Freie Universität Berlin

Numerical Algorithms for Algebraic Stabilizations of Scalar Convection-Dominated Problems

Dissertation zur Erlangung des Grades eines Doktors der
Naturwissenschaften (Dr. rer. nat.) am Fachbereich Mathematik und
Informatik der Freie Universität Berlin

Abhinav Jha

Berlin, October 30, 2020

1. Gutachter: Prof. Dr. Volker John,
Freie Universität Berlin and
Weierstraß-Institut für Angewandte Analysis und Stochastik

2. Gutachter: Dr. Gabriel Barrenechea,
University of Strathclyde, Glasgow

Date of Disputation: 16th October 2020

For Maa and Papa

Acknowledgments

> *“Be the change you want to see.”*
>
> Mahatma Gandhi

I would like to express my deepest gratitude to my supervisor, Prof. Dr. Volker John, who allowed me to be a part of his research group at Freie Univerität and Weierstrass Institute (WIAS). His patience, encouragement, constructive feedback, useful suggestions, and guidance have helped me be a better mathematician and above that, a better human.

I would like to thank the Berlin Mathematical School (BMS), who funded my studies during my stay in Berlin. I would especially like to thank the BMS office, which has not only helped me with my study-related queries but my general queries about Germany as well. I would also like to thank Weierstrass Institute for hosting me during my studies and provide me with all the facilities that I needed.

In addition, I would want to deeply appreciate my colleagues in the research group *Numerical Mathematics and Scientific Computing* at WIAS, who have always helped me. I would especially like to mention Dr. Ulrich Wilbrandt, who helped me at all the stages of my studies and my personal life. I have learnt so much more under his guidance. My sincerest regards to Dr. Clemens Bartsch who helped me tremendously to get familiar with the AFC topic and ParMooN. I have been lucky to share my office with many colleagues. I would like to thank, Felix, Mine, Baptiste, and Derk for all the little help and banter they provided in the office.

My journey of becoming the person I am today has been because of the great teachers throughout my life. I would like to especially thank Dr. Ankik Giri, Dr. A. Swaminathan, and Dr. Prashanto Chatterjee who guided me during my Bachelor and Masters studies. My heartfelt gratitude is for Dr. Naveed Ahmed, who played many roles from a teacher, a guide, a big brother, and a friend during my studies. He has always encouraged me to push myself to new limits,.

Life in Germany has not just been about all work. I also got to make new friends with whom I have made amazing memories. I am grateful to my friends at BMS and back in India who have kept me sane and talked to me whenever I needed someone. This list is to mention a few of them, Sanju, Pulkit, Vrinda, Divyansh, Shah, Duca, Marco, Levent, Ben, Hari, Kira,

Riccardo, Candan, Gari, Oghuzan, Laurie, Oli, and Luke for all the crazy memories we have made together. I would like to thank Shah Faisal, for proof-reading part of my thesis and helping me to construct it in a better way.

Last but not the least, the people without him this journey of my life would not have been possible, my parents. They have always supported my decisions. My sister who at the right time, introduced me to C++ and persuaded me to pursue career in computers. Juju, without you I would have never been able to love mathematics and computer science. And before I finish, I would like to thank my God, under whose faith, I have never lost sight of my aim. As he says, faith and patience get you everything.

Contents

1 Introduction **1**
- 1.1 Motivation . . . 2
- 1.2 Outline . . . 4
- 1.3 Function Spaces . . . 6

2 Convection-Diffusion-Reaction-Equations **9**
- 2.1 Derivation . . . 9
- 2.2 Maximum Principles . . . 12
- 2.3 Weak Solution Theory . . . 16
- 2.4 Summary . . . 20

3 Stabilized Finite Element Methods **23**
- 3.1 Discrete Maximum Principle . . . 24
- 3.2 Galerkin Approximation . . . 27
- 3.3 Streamline-Upwind Petrov-Galerkin Method (SUPG) . . . 30
  - 3.3.1 Spurious Oscillations at Layers Diminishing Methods (SOLD) . . . 34
- 3.4 Algebraic Flux Correction Schemes . . . 35
  - 3.4.1 Derivation . . . 35
  - 3.4.2 Limiters . . . 37
  - 3.4.3 Review of Analytical Results . . . 40
- 3.5 Summary . . . 44

4 Iterative Solvers for Steady-State Convection-Diffusion-Reaction Equation **47**
- 4.1 Iterative Schemes . . . 47
  - 4.1.1 The Mixed Fixed Point Iteration . . . 48
  - 4.1.2 A Formal Newton Method . . . 50
- 4.2 Algorithmic Components . . . 59
  - 4.2.1 Adaptive Choice of Damping Parameter . . . 59
  - 4.2.2 Anderson Acceleration . . . 60
  - 4.2.3 Projection to Admissible Values . . . 60
  - 4.2.4 Choosing the Initial Condition . . . 61
- 4.3 Numerical Studies . . . 61
  - 4.3.1 Examples in Two Dimensions . . . 63
  - 4.3.2 Examples in Three Dimensions . . . 77

4.4 Summary 86

5 A Posteriori Error Estimation for AFC Schemes 87

5.1 Preliminaries 88
  5.1.1 Definitions and Notations 89
  5.1.2 Auxiliary Results 95
5.2 A Posteriori Error Estimators 100
  5.2.1 Residual-Based Estimator 101
  5.2.2 AFC-SUPG Estimator 113
5.3 Numerical Studies 114
  5.3.1 A Known 2d Solution with a Boundary Layer 116
  5.3.2 Example with Interior and Boundary Layers 119
5.4 Summary 121

6 Hanging Nodes in Context of AFC Schemes 123

6.1 Hanging Nodes in Theory of Lagrange Finite Elements 124
6.2 Hanging Nodes in Theory of AFC Schemes 127
  6.2.1 Implementation of hanging nodes 127
  6.2.2 Kuzmin Limiter 129
  6.2.3 BJK Limiter 130
  6.2.4 Limiter Definition 132
6.3 Numerical Studies 132
6.4 Summary 137

7 Summary and Outlook 139

7.1 Summary 139
7.2 Outlook 141

A Algorithms 143

A.1 Algorithm for Dynamic Damping 143
A.2 Algorithm for Anderson Acceleration 145
A.3 Algorithm for Newton Dynamic Damping 146

B Numerical Values for a Posteriori Error Estimators 147

List of Notations 153

Bibliography 159

1 Introduction

> *"Mathematics is the queen of sciences."*
>
> Carl Friedrich Gauß

One fundamental equation of fluid dynamics is the Convection-Diffusion-Reaction equation given by

$$-\varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u + cu = f,$$

in a bounded domain $\Omega \subset \mathbb{R}^d$. Such equations appear in a lot of physical processes such as the Oseen equations [Ose11], water pollution problems [REI$^+$07], simulation of oil extraction from underground reservoirs [Ewi83], and convective heat transport [JK49]. These equations can also be found in semiconductor applications where the *continuity equation* for electrons in steady-state scaled models of a semiconductor is a convection-diffusion equations (see [PCSM87]). Their application can also be seen in biology as they model chemotaxis (movement of an organism in response to chemical stimulus) observed in bacteria [KS71], population migration [AS02], and the study of VEGFC (Vascular endothelial growth factor C) patterning in the context of lymphangiogenesis [WR17].

These equations are referred to as *singular perturbation problems* as they depend on a small positive parameter ε and whose solution (or derivatives) approaches a discontinuous limit as ε tends to zero. In this work, we are interested in the case when the convection ($\boldsymbol{b}$) dominates diffusion (ε), i.e., $\|\boldsymbol{b}\|_{L^\infty(\Omega)} \gg \varepsilon$, the reason being this case gives rise to layers in the interior and boundary. The terminology *boundary layer* was first introduced by Ludwig Prandtl at the 3rd International Congress of Mathematicians in Heidelberg, 1904. Layers can be defined as narrow regions where the solution has a steep gradient.

The numerical solution of such problems also depends on the singular perturbation parameter ε. If standard numerical approximations such as central finite difference method (FDM) or the Galerkin finite element method (FEM) are used, then for a critical value of ε, i.e., $\varepsilon \ll 1$, the numerical solution cannot be used in practice. The reason for such behavior is that the layers are so narrow that they cannot be properly resolved on the grid. Hence, one would prefer techniques that are robust with respect to ε. Since the analytic solution depends on ε, also the numerical solution might depend on this parameter. By robustness we mean that the numerical solution does not blow up if $\varepsilon \to 0$. . Along with robustness, we would also like the numerical solution to be physically consistent, i.e., it possesses the same property as that of the analytical solution. For Convection-Diffusion-Reaction equations this translates to the satisfaction of maximum principles or for numerical solution, the discrete analogue, the discrete maximum principles (DMP). The starting point of numerical algorithms satisfying the DMP was presented in the work of Peter Lax, in [Lax54]. The *Lax-Friedrichs* scheme is based on his work on the finite difference algorithm for one-dimensional fluid flow problem. It guarantees the boundedness of the solution in terms of initial values. Robustness of the method and satisfaction of DMP leads to the research area of stabilized discretizations of Convection-Diffusion-Reaction equations. In this work we will concentrate on stabilized finite element methods.

Many of the stabilized schemes follow the idea of upwind methods or Petrov-Galerkin methods (see [RST08]). The most known stabilized scheme in the context of FEM is the *Streamline Upwind Petrov-Galerkin* (SUPG) scheme introduced by Hughes and Brooks in [HB79, BH82], which is a linear scheme. The basic idea is to use different tests and trial space for the problem and introduce certain stabilization parameters known as the SUPG parameters. The SUPG scheme can be applied to different shapes of elements and different orders of polynomials. They compute the layers sharply but the solution possesses oscillations of a considerable order of magnitude near the layer.

Another approach is to use a nonlinear stabilized method instead of a linear method. One of the early works in this direction is the Mizukami-Hughes method introduced in [MH85] for linear triangular elements. It belongs to a very small class of stabilized methods which satisfy the DMP. Results relating to the existence, uniqueness, and convergence of the solution are not available. Also, because of the nonlinear nature of the problem it's difficult to solve the arising nonlinear system of equations. Hence, one asks, "What is an ideal stabilized technique?". In our opinion some of the important features of a stabilized FEM are:

1. Computation of accurate and sharp layers,
2. Satisfaction of discrete maximum principle,
3. Easy solvability of the arising system of equations.

1.1 Motivation

The origins of the algebraic flux correction schemes dates back to the work of Lax [Lax54]. According to the Godunov theorem [God59], linear bound preserving schemes of Lax-Freidrich kinds can be at most first order accurate. Consequently, more accurate constrained solutions can only be produced by nonlinear algorithms. To work around Godunov's order constraint

Figure 1.1: Ludiwg Prandtl (1875-1953) and Peter Lax (1926-).[1]

with the aim of achieving sharp and non-oscillatory resolution of shock waves, Boris and Book in [BB97] applied nonlinear conservative anti-diffusive corrections to a low order predictor. This is the main idea of the *flux corrected transport* (FCT) algorithm and many other nonlinear high resolution schemes. A fully multi-dimensional version of FCT was given in [Zal79] and was combined with finite element discretization given by [PC86]. A general framework for the design of bound-preserving finite element approximation was introduced by Kuzmin in [Kuz07] and was named *algebraic flux correction schemes* (AFC). AFC schemes are an approach for the stabilization of the Convection-Diffusion-Reaction equations which work on an algebraic level rather than on a variational level.

The main ideology of satisfaction of DMP by AFC schemes is the use of the M-matrix property. In the past few decades, a lot of research has been done in the implementation and modification of these methods. We refer to following literature for an overview [Kuz06, Kuz07, Kuz09, Kuz12, LKSM17, Loh]. The analysis of these schemes was developed recently in [BJK16, BJK17]. The first results regarding convergence and solvability were presented in [BJK16]. In [BJK17] a new limiter was proposed that made the AFC schemes linearity preserving, i.e., the modification vanishes whenever we have a linear solution. The study is also supported by the work in [BBK17] where a link between the nonlinear edge-based diffusion and the AFC schemes is presented. A review of the analysis can be found in [BJKR18].

One of the major drawbacks of the AFC schemes is their nonlinear nature. Even if one works with linear PDEs, after applying the AFC schemes, one gets a system of nonlinear equations. A brief overview on this topic was given in [BJKR18]. This thesis presents a comprehensive study towards the solving of these nonlinear equations. In this work we investigate different iterative schemes along with certain algorithmic components.

Another approach while solving singularly pertubed problems is the use of adaptive methods controlled by a posteriori error estimation. The first step towards solving a posteriori error estimation problem was done by Babuška and Rheinboldt, in [BR78]. And after that from $1978-1983$, several results for explicit error estimator techniques were obtained, see [BR81].

---

[1] Images from: *Ludwig Prandtl A Personal Biography Drawn from Memories and Correspondence* and Oberwolfach Photo Collection , `https://opc.mfo.de/detail?` `photoID=` 2458

Figure 1.2: Ivo Babuška and Werner Rheinboldt.[2]

By the early 1990s basic techniques of a posteriori error estimation were established and then the focus shifted to its real-life applications problems, see [Ver94, Ver98]. In the review [Sty05], the author predicts the success of adaptive methods over other methods for solving Convection-Diffusion-Reaction equations. In this thesis, we combine the idea of algebraic stabilization with adaptive refinements.

In many stabilization techniques one assumes certain assumptions on the grids, such as Delaunay triangulation or weakly acute triangulations, see [XZ99, Kno06, BJK16]. While using adaptive refinements one has to close the grids using conforming closures. This leads to the subsequent grids losing the initial grid properties. One way around this is the use of hanging nodes. Also, in three dimensions after the refinement process, the conforming closure of the grid leads to problematic or non-admissible elements such as prisms and pyramids. In this thesis we study the interplay of hanging nodes and AFC schemes.

1.2 Outline

The workflow of the thesis is as follows:

---

[2] Images from: `https://users.oden.utexas.edu/ babuska/` and `https://www.professoren.tum.de/en/honorary-professors/r/rheinboldt-werner`

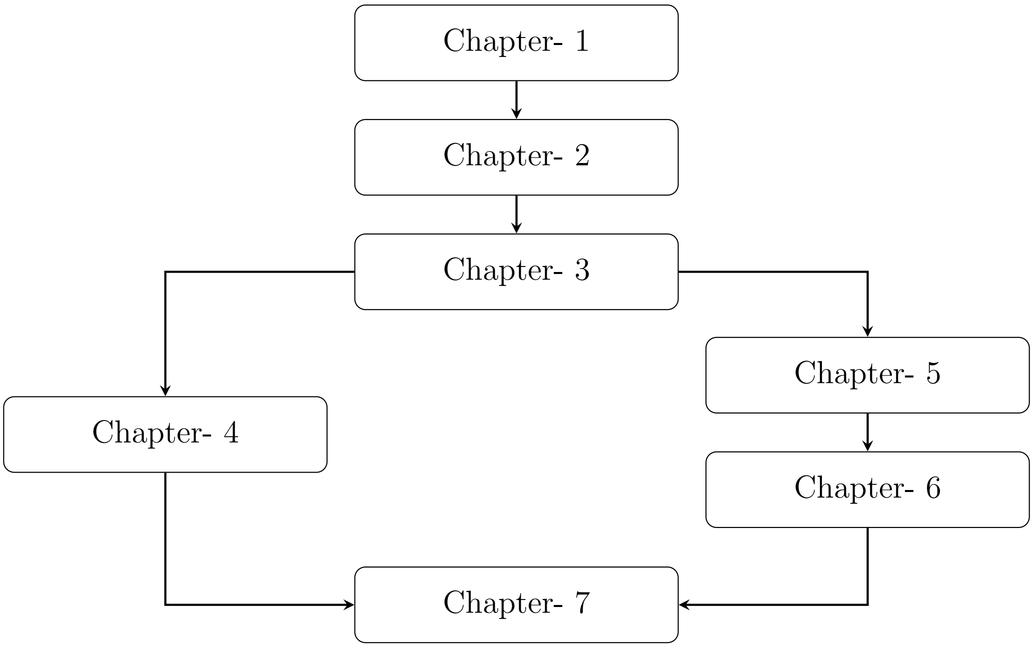


Figure 1.3: Workflow of the thesis

The first three chapters of the thesis contain introductory material. Chapter 2 formally introduces and derive the steady-state and the evolutionary Convection-Diffusion-Reaction equations. It is shown here that the analytic solution satisfies the maximum principles (both the weak and the strong form). The chapter closes with the study of the existence and uniqueness of the weak solution of these equations.

In Chapter 3 we introduce the idea of stabilized finite element methods. The formal definition of the discrete maximum principle (DMP) is given along with the necessary and sufficient conditions for the satisfaction of DMP. A brief overview of some of the standard discretizations and stabilization methods such as the Galerkin and the SUPG methods are given with certain analytic results. Then, the main topic of this thesis, the *algebraic flux correction* (AFC) schemes are discussed. A brief introduction alongside with the definition of different limiters and a review of the analytical results are presented.

The main work of the thesis starts from Chapter 4. Here, different iterative solvers for the steady-state Convection-Diffusion-Reaction equations are presented along with certain algorithmic components such as dynamic damping [JK08], Anderson acceleration, [WN11] etc. Finally, numerical simulations validating the results are presented in two and three dimensions. The different techniques are compared on the basis of the number of nonlinear steps and their solving time. It was found out that the most efficient approach, from computing time, is a simple fixed point approach, because the solvers can exploit the properties of the iteration matrix. This part of the thesis has already been published in [JJ20, JJ19].

Chapter 5 and Chapter 6 present the study of a posteriori error estimation for AFC schemes.

In Chapter 5, we introduce some standard definitions, refinement techniques, and auxiliary results that are used in the literature of a posteriori estimation. Then two different proposals are introduced for the global upper bound in the energy norm of the error, one using a residual-based approach while the other uses the SUPG solution along with the SUPG estimators presented in [JN13]. Numerical studies are done for examples in two dimensions where the performance of the estimator is compared on the effectivity index, adaptive grid refinement, and the individual terms of the estimators. It turns out that the residual-based approach gave better results with adaptive grid refinements whereas the SUPG approach gave better results with respect to the effectivity index. This chapter has been submitted for publication and a pre-print version is available [Jha20].

Chapter 6 studies the error estimators presented in the previous chapter on grids with hanging nodes. A preliminary including definitions and auxiliary results are presented. This chapter also presents certain results for Lagrange elements on grids with hanging nodes. First, we extend the idea of hanging nodes from lower order Lagrange elements to higher-order elements. Then, we study the hanging nodes in the context of AFC schemes and present examples of how certain limiters fail to satisfy DMP and while what modifications are required for others. Numerical simulations are presented in two dimensions which validate the theoretical findings. It turns out that the grids with hanging nodes do not perform better than the grids with conforming closure and hence should not be used with AFC schemes.

Finally, Chapter 7 summarizes all the results that have been presented in this thesis and provides an outlook for the future work.

This thesis also includes two appendices. Appendix A gives the flowchart diagram of the algorithms that are used in Chapter 4. Appendix B gives the numerical values for the effectivity index for the example studied in Chapter 5.

1.3 Function Spaces

In this section we present a brief overview of the function spaces that will be used throughout this thesis. We define the spaces over a bounded domain $\Omega \subset \mathbb{R}^d, \ \ d \in \{2,3\}$. For a detailed review on functional analysis, we refer to [Ada75].

Definition 1.1. (**Space of Lebesgue $p-$integrable functions**, $L^p(\Omega)$) The Lebesgue spaces are defined by

$$L^p(\Omega) := \left\{ f : \Omega \to \mathbb{R}^d : \int_\Omega |f(x)|^p dx < \infty \right\}, \quad p \in [1, \infty),$$

where the integral is to be understood in the sense of Lebesgue. The space $L^\infty(\Omega)$ is the space of all functions that are bounded for almost all $x \in \Omega$, i.e.,

$$L^\infty(\Omega) := \left\{ f : \Omega \to \mathbb{R}^d : |f(x)| < \infty \text{ for almost all } x \in \Omega \right\}.$$

Remark 1.2. The space $L^p(\Omega)$ is a Banach space with norm

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f(x)|^p dx \right)^{1/p}, \quad p \in [1, \infty)$$

and $L^\infty(\Omega)$ is a Banach space with the norm

$$\|f\|_{L^\infty(\Omega)} := \operatorname*{ess\,sup}_{x \in \Omega} |f(x)|,$$

where $\operatorname{ess\,sup}_{x \in \Omega}$ is the essential supremum.

Remark 1.3. For $p = 2$, the space $L^2(\Omega)$ is a Hilbert space with inner product defined as

$$(f, g)_{L^2(\Omega)} := \int_\Omega f(x) g(x) dx$$

and the induced norm

$$\|f\|_{L^2(\Omega)} := (f, f)^{1/2}_{L^2(\Omega)}.$$

Definition 1.4. (**Multi-index**) A multi-index α is a vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ with $\alpha_i \in \mathbb{N} \cup \{0\},\ i = 1, \ldots, n$. Derivatives are denoted by

$$D^{\boldsymbol{\alpha}} = \frac{\partial^{|\boldsymbol{\alpha}|}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \quad \text{with} \quad |\boldsymbol{\alpha}| = \sum_{i=1}^n \alpha_i.$$

Definition 1.5. (**Sobolev spaces**, $W^{k,p}(\Omega)$) Let $k \in \mathbb{N}$ and $p \in [1, \infty]$. The Sobolev space $W^{k,p}(\Omega)$ consists of all integrable functions $f : \Omega \to \mathbb{R}$ such that for each multi-index $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq k$, the derivative $D^{\boldsymbol{\alpha}} f$ exists in the weak sense and it belongs to $L^p(\Omega)$.

Remark 1.6. A norm is defined on the space $W^{k,p}(\Omega)$, as

$$\|f\|_{k,p,\Omega} := \left(\sum_{|\boldsymbol{\alpha}| \leq k} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \right)^{1/p}, \quad p \in [1, \infty)$$

and for $p = \infty$ the norm is defined by

$$\|f\|_{k,\infty,\Omega} := \sum_{|\boldsymbol{\alpha}| \leq k} \operatorname{ess\,sup}_{x \in \Omega} |D^{\boldsymbol{\alpha}} f|.$$

Sobolev spaces equipped with these norms are Banach Spaces (see [Eva10]).

We define a semi-norm on $W^{k,p}(\Omega)$ as

$$|f|_{k,p,\Omega} := \left(\sum_{|\boldsymbol{\alpha}|=k} \|D^{\boldsymbol{\alpha}} f\|^p_{L^p(\Omega)} \right)^{1/p}, \qquad p \in [1, \infty)$$

and for $p = \infty$ the semi-norm is defined by

$$|f|_{k,\infty,\Omega} := \sum_{|\boldsymbol{\alpha}|=k} \text{ess sup}_{x\in\Omega} |D^{\boldsymbol{\alpha}} f|.$$

Remark 1.7. For $p = 2$, the Sobolev spaces are Hilbert spaces. They are often denoted by $H^k(\Omega)$, i.e., $W^{k,2}(\Omega) = H^k(\Omega)$ and they are equipped with the inner product

$$(f, g)_{H^k(\Omega)} = \sum_{|\boldsymbol{\alpha}|\leq k} (D^{\boldsymbol{\alpha}} f, D^{\boldsymbol{\alpha}} g)_{L^2(\Omega)} \,.$$

We define the norm and semin-norm in the same way as for $W^{k,p}(\Omega)$ and denote them by $\| \cdot \|_{k,\Omega}$, $| \cdot |_{k,\Omega}$ respectively.

Definition 1.8. (**Sobolev Spaces**, $H^k_0(\Omega)$) The Sobolev spaces $H^k_0(\Omega)$ are defined as the closure of $C^\infty_0(\Omega)$ with respect to the norm of $H^k(\Omega)$.

Remark 1.9. The space $H^{-1}(\Omega)$ is the dual space of $H^1_0(\Omega)$ and not $H^1(\Omega)$. Also, the definition of Sobolev spaces can be extended to $k \in \mathbb{R}$ (see [Ada75]).

Remark 1.10. Throughout this work, if not mentioned otherwise a mathematical symbol with an underline, like $\underline{u}$, denotes a vector in $\mathbb{R}^N$.

2 Convection-Diffusion-Reaction-Equations

Imagine an industry having a big chimney on its top emitting smoke. How does the pattern of the smoke behave? One observes three things: first the smoke moves in the direction of wind flow, second the smoke diffuses and goes from a region of high concentration to low concentration, and finally, the smoke reacts with the air (Fig.2.1). The physical processes corresponding to the above three phenomena are known as *convection*, *diffusion*, and *reaction*, respectively. The same phenomena can be observed when we add some color to a flowing river. The partial differential equations (PDEs) that models such process is called Convection-Diffusion-Reaction equations. This chapter introduces the Convection-Diffusion-Reaction equations and discusses their properties.

The contents of the chapter are as follows: the equations is derived in Sec. 2.1 using principle of superposition. Next, results are presented in Sec. 2.2 regarding the physical properties that are satisfied by the equations namely the maximum principles. Sec. 2.3 gives results on the existence and uniqueness of the weak solution using Lax-Milgram lemma. Lastly, Sec. 2.4 summarizes the content of the chapter.

2.1 Derivation

The Convection-Diffusion-Reaction equations can be derived using the principle of superposition, i.e., convection and diffusion can be added together if they are linearly independent. It is known that these processes are independent because the only way they can be dependent is if one process feeds back to the other, which is not true in our case. It is known that diffusion is a random process due to molecular motion. Due to diffusion, each molecule moves with the same probability in an arbitrary direction and due to convection, each molecule will also move in the flow direction. These processes are clearly additive and independent; because the presence of flow does not affect the probability that the molecule will take a diffusive step in an arbitrary direction, it just adds something.

Let $\Omega \subseteq \mathbb{R}^d$, $d \in \{1,2,3\}$, be a bounded domain with a regular boundary Γ with outward

Figure 2.1: Smoke from an industry depicting convection, diffusion, and reaction.

pointing unit normal $\boldsymbol{n}$. Let $u = u(t,x)$ $[mol/m^d]$ be the concentration of the reactant. Using the conservation of mass, we have

$$\frac{d}{dt}\int_{\Omega} u dV = -\int_{\Gamma} \underline{f}_{\text{flux}} \cdot \boldsymbol{n} ds + \int_{\Omega} \hat{f} dV,$$

where, $\underline{f}_{\text{flux}}$:= total flux and $\hat{f} := \hat{f}(u(t,x))$ $[mol/(m^d s)]$ is the source term.

Since all functions and Γ are assumed to be sufficiently smooth, we can simplify the above equations by using the divergence theorem and changing differentiation with respect to time and integration with respect to space which leads to

$$\int_{\Omega} \frac{du}{dt} dV = -\int_{\Omega} \nabla \cdot \underline{f}_{\text{flux}} dV + \int_{\Omega} \hat{f} dV.$$

As Ω was an arbitrary volume it follows

$$\frac{du}{dt} + \nabla \cdot \underline{f}_{\text{flux}} = \hat{f}. \tag{2.1}$$

Now, as we have already noted that our convection and diffusion are independent events, we can write our flux as the sum of convection and diffusion, i.e.,

$$\underline{f}_{\text{flux}} := \underline{f}^{\text{conv}}_{\text{flux}} + \underline{f}^{\text{diff}}_{\text{flux}}.$$

Now, using Fick's law[1] for $\underline{f}_{\text{flux}}^{\text{diff}}$ and using convective transport for $\underline{f}_{\text{flux}}^{\text{conv}}$ we have,

$$\begin{aligned} \underline{f}_{\text{flux}}^{\text{diff}} &= -\varepsilon \nabla u, \\ \underline{f}_{\text{flux}}^{\text{conv}} &= \boldsymbol{b} u, \end{aligned} \tag{2.2}$$

where, $\varepsilon :=$ diffusion coefficient$[m^2/s]$ and $\boldsymbol{b} :=$ convective velocity $[m/s]$.

Substituting (2.2) into (2.1) and simplifying the equation we get

$$\frac{\partial u}{\partial t} - \varepsilon \Delta u - u \nabla \cdot \varepsilon + \nabla \cdot (\boldsymbol{b} u) = \hat{f}. \tag{2.3}$$

If we have the case of *incompressible flow* and constant diffusion, then we can further simplify our equation. For incompressible flow, our mass equation reduces to the continuity equation as,

$$\nabla \cdot \boldsymbol{b} = 0.$$

And for constant diffusion we have $\nabla \cdot \varepsilon = 0$. So, our equation reduces to

$$\frac{\partial u}{\partial t} - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u = \hat{f}. \tag{2.4}$$

Separating the reaction term $(c(t, x))$ with the source/sink term (f) and introducing appropriate boundary and initial conditions, our Evolutionary Convection-Diffusion-Reaction equations are

$$\begin{aligned} \frac{\partial u}{\partial t} - \varepsilon \Delta u + \boldsymbol{b} \cdot \nabla u + c u = f \quad & \text{in } (0, T] \times \Omega, \\ u = u^b \quad & \text{on } (0, T] \times \Gamma_{\mathrm{D}}, \\ -\varepsilon \nabla u \cdot \boldsymbol{n} = g \quad & \text{on } (0, T] \times \Gamma_{\mathrm{N}}, \\ u(0, x) = u_0(x) \quad & \text{in } \Omega, \end{aligned} \tag{2.5}$$

where $[0, T] :=$ time interval, $\Gamma_{\mathrm{D}} :=$ Dirichlet boundary, $\Gamma_{\mathrm{N}} :=$ Neumann boundary, $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$, and $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$.

---

[1]

Definition 2.1. Fick's first law of diffusion is as follows,

$$N = -D \nabla c,$$

where, $N =$ the flux $[mol/m^{d-1}s]$, $D =$ the diffusion coefficient m^2/s and $c =$ the concentration mol/m^d.

In case of steady-state concentration field our equations reduces to:

$$\begin{aligned} -\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + cu &= f && \text{in } \Omega,\\ u &= u^b && \text{on } \Gamma_D,\\ -\varepsilon\nabla u\cdot\boldsymbol{n} &= g && \text{on } \Gamma_N. \end{aligned} \tag{2.6}$$

This chapter focuses on the steady state Convection-Diffusion-Reaction equations. Maximum principles and weak solution theory also exist for the Evolutionary Convection-Diffusion-Reaction equations (see [Eva10, Chapter 7, Sec. 7.1]).

2.2 Maximum Principles

As we have already noted that PDEs model physical processes, hence we expect them to satisfy their physical properties as well. For Convection-Diffusion-Reaction equations, the physical property that we are interested in are the maximum principles. In this section, we will first introduce some definitions and then prove the weak maximum principle (see [Eva10, Sec. 6.4.1]) for Convection-Diffusion-Reaction equations followed by the strong maximum principle (see [Eva10, Sec. 6.4.2]).

Definition 2.2. Let $\Omega \subset \mathbb{R}^d$ be an open connected set with boundary Γ. Let L be the second order differential operator defined by,

$$\begin{aligned} L &:= -\varepsilon\Delta + \boldsymbol{b}\cdot\nabla + c\\ &= -\sum_{i,j=1}^{N}\varepsilon_{ij}(x)D_{ij} + \sum_{i=1}^{N} b_i(x)D_i + c(x), \end{aligned} \tag{2.7}$$

with $\varepsilon_{ij}(x) \in L^\infty_{\text{loc}}(\Omega)$, $b_i,\ c \in L^\infty(\Omega)$, $D_i = \frac{\partial}{\partial x_i}$, and $D_{ij} = \frac{\partial^2}{\partial x_i \partial x_j}$. Without loss of generality we also assume the symmetric condition, $\varepsilon_{ij} = \varepsilon_{ji}$.

1. We say that the operator L is *elliptic* on Ω if for every $x \in \Omega$ there is $C_{\text{elliptic}} > 0$ such that
$$\sum_{i,j=1}^{N}\varepsilon_{ij}(x)z_i z_j \geq C_{\text{elliptic}}\|z\|_{l^2}^2 \quad \forall\ z \in \mathbb{R}^d.$$
2. We say that the operator L is *strictly elliptic* on Ω if there is $C_{\text{elliptic}} > 0$ such that
$$\sum_{i,j=1}^{N}\varepsilon_{ij}(x)z_i z_j \geq C_{\text{elliptic}}\|z\|_{l^2}^2 \quad \forall\ z \in \mathbb{R}^d,$$

where $\| \cdot \|_{l^2}$ is the Euclidean norm.

Lemma 2.3. *Let L be an elliptic operator of the form of (2.7), such that $c \geq 0$. If $u \in C^2(\Omega)$ and $Lu < 0$ in Ω, then u cannot attain a non-negative maximum in Ω.*

Proof. We will prove this using contradiction. Let u has a non-negative maximum in Ω, i.e., $\exists\ x_0 \in \Omega$ such that $u(x_0) \geq 0$. As $u(x)$ has a maximum at x_0, therefore $D_i(x_0) = 0\ \ \forall i \in \{1, \dots, n\}$. Then

$$\sum_{i=1}^{N} b_i(x) D_i u(x_0) + c(x) u(x_0) \geq 0.$$

As we assumed $\varepsilon = \{\varepsilon_{ij}\}_{i,j=1}^{N}$ to be symmetrical, we can diagonalize it. Hence, $T\varepsilon T^T = D$ where T^T is the transpose of T and D is the diagonal matrix. Note, as L is elliptic, we have $d_{ii} \geq C_{\text{elliptic}}(x_0) > 0$.

Using the substitution $y = Tx$ and $U(y) = u(x)$, with the help of chain rule we get

$$\begin{aligned} \sum_{i,j=1}^{N} \varepsilon_{ij}(x_0) D_{ij} u(x_0) &= \sum_{i,j=1}^{N} \varepsilon_{ij}(x_0) \sum_{l=1}^{N} T_{li} \sum_{k=1}^{N} T_{kj} \left(\frac{\partial^2 U}{\partial x_i \partial x_j} \right) (Tx_0) \\ &= \sum_{i=1}^{N} d_{ii} \frac{\partial^2 U}{\partial y_i^2} (Tx_0) \,. \end{aligned}$$

Since U has a maximum at Tx_0 we have $\frac{\partial^2 U}{\partial y_i^2}(Tx_0) \leq 0\ \ \forall\ i \in \{1, \dots, N\}$, implying

$$-\sum_{i=1}^{N} d_{ii} \frac{\partial^2 U}{\partial y_i^2}(Tx_0) \geq 0.$$

Hence, $Lu(x_0) \geq 0$ which is a contradiction. □

Theorem 2.4. **(Weak maximum principle)** ([Eva10, Theorem 2, Sec. 6.4.1]) *Suppose that Ω is a bounded domain and L is a strictly elliptic operator with $c \geq 0$. If $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $Lu \leq 0$ in Ω, then a non-negative maximum is obtained at the boundary.*

Proof. Suppose that $\Omega \subset \{x = (x_1, \dots, x_d) : |x_1| < C_3\}$ for some $C_3 > 0$. Consider $w(x) = u(x) + C_1 e^{C_2 x_1}$ with $C_1,\ C_2 > 0$. Then by using strictly elliptic property of L, we have

$$\begin{aligned} Lw &= Lu + C_1 \left(-C_2^2 \varepsilon_{11}(x) + C_2 b_1(x) + c(x) \right) e^{C_2 x_1} \\ &\leq C_1 \left(-C_2^2 C_{\text{elliptic}} + C_2 \|b\|_\infty + \|c\|_\infty \right) e^{C_2 x_1}. \end{aligned}$$

We can choose C_2 large enough so that $Lw < 0$. By Lemma 2.3 w cannot have a non-negative maximum in Ω. Hence using the properties of supremum and boundedness of Ω, we have

$$\sup_{\Omega} u \leq \sup_{\Omega} w \leq \sup_{\Omega} w^+ = \sup_{\Gamma} w^+ \leq \sup_{\Gamma} u^+ + C_1 e^{C_2 C_3},$$

where $w^+ = \max\{w, 0\}$.

For $C_1 \to 0$ we get the result. □

Lemma 2.5. (Hopf's lemma) *Assume $u \in C^2(\Omega) \cap C^1(\overline{\Omega})$ and $c \geq 0$ in Ω. Suppose further $Lu \leq 0$ in Ω, and $\exists\ x_0 \in \Gamma$ such that $u(x_0) \geq 0$. Finally assume that Ω satisfies the interior ball condition at x_0 (i.e. $\exists$ an open ball $B \subset \Omega$ with $x_0 \in \partial B$). Then*

$$\frac{\partial u}{\partial \boldsymbol{n}_B}(x_0) > 0, \tag{2.8}$$

where $\boldsymbol{n}_B$ is the outer normal to B at x_0.

Proof. Refer to [Eva10, Sec. 6.4.2]. □

Theorem 2.6. (Strong maximum principle) ([Eva10, Theorem 3, Sec. 6.4.2]) *Assume $u \in C^2(\Omega) \cap C(\overline{\Omega})$ and $c \geq 0$ in Ω. Also assume Ω is an open, connected and bounded domain. If $Lu \leq 0$ in Ω and u attains its maximum over $\overline{\Omega}$ at an interior point, then u is constant within Ω.*

Proof. We will use method of contradiction to prove the theorem.

Let $\max_{\overline{\Omega}} u = M$ and $C = \{x \in \Omega : u(x) = M\}$. Suppose u is not constant, i.e., $u \not\equiv M$, and set $V = \{x \in \Omega : u(x) < M\}$.

Choose $y \in V$ such that $\operatorname{dist}(y, C) < \operatorname{dist}(y, \Gamma)$ and let B denote the largest ball with center y whose interior lies in V. Then $\exists\ x_0 \in C$ with $x_0 \in \partial B$ (See Fig. 2.2). Hence V satisfies the interior ball condition at x_0, so by Hopf's lemma,

$$\frac{\partial u(x_0)}{\partial \boldsymbol{n}_B} > 0.$$

Since $x_0 \in \Omega$ is maximum of u, therefore $Du(x_0) = 0$ which is a contradiction. □

Remark 2.7. (Minimum principles) One can also obtain the corresponding weak and strong minimum principles by replacing u with $-u$ in the above statements.

Example 2.8. (Standard 1d example) Let us take a simple example for (2.6) i.e. $d = 1$, $\boldsymbol{b} = 1$, $c = 0$, and $f = 1$ with homogeneous Dirichlet boundary conditions. Then our

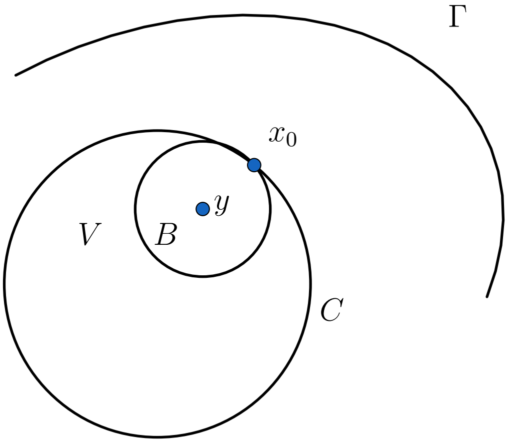


Figure 2.2: Proof of strong maximum principle.

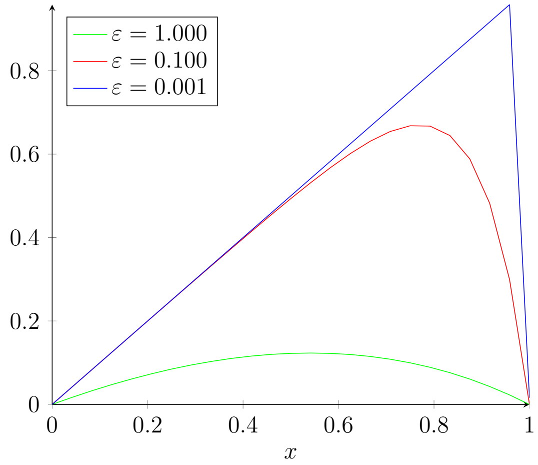


Figure 2.3: Solution of Example 2.8 for different values of ε.

equations reduce to

$$-\varepsilon u'' + u' = 1, \qquad u(0) = u(1) = 0,$$

which has the solution

$$u(x) = x - \frac{\exp\left(\frac{x-1}{\varepsilon}\right) - \exp\left(\frac{-1}{\varepsilon}\right)}{1 - \exp\left(\frac{-1}{\varepsilon}\right)}.$$

We see from the Fig. 2.3 that the solution satisfies the minimum principle for different values of ε. We also note that the solution becomes steep close to the right boundary as ε becomes small.

2.3 Weak Solution Theory

The existence and uniqueness of the solution for a PDE is quite a big research area. While proving the existence of a classical solution of a PDE requires all coefficients to be sufficiently smooth. In higher dimensions also the domain has to satisfy certain regularity conditions. Such smoothness (or regularity) conditions are in practice often not satisfied and hence, in general, one cannot expect the PDE to possess a classical solution. Nevertheless, the process that is modeled with PDE occurs and hence a different notation of solution is required. However, this solution will not possess the smoothness properties of the classical solution.

In this section, we present the notion of weak solution for Convection-Diffusion-Reaction equations and show its existence and uniqueness.

Remark 2.9. Multiply (2.6) with an appropriate function $v(x)$ with $v = 0$ on Γ and then integrate the resulting equation on Ω. Integration by parts leads to

$$\begin{aligned}\int_\Omega (-\varepsilon\Delta u + \boldsymbol{b}\cdot\nabla u + cu)\, v dx &= \int_\Gamma -\varepsilon\,(\nabla u\cdot \boldsymbol{n})\,(s) v ds + \int_\Omega \Big(\varepsilon\nabla u\cdot\nabla v + (\boldsymbol{b}\cdot\nabla u)\, v \\ &\quad +cuv\Big) ds \\ &= \int_\Omega (\varepsilon\nabla u\cdot\nabla v + (\boldsymbol{b}\cdot\nabla u)\, v + cuv)\, ds.\end{aligned}$$

The integral on the boundary vanishes because of the boundary conditions on v. Denoting the $L^2(\Omega)$ inner product by $(\cdot,\cdot)$, we can write the above equation in compact form.

Definition 2.10. (Weak or Variational formulation) Let $\boldsymbol{b}, c \in L^\infty(\Omega)$ and $f \in H^{-1}(\Omega)$ with $u = 0$ on Γ. The *weak or variational formulation* of a Convection-Diffusion-Reaction problem is given by: Find $u \in H^1_0(\Omega)$ such that $\forall v \in H^1_0(\Omega)$,

$$\varepsilon\,(\nabla u, \nabla v) + (\boldsymbol{b}\cdot\nabla u + cu, v) = \langle f, v\rangle \tag{2.9}$$

where $\langle\cdot,\cdot\rangle$ denote the duality pairing between $V (= H^1_0(\Omega))$ and its dual $V' (= H^{-1}(\Omega))$. A solution of (2.9) is called *weak solution.* The space in which the solution is searched is called *solution* or *ansatz* space. The function $v(x)$ is called *test function* and the space from which it comes is the *test space.* Note, $H^{-1}(\Omega)$ is dual of $H^1_0(\Omega)$ and not $H^1(\Omega)$.

Remark 2.11. **Boundary conditions**

1. *Essential boundary condition:* Consider inhomogeneous Dirichlet boundary conditions

$$u(x) = u^b(x) \quad \text{ on } \Gamma.$$

Such boundary conditions are included into the definition of the ansatz space,

$$V_{u^b} = \{v \in H^1(\Omega)\ :\ v|_\Gamma = u^b\},$$

where the restriction to the boundary is understood in the sense of traces. As they are required for the definition of the space they are called as essential boundary condition. The test space is still $V = H_0^1(\Omega)$. Then the weak formulation reads as follows: Find $u \in V_{u^b}$ such that

$$\varepsilon\left(\nabla u, \nabla v\right) + \left(\boldsymbol{b} \cdot \nabla u + cu, v\right) = \langle f, v\rangle \quad \forall v \in V.$$

A different way of writing the variational problem uses an extension $u_{u^b} \in H^1(\Omega)$ of $u^b(x)$ to Ω. The existence of such extension follows from the trace theorem. Then one seeks $u \in H^1(\Omega)$ such that $\tilde{u} = u - u_{u^b} \in V$ and

$$\begin{aligned}\varepsilon\left(\nabla \tilde{u}, \nabla v\right) + \left(\boldsymbol{b} \cdot \nabla \tilde{u} + c\tilde{u}, v\right) &= \langle f, v\rangle \\ &\quad + \varepsilon\left(\nabla u_{u^b}, \nabla v\right) + \left(\boldsymbol{b} \cdot \nabla u_{u^b} + cu_{u^b}, v\right) \quad \forall v \in V.\end{aligned}$$

In this formulation, one has the same ansatz and test space.

2. *Natural boundary conditions:* Neumann boundary conditions appear in a straightforward way in the variational formulation. Since the integral on Neumann boundaries appears in the integration by parts they are called natural boundary conditions. For simplicity assume $u(x) = 0 \ \ \forall x \in \Gamma_{\mathrm{D}}$. Let $V_D = \{v \in H^1(\Omega) \ : \ v|_{\Gamma_D} = 0\}$, then the variational formulation has the form. Find $u \in V_D$ such that

$$\varepsilon\left(\nabla u, \nabla v\right) + \left(\boldsymbol{b} \cdot \nabla u + cu, v\right) = \langle f, v\rangle + \int_{\Gamma_N} \varepsilon\left(\nabla u \cdot \boldsymbol{n}\right)(s) v(s) ds \quad \forall\ v \in V_D.$$

Definition 2.12. (Properties of bilinear form) Let $(V, \|\cdot\|_V)$ be a Banach space. A map $a : V \times V \to \mathbb{R}$ is called

1. bilinear, if $a(\cdot,\cdot)$ is linear in both arguments,
2. symmetric, if $a(u, v) = a(v, u)$ for all $u, v \in V$,
3. positive, if $a(u, u) \geq 0$ for all $u \in V$,
4. coercive or V-elliptic or positive definite, if there is a $C_{\text{elliptic}} > 0$ such that $a(u, u) \geq C_{\text{elliptic}}\|u\|_V^2$ for all $u \in V$.
5. bounded if there is a $C_{\text{bound}} > 0$ such that

$$|a(u, v)| \leq C_{\text{bound}}\|u\|_V\|v\|_V$$

for all $u, v \in V$.

Theorem 2.13. (Lax-Milgram lemma) ([Eva10, Theorem 1, Sec. 6.2.1]) *Let $a(\cdot,\cdot) : V \times V \to \mathbb{R}$ be a bounded and coercive bilinear form on the Hilbert space V. Then, for each* *bounded linear functional $f \in V'$ there is exactly one $u \in V$ with*

$$a(u,v) = \langle f, v\rangle \quad \forall v \in V. \tag{2.10}$$

Proof. Since, for $u \in V$ the mapping $v \mapsto a(u,v)$ is a bounded linear functional on V, by Riesz representation theorem[2] we know that there exists an unique element $w \in V$ such that

$$a(u,v) = (w,v) \quad \forall\ v \in V.$$

Let us denote w by Au whenever the above equation holds; so that

$$a(u,v) = (Au,v) \quad \forall\ v \in V.$$

We claim that $A : V \to V$ is a bounded linear operator. Let $C_1, C_2 \in \mathbb{R}$ and $u_1, u_2 \in V$, then for each $v \in V$

$$\begin{aligned}(A(C_1u_1 + C_2u_2), v) &= a(C_1u_1 + C_2u_2, v)\\ &= C_1a(u_1,v) + C_2a(u_2,v)\\ &= C_1(Au_1,v) + C_2(Au_2,v)\\ &= (C_1Au_1 + C_2Au_2, v).\end{aligned}$$

As the equality holds for all $v \in V$, and so we get that A is linear. Furthermore,

$$\|Au\|^2 = (Au,Au) = a(u,Au) \le C_{\text{bound}}\|u\|\|Au\|.$$

Consequently $\|Au\| \le C_{\text{bound}}\|u\|$ for all $u \in V$, and so A is bounded and which implies continuity as well.
Next we claim that A is injective, and $R(A)$, the range of A is closed in V. To show this, we compute

$$C_{\text{elliptic}}\|u\|^2 \le a(u,u) = (Au,u) \le \|Au\|\|u\|.$$

Hence, $C_{\text{elliptic}}\|u\| \le \|Au\| \quad \forall\ u \in\ V$. We will use this to show that A is injective, and $R(A)$ is closed in V.

Let $u_1, u_2 \in V$ such that $Au_1 = Au_2$, then by linearity of A we have $C_{\text{elliptic}}\|u_1 - u_2\| \le \|Au_1 - Au_2\| = 0$, hence $u_1 = u_2$. For closeness, let $\{Au_n\}_{n=1}^{\infty}$ be a convergent sequence in

---

[2]

Theorem 2.14. (**Riesz Representation Theorem**) ([Eva10, Theorem 2, Appendix D]) *Let H be a real Hilbert space, with inner product $(\cdot,\cdot)$, and H' denote it's dual space. Then, for each $u' \in H'$ there exists a unique element $u \in H$ such that*

$$\langle u', v\rangle = (u,v) \quad \forall\ v \in H.$$

The mapping $u' \mapsto u$ is a linear isomorphism of H' onto H.

$R(A)$. We want to show its limit, say Au, belongs to V. Since, Au_n is a convergent sequence we have it is a Cauchy sequence as well. Now,

$$\|u_n - u_m\| \leq \frac{\|Au_n - Au_m\|}{C_{\text{elliptic}}}$$

so, we have $\{u_n\}_{n=1}^{\infty}$ is a Cauchy sequence and as V is complete we have $\{u_n\}_{n=1}^{\infty}$ is convergent as well. Now, A is a continuous operator and hence $Au_n \to Au$. By uniqueness of the limit we have the limit of $\{Au_n\}_{n=1}^{\infty}$ belongs in $R(A)$.

Finally, we claim that $R(A) = V$. Suppose not, then $\exists\ \overline{u}(\neq 0) \in R(A)^{\perp}$ (orthogonal complement of $R(A)$) but its a contradiction as $C_{\text{elliptic}}\|\overline{u}\|^2 = a(\overline{u}, \overline{u}) = (A\overline{u}, \overline{u}) = 0$ implies $\overline{u} = 0$.

By Riesz representation theorem for $f \in V'$ we have some $\overline{w} \in V$ such that

$$\langle f, v\rangle = (\overline{w}, v) \quad \forall\ v \in V.$$

Hence, our problem reduces to finding a solution $u \in V$ such that $Au = \overline{w}$. Now, as A is bijective we have the existence of the inverse and so there exist a unique solution to the above problem. ☐

Corollary 2.15. (Existence and uniqueness of a solution of (2.9)) *Let $V = H_0^1(\Omega)$ and assume $f \in V'$, $\boldsymbol{b}$, $\nabla\boldsymbol{b}$, $c \in L^{\infty}(\Omega)$, and*

$$\left(c - \frac{1}{2}\nabla\cdot\boldsymbol{b}\right)(x) \geq \sigma_0 > 0 \quad \forall x \in \Omega \tag{2.11}$$

(as almost everywhere on Ω). Then, (2.9) has an unique solution.

Proof. Let us define,

$$a(u, v) := \int_{\Omega} \left(\varepsilon\nabla u(x)\cdot\nabla v(x) + \boldsymbol{b}\cdot\nabla u(x)v(x) + c(x)u(x)v(x)\right)dx \tag{2.12}$$

then it is a bilinear form in the space $V = H_0^1(\Omega)$ with $\|v\|_V = \|\nabla v\|_{L^2(\Omega)}$ for $v \in V$. This follows directly from the linearity of integration and differentiation.

Now, if we show this bilinear form is coercive and bounded then we have existence and uniqueness of the solution to (2.9) by the Lax-Milgram lemma.

1. *Coercitivity of $a(\cdot,\cdot)$.* Using integration by parts and the product rule, one obtains

$$\frac{1}{2}\int_{\Omega}\left(\boldsymbol{b}(x)\cdot\nabla v(x)\right)v(x)dx \quad = \quad -\frac{1}{2}\int_{\Omega}\nabla\cdot\left(\boldsymbol{b}(x)v(x)\right)v(x)dx$$

$$
\begin{aligned}
&= -\frac{1}{2}\int_\Omega (\nabla\cdot\boldsymbol{b}(x))\, v(x)v(x)dx \\
&\quad -\frac{1}{2}\int_\Omega (\boldsymbol{b}(x)\cdot\nabla v(x))\, v(x)dx.
\end{aligned}
$$

Inserting the above expression in (2.12) with $u(x) = v(x)$ and using (2.11) one has,

$$a(v,v) \geq \int_\Omega \varepsilon(\nabla v(x))^2 dx = \varepsilon\|\nabla v\|^2_{L^2(\Omega)} = \varepsilon\|v\|^2_V.$$

Hence, $a(\cdot,\cdot)$ is coercive.

2. *Boundedness of* $a(\cdot,\cdot)$. Using the Cauchy-Schwarz inequality, Hölder's inequality, and the Poincaré-Friedreichs inequality[3] we get,

$$
\begin{aligned}
|a(u,v)| &\leq \varepsilon\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + \|\boldsymbol{b}\|_{L^\infty(\Omega)}\|\nabla u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\
&\quad +\|c\|_{L^\infty(\Omega)}\|u\|_{L^2(\Omega)}\|v\|_{L^2(\Omega)} \\
&\leq \varepsilon\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} + C_{\mathrm{PF}}\|\boldsymbol{b}\|_{L^\infty(\Omega)}\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \\
&\quad +C^2_{\mathrm{PF}}\|c\|_{L^\infty(\Omega)}\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)} \\
&= C\|\nabla u\|_{L^2(\Omega)}\|\nabla v\|_{L^2(\Omega)},
\end{aligned}
$$

where C_{PF} is the constant appearing in Poincaré-Friedreichs inequality. Hence, the bilinear form is bounded and (2.9) has a unique solution.

□

2.4 Summary

This chapter introduced the Convection-Diffusion-Reaction equations. An evolutionary and its steady-state counterpart were derived. Sec. 2.2 and Sec. 2.3 concentrated on the maximum principles and the weak solution for the steady-state Convection-Diffusion-Reaction equations, respectively. We noted that for the steady-state equation an important property is that it satisfies the strong and weak maximum principles. While solving the equation numerically we want our numerical solutions to obey these properties as well. Chapter 3 is

---

[3] **Theorem 2.16.** (**Poincaré-Friedreichs inequality**) [BS08, Proposition 5.3.5] *Let* $f \in W^{1,p}_0(\Omega)$, *then*

$$\|f\|_{L^p(\Omega)} \leq \left(\frac{|\Omega|}{\omega_d}\right)^{1/d} \|\nabla f\|_{L^p(\Omega)} = C_{\mathrm{PF}}\|\nabla f\|_{L^p(\Omega)}, \quad p\in[1,\infty),$$

where ω_d *is the volume of the unit ball in* R^d.

dedicated to the study of such numerical methods.

Lastly, we discussed the weak solution theory for the equations and noted that we have existence and uniqueness of a weak solution.

3 Stabilized Finite Element Methods

Numerical solutions that we get sometimes do not behave in the way we want them to behave. The solution might be polluted with oscillations or in some cases might not approximate our solution correctly. To illustrate this let us take a 1d example of (2.6).

Remark 3.1. (**Example of two-point boundary value problem**) Let $\boldsymbol{b} = 1,\ c = 0$, and $f = 0$ in (2.6) then we get an ordinary differential equation

$$-\varepsilon u'' + u' = 0 \quad \text{in} \quad (0,1) \quad , u(0) = 0 \quad , u(1) = 1. \tag{3.1}$$

The analytical solution of the above problem is given by,

$$u(x) = \frac{\exp\left(\frac{x-1}{\varepsilon}\right) - \exp\left(\frac{-1}{\varepsilon}\right)}{1 - \exp\left(\frac{-1}{\varepsilon}\right)}.$$

By applying the central finite difference scheme on the above equation we get our numerical solution as

$$u^i = \frac{r^i - 1}{r^N - 1} \qquad i = \{0, \dots, N\} \qquad \text{with} \quad r = \frac{2\varepsilon + h}{2\varepsilon - h},$$

where h is the mesh width.

If $h \gg 2\varepsilon$, then $r \approx -1$ and hence we have $u^i \approx \frac{(-1)^i - 1}{(-1)^N - 1}$, then for even value of N we get oscillatory solutions which is not correct. Fig. 3.1 depicts this for $\varepsilon = 10^{-6}$ and $N = 32$.

If $h < 2\varepsilon$ then we have useful approximation which can be seen in Fig. 3.1 for $\varepsilon = 10^{-3}$ and $N = 128$.

But in applications $\varepsilon \approx 10^{-6}$, so for useful approximations, we need small mesh width which is not affordable, especially for higher dimension cases. Hence, we note that some kind of modification is required. In numerical analysis terminology, this kind of modification is referred to as *stabilization.* The simplest stabilization technique for finite difference method is the upwind method in which the finite difference approximation of the convective term is computed with values from upwind direction (For a detailed explanation refer to [RST08, Sec. 2.1.2]).

Finite Element Methods (FEM) are a different method of solving PDEs. It has its advantages over the simple finite difference methods like it can be applied to more variety of problems

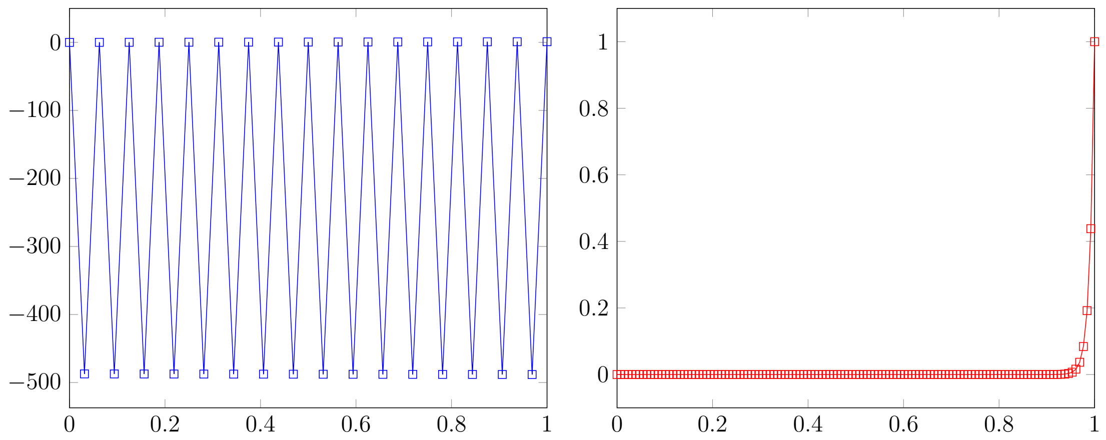


Figure 3.1: Numerical solution for Eq. (3.1) Left: $\varepsilon = 10^{-6}$, $N = 32$, Right: $\varepsilon = 10^{-3}$, $N = 128$.

with complicated domains, it requires less regularity conditions, and incorporates Neumann boundary conditions better. We will focus our attention to the most commonly used FEM scheme, the *Galerkin approximation* (For an overview refer to [GT17]). It is observed that the standard Galerkin method behaves in the same way as the standard finite difference method, i.e., it gives inaccurate solutions when the convection dominates the diffusion [RST08]. This chapter deals with the stabilization techniques for the FEM when applied to (2.6).

The contents of the chapter are as follows: Sec. 3.1 introduces the discrete counterpart of the maximum principles introduced in Chapter 2 namely the Discrete Maximum Principle (DMP). Sec. 3.2 gives a brief overview of the Galerkin method and states the reason why the method fails to give accurate solutions. Sec. 3.3 gives an introduction and some results for one of the commonly used stabilization method namely the *Streamline-Upwind Petrov-Galerkin Method* (SUPG). It also mentions the drawbacks of the scheme and some improvement so as to get slightly better schemes known as *Spurious Oscillations at Layers Diminishing Methods* (SOLD). Sec. 3.4 introduces the main topic of the thesis, the Algebraic Flux Correction schemes (AFC), it mentions the different limiters that are used for simulations and gives a summary of the theoretical results for the existence and uniqueness of the solution.

3.1 Discrete Maximum Principle

Remark 3.2. Discretization of a PDE should provide a numerical solution that is not only a good approximation of the analytical solution but is also physically consistent, i.e., it possesses the same physical properties as that of the solution of the continuous problem. One of the important properties for Convection-Diffusion-Reaction equations is that the numerical solution should be in the range of admissible physical values. From a mathematical viewpoint, this requirement can be formulated as the Discrete Maximum Principle (DMP) which is analogous to the maximum principle introduced in Sec. 2.2.

Definition 3.3. (Discrete maximum principle) Let V_h be a finite dimensional subset of the function space V with dimension N. Let us assume that the last $N-M$ of any $v \in V_h$ corresponds to the nodes on boundary. We say that an operator L_h satisfies the (global) DMP if and only if for any $v \in V_h$ with

$$(L_h v)_i \leq 0, \quad 1 \leq i \leq M,$$

it holds that

$$\max_{i=1,\ldots,M} v_i \leq \max\left\{0, \max_{j=M+1,\ldots,N} v_j\right\}.$$

In other words, the following relation is valid for $i = 1, \ldots, M$

$$(L_h v)_i \leq 0 \Longrightarrow \left(\text{if } v_i \geq 0, \text{ then } v_i \leq \max_{j=M+1,\ldots,N} v_j\right).$$

Definition 3.4. (Monotone matrix) A square matrix A is called *monotone* or *inverse-monotone* matrix if A is non-singular and $A^{-1} \geq 0$.

Remark 3.5. The notation $A \geq 0$ (or $A > 0$) means for a matrix $A = (a_{ij}) \in \mathbb{R}^{N \times N}$ that $a_{ij} \geq 0$ (or $a_{ij} > 0$ $\forall i, j = 1, \ldots, N$).

Theorem 3.6. (**Sufficient and necessary condition for the satisfaction of the DMP**) ([Cia70, Theorem 1]) *Let A be the matrix representation of the operator L_h, then A satisfies DMP if and only if the following conditions are satisfied:*

1. *A is monotone,*
2. *$-(A_i)^{-1}A^b 1_{N-M} \leq 1_M$ i.e., the row sums of $-(A^i)^{-1}A^b$ are smaller than 1,*

where

$$A = \begin{bmatrix} A^i & A^b \\ 0 & I \end{bmatrix},$$

and 1_M is a vector consisting of only 1.

Proof. See [Cia70, Theorem 1]. □

Remark 3.7. The conditions in Theorem 3.6 are based on the inverse of A^i, which is not available in practice. From implementation point of view, sufficient conditions are required which can be used to check easily whether a discretizations satisfies the DMP or not.

Theorem 3.8. (**Sufficient condition for the satisfaction of the DMP**) ([Cia70, Theorem 2]) *Let A be a monotone matrix and let*

$$\sum_{j=1}^{N} a_{ij} \geq 0, \qquad 1 \leq i \leq M \tag{3.2}$$

then the operator L_h that corresponds to the matrix A satisfies the DMP.

Proof. Let $v = 1$, then from (3.2) we get

$$0 \leq \sum_{j=1}^{N} a_{ij} = A^i v^i + A^b v^b.$$

As A is monotone, we have A^i is invertible and non-negative. Applying the inverse $(A^i)^{-1}$ from left, we get

$$v^i + (A^i)^{-1} A^b v^b \geq 0 \implies 1 + (A^i)^{-1} A^b \geq 0$$

which is the second condition for Theorem 3.6. Hence, L_h satisfies the DMP. □

Remark 3.9. The converse of the above theorem fails. Consider the operator L_h whose matrix is given by

$$A = \begin{bmatrix} -1 & 2 & 0 \\ 2 & -3 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(Ref. [Cia70]). Obviously (3.2) fails to hold. But

$$A^{-1} = \begin{bmatrix} 3 & 2 & 0 \\ 2 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \geq 0,$$

i.e., it is monotone. And $-(A^i)^{-1} A^b = 0 \leq 1$. Hence, Theorem 3.6 is satisfied and so it gives that L_h satisfies the DMP.

Now, we are going to discuss the so-called M-matrix and its relation with the DMP. From a numerical analysis point of view for the Convection-Diffusion-Reaction equations, M-matrices play an important role. They are a subset of Monotone matrices who are in some sense diagonally dominant.

Definition 3.10. (M-matrix, [Ost37]) A matrix $A = (a_{ij})_{i,j=1}^{N}$ is an M-matrix if:

1. $a_{ii} \geq 0, \quad i = 1, \ldots, N.$
2. $a_{ij} \leq 0, \quad i, j = 1, \ldots, N, \ i \neq j.$

3. All principal minors of A are non-negative.

4. $\det(A) > 0$.

There are a variety of equivalent definitions for M-matrix [Ple77], but the one that is most used in the literature for Convection-Diffusion-Reaction equations is,

Definition 3.11. (M-matrix) A matrix $A = (a_{ij})_{i,j=1}^{N}$ is an M-matrix if:

1. $a_{ij} \leq 0,\ i,j = 1,\ldots,N,\ i \neq j$.

2. $\det(A) > 0$.

3. $A^{-1} \geq 0$.

From the above definition it is clear that the set of M-matrices form a subset of monotone matrices. The following result gives a relation between an M-matrix and the DMP.

Theorem 3.12. (M-matrices and the DMP) *A discretization leading to an M-matrix that has the additional property*

$$\sum_{j=1}^{N} a_{ij} \geq 0, \qquad 1 \leq i \leq M,$$

gives a discrete solution that satisfies the DMP.

Proof. The statement follows from Theorem 3.8 and the fact that the set of M-matrices form a subset of the set of monotone matrices. □

3.2 Galerkin Approximation

The standard Galerkin method applied to (2.6) with homogeneous boundary conditions, replaces the infinite-dimensional space V by a finite-dimensional space V_h and then states : Find $u_h \in V_h$, such that for all $v_h \in V_h$

$$\varepsilon(\nabla u_h, \nabla v_h) + (\boldsymbol{b} \cdot \nabla u_h + c u_h, v_h) = \langle f, v_h \rangle. \tag{3.3}$$

Using appropriate regularity assumptions and under the condition

$$\left(c\,(x) - \frac{1}{2}\nabla \cdot \boldsymbol{b}\,(x)\right) \geq \sigma_0 > 0,$$

we have by Lemma of Céa[1] the error estimate as

$$\|u - u_h\|_V \leq C\frac{\max\{\|\boldsymbol{b}\|_\infty, \|c\|_\infty\}}{\varepsilon} \inf_{v_h \in V_h} \|u - v_h\|_V, \quad C \in \mathbb{R}.$$

In the convection-dominated case, i.e., $\varepsilon \ll \|\boldsymbol{b}\|_\infty$, the first factor of this estimate becomes very large and hence for the error estimate to be accurate we need the second factor, which is the best approximation error, to be small. On uniform grids, this best approximation error becomes very small only if the dimension of V_h is very large. The reason why the Galerkin method fails is that the solution possesses important scales that cannot be resolved by the grids. For convection-dominated problems such scales are layers that are present in the interior as well as the boundary and interior layers are of more importance from an application point of view. Hence, for solutions with sharp layers, the residue becomes very large. Many methods have been proposed for the stabilization of these discretizations, so as to get accurate results at the layers. The idea is to modify the Galerkin method by adding some sort of artificial diffusion to make it more stable and get improved results.

3.2.0.1 Conditions for M-matrix

Let us denote the matrix formulation of (3.3) by $Ax = b$ where A is the stiffness matrix. We want our solution to respect the DMP and for which according to Theorem 3.12 the sufficient condition is that the matrix A should correspond to an M-matrix. The Poisson problem is a special case for (2.6) when we only have the presence of diffusion. In [XZ99] geometrical conditions on the grid were introduced so as to make the stiffness matrix for the Poisson problem an M-Matrix.

For writing the necessary and sufficient conditions for the M-matrix property, let us introduce some geometrical notations for the triangulation K (see Fig. 3.2). Let K be a simplex with n number of vertices. Let us denote by,

- V_j : The vertices of K,

---

[1] **Lemma of Céa:** *Let $V_h \subset V$ and assume the conditions of Lax-Milgram Theorem are satisfied. Then there is a unique solution of the problem to find $u_h \in V_h$ such that*

$$a\left(u_h, v_h\right) = \langle f, v_h \rangle \quad \forall\ v_h \in V_h$$

and it holds the error estimate

$$\|u - u_h\|_V \leq \frac{M}{m} \inf_{v_h \in V_h} \|u - v_h\|_V,$$

where u is the unique solution of the continuous problem, m is the coercitivity constant, and M is the continuity constant of $a\left(\cdot, \cdot\right)$.

Proof. See [GT17, Lemma 2.1]. □

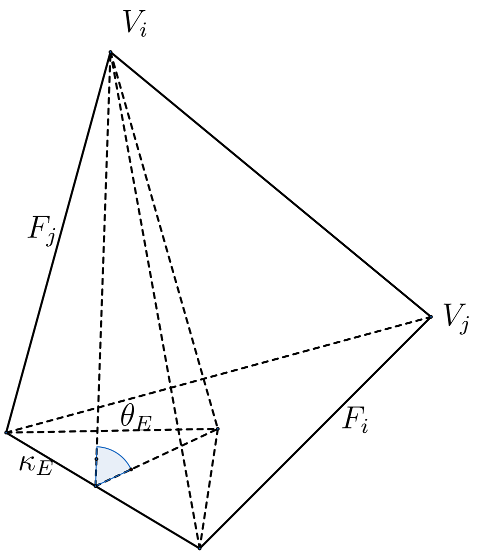


Figure 3.2: Geometric notations for a simplex K.

- E : The edge connecting two vertices,
- F_j : The $(n-1)$ dimensional simplex opposite vertex V_j,
- κ_E^K : $F_i \cap F_j$, the $(n-2)$-dimensional simplex opposite to the edge E,
- θ_E^K : The angle between faces F_i and F_j.

[XZ99, Lemma 2.1] gives the necessary and sufficient condition for the stiffness matrix of the Poisson problem to be an M-matrix.

Lemma 3.13. ([XZ99, Lemma 2.1]) *The stiffness matrix for the Poisson equation is an M-matrix if and only if for any fixed face edge E the following inequality holds:*

$$\omega_E \equiv \frac{1}{n(n-1)} \sum_{E\subset K} |\kappa_E^K| \cot \theta_E^K \geq 0, \tag{3.4}$$

where $\sum_{E\subset K}$ means summation over all simplices K containing E.

Remark 3.14. For $n = 2$, the condition (3.4) means that the sum of the angles opposite to any edge is less than or equal to π, i.e., if $K_1 \cap K_2 = \{E\}$ then $\theta_E^{K_1} + \theta_E^{K_2} \leq \pi$. This condition implies that the triangulation is a so-called Delaunay triangulation. It follows therefore that in $\mathbb{R}^2$ the stiffness matrix for the Poisson equation is an M-matrix if the triangulation is a Delaunay triangulation.

Example 3.15. Let us look at an example of (2.6) with $\boldsymbol{b} = (2,1)$, $c = 0$ and homogeneous Dirichlet boundary condition on $\Omega = (0,1)^2$. Then our problem reduces to

$$-\varepsilon\Delta u + 2u_x + u_y = f \quad \text{in} \quad \Omega = (0,1)^2, \quad u = 0 \quad \text{on} \quad \Gamma.$$

Using piece-wise linear elements on a uniform square mesh of Friedrichs-Keller type (i.e., the grid consists of three set of parallel lines) and denoting the mesh width by h after scaling, we get the difference stencil as

$$\frac{\varepsilon}{h^2}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} + \frac{1}{2h}\begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix}.$$

We see presence of positive off-diagonal matrix elements, so the sufficient condition of Theorem 3.12 in not satisfied and hence Galerkin method doesn't satisfy DMP.

3.3 Streamline-Upwind Petrov-Galerkin Method (SUPG)

Petrov-Galerkin methods are a subclass of Galerkin methods where the ansatz space (trial space) and the test space are not the same [Joh87]. Let A_h be the ansatz space and T_h be the test space with dim(A_h)=dim(T_h) then we have Petrov-Galerkin method as: Find $u_h \in A_h$ such that

$$a(u_h, v_h) = \langle f, v_h \rangle \qquad \forall v_h \in T_h.$$

The *Streamline-Upwind Petrov-Galerkin Method* (SUPG) or *Streamline-Diffusion FEM* (SDFEM) is one of the most commonly used methods for stabilization. Introduced in [BH82, HB79], the main idea of the SUPG is to add artificial diffusion in the streamline direction. This leads to a system of linear equations which can be easily solved.

A brief overview of the method is: Find $u_h \in V_h$, such that

$$a_h(u_h, v_h) = \langle f_h, v_h \rangle \qquad \forall v_h \in V_h \tag{3.5}$$

where

$$\begin{aligned} a_h(v, w) \; &:= \; a(v, w) \\ &\quad + \sum_{K \in T_h} \int_K \delta_K \left(-\varepsilon \Delta v(x) + \boldsymbol{b}(x) \cdot \nabla v(x) + c(x) v(x)\right) \left(\boldsymbol{b}(x) \cdot \nabla w(x)\right) dx, \\ \langle f_h, w \rangle \; &:= \; \langle f, w \rangle + \sum_{K \in T_h} \int_K \delta_K f(x) \left(\boldsymbol{b}(x) \cdot \nabla w(x)\right) dx. \end{aligned} \tag{3.6}$$

Here, $\{\delta_K\}$ are user-chosen weights, which are called stabilization parameters or SUPG parameters and $a(v, w)$ is the LHS of (3.3).

The method is called as SUPG because it can be considered as a Petrov-Galerkin method where the test space is given by

$$\operatorname{span}\left\{ w(x) + \sum_{K \in T_h} \delta_K \boldsymbol{b}(x) \cdot \nabla w(x) \right\}.$$

Lemma 3.16. ([RST08, Sec. 3.2.1]) *The SUPG method* (3.5)-(3.6) *is consistent.*

Proof. Let $u(x)$ be a sufficiently smooth solution of (2.6) then it satisfies the strong form of the equation point wise. Inserting this solution in (3.5)-(3.6) leads to vanishing of the stabilization term and hence our equations reduce to

$$a(u, v_h) = \langle f, v_h \rangle \qquad \forall\ v_h \in V_h,$$

which holds true by any weak solution as we have conforming finite elements (i.e. $V_h \subset V$).

Hence, we get

$$a_h(u, v_h) = \langle f_h, v_h \rangle \qquad \forall\ v_h \in V_h.$$

□

For showing the existence and uniqueness of the solution we first define the SUPG norm.

Definition 3.17. (SUPG norm) Let for almost all $x \in \Omega$ the following condition hold

$$\left(c(x) - \frac{1}{2} \nabla \cdot \boldsymbol{b}(x) \right) \geq \sigma_0 > 0. \tag{3.7}$$

For $v_h \in V_h$, the SUPG norm is defined by

$$\|v_h\|_{\text{SUPG}} := \left(\varepsilon |v_h|^2_{H^1(\Omega)} + \sigma_0 \|v_h\|^2_{L^2(\Omega)} + \sum_{K \in T_h} \|\delta_K^{1/2} (\boldsymbol{b} \cdot \nabla v_h)\|^2_{L^2(K)} \right)^{1/2}. \tag{3.8}$$

Theorem 3.18. (Coercitivity of the SUPG bilinear form) ([RST08, Lemma 3.25]) *Assume that* $\boldsymbol{b} \in W^{1,\infty}(\Omega)$, $c \in L^\infty(\Omega)$, *(3.7), and let*

$$0 < \delta_K \leq \frac{1}{2} \min \left\{ \frac{h_k^2}{\varepsilon C_{\text{inv}}^2}, \frac{\sigma_0}{\|c\|^2_{L^\infty(K)}} \right\}, \tag{3.9}$$

where C_{inv} *is the constant in the inverse estimate* (5.11)*. Then, the SUPG bilinear form is coercive with respect to the SUPG norm, i.e.,*

$$a_h(v_h, v_h) \geq \frac{1}{2} \|v_h\|^2_{\text{SUPG}} \qquad \forall\ v_h \in V_h.$$

Proof. By using integration by parts and the product rule we have,

$$\begin{aligned}\int_\Omega \boldsymbol{b}(x)\cdot\nabla v(x)\,v(x)\,dx &= -\int_\Omega \nabla\cdot(\boldsymbol{b}(x)\,v(x))\,v(x)\,dx\\ &= -\int_\Omega (\nabla\cdot\boldsymbol{b}(x))\,v(x)\,v(x)\,dx\\ &\quad -\int_\Omega \boldsymbol{b}(x)\cdot\nabla v(x)\,v(x)\,dx.\end{aligned}$$

Using the above equality we have

$$(\boldsymbol{b}\cdot\nabla v_h + cv_h, v_h) = \left(\left(-\frac{1}{2}\nabla\cdot\boldsymbol{b}+c\right)v_h, v_h\right) \qquad \forall\ v_h\in V_h.$$

With the definition of σ_0, we have

$$\begin{aligned}&a_h(v_h,v_h)\\ &= \varepsilon|v_h|_1^2 + \int_\Omega \underbrace{\left(c(x)-\frac{1}{2}\nabla\cdot\boldsymbol{b}(x)\right)}_{\geq\sigma_0>0}(v_h)^2(x)\,dx + \sum_{K\in T_h}\|\delta_k^{1/2}(\boldsymbol{b}\cdot\nabla v_h)\|_{L^2(K)}^2\\ &\quad+\sum_{K\in T_h}\int_K \delta_K(-\varepsilon\Delta v_h(x)+c(x)v_h(x))(\boldsymbol{b}(x)\cdot\nabla v_h(x))\,dx\\ &\geq \|v_h\|_{\text{SUPG}}^2 - \left|\sum_{K\in T_h}\int_K \delta_K(-\varepsilon\Delta v_h(x)+c(x)v_h(x))(\boldsymbol{b}(x)\cdot\nabla v_h(x))\,dx\right|.\end{aligned}$$

Now if we are able to approximate the last term from above then we get our result. Here we will use Cauchy-Schwarz inequality, the inverse estimate, Young's inequality, and condition (3.9) on the SUPG parameter, to get our result. For each $K\in T_h$

$$\begin{aligned}&\left|\int_K \delta_K(-\varepsilon\Delta v_h(x)+c(x)v_h(x))(\boldsymbol{b}(x)\cdot\nabla v_h(x))\,dx\right|\\ &\leq \int_K\left(\delta_K^{1/2}\varepsilon|\Delta v_h(x)|\right)\left(\delta_K^{1/2}|\boldsymbol{b}\cdot\nabla v_h(x)|\right)dx\\ &\quad+\int_K\left(\delta_K^{1/2}|c(x)||v_h(x)|\right)\left(\delta_K^{1/2}|\boldsymbol{b}\cdot\nabla v_h(x)\right)dx\\ &\leq \left(\delta_K^{1/2}\varepsilon\|\Delta v_h\|_{L^2(K)}+\delta_K^{1/2}\|c\|_{L^\infty(K)}\|v_h\|_{L^2(K)}\right)\left\|\delta_K^{1/2}(\boldsymbol{b}\cdot\nabla v_h)\right\|_{L^2(K)}\\ &\leq \left(\delta_K^{1/2}\frac{\varepsilon C_{\text{inv}}}{h_K}\|\nabla v_h\|_{L^2(K)}+\delta_K^{1/2}\|c\|_{L^\infty(K)}\|v_h\|_{L^2(K)}\right)\left\|\delta_K^{1/2}(\boldsymbol{b}\cdot\nabla v_h)\right\|_{L^2(K)}\end{aligned}$$

$$
\begin{aligned}
\leq\ & \left(\frac{h_k}{\sqrt{2\varepsilon}C_{\text{inv}}} \frac{\varepsilon C_{\text{inv}}}{h_K} \|\nabla v_h\|_{L^2(K)} + \frac{\sqrt{\sigma_0}}{\sqrt{2}\|c\|_{L^\infty(K)}} \|c\|_{L^\infty(K)} \|v_h\|_{L^2(K)} \right) \\
& \times \left\| \delta_K^{1/2} \left(\boldsymbol{b} \cdot \nabla v_h\right) \right\|_{L^2(K)} \\
=\ & \left(\sqrt{\frac{\varepsilon}{2}} \|\nabla v_h\|_{L^2(K)} + \sqrt{\frac{\sigma_0}{2}} \|v_h\|_{L^2(K)} \right) \| \left\| \delta_K^{1/2} \left(\boldsymbol{b} \cdot \nabla v_h\right) \right\|_{L^2(K)} \\
\leq\ & \frac{\varepsilon}{2} \|\nabla v_h\|^2_{L^2(K)} + \frac{1}{4} \left\| \delta_K^{1/2} (\boldsymbol{b} \cdot \nabla v_h) \right\|^2_{L^2(K)} + \frac{\sigma_0}{2} \|v_h\|^2_{L^2(K)} \\
& + \frac{1}{4} \left\| \delta_K^{1/2} \left(\boldsymbol{b} \cdot \nabla v_h\right) \right\|_{L^2(K)} \\
=\ & \frac{1}{2} \|v_h\|_{\text{SUPG},K}.
\end{aligned}
$$

Now, after summing over all the mesh cells and then subtracting the last estimate from the first estimate we get our result. □

Corollary 3.19. (Existence and Uniqueness of the solution of SUPG method) *Let the assumptions of Theorem 3.18 be valid. Then, the SUPG finite element method* (3.5)-(3.6) *has a unique solution.*

Proof. The corollary is proved using the Lax-Milgram theorem. We already have the coercitivity of bilinear form by Theorem 3.18 and hence we only have to show the boundedness of the bilinear form which can be shown in a similar way as the coercitivity of the SUPG method. □

Finally, we will mention the theorem which states the convergence of the SUPG method.

Theorem 3.20. (Convergence of the SUPG method) ([RST08, Theorem 3.27]) *Let the solution of* (2.6) *satisfy* $u \in H^{k+1}(\Omega),\ k \geq 1,$ *let* $\boldsymbol{b} \in W^{1,\infty}(\Omega),\ c \in L^\infty(\Omega)$, *let the assumptions of Theorem 3.18 be satisfied, and consider the SUPG method for* P_k *finite elements. Let the SUPG parameter be given as follows*

$$
\delta_K = \begin{cases} C_0 \frac{h_K^2}{\varepsilon} & \text{for} \quad \|\boldsymbol{b}\|_{L^\infty(\Omega)} h_K \leq \varepsilon, \\ C_0 h_k & \text{for} \quad \|\boldsymbol{b}\|_{L^\infty(\Omega)} h_K > \varepsilon, \end{cases} \tag{3.10}
$$

where the constant $C_0 > 0$ *is sufficiently small such that* (3.9) *is satisfied for* $k \geq 2$. *Then, the solution* $u_h \in \mathbb{P}_k$ *of the SUPG method* (3.5) *satisfies the following error estimate*

$$
\|u - u_h\|_{\text{SUPG}} \leq C \left(\varepsilon^{1/2} h^k + h^{k+1/2} \right) |u|_{H^{k+1}(\Omega)},
$$

where the constant C *is independent of* h *and* ε.

Proof. See [RST08, Theorem 3.27]. □

Remark 3.21. *Concerning the error estimate*

- In the convection dominated regime $\varepsilon \ll h$, the order of error reduction in the SUPG norm is $k + 1/2$ and in the diffusion-dominated case it is of order k.
- A fundamental problem in the application of the SUPG method is the choice of constant C_0 in the definition of the parameter (3.10). If C_0 is too large, then the layer is smeared, for an appropriate value of C_0 one obtains a solution which is almost exact in the nodes, and if C_0 is too small, then one can observe spurious (non-physical) oscillations in the layer. Hence, we obtain physically inconsistent results.

Remark 3.22. Let us look at the SUPG stabilization for the example presented in Example 3.15. With SUPG stabilization for convection-dominated problem our stiffness matrix stencil becomes,

$$\frac{\varepsilon}{h^2}\begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} + \frac{1}{2h}\begin{bmatrix} 0 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 0 \end{bmatrix} + \frac{C_0}{h}\begin{bmatrix} 0 & 1 & -2 \\ -2 & 6 & -2 \\ -2 & 1 & 0 \end{bmatrix},$$

for $\delta_K = C_0 h$.

We note that for small values of C_0 we have presence of positive off diagonal elements and hence sufficient condition for Theorem 3.12 is not satisfied and the method fails to satisfy the DMP.

To end this section we are going to briefly mention an improvement to the existing SUPG method.

3.3.1 Spurious Oscillations at Layers Diminishing Methods (SOLD)

One of the other used methods are *Spurious Oscillations at Layers Diminishing Methods* (SOLD). Because of the presence of oscillatory solutions, SUPG cannot be used to model physical systems. The idea of the SOLD scheme is to extend the SUPG method by introducing some numerical diffusion orthogonal to the streamline direction. To achieve higher-order methods the numerical diffusion has to depend on the finite element solution. Hence, we get a nonlinear term which leads to a nonlinear system of equations. The Mizukami-Hughes method introduced in [MH85] falls under the SOLD schemes. Results on the existence and uniqueness of the SOLD schemes can be found in [JK07a, JK08]. Most SOLD schemes reduce the oscillations but still the oscillations are considerably large and hence these methods also fail to give desired results.

3.4 Algebraic Flux Correction Schemes

After seeing some stabilization methods in the previous section we note that the ideal stabilization for convection dominated problems should possess the following properties:

- Satisfy the discrete maximum principle and hence, should have physically consistent results, i.e., no spurious oscillations.
- Gives accurate and sharp solutions near the layers.
- Provides an efficient solution for the system of equations obtained after the discretization.

We believe that the first property is of significance as it gives solutions that are accepted in practice. Algebraic Flux Correction (AFC) scheme proposed in [Kuz07] satisfies the first two properties. The idea of the AFC schemes is to add artificial diffusion to the algebraic system of equations and then limit that diffusion by using solution-dependent limiters. This method directly works on the system of equations rather on the variational formulation.

3.4.1 Derivation

Consider a linear boundary value problem for which the maximum principle holds. We can discretize the problem using a conforming finite element method. Then the discrete solution can be represented by a vector $u \in \mathbb{R}^N$ of its coefficients with respect to a basis of the respective finite element space. Let us assume that the last $N - M$ components of u $(0 < M < N)$ correspond to nodes where Dirichlet boundary conditions are prescribed, whereas the first M components of u are computed using the finite element discretization of the underlying partial differential equation. Using the Galerkin finite element discretizaion $\underline{u} \equiv (u_1, u_2, \ldots, u_N)$ satisfies a system of linear equations of the form

$$\sum_{j=1}^{N} a_{ij} u_j = g_i, \qquad i = 1, \ldots, M, \tag{3.11}$$

$$u_i = u_i^b, \qquad i = M+1, \ldots, N, \tag{3.12}$$

where, u_i^b are the Dirichlet boundary conditions.

We assume that the matrix $(a_{ij})_{i,j=1}^{M}$ is positive definite, i.e.,

$$\sum_{i,j=1}^{M} v_i a_{ij} v_j > 0 \qquad \forall (v_1, \ldots, v_M) \in \mathbb{R}^M \setminus \{0\}. \tag{3.13}$$

The starting point of the AFC schemes consists of modifying system (3.11) equivalently such that one gets formally a system with a different matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$. One idea is to simply use the finite element matrix corresponding to the above discretization in the case when homogeneous boundary conditions are used instead of Dirichlet conditions.

Using the matrix $\mathbb{A} = (a_{ij})_{i,j=1}^N$, we introduce a symmetric artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1}^N$, having the entries

$$d_{ij} = d_{ji} = -\max\{a_{ij}, 0, a_{ji}\} \qquad \forall i \neq j, \quad d_{ii} = -\sum_{i \neq j} d_{ij}. \tag{3.14}$$

This definition ensures that the matrix $\tilde{\mathbb{A}} := \mathbb{A} + \mathbb{D}$ has positive diagonal entries and non-positive off diagonal entries. If, in addition

$$\sum_{j=1}^{N} a_{ij} \geq 0, \qquad i = 1, \ldots, M, \tag{3.15}$$

then the matrix $\tilde{\mathbb{A}}$ satisfies sufficient conditions to preserve the discrete maximum principle. The property (3.15) is usually satisfied by finite element discretizations of elliptic equations arising in applications.

Going back to the solution of (3.11), this system is equivalent to

$$(\tilde{\mathbb{A}}\underline{u})_i = g_i + (\mathbb{D}\underline{u})_i, \qquad i = 1, \ldots, M. \tag{3.16}$$

Since the row sums of the matrix $\mathbb{D}$ vanish, it follows that

$$(\mathbb{D}\underline{u})_i = \sum_{i \neq j} f_{ij}, \qquad i = 1, \ldots, N,$$

where $f_{ij} = d_{ij}(u_j - u_i)$. Clearly it is $f_{ij} = -f_{ji}$ for all $i, j = 1, \ldots, N$.

Now, the idea of the AFC schemes is to limit those anti-diffusive fluxes f_{ij} that would otherwise cause spurious oscillations. To this end, system (3.11) (or, equivalently (3.16)) is replaced by

$$(\tilde{\mathbb{A}}\underline{u})_i = g_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \qquad i = 1, \ldots, M, \tag{3.17}$$

where $\alpha_{ij} \in [0, 1]$ are solution-dependent limiters. For $\alpha_{ij} = 1$, we move back to system (3.11). Hence, intuitively the coefficients α_{ij} should be as close to 1 as possible to limit the modification of the original discrete problem. They can be chosen in multiple ways but the idea is always based on using the fluxes f_{ij}.

For the scheme to be conservative and to show the existence of a solution, we require that the coefficients $\{\alpha_{ij}\}$ are symmetric, i.e.

$$\alpha_{ij} = \alpha_{ji}, \qquad i, j = 1, \dots, M. \tag{3.18}$$

Rewriting equation (3.17) using the definition of $\tilde{\mathbb{A}}$, we obtain the following nonlinear system of equations

$$\sum_{j=1}^{N} a_{ij} u_j + \sum_{j=1}^{N} (1 - \alpha_{ij}) d_{ij} (u_j - u_i) = g_i, \qquad i = 1, \dots, M, \tag{3.19}$$

$$u_i = u_i^b, \qquad i = M+1. \dots, N, \tag{3.20}$$

where $\alpha_{ij} = \alpha_{ij}(u_1, \dots, u_N) \in [0,1], i = 1, \dots, M, j = 1, \dots, N$ satisfy (3.18). A more detailed review of the AFC schemes can be found in [BJKR18].

Here, for the choice of limiters, we will present three different proposals but only two of them will be used for simulations.

3.4.2 Limiters

3.4.2.1 The Kuzmin Limiter

The first limiter we are going to consider is the Kuzmin limiter proposed in [Kuz06]. The idea of this limiter originates from [Zal79]. It starts by computing

$$P_i^+ = \sum_{\substack{j=1 \\ a_{ji} \le a_{ij}}}^{N} f_{ij}^+, \; P_i^- = \sum_{\substack{j=1 \\ a_{ji} \le a_{ij}}}^{N} f_{ij}^-, \; Q_i^+ = -\sum_{j=1}^{N} f_{ij}^-, \; Q_i^- = -\sum_{j=1}^{N} f_{ij}^+, \tag{3.21}$$

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

$$R_i^+ = \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \quad R_i^- = \min\left\{1, \frac{Q_i^-}{P_i^-}\right\}, \quad i = 1, \dots, M. \tag{3.22}$$

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

$$R_i^+ = 1, \quad R_i^- = 1, \quad i = M+1, \dots, N. \tag{3.23}$$

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

$$\alpha_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \\ 1 & \text{if } f_{ij} = 0 \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}, \quad \alpha_{ji} = \alpha_{ij}. \tag{3.24}$$

The Kuzmin limiter can be applied to P_1 and Q_1 finite elements, see [BJK16] for some details of its implementation. For P_1 finite elements, the satisfaction of the DMP for the solution of (3.19) is proved in [BJK16] with some restrictions on the grid. 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 [KK05].

3.4.2.2 The BJK Limiter

The second limiter we are going to study is the so-called BJK limiter proposed in [BJK17] for P_1 finite elements. This limiter makes the method linearity preserving, i.e. the modification added to the formulation (3.11) vanishes if the solution is a polynomial of degree 1. As first step, one defines for $i = 1, \ldots, N$

$$u_i^{\max} = \max_{j \in S_i \cup \{i\}} u_j, \quad u_i^{\min} = \min_{j \in S_i \cup \{i\}} u_j, \quad q_i = \gamma_i \sum_{j \in S_i} d_{ij}, \tag{3.25}$$

where γ_i is a positive constant computed according to Remark 3.23 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, i.e,

$$\{j \in \{1, \ldots, N\} \setminus \{i\} : a_{ij} \neq 0 \text{ or } a_{ji} > 0\} \subset S_i \subset \{1, \ldots, N\}.$$

As next step, one computes for $i = 1, \ldots, M$

$$P_i^+ = \sum_{j \in S_i} f_{ij}^+, \; P_i^- = \sum_{j \in S_i} f_{ij}^-, \; Q_i^+ = q_i(u_i - u_i^{\max}), \; Q_i^- = q_i(u_i - u_i^{\min}), \tag{3.26}$$

and then, one sets

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

If P_i^+ or P_i^- vanishes, one sets $R_i^+ = 1$ or $R_i^- = 1$, respectively. Then, (3.23) is applied for the Dirichlet nodes and the quantities

$$\bar{\alpha}_{ij} = \begin{cases} R_i^+ & \text{if } f_{ij} > 0 \\ 1 & \text{if } f_{ij} = 0 \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}, \quad i = 1, \ldots, M,\ j = 1, \ldots, N, \tag{3.27}$$

are calculated. Finally, one sets

$$\alpha_{ij} = \min\{\bar{\alpha}_{ij}, \bar{\alpha}_{ji}\}, \quad i, j = 1, \ldots, M, \tag{3.28}$$

$$\alpha_{ij} = \bar{\alpha}_{ij}, \quad i = 1, \ldots, M,\ j = M+1, \ldots, N. \tag{3.29}$$

It is proved in [BJK17] that the corresponding solution of the AFC method (3.19) satisfies the DMP and it is linearity preserving on arbitrary simplicial grids. The uniqueness of the solution and the study of mixed boundary conditions are open questions.

Remark 3.23. (Computation of γ_i) Let $\Delta_i = \operatorname{supp} \varphi_i$ for any interior vertex x_i and let Δ_i^{conv} be its convex hull. Define

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial\Delta_i^{\text{conv}})}, \quad i = 1, \ldots, M,$$

then we have linearity preservation for our AFC scheme [BJK17, Theorem 6.1]. We note that we require the geometrical information of the grid and that's why we can't regard the BJK limiter as a purely algebraic approach.

3.4.2.3 The BBK Limiter

The next limiter we are going to present is the BBK limiter introduced in [BBK17]. This is also referred to as smoothness-based viscosity and had its origin in the finite volume literature. The first difference from other limiters is in the definition of d_{ij}. Here, $d_{ij} = \gamma_0 h_{ij}^{d-1}$ where γ_0 is a fixed parameter depending on data of (2.6). The limiters α_E, for $E \in \mathcal{E}_h$ are given by the following algorithm: for $v_h \in V_h$, one defines ξ_{v_h} as the unique element in V_h whose nodal values are given by

$$\xi_{v_h}(x_i) := \begin{cases} \frac{|\sum_{j \in S_i} (v_h(x_i) - v_h(x_j))|}{\sum_{j \in S_i} |v_h(x_i) - v_h(x_j)|}, & \text{if } \sum\limits_{j \in S_i} |v_h(x_i) - v_h(x_j)| \neq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{3.30}$$

Then, on each edge $E \in \mathcal{E}_h$, α_E are defined by

$$\alpha_E(v_h) := 1 - \max_{x \in E} [\xi_{v_h}(x)]^p, \quad p \in [1, +\infty). \tag{3.31}$$

The value of p determines the rate of decay of the numerical diffusion with the distance to the critical point. For p closer to 1, it adds more diffusion in the far field, while a larger value makes the diffusion vanish faster, but on the other hand, larger value of p makes the nonlinear system difficult to solve.

Remark 3.24. Results with respect to the BBK limiter are not presented in this thesis. Analytical results with respect to this limiter can be found in [BBK17]. For a detailed comparison of the limiters, we refer to [BJKR18].

Remark 3.25. There is one more limiter in the literature which is of upwind type proposed in [Kno19]. It is a linearity preserving limiter which satisfies the DMP on arbitrary meshes. It combines the advantage of the Kuzmin limiter and the BJK limiter. Results with this limiter will not be presented in this thesis.

3.4.3 Review of Analytical Results

In this section, we will mention some results on the existence and uniqueness of the solution of nonlinear problem (3.19) and (3.20). We will also mention results which show that the AFC schemes satisfy the discrete maximum principle (DMP).

First, a lemma for the continuity of $\Phi_{ij}(\underline{u}) := \alpha_{ij}(\underline{u})(u_j - u_i)$ is stated. The sufficient condition for the continuity of Φ_{ij} is given by,

Lemma 3.26. ([BJK16, Lemma 6, Sec. 3]) *Consider any $i, j \in \{1, \cdots, N\}$, and let α_{ij} $: \mathbb{R}^N \to [0,1]$ satisfy*

$$\alpha_{ij}(\underline{u}) = \frac{A_{ij}(\underline{u})}{|u_j - u_i| + B_{ij}(\underline{u})} \quad \forall\ u \equiv (u_1, \cdots, u_N) \in \mathbb{R}^N, u_i \neq u_j, \tag{3.32}$$

where $A_{ij}, B_{ij} : \mathbb{R}^N \to [0, \infty)$ are non negative functions that are continuous at any point $\underline{u} \in \mathbb{R}^N$ with $u_i \neq u_j$. Then $\Phi_{ij}(\underline{u}) := \alpha_{ij}(\underline{u})(u_j - u_i)$ is a continuous function of $u_1, \ldots, u_N$ on $\mathbb{R}^N$. Moreover, if the functions A_{ij}, B_{ij} are Lipschitz-continuous with the constant C_L in the sets $\{u \in \mathbb{R}^N; u_i < u_j\}$ and $\{\underline{u} \in \mathbb{R}^N; u_i > u_j\}$, then the functions Φ_{ij} is Lipschitz-continuous on $\mathbb{R}^N$, with the constant $2C_L + \sqrt{2}$.

Proof. First we will show the continuity of the functions $\Phi_{ij}(\underline{u})$. Let $\underline{\tilde{u}} \equiv (\tilde{u}_1, \ldots, \tilde{u}_N) \in \mathbb{R}^N$.

If $\tilde{u}_i \neq \tilde{u}_j$ then $\exists$ a neighborhood $V_{\underline{\tilde{u}}}$ of $\underline{\tilde{u}}$ such that the denominator (3.32) does not vanish in $V_{\underline{\tilde{u}}}$ and as A_{ij}, B_{ij} are continuous we get α_{ij} is continuous at $\underline{\tilde{u}}$.

If $\tilde{u}_i = \tilde{u}_j$ then as $\alpha_{ij} \in [0,1]$ we get,

$$\begin{aligned} |\alpha_{ij}(u)(\underline{u}_j - u_i)| &\leq |u_j - u_i| \\ &\leq |u_j - \tilde{u}_j| + |u_i - \tilde{u}_i| \end{aligned}$$

$$\leq \sqrt{2}\|\underline{u}-\underline{\tilde{u}}\|_{l^2}$$

for any $\underline{u} \equiv (u_1, \dots, u_N) \in \mathbb{R}^N$ and $\|\cdot\|_{l^2}$ defines the Euclidean norm in $\mathbb{R}^N$. Therefore, α_{ij} $(u_j - u_i)$ is continuous at $\underline{u}$.

For Lipschitz-continuity of Φ_{ij}, let $\underline{u}, \underline{\tilde{u}} \in \mathbb{R}^N$. Set $v = u_j - u_i$ and $\tilde{v} = \tilde{u}_j - \tilde{u}_i$.
If $v\tilde{v} \leq 0$ then

$$\begin{aligned}
|\Phi_{ij}(\underline{u}) - \Phi_{ij}(\underline{\tilde{u}})| &= |\alpha_{ij}(\underline{u})(u_j - u_i) - \alpha_{ij}(\underline{\tilde{u}})(\tilde{u}_j - \tilde{u}_i)| \\
&\leq |v| + |\tilde{v}|,
\end{aligned}$$

and

$$\begin{aligned}
(|v| + |\tilde{v}|)^2 &= |v|^2 + |\tilde{v}|^2 + 2|v||\tilde{v}| \\
&= v^2 + \tilde{v}^2 - 2v\tilde{v} \\
&= |v - \tilde{v}|^2.
\end{aligned}$$

Hence,

$$|\Phi_{ij}(\underline{u}) - \Phi_{ij}(\underline{\tilde{u}})| \leq |v - \tilde{v}|.$$

If $v\tilde{v} > 0$, then

$$\begin{aligned}
\Phi_{ij}(\underline{u}) - \Phi_{ij}(\underline{\tilde{u}}) &= \frac{A_{ij}(\underline{u})v}{|v| + B_{ij}(\underline{u})} - \frac{A_{ij}(\underline{\tilde{u}})\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} + \frac{A_{ij}(\underline{u})\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} - \frac{A_{ij}(\underline{u})\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} \\
&= (A_{ij}(\underline{u}) - A_{ij}(\underline{\tilde{u}}))\frac{\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} + A_{ij}(\underline{u})\left[\frac{v}{|v| + B_{ij}(\underline{u})} - \frac{\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})}\right] \\
&= (A_{ij}(\underline{u}) - A_{ij}(\underline{\tilde{u}}))\frac{\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} + \frac{A_{ij}(\underline{u})}{|v| + B_{ij}(\underline{u})}\left[v - \frac{\tilde{v}\,(|v| + B_{ij}(\underline{u}))}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})}\right] \\
&= (A_{ij}(\underline{u}) - A_{ij}(\underline{\tilde{u}}))\frac{\tilde{v}}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})} \\
&\quad + \frac{\alpha_{ij}(\underline{u})}{|\tilde{v}| + B_{ij}(\underline{\tilde{u}})}\left[(B_{ij}(\underline{\tilde{u}}) - B_{ij}(\underline{u}))\,\tilde{v} + (v - \tilde{v})B_{ij}(\underline{\tilde{u}})\right].
\end{aligned}$$

Therefore, we get

$$\begin{aligned}
|\Phi_{ij}(\underline{u}) - \Phi_{ij}(\underline{\tilde{u}}) &\leq |A_{ij}(\underline{u}) - A_{ij}(\underline{\tilde{u}})| + |B_{ij}(\underline{u}) - B_{ij}(\underline{\tilde{u}})| + |v - \tilde{v}| \\
&\leq (2C_L + \sqrt{2})\|\underline{u} - \underline{\tilde{u}}\|_{l^2}.
\end{aligned}$$

Hence $\Phi_{ij}(\underline{u})$ is Lipschitz-continuous. □

Using this lemma we can state the existence result for the system (3.19), (3.20).

Theorem 3.27. ([BJK16, Theorem 3, Sec. 3]) *Let* (3.15) *hold. For any* $i, j \in \{1, \dots, N\}$, *let* $\alpha_{ij} : \mathbb{R}^N \to [0, 1]$ *be such that* $\alpha_{ij}\ (u_1, \dots, u_N)(u_j - u_i)$ *is a continuous function of* $u_1, \dots, u_N$. *Finally, let the functions* α_{ij} *satisfy* (3.18). *Then there exists a solution of the nonlinear problem* (3.19), (3.20).

Proof. Let us denote by $\tilde{\underline{v}} \equiv (v_1, \dots, v_M)$ as an element of $\mathbb{R}^M$, and $v_i = u_i^b$ if $i \in \{M + 1, \dots, N\}$.

For $\tilde{\underline{v}} \in \mathbb{R}^M$ we set $\underline{v} := (v_1, \dots, v_N)$ and $\underline{G} = (g_1, \dots, g_M)$.
Let $C_M = \inf\limits_{\|\tilde{\underline{v}}\|_{l^2}=1} \sum_{i,j=1}^{M} v_i a_{ij} v_j > 0$ which holds by (3.13). Then we can write

$$\sum_{i,j=1}^{M} v_i a_{ij} v_j \geq C_M \|\tilde{\underline{v}}\|_{l^2}^2 \qquad \forall\ \tilde{\underline{v}} \in \mathbb{R}^M. \tag{3.33}$$

Let us define the operator $T : \mathbb{R}^M \to \mathbb{R}^M$ by

$$(T\tilde{\underline{v}})_i = \sum_{j=1}^{N} a_{ij} v_j + \sum_{j=1}^{N} [1 - \alpha_{ij}(\underline{v})]\, d_{ij}(v_j - v_i) - g_i, \qquad i = 1, \dots, M.$$

Then $\underline{u}$ is a solution of (3.19), (3.20) if and only if $T\tilde{\underline{u}} = 0$. The operator T is continuous and we get

$$\begin{aligned} (T\tilde{\underline{v}}, \tilde{\underline{v}}) \ &= \ \sum_{i,j=1}^{M} v_i a_{ij} v_j + \sum_{i=1}^{M} \sum_{j=M+1}^{N} v_i a_{ij} u_i^b + \sum_{i=1}^{M} \sum_{j=1}^{M} v_i\, [1 - \alpha_{ij}(\underline{v})]\, d_{ij}(v_j - v_i) \\ &\quad - (G, \tilde{\underline{v}}) + \sum_{i=1}^{M} \sum_{j=M+1}^{N} v_i\, [1 - \alpha_{ij}(\underline{v})]\, d_{ij}(u_j^b - v_i). \end{aligned}$$

Using (3.33), Hölder's inequality and Lemma A [2] we get

$$\begin{aligned} (T\tilde{\underline{v}}, \tilde{\underline{v}}) \ &\geq \ C_M \|\tilde{\underline{v}}\|_{l^2}^2 - 0 - C_1 \|\tilde{\underline{v}}\|_{l^2} - C_0 \\ &\geq \ \frac{C_M}{2} \|\tilde{\underline{v}}\|_{l^2}^2 - C_2, \end{aligned}$$

where the last inequality comes from Young's inequality, and C_0, C_1, and C_2 are positive

---

[2] **Lemma A:** *Consider any* $\mu_{ij} = \mu_{ji} \leq 0,\ i, j = 1, \dots, N$. *Then*

$$\sum_{i,j=1}^{N} v_i \mu_{ij}(v_j - v_i) = -\sum_{\substack{i,j=1 \\ i<j}}^{N} \mu_{ij}(v_i - v_j)^2 \geq 0 \quad \forall v_1, \dots, v_N \in \mathbb{R}.$$

Proof. See Lemma 1 in [BJK16]. □

constants independent of $\tilde{\underline{v}}$. Then for any $\tilde{\underline{v}} \in \mathbb{R}^M$ satisfying $\|\tilde{\underline{v}}\|_{l^2} = (3C_2/C_M)^{1/2}$, we have $(T\tilde{\underline{v}}, \tilde{\underline{v}}) > 0$ and hence by Lemma B[3] $\exists\ \tilde{\underline{u}} \in \mathbb{R}^M$ such that $T\tilde{\underline{u}} = 0$. ☐

The following corollary gives the uniqueness of the linearized problem,

Corollary 3.28. *Let* (3.15) *hold. Consider any fixed* $\alpha_{ij} \in [0,1]$, $i,j = 1,\dots,N$, *satisfying* (3.18). *Then the system* (3.19), (3.20) *has a unique solution for any* $g_1,\dots,g_M \in \mathbb{R}$ *and* $u^b_{M+1},\dots,u^b_N \in \mathbb{R}$.

Proof. According to Theorem 3.27, for any values of $g_1,\dots,g_M$ and $u^b_{M+1},\dots,u^b_N$, there exists a solution of the considered linear system. Consequently, the solutions have to be unique. ☐

Lemma 7 from [BJK16] and Lemma 4.2 from [BJK17] shows that the Kuzmin limiter and the BJK limiter satisfy Lemma 3.26 and hence we get the existence of solution for the nonlinear problem. If we linearize the problem then we have the uniqueness of the solution as well.

Finally, we want our discrete solution to satisfy the DMP, which is stated as,

Theorem 3.29. *Consider any* $i \in \{1,\dots,M\}$. *Then*

$$g_i \le 0 \Rightarrow\ u_i \le \max_{j\neq i,\ a_{ij}\neq 0} u_j \ for\ u_i \ge 0\ \Rightarrow u_i \le \max_{j\neq i,\ a_{ij}\neq 0} u_j^+, \tag{3.34}$$

$$g_i \ge 0 \Rightarrow\ u_i \ge \min_{j\neq i,\ a_{ij}\neq 0} u_j \ for\ u_i \le 0\ \Rightarrow u_i \ge \min_{j\neq i,\ a_{ij}\neq 0} u_j^-. \tag{3.35}$$

Proof. See Corollary 11 in Ref. [BJK16] for Kuzmin limiter. ☐

With some more restrictions on Theorem 3.29, we get that the solution obtained with the BJK limiter also satisfies the DMP ([BJK17, Theorem 3.1]).

Finally, we mention the error analysis of the AFC schemes proved in [BJK16].

Theorem 3.30. ([BJK16, Corollary 17]) *Let* $u \in H^2(\Omega)$ *be a solution of* (2.6), *and let* $u_h \in W_h$ *be a solution of the discrete problem*

$$a_h(u_h, v_h) + d_h(u_h; u_h, v_h) = \langle f, v_h\rangle \qquad \forall v_h \in V_h, \tag{3.36}$$

---

[3] **Lemma B:** *Let* X *be a finite-dimensional Hilbert space with inner product* $(\cdot,\cdot)_X$ *and norm* $\|\cdot\|_X$. *Let* $T : X \to X$ *be a continuous mapping, and let* $K > 0$ *be a real number such that* $(Tx, x)_X > 0$ *for any* $x \in X$ *with* $\|x\|_X = K$. *Then there exists* $x \in X$ *such that* $\|x\|_X < K$ *and* $Tx = 0$.

Proof. See Lemma 1.4, p.164 in [Tem77]. ☐

where $a_h(\cdot,\cdot)$ *is the approximate bilinear form of* $a(\cdot,\cdot)$*, the bilinear form of convection-diffusion equations,* $W_h \subset C(\overline{\Omega}) \cap H^1(\Omega)$*,* $V_h := W_h \cap H_0^1(\Omega)$*, and*

$$d_h(u_h; u_h, v_h) = \sum_{i,j=1}^{N} (1 - \alpha_i j(u_h)) d_{ij}(u_j - u_i) v_i .$$

Then if $\sigma_0 > 0$*, there exists a constant* $C > 0$ *independent of* h *and the data of* (2.6) *such that*

$$\begin{aligned} \|u - u_h\|_{\mathrm{AFC}} \quad \leq \quad & C \left(\varepsilon \sigma_0^{-1} \left\{ \|\boldsymbol{b}\|_{0,\infty,\Omega}^2 + \|c\|_{0,\infty,\Omega}^2 \right\} + \sigma_0 h^2 \right)^{1/2} h \|u\|_{2,\Omega} \\ & + C \left(\varepsilon + \|\boldsymbol{b}\|_{0,\infty,\Omega} h \right)^{1/2} |i_h u|_{1,\Omega}, \end{aligned} \tag{3.37}$$

where $\|\cdot\|_{\mathrm{AFC}}$ *is the natural norm on* V_h *corresponding to the left-hand side of* (3.36) *and* $i_h u$ *is the Lagrange interpolation operator.*

Proof. See [BJK16, Lemma 13, 15, 16]. □

Remark 3.31. The error estimate in Theorem 3.30 shows that the order of convergence depends on the relation of ε and $\|\boldsymbol{b}\|_{0,\infty,\Omega} h$ and on the geometrical properties of the triangulations. For a convection dominated regime, i.e., $\varepsilon < \|\boldsymbol{b}\|_{0,\infty,\Omega} h$ we have $\mathcal{O}(h^{1/2})$ for any choice of limiters. From [BJK16, Remark 18] one also gets improved order if the triangulation contains all acute triangles (all angles $< \pi/2$). But if the triangulation is not of Delaunay type, i.e., it contains some obtuse triangles (i.e., an angle $> \pi/2$), then we lose convergence of the method for the Kuzmin limiter.

But this method also has its drawbacks. Because of the presence of α_{ij} we get a system of nonlinear equations even for linear PDEs, hence they are difficult to solve. But this problem is not of significance as the PDEs that we encounter are in general nonlinear. The second drawback of the AFC schemes is that they have been developed only for lowest order finite elements, which limits the accuracy of the numerical solutions. It should be noted that an extension to P_2 elements has been developed in [Kuz08], but it was mentioned that the extension gave rise to many new challenging open problems.

3.5 Summary

This chapter introduced the notion of stabilization for finite element methods. A discrete analog of the maximum principle was introduced. Necessary and sufficient conditions were given for the stabilization methods to satisfy DMP using the notion of an M-matrix. The most commonly used FEM, namely the Galerkin method was studied and it was found that the method fails to give solutions that satisfy the DMP because of the presence of positive off-diagonal entries in the stiffness matrix A.

Results regarding improvement over the Galerkin method were presented. The most commonly used finite element stabilization method namely the SUPG method was studied. The existence and uniqueness of the solution were proved under the assumption (3.7). With appropriate choice of the parameters $\{\delta_K\}$, it's shown that we get satisfactory results. But the choice of $\{\delta_K\}$ was not easy. The section ended with mentioning the drawbacks of the scheme and an improvement over it namely the SOLD schemes were defined.

Hence, we noted that an ideal stabilization of FEM must satisfy the DMP and compute steep layers. Lastly, the Algebraic Flux Correction scheme was introduced. The construction of the AFC scheme allowed it to satisfy the DMP and approximate the solution near the layers properly. Three different definitions of the limiters were stated out of which the BJK limiter makes the scheme linearity preserving. Under the assumption of [XZ99], results for the existence of the solution and the satisfaction of the DMP were proved. As the AFC solutions do not possess spurious oscillations we prefer this method over other methods. But the nonlinearity nature of the scheme produces new challenges, namely the efficient solvers for the scheme. Chapter 4 deals with different solvers for the AFC scheme and some optimization tools to reduce the number of iterations.

4 Iterative Solvers for Steady-State Convection-Diffusion-Reaction Equation

In Chapter 3 we introduced stabilized finite element methods and it was noted that one of the issues for stabilization techniques is the efficient solution of the system of equations. This issue is prominent for nonlinear schemes such as Algebraic Flux Correction and Mizukami-Hughes method [MH85]. It was also noted in [BJKR18] and [JK07a] that non-efficient solution of the system of equations is one of the drawbacks of these methods. This chapter will discuss this issue and present a comprehensive study for the solution of the nonlinear problem of the AFC schemes.

The contents of the chapter are as follows: Sec. 4.1 introduces the various iterative schemes that will be considered. Sec. 4.2 will introduce some tools such as Anderson acceleration, dynamic damping, etc, which could be beneficial while solving the system. Numerical studies are presented in Sec. 4.3 with various examples from 2d as well as 3d. And lastly, all the results are summarized in Sec. 4.4.

4.1 Iterative Schemes

Definition 4.1. Let S be a non-empty closed subset of $\mathbb{R}^N$ and $f : S(\subset \mathbb{R}^N) \to \mathbb{R}^N$ be a continuous function defined on S. We need to find a $\underline{\zeta} \in S$ such that $f(\underline{\zeta}) = 0$. We try to transform the above problem into an equivalent problem of the form $g_{\text{ite}}(\underline{x}) = \underline{x}$, where $g_{\text{ite}} : \mathbb{R}^N \to \mathbb{R}^N$ is a continuous function. More in general, we choose

$$g_{\text{ite}}(\underline{x}) = \underline{x} - h(\underline{x})f(\underline{x}),$$

where $h(\underline{x}) \neq 0$ is a damping parameter.

We compute the next iterate by replacing $g_{\text{ite}}(\underline{x}^{(\nu)})$ with $\underline{x}^{(\nu+1)}$ and hence, we get

$$\underline{x}^{(\nu+1)} = \underline{x}^{(\nu)} - h(\underline{x}^{(\nu)})f(\underline{x}^{(\nu)}).$$

Depending on how we choose $h(\underline{x}^{(\nu)})$ we get different iteration schemes.

Remark 4.2. The iterative schemes that we are going to discuss relies on the same ideology of finding a fixed point and then solving the system. The methods considered here have already been outlined in [BJKR18, Sec. 5].

Remark 4.3. As, we are dealing with system of equations, we know that $f\left(\underline{x}^{(\nu)}\right)$ will be of the form $M_{\text{ite}}^{-1}\tilde{f}(\underline{x}^{(\nu)})$, where $M_{\text{ite}} \in \mathbb{R}^{N\times N}$ is an invertible matrix. Hence, our iteration can be written as

$$\underline{x}^{(\nu+1)} = \underline{x}^{(\nu)} - \omega^{(\nu)} M_{\text{ite}}^{-1}\tilde{f}(\underline{x}^{(\nu)}). \tag{4.1}$$

Consider the nonlinear problem (3.19), (3.20) in the form

$$\begin{aligned} F(\underline{u}) &= 0 \quad \text{with} && (4.2)\\ F_i(\underline{u}) &= \sum_{j=1}^{N} a_{ij}u_j + \sum_{j=1}^{N}(1-\alpha_{ij}(\underline{u}))d_{ij}(u_j - u_i) - f_i = 0, \quad i = 1,\ldots,M, \\ F_i(\underline{u}) &= u_i - u_i^b = 0, \quad i = M+1,\ldots,N. \end{aligned}$$

Then, a damped iteration for solving (4.2) is given by

$$\underline{u}^{(\nu+1)} = \underline{u}^{(\nu)} - \omega^{(\nu)} M_{\text{ite}}^{-1} F\left(\underline{u}^{(\nu)}\right), \quad \nu = 0,1,\ldots, \tag{4.3}$$

where $M_{\text{ite}} \in \mathbb{R}^{N\times N}$ is a non-singular matrix. A vector $\underline{u}$ is a solution of the nonlinear problem (3.20) if and only if it is a fixed point of (4.3). The choice of the damping parameter $\omega^{(\nu)}$ is briefly discussed in Sec. 4.2.1.

4.1.1 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 (4.2). A straightforward idea consists in using for the construction of the left-hand side of (4.2) the currently available values for the limiter, leading in the iteration step $(\nu+1)$ to a linear system of equations of the form

$$\begin{aligned} \sum_{j=1}^{N} a_{ij}u_j^{(\nu+1)} + \sum_{j=1}^{N}\left(1-\alpha_{ij}^{(\nu)}\right)d_{ij}\left(u_j^{(\nu+1)} - u_i^{(\nu+1)}\right) &= f_i, \quad i = 1,\ldots,M, \\ u_i^{(\nu+1)} &= u_i^b, \quad i = M+1,\ldots,N, \end{aligned} \tag{4.4}$$

with $\alpha_{ij}^{(\nu)} = \alpha_{ij}\left(\underline{u}^{(\nu)}\right)$. This method is called *fixed point matrix*. From Corollary 3.28, Chapter 3, it is shown that in the case of Dirichlet boundary conditions, the linear system (4.4) 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 $\mathbb{D}$ 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

$$\begin{aligned}\sum_{j=1}^{N}(a_{ij}+d_{ij})u_j^{(\nu+1)} &= f_i+\sum_{j=1}^{N}\alpha_{ij}^{(\nu)}f_{ij}^{(\nu)}, \quad i=1,\ldots,M,\\ u_i^{(\nu+1)} &= u_i^b, \qquad\qquad\qquad\quad i=M+1,\ldots,N,\end{aligned}\tag{4.5}$$

where $f_{ij}^{(\nu)}$ is the flux computed with the limiter $\alpha_{ij}^{(\nu)}$. We refer to this method as *fixed point rhs*. A distinct feature of *fixed point rhs* is that the matrix $\mathbb{A}+\mathbb{D}=\tilde{\mathbb{A}}$ 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.

The numerical studies in Sec. 4.3 also consider 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 was that the method *fixed point matrix* might need in general 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. In addition to methods (4.4) and (4.5), we define the mixed fixed point iteration

$$\begin{aligned}&\sum_{j=1}^{N}\left(a_{ij}+d_{ij}\right)u_j^{(\nu+1)}-\omega_{\mathrm{fp}}\sum_{j=1}^{N}\alpha_{ij}^{(\nu)}d_{ij}\left(u_j^{(\nu+1)}-u_i^{(\nu+1)}\right)\\ &\quad= f_i+(1-\omega_{\mathrm{fp}})\sum_{j=1}^{N}\alpha_{ij}^{(\nu)}f_{ij}^{(\nu)}, \quad i=1,\ldots,M,\\ u_i^{(\nu+1)} &\quad= u_i^b, \qquad\qquad\qquad\qquad\qquad i=M+1,\ldots,N,\end{aligned}\tag{4.6}$$

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 (4.3), method (4.6) uses the matrix M_{ite} with

$$M_{\text{ite}}\left(u^{(\nu)}\right)_{ij} = \begin{cases} a_{ij} + d_{ij} - \omega_{\text{fp}}\alpha_{ij}^{(\nu)} d_{ij} & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \omega_{\text{fp}} \sum\limits_{j=1, j\neq i}^{N} \alpha_{ij}^{(\nu)} d_{ij} & \text{if } i = j, \end{cases}$$

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 (4.6) are presented in Sec. 4.3, Examples 4.3.1.7, 4.3.2.1, and 4.3.2.4.

4.1.2 A Formal Newton Method

This section presents a formal Newton method for solving (4.2). 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.

4.1.2.1 Derivation

For Newton's method, the matrix M_{ite} in (4.3) is the Jacobian of F. Considering (4.2) for $i = 1, \ldots, M$, one can compute the Jacobian formally, using standard calculus, as

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

Hence, the entries of the matrix that has to be inverted in (4.3) are given by

$$M_{\text{ite}}\left(\underline{u}^{(\nu)}\right)_{ij} = DF\left(\underline{u}^{(\nu)}\right)_{ij}$$

$$= \begin{cases} a_{ij} + d_{ij} - \alpha_{ij}^{(\nu)} d_{ij} - \sum_{k=1}^{N} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_j} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) & \text{if } i \neq j, \\ a_{ii} + d_{ii} + \sum_{k=1, k\neq i}^{N} \alpha_{ik}^{(\nu)} d_{ik} - \sum_{k=1}^{N} \frac{\partial \alpha_{ik}^{(\nu)}}{\partial u_i} d_{ik} \left(u_k^{(\nu)} - u_i^{(\nu)} \right) & \text{if } i = j, \end{cases} \tag{4.7}$$

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 (4.7) 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 (4.7) of the Jacobian are given as follows

$$M_{\text{ite}} \left(\underline{u}^{(\nu)} \right)_{ij} = DF \left(\underline{u}^{(\nu)} \right)_{ij} = a_{ij} + d_{ij} - \sum_{k=1}^{N} \frac{\partial \beta_{ik}^{(\nu)}}{\partial u_j} \tag{4.8}$$

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

4.1.2.2 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.

4.1.2.2.1 Approach with separate treatment of the non-smooth points This approach uses the representation (4.7) 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 (3.21) and the definition of f_{ik}, one obtains, e.g.,

$$\begin{aligned} \frac{\partial Q_i^+}{\partial u_j} &= -\frac{\partial}{\partial u_j} \sum_{l=1}^{N} f_{il}^- = -\frac{\partial}{\partial u_j} \sum_{l=1}^{N} \min\left\{0, d_{il}(u_l - u_i)\right\}, \\ &= \begin{cases} 0 & \text{if } f_{ij} \geq 0, i \neq j, \\ -d_{ij} & \text{if } f_{ij} < 0, i \neq j, \\ \displaystyle\sum_{l=1, f_{il}<0}^{N} d_{il} & \text{if } i = j, \end{cases} \end{aligned} \tag{4.9}$$

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 (3.24). 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.

4.1.2.2.2 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 [BB17]

$$\max_\sigma(x,y) = \frac{1}{2}\left(x + y + \sqrt{(x-y)^2+\sigma}\right) \tag{4.10}$$

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 (4.8) of the Jacobian is utilized. In the case $a_{ki} \le a_{ik}$, 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

$$\tilde{\beta}_{ik} = \min_\sigma\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}, 1\right)\max_\sigma(f_{ik},0) + \min_\sigma\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}, 1\right)\min_\sigma(f_{ik},0). \tag{4.11}$$

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

$$\begin{aligned}
&\frac{\partial\tilde{\beta}_{ik}}{\partial u_j}\\
&= \frac{1}{2}\left(1 - \frac{\tilde{Q}_i^+/\tilde{P}_i^+ - 1}{\sqrt{\left(\tilde{Q}_i^+/\tilde{P}_i^+ - 1\right)^2 + \sigma}}\right)\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}\right)\frac{1}{2}\left(f_{ik} + \sqrt{f_{ik}^2 + \sigma}\right)\\
&\quad + \frac{1}{2}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+} + 1 - \sqrt{\left(\tilde{Q}_i^+/\tilde{P}_i^+ - 1\right)^2 + \sigma}\right)\frac{1}{2}\left(1 + \frac{f_{ik}}{\sqrt{f_{ik}^2+\sigma}}\right)\frac{\partial f_{ik}}{\partial u_j}\\
&\quad + \frac{1}{2}\left(1 - \frac{\tilde{Q}_i^-/\tilde{P}_i^- - 1}{\sqrt{\left(\tilde{Q}_i^-/\tilde{P}_i^- - 1\right)^2 + \sigma}}\right)\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}\right)\frac{1}{2}\left(f_{ik} - \sqrt{f_{ik}^2 + \sigma}\right)\\
&\quad + \frac{1}{2}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-} + 1 - \sqrt{\left(\tilde{Q}_i^-/\tilde{P}_i^- - 1\right)^2 + \sigma}\right)\frac{1}{2}\left(1 - \frac{f_{ik}}{\sqrt{f_{ik}^2+\sigma}}\right)\frac{\partial f_{ik}}{\partial u_j}.
\end{aligned} \tag{4.12}$$

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

$$\frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^+}{\tilde{P}_i^+}\right) = \frac{\frac{\partial \tilde{Q}_i^+}{\partial u_j}\tilde{P}_i^+ - \tilde{Q}_i^+\frac{\partial \tilde{P}_i^+}{\partial u_j}}{\left(\tilde{P}_i^+\right)^2}, \quad \frac{\partial}{\partial u_j}\left(\frac{\tilde{Q}_i^-}{\tilde{P}_i^-}\right) = \frac{\frac{\partial \tilde{Q}_i^-}{\partial u_j}\tilde{P}_i^- - \tilde{Q}_i^-\frac{\partial \tilde{P}_i^-}{\partial u_j}}{\left(\tilde{P}_i^-\right)^2}. \tag{4.13}$$

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

$$\begin{aligned} \frac{\partial \tilde{Q}_i^+}{\partial u_j} &= -\sum_{l=1}^{N} \frac{\partial \min_\sigma(f_{il}, 0)}{\partial u_j} = -\frac{1}{2}\sum_{l=1}^{N}\left(1 - \frac{f_{il}}{\sqrt{f_{il}^2 + \sigma}}\right) d_{il}\frac{\partial (u_l - u_i)}{\partial u_j} \\ &= \begin{cases} -\dfrac{1}{2}\left(1 - \dfrac{f_{ij}}{\sqrt{f_{ij}^2 + \sigma}}\right) d_{ij} & \text{if } i \neq j, \\ \dfrac{1}{2}\displaystyle\sum_{l=1, l\neq i}^{n}\left(1 - \dfrac{f_{il}}{\sqrt{f_{il}^2 + \sigma}}\right) d_{il} & \text{if } i = j. \end{cases} \end{aligned} \tag{4.14}$$

This expression is compared with the corresponding expression (4.9) 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 (4.14) is very close to zero, which holds also for the value of (4.14). If $f_{ij} < 0$ is sufficiently small, then the expression in the parentheses is close to two and the value of (4.14) is close to $-d_{ij}$. In both cases, the values of (4.9) and (4.14) are practically the same. In the situation $f_{ij} = 0$, the value of (4.14) is $-d_{ij}/2$, which is different to the value 0 of (4.9) if $d_{ij} \neq 0$.

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

If $a_{ki} > a_{ik}$, one gets with (3.24) 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 [BB17] by $\sigma = 10^{-8} \cdot h^4$, where h is the maximal diameter of the mesh cells of the current triangulation. In [BB17], 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.

4.1.2.3 BJK Limiter

For the BJK limiter, the numerical studies are done only for a formal Newton method with separate treatment of the non-smooth points. A regularization of the BJK limiter is presented here but it should be noted here that the method fails to give results numerically.

4.1.2.3.1 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 (4.7) 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 (3.26) 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.

4.1.2.3.2 Approach with regularization of non-smooth points The regularization of BJK limiter is somewhat involved. The issues that one faces arises in definition of limiter α_{ij} (3.28) and the regularization of Q_i^+, Q_i^-.

For the regualrization of α_{ik}, we first note that,

$$\begin{aligned}\beta_{ik} &= \alpha_{ik} f_{ik} \\ &= \min\{R_i^+, R_k^-\} f_{ik}^+ + \min\{R_i^-, R_k^+\} f_{ik}^-.\end{aligned}$$

Now, we can regularize these functions using (4.10) and proceed in same way as (4.12) to compute the derivative as,

$$\begin{aligned}\frac{\partial \tilde{\beta}_{ik}}{\partial u_j} &= \left[\frac{\partial \tilde{R}_i^+}{\partial u_j} h(\tilde{R}_k^-, \tilde{R}_i^+) + \frac{\partial \tilde{R}_k^-}{\partial u_j} h(\tilde{R}_i^+, \tilde{R}_k^-)\right] \tilde{f}_{ik}^+ + \min_\sigma\{\tilde{R}_i^+, \tilde{R}_k^-\} \frac{\partial f_{ik}}{\partial u_j} h(f_{ik}, 0) \\ &+ \left[\frac{\partial \tilde{R}_i^-}{\partial u_j} h(\tilde{R}_k^+, \tilde{R}_i^-) + \frac{\partial \tilde{R}_k^+}{\partial u_j} h(\tilde{R}_i^-, \tilde{R}_k^+)\right] \tilde{f}_{ik}^- + \min_\sigma\{\tilde{R}_i^-, \tilde{R}_k^+\} \frac{\partial f_{ik}}{\partial u_j} h(0, f_{ik}),\end{aligned} \tag{4.15}$$

where

$$h(x, y) = 1 + \frac{x - y}{\sqrt{(x - y)^2 + \sigma}}.$$

We can find the intermediate derivatives of $\tilde{R}_i^+, \tilde{R}_i^-$ similarly to the previous case of the Kuzmin limiter.

For computing the quantities Q_i^+ and Q_i^- of the BJK limiter, one has to take the maximum of a set of numbers whose cardinality is larger than two, compare (3.25), (3.28). This operation has to be regularized. A straightforward idea consists in extending the regularization (4.10) to more than two arguments by using

$$\max\{x, y, z\} = \max\{x, \max\{y, z\}\}$$

and replacing the maximum by its regularized version. However, it is in general

$$\max_\sigma(x, \max_\sigma(y, z)) \neq \max_\sigma(\max_\sigma(x, y), z) \neq \max_\sigma(y, \max_\sigma(x, z)). \tag{4.16}$$

Thus, this approach leads to a regularization whose value depends on the sequence of the arguments. This situation is not desirable since it would not be possible for somebody else to reproduce the simulations unless the sequence of arguments is specified for each call of the regularization. This issue can be resolved by computing all three possible values of (4.16) and taking the arithmetic mean. This idea can be extended to more than three arguments. However, the number of possible sequences of arguments as well as the number of calls to $\max_\sigma$ increases considerably such that this approach is inefficient.

Another regularization of the maximum can be derived from the following relation

$$\max\{u_1,\ldots,u_q\} = \lim_{p\to\infty}\left(\sum_{i=1}^{q} u_i^p\right)^{1/p},$$

where $\{u_i\}_{i=1}^q$ are assumed to be non-negative. If all arguments are even positive, the p^{th} root is differentiable at the arguments. For the BJK limiter, one has to compute the maximum of values from a finite element function that is an approximation of the discrete solution at some intermediate iteration. The values of this function cannot be assumed to be positive. However, usually, one can choose a lower bound κ_{low} such that all expected finite element values are larger than κ_{low}, e.g., by using the approximation from the previous iteration. We do not see the necessity that κ_{low} is in some sense a tight strict lower bound. Then, a regularization can be derived in the following way:

$$\begin{aligned}\max\{u_1,\ldots,u_q\} &= \max\{u_1-\kappa_{\text{low}},\ldots,u_q-\kappa_{\text{low}}\}+\kappa_{\text{low}}\\ &= \lim_{p\to\infty}\left(\sum_{i=1}^{q}(u_i-\kappa_{\text{low}})^p\right)^{1/p}+\kappa_{\text{low}}\\ &\approx \left(\sum_{i=1}^{q}(u_i-\kappa_{\text{low}})^{p_0}\right)^{1/p_0}+\kappa_{\text{low}} \qquad (4.17)\end{aligned}$$

for some sufficiently value p_0. Since by construction $u_i - \kappa_{\text{low}} > 0$, this regularization is differentiable. In the same way, one derives

$$\min\{u_1,\ldots,u_q\} = -\max\{-u_1,\ldots,-u_q\} \approx -\left(\sum_{i=1}^{q}(-u_i+\kappa_{\text{upp}})^{p_0}\right)^{1/p_0} - \kappa_{\text{upp}},$$

where κ_{upp} is chosen such that $\kappa_{\text{upp}} > u_i$ for all expected values of approximations of the finite element solution. Besides κ_{low} and κ_{upp}, this approach requires to choose the power p_0. Now, proceeding in same way as for the Kuzmin limiter, leads to the formulas (4.11), (4.12), and (4.13). In (4.13), a division by zero cannot occur since $\tilde{P}_i^+ > 0$ and $\tilde{P}_i^- < 0$ because the set S_i is not empty.

One of the issues with this regularization is the choice of p_0. This issue is twofold:

1. We were not able to find in the literature the appropriate choice of p_0 which has to be used as a power in (4.17). We performed simulations with different values of p_0 and were not able to find an appropriate choice of p_0.

2. The second issue comes in taking the p_0^{th} root in (4.17). The system we worked on was not able to properly approximate for higher values of p_0, say $p_0 > 10$.

For completeness we are providing the details on how would the derivatives look after the computation but, no numerical results would be provided here.

For the derivatives in (4.13), one obtains with direct calculations

$$\begin{aligned}
\frac{\partial \tilde{Q}_i^+}{\partial u_j} &\approx \frac{\partial}{\partial u_j}\left[q_i \left(u_i - \left(\sum_{j\in S_i\cup\{i\}} (u_j-\kappa_{\mathrm{low}})^{p_0} \right)^{1/p_0} \right) \right] \\
&\approx q_i \begin{cases} -\left(\displaystyle\sum_{j\in S_i\cup\{i\}} (u_j-\kappa_{\mathrm{low}})^{p_0} \right)^{1/p_0-1} (u_j-\kappa_{\mathrm{low}})^{p_0-1} & \text{if } j\in S_i, \\ 1-\left(\displaystyle\sum_{j\in S_i\cup\{i\}} (u_j-\kappa_{\mathrm{low}})^{p_0} \right)^{1/p_0-1} (u_i-\kappa_{\mathrm{low}})^{p_0-1} & \text{if } i=j, \\ 0 & \text{else.} \end{cases}
\end{aligned}$$

The other derivatives can be computed in the same way.

4.1.2.4 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 (4.3)

$$\begin{aligned}
&M_{\mathrm{ite}}\left(\underline{u}^{(\nu)}\right)_{ij} \\
&= \begin{cases} a_{ij}+d_{ij}-\omega_{\mathrm{fp}}\alpha_{ij}^{(\nu)}d_{ij}-\omega_{\mathrm{Newt}}\displaystyle\sum_{k=1}^{N}\frac{\partial\alpha_{ik}^{(\nu)}}{\partial u_j}d_{ik}\left(u_k^{(\nu)}-u_i^{(\nu)}\right) & \text{if } i\neq j, \\ a_{ii}+d_{ii}+\omega_{\mathrm{fp}}\displaystyle\sum_{k=1,k\neq i}^{N}\alpha_{ik}^{(\nu)}d_{ik}-\omega_{\mathrm{Newt}}\displaystyle\sum_{k=1}^{n}\frac{\partial\alpha_{ik}^{(\nu)}}{\partial u_i}d_{ik}\left(u_k^{(\nu)}-u_i^{(\nu)}\right) & \text{if } i=j, \end{cases}
\end{aligned} \tag{4.18}$$

with ω_{fp} being the damping parameter already introduced for the mixed fixed point iteration (4.6) and $\omega_{\mathrm{Newt}} \in [0,1]$ being a second damping parameter. The last $N-M$ rows have just the diagonal entry 1.

Remark 4.4. Because of the conditions for achieving symmetry of the limiters, usually terms occur in the sums containing the derivatives of the limiters in (4.18) 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 4.5. It is expected that the convergence radius of Newton-type methods is in general smaller than of simple fixed point iterations. Thus, it is advisable to start the solution process for the nonlinear problem (4.2) with a simple fixed point iteration and then switch to a Newton-type method. This approach is studied for Example 4.3.1.1. 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 Sec. 4.3.

In performing preliminary simulations for the examples considered in Sec. 4.3, we observed that the formal Newton method with parameters $\omega_{\text{fp}} = 1$ and $\omega_{\text{Newt}} = 1$, for simple academic test problems in two dimensions did often not worked (Example 4.3.1.1). 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. Algorithm A.3 in Appendix A gives an overview of the above method.

Concerning the calculation of the entries of the formal Jacobian, we like to note that computing the sum after the factor ω_{Newt} in (4.18) is considerably more costly than evaluating the other terms in (4.18), because of the many cases that have to be distinguished for computing the derivatives of $\alpha_{ik}^{(\nu)}$.

4.2 Algorithmic Components

4.2.1 Adaptive Choice of Damping Parameter

It is our experience that an appropriate choice of the damping parameters $\{\omega^{(\nu)}\}$ in (4.3) 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 in general 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 [JK08], which includes also the rejection of iterates. In the numerical studies presented in the current paper, exactly this algorithm was used. The algorithm is presented in Appendix A, Algorithm A.1.

4.2.2 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 [WN11]. 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 [WN11], 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 [WN11] 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 [ACF$^+$11, BJKR18]. In these papers, method (4.5) was applied and a constant damping parameter was used. Whereas in [ACF$^+$11], a certain improvement compared with using method (4.5) with adaptive damping parameter is reported, the results in [BJKR18] 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 (4.5). In the simulations presented here, Anderson acceleration was used in combination with the adaptive damping strategy from [JK08], but without rejection of steps.

In addition to Algorithm AA from [WN11], we implemented also the Anderson acceleration with the new iterate [WN11, (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. The algorithm is presented in Appendix A, Algorithm A.2.

4.2.3 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 [BB17], 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 [JMR$^+$09] for a concrete example.

In the examples presented here, lower and upper admissible values of the solution are known. Therefore, the idea from [BB17] 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 [BJK16]. This aspect will be discussed for Examples 4.3.1.7, 4.3.2.1, and 4.3.2.4 in Sec. 4.3.

4.2.4 Choosing the Initial Condition

In Sec. 4.3, Example 4.3.1.4, studies concerning the impact of the initial iterate on the number of iterations are 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 [RST08], and the solution of the SUPG (Streamline-Upwind Petrov–Galerkin) finite element method from Chapter 3. It was observed that there was only a minor impact. In general, the solution of the SUPG finite element method with the choice of the stabilization parameter as given in [JK07b] was a good choice and it was used in all simulations presented below.

4.3 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 $\mathcal{O}(1)$. Different convection fields are considered, a mildly convection-dominated case, $\varepsilon = 10^{-3}$ or $\varepsilon = 10^{-4}$, and a more strongly convection-dominated case, $\varepsilon = 10^{-6}$. In these studies, the following methods were involved:

- *mixed fixed point*(ω_{fp}): mixed fixed point iteration (4.6) with the parameter ω_{fp}. Note that *mixed fixed point*(0) corresponds to the method *fixed point rhs*, see also (4.5), and *mixed fixed point*(1) to the method *fixed point matrix*, compare (4.4).

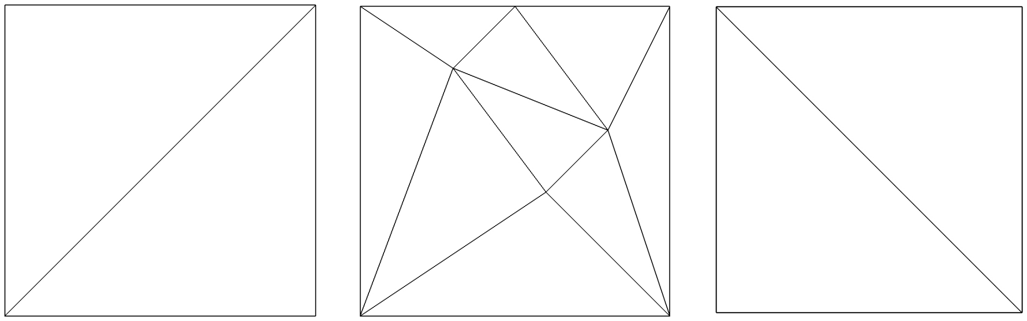


Figure 4.1: Grid 1, 2 and 3, level 0.

- *mixed fixed point with Anderson acceleration*$(\omega_{\mathrm{fp}}, \kappa)$: *mixed fixed point*$(\omega_{\mathrm{fp}})$ with Anderson acceleration and κ Anderson vectors, see Sec. 4.2.2.

- *formal Newton* $(\omega_{\mathrm{fp}} = 1, \omega_{\mathrm{Newt}} = 1)$ with separate treatment of the non-smooth points and fixed ω_{fp} and ω_{Newt}.

- *formal Newton* $(\omega_{\mathrm{fp}}, \omega_{\mathrm{Newt}})$ with separate treatment of the non-smooth points and adaptive change of ω_{Newt}, see Secs. 4.1.2.2 and 4.1.2.3,

- *formal Newton* $(\omega_{\mathrm{fp}}, \omega_{\mathrm{Newt}})$ with regularization and adaptive change of ω_{Newt} (only for the Kuzmin limiter), see Sec. 4.1.2.2.

For all *formal Newton* methods apply the approaches discussed in Remarks 4.4 and 4.5.

Stopping criteria for solving the nonlinear equations were as follows:

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

- 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 [GJM$^+$16, WBA$^+$16] at compute servers HP BL460c Gen9 2xXeon, Fourteen-Core 2600MHz.

4.3.1 Examples in Two Dimensions

4.3.1.1 Example with a Smooth Solution

In this example, the prescribed solution is

$$u(x, y) = 100x^2(1-x)y(1-2y)(1-y),$$

the convection field is $\boldsymbol{b} = (3, 2)^T$, and the reaction coefficient $c = 1$. The domain is $\Omega = (0, 1)^2$. Homogeneous Dirichlet boundary conditions are applied on the whole boundary. Results will be presented for two values of the diffusion coefficient: the moderately small value $\varepsilon = 10^{-3}$ and the much smaller value $\varepsilon = 10^{-6}$. This example serves for obtaining first impressions on the behavior of the iterative schemes. Various meshes were used in the simulations, whose coarsest level (level 0) are shown in Fig. 4.1. Simulations were performed on Grid 1 and Grid 2 from Fig. 4.1. Note that Grid 2 is not a Delaunay triangulation. For the initial iterate, all values were set to be zero.

This example studies the basic iteration schemes *fixed point rhs*, *fixed point matrix* and *formal Newton* ($\omega_{\text{Newt}} = 1$ and $\omega_{\text{fp}} = 1$). Studies related to switch from fixed point iteration to Newton method, as discussed in Remark 4.5 is also considered here.

4.3.1.2 Fixed Point Iterations

In a first study, only the fixed point iterations *fixed point rhs* and *fixed point matrix* were considered. For $\varepsilon = 10^{-3}$, the numbers of iteration steps are presented in Fig. 4.2. One can already observe that the behavior of the methods is somewhat different for the different limiters. For the Kuzmin limiter, the method *fixed point rhs* had no difficulties to solve the nonlinear problems and the number of iterations decreased with refinement of the grids. A similar behavior can be observed for *fixed point matrix*, often with a similar number of iterations. For the BJK limiter, in contrast, the method *fixed point matrix* needed consistently much fewer iterations than *fixed point rhs*, apart of the coarsest uniform grid. Altogether, the nonlinear problems in the case of a moderately small value of the diffusion could be solved without real difficulties.

Results for $\varepsilon = 10^{-6}$ are shown in Figs. 4.3 and 4.4. Fig. 4.3 presents the reduction of the error $\|\nabla(u - u^h)\|_{L^2}$. On the uniform grid, the order of error convergence is similar for both limiters, with the solution of the Kuzmin limiter being somewhat more accurate. For the unstructured grid, it can be observed that the BJK limiter worked well on this grid with an order of convergence of about 1. In contrast, the application of the Kuzmin limiter led to a clear reduction of this order. The behavior of the iterative methods is presented in Fig. 4.4. Now, there are fundamental differences considering both limiters. For the Kuzmin limiter,

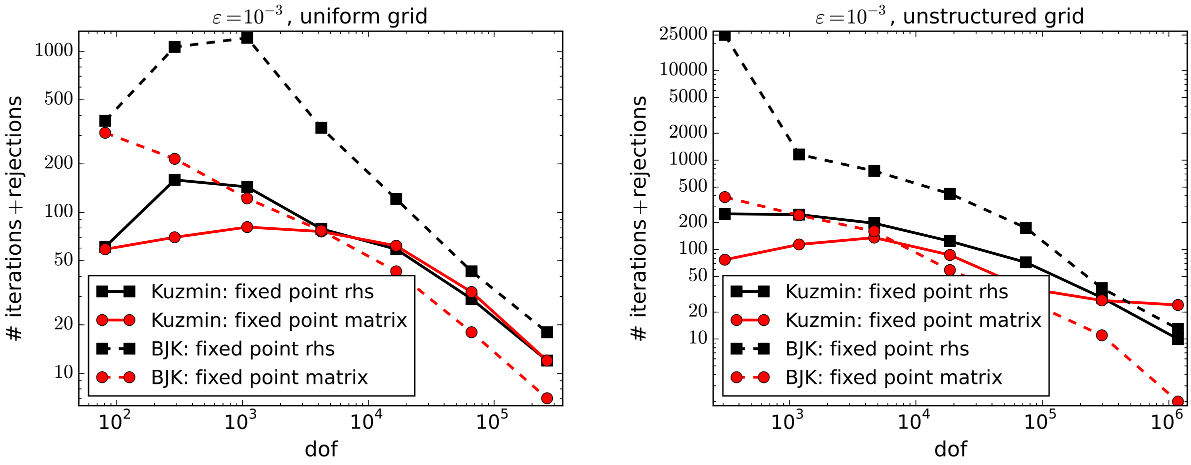


Figure 4.2: 2d smooth solution. Number of iterations and rejections for $\varepsilon = 10^{-3}$, left: uniform grid, right: unstructured grid.

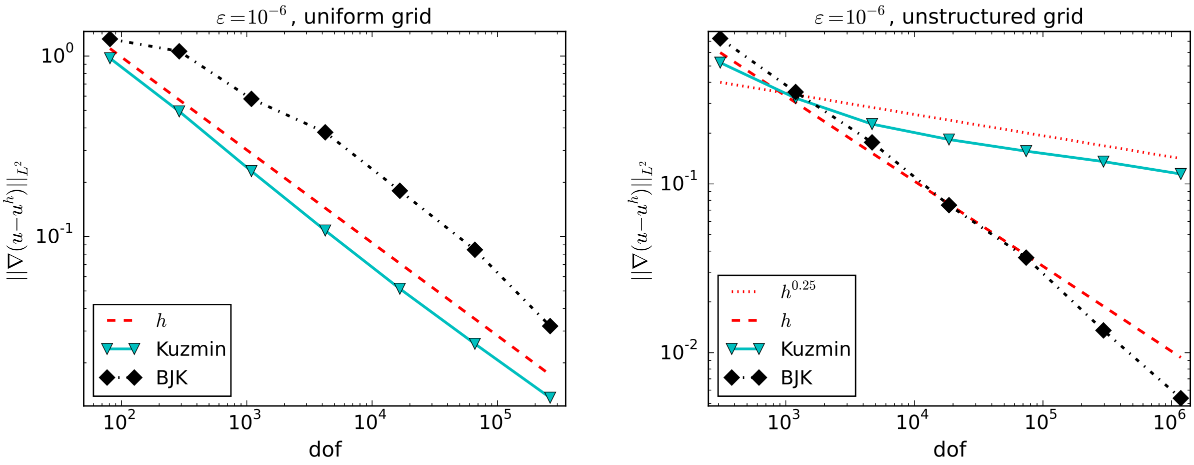


Figure 4.3: 2d smooth solution. Errors of the computed solutions.

fixed point rhs worked satisfactory, all problems were solved within the prescribed maximal number of iterations. But even on the uniform grid, *fixed point matrix* failed to converge on fine grids. In case of the BJK limiter, *fixed point rhs* did not converge on many grids, but *fixed point matrix* performed usually quite well.

Since the application of the Kuzmin limiter on the unstructured grid led to quite inaccurate numerical solutions, this limiter should not be used on this grid. This combination will not be considered in the further studies.

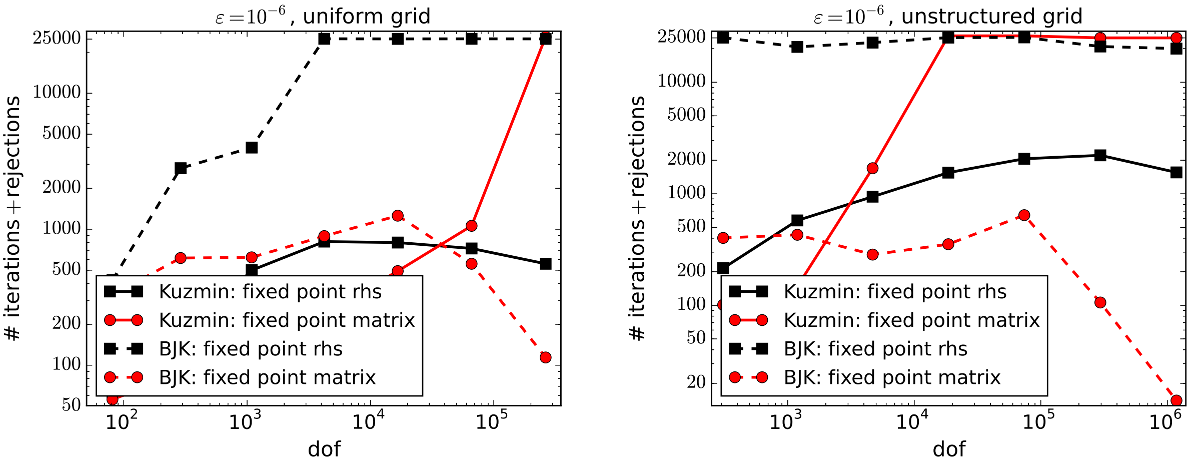


Figure 4.4: 2d smooth solution. Number of iterations and rejections for $\varepsilon = 10^{-6}$, left: uniform grid, right: unstructured grid.

4.3.1.3 Formal Newton Methods ($\omega_{\text{fp}} = 1, \omega_{\text{Newt}} = 1$)

Next, the *formal Newton* method will be included in the studies. It is well known that Newton-type methods possess in general a smaller domain of convergence than simpler fixed point iterations. We could observe this behavior also here: applying *formal Newton* from the first step of the iteration led usually to unsatisfactory results concerning the number of steps. For brevity, those results are not presented here.

The first approach for involving the *formal Newton* method was quite simple. In the first part of the iteration, a fixed point method was applied until the Euclidean norm of the residual vector was below a switching tolerance tol_{sw}. Then, *formal Newton* was performed without any possibility of switching back. The current damping parameter ω was used in the first step of the *formal Newton* method. For the first part, we applied as well *fixed point rhs* as *fixed point matrix*. From the results obtained with these methods, Fig. 4.4, it can be expected that *fixed point rhs* is a better choice for the Kuzmin limiter and *fixed point matrix* for the BJK limiter. In fact, the numerical results confirmed these expectations. Thus, for brevity, only the corresponding results are presented in Figs. 4.5 and 4.6.

For the Kuzmin limiter, Fig. 4.5, it can be seen that *formal Newton* worked well only on coarse grids. On finer grids, it did not converge even for small switching tolerances tol_{sw}. The observations for the BJK limiter are different. On some levels, *formal Newton* worked well, at least for sufficiently small tol_{sw}, but on other levels, this method failed to converge.

Examining the non-convergent simulations more closely, we found that often the Euclidean norm of the residual increased within a few steps after having switched to the *formal Newton* method, sometimes it increased considerably. A straightforward idea to mitigate this behav-

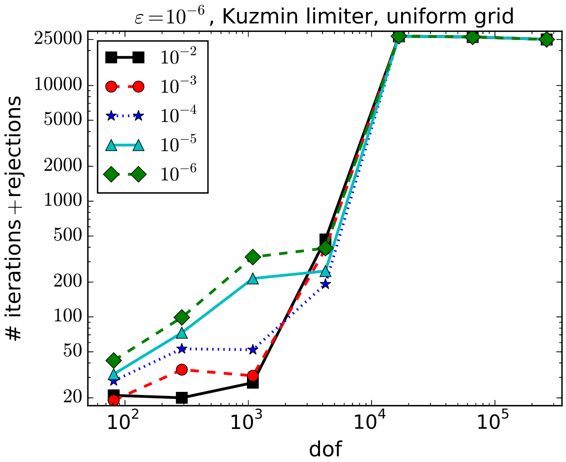


Figure 4.5: 2d smooth solution. Number of iterations and rejections for $\varepsilon = 10^{-6}$, Kuzmin limiter and *formal Newton* method with *fixed point rhs* in the first part, different values for the parameter $\mathrm{tol_{sw}}$.

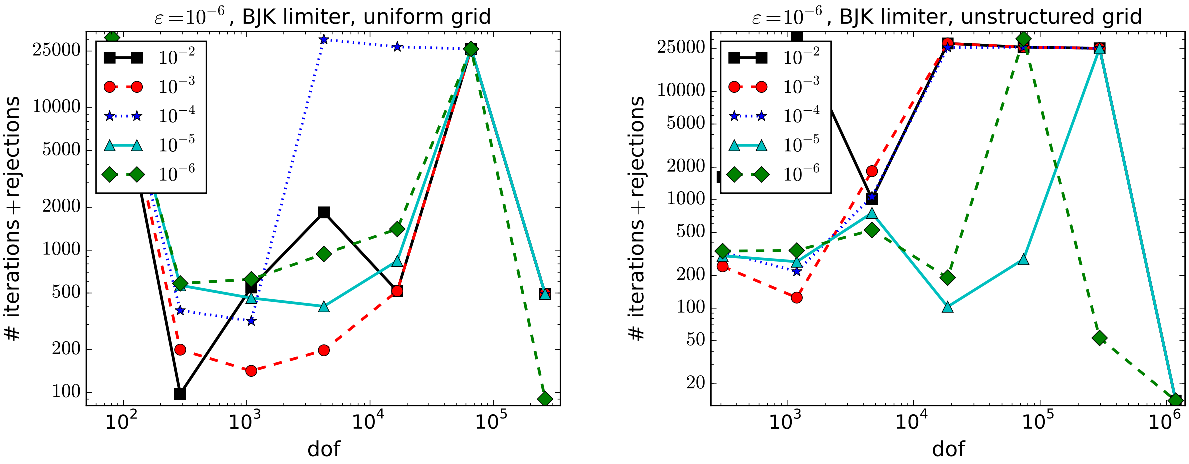


Figure 4.6: 2d smooth solution. Number of iterations and rejections for $\varepsilon = 10^{-6}$, BJK limiter and *formal Newton* method with *fixed point matrix* in the first part, different values for the parameter $\mathrm{tol_{sw}}$, left: Grid 1, right: Grid 2.

ior consists in switching back to the fixed point iteration that was used in the first part after the norm of the residual exceeds a certain limit. This approach was implemented in the form that the back switch to the method from the first part took place always if the Euclidean norm of the residual became larger than $100 \cdot \mathrm{tol_{sw}}$. While switching between the methods, the current damping parameter ω was not changed. However, the behavior of the *formal Newton* method in general did not improve. The only exception is presented in Fig. 4.7, where it can be seen that the choice $\mathrm{tol_{sw}} = 10^{-5}$ led to a convergent method for the BJK limiter on all levels of the unstructured grid.

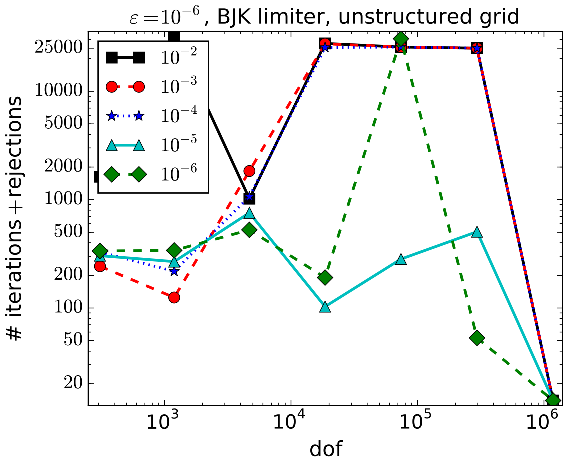


Figure 4.7: 2d smooth solution. Number of iterations and rejections for $\varepsilon = 10^{-6}$, BJK limiter and *formal Newton* method with *fixed point matrix* in the first part and switching back to *fixed point matrix* if the norm of the residual became too large, different values for the parameter $\mathrm{tol}_{\mathrm{sw}}$, Grid 2.

Already for an example with a smooth solution, there were only few of the considered methods that converged in the convection-dominated case on every refinement level. On the uniform grid, for the Kuzmin limiter only *fixed point rhs* worked well and for the BJK limiter only *fixed point matrix*. There were two satisfactory performing approaches for the BJK limiter on the unstructured grid: *fixed point matrix* and *formal Newton* with $\mathrm{tol}_{\mathrm{sw}} = 10^{-5}$, where *fixed point matrix* was used as starting method and it was switched back to *fixed point matrix* if the norm of the residual became too large.

4.3.1.4 Example with Interior and Boundary Layers

This example, proposed in [HMM86], is a standard academic example for numerical studies of steady-state convection-diffusion equation. It is given in $\Omega = (0,1)^2$ with $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))$, $c = f = 0$ and the Dirichlet boundary condition

$$u = \begin{cases} 1 & (y = 1 \wedge x > 0) \text{ or } (x = 0 \wedge y > 0.7), \\ 0 & \text{else.} \end{cases}$$

Again, the strongly convection-dominated case $\varepsilon = 10^{-6}$ is considered. Then, the solution exhibits an internal layer in the direction of the convection starting from the jump of the boundary condition at the left boundary and two exponential layers at the right and the lower boundary, see Fig. 4.8. This example studies the impact of initial iterate on the iterative schemes as discussed in Sec. 4.2.4.

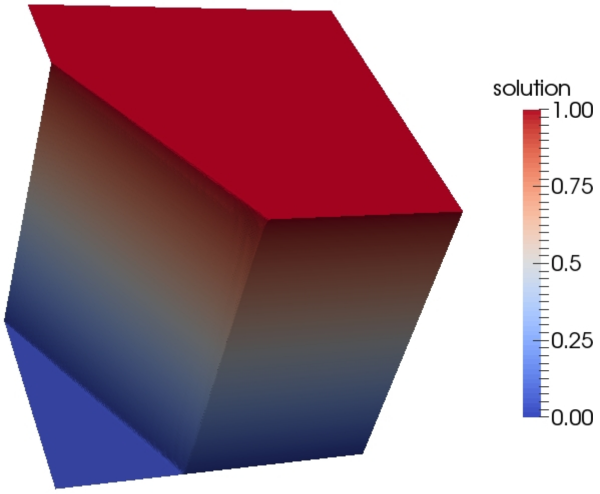


Figure 4.8: 2d Interior and boundary layer example. Solution (computed with the BJK limiter, Grid 3, level 9).

4.3.1.5 Impact of Initial Iterate

In this example, a study of the impact on choosing the initial iterate in different ways will be presented. For the initial iterate, we considered the following options:

- setting all non-Dirichlet degrees of freedom to zero (zero),
- using the solution of the upwind finite element method from [RST08] (upwind),
- using the solution of the SUPG method from [HB79, BH82] (SUPG),
- using the solution of the Galerkin method (Galerkin).

Starting with the zero initial iterate is a usual approach if no information about the expected solution are available. With the upwind method as initial iterate, the positions of the layers are known from the beginning, but the layers are strongly smeared. The positions of the layers are also known with the SUPG method, the layers are sharp, but there are considerable spurious oscillations in a vicinity of the layers. The incorporation of the Galerkin finite element method in this study is just for completeness.

4.3.1.6 Fixed Point Iterations

First, again the behavior of the fixed point iterations was studied, see Fig. 4.9, left picture. All simulations presented in this figure were started with the SUPG solution as initial iterate. In this example, *fixed point rhs* converged for both limiters on all grids, whereas *fixed point matrix* did not converge for both limiters on fine grids. For the Kuzmin limiter, the *fixed*

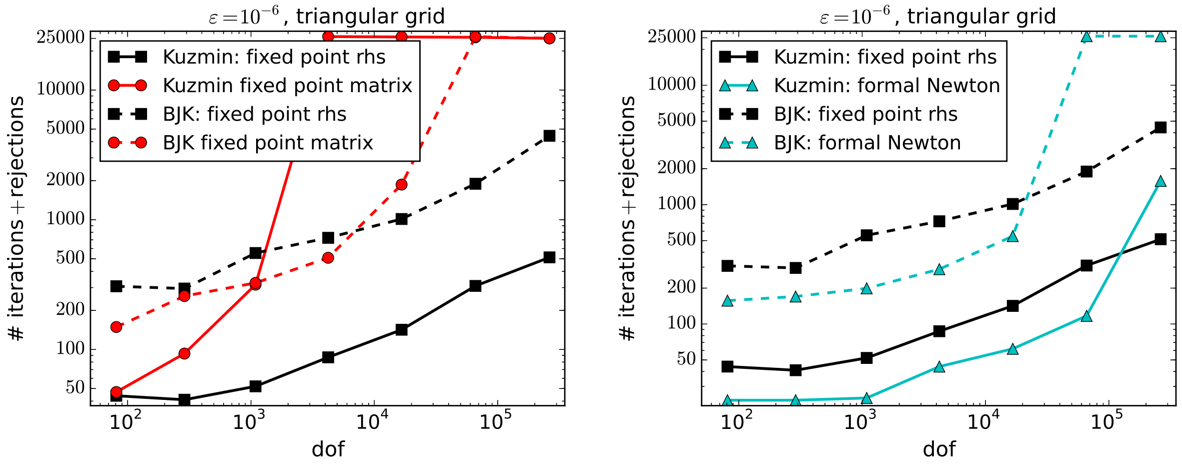


Figure 4.9: 2d Interior and boundary layer example. Number of iterations and rejections.

point rhs method needed considerably less iterations. Representative results for the *formal Newton* method, with *fixed point rhs* as scheme that was used if the norm of the residual was too large and $\mathrm{tol}_{\mathrm{sw}} = 10^{-5}$, are displayed in Fig. 4.9, right picture. On coarser grids, this approach needed less iterations than *fixed point rhs*, but on finer grids, it even failed in two cases.

The dependency of the number of iterations and rejections on the initial iterate is illustrated in Fig. 4.10. Generally, there are only minor differences between the four initial iterates. Often, using the SUPG solution proved to be a good choice.

Remark 4.6. After performing simulations for simple academic examples in two dimensions, we have an idea of how the basic methods behave. Now, in the further examples we will look at *mixed fixed point* iterations and the algorithmic components defined in Sec. 4.2.

4.3.1.7 The 2d Hemker Problem

This example, defined in [Hem96], is a standard benchmark problem for steady-state convection-diffusion equation. 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 (2.6). 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 [BJKR18] 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 Fig. 4.11. Simulations were performed on a triangular grid and a quadrilateral grid, see Fig. 4.12 for the coarsest grids (level 0) and Table 4.1 for information on the number of degrees of freedom.

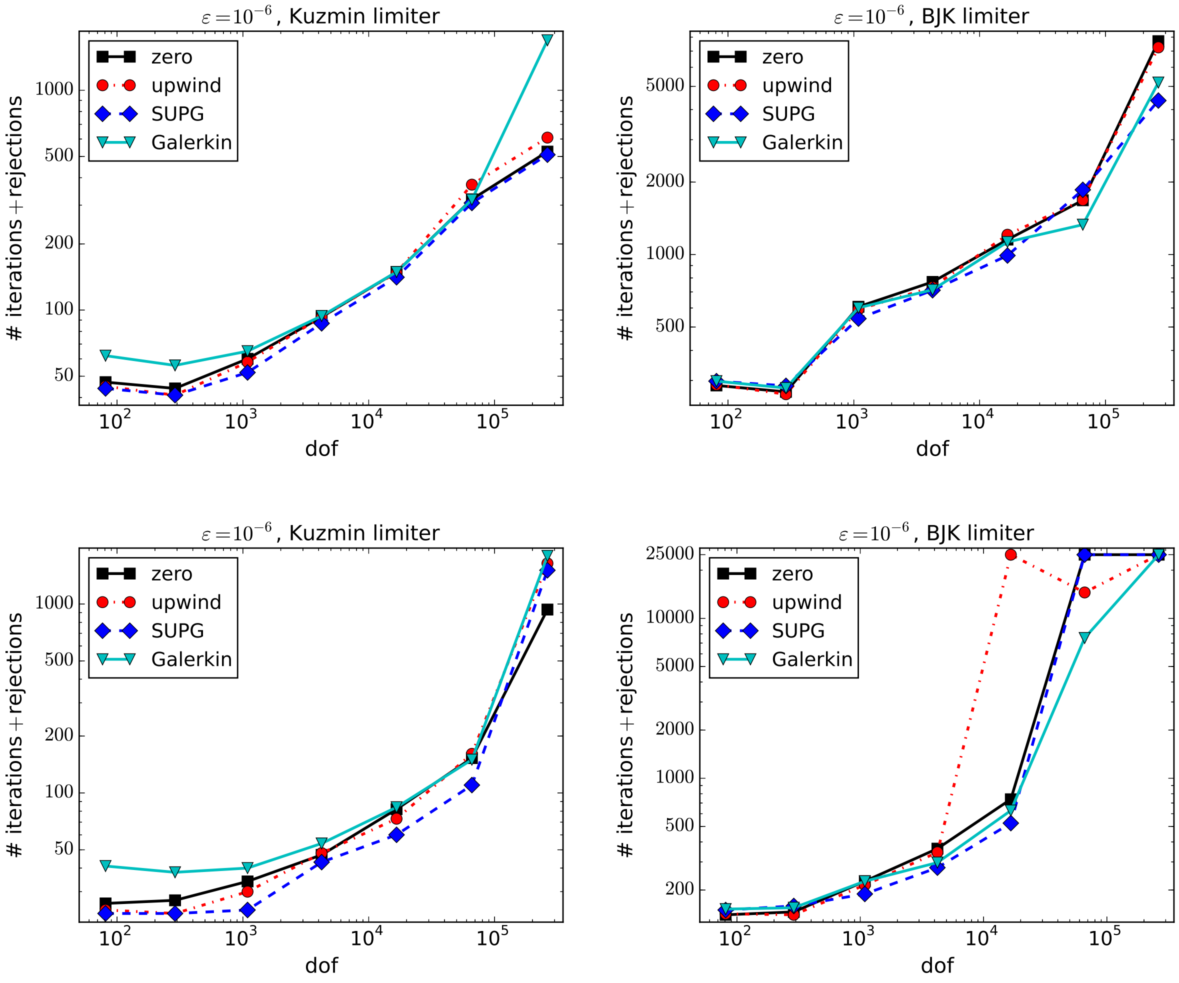


Figure 4.10: 2d Interior and boundary layer example. Number of iterations and rejections depending on the initial iterate, top: *fixed point rhs*, bottom: *formal Newton*.

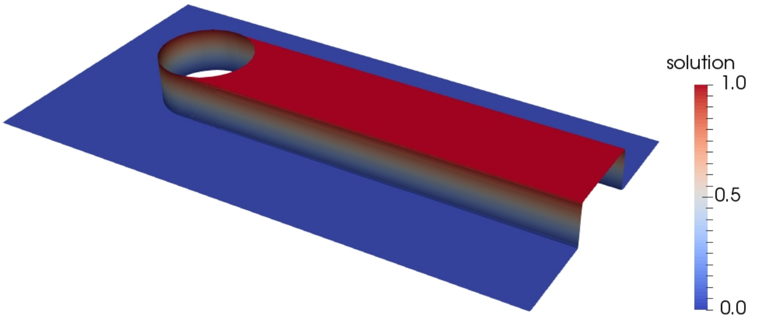


Figure 4.11: 2d Hemker problem. Solution for $\varepsilon = 10^{-6}$, computed with the BJK limiter, P_1, level 6.

Concerning the satisfaction of the DMP, both grids from Fig. 4.12 are not covered by the available analysis for the Kuzmin limiter. However, we could observe in preliminary simula-

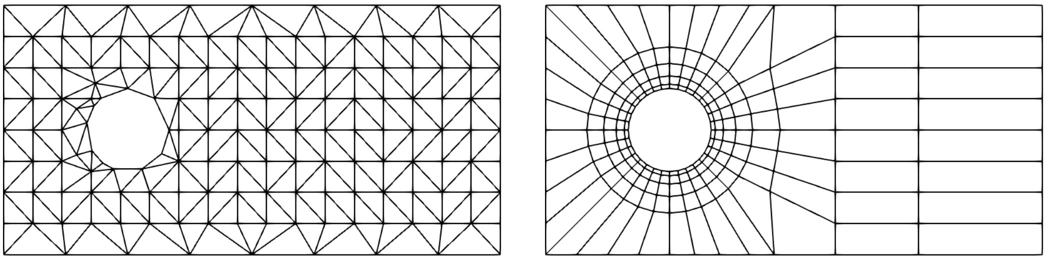


Figure 4.12: 2d Hemker problem. Triangular grid and quadrilateral grid (level 0).

Table 4.1: 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

tions that the computed solutions with the Kuzmin limiter take values in $[0, 1]$. Here the first comprehensive study for *mixed fixed point* in 2d is presented. This example also presents the impact of Anderson acceleration Sec. 4.2.2 and the projection to admissible values Sec. 4.2.3. Finally, results regarding the efficiency of the methods in 2d are presented here.

4.3.1.8 Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). First, the behavior of *mixed fixed point*(ω_{fp}) for $\omega_{\text{fp}} \in \{0, 0.05, \dots, 0.95, 1\}$ is illustrated in Fig. 4.13. The simulations were performed with and without the projection to admissible values as described in Sec. 4.2.3. One can see that there are only small differences with respect to the behavior of this method in both cases. A good value for the mixing parameter is $\omega_{\text{fp}} = 0.85$.

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 $\omega_{\text{fp}} = 0.85$, the impact of using Anderson acceleration with different numbers of Anderson vectors is presented in Fig. 4.14. For the moderately convection-dominated case, the use of

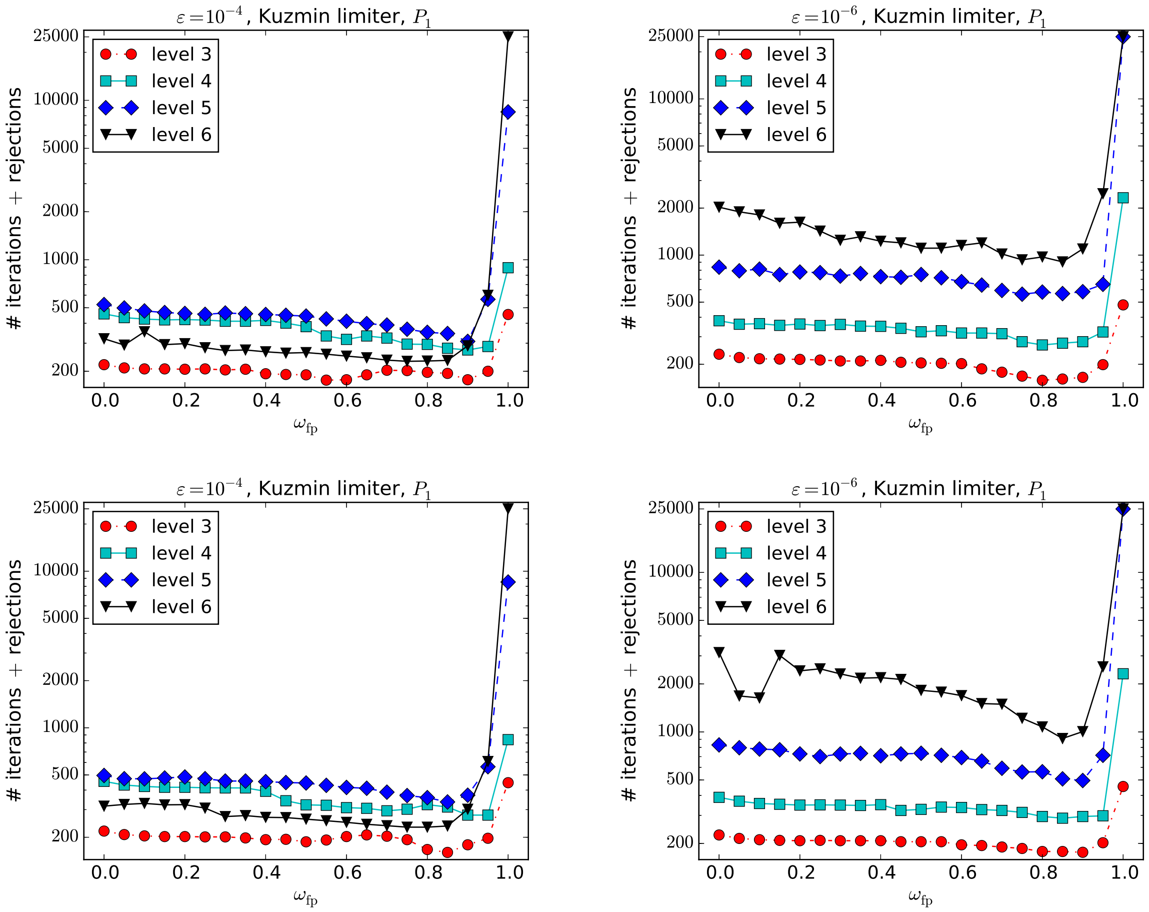


Figure 4.13: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), top: without projection to admissible values, bottom: with projection to admissible values.

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 Fig. 4.15. 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*

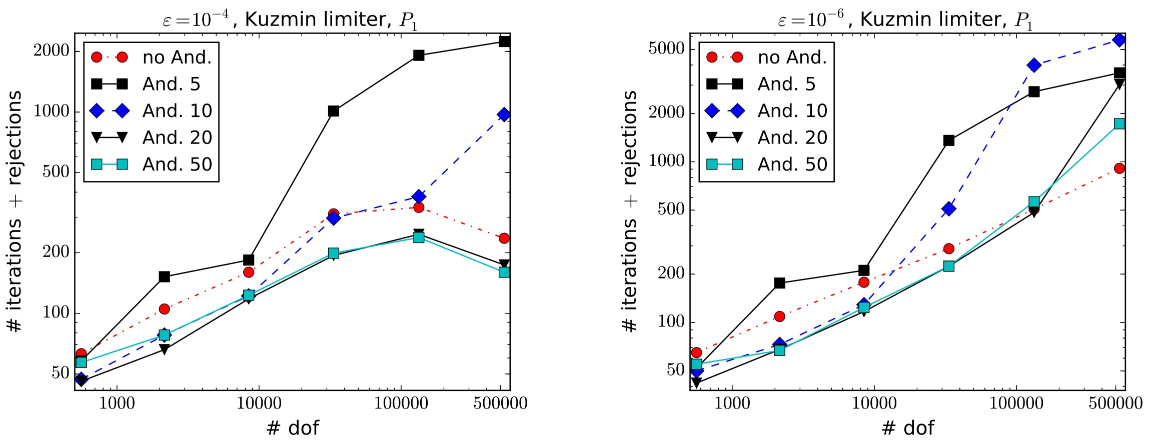


Figure 4.14: 2d Hemker problem. Results for *mixed fixed point with Anderson acceleration*$(0.85, \kappa)$, where κ is the number in the legends, with projection to admissible values.

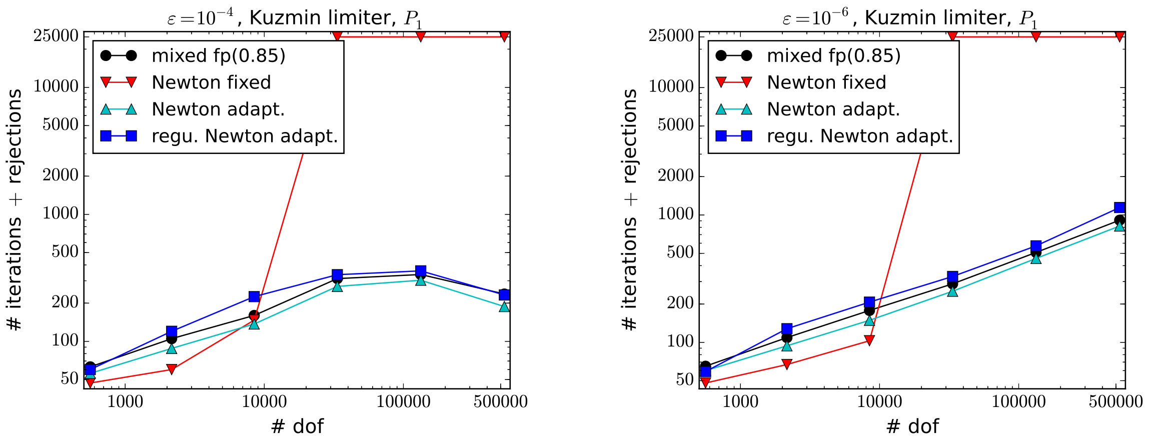


Figure 4.15: 2d Hemker problem. Results for the formal Newton methods, with projection to admissible values. The adaptive methods were used with $\omega_{\mathrm{fp}} = 0.85$ and $\omega_{\mathrm{Newt}} = 0.0625$.

Newton methods does not significantly reduce the number of iterations+rejections.

4.3.1.9 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 Figs. 4.16 and 4.17, which should be compared with Figs. 4.13 and 4.15, respectively.

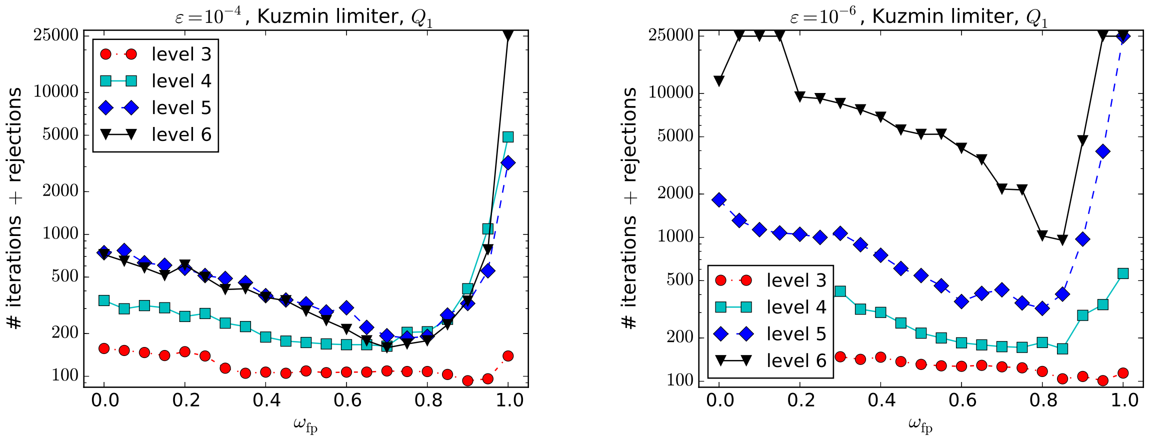


Figure 4.16: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values.

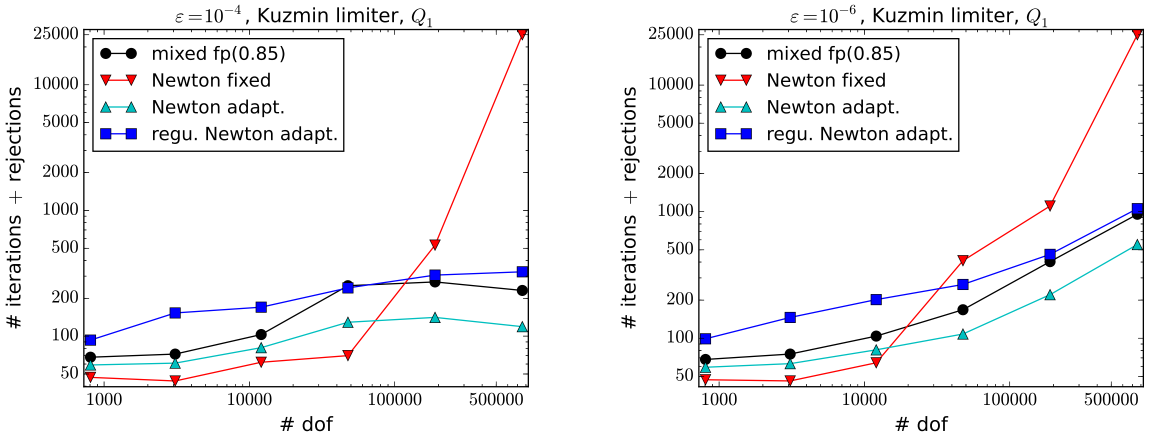
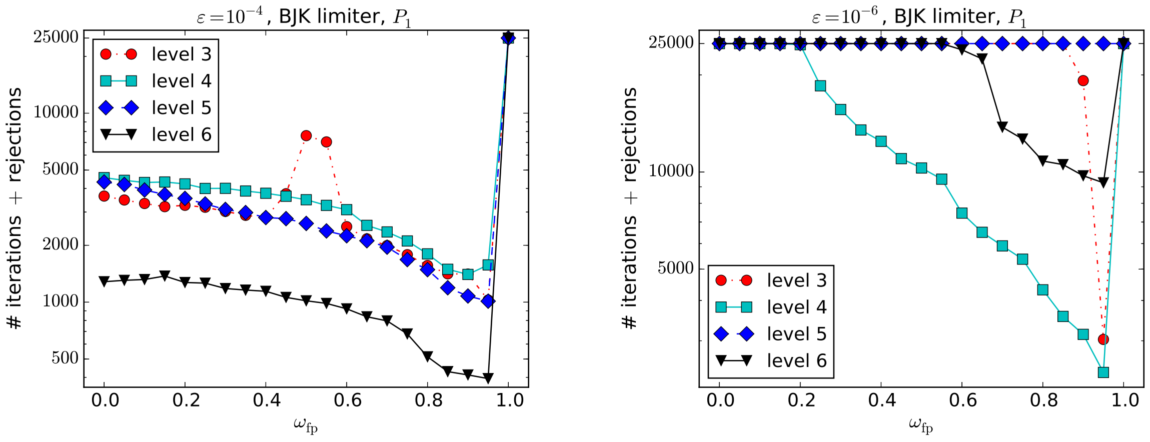


Figure 4.17: 2d Hemker problem. Results for the formal Newton methods, with projection to admissible values. The adaptive methods were used with $\omega_{\mathrm{fp}} = 0.85$ and $\omega_{\mathrm{Newt}} = 0.0625$.

4.3.1.10 BJK Limiter with P_1 Finite Elements

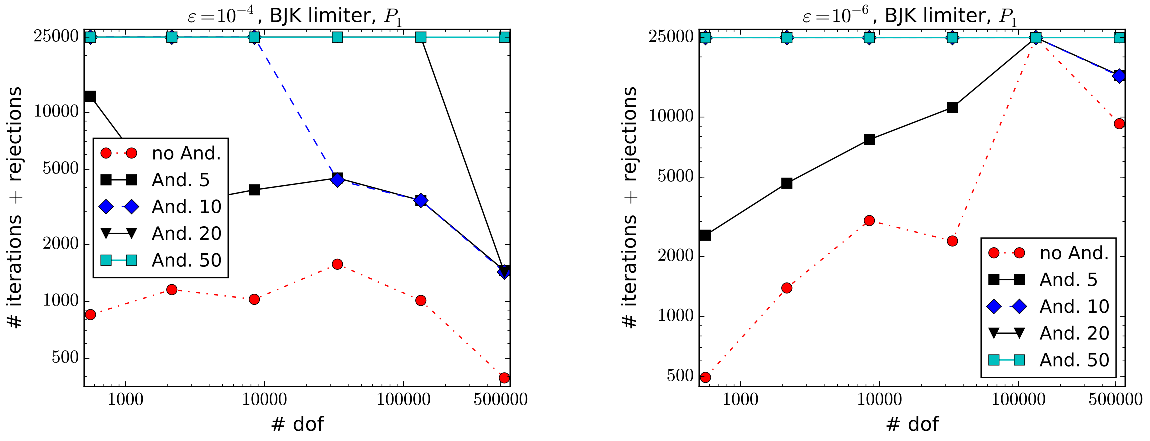
Studies for mixed fixed point(ω_{fp}). The results for this method are presented in Fig. 4.18. In the moderately convection-dominated regime, it can be observed that choosing $\omega_{\mathrm{fp}} = 0.95$ leads always to a comparatively small number of iterations, whereas the method does not converge for $\omega_{\mathrm{fp}} = 1$. 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 $\omega_{\mathrm{fp}} = 0.95$.

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


$\varepsilon = 10^{-4}$, BJK limiter, P_1

$\varepsilon = 10^{-6}$, BJK limiter, P_1

Figure 4.18: 2d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values.


$\varepsilon = 10^{-4}$, BJK limiter, P_1

$\varepsilon = 10^{-6}$, BJK limiter, P_1

Figure 4.19: 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.

Fig. 4.19.

Studies for formal Newton methods. Results obtained for *formal Newton* methods are presented in Figs. 4.20 and 4.21. 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 $\omega_{\mathrm{fp}} = 1$, $\omega_{\mathrm{Newt}} = 1$, starting from the first iteration (Newton wo damp. in Fig. 4.21) 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

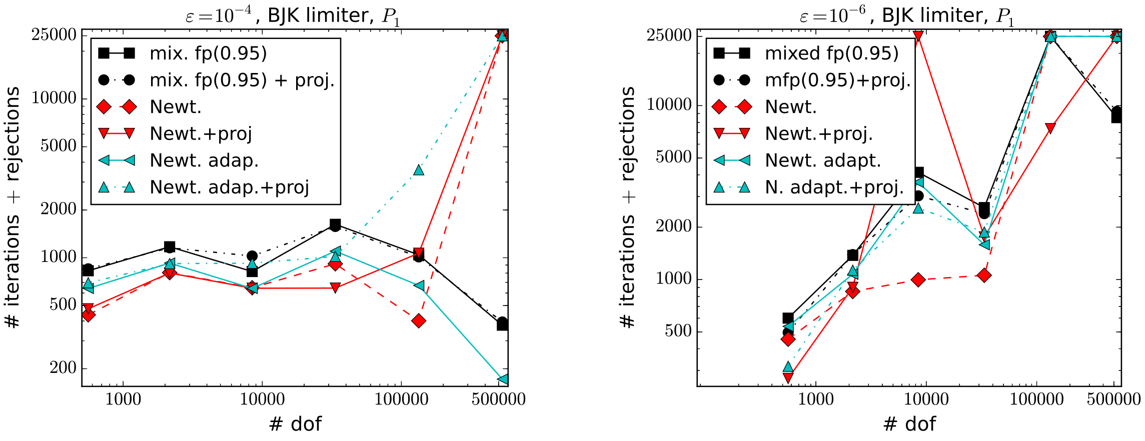


Figure 4.20: 2d Hemker problem. Results for the formal Newton methods, with and without projection to admissible values. The adaptive methods were used with $\omega_{\mathrm{fp}} = 0.95$ and $\omega_{\mathrm{Newt}} = 0.0625$.

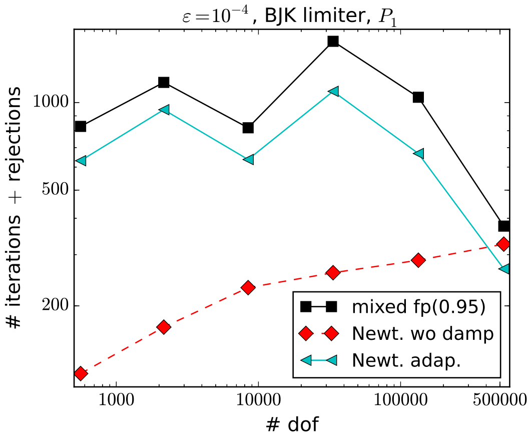


Figure 4.21: 2d Hemker problem. Results for the formal Newton methods, without projection to admissible values.

differently with and without projection to admissible values.

4.3.1.11 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 [Dav04]. 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 Fig. 4.22.

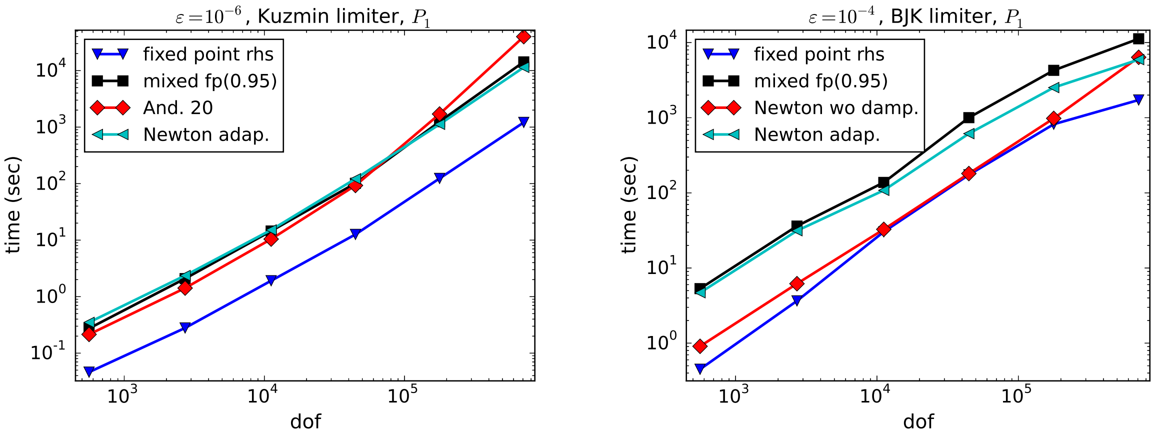


Figure 4.22: 2d Hemker problem. Efficiency of several methods.

Fig. 4.22 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.

4.3.2 Examples in Three Dimensions

4.3.2.1 The 3d Hemker Problem

The 3d Hemker problem is a natural extension of the 2d Hemker problem, which was proposed in [WBA$^+$16]. The domain is defined by

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

and the convection vector in (2.6) 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 Fig. 4.23. This example presents studies with respect to Anderson acceleration Sec. 4.2.2 and projection to admissible values Sec. 4.2.3. Comprehensive studies for *mixed fixed point* is also presented.

Simulations were performed for P_1 and Q_1 (only Kuzmin limiter) finite elements, see Fig. 4.23 for the coarsest tetrahedral grid and Table 4.2 for information on the number of degrees of freedom. It turned out that the solutions computed with the Kuzmin limiter on the

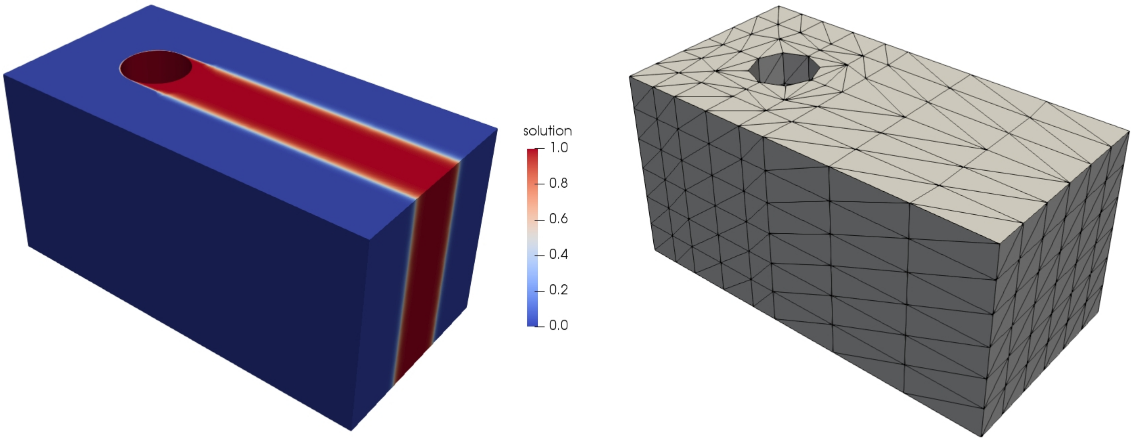


Figure 4.23: 3d Hemker problem. Solution for $\varepsilon = 10^{-6}$, computed with the Kuzmin limiter, P_1, level 4, and sketch of the coarsest grid (level 0).

Table 4.2: 3d Hemker problem. Number of degrees of freedom, including Dirichlet nodes.

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

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 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.

4.3.2.2 Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). The results of these studies are displayed in Fig. 4.24. It can be seen that *mixed fixed point*(ω_{fp}) converged only for sufficiently small mixing parameters

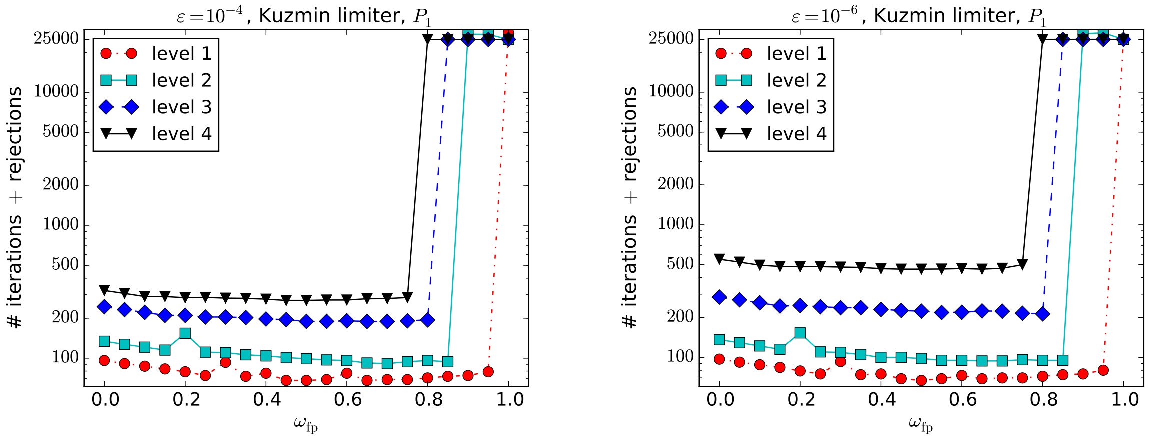


Figure 4.24: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$; $\varepsilon = 10^{-6}$: level 1 with $\omega_{\mathrm{fp}} = 1$, level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

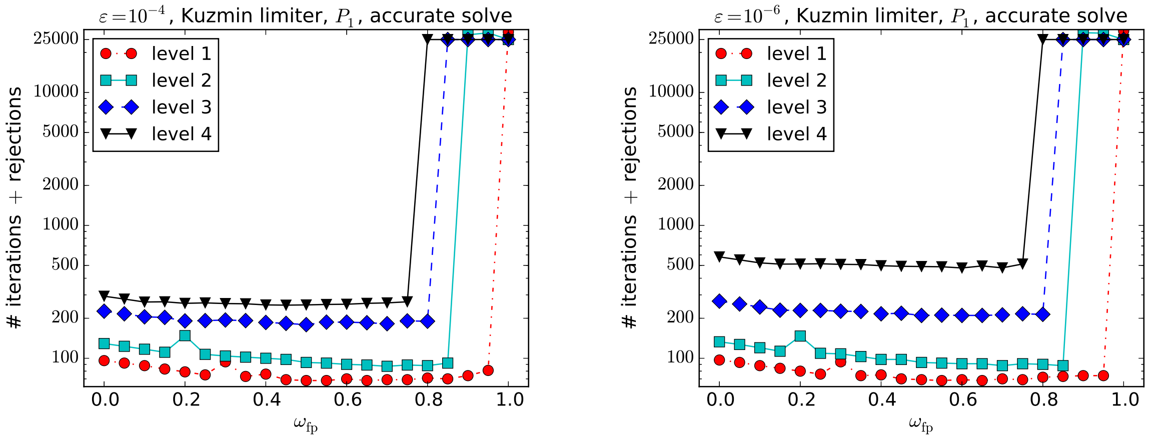


Figure 4.25: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values, with accurate solution of the linear problems. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 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 $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

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 Sec. 4.3.2.7 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 Fig. 4.25. One can see by comparing with Fig. 4.24 that the number of iterations is in all situations almost the same.

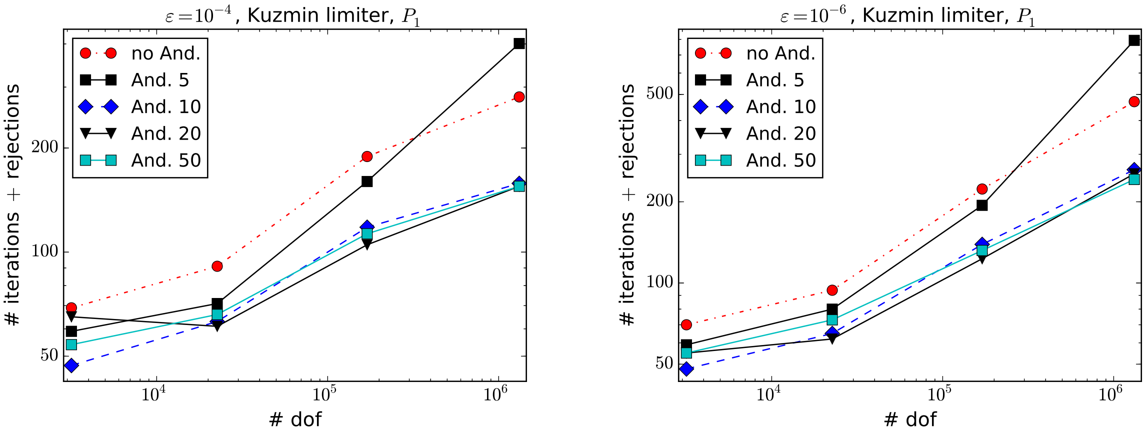


Figure 4.26: 3d Hemker problem. Results for *mixed fixed point with Anderson acceleration*$(0.7, \kappa)$, where κ is the number in the legends, without projection to admissible values.

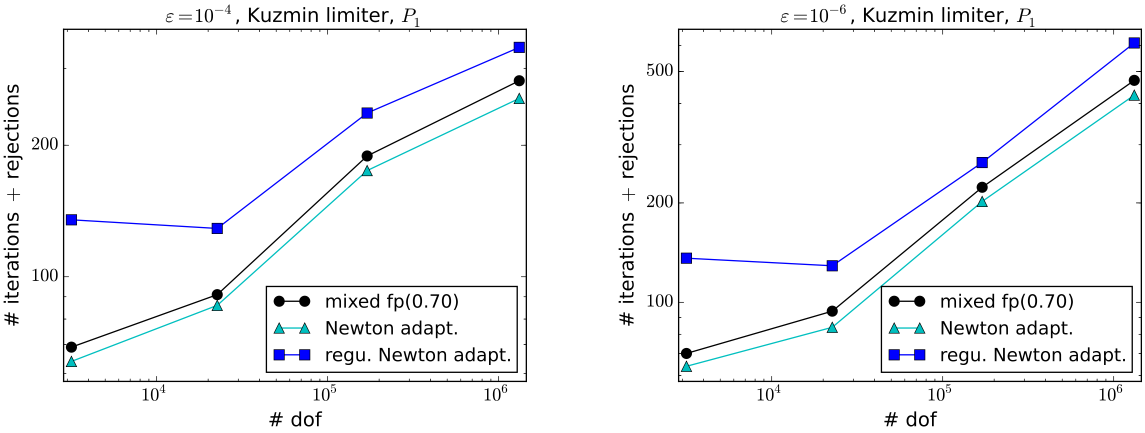


Figure 4.27: 3d Hemker problem. Results for the formal Newton methods, without projection to admissible values. The adaptive methods were used with $\omega_{\mathrm{fp}} = 0.7$ and $\omega_{\mathrm{Newt}} = 0.1$.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* The impact of using Anderson acceleration is demonstrated in Fig. 4.26. For both convection-dominated regimes, the application of the Anderson acceleration reduces the needed 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 Fig. 4.27. As can be seen, the *formal Newton* method without regularization sometimes reduces the number of iterations+rejections slightly, but in general do not lead to a notable improvement.

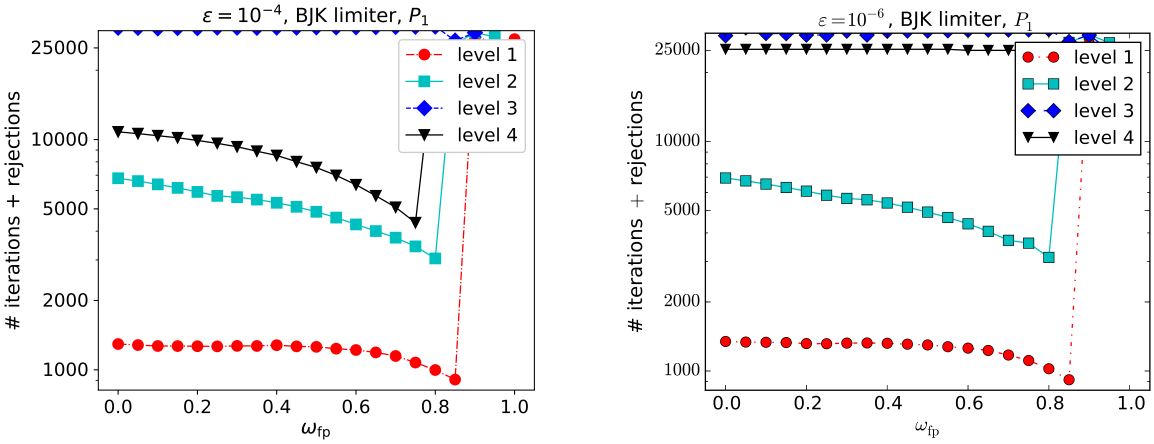


Figure 4.28: 3d Hemker problem. Results for the method *mixed fixed point*(ω_{fp}), with projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$; $\varepsilon = 10^{-6}$: level 1 with $\omega_{\mathrm{fp}} = 1$, level 2 with $\omega_{\mathrm{fp}} = 1$, level 3 with $\omega_{\mathrm{fp}} = 1$, level 4 with $\omega_{\mathrm{fp}} \in \{0.95, 1\}$.

4.3.2.3 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 Fig. 4.28. 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.

4.3.2.4 A 3d Problem with Non-Constant Convection

This example was proposed in [BJKR18]. 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 Fig. 4.29. All simulations were performed for P_1 finite elements on unstructured tetrahedral grids, whose coarsest grid was obtained with the mesh generator GMSH [GR09], see Fig. 4.29. Information concerning the degrees of freedom are provided in Table 4.3. This example presents the efficiency of different methods in 3d.

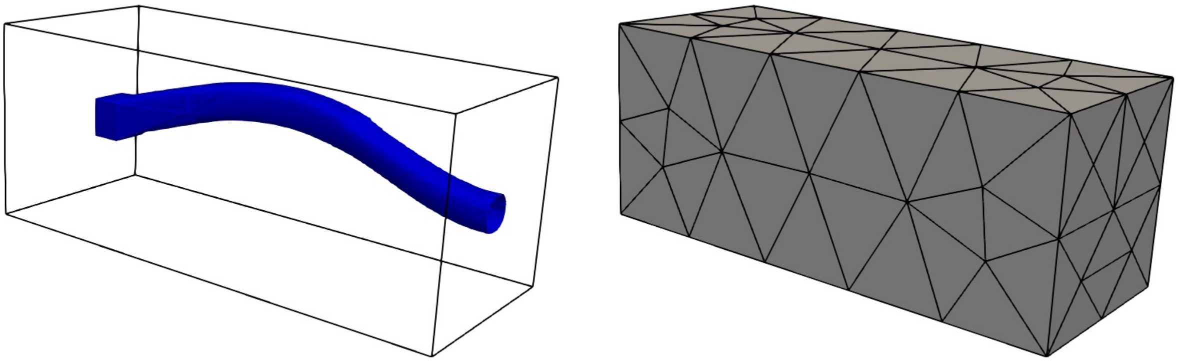


Figure 4.29: 3d problem with non-constant convection. Solution for $\varepsilon = 10^{-6}$, isosurface for $u = 0.05$, computed with the Kuzmin limiter, P_1, level 5, and sketch of the coarsest grid (level 0).

Table 4.3: 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

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 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 Sec. 4.2.3 behaved in general similarly. For the sake of brevity, only results without projection are presented below.

4.3.2.5 Kuzmin Limiter with P_1 Finite Elements

Studies for mixed fixed point(ω_{fp}). The results of these studies are displayed in Fig. 4.30. 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 $\omega_{\text{fp}} = 0.6$.

Studies for mixed fixed point(ω_{fp}) *with Anderson acceleration.* Fig. 4.31 shows the effect of using Anderson acceleration. For sufficiently many Anderson vectors, $\kappa \in \{10, 20, 50\}$, there is in general a notable reduction of the number of iterations+rejections compared with *mixed fixed point*(0.6).

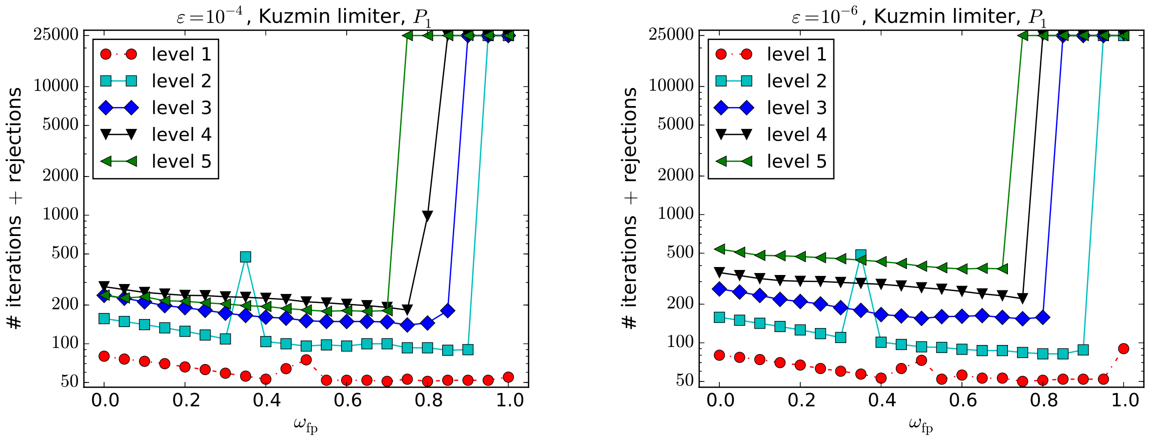


Figure 4.30: 3d problem with non-constant convection. Results for the method *mixed fixed point*$(\omega_{\rm fp})$, without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-4}$: level 4 with $\omega_{\rm fp} = 1$, level 5 with $\omega_{\rm fp} = 1$; $\varepsilon = 10^{-6}$: level 3 with $\omega_{\rm fp} = 1$, level 4 with $\omega_{\rm fp} = 1$, level 5 with $\omega_{\rm fp} \in \{0.95, 1\}$.

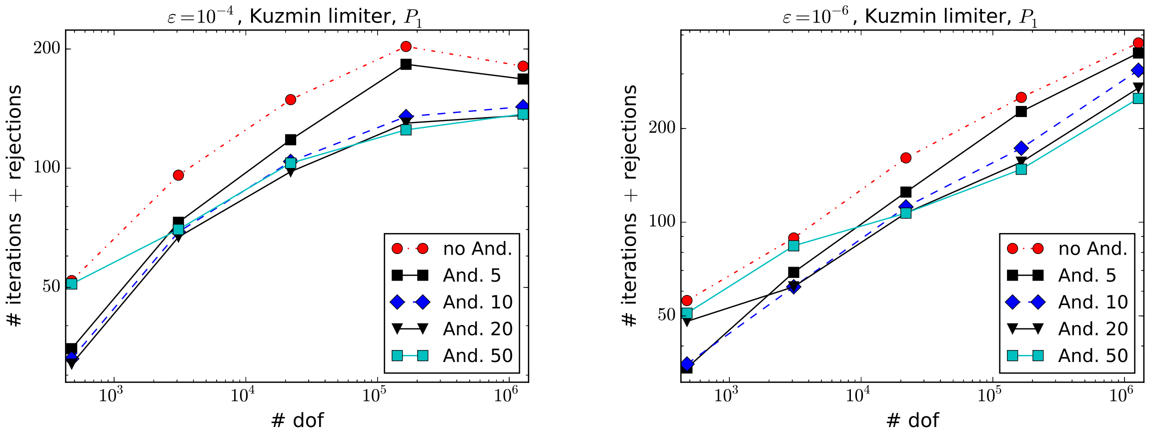


Figure 4.31: 3d problem with non-constant convection. Results for *mixed fixed point with Anderson acceleration*$(0.6, \kappa)$, where κ is the number in the legends, without projection to admissible values.

Studies for formal Newton methods. The results for this approach, displayed in Fig. 4.32, are similar as for the 3d Hemker problem, Fig. 4.27. Also here, the *formal Newton* methods usually do not show a notably better behavior than the *mixed fixed point*(0.6) method.

4.3.2.6 BJK Limiter with P_1 Finite Elements

For the BJK limiter, results for the method *mixed fixed point*$(\omega_{\rm fp})$ are presented in Fig. 4.33. On the one hand, there is a similar behavior as for the Kuzmin limiter, because the method

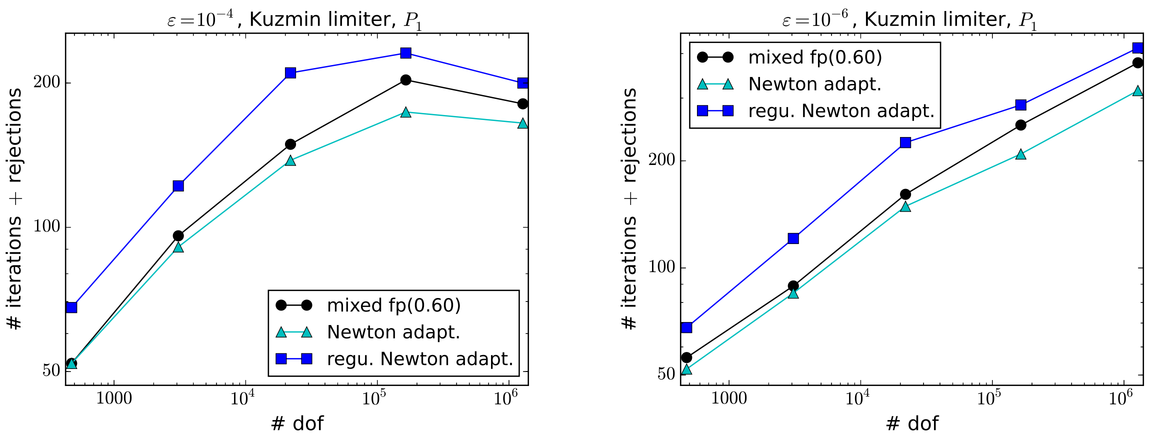


Figure 4.32: 3d problem with non-constant convection. Results for the formal Newton methods, without projection to admissible values. The adaptive methods were used with $\omega_{\text{fp}} = 0.6$ and $\omega_{\text{Newt}} = 0.1$.

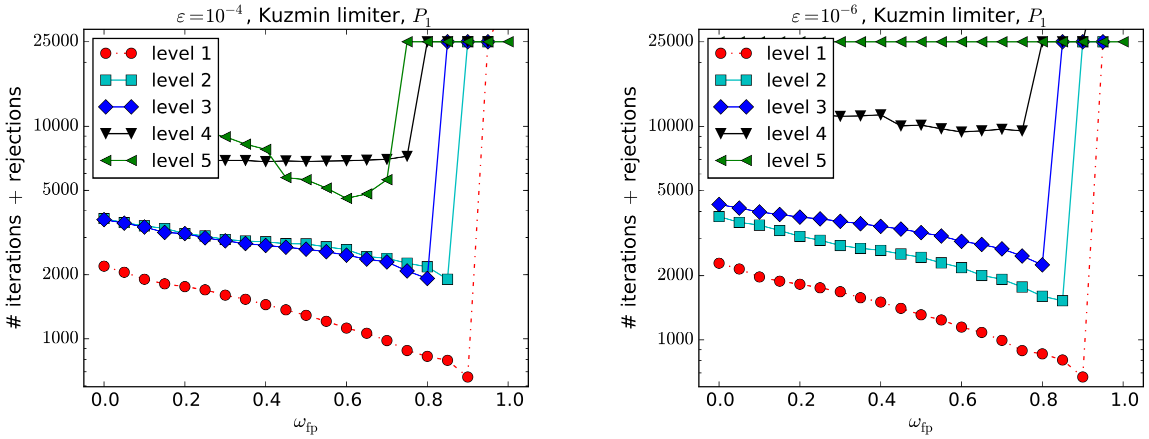


Figure 4.33: 3d problem with non-constant convection. Results for the method *mixed fixed point*(ω_{fp}), without projection to admissible values. Diverged iterations: $\varepsilon = 10^{-6}$: level 1 with $\omega_{\text{fp}} = 1$; $\varepsilon = 10^{-4}$ and $\varepsilon = 10^{-6}$: level 2 with $\omega_{\text{fp}} = 1$, level 3 with $\omega_{\text{fp}} = 1$, level 4 with $\omega_{\text{fp}} = 1$, level 5 with $\omega_{\text{fp}} \in \{0.95, 1\}$.

converges if the mixing parameter ω_{fp} is sufficiently small. On the other hand, much more iterations are needed than for the Kuzmin limiter.

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 numbers are for levels 1–3: 171, 401, 598 in comparison with the best numbers from Fig. 4.33: 706, 1574, 2298. Thus, on levels 2 and 3 there is a

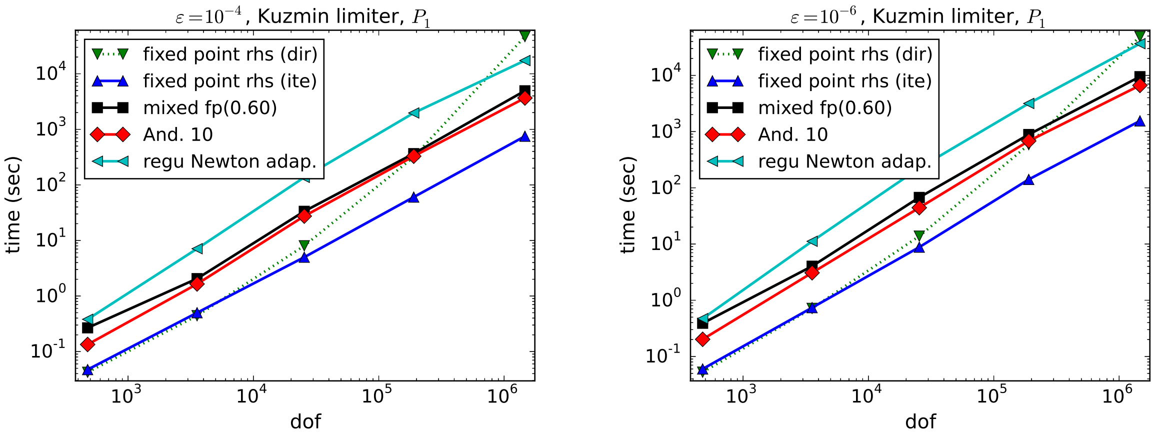


Figure 4.34: 3d problem with non-constant convection. Efficiency for several methods.

considerable reduction of these numbers.

4.3.2.7 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 [SS86] was used with right preconditioner. The preconditioner was SSOR with relaxation parameter 1.0. In our experience, it is in general 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 stronger stopping criterion has been already provided in Sec. 4.3.2.2. 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 Sec. 4.3.1.11.

Results are displayed in Fig. 4.34. 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 $\tilde{\mathbb{A}} = \mathbb{A} + \mathbb{D}$, much more efficient than for the matrices from the other methods.

4.4 Summary

This chapter presented comprehensive numerical studies for solving the nonlinear problems arising in AFC discretizations of steady-state convection-diffusion-reaction equation.

Taking the simplest fixed point method *fixed point rhs*, or equivalently *mixed fixed point*(0), as a reference method, 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 Secs. 4.3.1.11 and 4.3.2.6. 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 Sec. 4.2.3 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. For all methods, the choice of the initial iterate did in general not possess a big impact on the number of iterations. Usually, using the SUPG solution was an appropriate choice.

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.

5 A Posteriori Error Estimation for AFC Schemes

In Chapter 3 it was noted that the solution of Convection-Diffusion-Reaction equations changes abruptly in the layers and hence some kind of stabilization was required. An approach to approximate the layers properly and reduce the number of unknowns is the use of highly non-equidistant meshes instead of equidistant (or uniform) meshes. Now, one can use a priori non-equidistant meshes based on the knowledge of the exact solution (e.g. graded meshes [Bah69], Shishkin meshes [MOS96, FHM$^+$00]), or one may begin with some uniform mesh, compute a numerical solution on it, and then use information from this to adapt the grid in an a posteriori way, thereby obtaining a grid more suited to the problem. This technique is referred as *adaptive methods based on a posteriori error estimation*. Interest in a posteriori error estimation for FEMs for two point boundary value problems began with the pioneering work of Babuška and Rheinboldt [BR78]. In the review [Sty05] the author prophesizes that adaptive methods will triumph over other methods to solve Convection-Diffusion-Reaction equations.

From the past three decades, a posteriori error estimation for Convection-Diffusion-Reaction equations has received a lot of attention. A review of all the estimators proposed for these equations is beyond the scope of this work, but some examples of estimators obtained using different techniques can be found in [Ver98, APS05, San08, JN13]. One of the initial studies for the comparison of different estimators using the SUPG solution of Convection-Diffusion-Reaction equations was done in [Joh00] and it was shown that none of the estimators was robust with respect to the diffusion coefficient, ε. By robustness, we mean that the equivalence constants between the estimator and the error should be independent of how much convection-dominated the problem is. Work towards deriving a robust estimator was proposed in [Ver05] where the analysis from [Ver98] was extended by adding a dual norm of the convective derivative to the energy norm, but the additional term in the norm can only be approximated. A generalization of the robust estimators was considered in [TV15], where the analysis was applied to linear stabilized schemes. Robust a posteriori error estimators for $L^1(\Omega)$ and $L^2(\Omega)$ norm of the error can be found in [HDF$^+$06, HFD08, HDF11]. In [JN13] a robust estimator is proposed in the same norm in which the a priori analysis is performed for the SUPG method, namely the SUPG norm. Here the analysis relied on certain hypotheses including an interpolation of the solution.

One of the drawbacks of all the above-mentioned estimators is the presence of certain constants which can only be approximated. Results related to find a fully computable upper bound for the error of convection-diffusion equations have gained attention recently and can be found in [AABR13, ESV10]. For the algebraic flux correction schemes (AFC), a fully computable estimator was proposed in [ABR17] with respect to the energy norm. To the best of our knowledge, this was the first work, where an a posteriori error estimator has been derived for the AFC schemes. It is shown that the estimator is not robust with respect to ε and also the local efficiency of the scheme relied on certain assumptions including the Lipschitz continuity of the nonlinear term and the linearity preservation of the scheme.

In this chapter, our focus will be on the study of adaptive methods in the context of AFC schemes with respect to the energy norm. The contents of the chapters are as follows: Sec. 5.1 introduces certain notations and definitions used in a posteriori analysis, namely the properties of the triangulation, refinement techniques, and certain auxiliary results. In Sec. 5.2 a global upper bound and a local lower bound are derived for the error in the energy norm. Here, we also present another strategy for deriving an upper bound using the SUPG solution. Lastly, numerical simulations validating the results are presented in Sec. 5.3.

5.1 Preliminaries

Let us recall the Algebraic Flux Correction (AFC) scheme introduced in Sec. 3.4, Chapter 3. The AFC scheme for (2.6) reads as (see [BJK16]): Find $u_h \in W_h(\subseteq C(\overline{\Omega}) \cap H^1_D(\Omega))$ such that

$$a_{\mathrm{AFC}}(u_h; u_h, v_h) = \langle f, v_h\rangle + \langle g, v_h\rangle_{\Gamma_N} \qquad \forall v_h \in V_h(\subseteq C(\overline{\Omega}) \cap H^1_0(\Omega)), \tag{5.1}$$

with $a_{\mathrm{AFC}}(\cdot,\cdot) : H^1_D(\Omega) \times H^1_0(\Omega) \to \mathbb{R}$ such that

$$a_{\mathrm{AFC}}(u_h; u_h, v_h) := a(u_h, v_h) + d_h(u_h, u_h, v_h),$$

where

$$d_h(w; u, v) = \sum_{i,j=1}^{N} (1 - \alpha_{ij}(w)) d_{ij}(u(x_j) - u(x_i)) v(x_i) \qquad \forall u, v, w \in C(\overline{\Omega}), \tag{5.2}$$

$a(u_h, v_h)$ is given by (2.12), and $H^1_D(\Omega) = \{v \in H^1(\Omega) : v|_{\Gamma_D} = u^b\}$. For our analysis we will be assuming homogeneous Dirichlet conditions, i.e., $u^b = 0$.

In [BJKR18] a different representation of $d_h(\cdot;\cdot,\cdot)$ is given for conforming piecewise linear finite element functions u and v, which reads as

$$d_h(w; u, v) = \sum_{E \in \mathcal{E}_h} (1 - \alpha_E(w)) |d_E| h_E (\nabla u \cdot \boldsymbol{t}_E,\ \nabla v \cdot \boldsymbol{t}_E)_E, \tag{5.3}$$

where $\mathcal{E}_h$ is the set of all edges and $\boldsymbol{t}_E$ is the tangential unit vector on edge E.

For $u, v, w, u_1, u_2 \in C(\overline{\Omega})$ we have the following properties of $d_h(\cdot;\cdot,\cdot)$ (see [BJK16]),

1. *Non-negativity*: $0 \leq d_h(w; v, v)$.

2. *Linearity*:
$$\begin{aligned} d_h(w; u_1 + u_2, v) &= d_h(w; u_1, v) + d_h(w; u_2, v), \\ d_h(w; v, u_1 + u_2) &= d_h(w; v, u_1) + d_h(w; v, u_2). \end{aligned} \tag{5.4}$$

3. *Semi-Norm property, Cauchy-Schwarz inequality*:
$$d_h(w; u, v) \leq d_h^{1/2}(w; u, u) d_h^{1/2}(w; v, v). \tag{5.5}$$

We will present our result with respect to the energy norm given by

$$\|v\|_a^2 = \varepsilon |v|_{1,\Omega}^2 + \sigma_0 \|v\|_{0.\Omega}^2 \quad \forall v \in H^1(\Omega). \tag{5.6}$$

We would also like to mention the induced AFC norm of the system which is used for its a priori analysis ([BJK16, BJK17]) and which is the starting point of our a posteriori analysis,

$$\|v\|_{\mathrm{AFC}}^2 = \|v\|_a^2 + d_h(v_h, v, v) \quad \forall v \in H^1(\Omega). \tag{5.7}$$

5.1.1 Definitions and Notations

For $d > 1$, the domain $\Omega \subset \mathbb{R}^d$ is decomposed into (simple) subdomains for which local polynomials are defined. These decompositions are referred as *grids* or *meshes*. A simplicial decomposition, i.e., a decomposition consisting only of triangles or tetrahedron is called a triangulation.

Definition 5.1. (**Grid or Mesh**) ([DW11, Definition 4.11]) Let $\Omega \subset \mathbb{R}^d$, $d \in \{2, 3\}$, denote a domain and $\mathbb{S}$ be a finite system of closed connected sets of subdomains of $\overline{\Omega}$. A subset $\mathcal{T} \subset \mathbb{S}$ is called *conformal* if for all $K_1, K_2 \in \mathcal{T}$ with $K_1 \cap K_2 \in \mathcal{T}$ also $K_1 = K_2$ holds. Let $\mathcal{T} = \{K \in \mathbb{S} : \mathrm{int}(K) \neq \emptyset\}$ denotes the set of all elements. $\mathbb{S}$ is called a grid or mesh whenever the following property holds:

1. $\mathcal{T}$ covers Ω, i.e., $\overline{\Omega} = \underset{K \in \mathcal{T}}{\cup} K$,

2. $\mathcal{T}$ is conformal,

3. $\mathbb{S} \cup \partial\Omega$ is closed under intersection of sets, i.e., $K_1, K_2 \in \mathbb{S} \cup \partial\Omega \Rightarrow K_1 \cap K_2 \in \mathbb{S}$.

If for $F \in \mathbb{S}$ there exist exactly two $K_1, K_2 \in \mathcal{T}$ with $F = K_1 \cap K_2$, then F is called a *face*. We denote by $\mathcal{F}_h$ the set of all faces which are $m-$dimensional linear manifolds, $0 \leq m \leq d-1$. For $d = 3$ an element $E \in \mathbb{S}$ is called an edge of some $K \in \mathcal{T}$ if exactly two $F_1, F_2 \in \mathcal{F}_h,\ F_1, F_2 \subset K$ exist so that $E = F_1 \cap F_2$. We denote by $\mathcal{E}_h$ the set of all edges. Finally, $\mathcal{N}_h = \{E_1 \cap E_2 : E_1, E_2 \in \mathcal{E}_h,\ E_1 \neq E_2\}$ is the set of *vertices*. A grid is conformal if $\mathcal{E}_h$ and $\mathcal{F}_h$ are conformal. Note that $\mathcal{F}_h = \mathcal{F}_{h,\Omega} \cup \mathcal{F}_{h,D} \cup \mathcal{F}_{h,N}$, where $\mathcal{F}_{h,\Omega}$, $\mathcal{F}_{h,D}$, and $\mathcal{F}_{h,N}$ denote the interior, Dirichlet, and Neumann faces respectively. In 2d, it holds that $\mathcal{E}_h = \mathcal{F}_h$. The set of mesh cells having a common face F is denoted by $\omega_F = \cup_{F \subset \partial K'} K'$ and ω_K denotes the patch of mesh cells that have a joint face with K.

Let $P(\mathcal{T})$ define a finite element space on our triangulation, then the functionals that define our finite element space are referred as *nodal functionals*. We denote by $N_F(\mathcal{T})$ the set of nodal functionals.

Remark 5.2. As we are concentrating on Lagrange elements each nodal functional can be determined by a point on the simplex, i.e., there is a one-to-one map between the functionals and the nodes on a simplex. By abuse of notations, we are denoting them by the same notation. If we need to make a distinction between the two, it will be explicitly stated.

Remark 5.3. A finite element space can have different number of nodes and vertices, for, e.g., P_2 Lagrange elements on simplices.

Definition 5.4. (**Conforming triangulation**) A triangulation $\mathcal{T}$ of Ω is called conforming if for $K_1, K_2 \in \mathcal{T}$ with $K_1 \neq K_2$ the intersection $K_1 \cap K_2$ is either empty, a vertex, an edge, or a $2-$ face of K_1 and K_2.

Definition 5.5. (**Refinement**) ([Grä11, Definition 3.3]) Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be triangulations of Ω. Then $\mathcal{T}_2$ is called a refinement of $\mathcal{T}_1$ if for all $K \in \mathcal{T}_1$ the set

$$\{K' \in \mathcal{T}_2 : K' \cap K \neq \emptyset\}$$

is a triangulation of K.

Definition 5.6. (**Grid hierarchy**) ([Grä11, Definition 3.4]) A family $\{\mathcal{T}_i\}_{i=0}^{j}$ is called a grid hierarchy on Ω if $\mathcal{T}_0$ is a conforming triangulation of Ω, and if each $\mathcal{T}_i, i = 1, \cdots, j$, is a refinement of $\mathcal{T}_{i-1}$. If the grid $\mathcal{T}_i$ is conforming we call it conforming grid refinement otherwise non-conforming grid refinement.

Remark 5.7. An interesting property of the grid hierarchy is the embedding of the set of vertices $\mathcal{N}_{hi} \subset \mathcal{N}_{h(i+1)}$.

Definition 5.8. (**Shape regularity**) Let $\mathcal{T}$ be a triangulation of Ω into simplices. We say it is *shape regular*, if there exists a constant $C_{\text{shrg}} > 0$ such that for each mesh cell $K \in \mathcal{T}$, it holds

$$\rho_K \geq C_{\text{shrg}} h_K, \tag{5.8}$$

where h_K and ρ_K denote the diameter of K and the diameter of the largest ball inside K, respectively.

Remark 5.9. The characteristic parameter of the triangulation is given by $h = \max_{K\in\mathcal{T}_h} h_K$. We use $|K|$ as symbol for the volume of a mesh cell K.

Remark 5.10 (Consequences of the shape regularity assumption (5.8)). The 2d and 3d case will be discussed separately.

2d case. Denote the edges of an arbitrary triangle K by E_1, E_2, and E_3, the angle opposite the edge E_i by θ_i, and the length of E_i by h_{E_i}, $i = 1, 2, 3$. Then, the diameter of the largest ball inside K can be computed by

$$\rho_K = \frac{2|K|}{h_{E_1} + h_{E_2} + h_{E_3}}.$$

Hence, for a given triangulation, one can compute ρ_K/h_K for each mesh cell, such that on gets information on the constant C_{shrg}. Likewise, it is

$$\rho_K = \frac{h_{E_1}}{\cot\frac{\theta_2}{2} + \cot\frac{\theta_3}{2}}$$

and similarly for the other edges. Since $\theta_2 > 0$, $\theta_3 > 0$, and $\theta_2 + \theta_3 < \pi$, one can check that the denominator is larger than 2 such that $\rho_K < h_{E_1}$ and similarly for the two other edges:

$$h_{E_i} > \rho_K, \quad h_{E_i} \geq C_{\text{shrg}} h_K, \quad i = 1, 2, 3. \tag{5.9}$$

In 2d, the shape regularity condition (5.8) is equivalent with the minimal angle condition, i.e., there is a minimal angle $\theta_0 > 0$ for all triangles and all triangulations from the family of triangulations (see [Cia78, Pg. 130, 3.1.3]). The minimal angle condition implies a maximal angle condition. Altogether, there is a positive constant $C_{\cos} < 1$ such that for all $\mathcal{T}_h$ and all $K \in \mathcal{T}_h$

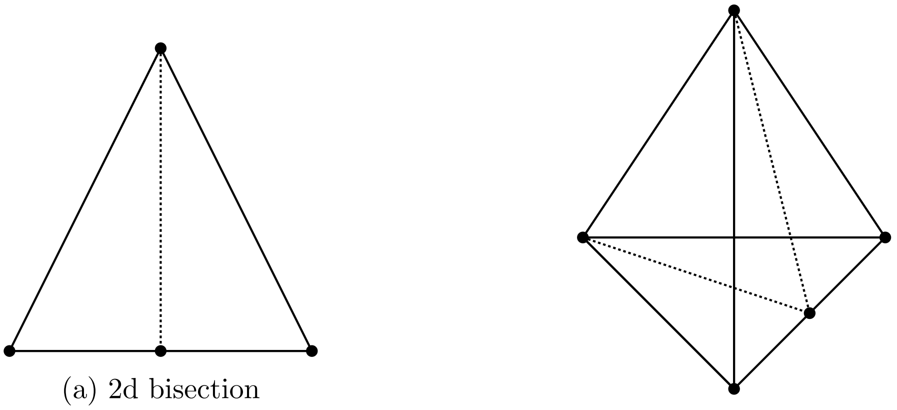
$$\cos(\theta_i) \leq C_{\cos} \quad i = 1, 2, 3. \tag{5.10}$$

For a given triangulation, $C_{\cos}$ can be computed.

3d case. In [BKK08] it is shown that for 3d, shape regularity implies a minimum angle condition and hence (5.10) holds in 3d as well with θ_i replaced by θ_{ij}, the dihedral angle between the faces F_i, F_j. Also $\rho_K < h_E$, because the projection of the ball onto the surface of the tetrahedron gives the result.

Now we will discuss some strategies for grid refinement. After the application of an a posteriori error estimator we have some *marked cells* that need to be refined. The marking of the cells is done using marking strategies which will be described later in Sec. 5.3. One of the popular ways of refining a triangle is by dividing its edges.

Definition 5.11. (**Bisection method**) The simplest method introduced by Rivara in 1984 [Riv84] is the decomposition of a cell into two neighboring cells by bisecting an edge of the marked cell and joining it to its opposing vertex. In order to maintain shape regularity, the


(a) 2d bisection

(b) 3d bisection

Figure 5.1: Bisection in 2d and 3d.

longest edge is usually selected for bisection. To close the grid, refinement is continued as long as we don't have hanging vertices. In general, the refinement process terminates before the mesh has been uniformly refined. If a hanging vertex is situated on the *longest edge*, then the further bisection does not guarantee a new node. Otherwise, a new hanging vertex is generated on a long edge. Hence, the continuation of the refinement can only affect elements with longer edges than the currently considered longest edge. Fig. 5.1 shows the method in 2d as well as 3d.

Definition 5.12. (**Red-Green refinement**) A more complicated method dates back to Bank et.al [BSW83] who have introduced some *red-green* refinement strategy in their adaptive finite element package.

Red refinement: In this kind of refinement the elements are subdivided into 2^d smaller simplicies, where exactly the midpoint of all edges are introduced as new vertices. In 2d this yields an unique decomposition of a triangle into four similar smaller triangles. As the interiors do not change, the shape regularity is preserved.

In 3d, things become complex geometrically: first one gets four tetrahedron at the vertices of the original tetrahedron as well as an octahedron in the center. By selecting one of the diagonals as common new edge the octahedron may be decomposed into four further tetrahedron, which, however are not similar to the original tetrahedron. The selection of the diagonal is important so as to preserve shape regularity. As a rule the shortest diagonal is selected. In order to preserve conformity of the grid or to close the grid, elements neighboring already subdivided elements must be subdivided. This strategy is not useful as this will refine the whole grid uniformly. One remedy is to red refine locally only if (in 2d) at least two edges of a triangle are marked.

Green refinement: We now assume (in 2d) that in a local triangle one edge has a hanging vertex and the grid needs to be closed. In this case we introduce a new so-called green edge and subdivide the neighboring triangle into two triangles. In this way the refinement is not

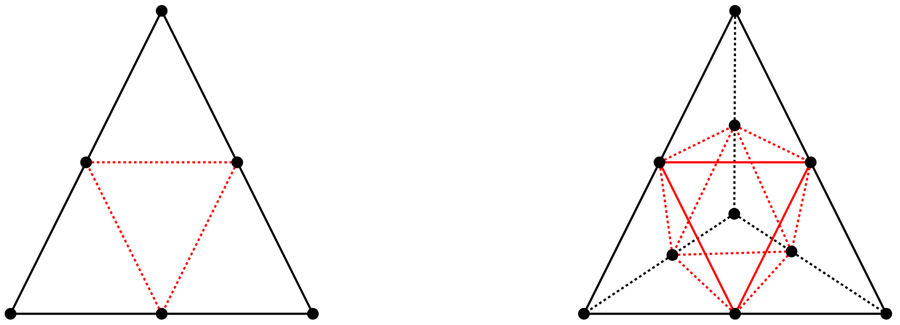

(a) 2d red refinement
(b) 3d red refinement

Figure 5.2: Red refinement in 2d and 3d.

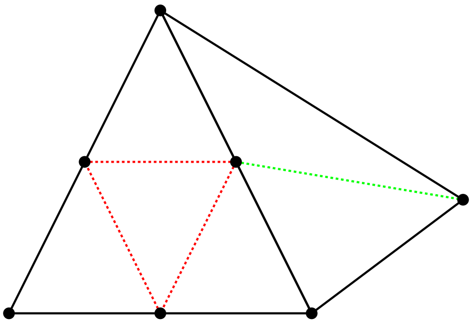


Figure 5.3: Green completion not preserving shape regularity.

continued further which is why the green edges are called *green completion*. One drawback of green refinement is that the interior angles deteriorate from the original triangle. Hence by continuing green refinement the shape regularity may suffer, see Fig. 5.3. In order to preserve shape regularity for subsequent refinements the green refinements are removed before the next refinement process and then performing green completion once the red refinement has been performed, see Fig. 5.4. In 3d, the process runs in a similar way, but requires three different type of green completion depending on whether one, two, or three edges are marked.

Definition 5.13. (**Blue refinement**) Introduced in [KR89] blue refinement (in 2d) is performed by bisecting exactly two edges. To avoid too acute or obtuse triangles, the longest one of the refinement edges is bisected first. Fig. 5.5 shows blue refinement in 2d.

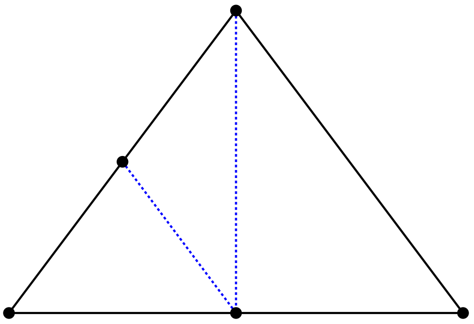
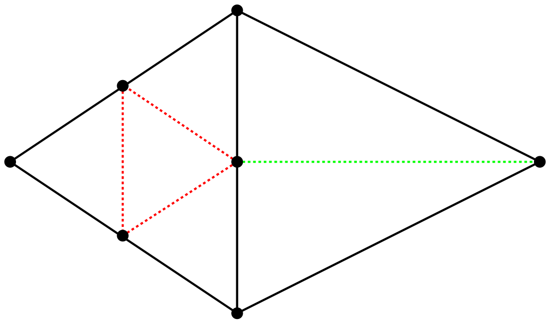
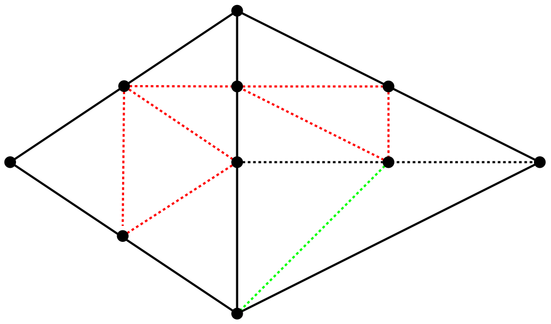


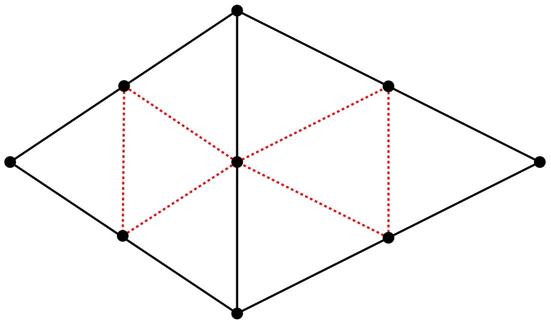
Figure 5.5: Blue refinement.



(a) Initial green refinement



(b) Green refinement leading to bad angles



(c) Substituted red refinement

Figure 5.4: Substituted red refinement before green completion.

Remark 5.14. The preservation of shape regularity by red refinement of simplices by arbitrary space dimension in finitely many similar classes was proven in 1942 by H. Freundenthal [Fre42] for the adaptive mesh in 3d, however rediscovered only 50 years later [Bey95, Ong89]. The stability for selecting the shortest diagonal has been shown in [KP08, Zha95].

Lastly, we would like to define *bubble functions* on our finite element space as they play an important role while finding local lower bounds.

Definition 5.15. (**Bubble functions**) ([AO00, Sec. 2.3.1]) Let the triangular reference element be chosen as

$$\hat{K}\{(\hat{x},\hat{y}) : 0 \leq \hat{x} \leq 1;\ 0 \leq \hat{y} \leq 1 - \hat{x}\}$$

and introduce the barycentric coordinates on the reference elements defined by

$$\hat{\varphi}_1 = \hat{x}; \quad \hat{\varphi}_2 = \hat{y}, \quad \hat{\varphi}_3 = 1 - \hat{x} - \hat{y}.$$

The *interior bubble function* $\hat{\psi}_{\hat{K}}$ is defined by

$$\hat{\psi}_{\hat{K}} = 27\hat{\varphi}_1\hat{\varphi}_2\hat{\varphi}_3$$

and the three *edge bubble functions* are given by

$$\hat{\psi}_{\hat{F}1} = 4\hat{\varphi}_2\hat{\varphi}_3; \quad \hat{\psi}_{\hat{F}2} = 4\hat{\varphi}_1\hat{\varphi}_3; \quad \hat{\psi}_{\hat{F}3} = 4\hat{\varphi}_1\hat{\varphi}_2.$$

For each element $K \in \mathcal{T}$ let $\mathcal{F}_K : \hat{K} \to K$ be the affine mapping [BS08, Sec. 3.4], then define the *bubble functions* on element K by

$$\psi_K = \hat{\psi}_{\hat{K}} \circ \mathcal{F}_K^{-1}; \quad \psi_F = \hat{\psi}_{\hat{F}} \circ \mathcal{F}_K^{-1}.$$

The concept of bubble functions can be extended to quadrilaterals (see [AO00, Sec. 2.3.1]) and higher-dimensional simplices and cubes (see [Ver13, Sec. 3.6]).

5.1.2 Auxiliary Results

In this subsection we would mention certain standard results used for a posteriori error estimation. We would also give some concrete choices of constants in certain trace results. We will assume that the triangulations are regular.

Lemma 5.16. (**Cauchy-Schwarz inequality**) *Let $(\cdot,\cdot)_V$ be an inner product on V and $\|\cdot\|_V$ be the induced norm on V, then for $u, v \in V$*

$$(u, v)_V \leq \|u\|_V \|v\|_V.$$

Lemma 5.17. (**Generalized Young's inequality**) *Let $a, b \in \mathbb{R}^+ \cup \{0\}$ and $p, q > 1$ such that $1/p + 1/q = 1$. Then*

$$ab \leq \frac{a^p}{C_Y p} + \frac{C_Y b^q}{q},$$

where $C_Y > 0$.

Lemma 5.18. (**Inverse estimate**) ([BS08, Lemma 4.5.3]) *Let $\rho h \leq h_K \leq h$, where $0 < h \leq 1$, and V_h be a finite-dimensional subspace of $H^m(K)$. Then for $0 \leq l \leq m$ there exists a constant C_{inv} such that for all $v \in V_h$ and $K \in \mathcal{T}_h$, we have*

$$\|v_h\|_{m,K} \leq C_{\text{inv}} h_K^{l-m} \|v_h\|_{l,K}. \tag{5.11}$$

Theorem 5.19. (**Interpolation estimate**) ([Cia78, Theorem 3.1.6]) *Let $q \in [1, \infty]$, $s \in$*

$\{0,1\}$ *and* $s \le t$. *Let,* $I_h : W^{t,q}(\Omega) \to V_h$ *denote a bounded linear interpolation operator. Then, it satisfies* $\forall v \in W^{t,q}(\Omega)$ *and all mesh cells* $K \in \mathcal{T}_h$

$$|v - I_h v|_{s,q,K} \le C_I h_K^{t-s} |v|_{t,q,K}. \tag{5.12}$$

Remark 5.20. For $s = t$ in (5.12), one gets with $u_h = I_h u_h$

$$\begin{aligned} \|u - I_h u\|_{s,q,K} &\le \|u - u_h\|_{s,q,K} + \|I_h u - I_h u_h\|_{s,q,K} \\ &\le \|u - u_h\|_{s,q,K} + C_I \|u - u_h\|_{s,q,K} \\ &= (1 + C_I)\|u - u_h\|_{s,q,K}. \end{aligned} \tag{5.13}$$

Remark 5.21. We assume a stable quasi-interpolation (similar to [JN13, Eq. (6)]) which is identity on the finite element space, i.e.,

$$I_h u_h = u_h \quad \forall\ u_h \in V_h.$$

A trace inequality which relates the $L^2(F)$ norm on a face of a mesh cell K to norms defined on K was proved in [Ver98].

Lemma 5.22. ([Ver98, Lemma 3.1]) *Let* $v \in H^1(K)$ *and* $F \subset \partial K$, *then it holds*

$$\|v\|_{L^2(F)} \le C \left(h_F^{-1/2} \|v\|_{L^2(K)} + \|v\|_{L^2(K)}^{1/2} \|\nabla v\|_{L^2(K)}^{1/2} \right). \tag{5.14}$$

Lemma 5.23. *Let* E *be an edge with length* h_E *and* v *be a linear function on* E, *then*

$$\|\nabla v \cdot \boldsymbol{t}_E\|_{0,E}^2 \le \|\nabla v\|_{0,E}^2, \tag{5.15}$$

where $\boldsymbol{t}_E$ *is the tangent unit vector to* E.

Proof. From orthogonal decomposition one has

$$\nabla v = (\nabla v \cdot \boldsymbol{t}_E)\boldsymbol{t}_E + (\nabla v \cdot \boldsymbol{n}_E)\boldsymbol{n}_E$$

where $\boldsymbol{n}_E$ is the normal unit vector to E. Now, one knows $\boldsymbol{n}_E \cdot \boldsymbol{t}_E = 0$.

So,

$$\nabla v \cdot \nabla v = (\nabla v \cdot \boldsymbol{t}_E)^2 + (\nabla v \cdot \boldsymbol{n}_E)^2 + 2(\nabla v \cdot \boldsymbol{t}_E)(\nabla v \cdot \boldsymbol{n}_E)\boldsymbol{n}_E \cdot \boldsymbol{t}_E.$$

Integrating on both sides along the edge E,

$$\|\nabla v\|_{0,E}^2 = \|\nabla v \cdot \boldsymbol{t}_E\|_{0,E}^2 + \|\nabla v \cdot \boldsymbol{n}_E\|_{0,E}^2.$$

Hence, (5.15) follows. □

Lemma 5.24 (Estimate of the trace on an edge by the norm on the mesh cell)**.** *Let $K \in \mathcal{T}$ be a mesh cell, $\mathcal{E}_h(K)$ the set of all edges of K and $\varphi_h \in P_1(K)$. Then, it holds*

$$\sum_{E \in \mathcal{E}_h(K)} \|\nabla \varphi_h \cdot \boldsymbol{t}_E\|^2_{L^2(E)} \le C_{\mathrm{edge}} h_K^{1-d} \|\nabla \varphi_h\|^2_{L^2(K)}, \tag{5.16}$$

with C_{edge} independent of K.

Proof. The principal way for proving the statement of the lemma is the same for two and three dimensions. It uses the mapping to the reference cell. First, the proof for $d = 2$ will be presented.

Relating the norms on E and $\hat{E}$. This step is just a one-dimensional consideration for an edge. Thus, one has to do the same calculations in 2d and 3d. For brevity, the presentation below is performed for the 2d case.

Let $\hat{K}$ be the reference triangle with the vertices $\hat{V}_0 = (0,0)$, $\hat{V}_1 = (1,0)$, and $\hat{V}_2 = (0,1)$. Since a additive constant does not play any role, it will be assumed that $\hat{\varphi}_h(\hat{V}_0) = 0$, $\hat{\varphi}_h(\hat{V}_1) = \alpha$, and $\hat{\varphi}_h(\hat{V}_2) = \beta$ with $\alpha, \beta \in \mathbb{R}$. Consequently, it is $\nabla \hat{\varphi}_h = (\alpha, \beta)^T$. One obtains for $\hat{E} = \overline{\hat{V}_0 \hat{V}_1}$ and $h_{\hat{E}} = |\hat{E}| = 1$

$$\int_{\hat{V}_0}^{\hat{V}_1} (\nabla \hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}})^2 \, ds = \left(\frac{(\hat{\varphi}_h(\hat{V}_1) - \hat{\varphi}_h(\hat{V}_0))^2}{h_{\hat{E}}^2} \right) h_{\hat{E}} = \alpha^2. \tag{5.17}$$

Analogously, one finds

$$\int_{\hat{V}_0}^{\hat{V}_2} (\nabla \hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}})^2 \, ds = \beta^2, \quad \int_{\hat{V}_0}^{\hat{V}_2} (\nabla \hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}})^2 \, ds = \frac{1}{\sqrt{2}} (\alpha - \beta)^2. \tag{5.18}$$

Let the reference map $\mathcal{F}_K : \hat{K} \to K$ map $\hat{V}_0$ to V_0 and $\hat{V}_1$ to V_1, where V_0 and V_1 are vertices of K. Then it holds that $\hat{\varphi}_h(\hat{V}_0) = \varphi_h(V_0)$ and $\hat{\varphi}_h(\hat{V}_1) = \varphi_h(V_1)$. Denote $E = \overline{V_0 V_1}$, then it is

$$\int_{V_0}^{V_1} (\nabla \varphi_h \cdot \boldsymbol{t}_E)^2 \, ds = \left(\frac{(\varphi_h(V_0) - \varphi_h(V_1))^2}{h_E^2} \right) h_E.$$

The value of this integral has to be equal to (5.17), from what follows that

$$\|\nabla \varphi_h \cdot \boldsymbol{t}_E\|^2_{L^2(E)} = \frac{h_{\hat{E}}}{h_E} \|\nabla \hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|^2_{L^2(\hat{E})}.$$

Performing the same considerations for the other two edges, one obtains with (5.18)

$$\|\nabla \varphi_h \cdot \boldsymbol{t}_E\|^2_{L^2(E)} \le \frac{\sqrt{2}}{h_E} \|\nabla \hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|^2_{L^2(\hat{E})}. \tag{5.19}$$

2d: Estimate on the reference cell. Using (5.17), (5.18) and Young's inequality yields

$$\begin{aligned}\sum_{\hat{E}\subset\partial\hat{K}} \|\nabla\hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|^2_{L^2(\hat{E})} &= \alpha^2+\beta^2+\frac{1}{\sqrt{2}}(\alpha-\beta)^2 \\ &\le \left(1+\sqrt{2}\right)(\alpha^2+\beta^2).\end{aligned}$$

Since

$$\int_{\hat{K}} (\nabla\hat{\varphi}_h \cdot \nabla\hat{\varphi}_h)\, d\boldsymbol{x} = \frac{1}{2}(\alpha^2+\beta^2), \tag{5.20}$$

one obtains

$$\sum_{\hat{E}\subset\partial\hat{K}} \|\nabla\hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|^2_{L^2(\hat{E})} \le 2\left(1+\sqrt{2}\right)\|\nabla\hat{\varphi}_h\|^2_{L^2(\hat{K})}. \tag{5.21}$$

3d: estimate on the reference cell. The reference cell is given by the vertices $\hat{V}_0=(0,0,0)$, $\hat{V}_1=(1,0,0)$, $\hat{V}_2=(0,1,0)$, and $\hat{V}_2=(0,0,1)$. A linear function $\hat{\varphi}_h$ is considered with $\hat{\varphi}_h(\hat{V}_0)=0$, $\hat{\varphi}_h(\hat{V}_1)=\alpha$, $\hat{\varphi}_h(\hat{V}_2)=\beta$, and $\hat{\varphi}_h(\hat{V}_3)=\gamma$. Performing very similar calculations as in the 2d case leads to the estimate

$$\sum_{\hat{E}\subset\partial\hat{K}} \|\nabla\hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|^2_{L^2(\hat{E})} \le 6\left(1+\sqrt{2}\right)\|\nabla\hat{\varphi}_h\|^2_{L^2(\hat{K})}. \tag{5.22}$$

Relating the norms on $\hat{K}$ and K. From the standard numerical analysis it is known that there is a constant C which is independent of K, such that

$$\|\nabla\hat{\varphi}_h\|^2_{L^2(\hat{K})} \le C h_K^{2-d}\|\nabla\varphi_h\|^2_{L^2(K)}. \tag{5.23}$$

Estimate (5.16) is now obtained by combining (5.19), (5.21) or (5.22), and (5.23), and using the shape regularity of the mesh cell (5.9). □

Remark 5.25 (More detailed estimate in 2d). Let φ_h be a linear function on K with $\varphi_h(V_0)=0$, $\varphi_h(V_1)=\alpha$, and $\varphi_h(V_2)=\beta$, and (x_0,y_0), (x_1,y_1), and (x_2,y_2) be the coordinates of V_0, V_1, and V_2 respectively. Then the standard Hessian form of the plane on K is given by

$$\varphi_h = -\left(a_4+\frac{a_1x}{a_3}+\frac{a_2y}{a_3}\right),$$

where $a_1=(y_1-y_0)\beta-(y_2-y_0)\alpha$, $a_2=(x_2-y_0)\alpha-(x_1-x_0)\beta$, $a_3=(x_1-x_0)(y_2-y_0)-(x_2-x_0)(y_1-y_0)$, and a_4 is a constant which can be computed by a point on the plane. Now

$$\nabla\varphi_h = -\frac{1}{a_3}\begin{pmatrix}a_1\\a_2\end{pmatrix} = -\frac{1}{2|K|}\begin{pmatrix}a_1\\a_2\end{pmatrix}$$

A direct calculation gives that

$$\nabla\varphi_h \cdot \nabla\varphi_h = \frac{1}{4|K|^2}\left(\alpha^2 h_{E_2}^2 + \beta^2 h_{E_1}^2 - 2\alpha\beta h_{E_1} h_{E_2}\cos(\theta_0)\right),$$

where E_1 and E_2 are the edges joining (x_0, y_0) with (x_1, y_1) and (x_2, y_2), respectively and θ_0 is the angle between the two edges.

Using the condition (5.10) on the maximal cosine, Young's inequality, the shape regularity (5.9), and (5.20) yields

$$\begin{aligned}
\|\nabla\varphi_h\|_{L^2(K)}^2 &\geq \frac{1}{4|K|}\left(\alpha^2 h_{E_2}^2 + \beta^2 h_{E_1}^2 - 2C_{\cos}|\alpha||\beta| h_{E_1} h_{E_2}\right) \\
&\geq \frac{1}{4|K|}\left(\alpha^2 h_{E_2}^2 (1 - C_{\cos}) + \beta^2 h_{E_1}^2 (1 - C_{\cos})\right) \\
&\geq \frac{1 - C_{\cos}}{4|K|}\rho_K^2\left(\alpha^2 + \beta^2\right) \\
&= \frac{1 - C_{\cos}}{2|K|}\rho_K^2 \|\nabla\hat{\varphi}_h\|_{L^2(\hat{K})}^2.
\end{aligned}$$

Combining this estimate with (5.19), (5.9), and (5.21) leads to

$$\begin{aligned}
\sum_{E\in\mathcal{E}_h(K)} \|\nabla\varphi_h \cdot \boldsymbol{t}_E\|_{L^2(E)}^2 &\leq \frac{\sqrt{2}}{\rho_K}\sum_{\hat{E}\subset\partial\hat{K}} \|\nabla\hat{\varphi}_h \cdot \boldsymbol{t}_{\hat{E}}\|_{L^2(\hat{E})}^2 \\
&\leq \frac{2\sqrt{2}\left(1+\sqrt{2}\right)}{\rho_K}\|\nabla\hat{\varphi}_h\|_{L^2(\hat{K})}^2 \\
&\leq \frac{4\sqrt{2}\left(1+\sqrt{2}\right)|K|}{(1 - C_{\cos})\rho_K^3}\|\nabla\varphi_h\|_{L^2(K)}^2.
\end{aligned}$$

The first factor on the right-hand side scales like h_K^{-1} since $\rho_K \sim h_K$ and $|K| \sim h_K^2$. For a given triangulation, it is computable.

Remark 5.26. (More detailed estimate in 3d). Let φ_h be a linear function on K with $\varphi_h(V_0) = 0, \varphi_h(V_1) = \alpha, \varphi_h(V_2) = \beta$, and $\varphi_h(V_3) = \gamma$. For linear simplical elements on K we have

$$(\nabla v_i)^\top \nabla v_j = -\frac{|F_i||F_j|}{(3|K|)^2}\cos(\theta_{ij}), \quad \text{if } i \neq j, \tag{5.24}$$

where v_i are basis functions and θ_{ij} is the dihedral angle between two faces F_i and F_j (see [KQ95]). For the case $i = j$ we can follow the same steps as for 2d, i.e., use the standard Hessian form to get the value as

$$(\nabla v_i)^\top \nabla v_i = \frac{|F_i|^2}{(3|K|)^2}.$$

Using (5.24) a direct calculation gives

$$\begin{aligned}\nabla\varphi_h \cdot \nabla\varphi_h = \frac{1}{9|K|^2}[&\alpha^2|F_1|^2 + \beta^2|F_2|^2 + \gamma^2|F_3|^2 \\ &- 2\alpha\beta|F_1||F_2|\cos(\theta_{12}) - 2\alpha\gamma|F_1||F_3|\cos(\theta_{13}) \\ &- 2\gamma\beta|F_2||F_3|\cos(\theta_{23})].\end{aligned}$$

Using Young's inequality, the shape regularity (5.9), and $\int_{\hat{K}}(\nabla\hat{\varphi}_h \cdot \nabla\hat{\varphi}_h)dx = \frac{1}{6}(\alpha^2+\beta^2+\gamma^2)$ yields

$$\|\nabla\varphi_h\|^2_{L^2(K)} \geq \frac{6}{9|K|}\rho_K^4(1-2C_0)\|\nabla\hat{\varphi}_h\|^2_{L^2(\hat{K})},$$

where

$$C_0 = \max_{1\leq i\leq j\leq 3}\cos(\theta_{ij}).$$

Combining this estimate with (5.19), (5.9), (5.22), and assuming $C_0 < 1/2$ leads to

$$\sum_{E\in\mathcal{E}_h(K)} \|\nabla\varphi_h \cdot \boldsymbol{t}_E\|^2_{L^2(E)} \leq \frac{9\sqrt{2}(1+\sqrt{2})|K|}{(1-2C_0)\rho_K^5}\|\nabla\varphi_h\|^2_{L^2(K)}. \tag{5.25}$$

The first factor scales as h_K^{-2} as $|K| \sim h_K^3$ and $\rho_K \sim h_K$.

Remark 5.27. In Remark 5.26 we assumed that $C_0 < 1/2$. This condition is not a consequence of the shape regularity but an essential argument arising in the proof. Another reformulation of this condition is that all the dihedral angles in a tetrahedron are greater than $\pi/3$. In general this condition is not satisfied, for e.g., for reference unit tetrahedron we have dihedral angles less than $\pi/3$. One example where this condition is satisfied is a regular tetrahedron with edges of equal length where the dihedral angle is 0.3918π.

5.2 A Posteriori Error Estimators

In this section, we propose a new residual-based a posteriori error estimator for the AFC schemes in the energy norm. To the best of our knowledge only one work has been done in the context of a posteriori error estimation and the AFC schemes (see [ABR17]). A fully computable upper bound has been derived under certain assumptions on the nonlinear stabilization term. In this work ideas from [AABR13] have been extended to the AFC schemes. The design of the estimator relies on introducing certain first-order consistent equilibrated fluxes and then solving a local Neumann problem to get explicit bounds. To show the local efficiency of the estimator two assumptions are made on the nonlinear stabilization ($d_h(\cdot;\cdot,\cdot)$) namely the local Lipschitz continuity and the linearity preservation. Because of the last assumption, this estimator is not applicable to the Kuzmin limiter (see [BJK16]).

The derivation of an estimator presented in this section follows the standard residual-based approach. We start with the variational formulation and use standard interpolation estimates to bound the terms. We also propose an estimator later in this section which uses the SUPG solution for bounding the error.

5.2.1 Residual-Based Estimator

5.2.1.1 Global Upper Bound

In this section we will present a global upper bound for the AFC scheme in the energy norm (5.6).

Let $u \in H^1_D(\Omega)$ be a continuous solution of (2.12) and $u_h \in W_h$ be a solution for (5.1), then for $v_h \in V_h$ one obtains with (2.12) and (5.1)

$$\begin{aligned} a_{\mathrm{AFC}}(u_h; u - u_h, v_h) &= a(u - u_h, v_h) + d_h(u_h; u - u_h, v_h) \\ &= \langle f, v_h\rangle + \langle g, v_h\rangle_{\Gamma_N} - \langle f, v_h\rangle - \langle g, v_h\rangle_{\Gamma_N} + d_h(u_h; u, v_h) \\ &= d_h(u_h; u, v_h). \end{aligned} \tag{5.26}$$

For any $v \in H^1_0(\Omega)$, the application of (5.1), (5.2), and (5.26) yields

$$\begin{aligned} & a_{\mathrm{AFC}}(u_h; u - u_h, v) \\ &= a_{\mathrm{AFC}}(u_h; u - u_h, v - I_h v) + a_{\mathrm{AFC}}(u_h; u - u_h, I_h v) \\ &= a(u - u_h, v - I_h v) + d_h(u_h; u - u_h, v - I_h v) + d_h(u_h; u, I_h v) \\ &= \langle f, v - I_h v\rangle + \langle g, v - I_h v\rangle_{\Gamma_N} + d_h(u_h; u - u_h, v - I_h v) \\ &\quad + d_h(u_h; u, I_h v) - a(u_h, v - I_h v). \end{aligned}$$

Taking $v = u - u_h$ in this equation, using $u_h = I_h u_h$, and applying integration by parts, one gets

$$\begin{aligned} & \|u - u_h\|^2_{\mathrm{AFC}} \\ &= \|u - u_h\|^2_a + d_h(u_h; u - u_h, u - u_h) \\ &= a_{\mathrm{AFC}}(u_h; u - u_h, u - u_h) \\ &= \langle f, u - I_h u\rangle + \langle g, u - I_h u\rangle_{\Gamma_N} + d_h(u_h; u - u_h, u - u_h - I_h(u - u_h)) \\ &\quad + d_h(u_h; u, I_h u - I_h u_h) - a(u_h, u - I_h u) \qquad (5.27) \\ &= \sum_{K \in \mathcal{T}_h} (R_K(u_h), u - I_h u)_K + \sum_{F \in \mathcal{F}_h} \langle R_F(u_h), u - I_h u\rangle_F \\ &\quad + d_h(u_h; u, I_h u - u_h) + d_h(u_h; u - u_h, u - u_h - I_h(u - u_h)) \end{aligned}$$

with

$$R_K(u_h) := f + \varepsilon\Delta u_h - \boldsymbol{b}\cdot\nabla u_h - cu_h|_K,$$

$$R_F(u_h) := \begin{cases} -\varepsilon[|\nabla u_h \cdot \boldsymbol{n}_F|]_F & \text{if } F \in \mathcal{F}_{h,\Omega}, \\ g - \varepsilon(\nabla u_h \cdot \boldsymbol{n}_F) & \text{if } F \in \mathcal{F}_{h,N}, \\ 0 & \text{if } F \in \mathcal{F}_{h,D}. \end{cases}$$

The terms on the right-hand side of (5.27) have to be bounded.

For the first term in (5.27), using the Cauchy–Schwarz inequality, $u_h = I_h u_h$, the interpolation estimate (5.12) with $s = 0,\ t = 0$, and the generalized Young's inequality gives

$$\begin{aligned} \sum_{K\in\mathcal{T}_h} (R_K(u_h), u - I_h u)_K &\leq \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)} \|u - I_h u\|_{L^2(K)} \\ &= \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)} \|(u-u_h) - I_h(u-u_h)\|_{L^2(K)} \\ &\leq \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)} C_I \|u-u_h\|_{L^2(K)} \\ &\leq \frac{C_Y C_I^2}{2\sigma_0} \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)}^2 + \frac{\sigma_0}{2C_Y}\|u-u_h\|_{L^2(\Omega)}^2, \end{aligned} \tag{5.28}$$

where C_Y is the Young's inequality constant.

One can also approximate the interpolation error with (5.12) and $s = 0,\ t = 1$, leading to

$$\begin{aligned} \sum_{K\in\mathcal{T}_h} (R_K(u_h), u - I_h u)_K &\leq \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)} \|u - I_h u\|_{L^2(K)} \\ &\leq \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)} C_I h_K |u-u_h|_{H^1(K)} \\ &\leq \frac{C_Y C_I^2 h_K^2}{2\varepsilon} \sum_{K\in\mathcal{T}_h} \|R_K(u_h)\|_{L^2(K)}^2 \\ &\quad + \frac{\varepsilon}{2C_Y}|u-u_h|_{H^1(\Omega)}^2. \end{aligned} \tag{5.29}$$

Hence, combining (5.28) and (5.29) gives

$$\begin{aligned} &\sum_{K\in\mathcal{T}_h} (R_K(u_h), u - I_h u)_K \\ &\quad \leq \frac{C_Y}{2} \sum_{K\in\mathcal{T}_h} \min\left\{\frac{C_I^2}{\sigma_0}, \frac{C_I^2 h_K^2}{\varepsilon}\right\} \|R_K(u_h)\|_{L^2(K)}^2 + \frac{1}{2C_Y}\|u-u_h\|_a^2. \end{aligned} \tag{5.30}$$

The estimate of the second term in (5.27) starts also with the Cauchy–Schwarz inequality and using $u_h = I_h u_h$

$$\begin{aligned}\sum_{F\in\mathcal{F}_h} \langle R_F(u_h), u - I_h u\rangle_F &\le \sum_{F\in\mathcal{F}_h} \|R_F(u_h)\|_{L^2(F)} \|u - I_h u\|_{L^2(F)} \\ &= \sum_{F\in\mathcal{F}_h} \|R_F(u_h)\|_{L^2(F)} \|(u-u_h) - I_h(u-u_h)\|_{L^2(F)}.\end{aligned}$$

Now, the local trace estimate (5.14) is applied to the second factor on the right-hand side. After this, one proceeds essentially as for the mesh cell residual by using the interpolation estimate (5.12), considering the cases $s = t = 0$ and $s = 0, t = 1$ for the interpolation error in $L^2(K)$, performing some straightforward calculations, compare [JN13], and using the shape regularity of the mesh cell, to find

$$\|(u-u_h) - I_h(u-u_h)\|_{L^2(F)} \le C_F \min\left\{\frac{h_F^{1/2}}{\varepsilon^{1/2}}, \frac{1}{\sigma_0^{1/4}\varepsilon^{1/4}}\right\} \|u-u_h\|_a,$$

where the constant C_F depends on the constant from (5.14) and the interpolation constant. Applying now the generalized Young's inequality, one gets for the face residuals

$$\begin{aligned}&\sum_{F\in\mathcal{F}_h} \langle R_F(u_h), u - I_h u\rangle_F \\ &\le \frac{C_Y}{2} \sum_{F\in\mathcal{F}_h} \min\left\{\frac{C_F^2 h_F}{\varepsilon}, \frac{C_F^2}{\sigma_0^{1/2}\varepsilon^{1/2}}\right\} \|R_F(u_h)\|_{L^2(F)}^2 + \frac{1}{2C_Y}\|u-u_h\|_a^2. \qquad (5.31)\end{aligned}$$

As intermediate result, one obtains from (5.27), (5.30), and (5.31)

$$\begin{aligned}&\|u-u_h\|_a^2 + \frac{C_Y}{C_Y-1} d_h(u_h; u-u_h, u-u_h) \\ &\le \frac{C_Y^2}{2(C_Y-1)} \sum_{K\in\mathcal{T}_h} \min\left\{\frac{C_I^2}{\sigma_0}, \frac{C_I^2 h_K^2}{\varepsilon}\right\} \|R_K(u_h)\|_{L^2(K)}^2 \\ &\quad + \frac{C_Y^2}{2(C_Y-1)} \sum_{F\in\mathcal{F}_h} \min\left\{\frac{C_F^2 h_F}{\varepsilon}, \frac{C_F^2}{\sigma_0^{1/2}\varepsilon^{1/2}}\right\} \|R_F(u_h)\|_{L^2(F)}^2 \\ &\quad + \frac{C_Y}{C_Y-1} d_h(u_h; u, I_h u - u_h) + \frac{C_Y}{C_Y-1} d_h(u_h; u-u_h, u-u_h-I_h(u-u_h)). \qquad (5.32)\end{aligned}$$

We estimate the last two term in (5.32), by using (5.4) and Remark 5.21, leading to

$$\begin{aligned}&d_h(u_h; u-u_h, u-u_h-I_h(u-u_h)) + d_h(u_h; u, I_h(u-u_h)) \\ &= d_h(u_h; u-u_h, u-u_h) - d_h(u_h; u, I_h(u-u_h)) \\ &\quad + d_h(u_h; u_h, I_h(u-u_h)) + d_h(u_h; u, I_h(u-u_h))\end{aligned}$$

$$= \quad d_h(u_h; u-u_h, u-u_h) + d_h(u_h; u_h, I_h(u-u_h)). \tag{5.33}$$

Inserting this relation in (5.32) reveals that the stabilization term on the left-hand side cancels with the first term on the right-hand side of (5.33). Consequently, only the energy norm is left to be estimated.

Since $I_h u - u_h$ is linear on each edge, the second term on the right-hand side of (5.33) can be rewritten as integral over the edges, see (5.3), and estimated with the Cauchy–Schwarz inequality and the generalized Young's inequality

$$\begin{aligned}
& d_h(u_h; u_h, I_h u - u_h) \\
&= \sum_{E\in\mathcal{E}_h} (1-\alpha_E)|d_E| h_E (\nabla u_h \cdot \boldsymbol{t}_E, \nabla(I_h u - u_h)\cdot \boldsymbol{t}_E)_E \\
&\le \sum_{E\in\mathcal{E}_h} (1-\alpha_E)|d_E| h_E \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)} \\
&\le \frac{1}{2C_Y\kappa_1} \sum_{E\in\mathcal{E}_h} \varepsilon h_E^{d-1} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)}^2 \\
&\quad + \frac{C_Y\kappa_1}{2} \sum_{E\in\mathcal{E}_h} \varepsilon^{-1}(1-\alpha_E)^2 |d_E|^2 h_E^{3-d} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)}^2 .
\end{aligned} \tag{5.34}$$

The parameter κ_1 will be defined later. The second term is computable.

Consider the first term in (5.34). Denoting

$$C_{\text{edge,max}} = \max_{K\in\mathcal{T}_h} C_{\text{edge}},$$

using $h_E \le h_K$, $d-1>0$, (5.16), the triangle inequality, and (5.13) yields

$$\begin{aligned}
& \frac{1}{\kappa_1} \sum_{E\in\mathcal{E}_h} \varepsilon h_E^{d-1} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)}^2 \\
&\le \frac{\varepsilon}{\kappa_1} \sum_{K\in\mathcal{T}_h} \left(\sum_{E\in\partial K} h_E^{d-1} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)}^2 \right) \\
&\le \frac{\varepsilon}{\kappa_1} \sum_{K\in\mathcal{T}_h} C_{\text{edge}} \|\nabla(I_h u - u_h)\|_{L^2(K)}^2 \\
&\le \frac{2\varepsilon C_{\text{edge,max}}}{\kappa_1} \sum_{K\in\mathcal{T}_h} \left(\|\nabla(u-u_h)\|_{L^2(K)}^2 + \|\nabla(u - I_h u)\|_{L^2(K)}^2 \right) \\
&\le \frac{2C_{\text{edge,max}}(1+(1+C_I)^2)}{\kappa_1} \|u-u_h\|_a^2 .
\end{aligned} \tag{5.35}$$

Choosing

$$\kappa_1 = C_{\text{edge,max}}(1+(1+C_I)^2), \tag{5.36}$$

then this term multiplied with $(2C_Y)^{-1}$ can be absorbed in the left-hand side of (5.32).

An alternative estimate proceeds similarly to (5.34)

$$\begin{aligned} d_h(u_h; u_h, I_h u - u_h) \ \le\ & \frac{1}{2C_Y\kappa_2} \sum_{E\in\mathcal{E}_h} \sigma_0 h_E^{d+1} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)}^2 \\ & + \frac{C_Y\kappa_2}{2} \sum_{E\in\mathcal{E}_h} \sigma_0^{-1}(1-\alpha_E)^2 |d_E|^2 h_E^{1-d} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)}^2. \end{aligned} \tag{5.37}$$

Continuing similarly to (5.35) and using in addition the inverse inequality (5.11) leads to

$$\begin{aligned} & \frac{1}{\kappa_2} \sum_{E\in\mathcal{E}_h} \sigma_0 h_E^{d+1} \|\nabla(I_h u - u_h)\cdot \boldsymbol{t}_E\|_{L^2(E)}^2 \\ \le\ & \frac{\sigma_0}{\kappa_2} \sum_{K\in\mathcal{T}_h} C_{\text{edge}} C_{\text{inv}}^2 \|I_h u - u_h\|_{L^2(K)}^2 \\ \le\ & \frac{2C_{\text{inv}}^2 C_{\text{edge,max}}(1+(1+C_I)^2)}{\kappa_2} \|u-u_h\|_a^2. \end{aligned} \tag{5.38}$$

Choosing

$$\kappa_2 = C_{\text{inv}}^2 C_{\text{edge,max}}(1+(1+C_I)^2) \tag{5.39}$$

enables again to absorb this term multiplied with $(2C_Y)^{-1}$ in the left-hand side of (5.32). Inserting (5.33) – (5.39) in (5.32) one gets

$$\begin{aligned} & \|u-u_h\|_a^2 \\ \le\ & \frac{C_Y^2}{2(C_Y-2)} \sum_{K\in\mathcal{T}_h} \min\left\{ \frac{C_I^2}{\sigma_0}, \frac{C_I^2 h_K^2}{\varepsilon} \right\} \|R_K(u_h)\|_{L^2(K)}^2 \\ & + \frac{C_Y^2}{2(C_Y-2)} \sum_{F\in\mathcal{F}_h} \min\left\{ \frac{C_F^2 h_F}{\varepsilon}, \frac{C_F^2}{\sigma_0^{1/2}\varepsilon^{1/2}} \right\} \|R_F(u_h)\|_{L^2(F)}^2 \\ & + \frac{C_Y^2}{2(C_Y-2)} \sum_{E\in\mathcal{E}_h} \min\left\{ \frac{\kappa_1 h_E^2}{\varepsilon}, \frac{\kappa_2}{\sigma_0} \right\} (1-\alpha_E)^2 |d_E|^2 h_E^{1-d} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)}^2. \end{aligned} \tag{5.40}$$

Using standard calculus arguments one gets an optimal value of $C_Y = 4$.

The estimates are summarized in the following theorem.

Theorem 5.28 (Global a posteriori error estimate). *A global a posteriori error estimate for the energy norm is given by*

$$\|u-u_h\|_a^2 \le \eta_1^2 + \eta_2^2 + \eta_3^2, \tag{5.41}$$

where

$$\begin{aligned}
\eta_1^2 &= \sum_{K\in\mathcal{T}_h} \min\left\{\frac{4C_I^2}{\sigma_0}, \frac{4C_I^2 h_K^2}{\varepsilon}\right\} \|R_K(u_h)\|_{L^2(K)}^2,\\
\eta_2^2 &= \sum_{F\in\mathcal{F}_h} \min\left\{\frac{4C_F^2 h_F}{\varepsilon}, \frac{4C_F^2}{\sigma_0^{1/2}\varepsilon^{1/2}}\right\} \|R_F(u_h)\|_{L^2(F)}^2,\\
\eta_3^2 &= \sum_{E\in\mathcal{E}_h} \min\left\{\frac{4\kappa_1 h_E^2}{\varepsilon}, \frac{4\kappa_2}{\sigma_0}\right\}(1-\alpha_E)^2|d_E|^2 h_E^{1-d}\|\nabla u_h\cdot \boldsymbol{t}_E\|_{L^2(E)}^2,
\end{aligned}$$

with κ_1 and κ_2 defined in (5.36) *and* (5.39), *respectively.*

Proof. The proof follows by inserting $C_Y = 4$ in (5.40). □

5.2.1.2 Local Lower Bound

The posteriori estimator implied by the equation (5.41)

$$\|u-u_h\|_a^2 \le C\sum_{K\in\mathcal{T}} \eta_K^2,$$

provides a global upper bound on the discretization error up to the constant C. For using this estimator as the basis of an adaptive refinement algorithm, one wants the estimator to be efficient in the sense that C is independent of the mesh size such that

$$\eta_K^2 \le C\|u-u_h\|_{a,\omega_K}^2,$$

where ω_K is some neighborhood of K. This type of bound is important as in conjunction with (5.41) it confirms that the rate of change of estimator as the mesh size is reduced matches the behavior of the actual error. If no such estimate is available, the performance of the estimator is not optimal, and its use in the applications may result in poorly designed meshes.

To derive such a lower bound we will use the standard bubble functions argument. The idea was introduced by Verfürth in [Ver94]. Let ψ_K be the interior bubble function associated with the mesh cell K which vanish on ∂K, and let ψ_F be the face bubble function associated to the face F which vanishes on the boundary of $\omega_F = K\cup K'$, where K and K' are two mesh cells sharing the face F.

Theorem 5.29. ([AO00, Theorem 2.2]) *There exists a constant C_K such that for all $v\in V_h$*

$$C_K^{-1}\|v\|_{0,K}^2 \le (v, v\psi_K)_{0,K} \le C_K\|v\|_{0,K}^2, \tag{5.42}$$

and

$$C_K^{-1}\|v\|_{0,K} \leq \|v\psi_K\|_{0,K} + h_K\|\nabla(v\psi_K)\|_{0,K} \leq C_K\|v\|_{0,K}. \tag{5.43}$$

One can find similar estimates for the face bubble function.

Theorem 5.30. ([AO00, Theorem 2.4]) *Let $F \subset \partial K$ be a face and let ψ_F be the corresponding face bubble function. Let $V_h(F)$ be the finite-dimensional space of functions defined on F obtained by mapping $V_h(\hat{F}) \subset H^1(\hat{F})$. Then there exists a constant C_{FB} such that*

$$C_{FB}^{-1}\|v\|_{0,F}^2 \leq (v, v\psi_F)_{0,F} \leq C_{FB}\|v\|_{0,F}^2, \tag{5.44}$$

$$h_K^{-1/2}\|v\psi_F\|_{0,K} + h_K^{1/2}\|\nabla(v\psi_F)\|_{0,K} \leq C_{FB}\|v\|_{0,F}, \tag{5.45}$$

where the constant C_{FB} is independent of v and h_K.

Consider a mesh cell K. Now the local estimator for mesh cell K is defined as

$$\eta_K^2 = \eta_{\text{Int,K}}^2 + \sum_{F\in\mathcal{F}_h(K)} \eta_{\text{Face},F}^2 + \sum_{E\in\mathcal{E}_h(K)} \eta_{d_h,E}^2 \tag{5.46}$$

with

$$\begin{aligned}
\eta_{\text{Int},K}^2 &= \min\left\{\frac{4C_I^2}{\sigma_0}, \frac{4C_I^2 h_K^2}{\varepsilon}\right\}\|R_{K,h}(u_h)\|_{L^2(K)}^2,\\
\eta_{\text{Face},F}^2 &= \frac{1}{N_F}\min\left\{\frac{4C_F^2 h_F}{\varepsilon}, \frac{4C_F^2}{\sigma_0^{1/2}\varepsilon^{1/2}}\right\}\|R_F(u_h)\|_{L^2(F)}^2,\\
\eta_{d_h,E}^2 &= \min\left\{\frac{4\kappa_1 h_E^2}{\varepsilon}, \frac{4\kappa_2}{\sigma_0}\right\}(1-\alpha_E)^2|d_E|^2 h_E^{1-d}\|\nabla u_h\cdot \boldsymbol{t}_E\|_{L^2(E)}^2,
\end{aligned} \tag{5.47}$$

where $\mathcal{F}_h(K)$ is the set of all facets of K, $E \in \mathcal{E}_h(K)$ the set of all edges belonging to K, and N_F the number of mesh cells where the face F belongs to. Each inner facet belongs to two mesh cells, that's why $N_F = 2$ for faces that do not lie on the boundary of the domain. We bound each term individually.

Interior Residual: In (5.46) define

$$R_{K,h}(u_h) = f_h + \varepsilon\Delta u_h - \boldsymbol{b}_h\cdot\nabla u_h - c_h u_h$$

as a polynomial approximation of the mesh cell residual, with suitable polynomial approximations $\boldsymbol{b}_h, c_h$, and f_h of the coeffecients (2.6).

Let $v = R_{K,h}(u_h)\psi_K$, then this function is a polynomial on K, that vanishes on the boundary of K and it can be extended by zero to the whole domain Ω. This function belongs to $H_0^1(\Omega)$, thus in particular to $H_D^1(\Omega)$ and hence, it can be used as test function in (3.3). Let $e = u - u_h$, then one obtains with integration by parts of the diffusion term, (3.3), and the definition of $R_{K,h}(u_h)$

$$a(e, R_{K,h}(u_h)\psi_K) = (R_{K,h}(u_h), R_{K,h}(u_h)\psi_K)_K + (R_K(u_h) - R_{K,h}(u_h), R_{K,h}(u_h)\psi_K)_K. \quad (5.48)$$

Using (5.42), (5.48), Hölder's inequality, (5.43), and $\|\psi_K\|_{L^\infty(K)} = 1$ yields

$$\begin{aligned}
\|R_{K,h}(u_h)\|^2_{L^2(K)} &\le C_K(R_{K,h}(u_h), R_{K,h}(u_h)\psi_K)_K \\
&\quad -C_K a(e, R_{K,h}(u_h)\psi_K)_K - C_K(R_K(u_h) - R_{K,h}(u_h), R_{K,h}(u_h)\psi_K)_K \\
&\le C_K\Big[\varepsilon\|\nabla e\|_{L^2(K)}\|\nabla(R_{K,h}(u_h)\psi_K)\|_{L^2(K)} \\
&\quad +\|\boldsymbol{b}\|_{L^\infty(K)}\|\nabla e\|_{L^2(K)}\|R_{K,h}(u_h)\psi_K\|_{L^2(K)} \\
&\quad +\|c\|_{L^\infty(K)}\|e\|_{L^2(K)}\|R_{K,h}(u_h)\psi_K\|_{L^2(K)}\Big] \\
&\quad +C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)}\|R_{K,h}(u_h)\|_{L^2(K)} \\
&\le C_K\left[C_K h_K^{-1}\varepsilon\|\nabla e\|_{L^2(K)} + \|\boldsymbol{b}\|_{L^\infty(K)}\|\nabla e\|_{L^2(K)} + \|c\|_{L^\infty(K)}\|e\|_{L^2(K)}\right] \\
&\quad \times\|R_{K,h}(u_h)\|_{L^2(K)} + C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)}\|R_{K,h}(u_h)\|_{L^2(K)}.
\end{aligned}$$

Hence, one obtains

$$\begin{aligned}
\|R_{K,h}(u_h)\|^2_{L^2(K)} &\le C_1\varepsilon^{1/2}\|\nabla e\|_{L^2(K)} + C_2\sigma_0^{1/2}\|e\|_{L^2(K)} + C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)} \\
&\le \max\{C_1, C_2\}(\varepsilon^{1/2}\|\nabla e\|_{L^2(K)} + \sigma_0^{1/2}\|e\|_{L^2(K)}) \\
&\quad +C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)} \\
&\le 2\max\{C_1, C_2\}\left(\varepsilon\|\nabla e\|^2_{L^2(K)} + \sigma_0\|e\|^2_{L^2(K)}\right)^{1/2} \\
&\quad +C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)} \\
&= 2\max\{C_1, C_2\}\|e\|_{a,K} + C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)}, \qquad (5.49)
\end{aligned}$$

with

$$C_1 = \frac{C_K^2\varepsilon^{1/2}}{h_K} + \frac{C_K\|\boldsymbol{b}\|_{L^\infty(K)}}{\varepsilon^{1/2}}, \qquad C_2 = \frac{C_K\|c\|_{L^\infty(K)}}{\sigma_0^2}.$$

Let $1/\sigma_0 > h_K^2/\varepsilon$, then one gets with (5.49)

$$\begin{aligned}
\eta_{\mathrm{Int},K} &= C\frac{h_K}{\varepsilon^{1/2}}\|R_{K,h(u_h)}\|_{L^2(K)} \\
&\le C\max\left\{C_K^2 + \frac{C_K h_K}{\varepsilon}\|\boldsymbol{b}\|_{L^\infty(K)}, \frac{C_K h_K}{\varepsilon^{1/2}\sigma_0^{1/2}}\|c\|_{L^\infty(K)}\right\}\|u - u_h\|_{a,K} \\
&\quad +\frac{h_K}{\varepsilon^{1/2}}C_K\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)}
\end{aligned}$$

$$
\begin{aligned}
&\leq C\max\left\{C_K^2+\frac{C_K h_K}{\varepsilon}\|\boldsymbol{b}\|_{L^\infty(K)},\frac{C_K}{\sigma_0}\|c\|_{L^\infty(K)}\right\}\|u-u_h\|_{a,K}\\
&\quad+\frac{h_K}{\varepsilon^{1/2}}C_K\|R_K(u_h)-R_{K,h}(u_h)\|_{L^2(K)}.
\end{aligned}
$$

If $1/\sigma_0 \leq h_K^2/\varepsilon$, then one obtains in the same way

$$
\begin{aligned}
\eta_{\mathrm{Int},K} &= \frac{C}{\sigma_0^{1/2}}\|R_{K,h(u_h)}\|_{L^2(K)}\\
&\leq C\max\left\{\frac{C_K^2\varepsilon^{1/2}}{\sigma_0^{1/2}h_K}+\frac{C_K}{\varepsilon^{1/2}\sigma_0^{1/2}}\|\boldsymbol{b}\|_{L^\infty(K)},\frac{C_K}{\sigma_0}\|c\|_{L^\infty(K)}\right\}\|u-u_h\|_{a,K}\\
&\quad+\frac{C}{\sigma_0^{1/2}}C_K\|R_K(u_h)-R_{K,h}(u_h)\|_{L^2(K)}\\
&\leq C\max\left\{C_K^2+\frac{C_K h_K}{\varepsilon}\|\boldsymbol{b}\|_{L^\infty(K)},\frac{C_K}{\sigma_0}\|c\|_{L^\infty(K)}\right\}\|u-u_h\|_{a,K}\\
&\quad+\frac{Ch_K}{\varepsilon^{1/2}}C_K\|R_K(u_h)-R_{K,h}(u_h)\|_{L^2(K)}.
\end{aligned}
$$

Hence, this leads to

$$
\begin{aligned}
\eta_{\mathrm{Int},K} \leq\ & C\Bigg(\max\left\{C_K^2+\frac{C_K h_K}{\varepsilon}\|\boldsymbol{b}\|_{L^\infty(K)},\frac{C_K}{\sigma_0}\|c\|_{L^\infty(K)}\right\}\|u-u_h\|_{a,K}\\
&+\frac{h_K}{\varepsilon^{1/2}}C_K\Big(\|f-f_h\|_{0,K}+\|(\boldsymbol{b}-\boldsymbol{b}_h)\cdot\nabla u_h\|_{0,K}+\|(c-c_h)u_h\|_{0,K}\Big)\Bigg).
\end{aligned}
\tag{5.50}
$$

Face Residuals: The analysis of the face residuals follows the same idea as that of the interior residuals. Let $R_{F,h}(u_h)$ be an approximation to the face residual from a suitable finite-dimensional space and ψ_F be the face bubble function that vanishes on the boundary of $\omega_F = K \cup K'$, where K and K' are two mesh cells sharing the face F. Then one obtains with (5.44)

$$
\|R_{F,h}(u_h)\|_{0,F}^2 \leq C_{FB}(R_{F,h}(u_h),R_{F,h}(u_h)\psi_F)_F. \tag{5.51}
$$

The function $v = R_{F,h}(u_h)\psi_F$, which vanishes on all the nodes, belongs to $H_D^1(\Omega)$. Hence, using this as test function in (3.3) and using the same arguments as that for the interior residual, shows that

$$
a(e,R_{F,h}(u_h)\psi_F) = \sum_{K\in\omega_F}(R_K(u_h),R_{F,h}(u_h)\psi_F)_K
$$

$$+(R_{F,h}(u_h), R_{F,h}(u_h)\psi_F)_F + (g - g_h, R_{F,h}(u_h)\psi_F)_F. \qquad (5.52)$$

Using (5.44) and (5.52) leads to

$$\begin{aligned}
\|R_{F,h}(u_h)\|^2_{L^2(F)} &\leq C_{FB}(R_{F,h}(u_h), R_{F,h}(u_h)\psi_F)_F \\
&= C_{FB}a(e, R_{F,h}(u_h)\psi_F) - C_{FB}\sum_{K\in\omega_F}(R_K(u_h), R_{F,h}(u_h)\psi_F)_K \\
&\quad -C_{FB}(g - g_h, R_{F,h}(u_h)\psi_F)_F. \qquad (5.53)
\end{aligned}$$

The first term is estimated similarly to the cell residual, using (5.45) and Young's inequality

$$\begin{aligned}
a(e, R_{F,h}(u_h)\psi_F) &\leq \sqrt{2}\left(\sum_{K\in\omega_F}\max\{C_{1,K}, C_{2,K}\}\|e\|_{a,K}\right)\|R_{F,h}(u_h)\|_{L^2(F)} \\
&\leq 2\sqrt{2}\,\max\{C_{1,\omega_F}, C_{2,\omega_F}\}\|e\|_{a,\omega_F}\|R_{F,h}(u_h)\|_{L^2(F)},
\end{aligned}$$

with

$$C_{1,K} = \frac{C_{FB}\varepsilon^{1/2}}{h_F^{1/2}} + \frac{C_{FB}h_F^{1/2}\|\boldsymbol{b}\|_{L^\infty(K)}}{\varepsilon^{1/2}}, \qquad C_{2,K} = \frac{C_{FB}h_F^{1/2}\|c\|_{L^\infty(K)}}{\sigma_0^{1/2}}$$

and C_{1,ω_F}, C_{2,ω_F} defined similarly with the norms on K replaced with the norms on ω_F. Applying the Cauchy-Schwarz inequality and (5.45) yields

$$\sum_{K\in\omega_F}(R_K(u_h), R_{F,h}(u_h)\psi_F)_K \leq C_{FB}h_F^{1/2}\left(\sum_{K\in\omega_F}\|R_K(u_h)\|_{L^2(K)}\right)\|R_{F,h}(u_h)\|_{L^2(F)}\|.$$

The term with the data approximation error of the Neumann data appears of course only if $F \in \mathcal{F}_{h,N}$. Then, one obtains with the Cauchy-Schwarz inequality and $\|\psi_F\|_{L^\infty(F)} = 1$

$$(g - g_h, R_{F,h}(u_h)\psi_F)_F \leq \delta_{F,\mathcal{F}_{h,N}}\|g - g_h\|_{L^2(F)}\|R_{F,h}(u_h)\|_{L^2(F)},$$

with

$$\delta_{F,\mathcal{F}_{h,N}} = \begin{cases} 1 & \text{if } F \in \mathcal{F}_{h,N} \\ 0 & \text{else.} \end{cases}$$

Inserting the last three bounds in (5.53) leads to

$$\begin{aligned}
\|R_{F,h}(u_h)\|_{L^2(F)} &\leq 2\sqrt{2}\,C_{FB}\,\max\{C_{1,\omega_F}, C_{2,\omega_F}\}\|e\|_{a,\omega_F} \\
&\quad +C_{FB}^2 h_F^{1/2}\left(\sum_{K\in\omega_F}\|R_{K,h}(u_h)\|_{L^2(K)}\right) \\
&\quad +C_{FB}^2 h_F^{1/2}\Big(\sum_{K\in\omega_F}\|R_K(u_h) - R_{K,h}(u_h)\|_{L^2(K)}\Big) \\
&\quad +C_{FB}\delta_{F,\mathcal{F}_{h,N}}\|g - g_h\|_{L^2(F)}.
\end{aligned}$$

The second term was already estimated by the error in the energy norm in (5.50).

If $h_F/\varepsilon^{1/2} \leq 1/\sigma_0^{1/2}$, one obtains with $h_F \leq h_K$

$$\begin{aligned}
\eta_{\text{Face},F} &\leq C\frac{h_F^{1/2}}{\varepsilon^{1/2}}\|R_{F,h}(u_h)\|_{L^2(F)}\\
&\leq C\max\left\{C_{FB}+\frac{C_{FB}h_F\|\boldsymbol{b}\|_{L^\infty(\omega_F)}}{\varepsilon},\frac{C_{FB}h_F\|c\|_{L^\infty(\omega_F)}}{\varepsilon^{1/2}\sigma_0^{1/2}}\right\}\|u-u_h\|_{a,\omega_F}\\
&\quad +C\sum_{K\in\omega_F}\max\left\{C_K^2+\frac{C_Kh_K}{\varepsilon}\|\boldsymbol{b}\|_{L^\infty(K)},\frac{C_K}{\sigma_0}\|c\|_{L^\infty(K)}\right\}\|u-u_h\|_{a,K}\\
&\quad +C\sum_{k\in\omega_F}\frac{h_K}{\varepsilon^{1/2}}\|R_K(u_h)-R_{K,h}(u_h)\|_{L^2(K)}\\
&\quad +C\delta_{F,\mathcal{F}_{h,N}}\frac{h_F^{1/2}}{\varepsilon^{1/2}}\|g-g_h\|_{L^2(F)}.
\end{aligned}$$

With a straightforward calculation, one can derive the same bound also in the case $h_F/\varepsilon^{1/2} > 1/\sigma_0^{1/2}$.

Hence, this leads to

$$\begin{aligned}
\eta_{\text{Face},F} &\leq C\Bigg(\max\left\{C_{FB}+\frac{C_{FB}h_F\|\boldsymbol{b}\|_{L^\infty(\omega_F)}}{\varepsilon},\frac{C_{FB}h_F\|c\|_{L^\infty(\omega_F)}}{\varepsilon^{1/2}\sigma_0^{1/2}}\right\}\|u-u_h\|_{a,\omega_F}\\
&\quad +\delta_{F\in\mathcal{F}_{h,N}}\frac{h_F^{1/2}}{\varepsilon^{1/2}}\|g-g_h\|_{L^2(F)}\\
&\quad +\sum_{K\in\omega_F}\Big[\eta_{\text{Int},K}+\frac{h_K}{\varepsilon^{1/2}}\Big(\|f-f_h\|_{0,K}\\
&\quad +\|(\boldsymbol{b}-\boldsymbol{b}_h)\cdot\nabla u_h\|_{0,K}+\|(c-c_h)u_h\|_{0,K}\Big)\Big]\Bigg).
\end{aligned}\tag{5.54}$$

Edge Residuals: The final term one wants to bound in η_K is the AFC contribution. A similar term can be observed in [ABR17, Theorem 2]. Based on certain assumptions on the nonlinear stabilization namely the Lipschitz continuity and linearity preservation that term is bounded there. We will not use such assumptions as they do not encompass the limiters presented in Sec. 3.4 namely the Kuzmin limiter.

From the proof of [BJKR18, Lemma 2] we have

$$|d_E| \leq C\left(\varepsilon+\|\boldsymbol{b}\|_{L^\infty(\Omega)}h+\|c\|_{L^\infty(\Omega)}h^2\right)h_E^{d-2}. \tag{5.55}$$

We have

$$\eta_{d_h,E} \;\leq\; C \sum_{E\in\mathcal{E}_h} (1-\alpha_E)|d_E| h_E^{(1-d)/2} \min\left\{ \frac{h_E}{\varepsilon^{1/2}}, \frac{1}{\sigma_0^{1/2}} \right\} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)}.$$

Hence, we get from (5.55)

$$\begin{aligned}
\eta_{d_h,E} \;&\leq\; C \sum_{E\in\mathcal{E}_h} (1-\alpha_E) \left(\varepsilon + \|\boldsymbol{b}\|_{L^\infty(\Omega)} h + \|c\|_{L^\infty(\Omega)} h^2\right) \frac{h_E^{(3-d)/2}}{\varepsilon^{1/2}} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)} \\
&=\; C \sum_{E\in\mathcal{E}_h} (1-\alpha_E) \left(\varepsilon^{1/2} + \frac{\|\boldsymbol{b}\|_{L^\infty(\Omega)} h}{\varepsilon^{1/2}} + \frac{\|c\|_{L^\infty(\Omega)} h^2}{\varepsilon^{1/2}} \right) \\
&\quad \times h_E^{(3-d)/2} \|\nabla u_h \cdot \boldsymbol{t}_E\|_{L^2(E)}.
\end{aligned} \tag{5.56}$$

For a fixed ε, we consider the convection-dominated regime, i.e., $\varepsilon \leq h$, then we get

$$\eta_{d_h,E} = \mathcal{O}(h)$$

in 2d, and

$$\eta_{d_h,E} = \mathcal{O}(h^{1/2})$$

in 3d, whereas, for diffusion-dominated case we get $\mathcal{O}(h^{1/2})$ in 2d. This term is not exactly an oscillation. It is noted in [BJK16] that the average rate of decay for the first factor in parentheses is one but no concrete analysis has been provided. Altogether this term has to be studied numerically. Also for shock-capturing methods a priori estimates usually give $\mathcal{O}(h^{1/2})$ convergence (see [BJK16, Corollary 17]), then we can expect the last term to behave as an oscillation (see [ABR17, Remark 5]).

Remark 5.31. To simplify the notation we will denote $\eta_{d_h,E}$ by η_{d_h} whenever we don't have ambiguity for E. Numerical examples will be presented in Sec. 5.3 to show the behavior of η_{d_h}.

Theorem 5.32. *There exists a constant $C > 0$, independent of the size of elements of $\mathcal{T}$, such that, for every $K \in \mathcal{T}$, the following local lower bound holds*

$$\begin{aligned}
&\eta_{\mathrm{Int},K} + \sum_{K\in\mathcal{F}_h(K)} \eta_{\mathrm{Face},F} + \sum_{E\in\mathcal{E}_h(K)} \eta_{d_h,E} \\
&\leq\; \max\left\{ C_K^2 + \frac{C_K h_K}{\varepsilon} \|\boldsymbol{b}\|_{L^\infty(K)}, \frac{C_K}{\sigma_0} \|c\|_{L^\infty(K)} \right\} \|u - u_h\|_{a,\omega_K} \\
&\quad + C \sum_{K\in\omega_K} \frac{h_K}{\varepsilon^{1/2}} \Big(\|f - f_h\|_{0,K} + \|(\boldsymbol{b} - \boldsymbol{b}_h)\cdot\nabla u_h\|_{0,K} + \|(c - c_h) u_h\|_{0,K} \Big) \\
&\quad + C \sum_{F\in\mathcal{F}_h(K)} \delta_{F\in\mathcal{F}_{h,N}} \frac{h_F^{1/2}}{\varepsilon^{1/2}} \|g - g_h\|_{L^2(F)}
\end{aligned}$$

$$+\sum_{E\in\mathcal{E}_h(K)} h^{1-d/2}\frac{h^{1/2}}{\varepsilon^{1/2}}\Big(\varepsilon+\|b\|_{L^\infty(\Omega}h+\|c\|_{L^\infty(\Omega}h^2\Big)\|\nabla u_h\cdot \boldsymbol{t}_E\|_{L^2(E)}. \qquad (5.57)$$

Proof. This estimate can be obtained by combining (5.50), (5.54), and (5.56). □

Remark 5.33. We note that the estimator is not robust with respect to ε. However, this is the usual case for a posteriori error estimators for the error measured in the energy norm. In [TV15] residual-based a posteriori estimators for the error were proved to be robust with respect to a norm that includes a dual norm of the convective term. However, all the methods considered in [TV15] were linear and application of those techniques to nonlinear discretizations such as AFC does not seem to be feasible.

5.2.2 AFC-SUPG Estimator

An alternative way of finding a global upper bound for the error in the energy norm for the AFC scheme is to use the estimator proposed in [JN13]. An upper bound which is robust with respect to the diffusion coefficient, ε, was derived for the error in the SUPG norm [JN13, Eq. (11)] for the SUPG scheme. It has been noted in Chapter 4 that choosing the initial solution as the SUPG solution for the nonlinear system of equations was most appropriate. We exploit this fact to bound our error.

Let u_{AFC}, u_{SUPG} denote the AFC and SUPG solution, respectively. Then by the triangle inequality

$$\begin{aligned}\|u-u_{\mathrm{AFC}}\|_a^2 &\le 2\left(\|u-u_{\mathrm{SUPG}}\|_a^2+\|u_{\mathrm{SUPG}}-u_{\mathrm{AFC}}\|_a^2\right)\\ &\le 2\left(\|u-u_{\mathrm{SUPG}}\|_{\mathrm{SUPG}}^2+\|u_{\mathrm{SUPG}}-u_{\mathrm{AFC}}\|_a^2\right).\end{aligned}$$

The first term can be bounded by [JN13, Theorem 2.1] and the second term is computable. Let

$$\|u-u_{\mathrm{SUPG}}\|_{\mathrm{SUPG}}^2\le \eta_{\mathrm{SUPG}}^2,$$

where η_{SUPG}^2 is given by [JN13, Eq. (36)] and

$$\eta_{\mathrm{AFC-SUPG}} := \|u_{\mathrm{AFC}}-u_{\mathrm{SUPG}}\|_a,$$

then

$$\|u-u_{\mathrm{AFC}}\|_a^2\le \eta^2,$$

where

$$\eta^2 = 2\left(\eta_{\mathrm{SUPG}}^2+\eta_{\mathrm{AFC-SUPG}}^2\right).$$

Numerical simulations depicting the behavior of $\eta_{\mathrm{SUPG}}, \eta_{\mathrm{AFC-SUPG}}$ along with the adaptive refinement of grids will be presented in Sec. 5.3.

5.3 Numerical Studies

The standard strategy for numerically solving a partial differential equation on adaptively refined grids using an a posteriori error estimator is

$$\textbf{SOLVE} \rightarrow \textbf{ESTIMATE} \rightarrow \textbf{MARK} \rightarrow \textbf{REFINE}.$$

We note that to refine a grid adaptively, two important things are required:

- *Marking strategy*, that decides which mesh cells should be refined,
- *Refinement rules*, which determines the actual subdivision of a mesh cell.

We have already discussed the refinement rules in Sec. 5.1. There are two marking strategies that are widely used in a posteriori packages, namely the *maximum marking strategy* and the *equilibration marking strategy* (see [Ver13]). It is noted in [Ver13] that both the strategies produce comparable results but it is computationally cheaper to implement the maximum marking strategy and hence it is used in our simulations. Algorithm 1 details the aforementioned strategy.

Algorithm 1 (**Maximum Strategy**) [Ver13, Algorithm 2.1]

Given: partition $\mathcal{T}$, error indicators $(\eta_K)_{K\in\mathcal{T}}$, threshold $\theta \in (0,1)$.
Find: subset $\tilde{\mathcal{T}}$ of *marked* elements that should be refined.

```
$\mathcal{T} \leftarrow \emptyset$
$\eta_{\mathcal{T},\max} \leftarrow \max_{K\in\mathcal{T}} \eta_K$
for $K \in \mathcal{T}$ do
    if $\eta_K \geq \theta\eta_{\mathcal{T},\max}$ then
        $\tilde{\mathcal{T}} \leftarrow \tilde{\mathcal{T}} \cup \{K\}$
    end if
end for
```

Remark 5.34. An issue that arises while marking of cells for convection-dominated problems is that only a few mesh cells with high error are marked, which deteriorates the performance of the algorithm. To ensure that enough cells are marked, we follow the strategy prescribed in [Joh00, Sec. 4]. Flowchart 5.6 describes this strategy.

The quality of an estimator is usually judged by its global effectivity index that is given by,

$$\eta_{\text{eff}} = \frac{\eta}{\|u - u_h\|_a}.$$

This index can be used to measure the quality of an estimator when the exact or a good approximation is known to the solution.

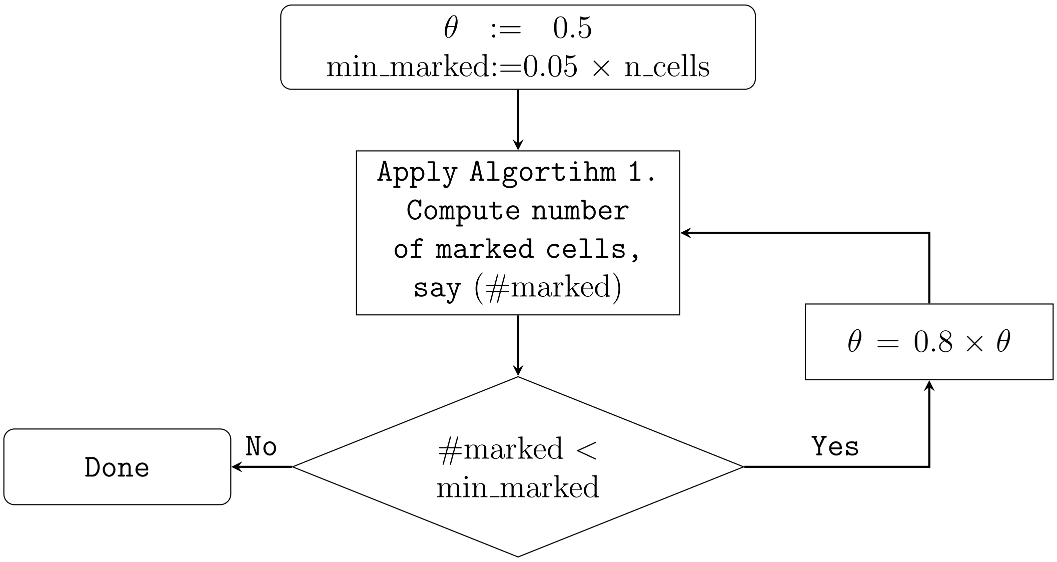

θ := 0.5
min_marked:=0.05 × n_cells
Apply Algortihm 1. Compute number of marked cells, say (#marked)
$\theta = 0.8 \times \theta$
#marked < min_marked
No
Yes
Done

Figure 5.6: Adaptive choice of θ in Algorithm 1

We note that we have the presence of certain constants in our estimators. We chose the value of these constants to be unity.

Remark 5.35. We have discussed two different strategies for finding a global upper bound for the AFC error in the energy norm. Further in this section we will refer to the idea from Sec. 5.2.1.1 as *AFC-energy* technique and from Sec. 5.2.2 as *AFC-SUPG-energy* technique.

Numerical studies presented further in this section will comprehend the results for the two different techniques on the following conditions:

1. Compare the *AFC-energy* and *AFC-SUPG-energy* techniques:
    a) with respect to the effectivity index in the energy norm
    b) with respect to adaptive grid refinement.
2. Study the behavior of η_{d_h} defined in (5.56), on uniformly and adaptively refined grids.
3. Study the behavior of η_{SUPG} and $\eta_{\mathrm{AFC-SUPG}}$ for the *AFC-SUPG-energy* technique.

The matrices were assembled exactly and the linear systems were solved using the direct solver UMFPACK [Dav04]. The method fixed point right-hand side was used for solving the nonlinear problems with the damping parameters as described in Chapter 4. The stopping criteria for the adaptive algorithm was either $\#\mathrm{dof} \gtrsim 10^6$ or $\eta < 10^{-3}$. All the simulations were performed with the in-house code ParMooN [WBA$^+$16].

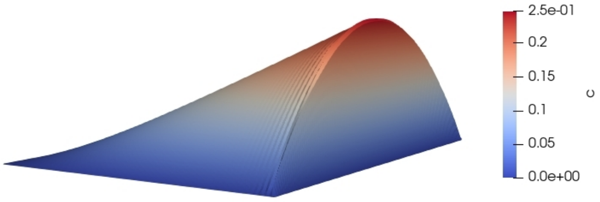


Figure 5.7: 2d Boundary layer example. Solution (computed with the BJK limiter, level 7).

5.3.1 A Known 2d Solution with a Boundary Layer

This example was proposed in [ABR17, Example 1]. Consider $\varepsilon = 10^{-3}$, $\boldsymbol{b} = (2,1)^T$, $c = 1$, $g = 0$, $u_b = 0$, and the right-hand side f such that the exact solution is given by

$$u(x,y) = y(1-y)\left(x - \frac{e^{(x-1)/\varepsilon} - e^{-1/\varepsilon}}{1 - e^{-1/\varepsilon}}\right),$$

on the domain $\Omega = (0,1)^2$ (see Fig. 5.7). An initial grid was defined with two triangles by joining the points $(0,0)$ and $(1,1)$. The simulations were started with a level 2 grid (i.e., $\#\mathrm{dof} = 25$), initially uniform refinement was performed till level 4 (i.e., $\#\mathrm{dof} = 289$). After that adaptive refinement was performed.

First, we compare the behavior of effectivity indices for the *AFC-energy* and *AFC-SUPG-energy* techniques. For the *AFC-energy* technique, we note that as the adaptive refinement starts the effectivity index is high and as the grid becomes refined the value decreases (see left Fig. 5.8). For the Kuzmin limiter on grids with fine adaptive regions $\eta_{\text{eff}} \approx 232$ and for the BJK limiter $\eta_{\text{eff}} \approx 12$. For the *AFC-SUPG-energy* technique the values of the effectivity index are better than for the *AFC-energy* technique (see right Fig. 5.8). One interesting observation to make is that the limiter does not play an important role in this technique. The values of effectivity indices are comparable for both the limiters. If the adaptive refinement is sufficiently fine, then for the Kuzmin limiter $\eta_{\text{eff}} \approx 2$ and for the BJK limiter $\eta_{\text{eff}} \approx 5$.

Next, we look at the individual behavior of η_{SUPG} and $\eta_{\mathrm{AFC-SUPG}}$. It can be seen in Fig. 5.9 that the dominating term is η_{SUPG} and hence, the AFC contribution, $\eta_{\mathrm{AFC-SUPG}}$, does not play a pivotal role in the effectivity index and the refinement of the grid.

Lastly, we study the behavior of the error in the energy norm, its relation to the a posteriori error estimates, and the behavior of the part η_{d_h} of the error estimators in some detail. One can observe that the error as well as η_{d_h} and η for the *AFC-energy* technique decay optimally on adaptive grids for the BJK limiter (see Fig. 5.10). For the Kuzmin limiter one observes that as the grid becomes fine the optimal rate is not obtained for the error as well as for η_{d_h} and η. It has been noted in [BJK16, Remark 18] that if the grid is non-Delaunay and the problem becomes diffusion-dominated then the AFC method with the Kuzmin limiter fails to converge. With successive refinement of the grid, the problem becomes locally diffusion-

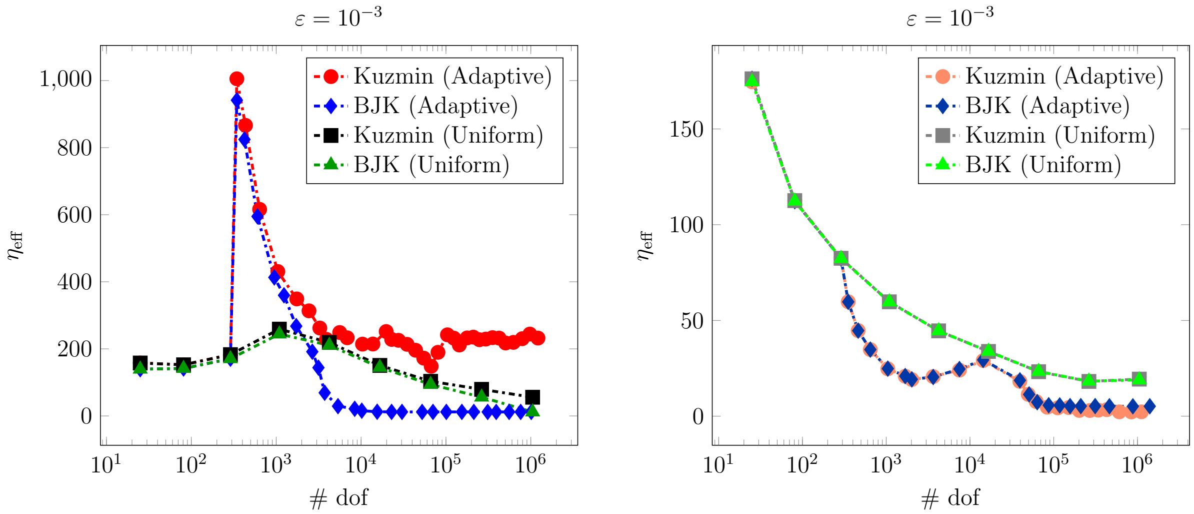


Figure 5.8: Example 5.3.1: Effectivity index in the energy norm with *AFC-energy* technique defined in Sec. 5.2.1.1 (left) and *AFC-SUPG-energy* technique defined in Sec. 5.2.2 (right).

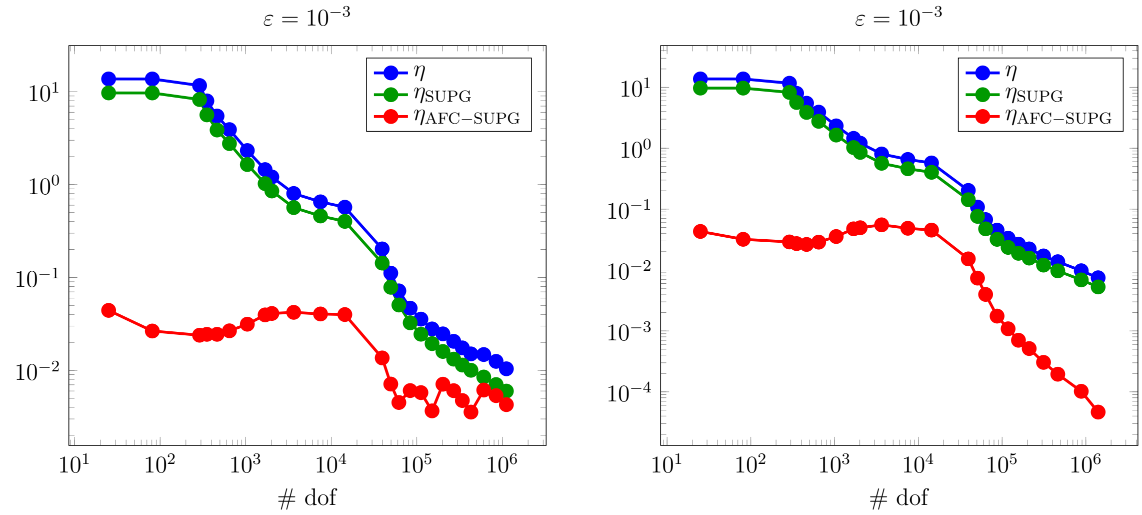


Figure 5.9: Example 5.3.1: Comparison of η_{SUPG} and $\eta_{\text{AFC-SUPG}}$ for *AFC-SUPG-energy* technique. Kuzmin limiter (left) and BJK limiter (right).

dominated (in the sense of a small grid Peclet number) and one has to expect, because of the conforming closure and the resulting obtuse angles, that there is no convergence. The error estimator with the *AFC-energy* technique predicts this irregular behavior of the error. This reduction of the rate of convergence is not observed while using BJK limiter.

For the *AFC-SUPG-energy* technique the error and η values are shown in Fig. 5.10 (right). For the Kuzmin limiter, similar observation to the *AFC-energy* technique can be made. One issue to note is that the estimator(η) with *AFC-SUPG-energy* technique does not predict this irregular behavior as it has already been mentioned that the AFC contribution does not

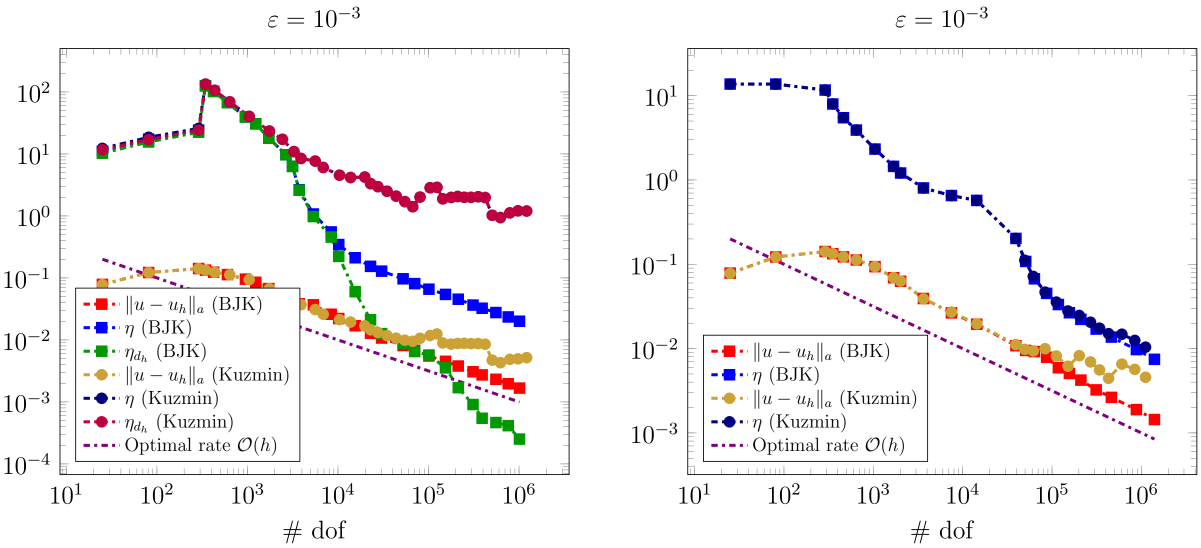


Figure 5.10: Example 5.3.1: Error in energy norm with *AFC-energy* technique defined in Sec. 5.2.1.1 (left) and *AFC-SUPG-energy* technique defined in Sec. 5.2.2 (right). The line corresponding to η (Kuzmin) is below η_{d_h} (Kuzmin) in the left figure.

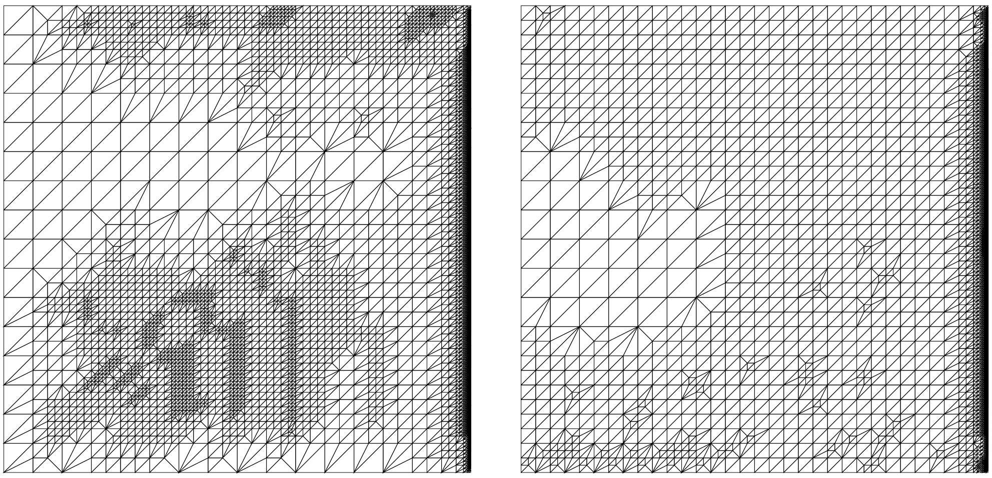


Figure 5.11: Example 5.3.1: 14th adaptively refined grid with *AFC-energy* technique. Kuzmin limiter (#dof = 22962) (left) and BJK limiter (#dof = 23572)(right)

play an important role here.

Fig. 5.11 shows the 14th adaptively refined grid with *AFC-energy* technique. One can observe obtuse angles in the adaptive grids. In Fig. 5.10 (left) for the Kuzmin limiter, we also note that η_{d_h} is comparable with η and hence is the leading term in the adaptive refinement of the grid. For the BJK limiter, as the grid becomes finer, η_{d_h} is small as compared to η.

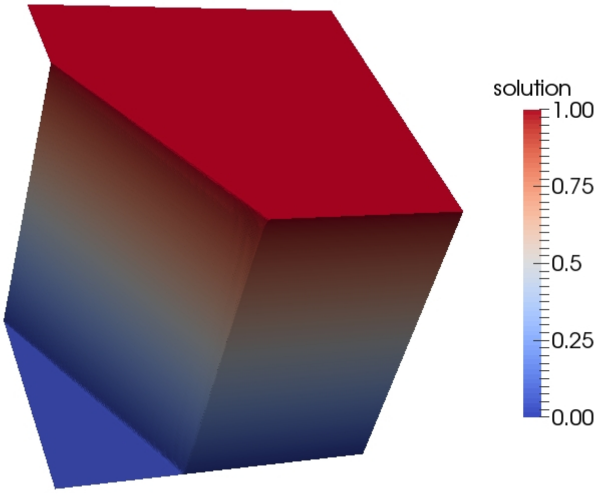


Figure 5.12: Example 5.3.2. Solution (computed with the BJK limiter, level 9).

5.3.2 Example with Interior and Boundary Layers

Let us recall this example. It is given in $\Omega = (0,1)^2$ with $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))$, $c = f = 0$ and the Dirichlet boundary condition

$$u_D = \begin{cases} 1 & (y = 1 \wedge x > 0) \text{ or } (x = 0 \wedge y > 0.7), \\ 0 & \text{else.} \end{cases}$$

Here, $\varepsilon = 10^{-4}$ is considered. It is known that the solution exhibits an internal layer in the direction of the convection starting from the jump of the boundary condition at the left boundary and two exponential layers at the right and the lower boundary (see Fig. 5.12). A known solution to this problem is not available but we know that $u \in [0,1]$. This example serves for studying the adaptive grid refinement in the presence of different kinds of layers.

An initial mesh was defined similar to the previous example, i.e., with two triangles by joining the points $(0,0)$ and $(1,1)$. The simulations were started with a level 2 grid (i.e., #dof $= 25$), uniform refinement was performed till level 4 (i.e., #dof $= 289$) and then the adaptive grid refinement was started. For this example, we do not have the presence of regions where the problem becomes locally diffusion-dominated because the refinement does not make the grid sufficiently fine for the considered diffusion parameter.

The 14th adaptively refined grids with conforming closure and *AFC-energy* technique are shown in Fig. 5.13 for the Kuzmin limiter (left) and the BJK limiter (right), respectively. Here we see that we have the presence of non-Delaunay triangulation but we could note that the DMP was satisfied for both the limiters. This result shows that using the Kuzmin limiter might lead to solutions that satisfy he DMP even if an essential assumption of the analysis (Delaunay triangluation [BJK16, Remark 14]) is not satisfied. Comparing the refinement

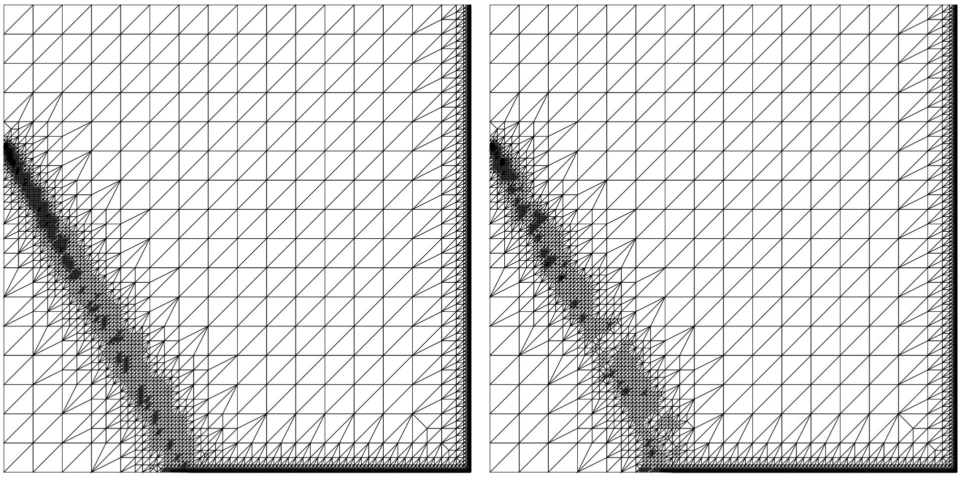


Figure 5.13: Example 5.3.2: 14^{th} adaptively refined grid with *AFC-energy* technique and with conforming closure.
Kuzmin limiter (14^{th} grid: #dof = 28548 (left) and BJK limiter (14^{th} grid: #dof = 28120) (right).

for both the limiters, we observe that the number of mesh cells is comparable for both the limiters (see Fig. 5.13 for #dof).

Next, we study the adaptive grid refinement for the *AFC-SUPG-energy* technique. The 14^{th} adaptively refine grids with conforming closure are shown in Fig. 5.14 for the Kuzmin limiter (left) and the BJK limiter (right), respectively. Here we observe that the mesh cells near the internal layer are not refined that much as compared to the *AFC-energy* technique. Also, we see that the limiters do not play an important role in the adaptive refinement. To be precise, the #dof are comparable for both the limiters and the meshes look much more similar than in Fig. 5.13.

To check the thickness of the interior layer we follow the idea described in [JK07a, Eq. (48)]. We define

$$smear_{\text{int}} = x_2 - x_1, \tag{5.58}$$

where x_1 is the $x-$coordinate of the first point on the cut line $(x, 0.25)$ with $u_h(x_1, 0.25) \geq 0.1$ and x_2 is the $x-$coordinate of the first point with $u_h(x_1, 0.25) \geq 0.9$. We note that in Fig. 5.15, the layers are most properly resolved for *AFC-energy* technique as compared to the *AFC-SUPG-energy* technique irrespective of the choice of limiters. Overall, for adaptive grid refinement, the *AFC-energy* technique does a much better job since all layers are refined properly, not only the strongest layer.

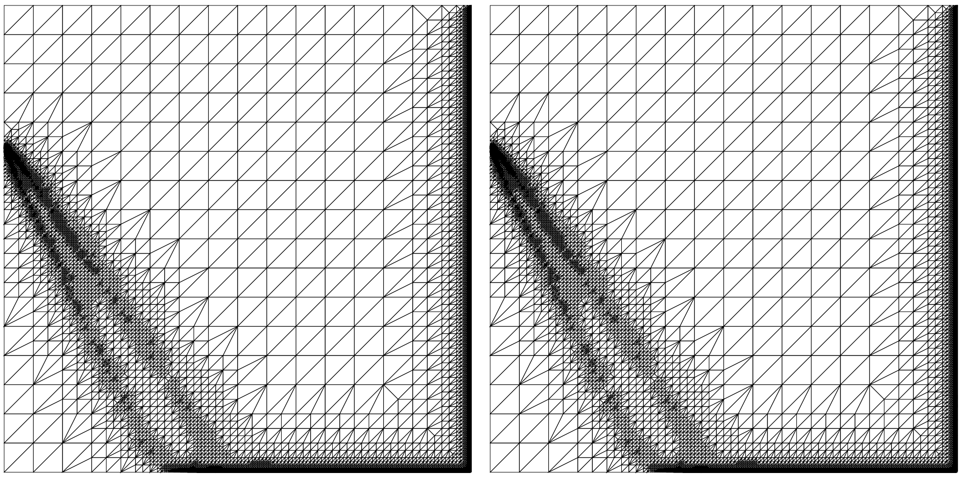


Figure 5.14: Example 5.3.2: 14th adaptively refined grid with *AFC-SUPG-energy* technique and with conforming closure.
Kuzmin limiter (14th grid: #dof = 100620 (left) and BJK limiter (14th grid: #dof = 100538) (right).

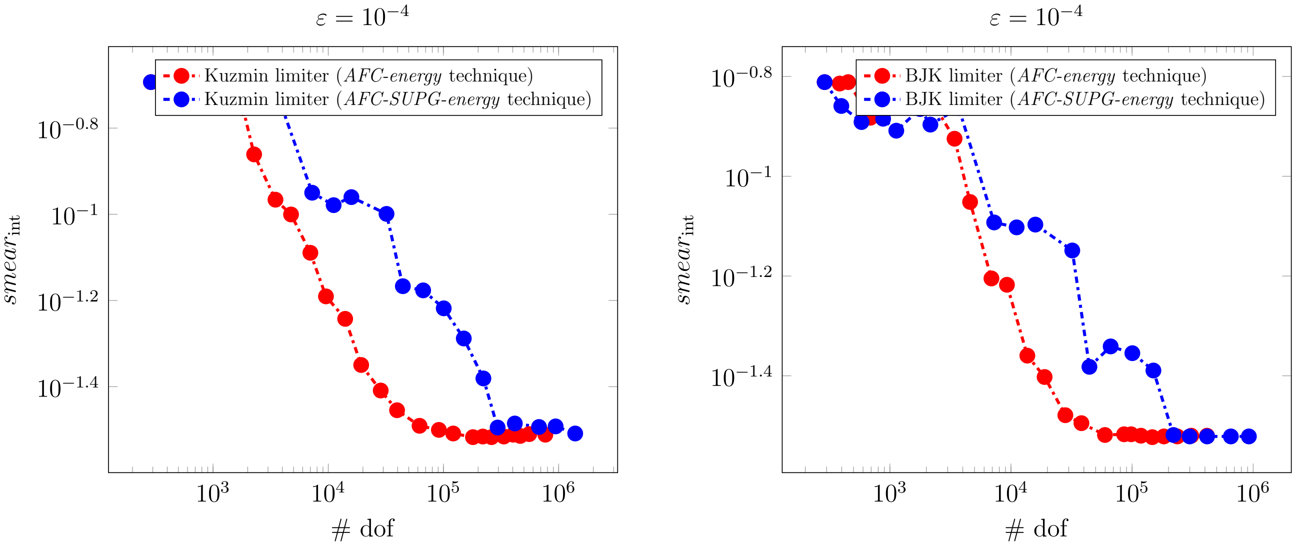


Figure 5.15: Example 5.3.2: Thickness of interior layer. Kuzmin limiter (left), BJK limiter (right)

5.4 Summary

This chapter presented a posteriori error estimators for the AFC scheme in the energy norm. Different refinement techniques and conforming closure for triangular elements have been discussed. Certain results regarding the relationship between the gradient and tangential component of the gradient on the edges and the mesh cell have been derived. A concrete value of constant has also been given for the aforementioned result.

The following conclusions can be made from the numerical simulations.

1. The effectivity index of the error estimator with *AFC-energy* was not robust with respect to ε. For a strongly convection-dominated case, the effectivity index was quite large which eventually decreased as the mesh became finer.

2. For the *AFC-SUPG-energy* technique, the effectivity index was better as compared with the *AFC-energy* technique.

3. The choice of limiter did not play an important role in *AFC-SUPG-energy* technique as the dominating term was η_{SUPG}. Because of this dominating nature, one gets very similar refined grids and effectivity indices for both the limiters.

4. For the Kuzmin limiter and the *AFC-energy* technique, a reduced order of convergence can be observed with conforming closure using red-green refinements as adaptive refinement leads to locally diffusion-dominated problems. This kind of reduction of order of convergence is not observed with the BJK limiter.

5. The AFC contribution η_{d_h} is the dominating term in the estimator η for the Kuzmin limiter whereas for the BJK limiter in the convection-dominated situation it is the dominating term but if the layer becomes to be resolved, then no longer.

6. With adaptive grid refinement, the problem could become locally diffusion-dominated then one has to use the BJK limiter because, with the Kuzmin limiter, the finite element solution does not converge. This situation might only happen if the diffusion coefficient is comparably large with respect to the mesh size.

7. For a small diffusion coefficient, one does not run into the issues of the previous point and one has to use the Kuzmin limiter because of the difficulties encountered while solving the nonlinear problems with the BJK limiter, see Chapter 4.

8. For adaptive grid refinement and problems with different kinds of layers, the *AFC-energy* technique refines the grid much better as compared to the *AFC-SUPG-energy* technique.

In summary, the *AFC-SUPG-energy* technique gave better results as compared to the *AFC-energy* technique with respect to the effectivity index, whereas the *AFC-energy* technique gave better results with adaptive grid refinement. For convection-dominated problems, the BJK limiter gave a better effectivity index as compared to the Kuzmin limiter but for a small diffusion, difficulties arise in solving the nonlinear problem associated with the BJK limiter. Future work of the research relates to the development of robust estimators and extending the analysis for the local lower bound.

6 Hanging Nodes in Context of AFC Schemes

If adaptively refined grids based on a posteriori error estimators should be used, then one has to define the actual grid refinement. One would prefer the subsequent grids to hold the same geometrical properties as that of the initial grid, e.g., preservation of angles. We saw certain grid refinement techniques in Chapter 5. The first step of the refinement of a grid, i.e., the refinement of the marked cells, leads to the formation of hanging vertices which can be described as the non-trivial linear combination of the endpoints of the edge to which they belongs. In the framework of discontinuous finite elements, the handling of grids with a hanging vertex is rather easy to understand (see [AR10]). We would like to explicitly point that in existing literature, what we have mentioned as the hanging vertex is referred to as a hanging node (see [CH09]). The distinction between the two will be made clear in this chapter. For continuous finite elements, the framework becomes a little involved. We saw one easy way around this is to use conforming closure or red-green refinements but this leads to the deterioration of angles. Also, while using hexahedral mesh cells in 3d, the green completion leads to formation of pyramids or prisms, which are not easy to handle by the finite element code and hence one would like to work with hanging vertices.

Apart from AFC schemes, there are certain finite element discretizations that rely on the geometrical properties of the grid such as angle preservation (see [MH85, XZ99]). Hence, one would like to study a continuous finite element in the framework of grids with hanging nodes. Angle preservation is also an important property for a certain class of stabilization methods for Convection-Diffusion-Reaction equations as they provide a sufficient condition for the satisfaction of discrete maximum principle (DMP) (see Chapter 3). To the best of our knowledge, no work has been done in the context of hanging nodes and nonlinear stabilization such as algebraic flux correction schemes (AFC). Some work in the area of hanging nodes can be found in [Grä11] where results have been provided for the lowest order Lagrange elements in the framework of multigrid methods and [CH09] where a unified error analysis for a posteriori error estimation has been provided.

In this chapter, we present the first work regarding the interplay of AFC schemes and grids with hanging nodes. The chapter is divided as follows: In Sec. 6.1 we extend the results from [Grä11] to higher-order Lagrange elements. Next, in Sec. 6.2 we present results concerning

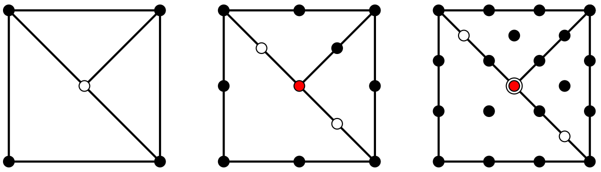


Figure 6.1: Hanging nodes and vertices for P_1, P_2, and P_3 Lagrange elements. Hanging nodes in white and hanging vertices in red.

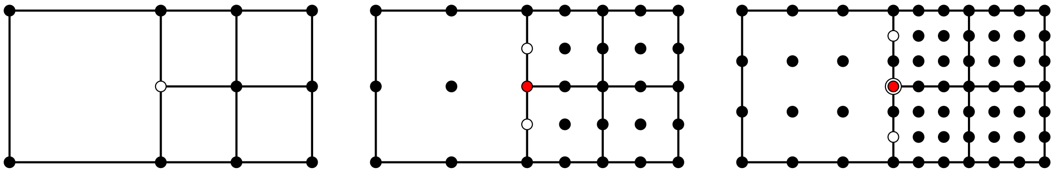


Figure 6.2: Hanging nodes and vertices for Q_1, Q_2, and Q_3 Lagrange elements. Hanging nodes in white and hanging vertices in red.

the behavior of AFC schemes concerning grids with hanging nodes. Finally, numerical simulations illustrating the results provided in the previous section will be presented in Sec. 6.3.

6.1 Hanging Nodes in Theory of Lagrange Finite Elements

In this subsection we extend the results from [Grä11] for hanging nodes from P_1/Q_1 elements to P_k/Q_k elements. First we present some definitions that are used in the terminology of hanging nodes.

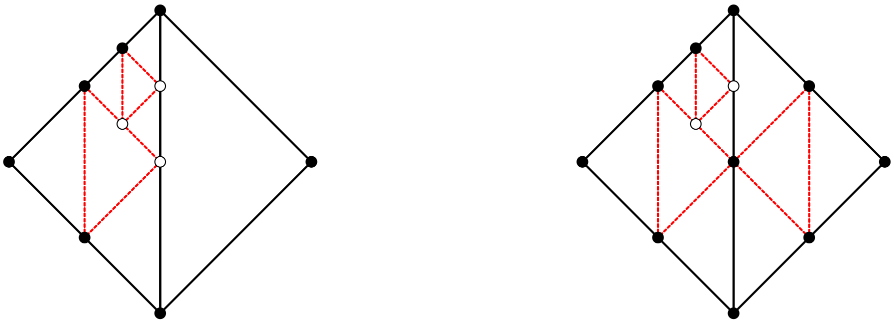
Definition 6.1. (**Hanging vertex**) ([Grä11, Definition 3.6]) Let $\mathcal{T}$ be a triangulation of Ω. Then a vertex $p \in \mathcal{N}_h(\mathcal{T})$ of $\mathcal{T}$ is called a hanging vertex if there is an element $K \in \mathcal{T}$ with $p \in \partial K$ but p is not a vertex of K.

Definition 6.2. (**Hanging node**) Let $\mathcal{T}$ be a triangulation of Ω and $P(\mathcal{T})$ be a Lagrange finite element space defined on $\mathcal{T}$. Then a node $p \in N_F(\mathcal{T})$ of $\mathcal{T}$ is called a hanging node if there is an element $K \in \mathcal{T}$, such that, $p \in K \cap K'$ and $p \in N_F(K)$ but $p \notin N_F(K')$ where K' is a neighbor of K. The set of all hanging nodes is denoted by $H(\mathcal{T})$.

Remark 6.3. Note that for P_1 and Q_1 elements the concepts of hanging vertex and hanging node match. But for P_k or Q_k elements, $k > 1$, they don't match see Fig. 6.1 and Fig. 6.2, where hanging nodes are shown with white color and hanging vertices by red.

Definition 6.4. (**k-irregular triangulation**) ([CH09, Definition 2.4]) If an edge $E \in \mathcal{E}_h$ contains at most k hanging nodes in its inside, we call $\mathcal{T}$ a k-irregular triangulation.

Remark 6.5. In Sec. 6.3 we will work with 1-*irregular* triangulations.



(a) 2-irregular triangulation (b) Red refinement of neighboring cell

Figure 6.3: Refinement rule for avoiding 2-irregular triangulation.

Remark 6.6. To avoid k-irregular triangulations for $k > 1$, the neighboring element is first red refined before the formation of the new hanging node. Fig. 6.3 shows the process in 2d.

Let $\mathcal{T}$ be a conforming triangulation of Ω. For such a triangulation the k^{th} order Lagrangian finite element functions are continuous functions on $\overline{\Omega}$ such that the restrictions to all elements $K \in \mathcal{T}$ are polynomials with degree at most k. It is known that these function spaces are conforming subspaces with respect to $H^1(\Omega)$. However, the same definition also leads to conforming spaces if it is used on non-conforming triangulations.

Definition 6.7. (**k^{th} order conforming space**) Let $\mathcal{T}$ be a triangulation of Ω. The k^{th} order conforming finite element space is defined as

$$S(\mathcal{T}) := \{v \in \mathcal{C}(\overline{\Omega}) : v|_K \in P_k(K) \quad \forall K \in \mathcal{T}\} \subset H^1(\Omega).$$

For conforming triangulations a basis of $S(\mathcal{T})$ is given by the well-known nodal basis functions. To deal with conforming finite element spaces in non-conforming triangulation we first introduce the non-conforming nodal basis functions.

Definition 6.8. (**Non-conforming nodal basis functions**) Let $\mathcal{T}$ be a triangulation of Ω. Then the non-conforming nodal basis function $\varphi_p^{\text{nc}} \in L^2(\Omega)$ associated with $p \in N_F(\mathcal{T})$ is defined as follows: For all $K \in \mathcal{T}$ there is a representative $\varphi_p^{\text{nc}}|_K = \mu_{p,K} \in \mathcal{C}(K)$ with $\mu_{p,K} = \delta_{pq}$ for all nodes q of K.

For a conforming triangulation $\mathcal{T}$ this reduces to $\varphi_p^{\text{nc}} \in S(\mathcal{T})$ and

$$\varphi_p^{\text{nc}}(q) = \delta_{pq} \quad \forall p, q \in N_F(\mathcal{T}),$$

i.e., the set $\{\varphi_p^{\text{nc}}\}_{p \in N_F(\mathcal{T})}$ is the conforming nodal basis of $S(\mathcal{T})$. For a conforming triangulation, $S(\mathcal{T})$ is in general only a subspace of the non-conforming finite element space,

$$S^{\text{nc}}(\mathcal{T}) := \text{span}\{\varphi_p^{\text{nc}} : p \in N_F(\mathcal{T})\}.$$

However, for a non-conforming triangulation it is possible to construct a basis of $S(\mathcal{T})$ from the non-conforming nodal basis $S^{\mathrm{nc}}(\mathcal{T})$ that resembles the usual nodal basis functions when $\mathcal{T}$ is conforming.

Lemma 6.9. *Let $\mathcal{T}$ be a non-conforming triangulation of Ω, i.e., $\mathcal{T}$ has hanging nodes. Then, if $v \in S(\mathcal{T})$, then $\forall q \in H(\mathcal{T})$ there are coefficients a_{qp} with $p \in N_F(\mathcal{T}) \setminus H(\mathcal{T})$ such that,*

$$v(q) = \sum_{p \in N_F(\mathcal{T}) \setminus H(\mathcal{T})} a_{qp} v(p).$$

Proof. Let $q \in H(\mathcal{T})$. Suppose there does not exist any a_{qp} such that

$$v(q) = \sum_{p \in N_F(\mathcal{T}) \setminus H(\mathcal{T})} a_{qp} v(p).$$

As $q \in H(\mathcal{T})$, therefore there exists $K, K' \in \mathcal{T}$ such that $q \in K \cap K'$ and $q \in N_F(K)$ but $q \notin N_F(K')$.

Now,

$$v|_{K'}(x) = \sum_{p_0 \in N_F(K')} v(p_0) \varphi^{\mathrm{nc}}_{p_0}(x),$$

as $q \in K'$

$$\Rightarrow v|_{K'}(q) = \sum_{p_0 \in N_F(K')} v(p_0) \varphi^{\mathrm{nc}}_{p_0}(q).$$

Also, as $q \in K$ and $q \in N_F(K)$,

$$\Rightarrow v|_K(q) = v(q).$$

By continuity of v we have

$$v|_K(q) = v|_{K'}(q) \Rightarrow v(q) = \sum_{p_0 \in N_F(K')} v(p_0) \varphi^{\mathrm{nc}}_{p_0}(q),$$

which is a contradiction and hence the result holds. ☐

Remark 6.10. The proof of the Lemma 6.9 gives a concrete choice for the definition of a_{qp}. Namely $a_{qp} = \varphi^{\mathrm{nc}}_p(q)$ where $p \in N_F(K')$ if $q \in N_F(K)$ and $q \in K \cap K'$.

Remark 6.11. If one would like the solution to be in $S(\mathcal{T})$, then one notes from Lemma 6.9 that the hanging nodes are not free but are dependent.

Theorem 6.12. ([Grä11, Theorem 3.1]) *Let $(\mathcal{T}_0, \cdots, \mathcal{T}_j)$ be a grid hierarchy on Ω with $\mathcal{T}_0$ being conforming. Let us denote $\mathcal{T} = \mathcal{T}_j$, i.e., the final refinement level. Then a basis of $S(\mathcal{T})$ is given by*

$$B(\mathcal{T}) := \left\{ \varphi_p = \varphi^{\mathrm{nc}}_p + \sum_{q \in H(\mathcal{T})} a_{qp} \varphi^{\mathrm{nc}}_q : p \in N_F(\mathcal{T}) \setminus H(\mathcal{T}) \right\}.$$

Proof. The proof from [Grä11] can be extended to higher order elements without any changes. □

6.2 Hanging Nodes in Theory of AFC Schemes

In this subsection we discuss the implementation of hanging nodes for the AFC schemes, the failure of satisfaction of DMP with hanging nodes for the Kuzmin limiter and the modification for the BJK limiter.

6.2.1 Implementation of hanging nodes

The implementation of hanging nodes is a little bit similar to the implementation of Dirichlet nodes, i.e., it works on an algebraic level. Let us denote our finite element matrix on a non-conforming grid by A^{nc} and the corresponding right-hand side by b^{nc}. Hence, our finite problem is to find $u \in S^{\mathrm{nc}}(\mathcal{T})$ such that

$$A^{\mathrm{nc}} u = b^{\mathrm{nc}},$$

where $S^{\mathrm{nc}}(\mathcal{T})$ is a finite element space defined on $\mathcal{T}$. Here A^{nc} and b^{nc} are derived using discontinuous elements from $S^{\mathrm{nc}}(\mathcal{T})$ and hence our solution is discontinuous as well. To restore the continuity of the finite element solution, we look at the variational form of the problem. Let $a_h^{\mathrm{nc}} : S^{\mathrm{nc}}(\mathcal{T}) \times S^{\mathrm{nc}}(\mathcal{T}) \to \mathbb{R}$ be the corresponding bilinear form and f_h be the right-hand side, then our problem is

$$a_h^{\mathrm{nc}}(u, v) = \langle f_h, v \rangle \quad \forall v \in S^{\mathrm{nc}}(\mathcal{T}).$$

First we modify our test space and replace $S^{\mathrm{nc}}(\mathcal{T})$ by $S(\mathcal{T})$. Then

$$a_h^{\mathrm{nc}}(u, v) = \langle f_h, v \rangle \quad \forall v \in S(\mathcal{T}).$$

The algebraic form of the above problem can be written as

$$\bar{A} u = \bar{b},$$

where the right-hand side is assembled using continuous elements. To enforce continuity on the solution we modify the stiffness matrix for the hanging nodes in the same way as that for the Dirichlet nodes, i.e., we modify the rows corresponding to hanging nodes such that the solution at hanging node is continuous with respect to the coupling nodes and set the corresponding right-hand side to zero. Till this point the implementation of hanging nodes is general and can be applied to any higher order elements.

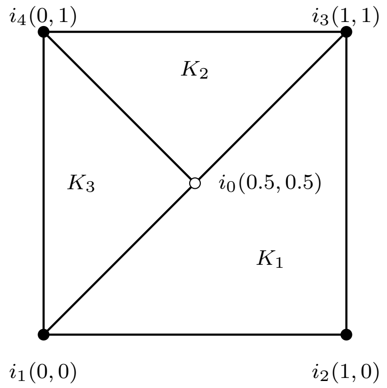


Figure 6.4: Example of a patch failing non-positivity condition for the Kuzmin limiter.

For the AFC scheme, in the first step, a system is assembled that corresponds to a Galerkin finite element discretization of the given equations but with Neumann boundary conditions on the whole boundary. For implementation with hanging nodes, $\bar{A}$ is used to define D and after the computation of limiters, $\bar{A}$ is modified to A with correct entries for hanging rows, where the rows of non-hanging nodes get entries from the rows of the hanging nodes.

Example 6.13. Implementation for P_1 elements We will take a patch as defined in Fig. 6.4 and see how does the stiffness matrix and the right-hand side modify. Initially the system is

$$A^{\mathrm{nc}}u = b^{\mathrm{nc}},$$

where

$$A^{\mathrm{nc}} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ a_{10} & a_{11} & a_{12} & a_{13} & a_{14} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{30} & a_{31} & a_{32} & a_{33} & a_{34} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}, \quad b^{\mathrm{nc}} = \begin{bmatrix} b_0 \\ b_1 \\ b_2 \\ b_3 \\ b_4 \end{bmatrix}.$$

First, we modify A^{nc} and b^{nc} to $\bar{A}$ and $\bar{b}$ by performing row transformations $R_1 \to R_1 + 0.5R_0$ and $R_3 \to R_3 + 0.5R_0$, then

$$\bar{A} = \begin{bmatrix} a_{00} & a_{01} & a_{02} & a_{03} & a_{04} \\ a_{10} + \frac{a_{00}}{2} & a_{11} + \frac{a_{01}}{2} & a_{12} + \frac{a_{02}}{2} & a_{13} + \frac{a_{03}}{2} & a_{14} + \frac{a_{04}}{2} \\ a_{20} & a_{21} & a_{22} & a_{23} & a_{24} \\ a_{30} + \frac{a_{00}}{2} & a_{31} + \frac{a_{01}}{2} & a_{32} + \frac{a_{02}}{2} & a_{33} + \frac{a_{03}}{2} & a_{34} + \frac{a_{04}}{2} \\ a_{40} & a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix}, \quad \bar{b} = \begin{bmatrix} b_0 \\ b_1 + \frac{b_0}{2} \\ b_2 \\ b_3 + \frac{b_0}{2} \\ b_4 \end{bmatrix}.$$

At this step the computation of the limiters in AFC is performed using $\bar{A}$ and $\bar{b}$. Once, the computation is done we modify the hanging row to $(1, -0.5, 0, -0.5, 0)$, where -0.5 appears on the coupling nodes and correspondingly set the right-hand side to 0. Finally, our system of equations is

$$Au = b$$

where

$$A = \begin{bmatrix} 1 & -\frac{1}{2} & 0 & -\frac{1}{2} & 0 \\ \check{a}_{10} + \frac{\check{a}_{00}}{2} & \check{a}_{11} + \frac{\check{a}_{01}}{2} & \check{a}_{12} + \frac{\check{a}_{02}}{2} & \check{a}_{13} + \frac{\check{a}_{03}}{2} & \check{a}_{14} + \frac{\check{a}_{04}}{2} \\ \check{a}_{20} & \check{a}_{21} & \check{a}_{22} & \check{a}_{23} & \check{a}_{24} \\ \check{a}_{30} + \frac{\check{a}_{00}}{2} & \check{a}_{31} + \frac{\check{a}_{01}}{2} & \check{a}_{32} + \frac{\check{a}_{02}}{2} & \check{a}_{33} + \frac{\check{a}_{03}}{2} & \check{a}_{34} + \frac{\check{a}_{04}}{2} \\ \check{a}_{40} & \check{a}_{41} & \check{a}_{42} & \check{a}_{43} & \check{a}_{44} \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ \check{b}_1 + \frac{\check{b}_0}{2} \\ \check{b}_2 \\ \check{b}_3 + \frac{\check{b}_0}{2} \\ \check{b}_4 \end{bmatrix}.$$

Remark 6.14. Depending on the iterative scheme for solving the nonlinear problem, the matrix or the right-hand side change because of the contribution from the limiter (see Chapter 4). Hence, instead of $\{a_{ij}\}$ or $\{b_i\}$, one gets $\{\check{a}_{ij}\}$ or $\{\check{b}_i\}$.

6.2.2 Kuzmin Limiter

In [BJK16], the proof of discrete maximum principle (DMP) for the Kuzmin limiter relies on the assumption of the type $a_{kl} + a_{lk} \leq 0$ where a_{kl} belongs to the stiffness matrix $\bar{A}$ defined previously. For Convection-Diffusion-Reaction equations on conforming grids this condition is satisfied if and only if $(\nabla\varphi_l, \nabla\varphi_k) \leq 0$, which leads to the Delaunay condition (see [BJK16, Remark 14]). For non-conforming grids we don't have such a generalization. One needs to check the condition, $a_{kl} + a_{lk} \leq 0$ individually for all nodes. The next example presents a patch in 2d where this condition fails.

Example 6.15. Let's take a simple 2d example of Convection-Diffusion-Reaction equations with some diffusion, ε, convection, $\boldsymbol{b} = (b_1, b_2)$, and reaction, $c = 0$ on the patch as shown in Fig. 6.4. Then our non-conforming nodal basis functions $\{\varphi_{i_j}^{\mathrm{nc}}\}_{j=0}^{4}$ are given by

$$\varphi_{i_0}^{\mathrm{nc}} = \begin{cases} 0 & \text{in } K_1, \\ 2-2y & \text{in } K_2, \\ 2x & \text{in } K_3, \end{cases} \quad \varphi_{i_1}^{\mathrm{nc}} = \begin{cases} 1-x & \text{in } K_1, \\ 0 & \text{in } K_2, \\ 1-x-y & \text{in } K_3, \end{cases}$$

$$\varphi_{i_2}^{\mathrm{nc}} = \begin{cases} x-y & \text{in } K_1, \\ 0 & \text{in } K_2, \\ 0 & \text{in } K_3, \end{cases} \quad \varphi_{i_3}^{\mathrm{nc}} = \begin{cases} y & \text{in } K_1, \\ x+y-1 & \text{in } K_2, \\ 0 & \text{in } K_3, \end{cases}$$

$$\varphi_{i_4}^{\mathrm{nc}} = \begin{cases} 0 & \text{in } K_1, \\ -x+y & \text{in } K_2, \\ -x+y & \text{in } K_3. \end{cases}$$

After the coupling the conforming nodal basis functions, φ_{i_1} and φ_{i_3} look like

$$\varphi_{i_1} = \varphi_{i_1}^{\mathrm{nc}} + \frac{1}{2}\varphi_{i_0}^{\mathrm{nc}} = \begin{cases} 1-x & \text{in } K_1, \\ 1-y & \text{in } K_2, \\ 1-y & \text{in } K_3, \end{cases} \quad \varphi_{i_3} = \varphi_{i_3}^{\mathrm{nc}} + \frac{1}{2}\varphi_{i_0}^{\mathrm{nc}} = \begin{cases} y & \text{in } K_1, \\ x & \text{in } K_2, \\ x & \text{in } K_3. \end{cases}$$

For rest of the nodes there are not any contributions from the hanging nodes and hence $\varphi_{i_j} = \varphi_{i_j}^{\mathrm{nc}}$ for $j = 0, 2, 4$. We need to check the sign of $a_{i_1 i_3} + a_{i_3 i_1}$. From the bilinear form we have

$$\begin{aligned} a_{i_1 i_3} + a_{i_3 i_1} &= a_{31} + \frac{1}{2}a_{01} + a_{13} + \frac{1}{2}a_{03} \\ &= \varepsilon\left(\nabla\varphi_{i_3}^{\mathrm{nc}}, \nabla\varphi_{i_1}\right) + \left(\boldsymbol{b}\cdot\nabla\varphi_{i_3}^{\mathrm{nc}}, \varphi_{i_1}\right) \\ &\quad + \varepsilon\left(\nabla\varphi_{i_1}^{\mathrm{nc}}, \nabla\varphi_{i_3}\right) + \left(\boldsymbol{b}\cdot\nabla\varphi_{i_1}^{\mathrm{nc}}, \varphi_{i_3}\right). \end{aligned}$$

The gradients of the basis functions required in the above computation are given by

$$\nabla\varphi_{i_1}^{\mathrm{nc}} = \begin{cases} (-1,0) & \text{in } K_1, \\ (0,0) & \text{in } K_2, \\ (-1,-1) & \text{in } K_3, \end{cases} \quad \nabla\varphi_{i_3}^{\mathrm{nc}} = \begin{cases} (0,1) & \text{in } K_1, \\ (1,1) & \text{in } K_2, \\ (0,0) & \text{in } K_3, \end{cases}$$

$$\nabla\varphi_{i_1} = \begin{cases} (-1,0) & \text{in } K_1, \\ (0,-1) & \text{in } K_2, \\ (0,-1) & \text{in } K_3, \end{cases} \quad \nabla\varphi_{i_3} = \begin{cases} (0,1) & \text{in } K_1, \\ (1,0) & \text{in } K_2, \\ (1,0) & \text{in } K_3, \end{cases}$$

Finally consider the sum $a_{i_1 i_3} + a_{i_3 i_1}$,

$$\begin{aligned} a_{i_1 i_3} + a_{i_3 i_1} &= -\varepsilon\left(|K_2| + |K_3|\right) + b_2\int_{K_1}(1-x)ds + (b_1+b_2)\int_{K_2}(1-y)ds \\ &\quad - b_1\int_{K_1} y ds - (b_1+b_2)\int_{K_3} x ds \\ &= \frac{-\varepsilon}{2} + \frac{(b_2 - b_1)}{6}. \end{aligned}$$

For $\varepsilon \leq 0.1$ and $\boldsymbol{b} = (0,1)$ we have $a_{i_1 i_3} + a_{i_3 i_1} > 0$.

Remark 6.16. One can consider the situation of a hanging node in Fig. 6.4 as the limit of a non-Delaunay grid and in this respect, this property of the Kuzmin limiter is not surprising (see Fig. 6.5).

6.2.3 BJK Limiter

We have another definition for the limiter, where the proof of DMP holds for all conforming simplicial grids. This is the BJK limiter defined in [BJK17]. Here the condition of the DMP

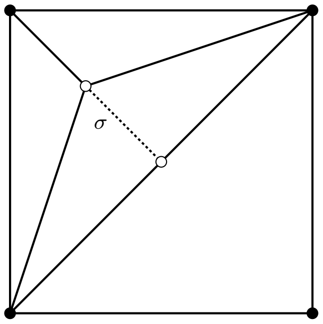


Figure 6.5: Hanging node being a limit of a non-Delaunay grid as $\sigma \to 0$.

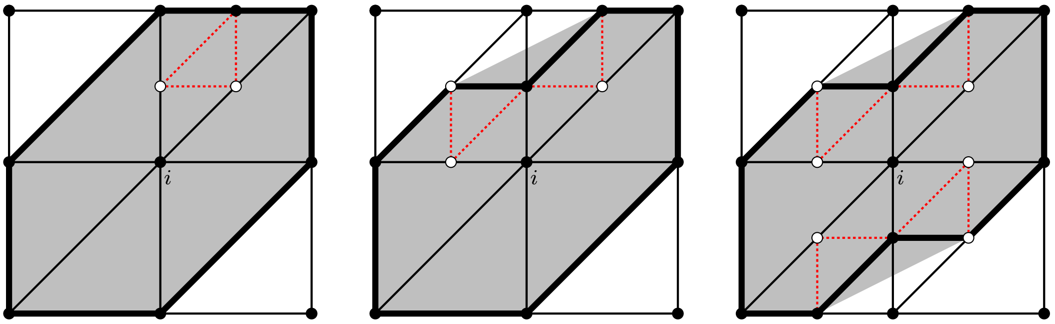


Figure 6.6: Examples of Δ_i for the node x_i with bold lines and Δ_i^{conv} with the shaded area.

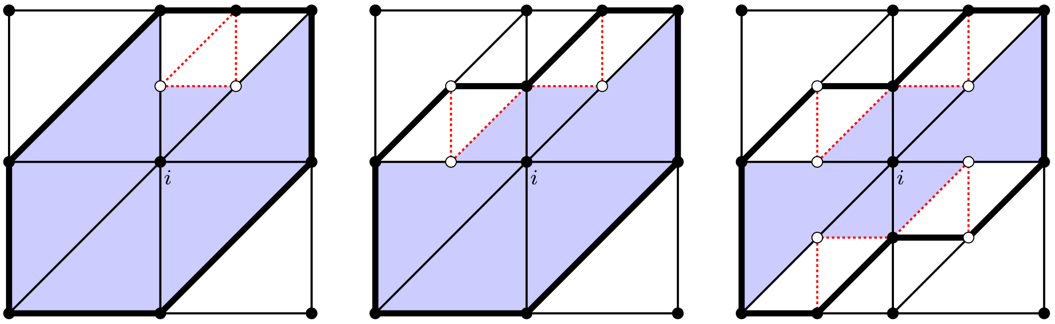


Figure 6.7: Examples of $\Delta_i^{T,\text{conv}}$ for the node x_i.

relies on the properties of the stiffness matrix instead on the triangulation. Let us assume that the condition

$$\sum_{j=1}^{N} a_{ij} \geq 0, \qquad i = 1, \ldots, M$$

is satisfied. Let Δ_i denote $\operatorname{supp}(\varphi_i)$. Examples of Δ_i for x_i are shown in Fig. 6.6 with bold lines and their convex hull, Δ_i^{conv}, by the shaded area. In [BJK17], for conforming grids, Δ_i denoted the patch having the node x_i.

Remark 6.17. One of the main assumptions for AFC schemes is the positivity of the row sum, i.e., $\sum_{j=1}^{N} a_{ij} \geq 0$ (see [BJK17, Eq. (2.6)], [BJK16, Eq. (8)]). With the use of hanging nodes, this condition is still satisfied, as the the positivity of row sum is not affected by adding a positive multiple of a row to another row.

Remark 6.18. For the computation of the limiters a certain constant γ_i is required to show linearity preservation. From [BJK17, Theorem 6.1] the definition of γ_i is given by

$$\gamma_i = \frac{\max\limits_{x_j \in \partial\Delta_i} |x_i - x_j|}{\operatorname{dist}(x_i, \partial\Delta_i^{\text{conv}})}, \qquad i = 1, \ldots, M.$$

The computation of the numerator is easy as compared to the computation of the denominator. For simplices, ideas on computation of Δ_i are given by [BJK17, Remark 6.2]. With hanging nodes the shape of Δ_i is not a polygon made of a union of triangles sharing the node x_i but a generalized polygon (in 2d). This computation is more involved. In our simulations for the denominator we consider Δ_i as all those triangles which share the vertex x_i. Let us denote it by $\Delta_i^{T,\text{conv}}$, see Fig. 6.7. This definition leads to

$$\operatorname{dist}(x_i, \partial\Delta_i^{T,\text{conv}}) \le \operatorname{dist}(x_i, \partial\Delta_i^{\text{conv}}),$$

hence, the value used in simulations might be larger than γ_i. From the theory of conforming grids, it is known that the DMP is satisfied if this parameter is larger than γ_i.

Remark 6.19. The example patches that we have shown are for structured grids that will be used in our simulations. As the BJK limiter can be applied to unstructured grids, we may have presence of triangles of varying sizes and hence requiring generalizations. We would not consider that case in this work, as we need to assume certain shape regularity on the initial grid for the underlying a posteriori error estimates. The analysis for AFC schemes with anisotropic grids remain an open problem [BJKR18].

6.2.4 Limiter Definition

One last thing we want to note is what should be α_{ij} for a hanging node x_i. First, the idea for Dirichlet nodes was used, i.e., $\alpha_{ij} = 1$ for each hanging node x_i. This choice leads to some overshoots. The possible reason being the presence of hanging nodes in the layer and the absence of the artificial diffusion as $\alpha_{ij} = 1$ leads to standard Galerkin method. Hence, we choose $\alpha_{ij} = 0$ for hanging node x_i. This is an overly diffusive approach at least locally. This issue will be studied in the numerical simulations. One should note that none of the applied estimators was derived for grids with hanging nodes.

6.3 Numerical Studies

The numerical studies presented in this section validates the results presented in the previous section. We will use a posteriori error estimators defined in Chapter 5. Let us recall, in Chapter 5 two different techniques for the upper bound were proposed in the energy norm of the error, one used a residual-based approach which we refer to as *AFC-energy* technique and the second approach used the SUPG solution and the SUPG estimators from [JN13], which will be referred as *AFC-SUPG-energy* technique.

For simulations the matrices were assembled exactly and the linear systems were solved using the direct solver UMFPACK [Dav04]. The method fixed point rhs was used for solving the nonlinear problems with the damping parameters as described in Chapter 4.

Example with Interior and Boundary Layers

Let us recall this example. It is given in $\Omega = (0,1)^2$ with $\boldsymbol{b} = (\cos(-\pi/3), \sin(-\pi/3))$, $c = f = 0$ and the Dirichlet boundary condition

$$u_b = \begin{cases} 1 & (y = 1 \wedge x > 0) \text{ or } (x = 0 \wedge y > 0.7), \\ 0 & \text{else.} \end{cases}$$

Here, the convection-dominated case $\varepsilon = 10^{-4}$ is considered. An analytic solution to this problem is not available but we know that $u \in [0,1]$. Hence, this example will help us in showing the violation or satisfaction of the DMP with grids containing hanging nodes. This example will also help in checking the quality of the adaptively refined grids.

To show the violation of DMP, we define a function

$$\mathrm{Var}(u_h) := u_h^{\max} - u_h^{\min}. \tag{6.1}$$

Due to boundary conditions, $u_h^{\max} \geq 1$ and $u_h^{\min} \leq 0$, hence $\mathrm{Var}(u_h) \geq 1$. As, the solution $u \in [0,1]$, one would expect $\mathrm{Var}(u_h) \approx 1$ for all grids.

An initial mesh was defined with two triangles by joining the points $(0,0)$ and $(1,1)$. The simulations were started with a level 2 grid (i.e., #dof = 25), initially uniform refinement was performed till level 4 (i.e., #dof = 289). After that adaptive refinement was performed. AFC schemes are applicable to first order elements, hence P_1 finite elements were used.

First, we study the behavior of $\mathrm{Var}(u_h)$ for the *AFC-energy* technique. For the Kuzmin limiter we see a violation of DMP on grids with hanging nodes but almost satisfaction on grids with conforming closure (see Fig. 6.8 (left)). The failure of DMP is not surprising as this behavior was predicted in Sec. 6.2. Whereas, for the BJK limiter, we observe the satisfaction of DMP on both kinds of grids (see Fig. 6.9 (left)). Next, we study the behavior of $\mathrm{Var}(u_h)$ for the *AFC-SUPG-energy* technique. The results are similar to the results for *AFC-energy* technique for grids with hanging nodes, that is, failure of the DMP with Kuzmin limiter (see Fig. 6.8 (right)) and satisfaction of the DMP with the BJK limiter (see Fig. 6.9 (right)). For grids with conforming closures the results are similar to the *AFC-energy* technique.

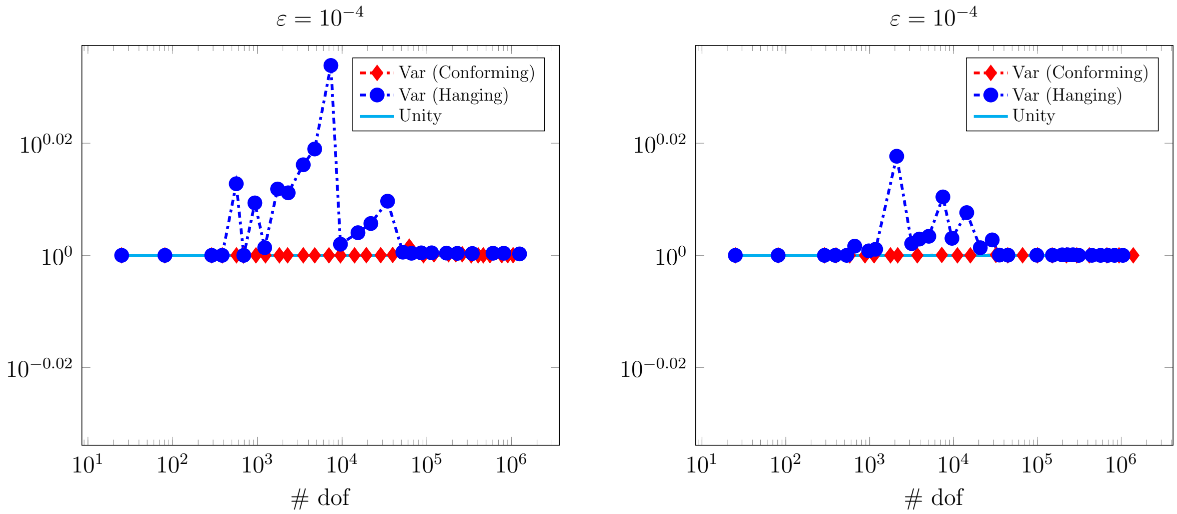


Figure 6.8: Variation for the Kuzmin limiter as defined in (6.1). *AFC-energy* technique (left), *AFC-SUPG-energy* technique (right)

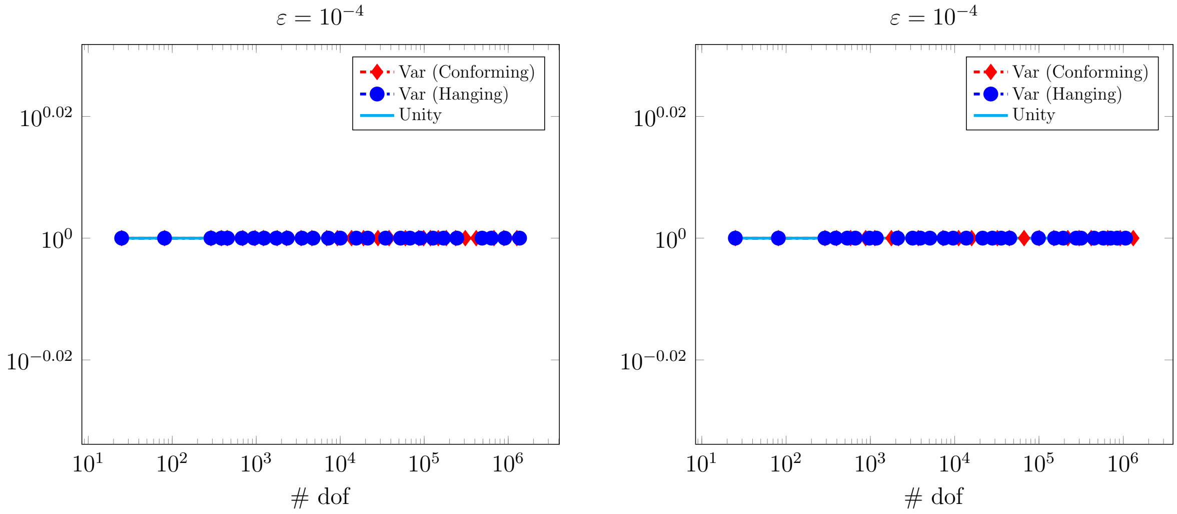


Figure 6.9: Variation for the BJK limiter as defined in (6.1). *AFC-energy* technique (left), *AFC-SUPG-energy* technique (right)

Now, we consider the adaptive grid refinements. The 14th adaptively refined grids with conforming closure are shown in Fig. 6.10 for both the techniques. Comparing the refinement for both the limiters, we observe that more mesh cells are refined for the BJK limiter as compared to the Kuzmin limiter (see Fig. 6.10 for #dof). For the *AFC-SUPG-energy* technique (see Fig. 6.10 (bottom left) for the Kuzmin limiter and (bottom right) for the BJK limiter) we observe that the mesh cells near the internal layer are not refined that much as compared to the *AFC-energy* technique. Also, we see that the limiters do not play an important role in the adaptive refinement.

The 14th adaptively refine grids with hanging nodes are shown in Fig. 6.11. For the *AFC-*

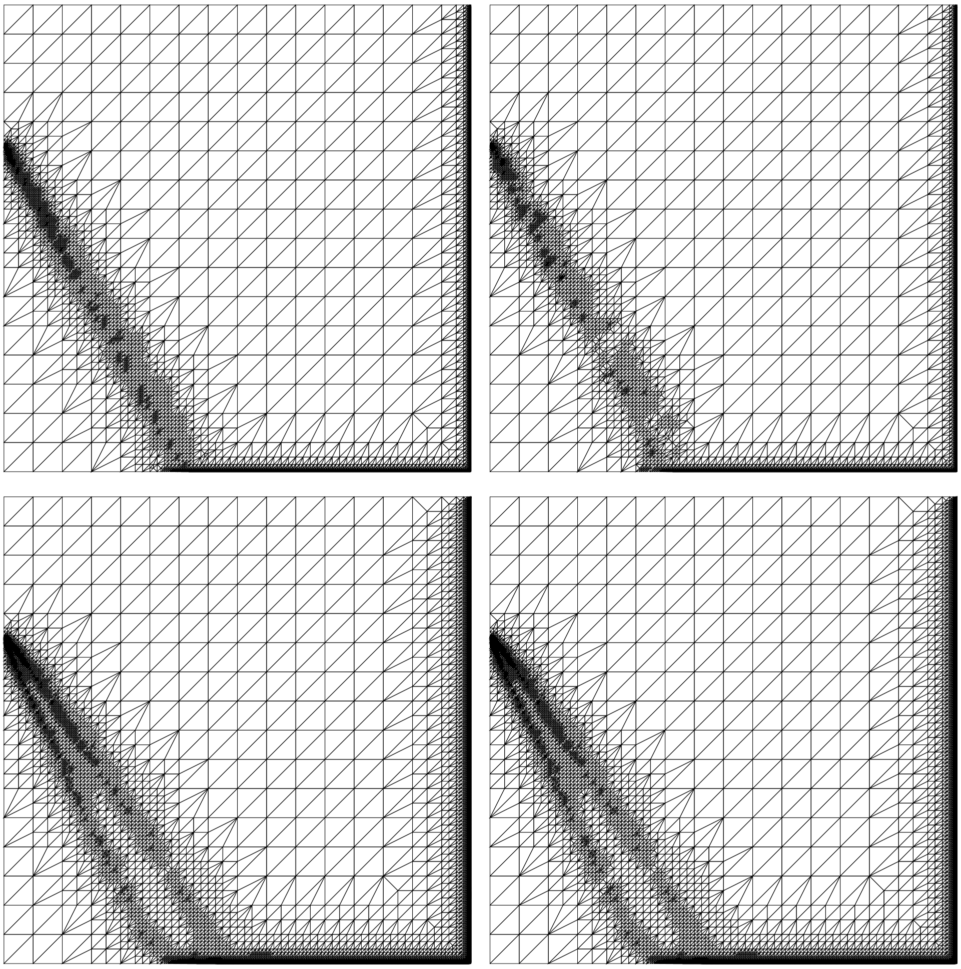


Figure 6.10: 14th adaptively refined grid with conforming closure. Kuzmin limiter+*AFC-energy* technique (14th grid: #dof = 19325 (top left); BJK limiter+*AFC-energy* technique (14th grid: #dof = 28120 (top right) and Kuzmin limiter+*AFC-SUPG-energy* technique (14th grid: #dof = 100620 (bottom left); BJK limiter+*AFC-SUPG-energy* technique (14th grid: #dof = 100538 (bottom right).

energy technique comparing the refinement for both the limiters, we observe that both the meshes are comparable (see Fig. 6.11 for all and hanging #dof). Here, all #dof refer to boundary+hanging+interior degrees of freedom, whereas hanging #dof refers to the hanging nodes. With the *AFC-SUPG-energy* technique we observe that the mesh cells near the internal layer are not refined that much as compared to the *AFC-energy* technique. Similar to conformally closed grids, the limiters do not play an important role in the refinement of the grid.

To check the thickness of the interior layer we follow the idea as in Chapter 5, i.e., compute $smear_{\text{int}}$ (see Eq. (5.58)). We note that in Fig. 6.12, the layers are most properly resolved on conformally closed grids for both the techniques. Overall, for adaptive grid refinement,

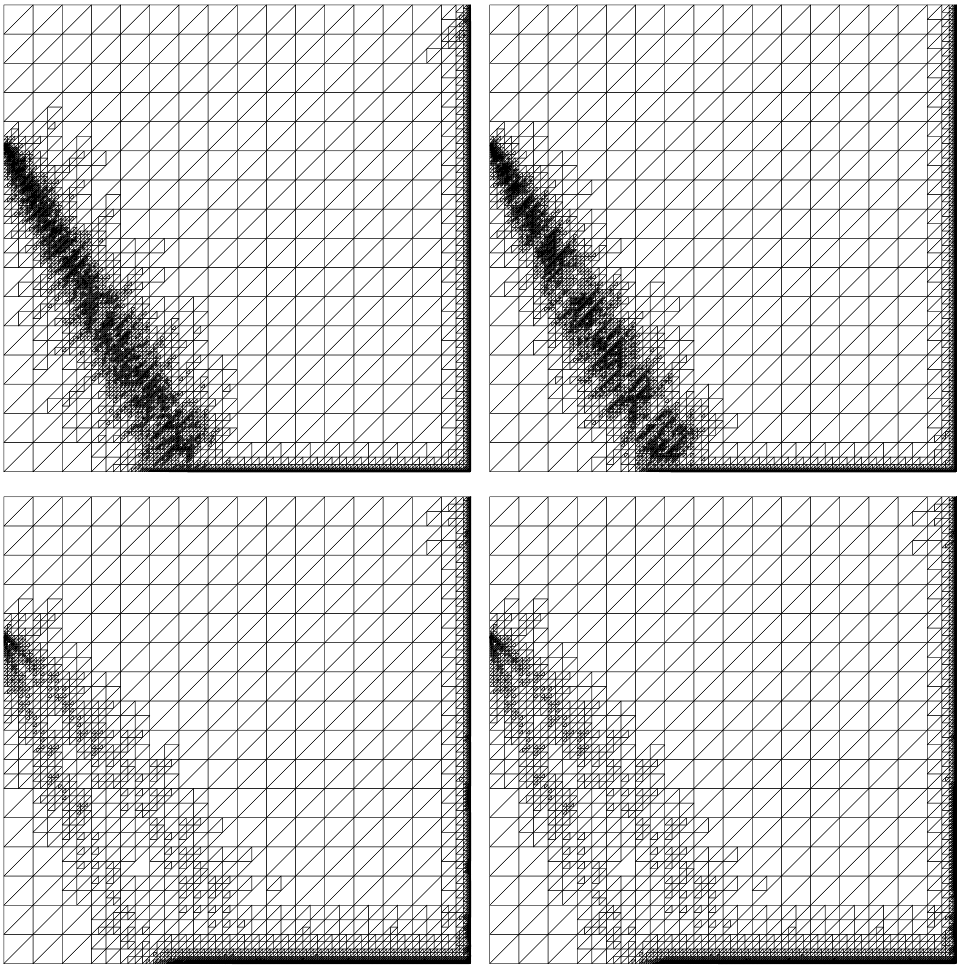


Figure 6.11: 14th adaptively refined grid with hanging nodes. Kuzmin limiter+*AFC-energy* technique (14th grid: all #dof = 34418 , hanging #dof = 10493 (top left); BJK limiter+*AFC-energy* technique (14th grid: all #dof = 34633, hanging #dof = 11029 (top right) and Kuzmin limiter+*AFC-SUPG-energy* technique (14th grid: all #dof = 28961 , hanging #dof = 7027 (top left); BJK limiter+*AFC-SUPG-energy* technique (14th grid: all #dof = 28027, hanging #dof = 6657 (top right).

the *AFC-energy* technique does a much better job since all layers are refined properly, not only the strongest layer. We also note that the layers are better approximated on conformally closed grids as compared to the grids with hanging nodes for the *AFC-SUPG-energy* technique.

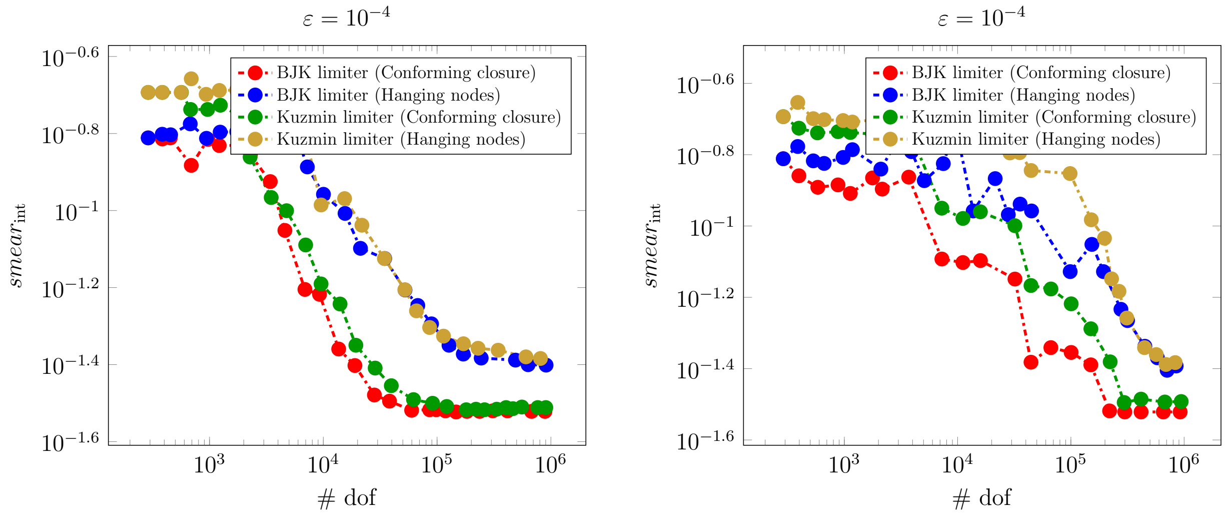


Figure 6.12: Thickness of internal layer. *AFC-energy* technique (left), *AFC-SUPG-energy* technique (right)

6.4 Summary

This is the first work in the direction of hanging nodes in context of non-linear stabilization for convection-diffusion equations. This chapter dealt with two aspects of hanging nodes. First, results have been extended from lower-order Lagrange elements to higher-order elements. Second, we studied the behavior of AFC schemes on grids with hanging nodes.

The following conclusions can be made from the numerical simulations

1. The Kuzmin limiter fails to satisfy the DMP for both the estimators on grids with hanging nodes. A concrete example was provided which justified this behavior.
2. The Kuzmin limiter almost satisfies DMP for both the estimators on grids with conforming closure.
3. The BJK limiter satisfies the DMP for both the estimators on all kinds of grids, i.e., conformally closed grids as well as grids with hanging nodes.
4. The layers were better resolved on conformally closed grids as compared to grids with hanging nodes, irrespective of the choice of limiters.

In summary, the numerical results on grids with hanging nodes are not satisfactory and one should find alternative ways for grid refinements in three dimension and should not continue to work in this direction.

7 Summary and Outlook

7.1 Summary

This thesis presented results for the Algebraic Flux Correction schemes in the framework of iterative solvers and a posteriori error estimation.

We started the thesis with a brief introduction to the analytical and numerical solutions of the Convection-Diffusion-Reaction equations. First we showed that the analytical solutions satisfy the maximum principles in weak form as well as strong from. Then we proved existence and uniqueness of the weak solution to the Convection-Diffusion-Reaction and Evolutionary Convection-Diffusion-Reaction equation. Then we studied the standard finite element approximation, i.e., the Galerkin formulation for the Convection-Diffusion-Reaction equations and it was shown that the Galerkin method fails to give physically consistent results for a small ε. We ended the preliminaries with an overview of a few stabilized FEMs, namely the SUPG and the AFC schemes. It was shown that both the methods compute the layers properly but the SUPG method fails to satisfy the DMP and hence some under and over shoots can be observed. Whereas the AFC schemes satisfy the DMP but because of the nonlinear nature, the system of equations are not easy to solve.

Chapter 4 dealt with the solvers for the AFC schemes. Several iterative solvers were studied including fixed point approaches and Newton-type methods. Advanced methods such as the Newton methods reduced the number of iterations for certain examples but as the computational cost involved in computing the Jacobian matrix, made the method inefficient in terms of computing time. The most simple fixed-point approach referred to as *fixed point rhs* has a structural advantage over other methods. Because of the fixed matrix structure, one can use a direct solver, compute the factorization only once, store it, and use it for subsequent iterations making the method quite efficient. Several algorithmic components such as dynamic damping and Anderson acceleration were also investigated. Anderson acceleration decreased the number of iterations for the Kuzmin limiter, with an appropriate choice of Anderson vectors (namely 10-20). But, it failed to give results for the BJK limiter. In terms of efficiency, the *fixed point rhs* was still more efficient in terms of computing time. Dynamic damping improved the convergence of the nonlinear scheme and is suggested to use with the *fixed point rhs* method. For three dimensional problems, the *fixed point rhs* was still the most efficient method but one needs to use an iterative solver instead of a direct solver. For the fine meshes, an iterative solver such as GMRES along with a proper pre-conditioner was most efficient to solve the problem. Irrespective of the dimension, it was comparatively easier to solve the problem with the Kuzmin limiter as that of the BJK limiter. Altogether, even though one might get fewer iterations, with advanced methods such as Newton methods or with the use of algorithmic components, the simple *fixed point rhs* method with dynamic damping is the most efficient as the advantage of either needing only one factorization of the matrix, in two dimensions or of the high efficiency of the iterative solver in three-dimension compensated the drawback effectively.

Next, in Chapter 5 a posteriori error estimation for the AFC schemes was considered. Here we studied two different approaches for finding a global upper bound in the energy norm of the error. One was the standard residual-based approach referred to as the *AFC-energy* approach and one used the SUPG norm along with the SUPG estimators referred to as *AFC-SUPG-energy* approach. Results were compared based on the effectivity index and adaptive grid refinements. The *AFC-energy* estimator was shown not to be robust with respect to ε and hence for the convection-dominated regime, the *AFC-SUPG-energy* approach gave a better effectivity index. For the BJK limiter, the effectivity was better than the Kuzmin limiter with *AFC-energy* approach, whereas in for the *AFC-SUPG-energy* approach the choice of limiter did not play an important role because of the dominating nature of the SUPG estimators. With adaptive grid refinement, the problem could become locally diffusion dominated and hence one has to use the BJK limiter as one can observe reduced order of convergence for the Kuzmin limiter. This situation is only possible when ε is comparatively larger than the mesh size. But for small ε, one has to use the Kuzmin limiter because of problems arising to solve the nonlinear problem with the BJK limiter. In regards to adaptive grid refinement, the *AFC-energy* approach approximated the layer much better as compared to the *AFC-SUPG-energy* approach. Altogether, the *AFC-SUPG-energy* approach gave better results in terms of effectivity index whereas the *AFC-energy* approach has better results in terms of adaptive grid refinements.

Lastly, Chapter 6 is an extension of results of Chapter 5 where grids with hanging node are considered. Here we also present a brief overview of the hanging nodes theory for Lagrange elements and the results are extended from lower-order elements to higher-order elements. Then we move to the interplay of hanging nodes and AFC schemes. An example is presented in two dimensions which shows the failure of the Kuzmin limiter to satisfy the DMP on grids with hanging nodes which are verified numerically. Numerical studies compare the results based on adaptive grid refinements and satisfaction of the DMP. The BJK limiter satisfies the DMP on all kinds of grids irrespective of the choice of estimators whereas the Kuzmin limiter almost satisfies the DMP on conformally closed grids. The adaptive grid refinement was better on conformally closed grids and the layers were computed sharply as compared to grids with hanging nodes. Altogether, one should not use the AFC schemes for grids with hanging nodes and find alternative grid refinement strategies.

7.2 Outlook

We believe research is an ongoing journey and when you think you found the solution to a problem, ten different questions arise from them. The research presented here is in no way an exception to this rule. Now, we will mention some open questions for the work that has been provided here.

The results presented in Chapter 4 start as an initial point of research in the efficient computation of the nonlinear problem. Although, the *fixed point rhs* method was time-efficient one would like to examine certain quasi-Newton approaches or a combination of the Newton and *fixed point rhs* approach such as the Newton-Dogleg methods (see [PSWS08]) so that one can switch between the two methods. In three dimension, we saw the advantage of using an iterative solver over a direct solver but the question remains, *What is the most appropriate iterative solver?* The future work in this area includes the investigation of different iterative solvers such as left-GMRES, right-GMRES, conjugate gradient, multigrid, SSOR, etc. along with a proper preconditioner such as Jacobi, SSOR, SOR, etc.

In Chapter 5 a non-robust residual-based estimator was proposed in the energy norm. Also, a SUPG approach was discussed. Only one work has been done in this direction (see [ABR17]) where also the estimator was not robust. Hence, one would like to develop robust estimators, preferably in the natural norm of the system, i.e., the AFC norm. Also, one would prefer to have the estimators independent of the choice of limiters. The local efficiency that was proved needs to be extended for the edge estimates. Finally, simulations in three-dimension should be performed to understand the estimators better.

Chapter 6 gave results on grids with hanging nodes. Initially, one should analyze the hanging nodes in the framework of non-conforming elements. Even though this work does not lie in the scope of AFC schemes, we believe this is an interesting topic of research. We treated the hanging nodes as Dirichlet nodes and assigned the value zero to the limiters, which lead to a diffusive solution. Hence, one should define the limiters properly and give some concrete results in this direction. Grids with hanging nodes are mostly used in three dimensions so as to avoid non-admissible or problematic elements, hence one needs to study the behavior of BJK limiter in 3d. The results provided here work as a stepping stone in this direction.

We would like to finish this section by mentioning certain open questions for the AFC schemes in general.

1. Analysis for improved order of convergence for the AFC schemes. In [BJK16] error bounds of $\mathcal{O}(h^{1/2})$ were derived but it can be seen numerically that better convergence rates are available. The reason being the analysis relied on general assumptions of a limiter.

2. Stability and error analysis of the time-dependent counterpart of AFC schemes, the Flux-Corrected transport (FEM-FCT). The ideas of AFC schemes originate from FEM-FCT methodology which is applied to Evolutionary Convection Diffusion Reaction equations. There stability and convergence order remains an open question.

3. Efficient solution of the nonlinear problems in FEM-FCT. Currently, only a preliminary work is available in this direction (see [JN12]). Different iterative schemes need to be investigated.

4. Convergence analysis on anisotropic meshes and for mixed boundary conditions. In [BJK16, BJK17] the analysis was performed on non-anisotropic meshes and Dirichlet boundary conditions were prescribed. One would like to extend the analysis from these papers to the aforementioned cases.

A lot of questions still remain open in the area of stabilized schemes for Convection-Diffusion-Reaction equations (see [JKN18]).

Appendix A

Algorithms

This appendix summarizes all the algorithms that were presented in Chapter 4.

Remark A.1. **Notations** We denote the residue of the solution $\underline{u}^k$ by r^k for $k \in \mathbb{N}$. The residue for solving a system of equations $A\underline{x} = \underline{b}$, $A \in \mathbb{R}^{N \times N}, \underline{b} \in \mathbb{R}^N$, is defined as

$$r^k = \|A\underline{u}^k - \underline{b}\|_{l^2},$$

if $\underline{u}^k$ is the numerical solution of the system.

A.1 Algorithm for Dynamic Damping

Flowchart A.1 summarizes the Dynamic damping parameter $\omega^{(\nu)}$ used in (4.3). Detailed explanation of the algorithm can be found in [JK08, Sec. 5]. For simplicity we denote $\omega^{(\nu)}$ by ω.

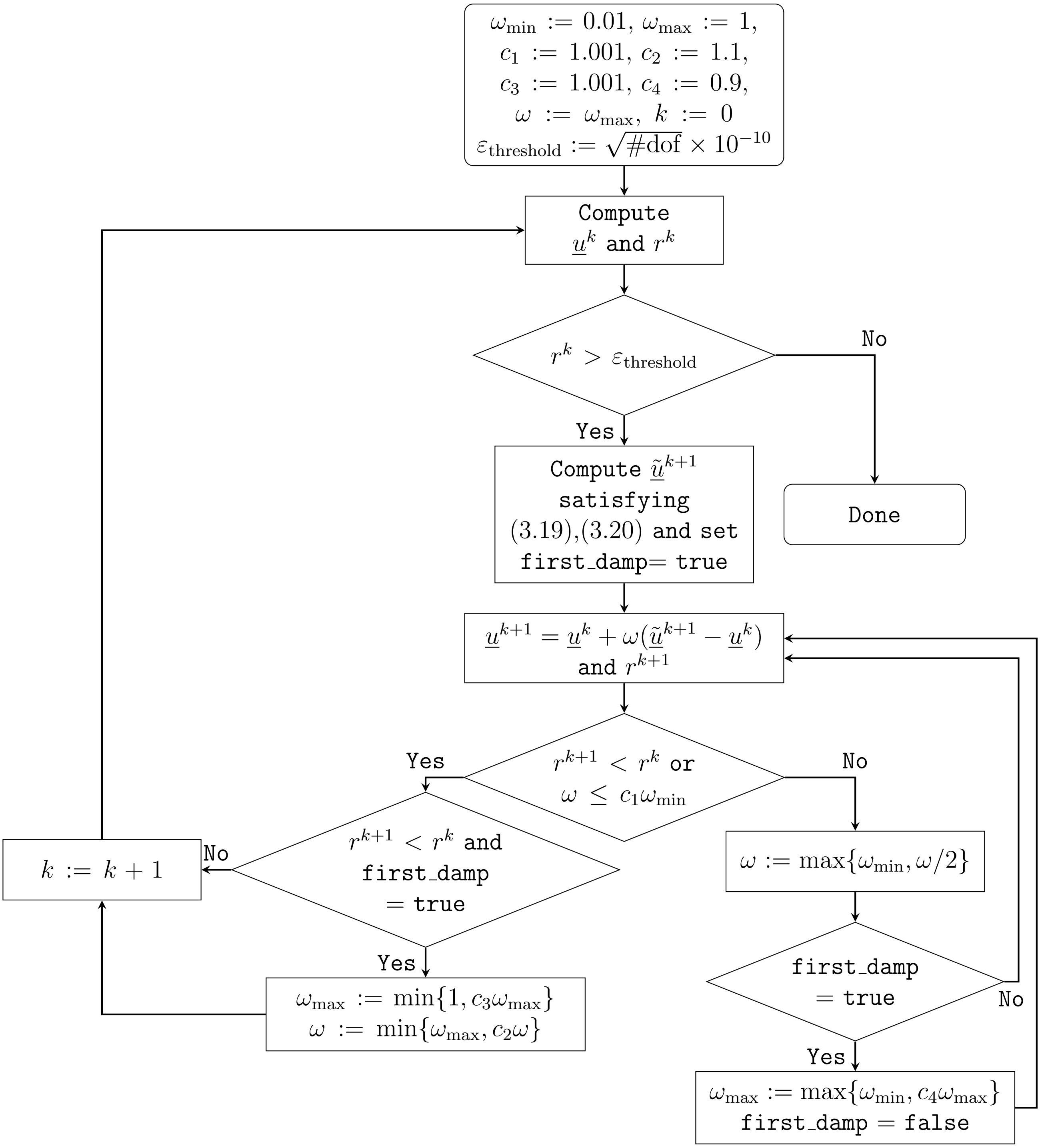

$\omega_{\min} := 0.01$, $\omega_{\max} := 1$,
$c_1 := 1.001$, $c_2 := 1.1$,
$c_3 := 1.001$, $c_4 := 0.9$,
$\omega := \omega_{\max}$, $k := 0$
$\varepsilon_{\text{threshold}} := \sqrt{\#\text{dof}} \times 10^{-10}$
Compute $\underline{u}^k$ and r^k
$r^k > \varepsilon_{\text{threshold}}$
No
Yes
Done
Compute $\tilde{\underline{u}}^{k+1}$ satisfying (3.19),(3.20) and set first_damp= true
$\underline{u}^{k+1} = \underline{u}^k + \omega(\tilde{\underline{u}}^{k+1} - \underline{u}^k)$ and r^{k+1}
$r^{k+1} < r^k$ or $\omega \leq c_1\omega_{\min}$
Yes
No
$r^{k+1} < r^k$ and first_damp = true
No
$k := k + 1$
Yes
$\omega_{\max} := \min\{1, c_3\omega_{\max}\}$
$\omega := \min\{\omega_{\max}, c_2\omega\}$
$\omega := \max\{\omega_{\min}, \omega/2\}$
first_damp = true
No
Yes
$\omega_{\max} := \max\{\omega_{\min}, c_4\omega_{\max}\}$
first_damp = false

Figure A.1: Adaptive choice of damping parameter

A.2 Algorithm for Anderson Acceleration

Flowchart A.2 summarizes the Anderson acceleration (Algorithm AA) described in [WN11]. ε_{And} denotes the tolerance. In our studies it was set equal to the tolerance used in Sec. 4.3 i.e., $\sqrt{\#\text{dof}} \cdot 10^{-10}$.

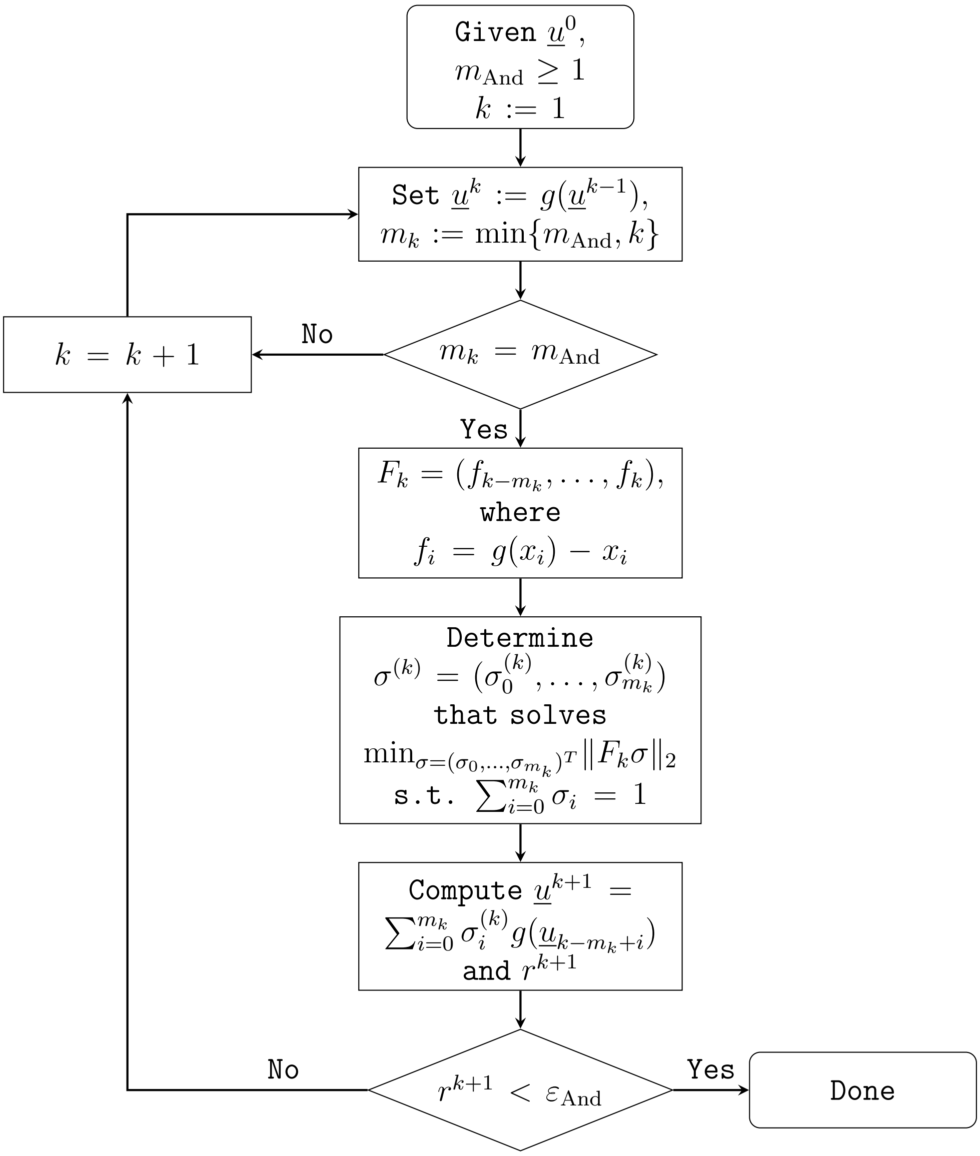

Given $\underline{u}^0$, $m_{\text{And}} \geq 1$ $k := 1$
Set $\underline{u}^k := g(\underline{u}^{k-1})$, $m_k := \min\{m_{\text{And}}, k\}$
$m_k = m_{\text{And}}$
No
$k = k + 1$
Yes
$F_k = (f_{k-m_k}, \dots, f_k)$, where $f_i = g(x_i) - x_i$
Determine $\sigma^{(k)} = (\sigma_0^{(k)}, \dots, \sigma_{m_k}^{(k)})$ that solves $\min_{\sigma=(\sigma_0,\dots,\sigma_{m_k})^T} \|F_k\sigma\|_2$ s.t. $\sum_{i=0}^{m_k} \sigma_i = 1$
Compute $\underline{u}^{k+1} = \sum_{i=0}^{m_k} \sigma_i^{(k)} g(\underline{u}_{k-m_k+i})$ and r^{k+1}
$r^{k+1} < \varepsilon_{\text{And}}$
No
Yes
Done

Figure A.2: Anderson Acceleration

Remark A.2. In the original paper by Anderson [And65], a general step was given for finding $\underline{u}^{k+1}$ [WN11, Eq (1.2)]),

$$\underline{u}^{k+1} = (1 - \Gamma_k) \sum_{i=0}^{m_k} \sigma_i^{(k)} \underline{u}^{k-m_k+i} + \Gamma_k \sum_{i=0}^{m_k} \sigma_i^{(k)} g(\underline{u}^{k-m_k+i}).$$

We performed our simulations with $\Gamma_k = 1$. It has been noted in [PE13], that, for nonlinear problems the value of Γ_k is really sensitive to the solution and the method may diverge for small Γ_k. We observed similar results in our simulations and hence the above formulation was not studied.

A.3 Algorithm for Newton Dynamic Damping

Flowchart A.3 summarizes the Newton dynamic damping used for formal Newton's method with an adaptive damping parameter ω_{Newt} a described in Chapter 4. In our simulations $\varepsilon_{\text{Newt-tol}}$ was set as 10^{-3} and initial value of $\omega_{\text{Newt}} \in [0.1, 1.0]$ was set to 0.25.

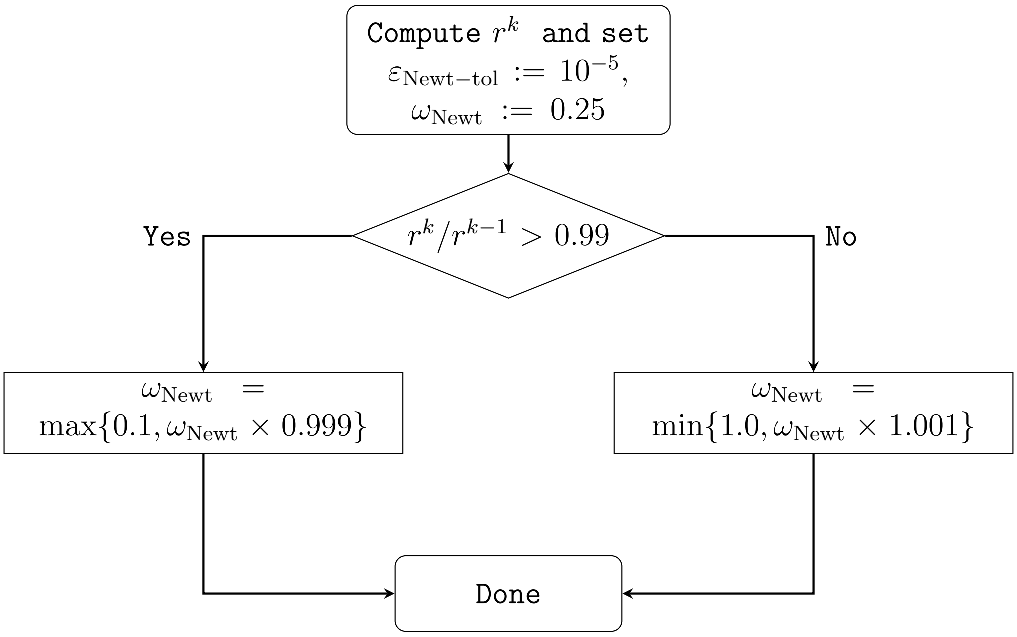

Compute r^k and set $\varepsilon_{\text{Newt-tol}} := 10^{-5}$, $\omega_{\text{Newt}} := 0.25$
$r^k/r^{k-1} > 0.99$
Yes
No
$\omega_{\text{Newt}} = \max\{0.1, \omega_{\text{Newt}} \times 0.999\}$
$\omega_{\text{Newt}} = \min\{1.0, \omega_{\text{Newt}} \times 1.001\}$
Done

Figure A.3: Adaptive choice of Newton damping parameter

Appendix B

Numerical Values for a Posteriori Error Estimators

This appendix provides numerical values for the effectivity indices presented in Sec. 5.3.

Example 5.3.1: A Known 2d Solution with a Boundary Layer

Table B.1: Effectivity index for $\varepsilon = 10^{-3}$ using *AFC-energy* technique and the BJK limiter

#dof	$\|u - u_h\|_a$	η_{d_h}	η	η_{eff}
25.0	0.0786	10.3	11.0	139
81.0	0.122	15.6	17.3	142
289	0.142	22.4	24.2	171
344	0.134	125	126	942
423	0.124	102	102	824
602	0.113	67.2	67.3	595
950	0.0955	39.4	39.5	413
1240	0.0850	30.6	30.6	360
1720	0.0673	18.0	18.0	268
2660	0.0507	9.68	9.71	192
3150	0.0439	6.29	6.31	144
3730	0.0382	2.60	2.64	69.1
5350	0.0369	0.977	1.07	29.1
8420	0.0259	0.456	0.552	21.3
10 200	0.0221	0.222	0.346	15.6
15 500	0.0170	0.0599	0.211	12.5
22 900	0.0128	0.0211	0.154	12.0

30 400	0.0107	0.0126	0.129	12.0
52 400	0.008 12	0.008 88	0.0968	11.9
70 400	0.006 80	0.006 49	0.0809	11.9
101 000	0.005 50	0.005 70	0.0658	12.0
154 000	0.004 52	0.003 54	0.0542	12.0
213 000	0.003 79	0.001 70	0.0454	12.0
310 000	0.003 07	0.000 907	0.0365	11.9
391 000	0.002 72	0.000 545	0.0324	11.9
555 000	0.002 32	0.000 464	0.0276	11.9
760 000	0.001 96	0.000 414	0.0235	12.0
1 010 000	0.001 67	0.000 251	0.0200	12.0

Table B.2: Effectivity index for $\varepsilon = 10^{-3}$ using *AFC-energy* technique and the Kuzmin limiter

#dof	$\|u - u_h\|_a$	η_{d_h}	η	η_{eff}
25.0	0.0779	11.7	12.3	158
81.0	0.122	17.0	18.6	152
289	0.141	24.3	25.8	183
344	0.133	134	134	1010
435	0.123	106	106	866
638	0.112	69.1	69.2	616
1050	0.0938	40.4	40.4	431
1750	0.0671	23.4	23.4	349
2430	0.0550	17.2	17.2	313
3270	0.0416	10.9	10.9	262
3910	0.0371	8.45	8.46	228
5600	0.0309	7.67	7.68	249
6870	0.0260	6.04	6.05	233
10 400	0.0212	4.54	4.54	214
13 800	0.0194	4.16	4.17	214
19 800	0.0168	4.24	4.24	252
22 900	0.0146	3.32	3.32	228
27 600	0.0132	2.96	2.97	225
34 800	0.0117	2.50	2.50	214
43 900	0.0107	2.08	2.09	196
54 800	0.009 89	1.70	1.70	172
67 000	0.009 52	1.41	1.41	148
80 600	0.0106	2.01	2.01	190
104 000	0.0118	2.85	2.85	242
124 000	0.0125	2.89	2.89	232
144 000	0.008 92	1.89	1.89	211
178 000	0.008 55	1.99	1.99	233

210 000	0.008 77	2.06	2.06	235
248 000	0.008 81	2.00	2.00	227
295 000	0.008 71	1.99	1.99	229
354 000	0.008 74	2.04	2.04	233
423 000	0.008 53	1.98	1.98	232
505 000	0.004 72	1.02	1.03	217
623 000	0.004 29	0.941	0.941	219
792 000	0.004 86	1.12	1.12	230
974 000	0.004 95	1.21	1.21	244
1 210 000	0.005 17	1.20	1.20	232

Table B.3: Effectivity index for $\varepsilon = 10^{-3}$ using *AFC-SUPG-energy* technique and the BJK limiter

#dof	$\|u - u_h\|_a$	η	η_{eff}
25.0	0.0786	13.7	175
81.0	0.122	13.7	112
289	0.142	11.7	82.3
352	0.133	7.96	59.7
463	0.122	5.47	44.7
646	0.112	3.91	34.8
1050	0.0937	2.33	24.8
1690	0.0692	1.45	21.0
2020	0.0629	1.21	19.2
3640	0.0393	0.804	20.5
7490	0.0265	0.653	24.6
14 400	0.0195	0.572	29.4
39 500	0.0108	0.203	18.7
50 800	0.009 50	0.108	11.4
63 500	0.009 22	0.0673	7.30
87 100	0.007 90	0.0452	5.72
118 000	0.005 92	0.0333	5.62
157 000	0.005 06	0.0266	5.26
211 000	0.004 23	0.0222	5.25
312 000	0.003 24	0.0170	5.26
462 000	0.002 64	0.0137	5.19
881 000	0.001 89	0.009 73	5.15
1 400 000	0.001 44	0.007 46	5.18

Table B.4: Effectivity index for $\varepsilon = 10^{-3}$ using *AFC-SUPG-energy* technique and the Kuzmin limiter

#dof	$\|u - u_h\|_a$	η	η_{eff}
25.0	0.0779	13.7	176
81.0	0.122	13.7	113
289	0.141	11.7	82.5
352	0.133	7.96	59.8
463	0.122	5.47	44.7
646	0.112	3.91	34.8
1050	0.0940	2.33	24.7
1690	0.0695	1.45	20.9
2020	0.0632	1.21	19.1
3640	0.0388	0.802	20.7
7490	0.0271	0.652	24.1
14 400	0.0196	0.571	29.1
39 400	0.0112	0.203	18.2
49 700	0.009 79	0.111	11.4
61 700	0.009 39	0.0718	7.65
83 700	0.009 97	0.0468	4.69
112 000	0.008 16	0.0356	4.36
151 000	0.006 18	0.0279	4.52
202 000	0.008 24	0.0247	3.00
270 000	0.006 95	0.0206	2.96
341 000	0.005 58	0.0175	3.13
430 000	0.004 44	0.0150	3.39
606 000	0.006 55	0.0148	2.26
845 000	0.005 67	0.0125	2.21

List of Notations

Table B.5: Greek Symbols

	Notation	Description	Section
α			
	$\boldsymbol{\alpha}$	Multi-index	1.3
	α_{ij} /α_E	Solution dependent limiters for AFC schemes	3.4
β			
	β_{ij}	$\alpha_{ij} d_{ij}(u_j - u_i)$	4.1
	$\tilde{\beta}_{ij}$	Regularized version of β_{ij}	4.1
φ			
	$\varphi_h(x_i)$ /φ_i	Nodal functional corresponding o x_i	3.4
	$\hat{\varphi}_i$	Nodal functional on reference element corresponding to x_i	5.1
	φ_i^{nc}	Non conforming nodal functional corresponding to x_i	6.1
Δ			
	δ_K	SUPG stabilization weights	3.3
	$\delta_{F\in\mathcal{F}_{h,N}}$	Kronecker delta function for Neumann faces	5.2.1.2
	Δ_i	Compact support of φ_i	3.4
	Δ_i^{conv}	Convex hull of Δ_i	3.4
	$\Delta_i^{T,\text{conv}}$	Union of all triangles sharing x_i	6.2
ε			
	ε	Diffusion coefficient	2.1
	$\varepsilon_{\text{threshold}}$	Threshold value for dynamic damping	A.1
	ε_{And}	Tolerance for Anderson acceleration	A.2
	$\varepsilon_{\text{Newt}-\text{tol}}$	Tolerance for ω_{Newt}	A.3
η			
	η_K	Global upper bound	5.2.1.2
	$\eta_{\text{Int},K}$	Interior local estimator of K	5.2.1.2
	$\eta_{\text{Face},K}$	Face local estimator of $F \subset \partial K$	5.2.1.2
	$\eta_{d_h,K}$	Edge local estimator of $E \subset \partial K$	5.2.1.2
	η_{SUPG}	Global upper bound from [JN13]	5.3
	$\eta_{\text{AFC}-\text{SUPG}}$	Norm of difference of u_{SUPG} and u_{AFC}	5.3
	η_{eff}	Effectivity index	5.3
γ			

	γ_i	Linearity preserving parameter in BJK limiter	3.4
	γ_0	Parameter in BBK limiter	3.4
	Γ	Boundary of Ω	2.1
	Γ_D	Dirichlet Boundary of Ω	2.1
	Γ_N	Neumann Boundary of Ω	2.1
κ			
	κ_E^K	$(n-2)$ dimensional simplex opposite E	3.2
	κ_{low}	Lower bound for solution	4.1.2.4
	κ_{upp}	Upper bound for solution	4.1.2.4
	κ	Number of Anderson vectors	4.2
Ω			
	Ω	Bounded domain $\subset \mathbb{R}^d$	2.1
	ω	Damping parameter for iteration	4.1
	ω_{fp}	Damping parameter for *fixed point matrix*	4.1.1
	ω_{Newt}	Damping parameter for *formal Newton*	4.1.2.4
	ω_F	Set of mesh cells having common face F	5.1
	ω_K	Set of mesh cells having a joint face with K	5.1
	$\omega_{\min}$	Lower bound for ω in dynamic damping	A.1
	$\omega_{\max}$	Upper bound for ω in dynamic damping	A.1
Φ			
	Φ_{ij}	$\alpha_{ij}(u_j - u_i)$	3.4
ρ			
	ρ_K	Diameter of largest ball inside K	5.1
σ			
	σ_0	Lower bound for reaction	3.2
	σ	Smoothness parameter in regularization	4.1.2.4
	$\sigma_i^{(k)}$	Variables used for constrained minimization algorithm in Anderson acceleration	A.2
θ			
	θ_i	Angle opposite faces F_i	5.1
	θ_{ij}/θ_E^K	Dihedral angle between faces F_i and F_j	5.1
ψ			
	ψ_K	Interior bubble function on K	5.1
	$\hat{\psi}_K$	Interior bubble function on reference element	5.1
	ψ_F	Face bubble function on F	5.1
	$\hat{\psi}_F$	Face bubble function on reference element	5.1

Table B.6: Latin Symbols

	Notation	**Description**	**Section**
$\mathcal{A}$			
	$a(\cdot,\cdot)$	Bilinear form of Convection-Diffusion-Reaction equations	2.3
	$a_h(\cdot,\cdot)$	Approximation of bilinear form $a(\cdot, cdot)$	3.4
	$a_{\mathrm{AFC}}(\cdot,\cdot)$	Bilinear form of AFC equations	5.1
	A_h	Ansatz space	3.3
	A	$\{a_{ij}\}_{i,j=1}^{N}$, Finite element matrix	3.4
	$\mathbb{A}$	FEM using homogeneous Dirichlet boundary condition	3.4
	$\tilde{\mathbb{A}}$	$\mathbb{A}+\mathbb{D}$	3.4
	A^{nc}	Finite element matrix using non-conforming basis function	6.2
$\mathcal{B}$			
	$\boldsymbol{b}$	Convective transport	2.1
	$\boldsymbol{b}_h$	Polynomial approximation of $\boldsymbol{b}$	5.2.1.2
	$B(\mathcal{T})$	Basis of $S(\mathcal{T})$	6.1
$\mathcal{C}$			
	c	Reaction coefficient	2.1
	c_h	Polynomial approximation of reaction c	5.2.1.2
	$C^k(\Omega)$	Space of functions having k continuous derivatives	2.2
	C_{elliptic}	Elliptic constant	2.2
	C_{bound}	Boundedness constant	2.3
	C_{PF}	Poincaré Friedrich's constant	2.3
	C_L	Lipschitz continuity constant	3.4
	C_{shrg}	Shape regularity constant	5.1
	$C_{\cos}$	$\max_{1\le i\le 3}\{\cos(\theta_i)\}$	5.1
	C_Y	Generalized Young's inequality constant	5.1
	C_{inv}	Inverse estimate constant	5.1
	C_I	Interpolation estimate constant	5.1
	$C_{\mathrm{T1}}, C_{\mathrm{T2}}$	Trace inequality constant	5.1
	C_{edge}	Constant appearing in Eq. (5.16)	5.1
	C_F	Constant depending on edge estimate constant and C_I	5.2.1.1
	$C_{\mathrm{edge,max}}$	$\max_{K\in\mathcal{T}_h} C_{\mathrm{edge}}$	5.2.1.1
	C_K	Constant appearing in Eq. (5.42)	5.2.1.2
	C_{FB}	Constant appearing in Eq. (5.44)	5.2.1.2
$\mathcal{D}$			
	$d_h(w;z,v)$	$\sum_{i,j=1}^{N}(1-\alpha_{ij}(w))d_{ij}(z(x_j)-z(x_i))v(x_i)$	3.4
	$\mathbb{D}$	$\{d_{ij}\}_{i,j=1}^{N}$, Artificial diffusion matrix	3.4
	DF	Jacobian matrix of F_i	4.1.2.4
$\mathcal{E}$			
	$E_i\ /E$	Edge of a simplex	3.2

	$\mathcal{E}_h$	Set of all edges of a triangulation	5.1
$\mathcal{F}$			
	$\underline{f}_{\text{flux}}$	Total flux	2.1
	$\underline{f}^{\text{conv}}_{\text{flux}}$	Convective flux	2.1
	$\underline{f}^{\text{diff}}_{\text{flux}}$	Diffusive flux	2.1
	$\hat{f}$	Source/Sink term with reaction	2.1
	f	Source/Sink term	2.1
	f_{ij}	$d_{ij}(u_j - u_i)$ Anti-diffusive fluxes	3.4
	f_h	Polynomial approximation of f	5.2.1.2
	F_i /F	$(n-1)$-dim simplex opposite k_i	3.2
	$\mathcal{F}_h$	Set of all faces of $\mathcal{T}_h$	5.1
	$\mathcal{F}_{h,\Omega}$	Set of all interior aces of $\mathcal{T}_h$	5.1
	$\mathcal{F}_{h,D}$	Set of all Dirichlet faces of $\mathcal{T}_h$	5.1
	$\mathcal{F}_{h,N}$	Set of all Neumann faces of $\mathcal{T}_h$	5.1
	$\mathcal{F}_K$	Affine map from $\hat{K} \to K$	5.1
$\mathcal{G}$			
	g	Neumann boundary conditions	2.1
	g_h	Polynomial approximation of g	5.2.1.2
$\mathcal{H}$			
	h_K	Width of mesh cell K	5.1
	h_E	Width of edge E	5.1
	h	Mesh width	5.1
	$H^k(\Omega)$	Sobolev spaces, $W^{k,2}(\Omega)$	1.3
	$H^k_0(\Omega)$	Closure of $C^\infty_0(\Omega)$ in H^k norm	1.3
	$H^{-1}(\Omega)$	Dual space of $H^1_0(\Omega)$	1.3
	$H^1_D(\Omega)$	$\{v \in H^1(\Omega) : v\|_{\Gamma_D} = u^b$	5.1
	$H(\mathcal{T})$	Set of hanging nodes	6.1
$\mathcal{I}$			
	$i_h u$	Lagrange interpolation of u	3.4
	$I_h u$	Quasi-interpolation of u	5.1
$\mathcal{K}$			
	k_i	Vertex of simplex K	3.2
	k	Iterate counter	4.1
	K_i /K	Simplex in $\mathcal{T}_h$	3.2
	$\hat{K}$	Reference Simplex	5.1
$\mathcal{L}$			
	$L_p(\Omega)$	Space of Lebesgue integrable functions for $1 \leq p\infty$	1.3
$\mathcal{M}$			
	$\max_\sigma\{x,y\}$	$\frac{x+y+\sqrt{(x-y)^2+\sigma}}{2}$	4.1.2.4
	$\min_\sigma\{x,y\}$	$\frac{x+y-\sqrt{(x-y)^2+\sigma}}{2}$	4.1.2.4
	m_{And}	Number of iterates that need to	

		be stored for Anderson acceleration	A.2
	M_{ite}	Matrix used in fixed point iteration	4.1
$\mathcal{N}$			
	$\boldsymbol{n}$	Outward unit normal	2.1
	$\mathcal{N}_h$	Set of vertices of $\mathcal{T}_h$	5.1
	$N_F(\mathcal{T})$	Set of nodal functionals	5.1
	N_F	Number of cells sharing a face	5.2.1.2
$\mathcal{P}$			
	P_k	Lagrange finite elements of order k on simplices	3.3
	$P_i^{\pm}$	Intermediate fluxes in Kuzmin and BJK limiter	3.4
	$\tilde{P}_i^{\pm}$	Regularized intermediate fluxes in Kuzmin and BJK limiter	3.4
	$P(\mathcal{T})$	Finite element function on triangulation $\mathcal{T}$	5.1
$\mathcal{Q}$			
	q_i	$\gamma_i \sum\limits_{j \in S_i} d_{ij}$	3.4
	$Q_i^{\pm}$	Intermediate fluxes in Kuzmin and BJK limiter	3.4
	$\tilde{Q}_i^{\pm}$	Regularized intermediate fluxes in Kuzmin and BJK limiter	3.4
	Q_k	Lagrange finite elements of order k on hexahedron	3.4
$\mathcal{R}$			
	r_k	Residue of solution u_k	A.2
	$R_i^{\pm}$	Intermediate fluxes in Kuzmin and BJK limiter	3.4
	$\tilde{R}_i^{\pm}$	Regularized intermediate fluxes in Kuzmin and BJK limiter	3.4
	$R_K(u_h)$	Interior residual on mesh cell K	5.2.1.1
	$R_F(u_h)$	Face residual on face F	5.2.1.1
	$R_{K,h}(u_h)$	Polynomial approximation of $R_K(u_h)$	5.2.1.2
	$R_{F,h}(u_h)$	Polynomial approximation of $R_F(u_h)$	5.2.1.2
$\mathcal{S}$			
	$smear_{\text{int}}$	Thickness of internal layer	5.3
	S_i	$S_i \supset \{j \in \{1, \ldots, N\} \backslash \{i\} : a_{ij} \neq 0 \text{ or } a_{ji} > 0\}$	3.4
	$\mathbb{S}$	Finite system of closed connected subsets of $\overline{\Omega}$	5.1
	$S(\mathcal{T})$	k^{th} order conforming finite element space	6.1
	$S^{\text{nc}}(\mathcal{T})$	k^{th} order non-conforming finite element space	6.1
$\mathcal{T}$			
	$\boldsymbol{t}_E$	Tangential unit vector on edge E	5.1
	tol_{sw}	Switching tolerance	A.1
	T	Final time	2.1
	T_h	Test space	3.3
	$\mathcal{T}_h$ / $\mathcal{T}$	Triangulation of domain Ω	5.1
$\mathcal{U}$			
	u	Concentration of the reactant	2.1
	u_i^b	Dirichlet boundary values	2.1
	u_0	Initial condition	2.1

	$\underline{u}$	Vector in $\mathbb{R}^N$	3.4
	$\tilde{\underline{u}}$	Vector in $\mathbb{R}^M$	3.4
	$u_i^{\max}$	$\max\limits_{j\in S_i\cup\{i\}} u_j$, Maximum of u_j in patch S_i	4.1.2.3
	$u_i^{\min}$	$\min\limits_{j\in S_i\cup\{i\}} u_j$, Minimum of u_j in patch S_i	4.1.2.3
	u_{AFC}	AFC solution	5.2.2
	u_{SUPG}	SUPG solution	5.2.2
$\mathcal{V}$			
	V	Banach space	2.3
	V'	Dual space of Banach space V	2.3
	V_{u_b}	$\{v \in H^1(\Omega) : v\|_\Gamma = u^b\}$	2.3
	V_D	$\{v \in H^1(\Omega) : v\|_{\Gamma_D} = 0\}$	2.3
	V_h	Finite dimensional subset of V	3.1
	V_i	Vertices of mesh cell K	5.1
	$\hat{V}_i$	Vertices of reference mesh cell $\hat{K}$	5.1
$\mathcal{W}$			
	$W^{k,p}(\Omega)$	Sobolev spaces	1.3

Table B.7: Norms and Semi-norms

Notation	**Description**
$\\|\cdot\\|_V$	Induced norm from $(\cdot,\cdot)_V$
$\\|\cdot\\|_{L^p(\Omega)}$	Norm on $L^p(\Omega)$
$\\|\cdot\\|_{l^2}$	Euclidean norm
$\\|\cdot\\|_{k,p,\Omega}$	Norm on $W^{k,p}(\Omega)$
$\|\cdot\|_{k,p,\Omega}$	Semi-norm on $W^{k,p}(\Omega)$
$\\|\cdot\\|_{k,\Omega}$	Norm on $H^k(\Omega)$
$\|\cdot\|_{k,\Omega}$	Semi-norm on $H^k(\Omega)$
$\\|\cdot\\|_a$	Energy norm $\left(:= \left(\varepsilon\|\cdot\|_{1,\Omega}^2 + \sigma_0\\|\cdot\\|_{0,\Omega}^2\right)^{1/2}\right)$
$\\|\cdot\\|_{\mathrm{AFC}}$	AFC norm $\left(:= (\\|\cdot\\|_a^2 + d_h(\cdot;\cdot,\cdot))^{1/2}\right)$
$\\|\cdot\\|_{\mathrm{SUPG}}$	SUPG norm

Bibliography

[AABR13] M. Ainsworth, A. Allendes, G. R. Barrenechea, and R. Rankin. Fully computable a posteriori error bounds for stabilised FEM approximations of convection-reaction-diffusion problems in three dimensions. *Internat. J. Numer. Methods Fluids*, 73(9):765–790, 2013.

[ABR17] A. Allendes, G. R. Barrenechea, and R. Rankin. Fully computable error estimation of a nonlinear, positivity-preserving discretization of the convection-diffusion-reaction equation. *SIAM J. Sci. Comput.*, 39(5):A1903–A1927, 2017.

[ACF$^+$11] M. Augustin, A. Caiazzo, A. Fiebach, J. Fuhrmann, V. John, A. Linke, and R. Umla. An assessment of discretizations for convection-dominated convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 200(47-48):3395–3409, 2011.

[Ada75] R. A. Adams. *Sobolev spaces.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[And65] D. G. Anderson. Iterative procedures for nonlinear integral equations. *J. Assoc. Comput. Mach.*, 12:547–560, 1965.

[AO00] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis.* Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[APS05] R. Araya, A. H. Poza, and E. P. Stephan. A hierarchical a posteriori error estimate for an advection-diffusion-reaction problem. *Math. Models Methods Appl. Sci.*, 15(7):1119–1139, 2005.

[AR10] M. Ainsworth and R. Rankin. Fully computable error bounds for discontinuous Galerkin finite element approximations on meshes with an arbitrary number of levels of hanging nodes. *SIAM J. Numer. Anal.*, 47(6):4112–4141, 2010.

[AS02] M. S. Adam and J. Sibert. Population dynamics and movements of skipjack tuna (katsuwonus pelamis) in the maldivian fishery: Analysis of tagging data from an advection-diffusion-reaction model. *Aquatic Living Resources*, 15:13–23, 03 2002.

[Bah69] N. S. Bahvalov. On the optimization of the methods for solving boundary value problems in the presence of a boundary layer. *Ž. Vyčisl. Mat i Mat. Fiz.*, 9:841–859, 1969.

[BB97] J. P. Boris and D. L. Book. Flux-corrected transport. I. SHASTA, a fluid transport algorithm that works [J. Comput. Phys. 11 (1973), no. 1, 38–69]. *J. Comput. Phys.*, 135(2):170–186, 1997. With an introduction by Steven T. Zalesak, Commemoration of the 30th anniversary {of J. Comput. Phys.}.

[BB17] S. Badia and J. Bonilla. Monotonicity-preserving finite element schemes based on differentiable nonlinear stabilization. *Comput. Methods Appl. Mech. Engrg.*, 313:133–158, 2017.

[BBK17] G. R. Barrenechea, E. Burman, and F. Karakatsani. Edge-based nonlinear diffusion for finite element approximations of convection–diffusion equations and its relation to algebraic flux-correction schemes. *Numerische Mathematik*, 135(2):521–545, Feb 2017.

[Bey95] J. Bey. Tetrahedral grid refinement. *Computing*, 55(4):355–378, 1995.

[BH82] A. N. Brooks and T. 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).

[BJK16] G. R. Barrenechea, V. John, and P. Knobloch. Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal.*, 54(4):2427–2451, 2016.

[BJK17] G. R. Barrenechea, V. John, and P. Knobloch. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci.*, 27(3):525–548, 2017.

[BJKR18] G. R. Barrenechea, V. John, P. Knobloch, and R. Rankin. A unified analysis of algebraic flux correction schemes for convection-diffusion equations. *SeMA J.*, 75(4):655–685, 2018.

[BKK08] J. Brandts, S. Korotov, and M. Křížek. On the equivalence of regularity criteria for triangular and tetrahedral finite element partitions. *Comput. Math. Appl.*, 55(10):2227–2233, 2008.

[BR78] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[BR81] I. Babuška and W. C. Rheinboldt. A posteriori error analysis of finite element solutions for one-dimensional problems. *SIAM J. Numer. Anal.*, 18(3):565–589, 1981.

[BS08] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008.

[BSW83] R. E. Bank, A. H. Sherman, and A. Weiser. Refinement algorithms and data structures for regular local mesh refinement. In *Scientific computing (Montreal, Que., 1982)*, IMACS Trans. Sci. Comput., I, pages 3–17. IMACS, New Brunswick, NJ, 1983.

[CH09] C. Carstensen and J. Hu. Hanging nodes in the unifying theory of a posteriori finite element error control. *J. Comput. Math.*, 27(2-3):215–236, 2009.

[Cia70] P. G. Ciarlet. Discrete maximum principle for finite-difference operators. *Aequationes Math.*, 4:338–352, 1970.

[Cia78] P. G. Ciarlet. *The finite element method for elliptic problems*. North-Holland Publishing Co., Amsterdam-New York-Oxford, 1978. Studies in Mathematics and its Applications, Vol. 4.

[Dav04] T. A. Davis. Algorithm 832: UMFPACK V4.3—an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Software*, 30(2):196–199, 2004.

[DW11] P. Deuflhard and M. Weiser. *Numerische Mathematik 3*. de Gruyter Lehrbuch. [de Gruyter Textbook]. Walter de Gruyter & Co., Berlin, 2011. Adaptive Lösung partieller Differentialgleichungen. [Adaptive solutions of partial differential equations].

[ESV10] A. Ern, A. F. Stephansen, and M. Vohralík. Guaranteed and robust discontinuous Galerkin a posteriori error estimates for convection-diffusion-reaction problems. *J. Comput. Appl. Math.*, 234(1):114–130, 2010.

[Eva10] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[Ewi83] R. Ewing. The mathematics of reservoir simulation, volume 1 (editor). *SIAM, Philadelphia, PA*, 1983.

[FHM$^+$00] P. Farrell, A. Hegarty, J. Miller, E. O'Riordan, and G. Shishkin. *Robust computational techniques for boundary layers*, volume 16 of *Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, 2000.

[Fre42] H. Freudenthal. Simplizialzerlegungen von beschränkter Flachheit. *Ann. of Math. (2)*, 43:580–582, 1942.

[GJM$^+$16] S. Ganesan, V. John, G. Matthies, R. Meesala, S. Abdus, and U. Wilbrandt. An object oriented parallel finite element scheme for computations of pdes: Design and implementation. *2016 IEEE 23rd International Conference on High Performance Computing Workshops (HiPCW)*, pages 2–11, 2016.

[God59] S. K. Godunov. A difference method for numerical calculation of discontinuous solutions of the equations of hydrodynamics. *Mat. Sb. (N.S.)*, 47 (89):271–306, 1959.

[GR09] C. Geuzaine and J. F. 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.

[Grä11] C. Gräser. *Convex minimization and phase field models.* PhD thesis, Freie Universität, Berlin, 2011.

[GT17] S. Ganesan and L. Tobiska. *Finite Elements: Theory and Algorithms.* Cambridge University Press, New York, NY, USA, 1st edition, 2017.

[HB79] 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.

[HDF$^+$06] G. Hauke, M. H. Doweidar, D. Fuster, A. Gómez, and J. Sayas. Application of variational a-posteriori multiscale error estimation to higher-order elements. *Comput. Mech.*, 38(4-5):356–389, 2006.

[HDF11] G. Hauke, M. H. Doweidar, and D. Fuster. A posteriori error estimation for computational fluid dynamics: the variational multiscale approach. In *Multiscale methods in computational mechanics*, volume 55 of *Lect. Notes Appl. Comput. Mech.*, pages 19–38. Springer, Dordrecht, 2011.

[Hem96] P. W. Hemker. A singularly perturbed model problem for numerical computation. *J. Comput. Appl. Math.*, 76(1-2):277–285, 1996.

[HFD08] G. Hauke, D. Fuster, and M. H. Doweidar. Variational multiscale a-posteriori error estimation for multi-dimensional transport problems. *Comput. Methods Appl. Mech. Engrg.*, 197(33-40):2701–2718, 2008.

[HMM86] T. J. R. Hughes, M. Mallet, and A. Mizukami. A new finite element formulation for computational fluid dynamics. II. Beyond SUPG. *Comput. Methods Appl. Mech. Engrg.*, 54(3):341–355, 1986.

[Jha20] A. Jha. A Residual Based A Posteriori Error Estimators for AFC Schemes for Convection-Diffusion Equations. *arXiv e-prints*, page arXiv:2005.02938, May 2020.

[JJ19] A. Jha and V. John. A study of solvers for nonlinear AFC discretizations of convection-diffusion equations. *Comput. Math. Appl.*, 78(9):3117–3138, 2019.

[JJ20] A. Jha and V. John. On basic iteration schemes for nonlinear afc discretizations. In Gabriel R. Barrenechea and John Mackenzie, editors, *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2018*, pages 113–128, Cham, 2020. Springer International Publishing.

[JK49] M. Jakob and S. P. Kezios. *Heat transfer.* N.Y : Wiley, 1949. V.2 with the technical and editorial assistance of S.P.Kezios.

[JK07a] V. John and P. 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.

[JK07b] V. John and P. Knobloch. On the performance of SOLD methods for convection-diffusion problems with interior layers. *Int. J. Comput. Sci. Math.*, 1(2-4):245–258, 2007.

[JK08] V. John and P. 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.

[JKN18] V. John, P. Knobloch, and J. 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.

[JMR$^+$09] V. John, T. Mitkova, M. Roland, K. Sundmacher, L. Tobiska, and A. Voigt. Simulations of population balance systems with one internal coordinate using finite element methods. *Chemical Engineering Science*, 64(4):733 – 741, 2009. 3rd International Conference on Population Balance Modelling.

[JN12] V. John and J. Novo. On (essentially) non-oscillatory discretizations of evolutionary convection-diffusion equations. *J. Comput. Phys.*, 231(4):1570–1586, 2012.

[JN13] V. John and J. Novo. A robust SUPG norm a posteriori error estimator for stationary convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 255:289–305, 2013.

[Joh87] C. Johnson. *Numerical solution of partial differential equations by the finite element method*. Cambridge University Press, Cambridge, 1987.

[Joh00] V. John. A numerical study of a posteriori error estimators for convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 190(5-7):757–781, 2000.

[KK05] 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.

[Kno06] P. Knobloch. Improvements of the Mizukami-Hughes method for convection-diffusion equations. *Comput. Methods Appl. Mech. Engrg.*, 196(1-3):579–594, 2006.

[Kno19] P. Knobloch. A linearity preserving algebraic flux correction scheme of upwind type satisfying the discrete maximum principle on arbitrary meshes. In Florin Adrian Radu, Kundan Kumar, Inga Berre, Jan Martin Nordbotten, and Iuliu Sorin Pop, editors, *Numerical Mathematics and Advanced Applications ENUMATH 2017*, pages 909–918, Cham, 2019. Springer International Publishing.

[KP08] T. Kröger and T. Preusser. Stability of the 8-tetrahedra shortest-interior-edge partitioning method. *Numer. Math.*, 109(3):435–457, 2008.

[KQ95] M. Křížek and L. Qun. On diagonal dominance of stiffness matrices in 3D. *East-West J. Numer. Math.*, 3(1):59–69, 1995.

[KR89] R. Kornhuber and R. Roitzsch. On adaptive grid refinement in the presence of internal or boundary layers. Technical Report SC-89-05, ZIB, Takustr. 7, 14195 Berlin, 1989.

[KS71] Evelyn F. K. and Lee A. S. Model for chemotaxis. *Journal of Theoretical Biology*, 30(2):225 – 234, 1971.

[Kuz06] D. Kuzmin. On the design of general-purpose flux limiters for finite element schemes. I. Scalar convection. *J. Comput. Phys.*, 219(2):513–531, 2006.

[Kuz07] D. Kuzmin. Algebraic flux correction for finite element discretizations of coupled systems. *Computational Methods for Coupled Problems in Science and Engineering II*, 01 2007.

[Kuz08] D. Kuzmin. On the design of algebraic flux correction schemes for quadratic finite elements. *Journal of Computational and Applied Mathematics*, 218(1):79 – 87, 2008. Special Issue: Finite Element Methods in Engineering and Science (FEMTEC 2006).

[Kuz09] D. Kuzmin. Explicit and implicit FEM-FCT algorithms with flux linearization. *J. Comput. Phys.*, 228(7):2517–2534, 2009.

[Kuz12] D. Kuzmin. Linearity-preserving flux correction and convergence acceleration for constrained Galerkin schemes. *J. Comput. Appl. Math.*, 236(9):2317–2337, 2012.

[Lax54] P. D. Lax. Weak solutions of nonlinear hyperbolic equations and their numerical computation. *Comm. Pure Appl. Math.*, 7:159–193, 1954.

[LKSM17] C. Lohmann, D. Kuzmin, J. N. Shadid, and S. Mabuza. Flux-corrected transport algorithms for continuous Galerkin methods based on high order Bernstein finite elements. *J. Comput. Phys.*, 344:151–186, 2017.

[Loh] C. Lohmann. *Physics-compatible finite element methods for scalar and tensorial advection problems.* Springer.

[MH85] A. Mizukami and T. J. R. Hughes. A Petrov-Galerkin finite element method for convection-dominated flows: an accurate upwinding technique for satisfying the maximum principle. *Comput. Methods Appl. Mech. Engrg.*, 50(2):181–193, 1985.

[MOS96] J. Miller, E. O'Riordan, and G. Shishkin. *Fitted numerical methods for singular perturbation problems.* World Scientific Publishing Co., Inc., River Edge, NJ, 1996. Error estimates in the maximum norm for linear problems in one and two dimensions.

[Ong89] M. E. G. Ong. Hierarchical basis preconditioners for second order elliptic problems in three dimensions. Technical Report 89–3, Dept. of Applied Math. University of Washington, Seattle, 1989.

[Ose11] C. W. Oseen. *Über die Stoke'sche Formel und über eine verwandte Aufgabe in der Hydrodynamik: Mitteilung 2.* Number v. 1 in Arkiv för matematik, astronomi och fysik. Almqvist & Wiksell, 1911.

[Ost37] A. Ostrowski. Über die determinanten mit überwiegender Hauptdiagonale. *Comment. Math. Helv.*, 10(1):69–96, 1937.

[PC86] A. K. Parrott and M.A. Christie. Fct applied to the 2-d finite element solution of tracer transport by single phase flow in a porous medium. In *Proc. ICFD Conf. on Numerical Methods in Fluid Dynamics, Oxford University Press*, volume 609, 1986.

[PCSM87] S. J. Polak, Den H. C., W. H. A. Schilders, and P. Markowich. Semiconductor device modelling from the numerical point of view. *International Journal for Numerical Methods in Engineering*, 24(4):763–838, 1987.

[PE13] F. A. Potra and H. Engler. A characterization of the behavior of the Anderson acceleration on linear problems. *Linear Algebra Appl.*, 438(3):1002–1011, 2013.

[Ple77] R. J. Plemmons. M-matrix characterizations. I. Nonsingular M-matrices. *Linear Algebra and Appl.*, 18(2):175–188, 1977.

[PSWS08] R. P. Pawlowski, J. P. Simonis, H. F. Walker, and J. N. Shadid. Inexact Newton dogleg methods. *SIAM J. Numer. Anal.*, 46(4):2112–2132, 2008.

[REI$^+$07] A. Rap, L. Elliott, D. B. Ingham, D. Lesnic, and X. Wen. The inverse source problem for the variable coefficients convection-diffusion equation. *Inverse Probl. Sci. Eng.*, 15(5):413–440, 2007.

[Riv84] M. C. Rivara. Mesh refinement processes based on the generalized bisection of simplices. *SIAM J. Numer. Anal.*, 21(3):604–613, 1984.

[RST08] H. G. Roos, M. Stynes, and L. 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.

[San08] G. Sangalli. Robust a-posteriori estimator for advection-diffusion-reaction problems. *Math. Comp.*, 77(261):41–70, 2008.

[SS86] Y. Saad and M. H. Schultz. GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM J. Sci. Statist. Comput.*, 7(3):856–869, 1986.

[Sty05] M. Stynes. Steady-state convection-diffusion problems. *Acta Numer.*, 14:445–508, 2005.

[Tem77] R. Temam. *Navier-Stokes equations. Theory and numerical analysis.* North-Holland Publishing Co., Amsterdam-New York-Oxford, 1977. Studies in Mathematics and its Applications, Vol. 2.

[TV15] L. Tobiska and R. Verfürth. Robust a posteriori error estimates for stabilized finite element methods. *IMA J. Numer. Anal.*, 35(4):1652–1671, 2015.

[Ver94] R. Verfürth. A posteriori error estimation and adaptive mesh-refinement techniques. In *Proceedings of the Fifth International Congress on Computational and Applied Mathematics (Leuven, 1992)*, volume 50, pages 67–83, 1994.

[Ver98] R. Verfürth. A posteriori error estimators for convection-diffusion equations. *Numer. Math.*, 80(4):641–663, 1998.

[Ver05] R. Verfürth. Robust a posteriori error estimates for nonstationary convection-diffusion equations. *SIAM J. Numer. Anal.*, 43(4):1783–1802, 2005.

[Ver13] R. Verfürth. *A posteriori error estimation techniques for finite element methods.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013.

[WBA$^+$16] U. Wilbrandt, C. Bartsch, N. Ahmed, N. Alia, F. Anker, L. Blank, A. Caiazzo, S. Ganesan, S. Giere, G. Matthies, R. Meesala, A. Shamim, J. Venkatesan, and V. John. Parmoon – a modernized program package based on mapped finite elements. *Computers and Mathematics with Applications*, 74:74–88, 2016.

[WN11] H. F. Walker and P. Ni. Anderson acceleration for fixed-point iterations. *SIAM J. Numer. Anal.*, 49(4):1715–1735, 2011.

[WR17] K. Y. Wertheim and T. Roose. A mathematical model of lymphangiogenesis in a zebrafish embryo. *Bull. Math. Biol.*, 79(4):693–737, 2017.

[XZ99] J. Xu and L. T. Zikatanov. A monotone finite element scheme for convection-diffusion equations. *Mathematics of Computation*, 68(228):1429–1446, 10 1999.

[Zal79] S. T. Zalesak. Fully multidimensional flux-corrected transport algorithms for fluids. *J. Comput. Phys.*, 31(3):335–362, 1979.

[Zha95] S. Zhang. Successive subdivisions of tetrahedra and multigrid methods on tetrahedral meshes. *Houston J. Math.*, 21(3):541–556, 1995.

Zusammenfassung

In dieser Arbeit wurden Finite-Elemente-Verfahren mit algebraischer Flusskorrektur (AFC) für stationäre Konvektions-Diffusions-Reaktions Gleichungen untersucht. Die beiden Hauptaspekte, die studiert wurden, sind iterative Löser für die auftretenden nichtlinearen Gleichungen und adaptive Gitterverfeinerung basierend auf a posteriori Fehlerschätzern. Die wichtigsten Ergebnisse der Arbeit sind im Folgenden zusammengefasst.

Zunächst wurden Studien zu den Lösern vorgestellt. Es wurden mehrere iterative Löser untersucht, darunter Fixpunktansätze und Methoden vom Newton-Typ. Die Newton Methoden reduzierten die Anzahl der Iterationen für bestimmte Beispiele, aber sie waren ineffizient bezüglich der Rechenzeit. Der einfachste Fixpunktansatz, nämlich *fixed point rhs*, war auf Grund seiner Matrixeigenschaften am effizientesten. Algorithmische Komponenten, wie die Anderson-Beschleunigung, reduzierten die Anzahl der Iterationen in einigen Beispielen, aber sie lieferte keine Ergebnisse für den BJK-Limiter. In drei Dimensionen wurde ein iterativer Löser für feinere Gitter benötigt, aber auch hier war *fixed point rhs* die effizienteste Herangehensweise. Unabhängig von der Dimension war es einfacher, die Probleme mit dem Kuzmin-Limiter als mit dem BJK-Limiter zu lösen.

Der zweite Hauptaspekt sind Studien zur a posteriori Fehlerschätzung. Es wurden zwei Ansätze zur Bestimmung einer oberen Schranke in der Energienorm untersucht, ein auf Residuen basierender Ansatz (*AFC-Energie* Technik) und ein anderer mit der SUPG-Lösung (*AFC-SUPG-Energie* Technik). Beide Techniken liefern keine robusten Schätzungen bezüglich ε, aber es zeigte sich, dass der *AFC-SUPG Energie* Ansatz einen besseren Effektivitätsindex besaß. Für den BJK-Limiter war die Effektivität besser als für den Kuzmin-Limiter mit dem *AFC-Energie* Ansatz, während beim *AFC-SUPG Energie* Ansatz die Wahl des Limiters keine Rolle spielte. Im Zuge der adaptiven Gitterverfeinerung kann das Problem lokal diffusions-dominant werden. In diesem Falle muss man den BJK-Limiter verwenden, da man beim Kuzmin-Limiter eine reduzierte Konvergenzordnung beobachten kann. Im Hinblick auf die adaptive Gitterverfeinerung wurden Grenzschichten unterschiedlichen Typs besser mit dem *AFC-Energie* Ansatz verfeinert als mit dem *AFC-SUPG Energie* Ansatz.

Schließlich wurden die Ergebnisse für die a posteriori Fehlerschätzung auf Gitter mit hängenden Knoten angewandt. Zunächst wurden Ergebnisse bezüglich hängender Knoten von Lagrange-Elementen niedriger Ordnung auf Elemente höherer Ordnung erweitert. Es zeigte sich in numerischen Studien, dass der Kuzmin-Limiter auf Gittern mit hängenden Knoten dem DMP nicht genügt, während der BJK-Limiter Ergebnisse lieferte, die dem DMP entsprachen. Die Grenzschichten wurden auf konform abgeschlossenen Gittern wesentlich besser approximiert als auf Gittern mit hängenden Knoten. Insgesamt sollte man Gitter mit hängenden Knoten nicht für AFC Verfahren verwenden.

Curriculum Vitae

Name:	Abhinav Jha
Date of Birth:	2nd July, 1994
Place of Birth:	New Delhi, India
2017 – 2020:	Doctoral Student, Berlin Mathematical School, Freie Universität, Berlin
2015 – 2017:	Master of Science, Mathematics, Indian Institute of Technology, Roorkee, India
2012 – 2015:	Bachelor of Science, Mathematics, St. Stephen's College, University of Delhi, India

Selbstständigkeitserklärung

Ich versichere hiermit, alle Hilfsmittel und Hilfen angegeben zu haben, und die Arbeit selbstständig und ausschließlich auf Grundlage der angegebenen Hilfsmittel und Hilfen angefertigt zu haben. Des Weiteren versichere ich, die Arbeit oder Teile der Arbeit nicht schon einmal in einem früheren Promotionsverfahren eingereicht zu haben.

Berlin, October 30, 2020

Abhinav Jha