EÖTVÖS LORÁND UNIVERSITY

INSTITUTE OF MATHEMATICS

ADAPTIVE FINITE ELEMENT METHODS FOR ELLIPTIC EQUATIONS

Ph.D. thesis

Tamás Horváth

Supervisor: Péter L. Simon,

Associate Professor, PhD

Mathematical Doctoral School

Director: Professor Miklós Laczkovich

Member of the Hungarian Academy of Sciences

Doctoral Program: Applied Mathematics

Director of Program: Professor György Michaletzky

Doctor of the Hungarian Academy of Sciences

Department of Applied Analysis and Computational Mathematics

Budapest, 2013

Contents

Acknowledgement

I would like to express my gratitude to my esteemed supervisor Dr. Péter L. Simon for the valuable discussions and for his inspiring lectures. This thesis could not have been done without his endless support and encouragement.

I would also like to express my gratitude to my colleagues Dr. Ferenc Izsák, Miklós Mincsovics for the common work; Dr. János Karátson, Balázs Kovács and Tímea Haszpra for their valuable comments and help.

Throughout the recent years I had the opportunity to participate in several conferences and summer schools where I was inspired by Professor Endre Süli (Oxford), William Mitchell (NIST) and many others who helped me with my questions, regardless of their busy schedule.

Finally, I would like to express my deepest gratitude to my Mother for her support, understanding and unceasing patience.

Introduction

Originated from potential theory, elliptic partial differential equations have been studied for more than two hundred years, hence they have a huge literature. Its history and the theory from the basic to the advanced level is presented among many others in Evans' monograph [26]. Potential theory was first developed for describing gravity and electrostatics, but today elliptic partial differential equations can be found in many branches of physics, chemistry and biology. For example, even the simplest elliptic equation,

$$-\mathrm{div}\,(\mathcal{K}\nabla u) = f$$

can describe electric conduction ($\mathcal{K}$ is the electric conductivity), or fluid flow through a porous medium ($\mathcal{K}$ is the porosity), or simply stationary diffusion ($\mathcal{K}$ is the diffusion coefficient).

An elliptic partial differential equation cannot be solved analytically in general, i.e. there is no explicit formula for the solution. Therefore, nowadays they are solved numerically when the main tasks are to estimate the error of the computational method and to examine the preservation of qualitative properties.

A wide range of numerical methods have been developed in the last few decades, such as finite difference, finite element or finite volume techniques. Throughout this thesis we will focus on finite element methods. On the one hand, it is motivated by the fact that finite element methods are widely used, however, which is even more important, we are interested in higher order methods that can be more easily constructed in the case of finite elements.

The standard version of the finite element method is a popular and well-known procedure to solve an elliptic partial differential equation. However, recently a new class of Galerkin methods has been developed, the so-called discontinuous Galerkin method. There are several types of it, see e.g. [21, 25, 61] or the review article [5]. The most important advantages of using the discontinuous Galerkin method are the following: the polynomial degree can be different in the neighbouring elements and hanging nodes can be used in the mesh. We will use and investigate the so-called Interior Penalty Discontinuous Galerkin (IPDG) method.

One of the main goals of our investigation is to study adaptivity. Adaptivity is a useful tool to solve an equation in a fast manner. When using adaptivity we solve the discrete problem not only once, instead the discrete approximation space is modified in an appropriate way in order to decrease the error. Therefore, an adaptive method contains an error estimation in every step. If the error is bigger than a given tolerance then some refinements are applied until the error becomes small enough. The refinement can be one of the following: mesh modification (h-adaptivity) or modification of the used polynomial degree (p-adaptivity) or both (hp-adaptivity). Using adaptivity, the above mentioned advantages (varying polynomial degrees and hanging nodes) make the DG methods more effective and easy-to-implement in comparison with standard methods.

The second area of our research is the maximum principle both in the continuous and in the discrete case. Continuous maximum principles are important tools when modeling physical phenomena. For example, when modeling heat distribution without internal heat source it is natural that the maximum and the minimum of the temperature occurs at the boundary. Recently it has become more and more important and widely studied how the relations prescribed by physical principles are preserved by numerical methods. Using standard or discontinuous Galerkin methods discrete maximum principles can be investigated, i.e. the minimum and the maximum of the discrete solution have to be on the boundary.

There are several results on maximum principles when first order elements are used. The case of higher order elements is slightly more complicated, namely even the second order basis functions can have negative values, hence the nonnegativity of the discrete solution cannot be guaranteed via the nonnegativity of the representing vector.

Our two topics, adaptivity and maximum principles, are typically dealt with separately. However, they are related in a way, namely, during mesh refinements we can create such meshes that fulfil the mesh conditions arising from the discrete maximum principles. Maximum principles are usually examined only when first order elements are used in the discretisation. On the other hand, the simplest way of h-adaptivity is the same: using the lowest order elements and refining only the mesh, until the error is small enough in some sense.

In this thesis we will present two results in connection with adaptivity: an improvement of the implicit a-posteriori error estimation technique and the correction of the reference solution based hp-adaptivity. Two results will be shown in connection with maximum principles: firstly, the collection of various definitions of different maximum principles with proven necessary and sufficient conditions for the missing ones; and secondly, the discrete maximum principle investigation for IPDG in one dimension.

In Chapter 1 we will go through the basics of standard and discontinuous Galerkin methods, as well as adaptivity and maximum principles. We will establish the theoretical background for the latter chapters and we will introduce the notations that will be used throughout this thesis.

In Chapter 2 we will examine the implicit a-posteriori error estimator. This method solves a local Neumann problem in every subdomain. One of the main questions related to this method is the efficient approximation of the derivative of the unknown solution on the boundary. We will combine the method with the gradient averaging, based on [39], which gives us the possibility to derive a new type of error bound that can be used for adaptivity. In the literature usually the norm of the computational error $e_{h,p}$ and the norm of estimator $\hat{e}_{h,p}$ are related. We will show, that the norm of their difference can be estimated by using a new construction for the boundary conditions.

In Chapter 3 the well-known reference solution based hp-adaptive method will be investigated, according to [38]. We will emphasize some of its drawbacks and will give a method that can produce counterexamples. Such examples force the algorithm to terminate, because the error estimation process identifies the error to be zero, however, this can be arbitrary. We will also show possible corrections.

Chapter 4 will list the different weak and strong maximum principles for elliptic operators and discrete operators (matrices), this chapter is based on [53]. We will prove that the strong one depends on the connectivity of the computational domain in the continuous case; and in the discrete case it depends on the irreducibility of the matrices. Finally, numerical examples will be presented, in which some maximum principles are, while some are not preserved by the discretisation.

Chapter 5 will focus on IPDG methods for one dimensional boundary value problems and it is based on [40]. We will give sufficient mesh conditions that ensure the discrete maximum principle for the discrete solution. We will show, that these conditions are not necessary, but they are sharp in some sense. Such an investigation for IPDG methods have never been done before.

Finally, a short Appendix will close the thesis. It contains some chapters from elementary functional analysis, some proofs from Chapter 1 and a short review of M-matrix theory.

Chapter 1

Finite Element Methods

1.1 Second order PDEs - weak form

Throughout this chapter we will build up numerical methods for solving second order elliptic partial differential equations. It will be done step-by step and finally we will develop tools to solve the general types of boundary value problems.

First, we present the method for the simple case of second order elliptic equations, then for the general case of diffusion-reaction equations. Although the first, simpler derivation is a special case of the second, we use it to convey our ideas and arguments in a clearer and more concise way.

Let $\Omega \subset \mathbb{R}^d$ be an open bounded domain, $\Gamma = \partial\Omega$ be the boundary of Ω, that is decomposed into two parts: $\Gamma = \Gamma_{\mathrm{D}} \cup \Gamma_{\mathrm{N}}$ such that $\Gamma_{\mathrm{D}} \cap \Gamma_{\mathrm{N}} = \emptyset$. Let us examine the following second order elliptic partial differential equation

$$
\begin{aligned}
-\text{div}\,(\mathcal{K}\nabla u) + \mu u = f \qquad & \text{in } \Omega, && (1.1)\\
u = g \qquad & \text{on } \Gamma_{\mathrm{D}}, && (1.2)\\
\partial_\nu \mathcal{K} u = u_N \quad & \text{on } \Gamma_{\mathrm{N}}, && (1.3)
\end{aligned}
$$

where $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is a symmetric uniformly positive definite matrix valued function, see Definition 1.1, $\mathcal{K}_{i,j} \in L^\infty(\Omega)$ for all $i, j \in \{1, 2, \dots, d\}$, $\mu : \mathbb{R}^d \to \mathbb{R}$ is a nonnegative function, $\mu \in L^\infty(\Omega)$, $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$, $u_N \in L^2(\partial\Omega)$ are given functions and ν is the outward normal of Ω. Classical solution means that we seek $u \in C^2(\Omega) \cap C(\overline{\Omega})$, satisfying (1.1)-(1.3). Throughout this thesis we will suppose that $\mathrm{meas}(\Gamma_{\mathrm{D}}) \neq 0$ where meas is the $d-1$ dimensional Lebesgue measure.

Definition 1.1 *A matrix valued function $\mathcal{K} : \mathbb{R}^d \to \mathbb{R}^{d\times d}$ is called uniformly positive definite if there exist $\lambda, \Lambda > 0$ such that $\lambda\|\xi\|_E^2 \leq \mathcal{K}(x)\xi \cdot \xi \leq \Lambda\|\xi\|_E^2$, for all $x \in \Omega$, $\xi \in \mathbb{R}^d$, where $\|\cdot\|_E$ denotes the Euclidean norm in $\mathbb{R}^d$.*

Definition 1.2 *The partial differential operator $\mathcal{A}u = -div\,(\mathcal{K}\nabla u) + \mu u$ is called uniformly elliptic if there exist $\lambda > 0$ such that $\lambda||\xi||_E^2 \leq \mathcal{K}(x)\xi \cdot \xi$, for all $x \in \Omega$, $\xi \in \mathbb{R}^d$, where $||\cdot||_E$ denotes the Euclidean norm in $\mathbb{R}^d$.*

Remark 1.3 *If $\mathcal{K}$ is uniformly positive definite, then the corresponding partial differential operator is uniformly elliptic.*

There are many physical examples where the classical solution does not exist, hence weak solutions are considered. For example, when modeling heat distribution in a stick that is made of two different materials and the temperature is different on the two sides of the stick the solution will not be differentiable at the point where the different materials are joined, therefore it is impossible to expect a C^2 solution.

1.1.1 Dirichlet boundary condition

Suppose that $\mu \equiv 0$, and $\Gamma_{\mathrm{N}} = \emptyset$ that means we have pure Dirichlet boundary condition. Let us start with the simplest case: $u_g = 0$. Mathematically we have

$$-\mathrm{div}\,(\mathcal{K}\nabla u) = f \quad \text{in } \Omega, \tag{1.4}$$

$$u = 0 \quad \text{on } \Gamma. \tag{1.5}$$

Multiplication of (1.4) by a test function $v \in C^2(\Omega)$, such that $v|_\Gamma = 0$, according to (1.5), integrating over Ω and using Green's Theorem gives that

$$\int_\Omega -\mathrm{div}\,(\mathcal{K}\nabla u) \cdot v = \int_\Omega fv$$

$$\int_\Omega \mathcal{K}\nabla u \cdot \nabla v - \int_\Gamma \partial_\nu \mathcal{K} uv = \int_\Omega fv \qquad \text{(Green's Theorem)}$$

$$\int_\Omega \mathcal{K}\nabla u \cdot \nabla v = \int_\Omega fv, \tag{1.6}$$

where in the last equality we used that $v|_\Gamma = 0$. However, (1.6) requires lower regularity on u and v, namely they only have to be in $H_0^1(\Omega) = \{u \in H^1(\Omega) : u|_\Gamma = 0\}$. Using these the weak form reads as follows

Problem Set 1.4

$$\begin{cases} \textit{Seek } u \in H_0^1(\Omega) \textit{ such that} \\ a(u,v) = L(v) \quad \forall v \in H_0^1(\Omega). \end{cases}$$

where we used the notations

$$a(w,v) := \int_\Omega \mathcal{K}\nabla w \cdot \nabla v, \tag{1.7}$$

$$L(v) := \int_\Omega fv. \tag{1.8}$$

Using the above notations we have $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ and $L : H_0^1(\Omega) \to \mathbb{R}$. To prove the existence and uniqueness of the solution to Problem Set 1.4 we have to show some properties of $a(\cdot,\cdot)$ and $L(\cdot)$, namely that $a(\cdot,\cdot)$ is a bounded and coercive bilinear form, and L is a bounded linear functional.

Definition 1.5 *The bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is **bounded** (in the $H_0^1(\Omega)$ norm) if there exists a positive constant C_b: $|a(w,v)| \leq C_b \|w\|_{H_0^1(\Omega)} \|v\|_{H_0^1(\Omega)}$, $\forall w, v \in H_0^1(\Omega)$.*

*The bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is **coercive** (in the $H_0^1(\Omega)$ norm) if there exists a positive constant C_c: $a(v,v) \geq C_c \|v\|^2_{H_0^1(\Omega)}$, $\forall v \in H_0^1(\Omega)$.*

*The linear form $L : H_0^1(\Omega) \to \mathbb{R}$ is **bounded** (in the $H_0^1(\Omega)$ norm) if there exists a positive constant C_{L_b}: $L(v) \leq C_{L_b} \|v\|_{H_0^1(\Omega)}$, $\forall v \in H_0^1(\Omega)$.*

Remark 1.6 (Non-homogeneous Dirichlet boundary) *Suppose that $g \neq 0$ and u_g satisfies the Dirichlet boundary condition, i.e. $u_g|_\Gamma = g$ (it is usually called "Dirichlet-lift"). By subtracting this we can reduce our problem to homogeneous boundary condition which means we seek $u = w + u_g$, where $w|_\Gamma = 0$ and derive the weak form for u. Seeking $u \in H^1(\Omega)$ such that $u = w + u_g, w \in H_0^1(\Omega) : a(w,v) = L(v) - a(u_g, v), \forall v \in H_0^1(\Omega)$.*

Remark 1.7 *We will use the following notations for norms and seminorms:*

- $\|u\|^2_{0,T} := \|u\|^2_{L^2(T)} = \int_T |u|^2$,
- $\|u\|^2_{k,T} := \|u\|^2_{H^k(T)} = \sum_{|\alpha| \leq k} \int_T |\partial^\alpha u|^2$,
- $|u|^2_{k,T} := \sum_{|\alpha| = k} \int_T |\partial^\alpha u|^2$,

where α is a multi-index and T is an arbitrary domain of integration. For more details on the corresponding functional spaces, see Appendix A.1. We will omit the second subscript if the integration domain is Ω (i.e.: $\|u\|_0^2 := \|u\|^2_{0,\Omega}$).

The $H_0^1(\Omega)$ norm is: $\|u\|^2_{H_0^1(\Omega)} = \|\nabla u\|_0^2 = \sum_{i=1}^d \|\partial_i u\|_0^2$.

Theorem 1.8 (Poincare-Friedrichs-Sztyeklov inequality) *There exists a $C_{\mathrm{PFS}} > 0$ constant, such that for all $u \in H_0^1(\Omega)$: $\|u\|_1 \leq C_{\mathrm{PFS}} \|u\|_{H_0^1(\Omega)}$*

First of all we will show, that this is really a norm (not only a seminorm) and this is equivalent to the standard $H^1(\Omega)$ norm.

According to Theorem 1.8 we can conclude

$$\|u\|_1^2 \leq C^2_{\mathrm{PFS}} \|u\|^2_{H_0^1(\Omega)} = C^2_{\mathrm{PFS}} \|\nabla u\|_0^2 \leq C^2_{\mathrm{PFS}} \left(\|\nabla u\|_0^2 + \|u\|_0^2 \right) \leq C^2_{\mathrm{PFS}} \|u\|_1^2.$$

If $a(\cdot,\cdot)$ is a symmetric, coercive bilinear form we can introduce the energy inner product: $\langle u, v \rangle_a := a(u,v)$ and the energy norm $|||u|||_a^2 := a(u,u)$.

Remark 1.9 *For simplicity we will use the $H_0^1(\Omega)$ (or equivalently the $H^1(\Omega)$) norm, however, it should be noted that we can achieve better theoretical results if we use the energy norm instead of the $H_0^1(\Omega)$ norm.*

Theorem 1.10 (Riesz Representation Theorem) *Let H be a Hilbert space. For all bounded linear functionals $L : H \to \mathbb{R}$ there exists a unique $u \in H$ such that $L(v) = \langle v, u \rangle$ for all $v \in H$, where $\langle \cdot, \cdot \rangle$ is an inner product on H.*

Using the above theorem we can prove existence and uniqueness for Problem Set 1.4.

Theorem 1.11 *Suppose that $a(u, v)$ is symmetric, coercive, bilinear and $L(v)$ is bounded. Then there exists a unique solution to Problem Set 1.4.*

proof: Simple consequence of the Riesz Representation Theorem is that using the energy inner product we have $\exists ! u$: such that $L(v) = \langle v, u \rangle_a = a(v, u) = a(u, v)$. In the last equality we used the symmetry of $a(\cdot, \cdot)$. □

The proof of the required properties of $a(\cdot, \cdot)$ and $L(\cdot)$, defined in (1.7)-(1.8), can be found in Appendix A.2.

Riesz Representation Theorem can only be used if $a(\cdot, \cdot)$ is symmetric as in our case. However, if first order derivatives appear a more powerful theorem is required: the Lax-Milgram Lemma.

Theorem 1.12 (Lax-Milgram Lemma) *Let H be real Hilbert space, $a : H \times H \to \mathbb{R}$ is a bounded, coercive bilinear form. For any bounded linear functional, $L : H \to \mathbb{R}$ there exists a unique $u \in H$ such that $L(v) = a(u, v)$ for all $v \in H$.*

The proof of the existence and uniqness of Problem Set 1.4 by using the Lax-Milgram Lemma can be carried out similarly as by using the Riesz Representation Theorem.

Remark 1.13 *In many applications there is also a linear reaction term, then equation* (1.1) *becomes (with $0 \leq \mu \in L^2(\Omega)$)*

$$-div\,(\mathcal{K} \nabla u) + \mu u = f \quad \text{in } \Omega,$$

subject to proper boundary conditions. The only difference, appearing in the bilinear form, is an additional term: $\int_\Omega \mu u v$.

1.2 Reduction to finite dimension

Problem Set 1.4 cannot be handled numerically, because $H_0^1(\Omega)$ is infinite dimensional. To construct a numerical method we should reduce it to a finite dimensional problem. The simplest way is to define a finite dimensional subspace $V_{h,p} \subset H_0^1(\Omega)$ and

Problem Set 1.14

$$\begin{cases} \textit{Seek } u_{h,p} \in V_{h,p} \textit{ such that} \\ a(u_{h,p}, v_{h,p}) = L(v_{h,p}) \quad \forall v_{h,p} \in V_{h,p}. \end{cases}$$

$a(\cdot,\cdot)$ and $L(\cdot)$ inherit boundedness and coercivity from $H_0^1(\Omega)$ to $V_{h,p}$ hence the existence and uniqueness can be proved similarly as we did in the case of Problem Set 1.4.

We shall define a suitable finite dimensional space $V_{h,p}$. First of all we have to decompose the domain Ω into elements: typically triangles in two dimensions and tetrahedrons in three dimensions. In some cases other elements are also included: quadrilaterals, cubes or prisms. In this thesis we will consider only one and two dimensional examples, therefore three dimensional meshes will be examined very briefly.

The set of the elements will be denoted by $\mathcal{T}_h = \{E_i, i = 1, \ldots, N_{el}\}$, where $\cup_i E_i = \Omega$, and $int\ E_i \cap int\ E_j = \emptyset$ whenever $i \neq j$. At this point we have an extra restriction: two neighbouring elements should share a common edge. This means that hanging nodes are excluded. On the left side of Figure 1.1 the mesh satisfies this, however, in the middle there is a hanging node. Hanging nodes are such nodes that lie on an edge of the neighbouring triangle ($\partial E \cap \partial F$ is not an edge of E). The meshes without hanging node are called regular, otherwise n-irregular, where n is the maximum number of the hanging nodes over an edge. In the middle of Figure 1.1 there is a 1-irregular mesh while in the right hand side there is a 2-irregular.

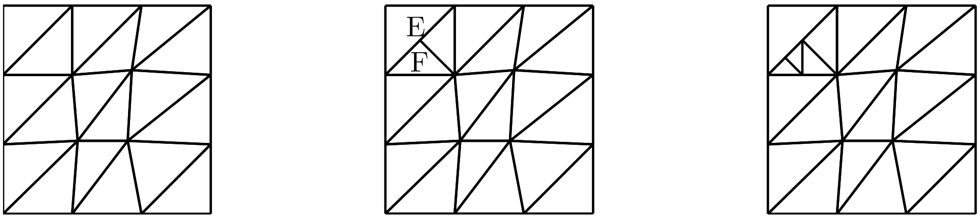

Figure 1.1: Left: regular mesh, middle: 1-irregular mesh, right: 2-irregular mesh.

In the case of standard finite element techniques the irregular meshes are excluded. However, there are some papers on irregular meshes, see e.g. [65].

The space $V_{h,p}$ contains continuous piecewise polynomials of degree p over the elements. Let us denote by $\Phi_1, \ldots, \Phi_N$ a basis of $V_{h,p}$. Using these notations we seek $u_{h,p}$ as

$$u_{h,p} = \sum_{i=1}^{N} c_i \Phi_i. \tag{1.9}$$

Due to the bilinearity of $a(\cdot,\cdot)$ the equation in Problem Set 1.14 has to be satisfied only for the basis functions, leading to a system of linear equations: $\mathbf{Ac} = \mathbf{L}$, where $(\mathbf{A})_{i,j} = a(\Phi_j, \Phi_i)$, $\mathbf{c} = (c_1, \ldots, c_N)^T$ and $(\mathbf{L})_i = L(\Phi_i)$.

We will consider only the case of Lagrange basis functions (sometimes called Lagrange elements). Such a function is associated with a point in the mesh - we will see later that this point is not necessarily a mesh node.

For example $V_{h,1}$ contains piecewise linear functions. Let us denote by x_i an interior node and let us introduce the basis function Φ_i as follows: $\Phi_i(x_j) = \delta_{i,j}$ ($\delta_{i,j}$ is the Kronecker-delta). This property and the linearity of Φ_i determine the function. The support of Φ_i will be the union of those mesh elements that have x_i as a node. The function Φ_i will be referred to as the basis function associated to the mesh node x_i.

Higher order spaces can be constructed similarly, although, the basis functions are associated not only to the nodes, but to the edges and to the elements. That is $V_{h,2}$ contains piecewise quadratic functions, and the basis functions will belong to the nodes of $\mathcal{T}_h$ (nodal functions) or to the midpoint of the edges (edge functions) while in $V_{h,3}$ we have piecewise cubic functions, and the basis functions will belong either to the nodes of $\mathcal{T}_h$ (nodal functions) or to the edges (edge functions) or to the element itself (bubble functions). Usually the bubble functions are chosen such that they belong to some interior points, see Figure 1.3 and 1.4.

The set of points that are associated with basis functions will be denoted by $\mathcal{DOF}$ and it is called degree of freedom. If first degree polynomials are used then $\mathcal{DOF}$ equals to the set of the mesh nodes, in the case of second degree polynomials it is enriched with the edge midpoints, etc. It is important to note that in the case of (homogenous) Dirichlet boundary condition basis functions that are related to boundary points in $\mathcal{DOF}$ can be excluded from the computations, due to the fact that the solution is expressed in the terms of the basis functions, see Section 1.2.1. This means that in (1.9) N is equal to the number of interior points in $\mathcal{DOF}$.

If the mesh contains only triangles/tetrahedrons (or parallelograms/parallelepipeds) it is very comfortable to define the basis functions over a reference element due to the fact that the above elements can be transformed into each other by using an affine linear mapping. We will discuss the case of triangles.

Let us introduce the triangle Ω_0 with nodes $(0, 0), (1, 0)$ and $(0, 1)$. This will be called as reference triangle. Let us take an element $E \in \mathcal{T}_h$, this will be called as physical element, with nodes $(x_1, y_1), (x_2, y_2)$ and (x_3, y_3), see Figure 1.2. The affine linear mapping $\mathcal{J}_E : \Omega_0 \to E$ will be an important tool to define the basis functions. It maps nodes to nodes, more precisely it maps $(0, 0)$ to (x_1, y_1), $(1, 0)$ to (x_2, y_2) and finally $(0, 1)$ to (x_3, y_3). $\mathcal{J}_E(x, y) = J_E \cdot (x, y)^T + C_E$, where J_E is a 2 × 2 matrix and C_E is a constant vector.

$$J_E = \begin{pmatrix} x_2 - x_1 & x_3 - x_1 \\ y_2 - y_1 & y_3 - y_1 \end{pmatrix} \quad C_E = \begin{pmatrix} x_1 \\ y_1 \end{pmatrix}.$$

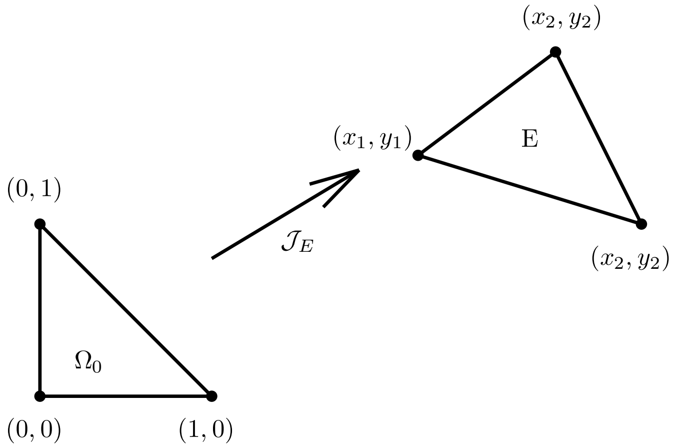

Figure 1.2: The reference element Ω_0, the physical element E and the mapping $\mathcal{J}_E$.

Let us denote the basis functions defined over Ω_0 by $\Phi_i^{\Omega_0}$ (while Φ_i^E denotes the basis functions over E). The following formula establishes the relation between Φ_i^E and $\Phi_i^{\Omega_0}$

$$\Phi_i^{\Omega_0} = \Phi_i^E \circ \mathcal{J}_E. \tag{1.10}$$

Remark 1.15 *We note that the determinant of J_E can easily be computed: $|\det(J_E)| = 2|E|$, where $|E|$ is the area of E, namely*

$$|E| = \int_E 1 = |\det(J_E)| \int_{\Omega_0} 1 = \frac{|\det(J_E)|}{2}$$

using the integral transform.

With these notations the basis functions for $V_{h,1}$ over Ω_0 are:

- $\Phi_1^{\Omega_0}(x, y) = 1 - x - y,$
- $\Phi_2^{\Omega_0}(x, y) = x,$
- $\Phi_3^{\Omega_0}(x, y) = y.$

In Figure 1.3 there are the $\mathcal{DOF}$ points over Ω_0 for polynomial degree one, two and three. The nodes are denoted by $\bullet$ the corresponding functions are the nodal functions. The functions that are belonging to the points on the edges (denoted by $\circ$) are the edge functions. Finally, there are the bubble functions: these functions are associated with the element, although it is convenient to define point(s) inside the element (denoted by $\times$) and set the functions to be equal to 1 at a given interior point (and 0 at the others).

For higher order elements the $\mathcal{DOF}$ points are shown in Figure 1.4.

The number of the basis functions over Ω_0 is $D = \binom{p+2}{2}$, over the three dimensional reference tetrahedron it is $\binom{p+3}{3}$. Similar argumentation can be used over parallelograms/parallelepipeds, the number of the basis functions is $(p+1)^2$ in two dimensions, and $(p+1)^3$ in three dimensions.

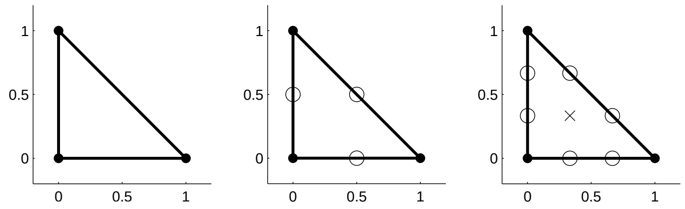

Figure 1.3: $\mathcal{DOF}$ for Lagrange basis functions for $p = 1, 2, 3$.

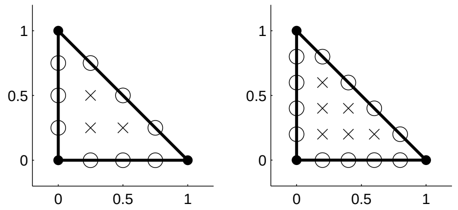

Figure 1.4: $\mathcal{DOF}$ for Lagrange basis functions for $p = 4, 5$.

1.2.1 Non-homogeneous Dirichlet boundary condition

In the case of non-homogeneous boundary condition, $u|_\Gamma = g$, we have seen that we need a function u_g such that $u_g|_\Gamma = g$, by using which we can reduce the problem to the homogeneous one, see Remark 1.6. The function $u_g|_\Gamma$ also has to be approximated. The simplest way is to use pointwise approximation using the basis functions that are associated to boundary points in $\mathcal{DOF}$. Let us denote this approximation by $u_{D_{h,p}}$. In other terms if $x_i \in \mathcal{DOF}$, $i = 1, \ldots, n_D$ are the boundary points in $\mathcal{DOF}$ and Φ_{x_i} $(i = 1, \ldots, n_D)$ are the corresponding basis functions, then

$$u_{D_{h,p}} = \sum_{i=1}^{n_D} g(x_i)\Phi_{x_i}.$$

1.3 Convergence

1.3.1 Convergence in the $H_0^1(\Omega)$ norm

The main goal of this subsection is to estimate the discretisation error, i.e. the distance between u (the solution of Problem Set 1.4) and $u_{h,p}$ (the solution of Problem Set 1.14). First of all, we need the following lemma.

Lemma 1.16 *The discrerisation error is a-orthogonal to the finite element space, that is $a(u - u_{h,p}, v_{h,p}) = 0$, $\forall v_{h,p} \in V_{h,p}$.*

proof: According to (1.4) we have: $a(u, v) = L(v)\ \forall v \in H_0^1(\Omega)$. The relation $V_{h,p} \subset H_0^1(\Omega)$ implies

$$a(u, v_{h,p}) = L(v_{h,p}) \quad \forall v_{h,p} \in V_{h,p}. \tag{1.11}$$

On the other hand Problem Set 1.14 means $a(u_{h,p}, v_{h,p}) = L(v_{h,p})\ \forall v_{h,p} \in V_{h,p}$. Subtracting the two equations we get the desired statement. □

Remark 1.17 *The above property is called Galerkin orthogonality. If equation* (1.11) *holds the finite element method is called consistent.*

Definition 1.18 *The mesh $\mathcal{T}_h = \{E_i, i = 1, \ldots, N_{el}\}$ (E_i is a triangle for all $i = 1, \ldots, N_{el}$) is called shape regular if there exists a constant c_0 such that*

$$h_i \leq c_0 \rho_i,$$

holds $\forall i = 1, 2, \ldots, N$, where h_i is the diameter and ρ_i is the radius of the inner circle of E_i.

In the following we always suppose that the mesh is shape regular.

We will develop an a-priori error estimation using the following four conditions:

- $a(\cdot, \cdot)$ is coercive,
- $a(\cdot, \cdot)$ is bounded,
- $a(\cdot, \cdot)$ possesses the Galerkin orthogonality: $a(u - u_{h,p}, v_{h,p}) = 0\ \forall v_{h,p} \in V_{h,p}$,
- we have an approximation result: an arbitrary function $u \in H_0^1(\Omega) \cap H^{l+1}(\Omega)$ $(l \geq p)$ can be approximated by piecewise polynomials of degree p in order $p : ||u - u_{\mathcal{I}_p}||_{H_0^1(\Omega)} \leq C_a h^p |u|_{p+1}$ (where h is the maximal diameter of the mesh elements and the interpolant is denoted by $u_{\mathcal{I}_p}$),

where we used the seminorm: $|u|_p^2 = \sum_{|\alpha|=p} \int_\Omega |\partial^\alpha u|^2$.

Theorem 1.19 *Suppose that the four above mentioned conditions are satisfied and the exact solution of Problem Set 1.4 is smooth enough, $u \in H_0^1(\Omega) \cap H^{l+1}(\Omega)$ ($l \geq p$). Then there exists $C > 0$ (independent of h) such that for any shape regular mesh with h as the maximal diameter of the elements the following estimation holds*

$$||u - u_{h,p}||_{H_0^1(\Omega)} \leq C h^p |u|_{p+1}.$$

This means that if the solution u is in $H_0^1(\Omega) \cap H^{l+1}(\Omega)$ for some $l \geq p$ and it can be interpolated using polynomials of degree p in order p (as it is stated in the last assumption) then the error of the finite element discretisation is in the same order. The proof can be easily illustrated by Figure 1.5. We will show, that the error $||u_{\mathcal{I}_p} - u_{h,p}||_{H_0^1(\Omega)}$ is in the same order as $||u - u_{\mathcal{I}_p}||_{H_0^1(\Omega)}$ then the triangle inequality completes the proof.

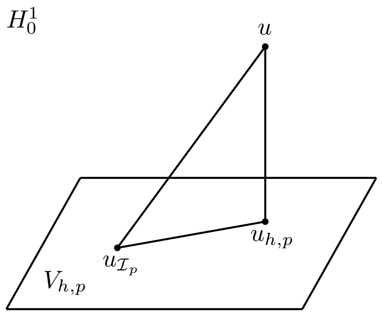

Figure 1.5: Orthogonality of the error to $V_{h,p}$.

proof: Using the first three assumption we have

$$\begin{aligned}
C_c\|u_{\mathcal{I}_p}-u_{h,p}\|^2_{H_0^1(\Omega)} &\leq a(u_{\mathcal{I}_p}-u_{h,p},u_{\mathcal{I}_p}-u_{h,p}) && \text{(coercivity)}\\
&= a(u_{\mathcal{I}_p}-u_{h,p},u_{\mathcal{I}_p}-u_{h,p})-a(u-u_{h,p},u_{\mathcal{I}_p}-u_{h,p})\\
& && \text{(Galerkin orthogonality)}\\
&= a(u_{\mathcal{I}_p}-u,u_{\mathcal{I}_p}-u_{h,p}) \leq C_b\|u_{\mathcal{I}_p}-u\|_{H_0^1(\Omega)}\|u_{\mathcal{I}_p}-u_{h,p}\|_{H_0^1(\Omega)}\\
& && \text{(boundedness)}\\
&\leq C_bC_ah^p|u|_{p+1}\|u_{\mathcal{I}_p}-u_{h,p}\|_{H_0^1(\Omega)}.
\end{aligned}$$

If $\|u_{\mathcal{I}_p}-u_{h,p}\|_{H_0^1(\Omega)}=0$ then the proof follows from the approximation result, otherwise

$$\|u_{\mathcal{I}_p}-u_{h,p}\|_{H_0^1(\Omega)} \leq \frac{C_bC_a}{C_c}h^p|u|_{p+1}.$$

Finally the triangle inequality completes the proof

$$\|u-u_{h,p}\|_{H_0^1(\Omega)} \leq \|u-u_{\mathcal{I}_p}\|_{H_0^1(\Omega)}+\|u_{\mathcal{I}_p}-u_{h,p}\|_{H_0^1(\Omega)} \leq \left(C_a+\frac{C_bC_a}{C_c}\right)h^p|u|_{p+1}.$$

□

For some comments on the approximation results see Appendix A.4.

1.3.2 Error estimation in the $L^2(\Omega)$ norm

Now we derive error estimation in the weaker $L^2(\Omega)$ norm that will yield a better convergence rate than that in the $H_0^1(\Omega)$ norm.

Definition 1.20 *[5] Let $\psi \in H_0^1(\Omega)$ solve the following problem*

$$\begin{aligned}
-div\,(\mathcal{K}\nabla\psi) &= g \quad \text{in } \Omega\\
\psi &= 0 \quad \text{on } \Gamma.
\end{aligned}$$

where g is a given function. If $a(v,\psi)=\int_\Omega vg\ \forall v \in H_0^1(\Omega)$, then the problem is called adjoint consistent.

The adjoint consistency is a simple consequence of two facts: the original problem is self-adjoint and the bilinear form is symmetric.

Theorem 1.21 *Assume the hypotheses of Theorem 1.19, and additionally that the problem is adjoint consistent. Then there exists $C > 0$ (independent of h) such that*

$$||u - u_{h,p}||_0 \leq Ch^p|u|_{p+1}.$$

proof: Let us choose $g := u - u_{h,p}$ and $v := u - u_{h,p}$. With these we have

$$||u - u_{h,p}||_0^2 = \int_\Omega (u - u_{h,p})^2 = a(u - u_{h,p}, \psi) = a(u - u_{h,p}, \psi - \psi_{\mathcal{I}_p}) \leq$$
$$C_b||u - u_{h,p}||_{H_0^1(\Omega)}||\psi - \psi_{\mathcal{I}_p}||_{H_0^1(\Omega)} \leq C||u - u_{h,p}||_{H_0^1(\Omega)}h|\psi|_2$$

In the last step we used the following approximation property: $||\psi - \psi_{\mathcal{I}_p}||_{H_0^1(\Omega)} = |\psi - \psi_{\mathcal{I}_p}|_1 \leq Ch|\psi|_2$. For more details on the approximation results see Appendix A.4.

The continuous dependence of ψ on the datum g gives

$$|\psi|_2 \leq C||u - u_{h,p}||_0.$$

For the proof see i.e. [52, Th. 4.10.].

Summing up we have the final estimation

$$||u - u_{h,p}||_0 \leq C||u - u_{h,p}||_{H_0^1(\Omega)}h \leq \widetilde{C}h^{p+1}|u|_{p+1}.$$

□

1.4 Mixed boundary conditions

In the following we will discuss other boundary conditions. Let us consider the more general case, the mixed boundary conditions

$$\begin{aligned} -\text{div}\,(\mathcal{K}\nabla u) &= f && \text{in } \Omega, \\ u &= g && \text{on } \Gamma_\text{D}, \\ \partial_\nu \mathcal{K} u &= u_N && \text{on } \Gamma_\text{N}. \end{aligned}$$

The non-homogeneous Dirichlet case can be handled as in Remark 1.6 therefore we will consider only the homogeneous one ($g = 0$). As in Section 1.1.1 we multiply with a test function $v \in C^2(\Omega)$. The difference is that v should vanish only on the Dirichlet part of the boundary, therefore

$$\int_\Gamma \partial_\nu \mathcal{K} uv = \int_{\Gamma_\text{N}} \partial_\nu \mathcal{K} uv = \int_{\Gamma_\text{N}} u_N v.$$

Problem Set 1.22

$$\begin{cases} \textit{Seek } u \in H^1_{\Gamma_\mathrm{D}}(\Omega) \textit{ such that} \\ a_N(u,v) = L_N(v) \quad \forall v \in H^1_{\Gamma_\mathrm{D}}(\Omega), \end{cases}$$

where $a_N(u,v) := \int_\Omega \mathcal{K}\nabla u \cdot \nabla v$ is formally the same as in (1.7) although it is defined over $H^1_{\Gamma_\mathrm{D}}(\Omega) \times H^1_{\Gamma_\mathrm{D}}(\Omega)$, where

$$H^1_{\Gamma_\mathrm{D}}(\Omega) = \{u \in H^1(\Omega) : u|_{\Gamma_\mathrm{D}} = 0\}.$$

The linear functional is $L_N(v) := \int_\Omega fv + \int_{\Gamma_\mathrm{N}} u_N v$. Just as before it can be shown that $a_N : H^1_{\Gamma_\mathrm{D}}(\Omega) \times H^1_{\Gamma_\mathrm{D}}(\Omega) \to \mathbb{R}$ is bounded and coercive, $L_N : H^1_{\Gamma_\mathrm{D}}(\Omega) \to \mathbb{R}$ is bounded. These properties ensure the solvability of Problem Set 1.22, for more details see the proof of Theorem 1.11.

1.4.1 Reduction to finite dimension

As in Section 1.2 we have to define a finite dimensional subspace of $H^1_{\Gamma_\mathrm{D}}(\Omega)$. It can be done similarly as before, using piecewise polynomials, the only difference is that the points in $\mathcal{DOF}$ that lie on the Neumann part of the boundary are also included in the computation, therefore in (1.9) N is equal to the total number of interior points and Neumann boundary points in $\mathcal{DOF}$.

1.5 Discontinuous Galerkin Methods

Discontinuous Galerkin (DG) methods are similar to the above finite element method, but they have some advantages:

- built-in stability for time-dependent advection-convection equations,
- adaptivity can be done easily (the basis function do not have to be continuous over the interfaces),
- the mesh does not have to be regular, hanging-nodes can be handled easily,
- conservation laws can be achieved by the numerical solutions.

From now on let us suppose $u \in V_* := H^1_{\Gamma_\mathrm{D}}(\Omega) \cap H^2(\Omega)$. We will introduce the DG methods following [21]. In this section we will use a finite dimensional space $V_{DG} \not\subset V_*$. In such cases the finite element method is called nonconforming. In our case V_{DG} will contain piecewise polynomials, that are not necessarily continuous over the element interfaces. There is one more important function space: $V_{*DG} = V_* + V_{DG}$, where $+$ stands for Minkowski addition.

Also from now on let us consider the following simplification in the diffusion coefficient: we suppose that $\mathcal{K}(\mathbf{x}) = \kappa(\mathbf{x}) \cdot I$, where $\kappa : \mathbb{R}^d \to \mathbb{R}$, $\kappa \in L^\infty(\Omega)$ and it is bounded from below by a positive constant, and I is the identity matrix ($I \in \mathbb{R}^{d\times d}$). The general case is more difficult to handle and not relevant in this thesis.

Additionally we suppose that $d > 1$. For the one dimensional case see Section 1.5.6.

First of all we decompose the domain Ω into elements, just as before. We will use the same notations as in Section 1.2: the set of the elements will be denoted by $\mathcal{T}_h = \{E_i, i = 1, \dots, N_{el}\}$, where $\cup_i E_i = \Omega$, and $int\, E_i \cap int\, E_j = \emptyset$ whenever $i \neq j$. However, the mesh can be irregular i.e. it can contain hanging nodes. Let us denote by Γ_0 the interior interfaces, i.e.: $\Gamma_0 := \{E \cap F : \forall E, F \in \mathcal{T}_h, E \neq F\}$. Before building up the bilinear and linear forms let us define the jumps and averages.

Definition 1.23 *Suppose that the interface $e \in \Gamma_0$ lies on the boundary of E and F and the normal vector of e denoted by ν is oriented from E to F. The jump and the average of u is defined as*

$$[\![u]\!]_e := u|_E - u|_F, \qquad \{\!\{u\}\!\}_e := \frac{1}{2}\left(u|_E + u|_F\right).$$

If e is on the boundary Γ, then $[\![u]\!]_e = \{\!\{u\}\!\}_e = u|_E$. If u is vector valued, then jump and average operator act componentwise.

Remark 1.24 *At this point is seems that the jump of u depends on the orientation of ν. However, in the definition of the bilinear form it will not be confusing.*

Remark 1.25 *It is important to note that $u|_E$ is understood as the trace of u defined over the interior of E.*

Remark 1.26 *There is a different definition of jumps [25, 42]. According to that for an interior edge $e \in \Gamma_0$ (see Figure 1.6)*

$$[\![u]\!]_e = u|_E \cdot \nu_E + u|_F \cdot \nu_F.$$

On the boundary: $[\![u]\!]_e = u \cdot \nu_E$.

It can be seen that using Definition 1.23 the jump of a vector valued function is a vector and the jump of a scalar valued function is a scalar. However, using Remark 1.26 the jump of a vector valued function is a scalar and vice-versa.

Remark 1.27 *Later, when we have to deal with terms like $\int_e [\![u]\!]_e [\![v]\!]_e$ we will omit the subscripts e from the jumps (and also from the averages) because it will be clear from the integral.*

The finite dimensional space V_{DG} will be the broken polynomial space over $\mathcal{T}_h$ that is defined as follows.

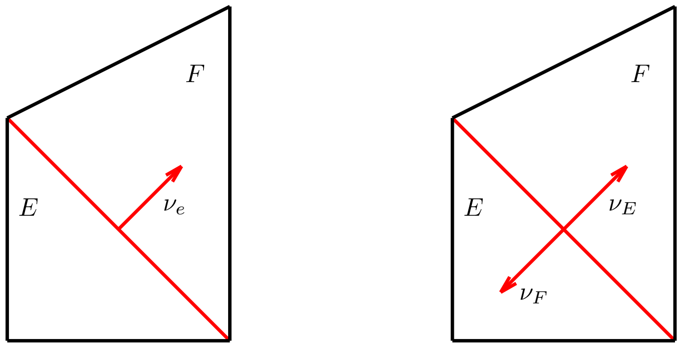

Figure 1.6: Jumps according to Definition 1.23 and Remark 1.26.

Definition 1.28 *Let us denote by $\mathcal{P}_d^p(\mathcal{T}_h)$ the broken polynomial space over $\mathcal{T}_h$:*

$$\mathcal{P}_d^p(\mathcal{T}_h) := \{v \in L^2(\Omega) : \forall E \in \mathcal{T}_h, v|_E \in \mathcal{P}_d^p(E)\},$$

where $\mathcal{P}_d^p(E)$ contains the d-variable polynomials of degree p, defined over E.

If we use piecewise polynomials then neither the standard Sobolev spaces nor the usual gradient can be used. Hence the following definitions are introduced.

Definition 1.29 *Let us denote by $W^{m,l}(\mathcal{T}_h)$ the space of piecewise $W^{m,l}$ functions:*

$$W^{m,l}(\mathcal{T}_h) := \{v \in L^2(\Omega) : \forall E \in \mathcal{T}_h, v|_E \in W^{m,l}(E)\},$$

where $m \geq 0$, $1 \leq p \leq \infty$. For the definition of $W^{m,l}(E)$ see Appendix A.1. Similarly as in Appendix A.1 $H^m(\mathcal{T}_h) := W^{m,2}(\mathcal{T}_h)$, and the corresponding norms and seminorms are $||u||^2_{m,\mathcal{T}_h} = \sum_{E\in\mathcal{T}_h} ||u||^2_{m,E}$ and $|u|^2_{m,\mathcal{T}_h} = \sum_{E\in\mathcal{T}_h} |u|^2_{m,E}$, respectively.

Definition 1.30 *Let us denote by $\nabla_h : W^{1,l}(\mathcal{T}_h) \to [L^l(\Omega)]^d$ the piecewise gradient operator, which is defined as follows:*

$$\forall E \in \mathcal{T}_h : \quad (\nabla_h v)|_E := \nabla(v|_E).$$

Remark 1.31 *It is important to note, that in the literature ∇_h is usually called the* discrete gradient operator, *however, we want to preserve this name for later purposes.*

1.5.1 Construction of the bilinear form

As in the previous Section suppose that $g = 0$ and $\Gamma_\mathrm{N} = \emptyset$, i.e. we have pure homogeneous Dirichlet boundary condition all over the boundary, and for simplicity we omit the reaction term. In this case $L(v) = \int_\Omega fv$.

Let us start with the bilinear form associated with the weak form (see equations (1.7)). Extend it to $V_{*,DG} \times V_{DG}$ in the following sense

$$a^0_{DG}(v, w_h) = \int_\Omega \kappa \nabla_h v \cdot \nabla_h w_h = \sum_{E \in \mathcal{T}_h} \int_E \kappa \nabla v \cdot \nabla w_h.$$

As we have seen in the error estimation for the classical Galerkin method consistency is an important property. It means we have to develop the bilinear and the linear form in such a way that for the exact solution $u \in V_*$ the relation $a_{DG}(u, w_h) = L(w_h)$ holds for all $w_h \in V_{DG}$.

Using Green's Theorem on each element we have

$$a^0_{DG}(v, w_h) = \sum_{E \in \mathcal{T}_h} \int_E -\text{div}\,(\kappa \nabla v) w_h + \sum_{E \in \mathcal{T}_h} \int_{\partial E} (\kappa \partial_{\nu_T} v) w_h. \tag{1.12}$$

Using the fact that for $e \in \Gamma_{_0}$ (where $e = E \cap F$) $\nu_e = \nu_{_E} = -\nu_{_F}$ we can modify the second term of (1.12).

$$\sum_{E \in \mathcal{T}_h} \int_{\partial E} \kappa \partial_\nu v w_h = \sum_{e \in \Gamma_0} \int_e [\![\kappa \partial_\nu v w_h]\!] + \sum_{e \in \Gamma} \int_e \kappa \partial_\nu v w_h.$$

Moreover, $[\![\kappa \partial_\nu u v]\!] = [\![\kappa \nabla_h u \cdot \nu v]\!] = [\![\kappa \nabla_h u v]\!] \cdot \nu$ which means that over $e \in \Gamma_{_0}$ we have

$$\begin{aligned}
[\![\kappa \nabla_h v w_h]\!] =& (\kappa \nabla v w_h)|_E - (\kappa \nabla v w_h)|_F \\
=& \frac{1}{2} \left((\kappa \nabla v)|_E + (\kappa \nabla v)|_F \right) (w_h|_E - w_h|_F) + \\
& \left((\kappa \nabla v)|_E - (\kappa \nabla v)|_F \right) \frac{1}{2} (w_h|_E + w_h|_F) \\
=& \{\!\{\kappa \nabla_h v\}\!\} [\![w_h]\!] + [\![\kappa \nabla_h v]\!] \{\!\{w_h\}\!\}.
\end{aligned}$$

The definition of jumps and averages on the boundary yields

$$\sum_{E \in \mathcal{T}_h} \int_{\partial E} \kappa \partial_\nu v w_h = \sum_{e \in \Gamma_0 \cup \Gamma} \int_e \{\!\{\kappa \nabla_h v\}\!\} \cdot \nu [\![w_h]\!] + \sum_{e \in \Gamma_0} \int_e [\![\kappa \nabla_h v]\!] \cdot \nu \{\!\{w_h\}\!\}.$$

Using this (1.12) can be rewritten as

$$\begin{aligned}
a^0_{DG}(v, w_h) = \sum_{E \in \mathcal{T}_h} \int_E (-\text{div}\,(\kappa \nabla v)\, w_h + \sum_{e \in \Gamma_0 \cup \Gamma} \int_e \{\!\{\kappa \nabla_h v\}\!\} \cdot \nu [\![w_h]\!] \\
+ \sum_{e \in \Gamma_0} \int_e [\![\kappa \nabla_h v]\!] \cdot \nu \{\!\{w_h\}\!\}.
\end{aligned} \tag{1.13}$$

Next we have to check consistency. Since $u \in H^1_0(\Omega) \cap H^2(\Omega)$ we have that $[\![\kappa \nabla_h u]\!] \cdot \nu_e = 0$, $\forall e \in \Gamma_{_0}$ and substituting $u = v$ into (1.13)

$$a^0_{DG}(u, w_h) = \sum_{E \in \mathcal{T}_h} \int_E f w_h + \sum_{e \in \Gamma_0 \cup \Gamma} \int_e \{\!\{\kappa \nabla_h u\}\!\} \cdot \nu [\![w_h]\!].$$

To achieve consistency we should modify the bilinear form

$$a^1_{DG}(v, w_h) = \sum_{E\in\mathcal{T}_h} \int_E \kappa\nabla v \cdot \nabla w_h - \sum_{e\in\Gamma_0\cup\Gamma} \int_e \{\!\{\kappa\nabla_h v\}\!\} \cdot \nu\, [\![w_h]\!]\,.$$

However it can be seen, that this form is neither symmetric nor coercive. To ensure these properties as well, we add two terms to both sides. These terms contain the jump of u. On the left hand side we will use the fact, that these are zeros over the interior faces. Hence,

$$\begin{aligned} a^{hD}_{DG}(v, w_h) &= \sum_{E\in\mathcal{T}_h} \int_E \kappa\nabla v \cdot \nabla w_h - \sum_{e\in\Gamma_0\cup\Gamma} \int_e \{\!\{\kappa\nabla v\}\!\} \cdot \nu_e\, [\![w_h]\!] \\ &\quad + \varepsilon \sum_{e\in\Gamma_0\cup\Gamma} \int_e \{\!\{\kappa\nabla w_h\}\!\} \cdot \nu_e\, [\![v]\!] + \sum_{e\in\Gamma_0\cup\Gamma} \int_e \frac{\sigma}{|e|}\, [\![v]\!]\,, [\![w_h]\!] \qquad (1.14) \\ L^{hD}_{DG}(w_h) &= \int_\Omega f w_h, \end{aligned}$$

where superscript hD stands for homogeneous Dirichlet.

Naturally $a^{hD}_{DG}(\cdot,\cdot)$ is symmetric only if $\varepsilon = -1$. Theoretically ε can be any arbitrary number although in most cases $\varepsilon \in \{-1, 0, 1\}$. The three different choices lead to three different Interior Penalty DG (IPDG) methods:

- $\varepsilon = -1$: the Symmetric IPDG Method,
- $\varepsilon = 0$: the Incomplete IPDG Method,
- $\varepsilon = 1$: the Nonsymmetric IPDG Method.

The effect of non-homogeneous Dirichlet boundary condition comes into play when we add the extra terms to achieve consistency and symmetry. The jump of u along the Dirichlet part of Γ is nonzero, therefore a term caused by the inhomogeneity of the problem appears in the linear form. Neumann boundary condition means that when using Green's Theorem we know the integral $\int_{\Gamma_\mathrm{N}} u_N w_h$, therefore, this term also appears in the linear form. Summing them up we have

$$\begin{aligned} a_{DG}(v, w_h) &= \sum_{E\in\mathcal{T}_h} \int_E \kappa\nabla v \cdot \nabla w_h - \sum_{e\in\Gamma_0\cup\Gamma_\mathrm{D}} \int_e \{\!\{\kappa\nabla v\}\!\} \cdot \nu_e\, [\![w_h]\!] \\ &\quad + \varepsilon \sum_{e\in\Gamma_0\cup\Gamma_\mathrm{D}} \int_e \{\!\{\kappa\nabla w_h\}\!\} \cdot \nu_e\, [\![v]\!] + \sum_{e\in\Gamma_0\cup\Gamma_\mathrm{D}} \int_e \frac{\sigma}{|e|}\, [\![v]\!]\, [\![w_h]\!]\,, \qquad (1.15) \\ L_{DG}(w_h) &= \int_\Omega f w_h + \sum_{e\in\Gamma_\mathrm{D}} \int_e \left(\varepsilon\partial_\nu \kappa w_h + \frac{\sigma}{|e|} w_h\right) g + \sum_{e\in\Gamma_\mathrm{N}} \int_e u_N w_h. \end{aligned}$$

Remark 1.32 *Using the definition of jumps and averages from Remark 1.26 a similar bilinear form can be derived*

$$\begin{aligned}a_{DG}(v,w_h) &= \sum_{E\in\mathcal{T}_h}\int_E \kappa\nabla v\cdot\nabla w_h - \sum_{e\in\Gamma_0\cup\Gamma_D}\int_e \{\!\{\kappa\nabla v\}\!\}\,[\![w_h]\!]\\ &\quad+\varepsilon\sum_{e\in\Gamma_0\cup\Gamma_D}\int_e \{\!\{\kappa\nabla w_h\}\!\}\,[\![\nabla v]\!]+\sum_{e\in\Gamma_0\cup\Gamma_D}\int_e\frac{\sigma}{|e|}\,[\![v]\!]\,[\![w_h]\!]\,,\\ L_{DG}(w_h) &= \int_\Omega fv+\sum_{e\in\Gamma_D}\int_e\left(\varepsilon\partial_\nu\kappa w_h+\frac{\sigma}{|e|}w_h\right)g+\sum_{e\in\Gamma_N}\int_e u_N w_h.\end{aligned}$$

Remark 1.33 *Remark 1.24 stated that even though the jump of a function depends on the orientation of ν it will not be confusing later. For example see* (1.15)*: it contains terms such as $\{\!\{\kappa\nabla v\}\!\}\cdot\nu_e\,[\![w]\!]$ and this is really independent of the orientation of ν. If we take $-\nu$ there will be two changes in the sign.*

Remark 1.34 *Similarly to Remark 1.13 if the linear reaction term is also included in the differential equation we only have to add $\int_\Omega \mu uv$ to the bilinear form and the proofs of the further lemmas become more technical.*

1.5.2 Reduction to finite dimension

As in Section 1.2 $V_{*,DG}$ is infinite dimensional therefore we use the finite dimensional subspace: $V_{DG}=\mathcal{P}_d^p(\mathcal{T}_h)$. The discrete problem is

Problem Set 1.35

$$\begin{cases}\text{Seek } u_{DG}\in V_{DG} \text{ such that}\\ a_{DG}(u_{DG},v_{DG})=L(v_{DG})\quad\forall v_{DG}\in V_{DG}.\end{cases}$$

The first question that can arise is the solvability of Problem Set 1.35. We can use the Lax-Milgram Lemma again, all we have to prove is that $a_{DG}(\cdot,\cdot)$ is bounded over $V_{*,DG}\times V_{DG}$ and coercive over V_{DG}. This will be done in the next Section.

The equation in Problem Set 1.35 has to be satisfied only for the basis of V_{DG}. This leads to a system of linear equations again, although the matrix is symmetric only if $\varepsilon=-1$. In any other cases we have to solve a system of linear equations with a nonsymmetric matrix.

We should define a basis for V_{DG}. We can use the Lagrange elements from Section 1.2, however, thanks to the discontinuity we can use "simpler" basis functions: for example the monomials will be enough. Although, if we want to use basis functions that are equal to zero at the Dirichlet boundary it is necessary to use different functions over the elements that lie on the boundary. We should note that in the DG case the exact matching to the Dirichlet condition is usually dropped (i.e. (1.15) is used instead of (1.14)).

1.5.3 Convergence

If we ought to study convergence we cannot use the $H_0^1(\Omega)$ norm. Instead let us introduce the following norm for $v \in V_{DG}$

$$|||v|||_{DG}^2 := \|\nabla_h v\|_0^2 + \sum_{e \in \Gamma_0 \cup \Gamma_N} \frac{1}{|e|} \| [\![v]\!] \|_{0,e}^2, \tag{1.16}$$

and for $v \in V_{*,DG}$

$$|||v|||_{*,DG}^2 := |||v|||_{DG}^2 + \sum_{E \in \mathcal{T}_h} h_E \|\nabla v|_E \cdot \nu_E\|_{0,\partial E}^2. \tag{1.17}$$

The proper proof of the following Lemmas are exceptionally technical. We refer to [21, Sect. 4.1] for the details.

Lemma 1.36 (Lemma 4.12 [21]) *There exists $\sigma_0 \geq 0$ such that for all $\sigma > \sigma_0$ the bilinear form defined by* (1.15) *is coercive on V_{DG} in the $|||\cdot|||_{DG}$ norm, i.e. $\exists C_c > 0$*

$$a_{DG}(v_{DG}, v_{DG}) \geq C_c \, |||v_{DG}|||_{DG}^2, \qquad \forall v_{DG} \in V_{DG}.$$

Remark 1.37 *It is important to note that in the nonsymmetric case $\sigma_0 = 0$, therefore any $\sigma > 0$ can guarantee coercivity.*

Lemma 1.38 *According to Lemma 1.36 Problem Set 1.35 can be solved in the finite dimensional space V_{DG} if $\sigma > \sigma_0$.*

Lemma 1.39 (Lemma 4.16 [21]) *There exists a constant $C_b > 0$ (independent of h) such that*

$$a_{DG}(v, w_{DG}) \leq C_b \, |||v|||_{*,DG} \, |||w_{DG}|||_{DG}, \qquad \forall (v, w_{DG}) \in V_{*,DG} \times V_{DG}.$$

Lemma 1.40 (Lemma 4.20 [21]) *The $|||\cdot|||_{*,DG}$ and $|||\cdot|||_{DG}$ are (uniformly) equivalent on V_{DG}. $\exists C_{DG} > 0$*

$$C_{DG} \, |||v_{DG}|||_{*,DG} \leq |||v_{DG}|||_{DG} \leq |||v_{DG}|||_{*,DG}, \qquad \forall v_{DG} \in V_{DG}.$$

Remark 1.41 *Using Lemma 1.40 it can be shown that $a_{DG}(\cdot,\cdot)$ is bounded on $V_{DG} \times V_{DG}$.*

Lemma 1.42 (Approximation result) *There exists $C > 0$ such that $\forall u \in H^{p+1}(\Omega)$*

$$|||u - \pi_h u|||_{*,DG} = C h^p |u|_{p+1},$$

where h is the maximal diameter of the mesh elements, $\pi_h u$ is the $L^2(\Omega)$-orthogonal projection of $u \in L^2(\Omega)$ to $\mathcal{P}_d^p(\mathcal{T}_h)$.

Again, for more details on the approximation result see Appendix A.4.

In the continuous case we have seen that the mesh has to satisfy certain conditions, namely it has to be shape regular. In the discontinuous case the mesh conditions are more technical - for the readers' convenience we will skip these technical details and we refer to [21, Sect. 1.4.4-1.4.5].

Theorem 1.43 *Suppose that σ is chosen such that $\sigma > \sigma_0$ (see Lemma 1.36) and the exact solution of Problem Set 1.35 is smooth enough, $u \in V_* \cap H^{l+1}(\Omega)$ for some $l \geq p$. Then there exists $C > 0$ (independent of h) such that for any proper mesh with h as the maximal diameter of the elements the following estimation holds*

$$|||u - u_{DG}|||_{*,DG} \leq Ch^p|u|_{p+1}.$$

proof: Using the consistency, coercivity and boundedness of $a_{DG}(\cdot, \cdot)$ we can derive an error estimation as we did in Section 1.3.1

$$C_c\, |||\pi_h u - u_{DG}|||_{DG}^2 \leq a_{DG}(\pi_h u - u_{DG}, \pi_h u - u_{DG}) = a_{DG}(\pi_h u - u, \pi_h u - u_{DG}) \leq$$

$$C_b\, |||\pi_h u - u|||_{*,DG}\, |||\pi_h u - u_{DG}|||_{DG} \leq C_b C_a h^p |u|_{p+1}\, |||\pi_h u - u_{DG}|||_{DG}\, .$$

If $|||\pi_h u - u_{DG}|||_{DG} = 0$ then the proof is complete according to the approximation result. Otherwise

$$C_{DG}\, |||\pi_h u - u_{DG}|||_{*,DG} \leq |||\pi_h u - u_{DG}|||_{DG} \leq \frac{C_b C_a}{C_c} h^p |u|_{p+1}.$$

Finally the triangle inequality completes the proof

$$|||u - u_{DG}|||_{*,DG} \leq |||u - \pi_h u|||_{*,DG} + |||\pi_h u - u_{DG}|||_{*,DG} \leq \left(C_a + \frac{C_b C_a}{C_{DG} C_c}\right) h^p |u|_{p+1}.$$

□

1.5.4 Error estimation in the $L^2(\Omega)$ norm

Similarly as in Section 1.3.2 we can derive error estimation in the $L^2(\Omega)$ norm for the symmetric IPDG. For the details we refer to [21, Sect.4.2.4]. The final result is the following. Suppose that the weak solution $u \in V_* \cap H^{l+1}$ ($l \geq p$) and u_{DG} solve the discrete problem. In this case:

$$||u - u_{DG}||_0 \leq Ch^{p+1}|u|_{p+1}.$$

Unfortunately the other cases ($\varepsilon \neq -1$) are not adjoint consistent, hence the idea that was used in 1.3.2 cannot be used. The suboptimality of these methods in the $L^2(\Omega)$ norm have been investigated in the recent years. For years it seemed that suboptimality occurs only when polynomial with even degree are used, however, there are counterexamples for odd polynomial degrees as well, see [32] for the nonsymmetric case and [6] for the incomplete one.

1.5.5 Lifting operator

As we have seen $a_{DG}(\cdot,\cdot)$ is defined over $V_{*,DG} \times V_{DG}$. However, the weak solution belongs to $H_0^1(\Omega)$ (in the case of pure homogenous Dirichlet boundary condition).

Let us introduce the function space $\mathcal{V} = V_{DG} + H_0^1(\Omega)$ (where $+$ again denotes the Minkowski addition). We will construct a generalized bilinear form $\widetilde{a}_{DG} : \mathcal{V} \times \mathcal{V} \to \mathbb{R}$. The key ingredient to this is the lifting operator. Let us consider an arbitrary function $v \in \mathcal{V}$. The lifting operator $\mathcal{L} : \mathcal{V} \to \Sigma$ is defined as follows

$$\int_\Omega \mathcal{L}(v) \cdot q = \sum_{e \in \Gamma_0 \cup \Gamma_D} \int_e \{\!\{q\}\!\} \cdot \nu_e [\![v]\!] \qquad \forall q \in \Sigma$$

where $\Sigma = \{q \in [L^2(\Omega)]^d : q|_E \in [\mathcal{P}_d^p(E)]^d, \forall E \in \mathcal{T}_h\}$. It is important to note that $\mathcal{L}(v) \equiv 0$ for $v \in H_0^1(\Omega)$. For more details on this operator we refer to [21, Ch. 4.3]. Let us introduce the modified bilinear form

$$\begin{aligned}\widetilde{a}_{DG}(v,w) = \sum_{E \in \mathcal{T}_h} \int_E \kappa \nabla v \cdot \nabla w - \int_\Omega \mathcal{L}(w) \cdot (\kappa \nabla v) + \varepsilon \int_\Omega \mathcal{L}(v) \cdot (\kappa \nabla w) \\ + \sum_{e \in \Gamma_0 \cup \Gamma_D} \int_e \frac{\sigma}{|e|} [\![v]\!] [\![w]\!] .\end{aligned}$$

It is easy to see, that $\widetilde{a}_{DG}$ is an extension of both $a_{DG}(\cdot,\cdot)$ and $a(\cdot,\cdot)$, more precisely $\widetilde{a}_{DG}(u,v) = a_{DG}^{hD}(u,v) \, \forall u,v \in V_{DG}$ and $\widetilde{a}_{DG}(u,v) = a(u,v) \, \forall u,v \in H_0^1(\Omega)$ due to the fact that for an arbitrary function $u \in H_0^1(\Omega)$ we have that $\mathcal{L}(u) \equiv 0$ and $[\![u]\!] = 0$ over all the edges.

1.5.6 One dimensional case

Similar argument in one dimension can lead to a similar bilinear form. The first problem is the definition of the penalty terms. In (1.15) we have the edge length (interface area in 3D) in the dominator which makes no sense in one dimension.

Let $\Omega = (\alpha, \beta)$, ($\alpha, \beta \in \mathbb{R}$, $\alpha < \beta$), $f \in L^2(\Omega)$ be a given function. For simplicity suppose that the equation is subject to Dirichlet boundary conditions. Let us examine the following second order boundary value problem

$$-(\kappa u')' + \mu u = f \quad \text{in } \Omega, \tag{1.18}$$

$$u(\alpha) = u_\alpha, \quad u(\beta) = u_\beta \tag{1.19}$$

where $\kappa : \mathbb{R} \to \mathbb{R}$, $\kappa \geq \kappa_0 > 0$, $\mu : \mathbb{R} \to \mathbb{R}$ is a nonnegative function, $\kappa, \mu \in L^\infty(\Omega)$. Classical solution means we seek $u \in C^2(\Omega) \cap C(\overline{\Omega})$, for which (1.18) $-$ (1.19) is satisfied.

The mesh $\mathcal{T}_h$ is defined as follows: $\alpha = x_0 < x_1 < x_2 \ldots < x_N = \beta$, $I_i = [x_{i-i}, x_i]$ ($i = 1, \ldots, N$), $\mathcal{T}_h := \cup_i I_i$. Let us denote by $h_i = |I_i|$ the length of the interval. For the interior nodes ($i = 1, \ldots, N-1$) let us use one of the following definitions:

1. $h_{i,i+1} = \max\{h_i, h_{i+1}\}$ ([61, Ch. 1])

2. $h_{i,i+1} = \min\{h_i, h_{i+1}\}$ ([21, Ch. 4, Definition 4.5])

3. $h_{i,i+1} = \frac{1}{2}(h_i + h_{i+1})$ ([32, Sect.1]).

In all three cases we have $h_{0,1} = h_1$, $h_{N,N+1} = h_N$.

Definition 1.44 *Jumps and averages are defined at interior nodes as*

$$[\![v(x_i)]\!] := v|_{I_i} - v|_{I_{i+1}}, \qquad \{\!\{v(x_i)\}\!\} := \frac{1}{2}\left(v|_{I_i} + v|_{I_{i+1}}\right),$$

and on the boundary as

$$[\![v(\alpha)]\!] = -v(\alpha), \quad \{\!\{v(\alpha)\}\!\} = v(\alpha), \quad [\![v(\beta)]\!] = v(\beta), \quad \{\!\{v(\beta)\}\!\} = v(\beta).$$

Remark 1.45 *Strictly speaking the notations in the above definitions are not correct. Mathematically it would be more precise to write e.g. $[\![u]\!]_{x_i}$ instead of $[\![u(x_i)]\!]$, however, it would be less readable. The notations $u|_{I_i}$ should be understood as the trace of the function, or as one side limits as in Section 5.1.2.*

Using these notations

$$\begin{aligned} a_{DG}(u,v) = &\sum_{n=0}^{N-1} \int\limits_{x_n}^{x_{n+1}} \kappa u'(x)v'(x)\,\mathrm{d}x - \sum_{n=0}^{N} \{\!\{\kappa u'(x_n)\}\!\}\,[\![v(x_n)]\!] \\ &+ \varepsilon \sum_{n=0}^{N} \{\!\{\kappa v'(x_n)\}\!\}\,[\![u(x_n)]\!] + \sum_{n=0}^{N} \frac{\sigma}{h_{n,n+1}}\,[\![v(x_n)]\!]\,[\![u(x_n)]\!] + \int_{\alpha}^{\beta} \mu u(x)v(x)\,\mathrm{d}x, \\ L_{DG}(v) = &\int_a^b f(x)v(x)\,\mathrm{d}x - \varepsilon\kappa v'(\alpha)u_\alpha + \varepsilon\kappa v'(\beta)u_\beta + \frac{\sigma}{h_1}v(\alpha)u_\alpha + \frac{\sigma}{h_N}v(\beta)u_\beta. \end{aligned}$$

This will play an important role in Chapter 5.

1.6 Adaptivity

1.6.1 Adaptive finite element strategies

One way to achieve the smallest error is to use adaptivity: after solving the problem on a given mesh with a given polynomial degree it is possible to make an a-posteriori error estimation. If the error is smaller than a given tolerance, we accept the solution, otherwise we redefine the mesh and/or the degree of the polynomial and solve the discrete problem again. A-posteriori error estimation differs from the a-priori estimation: it only contains data that are available at hand. Briefly: it does not contain the unknown solution u.

There are four main versions of adaptivity:

h-**adaptivity**: in this case the polynomial degree do not change anywhere over the mesh, but the mesh itself changes. We refine the elements of the mesh where the error is huge (in some sense).

p-**adaptivity**: in this case the polynomial degree changes over the elements (it is increased) where the error is too big. In this case the mesh is fix.

hp-**adaptivity**: this one combines the two above methods. Nothing is fix and in an adaptive step both the mesh and the polynomial degree can change. In some methods they are varied separately: over a given element it is an h-adaptive step or a p-adaptive. However, there are methods where h- and p-adaptivity is done in parallel over the elements.

r-**adaptivity**: in this method the mesh changes, but on a different way as in h-adaptivity: in an r-adaptive method the mesh nodes move. This version is not included in this thesis.

The hp-adaptivity is the only one of them that gives exponential convergence [29, 30, 31, 55]. Obviously it is the most challenging because on one element of the mesh we have enormous number of possible refinements. In the h- or p-version there is only one possible step and the algorithm should only decide to do it or not. For this version all we need is the norm of the error (a number) to decide. However, in the hp-case some more information is needed. Some knowledge about the shape of the error.

The adaptive finite element algorithms are based on the following scheme:

Initialize: solve the initial problem with small polynomial degree p on a coarse grid

Repeat:

S1 estimate the error

S2 if the error is small then stop

S3 else determine on which elements in the grid and how to refine/derefine

S4 compute the new solution and go to S1

The main differences between the different methods are in the error estimation and refinement procedures. For a wide range of error estimation processes we refer to [3] and to the references in [55] to the original papers on the hp-adaptive methods.

Derefinement means that the polynomial degree is decreased or some mesh elements are melted into a bigger one. It is used on elements where the error is small and the aim is to control the growth of the number of unknowns. The basic idea behind this is the following: where the error is the smallest all over the domain we could use less unknowns without significantly losing accuracy.

1.6.2 A-posteriori error estimation techniques

The construction of accurate a-posteriori error estimators for the finite element solution of PDE's is of great importance. Besides providing a reliable stopping criterion for the successive refinements, a-posteriori error estimation also gives a solid basis of adaptive finite element algorithms [19], [64]. From this point of view, local a-posteriori error estimates are of particular importance. For a general overview on a-posteriori error estimators we refer to [3, 25, 60, 73].

The starting point of many error estimation techniques is the residual-based a-posteriori error estimator, which provides an explicit formula for the error. The original idea in [7] has been generalized for several types of equations, such as advection-diffusion [74], convection-diffusion-"reaction" [75] and Maxwell equations [63]. Accordingly, explicit error estimators have been provided for nonconforming finite element methods [4] and uniform approaches have been developed [14]. Moreover, the estimation methodology can be extended for nonlinear problems, see, *e.g.* [16] and [46].

Another approach is given by the *functional type* a-posteriori error estimates. These can provide both an upper and a lower bound for the exact error and are free of unknown constant (depending on the mesh geometry or interpolation inequalities). Usually, these estimates are independent of the numerical technique used to obtain approximate solutions, and they can be extended to nonlinear elliptic problems as well [47]. For more information and relevant references we refer to the monograph [59].

For the *implicit a-posteriori error estimators* Neumann type problems are formulated locally using the numerical solution at hand, and these are solved in certain local finite element spaces. In the simplest case, the boundary conditions for the local problems have been constructed with a simple averaging on element interfaces. To enforce the well-posedness of the local problems or enhance the quality of the estimators special equilibrated fluxes were defined and analyzed ([8], [51]) using the results for the residual-based explicit error estimators. Though it seems to be an involved approach, it pays off to compute an accurate error estimator which provides local error bounds and is sensitive to the shape of the subdomain or to the mesh geometry. Implicit a-posteriori error estimators have been applied and analyzed for elliptic boundary value problems (see an overview in [3]) and generalized for time-harmonic Maxwell equations [45].

Another family of powerful methods for a-posteriori error estimation can be obtained using *gradient averaging techniques* [13], which result in simple and computationally cheap estimates [50]. In another context, they are called recovery techniques, as the aim is to give an approximation to the gradient of the exact solution of the original problem [77], [78]. Gradient averaging techniques [2], [3] can deliver reliable a-posteriori error control [36] even on unstructured grids [15] and they can be used in goal-oriented error estimations [49]. The accuracy of the a-posteriori error indicators can be enhanced using a superconvergent gradient recovery technique, see [9] and [10].

1.7 Maximum principle

1.7.1 Introduction

When choosing a numerical method to approximate the solution of a continuous mathematical problem, we need to consider which method results in an approximation that is not only close to the solution of the original problem, but possesses the important qualitative properties of the original problem, too. For linear elliptic problems the main qualitative properties are the various maximum principles. The preservation of the weak maximum principle was extensively investigated in the last decades, but not the preservation of strong maximum principle. In Chapter 4 we focus on the latter property by giving its necessary and sufficient conditions, investigating the relation of the preservation of the strong and weak maximum principles and illustrating the differences between them with numerous examples.

In the early theory of PDE's maximum principles played an important role. They provide an efficient tool to prove uniqueness and stability for the classical solutions of linear elliptic and parabolic problems. Later, when the concept of weak solution had been introduced, they lost a little bit from their importance. Now, in the age of computers and numerical methods, the investigation of maximum principles came into fashion again.

A numerical method is a sequence of simpler problems, whose solutions hopefully tend to the solution of the original problem. When this holds, the numerical method is called convergent. However, convergence is a theoretical question, in the application we must choose some parameter settings, and not an infinite sequence of it. Thus, we need to decide between convergent numerical methods from another point of view. This leads to the investigation of what qualitative properties can be preserved when we apply certain numerical method, e.g. we usually prefer one in which the simpler problems possess the same important qualitative properties as the original problem.

Maximum principles are essential qualitative properties of linear elliptic problems. When for a simpler problem the maximum principle holds, we say that it possesses the discrete maximum principle, since the simpler problems are usually defined in finite dimensional spaces.

The first paper in which a discrete maximum principle was formulated is probably [70]. The definition of the discrete weak maximum principle which is used today appeared first in [17] (but it was named differently).

1.7.2 Continuous maximum principle for elliptic operators

We formulate the maximum principle for operators, following the book [26], instead of defining it for equations. Naturally, there are no important differences between the two approaches, but our choice is simpler to handle.

Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain with boundary $\partial\Omega$, and $\overline{\Omega} = \Omega \cup \partial\Omega$ its closure. We investigate the elliptic operator $\mathcal{A}$, $\operatorname{dom}\mathcal{A} = C^2(\Omega) \cap C(\overline{\Omega})$, defined in divergence form as

$$\mathcal{A}u = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(\mathcal{K}_{ij}\frac{\partial u}{\partial x_i}\right) + \mu u, \tag{1.20}$$

where $\mathcal{K}_{ij} \in C^1(\Omega)$, $0 \leq \mu \in C(\Omega)$. Note that the smoothness of the coefficient functions gives the opportunity to rewrite (1.20) to a non-divergence form that is more suitable for the investigation of maximum principles.

Definition 1.46 *We say that the operator $\mathcal{A}$ defined in* (1.20) *possesses the continuous weak maximum principle if for all $u \in C^2(\Omega) \cap C(\overline{\Omega})$ the following implication holds*

$$\mathcal{A}u \leq 0 \textit{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} u\}.$$

Theorem 1.47 *([Ch. 6.4, Th.2[26]]) If operator $\mathcal{A}$ defined in* (1.20) *is uniformly elliptic, see Definition 1.2, and $\mu \geq 0$, then it possesses the continuous weak maximum principle.*

1.8 Maximum principle for FEM elliptic operators – short overview

1.8.1 The construction of the FEM elliptic operator

When discretising the operator (1.20) with finite element method we have to define the corresponding bilinear form as before

$$a(u,v) = \int_{\Omega} \sum_{i,j=1}^{d} \mathcal{K}_{ij}\frac{\partial u}{\partial x_i}\frac{\partial v}{\partial x_j} + \mu uv, \tag{1.21}$$

where $u \in H^1(\Omega), v \in H_0^1(\Omega)$.

We note that this means we deal with non-homogeneous Dirichlet boundary condition, for the homogeneous one see Remark 1.52.

Remark 1.48 *It seems as if we would handle non-homogeneous Dirichlet boundary condition differently from the way we did in Remark 1.6, however, the two ways are the same. In that Remark we looked for $u \in H^1(\Omega)$ such that $u = w + u_g \quad w \in H_0^1(\Omega) : a(w,v) = L(v) - a(u_g,v) \quad \forall v \in H_0^1(\Omega)$, where u_g satisfies the non-homogeneous Dirichlet boundary condition, and w satisfies the homogeneous one. If we add $a(u_g,v)$ to both sides and we use the linearity of $a(\cdot,\cdot)$ we get $a(w+u_g,v) = a(u,v)$, where $u \in H^1(\Omega), v \in H_0^1(\Omega)$.*

The following step is to define a mesh on Ω. A 1D mesh consists of intervals. The discrete maximum principle literature focuses on regular triangle or hybrid meshes (containing both triangles and rectangles) in 2D and tetrahedron or block meshes in 3D. A given mesh determines the sets $\mathcal{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$ and $\mathcal{X}_\partial = \{\mathbf{x}_{N+1}, \mathbf{x}_{N+2}, \ldots, \mathbf{x}_{N+N_\partial}\}$ containing the vertices in Ω and on $\partial\Omega$, respectively. Let us introduce two more notations: $\overline{N} = N + N_\partial$ and $\overline{\mathcal{X}} = \mathcal{X} \cup \mathcal{X}_\partial$.

Next we can define a subspace of $H^1(\Omega)$ corresponding to the mesh. This can be done by giving a basis of this subspace. The basis functions are denoted by Φ_i, $i = 1, \ldots \overline{N}$. The discrete maximum principle literature investigates almost solely the case of hat-functions which are defined with the following properties:

1. the basis functions are continuous;

2. the basis functions are piecewise linear functions over intervals/triangles/tetrahedrons and multilinear over rectangles/blocks;

3. $\Phi_i(\mathbf{x}_i) = 1$ for $i = 1, \ldots \overline{N}$

4. $\Phi_i(\mathbf{x}_j) = 0$ for $i, j = 1, \ldots \overline{N}$, $i \neq j$.

Note that this choice has some consequences:

1. the subspace consists of continuous functions;

2. $\sum_{i=1}^{\overline{N}} \Phi_i(\mathbf{x}) = 1$ holds for all $\mathbf{x} \in \overline{\Omega}$;

3. $\Phi_i(\mathbf{x}) \geq 0$ holds for all $\mathbf{x} \in \overline{\Omega}$ and $i = 1, \ldots \overline{N}$;

4. in a linear combination of the basis functions the coefficients represent the values of the resulting function at the points of $\overline{\mathcal{X}}$.

We remark that for higher order elements the investigation is more difficult, and positive results are obtained only for a simple 1D problem, see [72]; for a higher dimensional case, in [37] negative results are obtained.

Finally, we can construct the so-called stiffness matrix $\mathbf{A} \in \mathbb{R}^{N \times \overline{N}}$ as

$$\mathbf{A}_{ij} = a\left(\Phi_j, \Phi_i\right).$$

that is the discrete operator corresponding to (1.20).

In the following it will be useful to introduce the partitioned form $\mathbf{A} = [\mathbf{A}_0 | \mathbf{A}_\partial]$, where $\mathbf{A}_0 \in \mathbb{R}^{N \times N}$, $\mathbf{A}_\partial \in \mathbb{R}^{N \times N_\partial}$, acting on the vector $\mathbf{u} = [\mathbf{u}_0 | \mathbf{u}_\partial]^T \in \mathbb{R}^{\overline{N}}$, $\mathbf{u}_0 \in \mathbb{R}^N$, $\mathbf{u}_\partial \in \mathbb{R}^{N_\partial}$, which is constructed by taking into consideration the separation of the (discrete) interior and boundary nodes.

1.8.2 Maximum principle for FEM elliptic operators

To formulate the corresponding discrete maximum principle we introduce some notations. The symbol $\mathbf{0}$ denotes the zero matrix (or vector), $\mathbf{e}$ is the vector all coordinates of which are equal to 1. The dimensions of these vectors and matrices will be clear from the context. Inequalities $\mathbf{B} \geq \mathbf{0}$ or $\mathbf{a} \geq \mathbf{0}$ means that all the elements of $\mathbf{B}$ or $\mathbf{a}$ are nonnegative. By $\max \mathbf{a}$ we denote the maximal coordinate of the vector $\mathbf{a}$.

Now we are ready to define the corresponding discrete maximum principle for the matrix $\mathbf{A}$.

Definition 1.49 *([17]) We say that a matrix $\mathbf{A}$ has the discrete weak maximum principle if the following implication holds:*

$$\mathbf{Au} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u} \leq \max\{0, \max \mathbf{u}_\partial\}.$$

Note that this definition is adequate only because the chosen basis functions have special properties, the nonnegativity of the numerical solution coincides with the nonnegativity of the solution vector of the linear system. For higher order basis functions this definition is not applicable, because those basis functions are not nonnegative, therefore there exists a solution vector which is nonnegative, but it results in a solution that can be negative between mesh nodes.

It is relatively easy to give sufficient and necessary conditions to ensure that a matrix possesses the discrete weak maximum principle.

Theorem 1.50 *([17]) The matrix $\mathbf{A}$ possesses the discrete weak maximum principle if and only if the following three conditions hold:*

(T1) $\mathbf{A}_0^{-1} \geq \mathbf{0}$; (T2) $-\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}$; (T3) $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \leq \mathbf{e}$.

Theorem 1.50 is a theoretical result and is difficult to apply directly. Usually these conditions are relaxed with the following practical conditions.

Theorem 1.51 *([17]) The matrix $\mathbf{A}$ possesses the discrete weak maximum principle if the following three conditions hold:*

(P1) $\mathbf{A}_0$ *is a nonsingular M-matrix;* (P2) $-\mathbf{A}_\partial \geq \mathbf{0}$; (P3) $\mathbf{Ae} \geq \mathbf{0}$.

For the definition of M-matrix see Definition A.14. The reader can find a detailed text and a plentiful reference list about the discrete maximum principle in [71]. For attempts to use less restrictive practical conditions we recommend the papers [62] and [34].

Remark 1.52 *Note that if we apply the homogeneous Dirichlet boundary condition – this is the case when we eliminate the boundary condition at the continuous level – then the matrix $\mathbf{A}_\partial$*

has no effect, which results in that we need to guarantee (T1) or (P1) only. This milder property has its own name, the so-called nonnegativity preservation property.

If we want to handle the homogeneous Dirichlet boundary condition abstractly, we have to introduce a new bilinear form a^{hD}*, formally the same as* (1.21) *with the exception that it is defined for* $u \in H_0^1(\Omega), v \in H_0^1(\Omega)$*. Then by the discretisation we simply do not have* $\mathbf{A}_\partial$*.*

Chapter 2

Implicit error estimation using higher order fitting

The estimators that will be examined in this chapter are the implicit error estimation and the gradient averaging, see Section 1.6 or [3]. We will combine the two approaches: we are going to prove that using an accurate approximation of the gradient - obtained with a feasible gradient averaging technique or patch recovery operator - as the Neumann type boundary condition for the local problems, results in a reliable implicit a-posteriori error estimation. The favor of our approach is that we do not need any link with the explicit estimators, which gives a freedom in the choice of the above operators. Moreover, the polynomial degree of the boundary data is related to the degree of elements in the local problems. We could also eliminate strict assumptions about the mesh geometry such as the need of parallel meshes. This chapter is based on [39].

Let us denote by $e_{h,p}$ and $\hat{e}_{h,p}$ the analytic error and the error estimator, respectively. In the literature their norms are usually related. At the same time, our result provides a better comparison: we derive an upper bound for $e_{h,p} - \hat{e}_{h,p}$ in the corresponding energy norm.

We will investigate (1.1)-(1.2) with no Neumann boundary condition and $\mathcal{K} = I$, $0 < \mu \in \mathbb{R}$. $\Omega \subset \mathbb{R}^d$, $f \in L^2(\Omega)$, $g \in H^{1/2}(\partial\Omega)$, $\Gamma = \partial\Omega$ is the boundary of Ω. Using these notations the boundary value problem takes the form

$$-\Delta u + \mu u = f \quad \text{in } \Omega, \tag{2.1}$$

$$u = g \quad \text{on } \Gamma. \tag{2.2}$$

The finite dimensional problem associated with (2.1)-(2.2) is

Problem Set 2.1

$$\begin{cases} \textit{Seek } u_{h,p} \in V_{h,p} \textit{ such that} \\ a(u_{h,p}, v_{h,p}) = L(v_{h,p}) \quad \forall v_{h,p} \in V_{h,p}, \end{cases}$$

where $a(u, v) = \int_\Omega \nabla u \cdot \nabla v + \int_\Omega \mu uv$, $L(v) = \int_\Omega fv - a(u_{D_{h,p}}, v)$. As in Section 1.2.1 $u_{D_{h,p}}$ is the approximation of g using the finite dimensional subspace $V_{h,p}$.

The implicit error estimation is based on the equation that could be derived for the computational error $e_{h,p} := u - u_{h,p}$ on every subdomain $T \in \mathcal{T}_h$. Throughout this chapter subdomain will always stand for a union of elements in the tessellation $\mathcal{T}_h$. Using the fact that $-\Delta e_{h,p} + \mu e_{h,p} = -\Delta(u - u_{h,p}) + \mu(u - u_{h,p})$ we have

$$-\Delta e_{h,p} + \mu e_{h,p} = f - (-\Delta u_{h,p} + \mu u_{h,p}) \qquad \text{in } T, \tag{2.3}$$

$$\partial_\nu e_{h,p} = \partial_\nu (u - u_{h,p}) = \partial_\nu u - \partial_\nu u_{h,p} \quad \text{on } \partial T. \tag{2.4}$$

The right hand side of the boundary condition (2.4) is, however, in general unknown. Therefore, we should approximate it. As a first attempt a simple average of $\partial_\nu u_{h,p}$ on the common face of two neighbouring subdomains may be used to approximate $\partial_\nu u$. The corresponding implicit error estimator can be related to explicit ones. This paves the way to prove its reliability and local efficiency up to the approximation of the data, see [3, 7, 73]. Similar results can be obtained for Maxwell equations, see [35, 45].

Some observations, however, motivated us to develop the above approach for the approximation of $\partial_\nu u$ (or, equivalently, that of $\partial_\nu e_{h,p}$) on the element interfaces.

- Using polynomials of degree p to solve the original problem in (2.1)-(2.2) the simple averaging on the element interfaces delivers a polynomial approximation for ∇u of degree $p - 1$. At the same time, it is advised that the local problems in (2.3) have to be solved using a higher order finite element space than the original one [1]. This would require a Neumann type boundary condition for the error of order p.

- The local problems in (2.3) could be ill-posed for $\mu = 0$, or the local error bound may lead to a crude overestimate of the error (see [3, Sect. 6.2]).

- On the other hand, in an automatic mesh refinement technique the mesh size of the neighbouring elements can be highly different. Then a simple average (or even a convenient averaging technique) does not provide an accurate approximation of the gradient.

We will construct an error estimator

$$\hat{e}_{h,p} : \bigcup_{T \in \mathcal{T}_h} \to \mathbb{R}$$

such that $\hat{e}_{h,p}$ is a polynomial of degree $p + 1$ for all subdomain $T \in \mathcal{T}_h$. Note that $\hat{e}_{h,p}$ is not necessarily continuous over the element interfaces. As it is usual for the local error indicators, we use the patch $\tilde{T} \in \mathcal{T}_h$ of T to the construction, where

$$\tilde{T} = int \bigcup_{\substack{T_j \in \mathcal{T}_h \\ T \cap T_j \neq \emptyset}} T_j$$

and we use the notation $V_{h,p}(\tilde{T})$ for the restriction of $V_{h,p}$ to the patch $\tilde{T}$. For a suitable approximation

$$G_{p,T}(u_{h,p}) \approx \nabla u_T$$

according to (2.3) the error estimator $\hat{e}_{h,p}$ is defined as the finite element solution of the boundary value problem

$$-\Delta\hat{e}_{h,p} + \mu\hat{e}_{h,p} = f + \Delta u_{h,p} - \mu u_{h,p} \qquad \text{in } T, \tag{2.5}$$

$$\partial_\nu \hat{e}_{h,p} = \nu \cdot G_{p,T}(u_{h,p}) - \partial_\nu u_{h,p} \quad \text{on } \partial T, \tag{2.6}$$

where the right hand side is known, see also Section 2.4.

2.1 Assumptions on the gradient averaging

We investigate the discrete gradient operator

$$G_{p,T} : W^{1,\infty}(\tilde{T}) \to [L^1(T)]^d,$$

where p denotes the dependence on the local polynomial degree of the finite element space $V_{h,p}$. Accordingly, we define also $G_p : W^{1,\infty}(\mathcal{T}_h) \to [L^1(\mathcal{T}_h)]^d$ with $G_p|_T = G_{p,T}$ for all $T \in \mathcal{T}_h$. While the first three assumptions are borrowed from [3, Sect. 4], the fourth one which streamlines the analysis at many places, is specific for our method:

(A1) $G_{p,T}(v)$ depends only on $v|_{\tilde{T}}$,

(A2) $G_{p,T} : W^{1,\infty}(\tilde{T}) \to [L^1(T)]^d$ is continuous,

(A3) If $u \in \mathcal{P}_{p+1}(\tilde{T})$ then $G_{p,T}(\mathcal{I}_{h,p,\tilde{T}}u) = \mathcal{I}_{h,p,T}\nabla u_T$,

(A4) $G_{p,T}(u_{h,p})$ is a gradient, i.e. there is a function $\mathcal{G}_p(u_{h,p}) \in W^{1,\infty}(\Omega)$ such that $G_{p,T}(u_{h,p}) = \nabla\mathcal{G}_p(u_{h,p})|_T$.

Similarly as before $\mathcal{I}_{h,p,T}$ is the interpolation operator, using polynomials of degree p on the subdomain T, h stands for the mesh size. $\mathcal{P}_{p+1}(T)$ denotes the piecewise polynomials over the subdomain T with degree $p + 1$.

An extra condition which can imply the superconvergence is the following.

(SC) There exists a constant $C(u)$ depending on u such that for some $\tau \geq 0$ we have

$$||\nabla(u_{h,p} - \mathcal{I}_{h,p}u)||_0 \leq C(u)h^{p+\tau}, \tag{2.7}$$

for all $h > 0$.

Remark 2.2 *If $\tau = 0$ then (SC) does not imply superconvergence and the constant C does not depend on u. This case should not be considered as an assumption, since the inequality* (2.7) *is a consequence of the standard finite element interpolation theory, see [12, Ch. 4]. At the same time, the error estimate in Theorem 1 still delivers an accurate upper bound.*

Remark 2.3 *Unlike in the flux equilibration technique we do not assume that the Neumann type boundary conditions would be continuous on the element interfaces.*

2.2 Convergence of the error estimation

In the consecutive analysis, we use the following result which can be obtained at once using a density argument.

Proposition 2.4 *For any $w \in H^1(\Omega)$ we have $\Delta w \in H^{-1}(\Omega)$ and the following estimate is valid*

$$||\Delta w||_{-1} \leq ||\nabla w||_0.$$

We also recall a continuity estimate for elliptic boundary value problems.

Proposition 2.5 *For an arbitrary Lipschitz domain $\Omega \subset \mathbb{R}^d$ with any function $f \in H^{-1}(\Omega)$ the boundary value problem*

$$\begin{aligned} -\Delta u + \mu u &= f \quad \text{in } \Omega, \\ \partial_\nu u &= 0 \quad \text{on } \partial\Omega, \end{aligned}$$

has a unique solution in $H^1(\Omega)$ and the following estimate holds

$$||u||_1 \leq c_{-1}||f||_{-1}, \tag{2.8}$$

where the constant c_{-1} depends only on Ω.

For the proof in a more general context, we refer to [52, Th. 4.10.], □

Remark 2.6 *Since $L^2(\Omega) \subset H^{-1}(\Omega)$ Proposition 2.5 also holds for every $f \in L^2(\Omega)$. Therefore, this allows us to extend the regularity requirements in* (2.1)-(2.2)*, it is enough to have $f \in H^{-1}(\Omega)$ and $g \in H^{-1/2}(\partial\Omega)$, see [39].*

The following proposition concerning the accuracy of the gradient averaging is proved in [3].

Proposition 2.7 *Assume that the gradient averaging operator G_p satisfies $(A1)$, $(A2)$ and $(A3)$, and also, $u \in H^{p+2}(\Omega)$ and (SC) hold. Then*

$$\|\nabla u - G_p(u_{h,p})\|_0 \leq C(u)h^{p+\tau}|u|_{p+2}$$

is valid, where the exponent τ is given in (SC).

Theorem 2.8 (Poincare inequality) *Let f be a linear form on $H^1(\Omega)$ whose restriction on constant functions is not zero. Then, there is $C_p > 0$ such that*

$$\|v\|_1 \leq C_p(\|v\|_0 + |f(v)|) \qquad \forall v \in H^1(\Omega).$$

For more details we refer to [25, Sect. B.3.7].

Using the above results we can state the main statement on the accuracy of our error estimator.

Theorem 2.9 *Assume that the conditions $(A1)$, $(A2)$, $(A3)$, $(A4)$ and (SC) hold. Then we have the following estimate about the precision of the error estimate in* (2.5)-(2.6)

$$\sum_{T\in\mathcal{T}_h} \|e_{h,p} - \hat{e}_{h,p}\|^2_{1,T} \leq 2(c_{-1}(1+\mu C_P) + C_P)C^2(u)h^{2(p+\tau)}|u|^2_{p+2},$$

where C_P comes from the Poincare inequality (see Theorem 2.8). Indeed, the constant c_{-1} depends also on μ.

proof: First we note that according to (A4) one can assume that $\mathcal{G}_{p,T}(u_{h,p})$ is chosen such that $\int_T u - \mathcal{G}_{p,T}(u_{h,p}) = 0$ and therefore, the Poincare inequality implies

$$\|u - \mathcal{G}_{p,T}(u_{h,p})\|_{1,T} \leq C_P\|\nabla(u - \mathcal{G}_{p,T}(u_{h,p}))\|_{0,T} = C_P\|\nabla u - G_{p,T}(u_{h,p})\|_{0,T}, \tag{2.9}$$

where C_P depends only on $T \in \mathcal{T}_h$. The corresponding linear form that fulfils the requirement in Theorem 2.8 is $f(v) = \frac{1}{|\Omega|}\int_\Omega v$.

Taking the difference of (2.3) and (2.5) we have that for any $T \in \mathcal{T}_h$

$$\Delta(e_{h,p} - \hat{e}_{h,p}) - \mu(e_{h,p} - \hat{e}_{h,p}) = 0 \quad \text{in } T,$$

and therefore, using (2.3) and (2.5) again, for every subdomain $T \in \mathcal{T}_h$ we obtain

$$\begin{aligned}
&\Delta[(e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p}))] - \mu((e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p})))\\
&= -\Delta(u - \mathcal{G}_{p,T}(u_{h,p})) + \mu(u - \mathcal{G}_{p,T}(u_{h,p})) \quad \text{in } T,\\
&\partial_\nu[(e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p}))] = 0 \quad \text{on } \partial T.
\end{aligned}$$

The estimates in (2.8), Proposition 2.4 and (2.9) give that

$$\begin{aligned}
&||(e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p}))||_{1,T} \\
&\quad \leq c_{-1}||\Delta(u - \mathcal{G}_{p,T}(u_{h,p}))||_{-1,T} + c_{-1}\mu||u - \mathcal{G}_{p,T}(u_{h,p})||_{-1,T} \\
&\quad \leq c_{-1}||\Delta(u - \mathcal{G}_{p,T}(u_{h,p}))||_{-1,T} + c_{-1}\mu||u - \mathcal{G}_{p,T}(u_{h,p})||_{1,T} \\
&\quad \leq c_{-1}(1 + \mu C_P)||\nabla u - G_{p,T}(u_{h,p})||_{0,T}.
\end{aligned}$$

Hence, the convergence result in Proposition 2.7 provides the estimate

$$\begin{aligned}
&\sum_{T\in\mathcal{T}_h} ||(e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p}))||_{1,T}^2 \leq c_{-1}(1 + \mu C_P) \sum_{T\in\mathcal{T}_h} ||\nabla u - G_{p,T}(u_{h,p})||_{0,T}^2 \\
&= c_{-1}(1 + \mu C_P)||\nabla u - G_p(u_{h,p})||_0^2 \leq c_{-1}(1 + \mu C_P)C^2(u)h^{2(p+\tau)}|u|_{p+2}^2.
\end{aligned} \tag{2.10}$$

With the aid of a triangle inequality, applying (2.10), (2.9) and Proposition 2.7 again we conclude that

$$\begin{aligned}
\sum_{T\in\mathcal{T}_h} ||e_{h,p} - \hat{e}_{h,p}||_{1,T}^2 &\leq 2 \sum_{T\in\mathcal{T}_h} ||(e_{h,p} - \hat{e}_{h,p}) - (u - \mathcal{G}_{p,T}(u_{h,p}))||_{1,T}^2 + ||u - \mathcal{G}_{p,T}(u_{h,p})||_{1,T}^2 \\
&\leq 2 \sum_{T\in\mathcal{T}_h} c_{-1}(1 + \mu C_P)||\nabla u - G_{p,T}(u_{h,p})||_{0,T}^2 + C_P||\nabla u - G_{p,T}(u_{h,p})||_{0,T}^2 \\
&= (2c_{-1}(1 + \mu C_P) + 2C_P)||\nabla u - G_p(u_{h,p})||_0^2 \\
&\leq (2c_{-1}(1 + \mu C_P) + 2C_P) \cdot C^2(u)h^{2(p+\tau)}|u|_{p+2}^2,
\end{aligned}$$

as stated in the theorem. □

Remark 2.10 *Observe that the estimator in Theorem 2.9 provides not only a relation between $\sum_{T\in\mathcal{T}_h} ||e_{h,p}||_{1,T}^2$ and $\sum_{T\in\mathcal{T}_h} ||\hat{e}_{h,p}||_{1,T}^2$ but also an upper bound for the difference $\sum_{T\in\mathcal{T}_h} ||e_{h,p} - \hat{e}_{h,p}||_{1,T}^2$.*

2.3 Gradient recovery using higher order fitting

We discuss in this section two-dimensional examples so that $\{\mathcal{T}_h\}$ denotes a shape-regular family of geometrically conforming triangular meshes of a polygon $\Omega \subset \mathbb{R}^2$. Let $T \in \mathcal{T}_h$ denote an arbitrary triangle for some h. Note that the approach could be extended to higher space dimensions as well.

The symbol Ω_0 stands for a reference triangle with $\mathcal{J}_T : \Omega_0 \to T$, an affine linear mapping, which is invertible and onto and has the form

$$\mathcal{J}_T = J_T + C_T,$$

where C_T is constant and J_T is linear.

The mapping between a reference patch $\tilde{\Omega}_0$ and $\tilde{T}$ is given by

$$[J_T + C_T, J_{T_1} + C_T, J_{T_2} + C_T, J_{T_3} + C_T], \tag{2.11}$$

where

$$J_T(T \cap T_j) = J_{T_j}(T \cap T_j),\ j = 1, 2, 3,$$

i.e. the affine linear mappings in (2.11) match continuously.

If the triangle T_i degenerates into a boundary edge e_i the transformation J_{T_i} is identified with J_T. This applies also in the proof of Lemma 2.11 and Lemma 2.16. Before introducing gradient recovery techniques which satisfy the assumptions (A1)-(A4), we provide sufficient conditions to verify (A2).

A natural requirement for the gradient recovery is that it is transformed as the gradient by changing the coordinate system, i.e. in precise terms, for any $T \in \mathcal{T}_h$ and $u \in L^1(T)$ we have

(B1) $J_T^{-1} \cdot [G_{p,K}(u \circ \tilde{J}_T)](J_T^{-1}(x)) = [G_{p,T}(u)](x) \quad x \in T.$

As we cannot provide in general a linear or affine linear bijection between patches, an extra condition is necessary, which ensures a continuity property of the gradient recovery.

(B2) Assume that for a sequence $(\tilde{T}_n) = \text{int } (T_n \cup T_{n,1} \cup T_{n,2} \cup T_{n,3})$ of patches and for the corresponding mappings we have the convergence

$$[J_{T_n}, J_{T_{n,1}}, J_{T_{n,2}}, J_{T_{n,3}}] \to [J_T, J_{T_1}, J_{T_2}, J_{T_3}].$$

Then for any polynomial $u \in \mathcal{P}_p(\tilde{T})$ we have the convergence

$$G_p(u_n)(J_{T_n}(x)) \to G_p(u)(J_T(x)),\ x \in K,$$

where the polynomial $u_n \in \mathcal{P}_p(\tilde{T}_n)$ is defined piecewise with $u_n|_{T_n} = u \circ J_T \circ J_{T_n}^{-1}$ and $u_n|_{T_{n,j}} = u \circ J_{T_j} \circ J_{T_{n,j}}^{-1}$, $j = 1, 2, 3$.

We point out that these two assumptions imply $(A2)$ and they are easy to verify.

Lemma 2.11 *Assume that* $(B1)$ *and* $(B2)$ *hold. Then assumption* $(A2)$ *is also valid.*

proof: We consider the orthogonal decomposition

$$V_{h,p}(\tilde{T}) = \mathbf{1} \oplus V_{h,p,0}(\tilde{T})$$

in the L^2-sense, where $\mathbf{1}$ denotes the subspace of constant functions in $V_{h,p}(\tilde{T})$ and $u \in V_{h,p,0}(\tilde{T})$ stands for the functions $u \in V_{h,p}(\tilde{T})$ with $\int_{\tilde{T}} u = 0$. Since the inequality in $(A2)$ is valid for all constant functions, it is sufficient to prove it for functions in $V_{h,p,0}(\tilde{T})$.

Proving by contradiction we assume that there is a sequence $T^1, T^2, \ldots$ of triangles and piecewise polynomials $v_1 \in V_{h,p,0}(\tilde{T^1}), v_2 \in V_{h,p,0}(\tilde{T^2}), \ldots$ with $\|\nabla v_j\|_{L^1(\tilde{T^j})} = 1$ such that the gradient averaging is not bounded; i.e. for each positive integer j we have the inequality

$$\|G_p(v_j)\|_{L^1(T^j)} \geq j\|\nabla v_j\|_{L^1(\tilde{T^j})} = j.$$

Using (B1) we obtain the equality

$$\|G_p(v_j)\|_{L^1(T^j)} = \det J_{T^j}^{-1} \cdot \|G_p(v_j \circ \tilde{J}_{T^j})\|_{L^1(K)} \cdot \det J_{T^j} \tag{2.12}$$

and in the same way

$$\|\nabla v_j\|_{L^1(\tilde{T^j})} = \det J_{T^j}^{-1} \cdot \|\nabla(v_j \circ \tilde{J}_{T^j})\|_{L^1(\tilde{K}_j)} \cdot \det J_{T^j}, \tag{2.13}$$

where $\tilde{K}_j = \tilde{J}_{T^j}^{-1}(\tilde{T}^j)$. Summarized, (2.12) and (2.13) give that

$$\|G_p(w_j)\|_{L^1(T^j)} \geq j\|\nabla w_j\|_{L^1(\tilde{T^j})} = j, \tag{2.14}$$

with $w_j := v_j \circ \tilde{J}_{T^j} : \tilde{K}_j \to \mathbb{R}$. The mapping between $\tilde{K}_j$ and $\tilde{K}$ corresponding to (2.11) is given by

$$[I, J_{n,1}, J_{n,2}, J_{n,3}] : \tilde{K}_j \to \tilde{K} \tag{2.15}$$

with I the identity operator. As the mesh is shape-regular and the edges of K are kept fixed, the series $(\|J_{n,j}\|)_n$ of the norms should be bounded and therefore, the series in (2.15) should (componentwise) converge to

$$[I, J_1, J_2, J_3] : \tilde{K}^* \to \tilde{K}$$

with some patches K^* of K. According to the assumption (B2) for all $x \in K$ we have

$$G_p(w_n)(x) \to G_p(w)(x),$$

where $w : K^* \to \mathbb{R}$ is defined with $w_n|_{T_{n,j}} = w \circ J_j \circ J_{n,j}^{-1}$. Since $J_j \circ J_{n,j}^{-1} \to I$, we obtain

$$\|G_p(w_n)\|_{L^1(K)} \to \|G_p(w)\|_{L^1(K)} \quad \text{and} \quad \|\nabla w_n\|_{L^1(\tilde{K}_n)} \to \|\nabla w\|_{L^1(\tilde{K}^*)}$$

and therefore, using (2.14) we get

$$\|G_p(w)\|_{L^1(K)} = \infty,$$

which is a contradiction. □

In general, if $u_{h,p}|_{\tilde{T}} \in \mathcal{P}_p(\tilde{T})$, we aim to construct $G(u_{h,p}) \in [\mathcal{P}_p(T)]^2$ such that $G(u_{h,p})$ should be a gradient of a polynomial of order $p+1$ on $\tilde{T}$. For this we denote with E_2, E_4 and E_6 the vertices of an arbitrary triangle T and with E_1, E_3 and E_5 the remaining vertices of the neighbouring triangles T_1, T_2 and T_3, respectively. If the adjacent triangle T_i degenerates into an edge e_i then E_{2i-1} is defined as the midpoint of e_i.

Example 1 - Gradient recovery for $u_{h,p} \in \mathcal{P}_1(T)$.

First we give a general construction.

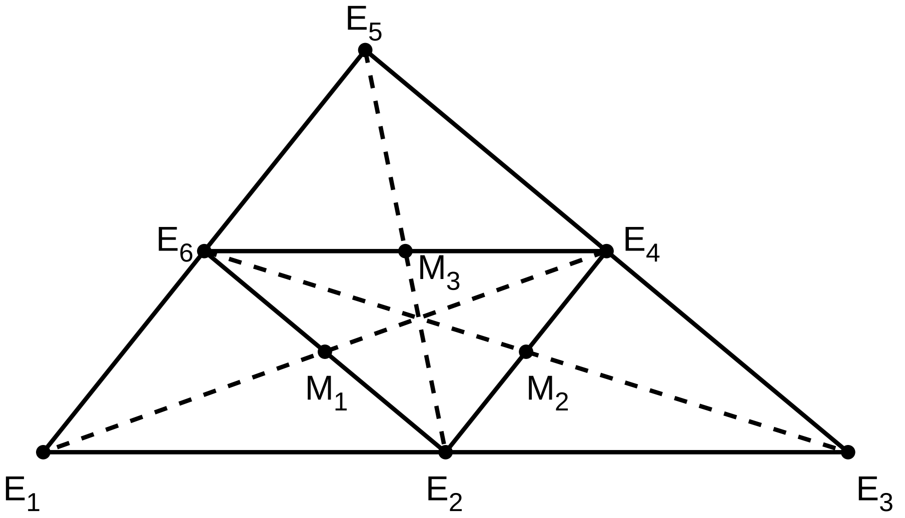

Figure 2.1: The patch $\tilde{T}$ in a uniform tessellation with the midpoints for the first order gradient averaging.

- We fit a second order polynomial $p_{2,\tilde{T}}(u_{h,p})$ to $\{(E_i, u_{h,p}(E_i)) : i = 1, 2, \ldots, 6\}$.
- The first order gradient average is $G_{1,T}(u_{h,p}) = \nabla p_{2,\tilde{T}}(u_{h,p})|_T$.

Remark 2.12 *In practice, we use the least square fit, but the particular fitting method has no importance in the analysis.*

To reduce the computational costs we simplify the above process in case of a special geometry of $\hat{K}$. If $\tilde{T}$ is a triangle and consists of four congruent triangles, called *uniform subdivision* henceforth, then the above fitting procedure can be simplified. For this, first we determine the gradient averages in the midpoints M_1, M_2 and M_3 of the edges of T.

Using the geometrical setup in Figure 2.1 we identify the vertices $E_j,\ j = 1, 2, \ldots, 6$ with their position vectors and introduce the notations

$$\mathbf{v}_{ij} = \frac{E_j - E_i}{|E_j - E_i|},\ i, j = 1, 2, \ldots, 6,\ i \neq j.$$

Example 1a - Gradient recovery for $u_{h,p} \in \mathcal{P}_1(T)$ on a uniform subdivision.

- We define certain directional gradient averages at M_1, M_2 and M_3 as follows

$$\mathbf{v}_{41} \cdot G_{1,T}(u_{h,p})(M_1) = \frac{u_{h,p}(E_1) - u_{h,p}(E_4)}{|E_1 - E_4|},\ \mathbf{v}_{26} \cdot G_{1,T}(u_{h,p})(M_1) = \frac{u_{h,p}(E_6) - u_{h,p}(E_2)}{|E_6 - E_2|}$$
$$\mathbf{v}_{63} \cdot G_{1,T}(u_{h,p})(M_2) = \frac{u_{h,p}(E_3) - u_{h,p}(E_6)}{|E_3 - E_6|},\ \mathbf{v}_{42} \cdot G_{1,T}(u_{h,p})(M_2) = \frac{u_{h,p}(E_2) - u_{h,p}(E_4)}{|E_2 - E_4|}$$
$$\mathbf{v}_{25} \cdot G_{1,T}(u_{h,p})(M_3) = \frac{u_{h,p}(E_5) - u_{h,p}(E_2)}{|E_5 - E_2|},\ \mathbf{v}_{64} \cdot G_{1,T}(u_{h,p})(M_3) = \frac{u_{h,p}(E_4) - u_{h,p}(E_6)}{|E_4 - E_6|}.$$

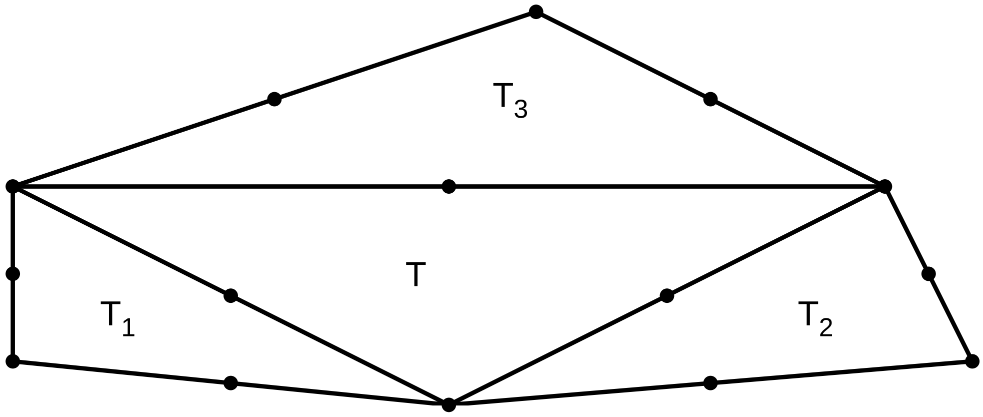

Figure 2.2: Basis points for the second order gradient averaging.

- These determine $G_{1,T}(u_{h,p})$ at M_1, M_2 and M_3.
- Since $G_{1,T}(u_{h,p})$ is a first order polynomial in both components, it can be obtained with a linear interpolation using $G_{1,T}(u_{h,p})(M_1), G_{1,T}(u_{h,p})(M_2)$ and $G_{1,T}(u_{h,p})(M_3)$.

Example 1b - Gradient recovery for $u_{h,p} \in \mathcal{P}_1(T)$ on a uniform subdivision.
For the following construction, we note that $\nabla u_{h,p}$ is piecewise constant.

- We define the gradient averages at M_1, M_2 and M_3 as follows

$$G_{1,T}(u_{h,p})(M_1) = \frac{\nabla u_{h,p}|_{\overline{T}_1} + \nabla u_{h,p}|_{\overline{T}}}{2}, \; G_{1,T}(u_{h,p})(M_2) = \frac{\nabla u_{h,p}|_{\overline{T}_2} + \nabla u_{h,p}|_{\overline{T}}}{2}$$
$$G_{1,T}(u_{h,p})(M_3) = \frac{\nabla u_{h,p}|_{\overline{T}_3} + \nabla u_{h,p}|_{\overline{T}}}{2}.$$

- Since $G_{1,T}(u_{h,p})$ is a first order polynomial in both components, it can be obtained with a linear interpolation using $G_{1,T}(u_{h,p})(M_1), G_{1,T}(u_{h,p})(M_2)$ and $G_{1,T}(u_{h,p})(M_3)$ defined above.

Example 2 - Gradient recovery for $u_{h,p} \in \mathcal{P}_2(T)$.

The second order approximation $u_{h,p}$ is determined by the nodal values at the vertices and the midpoints of the edges of the triangles (see [12, p. 73]). These 15 nodal points in $\tilde{T}$ are depicted in Figure 2.2. If the adjacent triangle T_i degenerates into an edge e_i then we take instead four equidistributed points on e_i.

- We fit the above 15 data points with a full 3rd order polynomial in $\tilde{T}$, which is denoted with $p_{3,\tilde{T}}(u_{h,p})$.
- The second order gradient averaging is $G_{2,T}(u_{h,p}) = \nabla p_{3,\tilde{T}}(u_{h,p})|_{\overline{T}}$.

Remark 2.13 *It would be easier to fit the 3rd order polynomial to 10 data points. The advantage of the setup in Example 2 is that the distribution of the basis points is symmetric with respect to the triangles.*

Remark 2.14 *One can generalize the procedures in Examples 1 and 2 to provide a gradient recovery of an arbitrary order.*

Lemma 2.15 *The gradient recovery techniques in Example 1, Example 1a and Example 1b are identical on uniform tessellations.*

proof: On a uniform subdivision we can exactly fit a second order polynomial q to $E_j, j = 1, 2, \dots, 6$ with $u_{h,p}(E_j) = q(E_j), j = 1, 2, \dots, 6$. We define then $\hat{q} : \mathbb{R} \to \mathbb{R}$ by

$$\hat{q}(\lambda) = q(E_1 - \lambda|E_1 - E_4|\mathbf{v}_{41}).$$

It is clear that $\hat{q}$ is second order with

$$\hat{q}(0) = q(E_1), \quad \hat{q}\,(0.5) = q(M_1) \quad \text{and} \quad \hat{q}(1) = q(E_4).$$

Moreover, $\hat{q}'\,(0.5) = \hat{q}(1) - \hat{q}(0)$ and therefore

$$\begin{aligned}\partial_{\mathbf{v}_{41}} q(M_1) &= -\hat{q}'\,(0.5) \cdot \frac{1}{|E_1 - E_4|} = -(\hat{q}(1) - \hat{q}(0)) \cdot \frac{1}{|E_1 - E_4|} \\ &= \frac{q(E_1) - q(E_4)}{|E_1 - E_4|} = \frac{u_{h,p}(E_1) - u_{h,p}(E_4)}{|E_1 - E_4|}.\end{aligned}$$

A similar derivation gives that

$$\partial_{\mathbf{v}_{26}} q(M_1) = \frac{q(E_6) - q(E_2)}{|E_6 - E_2|} = \frac{u_{h,p}(E_6) - u_{h,p}(E_2)}{|E_6 - E_2|}$$

and in the same way

$$\begin{aligned}\partial_{\mathbf{v}_{63}} q(M_2) &= \frac{u_{h,p}(E_3) - u_{h,p}(E_6)}{|E_3 - E_6|}, \partial_{\mathbf{v}_{42}} q(M_2) = \frac{u_{h,p}(E_2) - u_{h,p}(E_4)}{|E_2 - E_4|}, \\ \partial_{\mathbf{v}_{25}} q(M_3) &= \frac{u_{h,p}(E_5) - u_{h,p}(E_2)}{|E_5 - E_2|}, \partial_{\mathbf{v}_{64}} q(M_3) = \frac{u_{h,p}(E_4) - u_{h,p}(E_6)}{|E_4 - E_6|}.\end{aligned}$$

This gives that the procedures in Example 1 and Example 1a results in the same averages at M_1, M_2 and M_3.

For proving that the averages in Example 1a and Example 1b are equivalent, we note that the gradient of an arbitrary function $q : T \to \mathbb{R}$ or $q : T_1 \to \mathbb{R}$ are determined by $\partial_{\mathbf{v}_{41}} q$ and $\partial_{\mathbf{v}_{26}} q$

$$\nabla q = A^{-1}(\partial_{\mathbf{v}_{41}} q, \partial_{\mathbf{v}_{26}} q)^T, \text{ where } \quad A = \begin{pmatrix} \mathbf{v}_{41}^T \\ \mathbf{v}_{26}^T \end{pmatrix} \in \mathbb{R}^{2\times 2}. \tag{2.16}$$

In this way, the gradient corresponding to the procedure in Example 1b is

$$\begin{aligned}
G_{1,T}(u_{h,p})(M_1) &= \frac{1}{2}\left(\nabla u_{h,p}|_{T_1} + \nabla u_{h,p}|_{T}\right)(M_1)\\
&= \frac{1}{2}A^{-1}\left((\partial_{\mathbf{v}_{41}} u_{h,p}|_{T_1}, \partial_{\mathbf{v}_{26}} u_{h,p}|_{T_1})^T + (\partial_{\mathbf{v}_{41}} u_{h,p}|_{T}, \partial_{\mathbf{v}_{26}} u_{h,p}|_{T})^T\right)(M_1)\\
&= A^{-1}\left(\frac{u_{h,p}(M_1)-u_{h,p}(E_1)}{|E_1-E_4|} + \frac{u_{h,p}(E_4)-u_{h,p}(M_1)}{|E_1-E_4|},\right.\\
&\qquad \left.\frac{u_{h,p}(M_1)-u_{h,p}(E_2)}{|E_6-E_2|} + \frac{u_{h,p}(E_6)-u_{h,p}(M_1)}{|E_6-E_2|}\right)\\
&= A^{-1}\left(\frac{u_{h,p}(E_4)-u_{h,p}(E_1)}{|E_1-E_4|}, \frac{u_{h,p}(E_6)-u_{h,p}(E_2)}{|E_6-E_2|}\right).
\end{aligned}$$

This means, by (2.16), that in Example 1b we obtain the same directional derivatives $\partial_{\mathbf{v}_{41}}$ and $\partial_{\mathbf{v}_{26}}$ as in Example 1a. In this way the recovered gradients in Example 1a and Example 1b the will be the same, as well. □

Lemma 2.16 *The recovered gradient $G_1(u_{h,p})$ given in Example 1 satisfies the conditions in (A1)-(A4).*

Proof By the construction $G_{1,T}(u)$ depends only on $u|_{\tilde{T}}$, therefore (A1) is satisfied.

To verify (B1) we first give $G_1(u \circ J_T)$. Observe that the fitted second order polynomial $p_{2,T}(u \circ \tilde{J}_T)$ provides the same approximation at the basis points for $u \circ \tilde{J}_T$ as $p_{2,T}(u)$ at the basis points for u. Taking its gradient gives

$$\begin{aligned}
(J_T^{-1}(x)) &= [\nabla(p_{2,\tilde{T}}(u \circ \tilde{J}_T))](J_T^{-1}(x)) = J_T \nabla p_{2,\tilde{T}}(u \circ \tilde{J}_T J_T^{-1}(x))\\
&= J_T \nabla p_{2,\tilde{T}}(u(x)) = J_T G_{1,T}(u)(x)
\end{aligned}$$

such that (B1) is satisfied.

For the proof of (B2) we denote with $E_{n,1}, E_{n,2}, \ldots, E_{n,6}$ the vertices of the patches $\tilde{T}_n$. If the convergence

$$[J_{T_n}, J_{T_{n,1}}, J_{T_{n,2}}, J_{T_{n,3}}] \to [J_T, J_{T_1}, J_{T_2}, J_{T_3}].$$

holds, then obviously $J_{T_n}(x) \to J_T(x)$ and $E_{n,j} \to E_j, j = 1, 2, \ldots, 6$. Also, by definition $u_n(E_{n,j}) = u(E_j), j = 1, 2, \ldots, 6$. Since the result of the fitting depends continuously on the input data, we obtain the convergence

$$p_{2,T_n}(J_{T_n}(x)) \to p_{2,T}(J_T(x)).$$

Since here the range is finite dimensional, the gradients converge as well, i.e. for all $x \in K$ we have

$$G_{p,T_n}(u_n)(J_{T_n}(x)) = \nabla p_{2,T_n}(u_n \circ J_{T_n}(x)) \to \nabla p_{2,T}(u \circ J_T(x)) = G_{p,T}(u)(J_T(x)),$$

such that (B2) is satisfied.

Therefore, using Lemma 2.11 (A2) is satisfied, too.

If u is a second order polynomial and we fit a second order polynomial to some of its nodal values, we certainly get u itself so that $p_{2,\tilde{T}}(I_1 u) = u$. Taking its gradient gives

$$G_{1,T}(I_1 u) = \nabla p_{2,\tilde{T}}(I_1 u)|_T = \nabla u|_T = I_1 \nabla u|_T,$$

which proves (A3).

Obviously the last condition (A4) is also valid: $G_{1,T}(u)$ is a gradient, as it is defined by $\nabla p_{2,\tilde{T}}(u)|_T$. □

Lemma 2.17 *The recovered gradient $G_2(u_{h,p})$ satisfies the conditions in (A1)-(A4).*

proof: By the construction $G_2(u)|_K$ depends only on $u|_{\tilde{K}}$, therefore (A1) is satisfied.

To verify (B1) we first observe that the fitted third order polynomial $p_{3,\tilde{T}}(u \circ J_K)$ provides the same approximation for $u \circ J_T$ as $p_{3,T}(u)$ for u. Taking its gradient gives

$$\begin{aligned}(\tilde{J}_T^{-1}(x)) &= [\nabla(p_{3,\tilde{T}}(u \circ \tilde{J}_T))](J_T^{-1}(x)) = J_T \nabla p_{3,\tilde{T}}(u \circ \tilde{J}_T J_T^{-1}(x)) \\ &= J_T \nabla p_{3,\tilde{T}}(u(x)) = J_T G_{2,T}(u)(x)\end{aligned}$$

such that (B1) is satisfied.

We can verify (B2) using the same arguments as in Lemma 2.16 such that according to Lemma 2.11 (A2) is also satisfied.

If u is a third order polynomial then the second order interpolation is executed based on the 15 values such that $p_{3,T}(I_2 u) = u$. Therefore,

$$G_{2,T}(I_2 u) = \nabla p_{3,\tilde{T}}(I_2 u)|_T = \nabla u|_T = I_2 \nabla u|_T,$$

which proves (A3).

Obviously $G_{2,T}(u_{h,p})$ is a gradient, as it is defined by $\nabla p_{3,\tilde{T}}(u_{h,p})|_{\overline{T}}$. This completes the proof that the conditions in (A1)-(A4) are valid. □

Remark 2.18 *One can generalize the proof in Lemma 2.16 and Lemma 2.17 to prove that any higher order gradient recovery (corresponding to Examples 1 and 2) satisfies (A1)-(A4).*

Remark 2.19 *A standard finite element convergence theory implies that the estimate in assumption (SC) is always satisfied with $\tau = 0$, see [12], [25]. We do not verify here that it is also valid with some $\tau > 0$. The related topic,* superconvergence analysis *has an extended literature depending on the particular equations and finite element discretisations. For a detailed study of this condition for elliptic problems we refer to the monograph [76] and for some recent results to [33, 43].*

2.4 Numerical experiments

The performance of the a-posteriori error estimator and the corresponding estimate for the Neumann type boundary data introduced in Section 2.3 will be demonstrated by using three test cases indexed by $j = 1, 2, 3$.

In each case we investigated the finite element solution of the problem

$$\Delta u_j - \mu u_j = f_j \quad \text{in } \Omega = (0,1) \times (0,1) \tag{2.17}$$

$$u_j = g_j \quad \text{on } \Gamma = \partial\Omega, \tag{2.18}$$

using the constant $\mu = 1000$ on a uniform triangular tessellation of Ω.

For the computation of $u_{j,h,p}$ we have used Lagrange elements of first, second and third order on a uniform triangular mesh of Ω such that $u_{j,h,p} \in \mathcal{P}_p$ for all $T \in \mathcal{T}_h$ for some $p \in \mathbb{Z}^+$.

To solve the corresponding Neumann problems (2.5) for the error we rewrite these in a weak form and discretise as follows

Find $\hat{e}_{h,p} \in \mathcal{P}_{p+1}$ such that

$$\begin{aligned}&\int_T \nabla \hat{e}_{h,p} \cdot \nabla v_{h,p} + \int_T \mu \hat{e}_{h,p} v_{h,p} \\ &= -\int_T (f - \Delta u_{h,p} + \mu u_{h,p}) v_{h,p}) - \int_{\partial T} (\nu \cdot G_{p,T}(u_{h,p}) - \partial_\nu u_{h,p}) v_{h,p} \quad \forall\, v_{h,p} \in \mathcal{P}_{p+1}.\end{aligned}$$

The exact solution of (2.17)-(2.18) for $j = 1, 2, 3$ are given as follows:

- *Test case 1*: $u_1(x, y) = \sin(2\pi x)\sin(2\pi y)$.
- *Test case 2*: $u_2(x, y) = 1 - (x^2 + y^2)^{1/4}$.
- *Test case 3*: $u_3(x, y) = \arctan\left(60\sqrt{(x - 1.25)^2 + (y + 0.25)^2} - \frac{\pi}{3}\right)$.

These define the f_j and g_j in (2.17) for $j = 1, 2, 3$.

The methods we compare are the following:

- Standard approximation based on interface averages (hereafter FA): on each edge we approximate $\partial_\nu u$ with the average $\partial_\nu e_{h,p}$ from the both sides. For further details, see [3] for elliptic problems and [45] for Maxwell equations.
- Gradient averaging (hereafter GA): we apply the standard techniques given in [3, 77].
- Gradient recovery using higher order fitting (hereafter GR): described in Section 2.3.

2.4.1 Global error estimators for the Neumann boundary data and the energy norm

In the local error estimates the only unknown term is the Neumann type boundary condition, cf. with (2.5). Therefore, according to Proposition 2.5 the accuracy of the error estimate depends on the quality of the estimate of these boundary conditions. Accordingly, we first compute the following norm

$$d(L^2) := \left(\sum_{\substack{K \subset \Omega \\ \partial K \cap \partial\Omega = \emptyset}} ||\partial_\nu e_{h,p} - \partial_\nu \hat{e}_{h,p}||^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \tag{2.19}$$

which depends only on the computed data, and we compare our estimator with the classical ones. We also compare the local errors on the subdomains: the exact error $e_{h,p}$ on K is computed by using the exact boundary condition $\partial_\nu e_{h,p}$ on ∂K, while for the implicit error estimation $\hat{e}_{h,p}|_K$ the estimated boundary condition $\partial_\nu \hat{e}_{h,p}|_{\partial K}$ has been utilized using different approximations. We compute the total amount of these errors over all of the interior subdomains

$$d(H^1) := \left(\sum_{\substack{K \subset \Omega \\ \partial K \cap \partial\Omega = \emptyset}} ||e_{h,p} - \hat{e}_{h,p}||^2_{H^1(K)} \right)^{\frac{1}{2}} \tag{2.20}$$

and relate them with corresponding norm of the exact error $e_{h,p}$

$$||e_{h,p}||_1 := \left(\sum_{K \subset \Omega} ||e_{h,p}||^2_{H^1(K)} \right)^{\frac{1}{2}} .$$

The results for $u_{h,p} \in \mathcal{P}_1, \mathcal{P}_2$ and $\mathcal{P}_3$, i.e. using first, second and third order Lagrange elements are shown in Table 2.1, 2.2 and 2.3, respectively.

While in the quality of the Neumann boundary conditions no significant differences can be detected, the performance of the method GR, proposed here, seems to be substantially better than the classical ones FA and GA for the piecewise energy norm. The only exception is *Test case 3*. Here the large oscillations in the higher order approximation of steep gradients can make the estimator for the local boundary conditions rather inaccurate, which result in unsharp error estimators in each case. This could be avoided by using a local mesh refinement in this critical region. In general, an accurate finite element solution is necessary to obtain a proper error estimate. This is also shown also in case of first order elements where none of the listed methods provide an accurate error indicator.

2.4.2 Local performance of the error estimator

Since the adaptive FE solvers make use of local error indicators, we present the performance of our estimate locally on some subdomains shown in Figure 2.3. The graphs at the left and the

Table 2.1: Accuracy of the estimations of the elementwise Neumann type boundary data and for the energy norm of the errors. The approximations $u_{h,p}$ of u_1, u_2 and u_3, respectively, have been computed using *first order* Lagrange elements. The quantities in (2.19) (left) and (2.20) (right) are given for each test case using different methods. In the last column the exact error is given.

	$d(L^2)$				$d(H^1)$			$\|\|e_{h,p}\|\|_1$
	FA	GA	GR		FA	GA	GR	
u_1								
$n = 5$	18.9076	14.0775	18.5386		18.1201	6.0507	18.1210	7.6763
$n = 10$	20.2392	19.4027	19.5386		6.1978	6.1237	5.8245	5.0027
$n = 15$	15.2591	13.9985	13.9397		1.7778	1.5945	1.6463	3.5399
u_2								
$n = 5$	0.1259	0.0709	0.0560		0.0006	0.0005	0.0002	0.2448
$n = 10$	0.1559	0.0724	0.0611		0.0005	0.0006	0.0003	0.1740
$n = 15$	0.1646	0.0725	0.0617		0.0006	0.0008	0.0004	0.1423
u_3								
$n = 5$	52.3118	44.7910	51.9011		89.4993	54.3913	90.2840	19.5809
$n = 10$	60.8298	56.2916	62.8102		93.4484	78.4476	97.9891	13.4684
$n = 15$	44.1284	44.8308	44.6453		81.3571	87.6624	81.9278	6.2388

right hand side of Figures 2.4 - 2.6 exhibit the L^2 error in the Neumann boundary data

$$d(L^2_K) := \left(||\partial_\nu e_{h,p} - \partial_\nu \hat{e}_{h,p}||^2_{L^2(\partial K)} \right)^{\frac{1}{2}} \tag{2.21}$$

and the H^1 error of the implicit error estimation

$$d(H^1_K) := \left(||e_{h,p} - \hat{e}_{h,p}||^2_{H^1(K)} \right)^{\frac{1}{2}}, \tag{2.22}$$

respectively, on the subdomains in Figure 2.3.

The following observations confirm the favor of our method:

- The gradient recovery operator GR proposed here delivers significantly sharper results in the presented test cases than the classical techniques FA and GA.
- The estimator GR becomes even sharper in the case of higher order elements.
- The error estimator GR seems to be equally distributed over the finite element subdomains such that $e_{h,p}$ and $\hat{e}_{h,p}$ correlate perfectly.

Therefore, the error estimator presented in this chapter can maintain an accurate hp-adaptive refinement algorithm [19], [64].

Table 2.2: Accuracy of the estimations of the elementwise Neumann type boundary data and of the energy norm of the errors. The approximations $u_{h,p}$ of u_1, u_2 and u_3, respectively, have been computed using *second order* Lagrange elements. The quantities in (2.19) (left) and (2.20) (right) are given for each test case using different methods.

	$d(L^2)$				$d(H^1)$			$\|\|e_{h,p}\|\|_1$
	FA	GA	GR		FA	GA	GR	
u_1								
$n = 5$	5.5409	11.9002	9.0425		0.7820	5.8598	1.1150	3.8706
$n = 10$	5.6623	12.9412	3.5634		0.7143	7.2921	0.0714	1.2353
$n = 15$	4.4471	11.9992	1.4355		0.2653	5.5670	0.0083	0.5830
u_2								
$n = 5$	0.0354	0.0918	0.0178		0.0001	0.0004	$< 10^{-4}$	0.1446
$n = 10$	0.0366	0.1099	0.0178		0.0002	0.0006	$< 10^{-4}$	0.1022
$n = 15$	0.0367	0.1156	0.0178		0.0003	0.0009	$< 10^{-4}$	0.0834
u_3								
$n = 5$	33.8184	29.1245	29.8231		55.8889	28.3894	46.5715	10.2285
$n = 10$	24.0926	25.0117	24.7771		15.5860	33.2854	18.0114	4.1497
$n = 15$	17.7498	23.1537	19.8181		10.0317	54.6338	15.9303	2.4151

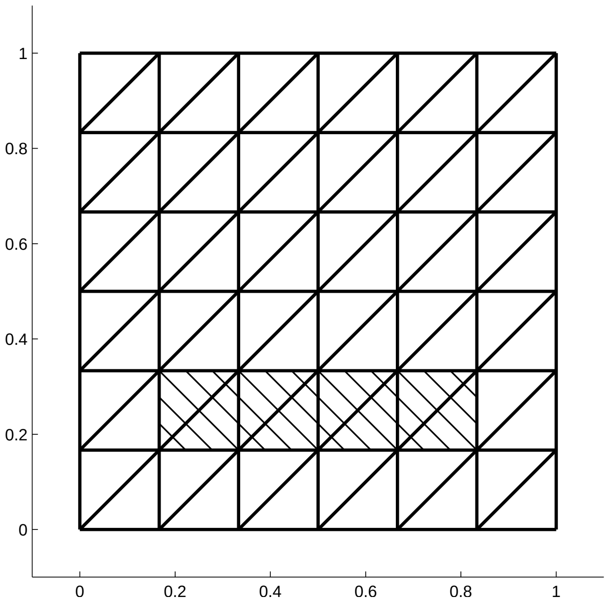

Figure 2.3: Uniform mesh for the computations. The comparison of the local accuracy has been performed on the shaded elements in the 2nd row.

Table 2.3: Accuracy of the estimations of the elementwise Neumann type boundary data and of the energy norm of the errors. The approximations $u_{h,p}$ of u_1, u_2 and u_3, respectively, have been computed using *third order* Lagrange elements. The quantities in (2.19) (left) and (2.20) (right) are given for each test case using different methods.

	$d(L^2)$				$d(H^1)$			$\|\|e_{h,p}\|\|_1$
	FA	GA	GR		FA	GA	GR	
u_1								
$n = 5$	4.8794	8.3113	5.1958		1.0356	1.4403	0.6440	1.4191
$n = 10$	1.4455	7.0024	1.0783		0.0196	0.5068	0.0086	0.2100
$n = 15$	0.4975	6.1275	0.3370		0.0009	0.3342	0.0006	0.0633
u_2								
$n = 5$	0.0113	0.0561	0.0061		$< 10^{-4}$	0.0001	$< 10^{-4}$	0.0961
$n = 10$	0.0114	0.0653	0.0061		$< 10^{-4}$	0.0001	$< 10^{-4}$	0.0680
$n = 15$	0.0115	0.0679	0.0061		$< 10^{-4}$	0.0001	$< 10^{-4}$	0.0555
u_3								
$n = 5$	34.2449	33.5215	25.6820		66.0944	54.5199	22.3025	9.3745
$n = 10$	16.087	17.3896	19.2871		9.2497	7.5523	5.5573	2.5407
$n = 15$	10.9046	12.5226	14.9013		6.1932	4.2965	7.7984	1.1588

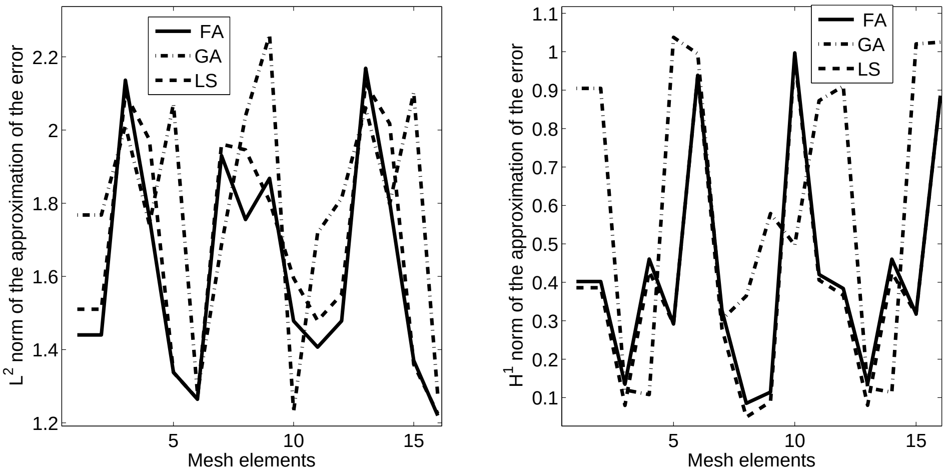

Figure 2.4: Local accuracy of the implicit error estimation technique using the gradient recovery operators FA, GA and GR in Test Case 1 with $u_1 \in P_1$. Left: L^2 norm $d(L^2_K)$ of the approximation of the error in the Neumann boundary data (see (2.21)) on the depicted elements in Fig. 2.3. Right: H^1 norm $d(H^1_K)$ of the approximation of the error (see (2.22)) on the depicted elements in Fig. 2.3.

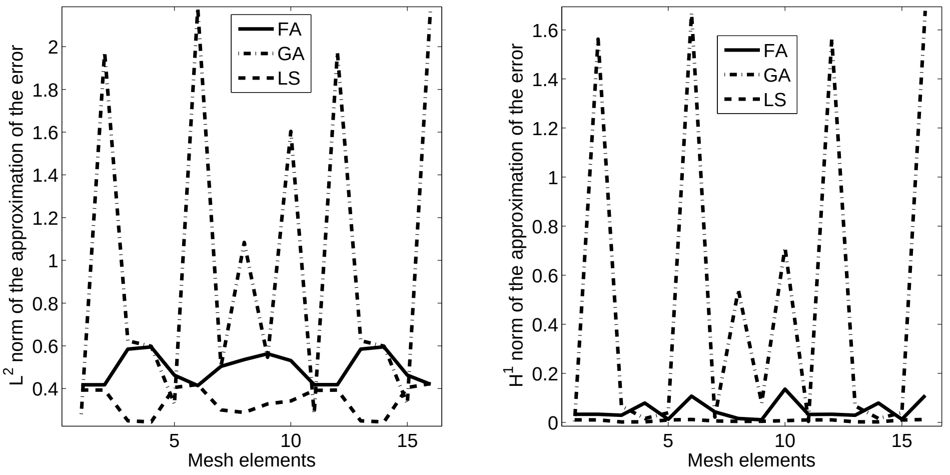

Figure 2.5: Local accuracy of the implicit error estimation technique using the gradient recovery operators FA, GA and GR in Test Case 1 with $u_1 \in P_2$. Left: L^2 norm $d(L^2_K)$ of the approximation of the error in the Neumann boundary data (see (2.21)) on the depicted elements in Fig. 2.3. Right: H^1 norm $d(H^1_K)$ of the approximation of the error (see (2.22)) on the depicted elements in Fig. 2.3.

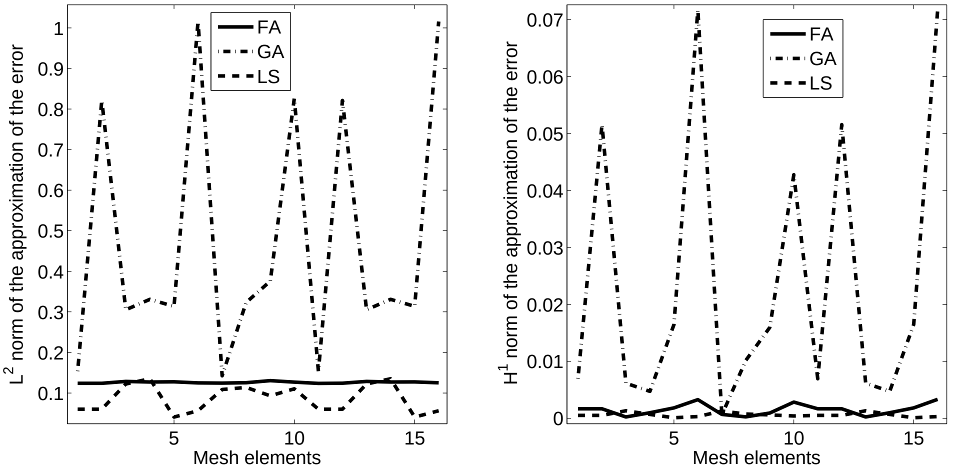

Figure 2.6: Local accuracy of the implicit error estimation technique using the gradient recovery operators FA, GA and GR in Test Case 1 with $u_1 \in P_3$. Left: L^2 norm $d(L^2_K)$ of the approximation of the error in the Neumann boundary data (see (2.21)) on the depicted elements in Fig. 2.3. Right: H^1 norm $d(H^1_K)$ of the approximation of the error (see (2.22)) on the depicted elements in Fig. 2.3.

Chapter 3

Adaptivity via reference solution

3.1 hp-adaptivity

In this chapter we suggest an improvement for a family of hp-adaptive finite element methods. We point out the necessity of the new procedure by constructing model problems for which certain standard hp-adaptive algorithms fail to work properly. It is verified also in the corresponding simulations that the algorithm can terminate even though the numerical solution still contains a sizable computational error. This chapter is based on [38].

There are several hp-adaptive finite element algorithms in the literature for the numerical solution of PDE's (see the collection [55] and the references therein). These are based on the following scheme:

```
Initialize: solve the initial problem with small polynomial degree $p$ on a coarse grid

Repeat:

  S1 estimate the error
  S2 if the error is small then stop
  S3 else determine on which elements in the grid and how to refine/derefine
  S4 compute the new solution and go to S1
```

The main differences between the different methods are in the error estimation and refinement procedures. Here we will focus on the method introduced by Demkowicz et al. [20] and also used (with a small modification) by Šolín et al. [65]. In step S1 this method uses the so called "reference solution" to compute the error. This reference solution is calculated using a uniform refinement in space and by increasing the degree of polynomials on *every* element in order to calculate a finer solution that will be considered as the reference solution. The error is defined as the difference between the reference solution and original solution.

Despite the fact that this method can be used for many different problems, such as time dependent problems [68, 23], Maxwell's equations [66, 67] or coupled highly nonlinear problems [24], the efficiency of the error estimation has not been justified rigorously. We illustrate this by the so-called antenna example [18, Section 15.3.] where the convergence of the method seems to be broken because the estimated error does not decrease during refinements. However, after further iteration steps it decreases again until the estimated error becomes small enough. In that case the error estimator works properly because it forces the refinements until the estimated error is small enough.

The aim of this chapter is to give a counterexample where the reference solution is the same as the original one, hence the computed error will be zero, whereas the error between the exact solution and the approximate solution is large. Furthermore we suggest a modification of the method that can be used to avoid these kinds of difficulties.

3.2 Notations

We investigate the elliptic boundary-value problem (1.1)-(1.2) with - for simplicity - fully Dirichlet boundary conditions

$$-\text{div}\,(\mathcal{K}\nabla u) + \mu u = f \quad \text{in } \Omega, \tag{3.1}$$

$$u = g \quad \text{on } \Gamma, \tag{3.2}$$

and its weak form. Find $u \in H^1(\Omega)$ such that $u = u_g + w$ where $u_g = g$ on $\partial\Omega$, $w \in H^1_0(\Omega)$

$$a(w,v) = \int_\Omega fv - a(u_g, v) \quad \forall v \in H^1_0(\Omega), \tag{3.3}$$

$$a(w,v) := \int_\Omega \mathcal{K}\nabla w \cdot \nabla v + \int_\Omega \mu w v,$$

where $\Omega \subset \mathbb{R}^d$, however, only $d = 2$ is used in this chapter, $\Gamma = \partial\Omega$, $f \in L^2(\Omega)$, $\mu \in L^\infty(\Omega)$, $\mathcal{K}$ is a symmetric uniformly positive definite matrix valued function just as before.

Let us denote by $\mathcal{T}_h$ a tessallation of Ω into elements. The discretisation of (3.3) takes the following form. Let us denote by $V_{h,p} \subset H^1(\Omega)$ a finite dimensional subspace. Find $u_{h,p} \in V_{h,p}$ such that $u_{h,p} = u_{D_{h,p}} + w_{h,p}$ where $u_{D_{h,p}} \in V_{h,p}$ approximates g on $\partial\Omega$, see Section 1.2.1, $w_{h,p} \in V_{h,p}$ then

$$a(w_{h,p}, v_{h,p}) = \int_\Omega f v_{h,p} - a(u_{D_{h,p}}, v_{h,p}) \quad \forall v \in V_{h,p}. \tag{3.4}$$

The space $V_{h,p}$ contains piecewise polynomials whose degrees can vary from element to element. In this framework it would be more precise to use the notation $V_{h,\mathbf{p}}$, emphasizing that p is not constant, where $\mathbf{p}$ is a vector that contains the polynomial degrees for every element in the mesh.

Similarly, $V_{h/2,p+1}$ denotes the approximation space where the mesh is refined and the degree of the local polynomials is increased. If we think about the vector representation of p than this becomes a bit technical: $p+1$ should be a vector that has four times more entries than p. There is a connection between the two vectors. Every mesh element is divided into four subelements, and these elements are inherits the polynomial degree from the "parent" element. $p+1$ contains these entries, although, all entries are increased by 1.

The reference solution based methods, described in [20] and [65], use the following basic idea:

S1 compute $u_{h,p} \in V_{h,p}$ by solving (3.4),

S2 compute the reference solution $u_{h/2,p+1} \in V_{h/2,p+1}$ by solving (3.4) in the enriched space $V_{h/2,p+1}$,

S3 use $u_{h/2,p+1}$ as a more accurate solution and define $u_{h,p} - u_{h/2,p+1}$ as an error indicator,

S4 compute one of the following quantities on all $T \in \mathcal{T}_h$:

- $\eta_T = \dfrac{|u_{h,p} - u_{h/2,p+1}|_{1,T}}{|u_{h/2,p+1}|_{1,T}}$
- $\eta_T = \dfrac{||u_{h,p} - u_{h/2,p+1}||_{1,T}}{||u_{h/2,p+1}||_{1,T}}$
- $\eta_T = |u_{h,p} - u_{h/2,p+1}|_{1,T}$
- $\eta_T = ||u_{h,p} - u_{h/2,p+1}||_{1,T}$

S5 if $\sqrt{\sum_T \eta_T^2} < \mathtt{TOL}$ stop, else refine where it is needed, and go to S1.

The difference between [20] and [65] lies in the choice of η_T and [20] does a refinement all over the edges not only all over the elements.

3.3 Counterexample

For simplicity suppose that we have homogeneous Dirichlet boundary conditions.

The construction of the counterexample is based on the following simple idea. Let us suppose that we can find a function $f \neq 0$ such that $\int_\Omega f v_{h,p} = 0$, $\forall v_{h,p} \in V_{h,p}$. In this case (3.4) simplifies to

$$a(w_{h,p}, v_{h,p}) = 0.$$

and it has an exact solution $u_{h,p} = 0$. However, if f differs from 0 the exact solution u also differes from 0. This can be extended to the case of non-homogeneous Dirichlet boundary condition. In that case the condition $f \neq 0$ has to be replaced by $f \neq -\text{div}\,(\mathcal{K}\nabla u_{D_{h,p}}) + \mu u_{D_{h,p}}$.

If $\int_\Omega f v_{h/2,p+1} = 0$, $\forall v_{h/2,p+1} \in V_{h/2,p+1}$ then using the fact $V_{h,p} \subset V_{h/2,p+1}$ we have that $\int_\Omega f v_{h,p} = 0$ also holds. In this case $u_{h,p} = u_{h/2,p+1} = u_0$, and therefore the computed error is zero.

Now we show how to create a function f satisfying the above assumptions. For any $T \in \mathcal{T}_h$ we define $u_c : \Omega \to \mathbb{R}$ such that $\operatorname{supp}(u_c) = T$, $u_c(x,y) = \chi_T(x,y) b_T^2(x,y) p(x,y)$, where χ_T is the characteristic function of T, $b_T : \Omega \to \mathbb{R}$ is a bubble function on T, so $b_T = 0$ on ∂T and

$$p(x,y) = \sum_{k=0}^{m-1} c_k x^{a_k} y^{b_k}. \tag{3.5}$$

Here $a_k, b_k \in \mathbb{N} \cup \{0\}$ and $c_k \in \mathbb{R}$, m is a fixed integer (that will be determined later). The polynomial $p(x,y)$ is chosen so that

$$\int_\Omega (-\text{div}\,(\kappa \nabla u_c) + \mu u_c) v = 0 \qquad \forall v : v \in V_{h/2,p+1}, \operatorname{supp}(v) = T, \tag{3.6}$$

or, since $\operatorname{supp}(u_c) = T$

$$\int_T (-\text{div}\,(\kappa \nabla u_c) + \mu u_c) v = 0 \qquad \forall v : v \in V_{h/2,p+1}. \tag{3.7}$$

In order to compute the coefficient vector $c = (c_0, \ldots, c_{m-1})$ we have to solve the linear system $Ac = 0$ where the entries of $A \in \mathbb{R}^{n \times m}$ are given by

$$A_{i,j} = \int_T (-\text{div}\,(\kappa \nabla b_T^2 c_j x^{a_j} y^{b_j}) + \mu b_T^2 c_j x^{a_j} y^{b_j}) v_i \; dT \qquad \forall i \in \{1, 2, \ldots, n\}.$$

Here we have used the notation $\dim V_h|_T = n$. To find a nonzero solution c, A must have more columns than rows, so $m := n + 1$. It is sufficient to find a submatrix $\overline{A} \in \mathbb{R}^{r(A) \times r(A)+1}$ where $r(A)$ is the rank of A, and solve the reduced system $\overline{A}\overline{c} = 0$.

If we have such a solution then we have at least one free parameter to define c. We set this parameter to an arbitrary number, i.e. 1. This can be used as coefficients c in the definition of u_c.

For any $C_0 \in \mathbb{R}$ the test function $u_0 + C_0 u_c$ will give $u_{h,p} = u_{h/2,p+1} = u_0$ and the algorithm will terminate, even though the error $C_0 u_c$ can be arbitrarily large.

Remark 3.1 *If we use implicit a-posteriori error estimation first we solve* (3.4) *which gives* $u_{h,p} = u_0$. *Then a local Neumann problem is solved on every element* $\widetilde{T}$

$$-div\,(\kappa \nabla e) + \mu e = f + div\,(\kappa \nabla u_{h,p}) - \mu u_{h,p} \quad \text{in } \widetilde{T}, \tag{3.8}$$

$$\partial_\nu e = -\frac{1}{2}\,[\partial_\nu u_{h,p}] \quad \text{on } \partial\widetilde{T} \setminus \partial\Omega, \tag{3.9}$$

$$e = 0 \quad \text{on } \partial\widetilde{T} \cap \partial\Omega, \tag{3.10}$$

where e *is an estimator of* $u - u_{h,p}$, $[\partial_\nu u_{h,p}]$ *is the jump of the outward normal derivative of the numerical solution on an interior edge (see [3, Ch. 3] for details). If* $u_0 = 0$ *on T then the* *r.h.s. of* (3.8) *will be f. The estimated boundary condition will also be zero. The solution of* (3.8)-(3.10) *depends on the local finite element space $W_{\widetilde{T}}$. If $W_{\widetilde{T}} \subseteq V_{h/2,p+1}$, then we will again have $e = 0$ on $\widetilde{T}$.*

3.4 Numerical results

By courtesy of William F. Mitchell the procedure described above was tested numerically using his PHAML code [54]. The code was supplied with the following initial mesh:

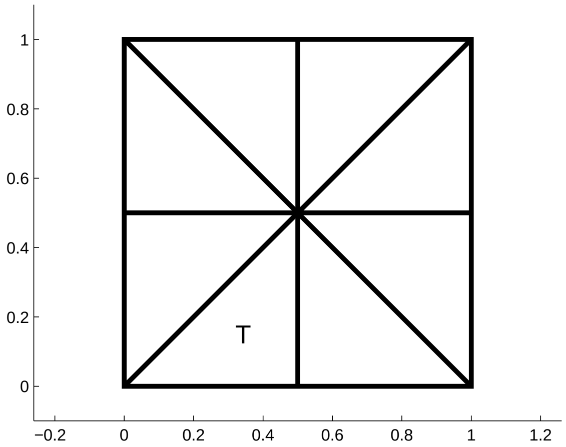

Figure 3.1: The initial mesh and triangle T.

The problem was a simple Poisson equation, $\mathcal{K} \equiv 1$, $\mu \equiv 0$, with homogenous Dirichlet boundary condition in (3.1)-(3.2)

$$\begin{aligned} -\triangle u(x,y) &= f(x,y) \qquad \text{in } \Omega, \\ u(x,y) &= u_D(x,y) \quad \text{on } \Gamma, \end{aligned}$$

with $\Omega = (0,1)^2$, and $b_k = 0$ (see (3.5)). The polynomial degree p was 1 at the initial step. For computing the reference solution PHAML used bisected triangles.

We used as our counterexample

$$\begin{aligned} u_c(x,y) = x + C_0\chi_T(x,y)(&-3200(x-y)^2y^2(2x-1)^2(-1+4x)^2\cdot \\ &(1810432x^7 - 4313088x^6 + 4323072x^5 - 2356224x^4 \\ &+ 751088x^3 - 139176x^2 + 13747x - 549)) \end{aligned} \tag{3.11}$$

and $f_c(x,y) := -\triangle u_c(x,y)$, $u_D(x,y) = x|_\Gamma$, $C_0 = 10^5$.

The second term of u_c was calculated by the method described above. Theoretically we should have $u_{h,p} = u_{h/2,p+1} = x$ according to the previous section, yielding that the real error $||u - u_{h,p}||_1$ can be arbitrary and we can control it by our choice of C_0.

When we implemented this model problem in PHAML we encountered problems with numerical integration. One can verify that $\int_{T_j} f_c(x,y) \cdot x^k y^l = 0$ if $0 \leq k, l \leq 2$, $k + l \leq 2$ for all T_j that are a subtriangle of T and for any C_0. However, even when the highest available order quadrature was used it was different from zero and the r.h.s. of (3.3) became nonzero.

We obtained $||u_{h/2,p+1} - u_{h,p}||_1 \approx 10^{-9}$, which was our main aim. The addition of x was necessary. Without it $||u_{h/2,p+1}||_1 \approx 10^{-9}$ and the relative error was $O(1)$. The addition of 1 would not make any difference if we used a seminorm instead of a norm.

For all possible stopping criteria mentioned in Section 3.3 we could achieve $\sqrt{\sum_T \eta_T^2} \approx 10^{-9}$. This means that the algorithm terminated at the initial step whenever $\texttt{TOL} > 10^{-8}$, even though the computational error can be almost arbitrary, depending only on C_0.

3.5 Possible corrections

We can easily fix this problem by building a back-up estimator into the code. For example, we can use the residual-based error estimator

$$|||u - u_{h,p}|||^2 \leq \eta_{\text{res}}^2 = C_{\text{res}} \left(\sum_{T \in \mathcal{T}_h} h_T^2 ||r||_{L_2(T)}^2 + \sum_{\gamma \in \partial T} h_T ||R||_{L_2(\gamma)}^2 \right), \tag{3.12}$$

where r is the interior residual $r = f + \text{div}\,(\mathcal{K} \nabla u_{h,p}) - \mu u_{h,p}$, $R = \left[\frac{\partial u_{h,p}}{\partial \eta}\right]$ is the jump of the derivative of the numerical solution on the interior edges, $|||u|||^2 = a(u,u)$ is the energy norm (see [3] for details), and C_{res} is a constant which does not depend on h.

It is well known that reference solution based methods form a very effective class of adaptive techniques. The inequality (3.12) supplies a guaranteed upper bound; on the other hand, its use for the purpose of hp-adaptivity is a little bit complicated. Therefore, we should modify our algorithm as follows:

Initialize: solve the initial problem with small polynomial degree p on a coarse grid

Repeat:

S1 compute the error using one of the quantities from Step 4 of the algorithm described at the end of Section 3.3.

S2 if the error is small then use (3.12)

S2a if $\eta_{\text{res}} < \texttt{TOL}$ terminate

S2b else do a brute-force adaptive step (both h and p) and go to step S4. (See Remark 3.2.)

S3 else determine on which elements in the grid and how to refine/derefine

S4 compute the new solution and go to S1

Remark 3.2 *By adopting this strategy we can avoid the need for building an effective hp-adaptive technique that impinges on a residual-based estimator. We propose to use a brute-force adaptive step in order to try to avoid a major modification of the original algorithm. It is possible to use the hp-adaptive technique from [41] that provides lower degrees of freedom however, the method would become more complicated.*

Remark 3.3 *There exists another possible correction that was published in [57, 58]. The key is the measurement of the oscillation of f, using which it is possible to prove that the adaptive method will be convergent. Each of the elements should be checked twice: first according to the error estimator and after that according to the oscillation of the right hand side. If the refinement involves all "problematic" elements the method will converge. In our case the only question would be how to decide which refining option should be used over the elements where there is considerable oscillation: they could be refined both in h and p.*

Chapter 4

Differences between the discrete weak and strong maximum principles for elliptic operators

The easiest way of adaptivity is using first order polynomials and h-adaptivity. In this case a natural requirement can be to use such a method that possess the discrete maximum principle, however, there are several different maximum principles: (strictly) weak/strong. While the discrete weak maximum principle was extensively investigated in the last decades, the discrete strong maximum principle has not been thoroughly analysed. In [44] and in [48] a sufficient algebraic condition was given, while in [22] the positivity of the discrete Green function was investigated (which is in a close relation with the discrete strong maximum principle) in a special case. However, a sufficient and necessary condition was missing. The scope of this chapter is to fill this gap. This chapter is based on [53].

4.1 Maximum Principles

In this section we list the definitions of continuous maximum principles for linear elliptic operators and the important theorems about them, based on [26]. We study elliptic operators, and not elliptic PDEs, since this way is more comfortable, and clearly the qualitative properties of some PDEs depend on the qualitative properties of the corresponding operators.

Let $\Omega \subset \mathbb{R}^d$ be an open and bounded domain with boundary $\partial\Omega$, and $\overline{\Omega} = \Omega \cup \partial\Omega$. We investigate the elliptic operator $\mathcal{A}$, $\operatorname{dom}\mathcal{A} = C^2(\Omega) \cap C(\overline{\Omega})$, defined in divergence form as

$$\mathcal{A}u = -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_j}\left(\mathcal{K}_{ij}\frac{\partial u}{\partial x_i}\right) + \mu u\,, \tag{4.1}$$

where $\mathcal{K}_{ij} \in C^1(\Omega)$, $0 \le \mu \in C(\Omega)$.

Definition 4.1 *We say that the operator $\mathcal{A}$ defined in (4.1) possesses*

- *the weak maximum principle (wMP) if the following implication holds*

$$\mathcal{A}u \leq 0 \textit{ in } \Omega \quad \Rightarrow \quad \max_{\overline{\Omega}} u \leq \max\{0, \max_{\partial\Omega} u\}\,;$$

- *the strong maximum principle (sMP) if it possesses the wMP, moreover, the following implication holds*

$$\mathcal{A}u \leq 0 \textit{ in } \Omega \quad \textit{and} \quad \max_{\Omega} u = \max_{\overline{\Omega}} u = m \geq 0 \quad \Rightarrow \quad u \equiv m \textit{ in } \overline{\Omega}\,;$$

- *the strictly weak maximum principle (WMP) if the following implication holds*

$$\mathcal{A}u \leq 0 \textit{ in } \Omega \quad \Rightarrow \quad \max_{\partial\Omega} u = \max_{\overline{\Omega}} u\,;$$

- *the strictly strong maximum principle (SMP) if it possesses the WMP, moreover, the following implication holds*

$$\mathcal{A}u \leq 0 \textit{ in } \Omega \quad \textit{and} \quad \max_{\Omega} u = \max_{\overline{\Omega}} u = m \quad \Rightarrow \quad u \equiv m \textit{ in } \overline{\Omega}\,.$$

Theorem 4.2 *If operator $\mathcal{A}$ defined in (4.1) is uniformly elliptic (there exist $\lambda > 0$ such that $\lambda||\xi||_E^2 \leq \mathcal{K}(x)\xi \cdot \xi$, for all $x \in \Omega$, $\xi \in \mathbb{R}^d$, where $||\cdot||_E$ denotes the Euclidean norm in $\mathbb{R}^d$, see Definition 1.2) and*

- *$\mu \geq 0$, then it possesses the wMP;*
- *$\mu \geq 0$, moreover, Ω is connected, then it possesses the sMP;*
- *$\mu = 0$, then it possesses the WMP;*
- *$\mu = 0$, moreover, Ω is connected, then it possesses the SMP.*

Remark 4.3 *Sometimes the case $\mu = 0$ is called strong elliptic maximum principle, but we wanted to reserve this name to another property.*

The requirements under which the operator possesses a weak maximum principle can be weakened, see, e.g., [17].

Finally, we mention that it is possible to define minimum principles and to get similar theorems, due to the linearity of operator $\mathcal{A}$. More information about maximum and minimum principles can be found in [26].

4.2 Discrete Maximum Principles

To obtain a simpler problem (or a sequence of simpler problems) from an elliptic PDE, usually a discretisation method is applied, in our case it will be a finite element method. This discretisation method leads to a system of linear algebraic equations, where the discrete operator corresponding to the operator $\mathcal{A}$ in (4.1) is the stiffness matrix.

In the following we define discrete maximum principles for such a discrete operator, i.e., for a matrix. We choose the natural way (independently of the original problem), which results in the adequate definition, corresponding to the definition of the last section, if finite element method with linear or multilinear elements is applied. However, it should be mentioned that in case of higher order elements this approach is not applicable.

We use the following typesetting, similarly as in Chapter 1: $\mathbf{0}$ denotes the zero matrix (or vector), $\mathbf{e}$ is the vector all coordinates of which are equal to 1. The dimensions of these vectors and matrices will be clear from the context. $\mathbf{B} \geq \mathbf{0}$ ($\mathbf{B} > \mathbf{0}$) or $\mathbf{a} \geq \mathbf{0}$ ($\mathbf{a} > \mathbf{0}$) means that all the elements of $\mathbf{B}$ or $\mathbf{a}$ are nonnegative (positive). $\max \mathbf{a}$ denotes the maximal element of the vector $\mathbf{a}$. The symbol $\max\{0, \mathbf{a}\}$ denotes $\max\{0, \max \mathbf{a}\}$.

We will investigate the matrix $\mathbf{A} = [\mathbf{A}_0 | \mathbf{A}_\partial] \in \mathbb{R}^{N \times \overline{N}}$, where $\mathbf{A}_0 \in \mathbb{R}^{N \times N}$, $\mathbf{A}_\partial \in \mathbb{R}^{N \times N_\partial}$, $\overline{N} = N + N_\partial$, acting on the vector $\mathbf{u} = [\mathbf{u}_0 | \mathbf{u}_\partial]^T \in \mathbb{R}^{\overline{N}}$, $\mathbf{u}_0 \in \mathbb{R}^N$, $\mathbf{u}_\partial \in \mathbb{R}^{N_\partial}$. The partitioned forms are constructed by taking into consideration the separation of the interior and boundary points. We assume that $N, N_\partial \geq 2$.

Definition 4.4 *We say that a matrix* $\mathbf{A}$ *possesses*

- *the discrete weak maximum principle (DwMP) if the following implication holds*

$$\mathbf{A}\mathbf{u} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u} \leq \max\{0, \mathbf{u}_\partial\}\,;$$

- *the discrete strong maximum principle (DsMP) if it possesses the DwMP, moreover, the following implication holds*

$$\mathbf{A}\mathbf{u} \leq \mathbf{0} \quad \textit{and} \quad \max \mathbf{u} = \max \mathbf{u}_0 = m \geq 0 \quad \Rightarrow \quad \mathbf{u} = m\mathbf{e}\,;$$

- *the discrete strictly weak maximum principle (DWMP) if the following implication holds*

$$\mathbf{A}\mathbf{u} \leq \mathbf{0} \quad \Rightarrow \quad \max \mathbf{u}_\partial = \max \mathbf{u}\,;$$

- *the discrete strictly strong maximum principle (DSMP) if it possesses the DWMP, moreover, the following implication holds*

$$\mathbf{A}\mathbf{u} \leq \mathbf{0} \quad \textit{and} \quad \max \mathbf{u} = \max \mathbf{u}_0 = m \quad \Rightarrow \quad \mathbf{u} = m\mathbf{e}\,.$$

4.3 Differences between the Discrete Weak and Strong Maximum Principles

The only difference between the conditions in Theorem 4.2 for the weak and strong maximum principles is the connectedness of the domain Ω. Next, we investigate this question in the discrete case.

The following theorem gives necessary and sufficient conditions for the discrete weak and strong maximum principles. The first part (DwMP) of it is from [17], the third (DwMP) is a joint result of Faragó and Mincsovics, published in [27]. The second (DsMP) and fourth (DSMP) parts of it are new.

Theorem 4.5 *The matrix* $\mathbf{A}$ *possesses*

- *the DwMP if and only if the following three conditions hold*

  $$\text{(w1) } \mathbf{A}_0^{-1} \geq \mathbf{0}\,; \qquad \text{(w2) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}\,; \qquad \text{(w3) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \leq \mathbf{e}\,.$$

- *the DsMP if and only if the following three conditions hold*

  $$\text{(s1) } \mathbf{A}_0^{-1} > \mathbf{0}\,; \qquad \text{(s2) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial > \mathbf{0}\,;$$
  $$\text{(s3) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} < \mathbf{e} \quad \textit{or} \quad -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}\,.$$

- *the DWMP if and only if the following three conditions hold*

  $$\text{(W1) } \mathbf{A}_0^{-1} \geq \mathbf{0}\,; \qquad \text{(W2) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial \geq \mathbf{0}\,; \qquad \text{(W3) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}\,.$$

- *the DSMP if and only if the following three conditions hold*

  $$\text{(S1) } \mathbf{A}_0^{-1} > \mathbf{0}\,; \qquad \text{(S2) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial > \mathbf{0}\,; \qquad \text{(S3) } -\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}\,.$$

proof:

- We begin with the DSMP case.

  - First, we assume (S1)–(S3), then

    $$\mathbf{u}_0 = \mathbf{A}_0^{-1}\mathbf{A}\mathbf{u} - \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{u}_\partial$$

    holds. (It follows from (S1) that $\mathbf{A}_0^{-1}$ exists.) Let us assume that $\mathbf{A}\mathbf{u} \leq \mathbf{0}$. We write $\mathbf{u}_0 = m\mathbf{e} - \mathbf{h}_0$, $\mathbf{u}_\partial = m\mathbf{e} - \mathbf{h}_\partial$, where both $\mathbf{h}_0, \mathbf{h}_\partial \geq \mathbf{0}$ have a 0 coordinate (i.e., $\max \mathbf{u} = \max \mathbf{u}_0 = m$). Thus

    $$m\mathbf{e} - \mathbf{h}_0 = \mathbf{A}_0^{-1}\mathbf{A}\mathbf{u} - \mathbf{A}_0^{-1}\mathbf{A}_\partial m\mathbf{e} + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{h}_\partial. \qquad (4.2)$$

Using (S3) we get

$$\mathbf{h}_0 = \mathbf{A}_0^{-1}(-\mathbf{A}\mathbf{u}) - \mathbf{A}_0^{-1}\mathbf{A}_\partial \mathbf{h}_\partial \,. \tag{4.3}$$

Using (S1), (S2) and the fact that $\mathbf{h}_0$ has a 0 coordinate yields that $-\mathbf{A}\mathbf{u} = \mathbf{0}$ and $\mathbf{h}_\partial = \mathbf{0}$. These imply $\mathbf{h}_0 = \mathbf{0}$.

- Second, we assume the DSMP. Then the DWMP holds, thus (W1)–(W3) hold. We can choose freely $\mathbf{A}\mathbf{u} \leq \mathbf{0}$, $\mathbf{h}_\partial \geq \mathbf{0}$ in (4.3).

  First, we set $\mathbf{h}_\partial = \mathbf{0}$ and we assume that $\mathbf{A}_0^{-1}$ has a 0 element, let it be the ij-th entry of the matrix. We choose the j-th coordinate of $-\mathbf{A}\mathbf{u}$ as 1, the others as 0, then the i-th coordinate of $\mathbf{h}_0$ is 0. If in the j-th column there is a positive entry, then $\mathbf{h}_0 \neq \mathbf{0}$, which is a contradiction. Otherwise, the matrix $\mathbf{A}_0^{-1}$ has a zero column, which is a contradiction, too, since it is invertible.

  Second, we set $\mathbf{A}\mathbf{u} = \mathbf{0}$, and we assume that $-\mathbf{A}_0^{-1}\mathbf{A}_\partial$ has a 0 element, let it be the ij-th entry of the matrix. We choose the j-th coordinate of $\mathbf{h}_\partial$ as 1, the others as 0, then the i-th coordinate of $\mathbf{h}_0$ is 0, but $\mathbf{h}_\partial \neq \mathbf{0}$, which is a contradiction.

- We finish with the DsMP case.

  - First, we assume (s1)–(s3). If $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}$ holds, then we can adopt the proof of the DSMP case. If $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} < \mathbf{e}$ holds, then (4.3) is modified as $\mathbf{h}_0 > \mathbf{A}_0^{-1}(-\mathbf{A}\mathbf{u}) - \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{h}_\partial$, which excludes the possibility that $\mathbf{h}_0$ has a 0 coordinate.
  - Second, we assume the DsMP. We get (s1), (s2) by putting $m = 0$ into (4.2), then the argumentation of the DSMP case can be repeated. To get (s3), we assume that $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \not< \mathbf{e}$ and $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \neq \mathbf{e}$, i.e., $\mathbf{e} + \mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e}$ has a 0 and a positive coordinate, too, let them be the i-th one and the j-th one, respectively. Choosing $m = 1$, $\mathbf{A}\mathbf{u} = \mathbf{0}$, $\mathbf{h}_\partial = \mathbf{0}$ yields that the i-th coordinate of $\mathbf{h}_0$ is 0 and the j-th one is positive, which is a contradiction.

□

If we compare the continuous and the discrete case, we can conclude that the condition (w3) $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \leq \mathbf{e}$ corresponds to $\mu \geq 0$, (W/S3) $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}$ corresponds to $\mu = 0$, moreover, (s3) can be shed more light on if we notice the fact that u constant implies $\mu = 0$. Note that $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} = \mathbf{e}$ is equivalent to $\mathbf{A}\mathbf{e} = \mathbf{0}$, and $\mathbf{A}\mathbf{e} \geq \mathbf{0}$ implies $-\mathbf{A}_0^{-1}\mathbf{A}_\partial\mathbf{e} \leq \mathbf{e}$, but here the reverse is not true.

(s/S1) and (s/S2) correspond to the connectedness of the domain Ω. One can easily see that (s/S1) implies the irreducibility of $\mathbf{A}_0$, which means that all the discrete interior points are in contact with each other, which is clearly a kind of discrete connectedness property. [17] gives practical conditions to satisfy the DwMP by introducing the notion of generalized nonnegative type. In [44] the DsMP was proved for this class of matrices and it was observed that if the matrix $\mathbf{A}_0$ is irreducibly diagonally dominant (the generalized nonnegative type contains this property), then (s/S1) is fulfilled, the proof can be found in [69]. However, for discrete weak maximum principles it is not needed, as it was observed in [34]. To ensure (s/S2), one possibility is to require $\mathbf{A}_\partial \leq \mathbf{0}$ and at least one nonzero element in every column (with (s/S1)), which can be interpreted as all of the discrete boundary points are in contact with the discrete interior points.

We can conclude that irreducibility is necessary for DsMP and DSMP (but it is not sufficient). Anyway, this would be the key-concept, if we want something to emphasize.

4.4 Numerical Examples

In this section we present some numerical examples, constructed with the help of Matlab. In all examples we used linear finite element discretisation. We focus on the irreducibility property, i.e., we give examples in which the discrete domain is not connected from a point of view. This can easily happen when the domain consists of two relatively large areas connected in the middle with a thin "path". For instance in this case the program package COMSOL can produce qualitatively incorrect meshes, too.

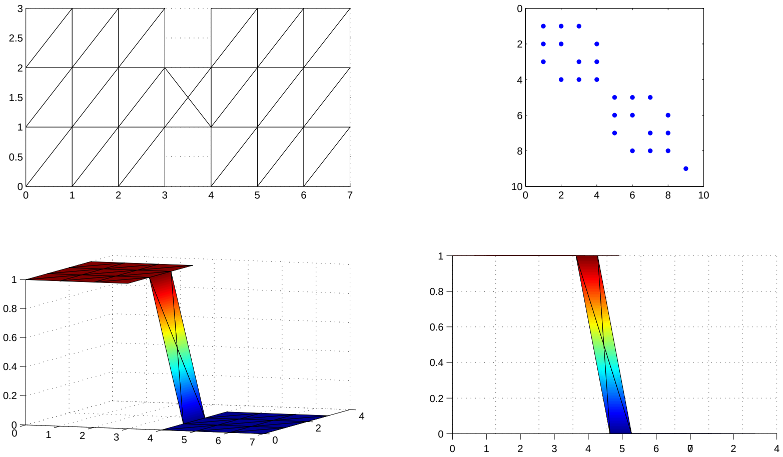

Figure 4.1: 1. Example: The mesh results in a reducible matrix. The DsMP failed while the DwMP was fulfilled.

In the first three examples $\mathcal{A} = -\Delta$, in the fourth it is defined as $\mathcal{A} = -\Delta u + 128u$. In all examples $\mathcal{A}u = 0$. In the first two u is defined as 1 on the boundary of the left square, 0 on the boundary of the right square and linearly decreasing from 1 to 0 on the boundary of the middle square. The boundary condition of the third example differs only on the middle part: on the left part of the boundary of it, i.e. on $\{(x, y) : x \in [3, 3.5], y \in \{1, 2\}\}$, u is 1, then linearly decreasing from 1 to 0 on the right part of the boundary of the middle square i.e. on

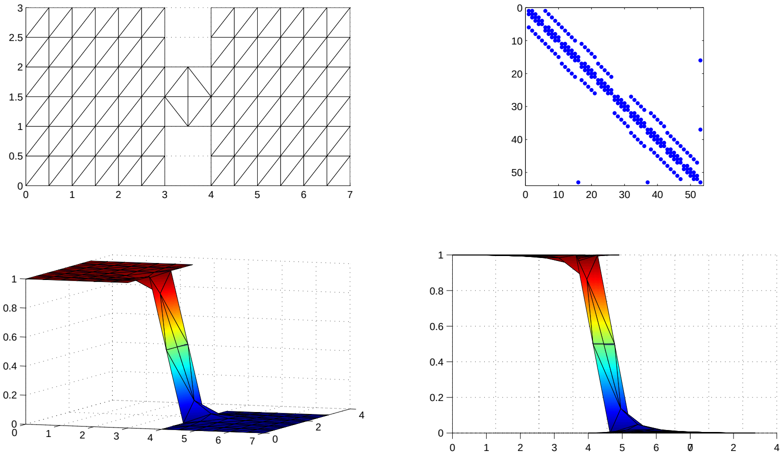

Figure 4.2: 2. Example: The mesh results in an irreducible matrix. Both of the DsMP and DwMP were fulfilled.

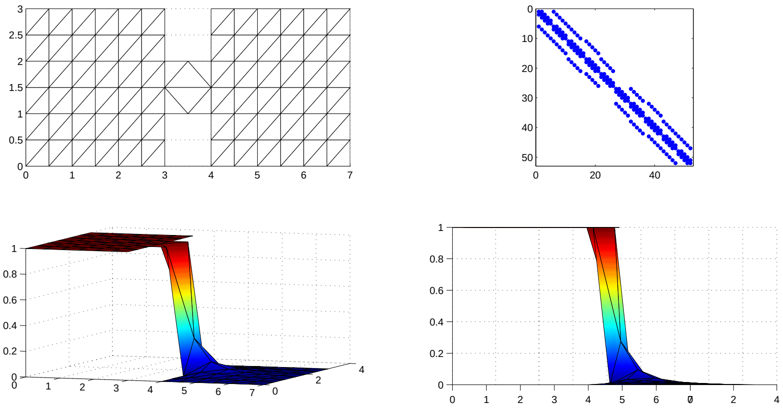

Figure 4.3: 3. Example: The mesh results in a reducible matrix. The DsMP failed while the DwMP was fulfilled.

$\{(x, y) : x \in [3.5, 4], y \in \{1, 2\}\}$. The fourth example is similar to the first two.

The arrangement within the figures is as follows. The top left panel presents the mesh, the top right panel presents the nonzero elements of the matrix $\mathbf{A}_0$, and in the bottom panels u_h is plotted from two different angles, the right one shows us better where the function is constant.

The first example shows us how an inadequate mesh can result in a reducible matrix and so losing the DSMP (but the DWMP is fulfilled). The second is the "good" example, here both discrete maximum principles are fulfilled. In [22] a mesh is presented, this is the third example here, which seems to be good at first sight, but the two right angles damage the connection of the two seemingly connected points in the middle, cf. [34], too.

The fourth example presents a mesh, which results in losing the DsMP, in addition to that the DwMP is fulfilled. It is caused surprisingly by the usage of equilateral triangles.

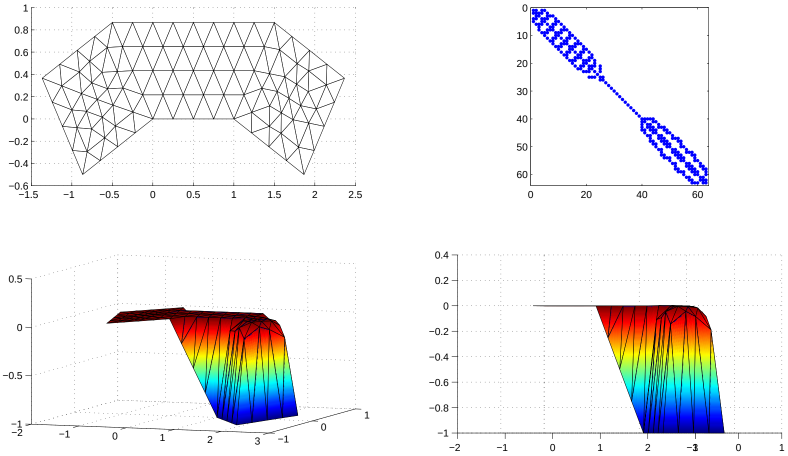

Figure 4.4: 4. Example: The mesh which contains equilateral triangles can results in a reducible matrix, too. The DsMP failed while the DwMP fulfilled.

4.4.1 Adaptivity

Examining the four above mentioned examples we can see that the first three cases can be handled using h-adaptivity. As the mesh gets denser the problematic part will disappear, there will be connections between the left and right squares.

However, the fourth case is not that easy. Using h-adaptivity we have several opportunities: divide a triangle into four smaller similar triangle or divide them into two using bisection. Naturally, as hanging nodes are not allowed, some mesh corrections are needed after the division of an element. Although in every refining techniques if more elements are needed to be refined at a given part, the mesh geometry will not change, therefore the stiffness matrix will be reducible.

Chapter 5

Discrete weak maximum principle for Discontinuous Galerkin methods

In this chapter we investigate how weak maximum principle can be preserved on the discrete level when an interior penalty discontinuous Galerkin method is applied for the discretisation of a 1D elliptic operator. We give mesh conditions for the symmetric and for the incomplete method that establish some connections between the mesh size and the penalty parameter. We investigate the sharpness of these conditions, too. This chapter is based on [40].

The preservation of the weak maximum principle was extensively investigated for finite difference methods (FDM) and for finite element methods (FEM) with linear and continuous elements, but not in the context of the discontinuous Galerkin method. In this chapter we take the first step to fill this gap. Namely, we investigate an interior penalty discontinuous Galerkin method (IPDG) applied to a 1D elliptic operator (containing diffusion and reaction terms) and we show that it is possible to give reasonable and sufficient conditions for the weak maximum principle on the discrete level.

5.1 Discontinuous Galerkin method in one dimension

5.1.1 Problem setting

Let us set $\Omega = (0,1)$ and consider the following special elliptic operator $\mathcal{A}$, $\operatorname{dom}\mathcal{A} = C^2(\Omega) \cap C(\overline{\Omega})$, defined as

$$\mathcal{A}u = -(\kappa u')' + \mu u\,, \tag{5.1}$$

where $\kappa, \mu \in \mathbb{R}$, $\kappa > 0$, $\mu \geq 0$. It is clear that for this operator the weak maximum principle holds due to Theorem 1.47.

Note that continuity is an important qualitative property and it cannot be preserved by the discontinuous Galerkin method. This is one reason why we need to be careful, especially with the preservation of some milder qualitative properties which are in connection with the continuity. This leads directly to the investigation of weak maximum principle for the interior penalty discontinuous Galerkin method.

5.1.2 The construction of the IPDG elliptic operator

Here we briefly introduce the discontinuous Galerkin methods for one dimensional problems, for more details see Section 1.5.6.

Let us denote the mesh by $\mathcal{T}_h$ it is defined in the following way: $0 = x_0 < x_1 < x_2 < \ldots < x_{N-1} < x_N = 1$. Let us use the following notations $I_n = [x_{n-1}, x_n]$, $h_n = |I_n|$, $h_{n-1,n} = \max\{h_{n-1}, h_n\}$, (with $h_{0,1} = h_1$, $h_{N,N+1} = h_N$).

The next step is to define the space $\mathcal{P}_l(\mathcal{T}_h) := \{v : v|_{I_n} \in \mathcal{P}_l(I_n), \forall n = 1, 2, \ldots, N\}$ - piecewise polynomials over every interval with maximal degree l. For these functions we introduce the right and left hand side limits $v(x_n^+) := \lim\limits_{t \to 0^+} v(x_n + t)$, $v(x_n^-) := \lim\limits_{t \to 0^+} v(x_n - t)$ and jumps and averages over the mesh nodes as

$$[\![u(x_n)]\!] := u(x_n^-) - u(x_n^+) , \quad \{\!\{ u(x_n) \}\!\} := \frac{1}{2}(u(x_n^-) + u(x_n^+)) .$$

At the boundary nodes these are defined as

$$[\![u(x_0)]\!] := -u(x_0^+) , \; \{\!\{ u(x_0) \}\!\} := u(x_0^+) , \; [\![u(x_N)]\!] := u(x_N^-) , \; \{\!\{ u(x_N) \}\!\} := u(x_N^-) .$$

We fix the penalty parameter σ as $\sigma \geq 0$ and ε which can be any arbitrary number, but it is usually chosen from the set $\{-1, 0, 1\}$. After these preparations we are ready to define the (discrete) IPDG bilinear form as

$$a_{DG}(u, v) = \sum_{n=0}^{N-1} \int\limits_{x_n}^{x_{n+1}} \kappa u'(x) v'(x) \,\mathrm{d}x - \sum_{n=0}^{N} \{\!\{ \kappa u'(x_n) \}\!\} [\![v(x_n)]\!] +$$

$$\varepsilon \sum_{n=0}^{N} \{\!\{ \kappa v'(x_n) \}\!\} [\![u(x_n)]\!] + \sum_{n=0}^{N} \frac{\sigma}{h_{n,n+1}} [\![v(x_n)]\!] [\![u(x_n)]\!] + \int_0^1 \mu u(x) v(x) \,\mathrm{d}x .$$

Note that fixing the parameters σ, ε and the mesh $\mathcal{T}_h$ can be done in parallel.

The crucial step is the following. We fix a basis in the space $\mathcal{P}_l(\mathcal{T}_h)$. First we need to choose $l = 1$ for the same reasons as in the FEM case discussed in Section 1.8. When choosing the basis functions we need to consider the following. If we want to use the Definition 1.49 and apply Theorems 1.50 and 1.51, then we need to choose basis functions that possess the important properties listed in Subsection 1.8.1. Let us recall them:

1. the subspace consists of continuous functions;

2. $\sum_{i=1}^{\overline{N}} \Phi_i(\mathbf{x}) = 1$ holds for all $\mathbf{x} \in \overline{\Omega}$;

3. $\Phi_i(\mathbf{x}) \geq 0$ holds for all $\mathbf{x} \in \overline{\Omega}$ and $i = 1, \dots \overline{N}$;

4. in a linear combination of the basis functions the coefficients represent the values of the resulting function at the points of $\overline{\mathcal{X}}$.

We already set aside continuity, but the next choice fulfils the second and third property and a milder version of the fourth, and this is enough for us.

We will use the notation Φ_1^i for the $(2(i-1)+1)^{\text{th}}$ basis functions, and Φ_2^i for the $(2(i-1)+2)^{\text{th}}$ basis functions, see Figure 5.1. On interval I_i the function Φ_1^i is the linear function with $\Phi_1^i(x_{i-1}^+) = 1$, $\Phi_1^i(x_i^-) = 0$ and Φ_2^i is the linear function with $\Phi_2^i(x_{i-1}^+) = 0$, $\Phi_2^i(x_i^-) = 1$, and these functions are zero outside I_i, see Figure 5.1.

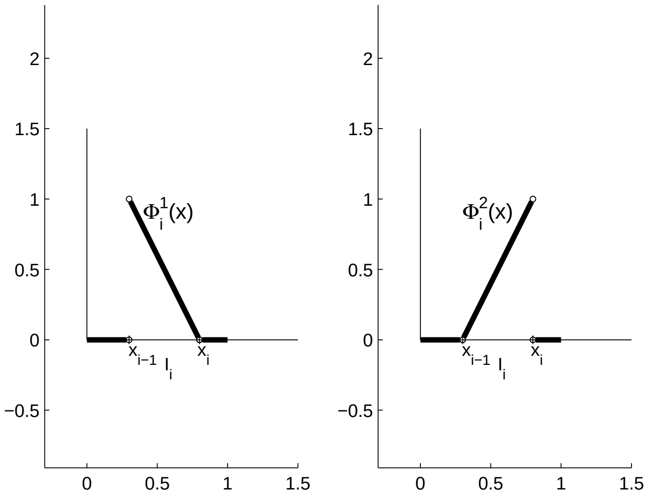

Figure 5.1: Φ_1^i and Φ_2^i

Finally, we construct the IPDG elliptic operator similarly to the way as we did in Section 1.8. However, there are slight differences. This matrix can be split in a partitioned form separating the (discrete) interior and boundary nodes as

$$\overline{\mathbf{A}} = \begin{bmatrix} \mathbf{A_0} & \mathbf{A}_\partial \\ \mathbf{B} & \mathbf{D} \end{bmatrix},$$

where $\overline{\mathbf{A}} \in \mathbb{R}^{(2N)\times(2N)}$, $\mathbf{A_0} \in \mathbb{R}^{(2N-2)\times(2N-2)}$, and the others are trivial. The $2N$ basis function are ordered as follows: the first $2N-2$ are the basis functions that belong to the interior nodes and they are numbered from left to right. The $(2N-1)^{\text{th}}$ belongs to the left boundary and the $2N^{\text{th}}$ belongs to the right boundary.

Note that the matrices $\mathbf{B}$ and $\mathbf{D}$ are not important from the point of view of the maximum principle, thus we can omit them. So, the matrix we need to investigate has the usual form $\mathbf{A} = [\mathbf{A}_0|\mathbf{A}_\partial]$.

Remark 5.1 *When working with the homogeneous Dirichlet boundary condition we could restrict a_{DG} to $D_1^0(\mathcal{T}_h) \times D_1^0(\mathcal{T}_h)$, where $D_1^0(\mathcal{T}_h) := \{v \in D_1(\mathcal{T}_h) : v(0) = v(1) = 0\}$ ($\Phi_1^1(x)$ and $\Phi_N^2(x)$ are excluded from the basis), although it is not a usual practice in the discontinuous Galerkin community. Let us denote the corresponding bilinear form by a_{DG}^{hD} and it is defined as*

$$a_{DG}^{hD}(u,v) = \sum_{n=0}^{N-1} \int\limits_{x_n}^{x_{n+1}} \kappa u'(x)v'(x)\,dx - \sum_{n=1}^{N-1} \{\!\{\kappa u'(x_n)\}\!\}\, [\![v(x_n)]\!] +$$

$$\varepsilon \sum_{n=1}^{N-1} \{\!\{\kappa v'(x_n)\}\!\}\, [\![u(x_n)]\!] + \sum_{n=1}^{N-1} \frac{\sigma}{h_{n,n+1}} [\![v(x_n)]\!]\, [\![u(x_n)]\!] + \int_0^1 \mu u(x)v(x)\,dx\,.$$

In this case the discrete operator simplifies to $\mathbf{A}_0$ and similarly to Remark 1.52. only (T1) or (P1) should be fulfilled.

In the following we calculate the elements of the matrix $\mathbf{A}$.

5.1.3 The exact form of the discrete operators

It is easy to check that

$$\partial_x \Phi_1^i(x) = -\frac{1}{h_i}\,, \qquad \partial_x \Phi_2^i(x) = \frac{1}{h_i}\,,$$

which means that the averages are

$$\left\{\!\!\left\{ \partial_x \Phi_1^i(x_k) \right\}\!\!\right\} = -\frac{1}{2h_i}\,, \qquad \left\{\!\!\left\{ \partial_x \Phi_2^i(x_k) \right\}\!\!\right\} = \frac{1}{2h_i}$$

at both endpoints x_k of I_i, with the exception of the boundary nodes, where there is no division by 2. Similarly the jumps are

$$\left[\!\left[\Phi_1^i(x_{i-1}) \right]\!\right] = -1\,, \qquad \left[\!\left[\Phi_2^i(x_i) \right]\!\right] = 1$$

and zero elsewhere. Using these facts we can calculate the matrix entries.

Summing them up we have the following discretisation matrices

$$\mathbf{A_0} = \begin{bmatrix} d_1 & r_1 & s_2 & & & & & & & \\ t_2 & e_2 & q_2 & w_2 & & & & & & \\ w_2 & q_2 & d_2 & r_2 & s_3 & & & & & \\ & s_2 & t_3 & e_3 & q_3 & w_3 & & & & \\ & & & & \ddots & & & & & \\ & & & w_i & q_i & d_i & r_i & s_{i+1} & & \\ & & & & s_{i-1} & t_i & e_i & q_i & w_i & \\ & & & & & & & \ddots & & \\ & & & & & & w_{N-1} & q_{N-1} & d_{N-1} & r_{N-1} \\ & & & & & & & s_{N-1} & t_N & e_N \end{bmatrix}, \quad \mathbf{A}_\partial = \begin{bmatrix} v_1 & 0 \\ s_1 & 0 \\ 0 & 0 \\ \vdots & \vdots \\ 0 & 0 \\ 0 & s_N \\ 0 & v_N \end{bmatrix},$$

where

$$\begin{aligned}
d_i &= \frac{\kappa}{2h_i} + \frac{\sigma}{h_{i,i+1}} + \frac{\kappa\varepsilon}{2h_i} + \mu\frac{h_i}{3}\,, && i = 1,\dots,N-1\,,\\
e_i &= \frac{\kappa}{2h_i} + \frac{\sigma}{h_{i-1,i}} + \frac{\kappa\varepsilon}{2h_i} + \mu\frac{h_i}{3}\,, && i = 2,\dots,N\,,\\
w_i &= \frac{\kappa\varepsilon}{2h_i}\,, && i = 2,\dots,N-1\,,\\
q_i &= -\frac{\kappa}{h_i} + \frac{\kappa}{2h_i} - \frac{\kappa\varepsilon}{2h_i} + \mu\frac{h_i}{6}\,, && i = 2,\dots,N-1\,,\\
r_i &= \frac{\kappa}{2h_{i+1}} - \frac{\sigma}{h_{i,i+1}} - \frac{\kappa\varepsilon}{2h_i}\,, && i = 1,\dots,N-1\,,\\
s_i &= -\frac{\kappa}{2h_i}\,, && i = 1,\dots,N\,,\\
t_i &= \frac{\kappa}{2h_{i-1}} - \frac{\sigma}{h_{i-1,i}} - \frac{\kappa\varepsilon}{2h_i}\,, && i = 2,\dots,N\,,\\
v_i &= -\frac{\kappa}{h_i} + \frac{\kappa}{2h_i} - \frac{\kappa\varepsilon}{h_i} + \mu\frac{h_i}{6}\,, && i = 1,\dots,N
\end{aligned}$$

and zero elsewhere.

5.2 Weak maximum principle for IPDG elliptic operators

Our aim is to get useful mesh conditions that guarantee the discrete weak maximum principle by using Theorem 1.51.

First we deal with (P1). To this aim we guarantee that the diagonal elements of the matrix $\mathbf{A_0}$ are nonnegative and the off-diagonal elements are nonpositive.

- d_i, e_i:

  we get the following conditions for ε

$$\varepsilon \geq -1 - \frac{2\sigma h_i}{\kappa h_{i,i+1}} - \frac{2\mu h_i^2}{3\kappa}\,, \quad i = 1,\dots,N-1$$

$$\varepsilon \geq -1 - \frac{2\sigma h_i}{\kappa h_{i-1,i}} - \frac{2\mu h_i^2}{3\kappa}\,, \quad i = 2,\dots,N\,.$$

- w_i:

  w_i should be nonpositive, which indicates

$$\varepsilon \leq 0 \tag{5.2}$$

in the case where we have more than two subintervals. See the third part of Remark 5.5 for the degenerate case. This means that $\varepsilon = 1$ is excluded generally.

- q_i:

  because of q_i we need to guarantee $-\frac{\kappa}{2h_i} - \frac{\kappa\varepsilon}{2h_i} + \mu\frac{h_i}{6} \le 0$, $i = 2, \ldots, N-1$, which means the following for ε

  $$\varepsilon \ge -1 + \frac{\mu h_i^2}{3\kappa}, \quad i = 2, \ldots, N-1\,.$$

  Or, rephrasing it for the mesh, we have

  $$h_i^2 \le \frac{3(1+\varepsilon)\kappa}{\mu}, \quad i = 2, \ldots, N-1$$

  in the case where $\mu \ne 0$. (In the case $\mu = 0$ we simply have $\varepsilon \ge -1$.)

- s_i:

  Inequality $s_i < 0$ always holds.

- r_i, t_i:

  we need to guarantee $\frac{\kappa}{2h_{i+1}} - \frac{\sigma}{h_{i,i+1}} - \frac{\kappa\varepsilon}{2h_i} \le 0$ and $\frac{\kappa}{2h_{i-1}} - \frac{\sigma}{h_{i-1,i}} - \frac{\kappa\varepsilon}{2h_i} \le 0$. After re-indexing t_i and reformulating we have

  $$\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \le \frac{2\sigma}{\kappa} \quad \text{and} \quad \frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \le \frac{2\sigma}{\kappa}, \quad i = 1, \ldots, N-1\,. \tag{5.3}$$

Finally, we show that there is no other restriction needed since the following Lemma is valid.

Lemma 5.2 *There exists a positive vector* $\mathbf{v}$ *with* $\mathbf{A_0 v} > \mathbf{0}$.

proof: Fist let us consider the case where $\mu = 0$ and $\kappa = 1$.

We choose the dominant vector $\mathbf{v}$ as the piecewise linear interpolation of the function $d(x) = c - x^2$ with the bases of Φ_j^i in the interior nodes and zero at $x = 0, 1$, where $c \ge 1$, see Figure 5.2. We prove that this choice is suitable.

Let us denote this interpolation by $\Pi_d(x)$ and $\mathbf{v}$ contains its coefficients, so $\Pi_d(x) = \sum_{(i,j)\in \mathrm{int}(\mathcal{T}_h)} v_{2(i-1)+j-1}\Phi_j^i(x)$, where the summation goes over all basis functions with exception of the two that belong to the boundary nodes, ($\Phi_1^1(x)$ and $\Phi_2^N(x)$). It is clear that $\mathbf{v} > 0$, and we need to prove that $\mathbf{A_0 v} > 0$ holds. The meaning of this inequality is that $a_{DG}(\Pi_d(x), \Phi_j^i(x)) > 0$ holds for all basis functions, since e.g. for the first coordinate of $\mathbf{A_0 v}$

$$(\mathbf{A_0 v})_1 = \sum_{(i,j)\in \mathrm{int}(\mathcal{T}_h)} v_{2(i-1)+j-1} a_{DG}\left(\Phi_j^i(x), \Phi_2^1(x)\right) =$$

$$a_{DG}\left(\sum_{(i,j)\in \mathrm{int}(\mathcal{T}_h)} v_{2(i-1)+j-1}\Phi_j^i(x), \Phi_2^1(x)\right) = a_{DG}\left(\Pi_d(x), \Phi_2^1(x)\right)\,.$$

Next we calculate these bilinear forms. Function is $\Pi_d(x)$ continuous, therefore its jumps are zero all over the nodes, which means we have to take into account neither the ε, nor the penalty terms.

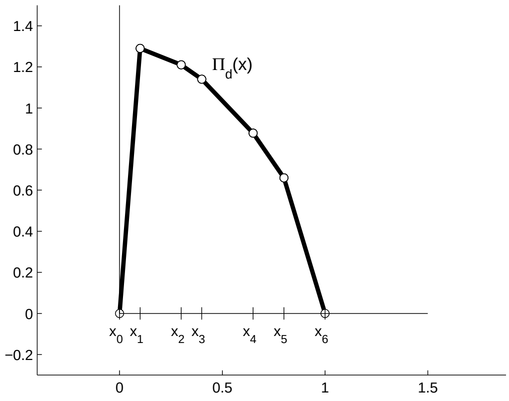

Figure 5.2: $\Pi_d(x)$ for $c = 1.3$

The derivative of $\Pi_d(x)$ can be calculated on every I_n. It is

- $\dfrac{c - x_1^2}{x_1}$ on I_1,
- $-\dfrac{x_i^2 - x_{i-1}^2}{x_i - x_{i-1}} = -(x_i + x_{i-1})$ on $I_i \quad i = 2, \dots, N-1$,
- $\dfrac{x_{N-1}^2 - c}{1 - x_{N-1}}$ on I_N.

This means

$$a_{DG}(\Pi_d(x), \Phi_2^1(x)) = \int_{I_1} \partial_x \Pi_d(x) \partial_x \Phi_2^1(x)\, dx - \{\!\{\partial_x \Pi_d(x_1)\}\!\}\ [\![\Phi_2^1(x_1)]\!] =$$

$$\left(\frac{c - x_1^2}{x_1}\right) \underbrace{\int_{I_1} \frac{1}{h_1}\, dx}_{=1} - \left(\frac{\frac{c - x_1^2}{x_1} - x_1 - x_2}{2}\right) \cdot 1 = \frac{c - x_1^2}{2x_1} + \frac{x_1 + x_2}{2}. \tag{5.4}$$

Similarly,

$$a_{DG}(\Pi_d(x), \Phi_1^2(x)) = \frac{c - x_1^2}{2x_1} + \frac{x_1 + x_2}{2}.$$

For $i \neq 1, N-1, N$

$$a_{DG}(\Pi_d(x), \Phi_2^i(x)) = \int_{I_i} \partial_x \Pi_d(x) \partial_x \Phi_2^i(x)\, dx - \{\!\{\partial_x \Pi_d(x_i)\}\!\}\ [\![\Phi_2^i(x_i)]\!] =$$

$$-(x_i + x_{i-1}) \int_{I_i} \frac{1}{h_i}\, dx - \left(-\frac{x_i + x_{i-1} + x_i + x_{i+1}}{2}\right) \cdot 1 = \frac{x_{i+1} - x_{i-1}}{2}. \tag{5.5}$$

For $i \neq 1, 2, N$

$$a_{DG}(\Pi_d(x), \Phi_1^i(x)) = \int_{I_i} \partial_x \Pi_d(x) \partial_x \Phi_i^i(x)\, dx - \{\!\{\partial_x \Pi_d(x_{i-1})\}\!\} \left[\!\left[\Phi_2^i(x_{i-1})\right]\!\right] =$$
$$-(x_i + x_{i-1}) \int_{I_i} -\frac{1}{h_i}\, dx - \left(-\frac{x_i + x_{i-1} + x_{i-1} + x_{i-2}}{2}\right) \cdot (-1) = \frac{x_i - x_{i-2}}{2}\,. \tag{5.6}$$

On I_{N-1}

$$a_{DG}(\Pi_d(x), \Phi_2^{N-1}(x)) = \int_{I_{N-1}} \partial_x \Pi_d(x) \partial_x \Phi_2^{N-1}(x)\, dx - \{\!\{\partial_x \Pi_d(x_{N-1})\}\!\} \left[\!\left[\Phi_2^{N-1}(x_{N-1})\right]\!\right] =$$
$$-(x_{N-2} + x_{N-1}) \int_{I_{N-1}} \frac{1}{h_{N-1}}\, dx - \left(\frac{-(x_{N-2} + x_{N-1}) + \frac{x_{N-1}^2 - c}{1 - x_{N-1}}}{2}\right) \cdot 1 =$$
$$-\frac{x_{N-2} + x_{N-1}}{2} + \frac{c - x_{N-1}^2}{2(1 - x_{N-1})}\,. \tag{5.7}$$

Finally,

$$a_{DG}(\Pi_d(x), \Phi_1^N(x)) = -\frac{x_{N-2} + x_{N-1}}{2} + \frac{c - x_{N-1}^2}{2(1 - x_{N-1})}\,.$$

We have to prove that these are positive values. The first three (5.4) - (5.6) are trivial. To prove that (5.7) is positive, some simple calculation is still needed. The following has to be satisfied

$$-\frac{x_{N-2} + x_{N-1}}{2} + \frac{c - x_{N-1}^2}{2(1 - x_{N-1})} > 0\,,$$
$$\frac{c - x_{N-1}^2}{1 - x_{N-1}} > x_{N-2} + x_{N-1}$$

and this holds, since

$$\frac{c - x_{N-1}^2}{1 - x_{N-1}} = \frac{(\sqrt{c} - x_{N-1})(\sqrt{c} + x_{N-1})}{1 - x_{N-1}} =$$
$$\frac{\sqrt{c} - x_{N-1}}{1 - x_{N-1}}(\sqrt{c} + x_{N-1}) > \sqrt{c} + x_{N-1} > 1 + x_{N-1} > x_{N-2} + x_{N-1}\,.$$

When $\kappa \neq 1$, we only have to multiply the matrix $\mathbf{A_0}$ with κ, which makes no difference in the sign of the product.

When $\mu > 0$, we have the extra terms $\int_{I_i} \mu \Phi_j^i(x) \cdot \Phi_l^i(x)$, where $j, l \in \{1, 2\}$. All functions are positive, so these integrals are also positive hence we have just increased the elements of $\mathbf{A_0}$, consequently increased the coordinates of $\mathbf{A_0 v}$. □

Accordingly, we can apply Theorem A.15, which completes the investigation of the condition (P1).

Property (P2) means that v_1 and v_N should be nonpositive, i.e.,

$$\varepsilon \geq \frac{-3\kappa + \mu h_i^2}{6\kappa} = -\frac{1}{2} + \frac{\mu h_i^2}{6\kappa} \geq -\frac{1}{2}\,, \qquad i = 1, N\,. \tag{5.8}$$

Note, it means that $\varepsilon = -1$ is excluded.

Property (P3) means that $\mathbf{0} \leq (\mathbf{A_0}|\mathbf{A}_\partial)\mathbf{e}$ should hold. It is equivalent to $a_{DG}(1, \Phi_j^i) \geq 0$ for $(i, j) \in \operatorname{int}(\mathcal{T}_h)$, for example, for the first coordinate of $(\mathbf{A_0}|\mathbf{A}_\partial)\mathbf{e}$

$$((\mathbf{A_0}|\mathbf{A}_\partial)\mathbf{e})_1 = \sum_{i=1}^{N}\sum_{j=1}^{2} 1 \cdot a_{DG}\left(\Phi_j^i(x), \Phi_2^1(x)\right) =$$

$$a_{DG}\left(\sum_{i=1}^{N}\sum_{j=1}^{2} 1 \cdot \Phi_j^i(x), \Phi_2^1(x)\right) = a_{DG}\left(1, \Phi_2^1(x)\right) .$$

The result of this matrix-vector product is

$$\begin{bmatrix} \frac{\mu h_1}{2} - \varepsilon\frac{\kappa}{h_1} \\ \frac{\mu h_2}{2} \\ \vdots \\ \frac{\mu h_{N-1}}{2} \\ \frac{\mu h_N}{2} - \varepsilon\frac{\kappa}{h_N} \end{bmatrix}$$

which is nonnegative if

$$\varepsilon \leq \frac{\mu h_i^2}{2\kappa}, \qquad i = 1, N . \tag{5.9}$$

We should note that we need to take it into consideration only in the degenerate case, when the interval is divided into two subintervals, since (5.2) is stricter.
Inequalities (5.8) and (5.9) can be pulled together as

$$-\frac{1}{2} + \frac{\mu h_i^2}{6\kappa} \leq \varepsilon \leq \frac{\mu h_i^2}{2\kappa}, \qquad i = 1, N \tag{5.10}$$

or rephrasing it for the mesh (if $\mu > 0$)

$$\frac{2\kappa\varepsilon}{\mu} \leq h_i^2 \leq \frac{3\kappa\,(2\varepsilon + 1)}{\mu}, \qquad i = 1, N . \tag{5.11}$$

5.2.1 The mesh conditions

In this subsection we sum up and systematize the conditions we obtained. Our plan is to give a “recipe” on how we should choose the parameters and the mesh to guarantee the discrete weak maximum principle. The trick is that we fix the order of the choices.

First we suppose that the interval $(0, 1)$ is divided into more than two subintervals.

Theorem 5.3 *Let* $\mathbf{A} = [\mathbf{A_0}|\mathbf{A}_\partial]$ *be the matrix constructed from* (5.1) *by the bilinear form* a_{DG} *as discribed in Subsection 5.1.2. This matrix possesses the discrete weak maximum principle if we choose*

- *ε as*

$$-\frac{1}{2} \leq \varepsilon \leq 0\,, \qquad \text{when } \mu = 0\,,$$
$$-\frac{1}{2} < \varepsilon \leq 0\,, \qquad \text{when } \mu > 0\,,$$

- *σ as*

$$\frac{\kappa(1-\varepsilon)}{2} \leq \sigma\,,$$

- *the mesh $\mathcal{T}_h$ as*

$$h_i^2 \leq \frac{3\kappa(2\varepsilon+1)}{\mu}\,, \qquad i = 1, N\,, \qquad \text{(fineness at the boundary)}$$

$$h_i^2 \leq \frac{3\kappa(\varepsilon+1)}{\mu}\,, \qquad i = 2, \ldots, N-1\,, \qquad \text{(fineness at the interior)}$$

$$\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \leq \frac{2\sigma}{\kappa} \quad \text{and} \quad \frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{\kappa}\,, \qquad i = 1, \ldots, N-1\,. \qquad \text{(uniformity)}$$

proof: Almost all of the conditions are simple consequences of the above calculations. The condition for σ can be derived from (5.3) by taking its minimum

$$\frac{2\sigma}{\kappa} \geq \frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \geq 1 - \varepsilon$$

□

Note that we have two types of mesh conditions, one is about the fineness of the mesh and the other is about the uniformity. The first determines the maximum size of the subintervals and it depends on the choice of ε, $\varepsilon = 0$ is the least restrictive. The second determines the maximum ratio of the size of the neighbouring subintervals and it depends on the choice of σ, $\sigma = \frac{\kappa(1-\varepsilon)}{2}$ is the most restrictive.

When working with homogeneous Dirichlet boundary conditions, we only have to fulfil (P1), see Remark 5.1. This leads to the following conditions.

Theorem 5.4 *Let* $\mathbf{A} = \mathbf{A_0}$ *be the matrix constructed from* (5.1) *by the bilinear form* a_{DG}^{hD} *as described in Subsection 5.1.2. This matrix possesses the discrete weak maximum principle if we choose*

- *ε as*

$$-1 \leq \varepsilon \leq 0\,, \qquad \text{when } \mu = 0\,,$$
$$-1 < \varepsilon \leq 0\,, \qquad \text{when } \mu > 0\,,$$

- *σ as*

$$\frac{\kappa(1-\varepsilon)}{2} \leq \sigma \,,$$

- *the mesh $\mathcal{T}_h$ as*

$$h_i^2 \leq \frac{3\kappa(\varepsilon+1)}{\mu} \,, \qquad i = 2, \ldots, N-1 \,, \qquad \text{(fineness at the interior)}$$

$$\frac{h_{i,i+1}}{h_{i+1}} - \frac{\varepsilon h_{i,i+1}}{h_i} \leq \frac{2\sigma}{\kappa} \quad \text{and} \quad \frac{h_{i,i+1}}{h_i} - \frac{\varepsilon h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{\kappa} \,, \quad i = 1, \ldots, N-1 \,. \qquad \text{(uniformity)}$$

Remark 5.5 *We investigate the most popular cases: $\varepsilon \in \{-1, 0, 1\}$, too.*

- $\varepsilon = -1$

  We can guarantee the discrete weak maximum principle in this case only if $\mu = 0$ holds and a_{DG}^{hD} is used as a discretisation.

  In this case (5.3) *simplified to the following*

$$\frac{h_{i,i+1}}{h_i} + \frac{h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{\kappa} \,, \quad i = 1, \ldots, N-1 \,. \tag{5.12}$$

  This has the consequence that σ needs to be chosen as $\sigma \geq \kappa$.

- $\varepsilon = 0$

  We have no additional restrictions in this case. The conditions simplified as

$$\frac{h_{i,i+1}}{h_{i+1}} \leq \frac{2\sigma}{\kappa} \quad \text{and} \quad \frac{h_{i,i+1}}{h_i} \leq \frac{2\sigma}{\kappa} \,, \quad i = 1, \ldots, N-1$$

  which can be pulled together as

$$\frac{h_{i,i+1}}{\min\{h_i, h_{i+1}\}} \leq \frac{2\sigma}{\kappa} \,, \quad i = 1, \ldots, N-1 \tag{5.13}$$

  since it is enough to guarantee that the inequality holds for the greater left-hand side. Thus σ needs to be chosen as $\sigma \geq \dfrac{\kappa}{2}$.

- $\varepsilon = 1$

  We can guarantee the discrete weak maximum principle in this case only if $(0, 1)$ is subdivided into two subintervals. Then (5.3) *leads to the following conditions*

$$\frac{h_{1,2}}{h_1} - \frac{h_{1,2}}{h_2} \leq \frac{2\sigma}{\kappa} \quad \text{and} \quad \frac{h_{1,2}}{h_2} - \frac{h_{1,2}}{h_1} \leq \frac{2\sigma}{\kappa} \,.$$

They can be pulled together as

$$\frac{h_{1,2} - \min\{h_1, h_2\}}{\min\{h_1, h_2\}} \leq \frac{2\sigma}{\kappa}\,. \tag{5.14}$$

When discretisation a_{DG} *is used, we have more conditions, namely* $\mu > 0$ *and*

$$\frac{2\kappa}{\mu} \leq h_i^2 \leq \frac{9\kappa}{\mu}\,, \qquad i = 1, 2\,.$$

Remark 5.6 *If we choose a different definition for* $h_{n-1,n}$*, namely,* $h_{n-1,n} = \min\{h_{n-1}, h_n\}$ *(c.f. [21, Ch. 4, Definition 4.5] and [61, Ch. 1]) the condition for* σ *will coincide with the condition that describes the relation between the neighbouring subintervals.*

5.3 Numerical examples and conclusion

5.3.1 Numerical examples - on the sharpness of the conditions

In this section we will investigate the mesh conditions we derived. Naturally, these cannot be sharp since we applied Theorem 1.51, whose conditions are only sufficient and not necessary. However, we will show that we obtain sharpness in some sense.

Example 5.7 *Let us set* $\kappa = 1$*,* $\varepsilon = 0$*,* $\sigma = 5$*,* $\mu = 0$*. First of all it is clear that condition* (5.10) *holds for* ε *and* (5.11) *is out of view. In this case for the mesh*

$$\mathcal{T}_h = \{0,\ 0.02,\ 0.22,\ 0.8,\ 1\}$$

the condition (5.13) *is sharp in the following sense. Let us modify this mesh as*

$$\mathcal{T}_h^m = \{0,\ 0.02,\ 0.22 + \frac{1}{10^m},\ 0.8,\ 1\}\,.$$

Let us consider the vector $\mathbf{v} = [-1,\ \frac{1}{10^m},\ 0,\ 0,\ 0,\ 0]^T$*, see Figure 5.3. The following calculation shows that the resulting right-hand side is nonpositive, which means that the weak maximum principle fails.*

The product $\mathbf{Av}$ *has only four nonzero coordinates:* $[-d_1 + r_1/10^m, -t_2 + e_2/10^m, -w_2 + q_2/10^m, s_2/10^m, 0, 0, 0, 0]^T$*. In this case* $h_{1,2} = h_2$*. Let us examine these terms.*

$$-d_1 + \frac{r_1}{10^m} = -\frac{1}{2h_1} - \frac{5}{h_2} + \frac{1}{10^m}\left(\frac{1}{2h_2} - \frac{5}{h_2}\right) = -\frac{1}{2h_1} - \frac{5}{h_2} - \frac{1}{10^m}\cdot\frac{9}{2h_2} < 0\,.$$

The second one is

$$-t_2 + \frac{e_2}{10^m} = -\frac{1}{2h_1} + \frac{5}{h_2} + \frac{1}{10^m}\left(\frac{1}{2h_2} + \frac{5}{2h_2}\right) = -25 + \frac{1}{h_2}\left(5 + \frac{11}{2\cdot 10^m}\right)\,.$$

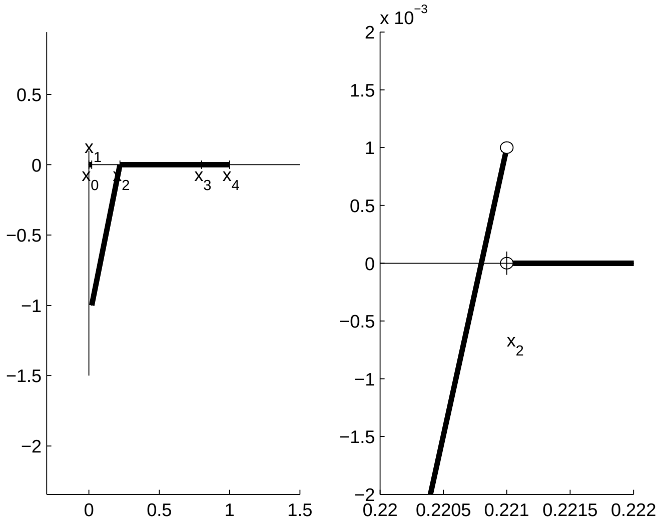

Figure 5.3: Left: the counterexample with $m = 3$. Right: the positive value at the node 0.221

Simple computations prove its negativity

$$\frac{5(10^{n+1}+11)}{2(10^n+5)} = \frac{5+\frac{11}{2\cdot 10^m}}{\frac{1}{5}+\frac{1}{10^m}} = \frac{1}{h_2}\left(5+\frac{11}{2\cdot 10^m}\right) < 25$$

$$5(10^{n+1}+11) < 50(10^n+5)$$

$$5\cdot 10^{n+1}+55 < 5\cdot 10^{n+1}+250\,.$$

The last two terms are easier to handle

$$-w_2 + q_2/(10^m) = 0 + \frac{q_2}{10^m} = \frac{1}{10^m}\left(-\frac{1}{h_1}+\frac{1}{2h_1}\right) = -\frac{1}{2\cdot 10^m\cdot h_1} < 0\,,$$

$$s_2/(10^m) = \frac{1}{10^m}\cdot\left(-\frac{1}{2h_1}\right) < 0\,.$$

Example 5.8 *Let us set* $\kappa = 1$, $\varepsilon = 1$, $\sigma = 5$, $\mu = 0$ *and use* a_{DG}^{hD}. *In the case that was discussed in the third part of Remark 5.5 the mesh* $\mathcal{T}_h = \{0,\, 1/12,\, 1\}$ *is sharp in the same sense as in the last example with respect to* (5.14). *Similarly as above, we modify the mesh as* $\mathcal{T}_h^m = \{0,\, 1/12 - 1/10^m,\, 1\}$ *and choose* $\mathbf{v} = [-1, \frac{1}{10^m}]^T$ *which breaks the weak maximum principle.*

Then $\mathbf{A_0 v} = [-d_1 + r_1/10^m, -t_2 + e_2/10^m]^T$, *where*

$$-d_1 + \frac{r_1}{10^m} = -\frac{2}{h_1} - \frac{5}{h_2} - \frac{1}{2h_1} + \frac{1}{10^m}\left(\frac{1}{2h_2} - \frac{5}{h_2} - \frac{1}{h_1}\right) =$$

$$-\frac{2}{h_1} - \frac{5}{h_2} - \frac{1}{2h_1} + \frac{1}{10^m}\left(-\frac{9}{2h_2} - \frac{1}{h_1}\right) < 0$$

and

$$-t_2 + \frac{e_2}{10^m} = -\frac{2}{h_1} + \frac{5}{h_2} + \frac{1}{2h_2} + \frac{1}{10^m}\left(\frac{1}{2h_2} + \frac{5}{h_2} + \frac{1}{2h_2}\right) = -\frac{2}{h_1} + \frac{5}{h_2} + \frac{1}{2h_2} + \frac{1}{10^m}\frac{6}{h_2}.$$

Similar calculations as before give

$$\begin{aligned}\frac{5}{h_2} + \frac{1}{2h_2} + \frac{1}{10^m}\frac{6}{h_2} &< \frac{2}{h_1}\\ h_1\left(11 + \frac{12}{10^m}\right) &< h_2\\ \left(\frac{1}{12} - \frac{1}{10^m}\right)\cdot\left(11 + \frac{12}{10^m}\right) &< \frac{11}{12} + \frac{1}{10^m}\\ \frac{11}{12} + \frac{1}{10^m} - \frac{11}{10^m} - \frac{12}{10^{2m}} &< \frac{11}{12} + \frac{1}{10^m}\end{aligned}$$

which holds for all $m > 0$*.*

5.3.2 Overview and outlook

First of all, we have shown that it is possible to guarantee the discrete weak maximum principle when IPDG discretisation is used. However, we should mention that our conditions are restrictive at the following points:

- the choice of the basis functions,
- $\varepsilon = 1$ is excluded from a practical point of view,
- we can handle $\varepsilon = -1$ only in special cases.

On the other hand, we could state that $\varepsilon = 0$ works very well from the discrete weak maximum principle point of view and the conditions suggest that we need to take into consideration a non-integer $\varepsilon \in \left(-\frac{1}{2}, 0\right)$, too.

We have shown with numerical examples that our conditions are sharp in some sense. The numerical examples and computational tests suggest the following points of interest:

- for the symmetric IPDG (5.12) does not seem to be sharp,
- the mesh condition (5.13) seems to be sharp only at the boundary, it could be slightly broken in the interior intervals without losing the weak maximum principle,
- for meshes that consist of more than two subintervals, the condition (5.14) seems to be irrelevant for the neighbouring elements.

5.4 Connection to the solvability of the system

For simplicity in this section we suppose that we are dealing with homogeneous Dirichlet boundary conditions.

We have seen in Chapter 1 that the solvability of the linear system in the DG case is the consequence of the coercivity of the bilinear form, that can only be guaranteed if the penalty parameter is large enough. For more details see Section 1.5.3 and Lemma 1.36.

On the other hand, if the linear system is an M-matrix then all of its eigenvalues have positive real parts, therefore the system is solvable. During this Chapter we have developed some conditions on the mesh which can ensure that the matrix is an M-matrix. Suppose that we have a given function $v \in \mathcal{P}_1(\mathcal{T}_h)$ and let us denote by $\tilde{v}$ the coefficients of v in the Lagrange basis functions. If the mesh conditions are fulfilled then we have

$$a_{DG}^{hD}(v,v) = \langle \mathbf{A_0}\tilde{v}, \tilde{v}\rangle \geq C_1 \|\tilde{v}\|_E^2 \geq C_1 C_2 \left\| v \right\|_{DG}^2 ,$$

which is the coercivity of the bilinear form. $\|\cdot\|_E$ is the Euclidian norm of the vector. Although, there are at least three questions that should be answered:

- C_1 is the smallest real part of the eigenvalues, does it depend on h?
- C_2 comes from the equivalency of the DG norm on V_{DG} and the Euclidian norm. Does it exist?
- If C_2 exists, then how does it depend on h?

It is well known that every two norms are equivalent on a finite dimensional space, see Lemma A.4. Let us define the following family of norms: $\|\tilde{u}\|_{E,*}^2 = \left\| u \right\|_{DG}$ where $\tilde{u} \in \mathbb{R}^{2m}$ is a vector, and u is the corresponding function, $u(x) = \sum_{i=1}^{m}\sum_{j=1}^{2} \tilde{u}_{2(i-1)+j-1}\Phi_j^i(x)$. With these notations we have two norms on $\mathbb{R}^{2m}$, $\|\cdot\|_E$ and $\|\cdot\|_{E,*}$ and they are equivalent. This gives an answer to the second question. The first one had been observed numerically, and unfortunately $C_1 \approx Ch$, hence the third one would be important only if C_2 would be $O(h^{-1})$. Let us consider the following case: divide $[0,1]$ into N equal subintervals, and let v be the following function: it is zero over the most left and most right subintervals and one otherwise (we suppose that the interval is divided into more than three subintervals) see Figure 5.4. In this case $\tilde{v} = (0,1,\ldots,1,0) \in \mathbb{R}^{2N-2}$. With these notations

$$\|\tilde{v}\|_2^2 = \|(0,1,\ldots,1,0)\|^2 = 2N-4,$$

$$\left\| v \right\|_{DG}^2 = \|\kappa\partial_x v\|_{[L^2(\Omega)]^d}^2 + \sum_{n=0}^{N} \frac{1}{h_{n,n+1}} [\![v]\!]^2 = \frac{1}{h_{1,2}} + \frac{1}{h_{N-1,N}} = \frac{1}{h} + \frac{1}{h} = 2N.$$

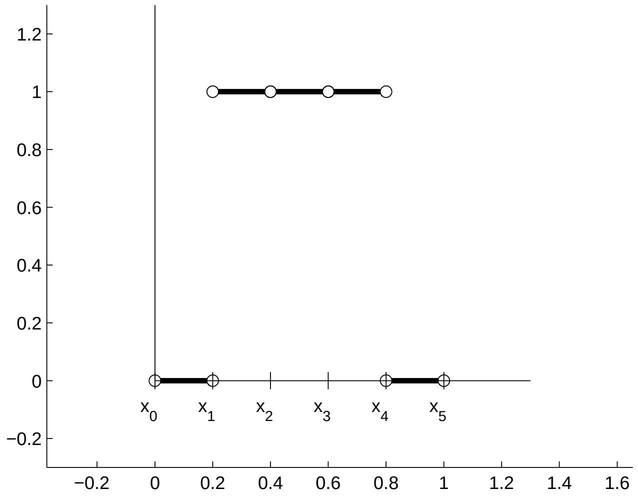

Figure 5.4: The function v.

Therefore C_2 cannot be $O(h^{-1})$. This means we can derive coercivity if the mesh conditions are fulfilled, but only with a constant that depends on h. This means, using the similar argument as in Section 1.5.3, we can have convergence, but with lower rate.

The interesting part is that in the $\varepsilon = 0$ case nothing changes if we are using higher order polynomials. The key is: we can define basis functions for arbitrary polynomial degree, in which the stiffness matrix has a special form

$$\overline{\mathbf{A}}_p = \begin{bmatrix} \mathbf{A}_0 & \mathbf{M} \\ 0 & I \end{bmatrix},$$

where $\mathbf{A}_0$ is the matrix, belonging to the first order polynomials, I is the identity matrix with the proper size, p is the maximal polynomial order. $\mathbf{M}$ is out of interest, because the eigenvalues of $\overline{\mathbf{A}}_p$ are the same as those of $\mathbf{A}$'s and 1.

Over the interval I_i, we will use $\Phi_1^i(x)$ and $\Phi_2^i(x)$. We will construct the second order polynomial $\Phi_3^i(x)$ such that it will possess the following properties:

1. $\Phi_3^i(x_{i-1}) = 0,$
2. $\Phi_3^i(x_i) = 0,$
3. $\int_{I_i} \partial_x \Phi_3^i(x)\, dx = 0,$
4. $\int_{I_i} (\partial_x \Phi_3^i(x))^2\, dx = 1.$

We seek Φ_3^i in the following form $\Phi_3^i(x) = a_0 + a_1 x + a_2 x^2$. The above four requirements could be interpreted as a system of linear equations. It can be seen, that the fourth row is the linear combination of the second and the third, therefore it can be omitted.

First we should check that the first two conditions implies the third one:

$$(1) : a_0 + a_1 x_{i-1} + a_2 x_{i-1}^2,$$
$$(2) : a_0 + a_1 x_i + a_2 x_i^2,$$
$$(3) : \int_{x_{i-1}}^{x_i} a_1 + 2a_2 x \, dx = a_1(x_i - x_{i-1}) + a_2(x_i^2 - x_{i-1}^2) = (2) - (1).$$

Hence $\Phi_3^i(x) = c(x - x_{i-1})(x - x_i)$ fulfils the first three conditions, and c can be chosen such that Φ_3^i satisfies the last requirement.

Similarly we can define $\Phi_n^i(x)$ in such a way that

1. $\Phi_n^i(x_{i-1}) = 0,$
2. $\Phi_n^i(x_i) = 0,$
3. $\int_{I_i} x^i \partial_x \Phi_n^i(x) \, dx = 0, \forall i = 0, \dots, n-2,$
4. $\int_{I_i} (\partial_x \Phi_n^i(x))^2 \, dx = 1.$

Again, the first two imply that $\int_{I_i} \partial_x \Phi_n^i(x) \, dx = 0$, therefore, the third one should be required only for $i = 1, \dots, n-2$, which means we have a linear system with a coefficient matrix: $K \in R^{(n-1)\times n}$. It has infinite number of solutions, we pick the solution that satisfies the last requirement.

See Table 5.1 for the first nine polynomials over the reference domain $(0, 1)$. Over a subinterval $I_i = (x_{i-1}, x_i)$ there is a small difference in the coefficient, the functions have to be multiplied by the factor c_i which will be determined below.

Using the connection between the reference element basis functions and the physical element basis functions $\Phi_j^i(x) := \Phi_j^{\Omega_0}\left(\frac{x - x_{i-1}}{h_i}\right)$ we get that for all $k, j \in \{1, \dots, n\}$ the following holds

$$\int_{I_i} \left(c_i \partial_x \Phi_j^i(x)\right) \left(c_i \partial_x \Phi_k^i(x)\right) \, dx =$$
$$c_i^2 \int_{I_i} \left(\partial_x \left(\Phi_j^{\Omega_0}\left(\frac{x - x_{i-1}}{h_i}\right)\right)\right) \left(\partial_x \left(\Phi_k^{\Omega_0}\left(\frac{x - x_{i-1}}{h_i}\right)\right)\right) \, dx =$$
$$\frac{c_i^2}{h_i^2} \int_{I_i} \left(\partial_x \Phi_j^{\Omega_0}\left(\frac{x - x_{i-1}}{h_i}\right)\right) \left(\partial_x \Phi_k^{\Omega_0}\left(\frac{x - x_{i-1}}{h_i}\right)\right) \, dx =$$
$$\frac{c_i^2 h_i}{h_i^2} \int_{\Omega_0} \left(\partial_x \Phi_j^{\Omega_0}(\xi)\right) \left(\partial_x \Phi_k^{\Omega_0}(\xi)\right) \, d\xi = \frac{c_i^2}{h_i} \delta_{jk},$$

where we used the substitution $\xi = \frac{x - x_{i-1}}{h_i}$. To fulfil requirement 4. c_i has to be chosen as $c_i = \sqrt{h_i}$. Requirements 1.-3. are fulfilled automatically.

After defining these polynomials let us examine the stiffness matrix.

First let us discuss the identity part. According to the first and second requirements u and v are bubble functions, therefore $a_{DG}^{hD}(u, v)$ simplifies only to the integrals over the elements,

p	polynomial
2	$\sqrt{3}x(1-x)$
3	$\sqrt{5}x(1-x)(-2x+1)$
4	$\sqrt{7}x(1-x)(-5x^2+5x+1)$
5	$3x(1-x)(-14x^3+21x^3-9x+1)$
6	$\sqrt{11}x(1-x)(42x^4-84x^3+56x^2-14x+1)$
7	$\sqrt{13}x(1-x)(-132x^5+330x^4-300x^3+120x^2-20x+1)$
8	$\sqrt{15}x(1-x)(429x^6-1287x^5+1485x^4-825x^3+225x^2-27x+1)$
9	$\sqrt{17}x(1-x)(-1430x^7+5005x^6-7007x^5+5005x^4-1925x^3+385x^2-35x+1)$

Table 5.1: Orthogonal polynomials (in the case of $\varepsilon = 0$) defined over the reference domain $(0, 1)$ up to degree 9.

since the jumps of the bubbles are 0. The last requirement ensures that if $u = v$, then the matrix entry is equal to 1, the third ensures that if u and v are defined on the same subinterval but their degrees are different, then they are orthogonal, hence the integral of their derivatives is 0. (If their supports are different subintervals then the $a_{DG}^{hD}(u, v)$ is obviously zero.)

Let us suppose that we are in the 0 part, which means we ought to calculate $a_{DG}^{hD}(u, v)$ when the degrees of u and v are one and higher than one, respectively. $\varepsilon = 0$ means we have only those terms that contain the jumps of v, which are zero, and the integral of the derivatives, which is also zero.

This approach has two advantages. First of all, if we design a mesh, that fulfils the mesh conditions we derived earlier, the linear system will be solvable due to the fact that all of its eigenvalues have positive real parts. Second, the computational costs of solving the linear system do not increase dramatically with the polynomial degrees, owing to the identity part of the matrix. The higher order terms are known from the right hand side, and the system can be reduced to the size of $\mathbf{A}_0$.

Appendix A

Mathematical supplement

A.1 Banach and Hilbert spaces

Let V be a real vector space.

Definition A.1 *The mapping* $||\cdot|| : V \to \mathbb{R}_{\geq 0}$ *is called norm, if it satisfies the following three conditions:*

1. $||v|| = 0 \Leftrightarrow v = 0$,
2. $||\lambda v|| = |\lambda|||v||$, $\forall v \in V$, $\forall \lambda \in \mathbb{R}$,
3. $||v + w|| \leq ||v|| + ||w||$, $\forall v, w \in V$ *(triangle inequality).*

Definition A.2 *The mapping* $||\cdot||_* : V \to \mathbb{R}_{\geq 0}$ *is called seminorm, if it satisfies 2. and 3. from the previous definition and*

1.' $v = 0 \Rightarrow ||v|| = 0$.

Definition A.3 (Equivalent norms) *Let us have two norms on* V*:* $||\cdot||_1$*, and* $||\cdot||_2$*. We say that these two norms are equivalent if there exist* $M > m > 0$ *constants such that* $\forall v \in V$*:*

$$m||\cdot||_1 \leq ||\cdot||_2 \leq M||\cdot||_1.$$

Lemma A.4 *If* V *is finite dimensional then any two norms are equivalent.*

Definition A.5 *The bilinear mapping* $\langle\cdot,\cdot\rangle : V \times V \to \mathbb{R}$ *is called inner product (or scalar product), if it satisfies the following three conditions:*

1. $\langle v, w\rangle = \langle w, v\rangle$, $\forall v, w \in V$ *(symmetry),*
2. $\langle v, v\rangle \geq 0$, $\forall v \in V$ *(positivity),*

3. $\langle v, v\rangle = 0 \leftrightarrow v = 0$.

Remark A.6 *Let* $\langle\cdot,\cdot\rangle$ *be an inner product, then* $\|v\|_V := \langle v, v\rangle^{1/2}\ \forall v \in V$ *defines a norm on* V.

Lemma A.7 (Cauchy-Schwartz inequality) $\forall v, w \in V$*:* $|\langle v, w\rangle| \leq \|v\|_V\|w\|_V$.

Definition A.8 *A Hilbert space is an inner product space that is complete with the norm defined by the inner product.*

Let us recall Theorem (1.10) from page 4.

Theorem 1.10 (Riesz Representation Theorem) *Let* H *be a Hilbert space. For all bounded linear functionals* $L : H \to \mathbb{R}$ *there exists a unique* $u \in H$ *such that* $L(v) = \langle v, u\rangle$ *for all* $v \in H$*, where* $\langle\cdot,\cdot\rangle$ *is an inner product on* H.

In the following $\alpha = (\alpha_1, \dots, \alpha_d)$ (where α_i is a non-negative integer $\forall i = 1, \dots, d$) will denote a multi-index, and for a function with d variable $\partial^\alpha v := \partial_1^{\alpha_1} \dots \partial_d^{\alpha_d} v$. The absolute value of α is defined as $|\alpha| := \alpha_1 + \dots + \alpha_d$.

Throughout this thesis we have used the following function spaces ($\Omega \subset \mathbb{R}^d$ in every case).

- $L^p(\Omega) := \{v : \Omega \to \mathbb{R} : \int_\Omega |v|^p < \infty\}, 1 \leq p \leq \infty$.
- $L^\infty(\Omega) := \{v : \Omega \to \mathbb{R} : \inf\{\sup_{N\subset\Omega, \mathrm{meas}(N)=0} |v|\} < \infty\}, 1 \leq p \leq \infty$.
- $W^{m,p}(\Omega) := \{v : \Omega \to \mathbb{R} : (\partial^\alpha v) \in L^p(\Omega), \forall\alpha : |\alpha| \leq m\}$
- $H^i(\Omega) := W^{i,2}(\Omega)$
- $H_0^1(\Omega) := \{v \in H^1(\Omega) : v|_{\partial\Omega} = 0\}$

The fractional Hilbert space $H^{1/2}(\partial\Omega)$ was used when we worked with non-homogeneous Dirichlet boundary condition. This space can be characterized as follows.

Definition A.9

$$H^{1/2}(\partial\Omega) := \left\{ u \in L^2(\partial\Omega); \frac{u(x) - u(y)}{|x - y|^{\frac{1+d}{2}}} \in L^2(\partial\Omega \times \partial\Omega) \right\}.$$

Remark A.10 $H^{1/2}(\partial\Omega)$ *can be defined using trace operators:* $H^{1/2}(\partial\Omega) := \{u \in L^2(\partial\Omega) : u = U|_{\partial\Omega}$ *(in trace sence) for some* $U \in H^1(\Omega)\}$*. For more details see [56, p.58].*

In Chapter 2 especially in Proposition 2.4 we had to deal with the $H^{-1}(\Omega)$ norm. The corresponding space can be introduced via duality.

Definition A.11 *Let X be a real Banach space. A bounded linear operator $u^* : X \to \mathbb{R}$ is called bounded linear functional on X. The set of bounded linear functionals on X is denoted by X^* and it is called the dual space of X. If $u \in X$, $u^* \in X^*$ we use the notation $\langle u^*, u\rangle$ to denote the real number $u^*(u)$. We can define a norm on X^* with $||u^*||_{X^*} := \sup\{\langle u^*, u\rangle : ||u||_X = 1\}$.*

Definition A.12 *We denote by $H^{-1}(\Omega)$ the dual space of $H_0^1(\Omega)$. It is equipped with the following norm:*

$$||f||_{-1} := ||f||_{H^{-1}(\Omega)} = \sup\left\{\langle f, u\rangle : u \in H_0^1(\Omega), ||u||_{H_0^1(\Omega)} = 1\right\}.$$

A.2 Proof of coercivity and boundedness

In the first Chapter we skipped all the proofs about the properties of the bilinear and the linear forms. In this Section we will prove coercivity and boundedness for the readers' convenience in the simplest case. Suppose that we have the linear reaction term and pure homogeneous Dirichlet boundary condition. In this case the linear and bilinear forms are

- $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$, $a(u, v) = \int_\Omega \mathcal{K}\nabla u \cdot \nabla v + \int_\Omega \mu uv$,
- $L : H_0^1(\Omega) \to \mathbb{R}$, $L(v) = \int_\Omega fv$.

We have to show the followings

- $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is **bounded** (in the $H_0^1(\Omega)$ norm): there exist a positive constant C_b: $|a(u, v)| \leq C_b||u||_{H_0^1}||v||_{H_0^1(\Omega)}$,
- $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is **coercive** (in the $H_0^1(\Omega)$ norm): there exist a positive constant C_c: $a(u, u) \geq C_c||u||^2_{H_0^1(\Omega)}$,
- $L : H_0^1(\Omega) \to \mathbb{R}$ is **bounded** (in the $H_0^1(\Omega)$ norm): there exist a positive constant $C_{_{L_b}}$: $L(v) \leq C_{L_b}||v||_{H_0^1(\Omega)}$.

In the proofs we will use Cauchy-Schwartz inequality in the $L^2(\Omega)$ inner product

$$\left|\int_\Omega vw\right| = |\langle v, w\rangle| \leq ||v||_0||w||_0 = \sqrt{\int_\Omega v^2}\sqrt{\int_\Omega w^2}.$$

We will also use the fact, that $\mathcal{K}$ in uniformly positive definite, see Definition 1.1, i.e.: there exist $\lambda, \Lambda > 0$ such that $\lambda||\xi||_E^2 \leq \mathcal{K}(x)\xi \cdot \xi \leq \Lambda||\xi||_E^2$, for all $x \in \Omega$, $\xi \in \mathbb{R}^d$, where $||\cdot||_E$ denotes the Euclidean norm in $\mathbb{R}^d$. This yields

$$|\langle\mathcal{K}\nabla u, \nabla v\rangle| \leq ||\mathcal{K}\nabla u||_E||\nabla v||_E \leq \Lambda||\nabla u||_E||\nabla v||_E.$$

Finally we recall the Poincare-Friedrichs-Sztyeklov inequality (see Theorem 1.8) stating: there exists a $C_{\text{PFS}} > 0$ constant, such that for all $u \in H_0^1(\Omega)$

$$||u||_0 \leq ||u||_1 \leq C_{\text{PFS}}||u||_{H_0^1(\Omega)} = C_{\text{PFS}}||\nabla u||_0.$$

It is important to note that the $L^2(\Omega)$ norm of the derivative can be interpreted using the Euclidean norm:

$$||u||^2_{H_0^1(\Omega)} = \int_\Omega ||\nabla u||_E^2.$$

Let us start with the easiest one.

$$a(u,u) = \int_\Omega \mathcal{K}\nabla u \cdot \nabla u + \int_\Omega \mu uu \geq \int_\Omega \mathcal{K}\nabla u \cdot \nabla u \geq \lambda \int_\Omega ||\nabla u||_E^2 \geq \lambda ||u||^2_{H_0^1(\Omega)}$$

Therefore $a(\cdot,\cdot)$ is coercive. The proof of its boundedness is a bit more technical.

$$\begin{aligned}
|a(u,v)| =& \left| \int_\Omega \mathcal{K}\nabla u \cdot \nabla v + \int_\Omega \mu uv \right| \\
\leq& \left| \int_\Omega \mathcal{K}\nabla u \cdot \nabla v \right| + \left| \int_\Omega \mu uv \right| \\
\leq& \int_\Omega |\mathcal{K}\nabla u \cdot \nabla v| + \int_\Omega |\mu uv| \\
\leq& \max\{\Lambda, \sup \mu\} \left(\int_\Omega ||\nabla u||_E \, ||\nabla v||_E + \int_\Omega |u||v| \right) \\
\leq& \max\{\Lambda, \sup \mu\} \left(\sqrt{\int_\Omega ||\nabla u||_E^2} \sqrt{\int_\Omega ||\nabla v||_E^2} + \sqrt{\int_\Omega |u|^2} \sqrt{\int_\Omega |v|^2} \right) \\
=& \max\{\Lambda, \sup \mu\} \left(||u||_{H_0^1(\Omega)} ||v||_{H_0^1(\Omega)} + ||u||_0 ||v||_0 \right) \\
\leq& \max\{\Lambda, \sup \mu\} \left(||u||_{H_0^1(\Omega)} ||v||_{H_0^1(\Omega)} + C_{PFS}^2 ||u||_{H_0^1(\Omega)} ||v||_{H_0^1(\Omega)} \right) \\
=& C ||u||_{H_0^1(\Omega)} ||v||_{H_0^1(\Omega)}
\end{aligned}$$

Finally the proof of the boundedness of the linear form.

$$\begin{aligned}
|L(v)| =& \left| \int_\Omega fv \right| \leq \sqrt{\int_\Omega f^2} \sqrt{\int_\Omega v^2} = ||f||_0 ||v||_0 \\
\leq& ||f||_0 ||v||_1 \leq C_{PFS} ||f||_0 ||v||_{H_0^1(\Omega)}
\end{aligned}$$

A.3 M-matrices

The M-matrix theory provides a powerful tool to prove that the inverse of a matrix is nonnegative. This section is based on [11, Ch.6] with small changes.

Definition A.13 *We call a real matrix Z-matrix if its off-diagonal entries are nonpositive.*

Definition A.14 *We call a real matrix M-matrix if it can be represented as* $s\mathbf{I} - \mathbf{B}$*, where* $\mathbf{I}$ *is the identity matrix and* $\mathbf{B} \leq \mathbf{0}$*, moreover* $s \geq \varrho(\mathbf{B})$*, where* ϱ *denotes the spectral radius of a matrix.*

It is obvious that an M-matrix is a Z-matrix, too.

Theorem A.15 *[11, Ch.6, Th.2.3] We assume that the matrix* $\mathbf{A}$ *is a Z-matrix. Then the following statements are equivalent.*

1. $\mathbf{A}$ *is a nonsingular M-matrix.*
2. *There exists* $\mathbf{u} > \mathbf{0}$ *with* $\mathbf{A}\mathbf{u} > \mathbf{0}$.
3. *There exists* $\mathbf{A}^{-1}$ *and* $\mathbf{A}^{-1} \geq \mathbf{0}$.

A.4 Polynomial approximation in Hilbert spaces

A.4.1 The continuous case

Throughout the convergence analysis of continuous and discontinuous Galerkin methods we have seen that one of the key ingredients is the approximation of a given function using (piecewise) polynomials of degree p. In this section we will collect the most important theorems and we will give references for the proofs.

Let us denote by $u_{\mathcal{I}_p}$ the interpolant of $u \in H^{l+1}(\Omega)$ $(1 \leq l \leq p)$ that can be calculated using Lagrange elements of degree p for the interpolation. Then for all $u \in H^{l+1}(\Omega)$ $(1 \leq l \leq p)$ we have

$$||u - u_{\mathcal{I}_p}||_0 + h|u - u_{\mathcal{I}_p}|_1 \leq ch^{l+1}|u|_{l+1}. \tag{A.1}$$

For the proof see i.e. [25, Sect. 1.5.1].

If we are using the $H_0^1(\Omega)$ norm we have

$$||u - u_{\mathcal{I}_p}||_{H_0^1(\Omega)} = |u - u_{\mathcal{I}_p}|_1 \leq ch^l|u|_{l+1}.$$

Taking l to its maximal possible value ($l = p$) we receive the approximation result we used in Section 1.3.1

$$||u - u_{\mathcal{I}_p}||_{H_0^1(\Omega)} \leq ch^p|u|_{p+1}.$$

Taking $l = 1$ we end up with the approximation result we used in Section 1.3.2.

$$|u - u_{\mathcal{I}_p}|_1 = ||u - u_{\mathcal{I}_p}||_{H_0^1(\Omega)} \leq ch|u|_2.$$

A.4.2 The discontinuous case

For the discontinuous problem let us recall the $|||\cdot|||_{*,DG}$ norm, see (1.16)-(1.17)

$$|||v|||_{*,DG}^2 = ||\nabla_h v||_0^2 + \sum_{e\in\Gamma_0\cup\Gamma_N} \frac{1}{|e|}|| \, [\![v]\!] \, ||_{0,e}^2 + \sum_{E\in\mathcal{T}_h} h_E||\nabla v|_E \cdot \nu_E||_{0,\partial E}^2.$$

Definition A.16 *The projection $\pi_h : L^2(\Omega) \to \mathcal{P}_d^p(\mathcal{T}_h)$ (broken polynomial space over $\mathcal{T}_h$, see Definition 1.28) is called $L^2(\Omega)$-orthogonal projection, if for all $v \in L^2(\Omega)$, $\pi_h v \in \mathcal{P}_d^p(\mathcal{T}_h)$ with*

$$\langle \pi_h v, y_h \rangle_{L^2(\Omega)} = \langle v, y_h \rangle_{L^2(\Omega)} \qquad \forall y_h \in \mathcal{P}_d^p(\mathcal{T}_h).$$

It is important to note that the restriction of $\pi_h v$ to a given mesh element $E \in \mathcal{T}_h$ can be computed independently from other mesh elements.

For the approximation result we need three estimations. They can be derived locally, on a proper mesh - for the readers' convenience we will skip the technical details on the meshes (see i.e. [21, Sect. 1.4.4-1.4.5] for these details).

Lemma A.17 (Lemma 1.58 [21]) *Let π_h be the L^2-orthogonal projection onto $\mathcal{P}_d^p(\mathcal{T}_h)$. Then for all $s \in \{0, \dots, p+1\}$ and for all $v \in H^s(E)$, there holds*

$$|v - \pi_h v|_{m,E} \le C_1 h_E^{s-m} |v|_{s,E} \qquad \forall m \in \{0, \dots, s\},$$

where C_1 is independent of both E and h.

Lemma A.18 (Lemma 1.59 [21]) *Suppose that the hypotheses of Lemma A.17 are valid and $s \ge 1$. Then for all $E \in \mathcal{T}_h$, for all e edge of E, there holds*

$$||v - \pi_h v||_{0,e} \le C_2 h_E^{s-1/2} |v|_{s,E},$$

and if $s \ge 2$ then

$$||\nabla \left(v - \pi_h v\right)|_E \cdot \nu_E||_{0,e} \le C_3 h_E^{s-3/2} |v|_{s,E},$$

where C_2 and C_3 are independent of both E and h.

Using Lemma A.17 with $m = 1$ and supposing that $v \in H^{p+1}(\mathcal{T}_h)$ $(s = p+1)$ we get

$$||\nabla_h \left(v - \pi_h v\right)||_{0,E}^2 = |v - \pi_h v|_{1,E}^2 = C_1^2 h_E^{2p} |v|_{p+1,E}^2.$$

Lemma A.18 with $s = p+1$ results in (we suppose that there exists C_h such that $C_h \le h/h_E$ for all faces of every elements)

$$\frac{1}{|e|}||v - \pi_h v||_{0,e}^2 \le C_2^2 \frac{1}{h} h_E^{2p+1/2} |v|_{p+1,E} = \widetilde{C_2^2} h_E^{2p} |v|_{p+1,E}^2,$$

and for the gradient

$$h_E||\nabla \left(v - \pi_h v\right)|_E \cdot \nu_E||_{0,e}^2 \le C_3 h_E h_E^{2p-1} |v|_{p+1,E} = C_3^2 h_E^{2p} |v|_{p+1,E}^2.$$

Using these estimations we can handle the three terms of $|||(v - \pi_h v)|||^2_{*,DG}$

$$\|\nabla_h (v - \pi_h v)\|_0^2 \leq \sum_{E \in \mathcal{T}_h} C_1^2 h_E^{2p} |v|^2_{p+1,E} \leq C_1^2 h_{\max}^{2p} |v|^2_{p+1,\mathcal{T}_h},$$

$$\sum_{e \in \partial E} \frac{1}{|e|} \| [\![(v - \pi_h v)]\!] \|^2_{0,e} \leq 2 \sum_{E \in \mathcal{T}_h} \sum_{e \in \partial E} \frac{1}{|e|} \| (v - \pi_h v) \|^2_{0,e} \leq 2\widetilde{C_2^2} h_{\max}^{2p} |v|^2_{p+1,\mathcal{T}_h},$$

$$\sum_{E \in \mathcal{T}_h} h_E \|\nabla (v - \pi_h v)|_E \cdot \nu_E\|^2_{0,\partial E} \leq \sum_{E \in \mathcal{T}_h} C_3^2 h_E^{2p} |v|^2_{p+1,E} \leq C_3^2 h_{\max}^{2p} |v|^2_{p+1,\mathcal{T}_h}.$$

Summing them up we have that there exists C such that $\forall v \in H^{p+1}(\mathcal{T}_h)$

$$|||v - \pi_h v|||_{*,DG} = C h_{\max}^p |v|_{p+1,\mathcal{T}_h}.$$

Finally using the facts that $H^{p+1}(\Omega) \subset H^{p+1}(\mathcal{T}_h)$ and $|v|_{p+1,\mathcal{T}_h} = |v|_{p+1}$ holds for all $v \in H^{p+1}(\Omega)$ we have that there exists C such that $\forall v \in H^{p+1}(\Omega)$

$$|||v - \pi_h v|||_{*,DG} = C h_{\max}^p |v|_{p+1}.$$

Summary

In the thesis the finite element method as one of the most frequently used tool for solving partial differential equations numerically was studied. We have seen that its different versions are based on the idea of solving the (approximate) weak form in a finite dimensional subspace of certain originally infinite dimensional functional space. The effectiveness of the method can be considerably increased by using adaptivity. It raises, however, the delicate question of finding an accurate a-posteriori error estimation.

Two different error estimation procedures have been dealt with. One of them is the implicit a-posteriori one, in the course of which we have to solve a local Neumann problem in every subdomain. In Chapter 2 we have shown a new construction of the Neumann boundary condition that enabled us to prove a bound on the norm of the difference between the computational error $e_{h,p}$ and the estimated error $\hat{e}_{h,p}$. We have shown numerical results justifying the usage of the method.

Another way of estimating the error is to use a reference solution, which is computationally relatively expensive. However, the method can be used for a wide range of partial differential equations. In Chapter 3 we have pointed out a drawback of the method. We were able to give an example that forces the algorithm to terminate, because the estimated error is zero, although, in fact it can be arbitrary. We have seen, that this difficulty can be overcome by using the residual based error estimator as a back-up estimator.

The simplest adaptive method uses first order elements and where it is necessary it refines the mesh only. In this case the preservation of maximum principles is almost a natural requirement (if the continuous problem possesses them). Chapter 4 has described first the differences between the weak and strong maximum principles, then numerical examples have been shown to illustrate that some of the maximum principles are preserved by the discrete solution, while others are not.

Finally, in Chapter 5 we have presented the maximum principle analysis of a class of one dimensional boundary value problem discretised by the interior penalty discontinuous Galerkin method. We have derived sufficient mesh conditions that guarantee discrete maximum principle preservation. We have justified, by numerical examples, that these conditions are sharp in some sense.

Magyar nyelvű összefoglaló

Jelen dolgozatban a parciális differenciálegyenletek egyik legelterjedtebb numerikus megoldási módszerét, a végeselem módszert tanulmányoztuk. Ennek különféle változatait láttuk, melyek mindegyike ugyanazon az ötleten alapszik: a gyenge alak (vagy annak egy közelítésének) megoldását keressük az eredeti végtelen dimenziós függvénytér véges dimenziós alterében. A módszer hatékonysága nagyban javítható az adaptivitással, ugyanakkor ennek használata további nehézségeket vet fel, példának okáért ilyen a hibabecslés kérdése.

Két különböző hibabecslést jártunk körül a dolgozatban. Az egyik az implicit utólagos hibabecslés volt, melynek során lokális Neumann-feladatokat oldunk meg résztartományonként. A 2. Fejezetben a Neumann-peremfeltétel újfajta konstrukcióját mutattuk meg, melynek segítségével a módszerből származó $e_{h,p}$ hiba és a becsült $\hat{e}_{h,p}$ hiba különbségének normáját tudtuk becsülni. A fejezet végén numerikus példákkal illusztráltuk a módszer hatékonyságát.

Az utólagos hibabecslés egy másik módja az ún. referencia megoldás módszere, mely viszonylag magas számításigényű, de előnye, hogy a feladatok széles skáláján alkalmazható. A 3. Fejezetben a módszer egy gyenge pontjára világítottunk rá. Olyan példák elkészítésének módszerét mutattuk meg, melyekkel leállásra kényszeríthetjük az algoritmust, mert nullának érzékeli a hibát, holott az tetszőlegesen nagy lehet. Azt is láthattuk, hogy a módszer hibája kiküszöbölhető pl. a reziduális-alapú hibabecslő tartalék hibabecslőként való beépítésével.

A legegyszerűbb adaptív módszer elsőfokú polinomokat használ és csak a rácsot finomítja. Ilyen esetekben jogos elvárás lehet, hogy a diszkrét módszer is megőrizze a maximum elvet (amennyiben a folytonos is megőrizte). A 4. Fejezetben megismerhettük a különféle gyenge és erős maximum elveket. Láthattunk olyan numerikus példákat, melyek során bizonyos maximum elvek megőrződtek, míg mások nem.

Végül a 5. Fejezetben a nemfolytonos végeselem módszerek közül a belső büntetésen alapuló módszert vizsgáltuk a maximum elv szempontjából egydimenziós peremérték problémák megoldásához. Elégséges feltételeket adtunk a rácsra, melyek garantálják a maximum elv megőrzést. Numerikus példákon keresztül megmutattuk, hogy ezek a feltételek, bár csak elégségesek, valamilyen értelemben mégis élesek.

Bibliography

[1] M. Ainsworth. The influence and selection of subspaces for a posteriori error estimators. *Numer. Math.*, 73(4):399–418, 1996.

[2] M. Ainsworth and A. Craig. A posteriori error estimators in the finite element method. *Numer. Math.*, 60(4):429–463, 1992.

[3] M. Ainsworth and J. T. Oden. *A posteriori error estimation in finite element analysis*. Pure and Applied Mathematics (New York). Wiley-Interscience [John Wiley & Sons], New York, 2000.

[4] M. Ainsworth and R. Rankin. Fully computable bounds for the error in nonconforming finite element approximations of arbitrary order on triangular elements. *SIAM J. Numer. Anal.*, 46(6):3207–3232, 2008.

[5] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini. Unified analysis of discontinuous Galerkin methods for elliptic problems. *SIAM J. Numer. Anal.*, 39(5):1749–1779, 2001/02.

[6] B. Ayuso de Dios, F. Brezzi, O. Havle, and L. D. Marini. L^2-estimates for the DG IIPG-0 scheme. *Numer. Methods Partial Differential Equations*, 28(5):1440–1465, 2012.

[7] I. Babuška and W. C. Rheinboldt. Error estimates for adaptive finite element computations. *SIAM J. Numer. Anal.*, 15(4):736–754, 1978.

[8] R. E. Bank and A. Weiser. Some a posteriori error estimators for elliptic partial differential equations. *Math. Comp.*, 44(170):283–301, 1985.

[9] R. E. Bank and J. Xu. Asymptotically exact a posteriori error estimators. I. Grids with superconvergence. *SIAM J. Numer. Anal.*, 41(6):2294–2312 (electronic), 2003.

[10] R. E. Bank and J. Xu. Asymptotically exact a posteriori error estimators. II. General unstructured grids. *SIAM J. Numer. Anal.*, 41(6):2313–2332 (electronic), 2003.

[11] A. Berman and R. J. Plemmons. *Nonnegative matrices in the mathematical sciences*. Academic Press [Harcourt Brace Jovanovich Publishers], New York, 1979. Computer Science and Applied Mathematics.

[12] S. C. Brenner and L. R. Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer-Verlag, New York, second edition, 2002.

[13] C. Carstensen. Some remarks on the history and future of averaging techniques in a posteriori finite element error analysis. *ZAMM Z. Angew. Math. Mech.*, 84(1):3–21, 2004.

[14] C. Carstensen. A unifying theory of a posteriori finite element error control. *Numer. Math.*, 100(4):617–637, 2005.

[15] C. Carstensen and S. Bartels. Each averaging technique yields reliable a posteriori error control in FEM on unstructured grids. I. Low order conforming, nonconforming, and mixed FEM. *Math. Comp.*, 71(239):945–969 (electronic), 2002.

[16] C. Carstensen, A. Orlando, and J. Valdman. A convergent adaptive finite element method for the primal problem of elastoplasticity. *Internat. J. Numer. Methods Engrg.*, 67(13):1851–1887, 2006.

[17] P. G. Ciarlet. Discrete maximum principle for finite-difference operators. *Aequationes Math.*, 4:338–352, 1970.

[18] L. Demkowicz. *Computing with hp-adaptive finite elements. Vol. 1.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2007. One and two dimensional elliptic and Maxwell problems, With 1 CD-ROM (UNIX).

[19] L. Demkowicz. *Computing with hp-adaptive finite elements. Vol. 2.* Chapman & Hall/CRC Applied Mathematics and Nonlinear Science Series. Chapman & Hall/CRC, Boca Raton, FL, 2007. Frontiers: Three dimensional elliptic and Maxwell problems with applications.

[20] L. Demkowicz, W. Rachowicz, and P. Devloo. A fully automatic hp-adaptivity. In *Proceedings of the Fifth International Conference on Spectral and High Order Methods (ICOSAHOM-01) (Uppsala)*, volume 17, pages 117–142, 2002.

[21] D. A. Di Pietro and A. Ern. *Mathematical aspects of discontinuous Galerkin methods*, volume 69 of *Mathématiques & Applications (Berlin) [Mathematics & Applications]*. Springer, Heidelberg, 2012.

[22] A. Drăgănescu, T. F. Dupont, and L. R. Scott. Failure of the discrete maximum principle for an elliptic finite element problem. *Math. Comp.*, 74(249):1–23 (electronic), 2005.

[23] L. Dubcová, P. Šolín, J. Červený, and P. Kus. Space and time adaptive two-mesh hp-fem for transient microwave heating problems. *Electromagnetics*, 30(1-2):23–40, 2009.

[24] L. Dubcová, P. Šolín, G. Hansen, and H. Park. Comparison of multimesh hp-fem to interpolation and projection methods for spatial coupling of thermal and neutron diffusion calculations. *J. Comput. Phys.*, 77:1182–1197, 2011.

[25] A. Ern and J.-L. Guermond. *Theory and practice of finite elements*, volume 159 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.

[26] L. C. Evans. *Partial differential equations*, volume 19 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, second edition, 2010.

[27] I. Faragó. *Numerical treatment of linear parabolic problems*. PhD thesis, Doctoral Dissertation, Eötvös Loránd University, Budapest, 2008.

[28] M. S. Gockenbach. *Understanding and implementing the finite element method*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[29] W. Gui and I. Babuška. The $h,\ p$ and h-p versions of the finite element method in 1 dimension. I. The error analysis of the p-version. *Numer. Math.*, 49(6):577–612, 1986.

[30] W. Gui and I. Babuška. The $h,\ p$ and h-p versions of the finite element method in 1 dimension. II. The error analysis of the h- and h-p versions. *Numer. Math.*, 49(6):613–657, 1986.

[31] W. Gui and I. Babuška. The $h,\ p$ and h-p versions of the finite element method in 1 dimension. III. The adaptive h-p version. *Numer. Math.*, 49(6):659–683, 1986.

[32] J. Guzmán and B. Rivière. Sub-optimal convergence of non-symmetric discontinuous Galerkin methods for odd polynomial approximations. *J. Sci. Comput.*, 40(1-3):273–280, 2009.

[33] A. Hannukainen, S. Korotov, and M. Křížek. Nodal $O(h^4)$-superconvergence in 3D by averaging piecewise linear, bilinear, and trilinear FE approximations. *J. Comput. Math.*, 28(1):1–10, 2010.

[34] A. Hannukainen, S. Korotov, and T. Vejchodskỳ. On weakening conditions for discrete maximum principles for linear finite element schemes. *Numerical Analysis and Its Applications*, pages 297–304, 2009.

[35] D. Harutyunyan, F. Izsák, J. J. W. van der Vegt, and M. A. Botchev. Adaptive finite element techniques for the Maxwell equations using implicit a posteriori error estimates. *Comput. Methods Appl. Mech. Engrg.*, 197(17-18):1620–1638, 2008.

[36] I. Hlaváček and M. Křížek. Optimal interior and local error estimates of a recovered gradient of linear elements on nonuniform triangulations. *J. Comput. Math.*, 14:345–362, 1996.

[37] W. Höhn and H.-D. Mittelmann. Some remarks on the discrete maximum-principle for finite elements of higher order. *Computing*, 27(2):145–154, 1981.

[38] T. L. Horváth. A note on reference solution based hp-adaptive methods. unpublished.

[39] T. L. Horváth and F. Izsák. Implicit a posteriori error estimation using patch recovery techniques. *Cent. Eur. J. Math.*, 10(1):55–72, 2012.

[40] T. L. Horváth and M. E. Mincsovics. Discrete maximum principle for interior penalty discontinuous Galerkin methods. *Cent. Eur. J. Math.*, 11(4):664–679, 2013.

[41] P. Houston, B. Senior, and E. Süli. Sobolev regularity estimation for hp-adaptive finite element methods. In *Numerical mathematics and advanced applications*, pages 631–656. Springer Italia, Milan, 2003.

[42] P. Houston, E. Süli, and T. P. Wihler. A posteriori error analysis of hp-version discontinuous Galerkin finite-element methods for second-order quasi-linear elliptic PDEs. *IMA J. Numer. Anal.*, 28(2):245–273, 2008.

[43] Y. Huang and J. Xu. Superconvergence of quadratic finite elements on mildly structured grids. *Math. Comp.*, 77(263):1253–1268, 2008.

[44] K. Ishihara. Strong and weak discrete maximum principles for matrices associated with elliptic problems. *Linear Algebra Appl.*, 88/89:431–448, 1987.

[45] F. Izsák, D. Harutyunyan, and J. J. W. van der Vegt. Implicit a posteriori error estimates for the Maxwell equations. *Math. Comp.*, 77(263):1355–1386, 2008.

[46] H. Jin and S. Prudhomme. A posteriori error estimation of steady-state finite element solutions of the navier-stokes equations by a subdomain residual method. *Comput. Methods Appl. Mech. Engrg.*, 159:19–48, 1998.

[47] J. Karátson and S. Korotov. Sharp upper global a posteriori error estimates for nonlinear elliptic variational problems. *Appl. Math.*, 54(4):297–336, 2009.

[48] P. Knabner and L. Angermann. *Numerical methods for elliptic and parabolic partial differential equations*, volume 44 of *Texts in Applied Mathematics*. Springer-Verlag, New York, 2003.

[49] S. Korotov, P. Neittaanmäki, and S. Repin. A posteriori error estimation of goal-oriented quantities by the superconvergence patch recovery. *J. Numer. Math.*, 11(1):33–59, 2003.

[50] M. Křížek and P. Neittaanmäki. Superconvergence phenomenon in the finite element method arising from averaging gradients. *Numer. Math.*, 45(1):105–116, 1984.

[51] P. Ladevèze and D. Leguillon. Error estimate procedure in the finite element method and applications. *SIAM J. Numer. Anal.*, 20(3):485–509, 1983.

[52] W. McLean. *Strongly elliptic systems and boundary integral equations*. Cambridge University Press, Cambridge, 2000.

[53] M. E. Mincsovics and T. L. Horváth. On the differences of the discrete weak and strong maximum principles for elliptic operators. In *Large-scale scientific computing*, volume 7116 of *Lecture Notes in Comput. Sci.*, pages 614–621. Springer, Heidelberg, 2012.

[54] W. F. Mitchell. Phaml user's guide, 2006.

[55] W. F. Mitchell and M. A. McClain. A survey of hp-adaptive strategies for elliptic partial differential equations. *Recent Advances in Computational and Applied Mathematics*, pages 227–258, 2011.

[56] P. Monk. *Finite element methods for Maxwell's equations*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2003.

[57] P. Morin, R. H. Nochetto, and K. G. Siebert. Data oscillation and convergence of adaptive FEM. *SIAM J. Numer. Anal.*, 38(2):466–488 (electronic), 2000.

[58] P. Morin, R. H. Nochetto, and K. G. Siebert. Convergence of adaptive finite element methods. *SIAM Rev.*, 44(4):631–658 (electronic) (2003), 2002. Revised reprint of "Data oscillation and convergence of adaptive FEM" [SIAM J. Numer. Anal. **3**8 (2000), no. 2, 466–488 (electronic); MR1770058 (2001g:65157)].

[59] P. Neittaanmäki and S. Repin. *Reliable methods for computer simulation*, volume 33 of *Studies in Mathematics and its Applications*. Elsevier Science B.V., Amsterdam, 2004. Error control and a posteriori estimates.

[60] S. Repin. *A posteriori estimates for partial differential equations*, volume 4 of *Radon Series on Computational and Applied Mathematics*. Walter de Gruyter GmbH & Co. KG, Berlin, 2008.

[61] B. Rivière. *Discontinuous Galerkin methods for solving elliptic and parabolic equations*, volume 35 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2008. Theory and implementation.

[62] V. Ruas Santos. On the strong maximum principle for some piecewise linear finite element approximate problems of nonpositive type. *J. Fac. Sci. Univ. Tokyo Sect. IA Math.*, 29(2):473–491, 1982.

[63] J. Schöberl. A posteriori error estimates for Maxwell equations. *Math. Comp.*, 77:633–649, 2008.

[64] C. Schwab. *p- and hp-finite element methods*. Numerical Mathematics and Scientific Computation. The Clarendon Press Oxford University Press, New York, 1998. Theory and applications in solid and fluid mechanics.

[65] P. Šolín, J. Červený, and I. Doležel. Arbitrary-level hanging nodes and automatic adaptivity in the hp-fem. *Math. Comput. Simulation*, 77:117–132, 2008.

[66] P. Šolín, L. Dubcová, J. Červený, and I. Doležel. Adaptive hp-fem with arbitrary-level hanging nodes for maxwell's equations. *Math. Comput. Simul.*, 17:127–155, 2009.

[67] P. Šolín, L. Dubcová, and I. Doležel. Adaptive hp-FEM with arbitrary-level hanging nodes for Maxwell's equations. *Adv. Appl. Math. Mech.*, 2(4):518–532, 2010.

[68] P. Šolín, K. Segeth, and I. Doležel. Space-time adaptive hp-FEM: methodology overview. In *Programs and algorithms of numerical mathematics 14*, pages 185–200. Acad. Sci. Czech Repub. Inst. Math., Prague, 2008.

[69] R. S. Varga. *Matrix iterative analysis*. Prentice-Hall Inc., Englewood Cliffs, N.J., 1962.

[70] R. S. Varga. On a discrete maximum principle. *SIAM Journal on Numerical Analysis*, 3(2):355–359, 1966.

[71] T. Vejchodskỳ. *Discrete Maximum Principles*. PhD thesis, Habilitation Thesis, Institute of Mathematics of the Academy of Sciences and Faculty of Mathematics and Physics, Charles University in Prague, 2011.

[72] T. Vejchodský and P. Šolín. Discrete maximum principle for higher-order finite elements in 1D. *Math. Comp.*, 76(260):1833–1846 (electronic), 2007.

[73] R. Verfürth. *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*. Advances in Numerical Mathematics. Wiley - Teubner, Chichester - Stuttgart, 1996.

[74] R. Verfürth. A posteriori error estimators for convection-diffusion equations. *Numer. Math.*, 80(4):641–663, 1998.

[75] M. Vohralík. A posteriori error estimates for lowest-order mixed finite element discretizations of convection-diffusion-reaction equations. *SIAM J. Numer. Anal.*, 45(4):1570–1599 (electronic), 2007.

[76] L. B. Wahlbin. *Superconvergence in Galerkin finite element methods*, volume 1605 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 1995.

[77] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. I. The recovery technique. *Internat. J. Numer. Methods Engrg.*, 33(7):1331–1364, 1992.

[78] O. C. Zienkiewicz and J. Z. Zhu. The superconvergent patch recovery and a posteriori error estimates. II. Error estimates and adaptivity. *Internat. J. Numer. Methods Engrg.*, 33(7):1365–1382, 1992.