right)}{\partial u_j} \\ &= \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_j} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) + \alpha_{ik}^{(\nu)} d_{ik} \begin{cases} 1 & \text{if } k = j \neq i, \\ -1 & \text{if } i = j \neq k, \\ 0 & \text{else.} \end{cases} \end{aligned}$$

Now, the entries (20) of the Jacobian are given as follows

$$B \left(\underline{u}^{(\nu)} \right)_{ij} = DF \left(\underline{u}^{(\nu)} \right)_{ij} = a_{ij} + d_{ij} - \sum_{k=1}^{n} \frac{\partial \beta_{ik}^{(\nu)}}{\partial u_j} \tag{21}$$

for $i = 1, \ldots, m, j = 1, \ldots, n$. The last $n - m$ rows have only an entry on the diagonal that is 1.

3.2.2. Kuzmin Limiter

The non-smoothness of the Kuzmin limiter is introduced by computing minima and maxima of two values. For this limiter, we pursued two approaches. In the first one, the non-smooth situations are treated separately. The second approach uses a regularization.

Approach with Separate Treatment of the Non-Smooth Points. This approach uses the representation (20) of the Jacobian. In the minima and maxima contained in the Kuzmin limiter, one value is always constant. Thus, there is a one-sided derivative that vanishes. In this approach, the derivative that appears in the Jacobian is set to be zero in these situations.

Consider first the case $a_{ki} \leq a_{ik}$. Then, the entry of the Jacobian is set to be zero if $(f_{ik} > 0) \wedge R_i^+ = 1$, $f_{ik} = 0$, or $(f_{ik} < 0) \wedge R_i^- = 1$. Note that the situations $P_i^+ = 0$ and $P_i^- = 0$ are included in these cases.

In all other situations, the limiter is differentiable. With the product rule, one gets for the case $(f_{ik} > 0) \wedge R_i^+ < 1$

$$\frac{\partial \alpha_{ik}}{\partial u_j} = \frac{\frac{\partial Q_i^+}{\partial u_j} P_i^+ - Q_i^+ \frac{\partial P_i^+}{\partial u_j}}{\left(P_i^+ \right)^2},$$

and for the case $(f_{ik} < 0) \wedge R_i^- < 1$

$$\frac{\partial \alpha_{ik}}{\partial u_j} = \frac{\frac{\partial Q_i^-}{\partial u_j} P_i^- - Q_i^- \frac{\partial P_i^-}{\partial u_j}}{\left(P_i^- \right)^2}.$$

Hence, one has to compute the derivatives of $P_i^+, P_i^-, Q_i^+, Q_i^-$ with respect to u_j.

Using (6) and the definition of f_{ik}, one obtains, e.g.,

$$\begin{aligned}\frac{\partial Q_i^+}{\partial u_j} &= -\frac{\partial}{\partial u_j}\sum_{l=1}^{n} f_{il}^- = -\frac{\partial}{\partial u_j}\sum_{l=1}^{n}\min\left\{0, d_{il}(u_l-u_i)\right\},\\ &= \begin{cases} 0 & \text{if } f_{ij}\geq 0, i\neq j,\\ -d_{ij} & \text{if } f_{ij}<0, i\neq j,\\ \displaystyle\sum_{l=1, f_{il}<0}^{n} d_{il} & \text{if } i=j,\end{cases}\end{aligned} \tag{22}$$

and

$$\frac{\partial P_i^+}{\partial u_j} = \begin{cases} 0 & \text{if } f_{ij}\leq 0, i\neq j,\\ d_{ij} & \text{if } f_{ij}>0, i\neq j, a_{ji}\leq a_{ij},\\ 0 & \text{if } f_{ij}>0, i\neq j, a_{ji}>a_{ij},\\ \displaystyle -\sum_{\substack{l=1, f_{il}>0\\ a_{li}\leq a_{il}}}^{n} d_{il} & \text{if } i=j.\end{cases}$$

In a similar way, the other derivatives can be calculated.

In the case $a_{ki} > a_{ik}$, it is $\alpha_{ik} = \alpha_{ki}$, compare (9). Now, one can proceed in the same way as for the other case and one derives the same type of formulas: only the index i has to be replaced by the index k.

Approach with Regularization of the Non-Smooth Points. For the approximation of the maximum, a proposal is used that can be found, e.g., in [2]

$$\max_\sigma(x,y) = \frac{1}{2}\left(x+y+\sqrt{(x-y)^2+\sigma}\right) \tag{23}$$

with some small value $\sigma > 0$. Consequently, one has

$$\min_\sigma(x,y) = -\max_\sigma(-x,-y) = \frac{1}{2}\left(x+y-\sqrt{(x-y)^2+\sigma}\right).$$

In this approach, the formulation (21) of the Jacobian is utilized. In the case $a_{ki} \leq a_{ik}$, the starting point is the representation

$$\beta_{ik} = R_i^+ f_{ik}^+ + R_i^- f_{ik}^-,$$

where the superscript ν is neglected to simplify the notation. Regularizations of functions will be denoted with a tilde. Then, the following regularization is considered

$$\tilde{\beta}_{ik} = \min_\sigma\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}, 1\right)\max_\sigma(f_{ik}, 0) + \min_\sigma\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}, 1\right)\min_\sigma(f_{ik}, 0). \tag{24}$$

A straightforward calculation, using the definitions of the regularized maximum and

minimum, yields

$$\begin{aligned}
&\frac{\partial\tilde{\beta}_{ik}}{\partial u_j} \\
&= \frac{1}{2}\left(1-\frac{\tilde{Q}_i^+/\tilde{P}_i^+-1}{\sqrt{\left(\tilde{Q}_i^+/\tilde{P}_i^+-1\right)^2+\sigma}}\right)\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}\right)\frac{1}{2}\left(f_{ik}+\sqrt{f_{ik}^2+\sigma}\right) \\
&\quad+\frac{1}{2}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}+1-\sqrt{\left(\tilde{Q}_i^+/\tilde{P}_i^+-1\right)^2+\sigma}\right)\frac{1}{2}\left(1+\frac{f_{ik}}{\sqrt{f_{ik}^2+\sigma}}\right)\frac{\partial f_{ik}}{\partial u_j} \\
&\quad+\frac{1}{2}\left(1-\frac{\tilde{Q}_i^-/\tilde{P}_i^--1}{\sqrt{\left(\tilde{Q}_i^-/\tilde{P}_i^--1\right)^2+\sigma}}\right)\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}\right)\frac{1}{2}\left(f_{ik}-\sqrt{f_{ik}^2+\sigma}\right) \\
&\quad+\frac{1}{2}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}+1-\sqrt{\left(\tilde{Q}_i^-/\tilde{P}_i^--1\right)^2+\sigma}\right)\frac{1}{2}\left(1-\frac{f_{ik}}{\sqrt{f_{ik}^2+\sigma}}\right)\frac{\partial f_{ik}}{\partial u_j}. \qquad (25)
\end{aligned}$$

Note that the first part of each term does not depend on the summation index k. It holds

$$\frac{\partial f_{ik}}{\partial u_j}=\begin{cases}-d_{jk}=-d_{ik} & \text{if } j=i\neq k,\\ d_{ij} & \text{if } j=k\neq i,\\ 0 & \text{else,}\end{cases}$$

and

$$\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}\right)=\frac{\frac{\partial\tilde{Q}_i^+}{\partial u_j}\tilde{P}_i^+-\tilde{Q}_i^+\frac{\partial\tilde{P}_i^+}{\partial u_j}}{\left(\tilde{P}_i^+\right)^2},\quad \frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}\right)=\frac{\frac{\partial\tilde{Q}_i^-}{\partial u_j}\tilde{P}_i^--\tilde{Q}_i^-\frac{\partial\tilde{P}_i^-}{\partial u_j}}{\left(\tilde{P}_i^-\right)^2}. \qquad (26)$$

It is $\tilde{f}_{ik}^+=\max_\sigma(f_{ik},0)>0$ and hence $\tilde{P}_i^+>0$ because $\tilde{P}_i^+$ is a sum of $\tilde{f}_{ik}^+$ and at least $\tilde{f}_{ii}^+$ appears in this sum. With the same argument, one finds that $\tilde{P}_i^-<0$. One gets

$$\begin{aligned}
\frac{\partial\tilde{Q}_i^+}{\partial u_j} &= -\sum_{l=1}^n\frac{\partial\min_\sigma(f_{il},0)}{\partial u_j}=-\frac{1}{2}\sum_{l=1}^n\left(1-\frac{f_{il}}{\sqrt{f_{il}^2+\sigma}}\right)d_{il}\frac{\partial(u_l-u_i)}{\partial u_j} \\
&= \begin{cases}-\frac{1}{2}\left(1-\frac{f_{ij}}{\sqrt{f_{ij}^2+\sigma}}\right)d_{ij} & \text{if } i\neq j,\\ \frac{1}{2}\sum\limits_{l=1,l\neq i}^n\left(1-\frac{f_{il}}{\sqrt{f_{il}^2+\sigma}}\right)d_{il} & \text{if } i=j.\end{cases} \qquad (27)
\end{aligned}$$

This expression is compared with the corresponding expression (22) for the approach without regularization. Consider the case $i\neq j$. If $f_{ij}>0$ is sufficiently large, then the expression in the parentheses in (27) is very close to zero, which holds also for the value of (27). If $f_{ij}<0$ is sufficiently small, then the expression in the parentheses is close to two and the value of (27) is close to $-d_{ij}$. In both cases, the values of (22) and (27) are practically the same. In the situation $f_{ij}=0$, the value of (27) is $-d_{ij}/2$, which is different to the value 0 of (22) if $d_{ij}\neq 0$.

Again, the other derivatives can be computed in the same way.

If $a_{ki}>a_{ik}$, one gets with (9) that $\beta_{ik}=R_k^+f_{ik}^++R_k^-f_{ik}^-$. Now, one can proceed as in the other case for deriving formulas for the entries of the Jacobian.

The value of the regularization parameter was chosen similarly as in [2] by $\sigma = 10^{-8} \cdot h^4$, where h is the maximal diameter of the mesh cells of the current triangulation. In [2], also the limiter itself (shock detector) is regularized if the regularized Newton method is applied. Thus, strictly speaking, the discretization depends on the solution method. In our opinion, this situation is unusual and we decided not to use this approach but to apply the regularized Newton method to the standard Kuzmin limiter.

3.2.3. BJK Limiter

For the BJK limiter, only a formal Newton method with separate treatment of the non-smooth points is studied.

Approach with Separate Treatment of the Non-Smooth Points. The principal idea of this approach is the same as for the Kuzmin limiter. It is based on the representation (20) of the Jacobian. Again, several entries of this matrix are set to be zero in non-smooth points. This step is performed in the following cases, compare the definition of the α_{ik}: $(f_{ik} > 0) \wedge R_i^+ = 1$, $f_{ik} = 0$, and $(f_{ik} < 0) \wedge R_i^- = 1$.

Consider now the situation $(f_{ik} > 0) \wedge R_i^+ < 1$. Since $f_{ki} < 0$, one gets $\alpha_{ik} = \min\{R_i^+, R_k^-\}$. For $R_i^+ \leq R_k^-$, it follows that

$$\frac{\partial \alpha_{ik}}{\partial u_j} = \frac{\partial R_i^+}{\partial u_j} = \frac{P_i^+ \frac{\partial Q_i^+}{\partial u_j} - Q_i^+ \frac{\partial P_i^+}{\partial u_j}}{(P_i^+)^2},$$

and for $R_k^- < R_i^+$ that

$$\frac{\partial \alpha_{ik}}{\partial u_j} = \frac{\partial R_k^-}{\partial u_j} = \frac{P_k^- \frac{\partial Q_k^-}{\partial u_j} - Q_k^- \frac{\partial P_k^-}{\partial u_j}}{(P_k^-)^2}.$$

Using (11) for the definition of Q_i^+, one has

$$\frac{\partial Q_i^+}{\partial u_j} = \frac{\partial}{\partial u_j} q_i (u_i - u_i^{\max}) = \begin{cases} \begin{cases} -q_i & \text{if } u_i^{\max} = u_j, \\ 0 & \text{if } u_i^{\max} \neq u_j, \end{cases} & \text{if } i \neq j, \\ \begin{cases} 0 & \text{if } u_i^{\max} = u_j, \\ q_i & \text{if } u_i^{\max} \neq u_j, \end{cases} & \text{if } i = j. \end{cases}$$

In the same way, one gets the derivative of Q_k^-. The derivative of P_i^+ and P_i^- is obtained in the same way as for the Kuzmin limiter.

The second case that gives contribution to the Jacobian is $(f_{ik} < 0) \wedge R_i^- < 1$. This case can be treated analogously to the first one.

3.2.4. The General Iteration, Starting Newton's Method, Damping the Newton Contribution

A formal Newton method with damping is given by the following matrix in iteration (16)

$$B\left(\underline{u}^{(\nu)}\right)_{ij} = \begin{cases} a_{ij} + d_{ij} - \omega_{\text{fp}} \alpha_{ij}^{(\nu)} d_{ij} - \omega_{\text{Newt}} \sum_{k=1}^{n} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_j} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)}\right) & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \omega_{\text{fp}} \sum_{k=1, k \neq i}^{n} \alpha_{ik}^{(\nu)} d_{ik} - \omega_{\text{Newt}} \sum_{k=1}^{n} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_i} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)}\right) & \text{if } i = j, \end{cases} \tag{28}$$

with ω_{fp} being the damping parameter already introduced for the mixed fixed point iteration (19) and $\omega_{\text{Newt}} \in [0, 1]$ being a second damping parameter. The last $n - m$ rows have just the diagonal entry 1.

Remark 2. *Because of the conditions for achieving symmetry of the limiters, usually terms occur in the sums containing the derivatives of the limiters in* (28) *that do not fit into the sparsity pattern of the matrix A. This situation happens if* $\alpha_{ik}^{(\nu)}$ *is defined actually by* $\alpha_{ki}^{(\nu)}$ *and if there are nodes that are neighbors of the node k but not of the node i. All terms in the sums that do not fit into the sparsity pattern of A were neglected in our simulations.*

Remark 3. It is expected that the convergence radius of Newton-type methods is generally smaller than of simple fixed point iterations. Thus, it is advisable to start the solution process for the nonlinear problem (15) with a simple fixed point iteration and then switch to a Newton-type method. This approach was studied in [13]. It was found that a good criterion was to switch when the Euclidean norm of the residual vector was below 10^{-5}. Sometimes, one could observe that the norm of the residual vector increased after having switched to the formal Newton method. To avoid divergence, it was helpful to switch back to the simple fixed point iteration whenever the Euclidean norm of the residual vector was larger than 10^{-3}. Exactly this approach was used in the numerical studies presented in Section 5.

Remark 4. The formal Newton method studied in [13] used the fixed values $\omega_{\text{fp}} = 1$ and $\omega_{\text{Newt}} = 1$ and applied the strategies from Remarks 2 and 3.

In performing preliminary simulations for the examples considered in Section 5, we observed that the formal Newton method as used in [13] for simple academic test problems in two dimensions did often not worked. For this reason, we introduced the parameter ω_{Newt}. However, we found it sometimes complicated to fix an appropriate value for this parameter. For this reason, an initial value was chosen and

- ω_{Newt} was increased by the factor 1.001 after an iteration, if the Euclidean norm of the residual vector decreased at least by the factor 0.99,
- otherwise, ω_{Newt} was decreased by the factor 0.999.

Thus, in our adaptive formal Newton method, the parameter ω_{fp} is fixed (but usually not equal to 1) and ω_{Newt} changes accordingly to the progress of the iteration.

Concerning the calculation of the entries of the formal Jacobian, we like to note that computing the sum after the factor ω_{Newt} in (28) is considerably more costly than evaluating the other terms in (28), because of the many cases that have to be distinguished for computing the derivatives of $\alpha_{ik}^{(\nu)}$.

4. Further Algorithmic Components

4.1. Adaptive Choice of Damping Parameter

It is our experience that an appropriate choice of the damping parameters $\{\omega^{(\nu)}\}$ in (16) is often essential for the convergence of the iterative process and the number of iterations.

Choosing an appropriate damping parameter depends on a number of factors, like the problem and its data, the scheme used for discretizing the problem, the iterative scheme used to solve the system of equations, the grid, and the initial iterate. An a priori knowledge of all these information is generally not available. For this reason, an algorithm is desirable that chooses the damping parameter adaptively, e.g., based on the current behavior of the iterative scheme. Such an algorithm was proposed in [15], which includes also the rejection of iterates. In the

numerical studies presented in the current paper, exactly this algorithm was used. For the sake of brevity, this algorithm is not presented here in detail, but it is just referred to [15, Fig. 12].

4.2. Anderson Acceleration

Anderson acceleration is a process that tries to extract from the history of a linear fixed point iteration second order information. To this end, a parameter $\kappa \geq 1$ is chosen, which will be called here the number of Anderson vectors. The last κ iterates are stored and then, the new iterate is computed as a linear combination of the function values corresponding to these iterates, where the weights are computed by solving a least-squares problem.

The simulations presented in this paper utilized Algorithm AA from [27]. In the first κ steps, the linear fixed point iteration was performed and only after this, Anderson acceleration was started. The least-squares problem was solved with the LAPACK routine `dgglse`. The crucial parameter of this approach is the number of Anderson vectors. As already noted in [27], if κ is too small, then there might not be enough information to speed up the convergence sufficiently. But if κ is too large, the least-squares problem might be badly conditioned. The numerical studies in [27] used values in the range $\kappa \in [3, 50]$.

Anderson acceleration was already used for the solution of the nonlinear problem in AFC methods, e.g., in [1, 5]. In these papers, method (18) was applied and a constant damping parameter was used. Whereas in [1], a certain improvement compared with using method (18) with adaptive damping parameter is reported, the results in [5] show only small differences concerning the number of iterations. Note that in none of these papers, it was exploited that only one matrix factorization for the whole iteration is necessary for method (18). In the simulations presented here, Anderson acceleration was used in combination with the adaptive damping strategy from [15], but without rejection of steps.

In addition to Algorithm AA from [27], we implemented also the Anderson acceleration with the new iterate [27, (2.1)]. However, the results obtained with this approach were unsatisfactory, usually much worse than with Algorithm AA. For the sake of brevity, the corresponding results are not shown here.

4.3. Projection to Admissible Values

In the literature, the nonlinear problems from AFC discretizations are solved very accurately. The motivation for this approach is that the favorable properties, in particular the satisfaction of the DMP, hold only for the solution of the nonlinear problem.

In [2], it is proposed, for a time-dependent transport equation, to project each iterate to a space of admissible values. These values are given by a lower and an upper bound for the function values of the discrete solution. We like to note that such values are not always available in practice. For instance, in precipitation processes, particles grow by using the supersaturation of some species that are dissolved in a fluid. In this case, an upper bound for the concentration of the dissolution is not known, see [17] for a concrete example.

In the examples presented here, lower and upper admissible values of the solution are known. Therefore, the idea from [2] can be applied and we utilized exactly the same approach as in this paper: for each iterate, all values outside the admissible range are truncated to the closest border of this range before performing the next iteration step.

It has to be noted that the projection to admissible values only makes sense if it is clear a priori that the numerical solution satisfies the DMP. We like to recall that this property can be proved for the Kuzmin limiter only under restrictions on the mesh, see [3]. This aspect will be discussed for each numerical example in Section 5.

4.4. Choosing the Initial Condition

In [13], studies concerning the impact of the initial iterate on the number of iterations were performed. Four choices were investigated: choosing zero for all degrees of freedom, the solution of the Galerkin finite element method, the solution of the upwind finite element method from [25], and the solution of the SUPG (Streamline-Upwind Petrov–Galerkin) finite element method from [11, 6]. It was observed that there was only a minor impact. Generally, the solution of the SUPG finite element method with the choice of the stabilization parameter as given in [14] was a good choice. We performed similar studies for the examples considered in Section 5. For the sake of brevity, these studies are not presented. It turned out that also for these examples, the SUPG finite element solution was generally an appropriate starting iterate and it was used in all simulations presented below.

5. Numerical Studies

The numerical studies consider examples that model the transport of energy (temperature) in a flow field, a process which occurs in many applications. In all examples, the size of the convection field is of order $\mathcal{O}(1)$. A mildly convection-dominated case, $\varepsilon = 10^{-4}$, and a more strongly convection-dominated case, $\varepsilon = 10^{-6}$, were considered. In these studies, the following methods were involved:

- *mixed fixed point*(ω_{fp}): mixed fixed point iteration (19) with the parameter ω_{fp}. Note that *mixed fixed point*(0) corresponds to the method *fixed point rhs* from [13], see also (18), and *mixed fixed point*(1) to the method *fixed point matrix* from [13], compare (17).
- *mixed fixed point with Anderson acceleration*$(\omega_{\text{fp}}, \kappa)$: *mixed fixed point*$(\omega_{\text{fp}})$ with Anderson acceleration and κ Anderson vectors, see Section 4.2.
- *formal Newton* with separate treatment of the non-smooth points, as used in [13], compare Remark 4 for this method,
- *formal Newton* $(\omega_{\text{fp}}, \omega_{\text{Newt}})$ with separate treatment of the non-smooth points and adaptive change of ω_{Newt}, see Sections 3.2.2 and 3.2.3,
- *formal Newton* $(\omega_{\text{fp}}, \omega_{\text{Newt}})$ with regularization and adaptive change of ω_{Newt} (only for the Kuzmin limiter), see Section 3.2.2.

For all *formal Newton* methods apply the approaches discussed in Remarks 2 and 3.

Stopping criteria for solving the nonlinear equations were as follows:

- The Euclidean norm of the residual vector was smaller than $\sqrt{\#\text{ dof}} \cdot \text{tol}$, where # dof is the number of degrees of freedom (including Dirichlet nodes) and $\text{tol} = 10^{-10}$.
- A maximal number of 25000 accepted iterations was performed.

Below, the sum of accepted and rejected iterations is given since a rejected step has a similar computational cost as an accepted step. For simplicity of presentation, it is not distinguished in the pictures between simulations that did not converge within the prescribed maximal number of steps and simulations that diverged (with `inf` or `nan`); both are indicated by markers at 25000 or above. Diverged simulations are mentioned in the captions of the corresponding figures. The initial damping parameter was always set to be $\omega^{(0)} = 1$. All simulations were performed with the code ParMooN [8, 29] at compute servers HP BL460c Gen9 2xXeon, Fourteen-Core 2600MHz.

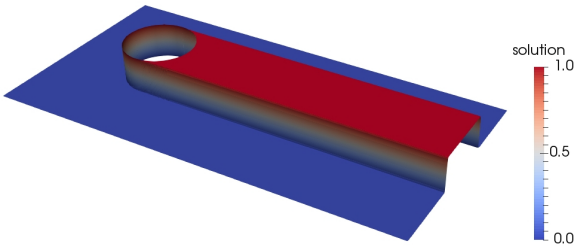

Figure 1: 2d Hemker problem. Solution for $\varepsilon = 10^{-6}$, computed with the BJK limiter, P_1, level 6.

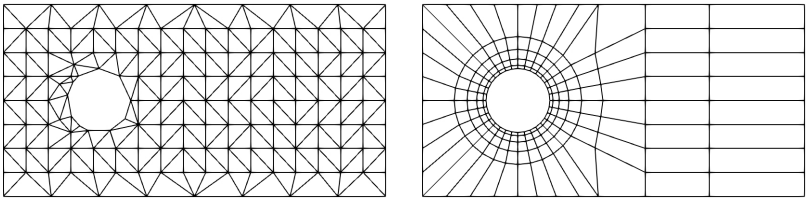

Figure 2: 2d Hemker problem. Triangular grid and quadrilateral grid (level 0).

Table 1: 2d Hemker problem. Number of degrees of freedom, including Dirichlet nodes.

level	P_1	Q_1
0	151	219
1	561	806
2	2158	3084
3	8460	12056
4	33496	47664
5	133296	189536
6	531808	755904

5.1. The 2d Hemker Problem

This example, defined in [10], is a standard benchmark problem for steady-state convection-diffusion equations. It is given by $\Omega = \{(-3,9) \times (-3,3)\} \setminus \{(x,y) \ : x^2 + y^2 \leq 1\}$, and $\boldsymbol{b} = (1,0)^T$ in (1). Dirichlet boundary conditions are set at $x = -3$, with $u^b = 0$, and at the circular boundary with $u^b = 1$. On all other boundaries, homogeneous Neumann conditions are prescribed. Reference values for the solution are available for $\varepsilon = 10^{-4}$. It was reported in [5] that in this case, the solutions obtained with the BJK limiter are more accurate than with the Kuzmin limiter, in particular the interior layers are sharper. The solution for $\varepsilon = 10^{-6}$ is illustrated in Figure 1. Simulations were performed on a triangular grid and a quadrilateral grid, see Figure 2 for the coarsest grids (level 0) and Table 1 for information on the number of degrees of freedom.

Concerning the satisfaction of the DMP, both grids from Figure 2 are not covered by the available analysis for the Kuzmin limiter. However, we could observe in preliminary simulations that the computed solutions with the Kuzmin limiter take values in $[0,1]$.

5.1.1. Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). First, the behavior of *mixed fixed point*(ω_{fp}) for $\omega_{\text{fp}} \in \{0, 0.05, \ldots, 0.95, 1\}$ is illustrated in Figure 3. The simulations were performed with and without the projection to admissible values as described in Section 4.3. One can see that there are only small differences with respect to the

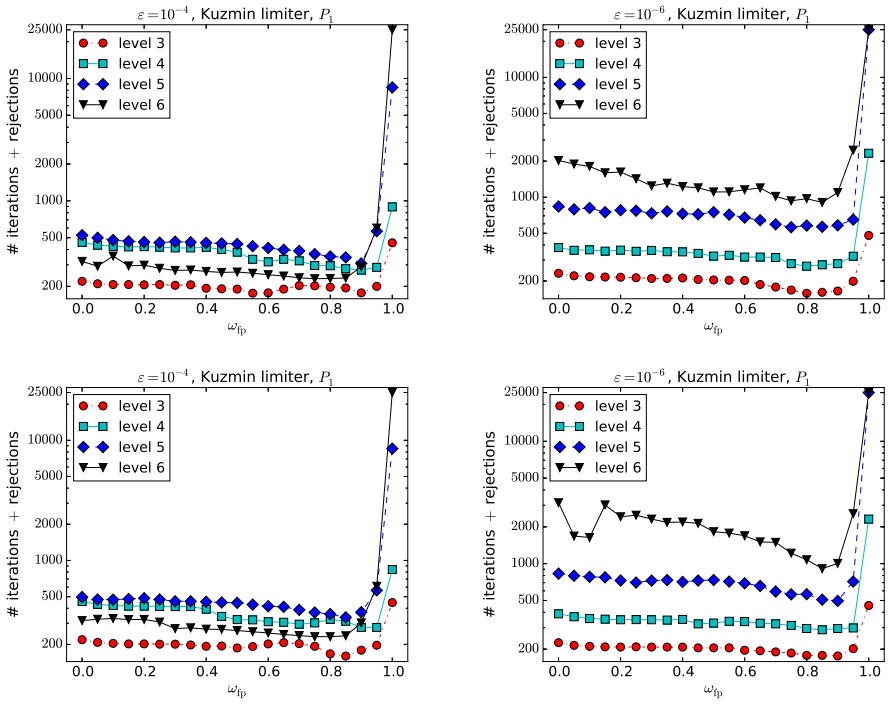

Figure 3: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), top: without projection to admissible values, bottom: with projection to admissible values.

behavior of this method in both cases. A good value for the mixing parameter is $\omega_{\mathrm{fp}} = 0.85$.

We already like to note here that the impact of the projection on the behavior of the iterative scheme was not always negligible. Usually, we performed simulations with and without projection. In cases where the impact of the projection is negligible, only the results with projection are presented for this example.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* For the best mixing parameter $\omega_{\mathrm{fp}} = 0.85$, the impact of using Anderson acceleration with different numbers of Anderson vectors is presented in Figure 4. For the moderately convection-dominated case, the use of 20 or 50 Anderson vectors reduces the needed number of iterations on all levels. However, each iteration requires the solution of an eigenvalue problem whose dimension equals the number of Anderson vectors. For $\varepsilon = 10^{-6}$, a reduction of the number of iterations can be seen only on coarse levels if sufficiently many Anderson vectors are used.

Studies for formal Newton methods. Representative results for several types of formal Newton methods are displayed in Figure 5. It can be seen that the approach with fixed damping parameters reduces the number of iterations+rejections considerably on coarse grids, but it fails to converge on fine grids. The *formal Newton* with adaptive parameter ω_{Newt} and separate treatment of the non-smooth points needed somewhat fewer iterations+rejections than *mixed fixed point*(0.85). Using instead the regularized *formal Newton* method, requires somewhat more iterations+rejections. We could observe that the behavior of the *formal Newton* methods is quite sensitive to the choice of ω_{Newt}. For instance, using $\omega_{\mathrm{Newt}} = 0.1$ increases the number of iterations+rejections such that it is on the two finest grids higher than for *mixed fixed point*(0.85). For the sake of brevity, we do not like to present a detailed study of this topic here. Altogether, one has to conclude that the application of the *formal Newton* methods does not significantly reduce the number

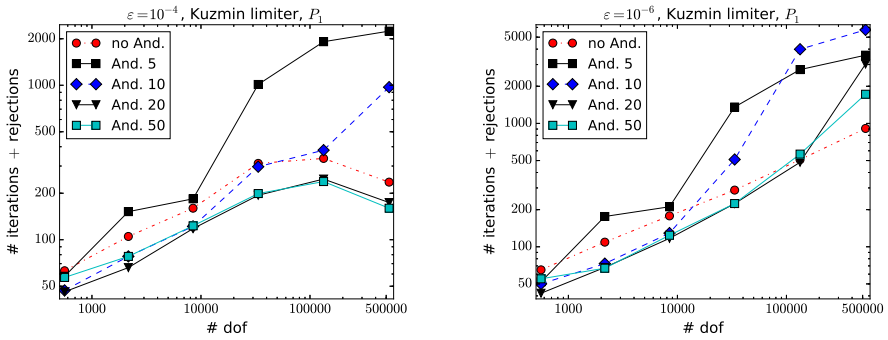

Figure 4: 2d Hemker problem. Results for *mixed fixed point with Anderson acceleration*$(0.85, \kappa)$, where κ is the number in the legends, with projection to admissible values.

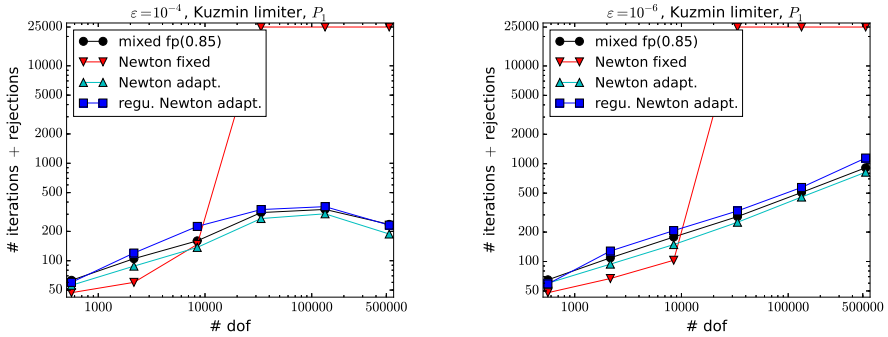

Figure 5: 2d Hemker problem. Results for the formal Newton methods, with projection to admissible values. The adaptive methods were used with $\omega_{\text{fp}} = 0.85$ and $\omega_{\text{Newt}} = 0.0625$.

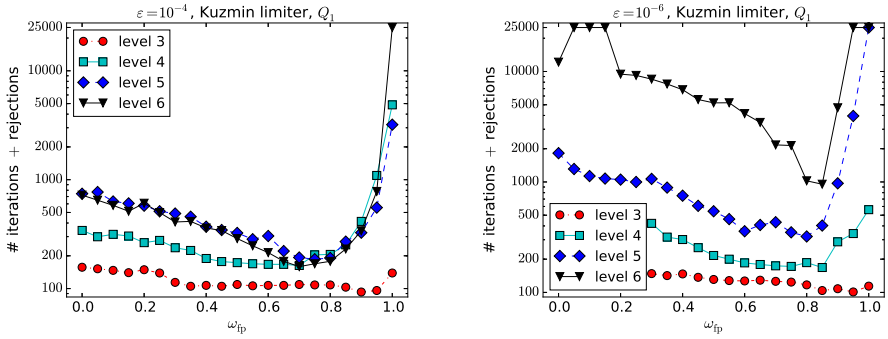

Figure 6: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values.

of iterations+rejections.

5.1.2. Kuzmin Limiter with Q_1 Finite Elements

The observations in this case are similar as for the Kuzmin limiter with P_1 finite elements. Some representative results are shown in Figures 6 and 7, which should be compared with Figures 3 and 5, respectively.

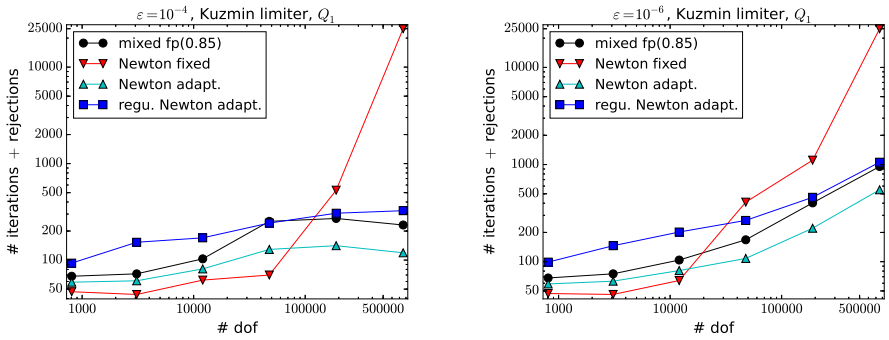

Figure 7: 2d Hemker problem. Results for the formal Newton methods, with projection to admissible values. The adaptive methods were used with $\omega_{\text{fp}} = 0.85$ and $\omega_{\text{Newt}} = 0.0625$.

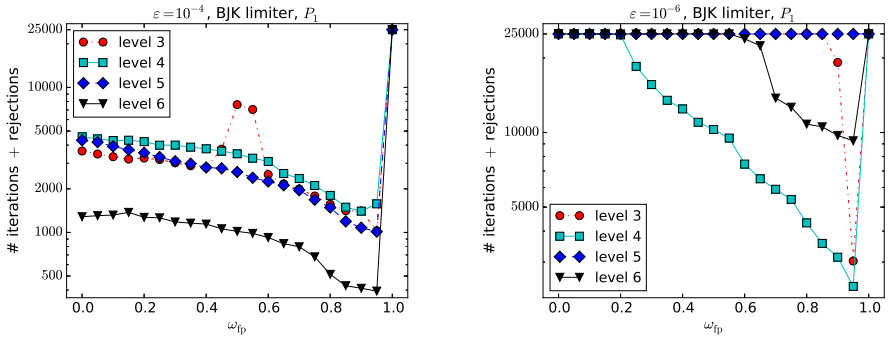

Figure 8: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values.

5.1.3. BJK Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). The results for this method are presented in Figure 8. In the moderately convection-dominated regime, it can be observed that choosing $\omega_{\text{fp}} = 0.95$ leads always to a comparatively small number of iterations, whereas the method does not converge for $\omega_{\text{fp}} = 1$. To achieve convergence in the strongly convection-dominated case is much harder. In fact, on level 5, *mixed fixed point*(ω_{fp}) does not converge for all used parameters. In case of convergence, an appropriate parameter is again $\omega_{\text{fp}} = 0.95$.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* The application of the Anderson acceleration worsens the convergence for all simulations with the BJK limiter, compare Figure 9.

Studies for formal Newton methods. Results obtained for *formal Newton* methods are presented in Figures 10 and 11. For $\varepsilon = 10^{-4}$, it can be seen that *formal Newton* with an adaptive choice of the damping parameter ω_{Newt} needs fewer iterations on all levels than *mixed fixed point*(0.95) if the projection to admissible values is not used. With this projection, the method does not converge on fine grids. The method *formal Newton* with fixed parameters converges quite well, apart on the finest level. For the mildly convection-dominated case, we observed that also a *formal Newton* method with $\omega_{\text{fp}} = 1$, $\omega_{\text{Newt}} = 1$, starting from the first iteration (Newton wo damp. in Figure 11) works quite well, at least on the coarse grids. In the strongly convection-dominated regime, some *formal Newton* methods needed fewer iterations than *mixed fixed point*(0.95) on coarse grids. Again, some methods behaved rather differently with and without projection to admissible values.

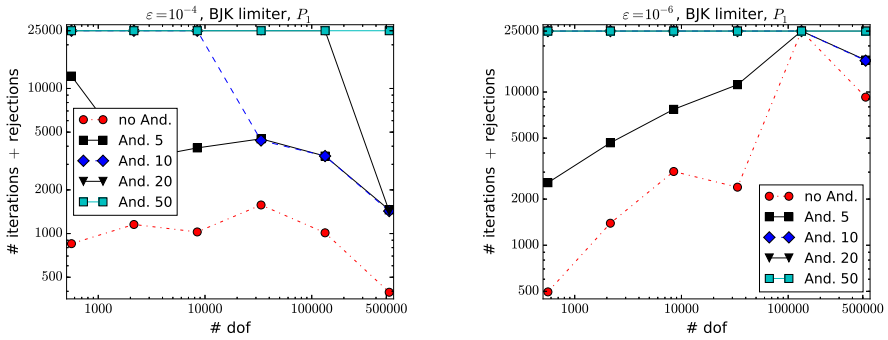

Figure 9: 2d Hemker problem. Results for *mixed fixed point with Anderson acceleration* $(0.95, \kappa)$, where κ is the number in the legends, with projection to admissible values.

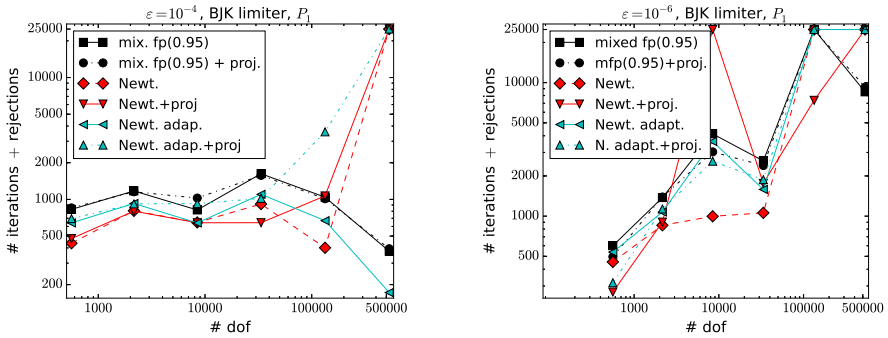

Figure 10: 2d Hemker problem. Results for the formal Newton methods, with and without projection to admissible values. The adaptive methods were used with $\omega_{\text{fp}} = 0.95$ and $\omega_{\text{Newt}} = 0.0625$.

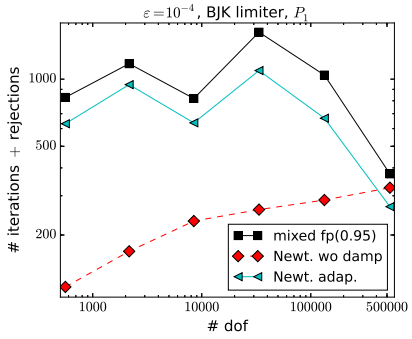

Figure 11: 2d Hemker problem. Results for the formal Newton methods, without projection to admissible values.

5.1.4. Efficiency

As final part of the 2d example, a study with respect to the efficiency, in terms of computing times, of the methods is presented. To this end, approaches for each type of method with a small number of iterations+rejections are taken and compared. The arising linear systems of equations were solved with the sparse direct solver UMFPACK [7]. All simulations were performed five times, then the fastest and slowest times were neglected and the average of the remaining three times is shown in Figure 12.

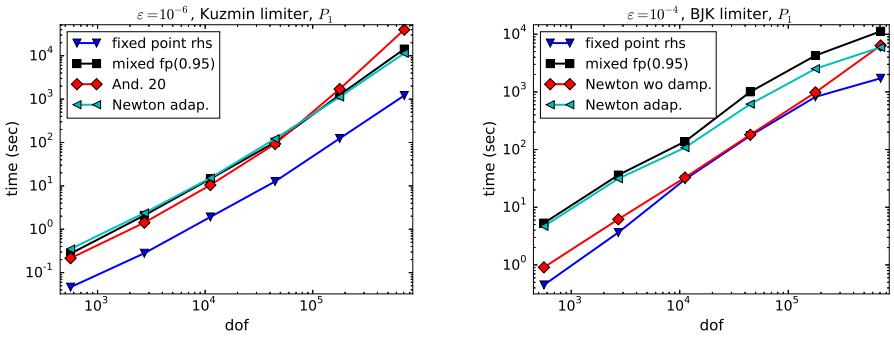

Figure 12: 2d Hemker problem. Efficiency of several methods.

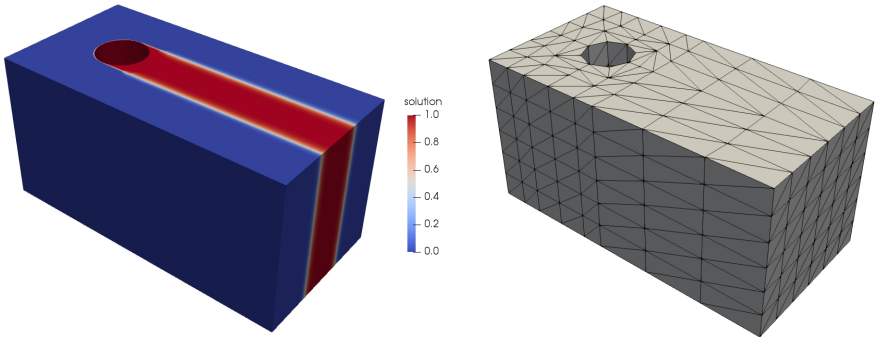

Figure 13: 3d Hemker problem. Solution for $\varepsilon = 10^{-6}$, computed with the Kuzmin limiter, P_1, level 4, and sketch of the coarsest grid (level 0).

Figure 12 shows some representative results. For both limiters, *fixed point rhs* (= *mixed fixed point*(0)) is the most efficient method. The advantage of needing just one matrix factorization for the whole iteration results in a gain of one order of magnitude concerning the simulation times compared with most of the other methods. Only Newton's method without damping for the BJK limiter is similarly efficient on coarse grids. Note that this method needs much fewer iteration steps than *fixed point rhs* for solving the nonlinear problem, e.g., on the grid with around 33000 degrees of freedom 260 iterations vs. 4199 iterations.

5.2. The 3d Hemker Problem

The 3d Hemker problem is a natural extension of the 2d Hemker problem, which was proposed in [29]. The domain is defined by

$$\Omega = \big\{\{(-3,9)\times(-3,3)\} \setminus \big\{(x,y) \;:\; x^2+y^2 \le 1\big\}\big\} \times (0,6)$$

and the convection vector in (1) is given by $\boldsymbol{b} = (1,0,0)^T$. Homogeneous Dirichlet boundary conditions $u^b = 0$ are prescribed at the inlet plane $x = -3$ and at the cylinder, the Dirichlet boundary condition is $u^b = 1$. At all other boundaries, homogeneous Neumann conditions are imposed. An illustration of the solution is provided in Figure 13.

Simulations were performed for P_1 and Q_1 (only Kuzmin limiter) finite elements, see Figure 13 for the coarsest tetrahedral grid and Table 2 for information on the number of degrees of freedom. It turned out that the solutions computed with the Kuzmin limiter on the tetrahedral grids showed small negative values. For example, on level 1, these values are $-2\cdot 10^{-6}$ ($\varepsilon = 10^{-4}$) and $-8\cdot 10^{-9}$ ($\varepsilon = 10^{-6}$) and on

Table 2: 3d Hemker problem. Number of degrees of freedom, including Dirichlet nodes.

level	P_1
0	490
1	3172
2	22600
3	170128
4	1319200

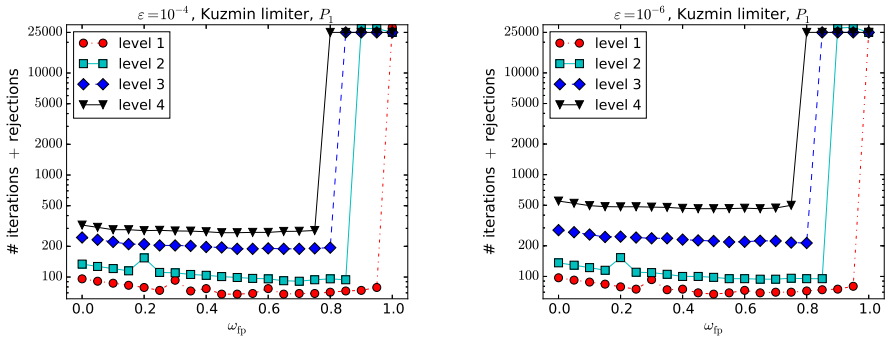

Figure 14: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$; $\varepsilon = 10^{-6}$: level 1 with $\omega_{\mathrm{fp}} = 1$, level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

level 3 they are $-7 \cdot 10^{-6}$ ($\varepsilon = 10^{-4}$) and $-8 \cdot 10^{-8}$ ($\varepsilon = 10^{-6}$). Although negative oscillations of this size might be still tolerable in applications, they do not allow to use the projection of the iterates to the admissible interval $[0, 1]$ since the Euclidean norm of the residual vector stalled at some value larger than the stopping tolerance. The values of the results obtained with the Kuzmin limiter on the hexahedral grids and the BJK limiter on the tetrahedral grids were always in $[0, 1]$. In these cases, both approaches, with and without projection to admissible values, led usually to a similar number of iterations. Since in the approach without projection to admissible values, the results found for Q_1 finite elements are also in this example qualitatively the same as for P_1 finite elements, only the investigations for P_1 finite elements are presented below, for the sake of brevity.

5.2.1. Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). The results of these studies are displayed in Figure 14. It can be seen that *mixed fixed point*(ω_{fp}) converged only for sufficiently small mixing parameters ω_{fp}. An appropriate mixing parameter for both regimes is $\omega_{\mathrm{fp}} = 0.7$.

If not mentioned otherwise, an iterative solver was used for the arising linear systems of equations in three dimensions and an inexact solve of these systems was performed, see Section 5.3.3 for details. Usually, we could not observe a qualitative difference with respect to the number of iterations+rejections concerning an accurate and an inexact solution of the linear systems. An example is given in Figure 15. One can see by comparing with Figure 14 that the number of iterations is in all situations almost the same.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* The impact of using Anderson acceleration is demonstrated in Figure 16. For both convection-dominated regimes, the application of the Anderson acceleration reduces the needed

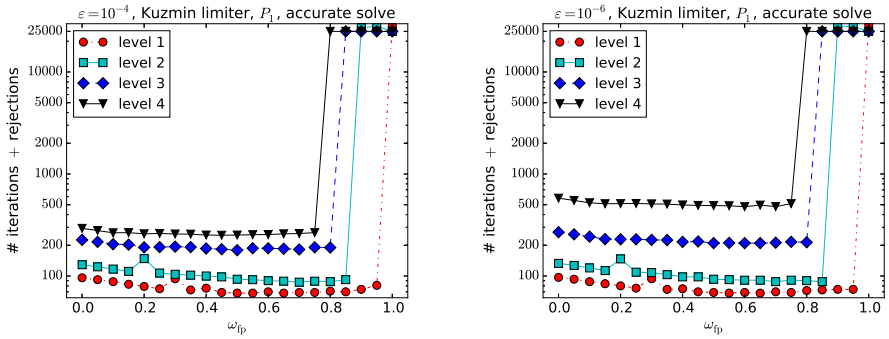

Figure 15: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values, with accurate solution of the linear problems. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$; $\varepsilon = 10^{-6}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

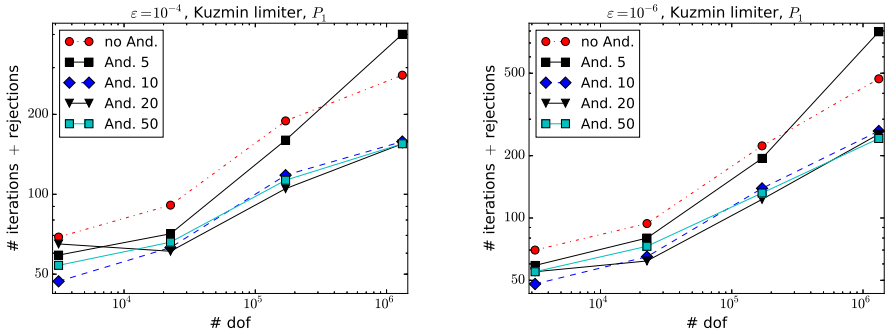

Figure 16: 3d Hemker problem. Results for *mixed fixed point with Anderson acceleration*($0.7, \kappa$), where κ is the number in the legends, without projection to admissible values.

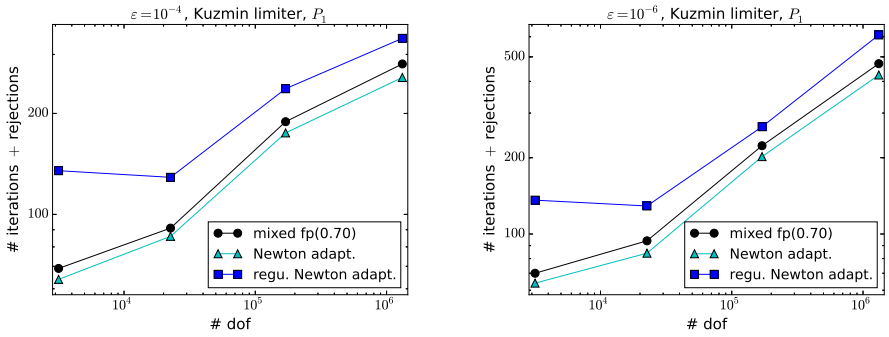

Figure 17: 3d Hemker problem. Results for the formal Newton methods, without projection to admissible values. The adaptive methods were used with $\omega_{\mathrm{fp}} = 0.7$ and $\omega_{\mathrm{Newt}} = 0.1$.

number of iterations+rejections on all levels if the number of Anderson vectors is chosen to be $\kappa \in \{10, 20, 50\}$. For these values, only little differences are observable.

Studies for formal Newton methods. Results for the *formal Newton* methods, in comparison with *mixed fixed point*(0.7), are presented in Figure 17. As can be seen, the *formal Newton* method without regularization sometimes reduces the number of iterations+rejections slightly, but generally do not lead to a notable improvement.

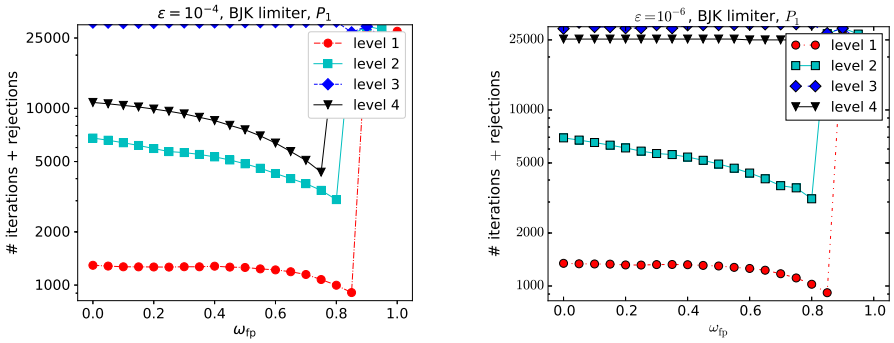

Figure 18: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\text{fp}} = 1$, level 3 with $\omega_{\text{fp}} = 1$, level 4 with $\omega_{\text{fp}} \in \{0.95, 1\}$; $\varepsilon = 10^{-6}$: level 1 with $\omega_{\text{fp}} = 1$, level 2 with $\omega_{\text{fp}} = 1$, level 3 with $\omega_{\text{fp}} = 1$, level 4 with $\omega_{\text{fp}} \in \{0.95, 1\}$.

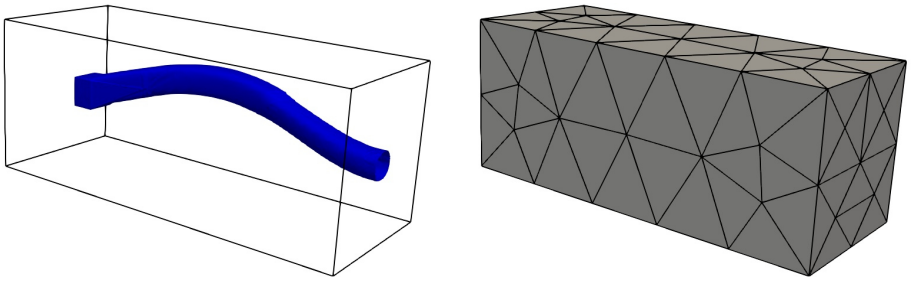

Figure 19: 3d problem with non-constant convection. Solution for $\varepsilon = 10^{-6}$, isosurface for $u = 0.05$, computed with the Kuzmin limiter, P_1, level 5, and sketch of the coarsest grid (level 0).

5.2.2. BJK Limiter with P_1 Finite Elements

Utilizing the method *mixed fixed point*(ω_{fp}) for the BJK limiter, one finds that also in this case the method converges only if the mixing parameter is sufficiently small, compare Figure 18. However, there are situations where the maximal number of 25000 iteration steps is not sufficient for the convergence of *mixed fixed point*(ω_{fp}) with any of the considered parameters: level 3 for both regimes and the finest grid for the strongly convection-dominated regime.

Without presenting detailed results, we like to note that, similar as for the 2d Hemker problem, the application of Anderson acceleration does not benefit for *mixed fixed point*(ω_{fp}) and the BJK limiter. The *formal Newton* method for this limiter will be discussed briefly in the next example.

5.3. A 3d Problem with Non-Constant Convection

This example was proposed in [5]. The domain is given by $\Omega = \Omega_1 \setminus \overline{\Omega}_2$ with $\Omega_1 = (0,5) \times (0,2) \times (0,2)$ and $\Omega_2 = (0.5, 0.8) \times (0.8, 1.2) \times (0.8, 1.2)$ and the convection field by $\boldsymbol{b} = (1, l(x), l(x))^T$ with $l(x) = (0.19x^3 - 1.42x^2 + 2.38x)/4$. At the interior cube, the Dirichlet boundary condition $u^b = 0$ is imposed, at the outlet $x = 5$ homogeneous Neumann boundary conditions are set, and at all other boundaries $u^b = 1$ is prescribed. An illustration of the solution is given in Figure 19. All simulations were performed for P_1 finite elements on unstructured tetrahedral grids, whose coarsest grid was obtained with the mesh generator Gmsh [9], see Figure 19. Information concerning the degrees of freedom are provided in Table 3.

On the used grids, the BJK limiter computed solutions with values in $[0, 1]$ whereas the Kuzmin limiter showed small overshoots on levels 3, 4, and 5. In

Table 3: 3d problem with non-constant convection. Number of degrees of freedom, including Dirichlet nodes.

level	P_1
0	86
1	476
2	3078
3	21898
4	164626
5	1275426

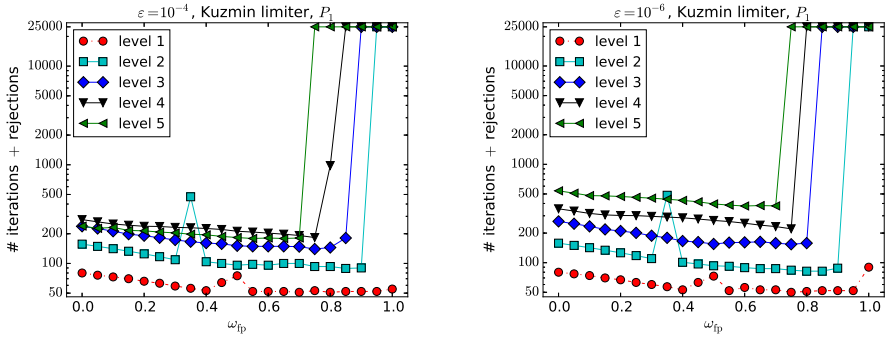

Figure 20: 3d problem with non-constant convection. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 4 with $\omega_{\mathrm{fp}} = 1$, level 5 with $\omega_{\mathrm{fp}} = 1$; $\varepsilon = 10^{-6}$: level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} = 1$, level 5 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

all situations where the numerical solution had values in $[0, 1]$, it turned out that the simulations without and with projecting to admissible values as described in Section 4.3 behaved generally similarly. For the sake of brevity, only results without projection are presented below.

5.3.1. Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). The results of these studies are displayed in Figure 20. As for the 3d Hemker example, it can be seen that *mixed fixed point*(ω_{fp}) converges if ω_{fp} is sufficiently small. The finer the grid, the smaller is the interval for which the method converges. An appropriate parameter for both regimes and for all levels is $\omega_{\mathrm{fp}} = 0.6$.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* Figure 21 shows the effect of using Anderson acceleration. For sufficiently many Anderson vectors, $\kappa \in \{10, 20, 50\}$, there is generally a notable reduction of the number of iterations+rejections compared with *mixed fixed point*(0.6).

Studies for formal Newton methods. The results for this approach, displayed in Figure 22, are similar as for the 3d Hemker problem, Figure 17. Also here, the *formal Newton* methods usually do not show a notably better behavior than the *mixed fixed point*(0.6) method.

5.3.2. BJK Limiter with P_1 Finite Elements

For the BJK limiter, results for the method *mixed fixed point*(ω_{fp}) are presented in Figure 23. On the one hand, there is a similar behavior as for the Kuzmin limiter, because the method converges if the mixing parameter ω_{fp} is sufficiently small. On the other hand, much more iterations are needed than for the Kuzmin limiter.

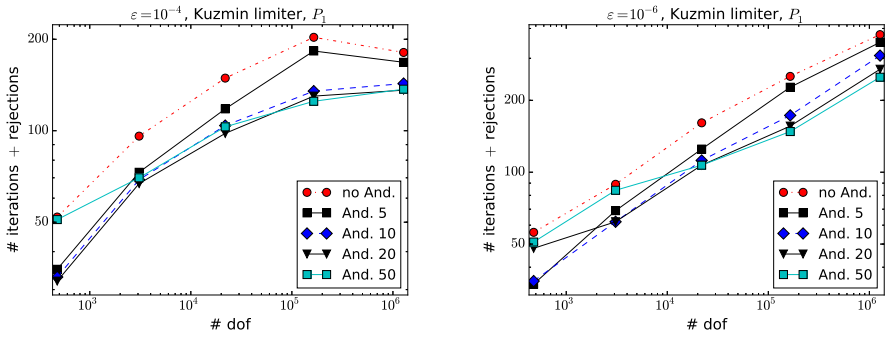

Figure 21: 3d problem with non-constant convection. Results for *mixed fixed point with Anderson acceleration*$(0.6, \kappa)$, where κ is the number in the legends, without projection to admissible values.

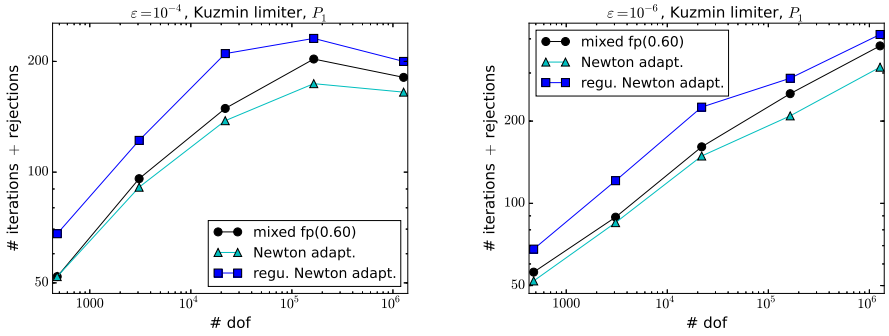

Figure 22: 3d problem with non-constant convection. Results for the formal Newton methods, without projection to admissible values. The adaptive methods were used with $\omega_{\text{fp}} = 0.6$ and $\omega_{\text{Newt}} = 0.1$.

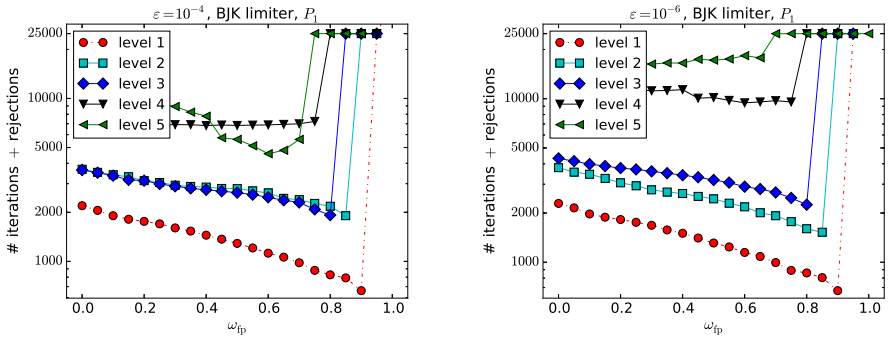

Figure 23: 3d problem with non-constant convection. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-6}$: level 1 with $\omega_{\text{fp}} = 1$; $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-6}$: level 2 with $\omega_{\text{fp}} = 1$, level 3 with $\omega_{\text{fp}} = 1$, level 4 with $\omega_{\text{fp}} = 1$, level 5 with $\omega_{\text{fp}} \in \{0.95, 1\}$.

For this example, the behavior of the *formal Newton* method without damping, which behaved quite well for the 2d Hemker problem, is discussed. First of all, we noticed that the used iterative solver did not work for this method, such that a sparse direct solver was utilized. With this solver, it was only possible to perform simulations on coarse grids. Concerning the number of iterations+rejections, the results are again quite good, e.g., in the strongly convection-dominated case, these

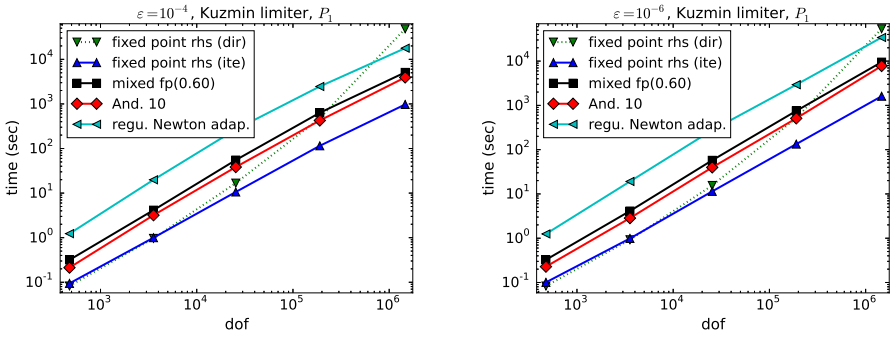

Figure 24: 3d problem with non-constant convection. Efficiency for several methods.

numbers are for levels 1–3: 171, 401, 598 in comparison with the best numbers from Figure 23: 706, 1574, 2298. Thus, on levels 2 and 3 there is a considerable reduction of these numbers.

5.3.3. Efficiency

Again, we selected a method from each approach with a small number of iterations+rejections for comparison. Usually, the arising linear systems of equations were solved with an iterative solver. To this end, GMRES [24] was used with right preconditioner. The preconditioner was SSOR with relaxation parameter 1.0. In our experience, it is generally not necessary to solve the linear systems of equations very accurately. Accordingly, the GMRES iteration was stopped if the Euclidean norm of the residual vector was reduced by the factor 100 or after 50 iterations. A comparison with the use of a much more stronger stopping criterion has been already provided in Section 5.2.1. For *fixed point rhs*, also the sparse direct solver UMFPACK was utilized for solving the linear system of equations, because for this method, only one factorization is necessary. The determination of the computing times was performed in the same way as described for the 2d Hemker problem in Section 5.1.4.

Results are displayed in Figure 24. Like in the 2d case, *fixed point rhs* (= *mixed fixed point*(0)) is the most efficient approach. On coarse grids, both the iterative or the direct solver can be used, but on finer grids, one has to apply the iterative solver. Compared with *mixed fixed point*(0.6) and *mixed fixed point with Anderson acceleration*(0.6, 10), the computing times of *fixed point rhs* are about half an order of magnitude smaller, even if the number of iterations+rejections is usually notably larger, e.g., for the strongly convection-dominated case on the finest grid 538 vs. 387 for *mixed fixed point*(0.6) and 308 for *mixed fixed point with Anderson acceleration*(0.6, 10). The reason is that the used iterative solver performed for the matrix from *fixed point rhs*, which is just $\hat{A} = A + D$, much more efficient than for the matrices from the other methods.

6. Summary

This paper presented comprehensive numerical studies for solving the nonlinear problems arising in AFC discretizations of steady-state convection-diffusion equations. Compared with the initial study [13], more approaches were considered and more challenging examples were studied.

Taking the simplest fixed point method *fixed point rhs*, or equivalently *mixed fixed point*(0), as reference method, then the numerical studies showed that it is

sometimes possible to reduce with advanced methods the number of iterations+rejections considerably, e.g., see the numbers given in Sections 5.1.4 and 5.3.2. The method *fixed point rhs* has, however, the structural advantage of having the same matrix in each iteration step. In two dimensions, due to the high efficiency of sparse direct solvers in 2d, it clearly outperforms all other approaches with respect to computing times, of course only in the case that *fixed point rhs* converges. A sparse direct solver can be applied in 3d only on very coarse grids. Usually, an iterative solver has to be utilized. However, also in 3d, the method *fixed point rhs* was most efficient, since the iterative solver worked much better than for other methods because of the favorable properties of the iteration matrix.

It was usually much easier to solve the problems for the Kuzmin limiter than for the BJK limiter. Especially in the strongly convection-dominated regime and on fine grids, the considered methods often did not converge for the BJK limiter within the prescribed maximal number of steps.

Whether or not the projection to admissible values as described in Section 4.3 should be performed depends on the example. If the numerical solution does not possess undershoots or overshoots, often only a minor impact on the behavior of the solver *mixed fixed point*(ω_{fp}) for the nonlinear problem could be observed.

In summary, the simplest fixed point iteration is the most efficient approach in terms of computing times, although it often needs considerably more iterations than other approaches. The gain of either needing only one matrix factorization in 2d or of the high efficiency of the iterative solver in 3d compensates this drawback more than enough.

Acknowledgment. The work of A. Jha has been supported by the Berlin Mathematical School (BMS). We would like to acknowledge also two unknown referees whose suggestions greatly helped to improve this paper.

References

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

[2] Santiago Badia and Jesús Bonilla. Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Comput. Methods Appl. Mech. Engrg.*, 313:133–158, 2017.

[3] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.*, 54(4):2427–2451, 2016.

[4] Gabriel R. Barrenechea, Volker John, and Petr Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.*, 27(3):525–548, 2017.

[5] Gabriel R. Barrenechea, Volker John, Petr Knobloch, and Richard Rankin. A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *SeMA J.*, 75(4):655–685, 2018.

[6] Alexander N. Brooks and Thomas J. R. Hughes. Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. *Comput. Methods Appl. Mech. Engrg.*, 32(1-3):199–259, 1982. FENOMECH '81, Part I (Stuttgart, 1981).

[7] Timothy A. Davis. Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Software*, 30(2):196–199, 2004.

[8] S. Ganesan, V. John, G. Matthies, R. Meesala, S. Abdus, and U. Wilbrandt. An object oriented parallel finite element scheme for computing pdes: Design and implementation. In *IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW) Hyderabad*, pages 106–115. IEEE, 2016.

[9] Christophe Geuzaine and Jean-François Remacle. Gmsh: A 3-D finite element mesh generator with built-in pre- and post-processing facilities. *Internat. J. Numer. Methods Engrg.*, 79(11):1309–1331, 2009.

[10] P. W. Hemker. A singularly perturbed model problem for numerical computation. *J. Comput. Appl. Math.*, 76(1-2):277–285, 1996.

[11] T. J. R. Hughes and A. Brooks. A multidimensional upwind scheme with no crosswind diffusion. In *Finite element methods for convection dominated flows (Papers, Winter Ann. Meeting Amer. Soc. Mech. Engrs., New York, 1979)*, volume 34 of *AMD*, pages 19–35. Amer. Soc. Mech. Engrs. (ASME), New York, 1979.

[12] Thomas J. R. Hughes, Michel Mallet, and Akira Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, 54(3):341–355, 1986.

[13] Abhinav Jha and Volker John. On basic iteration schemes for nonlinear afc discretizations. WIAS Preprint 2533, Weierstrass Institute for Applied Analysis and Stochastics, 2018. to appear in Proceedings of BAIL 2018.

[14] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations. I. A review. *Comput. Methods Appl. Mech. Engrg.*, 196(17-20):2197–2215, 2007.

[15] Volker John and Petr Knobloch. On spurious oscillations at layers diminishing (SOLD) methods for convection-diffusion equations. II. Analysis for P_1 and Q_1 finite elements. *Comput. Methods Appl. Mech. Engrg.*, 197(21-24):1997–2014, 2008.

[16] Volker John, Petr Knobloch, and Julia Novo. Finite elements for scalar convection-dominated equations and incompressible flow problems: a never ending story? *Comput. Vis. Sci.*, 19(5-6):47–63, 2018.

[17] Volker John, Teodora Mitkova, Michael Roland, Kai Sundmacher, Lutz Tobiska, and Andreas Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. *Chemical Engineering Science*, 64(4):733 – 741, 2009. 3rd International Conference on Population Balance Modelling.

[18] J. Karátson and S. Korotov. Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions. *Numer. Math.*, 99(4):669–698, 2005.

[19] Petr Knobloch. Improvements of the Mizukami-Hughes method for convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 196(1-3):579–594, 2006.

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

[21] Dmitri Kuzmin and Matthias Möller. Algebraic flux correction. I. Scalar conservation laws. In *Flux-corrected transport*, Sci. Comput., pages 155–206. Springer, Berlin, 2005.

[22] Akira Mizukami and Thomas J. R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Comput. Methods Appl. Mech. Engrg.*, 50(2):181–193, 1985.

[23] Hans-Görg Roos, Martin Stynes, and Lutz Tobiska. *Robust numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2008. Convection-diffusion-reaction and flow problems.

[24] Youcef Saad and Martin H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[25] Masahisa Tabata. A finite element approximation corresponding to the upwind finite differencing. *Mem. Numer. Math.*, 1(4):47–63, 1977.

[26] Richard S. Varga. *Matrix iterative analysis*, volume 27 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, expanded edition, 2000.

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

[28] Pieter Wesseling. *Principles of computational fluid dynamics*, volume 29 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 2001.

[29] Ulrich Wilbrandt, Clemens Bartsch, Naveed Ahmed, Najib Alia, Felix Anker, Laura Blank, Alfonso Caiazzo, Sashikumaar Ganesan, Swetlana Giere, Gunar Matthies, Raviteja Meesala, Abdus Shamim, Jagannath Venkatesan, and Volker John. ParMooN—A modernized program package based on mapped finite elements. *Comput. Math. Appl.*, 74(1):74–88, 2017.