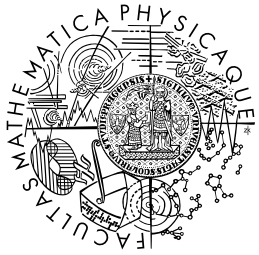


FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

MASTER THESIS

Ondřej Brichta

Properties of algebraic stabilizations for convection-diffusion equations

Department of Numerical Mathematics

Supervisor of the master thesis: prof. Mgr. Petr Knobloch, Dr., DSc.

Study programme: Computational Mathematics

Prague 2025

I declare that I carried out this master thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I would like to express my deepest gratitude to my supervisor, prof. Mgr. Petr Knobloch, Dr., DSc., for his patience, kindness, and the many valuable comments and ideas, without which this thesis would not have reached its current level of quality. I would also like to thank my wife Marcela and the rest of my family for their support and understanding, without which my studies would not have been possible.

Title: Properties of algebraic stabilizations for convection-diffusion equations

Author: Ondřej Brichta

Department: Department of Numerical Mathematics

Supervisor: prof. Mgr. Petr Knobloch, Dr., DSc., Department of Numerical Mathematics

Abstract: This thesis deals with the analysis of algebraic flux correction (AFC) schemes. This class of methods was developed for the discretization of elliptic boundary value problems (problems with convection-diffusion-reaction equations). The thesis provides a derivation of these methods for a model problem. Furthermore, results related to a priori error estimates are presented, and an error estimate in the L^2-norm is derived based on these results. The core of the thesis lies in the extension of the analysis of the BBK limiter to the case with weights, and in the construction of a novel limiter based on integral smoothness indicators. We focus on the so-called linearity preservation property of this new limiter, which may lead to improved convergence behavior. Finally, the results of numerical experiments with this limiter are presented.

Keywords: Galerkin method, discrete maximum principle, algebraic stabilization, limiter, a priori error estimates, linearity-preservation, smoothness indicator

Název práce: Vlastnosti algebraických stabilizací pro rovnice konvekce-difúze

Autor: Ondřej Brichta

Katedra: Katedra numerické matematiky

Vedoucí diplomové práce: prof. Mgr. Petr Knobloch, Dr., DSc., Katedra numerické matematiky

Abstrakt: Tato práce se zabývá analýzou numerických metod založených na algebraické stabilizaci (tzv. AFC schémata). Jedná se o třídu metod určených pro diskretizaci okrajových úloh s eliptickými parciálními diferenciálními rovnicemi (stacionárními rovnicemi konvekce-difuze-reakce). Práce obsahuje odvození těchto metod pro modelový problém. Dále jsou shrnuty výsledky týkající se a priorních odhadů chyb. Na základě těchto výsledků je představen odhad chyby v L^2-normě. Těžištěm práce je rozšíření analýzy tzv. BBK limiteru na případ s váhami a konstrukce nového limiteru založeného na integrálních indikátorech. U tohoto nového limiteru je studováno tzv. zachování linearity, což je vlastnost, která může vést k lepší konvergenci. Nakonec jsou uvedeny výsledky numerických experimentů s tímto novým limiterem.

Klíčová slova: Galerkinova metoda, diskrétní princip maxima, algebraická stabilizace, limiter, a priorní odhady chyb, zachování linearity, indikátor hladkosti

Contents

Introduction **6**

1 Algebraic Flux Correction Schemes **8**
1.1 Model Problem and Weak Formulation 8
1.2 Finite Element Discretization 9
1.2.1 Triangulation of the Domain 9
1.2.2 Special Classes of Simplicial Triangulations 11
1.2.3 Finite Element Spaces and Approximation of Sobolev Spaces 12
1.2.4 Discrete Variational Formulation 15
1.3 Discrete Maximum Principle for the Linear Problem 18
1.3.1 Validity of DMPs for Discrete Solution 21
1.4 Abstract Algebraic Flux Correction Setting 24
1.5 Discrete Maximum Principles for Nonlinear AFC Scheme 27

2 Error Analysis of AFC Schemes **29**
2.1 General Error Estimate 29
2.2 Improved Error Estimate 33
2.3 L^2-error Estimates 34

3 Limiters Based on Algebraic Fluxes **40**
3.1 Kuzmin Limiter 40
3.2 BJK Limiter 41

4 Limiters Based on Local Smoothness Indicators **43**
4.1 Abstract Formulation of the BBK Method 43
4.2 BBK Method Based on Integral Smoothness Indicators 49
4.2.1 Motivation 49
4.2.2 General Form of Integral Based BBK Limiter 50
4.2.3 Construction of the Test Function $\widetilde{\varphi}_i$ 51
4.2.4 Alternative Construction of the Test Function $\widetilde{\varphi}_i$ in Two Dimensions 54
4.2.5 General Remarks on the Construction of $\widetilde{\varphi}_i$ in Three Dimensions 59

5 Numerical Results **61**
5.1 Implementation 61
5.2 Problem with Smooth Solution 61

Conclusion **64**

Bibliography **66**

List of Figures **68**

Introduction

Boundary value problems involving steady-state convection–diffusion–reaction equations (or, more generally, elliptic partial differential equations) frequently arise in the natural sciences and engineering practice. For example, they can be used to model the conservation of energy in a fluid or the concentration of chemical species (see [1], Section 2.1).

A necessary requirement for such models is that they preserve the essential physical properties of the described system. For instance, the resulting solutions should attain only physically admissible values (e.g., the concentration of chemical species must be nonnegative). Mathematically, this requirement is formulated in terms of the so-called maximum principles. This principle often holds for the considered elliptic problems (see, e.g., [2], Section 6.4).

In practice, it is often difficult to construct weak solutions of these equations, and the only available approach is to approximate them using numerical methods. These approximate solutions are also expected to preserve the relevant physical properties.

In most applications, convection dominates diffusion. This dominance causes the exact solution to contain the so-called layers, narrow subregions where the gradients of the solution are very large. The presence of such layers makes it very difficult to obtain sufficiently accurate approximate solutions using discrete methods based on the classical Galerkin discretization, as the layers may not be well resolved by the chosen triangulation. Classical methods then typically produce approximate solutions polluted by spurious oscillations, potentially leading to values that violate physical constraints (see [1], Section 8.2).

A well-known remedy is to appropriately modify the discrete variational formulation (for details, see, e.g., [1], Chapter 8). One such method is the streamline upwind Petrov–Galerkin (SUPG) method, which is based on adding a linear stabilization term. An extension of the SUPG method using nonlinear terms leads to the class of methods known as spurious oscillations at layers diminishing (SOLD) methods (see, e.g., [3]).

However, many of these stabilization techniques still do not fully eliminate spurious oscillations. A possible solution to this problem is to apply methods that satisfy discrete maximum principles (DMPs). The validity of the DMP helps eliminate many types of oscillatory behavior in the numerical solution (see, e.g., [4]).

An example of a method based on a modified discrete variational formulation that satisfies the DMP is the Mizukami–Hughes method (see [1], Section 9.1).

A completely different class of methods satisfying the DMPs (generally, under some additional assumptions) was proposed for the steady-state case in [5]. These methods are called algebraic flux correction (AFC) schemes (or algebraic stabilization), and have been intensively developed in recent years (see, e.g., [6], [7], [8], [9], [4]).

The core idea is to modify the resulting linear system directly instead of changing the discrete variational formulation. This approach involves the so-called limiters that depend nonlinearly on the discrete solution. This thesis specifically focuses on this class of methods.

The thesis is divided into five chapters. The first chapter provides a brief introduction to the weak formulation of the problem and discusses the existence and uniqueness of the weak solution. It also presents a discrete variational form derived via the Galerkin method and includes an analysis of the validity of DMPs for this method. In particular, analysis is performed for the Galerkin method with numerical quadrature, a topic that is often overlooked in the literature. Based on this analysis, a general form of AFC schemes is derived.

The second chapter is dedicated to a priori error estimates. The first two sections present an overview of existing results in this area, extended slightly to the case involving a general diffusion tensor. The third section provides an error analysis in the L^2-norm for the diffusion-dominated regime, which represents a novel contribution in the context of AFC schemes.

The third chapter gives an overview of two fundamental limiters based on algebraic fluxes, specifically the Kuzmin limiter and the BJK limiter. Their basic properties are also briefly discussed.

The fourth chapter contains a detailed analysis of a limiter based on local smoothness indicators (the so-called BBK method). Here, the standard analysis of these limiters is extended to the general case incorporating weights. Finally, a new method is derived based on an integral variant of the smoothness indicator. It is also proven that the resulting limiter is linearity preserving. This property may lead to an improved convergence behavior of the method.

The fifth chapter is devoted to numerical experiments with the limiter introduced in Chapter 4.

1 Algebraic Flux Correction Schemes

In this chapter, we introduce our model problem, a basic setting of the finite element method (FEM) and the class of algebraic flux correction (AFC) schemes.

1.1 Model Problem and Weak Formulation

Let us investigate the stationary convection-diffusion-reaction equation. This elliptic boundary value problem is given by (see [2], Chapter 6 or [10], Chapter 8):

$$\begin{aligned} -\mathrm{div}(\mathbb{E}\nabla u) + b \cdot \nabla u + c\, u &= g \quad \text{in } \Omega, \\ u &= u_b \quad \text{on } \partial\Omega, \end{aligned} \tag{1.1}$$

where $\Omega \subset \mathbb{R}^d$ $(d = 2, 3)$ is a bounded domain with a Lipschitz boundary, $\mathbb{E} = (\varepsilon_{ij})_{i,j=1,\dots,d}$ is a so-called diffusion tensor, which is assumed to be symmetric and uniformly elliptic, i.e, there exists $\varepsilon > 0$ such that

$$\forall \xi \in \mathbb{R}^d \quad \xi^\top \mathbb{E} \xi \geq \varepsilon \left\|\xi\right\|_2^2, \quad \text{a.e. in } \Omega, \tag{1.2}$$

where $\left\|\cdot\right\|_2$ denotes the Euclidean norm. Furthermore, $b : \Omega \to \mathbb{R}^d$ denotes a convection field, $c : \Omega \to \mathbb{R}$ is a nonnegative reaction coefficient, and $g : \Omega \to \mathbb{R}$ describes sources and sinks of an unknown quantity $u : \Omega \to \mathbb{R}$. For simplicity, we assume that Ω is either a polygonal domain in two dimensions or a polyhedral domain in three dimensions.

We adopt the standard notation $W^{k,p}(\Omega)$ for Sobolev spaces, where $k \in \mathbb{N}_0$ and $p \in \langle 1, \infty \rangle$. The standard norm and seminorm on $W^{k,p}(\Omega)$ are denoted by $\left\|\cdot\right\|_{k,p,\Omega}$ and $\left|\cdot\right|_{k,p,\Omega}$, respectively. In particular, we define $H^k(\Omega) := W^{k,2}(\Omega)$. Furthermore, the space $H^{\frac{1}{2}}(\partial\Omega)$, known as the Sobolev-Slobodeckij space, denotes the range of the trace operator on $H^1(\Omega)$.

Throughout this thesis, we impose the following assumptions on the data:

1. $\mathbb{E} \in (L^\infty(\Omega))^{d \times d}$,
2. $b \in (W^{1,\infty}(\Omega))^d$ satisfies $\mathrm{div} b = 0$ in the weak sense in Ω,
3. $c \in L^\infty(\Omega)$, $c \geq 0$ almost everywhere in Ω,
4. $g \in L^2(\Omega)$,
5. $u_b \in H^{\frac{1}{2}}(\partial\Omega) \cap \mathcal{C}(\partial\Omega)$.

Finally, let us denote by $\widetilde{u}_b \in H^1(\Omega)$ an extension of u_b, i.e., $\widetilde{u}_b|_{\partial\Omega} = u_b$ in the sense of traces.

Now we have everything ready to introduce the concept of *weak formulation*.

Definition 1. *We say that* $u \in H^1(\Omega)$ *is a* weak solution *of problem (1.1) provided that* $u - \widetilde{u}_b \in H_0^1(\Omega)$ *and*

$$\forall v \in H_0^1(\Omega) \quad a(u,v) := (\mathbb{E}\nabla u, \nabla v) + (b \cdot \nabla u, v) + (cu, v) = (g, v), \tag{1.3}$$

where $(\cdot,\cdot)$ *denotes the scalar product in* $L^2(\Omega)$.

The next lemma states the basic properties of the bilinear form associated with the previous definition.

Lemma 1. *Let us consider the bilinear form* $a : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ *from Definition 1 and suppose that the assumptions stated above hold. Then following holds:*

1. $\exists C_e > 0 \quad \forall u \in H_0^1(\Omega) \quad a(u,u) \geq C_e \|u\|_{1,2,\Omega}^2$ (ellipticity),
2. $\exists C_c > 0 \quad \forall u, v \in H^1(\Omega) \quad |a(u,v)| \leq C_c \|u\|_{1,2,\Omega} \|v\|_{1,2,\Omega}$ (continuity).

Proof.

Ellipticity. Ellipticity of the bilinear form a restricted to the space $H_0^1(\Omega) \times H_0^1(\Omega)$ easily follows from the property (1.2), the Poincaré inequality and the Gauss integral theorem, e.g., see [2], Chapter 6 or [10], Chapter 8.

Continuity. Continuity of the bilinear form a is the consequence of the Hölder and the Cauchy-Schwarz inequalities, e.g., see [2], Chapter 6 or [10], Chapter 8.

□

The last statement in this section guarantees the existence and uniqueness of the weak solution.

Theorem 2. *Under the assumptions outlined above, there exists a unique weak solution of problem (1.1).*

Proof. The proof of the existence and uniqueness utilizes the Lax-Milgram lemma and Lemma 1, e.g., see [2], Chapter 6.

□

1.2 Finite Element Discretization

In this section, we briefly introduce the basic definitions, terminology, and results related to triangulations, finite element spaces, and the classical Galerkin finite element method, which will be frequently used throughout this thesis.

More detailed information on this topic can be found in [11], Section 1.3, and [1], Section 4.1.

1.2.1 Triangulation of the Domain

The first step in the finite element discretization of a partial differential equation is to construct a triangulation of the domain Ω. The following definition is taken from [11], Section 1.3 (Definition 1.49).

Definition 2. *Let $\mathcal{T}_h$ be a set consisting of a finite number of compact, connected, and convex subsets of $\overline{\Omega}$. If the set $\mathcal{T}_h$ satisfies the following conditions:*

1. $\forall T \in \mathcal{T}_h \quad \mathrm{Int}(T) \neq \emptyset$ *and T has a Lipschitz boundary,*
2. $\forall T, K \in \mathcal{T}_h, T \neq K \quad \mathrm{Int}(T) \cap \mathrm{Int}(K) = \emptyset$,
3. $\bigcup_{T \in \mathcal{T}_h} T = \overline{\Omega}$,

then this set is called a triangulation *of the domain Ω (this terminology will also be used for nonsimplicial cases).*

Furthermore, we define discretization parameter *h of the triangulation $\mathcal{T}_h$ by formula*

$$h := \max_{T \in \mathcal{T}_h} \mathrm{diam}(T).$$

In this thesis, we consider only triangulations composed of simplices (triangles and tetrahedra), parallelograms (in the two-dimensional case), or parallelepipeds (in the three-dimensional case).

Let $T_1, T_2 \in \mathcal{T}_h$ with $T_1 \neq T_2$ and $T_1 \cap T_2 \neq \emptyset$. We assume that $\partial T_1 \cap \partial T_2$ is either a common vertex, or a common edge in the two-dimensional case, and either a common vertex, a common edge, or a common face in the three-dimensional case. Triangulations with hanging vertices are not allowed.

Every considered triangulation $\mathcal{T}_h$ is generated from a so-called *reference element* $\widehat{T}$ and a set of invertible affine mappings. More precisely, for all $T \in \mathcal{T}_h$, there exist an invertible matrix $B_T \in \mathbb{R}^{d \times d}$ and a vector $b_T \in \mathbb{R}^d$ such that $F_T(\widehat{T}) = T$, where $F_T(\widehat{x}) := B_T \widehat{x} + b_T$ for $\widehat{x} \in \widehat{T}$.

We will consider the following particular reference elements:

1. In the two-dimensional case, $\widehat{T}$ is *unit triangle* with vertices $(0,0)$, $(1,0)$, $(0,1)$ or *unit square* $\langle 0,1 \rangle^2$.
2. In the three-dimensional case, $\widehat{T}$ is *unit simplex* with vertices $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ or *unit cube* $\langle 0,1 \rangle^3$.

The vertices of the triangulation $\mathcal{T}_h$ are denoted by $\{x_i\}_{i=1}^N$. Without loss of generality, they are numbered in such a way that $\{x_i\}_{i=1}^M \subset \Omega$ and $\{x_i\}_{i=M+1}^N \subset \partial\Omega$.

The following definition is taken from [1], Section 4.3 (Definition 4.11).

Definition 3. *A set of triangulations $\{\mathcal{T}_h\}$ is called a* family of triangulations*, if for every $h_0 > 0$ there exists a triangulation with $h < h_0$.*

As we shall see later, it is convenient to define a shape-regular family of triangulations. The following definition is taken from [1], Section 4.3 (Definition 4.12).

Definition 4. *A family of triangulations $\{\mathcal{T}_h\}$ is said to be* shape-regular *if there exists $\kappa > 0$ such that*

$$\forall T \in \bigcup_h \mathcal{T}_h \qquad \frac{h_T}{\rho_T} \leq \kappa, \tag{1.4}$$

where $h_T := \mathrm{diam}(T)$ and $\rho_T := \sup\{\mathrm{diam}(B);\ B \subset T \text{ is a ball}\}$.

At the end of this subsection, we introduce two definitions used at the end of this thesis (Definition 4.14 and Definition 4.2 in [1])

Definition 5. *Given a finite set S of n mutually different points in $\mathbb{R}^d$. The* Voronoi box $\mathcal{V}_p$ of a point $p \in S$ *is the following subset of $\mathbb{R}^d$:*

$$\mathcal{V}_p := \{x \in \mathbb{R}^d; \|x - p\|_2 \leq \|x - q\|_2 \ \forall q \in S \setminus \{p\}\}$$

Definition 6. *A triangulation of a domain Ω is called* locally symmetric *if, for any vertex $x_i \in \Omega$, the following holds: For any vertex x_j connected to x_i by an edge, there is a different vertex x_k connected to x_i by an edge such that*

$$\frac{1}{2}(x_j + x_k) = x_i$$

1.2.2 Special Classes of Simplicial Triangulations

In this subsection, we present some special classes of simplicial triangulations. As we shall see later, the geometrical structure may influence the properties of the resulting method.

Let $\mathcal{T}_h$ be a simplicial triangulation, let $T \in \mathcal{T}_h$ be an arbitrary element, and let $x_i \in T$ be a vertex of T. Then, F_i^T denotes the face of T opposite to the vertex x_i. For a different vertex $x_j \in T$, the edge with endpoints x_i and x_j is denoted by E_{ij}, and the unit outward normal vector to the face F_i^T with respect to T is denoted by n_i^T. Furthermore, a tangential unit vector to edge E_{ij} is given by formula

$$t_{ik} := \frac{x_k - x_i}{\|x_k - x_j\|_2}.$$

The Lebesgue measure (length or area) of F_i^T is denoted by $\left|F_i^T\right|$. Finally, the dihedral angle between the faces F_i^T and F_j^T is denoted by θ_{ij}^T.

From simple geometric reasoning, it follows that θ_{ij}^T is supplementary to the angle between the vectors n_i^T and n_j^T. Thus,

$$\cos(\theta_{ij}^T) = -n_i^T \cdot n_j^T.$$

For a better understanding of this notation in two dimensions, see Figure 1.1.

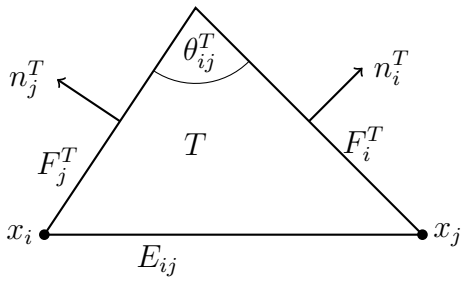


Figure 1.1 Illustration of the notation described in Subsection 1.2.2 for two-dimensional case.

The following type of simplicial triangulations was originally studied in [12].

Definition 7. *The triangulation $\mathcal{T}_h$ satisfies the* XZ-condition *(Xu-Zikatanov condition) if, for every edge E_{ij} with endpoints x_i and x_j, where at least one of them is not contained on $\partial\Omega$, the following condition holds:*

$$\sum_{T \subset \omega_{ij}} l_{ij}^T \cot(\theta_{ij}^T) \geq 0, \tag{1.5}$$

where

$$l_{ij}^T := \begin{cases} \left|F_i^T \cap F_j^T\right| & \text{if } d = 3, \\ 1 & \text{if } d = 2, \end{cases}$$

and

$$\omega_{ij} := \bigcup \{T \in \mathcal{T}_h;\ E_{ij} \subset T\}.$$

Let us introduce other important class of simplicial triangulations. This definition is taken from [1] (Definitions 4.4 and 4.6).

Definition 8. *A simplicial triangulation $\mathcal{T}_h$ is called* acute *if all dihedral angles of all elements in $\mathcal{T}_h$ are acute.*

Clearly, every acute triangulation satisfies the XZ-condition.

1.2.3 Finite Element Spaces and Approximation of Sobolev Spaces

This subsection provides a brief introduction to the linear and bilinear finite element spaces that will be consistently considered throughout this thesis, as well as some basic results from the theory of interpolation of Sobolev functions.

Our approach will be based on [13], Chapter 2 and Section 3.2, and [11], Sections 1.2, 1.4 and 1.5, where more detailed descriptions of finite element spaces can be found.

Let us begin with simplicial triangulations $\mathcal{T}_h$. By the notation $\mathcal{P}_k$ we will denote the space of polynomials of degree less than or equal to $k \in \mathbb{N}_0$ in the variables $x_1, x_2, \ldots, x_d$:

$$\mathcal{P}_k := \left\{ p : \mathbb{R}^d \to \mathbb{R};\ p(x) = \sum_{|\alpha| \le k} \gamma_\alpha x^\alpha,\ \gamma_\alpha \in \mathbb{R} \right\},$$

where the multi-index notation is used. Furthermore, we define the restriction of this space on arbitrary set $M \subset \mathbb{R}^d$ as

$$\mathcal{P}_k(M) := \{p|_M;\ p \in \mathcal{P}_k\}.$$

We then define the Lagrange finite element $\mathcal{P}_1$ (see, e.g. [11], Section 1.2.3).

Definition 9. *A* Lagrange $\mathcal{P}_1$ finite element *is defined as triple $\{T, \mathcal{P}_1(T), \Sigma\}$, where T is a simplex with vertices $a_1, \ldots, a_{d+1} \in \mathbb{R}^d$, and $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{d+1}\}$ is the set of linear functionals acting on $\mathcal{P}_1(T)$ such that $\sigma_i(p) = p(a_i)$ for $i = 1, 2, \ldots, d+1$, and for all $p \in \mathcal{P}_1(T)$. The functionals $\sigma_1, \sigma_2, \ldots, \sigma_{d+1}$ are called* local degrees of freedom *of the Lagrange $\mathcal{P}_1$ finite element.*

For triangulations consisting of simplices, we define the Lagrange $\mathcal{P}_1$ finite element space as follows:

$$W_h := \{v_h \in \mathcal{C}(\overline{\Omega});\ \forall T \in \mathcal{T}_h\ \ v_h|_T \in \mathcal{P}_1(T)\}.$$

It can be shown (see, e.g., [13], Theorem 2.1.1) that $W_h \subset H^1(\Omega)$. Such finite element spaces are called *conforming finite element spaces.*

In the same spirit, we can define the $\mathcal{Q}_1$ finite elements on a triangulation $\mathcal{T}_h$ consisting of parallelograms (in the two-dimensional case) or parallelepipeds (in the three-dimensional case). It is advantageous to first construct these finite elements on the reference element $\widehat{T} = \langle 0, 1\rangle^d$.

Similarly as in the previous case, $\mathcal{Q}_k$ will denotes the space of polynomials of degree less than or equal to $k \in \mathbb{N}_0$ with respect to each one of the d variables $x_1, x_2, \ldots, x_d$:

$$\mathcal{Q}_k := \left\{ p : \mathbb{R}^d \to \mathbb{R};\, p(x) = \sum_{0 \le \alpha_1, \ldots, \alpha_d \le k} \gamma_\alpha x^\alpha,\, \gamma_\alpha \in \mathbb{R} \right\}.$$

Let $M \subset \mathbb{R}^d$ be arbitrary, then

$$\mathcal{Q}_k(M) := \{ p|_M;\, p \in \mathcal{Q}_k \}.$$

Finally, we introduce the following auxiliary notation:

$$\mathcal{Q}(T) := \left\{ \widehat{p} \circ F_T^{-1};\, \widehat{p} \in \mathcal{Q}_1(\widehat{T}) \right\} \tag{1.6}$$

for all $T \in \mathcal{T}_h$

Definition 10. *A* Lagrange $\mathcal{Q}_1$ finite element *is defined as triple* $\{T, \mathcal{Q}(T), \Sigma\}$, *where* T *is a parallelogram or a parallelepiped with vertices* $a_1, \ldots, a_{2^d} \in \mathbb{R}^d$, *and* $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{2^d}\}$ *is the set of linear functionals acting on* $\mathcal{Q}(T)$ *such that* $\sigma_i(p) = p(a_i)$ *for all* $i = 1, 2, \ldots, 2^d$ *and for all* $p \in \mathcal{Q}(T)$. *The functionals* $\sigma_1, \sigma_2, \ldots, \sigma_{2^d}$ *are called* local degrees of freedom *of the Lagrange* $\mathcal{Q}_1$ *finite element.*

The definition of a Lagrange $\mathcal{Q}_1$ finite element space is following:

$$W_h := \{ v_h \in \mathcal{C}(\overline{\Omega});\, \forall T \in \mathcal{T}_h \;\; v_h|_T \in \mathcal{Q}(T) \}.$$

It can be shown that $W_h \subset H^1(\Omega)$, similarly to the previous case.

Moreover, for both types of triangulation, we denote $V_h := W_h \cap H_0^1(\Omega)$. It is easy to see that $\dim W_h$ and $\dim V_h$ are equal to N and M, respectively. We consider the standard basis $\{\varphi_i\}_{i=1}^N$ of the space W_h, satisfying

$$\varphi_i(x_j) = \delta_{ij}, \quad i, j = 1, 2, \ldots, N, \tag{1.7}$$

where δ_{ij} denotes the Kronecker delta. Furthermore, for every $v_h \in W_h$, there exists a uniquely determined vector of coordinates with respect to this basis

$$V := (v_1, \ldots, v_N)^\top \in \mathbb{R}^N \quad \longleftrightarrow \quad v_h = \sum_{i=1}^N v_i \varphi_i \in W_h. \tag{1.8}$$

Therefore, taking into account (1.7), it is easy to see that every function $v_h \in W_h$ is uniquely determined by its values at the points $x_1, \ldots, x_N$. This values are called *degrees of freedom*. We will interchange the elements of the space W_h with their coordinate representation with respect to the basis $\{\varphi_i\}_{i=1}^N$ whenever no ambiguity arises.

Note that the supports of such basis functions are local in the following sense:

$$\forall i \in \{1, \ldots, N\}, \quad \operatorname{supp} \varphi_i = \omega_i, \tag{1.9}$$

where

$$\omega_i := \bigcup \{T \in \mathcal{T}_h;\, x_i \in T\}. \tag{1.10}$$

Finally, we introduce the Lagrange interpolation operator. We begin by defining the local interpolation operator and then extend it to the entire domain and the space W_h (see [11], Subsections 1.4.1 and 1.4.5).

Definition 11. *Let $\mathcal{T}_h$ be a triangulation consisting of simplices and $T \in \mathcal{T}_h$ be an arbitrary element. Let $\{T, \mathcal{P}_1(T), \Sigma\}$ be the Lagrange $\mathcal{P}_1$ finite element, where $\Sigma = \{\sigma_1, \sigma_2, \ldots, \sigma_{d+1}\}$ is the set of local degrees of freedom corresponding to the vertices $a_1, \ldots, a_{d+1}$ of the element T. A mapping $I_T : \mathcal{C}(T) \to \mathcal{P}_1(T)$ defined by the formula*

$$I_T v := \sum_{l=1}^{d+1} \sigma_l(v)\, \varphi_l|_T = \sum_{l=1}^{d+1} v(a_l)\, \varphi_l|_T$$

for any $v \in \mathcal{C}(T)$, where $\{\varphi_l|_T\}_{l=1}^{d+1} \subset \mathcal{P}_1(T)$ is the set of basis functions satisfying $(\varphi_l|_T)(a_j) = \delta_{lj}$ for $l, j = 1, \ldots, d+1$, is called the local Lagrange interpolation operator*.*

Let W_h be a Lagrange $\mathcal{P}_1$ finite element space. Then, the global Lagrange interpolation operator *$I_h : \mathcal{C}(\overline{\Omega}) \to W_h$ is defined by the formula*

$$I_h v := \sum_{T \in \mathcal{T}_h} I_T v|_T = \sum_{l=1}^{N} v(x_l)\, \varphi_l$$

for any $v \in \mathcal{C}(\overline{\Omega})$.

In the case of Lagrange $\mathcal{Q}_1$ finite elements, local and global interpolation operators can be defined analogously.

Now, we present an important result from the theory of Sobolev function interpolation. A more general version of this theorem can be found in [13], Section 3.2.

Theorem 3. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces defined on a shape-regular family of triangulations $\{\mathcal{T}_h\}$ consisting of simplices, parallelograms, or parallelepipeds. Then, there exists a constant $C_{\mathrm{int}} > 0$ independent of h such that*

$$h\, |v - I_h v|_{1,2,\Omega} + \|v - I_h v\|_{L^2(\Omega)} \le C_{\mathrm{int}} h^2\, |v|_{2,2,\Omega} \tag{1.11}$$

for all $v \in H^2(\Omega)$.

Proof. This assertion is a special case of Theorem 3.2.1 in [13].

□

Note that $H^2(\Omega) \hookrightarrow \mathcal{C}(\overline{\Omega})$. Thus, the Lagrange interpolation operator can be applied to functions belonging to $H^2(\Omega)$.

Finally, we formulate three auxiliary statements that will be frequently used later.

Lemma 4. *Let W_h be a Lagrange finite element space defined on a triangulation $\mathcal{T}_h$, and let $\{\varphi_i\}_{i=1}^N \subset W_h$ be a basis of W_h satisfying condition (1.7). Then*

$$\sum_{i=1}^{N} \varphi_i = 1 \quad \text{in } \overline{\Omega}. \tag{1.12}$$

Proof. See the proof of Lemma 5.7 in [1]. □

Lemma 5. *Let W_h be a Lagrange $\mathcal{P}_1$ finite element space defined on a simplicial triangulation $\mathcal{T}_h$, and let $\{\varphi_i\}_{i=1}^N \subset W_h$ be a basis of W_h satisfying condition (1.7). Then, for any $T \in \mathcal{T}_h$ and any vertex $x_i \in T$, we have*

$$\nabla\varphi_i|_T = -\frac{\left|F_i^T\right|}{d\,|T|} n_i^T, \tag{1.13}$$

where $|T|$ denotes the d-dimensional Lebesgue measure of T, and the notations n_i^T and $\left|F_i^T\right|$ are introduced at the beginning of Section 1.2.2.

Proof. See the proof of Lemma 6.5 in [1]. □

Theorem 6. *Let W_h be a Lagrange $\mathcal{P}_1$ finite element space defined on a simplicial triangulation $\mathcal{T}_h$, and let $\{\varphi_i\}_{i=1}^N \subset W_h$ be a basis of W_h satisfying condition (1.7). Then, the triangulation $\mathcal{T}_h$ satisfies the XZ-condition if and only if, for all $i \in \{1,\dots,M\}$ and $j \in \{1,\dots,N\}$ with $i \neq j$, following holds:*

$$(\nabla\varphi_i, \nabla\varphi_j) \leq 0. \tag{1.14}$$

Moreover, if $\mathcal{T}_h$ is an acute triangulation, then the following holds:

$$\forall i \in \{1,\dots,M\}, \quad \forall j \in S_i, \quad \int_{\omega_i} \nabla\varphi_j \cdot \nabla\varphi_i \,\mathrm{d}x < 0, \tag{1.15}$$

where ω_i is given by (1.10) and

$$S_i := \{j \in \{1,\dots,N\} \setminus \{i\};\ x_i \text{ and } x_j \text{ are endpoints of the same edge}\}. \tag{1.16}$$

Proof. See the proof of Theorem 6.8 in [1]. □

1.2.4 Discrete Variational Formulation

In many cases, it is difficult to construct the weak solution of problem (1.1), even though its existence and uniqueness are guaranteed (see Theorem 2). Therefore, a numerical method is required. In this section, we provide a brief description of a standard discretization based on the Galerkin method. For details, see [11], Section 2.2.

The core idea of the Galerkin method is to approximate infinite-dimensional Sobolev spaces $H^1(\Omega)$ and $H_0^1(\Omega)$ in the weak formulation by finite-dimensional subspaces W_h and V_h, respectively. In this chapter, we consider both $\mathcal{P}_1$ and $\mathcal{Q}_1$ finite elements.

For simplicity of presentation, we assume that $\widetilde{u}_b \in \mathcal{C}(\overline{\Omega})$. This allows us to apply the Lagrange interpolation operator to this function.

A discrete approximation of the weak solution u can be obtained as the solution of a problem in a discrete variational form.

Definition 12. *The* discrete variational problem *reads as follows: Find* $u_h \in W_h$ *such that*

1. $u_h - u_{bh} \in V_h$,
2. $\forall v_h \in V_h \quad a_h(u_h, v_h) = \langle g_h, v_h \rangle_{W_h^*, W_h}$,

where $u_{bh} := I_h \widetilde{u}_b$, *the bilinear form* $a_h : W_h \times W_h \to \mathbb{R}$ *denotes a discrete approximation of the bilinear form on the left-hand side of (1.3),* $g_h \in W_h^*$ *denotes a discrete approximation of the right-hand side of (1.3), and* $\langle \cdot, \cdot \rangle_{W_h^*, W_h}$ *is a duality pairing. The solution* u_h *is called a* discrete solution.

Throughout this thesis we will suppose that a_h is V_h-elliptic:

$$\exists C_{eh} > 0 \quad \forall v_h \in V_h \quad a_h(v_h, v_h) \geq C_{eh} \left\| v_h \right\|_{1,2,\Omega}^2 . \tag{1.17}$$

Since $a_h(u_h, \cdot)$ and $\langle g_h, \cdot \rangle_{W_h^*, W_h}$ are linear functionals and $\{\varphi_i\}_{i=1}^M$ is a basis of V_h, we can express the second condition in Definition 12 in an equivalent form

$$a_h(u_h, \varphi_i) = \langle g_h, \varphi_i \rangle_{W_h^*, W_h}, \quad i = 1, \ldots, M.$$

According to the representation (1.8) of functions belonging to W_h, we have

$$u_h = \sum_{j=1}^{N} u_j \varphi_j.$$

Under the linearity of $a_h(\cdot, \varphi_i)$, $i = 1, \ldots, M$, we observe that u_h is the discrete solution from the Definition 12 if and only if $U = (u_1, \ldots, u_N)^\top \in \mathbb{R}^N$ is the solution of the linear system

$$\begin{aligned} \sum_{j=1}^{N} a_h(\varphi_j, \varphi_i) u_j &= \langle g_h, \varphi_i \rangle_{W_h^*, W_h}, \quad i = 1, \ldots, M, \\ u_i &= u_b(x_i), \quad i = M+1, \ldots N. \end{aligned} \tag{1.18}$$

By introducing the following notation:

$$a_{ij} := a_h(\varphi_j, \varphi_i), \quad i = 1, \ldots, M, j = 1, \ldots, N \tag{1.19}$$

and

$$\begin{aligned} g_i &:= \langle g_h, \varphi_i \rangle_{W_h^*, W_h}, \quad i = 1, \ldots, M, \\ u_i^b &:= u_b(x_i), \quad i = M+1, \ldots, N, \end{aligned}$$

we can rewrite the system (1.18) in the following form:

$$\begin{aligned} \sum_{j=1}^{N} a_{ij} u_j &= g_i, \quad i = 1, \ldots, M, \\ u_i &= u_i^b, \quad i = M+1, \ldots N. \end{aligned} \tag{1.20}$$

Finally, the system (1.20) can be writen in more compact matrix form

$$\begin{pmatrix} \mathbb{A}^i & \mathbb{A}^b \\ \mathbb{O} & \mathbb{I} \end{pmatrix} U = \begin{pmatrix} G \\ U^b \end{pmatrix}, \tag{1.21}$$

where $\mathbb{A}^i := (a_{ij})_{i,j=1,\dots,M}$, $\mathbb{A}^b := (a_{ij})_{j=M+1,\dots,N}^{i=1,\dots,M}$, $U^b := (u^b_{M+1}, \dots, u^b_N)^\top$, $G := (g_1, \dots g_M)^\top$, $\mathbb{O} \in \mathbb{R}^{(N-M)\times M}$ is the zero matrix and $\mathbb{I} \in \mathbb{R}^{(N-M)\times(N-M)}$ is the identity matrix.

The following lemma assures us that the problem from Definition 12 has a unique solution.

Theorem 7. *Under the condition (1.17), there exists a unique discrete solution $u_h \in W_h$ satisfying the problem stated in Definition 12.*

Proof. We follow the proof of Lemma 5.13 in [1]. According to the discussion above, it is sufficient to show the invertibility of the system matrix of (1.21). Since

$$\det \begin{pmatrix} \mathbb{A}^i & \mathbb{A}^b \\ \mathbb{O} & \mathbb{I} \end{pmatrix} = \det \mathbb{A}^i \det \mathbb{I} = \det \mathbb{A}^i,$$

it remains to show the invertibility of the matrix $\mathbb{A}^i$.

Let $V = (v_1, \dots, v_M)^\top \in \mathbb{R}^M \setminus \{0\}$ be arbitrary, then

$$V^\top \mathbb{A}^i V = \sum_{i,j=1}^{M} a_{ij} v_i v_j = \sum_{i,j=1}^{M} a_h(\varphi_j, \varphi_i) v_i v_j = a_h(v_h, v_h),$$

where we used (1.19), the bilinearity of the form a_h and the identification (1.8). Since the functions $\{\varphi_i\}_{i=1}^M$ are linearly independent, it follows that $v_h \neq 0$. By the condition (1.17) we obtain

$$V^\top \mathbb{A}^i V = a_h(v_h, v_h) \geq C_{eh} \|v_h\|^2_{1,2,\Omega} > 0.$$

Thus, the matrix $\mathbb{A}^i$ is positive definite and therefore invertible.

□

We conclude this section by describing a particular choice of a_h and g_h. For a practical implementation of the Galerkin method, a numerical quadrature is required. The following definition is taken from [11] (Definition 8.1 in Chapter 8), where more detailed information about numerical quadratures can be found.

Definition 13. *Let $\widehat{T}$ be a nonempty, compact, connected subset of $\mathbb{R}^d$ with a Lipschitz boundary, and let $l_q \in \mathbb{N}$. A* numerical quadrature *on $\widehat{T}$ with l_q points consist of:*

1. *a set of l_q real nonzero numbers $\{\widehat{\chi}_1, \dots, \widehat{\chi}_{l_q}\}$ called* quadrature weights*,*
2. *a set of l_q points $\{\widehat{\xi}_1, \dots, \widehat{\xi}_{l_q}\}$ in $\widehat{T}$ called* quadrature nodes*.*

The largest $k \in \mathbb{N}$ such that

$$\forall p \in \mathcal{P}_k(\widehat{T}) \quad \int_{\widehat{T}} p(\widehat{x}) \mathrm{d}\widehat{x} = \sum_{l=1}^{l_q} \widehat{\chi}_l \, p(\widehat{\xi}_l),$$

is called the algebraic order *of the quadrature formula and is denoted by k_q.*

Let $\widehat{T}$ be the reference element of a triangulation $\mathcal{T}_h$ of the domain Ω, and let $T \in \mathcal{T}_h$ be arbitrary. Then, for all $f \in \mathcal{C}(\overline{\Omega})$, we can write

$$\int_T f(x)\,\mathrm{d}x = \int_{\widehat{T}} f(F_T(\widehat{x}))\,|\det B_T|\;\mathrm{d}\widehat{x} \approx \sum_{l=1}^{l_q} |\det B_T|\;\widehat{\chi}_l f(F_T(\widehat{\xi}_l)),$$

where $F_T : \widehat{T} \to T$ denotes the affine mapping from the reference element onto T (see Section 1.2.1).

We will use the notation $\chi_l^T := |\mathrm{det}B_T|\,\widehat{\chi}_l$ and $\xi_l^T := F_T(\widehat{\xi}_l)$. Then

$$\int_T f(x)\mathrm{d}x \approx \sum_{l=1}^{l_q} \chi_l^T f(\xi_l^T) \tag{1.22}$$

and

$$\int_\Omega f(x)\mathrm{d}x = \sum_{T\in\mathcal{T}_h} \int_T f(x)\mathrm{d}x \approx \sum_{T\in\mathcal{T}_h} \sum_{l=1}^{l_q} \chi_l^T f(\xi_l^T). \tag{1.23}$$

Now, the discrete bilinear form a_h is obtained by approximating all integrals in the form a by the numerical quadrature. For this purpose, we have to assume that all elements of the diffusion tensor ε_{ij}, the convection field b and the reaction coefficient c are well-defined at the nodes of the numerical quadrature. First, we define a local form a_h^T, where $T \in \mathcal{T}_h$, by the following formula

$$a_h^T(u_h, v_h) := a_1^T(u_h, v_h) + a_2^T(u_h, v_h) + a_3^T(u_h, v_h), \tag{1.24}$$

where $u_h, v_h \in W_h$, and

$$\begin{aligned}
a_1^T(u_h, v_h) &:= \sum_{l=1}^{l_q} \chi_l^T (\mathbb{E}(\xi_l^T)\nabla u_h|_T) \cdot \nabla v_h|_T \approx \int_T (\mathbb{E}\nabla u_h)\cdot\nabla v_h \mathrm{d}x,\\
a_2^T(u_h, v_h) &:= \sum_{l=1}^{l_q} \chi_l^T v_h(\xi_l^T) b(\xi_l^T)\cdot\nabla u_h|_T \approx \int_T v_h b\cdot\nabla u_h \mathrm{d}x,\\
a_3^T(u_h, v_h) &:= \sum_{l=1}^{l_q} \chi_l^T c(\xi_l^T) u_h(\xi_l^T) v_h(\xi_l^T) \approx \int_T c u_h v_h \mathrm{d}x.
\end{aligned}$$

Then, the discrete form a_h is defined as follows:

$$a_h(u_h, v_h) := \sum_{T\in\mathcal{T}_h} a_h^T(u_h, v_h). \tag{1.25}$$

Finally, if g is well-defined at the nodes of the numerical quadrature, we define

$$\langle g_h, v_h\rangle_{W_h^*, W_h} := \sum_{T\in\mathcal{T}_h} \sum_{l=1}^{l_q} \chi_l^T g(\xi_l^T) v_h(\xi_l^T) \approx (g, v_h)\,. \tag{1.26}$$

1.3 Discrete Maximum Principle for the Linear Problem

It is well known that the (weak) solution u of equation (1.1) satisfies the so-called *maximum principle (MP)* (see [10], Chapter 8 or [2], Section 6.4). This

principle provides a mathematical formulation of the fact that the solution represents physically admissible values. For instance, if u denotes the concentration of a chemical species, then u should remain nonnegative in Ω.

From a practical point of view, it is natural to require that a similar principle holds for a discrete solution. This ensures that the discrete formulation remains consistent with the continuous one. Moreover, validity of this principle excludes many types of nonphysical solutions. In the following, we briefly present both the local and global forms of the *discrete maximum principles (DMPs)*. For further details, see [1], Section 3.1.

Let us start with the definition of a matrix of nonnegative type.

Definition 14. *A matrix* $(a_{ij})_{j=1,\dots,n}^{i=1,\dots,m} \in \mathbb{R}^{m\times n}$ *is said to be of* nonnegative type *if the following conditions hold:*

$$\forall i \in \{1,\dots,m\} \quad \forall j \in \{1,\dots,n\}, j \neq i \quad a_{ij} \leq 0, \tag{1.27}$$

$$\forall i \in \{1,\dots,m\} \quad \sum_{j=1}^{n} a_{ij} \geq 0. \tag{1.28}$$

Next, recall the definition of positive and negative part of real number a

$$a^+ := \max\{0, a\}, \quad a^- := \min\{0, a\}.$$

This definition enables us to formulate the following result about the local DMP.

Theorem 8. *Let* $a_{ii} > 0$ *for all* $i \in \{1,\dots,M\}$. *Then, the solution* $U = (u_1,\dots,u_N)^\top \in \mathbb{R}^N$ *of the linear system (1.21) satisfies*

$$\begin{aligned} g_i \leq 0 \quad &\Longrightarrow \quad u_i \leq \max_{j\neq i, a_{ij}\neq 0} u_j^+, \\ g_i \geq 0 \quad &\Longrightarrow \quad u_i \geq \min_{j\neq i, a_{ij}\neq 0} u_j^- \end{aligned}$$

for any $i \in \{1,\dots,M\}$ *if and only if the system matrix of the linear system (1.21) is of nonnegative type. The implications*

$$\begin{aligned} g_i \leq 0 \quad &\Longrightarrow \quad u_i \leq \max_{j\neq i, a_{ij}\neq 0} u_j, \\ g_i \geq 0 \quad &\Longrightarrow \quad u_i \geq \min_{j\neq i, a_{ij}\neq 0} u_j \end{aligned}$$

hold for any $i \in \{1,\dots,M\}$ *if and only if the system matrix of the linear system (1.21) satisfies condition (1.27) and following holds:*

$$\forall i \in \{1,\dots,M\} \quad \sum_{j=1}^{N} a_{ij} = 0. \tag{1.29}$$

Proof. See the proofs of Theorems 3.3 and 3.4 from [1].

□

Further, we present theorem about the global DMP.

Theorem 9. *Suppose that the system matrix of (1.21) is of nonnegative type and that the matrix $\mathbb{A}^i$ is invertible. Then, the (unique) solution $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ of this linear system satisfies*

$$\begin{aligned} \forall i \in \{1, \dots, M\} \quad g_i \le 0 \quad &\Longrightarrow \quad \max_{k=1,\dots,N} u_k \le \max_{l=M+1,\dots,N} u_l^+, \\ \forall i \in \{1, \dots, M\} \quad g_i \ge 0 \quad &\Longrightarrow \quad \min_{k=1,\dots,N} u_k \ge \min_{l=M+1,\dots,N} u_l^-. \end{aligned}$$

Moreover, if condition (1.29) holds, then

$$\begin{aligned} \forall i \in \{1, \dots, M\} \quad g_i \le 0 \quad &\Longrightarrow \quad \max_{k=1,\dots,N} u_k = \max_{l=M+1,\dots,N} u_l, \\ \forall i \in \{1, \dots, M\} \quad g_i \ge 0 \quad &\Longrightarrow \quad \min_{k=1,\dots,N} u_k = \min_{l=M+1,\dots,N} u_l. \end{aligned}$$

Proof. See the proof of Theorem 3.5 in [1].

□

Let W_h be the Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element space. The discrete maximum principles presented above take into account only the values of the discrete solution $u_h \in W_h$ at the vertices of a given triangulation. However, in practical applications, it is sometimes necessary to ensure that the values of the function u_h at points other than the vertices of the triangulation also satisfy certain constraints. We need versions of DMPs that also involve such values.

It can be shown (see [1], Theorem 5.16) that if the system matrix of the linear system (1.21) is of nonnegative type with positive diagonal entries, then the following version of the local DMP holds:

$$\begin{aligned} g \le 0 \text{ in } \omega_i \quad &\Longrightarrow \quad \max_{\omega_i} u_h \le \max_{\partial\omega_i} u_h^+, \\ g \ge 0 \text{ in } \omega_i \quad &\Longrightarrow \quad \min_{\omega_i} u_h \ge \min_{\partial\omega_i} u_h^-. \end{aligned}$$

In addition, if $c = 0$ in ω_i, then the following stronger local DMPs hold:

$$\begin{aligned} g \le 0 \text{ in } \omega_i \quad &\Longrightarrow \quad \max_{\omega_i} u_h = \max_{\partial\omega_i} u_h, \\ g \ge 0 \text{ in } \omega_i \quad &\Longrightarrow \quad \min_{\omega_i} u_h = \min_{\partial\omega_i} u_h. \end{aligned}$$

A counterpart of the global DMPs from Theorem 9 can be obtained in the same spirit (see [1], Section 5.3).

Finally, let us characterize the relations of local and global DMPs stated in Theorems 8 and 9.

Theorem 10. *The satisfaction of the global DMP does not necessarily imply the satisfaction of the corresponding local DMP.*

The satisfaction of the local DMP does not imply the satisfaction of the corresponding global DMP if and only if A^i is singular.

Proof. See the proof of Corollary 3.23 from [1].

□

1.3.1 Validity of DMPs for Discrete Solution

We conclude this section with a discussion on the validity of the conditions under which the DMPs hold for our system (1.21) based on the bilinear form a_h given by (1.25) and the functional g_h given by (1.26). In the literature, the influence of numerical integration on the validity of DMPs is not often discussed. Thus, in this section, we present a more detailed analysis.

Throughout the rest of this chapter, we consider the Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element space W_h with the standard basis $\{\varphi_i\}_{i=1}^N$, defined on a triangulation $\mathcal{T}_h$ consisting of simplices, parallelograms or parallelepipeds. According to Theorems 8 and 9, we need to determine when the system matrix of (1.21) is of nonnegative type with positive diagonal entries. We will follow the approach from [1], Sections 7.1, 8.1 and 6.4.

Let us begin by analyzing the row sums. For all $i \in \{1, \dots, M\}$, we have

$$\sum_{j=1}^{N} a_{ij} = \sum_{j=1}^{N} a_h(\varphi_j, \varphi_i) = a_h(1, \varphi_i) = \sum_{T \subset \omega_i} \sum_{l=1}^{l_q} \chi_l^T c(\xi_l^T) \varphi_i(\xi_l^T), \tag{1.30}$$

where we used identity (1.12) and the fact that a_1^T and a_2^T depend on $\nabla 1 = 0$. The following simple observation shows that the required inequality (1.28) may not always hold. Recall that the set ω_i is given by (1.10).

Lemma 11. *Let $i \in \{1, \dots, M\}$. Suppose that there exist $\widetilde{l} \in \{1, \dots, l_q\}$ and $\widetilde{T} \in \mathcal{T}_h$ such that $\xi_{\widetilde{l}}^{\widetilde{T}} \in \mathrm{Int}(\omega_i)$ and that*

$$\sum_{(T,l) \in \mathcal{M}} \chi_l^T < 0, \tag{1.31}$$

where $\mathcal{M} := \{(T, l);\, T \subset \omega_i, l = 1, \dots, l_q,\, \xi_l^T = \xi_{\widetilde{l}}^{\widetilde{T}}\}$. Then there exists a nonnegative function $c \in \mathcal{C}(\overline{\Omega})$ fulfilling

$$\sum_{T \subset \omega_i} \sum_{l=1}^{l_q} \chi_l^T c(\xi_l^T) \varphi_i(\xi_l^T) < 0. \tag{1.32}$$

Proof. Let us denote $\mathcal{N} := \{\xi_l^T; T \subset \omega_i, l = 1, \cdots l_q\}$. From the fact that

$$\mathrm{dist}\left(\mathcal{N} \setminus \{\xi_{\widetilde{l}}^{\widetilde{T}}\},\, \xi_{\widetilde{l}}^{\widetilde{T}}\right) > 0$$

and the openness of the set $\mathrm{Int}(\omega_i)$, it follows that there exists a neighborhood $B \subset \mathrm{Int}(\omega_i)$ of the point $\xi_{\widetilde{l}}^{\widetilde{T}}$ such that $B \cap \mathcal{N} = \{\xi_{\widetilde{l}}^{\widetilde{T}}\}$. This means that B isolates the point $\xi_{\widetilde{l}}^{\widetilde{T}}$ from all other points in $\mathcal{N}$. Now, consider a nonnegative function $c \in \mathcal{C}(\overline{\Omega})$ satisfying the following conditions:

1. $\mathrm{supp}\, c \subset B$,
2. $c(\xi_{\widetilde{l}}^{\widetilde{T}}) > 0$.

Then, we obtain

$$\sum_{T\subset\omega_i}\sum_{l=1}^{l_q}\chi_l^T c(\xi_l^T)\varphi_i(\xi_l^T)=\left(\sum_{(T,l)\in\mathcal{M}}\chi_l^T\right)c(\xi_{\widetilde{l}}^{\widetilde{T}})\varphi_i(\xi_{\widetilde{l}}^{\widetilde{T}})<0.$$

□

Let us assume that the sufficient condition for the validity of (1.28), which is stated in the following lemma, holds throughout this thesis.

Lemma 12. *Let the matrix $(a_{ij})_{j=1,\dots,N}^{i=1,\dots M}$ have entries given by (1.19), where a_h is defined by (1.25). If $\widehat{\chi}_l>0$ for all $l=1,\dots,l_q$, then*

$$\forall i\in\{1,\dots,M\}\quad \sum_{j=1}^{N}a_{ij}\geq 0. \tag{1.33}$$

Proof. The statement follows directly from the equality (1.30).

□

We will proceed by the positivity of the diagonal entries. According to (1.17), we have

$$\forall i\in\{1,\dots,M\}\quad a_{ii}>0. \tag{1.34}$$

Finally, we should discuss the validity of the remaining condition (1.27). For the sake of brevity, we consider only a simplicial triangulation $\mathcal{T}_h$ and $\mathcal{P}_1$ finite elements.

First, we introduce the following notation:

$$\mathbb{E}_T:=\frac{1}{\left|\widehat{T}\right|}\sum_{l=1}^{l_q}\widehat{\chi}_l\mathbb{E}(\xi_l^T), \tag{1.35}$$

where $T\in\mathcal{T}_h$ is arbitrary. Using this notation, we can rewrite the definition of the local form a_1^T as follows:

$$\begin{aligned}a_1^T(u_h,v_h)&=\sum_{l=1}^{l_q}\chi_l^T(\mathbb{E}(\xi_l^T)\nabla u_h|_T)\cdot\nabla v_h|_T=\\&=(\nabla u_h|_T)^\top\left(\sum_{l=1}^{l_q}\widehat{\chi}_l\left|\det B_T\right|\mathbb{E}(\xi_l^T)\right)\nabla v_h|_T\\&=|T|\,(\nabla u_h|_T)^\top\,\mathbb{E}_T\nabla v_h|_T,\end{aligned} \tag{1.36}$$

where we used the fact that

$$|\det B_T|=\frac{|T|}{\left|\widehat{T}\right|}.$$

Before presenting the final result of this section, we recall the following integral (see [14]):

$$\int_T\lambda_1^{A_1}\dots\lambda_{d+1}^{A_{d+1}}\mathrm{d}x=\frac{A_1!\dots A_{d+1}!d!}{(d+A_1+\dots+A_{d+1})!}\,|T|\,, \tag{1.37}$$

where $A_1, \dots, A_{d+1} \in \mathbb{N}_0$, $T \subset \mathbb{R}^d$ is a simplex with vertices $a_1, \dots, a_{d+1}$, and $\lambda_1, \dots, \lambda_{d+1} \in \mathcal{P}_1(T)$ are the barycentric coordinates, i.e., polynomials satisfying

$$\forall i, j \in \{1, \dots, d+1\} \quad \lambda_i(a_j) = \delta_{ij}.$$

Recall that the set S_i is given by (1.16).

Lemma 13. *Let $\mathcal{T}_h$ be a simplicial triangulation of the domain Ω. Consider a numerical quadrature on the reference element $\widehat{T}$ with $k_q \geq 2$. If*

$$\frac{\left|F_i^T\right|\left|F_j^T\right|}{d^2\,|T|}(n_j^T)^\top \mathbb{E}_T\, n_i^T + \frac{\|b\|_{L^\infty(T)}\left|F_j^T\right|}{d(d+1)} + \frac{\|c\|_{L^\infty(T)}\, d!\,|T|}{(d+2)!} \leq 0 \tag{1.38}$$

for all $i \in \{1, \dots, M\}$, $j \in S_i$, and $T \subset \omega_i \cap \omega_j$, then (1.27) holds true.

Proof. This proof is based on the combination of the approaches used in the proofs of Theorems 6.26, 7.4, and 8.5 in [1]. Consider $i \in \{1, \dots, M\}$ and $j \in \{1, \dots N\}$ with $i \neq j$. If $j \notin S_i$, then $\mathrm{supp}\varphi_i \cap \mathrm{supp}\varphi_j$ is a null set with respect to the d-dimensional Lebesgue measure and hence it follows that a_{ij} is equal to zero. Thus, we will consider only the case $j \in S_i$. Thanks to the decomposition

$$a_{ij} = a_h(\varphi_j, \varphi_i) = \sum_{T \in \mathcal{T}_h} a_1^T(\varphi_j, \varphi_i) + \sum_{T \in \mathcal{T}_h} a_2^T(\varphi_j, \varphi_i) + \sum_{T \in \mathcal{T}_h} a_3^T(\varphi_j, \varphi_i), \tag{1.39}$$

we can treat each term individually. Using the notation (1.35) and the formula (1.13), the first term can be written as follows:

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} a_1^T(\varphi_j, \varphi_i) &= \sum_{T \subset \omega_i \cap \omega_j} |T|\, (\nabla \varphi_j|_T)^\top\, \mathbb{E}_T\, (\nabla \varphi_i|_T) \\ &= \sum_{T \subset \omega_i \cap \omega_j} \frac{\left|F_i^T\right|\left|F_j^T\right|}{d^2\,|T|}(n_j^T)^\top \mathbb{E}_T\, n_i^T. \end{aligned} \tag{1.40}$$

The last two terms require a preparatory step. We assume that the numerical quadrature is at least of second (algebraic) order. Thus, using the integral (1.37), we obtain

$$\sum_{l=1}^{l_q} \chi_l^T (\varphi_i(\xi_l^T))^{A_1} (\varphi_j(\xi_l^T))^{A_2} = \frac{d!}{(A_1 + A_2 + d)!}\,|T|\,, \tag{1.41}$$

where $A_1, A_2 \in \{0, 1\}$, and $T \subset \omega_i \cap \omega_j$. Utilizing this equality, the Cauchy–Schwarz inequality, and the formula (1.13), the second term of the expansion (1.39) can be bounded as follows:

$$\begin{aligned} \sum_{T \in \mathcal{T}_h} a_2^T(\varphi_j, \varphi_i) &= \sum_{T \subset \omega_i \cap \omega_j} \sum_{l=1}^{l_q} \chi_l^T \varphi_i(\xi_l^T) b(\xi_l^T) \cdot \nabla \varphi_j|_T \\ &\leq \sum_{T \subset \omega_i \cap \omega_j} \frac{\|b\|_{L^\infty(T)}\left|F_j^T\right|}{|T|\,d} \sum_{l=1}^{l_q} \chi_l^T \varphi_i(\xi_l^T) \\ &= \frac{1}{d(d+1)} \sum_{T \subset \omega_i \cap \omega_j} \|b\|_{L^\infty(T)}\left|F_j^T\right|. \end{aligned} \tag{1.42}$$

Finally, the estimation of the third term in (1.39) relies only on (1.41). We proceed as follows:

$$\sum_{T\in\mathcal{T}_h} a_3^T(\varphi_j,\varphi_i) = \sum_{T\subset\omega_i\cap\omega_j}\sum_{l=1}^{l_q}\chi_l^T c(\xi_l^T)\varphi_i(\xi_l^T)\varphi_j(\xi_l^T) \le \sum_{T\subset\omega_i\cap\omega_j}\frac{\|c\|_{L^\infty(T)}\, d!\,|T|}{(d+2)!}. \tag{1.43}$$

By combining (1.39), (1.40), (1.42), and (1.43), we obtain

$$a_{ij} \le \sum_{T\subset\omega_i\cap\omega_j}\left(\frac{\left|F_i^T\right|\left|F_j^T\right|}{d^2\,|T|}(n_j^T)^\top\mathbb{E}_T\, n_i^T + \frac{\|b\|_{L^\infty(T)}\left|F_j^T\right|}{d(d+1)} + \frac{\|c\|_{L^\infty(T)}\, d!\,|T|}{(d+2)!}\right).$$

The proof is completed by using the assumption (1.38).

□

Further details about the anisotropic and heterogeneous diffusion problems can be found in [1], Section 6.4. Note that only the first term in (1.38) may be nonpositive. However, the fulfillment of the so-called *anisotropic nonobtuse angle condition*

$$\forall i\in\{1,\dots,M\}\quad \forall j\in S_i\quad \forall T\in\omega_i\cap\omega_j\quad (n_j^T)^\top\mathbb{E}_T\, n_i^T \le 0$$

may not necessarily be sufficient. This situation could occur if the remaining two terms in (1.38) are dominant.

1.4 Abstract Algebraic Flux Correction Setting

In the previous section, we observed that condition (1.27) may be violated for certain relevant data and triangulations. Consequently, the system matrix of the linear system (1.21) derived from the discrete variational form does not satisfy the conditions required for the validity of the DMPs.

The main idea of algebraic flux correction (AFC) schemes is to modify the linear system (1.20) in such a way that the assumptions, that ensure the validity of the DMPs, are met. The validity of this guarantees that many solutions containing spurious oscillations and nonphysical values are eliminated.

This idea for steady-state problems was originally proposed in [5], but we will follow the derivation in Section 10.1 of [1]. Note that almost all of the analysis has been developed in the last decade, and a brief overview of its development can be found at the beginning of Chapter 10 in [1].

First, we introduce an artificial diffusion matrix $\mathbb{D} = (d_{ij})_{i,j=1,\dots,N}$ with entries defined as follows:

$$\begin{aligned}\forall i,j\in\{1,2,\dots,N\}, i\neq j\quad d_{ij} = d_{ji} &:= -\max\{a_{ij},0,a_{ji}\},\\ \forall i\in\{1,2,\dots,N\}\quad d_{ii} &:= -\sum_{j\neq i} d_{ij},\end{aligned} \tag{1.44}$$

where $a_{ij} = a_h(\varphi_j,\varphi_i)$, $i,j=1,\dots,N$. It can be easily verified that the matrix $\widetilde{\mathbb{A}} = (\widetilde{a}_{ij})_{i,j=1,\dots,N}$ given by

$$\widetilde{\mathbb{A}} = \begin{pmatrix}\mathbb{A}^i & \mathbb{A}^b\\ \mathbb{O} & \mathbb{I}\end{pmatrix} + \mathbb{D}$$

is of nonnegative type and its diagonal entries are positive provided that $a_{ii} > 0$ and condition (1.33) holds.

Let us denote

$$\widetilde{\mathbb{A}^i} := (\widetilde{a}_{ij})_{i,j=1,\dots,M}, \quad \widetilde{\mathbb{A}^b} := (\widetilde{a}_{ij})_{j=M+1,\dots,N}^{i=1,\dots,M},$$

and turn our attention back to the linear system (1.21). If we replace $\mathbb{A}^i$ and $\mathbb{A}^b$ in the original system matrix with $\widetilde{\mathbb{A}^i}$ and $\widetilde{\mathbb{A}^b}$, respectively, the solution will satisfy the DMPs presented in the previous section. However, numerical experiments indicate that this modification leads to solutions that are excessively smeared. Therefore, the construction of an artificial diffusion should take into account the local behavior of the unknown solution.

A vector $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ solves (1.21) if and only if

$$\begin{aligned} \left(\widetilde{\mathbb{A}}U\right)_i &= g_i + (\mathbb{D}U)_i, \quad i = 1, \dots, M, \\ u_i &= u_b(x_i), \quad i = M+1, \dots, N. \end{aligned} \tag{1.45}$$

We introduce the so-called *algebraic fluxes*

$$f_{ij} := d_{ij}(u_j - u_i) \tag{1.46}$$

for $i, j = 1, \dots, N$. These quantities represent a discrete algebraic analogue of the usual fluxes of physical quantities like concentration or heat. Moreover, they satisfy the co-called *conservation property*

$$\forall i, j \in \{1, \dots N\} \quad f_{ij} = -f_{ji}.$$

Utilizing this notation, we can obtain the following identity for all $i = 1, \dots, M$:

$$\sum_{j \neq i} f_{ij} = \sum_{j=1}^{N} d_{ij} u_j - u_i \sum_{j=1}^{N} d_{ij} = \sum_{j=1}^{N} d_{ij} u_j = (\mathbb{D}U)_i,$$

where we used the fact that the row sums of $\mathbb{D}$ are zero. Inserting these expressions into (1.45), we obtain

$$\left(\widetilde{\mathbb{A}}U\right)_i = g_i + \sum_{j \neq i} f_{ij}, \quad i = 1, \dots, M. \tag{1.47}$$

The strategy of AFC schemes is to limit algebraic fluxes in this system using solution-dependent factors (so-called limiters) $\alpha_{ij} = \alpha_{ij}(u_h) \in \langle 0, 1 \rangle$. The first equation in (1.47) then takes the form

$$\left(\widetilde{\mathbb{A}}U\right)_i = g_i + \sum_{j \neq i} \alpha_{ij} f_{ij}, \quad i = 1, \dots, M. \tag{1.48}$$

Intuitively, the limiters α_{ij} should be close to 1 so that the system (1.48) remains similar to the original system (1.21). To ensure the conservative property of the product $\alpha_{ij} f_{ij}$, we assume the symmetry of the limiters

$$\forall i, j \in \{1, \dots, N\} \quad \alpha_{ij} = \alpha_{ji}. \tag{1.49}$$

Rewriting (1.48) using the fact that the row sums of the matrix $\mathbb{D}$ vanish, we obtain the following final form of the nonlinear *AFC scheme*:

$$\begin{aligned} \sum_{j=1}^{N} a_{ij}u_j + \sum_{j=1}^{N}(1-\alpha_{ij})d_{ij}(u_j-u_i) = g_i, \quad i = 1,\dots,M, \\ u_i = u_b(x_i), \quad i = M+1,\dots,N, \end{aligned} \tag{1.50}$$

where $\alpha_{ij} = \alpha_{ij}(u_h)$ and u_h is a discrete solution corresponding to U via (1.8). Note that the values α_{ij} for $i,j \in \{M+1,\dots,N\}$ do not appear in the system (1.50). We define them as

$$\forall i,j \in \{M+1,\dots,N\}, i \neq j \quad \alpha_{ij} := 1. \tag{1.51}$$

It is convenient to introduce a matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ defined as follows:

$$\begin{aligned} \forall i,j \in \{1,2,\dots,N\}, i \neq j \quad b_{ij}(u_h) := (1-\alpha_{ij}(u_h))d_{ij}, \\ \forall i \in \{1,\dots,N\} \quad b_{ii}(u_h) := -\sum_{k\neq i} b_{ik}(u_h). \end{aligned} \tag{1.52}$$

Furthermore, let us state the properties of this matrix.

Lemma 14. *Consider the matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ given by (1.52) with $\alpha_{ij}(u_h) = \alpha_{ji}(u_h) \in \langle 0,1\rangle$, $i,j = 1,\dots,N$. Then, this matrix possess following properties:*

$$\forall i,j \in \{1,\dots,N\} \quad \forall v_h \in W_h \quad b_{ij}(v_h) = b_{ji}(v_h), \tag{1.53}$$

$$\forall i,j \in \{1,\dots,N\}, i \neq j \quad \forall v_h \in W_h \quad b_{ij}(v_h) \leq 0, \tag{1.54}$$

$$\forall i \in \{1,\dots,N\} \quad \forall v_h \in W_h \quad \sum_{j=1}^{N} b_{ij}(v_h) = 0, \tag{1.55}$$

$$\forall i,j \in \{M+1,\dots,N\}, i \neq j \quad \forall v_h \in W_h \quad b_{ij}(v_h) = 0, \tag{1.56}$$

and

$$\forall i \in \{1,\dots,M\} \quad \forall j \notin S_i \cup \{i\} \quad \forall v_h \in W_h \quad b_{ij}(v_h) = 0. \tag{1.57}$$

Proof. All properties are straightforward consequences of the definition (1.52), the properties of the matrix $(a_{ij})_{ij=1,\dots,N}$, and the general properties of the limiters stated above.

□

Utilizing this notation, the system (1.50) can be written in the following form:

$$\begin{aligned} \sum_{j=1}^{N} \left(a_{ij} + b_{ij}(u_h)\right) u_j = g_i, \quad i = 1,\dots,M, \\ u_i = u_b(x_i), \quad i = M+1,\dots,N, \end{aligned} \tag{1.58}$$

which is also important for theoretical considerations, including the existence of the solution, which we will discuss now. The following theorem is adapted from [1], Subsection 3.2.2 (Theorem 3.81). The proof is based on the Brouwer fixed-point theorem.

Theorem 15. *Let us consider the matrix* $(b_{ij}(u_h))_{i,j=1,\dots,N}$ *given by (1.52) and based on the matrix* $(a_{ij})_{i,j=1,\dots,N}$ *with the entries* $a_{ij} = a_h(\varphi_j, \varphi_i)$*, and let us suppose that, for any* $i \in \{1, \dots M\}$*,*

$$v_h \in W_h \longmapsto \sum_{j=1}^{N} b_{ij}(v_h)(v_j - v_i) \tag{1.59}$$

is a continuous mapping and that, for any $i \in \{1, \dots M\}$ *and any* $j \in \{M+1, \dots N\}$*, the mapping* $v_h \in W_h \longmapsto b_{ij}(v_h)$ *is bounded. Then there exists a solution to the nonlinear problem (1.58).*

1.5 Discrete Maximum Principles for Nonlinear AFC Scheme

In this section, we summarize the results for the local and global DMPs in the context of the nonlinear problem (1.58) (or equivalently for (1.50)). For more details, see [1], Section 3.2.

The following definition formulates a key property under which the local DMPs can be proven.

Definition 15. *The system (1.58) is said to satisfy the* matrix-based DMP property for strict extrema *if the following condition holds: Consider any* $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ *and any* $i \in \{1, \dots, M\}$*. If one of the following options holds:*

1. $\forall j \in S_i \quad u_i < u_j$,
2. $\forall j \in S_i \quad u_i > u_j$,

then

$$\forall j \in S_i \quad a_{ij} + b_{ij}(u_h) \le 0,$$

where u_h *corresponds to* U *via the identification (1.8).*

Theorem 16. *Let conditions (1.33) and (1.34) hold for the system matrix of (1.21). Suppose that the system (1.58), with* $(b_{ij}(u_h))_{i,j=1,\dots,N}$ *given by (1.52), satisfies the matrix-based DMP property for strict extrema. Then, for any solution* $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ *of the nonlinear system (1.58), the local DMPs*

$$\begin{aligned} g_i \le 0 \quad &\Longrightarrow \quad u_i \le \max_{j \in S_i} u_j^+, \\ g_i \ge 0 \quad &\Longrightarrow \quad u_i \ge \min_{j \in S_i} u_j^-, \end{aligned}$$

hold for all $i \in \{1 \dots, M\}$*. If, in addition, the condition (1.29) holds, then the stronger local DMPs*

$$\begin{aligned} g_i \le 0 \quad &\Longrightarrow \quad u_i \le \max_{j \in S_i} u_j, \\ g_i \ge 0 \quad &\Longrightarrow \quad u_i \ge \min_{j \in S_i} u_j, \end{aligned}$$

hold for all $i \in \{1 \dots, M\}$ *and any solution* $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ *of the nonlinear system (1.58).*

Proof. See the proof of Theorem 3.72 from [1].

□

The proof of the global DMP requires a stronger assumption than that stated in Definition 15.

Definition 16. *The system (1.58) is said to satisfy the* matrix-based DMP property for nonstrict extrema *if the following condition holds: Consider any* $U = (u_1, \dots, u_N)^T \in \mathbb{R}^N$ *and any* $i \in \{1, \dots, M\}$*. If one of the following options holds:*

1. $\forall j \in S_i \quad u_i \leq u_j$,
2. $\forall j \in S_i \quad u_i \geq u_j$,

then

$$\forall j \in S_i, u_j \neq u_i \quad a_{ij} + b_{ij}(u_h) \leq 0,$$

where u_h *corresponds to* U *via the identification (1.8).*

Theorem 17. *Let the condition (1.33) holds for the system matrix of (1.21) and let* A^i *be positive definite. Suppose that the system (1.58), with* $(b_{ij}(u_h))_{i,j=1,\dots,N}$ *given by (1.52), satisfies the matrix-based DMP property for nonstrict extrema. Then, the global DMPs*

$$\begin{aligned} \forall i \in \{1, \dots M\} \quad g_i \leq 0 \quad &\Longrightarrow \quad \max_{j=1,\dots,N} u_j \leq \max_{j=M+1,\dots,N} u_j^+, \\ \forall i \in \{1, \dots M\} \quad g_i \geq 0 \quad &\Longrightarrow \quad \min_{j=1,\dots,N} u_j \geq \min_{j=M+1,\dots,N} u_j^- \end{aligned}$$

hold for any solution $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ *of the nonlinear system (1.58). If, in addition, the condition (1.29) is fulfilled, then the stronger global DMPs*

$$\begin{aligned} \forall i \in \{1, \dots M\} \quad g_i \leq 0 \quad &\Longrightarrow \quad \max_{j=1,\dots,N} u_j = \max_{j=M+1,\dots,N} u_j, \\ \forall i \in \{1, \dots M\} \quad g_i \geq 0 \quad &\Longrightarrow \quad \min_{j=1,\dots,N} u_j = \min_{j=M+1,\dots,N} u_j, \end{aligned}$$

hold for any solution $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ *of the nonlinear system (1.58).*

Proof. See the proof of Theorem 3.78 from [1].

□

2 Error Analysis of AFC Schemes

This chapter is dedicated to the derivation of a priori error estimates for AFC schemes. The main issue will be to derive an estimate of the term arising from the artificial diffusion. This estimate will be discussed under various assumptions. although we slightly extend these results on our case with the more general diffusion tensor. Nevertheless, the ideas of the proofs will be the same.

An analysis presented in this chapter was originally proposed in [6] and [7], but we will mostly follow [1], Section 10.2, where a detailed development of these techniques can be found, although we slightly extend these results on our case with the more general diffusion tensor. Nevertheless, the ideas of the proofs will be the same.

The first two sections contain a brief overview of error estimates in the special norm associated with AFC schemes. In the last section, the L^2-error estimate is derived for linearity-preserving AFC schemes in the diffusion-dominated regime. Let us mention that L^2-error analysis in the context of algebraic stabilization is a new result.

For simplicity, throughout this chapter, we assume exact integration in the definition of the discrete variational method

$$a_h(\cdot,\cdot) := a(\cdot,\cdot), \quad \langle g_h, \cdot \rangle_{W_h^*, W_h} := (g, \cdot)\,,$$

where $a(\cdot,\cdot)$ is the bilinear form from (1.3), and $(\cdot,\cdot)$ denotes the standard scalar product in $L^2(\Omega)$.

2.1 General Error Estimate

For an error analysis, it is convenient to rewrite the system of nonlinear equations (1.50) in a variational form. We define

$$b_h(u_h; v_h, z_h) := \sum_{i,j=1}^{N} (1 - \alpha_{ij}(u_h))\, d_{ij}\, (v_h(x_j) - v_h(x_i))\, z_h(x_i), \tag{2.1}$$

where $u_h, v_h, z_h \in W_h$. It is also useful to express this mapping equivalently in terms of the matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ defined in (1.52)

$$b_h(u_h; v_h, z_h) := \sum_{i,j=1}^{N} b_{ij}(u_h)\, (v_h(x_j) - v_h(x_i))\, z_h(x_i). \tag{2.2}$$

Then, the variational form of the AFC scheme reads as follows: Find $u_h \in W_h$ such that $u_h(x_i) = u_b(x_i)$, $i = M+1, \dots, N$, and

$$\forall v_h \in V_h \quad a(u_h, v_h) + b_h(u_h; u_h, v_h) = (g, v_h). \tag{2.3}$$

Now, we construct a natural norm for the error estimates of the AFC schemes. This norm will consist of two components. The first component is a so-called

energy norm in the space $H_0^1(\Omega)$

$$\|v\|_{\mathrm{en}} := \left(\left\| \sqrt{\mathbb{E}} \nabla v \right\|_{L^2(\Omega)}^2 + \mu \|v\|_{L^2(\Omega)}^2 \right)^{\frac{1}{2}},$$

where $\mu := \operatorname{essinf}_\Omega c$, and $\sqrt{\mathbb{E}}$ denotes a symmetric matrix, which satisfy

$$\sqrt{\mathbb{E}}\sqrt{\mathbb{E}} = \mathbb{E}.$$

Note that this matrix can be determined via the spectral decomposition for the symmetric positive definite matrices. The energy norm is associated with the symmetric part of the form $a(\cdot,\cdot)$. For the sake of brevity, we will consider only the case $\mu > 0$.

Next, we turn our attention to the mapping b_h.

Lemma 18. *Consider any $\mu_{ij} = \mu_{ji} \leq 0$, $i,j = 1,\dots N$. Then*

$$\sum_{i,j=1}^{N} \mu_{ij} v_i \left(v_j - v_i\right) = -\sum_{i=1}^{N} \sum_{j>i} \mu_{ij} (v_i - v_j)^2 \geq 0$$

for all $v_1,\dots v_N \in \mathbb{R}$.

Proof. See the proof of Lemma 1 from [6]. □

From this lemma, it obviously follows that the mapping $b_h(u_h;\cdot,\cdot) : W_h \times W_h \to \mathbb{R}$ is symmetric and nonnegative bilinear form on W_h. Thus, it satisfies the Cauchy-Schwarz inequality

$$\forall u_h, v_h, z_h \in W_h \quad |b_h(u_h; v_h, z_h)| \leq \left(b_h(u_h; v_h, v_h)\right)^{\frac{1}{2}} \left(b_h(u_h; z_h, z_h)\right)^{\frac{1}{2}}. \tag{2.4}$$

It is clear that the functional $\left(b_h(u_h;\cdot,\cdot)\right)^{\frac{1}{2}}$ is a seminorm on W_h. This is the second component of the final norm on V_h which we will use for a priori error estimate

$$\|v_h\|_h := \left(\|v_h\|_{\mathrm{en}}^2 + b_h(u_h; v_h, v_h) \right)^{\frac{1}{2}} \tag{2.5}$$

for all $v_h \in W_h$. This norm (which may only be a seminorm on W_h) obviously depends on the discrete solution u_h.

Now, we present the general error estimate for the AFC scheme with respect to this norm.

Theorem 19. *Let $u_h \in W_h$ be a solution of the nonlinear problem (2.3) defined using a matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ given by (1.52), and let $u \in H^2(\Omega)$ be a weak solution of (1.3). Then*

$$\|u - u_h\|_h \leq \|u - I_h u\|_{\mathrm{en}} + \sup_{v_h \in V_h} \frac{a(u - I_h u, v_h)}{\|v_h\|_h} + \left(b_h(u_h; I_h u, I_h u)\right)^{\frac{1}{2}}, \tag{2.6}$$

where $I_h : \mathcal{C}(\overline{\Omega}) \to W_h$ denotes the global Lagrange interpolation operator corresponding to the finite element space W_h.

Proof. See the proof of Theorem 10.5 from [1].

□

The first two terms on the right-hand side of (2.6) can be estimated using standard techniques based on the Hölder inequality and the interpolation error estimates (1.11).

Lemma 20. *Let $\{W_h\}$ be a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces defined on a shape-regular family of triangulations $\{\mathcal{T}_h\}$ consisting of simplices, parallelograms or parallelepipeds. Consider any $u \in H^2(\Omega)$. Then*

$$\|u - I_h u\|_{\text{en}} \leq C_{\text{int}} h \left(\|\mathbb{E}\|_{L^\infty(\Omega)} + \mu h^2 \right)^{\frac{1}{2}} |u|_{2,2,\Omega} , \tag{2.7}$$

where C_{int} is the constant from the interpolation error estimate (1.11), and

$$\sup_{v_h \in V_h} \frac{a(u - I_h u, v_h)}{\|v_h\|_h} \leq 2 C_{\text{int}} C_{\text{data}}^{\frac{1}{2}} h \, |u|_{2,2,\Omega} , \tag{2.8}$$

where

$$C_{\text{data}} := \|\mathbb{E}\|_{L^\infty(\Omega)} + \mu^{-1} \|b\|^2_{L^\infty(\Omega)} + \mu^{-1} \|c\|^2_{L^\infty(\Omega)} h^2.$$

Proof. See the proof of Lemma 10.6 from [1].

□

The last term on the right-hand side of (2.6) represents an estimate of the consistency error. The estimate of this term is given in the following lemma.

Lemma 21. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces defined on a shape-regular family of triangulations $\{\mathcal{T}_h\}$ consisting of simplices, parallelograms or parallelepipeds. Consider the mapping b_h given by (2.2) with matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ given by (1.52). Then*

$$\forall u \in \mathcal{C}(\overline{\Omega}) \quad \forall u_h \in W_h \quad (b_h(u_h; I_h u, I_h u))^{\frac{1}{2}} \leq C_{\text{co}} \left(A_h(u_h) \right)^{\frac{1}{2}} |I_h u|_{1,2,\Omega} , \tag{2.9}$$

where C_{co} depends only on the shape-regular constant κ (see Definition 4), and

$$A_h(u_h) := \max_{i,j=1,\dots,N, i \neq j} \left(|b_{ij}(u_h)| \, \|x_i - x_j\|_2^{2-d} \right). \tag{2.10}$$

Moreover, the following holds:

$$|A_h(u_h)| \leq C \left(\|\mathbb{E}\|_{L^\infty(\Omega)} + \|b\|_{L^\infty(\Omega)} h + \|c\|_{L^\infty(\Omega)} h^2 \right), \tag{2.11}$$

where C is independent of h and the data of the problem (1.1).

Proof. See the proof of Lemma 10.8 from [1].

□

As a conclusion, we present the following error estimate for the AFC scheme.

Theorem 22. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces defined on a shape-regular family of triangulations $\{\mathcal{T}_h\}$ consisting of simplices, parallelograms or parallelepipeds. Let $u_h \in W_h$ be a solution of the nonlinear problem (2.3) (or equivalently, a solution of (1.50)), and let $u \in H^2(\Omega)$ be the weak solution of (1.3). Then we have following a priori error estimate:*

$$\begin{aligned}\|u - u_h\|_h \leq 3C_{\mathrm{int}} h\Big(&\|\mathbb{E}\|_{L^\infty(\Omega)} + \mu^{-1}\|b\|^2_{L^\infty(\Omega)} + \mu^{-1}\|c\|^2_{L^\infty(\Omega)} h^2 \\ &+ \mu h^2\Big)^{\frac{1}{2}} |u|_{2,2,\Omega} + C\Big(\|\mathbb{E}\|_{L^\infty(\Omega)} + \|b\|_{L^\infty(\Omega)} h \\ &+ \|c\|_{L^\infty(\Omega)} h^2\Big)^{\frac{1}{2}} |I_h u|_{1,2,\Omega}\,,\end{aligned} \tag{2.12}$$

where C is independent of h and the data of problem (1.1).

Proof. This result follows directly from Theorem 19 and Lemmata 20 and 21. □

Let us conclude this section by a discussion of the above error estimate in the norm $\|\cdot\|_h$ for different regimes.

Definition 17. *A discrete problem is said to be in:*

1. *the* diffusion-dominated regime *if* $\|\mathbb{E}\|_{L^\infty(\Omega)} \gg h\|b\|_{L^\infty(\Omega)}, h^2\|c\|_{L^\infty(\Omega)}$,
2. *the* convection-dominated regime *if* $h\|b\|_{L^\infty(\Omega)} \gg \|\mathbb{E}\|_{L^\infty(\Omega)}, h^2\|c\|_{L^\infty(\Omega)}$,
3. *the* reaction-dominated regime *if* $h^2\|c\|_{L^\infty(\Omega)} \gg h\|b\|_{L^\infty(\Omega)}, \|\mathbb{E}\|_{L^\infty(\Omega)}$.

We first consider the convection-dominated regime. In this case, the consistency error estimate (the second term in (2.12)) is of order $\mathcal{O}(h^{\frac{1}{2}})$. Consequently, the estimate (2.12) predicts a convergence rate of order $\mathcal{O}(h^{\frac{1}{2}})$. Numerical experiments have shown that this bound (2.12) is generally sharp. However, in certain situations, a better convergence behavior can be observed (see [6] for more details).

Now, we turn our attention to the diffusion-dominated regime. The consistency error estimate in (2.12) reduces to order $\mathcal{O}(1)$, which does not imply convergence. If we consider a homogeneous and isotropic diffusion tensor $\mathbb{E} = \varepsilon\mathbb{I}$, one possible remedy for this problem is to consider only triangulations satisfying the XZ-condition. According to Theorem 6, we have

$$\forall i, j \in \{1, \dots M\}, i \neq j \quad a_{ij} \leq \widetilde{a}_{ij} := a_{ij} - \varepsilon\left(\nabla\varphi_j, \nabla\varphi_i\right). \tag{2.13}$$

For a more general diffusion tensor, assume that for any $i, j \in \{1, \dots, N\}$, with $i \neq j$, the following holds:

$$(\mathbb{E}\nabla\varphi_j, \nabla\varphi_i) \leq 0.$$

Then, we have

$$\forall i, j \in \{1, \dots M\}, i \neq j \quad a_{ij} \leq \widetilde{a}_{ij} := a_{ij} - \left(\mathbb{E}\nabla\varphi_j, \nabla\varphi_i\right). \tag{2.14}$$

Under the condition (2.13) or (2.14), respectively, it can be shown (see [1], Section 10.2) that

$$|A_h(u_h)| \leq Ch\left(\|b\|_{L^\infty(\Omega)} + \|c\|_{L^\infty(\Omega)} h\right), \tag{2.15}$$

Thus, the convergence rate is also of order $\mathcal{O}(h^{\frac{1}{2}})$.

For the reaction-dominated regime, the estimate (2.12) implies convergence of order $\mathcal{O}(h)$.

2.2 Improved Error Estimate

The estimate (2.11) uses minimal information about the behavior of the limiters α_{ij}, only the rough bound $\alpha_{ij} \leq 1$. This may lead to the issues described at the end of the previous section.

Another potential remedy for this problem is to employ additional information about the limiters. In the following, we present a result that combines Lipschitz continuity with respect to local variations and the linearity preservation property, leading to an improved error estimate (see [7]).

Definition 18. *Consider the limiters α_{ij}, $i, j \in \{1, \dots, N\}$, introduced in Section 1.4. These limiters possess* linearity-preservation property *if the following holds:*

$$\forall i \in \{1, \dots, M\} \quad \forall j \in S_i \quad \forall v_h \in W_h, v_h|_{\widetilde{\omega}_{ij}} \in \mathcal{P}_1(\widetilde{\omega}_{ij}) \quad \alpha_{ij}(v_h) = 1, \tag{2.16}$$

where

$$\widetilde{\omega}_{ij} := \bigcup \{T \in \mathcal{T}_h;\ x_i \in T \ \vee \ x_j \in T\} = \omega_i \cup \omega_j. \tag{2.17}$$

Before defining Lipschitz continuity with respect to local variations, we introduce the following seminorms for $v_h \in W_h$

$$|v_h|_i := \max_{j \in S_i} |v_h(x_j) - v_h(x_i)| = \max_{j \in S_i} |v_j - v_i|,$$

where $i = 1, \dots, M$.

Definition 19. *The functions $u_h \in W_h \longmapsto b_{ij}(u_h)(u_j - u_i)$, $i = 1, \dots, M$, $j \in S_i$, are called* Lipschitz continuous with respect to local variations *if for any $i \in \{1, \dots, M\}$ and $j \in S_i$, there exists constant $L_{ij} > 0$ such that for all $u_h, v_h \in W_h$ we have*

$$|b_{ij}(u_h)(u_j - u_i) - b_{ij}(v_h)(v_j - v_i)| \leq L_{ij} \max\{|u_h - v_h|_i, |u_h - v_h|_j\},$$

if $j \in \{1, \dots, M\} \cap S_i$, and

$$|b_{ij}(u_h)(u_j - u_i) - b_{ij}(v_h)(v_j - v_i)| \leq L_{ij} |u_h - v_h|_i,$$

if $j \in \{M+1, \dots, N\} \cap S_i$.

In this thesis, a detailed proof of the improved consistency error estimate will not be presented. However, the technique proposed in [7] and used in this proof will be demonstrated in the next section (see the proof of Lemma 27).

Lemma 23. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces defined on a shape-regular family of triangulations $\{\mathcal{T}_h\}$ consisting of simplices, parallelograms or parallelepipeds. Let the matrix $(b_{ij}(u_h))_{i,j=1,\dots,N}$ be given by (1.52) and satisfy Definition 19, and let the corresponding limiters possess the linearity-preservation property. Then, if there exists a constant $C_{\mathrm{lp}} > 0$ independent of h such that*

$$\forall u \in H^2(\widetilde{\omega}_{ij}) \qquad \inf_{v \in \mathcal{P}_1(\widetilde{\omega}_{ij})} |u - v|_{1,2,\widetilde{\omega}_{ij}} \leq C_{\mathrm{lp}} h |u|_{2,2,\widetilde{\omega}_{ij}} \tag{2.18}$$

for any $i \in \{1, \dots M\}$ and $j \in S_i$, then

$$\begin{aligned}(b_h(u_h; I_h u, I_h u))^{\frac{1}{2}} \leq \frac{1}{2} \|u - u_h\|_{\mathrm{en}} + \varepsilon^{\frac{1}{2}} h |u|_{2,2,\Omega} \\ + C(A_h(u_h) + L_h)\varepsilon^{-\frac{1}{2}} |I_h u|_{1,2,\Omega}\end{aligned}$$

for all $u \in H^2(\Omega)$ and $u_h \in W_h$, where

$$L_h := \max_{i=1,\dots M, j \in S_i} \left(L_{ij} \left\| x_i - x_j \right\|_2^{2-d} \right), \tag{2.19}$$

A_h is given by (2.10), and ε is from (1.2). The constant C depends only on the constant C_{int} introduced in Theorem 3, the shape-regularity constant κ, and the constant C_{lp}.

Proof. See the proof of Corollary 10.18 from [1].

□

Note that the inequality (2.18) holds under certain assumptions on the shape of the patches $\widetilde{\omega}_{ij}$, which are usually satisfied. For further details, see [15], Section 4.3.

The expression (2.19) has a structure similar to that of A_h (see (2.10)). Thus, we assume (as is often the case) that an estimate similar to (2.15) also holds for L_h

$$A_h(u_h) + L_h \leq C \left(\left\| b \right\|_{L^\infty(\Omega)} h + \left\| c \right\|_{L^\infty(\Omega)} h^2 \right). \tag{2.20}$$

Using this estimate, together with the result in Lemma 23, we obtain a convergence rate of order $\mathcal{O}(h)$ in $\left\| \cdot \right\|_h$.

Theorem 24. *Under the assumptions of Theorem 22 and Lemma 23, we have the error estimate*

$$\begin{aligned} \left\| u - u_h \right\|_h \leq 6 C_{\text{int}} h \Big(& \left\| \mathbb{E} \right\|_{L^\infty(\Omega)} + \mu^{-1} \left\| b \right\|_{L^\infty(\Omega)}^2 + \mu^{-1} \left\| c \right\|_{L^\infty(\Omega)}^2 h^2 \\ & + \mu h^2 \Big)^{\frac{1}{2}} \left| u \right|_{2,2\Omega} + 2 \varepsilon^{\frac{1}{2}} h \left| u \right|_{2,2,\Omega} \\ & + 2C \left(\left\| b \right\|_{L^\infty(\Omega)} h + \left\| c \right\|_{L^\infty(\Omega)} h^2 \right) \varepsilon^{-\frac{1}{2}} \left| I_h u \right|_{1,2,\Omega}, \end{aligned} \tag{2.21}$$

provided that the estimate (2.20) holds.

Proof. The estimate (2.21) follows directly from Theorem 19 and Lemmata 20 and 23.

□

2.3 L^2-error Estimates

In this section, we study an error estimate in the L^2-norm for the AFC scheme in the diffusion-dominated regime.

We restrict our analysis to a homogeneous Dirichlet boundary condition prescribed on the whole boundary, a shape-regular family of simplicial triangulations $\{\mathcal{T}_h\}$, and Lagrange $\mathcal{P}_1$ finite elements. We focus on linearity-preserving AFC schemes only, as these will be the primary subject of study in this thesis.

Our error analysis employs the Aubin-Nitsche method, which is based on a dual problem (see [11], Section 2.3). Recall that we consider the bilinear form $a(\cdot, \cdot)$ given by (1.3).

Definition 20. *The* dual problem *reads as follows: For a given $\psi \in L^2(\Omega)$, find $w_\psi \in H_0^1(\Omega)$ such that*

$$\forall v \in H_0^1(\Omega) \quad a(v, w_\psi) = (\psi, v). \tag{2.22}$$

We already know (see Theorem 2), that the following holds:

$$\forall \psi \in L^2(\Omega) \quad \exists! w_\psi \in H_0^1(\Omega) \quad \forall v \in H_0^1(\Omega) \quad a(v, w_\psi) = (\psi, v). \tag{2.23}$$

Moreover, we will suppose that $w_\psi \in H^2(\Omega)$ and that there exists a constant $C_R > 0$, independent of ψ, satisfying

$$\|w_\psi\|_{2,2,\Omega} \leq C_R \|\psi\|_{L^2(\Omega)}. \tag{2.24}$$

These assumptions hold for sufficiently regular problem data and/or a sufficiently smooth boundary $\partial\Omega$. For further details, see [10], Chapter 8.

Next, we present an important dual characterization of the norm in a Hilbert space.

Lemma 25. *Let X be a real Hilbert space with a scalar product $(\cdot,\cdot)_X$, then*

$$\forall v \in X \quad \|v\|_X = \sup_{x \in X,\, x \neq 0} \frac{(x, v)_X}{\|x\|_X}. \tag{2.25}$$

Proof. See [16], Chapter 3, Section 6.

□

Let $u_h \in V_h$ be a discrete solution of (2.3). By subtracting (2.3) from (1.3) and using the fact that $V_h \subset H_0^1(\Omega)$, we obtain

$$\forall v_h \in V_h \quad a(u - u_h, v_h) - b_h(u_h; u_h, v_h) = 0.$$

Employing the assumption (2.23) and this equality, we have

$$\begin{aligned}(u_h - u, \psi) &= a(u - u_h, w_\psi) \\ &= a(u - u_h, w_\psi) - a(u - u_h, I_h w_\psi) + b_h(u_h; u_h, I_h w_\psi) \\ &= a(u - u_h, w_\psi - I_h w_\psi) + b_h(u_h; u_h, I_h w_\psi),\end{aligned}$$

where $\psi \in L^2(\Omega)$ is arbitrary. We used the fact that $u - u_h \in H_0^1(\Omega)$ can be used as a test function for the dual problem.

Next, we apply the dual characterization of L^2-norm

$$\begin{aligned}\|u - u_h\|_{L^2(\Omega)} &= \sup_{\psi \in L^2(\Omega), \psi \neq 0} \frac{(u_h - u, \psi)}{\|\psi\|_{L^2(\Omega)}} \\ &= \sup_{\psi \in L^2(\Omega), \psi \neq 0} \left\{ \frac{a(u - u_h, w_\psi - I_h w_\psi)}{\|\psi\|_{L^2(\Omega)}} + \frac{b_h(u_h; u_h, I_h w_\psi)}{\|\psi\|_{L^2(\Omega)}} \right\}.\end{aligned} \tag{2.26}$$

The first term on the right-hand side of (2.26) can be treated as follows:

$$\begin{aligned}\frac{a(u - u_h, w_\psi - I_h w_\psi)}{\|\psi\|_{L^2(\Omega)}} &\leq 2C_{\mathrm{int}} h \Big(\|\mathbb{E}\|_{L^\infty(\Omega)} + \|b\|^2_{L^\infty(\Omega)} \varepsilon^{-1} h^2 \\ &\qquad + \mu^{-1} \|c\|^2_{L^\infty(\Omega)} h^2 \Big)^{\frac{1}{2}} \|u - u_h\|_{\mathrm{en}} \frac{|w_\psi|_{2,2,\Omega}}{\|\psi\|_{L^2(\Omega)}} \\ &\leq 2C_R C_{\mathrm{int}} h \Big(\|\mathbb{E}\|_{L^\infty(\Omega)} + \|b\|^2_{L^\infty(\Omega)} \varepsilon^{-1} h^2 \\ &\qquad + \mu^{-1} \|c\|^2_{L^\infty(\Omega)} h^2 \Big)^{\frac{1}{2}} \|u - u_h\|_{\mathrm{en}},\end{aligned} \tag{2.27}$$

where we used the Hölder inequality, the interpolation error estimate (1.11), ellipticity of the diffusion tensor (1.2), and the assumption (2.24).

The main issue of the Aubin-Nitsche method for the nonlinear AFC scheme is to estimate the second term on the right-hand side of (2.26). For this purpose, we need the following lemmata.

Lemma 26. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ finite element spaces defined on a shape-regular family of simplicial triangulations. Then there exists a constant C_{ed}, depending only on the shape-regularity constant κ, such that*

$$\forall v_h \in W_h \quad |v_i - v_j| \leq C_{\mathrm{ed}} h_{E_{ij}}^{\frac{2-d}{2}} |v_h|_{1,2,\widetilde{\omega}_{ij}} \tag{2.28}$$

and

$$\forall v_h \in W_h \quad |v_h|_i \leq C_{\mathrm{ed}} h_{E_{ij}}^{\frac{2-d}{2}} |v_h|_{1,2,\widetilde{\omega}_{ij}}, \tag{2.29}$$

where $i \in \{1, \ldots, M\}$ and $j \in S_i$, and $h_{E_{ij}}$ denotes the length of the edge E_{ij} with endpoints x_i, x_j (for notation, see Subsection 1.2.2).

Proof. The proof utilizes a transformation to the reference element, the equivalence of norms on the finite-dimensional space of polynomials, and the properties of shape-regular triangulations. For details, see [1], the proof of Lemma 10.16. □

Lemma 27. *Let W_h belong to a family of Lagrange $\mathcal{P}_1$ finite element spaces defined on a shape-regular family of simplicial triangulations. Suppose that the matrix $(b_{ij}(u_h))_{i,j=1,\ldots,N}$, given by (1.52), satisfies Definition 19, the corresponding limiters possess the linearity-preservation property, and the estimate (2.18) holds. Then, there exists a constant $C_{\mathrm{stab}} > 0$, independent of h, such that, for any $u_h, z_h \in W_h$ and $u \in H^2(\Omega)$, following holds:*

$$|b_h(u_h; u_h, z_h)| \leq C_{\mathrm{stab}} L_h \left(\varepsilon^{-\frac{1}{2}} \|u - u_h\|_{\mathrm{en}} + h\, |u|_{2,2,\Omega} \right) |z_h|_{1,2,\Omega}, \tag{2.30}$$

where L_h is given by (2.19).

Proof. This proof follows the ideas from [7] and [1], Section 10.2 (see the proofs of Lemmata 10.16 and 10.17). We begin by reformulating the expression

$$b_h(u_h; u_h, z_h) = \sum_{i,j=1}^{N} b_{ij}(u_h)(u_j - u_i) z_i.$$

Interchanging the indices and applying (1.53) yields

$$b_h(u_h; u_h, z_h) = \sum_{i,j=1}^{N} b_{ij}(u_h)(u_i - u_j) z_j.$$

Summing these expressions gives

$$b_h(u_h; u_h, z_h) = \frac{1}{2} \sum_{i,j=1}^{N} b_{ij}(u_h)(u_j - u_i)(z_i - z_j).$$

By decomposing this sum and using (1.56), (1.57), and (1.53), we obtain

$$\begin{aligned} b_h(u_h;u_h,z_h) = \frac{1}{2}\sum_{i=1}^{M}\sum_{j\in S_i\cap\{1,\dots,M\}} b_{ij}(u_h)(u_j-u_i)(z_i-z_j) \\ +\sum_{i=1}^{M}\sum_{j\in S_i\cap\{1+M,\dots,N\}} b_{ij}(u_h)(u_j-u_i)(z_i-z_j). \end{aligned} \tag{2.31}$$

Now, consider any $v\in\mathcal{P}_1(\overline{\Omega})$. Then, the linearity-preservation property of the limiters implies the following:

$$\forall i\in\{1,\dots,M\}\quad \forall j\in S_i\quad b_{ij}(v)=0. \tag{2.32}$$

Let us introduce an auxiliary quantity Λ_{ij} by the following formula:

$$\Lambda_{ij} := b_{ij}(u_h)(u_j-u_i) = b_{ij}(u_h)(u_j-u_i) - b_{ij}(v)\left(v(x_j)-v(x_i)\right).$$

Using Definition 19 and (2.29), we obtain

$$\begin{aligned} |\Lambda_{ij}| &\le L_{ij}\max\{|u_h-v|_i\,,|u_h-v|_j\} \\ &\le C_{\mathrm{ed}}L_{ij}h_{E_{ij}}^{\frac{2-d}{2}}\,|u_h-v|_{1,2,\widetilde{\omega}_{ij}}\,, \end{aligned}$$

provided that $j\in S_i\cap\{1,\dots,M\}$, and

$$|\Lambda_{ij}| \le L_{ij}\,|u_h-v|_i \le C_{\mathrm{ed}}L_{ij}h_{E_{ij}}^{\frac{2-d}{2}}\,|u_h-v|_{1,2,\widetilde{\omega}_{ij}}\,,$$

provided that $j\in S_i\cap\{1+M,\dots,N\}$. In both cases, we have

$$|\Lambda_{ij}| \le C_{\mathrm{ed}}L_{ij}h_{E_{ij}}^{\frac{2-d}{2}}\,|u_h-v|_{1,2,\widetilde{\omega}_{ij}}\,.$$

The next step is to apply the triangle inequality

$$|u_h-v|_{1,2,\widetilde{\omega}_{ij}} \le |u_h-u|_{1,2,\widetilde{\omega}_{ij}} + |u-v|_{1,2,\widetilde{\omega}_{ij}}\,.$$

Since v is arbitrary, it follows from (2.18) that

$$\begin{aligned} |\Lambda_{ij}| &\le C_{\mathrm{ed}}L_{ij}h_{E_{ij}}^{\frac{2-d}{2}}\left(|u_h-u|_{1,2,\widetilde{\omega}_{ij}} + \inf_{v\in\mathcal{P}_1(\overline{\Omega})}|u-v|_{1,2,\widetilde{\omega}_{ij}}\right) \\ &\le C_{\mathrm{ed}}L_{ij}h_{E_{ij}}^{\frac{2-d}{2}}\left(|u_h-u|_{1,2,\widetilde{\omega}_{ij}} + C_{\mathrm{lp}}h\,|u|_{2,2,\widetilde{\omega}_{ij}}\right) \end{aligned} \tag{2.33}$$

Combining the above relations (2.28), (2.31), and (2.33) yields the estimate

$$\begin{aligned} |b_h(u_h;u_h,z_h)| &\le \sum_{i=1}^{M}\sum_{j\in S_i}|\Lambda_{ij}|\,|z_i-z_j| \\ &\le \sum_{i=1}^{M}\sum_{j\in S_i} C_{\mathrm{ed}}h_{E_{ij}}^{\frac{2-d}{2}}\,|\Lambda_{ij}|\,|z_h|_{1,2,\widetilde{\omega}_{ij}} \\ &\le \sum_{i=1}^{M}\sum_{j\in S_i} C_{\mathrm{ed}}^2L_{ij}h_{E_{ij}}^{2-d}\,|u-u_h|_{1,2,\widetilde{\omega}_{ij}}\,|z_h|_{1,2,\widetilde{\omega}_{ij}} \\ &\quad+\sum_{i=1}^{M}\sum_{j\in S_i} C_{\mathrm{lp}}C_{\mathrm{ed}}^2L_{ij}h_{E_{ij}}^{2-d}h\,|u|_{2,2,\widetilde{\omega}_{ij}}\,|z_h|_{1,2,\widetilde{\omega}_{ij}}\,. \end{aligned}$$

Using the notation (2.19), the Cauchy–Schwarz inequality and (1.2) we obtain

$$|b_h(u_h;u_h,z_h)| \le C_{\mathrm{stab}} L_h \Big(\varepsilon^{-\frac{1}{2}} \left\| \sqrt{\mathbb{E}} \nabla (u-u_h) \right\|_{L^2(\Omega)} + h\,|u|_{2,2,\Omega} \Big)\, |z_h|_{1,2,\Omega}\,.$$

This finishes the proof.

□

The following lemma provides an estimate for the second term on the right-hand side of (2.26).

Lemma 28. *Under the assumptions of Lemma 27, we have the estimate*

$$\frac{b_h(u_h;u_h,I_h w_\psi)}{\|\psi\|_{L^2(\Omega)}} \le C_{\mathrm{stab}} C_R L_h \left(C_{\mathrm{int}} h + 1\right) \left(\varepsilon^{-\frac{1}{2}} \|u-u_h\|_{\mathrm{en}} + h\,|u|_{2,2,\Omega} \right). \quad (2.34)$$

Proof. From the estimate (2.30), we have

$$\frac{b_h(u_h;u_h,I_h w_\psi)}{\|\psi\|_{L^2(\Omega)}} \le C_{\mathrm{stab}} L_h \left(\varepsilon^{-\frac{1}{2}} \|u-u_h\|_{\mathrm{en}} + h\,|u|_{2,2,\Omega} \right) \frac{|I_h w_\psi|_{1,2,\Omega}}{\|\psi\|_{L^2(\Omega)}}.$$

Now, from the interpolation error estimate (1.11), definition of the norm in $H^2(\Omega)$, and the assumption (2.24), we get

$$\begin{aligned} |I_h w_\psi|_{1,2,\Omega} &\le |w_\psi - I_h w_\psi|_{1,2,\Omega} + |w_\psi|_{1,2,\Omega} \\ &\le \left(C_{\mathrm{int}} h + 1\right) C_R \|\psi\|_{L^2(\Omega)}\,. \end{aligned}$$

Combining both estimates, we obtain (2.34).

□

We summarize our conclusions with the following theorem.

Theorem 29. *Let the assumptions of Lemma 27 be satisfied, $u \in H^2(\Omega) \cap H_0^1(\Omega)$ be the weak solution of (1.1) with homogeneous Dirichlet boundary condition prescribed on the whole boundary, and let $u_h \in W_h$ be a discrete solution of the nonlinear problem (2.3). Then*

$$\begin{aligned} \|u-u_h\|_{L^2(\Omega)} \le\ & C_{\mathrm{stab}} C_R L_h \left(C_{\mathrm{int}} h + 1\right) \left(\varepsilon^{-\frac{1}{2}} \|u-u_h\|_{\mathrm{en}} + h\,|u|_{2,2,\Omega} \right) \\ & + 2 C_R C_{\mathrm{int}} M_h^{\frac{1}{2}} h \,\|u-u_h\|_{\mathrm{en}}\,, \end{aligned} \quad (2.35)$$

where

$$M_h = \|\mathbb{E}\|_{L^\infty(\Omega)} + \|b\|_{L^\infty(\Omega)}^2 \varepsilon^{-1} h^2 + \mu^{-1} \|c\|_{L^\infty(\Omega)}^2 h^2.$$

Proof. This estimate can be obtained directly by combining the estimates (2.26), (2.27), and the Lemmata 27 and 28.

□

We finish this section with a discussion of the estimate (2.35). If we admit the validity of the inequality (2.20) and assume that $\|u - u_h\|_{\text{en}} = \mathcal{O}(h)$ (see Theorem 24), then the estimate (2.35) predicts the convergence rate of order $\mathcal{O}(h^2)$.

A possible drawback is that the products

$$C_R \varepsilon^{-\frac{1}{2}} \quad \text{and} \quad C_R \, \|\mathbb{E}\|_{L^\infty(\Omega)}^{\frac{1}{2}}$$

may be significantly large if the constant ε from (1.2) is too small or $\|\mathbb{E}\|_{1,\infty,\Omega}$ is too large (for details, see the proof of (2.24) in [10], Section 8.3).

3 Limiters Based on Algebraic Fluxes

In this chapter, we briefly introduce two basic limiters based on algebraic fluxes. We present only the construction of these limiters and provide some comments on their properties. More detailed information can be found in [1], Sections 10.3 and 10.4.

Throughout this chapter, we consider a finite element space W_h belonging to a family of Lagrange $\mathcal{P}_1$ or $\mathcal{Q}_1$ finite element spaces, defined on a shape-regular family of triangulations $\mathcal{T}_h$, consisting of simplices, parallelograms or parallelepipeds. Furthermore, as in Section 1.4, we extend the matrix $(a_{ij})_{j=1,\dots,N}^{i=1,\dots,M}$, given by (1.19) to an $N \times N$ matrix as follows:

$$a_{ij} := a_h(\varphi_j, \varphi_i), \quad i, j = 1, \dots, N,$$

and construct the artificial diffusion matrix $(d_{ij})_{i,j=1,\dots,N}$ from this modified matrix.

3.1 Kuzmin Limiter

This limiter was originally proposed (for a steady-state case) in [5], and the first general theoretical analysis was presented in [6].

Let $u_h \in W_h$. For all $i \in \{1, \dots, M\}$, we first compute the following auxiliary quantities:

$$\begin{aligned} P_i^+ &:= \sum_{j \in S_i,\, a_{ji} \le a_{ij}} f_{ij}^+, \\ P_i^- &:= \sum_{j \in S_i,\, a_{ji} \le a_{ij}} f_{ij}^-, \\ Q_i^+ &:= -\sum_{j \in S_i} f_{ij}^-, \\ Q_i^- &:= -\sum_{j \in S_i} f_{ij}^+, \end{aligned}$$

where f_{ij} are the algebraic fluxes given by (1.46), and

$$f_{ij}^+ := \max\{0, f_{ij}\}, \quad f_{ij}^- := \min\{0, f_{ij}\}.$$

Then, we define

$$\begin{aligned} R_i^+ &:= \min\left\{1, \frac{Q_i^+}{P_i^+}\right\}, \\ R_i^- &:= \min\left\{1, \frac{Q_i^-}{P_i^-}\right\} \end{aligned}$$

for all $i \in \{1, \dots, M\}$. If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Moreover, we set

$$R_i^+ := 1, \quad R_i^- := 1$$

for all $i \in \{M+1, \dots, N\}$. Furthermore, we define

$$\widetilde{\alpha}_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}$$

for all $i, j \in \{1, \dots, N\}$. Finally, we define the *Kuzmin limiter* α_{ij} as follows:

$$\forall i, j \in \{1, \dots, N\} \quad \alpha_{ij}(u_h) := \alpha_{ji}(u_h) := \widetilde{\alpha}_{ij} \quad \text{if } a_{ji} \le a_{ij}.$$

Clearly, $\alpha_{ij} \in \langle 0, 1 \rangle$ for all $i, j \in \{1, \dots, N\}$.

It can be proven (see [1], Theorem 10.22) that the nonlinear problem (1.58) based on the limiter described above possesses the DMP properties formulated in Definitions 15 and 16, provided that the following condition holds:

$$\forall i \in \{1, \dots, M\}, \quad \forall j \in \{1, \dots, N\},\, i \neq j \quad \Rightarrow \quad \min\{a_{ij}, a_{ji}\} \le 0. \tag{3.1}$$

It can also be shown that the functions (1.59) associated with the Kuzmin limiter are continuous (see [1], Theorems 10.26 and 10.1). Thus, a discrete solution of the corresponding nonlinear problem exists. Furthermore, this limiter satisfies the Lipschitz condition from Definition 19 (see [1], Corollary 10.30).

Unfortunately, the Kuzmin limiter generally does not possess the linearity-preserving property. Sufficient conditions for this property involve the local symmetry of the simplicial triangulations used. For further details, see [1], at the end of Section 10.3.

3.2 BJK Limiter

Another limiter was proposed in [9]. This limiter is known as the *Barrenechea–John–Knobloch (BJK) limiter*, and its main advantage is that it is linearity preserving for general triangulations.

We start with modifying the matrix $(a_{ij})_{i,j=1,\dots,N}$ by setting

$$a_{ji} := 0 \quad \text{if} \quad a_{ij} < 0, \quad i = 1, \dots, M,\ j = M+1, \dots, N.$$

This modified matrix is used only for the construction of the artificial diffusion matrix $(d_{ij})_{i,j=1,\dots,N}$.

The construction of the BJK limiter proceeds similarly to the construction of the Kuzmin limiter. Let $u_h \in W_h$ have the coordinate representation $U = (u_1, \dots, u_N)^\top$ given by (1.8). First, we compute the following values:

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

Next, for all $i \in \{1, \dots, M\}$, we compute the following auxiliary quantities:

$$\begin{aligned} P_i^+ &:= \sum_{j \in S_i} f_{ij}^+, \\ P_i^- &:= \sum_{j \in S_i} f_{ij}^-, \\ Q_i^+ &:= q_i \left(u_i - u_i^{\max}\right), \\ Q_i^- &:= q_i \left(u_i - u_i^{\min}\right). \end{aligned}$$

Then, we define

$$R_i^+ := \min\left\{1, \frac{\mu_i Q_i^+}{P_i^+}\right\},$$
$$R_i^- := \min\left\{1, \frac{\mu_i Q_i^-}{P_i^-}\right\}$$

for all $i \in \{1, \dots, M\}$, where $\mu_i > 0$ are fixed constants. If P_i^+ or P_i^- vanishes, we set $R_i^+ := 1$ or $R_i^- := 1$, respectively. Moreover, we set

$$R_i^+ := 1, \quad R_i^- := 1$$

for all $i \in \{M+1, \dots, N\}$. Then, we again define

$$\widetilde{\alpha}_{ij} := \begin{cases} R_i^+ & \text{if } f_{ij} > 0, \\ 1 & \text{if } f_{ij} = 0, \\ R_i^- & \text{if } f_{ij} < 0 \end{cases}$$

for all $i, j \in \{1, \dots, N\}$. Finally, we construct the BJK limiter by following symmetrization:

$$\forall i, j \in \{1, \dots, N\} \quad \alpha_{ij}(u_h) := \alpha_{ji}(u_h) := \min\{\widetilde{\alpha}_{ij}, \widetilde{\alpha}_{ji}\}.$$

Note that an appropriate choice of the constants μ_i guarantees that this limiter possesses the linearity-preservation property.

Theorem 30. *Consider any constants μ_i satisfying*

$$\mu_i \geq \frac{\max_{x_j \in \partial\omega_i} \|x_i - x_j\|_2}{\operatorname{dist}(x_i, \partial\omega_i^{\text{conv}})}$$

for all $i = 1, \dots, M$, where ω_i^{conv} denotes the convex hull of ω_i. Then the BJK limiter is linearity preserving.

Proof. See the proof of Theorem 10.42 in [1].

□

Another important property of the BJK limiter is that the resulting nonlinear problem given by (1.58) possesses the DMP properties formulated in Definitions 15 and 16 without any additional assumptions on the matrix $(a_{ij})_{i,j=1,\dots,N}$.

All other properties satisfied by the Kuzmin limiter presented in Section 3.1 also hold for the BJK limiter. For further details, see [1], Section 10.4.

4 Limiters Based on Local Smoothness Indicators

This chapter is dedicated to presenting an alternative approach for introducing artificial diffusion within AFC schemes. The main idea is to measure the local variation of the approximated solution.

This approach, known as *Barrenechea-Burman-Karakatsani (BBK) method* (with the corresponding limiter known as *BBK limiter*), was originally proposed in [7] and slightly modified in [1], Section 10.8.

The BBK method generally does not possess the linearity-preservation property. In a paper [7] is suggested the possibility of enforcing the linearity-preservation property through suitable weighting. However, the described method requires the solution of a certain constrained minimization problem for each interior vertex of a triangulation, which is computationally expensive.

Another possibility was proposed in [17]. This choice of weights is based on so-called generalized barycentric coordinates associated with ω_i. However, this approach requires ω_i to be convex.

In this chapter, we extend the standard analysis of the BBK method to the case with weights and propose, to the best of our knowledge, a new approach to choosing the weights, which leads to linearity-preserving limiters.

Throughout this chapter, we restrict our considerations to Lagrange $\mathcal{P}_1$ finite elements on a triangulation $\mathcal{T}_h$ belonging to a shape-regular family of simplicial triangulations of the domain Ω.. The corresponding finite element space, consisting of piecewise linear functions, will be denoted by W_h, and the standard basis of W_h, denoted by $\{\varphi_i\}_{i=1}^N \subset W_h$, will satisfy condition (1.7).

4.1 Abstract Formulation of the BBK Method

For all $u_h \in W_h$ we introduce the following quantities:

$$\begin{aligned} A_i(u_h) &:= \sum_{j \in S_i} w_{ij} \, (u_i - u_j) \\ B_i(u_h) &:= \sum_{j \in S_i} |w_{ij}| \, |u_i - u_j|, \end{aligned} \tag{4.1}$$

where $i \in \{1, 2, \ldots, M\}$ and $w_{ij} \in \mathbb{R} \setminus \{0\}$ are called weights (the original BBK method can be obtained by setting $w_{ij} := 1$ for all $i \in \{1, 2, \ldots, M\}$ and $j \in S_j$). The general form of the *local smoothness indicator* $\alpha_i : W_h \to \mathbb{R}_0^+$ is then given by formula

$$\begin{aligned} &\forall i \in \{1, 2, \ldots, M\} \quad \alpha_i(u_h) := \begin{cases} \frac{|A_i(u_h)|}{B_i(u_h)} & \text{if } B_i(u_h) \neq 0, \\ 0 & \text{otherwise,} \end{cases} \\ &\forall i \in \{M+1, \ldots, N\} \quad \alpha_i(u_h) := 0. \end{aligned} \tag{4.2}$$

Let $p \in \langle 1, \infty)$ be a fixed parameter, then

$$\forall i, j \in \{1, 2, \ldots, N\} \quad \gamma_{ij}(u_h) := \max\{(\alpha_i(u_h))^p, (\alpha_j(u_h))^p\}. \tag{4.3}$$

Now we define the matrix of artificial diffusion $(b_{ij}(u_h))_{i,j=1}^N$ depending on a discrete solution u_h

$$\begin{aligned} \forall i,j \in \{1,2,\dots,N\}, i \neq j \quad & b_{ij}(u_h) := \gamma_{ij}(u_h) d_{ij} \\ \forall i \in \{1,\dots,N\} \quad & b_{ii}(u_h) := -\sum_{k \neq i} b_{ik}(u_h) \end{aligned} \tag{4.4}$$

By inserting this matrix into nonlinear problem (1.58), we obtain the desired scheme. Note that this matrix has the form of AFC scheme (1.52) with $\alpha_{ij} := 1 - \gamma_{ij}$.

In Section 2.2 we saw that the Lipschitz continuity with respect to local variations (see Definition 19) is a crucial property of the limiters. To prove this property for BBK limiter, we need the following two auxiliary lemmata. Both are adapted from [1], Chapter 10 (see Lemma 10.78 and Lemma 10.28). The proofs are adapted from the same monograph.

Lemma 31. *Let $i \in \{1,2,\dots,M\}$, and let α_i be defined by (4.2), then, for any $u_h, v_h \in W_h$, inequality*

$$|\alpha_i(u_h) - \alpha_i(v_h)| \leq 4 \frac{B_i(u_h - v_h)}{B_i(v_h) + B_i(u_h)} \tag{4.5}$$

holds provided that $B_i(v_h) + B_i(u_h) \neq 0$.

Proof. Let $u_h, v_h \in W_h$. By the triangle inequality, we obtain following estimates:

$$\begin{aligned} \Big| |A_i(u_h)| - |A_i(v_h)| \Big| &\leq |A_i(u_h) - A_i(v_h)| \\ &\leq \sum_{j \in S_i} |w_{ij} \left(u_i - u_j - (v_i - v_j)\right)| \\ &= B_i(u_h - v_h), \end{aligned} \tag{4.6}$$

and

$$|B_i(u_h) - B_i(v_h)| \leq \sum_{j \in S_i} |w_{ij}| \Big| |u_i - u_j| - |v_i - v_j| \Big| = B_i(u_h - v_h). \tag{4.7}$$

Assuming $B_i(u_h) \neq 0$ and $B_i(v_h) \neq 0$, we get

$$\alpha_i(u_h) - \alpha_i(v_h) = \frac{|A_i(u_h)| - |A_i(v_h)|}{B_i(u_h)} + \frac{|A_i(v_h)| \left(B_i(v_h) - B_i(u_h)\right)}{B_i(u_h) B_i(v_h)}. \tag{4.8}$$

By combining the estimates (4.6) and (4.7) with the expression (4.8) and using the inequality $|A_i(v_h)| \leq B_i(v_h)$ we have

$$|\alpha_i(u_h) - \alpha_i(v_h)| \leq 2 \frac{B_i(u_h - v_h)}{B_i(u_h)}.$$

Interchanging u_h and v_h yields

$$|\alpha_i(u_h) - \alpha_i(v_h)| \leq 2 \frac{B_i(u_h - v_h)}{B_i(v_h)}.$$

Combining these results, we obtain

$$|\alpha_i(u_h) - \alpha_i(v_h)| \leq 2B_i(u_h - v_h) \min\left\{\frac{1}{B_i(v_h)}, \frac{1}{B_i(u_h)}\right\}.$$

Finally, applying the standard inequality

$$\forall x, y > 0 \quad \min\left\{\frac{1}{x}, \frac{1}{y}\right\} \leq \frac{2}{x+y},$$

we conclude that

$$|\alpha_i(u_h) - \alpha_i(v_h)| \leq 4\frac{B_i(u_h - v_h)}{B_i(v_h) + B_i(u_h)}.$$

Now, suppose that $B_i(u_h) \neq 0$, but $B_i(v_h) = 0$. In this case $v_j = v_i$ for all $j \in S_i$, thus $A_i(v_h) = 0$. From estimate (4.6), it follows that

$$|\alpha_i(u_h) - \alpha_i(v_h)| = |\alpha_i(u_h)| \leq 4\frac{B_i(u_h - v_h)}{B_i(u_h)} = 4\frac{B_i(u_h - v_h)}{B_i(u_h) + B_i(v_h)}.$$

This completes the proof.

□

Lemma 32. *For any $k \in \mathbb{N}$ and any $a_1, \ldots, a_k, b_1, \ldots, b_k \in \mathbb{R}$, one has*

$$\left|\max_{i=1,\ldots,k} a_i - \max_{i=1,\ldots,k} b_i\right| \leq \max_{i=1,\ldots,k} |a_i - b_i|, \tag{4.9}$$

$$\left|\min_{i=1,\ldots,k} a_i - \min_{i=1,\ldots,k} b_i\right| \leq \max_{i=1,\ldots,k} |a_i - b_i|, \tag{4.10}$$

Proof. See the proof of Lemma 10.28 in [1].

□

The following theorem and its proof are also adapted from [1] (Theorem 10.79).

Theorem 33. *The functions $u_h \in W_h \longmapsto b_{ij}(u_h)(u_j - u_i)$, $i \in \{1, \ldots, M\}$, $j \in S_i$, defined by (4.1)-(4.4) are Lipschitz continuous with respect to local variations with*

$$L_{ij} = |d_{ij}| \left(4p \max\left\{\frac{1}{|w_{ij}|}\sum_{k\in S_i} |w_{ik}|, \frac{1}{|w_{ji}|}\sum_{k\in S_j} |w_{jk}|\right\} + 1\right) \tag{4.11}$$

for $i, j \in \{1, \ldots, M\}$, and

$$L_{ij} = |d_{ij}| \left(\frac{4p}{|w_{ij}|}\sum_{k\in S_i} |w_{ik}| + 1\right) \tag{4.12}$$

for $j \in S_i \cap \{M+1, \ldots, N\}$.

Proof. Consider $i = 1, \dots, M$, $j \in S_i \cap \{1, \dots, M\}$ and $u_h, v_h \in W_h$. Let us denote $r := u_j - u_i$ and $s := v_j - v_i$. First, suppose that $r \geq 0$ and $s \leq 0$, then the following holds:

$$\begin{aligned} |b_{ij}(u_h)r - b_{ij}(v_h)s| &= |d_{ij}|\,|\gamma_{ij}(u_h)r - \gamma_{ij}(v_h)s| \\ &= |d_{ij}|\,|\gamma_{ij}(u_h)\,|r| + \gamma_{ij}(v_h)\,|s|| \\ &\leq |d_{ij}|\,|r - s|\,, \end{aligned}$$

where we used the fact that $\gamma_{ij} \leq 1$. Now, the Lipschitz continuity with respect to local variations is easy consequence of the above inequality. If $s \geq 0$ and $r \leq 0$, then we can proceed analogously.

It remains to prove the case for $r, s > 0$ and $r, s < 0$. First, suppose that $r, s > 0$. Using the inequality (4.9), we obtain

$$\begin{aligned} &|b_{ij}(u_h)(u_j - u_i) - b_{ij}(v_h)(v_j - v_i)| \\ &\quad = |d_{ij}|\,|\gamma_{ij}(u_h)r - \gamma_{ij}(v_h)s| \\ &\quad = |d_{ij}|\,\Big|\max\{(\alpha_i(u_h))^p\,, (\alpha_j(u_h))^p\}r - \max\{(\alpha_i(v_h))^p\,, (\alpha_j(v_h))^p\}s\Big| \\ &\quad \leq |d_{ij}|\max\Big\{\,|(\alpha_i(u_h))^p\,r - (\alpha_i(v_h))^p\,s|\,, |(\alpha_j(u_h))^p\,r - (\alpha_j(v_h))^p\,s|\,\Big\}. \end{aligned} \tag{4.13}$$

Now, consider a function $g(z) = z^p$. Let $x, y \in \langle 0, 1\rangle$ be arbitrary, then from the mean value theorem follows the existence of $z \in (0, 1)$ such that

$$|x^p - y^p| \leq g'(z)\,|x - y| = pz^{p-1}\,|x - y| \leq p\,|x - y|\,.$$

By combining this inequality with the triangle inequality, we obtain

$$\begin{aligned} &|(\alpha_i(u_h))^p\,r - (\alpha_i(v_h))^p\,s| \\ &\quad = |\,(\alpha_i(u_h))^p\,r - (\alpha_i(v_h))^p\,s + (\alpha_i(v_h))^p\,r - (\alpha_i(v_h))^p\,r| \\ &\quad \leq p\,|\alpha_i(u_h) - \alpha_i(v_h)|\,r + (\alpha_i(v_h))^p\,|r - s| \\ &\quad \leq p\,|\alpha_i(u_h) - \alpha_i(v_h)|\,r + |r - s|\,. \end{aligned}$$

The first term can be estimated by the inequality (4.5)

$$\begin{aligned} p\,|\alpha_i(u_h) - \alpha_i(v_h)|\,r &\leq 4p\,(u_j - u_i)\,\frac{\sum_{k\in S_i}|w_{ik}|\,\Big|\,(u_k - u_i) - (v_k - v_i)\,\Big|}{\sum_{k\in S_i}|w_{ik}|\,\Big(|u_k - u_i| + |v_k - v_i|\Big)} \\ &\leq \frac{4p}{|w_{ij}|}\sum_{k\in S_i}|w_{ik}|\,\Big|\,(u_k - u_i) - (v_k - v_i)\,\Big| \\ &\leq \left(\frac{4p}{|w_{ij}|}\sum_{k\in S_i}|w_{ik}|\right)|u_h - v_h|_i\,. \end{aligned}$$

Altogether, we obtain

$$|(\alpha_i(u_h))^p\,r - (\alpha_i(v_h))^p\,s| \leq \left(\frac{4p}{|w_{ij}|}\sum_{k\in S_i}|w_{ik}| + 1\right)|u_h - v_h|_i\,.$$

Similarly, we have

$$|(\alpha_j(u_h))^p\,r - (\alpha_j(v_h))^p\,s| \leq \left(\frac{4p}{|w_{ji}|}\sum_{k\in S_j}|w_{jk}| + 1\right)|u_h - v_h|_j$$

Substituting these estimates into (4.13), we conclude that

$$|b_{ij}(u_h)(u_j - u_i) - b_{ij}(v_h)(v_j - v_i)| \le L_{ij} \max\{|u_h - v_h|_i, |u_h - v_h|_j\},$$

where the constants L_{ij} are given by the formula (4.11). The case $r, s < 0$ can be treated analogously (the inequality (4.10) must be used instead of (4.9)).

Now, assuming that $j \in \{M+1, \dots, N\}$, we can proceed similarly and arrive at the following estimate:

$$|b_{ij}(u_h)(u_j - u_i) - b_{ij}(v_h)(v_j - v_i)| \le L_{ij} |u_h - v_h|_i,$$

where the constants L_{ij} are given by the formula (4.12).

□

For completeness, let us briefly discuss the validity of (2.20) for the BBK method. It follows from Theorem 33 that the Lipschitz constants L_{ij} are proportional to $|d_{ij}|$. Thus, the estimate (2.20) can be proven, provided that the triangulation $\mathcal{T}_h$ satisfies the XZ-condition. For details of this proof, see [1], Remarks 10.31 and 10.32.

It is expected that Lipschitz continuity with respect to local variations ensures the continuity of the functions (1.59) and, consequently, guarantees the existence of a solution to the corresponding nonlinear problem. For details, see [1], Section 10.3.

Theorem 34. *The nonlinear problem (1.58) with matrix $(b_{ij}(u_h))_{i,j=1}^N$ given by (4.1)-(4.4) has a solution.*

Proof. This assertion follows immediately from Theorem 15. We only need to verify the continuity of the functions in (1.59).

Consider a sequence $\{u_h^k\}_{k=1}^\infty \subset W_h$ and an element $u_h \in W_h$ such that $u_h^k \to u_h$ as $k \to \infty$. Furthermore, let $\{U^k\}_{k=1}^\infty \subset \mathbb{R}^N$, with $U^k = (u_1^{(k)}, \dots, u_N^{(k)})^\top$ for all $k \in \mathbb{N}$, and let $U = (u_1, \dots, u_N)^\top \in \mathbb{R}^N$ be such that

$$\forall k \in \mathbb{N} \quad u_h^k = \sum_{l=1}^N u_l^{(k)} \varphi_l,$$

and

$$u_h = \sum_{l=1}^N u_l \varphi_l.$$

Then, we have $\lim_{k\to\infty} U^k = U$, which implies $\lim_{k\to\infty} u_i^{(k)} = u_i$ for all $i = 1, \dots, N$. Therefore, for any $l \in \{1, \dots, M\}$, we obtain

$$\begin{aligned}\lim_{k\to\infty} \left|u_h^k - u_h\right|_l &= \lim_{k\to\infty} \max_{j\in S_l} \left|(u_j^{(k)} - u_j) - (u_l^{(k)} - u_l)\right| \\ &\le \lim_{k\to\infty} \max_{j\in S_l} \left|u_j^{(k)} - u_j\right| + \lim_{k\to\infty} \left|u_l^{(k)} - u_l\right| = 0.\end{aligned}$$

It follows from Theorem 33 that

$$\lim_{k\to\infty} b_{lj}(u_h^k)(u_j^{(k)} - u_l^{(k)}) = b_{lj}(u_h)(u_j - u_l),$$

for any $l \in \{1,\dots,M\}$ and any $j \in \{1,\dots,N\} \cap S_l$. Thus, the validity of condition (1.59) follows as a direct consequence.

□

Next, we discuss the validity of discrete maximum principles for the BBK limiter. All conditions of Theorems 16 and 17, except for possession of the properties formulated in Definitions 15 and 16, are clearly satisfied.

Theorem 35. *The nonlinear problem (1.58) with the matrix* $(b_{ij}(u_h))_{i,j=1}^N$ *given by (4.1)-(4.4) possesses the DMP properties formulated in Definitions 15 and 16 provided that all weights* w_{ij} *have the same sign.*

Proof. The proof of this statement is adapted from [1], Section 10.8 (Theorem 10.75).

Since the validity of the matrix-based DMP property for nonstrict extrema implies its validity for strict extrema, it is sufficient to prove the former.

Consider any $i \in \{1,\dots,M\}$ and any $U = (u_1,\dots,u_N)^\top \in \mathbb{R}^N$ such that $u_i \le u_k$ for all $k \in S_i$. Suppose that $u_i \ne u_j$ for some $j \in S_i$, and assume that $w_{ik} > 0$ for all $k \in S_i$.

It is sufficient to show that $\alpha_i(u_h) = 1$, where $u_h \in W_h$ denotes the function corresponding to U via (1.8). Then we have

$$a_{ij} + b_{ij}(u_h) = a_{ij} + d_{ij} = a_{ij} - \max\{a_{ij}, 0, a_{ji}\} \le 0.$$

It follows from (4.1) that

$$\begin{aligned} A_i(u_h) &= \sum_{k\in S_i} w_{ik}(u_i - u_k) \\ &= w_{ij}(u_i - u_j) + \sum_{k\in S_i\setminus\{j\}} w_{ik}(u_i - u_k) \\ &= -\sum_{k\in S_i\setminus\{j\}} |w_{ik}|\,|u_k - u_i| - |w_{ij}|\,|u_j - u_i|\,. \end{aligned}$$

It is easy to see that

$$A_i(u_h) = -B_i(u_h) \le -|w_{ij}|\,|u_j - u_i| < 0.$$

This yields the desired equality $\alpha_i(u_h) = 1$.

In other cases, a similar approach can be applied.

□

Finally, we note that the BBK limiter is invariant under scaling of the weights by a nonzero constant.

Lemma 36. *Consider local smoothness indicators* α_i, $i \in \{1,2,\dots,N\}$, *based on weights* w_{kl}, $k \in \{1,2,\dots,M\}$, $l \in S_k$. *Let* $q \in \mathbb{R}\setminus\{0\}$ *be arbitrary and* $\overline{w}_{kl} := q w_{kl}$ *for all* $k \in \{1,2,\dots,M\}$ *and* $l \in S_k$. *Consider local smoothness indicators* $\overline{\alpha}_i$ *constructed using weights* $\overline{w}_{kl}$. *It follows that*

$$\forall i \in \{1,2,\dots,N\} \quad \forall v_h \in W_h \quad \alpha_i(v_h) = \overline{\alpha}_i(v_h).$$

Proof. From relation (4.2), it is clear that it is sufficient to consider $i \in \{1, \dots, M\}$. Furthermore, it is evident that the denominator of the fraction in (4.2) is nonzero if and only if the same denominator is nonzero for scaled weights. The statement thus follows by factoring out $|q|$ from both the numerator and the denominator of this fraction, and subsequently simplifying the expression.

□

4.2 BBK Method Based on Integral Smoothness Indicators

4.2.1 Motivation

As mentioned above, AFC schemes are activated in subdomains where spurious oscillations occur, i.e., where the DMP may be violated. Roughly speaking, we seek a solution $u_h \in W_h$ that minimizes $|u_h|_{1,2,\omega_i}$, for $i = 1, \dots, M$, as much as possible without smearing layers.

Consider the following motivating example (see [2], Chapter 8). Let $i \in \{1, \dots, M\}$ be arbitrary and define the local function space

$$W_h(\omega_i) := \{v_h \in \mathcal{C}(\omega_i)\,;\, v_h|_T \in \mathcal{P}_1(T)\ \forall T \subset \omega_i\}\,.$$

Choose $u_0 \in H^{1/2}(\partial\omega_i)$ such that there exists an extension $\widetilde{u}_0 \in W_h(\omega_i)$ of this function to the whole domain ω_i, i.e., $u_0 = \widetilde{u}_0|_{\partial\omega_i}$. Within this setting, we solve the following optimization problem:

$$\min_{v \in W_h(\omega_i),\, v|_{\partial\omega_i} = u_0} |v|^2_{1,2,\omega_i}.$$

This problem requires computing the Gâteaux derivative of the functional. As a result, we obtain the following necessary condition (Euler–Lagrange equation) for the minimizer v:

$$\forall \eta \in W_h(\omega_i),\ \eta|_{\partial\omega_i} = 0 \quad \langle D|v|^2_{1,2,\omega_i}, \eta\rangle_{(W_h(\omega_i))^*, W_h(\omega_i)} = 2\int_{\omega_i} \nabla v \cdot \nabla\eta \,\mathrm{d}x = 0.$$

Due to the construction of the space $W_h(\omega_i)$, this condition is equivalent to the following equation:

$$\int_{\omega_i} \nabla v \cdot \nabla\varphi_i \,\mathrm{d}x = 0. \tag{4.14}$$

This idea has an important consequence. If u_0 is the trace of a linear function $\widetilde{u}_0$ in ω_i, then $v = \widetilde{u}_0$ is the minimizer. Indeed, if $v \in \mathcal{P}_1(\omega_i)$, then by the Gauss integral theorem it follows that

$$\begin{aligned}\int_{\omega_i} \nabla v \cdot \nabla\varphi_i \,\mathrm{d}x &= \int_{\partial\omega_i} \varphi_i \nabla v \cdot n \,\mathrm{d}S - \int_{\omega_i} \varphi_i \Delta v \,\mathrm{d}x \\ &= -\int_{\omega_i} \varphi_i \Delta v \,\mathrm{d}x \\ &= 0,\end{aligned}$$

where n denotes the unit outward normal vector to $\partial\omega_i$. As we shall see later, this leads to the linearity-preservation property.

These considerations motivate the use of functionals that appear on the left-hand side of (4.14) as components of local smoothness indicators.

4.2.2 General Form of Integral Based BBK Limiter

Let $u_h \in W_h$ be arbitrary. First, we introduce auxiliary functionals

$$\begin{aligned} A_i(u_h) &:= \int_{\omega_i} \nabla u_h \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x, \\ B_i(u_h) &:= \sum_{j \in S_i} |A_i(\varphi_j)| \, |u_i - u_j|, \end{aligned} \tag{4.15}$$

where $i \in \{1, 2, \ldots, M\}$, and $\widetilde{\varphi}_i \in H_0^1(\omega_i)$ is a suitable piecewise polynomial test function such that $\mathrm{supp}\widetilde{\varphi}_i \subset \omega_i$ (a specific choice of this function will be discussed later).

Now, we show that this particular choice of the functionals A_i and B_i leads to the BBK limiter with special weights.

Lemma 37. *Let the functional A_i be given by (4.15). Then, for all $u_h \in W_h$, we have*

$$A_i(u_h) = \sum_{k \in S_i} \left(\int_{\omega_i} \nabla \varphi_k \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x \right) (u_k - u_i) \,.$$

Proof. Let $u_h \in W_h$ and $i \in \{1, \ldots, M\}$. Using property (1.9) and the identification given by (1.8), we derive the following identity:

$$u_h|_{\omega_i} = \sum_{k \in S_i \cup \{i\}} u_k \varphi_k|_{\omega_i}.$$

By differentiating (1.12), we obtain

$$\sum_{k \in S_i \cup \{i\}} \nabla \varphi_k|_{\omega_i} = 0,$$

and therefore,

$$\nabla \varphi_i|_{\omega_i} = - \sum_{k \in S_i} \nabla \varphi_k|_{\omega_i}.$$

Altogether, we obtain

$$\nabla u_h|_{\omega_i} = \sum_{k \in S_i} (u_k - u_i) \nabla \varphi_k|_{\omega_i}.$$

Inserting this expression into the functional A_i from (4.15) yields

$$A_i(u_h) = \sum_{k \in S_i} \left(\int_{\omega_i} \nabla \varphi_k \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x \right) (u_k - u_i) \,.$$

□

Let the test functions $\widetilde{\varphi}_i \in H_0^1(\omega_i)$ be chosen in such a way that the following condition holds:

$$\forall i \in \{1, \ldots, M\} \quad \forall j \in S_i \quad \int_{\omega_i} \nabla \varphi_j \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x < 0. \tag{4.16}$$

The nonlinear problem (1.58) with the matrix $(b_{ij}(u_h))_{i,j=1}^N$, constructed using (4.2)–(4.4) and based on the functionals A_i and B_i given by (4.15), can be considered a special case of the BBK method with weights

$$w_{ij} := -A_i(\varphi_j). \tag{4.17}$$

Note that under condition (4.16), the weights w_{ij} have the same sign. Therefore, the validity of DMPs can also be proven for this limiter.

Moreover, as already mentioned at the end of Subsection 4.2.1, the resulting BBK limiter possesses the linearity-preservation property.

Lemma 38. *Let α_i, for all $i \in \{1, \dots, M\}$, be given by (4.2) with A_i and B_i given by (4.15), then following holds:*

$$\forall i \in \{1, \dots, M\} \quad \forall v_h \in W_h, v_h|_{\omega_i} \in \mathcal{P}_1(\omega_i) \quad \alpha_i(v_h) = 0. \tag{4.18}$$

Proof. Let $i \in \{1, \dots, M\}$ and $v_h \in W_h$ be such that $v_h|_{\omega_i} \in \mathcal{P}_1(\omega_i)$. According to the construction of α_i, it suffices to show that $A_i(v_h) = 0$. Since $v_h|_{\omega_i} \in \mathcal{P}_1(\omega_i) \subset \mathcal{C}^2(\omega_i)$ and $\Delta v_h = 0$, we can apply the Gauss integral theorem to obtain

$$\begin{aligned} A_i(v_h) &= \int_{\omega_i} \nabla v_h \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x = \int_{\partial\omega_i} \widetilde{\varphi}_i \nabla v_h \cdot n \,\mathrm{d}S - \int_{\omega_i} \widetilde{\varphi}_i \Delta v_h \,\mathrm{d}x \\ &= -\int_{\omega_i} \widetilde{\varphi}_i \Delta v_h \,\mathrm{d}x = 0, \end{aligned}$$

where n denotes the unit outward normal vector to $\partial\omega_i$.

□

The validity of the linearity-preservation property on more general triangulations is the main advantage of this method compared to the standard BBK method. Consequently, the general theory of Sections 2.2 and 2.3 can be applied to this method. Therefore, this method may show a better convergence behavior.

4.2.3 Construction of the Test Function $\widetilde{\varphi}_i$

This subsection is devoted to a discussion of the proper choice of the test function $\widetilde{\varphi}_i$, which appears in the definition (4.15) of the functionals A_i and B_i. As already mentioned, this function must be chosen to satisfy the following key assumption:

$$\forall i \in \{1, \dots, M\} \quad \forall j \in S_i \quad \int_{\omega_i} \nabla \varphi_j \cdot \nabla \widetilde{\varphi}_i \,\mathrm{d}x < 0. \tag{4.19}$$

Lemma 39. *Condition (4.19) holds (irrespective of the dimension d) provided that $\mathcal{T}_h$ is an acute triangulation and $\widetilde{\varphi}_i := \varphi_i$.*

Proof. See the proof of Theorem 6.8 in [1]. □

Next, we turn our attention to the general triangulation $\mathcal{T}_h$ belonging to a shape-regular family of simplicial triangulations of the domain Ω. For notation, see Sections 1.2.1 and 1.2.2. Let $i \in \{1, \dots, M\}$. Consider a ball $B \subset \overline{B} \subset \mathrm{Int}(\omega_i)$

with the center located at x_i and a radius equal to r_i. A natural choice for the function $\widetilde{\varphi}_i$ is the following:

$$\widetilde{\varphi}_i(x) = \left(1 - \frac{\|x - x_i\|_2^2}{r_i^2}\right) \chi_B(x), \tag{4.20}$$

where χ_B denotes the characteristic function of the ball B. It can be easily verified that such radially symmetric function $\widetilde{\varphi}_i$ belongs to the space $H_0^1(\omega_i)$ with

$$\nabla \widetilde{\varphi}_i(x) = \frac{2}{r_i^2}(x_i - x) \tag{4.21}$$

for all $x \in B$, and $\widetilde{\varphi}_i(x_i) = 1$. It is easy to see that for all $x \in B$, there exists a constant $k \in (0, \infty)$ such that $\nabla \widetilde{\varphi}_i(x) = -k\, n|_{\partial B}(q)$, where $n|_{\partial B}(q)$ denotes the unit outward normal vector to ∂B at the intersection point q of ∂B and the ray determined by initial point x_i and passing through x.

If $j \in S_i$, then there exist finitely many elements $T_1, \ldots, T_P \in \mathcal{T}_h$ that contain both vertices x_i and x_j. Inserting (4.21) and (1.13) into the integral in (4.19), we obtain

$$\int_{\omega_i} \nabla \varphi_j \cdot \nabla \widetilde{\varphi}_i \, \mathrm{d}x = \frac{2}{r_i^2} \sum_{l=1}^{P} \frac{\left|F_j^{T_l}\right|}{d\,|T_l|} \int_{T_l \cap B} \|x - x_i\|_2 \cos(\zeta_l(x)) \, \mathrm{d}x, \tag{4.22}$$

where $\zeta_l(x)$ denotes the angle between the vectors $x - x_i$ and $n_j^{T_l}$. If $x \in \mathrm{Int}\,(\omega_{ij}) \cap B$, then the angles $\zeta_l(x)$ are larger than $\frac{\pi}{2}$ for all $l \in \{1, \ldots, P\}$ (see Figure 4.1 for the two-dimensional case). Thus, the right-hand side of (4.22) is less than zero.

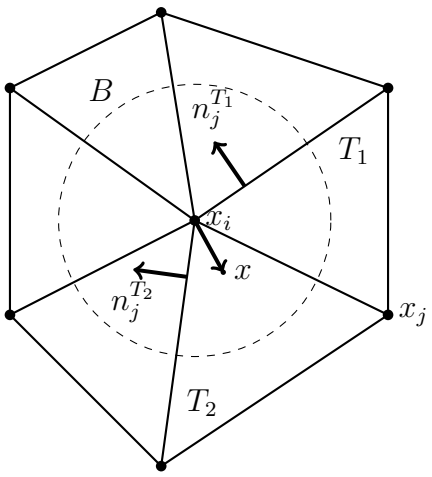


Figure 4.1 Illustration to the construction of radially symmetric $\widetilde{\varphi}_i$ in two-dimensional case.

We now present a practically applicable formula for computing the weights based on the test function (4.20) for the two-dimensional problem.

Theorem 40. *Let $d = 2$, and let, for all $i \in \{1, \ldots, M\}$, the function $\widetilde{\varphi}_i$ be given by formula (4.20). Then, the corresponding weights w_{ij}, $j \in S_i$, are given by formula*

$$w_{ij} = \frac{1}{|E_{ij}|} \left(\frac{1 - t_{ij} \cdot t_{ik}}{t_{ij} \cdot n_j^T} + \frac{1 - t_{ij} \cdot t_{il}}{t_{ij} \cdot n_j^{\widetilde{T}}} \right) \tag{4.23}$$

where $T, \widetilde{T} \in \mathcal{T}_h$ are such that $E_{ij} = T \cap \widetilde{T}$, and $k, l \in S_i$ are the indices of the remaining vertices $x_k \in T$, $x_l \in \widetilde{T}$ (see Figure 4.2).

Proof. Let $i \in \{1, \dots, M\}$ and $j \in S_i$ be arbitrary. Formula (4.22) reduces to the following form in two dimensions:

$$\begin{aligned} A_i(\varphi_j) = \int_{\omega_i} \nabla\varphi_j \cdot \nabla\widetilde{\varphi}_i \,\mathrm{d}x = \frac{1}{r_i^2} \frac{\left|F_j^T\right|}{|T|} \int_{T\cap B} (x - x_i) \cdot n_j^T \,\mathrm{d}x \\ + \frac{1}{r_i^2} \frac{\left|F_j^{\widetilde{T}}\right|}{\left|\widetilde{T}\right|} \int_{\widetilde{T}\cap B} (x - x_i) \cdot n_j^{\widetilde{T}} \,\mathrm{d}x. \end{aligned} \tag{4.24}$$

The entire situation is illustrated in Figure 4.2.

First, we define a Cartesian coordinate system (x^1, x^2) with its origin at x_i and the x^1-axis parallel to the edge E_{ij}. To evaluate the integrals in (4.24), we use the following standard polar substitution:

$$\begin{aligned} x^1(r, \theta) &= \Phi_1(r, \theta) = r\cos(\theta), \\ x^2(r, \theta) &= \Phi_2(r, \theta) = r\sin(\theta), \end{aligned}$$

where $r \in (0, \infty)$ and $\theta \in \langle -\pi, \pi)$. Straightforward calculations yield

$$|\det J_\Phi| = r > 0,$$

where J_Φ denotes the Jacobian matrix of the mapping Φ.

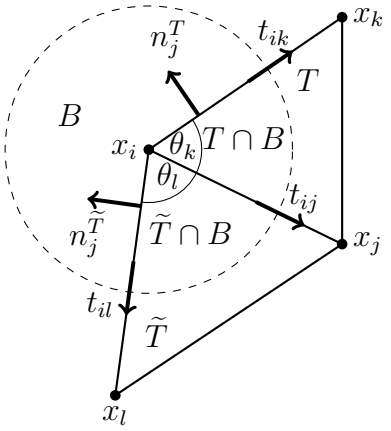


Figure 4.2 Illustration to the proof of Theorem 40.

By applying the Substitution theorem and Fubini's theorem, we can write

$$\int_{T\cap B} (x - x_i) \cdot n_j^T \,\mathrm{d}x = \int_0^{r_i} \int_0^{\theta_k} r^2(\cos\theta, \sin\theta) \cdot n_j^T \,\mathrm{d}\theta \,\mathrm{d}r, \tag{4.25}$$

$$\int_{\widetilde{T}\cap B} (x - x_i) \cdot n_j^{\widetilde{T}} \,\mathrm{d}x = \int_0^{r_i} \int_0^{\theta_l} r^2(\cos\theta, \sin\theta) \cdot n_j^{\widetilde{T}} \,\mathrm{d}\theta \,\mathrm{d}r, \tag{4.26}$$

where $\theta_k, \theta_l \in (0, \pi)$ are such that

$$t_{ik} = (\cos\theta_k, \sin\theta_k)^\top, \quad t_{il} = (\cos\theta_l, -\sin\theta_l)^\top.$$

Then, the following identities hold:

$$n_j^T = (-\sin\theta_k, \cos\theta_k)^\top, \quad n_j^{\widetilde{T}} = (-\sin\theta_l, -\cos\theta_l)^\top.$$

Let us now compute the integral on the right-hand side of (4.25)

$$\begin{aligned}\int_0^{r_i}\int_0^{\theta_k} r^2(\cos\theta,\sin\theta)\cdot n_j^T \,\mathrm{d}\theta\,\mathrm{d}r &= \frac{r_i^3}{3}\int_0^{\theta_k}(\cos\theta,\sin\theta)\cdot n_j^T\,\mathrm{d}\theta \\ &= \frac{r_i^3}{3}\int_0^{\theta_k}(\sin\theta\cos\theta_k - \cos\theta\sin\theta_k)\,\mathrm{d}\theta \\ &= \frac{r_i^3}{3}\int_0^{\theta_k}\sin(\theta-\theta_k)\,\mathrm{d}\theta = \frac{r_i^3}{3}(\cos\theta_k - 1) \\ &= \frac{r_i^3}{3}(t_{ij}\cdot t_{ik} - 1),\end{aligned} \tag{4.27}$$

where we used the standard sine difference formula in the third equality and the definition of the scalar product at the end. The integral on the right-hand side of (4.26) can be computed analogously

$$\int_0^{r_i}\int_0^{\theta_l} r^2(\cos\theta,\sin\theta)\cdot n_j^{\widetilde{T}}\,\mathrm{d}\theta\,\mathrm{d}r = \frac{r_i^3}{3}(t_{ij}\cdot t_{il} - 1). \tag{4.28}$$

Furthermore, using the standard expression for the area of a triangle, we obtain

$$\frac{\left|F_j^T\right|}{|T|} = \frac{2}{|E_{ij}|\sin\theta_k} = -\frac{2}{|E_{ij}|\ t_{ij}\cdot n_j^T}, \tag{4.29}$$

and

$$\frac{\left|F_j^{\widetilde{T}}\right|}{\left|\widetilde{T}\right|} = -\frac{2}{|E_{ij}|\ t_{ij}\cdot n_j^{\widetilde{T}}}. \tag{4.30}$$

Finally, by inserting equations (4.27), (4.28), (4.29), and (4.30) into (4.24), we obtain the following formula:

$$A_i(\varphi_j) = -\frac{2r_i}{3\,|E_{ij}|}\left(\frac{t_{ij}\cdot t_{ik}-1}{t_{ij}\cdot n_j^T} + \frac{t_{ij}\cdot t_{il}-1}{t_{ij}\cdot n_j^{\widetilde{T}}}\right).$$

By virtue of (4.17) and Lemma 36, we can obtain the desired formula (4.23).

□

It is important to note that the weights w_{ij} do not depend on the choice of the radius r_i.

4.2.4 Alternative Construction of the Test Function $\widetilde{\varphi}_i$ in Two Dimensions

In this subsection, we will describe an alternative approach to the construction of the test functions $\widetilde{\varphi}_i$. This approach is based on a suitable discrete counterpart of the method presented in Subsection 4.2.3. Throughout this subsection, we will also use the notation introduced in Subsections 1.2.1 and 1.2.2.

Let us suppose that $d = 2$. First, we describe the construction of supports of the functions $\widetilde{\varphi}_i$. Consider any $i \in \{1, \dots, M\}$. Then there exist finitely many elements $T_1, \dots, T_P \in \mathcal{T}_h$ such that

$$\bigcup_{l=1}^{P} T_l = \omega_i.$$

Let $B_i \subset \omega_i$ be an open disc with radius $r_i > 0$, with the center located at x_i, and such that $B_i \subset \overline{B_i} \subset \mathrm{Int}(\omega_i)$. Next, we construct a closed polygon D_i with P vertices $y_1, \dots, y_P$ such that B_i is inscribed in D_i, $y_l \in \mathrm{Int}(T_l)$ for $l = 1, \dots, P$, and E_{ij} is perpendicular to ∂D_i for any $j \in S_i$. This polygon exists for sufficiently small r_i (see Figure 4.3).

For the construction of the function $\widetilde{\varphi}_i$, we introduce some additional notation. First, it is clear from the above description that for any $k \in S_i$, the edge E_{ik} is perpendicular to some edge of D_i. We denote this edge of D_i by $\mathcal{E}_k$. This edge, together with the vertex x_i, form a triangle which we denote by D_{ik}. These triangles are depicted in Figure 4.3 using dashed lines. Clearly,

$$D_i = \bigcup_{k \in S_i} D_{ik}.$$

Now, the function $\widetilde{\varphi}_i$ is uniquely determined by the following conditions:

1. $\widetilde{\varphi}_i \in \mathcal{C}(\omega_i)$,
2. $\forall l \in S_i \quad \widetilde{\varphi}_i|_{D_{il}} \in \mathcal{P}_1(D_{il})$,
3. $\widetilde{\varphi}_i(x_i) = 1$,
4. $\forall l \in \{1, \dots, P\} \quad \widetilde{\varphi}_i(y_l) = 0$,
5. $\widetilde{\varphi}_i = 0$ in $\omega_i \setminus D_i$.

Note that $\mathrm{supp}(\widetilde{\varphi}_i) = D_i$.

Let $j \in S_i$ be fixed. We find elements $T, \widetilde{T} \in \mathcal{T}_h$ such that $E_{ij} = T \cap \widetilde{T}$ (or, equivalently, $T \cup \widetilde{T} = \omega_{ij}$). Then, there exist $o, s \in S_i$ such that x_o and x_s are the remaining vertices of the triangles $\widetilde{T}$ and T, respectively.

Let us denote the vertices of D_i lying in $\widetilde{T}$ and T by $\widetilde{y}$ and y, respectively. According to the construction above, the edge E_{ij} divides the triangle D_{ij} with vertices x_i, y and $\widetilde{y}$, into two subtriangles, denoted by $\mathcal{A}_1 \subset T$ and $\mathcal{A}_2 \subset \widetilde{T}$.

Furthermore, we define the following auxiliary sets:

$$\mathcal{A}_3 := D_{is} \cap T,$$
$$\mathcal{A}_4 := D_{io} \cap \widetilde{T}.$$

Clearly,

$$\bigcup_{l=1}^{4} \mathcal{A}_l = \omega_{ij} \cap D_i, \tag{4.31}$$

and

$$\forall k \neq l \quad \mathrm{Int}(\mathcal{A}_k) \cap \mathrm{Int}(\mathcal{A}_l) = \emptyset. \tag{4.32}$$

For a better understanding of this notation, see Figure 4.4.

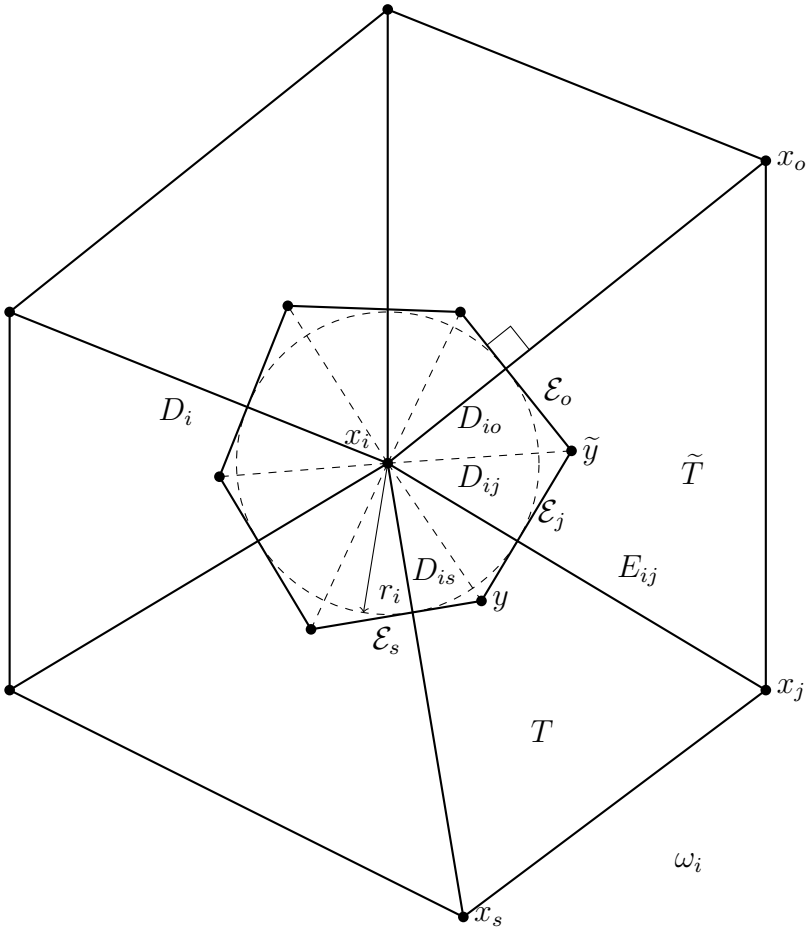


Figure 4.3 Illustration to the construction of final discrete test function $\widetilde{\varphi}_i$ in two dimensions.

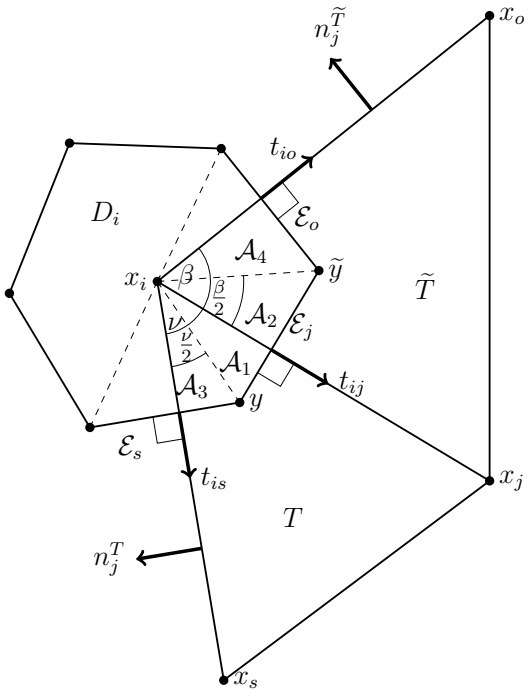


Figure 4.4 Decomposition of the polygon D_i.

We can observe that $n_i^{D_{io}} = t_{io}$, $n_i^{D_{ij}} = t_{ij}$ and $n_i^{D_{is}} = t_{is}$ (see Figure 4.4). According to Lemma 5 and this observation, we can easily compute the gradient of $\widetilde{\varphi}_i$ on $\omega_{ij} \cap D_i$:

$$\begin{aligned}
\nabla\widetilde{\varphi}_i|_{\mathcal{A}_1} &= -\frac{|\mathcal{E}_j|}{2\,|D_{ij}|}t_{ij},\\
\nabla\widetilde{\varphi}_i|_{\mathcal{A}_2} &= -\frac{|\mathcal{E}_j|}{2\,|D_{ij}|}t_{ij},\\
\nabla\widetilde{\varphi}_i|_{\mathcal{A}_3} &= -\frac{|\mathcal{E}_s|}{2\,|D_{is}|}t_{is},\\
\nabla\widetilde{\varphi}_i|_{\mathcal{A}_4} &= -\frac{|\mathcal{E}_o|}{2\,|D_{io}|}t_{io}.
\end{aligned}$$

Using these expressions, together with (4.31), (4.32), and Lemma 5, we obtain

$$\begin{aligned}
\int_{\omega_i} \nabla\varphi_j \cdot \nabla\widetilde{\varphi}_i \,\mathrm{d}x &= \sum_{l=1}^{4} \int_{\mathcal{A}_l} \nabla\varphi_j \cdot \nabla\widetilde{\varphi}_i \,\mathrm{d}x\\
&= |\mathcal{A}_1| \,\frac{|E_{is}|\,|\mathcal{E}_j|}{4\,|D_{ij}|\,|T|} t_{ij} \cdot n_j^T\\
&\quad + |\mathcal{A}_2| \,\frac{|E_{io}|\,|\mathcal{E}_j|}{4\,|D_{ij}|\,\left|\widetilde{T}\right|} t_{ij} \cdot n_j^{\widetilde{T}}\\
&\quad + |\mathcal{A}_3| \,\frac{|E_{is}|\,|\mathcal{E}_s|}{4\,|D_{is}|\,|T|} t_{is} \cdot n_j^T\\
&\quad + |\mathcal{A}_4| \,\frac{|E_{io}|\,|\mathcal{E}_o|}{4\,|D_{io}|\,\left|\widetilde{T}\right|} t_{io} \cdot n_j^{\widetilde{T}}.
\end{aligned} \tag{4.33}$$

Since t_{io} and t_{is} are perpendicular to $n_j^{\widetilde{T}}$ and n_j^T, respectively, the last two terms on the right-hand side of (4.33) are equal to zero. From straightforward reasoning, it follows that $t_{ij} \cdot n_j^T$ and $t_{ij} \cdot n_j^{\widetilde{T}}$ are strictly less than zero. Therefore, we have

$$\int_{\omega_i} \nabla\varphi_j \cdot \nabla\widetilde{\varphi}_i \,\mathrm{d}x = |\mathcal{A}_1| \,\frac{|E_{is}|\,|\mathcal{E}_j|}{4\,|D_{ij}|\,|T|} t_{ij} \cdot n_j^T + |\mathcal{A}_2| \,\frac{|E_{io}|\,|\mathcal{E}_j|}{4\,|D_{ij}|\,\left|\widetilde{T}\right|} t_{ij} \cdot n_j^{\widetilde{T}} < 0. \tag{4.34}$$

We have shown that the test functions $\widetilde{\varphi}_i$ possess the desired property (4.19). Moreover, based on the relation (4.34), we derive a practical formula for the computation of the corresponding weights.

Theorem 41. *Under the notation introduced above, the weights corresponding to $\widetilde{\varphi}_i$ are given by:*

$$w_{ij} = \frac{1}{|E_{ij}|}\left(\left(\frac{1 - t_{ij}\cdot t_{is}}{1 + t_{ij}\cdot t_{is}}\right)^{\frac{1}{2}} + \left(\frac{1 - t_{ij}\cdot t_{io}}{1 + t_{ij}\cdot t_{io}}\right)^{\frac{1}{2}}\right) \tag{4.35}$$

for all $i \in \{1, \ldots, M\}$ and $j \in S_i$.

Proof. Let us denote the angles between the vectors t_{ij}, t_{io} and t_{ij}, t_{is} by β and ν, respectively (see Figure 4.4). Then

$$\begin{aligned} t_{ij} \cdot n_j^T &= -\sin(\nu) \\ t_{ij} \cdot n_j^{\widetilde{T}} &= -\sin(\beta). \end{aligned}$$

Using this notation, we can write

$$\begin{aligned} |D_{ij}| &= \frac{1}{2} |\mathcal{E}_j| \, r_i, \\ |\mathcal{A}_1| &= \frac{1}{2} |\mathcal{E}_j \cap T| \, r_i, \\ |\mathcal{A}_2| &= \frac{1}{2} \left|\mathcal{E}_j \cap \widetilde{T}\right| r_i, \\ |T| &= \frac{1}{2} |E_{ij}| \, |E_{is}| \sin(\nu), \\ \left|\widetilde{T}\right| &= \frac{1}{2} |E_{ij}| \, |E_{io}| \sin(\beta). \end{aligned}$$

Inserting all the above relations into equation (4.34) yields the following formula for $A_i(\varphi_j)$:

$$A_i(\varphi_j) = \int_{\omega_i} \nabla\varphi_j \cdot \nabla\widetilde{\varphi}_i \, \mathrm{d}x = -\frac{|\mathcal{E}_j|}{2\,|E_{ij}|}. \tag{4.36}$$

It remains to compute the length of the line segment $\mathcal{E}_j$. We observe (see Figure 4.4) that

$$|\mathcal{E}_j| = |\mathcal{E}_j \cap T| + \left|\mathcal{E}_j \cap \widetilde{T}\right| = r_i \left(\tan\left(\frac{\nu}{2}\right) + \tan\left(\frac{\beta}{2}\right) \right). \tag{4.37}$$

From the standard double-angle formula for cosine, it follows that

$$t_{ij} \cdot t_{is} = \cos(\nu) = \cos^2\left(\frac{\nu}{2}\right) - \sin^2\left(\frac{\nu}{2}\right) = 1 - 2\sin^2\left(\frac{\nu}{2}\right),$$

$$t_{ij} \cdot t_{is} = \cos(\nu) = \cos^2\left(\frac{\nu}{2}\right) - \sin^2\left(\frac{\nu}{2}\right) = 2\cos^2\left(\frac{\nu}{2}\right) - 1.$$

Combination of these two relations with definition of the function tangent yields

$$\tan\left(\frac{\nu}{2}\right) = \left(\frac{1 - t_{ij} \cdot t_{is}}{1 + t_{ij} \cdot t_{is}}\right)^{\frac{1}{2}}. \tag{4.38}$$

Note that $0 < \nu < \pi$, therefore $\sin\left(\frac{\nu}{2}\right), \cos\left(\frac{\nu}{2}\right) > 0$. By an analogous procedure, we obtain

$$\tan\left(\frac{\beta}{2}\right) = \left(\frac{1 - t_{ij} \cdot t_{io}}{1 + t_{ij} \cdot t_{io}}\right)^{\frac{1}{2}}. \tag{4.39}$$

Inserting equations (4.38) and (4.39) into equation (4.37) yields

$$|\mathcal{E}_j| = r_i \left(\left(\frac{1 - t_{ij} \cdot t_{is}}{1 + t_{ij} \cdot t_{is}}\right)^{\frac{1}{2}} + \left(\frac{1 - t_{ij} \cdot t_{io}}{1 + t_{ij} \cdot t_{io}}\right)^{\frac{1}{2}} \right).$$

Finally, inserting this formula into equation (4.36) yields identity

$$A_i(\varphi_j) = -\frac{r_i}{2\,|E_{ij}|}\left(\left(\frac{1-t_{ij}\cdot t_{is}}{1+t_{ij}\cdot t_{is}}\right)^{\frac{1}{2}} + \left(\frac{1-t_{ij}\cdot t_{io}}{1+t_{ij}\cdot t_{io}}\right)^{\frac{1}{2}}\right).$$

By virtue of (4.17) and Lemma 36, we can obtain the desired formula (4.35).

□

4.2.5 General Remarks on the Construction of $\widetilde{\varphi}_i$ in Three Dimensions

Unfortunately, the procedure described in Subsection 4.2.4 cannot be directly applied in three dimensions. This limitation is illustrated by the following observation.

First, we turn our attention back to the construction of the support of $\widetilde{\varphi}_i$ in two dimensions. Consider an interior vertex $x_i \in \Omega$ of a given triangulation $\mathcal{T}_h$, and denote its neighboring vertices by $x_1, \ldots, x_P \in \omega_i$. We can construct the support of $\widetilde{\varphi}_i$ independently of these neighboring vertices. Thus, we choose points $\widetilde{x}_1, \ldots, \widetilde{x}_P \in \omega_i$ such that $\{\widetilde{x}_j\} = E_{ij} \cap \partial B_i$ for $j = 1, \ldots, P$, where $B_i \subset \omega_i$ is a disc of sufficiently small radius $r_i > 0$, centered at x_i, and satisfying $B_i \subset \overline{B_i} \subset \mathrm{Int}(\omega_i)$. The support of $\widetilde{\varphi}_i$ can then be constructed as the Voronoi box $\mathcal{V}_i$ of x_i (see Definition 5). Note that the vertex of $\mathcal{V}_i$ contained in the sector of a plane bounded by rays originating at x_i and passing through $\widetilde{x}_j$ and $\widetilde{x}_k$ is the center of the circumscribed circle of the triangle determined by x_i, $\widetilde{x}_j$, and $\widetilde{x}_k$. For better understanding, see Figure 4.5.

This construction is evidently equivalent to the one described in Subsection 4.2.4. Let us now attempt to proceed in a similar way in the three-dimensional case. Consider a tetrahedron with vertices x_i, $\widetilde{x}_j$, $\widetilde{x}_k$, and $\widetilde{x}_l$, where the points $\widetilde{x}_j$, $\widetilde{x}_k$, and $\widetilde{x}_l$ lie on the sphere ∂B_i of radius $r_i > 0$, centered at x_i. Next, denote by p_j, p_k, and p_l the planes that intersect the edges E_{ij}, E_{ik}, and E_{il}, respectively, perpendicularly at their midpoints. These planes form the faces of the corresponding Voronoi box of x_i. It is well known that the point $\{C\} = p_j \cap p_k \cap p_l$ is the center of the circumscribed sphere $\mathcal{S}$ of the tetrahedron.

If we wanted to proceed as in the two-dimensional case, the point C would need to lie within the region bounded by the planes of the faces that intersect at the vertex x_i. However, this condition is generally not satisfied.

Another possibility of extension to three dimensions is offered by formula (4.22). In theory, a procedure similar to the proof of Theorem 40 could be used.

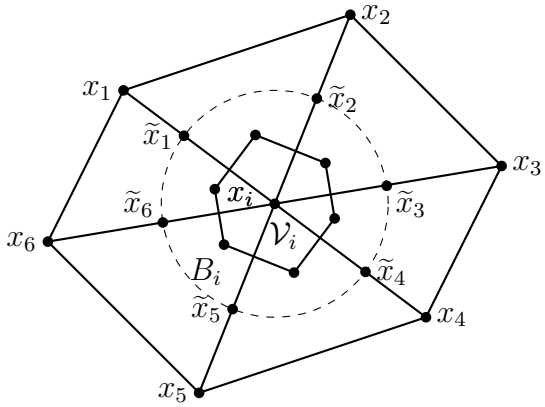


Figure 4.5 Illustration to the construction of the support of $\widetilde{\varphi}_i$ via Voronoi box.

5 Numerical Results

This chapter presents numerical results obtained using the BBK method (4.2)–(4.4), with A_i and B_i defined by (4.15), applied to the boundary value problem (1.1).

5.1 Implementation

All numerical demonstrations were performed using a MATLAB implementation. To assemble the linear system arising from the discrete variational problem, the *Dunavant numerical quadrature* of the fifth algebraic order is used. Note that this quadrature has positive weights (see [18]).

Both formulas (4.23) and (4.35) were tested for calculation of the weights of the limiter, but they exhibited the same convergence behavior on the model problem described below. Computations seems to be a little bit faster with (4.23).

The resulting nonlinear problem was solved using *fixed-point iteration with a dynamically selected damping parameter* (see [1], Section 10.12). Note that the parameter in (4.3) was chosen to be $p = 5$.

5.2 Problem with Smooth Solution

Example 5.2.1. Let $\Omega = (0,1)^2$. Consider the diffusion-convection-reaction equation (1.1) with $b = (2,3)^\top$, $c = 1$, and $\mathbb{E} = \varepsilon\mathbb{I}$, for various values of ε. The prescribed solution is given by

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

and the right-hand side g is determined by this function u, which also defines the homogeneous Dirichlet boundary condition.

♣

In this experiment, the convergence behavior of the AFC scheme with the BBK limiter, described at the beginning of this chapter, was studied for two different families of triangulations. The method was applied to problem described in Example 5.2.1.

The first type of triangulation is called the *triangulation of Friedrichs–Keller type* (see Figure 5.1, left). This triangulation satisfies the XZ-condition and is locally symmetric with respect to interior vertices (see Definition 6).

The second type of triangulation, a so-called *chevron triangulation*, is shown in Figure 5.1, right. This triangulation also satisfies the XZ-condition but does not possess local symmetry.

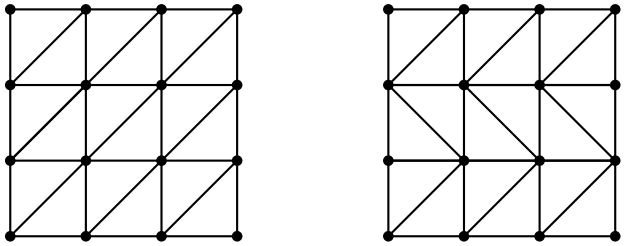
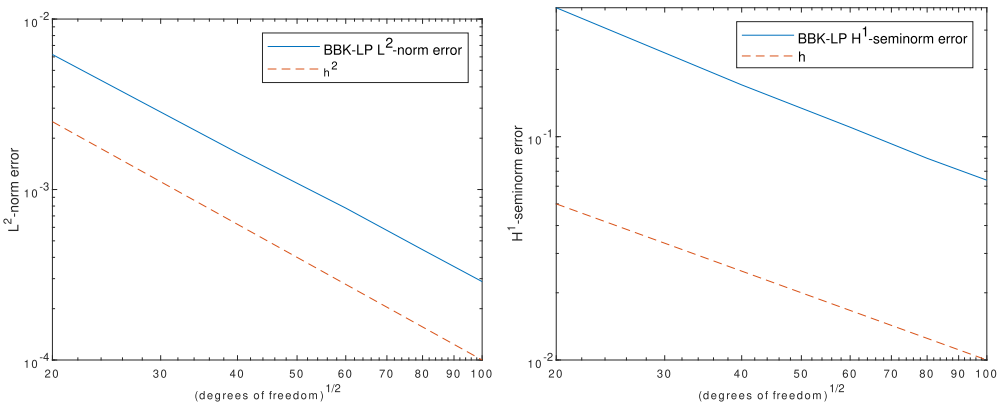


Figure 5.1 Triangulation of Friedrichs–Keller type (left) and chevron triangulation (right).

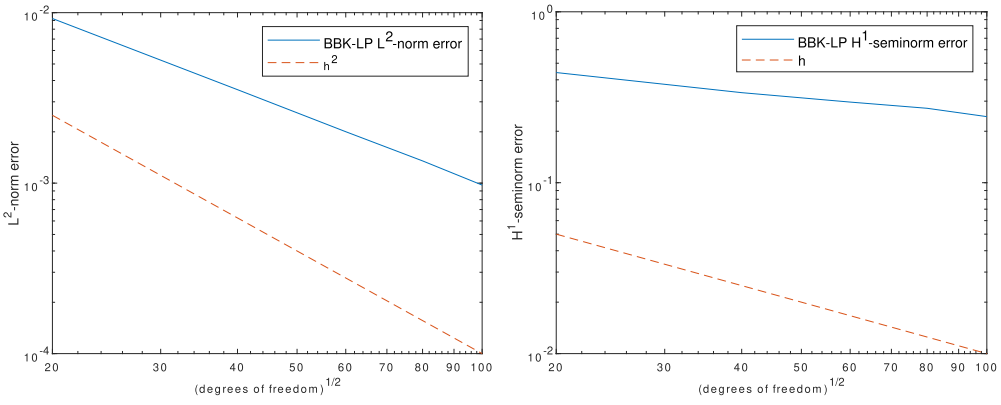
Figure 5.2 presents a convergence history obtained on a family of Friedrichs-Keller triangulations in the convection-dominated regime ($\varepsilon = 10^{-8}$). We can see optimal behavior for both norms.



(a) Convergence in L^2-norm. **(b)** Convergence in H^1-seminorm.

Figure 5.2 Convergence history on a family of Friedrichs-Keller triangulations in the convection-dominated regime ($\varepsilon = 10^{-8}$). The dependence of the error on the square root of the number of degrees of freedom is considered.

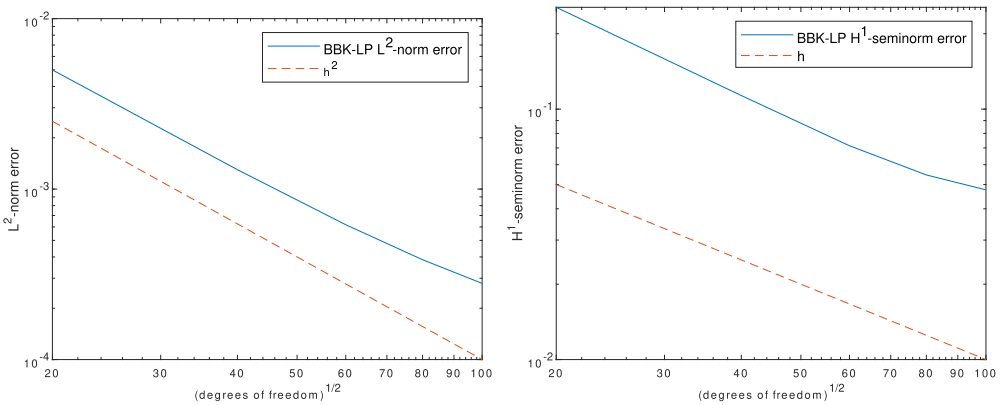
The behavior of the method on a family of chevron triangulations in the convection-dominated regime is noticeably worse, as we can see in Figure 5.3.



(a) Convergence in L^2-norm. **(b)** Convergence in H^1-seminorm.

Figure 5.3 Convergence history on a family of chevron triangulations in the convection-dominated regime ($\varepsilon = 10^{-8}$). The dependence of the error on the square root of the number of degrees of freedom is considered.

Figure 5.4 shows, that the method leads to almost optimal convergence rates also in the diffusion-dominated regime ($\varepsilon = 10$). This behavior was predicted by the theoretical results from Sections 2.2 and 2.3.



(a) Convergence in L^2-norm. **(b)** Convergence in H^1-seminorm.

Figure 5.4 Convergence history on a family of chevron triangulations in the diffusion-dominated regime ($\varepsilon = 10$). The dependence of the error on the square root of the number of degrees of freedom is considered.

Conclusion

We dealt with algebraic flux correction (AFC) schemes for boundary value problems governed by convection-diffusion-reaction equations. These methods provide an alternative to stabilization techniques that operate at the level of discrete variational form.

In the first chapter, we introduced the discrete variational problem based on the Galerkin discretization and discussed the validity of discrete maximum principles (DMPs) for this method. The analysis included the influence of numerical quadrature, which is often neglected in the literature. Based on this analysis, a general form of AFC schemes was derived. At the end of the chapter, we also presented results on the DMPs for nonlinear problems arising from AFC schemes.

The second chapter was devoted to a priori error estimates for AFC schemes. We slightly extended the standard results to the case involving a general diffusion tensor. An L^2-error estimate was derived for this class of nonlinear methods in the diffusion-dominated regime. This analysis represents a novel contribution to the theory of AFC schemes.

In the third chapter, we presented the construction of two standard limiters based on algebraic fluxes, namely, the Kuzmin limiter and the BJK limiter. Their basic properties were also briefly discussed.

The core of the thesis lies in Chapter 4. Here, an alternative approach to the construction of limiters was introduced, based on the so-called local smoothness indicators. The resulting method is known as the BBK method. However, we extended the standard analysis to the case involving weighted indicators. Moreover, a new method was derived based on integral indicators and it was shown that this method is a special case of the BBK method for a particular choice of weights.

An potential advantage of the new method is its linearity-preservation property on general triangulations, unlike the standard version of the BBK method. As indicated in Chapter 2, such methods can exhibit better asymptotic convergence behavior.

The new method described in Chapter 4 was derived for general triangulations only in two dimensions. It turned out that the construction approach used cannot be straightforwardly extended to three-dimensional problems. Therefore, the extension of the method to three dimensions remains an open problem.

In the fifth chapter, the behavior of the new method was demonstrated on a test problem in various regimes and on two different triangulations. A comparison of these results with results of a similar experiments in [1], Section 10.12, suggests that, unfortunately, no significant improvement in convergence was observed compared to the classical BBK method. However, further detailed research is needed in this direction. Nevertheless, this result is still valuable, as it highlights that even for the BBK method, the linearity-preservation property is not sufficient to obtain more accurate solutions. The local symmetry of triangulation may also be important.

For future research, it would be interesting to compare the behavior of AFC schemes based on the limiter from Chapter 4, the standard BBK limiter, and the BBK limiter constructed using generalized barycentric coordinates on various problems.

Another promising direction would be to apply the symmetrization procedure described in [4], Section 6, originally developed for the MUAS method, to the standard BBK limiter.

Bibliography

1. Barrenechea, G. R.; John, V.; Knobloch, P. *Monotone Discretizations for Elliptic Second Order Partial Differential Equations*. Springer, 2024.
2. Evans, L. C. *Partial differential equations*. Providence, R.I.: American Mathematical Society, 2010. Second edition.
3. John, V.; Knobloch, P. On spurious oscillations at layers diminishing (SOLD) methods for convection–diffusion equations: Part I– A review. *Comput. Methods Appl. Mech. Engrg. 196(17-20), 2197–2215*. 2007.
4. Knobloch, P. An algebraically stabilized method for convection-diffusion-reaction problems with optimal experimental convergence rates on general meshes. *Numerical Algorithms, 94, 547–580*. 2023.
5. Kuzmin, D. Algebraic flux correction for finite element discretizations of coupled systems. *M. Papadrakakis, E. Oñate, B. Schrefler (Eds.), Proceedings of the international confer ence on computational methods for coupled problems in science and engineering, pp. 653–656. CIMNE*. 2007.
6. Barrenechea, G. R.; John, V.; Knobloch, P. Analysis of algebraic flux correction schemes. *SIAM J. Numer. Anal. 54: 2427–2451*. 2016.
7. Barrenechea, G. R.; Burman, E.; Karakatsani, F. Edge based nonlinear diffusion for finite element approximations of convection diffusion equations and its relation to algebraic flux-correction schemes. *Numer. Math. 135: 521–545*. 2017.
8. Knobloch, P. On the discrete maximum principle for algebraic flux correction schemes with limiters of upwind type. *Boundary and Interior Layers, Computational and Asymptotic Methods BAIL 2016 (Z. Huang, M. Stynes Z. Zhang eds). Lect. Notes Comput. Sci. Eng., vol. 120. Cham: Springer-Verlag, pp. 129–139*. 2017.
9. Barrenechea, G. R.; John, V.; Knobloch, P. An algebraic flux correction scheme satisfying the discrete maximum principle and linearity preservation on general meshes. *Math. Models Methods Appl. Sci., 27, 525–548*. 2017.
10. Gilbarg, D.; Trudinger, N. S. *Elliptic Partial Differential Equations of Second Order*. Berlin: Springer-Verlag, 2001.
11. Ern, A.; Guermond, J.-L. *Theory and practice of finite elements*. New York: Springer-Verlag, 2004. Applied Mathematical Sciences, vol. 159.
12. Xu, J.; Zikatanov, L. A monotone finite element scheme for convection-diffusion equations. *Math. Comp., 68, 1429–1446*. 1999.
13. Ciarlet, P. G. *The finite element method for elliptic problems*. Amsterdam: North-Holland Publishing Co., 1978. Studies in Mathematics and its Applications, Vol. 4.
14. Cuvelier, F. Exact integration for products of power of barycentric coordinates over d-simplexes in Rn. *Hal-01816347*. 2018.
15. Brenner, S. C.; Scott, L. R. *The mathematical theory of finite element methods*. New York: Springer, 2008. Texts in Applied Mathematics, vol. 15, third edn.

16. Yosida, K. *Functional analysis*. Berlin: Springer-Verlag, 1980. sixth edition.

17. Guermond, J.-L.; Popov, B. Invariant domains and second-order continuous finite element approximation for scalar conservation equations. *SIAM J. Numer. Anal., 55, 3120–3146*. 2017.

18. Dunavant, D. A. High degree efficient symmetrical gaussian quadrature rules for the triangle. *Int. J. Numer. Methods Eng., 21:1129–1148*. 1985.

List of Figures

1.1 Illustration of the notation described in Subsection 1.2.2 for two-dimensional case. 11

4.1 Illustration to the construction of radially symmetric $\widetilde{\varphi}_i$ in two-dimensional case. 52
4.2 Illustration to the proof of Theorem 40. 53
4.3 Illustration to the construction of final discrete test function $\widetilde{\varphi}_i$ in two dimensions. 56
4.4 Decomposition of the polygon D_i. 56
4.5 Illustration to the construction of the support of $\widetilde{\varphi}_i$ via Voronoi box. 60

5.1 Triangulation of Friedrichs–Keller type (left) and chevron triangulation (right). 62
5.2 Convergence history on a family of Friedrichs-Keller triangulations in the convection-dominated regime ($\varepsilon = 10^{-8}$). The dependence of the error on the square root of the number of degrees of freedom is considered. 62
5.3 Convergence history on a family of chevron triangulations in the convection-dominated regime ($\varepsilon = 10^{-8}$). The dependence of the error on the square root of the number of degrees of freedom is considered. 62
5.4 Convergence history on a family of chevron triangulations in the diffusion-dominated regime ($\varepsilon = 10$). The dependence of the error on the square root of the number of degrees of freedom is considered. 63