On Solving Nonlinear Variational Inequalities by p-Version Finite Elements

Vom Fachbereich Mathematik
der Universität Hannover
zur Erlangung des Grades eines

DOKTORS DER NATURWISSENCHAFTEN

Dr. rer. nat.

genehmigte Dissertation
von

Dipl.-Math. Andreas Krebs
geboren am 30. April 1965 in Bückeburg

2004

Referent: Prof. Dr. E.P. Stephan, Universität Hannover
Korreferent: Prof. Dr. J. Gwinner, Universität der Bundeswehr München
Tag der Promotion: 1. Juli 2004

Zusammenfassung

hp-Finite-Elemente-Methoden (FEM) haben sich bei der Lösung von partiellen Differentialgleichungen (PDG) bewährt. Häufig liefern sie im Vergleich zur *h*-FEM höhere Konvergenzraten (s. Babuška und Guo, 1988). Ziel dieser Arbeit ist die Konstruktion, Analyse und Implementation einer *hp*-FEM zur numerischen Lösung von variationellen Ungleichungen, die zu *quasi-linearen elliptischen partiellen Differentialungleichungen* (PDU) *zweiter Ordnung* korrespondieren. Nichtlineare PDU spielen eine bedeutende Rolle in der Modellierung praktischer Probleme wie z.B. der *Mechanik elastischer und elasto-plastischer Körper* (s. Hlaváček, Haslinger, Nečas und Lovíšek, 1988) sowie der *Geometrie von Minimalflächen über ein Hindernis* (s. Kinderlehrer and Stampacchia, 1980).

Mit Hilfe der Variationsrechnung lassen sich PDU mathematisch als ein Minimierungsproblem auf der konvexen Teilmenge K eines Banachraumes V begreifen. Üblicherweise ist V ein Sobolewraum und K durch *Gleichungs- und Ungleichungsnebenbedingungen* (G&UB) definiert, die die Funktionen aus V erfüllen müssen.

In der Arbeit approximieren wir die Lösung eines Minimierungsproblems, indem wir das Minimum auf der diskreten Teilmenge K_p suchen. Hierbei stellt K_p die Teilmenge eines konformen p-Version-Finite-Elemente-Raumes V_p dar, die die Kontrolle der G&UB in geeigneten Punkten erlaubt. Die genannten Punkte sind auf dem Referenzquadrat $[-1,1]^2$ durch das Tensorprodukt der Gauß-Lobatto-Punkte gegeben. Die Bilder dieses Tensorprodukts auf die Rechtecke des Gitters im Sinne der FEM definieren die Kontrollpunkte. Für die Lösungen $u_p \in K_p$ des diskreten Minimierungsproblems wird die *Existenz*, die *Eindeutigkeit* und die *Konvergenz* gegen die Lösung u der PDU für $p \to \infty$ bzgl. der $\|\cdot\|_{H^1(\Omega)}$-Norm nachgewiesen sowie eine *A-priori-Abschätzung* für den Fehler $\|u-u_p\|_{H^1(\Omega)}$ angegeben.

Ferner wird für Dreieckgitter im Rekurs auf gewichtete Sobolew-Räume eine p-Diskretisierung vorgeschlagen, die eine Kontrolle der G&UB auf Dreiecken ermöglicht. Sie unterscheidet sich wesentlich von bekannten p-Diskretisierungen auf Dreiecken zur Behandlung *partieller Differentialgleichungen* mit der FEM. Numerische Experimente zeigen hohe Konvergenzraten sowohl auf Rechtecken als auch auf Dreiecken.

Die p-Diskretisierung auf quasi-uniformen Gittern wird so verallgemeinert, dass die G&UB auch auf Gittern mit hängenden Knoten und unterschiedlichen Polynomgraden auf den Rechtecken des Gitters kontrolliert werden können. Für die damit mögliche *adaptive hp-Verfeinerung* wird ein *dual gewichteter A-posteriori-Fehlerschätzer* angegeben.

Die computergestützte Berechnung der diskreten Minimalstelle $u_p \in K_p$ verlangt die Lösung eines *großskalierten nichtlinearen Minimierungsproblems mit* G&UB. Hierfür wird ein Löser angegeben, der das Minimum durch eine Kombination projizierter Gradientenschritte mit der Newtonmethode findet. Der Aufwand für die effiziente Lösung der dabei auftretenden linearen Probleme mit der konjugierten Gradientenmethode wird theoretisch und numerisch für verschiedene Präkonditionierer untersucht.

Bei der numerischen Berechnung von Minimalflächen über ein Hindernis erweist sich die p-Version gegenüber der h-Version als überlegen.

Schlagworte: Variationelle Ungleichungen, *hp*-Finite Elemente Methoden, a posteriori Fehlerschätzer, großskalierte Minimierung

Abstract

hp-finite element methods (FEM) have become a powerful tool in the treatment of second order quasi-linear elliptic partial differential equations (PDE). Frequently, their convergence rates are superior to *h*-FEM (Babuška and Guo, 1988). The objective of this dissertation is the design, analysis, and implementation of an *hp*-FEM for the treatment of *variational inequalities* which correspond to *second order quasi-linear elliptic partial differential inequalities* (PDI). Nonlinear PDI play an important role in the modeling of practical problems, for example in mechanics of elastic and elasto-plastic bodies (Hlaváček, Haslinger, Nečas, and Lovíšek, 1988) and in geometry of minimal surfaces (Kinderlehrer and Stampacchia, 1980).

Using the calculus of variations, PDI can be written mathematically as a *minimization problem on a closed convex subset K of a Banach space V*. Usually, V is a Sobolev space and K is defined by *equality and inequality constraints* (E&IC) which must be satisfied by the functions from V.

In this thesis, we approximate the solution of the minimum problem by searching the minimum on the discrete set K_p. Here, K_p is a subset of the conform p-FE space V_p which allows to control the E&IC at appropriate points of the domain. Namely, the points are given on the reference square $[-1,1]^2$ by the tensor product of Gauss-Lobatto points. The images of these points onto the quadrilaterals of the mesh define the control points. *Existence* and *uniqueness* of a minimum $u_p \in K_p$ can be proved. The *convergence* of u_p towards the minimum u on K with respect to $\|\cdot\|_{H^1(\Omega)}$ is shown and *an a priori bound* for the error $\|u-u_p\|_{H^1(\Omega)}$ is given.

Further, a p-discretization is suggested which allows to control the E&IC on *triangle meshes.* This discretization differs mutually from known p-discretizations for the treatment of PDE with FE on triangles. Numerical experiments yield high convergence rates on the square and on the triangle.

The p-discretization on quasi-uniform meshes is extended to quadrilateral meshes with hanging nodes and a non-uniform distribution of polygonal degrees such that the E&IC can be controlled again. This allows adaptive h- and p-refinements. A *dual-weighted residual error estimator* is introduced to drive these refinements.

The computation of the discrete minimum $u_p \in K_p$ demands to solve a *large-scale nonlinear convex minimum problem with* E&IC. This problem can be solved by a combination of the projected gradient method with Newton's method. Preconditioners for the efficient solution of the linear systems raised by Newton's method and their costs are discussed theoretically and numerically.

The p-version showed better results than the h-version when the minimal surface over an obstacle was computed numerically.

Key words. variational inequalities, *hp*-finite element methods, a posteriori error estimators, large-scale minimization.

Acknowledgements

It is my pleasure to express my sincere thanks to all who supported me during the time in which this thesis was written.

I am indebted to my supervisor, Prof. Dr. E. P. Stephan (University of Hannover, Germany) whom I thank at this point for his guidance over the last six years, mathematical and otherwise. He was an attentive listener to my ideas which sometimes were only vague in the very beginning. Moreover, Prof. Dr. Stephan was the claimant for the DFG project *Adaptive controls for the p- and hp-versions of the boundary element method with 2-level decompositions*, No. Ste 573/4-1, and for the DAAD-Project *Teaching and research partnership with the Departamento de Ingeniería Matemática of the Universidad de Concepción, Chile*, No. 412/HSPart, which founded this work theoretically. Thanks are also due to Dr. habil. M. Maischak for his constructive criticism concerning both, the mathematical foundation and the implementation. His software package `maiprogs` became a big resource of my programming experience and my programs.

Furthermore, I like to thank Prof. Dr. J. Hesthaven (Brown University, Providence, USA) for his encouragement and his helpful comments concerning the p-version on triangles. I would like to thank my co-referee Prof. Dr. J. Gwinner (Universität der Bundeswehr München, Germany) for his support and that he was ready to assess this thesis in a very short period of time.

I am also indebted to Prof. Dr. L. J. Cromme (Brandenburg University of Technology Cottbus, Germany) for his support, his interest, for the cheerful collaboration, and the stimulating atmosphere at the Chair of Numerical and Applied Mathematics. My special thanks go to Dr. F. Kaden for many fruitful discussions, to Frau S. Büttner for her efficient administration of the computing facilities, and to Frau M. Hein for reminding me of the administrative things which I tended to forget when I was implementing the numerical experiments.

Finally, my heartfelt thanks go to all who joined me in walking to the pyramids at the landscape park in Branitz and in visiting the Cottbus Staatstheater.

Cottbus, November 2004 Andreas Krebs

Contents

List of Tables

List of Figures

List of Algorithms

Glossary of Notations

Conventions

$i, j, k, l, m, n, p, N \in \mathbb{N}$ nonnegative integers

$\alpha, \beta, \ldots,$ small Greek letters scalars

$x, y, \ldots,$ small roman letters column vectors (except for integers)

$A, B, \ldots,$ capital roman letters matrices or operators

Common Notations

$\cong$ see p. 47

$\mathbb{N}$ set of natural numbers $1, 2, \ldots$

$\mathbb{R}, \mathbb{C}, \mathbb{Q}, \mathbb{Z},$ set of real, complex, rational, integer numbers

$\mathbb{R}_{>0}, \mathbb{R}_{\geq 0}$ positive and nonnegative real numbers, respectively

$\mathbb{R}^d$ Euclidean d-dimensional space

$\mathbb{R}^d_{\geq z}$ $\{x \in \mathbb{R}^d \,|\, x_i \geq z_i\}$

$x = (x_1, \ldots, x_d)^T$, $y = (y_1, \ldots, y_d)^T$, etc. column vector in $\mathbb{R}^d$

x^T transpose of x

$x^T y = x_1 y_1 + \cdots + x_d y_d = x \cdot y = (x, y)$ scalar product in $\mathbb{R}^d$

$|x| = (\sum_1^d x_i^2)^{1/2} = (x^T x)^{1/2}]$ length of $x \in \mathbb{R}^d$

$e_i = (0, \ldots, 0, 1, 0, \ldots, 0)^T$ unit vector with 1 in the coordinate i

$I = I_{d,d} \in \mathbb{R}^{d,d}$ identity matrix

$B_r(x)$ the open ball of radius r and center $x \in \mathbb{R}^d$

Ω an open, generally bounded and connected, subset of $\mathbb{R}^d$

$\partial\Omega$ the boundary of Ω

$\overline{\Omega} = \Omega \cup \partial\Omega$ the closure of Ω

$\operatorname{int} U = \overset{\circ}{U}$ the interior of U

$\operatorname{supp} u$	the support of the function u, which is the smallest compact set outside of which $u = 0$
$\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_V$	a duality pairing between a real Banach space V and its dual V'; $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{R}$
$\| \cdot \| = \| \cdot \|_V$	norm on a real Banach space V
$v_n \to v$	convergence in norm
$v_n \rightharpoonup v$	weak convergence

Function Spaces

$C(\Omega)$, $C(\overline{\Omega})$	the functions continuous in Ω and $\overline{\Omega}$, respectively
$C^k(\overline{\Omega})$, $C^{k,\lambda}(\overline{\Omega})$	the functions which are k times continuously differentiable in some neighborhood of $\overline{\Omega}$ and whose kth derivatives satisfy additionally a Hölder condition with exponent λ, respectively
$L^s(\Omega)$	the Lebesgue measurable functions u of Ω for which $\|u\|_{L^s(\Omega)} = (\int_\Omega \|u\|^s \, \mathrm{d}x)^{1/s}$, $1 \le s < \infty$
$L^\infty(\Omega)$	the Lebesgue measurable functions u of Ω which are essentially bounded, $\|u\|_\infty = \inf\{\|M : \|u\| \le M \text{ a.e. in } \Omega\}$
$H^m(\Omega)$, $H^1_{g_D}(\Omega)$, $H^1_{g_D, \ge \psi}(\Omega)$, $H^{1/2}(\Gamma)$, $H^{-1/2}(\Gamma)$	Sobolev spaces and subsets, see Section 1.2

Derivatives

$D^i A(u; v_1, \ldots, v_i)$	Fréchet derivatives of A in u
$\nabla A(u) = ((DA(u; e_i))_{1,\ldots,d}^T$	gradient of A in u
$\nabla^2 A(u) = ((D^2 A(u; e_i, e_j))_{\substack{i=1,\ldots,d \\ j=1,\ldots,d}}$	Hessian matrix of A in u

Introduction

The main purpose of this thesis is to present some *hp*-discretization techniques for the numerical treatment of variational inequalities corresponding to second order quasi-linear elliptic partial differential inequalities (PDI). We introduce subsets of conforming $p-version$ finite element spaces on quasi-uniform meshes, construct an adaptive *hp*-version based on a posteriori error estimation, and provide a solver for large-scale nonlinear minimum problems with equality and inequality constraints originated by the discretizations of variational inequalities.

Nonlinear PDI play an important role in the modeling of practical problems, for example:

(i) in mechanics of elastic and elasto-plastic bodies (cf. Hlaváček, Haslinger, Nečas, and Lovíšek [HHNL88]),

(ii) in geometry of minimal surfaces over obstacles (cf. Kinderlehrer and Stampacchia [KS80]).

The mathematical modeling exploits the fact that many processes in nature proceed according to extremal principles such as the principle of minimal potential energy in stable mechanical equilibrium states or the principle of stationary action in mechanics. The classical calculus of variations originated about 300 years ago in connection with extremal problems from mechanical problems by Euler.

(i) The link between an extremal problem without inequality side conditions, a variational equation, and a partial differential equation (PDE),

(ii) and the link between an extremal problem with equality and inequality side conditions, a variational inequality, and a partial differential inequality (PDI),

both are introduced in the context of convex functional analysis in Section 1.2, Theorem 1.22 and Theorem 1.23, respectively. The solutions of these extremal problems will be searched in a reflexive Banach space, typically a Sobolev and Hilbert space, when we have no side conditions. In case of side conditions, we look for the solution in a closed convex subset of a reflexive Banach space which guarantees that the side conditions are fulfilled. We give some model elliptic boundary problems and discuss the variational formulation for them in Section 1.2.

Variational equations can be solved approximately by searching the solution on *h*-, *p*-, and *hp*-discrete subspaces of Sobolev spaces called finite element (FE) spaces. Here, the *h*-version FEM achieves the convergence of the approximate solution in the Sobolev space

by mesh refinement, whereas the p-version FEM achieves the convergence by increasing the polynomial degree of the discrete subspace. The hp-version FEM names combinations of both methods which are tailored according to a priori knowledge of the solution, such as the distribution of singularities, or controlled adaptively by a posteriori error estimation.

The analysis of the relative advantages of the p-version approach over the classical h-version requires a careful look at the regularity of solutions of elliptic boundary value problems. The h-version gives asymptotically optimal approximations in terms of the number of degrees of freedom, if the regularity of the solution is measured in terms of $H^k(\Omega)$ spaces (see Babuška and Aziz [BA72]). For elliptic problems with singularities, however p- and hp-FEM with properly designed meshes are superior to standard h-version FEM with quasi-uniform meshes. The explanation is that the solutions to elliptic PDE are substantially smoother than $H^k(\Omega)$. For example, such solutions are analytic in the whole domain except in corners and edges on the boundary. The analysis of regularity demands the control of derivatives of any order of the solution in countably normed Sobolev spaces which take into account edge and corner singularities (see Babuška and Guo [BG88, BG89]). For the numerical treatment of elliptic PDE a proper combination of mesh refinement and increasing polynomial degree can achieve exponential convergence for elliptic problems with piecewise analytic solution, whereas h- or p-FEM converge at best algebraically.

For PDI we do not know such a sophisticated regularity analysis for the solution function. Kinderlehrer and Stampacchia give the example of an elliptic obstacle problem on a polyhedral domain which has a solution in $H^2(\Omega)$ in [KS80], when the obstacle is in $H^2(\Omega)$ (cf. Theorem 1.25, p. 30). We do not consider the regularity of PDI solutions in this thesis. Nevertheless, the situation seems intuitively to be comparable to that of PDE, when we decompose the domain into the *contact zone* where the solution u is equal to the obstacle ψ almost everywhere and into the *free zone* where u is greater than ψ. Then, u is governed by the material laws of a PDE on the free zone. On the contact zone, we have the regularity of the obstacle. Unfortunately, the problem with this intuitive approach is that we do not know the contact zone in advance, i.e., we can not tailor the mesh according to the boundary of the contact zone. To give an example: When we use p-version FE on quadrilaterals, we have to deal with the case that the boundary of the contact zone runs through quadrilaterals.

Due to Hlaváček, Haslinger, Nečas, and Lovíšek [HHNL88, Preface], the theory of variational inequalities is a relatively young mathematical discipline which main bases were developed by a paper from Fichera on the solution of the Signorini problem in the theory of elasticity in 1964 [Fic64]. The monograph *Numerical analysis of variational inequalities* from Glowinski, Lions, and Tremolieres [GLT81] presents an overview on inequality formulations and their approximate solution. In the book *An introduction to variational inequalities and their applications* [KS80] Kinderlehrer and Stampacchia present sophisticated theorems on the regularity of the solution for PDI and on the geometry of the contact zone. In the book *Solution of variational inequalities in mechanics* [HHNL88] Hlaváček et al. develop the solution of the Signorini generalized problem and discuss models of the theory of plastic flow theoretically and from the point of numerical analysis. They present h-version primal and dual discretizations and propose an a posteriori error estimator based on duality techniques. Gwinner and Stephan analyze the convergence of the h-version boundary element method for the treatment of a boundary integral in-

equality originated from the Signorini problem in [GS93]. The mentioned publications employ subsets of h-version FE with a polynomial degree ≤ 3 and achieve convergence by quasi-uniform or adaptive mesh refinement. A p-version boundary element approach for Signorini-type problems is presented by Maischak in [Mai01].

It is the intention of this thesis to extend the p-version on quasi-uniform meshes and the adaptive hp-version known for the treatment of PDE to the treatment of PDI. In the following, we want to outline the contents of the chapters.

In **Chapter 1**, we introduce the notations and the principles of convex functional analysis needed to answer the question of existence and uniqueness of solutions to a class of quasi-linear PDE and PDI. We define Sobolev spaces and closed convex subsets K of Sobolev spaces by equality and inequality constraint conditions on the contained functions. We give a mixed boundary value problem, an obstacle problem, and an obstacle problem with Signorini contact as model problems in Section 1.2.

In **Chapter 2**, we define subsets of p-version FE spaces which achieve the convergence of the discrete solutions against the Sobolev solution by increasing the polynomial degree. It is the basic idea of this p-version subsets to control the equality and inequality constraints of the variational solution only at certain points. Namely, the points are given on the reference square $[-1,1]^2$ by the tensor product of Gauss-Lobatto points. The images of these points onto the quadrilaterals of the mesh in the usual sense of FE define the control points. The p-version subsets are defined in Section 2.1. The convergence of the discrete solution u_p towards the solution u of the elliptic PDI with respect to the $\|\cdot\|_{H^1(\Omega)}$-norm follows by Theorem 2.8. For the case that the solution of the PDI is in $H^2(\Omega)$, we give an a priori error estimate by Theorem 2.11. Two approaches to the discretization on triangle meshes are sketched out in Section 2.2. Section 2.3 presents two simple numerical examples on a square and on a triangle which confirm the approximation properties of the p-version for PDI. Section 2.4 introduces a hp-version on non-uniform quadrilaterals including hanging nodes and a non-uniform distribution of polynomial degrees which can be used for adaptive refinement.

Chapter 3 addresses a posteriori error estimation and adaptive refinement of the hp-FE space. In Section 3.1 and Section 3.2, we take a look on the literature concerning a posteriori error estimation for the hp-FEM for PDE and for the h-FEM for PDI, respectively. Section 3.3 describes the hp-adaptive scheme based on dual-weighted a posteriori error estimation presented by Heuveline and Rannacher in [HR03]. In Section 3.4, we extend this approach to the treatment of variational inequalities using the work from Blum and Suttmeier on error estimation for h-FE solutions of variational inequalities.

Chapter 4 is devoted to implementation issues and the efficient solution of the large-scale nonlinear minimization problem with equality and inequality constraints originated by the discretization. We start with the definition of an appropriate basis in Section 4.1 which allows to write the minimization problems of Chapter 2 as nonlinear minimization problems with equality and inequality constraints on the vector components in $\mathbb{R}^N$. For the unconstrained nonlinear minimization, we present the *inexact Newton backtracking method* (see Algorithm 4.2) in Section 4.2. In Section 4.3, we specify the large-scale nonlinear minimizer given by Felkel in [Fel99] to our situation of a strictly convex minimization problem with equality and inequality constraints (see Algorithm 4.3). Here, the

minimum is searched by a combination of projected gradient steps with the mentioned inexact Newton backtracking method . The treatment of the unbounded and the bounded constrained discrete nonlinear problems demands preconditioned conjugated gradient iterations to solve linear systems. Section 4.4 and Section 4.5 analyze the influence of the polynomial degree p on the costs of the iterative solving of the linear systems. Here, we take a look on the literature on condition number estimates for the p-version (cf. Maitre, Pourquier [MP96], Melenk [Mel02]) and compare different strategies as diagonally preconditioning, static condensation, and hierarchical basis. The condition number estimates are confirmed numerically in Section 4.4.4. In Section 4.5 , we show how the preconditioners based on hierarchical bases (cf. Babuška et. al. [BCMP91], Ainsworth [Ain96]) can be used for preconditioning of linear systems originated by constrained minimization. In Section 4.6, we show how the continuity across inter-element boundaries between quadrilateral meshes with hanging nodes and with different polynomial degrees can be ensured by means of linear algebra. Thus, there is no need to introduce particular non-uniform basis functions for the non-uniform hp-version used by the adaptive scheme Algorithm 3.1. In Section 4.7, we apply the h-version and the p-version on uniform quadrilateral meshes to model obstacle problems given by the minimal surface operator with homogeneous and inhomogeneous Dirichlet boundary data. Here, the p-version turns out to be superior to the h-version concerning the convergence rate and the number of unknowns needed to reach a certain exactness of the approximation.

In **Chapter 5**, we present techniques which try to speed up the solving of the large-scale discrete nonlinear minimization problems by reducing them to nonlinear problems of lower dimension. In Section 5.1, we give a posteriori estimates for the first Newton iteration on a fine space, when the initial for the Newton method is determined on a subspace with lower polynomials degree and prolonged to the fine space (see Algorithm 5.1 and Proposition 5.1 for variational equalities, Algorithm 5.2 and Proposition 5.3 for variational inequalities). Section 5.2 describes a nonlinear solver for unconstrained minimization problems which decomposes the minimization space into a direct sum of subspaces and solves the corresponding low dimensional nonlinear minimization problems in parallel or sequentially (see Algorithm 5.3 and Algorithm 5.4, respectively). In Section 5.3, we extend the prolongation approach to a prolongation cascade heuristically and use this cascade to solve a nonlinear elliptic PDE with homogeneous boundary data. Further, we test the multiplicative nonlinear space decomposition method of Algorithm 5.4. The prolongation and the space decomposition approach, both are inferior to the straightforwardly applied global Newton method with respect to the totals of conjugate gradient iterations and computing time. We put this down to the fact that the number of Newton iterations needed for the minimization does not depend on the dimension of the problem in the numerical experiments.

The **Appendices A, B, C, and D** contain algorithms, tables, and technical proofs which are needed for notational and documentary reasons. During the implementation of the minimization algorithms, the computation of the Hessian of the minimization functional turned out as the bootle neck which slowed down the performance significantly. Here, the most efficient way, concerning cpu-time and memory management, to deal with this problem, is to avoid the computation of the Hessian at all. Since we employ an iterative solver, it suffices to implement the corresponding matrix-vector product. Appendix B presents algorithms for the computation of this matrix-vector product. Algorithm B.3 exploits the tensor product structure of the local basis function for an efficient numerical

evaluation of the occuring integrals by basis transformations which can be executed by highly optimized BLAS routines (cf. [ABB+95]). Proposition B.9 states that the number of floating point operations for the matrix-vector product grows with $\mathcal{O}(p^4)$.

Chapter 1

Variational inequalities in Hilbert space

In this chapter, we will present principles of convex functional analysis which, in particular, answer the question of existence and uniqueness of solutions to a class of nonlinear partial differential equations (PDE) and inequalities (PDI). For convenience and to introduce notations, we briefly recall a few basic concepts in Section 1.1. For a more detailed study, see [Alt91], [Yos94], [KS80] and [Zei85] among others. In Section 1.2, we take a second order quasi-linear differential operator as an example

(i) for the connection between an extremal problem without inequality side conditions, a variational equation, and a PDE,

(ii) and for the connection between an extremal problem with equality and inequality side conditions, a variational inequality, and a PDI.

1.1 Convexity and extremal principles

In this section, we introduce the concept of convexity of a functional on a reflexive Banach space. Further, we explain the link between Fréchet derivatives and extremal problems with side conditions.

Lemma 1.1. Let V be a normed linear space, V' its dual, and $V'' = (V')'$ its second dual. We note the duality pairings on V by

$$\langle v, u\rangle_V := v(u) \quad \text{for all } u \in V,\ v \in V'$$

and on V' by

$$\langle w, v\rangle_{V'} := w(v) \quad \text{for all } v \in V',\ w \in V''.$$

Then, the mapping $J_V : V \to V''$ given by

$$\langle J_V u, v\rangle_{V'} := \langle v, u\rangle_V$$

is well defined and an isometry.

Proof. see [Alt91, § 5.6]. □

Definition 1.2. With the notation of the above lemma V is said to be *reflexive* if the isometry J_V is surjective.

Remark 1.3. Every Hilbert space H is reflexive.

Definition 1.4. Let $(V, \|\cdot\|)$ be a *Banach space,* i.e., a normed space which is complete. A sequence $(u_n)_{n\in\mathbb{N}}$ in V is called *weakly convergent* to an element $u \in V$ if

$$\lim_{n\to\infty} \langle v, u_n\rangle = \langle v, u\rangle \quad \text{for all } v \in V'.$$

For a weakly convergent sequence we will write $u_n \rightharpoonup u$ as $n \to \infty$.

Definition 1.5. Let V be a linear space and let $A : M \subset V \to \mathbb{R}$ be a functional. The set M is said to be *convex* if

$$u, v \in M, t \in [0,1] \quad \text{implies} \quad (1-t)u + tv \in M.$$

If M is convex, then A is said to be *convex* if

$$A((1-t)u + tv) \le (1-t)A(u) + tA(v) \quad \text{for all } u, v \in M, \quad t \in (0,1). \tag{1.1}$$

A is called *strictly convex* if the last inequality holds with "$<$" instead of "$\le$".

Definition 1.6. Let $(V, \|\cdot\|)$ be a Banach space. A functional $A : V \to \mathbb{R}$ is called *coercive* if

$$\frac{A(u)}{\|u\|} \to \infty \quad \text{as } \|u\| \to \infty.$$

Lemma 1.7. The following assertions hold in a Banach space $(V, \|\cdot\|)$:

(i) $u_n \rightharpoonup u$ as $n \to \infty$ implies $u \in M$ when all u_n belong to M and M is a closed convex set in V.

(ii) If V is reflexive, then every bounded sequence in V has a weakly convergent subsequence.

Proof. see [Alt91, Satz 5.7 and Satz 5.10, respectively]. □

Definition 1.8. Let $(V, \|\cdot\|)$ be a Banach space and let $A : M \subset V \to [-\infty, \infty]$. The functional A is said to be *sequentially lower semicontinuous at* the point $u \in M$ if

$$A(u) \le \liminf_{n\to\infty} A(u_n) \tag{1.2}$$

holds for each sequence $(u_n)_{n\in\mathbb{N}}$ in M such that $u_n \to u$ as $n \to \infty$.
Similarly, A is said to be *weak sequentially lower semicontinuous at* $v \in M$ if (1.2) holds for each weak convergent sequence $(u_n)_{n\in\mathbb{N}}$ in M such that $u_n \rightharpoonup u$ as $n \to \infty$.

Lemma 1.9. For the functional $A : M \subset V \to \mathbb{R}$ with $M \neq \emptyset$, the minimum problem

$$\min_{v\in M} A(v) \tag{MP}$$

has a solution in case the following hold:

(i) $(V, \|\cdot\|)$ is a reflexive Banach space.

(ii) M is bounded and weak sequentially closed, i.e., by definition, for each sequence $(u_n)_{n\in\mathbb{N}}$ in M such that $u_n \rightharpoonup u$ as $n \to \infty$, we always have $u \in M$.

(iii) A is weak sequentially lower semicontinuous on M.

In particular, (ii) is fulfilled when M is bounded, closed, and convex.

Proof. Let $\gamma := \inf\{A(v) \,|\, v \in M\}$. We choose a sequence $(v_n)_{n\in\mathbb{N}}$ in M such that $A(v_n) \to \gamma$. Since M is bounded and V is reflexive, by Lemma 1.7 (ii), there exists a weak convergent subsequence $(v_{n_k})_{k\in\mathbb{N}}$ such that $v_{n_k} \rightharpoonup v$. From (ii) it follows that $v \in M$; therefore, $A(v) \le \liminf_{k\to\infty} A(v_{n_k}) = \gamma$ according to (iii). Since $\gamma \le A(v)$, we have $A(v) = \gamma$, i.e., v is a solution of (MP).

In particular, M is weak sequentially closed when it is closed and convex due to Lemma 1.7 (i). □

Lemma 1.10. The functional $A : M \subset V \to \mathbb{R}$ has at most one minimum on M in case the following hold:

(i) M is a convex subset of the linear space V.

(ii) A is strictly convex.

Proof. By (1.1) with "<" due to the strict convexity, we arrive at a contradiction for $A(u) = A(v) = \min\{A(w) \,|\, w \in M\}$ and $u \neq v$ when $t = \frac{1}{2}$. □

Definition 1.11. Let V be a reflexive Banach space with dual V'. Let $\langle\cdot,\cdot\rangle := \langle\cdot,\cdot\rangle_V$ denote a pairing between V' and V. Let M be a closed convex set. A mapping $F : M \to V'$ is said to be *monotone* if

$$\langle F(v) - F(u), v - u\rangle \ge 0 \quad \text{for all } u, v \in M.$$

The monotone mapping F is called *strictly monotone* if

$$\langle F(v) - F(u), v - u\rangle = 0 \quad \text{implies } u = v.$$

The monotone mapping F is called *uniformly monotone* if there exist fixed real numbers $p > 1$, $c > 0$, such that

$$\langle F(v) - F(u), v - u\rangle \ge c\|v - u\|^p \quad \text{for all } u, v \in M.$$

Lemma 1.12. Let $A : V \to \mathbb{R}$ be a functional on the real Banach space $(V, \|\cdot\|)$. Suppose the Fréchet derivative $DA : V \to V'$ exists on V. Then the following three assertions are equivalent:

(i) A is strictly convex on V.

(ii) DA is strictly monotone on V.

(iii) $A(v) - A(u) > DA(u; v - u)$ for all $u, v \in V$ such that $u \neq v$.

Assuming only convexity of A and monotonicity of DA instead of strict convexity and strict monotonicity, respectively, and $\geq$ instead of $>$ in (iii) the equivalence of the three assertions holds again.

If, in addition, the second Fréchet derivative exists for all $u, h \in V$, then one has the following criteria for convexity and coercivity, respectively:

(iv) If $D^2A(u;h,h) > 0$ for all $u, h \in V,\ h \neq 0$, then A is strictly convex on V.

(v) If $D^2A(u;h,h) \geq c\|h\|^p$ for all $u, h \in V$ and fixed $p > 1$, $c > 0$, and $t \mapsto D^2A(u+th;h,h)$ is continuous on $[0,1]$ for all $u, h \in V$, then DA is uniformly monotone and A is coercive on V.

Proof. For fixed $u, v \in V$ we define

$$\varphi(t) := A(u + t(v-u)) \quad \text{for all } t \in [0,1], \tag{1.3}$$

and write the derivatives

$$\varphi'(t) := DA(u + t(v-u); v-u)\,,$$

$$\varphi''(t) := D^2A(u + t(v-u); v-u, v-u)\,.$$

By definitions of convexity of A and monotonicity we have the following equivalences due to one dimensional analysis:

$$\begin{aligned} A \text{ is strictly convex on } V &\iff \varphi \text{ is strictly convex on } [0,1] \text{ for all } u, v \in V \\ &\iff \varphi' \text{ is strictly monotonely increasing on } [0,1] \\ &\qquad \text{for all } u, v \in V \\ &\iff \varphi'(1) - \varphi'(0) > 0 \text{ for all } u, v \in V,\ u \neq v \\ &\iff DA \text{ is strictly monotone on } V. \end{aligned}$$

We obtain the equivalence of (ii) and (iii) by noting the equivalence of both with

$$\varphi(1) - \varphi(0) = \varphi'(\tau) > \varphi'(0) \quad \text{for all } u, v \in V,\ u \neq v, \text{ for a } \tau \in (0,1).$$

Taking $v = u + h$ in $\varphi(t)$ leads to

$$\varphi''(t) > 0 \quad \text{for all } t \in [0,1],\ u, v \in V,\ u \neq v$$

which implies strict convexity of φ on $[0,1]$ and, hence, the strict convexity of A on V stated in (iv). We yield the uniform monotonicity of (v) from

$$DA(v;v-u) - DA(u;v-u) = \varphi'(1) - \varphi'(0) = \int_0^1 \varphi''(t)\ \mathrm{d}t \geq \int_0^1 c\|v-u\|^p\ \mathrm{d}t = c\|v-u\|^p\,.$$

Further, we write

$$\begin{aligned} \varphi(1) - \varphi(0) = \int_0^1 \varphi'(t)\ \mathrm{d}t\ &=\ \varphi'(0) + \int_0^1 (\varphi'(t) - \varphi'(0))\ \mathrm{d}t \\ &\geq\ \varphi'(0) + \int_0^1 ct\|v-u\|^p\ \mathrm{d}t\ =\ DA(u;v-u) + \tfrac{c}{2}\|v-u\|^p\,. \end{aligned}$$

As the Fréchet derivative $DA(u;\cdot) \in V'$ is a continuous linear functional, there exists a constant $\tilde{c}$ independent of v such that $\frac{1}{\|v\|}|DA(u;v)| \leq \tilde{c}$ for all $v \in V$. So, we obtain the coercivity of A from

$$\frac{A(v)}{\|v\|} \geq \frac{A(u)}{\|v\|} + \frac{DA(u;v)}{\|v\|} + \tfrac{c}{2}\|v\|^{p-1} \longrightarrow \infty \quad \text{as} \quad \|v\| \to \infty\,.$$

□

Theorem 1.13. Let $A : M \subset V \to \mathbb{R}$ be a functional on the convex nonempty set M of the real reflexive Banach space $(V, \|\cdot\|)$. Then:

(i) *Necessary condition:* If u is a solution of (MP), then

$$DA(u; v-u) \geq 0 \quad \text{for all } v \in M. \tag{VI}$$

(ii) *Equivalence.* If A is convex and DA exists as a Fréchet derivative on M, then the minimum problem (MP) and the variational inequality (VI) are mutually equivalent.

(iii) *Uniqueness.* If A is strictly convex on M, then (MP) and (VI) have at most one solution.

(iv) *Existence.* If A is weak sequentially lower semicontinuous and if M is closed, convex, bounded, and nonempty, then (MP) has a solution. For convex A, the solution set of (MP) is closed, convex, and bounded.

Proof. (i) Assuming u as a solution of (MP) and $v \in M$ arbitrarily, but fixed, we define φ as in the proof of Lemma 1.12 (1.3). For all $t \in [0,1]$, $\varphi(t) \geq \varphi(0)$, therefore $\varphi'(0) \geq 0$, but this is (VI).

(ii) Assuming u as a solution of (VI) we have $A(v) \geq A(u)$ for all $v \in M$ by the convexity of A due to Lemma 1.12 (iii) and therefore (MP).

(iii) and (iv) are special cases of Lemma 1.10 and Lemma 1.9, respectively. Alternatively, the uniqueness follows by Lemma 1.12. □

Corollary 1.14. The boundedness of M in Theorem 1.13 (iv) can be replaced by

$$A(u) \to \infty \quad \text{as} \quad \|u\| \to \infty\,.$$

Proof. By the assumption there exists an $R > 0$ such that $A(v) > A(u)$ holds for all v with $\|v-u\| > R$. Thus, we may replace M by the bounded set $M_R := \{v \in M \mid \|v-u\| \leq R\}$ which is also closed and convex. □

1.2 Some convex minimum formulations

In this section, we consider a second order quasi-linear elliptic PDE with mixed boundary conditions on a bounded Lipschitz domain Ω known as a scalar valued simplification of

the Hencky-Von Mises stress-strain relation. In Subsection 1.2.1, we give a weak minimum formulation and a variational formulation, both equivalent with the boundary value problem. Introducing an obstacle in the interior of Ω from below the material previously obeying the nonlinear PDE is forced upwards. In Subsection 1.2.2, this is described mathematically by searching the minimum of the mentioned weak minimum formulation on an appropriate subset of functions which take into consideration the obstacle. In the variational context this yields an equivalent variational inequality.

To start with, we recall the definitions of function spaces which will be essential to the analysis of PDE and PDI. A deep analysis of the Hölder continuous function spaces and the Sobolev spaces is given in the textbooks [Ada75] and [Mor66]. The relation "$\leq$"on a Sobolev space and the maximum of two Sobolev functions are introduce and analyzed in the book [KS80].

Definition 1.15. [KS80, Definition II.4.1] Let $\Omega \subset \mathbb{R}^d$, $d \in \mathbb{N}$, be a bounded open set with closure $\overline{\Omega}$ and boundary $\partial\Omega$. By $C^k(\overline{\Omega})$ we denote the space of real valued functions which are k times continuously differentiable on some neighborhood of $\overline{\Omega}$. By $C^{k,\lambda}(\overline{\Omega})$, $0 < \lambda < 1$, we indicate the functions k times differentiable in $\overline{\Omega}$ whose derivatives of order k are Hölder continuous with exponent λ, $0 < \lambda < 1$. Recall that $u \in C^{0,\lambda}(\overline{\Omega})$, if

$$[u]_\lambda := \sup_{x_1,x_2\in\overline{\Omega}} \frac{|u(x_1) - u(x_2)|}{|x_1 - x_2|^\lambda} \leq +\infty.$$

If we allow $\lambda = 1$, then u is called a *Lipschitz function.* The $d-$tuple of nonnegative integers $\alpha = (\alpha_1, \ldots, \alpha_d)$ is called a multi-index of length $|\alpha| = \alpha + \cdots + \alpha_d \geq 0$. We set $D^\alpha = \left(\frac{\partial}{\partial x_1}\right)^{\alpha_1} \ldots \left(\frac{\partial}{\partial x_1}\right)^{\alpha_d}$, a differential operator of order $|\alpha|$. Here $\left(\frac{\partial}{\partial x_i}\right)^0 u = u$, $1 \leq i \leq d$.

Definition 1.16. [KS80, Definition II.4.2] In the linear space $C^m(\overline{\Omega})$ we introduce the semi-norm

$$|v|_{H^{m,s}(\Omega)} = \sum_{|\alpha|=m} \|D^\alpha v\|_{L^s(\Omega)}, \quad 1 \leq s < \infty,$$

and the norm

$$\|v\|_{H^{m,s}(\Omega)} = \sum_{0\leq k\leq m} |v|_{H^{m,s}(\Omega)}, \quad 1 \leq s < \infty, \tag{1.4}$$

and we denote by $H^{m,s}(\Omega)$ the completion of $C^m(\overline{\Omega})$ in this norm. Usually, we write $H^m(\Omega) = H^{m,2}(\Omega)$.

Definition 1.17. [KS80, Definition II.4.3] Let $C_0^\infty(\Omega)$ denote the infinitely often differentiable functions having compact support in Ω. Then $H_0^{m,s}(\Omega)$ is defined as the closure of $C_0^\infty(\Omega)$ in the norm 1.4.

Definition 1.18. [KS80, Definition II.4.5] We denote the dual of $H_0^{m,s}(\Omega)$ by $H^{-m,s'}(\Omega)$, $\frac{1}{s} + \frac{1}{s'} = 1$, or simply $H^{-m}(\Omega)$ when $s = 2$.

Furthermore, we need the following boundary-function spaces $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$ to write the boundary conditions of PDE and PDI. A more sophisticated introduction of $H^s(\Gamma)$ spaces, $s \in \mathbb{R}$, and their connection to $H^1(\Omega)$ can be found in [LM72].

Definition 1.19. Let Γ be a nonempty, simply connected, and relatively open subset of the Lipschitz boundary $\partial\Omega$ and let γv be the trace of the function $v \in H^1(\Omega)$ on Γ, i.e., $\gamma v := v_{|\Gamma}$. Then, we denote the image of the space $H^1(\Omega)$ by

$$H^{1/2}(\Gamma) := \gamma\big(H^1(\Omega)\big).$$

The norm in $H^{1/2}(\Gamma)$ is defined by

$$\|v\|_{H^{1/2}(\Gamma)} = \inf_{\substack{w\in H^1(\Omega)\\ \gamma w = v}} \|w\|_{H^1(\Omega)}\,.$$

Further, we write $H^{-1/2}(\Gamma) := \big(H^{1/2}(\Gamma)\big)'$ for the dual space.

1.2.1 A mixed boundary value problem

Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain with boundary Γ. To describe mixed boundary conditions, let $\Gamma = \overline{\Gamma_D \cup \Gamma_N}$ where $\Gamma_D \neq \emptyset$ and Γ_N are simply connected, disjoint, and open in Γ. Let n denote the outward unit normal on Γ defined almost everywhere.

We consider the nonlinear PDE

$$-\operatorname{div}(\rho(|\nabla u|)\nabla u) + \sigma u - f = 0 \quad \text{in } \Omega, \tag{1.5}$$

where $\sigma \geq 0$ is real constant and $\rho : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ is a continuously differentiable function satisfying

$$\rho_0 \leq \rho(t) \leq \rho_1\,, \qquad \rho_2 \leq \rho(t) + t\rho'(t) \leq \rho_3 \tag{1.6}$$

with positive real constants ρ_i, $0 \leq i \leq 3$, for all $t = |\nabla u|$. We demand Dirichlet and Neumann boundary conditions

$$u_{|\Gamma_D} = g_D \quad \text{and} \quad \rho(|\nabla u|)\tfrac{\partial}{\partial n} u_{|\Gamma_N} = g_N \tag{1.7}$$

where $\frac{\partial}{\partial n} u$ denotes the derivative of u in direction of the outward normal on Γ.
Given data $f \in H^{-1}(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, and $g_N \in H^{-1/2}(\Gamma_N)$ we look for $u \in H^1(\Omega)$ satisfying (1.5)–(1.7) in a weak form.

We need functions of the affine subspace of $H^1(\Omega)$ which match with the Dirichlet data on $\Gamma_D \subset \Gamma$:

$$H^1_{g_D}(\Omega) := \{v \,|\, v \in H^1(\Omega),\, v_{|\Gamma_D} = g_D \text{ a.e. on } \Gamma_D\}\,.$$

Let the function $p : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ be given by ρ (see (1.6)) through the integral

$$p(t) := \int_0^t \tau\rho(\tau)\,\mathrm{d}\tau\,.$$

Now, we define the functional $A : H^1(\Omega) \to \mathbb{R}$ by

$$A(u) := \int_\Omega p(|\nabla u|)\,\mathrm{d}x + \tfrac{1}{2}\sigma \int_\Omega u^2\,\mathrm{d}x - \int_\Omega fu\,\mathrm{d}x - \int_{\Gamma_N} g_N u_{|\Gamma}\,\mathrm{d}s\,. \tag{1.8}$$

Lemma 1.20. For $0 < \rho_0 \le \rho(|\nabla u|)$ for all $u \in H^1(\Omega)$, there holds

$$A(u) \to +\infty \qquad \text{as } |u|_{H^1(\Omega)} \to +\infty\,.$$

Proof. With $t := |\nabla u|$ we have $p(t) \ge \int_0^t \tau\rho_0\, d\tau = \frac{1}{2}\rho_0 t^2$ and

$$\int_\Omega p(|\nabla u|)\ \mathrm{d}x \ge \tfrac{1}{2}\rho_0 |u|^2_{H^1(\Omega)}\,. \quad \square$$

Lemma 1.21.

(i) The Fréchet derivatives $D^i A$, $i = 1, 2, 3$, at $u \in H^1(\Omega)$ read as

$$DA(u; v) = \int_\Omega \Big(\rho(t)(\nabla u)^T \nabla v + \sigma\, uv - fv\Big)\ \mathrm{d}x - \int_{\Gamma_N} g_N v_{|\Gamma}\ \mathrm{d}s\,, \quad (1.9)$$

$$D^2 A(u; v_1, v_2) = \int_\Omega \Big(\rho(t)(\nabla v_1)^T \nabla v_2 + t\rho'(t) s_1 s_2 + \sigma\, v_1 v_2\Big)\ \mathrm{d}x\,, \quad (1.10)$$

$$D^3 A(u; v_1, v_2, v_3) = \int_\Omega \big(\rho'(t)(\hat{s}_{123} + \hat{s}_{213} + \hat{s}_{312} - s_1 s_2 s_3) + t\rho''(t)\, s_1 s_2 s_3\big)\ \mathrm{d}x \quad (1.11)$$

with $t := |\nabla u|$, $s_i := \frac{(\nabla u)^T}{t}\nabla v_i$, and $\hat{s}_{ijk} := s_i (\nabla v_j)^T \nabla v_k$, $\quad i, j, k \in \{1, 2, 3\}$, for all $v_1, v_2, v_3 \in H^1(\Omega)$.

In case of $t = 0$, we lose no generality, if we take $s_i := 0$ and $\hat{s}_{ijk} := 0$. Here, $D^3 A$ needs ρ two times continuously differentiable which is an additional demand concerning (1.5).

(ii) Assuming $\sigma = 0$ and positive constants ρ_1, ρ_2, ρ_3 with $\rho(t) \le \rho_1$ and $0 < \rho_2 \le \rho(t) + t\rho'(t) \le \rho_3$ for $t \in \mathbb{R}_{\ge 0}$, A is coercive and $D^2 A$ is continuous with respect to the semi-norm $|\cdot|_{H^1(\Omega)}$. There exist real constants κ_l, κ_u with $0 < \kappa_l \le \kappa_u$ such that

$$\kappa_l |u - v|^2_{H^1(\Omega)} \le DA(u; u - v) - DA(v; u - v) \le \kappa_u |u - v|^2_{H^1(\Omega)} \quad \text{for all } u, v \in H^1(\Omega)\,, \quad (1.12)$$

$$\kappa_l |v|^2_{H^1(\Omega)} \le D^2 A(u; v, v) \le \kappa_u |v|^2_{H^1(\Omega)} \qquad \text{for all } u, v \in H^1(\Omega)\,. \quad (1.13)$$

Assuming $\sigma > 0$, the last two inequalities hold, when we replace the semi-norm $|\cdot|_{H^1(\Omega)}$ by the norm $\|\cdot\|_{H^1(\Omega)}$.

(iii) Assuming constants ρ_4, ρ_5 with $|\rho'(t)| \le \rho_4$ and $|\rho'(t) + t\rho''(t)| \le \rho_5$ for $t \in \mathbb{R}_{\ge 0}$, the third derivative $D^3 A$ is continuous.

Proof. (i) The Fréchet derivatives follow by standard calculus.
(ii) We note that $s_i \le |\nabla v_i|$, $i = 1, 2$, almost everywhere. With the assumption $0 < \rho_2 \le \rho(t) + t\rho'(t)$ and taking $v_1 = v_2 = v$ in (1.10) we can write

$$D^2 A(u; v, v) = \int_\Omega \rho_2 |\nabla v|^2\ \mathrm{d}x + \int_\Omega \big((\rho(t) - \rho_2)|\nabla v|^2 + t\rho'(t) s_1^2 + \sigma\, v^2\big)\ \mathrm{d}x\,.$$

We estimate

$$0 \le \int_\Omega \big(\rho(t) - \rho_2 + t\rho'(t)\big)s_1^2 \,\mathrm{d}x \le \int_\Omega \big((\rho(t) - \rho_2)|\nabla v|^2 + t\rho'(t)s_1^2\big) \,\mathrm{d}x$$

and obtain (1.13) with $\kappa_l := \rho_2$. In case of $\sigma = 0$, (1.13) implies the uniform monotonicity of DA stated in the left inequality of (1.12) and the coercivity of A due to Lemma 1.12 (v). In case of $\sigma > 0$, (1.13), (1.12), and the coercivity of A follow by taking $\kappa_l := \min\{\sigma, \rho_2\}$.

Using $|t\rho'(t)| \le |\rho(t)| + |\rho(t) + t\rho'(t)| \le \rho_1 + \rho_3$ (see (1.6)), the Cauchy Schwarz inequality, and again $s_i \le |\nabla v_i|$ $(i, j = 1, 2)$, we obtain the continuity of D^2A in case of $\sigma = 0$ with

$$D^2A(u; v_1, v_2) \le (2\rho_1 + \rho_3)\,|v_1|_{H^1(\Omega)}|v_2|_{H^1(\Omega)} \quad \text{for all } u, v_1, v_2 \in H^1(\Omega)\,. \tag{1.14}$$

In case of $\sigma > 0$, there follows

$$D^2A(u; v, v) \le 2\max\{2\rho_1 + \rho_3, \sigma\}\,\|v\|^2_{H^1(\Omega)} \quad \text{for all } u, v \in H^1(\Omega). \tag{1.15}$$

The right inequalities of (1.12),(1.13) follow from (1.15) with $\kappa_u := 2\max\{2\rho_1 + \rho_3, \sigma\}$.
(iii) The continuity of D^3A follows completely analogously with

$$|D^3A(u; v_1, v_2, v_3)| \le (2\rho_4 + \rho_5)\,|v_1|_{H^1(\Omega)}|v_2|_{H^1(\Omega)}|v_3|_{H^1(\Omega)}\,,\ u, v_1, v_2, v_3 \in H^1(\Omega). \tag{1.16}$$

□

The following theorem connects a minimum of A, a variational equation given by the Fréchet derivative of A, and a PDE with mixed boundary conditions equivalently.

Theorem 1.22. (i) There exists a unique $u \in H^1_{g_D}(\Omega)$ which minimizes the functional A on $H^1_{g_D}(\Omega)$, i.e.,

$$A(u) \le A(v) \quad \text{for all } v \in H^1_{g_D}(\Omega)\,.$$

(ii) Furthermore, $u \in H^1_{g_D}(\Omega)$ solves the minimization problem of (i) if and only if u solves the variational equation

$$DA(u; v - u) = 0 \quad \text{for all } v \in H^1_{g_D}(\Omega)\,. \tag{VE}$$

(iii) Additionally, the solution of (VE) is the weak solution of the mixed boundary value problem given by (1.5)–(1.7) and vice versa.

Proof. With Lemma 1.21 A is a convex and differentiable functional on $H^1_{g_D}(\Omega)$ which is a affine subspace of $H^1(\Omega)$. Hence, $H^1_{g_D}(\Omega)$ is a convex, closed, and nonempty subset of the real reflexive Banach space $H^1(\Omega)$. Furthermore, we have $A(u) \to +\infty$ as $|u|_{H^1(\Omega)} \to +\infty$ with Lemma 1.20. Thus, the existence of a minimum of A, its uniqueness, and the equivalence (ii) follow with Theorem 1.13 and Corollary 1.14.

To show that (VE) implies (1.5)–(1.7) in the weak sense in (iii) we proceed as follows. We take $w \in C_0^\infty(\mathbb{R}^2)$ as test functions, i.e., w infinitely often differentiable with compact support, such that $v = u + w \in H^1_{g_D}(\Omega)$. Inserting v in $DA(u; v - u)$ yields

$$0 = \int_\Omega (-\operatorname{div}(\rho(|\nabla u|)\nabla u) + \sigma\, u - f)w \,\mathrm{d}x + \int_{\Gamma_N} (\rho(|\nabla u|)\tfrac{\partial}{\partial n}u_{|\Gamma} - g_N)\, w_{|\Gamma} \,\mathrm{d}s \tag{1.17}$$

by integration by parts. Since w is arbitrary, we obtain (1.5) in the weak sense by choosing $w \in C_0^\infty(\Omega) \subset C_0^\infty(\mathbb{R}^2)$. With $w \in C_0^\infty(\mathbb{R}^2)$ there follows $\rho(|\nabla u|)u_{|\Gamma}\frac{\partial}{\partial n} = g_N$. Thus, the Neumann condition in (1.7) is satisfied almost everywhere.

Conversely, we get (VE) from the problem given by (1.5)–(1.7) by multiplying (1.5) and $\rho(|\nabla u|)\frac{\partial}{\partial n}u_{|\Gamma_N} - g_N = 0$ with $(v-u)$ and integrating by parts. □

1.2.2 An obstacle problem

For brevity of notation, we define the relation $\leq$ on the set $M \subset \Omega$ in $H^1(\Omega)$ by $0 \leq u$ almost everywhere on M. Nevertheless, neglecting the "almost everywhere" we divide Ω sloppily into the set $\{x \in \Omega \,|\, u(x) > 0\}$ which is open, and its complement $\{x \in \Omega \,|\, u(x) \leq 0\}$. A more detailed analysis of different definitions of $\leq$ on M in $H^1(\Omega)$ given in [KS80, Chapter II] shows that this simplification yields mutually the same results.

We consider the following obstacle problem (P) and derive an equivalent variational inequality (VI). Again, let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain and $\Gamma := \partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$ where $\Gamma_D \neq \emptyset$ and Γ_N are simply connected and disjoint with outward directed unit normal n on Γ.

Problem (P).
For given data $f \in H^{-1}(\Omega)$, $g_D \in H^{1/2}(\Gamma_D)$, $g_N \in H^{-1/2}(\Gamma_N)$, and $\psi \in H^1(\Omega)$ with $\psi \leq g_D$ a.e. in a neighborhood of Γ_D, $\sigma \geq 0$, we look for $u \in H^1(\Omega)$ satisfying

$$\mathcal{P}(u) := -\operatorname{div}(\rho(|\nabla u|)\nabla u) + \sigma u - f \geq 0 \quad \text{in } \Omega\,, \tag{1.18}$$

$$u = g_D \text{ on } \Gamma_D\,, \qquad \mathcal{P}_{\Gamma_N}(u) := \rho(|\nabla u|)\tfrac{\partial}{\partial n}u - g_N \geq 0 \text{ on } \Gamma_N\,, \tag{1.19}$$

$$(u-\psi)\cdot\mathcal{P}(u) = 0 \text{ on } \Omega, \qquad (u-\psi)\cdot\mathcal{P}_{\Gamma_N}(u) = 0 \text{ on } \Gamma_N, \tag{1.20}$$

$$\text{and} \qquad u \geq \psi \quad \text{ on } \Omega \tag{1.21}$$

where $\rho \in C^1(\mathbb{R}_{\geq 0})$ satisfies (1.6) for all $t \in \mathbb{R}_{\geq 0}$ with constants $\rho_i > 0$, $0 \leq i \leq 3$.

Next, we introduce the cone

$$u \in K := \{v \in H^1_{g_D}(\Omega) \,|\, v \geq \psi \text{ a.e. on } \Omega\}$$

where

$$H^1_{g_D}(\Omega) := \{v \,|\, v \in H^1(\Omega),\, v_{|\Gamma_D} = g_D \text{ a.e. on } \Gamma_D\}\,,$$

and the coincidence set

$$\Psi := \{x \in \Omega \,|\, u(x) = \psi(x)\}. \tag{1.22}$$

The following equivalence result follows almost analogously to the proof of Theorem 1.22 with Theorem 1.13 and Corollary 1.14.

Theorem 1.23. Let the functional A be defined by (1.8).
(i) There exists a unique $u \in K$ which minimizes the functional A on K, i.e.,

$$A(u) \leq A(v) \qquad \text{for all } v \in K\,. \tag{MP}$$

(ii) Furthermore, $u \in K$ solves (MP) if and only if u solves the variational inequality

$$DA(u; v-u) \geq 0 \qquad \text{for all } v \in K\,. \tag{VI}$$

(iii) Additionally, the variational inequality (VI) and problem (P) are equivalent, i.e., the solution of (VI) is the weak solution of (P) and vice versa.

Proof. The existence of a minimum of A, its uniqueness, and the equivalence (ii) follow due to Theorem 1.13 and Corollary 1.14, if we prove that

(α) K is a *convex, closed, and nonempty subset* of the real reflexive Banach space $H^1(\Omega)$,

(β) A is strictly convex on K, i.e., $A((1-t)u+tv) < (1-t)A(u)+tA(v)$ for all $u, v \in K$, $t \in (0,1)$,

(γ) $A(u) \to \infty$ as $\|u\|_{H^1(\Omega)} \to \infty$.

(β) and (γ) follow from Lemma 1.21 and Lemma 1.20. (α) is verified as follows: Let w solve problem (P) with $\mathcal{P}(w) = 0$, $\mathcal{P}_{\Gamma_N}(w) = 0$, and without the obstacle condition (1.21), and set

$$\psi^+ := \max(\psi, w) := \begin{cases} \psi & \text{on } \{x \in \Omega \,|\, \psi(x) > w(x) \ \text{ a.e.}\} \\ w & \text{else.} \end{cases}$$

Due to [KS80, Theorem II.A.1] $\psi^+ \in H^1_{g_D}(\Omega)$ and $\psi^+ \geq \psi$. Therefore, K is nonempty. Let $w_n \to w$ strongly in $H^1(\Omega)$, where $w_n \in K$ and $w \in H^1(\Omega)$. Therefore, we have strong convergence in $L^2(\Omega)$ and the existence of a subsequence $\{w_{n_k}\}$ with $w_{n_k} \to w$ pointwise a.e. on Ω. Now, $w_{n_k} \geq \psi$ a.e. on Ω implies $w \geq \psi$ a.e. on Ω, i.e., $w \in K$. Hence, K is closed in $H^1(\Omega)$. Furthermore, K is obviously convex.

To show that (VI) implies (P) in (iii) we proceed as follows. We take $w \in C_0^\infty(\Omega)$ with $w \geq 0$. Inserting $v = u + w \in K$ in $DA(u; v-u)$ yields

$$0 \leq \int_\Omega \big(-\operatorname{div}(\rho(|\nabla u|)\nabla u) + \sigma u - f\big) w \,\mathrm{d}x$$

by integration by parts. This proves (1.18). To show the right inequality of (1.19), we take non-negative test functions $w \in C_0^\infty(B_\delta(x))$ for $x \in \Gamma_N$ with sufficiently small $\delta > 0$. Thus, $v := u + w \in K$ and integration by parts give by standard arguments

$$0 \leq \int_{\Gamma_N} (\rho(|\nabla u|)\tfrac{\partial}{\partial n} u_{|\Gamma} - g_N)\, w_{|\Gamma} \,\mathrm{d}s.$$

The equality $\mathcal{P}(u) = 0$ on $\Omega \setminus \Psi$ follows, if we assume $w \in C_0^\infty(\Omega)$ non-negative only on Ψ. The equality $\mathcal{P}_{\Gamma_N}(u) = 0$ on $\Gamma_N \cap \partial(\Omega\setminus\Psi)$ is obtained, if we take $w \in C_0^\infty(B_\delta(x)))$ non-negative only for $x \in \Gamma_N \setminus \partial(\Omega\setminus\Psi)$. This proves the complementary condition (1.20).

Conversely, we get (VI) from (P) by multiplying (1.18) and the inequality of (1.19) by $(v-u)$ and integrating by parts. □

Remark 1.24. Theorem 1.23 can be motivated as follows. Let the obstacle ψ be defined as in Problem (P). Let $\tilde{u}$ be a minimizer of A on $H^1_{g_D}(\Omega)$. It is known from nonlinear

functional analysis and it is a particular implication of Theorem 1.23 that $\tilde{u}$ solves the PDE $\mathcal{P}(\tilde{u}) = 0$ with Dirichlet and Neumann conditions

$$\tilde{u}_{|\Gamma_D} = g_D \quad \text{and} \quad \rho(|\nabla \tilde{u}|)\tfrac{\partial}{\partial n}\tilde{u}_{|\Gamma_N} = g_N\,.$$

Now, we assume that $\tilde{u}$ is pushed upwards by the obstacle ψ. This pushed function is said to be the solution u of the obstacle problem. Of course, u violates the PDE $\mathcal{P}(u) = 0$ when $\tilde{u} < \psi$. But it seems natural that the geometric or physical properties of u still minimize the functional A when the solution $u \in H^1_{g_D}(\Omega)$ satisfies $u \geq \psi$.

The smoothness of the solution u is studied by Kinderlehrer and Stampacchia by a method called *penalization.* The method of penalization consists in substituting the variational inequality by a family of nonlinear boundary value problems and demonstrating that their solutions converge to the solution of the variational inequality. A general guideline for Lipschitz domains in $\mathbb{R}^d$, $d \in \mathbb{N}$, is that the solution of the variational inequality is in $H^{2,s}(\Omega)$ whenever the solution of the associated obstacle-free boundary value has this property. We can cite the following result for the obstacle problem considered in Theorem 1.23 when we assume homogeneous Neumann conditions.

Theorem 1.25. Let $f \in L^s(\Omega)$ and $\psi, g \in H^{2,s}(\Omega)$ (see Definition 1.16), $\psi \leq g$, for some s, $d < s < \infty$. Further, let u be the unique solution of Theorem 1.23 with $g_D = g$ on Γ_D and $g_N \equiv 0$. In addition, assume that for a positive constant C independent of f there holds

$$\|w\|_{H^{2,s}(\Omega)} \leq C\big(\|f\|_{L^s(\Omega)} + \|g\|_{H^{2,s}(\Omega)}\big)$$

where w is the solution of the obstacle-free mixed boundary value problem. Then, the solution u of Theorem 1.23 satisfies $u \in H^{2,s}(\Omega) \cap C^{1,\lambda}(\overline{\Omega})$, $\lambda = 1 - \frac{d}{s}$.

Proof. See [KS80, Theorem IV.2.5]. □

1.2.3 An obstacle problem with Signorini contact

Now, we extend the obstacle problem (P) from the previous section to an obstacle problem (SP) with inequality boundary conditions also known as Signorini conditions. Inequality boundary conditions describe steady-state phenomena which arise for example in thermics, fluid mechanics, and elasto-statics.

Problem (SP).
Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain. To describe mixed boundary conditions, let $\Gamma := \partial\Omega = \overline{\Gamma_D \cup \Gamma_N \cup \Gamma_S}$ where $\Gamma_D \neq \emptyset$, Γ_N, Γ_S are simply connected, disjoint and open in Γ. For given data $f \in H^{-1}(\Omega)$, $g \in H^{1/2}(\Gamma)$, $\hat{g} \in H^{-1/2}(\Gamma)$, and $\psi \in H^1(\Omega)$ with $\psi \leq g$ in a neighborhood of $\Gamma_D \cup\ \Gamma_S$, $\sigma \geq 0$, we look for $u \in H^1(\Omega)$ satisfying

$$\mathcal{P}_S(u) := -\operatorname{div}(\rho(|\nabla u|)\nabla u) + \sigma u - f \geq 0 \quad \text{in } \Omega\,, \tag{1.23}$$

the Dirichlet and Neumann conditions

$$u_{|\Gamma_D} = g \quad \text{on } \Gamma_D, \qquad \rho(|\nabla u|)\tfrac{\partial}{\partial n}u_{|\Gamma_N} = \hat{g} \text{ on } \Gamma_N, \tag{1.24}$$

the Signorini conditions

$$u_{|\Gamma_S} \geq g \quad \text{on } \Gamma_S, \qquad \mathcal{P}_{\Gamma_S}(u) := \rho(|\nabla u|)\tfrac{\partial}{\partial n} u_{|\Gamma_S} - \hat{g} \geq 0 \quad \text{on } \Gamma_S, \tag{1.25}$$

and the complementary conditions

$$\big(u-\psi\big)\cdot \mathcal{P}_S(u) = 0 \quad \text{in } \Omega, \qquad \big(u_{|\Gamma_S} - g\big)\cdot \mathcal{P}_{\Gamma_S}(u) = 0 \quad \text{on } \Gamma_S \tag{1.26}$$

where $\rho \in C^1(\mathbb{R}_{\geq 0})$ satisfies (1.6).

We define the functional $A_S : H^1(\Omega) \to \mathbb{R}$ by

$$A_S(u) := A(u) - \int_{\Gamma_S} \hat{g}\, u_{|\Gamma}\, \mathrm{d}s \tag{1.27}$$

where A is given by (1.8) with $\hat{g}$ replacing g_N. Next, we introduce the cone

$$K_S := \{v \in H^1(\Omega) \,|\, v \geq \psi \text{ a.e. on } \Omega,\ v = g \text{ a.e. on } \Gamma_D, \text{ and } v \geq g \text{ a.e. on } \Gamma_S\}.$$

Theorem 1.26. Let the functional A_S be defined by (1.27).
(i) There exists a unique $u \in K_S$ which minimizes the functional A_S on K_S, i.e.,

$$A_S(u) \leq A_S(v) \qquad \text{for all } v \in K_S\,. \tag{MSP}$$

(ii) Furthermore, $u \in K_S$ solves (MSP) if and only if u solves the variational inequality

$$DA_S(u; v-u) \geq 0 \qquad \text{for all } v \in K_S\,. \tag{VSI}$$

(iii) Additionally, the variational inequality (VSI) and problem (SP) are equivalent, i.e., the solution of (VSI) is the weak solution of (SP) and vice versa.

Proof. The existence of a minimum of A_S, its uniqueness, and the equivalence (ii) follow due to Theorem 1.13 and Corollary 1.14, if we prove that

(α) K_S is a *convex, closed, and nonempty subset* of the real reflexive Banach space $H^1(\Omega)$,

(β) A_S is strictly convex on K_S,

(γ) $A_S(u) \to \infty$ as $\|u\|_{H^1(\Omega)} \to \infty$.

Since $D^2 A_S = D^2 A$ due to the linearity of $\int_{\Gamma_S} \hat{g}\, u_\Gamma \,\mathrm{d}s$, ($\beta$) and ($\gamma$) follow from Lemma 1.21 and Lemma 1.20. (α) is verified as follows: We have $\psi \in H^1(\Omega)$ with $\psi \leq g$ a.e. on $\Gamma_D \cup \Gamma_S$. Let w be a solution in the sense of Theorem 1.22 with $g_D := g_{|\Gamma_D \cup \Gamma_S}$ and Γ_D replaced by $\Gamma_D \cup \Gamma_S$. Let ψ^+ be defined by $\psi^+ := \max(\psi, w)$ almost everywhere on Ω. This implies $\psi^+ \in H^1_{g_D}(\Omega)$ and $\psi^+ \geq \psi$. Therefore, K_S is nonempty. Let $w_n \to w$ converge strongly in $H^1(\Omega)$, where $w_n \in K_S$ and $w \in H^1(\Omega)$. Therefore, we have strong convergence in $L^2(\Omega)$ and the existence of a subsequence (w_{n_k}) with $w_{n_k} \to w$ pointwise almost everywhere in Ω. Now, $w_{n_k} \geq \psi$ a.e. in Ω implies $w \geq \psi$ a.e. in Ω. Additionally, $w_{n_k} \geq g$ a.e. on Γ_S implies $w \geq g$ on Γ_S, and $w_{n_k} = g$ a.e. on Γ_D implies $w = g$ on Γ_D. As $w \in K_S$, K_S is closed in $H^1(\Omega)$. Furthermore, K_S is obviously convex.

To show that (VSI) implies (SP) in (iii) we proceed as follows. As in the proof of Theorem 1.23 we take nonnegative $w \in C_0^\infty(\Omega)$ and get the PDI (1.23) in the weak sense. The Dirichlet condition is satisfied due to construction of K_S. Again, as in the proof of Theorem 1.23, we take nonnegative $w \in C_0^\infty(B_\delta(x))$ for $x \in \Gamma_N$ and sufficiently small $\delta > 0$, and obtain the Neumann condition in (1.24). We have $u_{|\Gamma_S} \geq g$ almost everywhere on Γ_S by construction of K_S. Now, we take nonnegative $w \in C_0^\infty(B_\delta(x))$ for $x \in \Gamma_S$. Thus, $v := u + w \in K_S$ and integration by parts of $D^2 A_S(u; v - u)$ give by standard arguments

$$0 \leq \int_{\Gamma_S} w_{|\Gamma} \cdot \mathcal{P}_{\Gamma_S}(u) \, \mathrm{d}s$$

and the right inequality of (1.25).

The left equation of the complementary condition (1.26) is obtained analogously to the proof of the left complementary condition in (1.20) in the proof of Theorem 1.23. To show the right equation of (1.26) we note that there exists an extension $\bar{g} \in H^1(\Omega)$ of $g_{|\Gamma_D \cup \Gamma_S} \in H^{1/2}(\Gamma_D \cup \Gamma_S)$ with $\bar{g} \geq \psi$ in Ω due to Theorem 1.23. We insert $v = \bar{g}$ and $v = 2u - \bar{g}$ in $D^2 A_S(u; v - u)$, and get

$$0 \leq \pm \int_{\Gamma_S} (u - \bar{g})_{|\Gamma} \cdot \mathcal{P}_{\Gamma_S}(u) \, \mathrm{d}s.$$

Conversely, we obtain (VSI) from (SP) by multiplying (1.23), $\rho(|\nabla u|) \frac{\partial}{\partial n} u_{|\Gamma_N} - \hat{g}$, and the inequality of (1.25) by $(v - u)$ and integrating by parts. □

Chapter 2

Discretization

The *hp*-FEM for scalar elliptic problems in two-dimensional domains with a Lipschitz boundary is analyzed by many authors for quasi-uniform and geometric refined meshes (cf. Ainsworth, Babuška, Bernardi, Maday, Schwab). In the present chapter we extend the *hp*-approach to nonlinear scalar elliptic PDI mentioned in Chapter 1. For that we demand that the inequality constraint condition is fulfilled on a discrete set of points, namely the images of the tensorized Gauss-Lobatto points. We proof the existence and uniqueness of a discrete solution and its convergence towards the continuous solution with respect to the $\|\cdot\|_{H^1(\Omega)}$-norm.

It is known from Theorem 1.25 that the obstacle problem owns a solution $u \in H^2(\Omega)$ when the partial differential inequality and the obstacle ψ fulfill particular regularity properties. For this case the convergence result can be improved by an analysis of the convergence rate. In Section 2.1 the discrete subsets on quadrilateral meshes are introduced, and their approximation properties are discussed. Two approaches to the discretization on triangle meshes are sketched out in Section 2.2. Section 2.3 presents two simple numerical examples on a square and on a triangle. The experiments confirm a convergence rate of $\mathcal{O}(p^{-1})$.

A judicious combination of mesh refinement towards the corners of a polygon and increase of the polynomial degree p used in the approximation of partial differential equations with corner singularities were proved to achieve exponential convergence (cf. [BG88, BG89]). Also adaptive *hp*-refinement based on a posteriori error estimation (see Chapter 3) demands flexible ways of matching quadrilaterals and triangles with different polynomial degree. Section 2.4 is devoted to non-uniform *hp*-quadrilateral meshes including hanging nodes and a non-uniform distribution of polynomial degrees.

2.1 p-discretization of $K \subset H^1(\Omega)$ on quadrilaterals

Let $\tilde{Q} = (-1, 1)^2$ be the reference square. $\hat{v}_i$ and $\hat{\gamma}_i$ denote the vertices and sides, respectively, of $\tilde{Q}$.

The image of $\tilde{Q}$ under the mapping F_Q, $F_Q : \tilde{Q} \to Q = F_Q(\tilde{Q})$ is denoted by Q. We assume that F_Q is a bounded diffeomorphism, i.e., there exist positive constants C_i, $i = 1, 2, 3, 4$, with

$$\begin{aligned} |F_Q|_{1,\infty,\tilde{Q}} &\le C_1 h_Q, \qquad & |F_Q^{-1}|_{1,\infty,Q} &\le C_2 h_Q^{-1}, \\ |J_{F_Q}|_{0,\infty,\tilde{Q}} &\le C_3 h_Q^2, \qquad & |J_{F_Q^{-1}}|_{0,\infty,Q} &\le C_4 h_Q^{-2}, \end{aligned} \tag{2.1}$$

where $J_{F_Q} = \det DF_Q$ and $J_{F_Q^{-1}} = \det DF_Q^{-1}$ are the Jacobians of F_Q and F_Q^{-1}, respectively, and

$$|F_Q|_{k,\infty,\tilde{Q}} := \sup_{\substack{\tilde{x}\in\tilde{Q} \\ |l|=k}} |D^l F_Q(\tilde{x})| \,.$$

The parameters h_Q are numbers proportional to the diameter of Q.

Let $\Omega \subset \mathbb{R}^2$ be a domain bounded by a finite number of polynomials arcs with the end points at the vertices of Ω. Let $\mathcal{T}$ be a decomposition of Ω into a finite number of curvilinear quadrilaterals $Q = F_Q(\tilde{Q})$ with the same constants C_i, $i = 1, 2, 3, 4$, of (2.1) for all F_Q such that the following assumptions are fulfilled.

(i) For every vertex x of Ω there exists a $Q \in \mathcal{T}$ with a vertex $x_Q \in Q$ that coincides with x.

(ii) $\overline{\Omega} = \bigcup_{Q\in\mathcal{T}} \overline{Q}$ and the intersection $\overline{Q}_i \cap \overline{Q}_j$, $i \neq j$, $Q_i, Q_j \in \mathcal{T}$ is either a common vertex or common edge or empty.

(iii) For every inner edge γ_{ij} with $\gamma_{ij} = \overline{Q}_i \cap \overline{Q}_j$, $i \neq j$, the mappings $F_{Q_i}^{-1}{}_{|\gamma_{ij}}$ and $\hat{F} \circ F_{Q_j}^{-1}{}_{|\gamma_{ij}}$ coincide in the usual sense of finite elements, i.e., for an appropriately chosen isometric mapping $\hat{F} : \hat{\gamma}_j \to \hat{\gamma}_i$ between the edges of the reference square.

Now, we define the p-version finite element spaces on the reference square. By $\mathbb{P}_p^2(\overline{\tilde{Q}})$, $p \ge 1$, we denote the space of tensor product polynomials which are of degree at most p separately in the components x_1 and x_2 of x.

For the partition $\mathcal{T}$ we define the FE spaces

$$V_p := V_p(\mathcal{T}) := \{u \in H^1(\Omega) : u_{|Q} \circ F_Q \in \mathbb{P}_p^2(\tilde{Q}),\, Q \in \mathcal{T}\} \,. \tag{2.2}$$

To approximate the solutions u in the sense of Theorem 1.22 and Theorem 1.23, respectively, we introduce the affine subspace V_{p,g_D} and the subset K_{p,g_D} of V_p, respectively. We demand that the functions of V_{p,g_D} coincide with the boundary function g_D on a finite set $\Gamma_{D,p} \subset \Gamma$ of points. Concerning the obstacle problem, we demand further that the functions of K_{p,g_D} are greater equal the obstacle ψ on the set G_p. Before we can give the sets $\Gamma_{D,p}$ and G_p in Definition 2.2 we need some basics about Gauss-Lobatto quadrature.

Definition 2.1. The family of *Legendre polynomials* is the family $(L_p)_{p\ge0}$ of polynomials with one variable, which are orthogonal to each other with respect to the scalar product $\langle\cdot,\cdot\rangle_{L^2([-1,1])}$ and such that, for any integer $p \ge 0$, the polynomial L_p has degree p and satisfies: $L_p(1) = 1$.

Let us recall four basic properties that characterize these polynomials.

(i) For any positive integer p, the zeros of L_p are distinct real numbers ζ_i^p, $1 \le i \le p$ in $(-1,1)$, called *Gauss points of degree p.*

(ii) *Gauss quadrature.* There exist positive weight factors ω_i^p, $1 \le i \le p$, such that

$$\int_{-1}^{1} \phi(\zeta)\, \mathrm{d}\zeta = \sum_{i=1}^{p} \phi(\zeta_i^p)\, \omega_i^p$$

for all polynomials $\phi \in \mathbb{P}_{2p-1}([-1,1])$.

(iii) For any positive integer p, the zeros of $(1-\xi^2)L_p'(\xi)$ are distinct real numbers ξ_i^{p+1}, $0 \le i \le p$, in $[-1,1]$ called *Gauss-Lobatto points of degree p.*

(iv) *Gauss-Lobatto quadrature.* There exist positive weight factors ρ_i^{p+1}, $0 \le i \le p$, such that

$$\int_{-1}^{1} \phi(\zeta)\, \mathrm{d}\zeta = \sum_{i=0}^{p} \phi(\xi_i^{p+1})\rho_i^{p+1} \tag{2.3}$$

for all polynomials $\phi \in \mathbb{P}_{2p-1}([-1,1])$.

Now, we use the Gauss-Lobatto points to define the sets of points $G_{\tilde{Q},p}$, $G_{Q,p}$, G_p, and $\Gamma_{D,p}$.

Definition 2.2. Let ξ_i^{p+1}, $0 \le i \le p$, be the Gauss-Lobatto points of degree p. On the closed reference square $\overline{\tilde{Q}} = [-1,1]^2$ we define the set

$$G_{\tilde{Q},p} := \{(\xi_i^{p+1}, \xi_j^{p+1}) \mid 0 \le i,j \le p\}\,.$$

We assume that the mappings F_Q can be extended to the closed sets $\overline{\tilde{Q}}$, $\overline{Q}$ such that $F_Q : \overline{\tilde{Q}} \to \overline{Q}$. Now, using these transformations F_Q, we define

$$G_{Q,p} := \{F_Q(\xi) \mid \xi \in G_{\tilde{Q},p}\} \quad \text{and} \quad G_p := \bigcup_{Q \in \mathcal{T}} G_{Q,p}$$

on the curvilinear quadrilaterals $\overline{Q} = F_Q(\overline{\tilde{Q}})$, $Q \in \mathcal{T}$, and on $\overline{\Omega}$, respectively.
We denote the subset of G_p on the closed Dirichlet boundary by

$$\Gamma_{D,p} := \overline{\Gamma}_D \cap G_p\,.$$

The following approximation result concerning the polynomial interpolate $i_{Q^d,p}$ on the reference hypercube Q^d of dimension $d \in \mathbb{N}$ with respect to the tensorized Gauss-Lobatto points was proved by Bernardi and Maday. It will be used below to prove the convergence of the p-FE solution u_p towards u when $p \to \infty$.

Theorem 2.3. Let d the dimension of the reference element $Q^d := [-1,1]^d$. For any real numbers r and s satisfying $s > (d+r)/2$ and $0 \le r \le 1$, there exists a positive constant c depending only on s such that, for any function $v \in H^s(Q^d)$, the following estimate holds

$$\|v - i_{Q^d,p} v\|_{H^r(Q^d)} \le c p^{r-s} \|v\|_{H^s(Q^d)}.$$

The estimate also holds in the case $s = \frac{d+1}{2}$ and $r = 1$.

Proof. [BM97, Theorem 14.2]. □

We denote the *polynomial interpolate of degree p* of a continuous function v with respect to a set of points M by $i_M v$. Since we assumed $\mathcal{T}$ to be a quasi-uniform mesh it follows straightforwardly that Theorem 2.3 also holds for the interpolates $i_{G_{Q,p}} v$ on Q and $i_{G_p} v$ on Ω. For ease of notation we will write only $i_p v$ in the following and demand the appropriate sets of points $G_{\tilde{Q},p}$, $G_{Q,p}$, and G_p to be given implicitly due to Definition 2.2 by the domains $\tilde{Q}$, Q, and Ω, respectively.

With G_p and $\Gamma_{D,p}$ we have now sets of points which allow us to define appropriate subsets of V_p for the approximation of the solution u of the boundary value problems given in Theorem 1.22 and Theorem 1.23.

Definition 2.4. Let $V_p = V_p(\mathcal{T})$ be a FE space on the quasi-uniform decomposition $\mathcal{T}$ of the domain Ω with degree $p \geq 1$ as defined by equation (2.2). We define

$$V_{p,g_D} := \{w \in V_p \,|\, w(x) = g_D(x), x \in \Gamma_{D,p}\},$$

and

$$K_{p,g_D} := \{w \in V_{p,g_D} \,|\, w(x) \geq \psi(x), x \in G_p\}.$$

Lemma 2.5. V_{p,g_D} and K_{p,g_D} are nonempty closed convex subsets of V_p.

Proof. From interpolation theory it is known that there exists an interpolating polynomial $v \in V_p$ with $v(x) \geq \psi(x)$ for all $x \in G_p$ and $v(x) = g_D(x)$ for all $x \in \Gamma_{D,p}$. Thus K_{p,g_D} is *nonempty.* The *convexity* of K_{p,g_D} is trivial.

Let $v_n \to v$ strongly in $H^1(\Omega)$, where $v_n \in K_{p,g_D}$ and $v \in H^1(\Omega)$. With $v_n(x) \geq \psi(x)$ for all $x \in G_p$ and $v_n(x) = g(x)$ for all $x \in \Gamma_{D,p}$, there follows $v(x) \geq \psi(x)$ for all $x \in G_p$ and $v(x) = g(x)$ for all $x \in \Gamma_{D,p}$. Therefore $v \in K_{p,g_D}$. Dropping the above claims $\geq \psi(x)$, the statement concerning V_{p,g_D} follows. □

The unique p-version finite element approximations for the exact solution $u \in H^1_{g_D}(\Omega)$ from Theorem 1.22 and $u \in K$ from Theorem 1.23 are obtained as follows.

Theorem 2.6. Let the functional A be defined by (1.8).
(i) There exists a unique $u_p \in V_{p,g_D}$ which minimizes the functional A on V_{p,g_D}, i.e.,

$$A(u_p) \leq A(v) \qquad \text{for all } v \in V_{p,g_D}. \tag{2.4}$$

(ii) Furthermore, $u_p \in V_{p,g_D}$ solves (2.4) if and only if u_p solves the variational equation

$$DA(u_p; v - u_p) = 0 \qquad \text{for all } v \in V_{p,g_D}. \tag{2.5}$$

Theorem 2.7. Let the functional A be defined by (1.8).
(i) There exists a unique $u_p \in K_{p,g_D}$ which minimizes the functional A on K_{p,g_D}, i.e.,

$$A(u_p) \leq A(v) \qquad \text{for all } v \in K_{p,g_D}. \tag{2.6}$$

(ii) Furthermore, $u_p \in K_{p,g_D}$ solves (2.6) if and only if u_p solves the variational inequality

$$DA(u_p; v - u_p) \geq 0 \qquad \text{for all } v \in K_{p,g_D}. \tag{2.7}$$

Proof of Theorem 2.6 and Theorem 2.7. Analogously to the proof of Theorem 1.23 we obtain the existence of a minimum and its uniqueness since V_{p,g_D} and K_{p,g_D} are convex, closed, and nonempty subsets of the real reflexive Banach space V_p. The equivalence (ii) in Theorem 2.7 follows again with Theorem 1.13 and Corollary 1.14. The equivalence (ii) in Theorem 2.6 is yielded by noting that $v \in V_{p,g_D}$ implies $2u_p - v \in V_{p,g_D}$. Therefore, $DA(u_p; v - u_p) \geq 0$ implies $DA(u_p; v - u_p) \leq 0$. □

The convergence of the minimizer u_p of Theorem 2.7 towards the minimizer u of Theorem 1.23 is stated in the following theorem:

Theorem 2.8. Let $\psi \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ and $\psi \leq g_D$ in a neighborhood of Γ_D. With the above assumptions on K and K_{p,g_D}, there holds

$$\lim_{p\to\infty} \|u_p - u\|_{H^1(\Omega)} = 0$$

with u and u_p the minimizers of Theorem 1.23 and Theorem 2.7, respectively.

Proof. The form $D^2A(\hat{u}; \cdot, \cdot)$ is positive definite on $H^1_{g_D}(\Omega)$ and V_p for all $\hat{u} \in H^1(\Omega)$ due to Lemma 1.21 (1.13) and the Poincaré inequality. Thus, it suffices with Theorem [Glo84, Theorem I.5.2] to prove the following two hypotheses:

H1 If $(v_p)_p$ is such that $v_p \in K_{p,g_D}$ for all p and converges weakly to v as $p \to \infty$, then $v \in K$.

H2 There exists a dense subset χ of K and a family of mappings $r_p : \chi \to K_p$ such that $\lim_{p\to\infty} r_p v = v$ strongly in $H^1(\Omega)$ for all $v \in \chi$.

H1 is shown in Lemma 2.9, **H2** in Lemma 2.10. □

Lemma 2.9. If the sequence $(v_p)_p$ with $v_p \in K_p$ converges weakly to v for $p \to \infty$, then $v \in K$.

Proof. Firstly, we introduce a two-dimensional Bernstein operator on the closed reference square $\overline{\tilde{Q}}$. For a function $f \in C([-1, 1])$, the formula

$$B_p f(x) := \sum_{k=0}^{p} \tbinom{p}{k} \left(\tfrac{x-1}{2}\right)^k \left(\tfrac{1-x}{2}\right)^{p-k} f\left(\tfrac{2k}{p} - 1\right) \tag{2.8}$$

produces a linear map $f \to B_p f$ of $C([-1, 1])$ into $\mathbb{P}_p([-1, 1])$. This is the Bernstein polynomial of f and it is known that we have the uniform convergence

$$\lim_{p\to\infty} \|B_p f - f\|_{\infty,[-1,1]} = 0$$

(cf. [DL93, Chapter 1, Theorem 2.3]). For functions $f \in C(\overline{\tilde{Q}})$ on the reference square we write $B_p^{(i)}$, $i = 1, 2$, when the operator B_p is applied to the x_i variable and define the two-dimensional Bernstein operator by

$$B_p^2 f(x_1, x_2) := (B_p^{(1)} \circ B_p^{(2)} f)(x_1, x_2)$$

which maps f into $\mathbb{P}_p^2(\overline{\tilde{Q}})$ linearly. Here we note that the operators $B_p^{(1)}$ and $B_p^{(2)}$ commute. The uniform convergence

$$\lim_{p\to\infty} \|B_p^2 f - f\|_{\infty,\overline{\tilde{Q}}} = 0$$

follows due to $\|f - B_p^2 f\|_{\infty,\overline{\tilde{Q}}} \le \|f - B_p^{(1)} f\|_{\infty,\overline{\tilde{Q}}} + \|B_p^{(1)}(f - B_p^{(2)} f)\|_{\infty,\overline{\tilde{Q}}}$.

Now, we consider $\phi \in C^0(\overline{\Omega})$ and define its approximation ϕ_p by a combination of Bernstein polynomials on $\overline{Q}$, $Q \in \mathcal{T}$, i.e.,

$$B_{Q,p}^2 \phi(x) := B_p^2\Big(\phi \circ F_Q\Big)(F_Q^{-1}(x)) \quad \text{for } x \in \overline{Q}$$

and ϕ_p is given by $\phi_{p\,|Q} := B_{Q,p}^2 \phi$ for all $Q \in \mathcal{T}$.

It follows that

$$\phi_p \in V_p \qquad \text{and} \qquad \lim_{p\to\infty} \|\phi_p - \phi\|_\Omega = 0. \tag{2.9}$$

Further, if $\phi \ge 0$, we have $\phi_p \ge 0$ because the Bernstein operators $B_{Q,p}^2$ are monotone (cf. (2.8)).

Secondly, we define the interpolate $\psi_p := i_p \psi$. By Theorem 2.3 we know

$$\lim_{p\to\infty} \|\psi - \psi_p\|_{L^2(\Omega)} = 0\,.$$

Now, let the sequence $(v_p)_{p\in\mathbb{N}}$, $v_p \in K_p$, converge weakly to v in $H^1(\Omega)$ and let $\phi \ge 0$. Using the Gauss-Lobatto quadrature formula (2.3) and $\phi \in L^\infty(\overline{Q}) = (L^1(\overline{Q}))'$, we obtain that

$$\lim_{p\to\infty} \int_Q (v_p - \psi_p)\phi_{p-1}\,\mathrm{d}x = \int_Q (v - \psi)\phi\,\mathrm{d}x$$

due to the Lebesgue dominated convergence theorem. With Rellich's embedding theorem (cf. [Alt91, A 5.1]) there follows

$$\lim_{p\to\infty} v_p = v \quad \text{strongly in } L^2(\Omega)\,.$$

Thus, it is suffices to show that $v \ge \psi$ almost everywhere. With (2.3) and the definition of ψ_p we get for all $Q \in \mathcal{T}$

$$\begin{aligned}\int_Q (v_p - \psi_p)\phi_{p-1}\,\mathrm{d}x &\\ = \sum_{i=0}^{p}\sum_{j=0}^{p} \rho_i^{p+1}\rho_j^{p+1} &\left((v_p - \psi_p)\phi_{p-1}\right)\big(F_Q(\xi_i^{p+1}, \xi_j^{p+1})\big)\ |\det DF_Q(\xi_i^{p+1}, \xi_j^{p+1})| \\ &\ge 0. \end{aligned} \tag{2.10}$$

The inequality follows since $\phi_{p-1}(x) \ge 0$ for all $x \in Q$ and $(v_p - \psi_p)(x) \ge 0$ for all $x \in G_{Q,p}$ due to the definition of K_{p,g_D}. Furthermore, it is known that the weights ρ_i^{p+1}, $0 \le i \le p$, of the Gauss-Lobatto quadrature formula are positive.

Combining (2.9) and (2.10) we obtain that for all $\phi \in C^0(\overline{Q})$ with $\phi \geq 0$

$$\int_Q (v - \psi)\phi \, \mathrm{d}x \geq 0 \quad \text{for all } Q \in \mathcal{T},$$

hence $v \geq \psi$ almost everywhere on Ω, i.e., $v \in K$. □

Lemma 2.10. For ψ as in Theorem 2.8 there exists a dense subset χ of K and a sequence of mappings $r_p : \chi \to K_{p,g_D}$ such that $\lim_{p\to\infty} r_p v = v$ strongly in $H^1(\Omega)$ for all $v \in \chi$.

Proof. Consider $\chi := C^\infty(\Omega) \cap K$ and $r_p : H^1(\Omega) \cap C^0(\overline{\Omega}) \to V_p$ defined by

$$r_p v := i_p v \tag{2.11}$$

With Theorem 2.3 there exists a constant C independent of v and p such that

$$\|r_p v - v\|_{H^1(\Omega)} \leq C p^{-1} \|v\|_{H^2(\Omega)} \quad \text{for all } v \in H^2(\Omega)$$

and thus for all $v \in \chi$. With (2.11) it is obvious that $r_p v \in K_{p,g_D}$ for all $v \in \chi$. Thus the assertion of the lemma is fulfilled if $\overline{\chi} = K$. This follows with [Glo84, Lemma II.2.4] if $\psi \leq g_D$ in a neighborhood of Γ_D. □

With Theorem 2.8 the convergence of the p-version is proved. If we assume higher regularity of the solution u and of the obstacle ψ, i.e., $u, \psi \in H^2(Q)$, we obtain the following a priori error estimate which yields a convergence rate. This assumption of higher regularity of u and ψ is quite natural due to Theorem 1.25.

Note, in general, $V_{p,g_D} \not\subset H^1_{g_D}(\Omega)$ and $K_p \not\subset K$. This nonconformity of the approximation subset requests an extension of the analysis for the h-version given in [HHNL88] and [Fal74] for a Laplacian inequality. Particularly, we use the non-negativity of $\mathcal{P}(u)$ on the coincidence set Ψ and $\mathcal{P}_{\Gamma_N}(u)$ on Γ_N (cf. (1.18), (1.19)), the partition of Ψ into $\Psi \cap \Upsilon_p$ and $\Psi \setminus \Upsilon_p$ where

$$\Upsilon_p := \{x \in \Omega | u_p(x) \leq i_p \psi(x)\}, \tag{2.12}$$

and the set of functions

$$U_p^r := \{w \in H^r(\Omega) \,|\, w = u_p \text{ a.e. in } \Upsilon_p \text{ and } w(x) = \psi(x) \text{ for all } x \in G_p\} \tag{2.13}$$

for $r \geq 1$.

Theorem 2.11. Let u and u_p be the solutions of Theorem 1.23 and Theorem 2.7, respectively. Furthermore, suppose $u, \psi \in H^2(\Omega)$, $f \in L^2(\Omega)$, and $g_N \in H^{1/2}(\Gamma_N)$. Let Υ_p and U_p^r be defined by (2.12) and (2.13). Furthermore, let $\mathcal{P}(u)$ and $\mathcal{P}_{\Gamma_N}(u)$ be given by (1.18) and (1.19).

Then there exist constants $C_1, C_2(r) > 0$, independent of u, ψ, $\bar{u}_p$, and p, $C_2(r)$ only depending on $r \geq 1$ such that

$$\begin{aligned} \|u - u_p\|_{H^1(\Omega)} \;\leq\; & C_1\big(\|u\|_{H^2(\Omega)} + \|\mathcal{P}(u)\|_{L^2(\Psi)} + \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N \setminus \partial(\Omega\setminus\Psi))}\big)^{1/2} \\ & \cdot \big(\|u\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big)^{1/2} p^{-3/4} \\ & + C_2(r)\big(\|\mathcal{P}(u)\|_{L^2(\Psi)} + \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N \setminus \partial(\Omega\setminus\Psi))}\big)^{1/2} \|\bar{u}_p\|^{1/2}_{H^r(\Omega)} p^{-r/2+1/4}. \end{aligned} \tag{2.14}$$

For $u > \psi$ on Γ_N, there holds the improved result

$$\begin{aligned}\|u-u_p\|_{H^1(\Omega)} \;\le\; C_1\big(\|u\|_{H^2(\Omega)} + \|\mathcal{P}(u)\|_{L^2(\Psi)}\big)^{1/2}\,\big(\|u\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big)^{1/2} p^{-1} \\ + C_2(r)\|\mathcal{P}(u)\|^{1/2}_{L^2(\Psi)}\,\|\bar{u}_p\|^{1/2}_{H^r(\Omega)}\,p^{-r/2}.\end{aligned} \tag{2.15}$$

Remark 2.12. With [KS80, Theorem II.A.1] we know that $\bar{u}_p := \min\{u_p, i_p\psi\} \in H^1(\Omega)$ with $\|\bar{u}_p\|_{H^1(\Omega)} \le \|u_p\|_{H^1(\Omega)} + \|i_p\psi\|_{H^1(\Omega)}$, i.e., $\bar{u}_p \in U_p^r$ for $r = 1$. Further, $\bar{u}_p$ is bounded in $H^1(\Omega)$ for $p \to \infty$ due to the convergence of u_p towards u and the convergence of $i_p\psi$ towards ψ, i.e., there exists a real constant $c > 0$, such that $\|\bar{u}_p\|_{H^1(\Omega)} \le c(\|u\|_{H^1(\Omega)} + \|\psi\|_{H^1(\Omega)})$ for all $p \in \mathbb{N}$.

$\bar{u}_p$ can be constructed for $r > 1$ and finite p by continuous extension of $u_{p|\Upsilon_p}$ and its derivatives using Hermite interpolation polynomials in $x \in G_p$.

Our numerical experiments with nonempty coincidence sets and a Dirichlet boundary condition suggest convergence rates better than $\mathcal{O}(p^{-1})$, i.e., $\|\bar{u}_p\|_{H^r(\Omega)}$ is small for $p \le 20$ (see Experiment 2.25, Experiment 4.50, Experiment 4.51).

Proof of Theorem 2.11. Due to Theorem 1.23(ii) and Theorem 2.7(ii) we have

$$DA(u;u) \le DA(u;v) \;\text{ for all } v \in K; \quad DA(u_p;u_p) \le DA(u_p;v_p) \;\text{ for all } v_p \in K_{p,g_D}. \tag{2.16}$$

Setting $\varphi_w(t) := DA(u+tw;v)$, $t \in \mathbb{R}$, with Taylor's theorem we write

$$\begin{aligned}\varphi_w(1) &= \varphi_w(0) + \varphi_w'(\theta) \qquad \text{for a } \theta \in [0,1] \\ &= DA(u;v) + D^2A(u+\theta w; v, w)\end{aligned}$$

and get

$$DA(u_p;v) = DA(u;v) + D^2A(u+\theta(u_p-u); v, u_p-u) \quad \text{for a } \theta \in [0,1]$$

by setting $w := u_p - u$. Using Lemma 1.21 (1.12) we deduce

$$\begin{aligned}\kappa_l|u-u_p|^2_{H^1(\Omega)} &\le DA(u;u-u_p) - DA(u_p;u-u_p) \\ &= DA(u;u) + DA(u_p;u_p) - DA(u;u_p) - DA(u_p;u) \\ &\le DA(u;v-u_p) + DA(u_p;v_p-u) \\ &\le DA(u;v-u_p) + DA(u;v_p-u) \\ &\qquad + \max_{\theta\in[0,1]}\{D^2A(u+\theta(u_p-u); v_p-u, u_p-u)\}\end{aligned} \tag{2.17}$$

for all $v \in K$ and $v_p \in K_{p,g_D}$. We estimate *the last three terms of* (2.17) as follows:

$DA(u;v-u_p)$: Using the notations $\mathcal{P}(u)$ defined by (1.18), $\mathcal{P}_{\Gamma_N}(u)$ defined by (1.19) we observe $\mathcal{P}(u) \in L^2(\Omega)$, $\mathcal{P}_{\Gamma_N}(u) \in H^{1/2}(\Gamma_N)$. From Theorem 1.23 we know

$$\begin{aligned}\mathcal{P}(u) \;&\ge 0 \text{ on } \Omega \qquad \text{and} \quad \mathcal{P}(u)_{|\Omega\setminus\Psi} = 0, \\ \mathcal{P}_{\Gamma_N}(u) &\ge 0 \text{ on } \Gamma_N \qquad \text{and} \quad \mathcal{P}_{\Gamma_N}(u)_{|\Gamma_N\cap\partial(\Omega\setminus\Psi)} = 0.\end{aligned}$$

We recall the notation of the coincidence set $\Psi := \{x \in \Omega \,|\, u(x) = \psi(x)\}$ (see (1.22)) and define the consistency error set

$$\Omega_p^- := \{x \in \Omega \,|\, u_p(x) \le \psi(x)\}.$$

Now, let $v := \max\{u_p, \psi\}$. From [KS80, Chapter II, Theorem II.A.1] we know that $v \in H^1(\Omega)$. Therefore, $v \in K$. By partial integration of DA (cf. 1.17) we obtain

$$\begin{aligned} DA(u; v - u_p) &= \langle \mathcal{P}(u), v - u_p \rangle_{L^2(\Omega)} + \langle \mathcal{P}_{\Gamma_N}(u), v - u_p \rangle_{H^{-1/2}(\Gamma_N)} \\ &= \langle \mathcal{P}(u), \psi - u_p \rangle_{L^2(\Omega_p^- \cap \Psi)} + \langle \mathcal{P}_{\Gamma_N}(u), \psi - u_p \rangle_{H^{-1/2}(\Gamma_{N,p}^-)} \end{aligned} \tag{2.19}$$

with $\Gamma_{N,p}^- := \Gamma_N \setminus \partial(\Omega \setminus \Psi) \cap \partial\Omega_p^-$. Here, we use

$$0 \le v - u_p = \begin{cases} \psi - u_p & \text{a.e. on } \Omega_p^- , \\ 0 & \text{a.e. on } \Omega \setminus \Omega_p^- , \end{cases}$$

and the non-negativity of $\mathcal{P}(u)$ on Ψ, the non-negativity of $\mathcal{P}_{\Gamma_N}(u)$ on $\Gamma_N \setminus \partial(\Omega \setminus \Psi)$ to get (2.19). To estimate the right hand side of (2.19), we must cope with the consistency error $K_p \not\subset K$, i.e., there exist $x \in \Psi$ such that $u_p(x) < \psi(x)$. We know $u_p \ge i_p\psi$ in $\Omega \setminus \Upsilon_p$ by the definition of Υ_p (see (2.12)). This yields

$$\begin{aligned} \langle \mathcal{P}(u), \psi - u_p \rangle_{L^2(\Omega_p^- \cap \Psi)} \le{}& \langle \mathcal{P}(u), \psi - i_p\psi \rangle_{L^2((\Omega_p^- \cap \Psi) \setminus \Upsilon_p)} \\ &+ \langle \mathcal{P}(u), \psi - \bar{u}_p \rangle_{L^2(\Omega_p^- \cap \Psi \cap \Upsilon_p)} + \langle \mathcal{P}(u), \bar{u}_p - u_p \rangle_{L^2(\Omega_p^- \cap \Psi \cap \Upsilon_p)} \end{aligned}$$

with $\bar{u}_p \in U_p^r$. Due to the definition of U_p^r in (2.13) there holds $\bar{u}_p - u_p = 0$ a.e. in Υ_p. Additionally, we have $i_p\psi = i_p\bar{u}_p$ because of $\bar{u}_p(x) = \psi(x)$ for all $x \in G_p$ from the definition of U_p^r. Extending $\psi - \bar{u}_p = \psi - i_p\psi + i_p\bar{u}_p - \bar{u}_p$ yields

$$\langle \mathcal{P}(u), \psi - u_p \rangle_{L^2(\Omega_p^- \cap \Psi)} \le \langle \mathcal{P}(u), \psi - i_p\psi \rangle_{L^2(\Omega_p^- \cap \Psi)} + \langle \mathcal{P}(u), i_p\bar{u}_p - \bar{u}_p \rangle_{L^2(\Omega_p^- \cap \Psi \cap \Upsilon_p)} .$$

Using the corresponding estimate for $\langle \mathcal{P}_{\Gamma_N}(u), \psi - u_p \rangle_{H^{-1/2}(\Gamma_{N,p}^-)}$, we can write

$$\begin{aligned} DA(u; v - u_p) \le{}& \|\mathcal{P}(u)\|_{L^2(\Omega_p^- \cap \Psi)} \Big(\|\psi - i_p\psi\|_{L^2(\Omega)} + \|\bar{u}_p - i_p\bar{u}_p\|_{L^2(\Omega)} \Big) \\ &+ \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_{N,p}^-)} \Big(\|\psi - i_p\psi\|_{H^{-1/2}(\Gamma_N)} + \|\bar{u}_p - i_p\bar{u}_p\|_{H^{-1/2}(\Gamma_N)} \Big) \\ \le{}& \|\mathcal{P}(u)\|_{L^2(\Omega_p^- \cap \Psi)} \big(C_3 p^{-2} \|\psi\|_{H^2(\Omega)} + C_5(r)\, p^{-r} \|\bar{u}_p\|_{H^r(\Omega)} \big) \\ &+ \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_{N,p}^-)} \big(C_4 p^{-3/2} \|\psi\|_{H^{3/2}(\Gamma_N)} + C_6(r)\, p^{-r+1/2} \|\bar{u}_p\|_{H^{r-1/2}(\Gamma_N)} \big) . \end{aligned} \tag{2.20}$$

Here, C_3, C_4, $C_5(r)$ and $C_6(r)$ are the positive constants from Theorem 2.3 depending on the quasi-uniform mesh $\mathcal{T}$ and on r, $1 \le r < \infty$. Note, when estimating on Γ_N we only get $p^{-3/2}$, $p^{-1/2}$ (and not p^{-2}, p^{-2}) due to the restriction $r \ge 0$ in Theorem 2.3.

$DA(u; v_p - u)$: Let $v_p := i_p\, u \in K_{p,g_D}$ be the interpolate of u. Again, partial integration of DA, duality, and using the above constants C_3, C_4 from Theorem 2.3 yield

$$\begin{aligned} DA(u; v_p - u) &= \langle \mathcal{P}(u), v_p - u \rangle_{L^2(\Psi)} + \langle \mathcal{P}_{\Gamma_N}(u), v_p - u \rangle_{H^{-1/2}(\Gamma_N \setminus \partial(\Omega \setminus \Psi))} \\ &\le \|\mathcal{P}(u)\|_{L^2(\Psi)} \|u - i_p u\|_{L^2(\Omega)} + \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N \setminus \partial(\Omega \setminus \Psi))} \|u - i_p u\|_{H^{-1/2}(\Gamma_N)} \\ &\le \|\mathcal{P}(u)\|_{L^2(\Psi)}\, C_3 p^{-2} \|u\|_{H^2(\Omega)} + \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N \setminus \partial(\Omega \setminus \Psi))}\, C_4 p^{-3/2} \|u\|_{H^{3/2}(\Gamma_N)} . \end{aligned} \tag{2.21}$$

$D^2A(u+\theta(u_p-u);v_p-u,u_p-u)$: From the proof of Lemma 1.21 (1.14) we have

$$D^2A(\hat{u};v_1,v_2) \le \kappa_u \langle \nabla v_1, \nabla v_2\rangle_{L^2(\Omega)} \quad \text{for all } \hat{u}, v_1, v_2 \in H^1(\Omega).$$

Using

$$\langle \nabla v_1, \nabla v_2\rangle_{L^2(\Omega)} \le \tfrac{1}{2\mu}|v_1|^2_{H^1(\Omega)} + \tfrac{\mu}{2}|v_2|^2_{H^1(\Omega)} \quad \text{for all } \mu > 0$$

and setting $\hat{u} := u+\theta(u_p-u)$, $\mu := \kappa_l\kappa_u^{-1}$, and again, $v_p := i_p\, u \in K_{p,g_D}$ we estimate

$$\begin{aligned} D^2A(\hat{u};v_p-u,u_p-u) &\le \kappa_u \tfrac{\kappa_u}{2\kappa_l}|v_p-u|^2_{H^1(\Omega)} + \kappa_u \tfrac{\kappa_l}{2\kappa_u}|u_p-u|^2_{H^1(\Omega)} \\ &\le \tfrac{\kappa_u^2}{2\kappa_l}\left(C_3p^{-1}\right)^2\|u\|^2_{H^2(\Omega)} + \tfrac{\kappa_l}{2}|u_p-u|^2_{H^1(\Omega)} \end{aligned} \tag{2.22}$$

for all $0 \le \theta \le 1$ with Theorem 2.3.

Plugging in (2.20), (2.21), (2.22) into (2.17) we get

$$\begin{aligned} \tfrac{\kappa_l}{2}|u-u_p|^2_{H^1(\Omega)} \le\ & C_3^2\tfrac{\kappa_u^2}{2\kappa_l}p^{-2}\|u\|^2_{H^2(\Omega)} \\ &+ \|\mathcal{P}(u)\|_{L^2(\Psi)}\Big(C_3p^{-2}\big(\|u\|_{H^2(\Omega)} + \|\psi\|_{H^2(\Omega)}\big) + C_5(r)\,p^{-r}\|\bar{u}_p\|_{H^r(\Omega)}\Big) \\ &+ \|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N\setminus\partial(\Omega\setminus\Psi))}\Big(C_4p^{-3/2}\big(\|u\|_{H^{3/2}(\Gamma_N)} + \|\psi\|_{H^{3/2}(\Gamma_N)}\big) \\ &\qquad + C_6(r)\,p^{-r+1/2}\|\bar{u}_p\|_{H^{r-1/2}(\Gamma_N)}\Big). \end{aligned} \tag{2.23}$$

Since the trace operator onto Γ_N is continuous, there exists positive constants C_7, C_8 such that

$$\begin{aligned} \|v\|_{H^{3/2}(\Gamma_N)} &\le C_7\|v\|_{H^2(\Omega)} && \text{for all } v \in H^2(\Omega), \\ \|v\|_{H^{r-1/2}(\Gamma_N)} &\le C_8\|v\|_{H^r(\Omega)} && \text{for all } v \in H^r(\Omega),\ 1 \le r < \infty\,. \end{aligned}$$

Let $C_{\mathrm{P}} > 0$ be the constant of the Poincaré-Friedrich's inequality $\|v\|_{H^1(\Omega)} \le C_{\mathrm{P}}|v|_{H^1(\Omega)}$ which holds for all $v \in H^1_{g_D}(\Omega)$, $g_D \equiv 0$. As $i_pu = u_p$ on Γ_D, we can use the Poincaré-Friedrich's inequality and Theorem 2.3 to estimate

$$\begin{aligned} \|u-u_p\|^2_{L^2(\Omega)} &\le 2\|u-i_pu\|^2_{L^2(\Omega)} + 2\|i_pu-u_p\|^2_{L^2(\Omega)} \\ &\le 2C_3^2p^{-2}\|u\|^2_{H^2(\Omega)} + 2C_{\mathrm{P}}^2|i_pu-u_p|^2_{H^1(\Omega)} \\ &\le 2(1+C_{\mathrm{P}}^2)C_3^2p^{-2}\|u\|^2_{H^2(\Omega)} + 2C_{\mathrm{P}}^2|u-u_p|^2_{H^1(\Omega)}. \end{aligned}$$

Setting $C_{\mathrm{Q}} := (1+C_{\mathrm{P}}^2)\frac{4}{\kappa_l}$, $C_2^2(r) := C_{\mathrm{Q}}\max\{C_5(r), C_6(r)C_8\}$, and

$$C_1^2 := C_{\mathrm{Q}}\max\{C_3^2(\tfrac{\kappa_l}{2} + \tfrac{\kappa_u^2}{2\kappa_l}), C_3, C_4C_7\}$$

we obtain (2.14).

If $u > \psi$ on Γ_N, the Neumann condition (1.19) of Problem (P) holds with $\mathcal{P}_{\Gamma_N}(u) = 0$ on Γ_N (cf. (1.20)). Thus $\|\mathcal{P}_{\Gamma_N}(u)\|_{H^{1/2}(\Gamma_N\setminus\partial(\Omega\setminus\Psi))} = 0$ in (2.23) yields (2.15). □

Remark 2.13. Falk [Fal74] and Hlaváček et al. [HHNL88] both deduce a convergence rate of $\|u-u_h\|_{H^1(\Omega)} = \mathcal{O}(h)$ for the h-version for an obstacle problem with Dirichlet boundary conditions and the regularity assumptions of Theorem 2.11. Due to the piecewise affine linearity of h-version FEM functions, they can use the property $u_h \ge i_h\psi$. Here u_h denotes the h-version solution and $i_h\psi$ the linear interpolant of the obstacle. Unfortunately, the corresponding p-version analog $u_p \ge i_p\psi$ does not hold.

2.2 p discretization of $K \subset H^1(\Omega)$ on triangles

The use of high-order methods is traditionally in conflict with the need for significant geometric flexibility by being restricted to fairly simple geometries. The standard approach is to map the reference quadrilateral $\tilde{Q}$ onto curvilinear quadrilaterals in the sense of FE. Such techniques are powerful, but they do suffer from the need to tile the computational domain using only quadrilaterals. Unfortunately, automated grid generation using only such elements for general two-dimensional computational problems of a realistic complexity remains a nontrivial task and it becomes even worse in three dimensions. In contrast to this, automated grid generation employing a fully unstructured grid based on triangle elements, or tetrahedral elements in three dimension is significantly more mature, due mainly to extensive developments within the FE community.

2.2.1 An electrostatic approach

Typically, the high-order methods on triangles use the $\frac{1}{2}(p+1)(p+2)$ dimensional space $V_p(\tilde{T}) := \{\xi_1^i \xi_2^j \,|\, 0 \le i+j \le p\}$ on the reference triangle $\tilde{T}$ given by the vertices $(0,0)$, $(1,0)$, $(0,1)$. From [BCMP91] and [Dub91] we know hierarchical bases for $V_p(\tilde{T})$ deduced mainly from Legendre polynomials which lead to a mildly increasing condition number of the global stiffness matrix for growing p.

For higher-order collocation methods the question of how to choose $\frac{1}{2}(p+1)(p+2)$ good collocation points on $\hat{T}$ given by the vertices $(-1/2, 0)$, $(1/2, 0)$, and $(0, \sqrt{3}/2)$ was analyzed numerically by Hesthaven in [Hes98] for $p \le 16$ using different nodal schemes. The way of estimating the quality of alternative approximations corresponding to the nodal sets appears as a result of the following generalization of Lebesgue's lemma to $\hat{T}$ (cf. [DL93]).

Lemma 2.14 (*Lebesgue*). Let $\|g\|_\infty := \sup\{|g(x)| \,|\, x \in \hat{T}\}$ be the usual supremum-norm on $C(\hat{T})$. Assume that $f \in C(\hat{T})$ and that we consider the nodal set $G_{p,\hat{T}}$ and the interpolant $i_{G_{p,\hat{T}}} f$ of f with respect to $G_{p,\hat{T}}$. Then,

$$\|f - i_{G_{p,\hat{T}}} f\|_\infty \le \big(1 + \Lambda(G_{p,\hat{T}})\big) \min_{p \in V_p(\hat{T})} \|f - p\|_\infty$$

where

$$\Lambda(G_{p,\hat{T}}) := \sup_{\substack{f \in C(\hat{T}) \\ 0 \neq f}} \frac{\|i_{G_{p,\hat{T}}} f\|_\infty}{\|f\|_\infty} = \max_{x \in \hat{T}} \sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} |\lambda_i(x)| \tag{2.24}$$

is called the Lebesgue constant and $\lambda_i(x)$ denote the Lagrangian interpolation polynomials on $\hat{T}$, i.e., for the counted nodes $x_j \in G_{p,\hat{T}}$ we have $\lambda_i(x_j) = 1$ for $i = j$, and $\lambda_i(x_j) = 0$ for $i \neq j$.

Proof. As $i_{G_{p,\hat{T}}} p = p$ for all $p \in V_p(\hat{T})$, we obtain

$$\|f - i_{G_{p,\hat{T}}} f\|_\infty \le \|f - p\|_\infty + \|i_{G_{p,\hat{T}}}(p - f)\|_\infty \le (1 + \Lambda(G_{p,\hat{T}}))\|f - p\|_\infty \,.$$

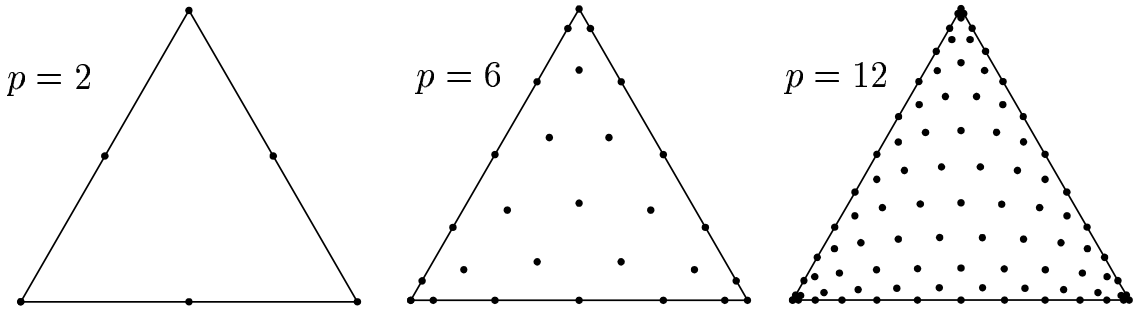

Figure 2.1: Chebyshev-Gauss-Lobatto points on $\hat{T}$ for $p = 2, 6, 12$

To proof the right equality in (2.24) we note firstly

$$|i_{G_{p,\hat{T}}} f(x)| \leq \sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} |f(x_i)\lambda_i(x)| \leq \|f\|_\infty \sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} |\lambda_i(x)|$$

for all $x \in \hat{T}$, $f \in C(\hat{T})$. This yields "$\leq$" of the right "$=$" in (2.24). Secondly, let $\hat{x} \in \hat{T}$ such that $\sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} |\lambda_i(\hat{x})|$ is maximal. There exist $s_i \in \{-1, 1\}$ such that $\sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} |\lambda_i(\hat{x})| = \sum_{i=1}^{\operatorname{card} G_{p,\hat{T}}} s_i\, \lambda_i(\hat{x})$. Further, let $\mathcal{S}$ be a triangulation on $\hat{T}$ with the interpolation points as vertices. Thereby, there exists a continuous and piecewise linear g on $\hat{T}$ with $s_i = g(x_i)$ for $i = 1, \ldots, \operatorname{card} G_{p,\hat{T}}$. Noting that $\|g\|_\infty = 1$ yields "$\geq$" of the right "$=$" in (2.24). □

Consequently, by computing the Lebesgue constant we obtain a measure of how close the approximation is to the best polynomial approximation.

In 1885, Stieltjes [Sti85a, Sti85b] revealed a very interesting connection between the Gauss-Lobatto points and the following electrostatic problem.

> *Problem:* Let two unit electrostatic mass charges $q_l, q_r > 0$ be concentrated at the positions $x = \pm 1$. Assume also that N_I unit charges, positioned at $x_1, \ldots, x_{N_I}$, are allowed to move freely along the line connecting the end point charges. What is the position of the charges that minimizes the electrostatic energy
>
> $$W(x_1, \ldots, x_{N_I}) := -\sum_{i=1}^{N_I} \left(q_l \log|x_i + 1| + q_r \log|x_i - 1| + \tfrac{1}{2} \sum_{\substack{j=1 \\ j \neq i}}^{N_I} \log|x_i + x_j| \right).$$

Stieltjes showed that the energy is minimized when $x_1, \ldots, x_{N_I}$ are given as the zeros of the Jacobi polynomials $J_{N_I}^{\alpha,\beta}$ with $\alpha = 2q_l - 1$ and $\beta = 2q_r - 1$ ($J_{N_I}^{\alpha,\beta}$ is given below, p. 46). In [Sze75, Theorem 6.7.1] it is shown that this minimum is the unique global minimum. Taking $\alpha = \beta = 1$ and $p = N_I + 1$ the zeros of the Jacobi polynomials $J_{N_I}^{\alpha,\beta}$ coincide with the interior Gauss-Lobatto points ξ_i^{p+1}, $1 \leq i \leq p - 1$.

Hesthaven [Hes98] generalizes this problem to $\frac{1}{2}(p + 1)(p + 2)$ unit mass charges on $\hat{T}$ and calculates their distribution $G_{p,\mathrm{CGJ}}$, termed *Chebyshev-Gauss-Lobatto nodes*, numerically (see Figure 2.1) for $p = 1, 2, \ldots, 16$. The respective interpolation operator owns

the Lebesgue constant $\Lambda(G_{p,\mathrm{CGJ}}) = 41.726$ for $p = 16$ which is a great improvement in comparison to the approximation properties of interpolation with respect to the equinodal set $G_{p,\mathrm{EQ}}$ characterized by the Lebesgue constant $\Lambda(G_{p,\mathrm{EQ}}) = 2418.5$ for $p = 16$. Furthermore, the $G_{p,\mathrm{CGJ}}$ nodes have the nice property that they coincide with the Gauss-Lobatto points on the edges of $\hat{T}$.

Recalling the definitions of V_{p,g_D} and K_{p,g_D} on quadrilateral meshes, it seems a nearby idea to take the Lagrangian interpolation polynomials λ_i with respect to the $G_{p,\mathrm{CGJ}}$ points as a basis and to control the obstacle condition in these points. The Lagrangian basis can be easily calculated by evaluating a known basis $(b_k)_{k=1,\dots,N_T}$, $N_T := \dim V_p(\hat{T})$, of $V_p(\hat{T})$ at all $x_i \in G_{p,\mathrm{CGJ}}$:

$$\begin{pmatrix} \lambda_1(x) \\ \vdots \\ \lambda_{N_T}(x) \end{pmatrix} = \begin{pmatrix} b_1(x_1) & \dots & b_1(x_{N_T}) \\ \vdots & & \vdots \\ b_{N_T}(x_1) & \dots & b_{N_T}(x_{N_T}) \end{pmatrix}^{-1} \begin{pmatrix} b_1(x) \\ \vdots \\ b_{N_T}(x) \end{pmatrix} .$$

The main drawback of the $G_{p,\mathrm{CGJ}}$ approach is that almost nothing is known from the theoretical point of view. This concerns particularly the numerical quadrature on $\hat{T}$. The proof of our Theorem 2.8 hangs mainly on the exactness of the Gauss-Lobatto quadrature rule for polynomials of degree $\leq 2p-1$ and the positivity of its weights ρ_i^{p+1}, $0 \leq i \leq p$. Although a quadrature rule with the $G_{p,\mathrm{CGJ}}$ points as quadrature points can be developed by integration of the respective Lagrangian interpolation polynomials, its exactness for polynomials of degree $> p$ and the positivity of its weights remain open questions. In the following we pursue another approach.

2.2.2 An approach by weighted Sobolev spaces

Instead of the usual $\frac{1}{2}(p+1)(p+2)$ dimensional local space $V_p(\tilde{T})$, we use a nonlinear mapping to transform the tensor product basis from $\tilde{Q}$ onto $\tilde{T}$ and yield a $(p+1)^2 - p$ dimensional basis. Let $\tilde{T}$ be the interior of the reference triangle defined by the vertices $(0,0)$, $(0,1)$, and $(0,1)$. The transformation $F_\triangleright$ defined by

$$F_\triangleright(\xi) = \tfrac{1+\xi_1}{4}\,(1-\xi_2, 1+\xi_2) \tag{2.25}$$

maps the interior of $\tilde{Q}$ upon the interior of $\tilde{T}$ diffeomorphly. We calculate the Jacobian matrix

$$DF_\triangleright(\xi) = \tfrac{1}{4}\begin{pmatrix} 1-\xi_2 & -(1+\xi_1) \\ 1+\xi_2 & 1+\xi_1 \end{pmatrix} \qquad \text{and its determinant } |\det DF_\triangleright(\xi)| = \tfrac{1+\xi_1}{8}\,.$$

Thus, we have

$$\int_{\tilde{T}} v^2(x)\,\mathrm{d}x = \int_{\tilde{Q}} \tfrac{1+\xi_1}{8} v^2(F(\xi))\,\mathrm{d}\xi \tag{2.26}$$

for smooth v and the respective analogies for $\frac{\partial}{\partial x_1}v$, $\frac{\partial}{\partial x_2}v$, $\frac{\partial^2}{\partial x_1^2}v$, $\frac{\partial^2}{\partial x_1 x_2}v$, $\frac{\partial^2}{\partial x_2^2}v$.

Example 2.15. Let u be given by $u(x) = (1-x_1)^{-3/2}$ on $\tilde{T}$ and $\hat{u} = u \circ F_\triangleright$ on $\tilde{Q}$. Elementary integration shows that $u \in L^2(\tilde{T})$ and due to (2.26) $(\frac{1+\xi_1}{8})^{1/2}\,\hat{u} \in L^2(\tilde{Q})$. But $\hat{u} \notin L^2(\tilde{Q})$ because of $\|\hat{u}\|^2_{L^2(\tilde{Q})} = \int_{\tilde{Q}} (1 - \frac{1}{4}(1+\xi_1)(1-\xi_2))^{-3}\,\mathrm{d}\xi = \infty$. So, the convergence result given in Theorem 2.3 can not be used.

As a work around for this problem we switch to weighted Sobolev spaces. We begin with the one-dimensional case. The weight on the interval $[-1,1]$ is defined by

$$w_{\alpha,\beta}(\zeta) := (1-\zeta)^\alpha\,(1+\zeta)^\beta \qquad \text{where } \alpha,\beta > -1$$

are real parameters. This last condition is necessary for the weight to be integrable. We define the basic space

$$L^2_{\alpha,\beta}((-1,1)) := \big\{v \in \mathcal{D}'(-1,1) \,\big|\, \int_{-1}^{1} w(\zeta)\,v^2(\zeta)\,\mathrm{d}\zeta < +\infty\big\}$$

which is provided with the norm

$$\|v\|_{L^2_{\alpha,\beta}((-1,1))} := \left(\int_{-1}^{1} w_{\alpha,\beta}(\zeta)\,v^2(\zeta)\,\mathrm{d}\zeta\right)^{1/2}.$$

Definition 2.16. Let m be a positive integer. The Sobolev space $H^m_{\alpha,\beta}((-1,1))$ is defined by

$$H^m_{\alpha,\beta}((-1,1)) = \big\{v \in L^2_{\alpha,\beta}((-1,1)) \,\big|\, \tfrac{\mathrm{d}^k v}{\mathrm{d}\zeta^k} \in L^2_{\alpha,\beta}((-1,1)),\ 1 \le k \le m\big\}.$$

It is provided with the norm

$$\|v\|_{H^m_{\alpha,\beta}((-1,1))} = \left(\int_{-1}^{1} \sum_{k=0}^{m} \big(\tfrac{\mathrm{d}^k v}{\mathrm{d}\zeta^k}\big)^2(\zeta)\; w_{\alpha,\beta}(\zeta)\;\mathrm{d}\zeta\right)^{1/2}.$$

In [BM92, BM97] Bernardi and Maday give a deep analysis of weighted Sobolev spaces with respect to weight functions given by $w_{\alpha,\alpha}(x) := (1-x^2)^\alpha$, $-\frac{1}{2} \le \alpha \le \frac{1}{2}$, defining the *ultraspherical Jacobi polynomials* $J_p^{\alpha,\alpha}$. There, also definitions of weighted Sobolev spaces $H^s_{\alpha,\alpha}((-1,1))$ for real numbers s yielded as interpolation spaces and extensions to higher dimensions can be found. We only give a sketch of the spaces and of the approximation properties of the Gauss-Lobatto-Jacobi interpolation operator i_p needed for our problem in the following.

By the theory of orthogonal polynomials (cf. [Sze75]) there exists polynomials $J_p^{\alpha,\beta}$, called *Jacobi polynomials*, of degree p for real α, β with $J_p^{\alpha,\beta}(1) = \binom{p+\alpha}{p}$, which are orthogonal to each other with respect to the scalar product $\langle\cdot,\cdot\rangle_{\alpha,\beta}$ given by $\langle f,g\rangle_{\alpha,\beta} := \int_{-1}^{1} w_{\alpha,\beta}(\zeta)\,f(\zeta)g(\zeta)\,\mathrm{d}\zeta$. Analogously to the Legendre polynomials on p. 35 we recall four basic properties that characterize these polynomials.

(i) For any positive integer p, the zeros of $J_p^{\alpha,\beta}$ are distinct real numbers $\zeta_i^{\alpha,\beta,p}$, $1 \le i \le p$ in $(-1,1)$, called *Gauss-Jacobi points of degree p.*

(ii) *Gauss-Jacobi quadrature.* There exist positive weight factors $\omega_i^{\alpha,\beta,p}$, $1 \le i \le p$, such that

$$\int_{-1}^{1} w_{\alpha,\beta}(\zeta)\,\phi(\zeta)\;\mathrm{d}\zeta = \sum_{i=1}^{p} \phi(\zeta_i^{\alpha,\beta,p})\,\omega_i^{\alpha,\beta,p}$$

for all polynomials $\phi \in \mathbb{P}_{2p-1}([-1,1])$.

(iii) For any positive integer p, the zeros of $(1-\zeta^2)\frac{\mathrm{d}}{\mathrm{d}\zeta}J_p^{\alpha,\beta}(\zeta)$ are distinct real numbers $\xi_i^{\alpha,\beta,p+1}$, $0 \le i \le p$, in $[-1,1]$ called *Gauss-Lobatto-Jacobi points of degree p.*

(iv) *Gauss-Lobatto-Jacobi quadrature.* There exist positive weight factors $\rho_i^{\alpha,\beta,p+1}$, $0 \le i \le p$, such that

$$\int_{-1}^{1} w_{\alpha,\beta}(\zeta)\,\phi(\zeta)\,\mathrm{d}\zeta = \sum_{i=0}^{p} \phi(\xi_i^{\alpha,\beta,p+1})\rho_i^{\alpha,\beta,p+1} \tag{2.27}$$

for all polynomials $\phi \in \mathbb{P}_{2p-1}([-1,1])$.

As a consequence of the differential equation satisfied by the Jacobi polynomials (see [Sze75, (4.2.2)])

$$\tfrac{\mathrm{d}}{\mathrm{d}\zeta}\Big((1-\zeta)^{\alpha+1}(1+\zeta)^{\beta+1}\tfrac{\mathrm{d}}{\mathrm{d}\zeta}J_p^{\alpha,\beta}(\zeta)\Big) + p(p+\alpha+\beta+1)(1-\zeta)^{\alpha}(1+\zeta)^{\beta}J_p^{\alpha,\beta}(\zeta) = 0,$$

we remark that the derivatives $\frac{\mathrm{d}}{\mathrm{d}\zeta}J_p^{\alpha,\beta}(\zeta)$ of the Jacobi polynomials are orthogonal with respect to the weight $(1-\zeta)^{\alpha+1}(1+\zeta)^{\beta+1}$. This leads to note that the interior nodes of a Gauss-Lobatto-Jacobi formula with $p+2$ nodes coincide with the nodes of a Gauss-Jacobi formula with p nodes, i.e.,

$$\xi_i^{\alpha,\beta,p+2} = \zeta_i^{\alpha+1,\beta+1,p}\,, \quad 1 \le i \le p\,; \tag{2.28}$$

besides the weights are linked by the following equality:

$$\rho_i^{\alpha,\beta,p+2} = \big(1-(\zeta_i^{\alpha+1,\beta+1,p})^2\big)^{-1}\,\omega_i^{\alpha+1,\beta+1,p}\,, \quad 1 \le i \le p\,. \tag{2.29}$$

In discussing the zeros of Jacobi polynomials, we define $\theta_i^{\alpha,\beta,p}$ by $\zeta_i^{\alpha,\beta,p} = \cos\theta_i^{\alpha,\beta,p}$ and enumerate the zeros in decreasing order:

$$+1 > \zeta_1^{\alpha,\beta,p} > \cdots > \zeta_p^{\alpha,\beta,p} > -1; \quad 0 < \theta_1^{\alpha,\beta,p} < \cdots < \theta_p^{\alpha,\beta,p} < \pi\,.$$

Assuming $\alpha, \beta > \frac{1}{2}$, an inspection of the proof of [Sze75, Theorem 6.3.1]) shows that

$$\frac{i}{p+(\alpha+\beta+1)/2}\pi < \theta_i < \frac{i+(\alpha+\beta-1)/2}{p+(\alpha+\beta+1)/2}\pi\,, \qquad i = 1,\ldots,p\,.$$

In particular, we obtain

$$\frac{i}{p+2}\pi < \arccos\xi_i^{0,1,p+2} < \frac{i+1}{p+2}\pi, \qquad i = 1,\ldots,p\,, \tag{2.30}$$

for the interior Gauss-Lobatto-Jacobi points by (2.28).

If two sequences a_n and b_n of real or complex numbers have the property that $b_n \neq 0$ and $a_n/b_n \to 1$ as $n \to \infty$, we write $\cong$.

Using the $\cong$-notation we can characterize the quadrature weights $\omega_i^{\alpha,\beta,p}$ of the Gauss-Jacobi quadrature by

$$\omega_i^{\alpha,\beta,p} \cong \frac{2^{\alpha+\beta+1}\pi}{p}\left(\sin\frac{\theta_i^{\alpha,\beta,p}}{2}\right)^{2\alpha+1}\left(\cos\frac{\theta_i^{\alpha,\beta,p}}{2}\right)^{2\beta+1}$$

due to [Sze75, (15.3.10)] for $\alpha, \beta > -1$. Using (2.29) and taking $\alpha = 0$, $\beta = 1$, we get

$$\rho_i^{0,1,p+2} \cong \frac{8\pi}{p}\left(\sin\frac{\theta_i^{1,2,p}}{2}\right)^2\left(\cos\frac{\theta_i^{1,2,p}}{2}\right)^4.$$

Thus, we can replace $(\sin\theta)^{\alpha+1/2}$ in the proof of [BM92, Theorem 3.1,(3.3)] by $(\sin\frac{\theta}{2})^1(\cos\frac{\theta}{2})^2$. Due to the partition (2.30) of the Gauss-Lobatto-Jacobi-points, collecting the above arguments should yield the following analogy of [BM92, Theorem 4.2].

Proposition 2.17. Let i_p be the interpolation operator with respect to the Gauss-Lobatto-Jacobi points $\xi_i^{0,1,p+1}$, $i = 0, 1, \ldots, p$. There exists a positive constant c depending on $s > 1/2$ such that the following approximation property holds for any function v on $H^s_{0,1}((-1,1))$:

$$\|v - i_p v\|_{L^2_{0,1}((-1,1))} \le cp^{-s}\|v\|_{H^s_{0,1}((-1,1))}.$$

For our problem we define a two-dimensional generalization using the weight function with respect to the variable ξ_1 as follows:

$$L^2_{\triangleright}(\tilde{Q}) := \left\{v \in \mathcal{D}'(\Omega) \,\middle|\, \int_{\tilde{Q}} w_{0,1}(\xi_1)\, v^2(\xi)\,\mathrm{d}\xi < +\infty\right\}$$

with the norm

$$\|v\|_{L^2_{\triangleright}(\tilde{Q})} := \left(\int_{\tilde{Q}} w_{0,1}(\xi_1)\, v^2(\xi)\,\mathrm{d}\xi\right)^{1/2}. \tag{2.31}$$

Remark 2.18. Let u and $\hat{u}$ be defined as in Example 2.15. With (2.26) it follows that $u \in L^2(\tilde{T})$ implies $\hat{u} \in L^2_{\triangleright}(\tilde{Q})$.

Using $L^2_{\triangleright}(\tilde{Q})$ as basic norm we can define appropriate Sobolev spaces.

Definition 2.19. Let m be a positive integer. The Sobolev space $H^m_{\triangleright}(\tilde{Q})$ is defined by

$$H^m_{\triangleright}(\tilde{Q}) = \left\{v \in L^2_{\triangleright}(\tilde{Q}) \,\middle|\, \tfrac{\partial^{k+l} v}{\partial\zeta_1^k \partial\zeta_2^l} \in L^2_{\triangleright}(\tilde{Q});\ 1 \le k+l \le m;\ 0 \le k,l\right\}.$$

It is provided with the norm

$$\|v\|_{H^m_{\triangleright}(\tilde{Q})} = \left(\int_{\tilde{Q}} \sum_{\substack{0\le k,l\\ 0\le k+l\le m}} \left(\tfrac{\partial^{k+l} v}{\partial\zeta_1^k\partial\zeta_2^l}\right)^2(\zeta)\; w_{0,1}(\zeta_1)\;\mathrm{d}\zeta\right)^{1/2}.$$

Analogously to Definition 2.2 of $G_{\tilde{Q},p}$, we use the Gauss-Lobatto-Jacobi points for the definition of points sets, and following, this for the definition of a p basis on the reference triangle $\tilde{T}$.

Definition 2.20. Let $\xi_i^{0,1,p+1}$ and ξ_j^{p+1}, $0 \le i,j \le p$, be the Gauss-Lobatto-Jacobi and Gauss-Lobatto points of degree p, respectively. On the reference elements $\tilde{Q} = [-1,1]^2$ and $\tilde{T}$ we define the sets

$$G_{0,1,\tilde{Q},p} := \{(\xi_i^{0,1,p+1}, \xi_j^{p+1}) \mid 0 \le i,j \le p\} \qquad \text{and} \qquad G_{\triangleright,p} := F_{\triangleright}(G_{w,\tilde{Q},p}).$$

Figure 2.2 shows examples of $G_{\triangleright,p}$. Further, we note the Lagrangian functions with respect to the Gauss-Lobatto points

$$\lambda_i^p(\xi) := \prod_{\substack{k=0 \\ k\neq i}}^{p} \frac{\xi - \xi_k^{p+1}}{\xi_i^{p+1} - \xi_k^{p+1}}, \qquad 0 \le i \le p,$$

and replacing ξ_i^{p+1} by $\xi_i^{0,1,p+1}$ with respect to the Gauss-Lobatto-Jacobi points

$$\lambda_i^{w,p}(\xi) := \prod_{\substack{k=0 \\ k\neq i}}^{p} \frac{\xi - \xi_k^{0,1,p+1}}{\xi_i^{0,1,p+1} - \xi_k^{0,1,p+1}}, \qquad 0 \le i \le p\,.$$

We define the basis

$$B_p^{\triangleright}(\tilde{Q}) := \left(\lambda_0^{w,p}\,; \quad (\lambda_i^{w,p} \cdot \lambda_j^p,\ 1 \le i \le p,\ 0 \le j \le p)\right)$$

and the $p(p+1)+1$ dimensional FE space

$$\mathbb{P}_p^{\triangleright,2}(\tilde{Q}) := \operatorname{span}(B_p^{\triangleright}(\tilde{Q})).$$

Using the extension $\widetilde{F_{\triangleright}^{-1}} : \tilde{T} \to \tilde{Q}$ of the inverse of $F_{\triangleright}$ given by

$$\widetilde{F_{\triangleright}^{-1}} := \begin{cases} F^{-1}(x_1,x_2) & \text{if } (x_1,x_2) \neq (0,0)\,, \\ (0,0) & \text{if } (x_1,x_2) = (0,0)\,, \end{cases}$$

we define the FE space on $\tilde{T}$

$$\mathbb{P}_p^{\triangleright}(\tilde{T}) := \left\{ v \circ \widetilde{F_{\triangleright}^{-1}} \,\middle|\, v \in \mathbb{P}_p^{\triangleright,2}(\tilde{Q}) \right\}.$$

For sake of simplicity and to avoid further canonical definitions we consider only the Dirichlet problem on $\Omega = \tilde{T}$, i.e., we assume $\Gamma_N = \emptyset$. We take the single element mesh $\mathcal{T} = \{\tilde{T}\}$ and state the triangle analogies to the above Theorems 2.6, 2.7, 2.8, and 2.11. Firstly, we define the appropriate subsets $V_{p,g_D}^{\triangleright}$ and $K_{p,g_D}^{\triangleright}$ of $\mathbb{P}_p^{\triangleright}(\tilde{T})$ which are necessary

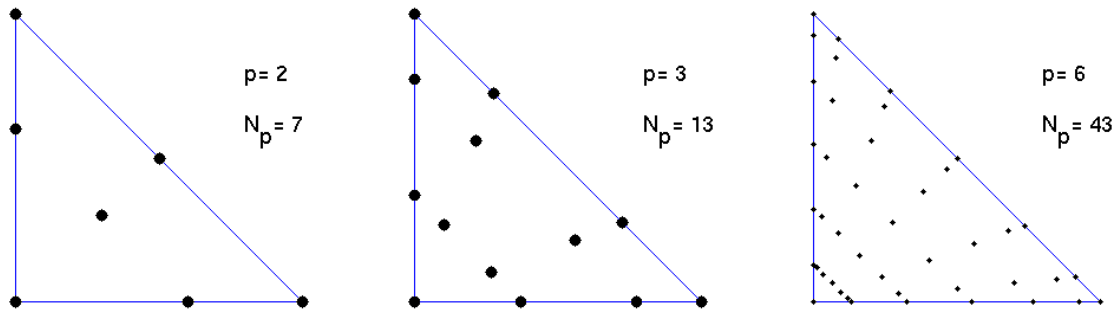

Figure 2.2: The point set $G_{\triangleright,p} \subset \tilde{T}$ for $p = 2, 3, 6$

to control the Dirichlet and the obstacle condition:

$$V_{p,g_D}^{\triangleright} := \{w \in \mathbb{P}_p^{\triangleright}(\tilde{T}) \,|\, w(x) = g_D(x),\ x \in \Gamma_{D,p}^{\triangleright}\}\,, \tag{2.32}$$

$$K_{p,g_D}^{\triangleright} := \{w \in \mathbb{P}_p^{\triangleright}(\tilde{T}) \,|\, w(x) \ge g_D(x), x \in G_{\triangleright,p},\ \text{and } w(x) = g_D(x), x \in \Gamma_{D,p}^{\triangleright}\} \tag{2.33}$$

with $\Gamma_{D,p}^{\triangleright} := \overline{\Gamma}_D \cap G_{\triangleright,p}$.

Analogously to [BM92, Section 5] (see [BM97, Sections 7, 14, 20] for a detailed analysis) Proposition 2.17 can be extended to two dimensions since weighted Sobolev spaces on tensorized domains satisfy the same tensorization properties as the standard ones. This should yield the following analogy of Theorem 2.3.

Proposition 2.21. Let i_p be the interpolation operator on $\tilde{Q}$ with respect to $G_{0,1,\tilde{Q},p}$ from Definition 2.20. For any real numbers r and s satisfying $s > (d+r)/2$, $d = 2$, and $0 \le r \le 1$, there exists a positive constant c depending only on $s > 1$ such that, for any function $v \in H^s_{\triangleright}(\tilde{Q})$, the following estimate holds

$$\|v - i_p v\|_{H^r_{\triangleright}(\tilde{Q})} \le c p^{r-s} \|v\|_{H^s_{\triangleright}(\tilde{Q})} .$$

The following three propositions are analogies to Theorem 2.6, Theorem 2.7, and Theorem 2.8 and state existence, uniqueness, and the convergence of the p-version approach by weighted Sobolev spaces. We do not give an a priori error estimate for the discrete solution u_p here. Nevertheless, Experiment 2.26 shows convergence rates better than $\|u - u_p\|_{H^1(\Omega)} \le \mathcal{O}(p^{-1})$ for the obstacle and the obstacle free problem.

Proposition 2.22. Let $\Omega := \tilde{T}$ and let $V^{\triangleright}_{p,g_D}$ be given by (2.32). Further, let there exist positive constants ρ_i, $i = 0,1,2,3$, as in (1.6). Then, there exists a unique u_p satisfying the following equivalences:

(i) u_p minimizes the functional A on $V^{\triangleright}_{p,g_D}$, i.e. $A(u_p) \le A(v)$ for all $v \in V^{\triangleright}_{p,g_D}$.

(ii) There exists a $u_p \in V^{\triangleright}_{p,g_D}$ satisfying $DA(u_p; v - u_p) = 0$ for all $v \in V^{\triangleright}_{p,g_D}$.

Proposition 2.23. Let $\Omega := \tilde{T}$ and let $K^{\triangleright}_{p,g_D}$ be given by (2.33). Let the obstacle $\psi \in H^1(\Omega)$ be defined with $\psi \le g_D$ almost everywhere in a neighborhood of Γ_D, let $u_p \in K^{\triangleright}_{p,g_D}$ be defined by Definition 2.4 and let there exist positive constants ρ_i, $i = 0,1,2,3$, as in (1.6). Then, there exists a unique u_p satisfying the following equivalences:

(i) u_p minimizes the functional A on $K^{\triangleright}_{p,g_D}$, i.e. $A(u_p) \le A(v)$ for all $v \in K^{\triangleright}_{p,g_D}$.

(ii) There exists a $u_p \in K^{\triangleright}_{p,g_D}$ satisfying $DA(u_p; v - u_p) \ge 0$ for all $v \in K^{\triangleright}_{p,g_D}$.

Proof of Proposition 2.22 and Proposition 2.23. Both theorems are analogies to Theorem 2.6 and Theorem 2.7. Thus, it suffices to note that $V^{\triangleright}_{p,g_D}$ and $K^{\triangleright}_{p,g_D}$ are convex, closed, and nonempty subsets of $\mathbb{P}^{\triangleright}_p(\tilde{T})$. □

Proposition 2.24. Let $\psi \in C^0(\overline{\tilde{T}}) \cap H^1(\tilde{T})$, $\psi \le g_D$ in a neighborhood of $\Gamma_D = \Gamma$, and $\Gamma_N = \emptyset$. With the above assumptions on K and $K^{\triangleright}_{p,g_D}$, there holds

$$\lim_{p\to\infty} \|u_p - u\|_{H^1(\tilde{T})} = 0$$

with u and u_p the minimizers of Theorem 1.23 and Proposition 2.23, respectively.

Sketch of the proof of Proposition 2.24. Since $u \in H^\mu(\tilde{T})$ implies $u \circ F_\triangleright \in H^\mu_\triangleright(\tilde{Q})$ for $\mu \geq 0$, it suffices to prove the convergence of u_p towards u and its rate on $\tilde{Q}$ with respect to the norm $H^1_\triangleright(\tilde{Q})$.

The proof of Theorem 2.8 depends basically on the approximation properties of the Bernstein operator B^2_p and the interpolation operator i_p given by Theorem 2.3. The approximation properties of the interpolation operator i_p with respect to $G_{0,1,\tilde{Q},p}$ were given in Proposition 2.21. This yields the approximation of the interpolant in $G_{\triangleright,p}$ on $\tilde{T}$. An appropriate Bernstein operator on $\tilde{T}$ can be defined using the transformation $F_\triangleright$.

Furthermore, the exactness of the Gauss-Lobatto quadrature formula (2.3) for the integration of polynomials of degree $\leq 2p-1$ and the positivity of the quadrature weights ρ_i^{p+1} play an essential role. By the Gauss-Lobatto-Jacobi quadrature (2.27) we have an analogy of the Gauss-Lobatto quadrature. □

2.3 Numerical experiments on the square and on the triangle

Experiment 2.25 (*p-version on a square*). We consider the Poisson equation

$$-\triangle u = f \tag{2.34}$$

with homogenous Dirichlet data on the square $Q := \tilde{Q} = [-1,1]^2$ and with $f = -\triangle w$,

$$w = (x+1)(y+1)(e^{(x-1)(y-1)} - 1). \tag{2.35}$$

The solution of (2.34) minimizes

$$A(v) := \int_T (\tfrac{1}{2}\nabla^T v \; \nabla v - fv) \; \mathrm{d}x \quad \text{ on } H_0^1(Q).$$

As obstacle functions ψ we introduce $\psi \equiv -1.5$ and $\psi \equiv -1$ on Q. The obstacle problem

$$-\triangle u \geq f, \qquad (u-\psi)\cdot \triangle u = 0, \qquad u \geq \psi \quad \text{ in } Q \tag{2.36}$$

is solved if we take the minimizer of A not on $H_0^1(Q)$, but on the admissible functions $K := \{v \in H_0^1(Q) | v \geq \psi\}$. As approximation space we choose V_p and its discrete subsets V_{p,g_D} and K_{p,g_D} given in Definition 2.4. The minimizer of the discrete minimization problem is called u_p.

The integrals of the discrete problem are calculated by a Gauss-Lobatto quadrature (cf. (2.3)) with $p+4$ points. The discrete problem leads to a quadratic programming problem which can be solved by relaxation methods (cf. [Glo84, Chapter V]) or a generalized conjugate gradient algorithm (cf. [O'L80]), known as Polyak algorithm. Here, the quadratic programming problem is solved using the matlab routine `quadprog` in large-scale mode which implements a subspace trust-region method based on the interior-reflective Newton method described by Coleman and Li in [CL96]. Each iteration involves the approximate solution of a large linear system using the method of diagonal preconditioned conjugate gradients (PCG). The interior PCG iteration is terminated when the relative residual is less than 10^{-14}. The algorithm stops when either the relative change of the solution vector representing u_p in the Euclidian norm is $\leq 10^{-13}$ or the relative change of $A(u_p)$ is $\leq 10^{-14}$.

The obstacle problem on the square is visualized in Figure 2.3 for the obstacle $\psi \equiv -1$ by 3d and contour plots. These plots give an insight on how the obstacle condition is violated by the p-FE solution for $p = 10$.

The numerical results on the square are listed in Table 2.1. While $u = w$ from (2.35) solves (2.34) and $A(u) = -14.95831706718053$, the exact solution u of the obstacle problem (2.36) is not known. Here, we apply Lemma 4.48 to estimate the error by

$$\|u - u_p\|^2_{H_1(Q)} \leq 2\,C\,|A(u_p) - A(u)| \quad \text{ with a constant } C > 0,$$

despite $u_p \not\subset K$. For $\psi \equiv -1.5$ the value $A(u) = -14.189$ and for $\varphi \equiv -1$ the value $A(u) = -12.109$ is obtained by extrapolation assuming $|A(u_p) - A(u)| \approx Cp^{-\beta}$ with

constants $C, \beta > 0$. Thus p, $p+2$, $p+4$ give three equations which allow to determine $A(u)$, C, and β by solving a nonlinear system.

The experimental convergence rates α_p with respect to the polynomial degrees are computed from $|A(u_p) - A(u)| \approx Cp^\alpha$ where C and α are constants independent of p. Note that in case of inhomogeneous boundary conditions and obstacle conditions the sequence $A(u_p)$ is not monotonously decreasing since $V_{p,g_D} \not\subset V_{q,g_D}$ and $K_p \not\subset K_q$ for $p < q$. Therefore, the experimental convergence rates are computed from

$$\alpha_{p+2} = \log\left|\frac{A(u_p) - A(u)}{A(u_{p+2}) - A(u)}\right| \Big/ \log\left(\frac{p}{p+2}\right)$$

in case of $\psi \equiv -\infty$ whereas for $\psi \equiv -1.5$, $\psi \equiv -1$ we take $p+4$ instead of $p+2$.

Experiment 2.26 (*p-version on a triangle*). The second numerical experiment plays on the triangle T given by the vertices $(0,0)$, $(1,0)$, $(0,1)$. Again, we take the Poisson equation with homogenous Dirichlet data, now with a right hand side f such that

$$u = xy\,(e^{(x+y)} - e)$$

is the exact solution of the unconstrained problem. As obstacle functions we introduce $\psi \equiv -0.06$ and $\psi \equiv -0.025$. For the approximation we use the subsets $V^{\triangleright}_{p,g_D}$ and $K^{\triangleright}_{p,g_D}$ defined on the reference triangle $\tilde{T}$ in (2.32) and (2.33).

The integrals $\int_{\tilde{T}} \nabla v_p \nabla w_p$ and $\int_{\tilde{T}} f v_p$ of the discrete problem are calculated on the reference square $\tilde{Q}$ using the transformation $F_{\triangleright}$ from (2.25). On $\tilde{Q}$ the quadrature is done by the weighted Gauss-Lobatto-Jacobi quadrature given in (2.27) wrt. to the first component and by the Gauss-Lobatto-Jacobi quadrature given in (2.3) wrt. to the second component with $p+4$ quadrature points in both directions.

The quadratic programming problem is calculated as described in Experiment 2.25. Figure 2.4 shows the situation with the obstacle $\psi \equiv -0.06$ for $u_p \in K^{\triangleright}_{p,0}$.

The analytical solution u yields the value $A(u) = -3.1712266254455625 \cdot 10^{-2}$ for the obstacle free problem. The extrapolations for the obstacle problems give $A(u) = -0.030443$ in case of $\varphi \equiv -0.06$ and $A(u) = -0.020738$ in case of $\varphi \equiv -0.02$, but it is only reliable in the first four significant digits due to oscillations. The results on the triangle are documented in Table 2.2.

The main problem with error analysis of the above mentioned numerical experiments is that we do not have a monotone decreasing sequence $A(u_p)$ in case of a real obstacle problem. As a work around the experimental convergence rate was calculated with respect to the numerical results for p, $p+4$ instead of p, $p+2$.

On the square the theoretical result $|A(u_p) - A(u)| \leq \mathcal{O}(p^{-2})$ for the p-version is confirmed. In case of the unconstrained problem the extrapolation of $A(u)$ equals the analytical calculated value in the first 15 digits (!) and the convergence rates are very high until $p = 13$. The values for $p > 13$ still give the correct solution, but since the working precision is already reached, the approximation error can not decrease any more. Also in case of the constrained problems convergence rates are high and the method performs

stable for $p \leq 20$ in the first four significant digits which are determined correctly already for $p = 14$.

On the triangle the extrapolation of the $A(u)$ for the unconstrained problem equals the analytical calculated value in the first 14 digits and the convergence rates are very high until $p = 9$. The values for $p > 9$ still give the correct solution but since working precision is already reached, the approximation error can not decrease any more. In case of the the constrained problems convergence rates are high and the method performs stable for $p \leq 20$. The first three significant digits of $A(u)$ are determined correctly for a small number of degree of freedoms. The triangle experiments confirm a convergence rate for the constrained and the unconstrained case better than $|A(u_p) - A(u)| \leq \mathcal{O}(p^{-2})$.

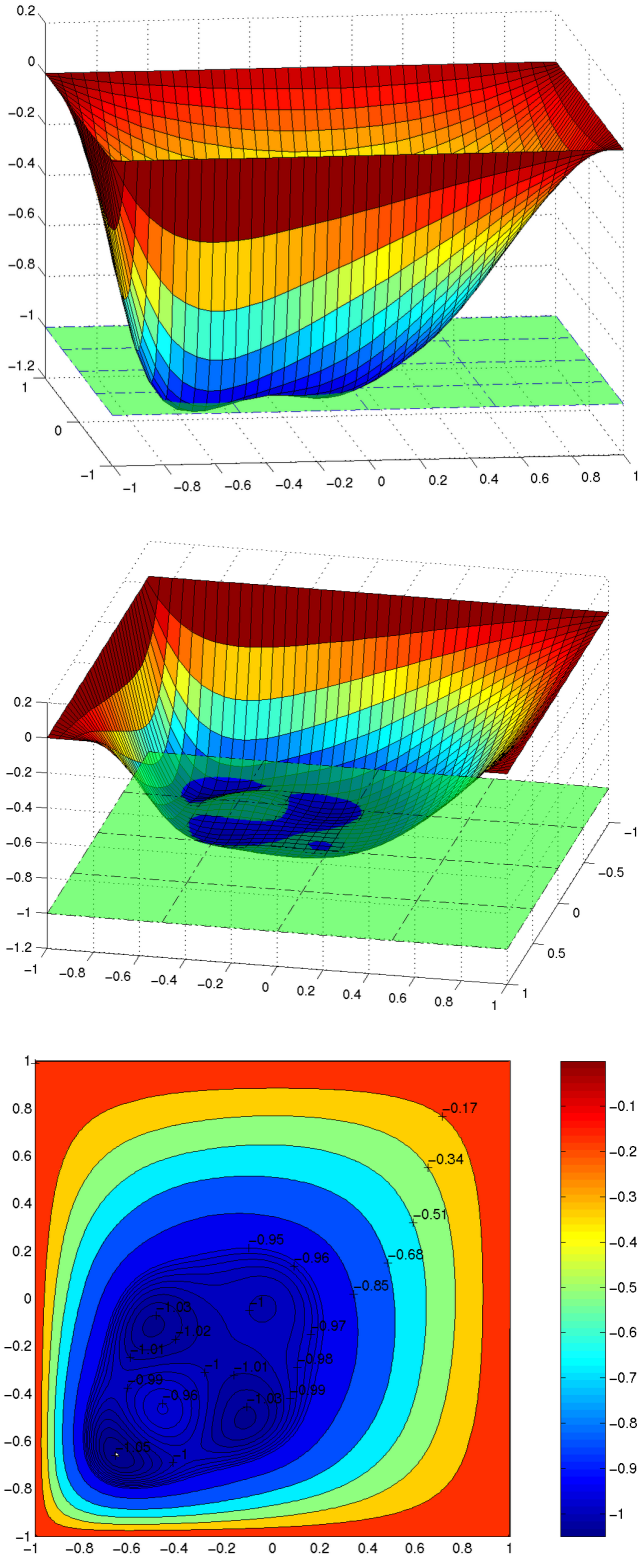

Figure 2.3: The obstacle problem on the square $[-1,1]^2$ with $\psi \equiv -1$. The plots visualize the FE solution $u_p \in K_{10,0}$.

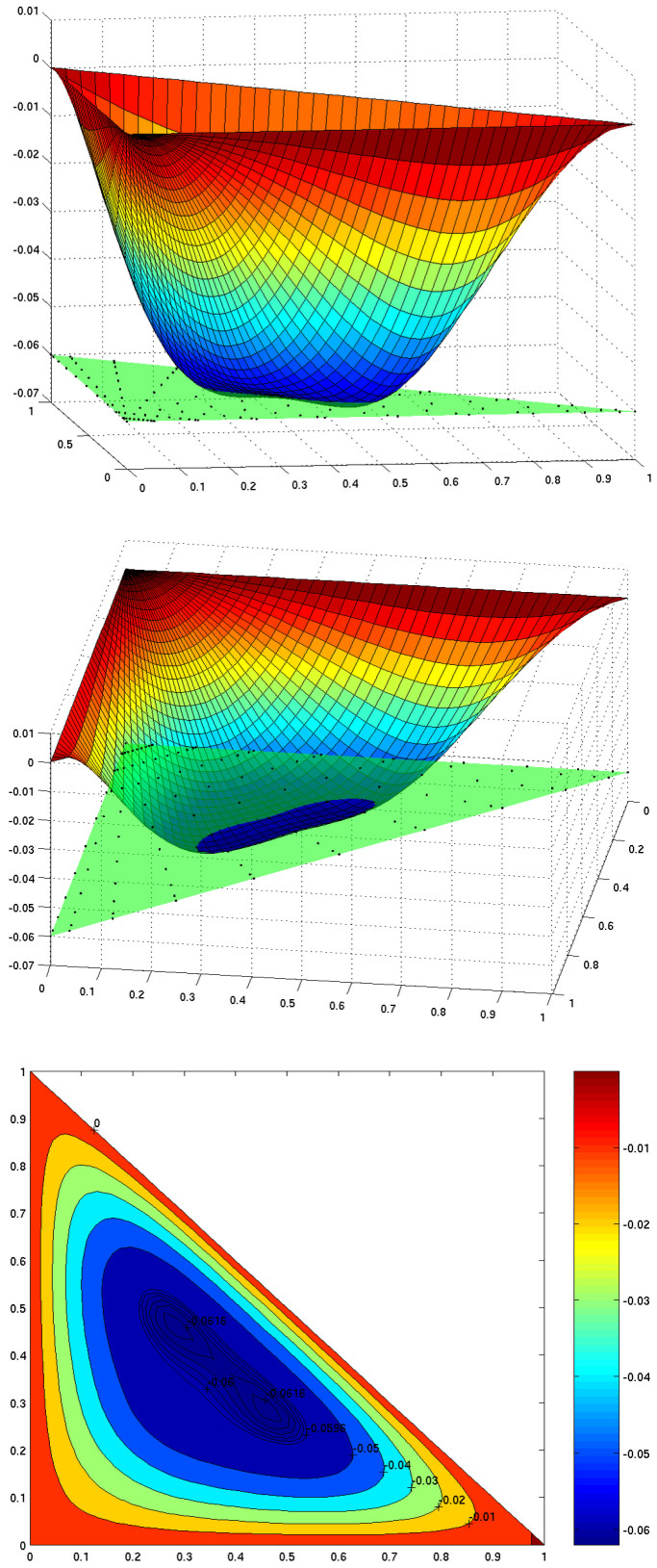

Figure 2.4: The obstacle problem on the triangle $\tilde{T}$ with $\psi \equiv -0.06$. The plots visualize the FE solution $u_p \in K^{\triangleright}_{10,0}$.

Table 2.1: Convergence on the square $[-1,1]^2$ for different obstacles, $A := A(u)$

$\varphi \equiv$	p	N_p	$A(u_p)$	$\lvert A(u_p) - A\rvert$	α_p
$-\infty$	2	1	-6.56805e+00	8.39e+00	–
	3	4	-1.37450e+01	1.21e+00	–
	4	9	-1.49633e+01	5.00e-03	-10.71
	5	16	-1.49643e+01	5.99e-03	-10.40
	6	25	-1.49587e+01	3.64e-04	-6.47
	7	36	-1.49583e+01	1.20e-05	-18.46
	8	49	-1.49583e+01	2.68e-07	-25.08
	9	64	-1.49583e+01	4.39e-09	-31.49
	10	81	-1.49583e+01	5.55e-11	-38.01
	11	100	-1.49583e+01	6.52e-13	-43.93
	12	121	-1.49583e+01	2.31e-14	-42.70
	13	144	-1.49583e+01	2.84e-14	-18.75
	14	169	-1.49583e+01	1.42e-13	–
	15	196	-1.49583e+01	4.44e-14	–
	16	225	-1.49583e+01	2.49e-14	–
	17	256	-1.49583e+01	1.21e-13	–
	18	289	-1.49583e+01	2.49e-14	–
	19	324	-1.49583e+01	1.78e-15	–
	20	361	-1.49583e+01	1.56e-13	–
-1.5	2	1	-6.56805e+00	7.62e+00	–
	3	4	-1.28368e+01	1.35e+00	–
	4	9	-1.47686e+01	5.80e-01	–
	5	16	-1.41591e+01	2.99e-02	–
	6	25	-1.41477e+01	4.13e-02	-4.75
	7	36	-1.42738e+01	8.48e-02	-3.27
	8	49	-1.42186e+01	2.96e-02	-4.29
	9	64	-1.41769e+01	1.21e-02	-1.53
	10	81	-1.41946e+01	5.63e-03	-3.90
	11	100	-1.42096e+01	2.06e-02	-3.13
	12	121	-1.41797e+01	9.29e-03	-2.86
	13	144	-1.41823e+01	6.72e-03	-1.61
	14	169	-1.41885e+01	4.59e-04	-7.45
	15	196	-1.41877e+01	1.32e-03	-8.86
	16	225	-1.41789e+01	1.01e-02	0.30
	17	256	-1.41856e+01	3.39e-03	-2.56
	18	289	-1.41855e+01	3.49e-03	8.07
	19	324	-1.41833e+01	5.65e-03	6.15
	20	361	-1.41825e+01	6.52e-03	-1.97
-1	2	1	-6.10945e+00	6.00e+00	–
	3	4	-1.04487e+01	1.66e+00	–
	4	9	-1.27217e+01	6.13e-01	–
	5	16	-1.24264e+01	3.17e-01	–
	6	25	-1.20966e+01	1.24e-02	-5.63

continued on next page

Table 2.1: *continued from previous page*					
$\varphi \equiv$	p	N_p	$A(u_p)$	$\lvert A(u_p) - A\rvert$	α_p
	7	36	-1.20735e+01	3.55e-02	-4.54
	8	49	-1.21574e+01	4.84e-02	-3.66
	9	64	-1.21550e+01	4.60e-02	-3.28
	10	81	-1.21047e+01	4.28e-03	-2.08
	11	100	-1.21054e+01	3.56e-03	-5.09
	12	121	-1.21217e+01	1.27e-02	-3.30
	13	144	-1.21230e+01	1.40e-02	-3.25
	14	169	-1.21057e+01	3.27e-03	-0.80
	15	196	-1.21051e+01	3.92e-03	0.32
	16	225	-1.21108e+01	1.84e-03	-6.72
	17	256	-1.21143e+01	5.34e-03	-3.58
	18	289	-1.21078e+01	1.25e-03	-3.83
	19	324	-1.21068e+01	2.22e-03	-2.41
	20	361	-1.21087e+01	3.05e-04	-8.05

Table 2.2: Convergence on the triangle $\tilde{T}$ for different obstacles, $A := A(u)$

$\varphi \equiv$	p	N_p	$A(u_p)$	$\lvert A(u_p) - A\rvert$	α_p
$-\infty$	2	1	-2.11640e-02	1.05e-02	–
	3	4	-3.15392e-02	1.73e-04	–
	4	9	-3.17109e-02	1.33e-06	-12.96
	5	16	-3.17123e-02	6.07e-09	-20.08
	6	25	-3.17123e-02	1.78e-11	-27.67
	7	36	-3.17123e-02	4.01e-14	-35.45
	8	49	-3.17123e-02	5.55e-17	-44.07
	9	64	-3.17123e-02	2.46e-15	-11.11
	10	81	-3.17123e-02	9.71e-17	2.51
	11	100	-3.17123e-02	1.03e-15	-4.31
	12	121	-3.17123e-02	1.32e-16	1.67
	14	169	-3.17123e-02	2.01e-16	-6.79
	15	196	-3.17123e-02	1.18e-16	–
	16	225	-3.17123e-02	4.72e-16	–
	17	256	-3.17123e-02	6.75e-14	–
	18	289	-3.17123e-02	6.52e-16	–
	19	324	-3.17123e-02	3.20e-13	–
	20	361	-3.17123e-02	5.34e-16	–
-0.06	2	1	-2.08958e-02	9.55e-03	–
	3	4	-3.13083e-02	8.65e-04	–
	4	9	-3.10520e-02	6.09e-04	–
	5	16	-3.04364e-02	6.56e-06	–
	6	25	-3.05239e-02	8.09e-05	-4.34
	7	36	-3.06197e-02	1.77e-04	-1.87
	8	49	-3.04235e-02	1.95e-05	-4.97
	9	64	-3.04702e-02	2.72e-05	2.42

continued on next page

Table 2.2: *continued from previous page*

$\varphi \equiv$	p	N_p	$A(u_p)$	$\lvert A(u_p) - A \rvert$	α_p
	10	81	-3.04981e-02	5.51e-05	-0.75
	11	100	-3.04484e-02	5.42e-06	-7.71
	12	121	-3.04416e-02	1.43e-06	-6.44
	13	144	-3.04690e-02	2.60e-05	-0.12
	14	169	-3.04429e-02	6.51e-08	-20.03
	15	196	-3.04379e-02	5.10e-06	-0.19
	16	225	-3.04575e-02	1.45e-05	8.05
	17	256	-3.04395e-02	3.51e-06	-7.47
	18	289	-3.04408e-02	2.18e-06	13.96
	19	324	-3.04487e-02	5.73e-06	0.50
	20	361	-3.04413e-02	1.75e-06	-9.49
-0.02	2	1	-1.27575e-02	7.98e-03	–
	3	4	-1.97011e-02	1.04e-03	–
	4	9	-2.18630e-02	1.13e-03	–
	5	16	-3.17123e-02	1.10e-02	–
	5	16	-2.12502e-02	5.12e-04	-3.00
	6	25	-2.05964e-02	1.42e-04	-2.87
	7	36	-2.06977e-02	4.03e-05	-5.95
	8	49	-2.08782e-02	1.40e-04	-9.28
	9	64	-2.08686e-02	1.31e-04	-2.33
	10	81	-2.07312e-02	6.83e-06	-5.93
	11	100	-2.07248e-02	1.32e-05	-2.46
	12	121	-2.07696e-02	3.16e-05	-3.68
	13	144	-2.07816e-02	4.36e-05	-2.98
	14	169	-2.07428e-02	4.82e-06	-1.04
	15	196	-2.07327e-02	5.29e-06	-2.96
	16	225	-2.07490e-02	1.10e-05	-3.65
	17	256	-2.07595e-02	2.15e-05	-2.63
	18	289	-2.07467e-02	8.68e-06	2.34
	19	324	-2.07375e-02	5.17e-07	-9.84
	20	361	-2.07432e-02	5.21e-06	-3.36

2.4 Non-uniform *hp*-refinements

The *hp*-version of the FEM is capable of producing approximations that converge exponentially fast in the energy norm to the solutions of elliptic boundary value problems which typically have singularities in the neighborhood of re-entrant corners and where the boundary conditions change type. Although the use of adaptive *h*-version methods improves convergence rates to solutions with singularities to some extent, it is still difficult to produce accurate solutions of elliptic PDE known from engineering applications like re-entrant corners in linear elasticity. Here, the *hp*-version is an attractive alternative. One can differentiate between two refinement schemes, both using h and p refinements of the approximation space.

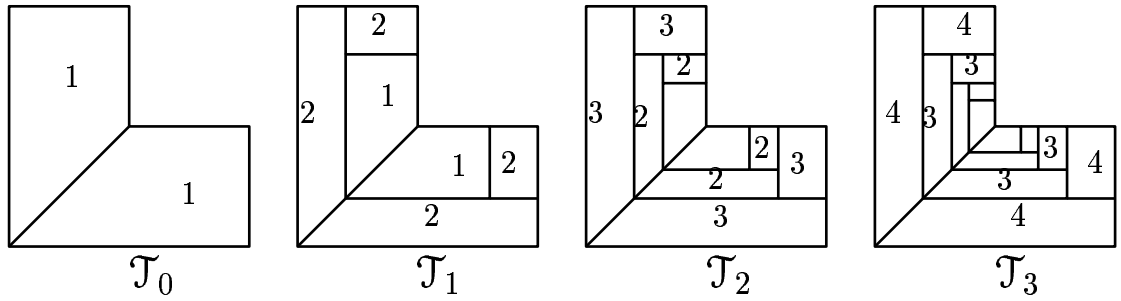

Figure 2.5: *hp*-refined mesh on an L-shape domain. The numbers on the quadrilaterals denote the polynomial degrees associated to the quadrilateral Q.

Algorithm 2.1 A *hp*-adaptive algorithm

Let $\mathcal{T}_0$ be an initial mesh on the domain Ω and $V_0 := V_{\vec{p}}$ an initial finite element subspace of $H^1(\Omega)$. Here, $\vec{p} := (p_Q)_{Q\in\mathcal{T}}$ denotes the distribution of the polynomials degrees wrt. the elements of the initial grid. Further, let $\theta \in (0,1)$ be the refinement parameter, $\tau > 0$ be the stopping parameter.

1. Set $i = 0$.
2. Compute the Galerkin approximation $u_i \in V_i$.
3. Compute a local a posteriori error estimator η_Q on each element $Q \in \mathcal{T}_i$. If $\eta < \tau$ then stop.
4. If $\eta_Q \geq \theta \max\{\eta_Q \,|Q \in \mathcal{T}_i\}$ then refine Q (h refinement) or increase the polynomial degree p_Q (p refinement).
5. Construct new subspace V_{i+1}. Increment i. Goto 2.

Firstly, we have a strict refinement scheme used to treat known singularities caused by re-entrant corners. Here, we use a geometrically refined mesh with the smallest triangles and quadrilaterals next to the re-entrant corner and assign increasing polynomial degrees to the elements when we go away from the the corner (see Figure 2.5). This concept is analyzed using countably normed Sobolev spaces for FEM, BEM, and the coupling of both methods (cf. [Sch98, HS96, HS98]).

Secondly, the *hp*-refinement is used in a more flexible way based on an a posteriori error estimator which should provide some indication of the distribution of the error on the elements. Assume that η_Q is an a posteriori estimator for the error on the element Q with the estimator for the global error obtained by local contributions $\eta^2 = \sum_{Q\in\mathcal{T}} \eta_Q^2$. Then, a *hp*-adaptive algorithm can be characterized by Algorithm 2.1. The idea of the algorithm is to equilibrate the errors in the elements. However, the *hp*-refinement requires a decision whether to perform *h*-refinement or *p*-refinement in *Step 4*. Here, Algorithm 3.1 presents a method where the decision between the two refinements is based on the refinement history. As we give a short overview on various *hp*-refinement strategies for the treatment of PDE in Section 3.1, we do not discuss this subject here.

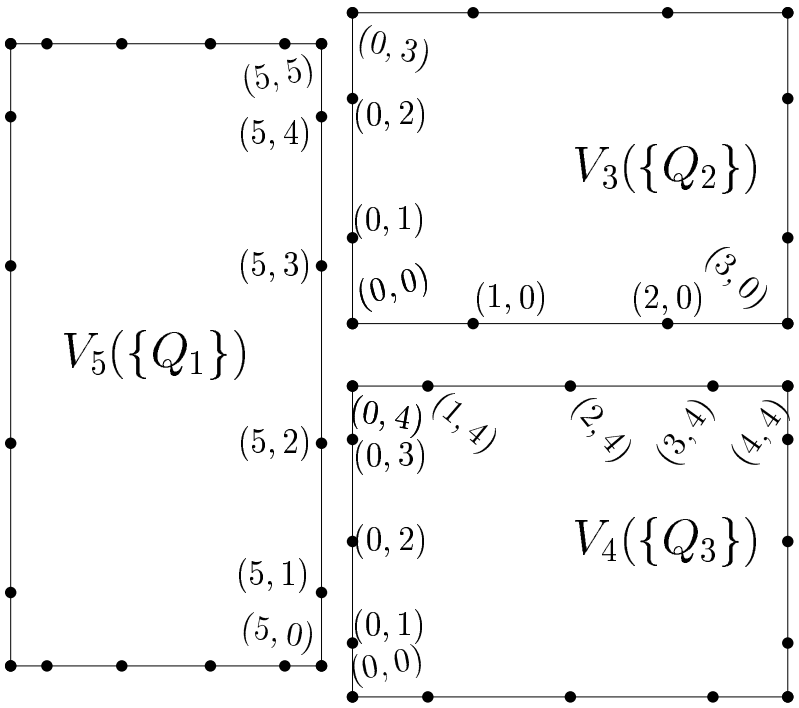

Figure 2.6: Connecting quadrilaterals with local p-FE spaces of different polynomial degree

The next section is devoted to the definition of hp-FE spaces which allow to control the inequality constraints raised by partial differential inequalities.

2.4.1 Conforming hp-finite elements

A typical situation obtained using adaptive h- and p-refinements is shown in Fig. 2.7. It contains differing polynomial orders and elements of differing sizes adjacent to one another. In the following we describe the basic ideas behind the realization of a conforming finite element space for such meshes.

Suppose that the single element spaces on the quadrilaterals Q_1, Q_2, and Q_3 are of polynomial degree 5, 3, and 4, respectively, as shown in Fig. 2.6. For ease of presentation only the degrees of freedom associated with the Gauss-Lobatto points on the boundary are marked by a dot. To maintain continuity across the inter-element boundary there are two alternatives:

(i) increase the polynomial degree of the approximation along the interface in elements Q_2, Q_3 to match the degree in element Q_1;

(ii) decrease the polynomial degree of approximation in elements Q_1, Q_2 along the interfaces to match the degree in element Q_3.

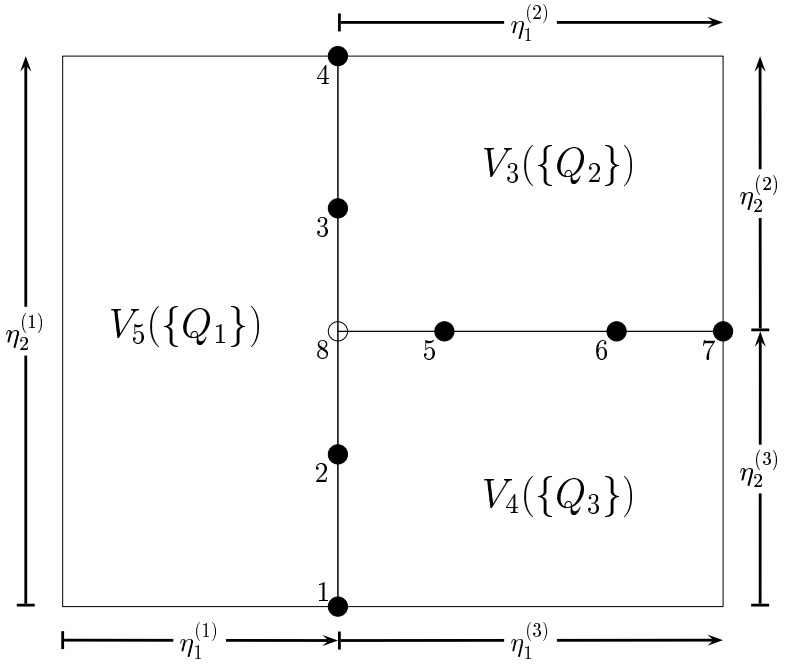

Figure 2.7: The continuity is obtained by edge associated degrees of freedom

An advantage of the second choice is that the information stored in element Q_3 is unchanged even though the neighbor has been enriched. In particular, the local Galerkin matrix and load vector associated with Q_3 from the previous mesh need not to be reconstructed, when the size of the element and the polynomial degree are left unchanged.

Let us denote the segments of the inter-element boundaries by

$$e_{12} := \overline{Q}_1 \cap \overline{Q}_2\,, \quad e_{13} := \overline{Q}_1 \cap \overline{Q}_3\,, \quad e_{23} := \overline{Q}_2 \cap \overline{Q}_3\,,$$

and let v be a continuous function with $v_{|Q_i} \in V_{p_i}(Q_i)$, $i = 1, 2, 3$. Assuming continuity on e_{12} we obtain that $v_{|e_{12}}$ is a polynomial of degree 3 due to $v_{|Q_2} \in V_3(\{Q_2\})$. $v_{|e_{13}}$ has to be of polynomial degree 4 due to $v_{|Q_3} \in V_4(\{Q_3\})$ and $v_{|e_{23}}$ of polynomial degree 3 due to $v_{|Q_2} \in V_3(\{Q_2\})$.

Since $v_{|e_{12}\cup e_{13}}$ must be a polynomial on the entire edge $\overline{e_{12} \cup e_{13}}$, it follows further that we have to lower the polynomial degree on $v_{|e_{13}}$ to 3. Thus, we may associate 4 global degrees of freedom with the scaled Gauss-Lobatto points on the edge $\overline{e_{12} \cup e_{13}}$ (see. Fig. 2.7). Analogously, we may associate the global degrees of freedom $5, 6, 7, 8$ with the scaled Gauss-Lobatto points on the edge e_{23}. But the node labeled by number 8 is ficticious in the sense that it depends on the global nodes $1, 2, 3, 4$. As a conclusion, we have that the local degrees of freedom denoted by $(\cdot, \cdot)$ in Fig, 2.6 depend on the edge associated global degrees of freedom $1, 2, \ldots, 7$.

In the treatment of variational inequalities we demanded that the discrete solution fulfills the obstacle condition in the Gauss-Lobatto points G_p. Up to now, this was no problem, since on p-uniform meshes without hanging nodes a global degree of freedom could be associated with each element of G_p (see Definition 2.2). But now we may have more local

degrees of freedom than global degrees of freedom which will lead to serious problems in the algebraic reformulation needed for the implementation. We omit the algebraic details here, since they will be discussed in Chapter 4.

In the analysis of the p-discretization on quadrilaterals and on triangles (see Sections 2.1, 2.2) we assumed $V_p \subset H^1(\Omega)$, but not $K_{p,g_D} \subset K$, i.e., we assumed conformity of the approximation subset wrt. the continuity of $H^1(\Omega)$ but not wrt. the obstacle condition. Accordingly, the violation of the obstacle condition was considered in the proofs of Theorem 2.8 and Theorem 2.11. Both proofs hold further, if we do not control the obstacle condition in the local degrees of freedom tagged by

$$\begin{aligned} (5,i), &\qquad i=0,\dots,5, &&\text{on } Q_1;\\ (0,i) \text{ and } (i,0), &\qquad i=0,\dots,3, &&\text{on } Q_2;\\ (0,i) \text{ and } (i,4), &\qquad i=0,\dots,4, &&\text{on } Q_3; \end{aligned}$$

but at the Gauss-Lobatto points $1,2,\dots,7$ on the edges in Figure 2.7. To make it general, we relieve the obstacle condition by associating it with the global degrees of freedom as follows.

Let $Q_1,\dots Q_n \in \mathcal{T}$ be the ordered quadrilateral elements defining a mesh in usual finite element sense, possibly with hanging nodes, and let $\mathcal{E}$ be the set of all inter-element and boundary edges e. Here, we demand that e is an entire edge of all adjacent quadrilaterals, e.g. $e_{12},\, e_{13} \notin \mathcal{E}$, but $e_{12} \cup e_{13},\, e_{23} \in \mathcal{E}$.

Let p_Q be the polynomial degree of the discrete space on the element Q and $\vec{p} := (p_Q)_{Q\in\mathcal{T}}$ the respective vector notation. For every Q there exists a diffeomorphism F_Q mapping the reference square $\tilde{Q}$ on Q. Noting the four edges of the reference square $\tilde{Q}$ by

$$e^{(1)} := (\eta_1,-1), \quad e^{(2)} := (1,\eta_2), \quad e^{(3)} := (\eta_1,1), \quad e^{(4)} := (-1,\eta_2) \tag{2.37}$$

where $\eta_1,\eta_2 \in [-1,1]$, we may write

$$e = F_Q(e^{(j)}) \qquad \text{for at least one } Q\in\mathcal{T} \text{ and one } j\in\{1,2,3,4\}. \tag{2.38}$$

If two Q fulfill (2.38) we take that with the lower ordering number to reach uniqueness. With every edge $e\in\mathcal{E}$ we associate the polynomial degree

$$p_e := \min\{p_Q \,|\, Q\in\mathcal{T},\, e\cap\overline{Q}\neq\emptyset\}\,,$$

i.e., the minimum of polynomial degrees of the adjacent quadrilaterals. Further, we associate the edge transformed Gauss-Lobatto points

$$x_i^{p_e+1,e} := F_Q((\xi_i^{p_e+1},-1)) \tag{2.39}$$

in case of $j=1$ in (2.38). In case of $j=2,3,4$, we get $x_i^{p_e+1,e}$ analogously by introducing $\xi_i^{p_e+1}$ into η_1 and η_2 of (2.37).

Now, we must consider the hanging nodes x_h. For all x_h there exist a pair $(e,i)\in\mathcal{E}\times\{0,p_e\}$ such that $x_h = x_i^{p_e+1,e}$ because they are end points of an edge e. Since x_h is a hanging node, there exists a second edge $\hat{e}\neq e$ with $x_h\in\hat{e}$ and x_h is not an end point of $\hat{e}$. Thus, we can differentiate between hanging nodes which own a representation

$$x_h = \xi_{\hat{i}}^{p_{\hat{e}}+1,\hat{e}} \qquad \text{for a pair } (\hat{e},\hat{i})\in\mathcal{E}\times\{1,\dots,p_{\hat{e}}-1\}\,, \tag{2.40}$$

and those which do not. The hanging nodes which do not fulfill (2.40) are called *ficticious nodes*, since only ficticious degrees of freedom are associated to them in the following.

We note the set of all non ficticious edge transformed Gauss-Lobatto points by

$$G_{\mathcal{E},\vec{p}} := \{x_i^{p_e+1,e} \,|\, 0 \le i \le p_e, e \in \mathcal{E} \text{ and } x_i^{p_e+1,e} \text{ is not a ficticious hanging node}\}\,. \quad (2.41)$$

Merging $G_{\mathcal{E},\vec{p}}$ with the tensor product of the Gauss-Lobatto points on the interior of the quadrilaterals, we get the generalization $G_{\vec{p}}$ of G_p from the p uniform mesh,

$$G_{\vec{p}} := G_{\mathcal{E},\vec{p}} \cup \left\{ F_Q\big((\xi_{j_1}^{p_Q+1}, \xi_{j_2}^{p_Q+1})\big) \,|\, Q \in \mathcal{T},\, 1 \le j_1, j_2 \le p_Q - 1 \right\}. \quad (2.42)$$

Analogously to $\Gamma_{D,p}$, we denote the subset of $G_{\vec{p}}$ on the closed Dirichlet boundary by

$$\Gamma_{D,\vec{p}} := \overline{\Gamma}_D \cap G_{\vec{p}}\,.$$

We conclude this section by defining the *hp*-FE spaces and its subsets for the minimization problem given in Theorem 2.7.

Definition 2.27. Let $\mathcal{T}$ be a triangulation of Ω into quadrilaterals Q, possibly with hanging nodes. Let the polynomial degree $p_Q \ge 1$ on Q be given by the vector $\vec{p}$. We define the *hp*-FE space

$$V_{\vec{p}} := V_{\vec{p}}(\mathcal{T}) := \{u \in H^1(\Omega) : u_{|Q} \circ F_Q \in \mathbb{P}^2_{p_Q},\, Q \in \mathcal{T}\}\,.$$

Further, we define the *hp*-FE subsets

$$V_{\vec{p},g_D} := \{w \in V_{\vec{p}} \,|\, w(x) = g_D(x), x \in \Gamma_{D,\vec{p}}\} \subset V_{\vec{p}}\,,$$

and

$$K_{\vec{p},g_D} := \{w \in V_{\vec{p},g_D} \,|\, w(x) \ge \psi(x), x \in G_{\vec{p}}\}.$$

Chapter 3

Error estimates and adaptivity

A posteriori error estimates are widely used in the solution of PDE. Such estimates provide useful indications of the accuracy of a calculation and also provide the basis of adaptive local mesh refinement or local increase of polynomial degree. To this end, they should have at least the two following properties: reliability and efficiency. Here, reliability means that the error estimator provides an upper bound for the error. An error estimator is called efficient, if it provides a lower bound for the error apart from data resolution. Local efficiency is of great interest for adaptive algorithms that involve local mesh refinements. In this chapter, we develop an adaptive *hp*-strategy for variational inequalities. In Section 3.1, we take a look on the literature concerning a posteriori error estimation for the *hp*-version. Section 3.2 is primarily devoted to a posteriori error estimators for *h*-discretizations of variational inequalities. The dual-weighted residual method of error estimation is described in Section 3.3 and applied to the *hp*-FEM. In Section 3.4 we generalize the dual-weighted residual method to the *hp*-version for variational inequalities.

3.1 A posteriori error estimation in the *hp*-FEM

Ainsworth and Oden give a survey of different approaches, namely of explicit residual and implicit residual error estimators, error estimators based on gradient recovery and on hierarchical bases in [AO00]. Comparable literature for high order methods like the *hp*-version FEM is rather scarce. Compared with adaptivity in *h*-FEM, the main difficulty in *hp*-adaptivity arises from the fact that accuracy can be improved by subdividing elements or by increasing the approximation order. Roughly speaking, increasing the polynomial degree is more efficient in regions where the solution is smooth and *h*-refinement is preferable in regions where the solution is non smooth. Accordingly, one looks not only for an error estimator, but also for a local regularity estimator.

In [AS97],[AS98] the a posteriori error estimation of the local error and the local regularity is based on the approximate solution of suitably defined local problems with Neumann type boundary conditions.

Another approach to error estimation is to bound local weighted residuals. An analysis

for meshes consisting of axi-parallel rectangles in two dimensions is given by Bernardi in [Ber96] and extended to meshes containing quadrilaterals and triangles by Melenk and Wohlmuth in [MW01]. Here, the decision whether to subdivide an element Q or to increase its polynomial degree depends on the refinement history of the element which is condensed into a number called *predicted error* η_Q^{pred} of the element Q. η_Q^{pred} can be viewed as an extrapolation of the element error of the preceding refinement step to the current one under the assumption that the solution is locally smooth. If the accuracy in the element Q has to be increased due to the error indicator and the indicator error is larger than the prediction η_Q^{pred}, then an h-refinement is performed since the assumption of local smoothness, which underlies the prediction is wrong. Conversely, if the indicated error is smaller than the predicted one, then the polynomial degree is increased. This hp-adaptive strategy thereby accommodates an implicit estimator of regularity based on the comparison between the error indicator and the predicted error.

A similar adaptive strategy based on the refinement history is employed by Heuveline and Rannacher in [HR03]. But here, the error is estimated by the dual-weighted residual method. Traditionally, a posteriori error estimation in FEM is done with respect to an energy norm $\|\cdot\|_A$ induced by the underlying differential operator. Energy error estimation seems rather generic as it is directly based on the variational formulation of the problem and allows to exploit its natural coercivity properties. Using duality arguments as is common from the a priori error analysis (the so-called *Aubin-Nitsche trick*) the dual-weighted residual method allows to control the error with respect to other quantities of physical interest, e.g. the error at a point $x_0 \in \Omega$ or the $\|\cdot\|_{L^2(\Omega)}$-norm error. The dual-weighted residual method is described in Section 3.3 and generalized to variational inequalities in Section 3.4. The adaptive scheme is reproduced in Algorithm 3.1.

A quite sophisticated hp-adaptive scheme is presented in [DRD02]. There, Demkowicz, Rachowicz, and Devloo observe that optimal refinements of neighboring elements are often in conflict with each other. The minimum rule requests that the order for an edge is set the minimum of the orders for the adjacent elements (cf. Section 2.4.1). If from two neighboring elements one wants to be h- and the other p-refined, the common edge is not refined at all. This implies that the mesh optimization problem may be more complicated than just the choice between the h- or p-refinements. In [DRD02] the error is estimated based on hierarchical bases. Each element is broken into four sons and the polynomial order on these sons is that of its mother increased by one. Having solved the problem on the fine mesh, a new, optimal mesh is constructed by minimizing the coarse grid hp-interpolant of the fine mesh solution. This minimization is done by comparing different coarse meshes generated by h- and p-refinements of the mother cells. The adaptive strategy starts by determining edge h- or p-refinements. Even anisotrop refinements are allowed, i.e., the element Q may be h- or p-refined across only one coordinate axis. With the edge refinement the topology of the mesh is determined and the interior degrees of freedom are adapted to the edges (cf. Section 2.4.1). This approach goes around the problem of estimating the local regularity by comparing h- and p-refinement using the hp-interpolant of the fine mesh solution.

3.2 A posteriori error estimation for variational inequalities

In the book [HHNL88] from Hlaváček et al., the primal formulation of the PDE $-\triangle u + u = f$ on a polygonal domain $\Omega \subset \mathbb{R}^2$ with boundary conditions of Signorini type $u \geq 0$, $\frac{\partial}{\partial n} u \geq 0$, $u \cdot \frac{\partial}{\partial n} u = 0$ on $\partial\Omega$, is approximated by the h-version. Its mutually equivalent dual formulation is approximated on an appropriate subset of the Raviart-Thomas lowest order FE space. The authors show that it is possible to evaluate an error bound for both the primal and the dual approximation using both approximations simultaneously. The disadvantage of this scheme is that an extra global problem must be solved. Further, they do not give local error estimates.

Ainsworth, Oden, and Lee [AOL93] also use the dual variational approach, but try to localize the estimators. However, their upper error bound depends on a sign condition for the element residuals which may fail in practice. Vester considers elliptic obstacle problems and develops an a posteriori error estimate for the h-version on simplex triangulations on domains $\Omega \subset \mathbb{R}^d$, $d \geq 1$. The error estimate is reliable, efficient, and identifies local error contributions. We refer to [CN00] for a short review. The mentioned error estimates are all *asymptotically sharp* in the sense that the convergence of the discrete solution $\|u - u_h\| \to 0$ implies the convergence of the estimate towards 0.

In [Mai01, Lemma 2.5, Theorem 3.7], Maischak generalizes hierarchical bases estimators for the h- and the hp-version to the BEM with Signorini contact using additive Schwarz operators known from preconditioning. Here, the numerical experiments show reasonable refinements towards singularities. But the hp-estimator is not asymptotically sharp.

As already mentioned in Section 3.1, the dual-weighted residual method allows to control the error for more general measures. In [BS00], Blum and Suttmeier carry over this technique to the case of variational inequalities by adapting the duality argument to a h-version for an obstacle problem. We extend this approach to the hp-version in Section 3.4.

3.3 Duality-based adaptivity in the hp-FEM for variational equalities

Now, we describe the approach presented by Heuveline and Rannacher in [HR03]. As we already defined hp-FE subsets for variational inequalities in Section 2.4.1, the adaptive scheme introduced in Algorithm 3.1 can be also used for the treatment of variational inequalities when the error indicator η_Q is adapted. This will be done in Section 3.4.

For sake of simplicity we consider the Poisson equation with mixed Dirichlet-Neumann conditions,

$$-\triangle u = f, \qquad u = 0 \text{ on } \Gamma_D, \qquad \frac{\partial}{\partial n} u = g_N \text{ on } \Gamma_N,$$

on a polygonal, not necessarily convex domain $\Omega \subset \mathbb{R}^2$ with boundary $\partial\Omega = \overline{\Gamma_D \cup \Gamma_N}$. Setting $\rho \equiv 1$ in (1.5) we know from Theorem 1.22 that there exists a unique $u \in V := H^1_{0,D}(\Omega)$ with

$$\langle \nabla u, \nabla v \rangle_{L^2(\Omega)} = \langle f, v \rangle_{L^2(\Omega)} + \langle g_N, \nabla v \rangle_{L^2(\Gamma_N)} \qquad \text{for all } v \in V. \tag{3.1}$$

We define the *hp*-finite element space

$$V_{hp} := V_{\vec{p}}(\mathcal{T}_h) := \{v \in H_0^1(\Omega) \,|\, v_{|Q} \circ F_Q \in \mathbb{P}^2_{p_Q},\ Q \in \mathcal{T}_h\}$$

on the mesh $\mathcal{T}_h$. Here, the vector $\vec{p}$ gives the distribution of the polynomial degrees p_Q on the elements Q. In particular, the meshes are not required to be conform, i.e., the elements Q may possess hanging nodes to facilitate local mesh refinement. The continuity of the discrete functions is yielded as discussed in Section 2.4 and Section 4.6. For practical purposes, we demand that for neighbor elements Q_1, Q_2, the element width h, and the polynomial degree p can only vary within certain limits. Then, the discrete solution $u_{hp} \in V_{hp}$ is determined by

$$\langle \nabla u_{hp}, \nabla v\rangle_{L^2(\Omega)} = \langle f, v\rangle_{L^2(\Omega)} + \langle g_N, v\rangle_{L^2(\Gamma_N)} \qquad \text{for all } v \in V_{hp}. \tag{3.2}$$

We will now derive a posteriori error estimates for the discrete solution of (3.2). Let the goal be to compute the quantity $J(u)$ from the solution u of (3.1) with a prescribed accuracy ε_{Tol} where $J(\cdot) \in V'$ denotes a linear functional. Due to the linearity of J the error $e = u - u_{hp}$ can be estimated by $J(e) = J(u) - J(u_{hp})$. To give examples we mention the following functionals of

$$\text{global average with weight function } \phi \in L^2(\Omega) \qquad J(v) := \langle \phi, v\rangle_{L^2(\Omega)}\,, \tag{3.3a}$$

$$\text{point evaluation} \qquad J_{x_0}(v) := v(x_0)\,, \tag{3.3b}$$

$$\text{square of } L^2 \text{ error} \qquad J_0(v) := \langle u - u_{hp}, v\rangle_{L^2(\Omega)}\,, \tag{3.3c}$$

$$\text{square of } H^1 \text{ error} \qquad J_1(v) := \langle u - u_{hp}, v\rangle_{H^1(\Omega)}\,. \tag{3.3d}$$

The task of computing $J(u)$ from the solution of (3.1) can be rewritten as a trivial constrained optimization problem for $u \in V$:

$$J(u) \le J(v) \quad \text{for all } v \in V \qquad \text{and (3.1) holds.} \tag{3.4}$$

All minima $J(u)$ correspond to stationary points $(u, z) \in V \times V$ of the Lagrangian

$$L(u; z) := J(u) + \langle f, z\rangle_{L^2(\Omega)} + \langle g_N, \nabla z\rangle_{L^2(\Gamma_N)} - \langle \nabla u, \nabla z\rangle_{L^2(\Omega)}$$

with the adjoint variable $z \in V$. Hence, we seek the solutions $(u, z) \in V \times V$ to the Euler-Lagrange system

$$\begin{aligned} \langle \nabla u, \nabla v\rangle_{L^2(\Omega)} &= \langle f, v\rangle_{L^2(\Omega)} + \langle g_N, \nabla v\rangle_{L^2(\Gamma_N)} && \text{for all } v \in V, \\ \langle \nabla v, \nabla z\rangle_{L^2(\Omega)} &= DJ(u; v) && \text{for all } v \in V. \end{aligned}$$

We note that the first equation of this system is the variational equation (3.1). As we assumed J to be a linear functional, we can rewrite the second equation

$$\langle \nabla v, \nabla z\rangle_{L^2(\Omega)} = J(v) \quad \text{for all } v \in V. \tag{3.5}$$

Here, we call $z \in V$ the *dual solution* of the variational equation (3.1). Taking $v = e = u - u_{hp}$ as test function and using Galerkin orthogonality, we obtain the error representation

$$J(e) = \langle \nabla e, \nabla z\rangle_{L^2(\Omega)} = \langle \nabla e, \nabla(z - v_{hp})\rangle_{L^2(\Omega)} \quad \text{for all } v_{hp} \in V_{hp}.$$

Integrating element-wise by parts, we can rewrite

$$J(e) = \sum_{Q \in \mathcal{T}_h} \left(\langle R_{hp}, z - v_{hp} \rangle_{L^2(Q)} + \langle r_{hp}, z - v_{hp} \rangle_{L^2(\partial Q)} \right) \tag{3.6}$$

with the following notation of cell and edge residuals:

$$R_{hp|Q} := f + \triangle u_{hp}, \qquad r_{hp|\Lambda} := \begin{cases} \frac{1}{2}\left[\frac{\partial}{\partial n} u_{hp}\right], & \text{if } \Lambda \subset \partial Q \backslash \partial\Omega, \\ g_N, & \text{if } \Lambda \subset \Gamma_N, \\ 0, & \text{if } \Lambda \subset \Gamma_D. \end{cases} \tag{3.7}$$

where $\left[\frac{\partial}{\partial n} u_{hp}\right]$ denotes the jump of the normal derivative across an element boundary.

We use the error representation (3.6) as the basis for error control. Its evaluation requires us to generate approximations to the dual solution z with accuracy better than that of $z_{hp} \in V_{hp}$ obtained from

$$\langle \nabla v, \nabla z_{hp} \rangle_{L^2(Q)} = J(v) \qquad \text{for all } v \in V_{hp}.$$

Usually we yield the approximation $\tilde{z}$ to z by post-processing $z_{hp} \in V_{hp}$ using locally higher-order interpolation or defect correction. Let the post-processed approximation to z denoted by $\tilde{z}$. Then the error is controlled by the error estimator

$$\eta_\omega(u_{hp}) := \Big| \sum_{Q \in \mathcal{T}_h} \bar{\eta}_Q \Big| \tag{3.8}$$

with the local contributions

$$\bar{\eta}_Q := \langle R_{hp}, \tilde{z} - i_p \tilde{z} \rangle_{L^2(K)} + \langle r_{hp}, \tilde{z} - i_p \tilde{z} \rangle_{L^2(\partial K)} .$$

The adaptation process defined by Algorithm 3.1 is driven by the absolute values of the local contributions $\eta_Q := |\bar{\eta}_Q|$. The philosophy of the algorithm is that it is more economical to raise the polynomial degree p_Q rather than to reduce the element size h_Q. Here, the decision whether to increase p_Q or to subdivide Q into four elements by bisection of the edges depends on the success of the last refinement (see *Step 3*). The mesh and the distribution of the polynomial degrees are constructed by a series of adaption cycles such that at the final stage the local error will be equilibrated, i.e., $\eta_Q \approx \frac{\varepsilon_{\mathrm{TOL}}}{\operatorname{card} \mathcal{T}}$.

Algorithm 3.1 Adaption process for the *hp*-FEM
Let a tolerance $\varepsilon_{TOL} > 0$, a refinement threshold σ, $0 < \sigma < 1$, an initial mesh $\mathcal{T}_0$ with mesh-size distribution h_Q, and polynomial degrees $p_Q \in \mathbb{N}$ for all $Q \in \mathcal{T}_0$ be given. Furthermore, let $V_{\vec{p}}(\mathcal{T}_0)$ the respective conforming FE space.

Starting with $i = 0$, the sequence of meshes $\mathcal{T}_i$, $i = 1, 2, \ldots$ with corresponding distributions h_Q and p_Q for $Q \in \mathcal{T}_i$ is constructed by the following process:

1. Compute the FEM solutions $u_i := u_{hp} \in V_{hp} := V_{\vec{p}}(\mathcal{T}_i), \qquad z_i := z_{hp} \in V_{hp}$.

2. Evaluate the local error contributions $\bar{\eta}_Q$ for all $Q \in \mathcal{T}_i$ and $\eta_\omega(u_{hp})$ (see (3.8)). If $\eta_\omega(u_{hp}) < \varepsilon_{TOL}$, then exit with u_{hp}.

3. Order the elements Q according to the size of $\eta_Q := |\bar{\eta}_Q|$. If

$$\eta_Q < \sigma \frac{\varepsilon_{TOL}}{\operatorname{card} \mathcal{T}_i},$$

   then skip to the next element Q, else increase $\dim V_{hp}$ according to the following scheme:

   (a) The element Q and its polynomial degree p_Q were left unchanged in the preceding cycle. Then, leave Q unchanged and increase p_Q to $p_Q + 1$.

   (b) The element Q was left unchanged in the preceding cycle, but p_Q was increased. If

$$\eta_Q < h_Q \eta_Q^{old},$$

   increase p_Q to $p_Q + 1$, else refine Q into 4 elements by bisection of the edges.

   (c) The element Q was obtained by refinement of a mother element $\bar{Q} \in \mathcal{T}_{i-1}$. If

$$\eta_Q < 2^{-p_Q} \eta_{\bar{Q}}^{old},$$

   increase p_Q to $p_Q + 1$, else refine Q into 4 elements by bisection of the edges.

4. Create the conforming FE space $V_{\vec{p}}(\mathcal{T}_{i+1})$ according to the new elements and their polynomial distribution (cf. Section 2.4.1). Increase i to $i+1$. Continue with *Step 1*.

3.4 Duality-based adaptivity in the *hp*-FEM for variational inequalities

We will now derive an a posteriori error estimate for variational inequalities. Again, we have to cope with the difficulty that $K_p \not\subseteq K$ is not satisfied in general. This inconsistency is treated analogously to the approach presented in [BS00]. Corollary 3.2 represents a statement which estimates the error by summing up local contributions. Thus, it allows local refinements.

For ease of presentation we consider the obstacle problem with the Laplacian on a poly-

gonal, not necessarily convex domain $\Omega \subset \mathbb{R}^2$,

$$-\triangle u - f \geq 0, \quad u \geq \psi, \quad (u-\psi)(-\triangle u - f) = 0 \quad \text{on } \Omega,$$

with mixed Dirichlet-Neumann conditions,

$$u = 0 \text{ on } \Gamma_D, \qquad \tfrac{\partial}{\partial n} u \geq g_N \text{ on } \Gamma_N, \qquad (u-\psi)(\tfrac{\partial}{\partial n} u - g_N) = 0 \text{ on } \Gamma_N.$$

Using the abstract notation

$$B(\cdot,\cdot) := \langle \nabla\cdot, \nabla\cdot \rangle_{L^2(\Omega)} \qquad \text{and} \qquad F(\cdot) := \langle f, v \rangle_{L^2(\Omega)} + \langle g_N, \nabla v \rangle_{L^2(\Gamma_N)} \tag{3.9}$$

we may write the obstacle problem equivalently to Theorem 1.23:

$$\text{Find } u \in K \text{ such that} \qquad B(u, v-u) \geq F(v-u) \quad \text{for all } v \text{ in } K \tag{3.10}$$

where $K := \{v \in V \,|\, v \geq \psi \text{ on } \Omega\}$ and $V := H_0^1(\Omega)$.
We introduce the *hp*-FE subset $K_{hp} := \{v \in V_{hp} \,|\, v(x) \geq \psi(x) \text{ for all } G_{\vec{p}}\}$ with $G_{\vec{p}}$ as defined in Section 2.4.1 and approximate $u \in K$ by the solution of the discrete problem:

$$\text{Find } u_{hp} \in K_{hp} \text{ such that} \qquad B(u_{hp}, v - u_{hp}) \geq F(v - u_{hp}) \quad \text{for all } v \in K_{hp}\,. \tag{3.11}$$

Motivated by the duality-based a posteriori error estimator for the variational equality (3.1) we look for the dual-like solution z of the variational problem:

$$\text{Find } z \in Z \text{ such that} \qquad J(v-z) \leq B(v-z, z) \quad \text{for all } v \in Z. \tag{3.12}$$

The idea is to give an upper bound of the error $J(e)$, $e := u - u_{hp}$, by defining an appropriate closed convex subset $Z \subset V$ below. If we assume $\bar{v} := z + e \in Z$, we get the estimate $J(e) \leq B(e, z)$.

To avoid misunderstandings, we note that the dual-like solution z from (3.12) should not be understood analogously to (3.4), (3.5) as the solution of a dual problem in the sense of generalized Kuhn-Tucker theory for optimization problems under inequality constraints. Nevertheless, we call the error estimator dual-based since it generalizes the concept of the dual weighted residual estimator to variational inequalities.

To handle the nonconformity $K_{hp} \not\subset K$ of the approximation subset, we modify the variational problem (3.12) and redefine z as the solution of the problem:

$$\text{Find } z \in Z \text{ such that} \qquad J(v + \delta - z) \leq B(v - z, z) \quad \text{for all } v \in Z. \tag{3.13}$$

Here, we allow $\delta := \psi - i_p\psi$ to be non zero. Redefining the test function $\bar{v} := z + e - \delta$ we obtain the estimate

$$\begin{aligned} J(e) &\leq B(u - u_{hp} - \delta, z) \\ &= B(u - u_{hp}, z - z_{hp}) + B(u - u_{hp}, z_{hp}) - B(\delta, z) \quad \text{for all } z_{hp} \in V_{hp}. \end{aligned} \tag{3.14}$$

In case of a variational equality the second term would be zero due to Galerkin orthogonality. We utilize this by introducing the auxiliary problem:

$$\text{Find } \bar{u} \in V \text{ such that} \qquad B(\bar{u}, v) = F(v) \quad \text{for all } v \in V. \tag{3.15}$$

This means that $\bar{u}$ is the solution of the obstacle free problem.

Now, using the modified variational inequality (3.13) and $\bar{u}$ from the auxiliary problem (3.15), we can state the following error estimate.

Theorem 3.1. We define the set of discrete contact by $\Psi_{hp} := \{x \in \Omega \,|\, u_{hp}(x) \le i_p\psi(x)\}$ and the set of admissible functions by

$$Z := \{v \in V \,|\, v \ge 0 \text{ on } \Psi_{hp} \text{ and } B(\bar{u} - u, v + u_{hp} - u + \delta) \ge 0\}\,.$$

Let $z \in Z$ be the solution of the variational inequality (3.13) and $\bar{u} \in V$ be the solution of the variational equality (3.15). Then, there exists a $z_{hp} \in V_{hp}$ such that

$$J(e) \le B(\bar{u} - u_{hp}, z - z_{hp}) - B(\psi - i_p\psi, z)\,.$$

Proof. We can use (3.14) when we assume that the test-function $\bar{v} = z + u - u_{hp} - \delta$ belongs to Z for all $z \in Z$. We insert the solution $\bar{u}$ of the auxiliary problem (3.15) into (3.14) to obtain

$$J(e) \le B(u - \bar{u}, z - z_{hp}) + B(\bar{u} - u_{hp}, z - z_{hp}) + B(u - u_{hp}, z_{hp}) - B(\delta, z). \tag{3.16}$$

Thus, noting $\delta = \psi - i_p\psi$, it remains to show that the first and the third term of (3.16) are non-positive for an appropriate $z_{hp} \in V_{hp}$, i.e.,

$$\begin{aligned} 0 &\ge B(u - \bar{u}, z - z_{hp}) + B(u - u_{hp}, z_{hp}) \\ &= B(\bar{u} - u_{hp}, z_{hp}) + B(u - \bar{u}, z - u + u_{hp} + \delta) + B(u - \bar{u}, u - (u_{hp} + \delta)). \end{aligned} \tag{3.17}$$

Replacing $F(\cdot)$ in the continuous and discrete variational inequalities (3.10), (3.11) by $B(\bar{u}, \cdot)$ from the auxiliary problem (3.15) yields

$$B(\bar{u} - u, v - u) \le 0 \qquad \text{for all } v \in K, \tag{3.18}$$

$$B(\bar{u} - u_{hp}, v - u_{hp}) \le 0 \qquad \text{for all } v \in K_{hp}. \tag{3.19}$$

Now, let $W^{\psi}_{hp} := \{v \in V \,|\, v(x) \ge i_p\psi(x) \text{ on } \Psi_{hp} \cap G_{\vec{p}}\} \cap V_{hp}$, and W^0_{hp} like W^{ψ}_{hp} using $\psi \equiv 0$. We consider the variational problem

$$\text{Find } \tilde{u}_{hp} \in W^{\psi}_{hp} \quad \text{such that} \qquad B(\bar{u} - \tilde{u}_{hp}, v - \tilde{u}_{hp}) \le 0 \qquad \text{for all } v \in W^{\psi}_{hp}. \tag{3.20}$$

and see that $\tilde{u}_{hp} = u_{hp}$, since the active constraints of (3.19) coincide with the constraints given by W^{ψ}_{hp}.

Choosing $z_{hp} \in W^0_{hp}$ yields $v = u_{hp} + z_{hp} \in W^{\psi}_{hp}$. It follows with (3.20) that the first term of (3.17) is non positive. The second term of (3.17) is non negative due to the definition of Z. Observing that $v = u_{hp} + \delta \in K$, it follows from (3.18) that the last term of(3.17) is non positive.

It remains to justify the assumption $\bar{v} \in Z$. We have $u \ge \psi$ and $u_{hp} + \delta \le i_p\psi + \delta = \psi$ on Ψ_{hp}, henceforth, $\bar{v} \ge 0$ on Ψ_{hp}.

Choosing $v = u_{hp} + \delta \in K$ in (3.18) we obtain $0 \le B(\bar{u} - u, u - (u_{hp} + \delta))$. From $z \in Z$ we know $0 \le B(\bar{u} - u, z + u_{hp} - u + \delta)$. Adding both inequalities with the substitution $z = \bar{v} - u + u_{hp} + \delta$ yields

$$0 \le B(\bar{u} - u, \bar{v} + u_{hp} - u + \delta)\,,$$

i.e., $\bar{v} \in Z$. □

Analogously to (3.6) element-wise integration by parts yields

$$B(\bar{u}-u_{hp},z-z_{hp})=\sum_{Q\in\mathcal{T}_h}\Big(\langle R_{hp},z-z_{hp}\rangle_{L^2(Q)}+\langle r_{hp},z-z_{hp}\rangle_{L^2(\partial Q)}\Big)$$

with the cell and edge residuals with respect to $\bar{u}_{hp}$,

$$R_{hp|Q}:=f+\triangle\bar{u}_{hp},\qquad r_{hp|\Lambda}:=\begin{cases}\frac{1}{2}\big[\frac{\partial}{\partial n}\bar{u}_{hp}\big], & \text{if } \Lambda\subset\partial Q\backslash\partial\Omega,\\ g_N, & \text{if } \Lambda\subset\Gamma_N,\\ 0, & \text{if } \Lambda\subset\Gamma_D.\end{cases}\tag{3.21}$$

Here, $\big[\frac{\partial}{\partial n}\bar{u}_{hp}\big]$ denotes the jump of the normal derivative across an element boundary. Using this notation we can rewrite the estimate of Theorem 3.1 as

Corollary 3.2. With the assumptions of Theorem 3.1 there holds

$$|J(e)|\le\sum_{Q\in\mathcal{T}_h}\Big(\langle R_{hp},z-z_{hp}\rangle_{L^2(Q)}+\langle r_{hp},z-z_{hp}\rangle_{L^2(\partial Q)}+Ch_Qp_Q^{-k}\|\psi\|_{H^2(Q)}\|z\|_{H^k(Q)}\Big)$$

for all $z_{hp}\in W_{hp}^0$ with $k=1$. Here, C denotes a positive constant independent of h_Q, p_Q, ψ, and z. Further, the estimate holds with $k=2$ when $z\in H^2(Q)$.

Proof. It remains to consider

$$-B(\psi-i_p\psi,z)=-\langle\nabla(\psi-i_p\psi),\nabla z\rangle_{L^2(\partial Q)}$$

from the estimate of Theorem 3.1. For $k=1$ the estimate follows using the Cauchy-Schwarz inequality and Theorem 2.3 for all $z\in Z$.

Noting that $\psi-i_p\psi=0$ on ∂Q due to the definition of the interpolation operator i_p, again, element-wise partial integration yields $-B(\psi-i_p\psi,z)=\langle\psi-i_p\psi,\triangle z\rangle_{L^2(Q)}$. Thus, the estimate for $k=2$ follows using the Cauchy-Schwarz inequality and Theorem 2.3 again. □

Corollary 3.2 provides a posteriori estimates for arbitrary functionals. Its evaluation requires us to generate approximations to $z\in Z$ given by the dual-based problem (3.13). In particular, this means that we have to approximate Z. A heuristic idea is to calculate approximations to u, $\bar{u}$, and z in a locally refined superspace $\tilde{V}_{hp}$ of V_{hp}. Denoting these approximations by $\tilde{u}_{hp}$, $\tilde{\bar{u}}_{hp}$, and $\tilde{z}_{hp}$, respectively, Z may be replaced by

$$\tilde{Z}:=\{v\in\tilde{V}_{hp}\,|\,v(x)\ge0\text{ on }\tilde{\Psi}_{hp}\cap\tilde{G}_{\vec{p}}\text{ and }B(\tilde{\bar{u}}_{hp}-\tilde{u}_{hp},v+u_{hp}-\tilde{u}_{hp}+\tilde{i}_p\psi-i_p\psi)\ge0\}.$$

Here, $\tilde{\Psi}_{hp}$, $\tilde{G}_{\vec{p}}$, and $\tilde{i}_p$ are defined analogously to Ψ_{hp}, $G_{\vec{p}}$, and i_p with respect to the refined locally superspace $\tilde{V}_{hp}$. Taking $z_{hp}=\max\{i_p\tilde{z}_{hp},0\}$ allows to calculate the a posteriori error according to Corollary 3.2.

If we assume that the contact zone $\Psi=\{x\in\Omega\,|\,u(x)=\psi(x)\}$ is simply connected, another heuristic approach can be employed to approximate z. Taking Υ_{hp} as the convex hull of the contact nodes $\{x\in G_{\vec{p}}\,|\,u_{hp}(x)=\psi(x)\}$ we may approximate z by the solution $\tilde{z}$ of the variational equation

$$B(\tilde{z},v)=\langle f,v\rangle_{L^2(\Omega\backslash\Upsilon_{hp})}.\tag{3.22}$$

To justify this heuristic, we assume that the contact zone is sufficiently well resolved by Υ_{hp}, i.e., $\Upsilon_{hp} \approx \Psi$. Using the complementary condition $-\triangle u - f > 0$ on Ψ, $-\triangle u - f = 0$ on $\Omega \backslash \Psi$ (cf. Theorem 1.23), the second condition in the definition of Z (see Theorem 3.1) may be approximated as follows

$$\begin{aligned} 0 \;&\leq\; B(\bar{u} - u, v + u_{hp} - u + \delta) \\ &\approx\; \int_{\Upsilon_{hp}} (f + \triangle u)\,(v + u_{hp} - u + \delta)\,\mathrm{d}x \quad \approx\; \int_{\Upsilon_{hp}} (f + \triangle u)\, v\,\mathrm{d}x\,. \end{aligned}$$

Using the assumption $-\triangle u - f > 0$ on Υ_{hp} this implies $v \leq 0$ on Υ_{hp}. Hence, together with the first condition in Z we can approximate Z by $\tilde{Z} = \{v \in V_{hp} \,|\, v(x) = 0 \text{ on } \Upsilon_{hp} \cap G_{\bar{p}}\}$.

This means that we only have to solve a Dirichlet problem on $\Omega \backslash \Upsilon_{hp}$ with homogenous boundary data to get an approximation to $z \in Z$. Equivalently, we can solve the variational equation (3.22).

Chapter 4

Solving discrete nonlinear problems

In Chapter 2, we defined FE-subsets which allow to solve variational inequalities by computing the minimum of a functional subject to equality and inequality constraints. This Chapter addresses the implementation of the discrete minimum problems given by Theorem 2.6 and Theorem 2.7. In Section 4.1, we introduce a basis B of V_p for the implementation which allows to control easily the constraint conditions of V_{p,g_D} and K_{p,g_D}. Furthermore, the discrete minimization problems of Theorem 2.6 and Theorem 2.7 can be transfered into the problem of minimizing a nonlinear function $A : \mathbb{R}^N \to \mathbb{R}$ under constant equality and inequality constraints. In the beginning, we consider only the quasi-uniform quadrilateral mesh without hanging nodes introduced in Section 2.1.

For the unconstrained minimization of the nonlinear function $A : \mathbb{R}^N \to \mathbb{R}$ we present the *inexact Newton backtracking method* in Section 4.2. In Section 4.3, we specify the large-scale nonlinear minimizer given by Felkel in [Fel99] to our strictly convex minimization problem with constant equality and inequality constraints. The treatment of the unbounded and the bounded constrained discrete nonlinear problems demands preconditioned conjugated gradient iterations to solve linear systems. Section 4.4 and Section 4.5 analyze the influence of the polynomial degree p on the costs of the iterative solving of the linear systems.

In Section 4.6, we show how the continuity across inter-element boundaries between quadrilateral meshes with hanging nodes and with different polynomial degrees can be ensured by means of linear algebra. In Section 4.7, we apply the h-version and the p-version on uniform quadrilateral meshes to model obstacle problems given by the minimal surface operator with homogeneous and inhomogeneous Dirichlet boundary data.

4.1 Basis functions

Let ξ_i^{p+1}, $0 \le i \le p$, be the Gauss-Lobatto points of degree p (see Definition 2.1) and

$$\lambda_i^p(\xi) := \begin{cases} \prod_{\substack{k=0 \\ k\neq i}}^{p} \frac{\xi-\xi_k^{p+1}}{\xi_i^{p+1}-\xi_k^{p+1}} & \text{for } \xi \in [-1,1], \\ 0 & \text{for } \xi \in \mathbb{R}\backslash[-1,1], \end{cases} \qquad 0 \le i \le p,$$

the respective Lagrangian interpolation polynomials. We define the basis $\tilde{B}_p := (b_{ij} \,|\, 0 \le i,j \le p)$ on the closed reference square $\overline{\tilde{Q}}$ by the tensor product polynomials $b_{ij}((\xi_1,\xi_2)) := \lambda_i^p(\xi_1)\lambda_j^p(\xi_2)$. Using the transformations F_Q we get the local bases

$$B_Q := (b_{Q,i,j} \,|\, 0 \le i,j \le p) \qquad \text{with } b_{Q,i,j} := \begin{cases} b_{ij}(F_Q^{-1}(x)) & \text{for } x \in \overline{Q}, \\ 0 & \text{for } x \in \Omega\backslash\overline{Q}, \end{cases} \tag{4.1}$$

on the quadrilaterals Q. Introducing a global counting $k = 1, \dots N$, $N := \operatorname{card} G_p$, for the $x_k \in G_p$ (see Definition2.2) we define the global basis functions

$$b_k := \sum_{(Q,i,j)\in X_k} b_{Q,i,j} \qquad \text{with } X_k := \{(Q,i,j) \,|\, Q \in \mathcal{T}, 0 \le i,j \le p, b_{Q,i,j}(x_k) = 1\}. \tag{4.2}$$

Using the vector notation $\underline{w} := (w_k)_{k=1\dots N}$, $\underline{b} := (b_k)_{k=1\dots N}$, and the product notation

$$\underline{w}^T\underline{b} = \sum_{k=1}^{N} w_k b_k$$

we can rewrite V_p, V_{p,g_D}, and K_{p,g_D} as

$$\begin{aligned} V_p &:= \{\underline{w}^T\underline{b} \,|\, \underline{w} \in \mathbb{R}^N\}, \\ V_{p,g_D} &:= \{\underline{w}^T\underline{b} \,|\, \underline{w} \in \mathbb{R}^N_{=\underline{g}_D}\}, \\ K_{p,g_D} &:= \{\underline{w}^T\underline{b} \,|\, \underline{w} \in \mathbb{R}^N_{=\underline{g}_D} \cap \mathbb{R}^N_{\ge\underline{\psi}}\} \end{aligned}$$

where

$$\begin{aligned} \mathbb{R}^N_{=\underline{g}_D} &:= \{\underline{w} \in \mathbb{R}^N \,|\, w_k = g_D(x_k) \text{ for all } x_k \in \Gamma_{D,p}\}, \\ \mathbb{R}^N_{\ge\underline{\psi}} &:= \{\underline{w} \in \mathbb{R}^N \,|\, w_k \ge \psi(x_k) \text{ for all } x_k \in G_p\}. \end{aligned}$$

Now, we define $A : \mathbb{R}^N_{=\underline{g}_D} \cap \mathbb{R}^N_{\ge\underline{\psi}} \to \mathbb{R}$ by

$$A(\underline{w}) := A(\underline{w}^T\underline{b}). \tag{4.3}$$

We can use the definitions of A on $\mathbb{R}^N$ and of A on V_p both simultaneously when we use the arguments $\underline{w} \in \mathbb{R}^N$ and $w \in V_p$ to separate between them. We reformulate the discrete obstacle problem from Theorem 2.7(i) equivalently as the minimization problem

$$\underline{u} \text{ minimizes } A \text{ on } \mathbb{R}^N_{=\underline{g}_D} \cap \mathbb{R}^N_{\ge\underline{\psi}}, \qquad \text{i.e.,} \quad A(\underline{u}) \le A(\underline{v}) \quad \text{for all } \underline{v} \in \mathbb{R}^N_{=\underline{g}_D} \cap \mathbb{R}^N_{\ge\underline{\psi}}. \tag{4.4}$$

Here, it is worth to note that the function A on $\mathbb{R}^N$ inherits the convexity from A on V_p.

The problem of minimizing a convex form subject to lower or upper bounds is an active field of research (cf. [NW99], [CGT96]). When A is of quadratic type

$$A(\underline{v}) = \tfrac{1}{2}\underline{v}^T B\, \underline{v} - \underline{v}^T \underline{f}$$

where $B \in \mathbb{R}^{N\times N}$ and $\underline{f} \in \mathbb{R}^N$, the problem is known as quadratic programming problem. It can be solved by relaxation methods (cf. [Glo84, Chapter V]) or a generalized conjugate gradient algorithm (cf. [O'L80]), known as Polyak algorithm. Here, it is worth to note that the classical active-set iterative schemes based on the dual reformulation of the quadratic programming problem change the active set, i.e., the index set of components with $\underline{u}_i = \underline{\psi}_i$, slowly, usually by a single index at each iteration. As a result, the number of iterations blows up on large-scale problems.

In the next two subsections we introduce a scheme tailored on the solution of the above problem. It can be viewed as a generalization of [O'L80] for nonlinear, but symmetric positive definite functions or as a specification of the algorithm given in [Fel99]. Felkel solves non convex nonlinear large-scale box-constrained problems by estimating the eigenvectors of the Hessian of A. In case of the above mentioned obstacle problem it suffices to consider functions with symmetric positive definite Hessians. This allows us to use preconditioned conjugate gradient iterations to solve the linear subproblems.

4.2 Unconstrained nonlinear problems

Looking back on Lemma 1.21 and Theorem 2.6 we know that the Hessian $\nabla^2 A(\underline{v})$ is symmetric positive definite on $\mathbb{R}_{=g_D}$ and it suffices to find $\underline{u} \in \mathbb{R}_{=g_D}$ with

$$\nabla A(\underline{u}) = 0$$

due to the uniqueness of $\underline{u}$. Thus, we formulate an inexact Newton method which uses preconditioned conjugate gradient iterations to solve the linear subproblems approximately. Since the initial $\underline{u}_0$ is not always near the solution $\underline{u}$ the usual local Newton method is globalized by backtracking.

To simplify notation, we write $\underline{g}(\underline{v}) := \nabla A(\underline{v})$ for the gradient and $H(\underline{v}) := \nabla^2 A(\underline{v})$ for the Hessian in the following. The inexact backtracking method is defined by Algorithm 4.2. Its backtracking line-search goes back on Goldstein and Armijo (cf. [NW99]), and makes use of the minimization context using zero order information of the problem. We introduce the Armijo-Goldstein line-search by the following algorithm.

Algorithm 4.1 $\sigma = \text{linesearch}(f, \sigma_0, \delta_1, \beta_1)$; Line-search

Given a differentiable function $f : \mathbb{R}\geq 0 \to \mathbb{R}$, an initial step width σ_0 and parameters $0 < \delta_1 < 1$, $0 < \beta_1 < 1$.

1. Set $j = 0$, $\sigma = \sigma_0$.
2. Do While $f'(0) \neq 0$ and $f(0) - f(\sigma_0\beta_1^j) \leq \delta_1\,\sigma_0\beta_1^j|f'(0)|$, i.e., non-sufficient decrease

   Set $\sigma = \sigma_0\,\beta_1^j$ and $j = j + 1$.
3. Exit with σ.

Algorithm 4.2 can be viewed as a usual Newton iteration scheme

$$\underline{u}_{k+1} = \underline{u}_k - \big(H(\underline{u})\big)^{-1}\underline{g}(\underline{u}_k), \qquad k \in \mathbb{N},$$

used to determine the zero of $\underline{g}(\underline{v})$, if we do not perform *Steps 4,5*, set $\sigma = 1$ in *Step 6* and omit the line search.

The algorithm is motivated by the goal of avoiding unnecessary conjugate gradient iterations and line-search steps in the beginning of the minimization process when high accuracy is not needed. *Step 2* performs conjugate gradient iterations until the linear problem $H(\underline{u}_k)\underline{y} = -\underline{g}(\underline{u}_k)$ is solved numerically with a relative error with respect to $\|\underline{g}(\underline{u}_k)\|$ less than η_k.

The solution $\underline{y}_i$ is taken as search direction $\underline{s}_k$ in *Step 3*. *Step 4* defines the one dimensional minimum problem and its derivatives in the starting point of the line-search. The derivatives are used to calculate an appropriate initial step-width σ_0 for the line-search with Algorithm 4.1. Further, $q'_{\underline{s}_k}(0)$ is needed for the stopping criterion of the line-search. *Step 7* determines how accurate the linear problem $H(\underline{u}_k)\underline{y} = -g(\underline{u}_k)$ must be solved by the conjugate gradient method in the next execution of *Step 2*.

In contrast to the inexact Newton method, the backtracking method checks if the Newton steps $\underline{y}_i$ have a good length and scales them according to line-search. This seems reasonable, since the line-search can be applied cheaply in comparison to the computation of $\underline{y}_i$. The choice of the line-search initial σ_0 results from the assumption that $q_{\underline{s}_k}(t) = A(\underline{u}_k + t\underline{s}_k)$ can be approximated by the quadratic model $\tilde{q}(t) = \alpha t^2 + \beta t + \gamma$ defined by $\tilde{q}(0) = q_{\underline{s}_k}(0)$, $\tilde{q}'(0) = q'_{\underline{s}_k}(0)$, and $\tilde{q}''(0) = q''_{\underline{s}_k}(0)$. If $\tilde{q}''(0) > 0$, the model $\tilde{q}(t)$ is minimal in $t = \sigma_0 = -q'_{\underline{s}_k}(0)/q''_{\underline{s}_k}(0)$. To analyze the convergence properties of the above algorithm we introduce the following concepts of convergence.

Definition 4.1. Let $(\underline{u}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$ be a sequence converging to $\underline{u} \in \mathbb{R}^N$. Then $(\underline{u}_k)$ converges to $\underline{u}$

(i) with q-order p, $p \geq 1$, if there exists a positive constant C such that

$$\|\underline{u}_{k+1} - \underline{u}\| \leq C\|\underline{u}_k - \underline{u}\|^p\,,$$

(ii) with r-order α if there exists a sequence $(\xi_k)_{k\in\mathbb{N}} \subset \mathbb{R}$ converging with q-order α to zero such that

$$\|\underline{u}_k - \underline{u}\| \leq \xi_k\,.$$

Algorithm 4.2 $\underline{u}_k = \text{inbm}(\underline{u}_0, \|\cdot\|, \epsilon, \eta_0, \gamma, \alpha, M)$; Inexact Newton backtracking method

Given the starting point $\underline{u}_0 \in \mathbb{R}^N$ and a norm $\|\cdot\|$ on $\mathbb{R}^N$. Furthermore, we have a termination parameter $\epsilon > 0$, an initial threshold parameter $0 \leq \eta_0 < 1$, and parameters $0 \leq \gamma \leq 1$, $1 < \alpha \leq 2$ for the calculation of the threshold parameters η_k. M denotes the preconditioner for the preconditioned conjugate gradient iterations.
To simplify notation, we write $\underline{g}(\underline{v}) := \nabla A(\underline{v})$ for the gradient and $H(\underline{v}) := \nabla^2 A(\underline{v})$ for the Hessian.

1. Set $k = 0$.

2. Set i=0.
   Set initial vector $\underline{y}_0$ (e.g. $= 0$) for preconditioned conjugate gradient method.
   Do While $\|\underline{g}(\underline{u}_k) + H(\underline{u}_k)\,\underline{y}_i\| > \eta_k \|\underline{g}(\underline{u}_k)\|$
   Compute the next M-preconditioned conjugate gradient iteration $\underline{y}_{i+1}$ (see Algorithm A.1).
   Set $i = i + 1$.
   End while

3. Set search direction $\underline{s}_k = \underline{y}_i$.

4. Set $q_{s_k}(t) := A(\underline{u}_k + t\underline{s}_k)$.
   Compute $q'_{s_k}(0) = \underline{s}_k^T \underline{g}(\underline{u}_k)$ and $q''_{s_k}(0) = \underline{s}_k^T H(\underline{u}_k)\underline{s}_k$.

5. Set $\sigma_0 = \begin{cases} \min\left\{1, -\frac{q'_{s_k}(0)}{q''_{s_k}(0)}\right\} & \text{if } q''_{s_k}(0) > 0 \\ 1 & \text{if } q''_{s_k}(0) = 0. \end{cases}$

6. Determine step length $\sigma_k = \sigma = \text{linesearch}(q_{s_k}, \sigma_0, \delta_1, \beta_1)$ with Algorithm 4.1.
   If $q_{s_k}(1) \leq q_{s_k}(\sigma)$, Set $\sigma = 1$.
   Set $\underline{u}_{k+1} = \underline{u}_k + \sigma \underline{s}_k$.

7. If $\|\underline{g}(\underline{u}_{k+1})\| > \epsilon$
   Set
   $$\eta_{k+1} = \gamma \min\left\{1, \left(\frac{\|\underline{g}(\underline{u}_k)\|}{\|\underline{g}(\underline{u}_{k-1})\|}\right)^{\alpha}\right\} \cdot \eta_k \,. \tag{4.5}$$
   Set $k = k + 1$. Continue with *Step 2*.

8. Exit with $\underline{u}_k$.

Choosing the forcing parameter η_{k+1} as in (4.5) Eisenstat and Walker proved the following theorem for the sequence $(\underline{u}_k)_{k\in\mathbb{N}}$ produced by Algorithm 4.2 with the substitution $\sigma = 1$ in *Step 6*.

Theorem 4.2. Let $\|\cdot\|$ be the norm of Algorithm 4.2 on $\mathbb{R}^N$ and the induced norm on $\mathbb{R}^{N\times N}$. Let $M := \max\{\|H(\underline{u})\|, \|(H(\underline{u}))^{-1}\|\}$. For $\delta > 0$, define

$$B_\delta(\underline{v}) := \{\underline{w} \in \mathbb{R}^N \mid \|\underline{w} - \underline{v}\| < \delta\}$$

and let $\delta_{\underline{u}} > 0$ sufficiently small that

(i) $\underline{g}$ is continuously differentiable and H is nonsingular on $B_{\delta_{\underline{u}}}(\underline{u})$,

(ii) $\|(H(\underline{v}))^{-1}\| \leq 2M$ for all $\underline{v} \in B_{\delta_{\underline{u}}}(\underline{u})$,

(iii) $\|H(\underline{v}) - H(\underline{u}\| \leq \kappa\|\underline{v} - \underline{u}\|$ for all $\underline{v} \in B_{\delta_{\underline{u}}}(\underline{u})$ holds for a constant $\kappa \geq 0$,

(iv) $\delta_{\underline{u}} \leq \frac{2}{\kappa M}$.

If $\underline{u}_0 \in B_{\delta_{\underline{u}}}(\underline{u})$, then the sequence $\{\underline{u}_k\}_k$ produced by the inexact Newton method, i.e., by Algorithm 4.2 with $\sigma = 1$ in *Step 6*, remains in $B_{\delta_{\underline{u}}}(\underline{u})$ and converges to $\underline{u}$.

Let γ and α be the parameters used in Algorithm 4.2. If $\gamma < 1$, then the convergence is of q-order α. If $\gamma = 1$, then the convergence is of r-order α and of q-order p for every $p \in [1, \alpha)$.

Proof. See [EW96, Theorem 2.3]. □

Corollary 4.3. Theorem 4.2 holds for $\{\underline{u}_k\}_k$ produced by Algorithm 4.2.

Proof. The line search in *Step 6* enhances the convergence of the Newton method because it guarantees a sufficient decrease of A. □

Remark 4.4. Controlling the assumptions of Theorem 4.2 in *Step 4* of the algorithm, the line searches of *Step 6* can be omitted, if $\underline{u}_0 \in B_{\delta_{\underline{u}}}(\underline{u})$, especially, if computation of the Hessian is cheap. If the gradient evaluations are much cheaper than the Hessian evaluations, line search is preferable.

Remark 4.5. Taking $\|\underline{v}\| := |v|_{1,\Omega}$ with $v = \sum_1^N v_i\lambda_i$ in Algorithm 4.2, we have $M = \max\{\kappa_u, \kappa_l^{-1}\}$ with Lemma 1.21. Due to (1.13) $|v|_{1,\Omega}$ can be estimated by the form $D^2A(u; v, v) = \underline{v}^T H(\underline{u})\, \underline{v}$.

In case that the assumptions of Theorem 4.2 are not fulfilled, r-linear convergence can still be proved using the convergence analysis on line search algorithms by [OR70]. With [OR70] we have the following definition and lemmas.

Lemma 4.6. Let $A : D \subset \mathbb{R}^N \to \mathbb{R}$ be continuously differentiable on the open convex set D, and assume that $L := L(A(\underline{u}_0)) := \{\underline{v} \in D \,|\, A(\underline{v}) \leq A(\underline{u}_0)\}$ is compact. Consider the iterations of Algorithm 4.2 producing the sequence $\underline{u}_k$, where

$$q'_{s_k}(0) = \underline{g}^T(\underline{u}_k)\underline{s}_k \leq 0\,, \quad \underline{s}_k \neq 0 \qquad \text{(cf. } \textit{Step 4}\text{)}.$$

Then $(\underline{u}_k)_k \subset L$ and

$$\lim_{k\to\infty} \frac{g^T(\underline{u}_k)\underline{s}_k}{\|\underline{s}_k\|} = 0\,.$$

Proof. See [OR70, 14.2.15]. □

Definition 4.7. A mapping σ is a *forcing function* (F-function) if for any sequence $(t_k)_k \subset [0, \infty)$

$$\lim_{k\to\infty} \sigma(t_k) = 0 \quad \text{implies} \quad \lim_{k\to\infty} t_k = 0\,.$$

Definition 4.8. Let $g : D \subset \mathbb{R}^N \to \mathbb{R}$ be Fréchet differentiable on D and let $(\underline{u}_k)_{k\in\mathbb{N}} \subset D$ be a given sequence. Then a sequence $(\underline{s}_k)_{k\in\mathbb{N}} \subset \mathbb{R}^N$of nonzero vectors is *gradient-related to* $(\underline{u}_k)_k$ if there is a F-Function φ such that

$$\frac{|\underline{g}^T(\underline{u}_k)\underline{s}_k|}{\|\underline{s}_k\|} \geq \varphi(\|\underline{g}(\underline{u}_k)\|) \quad \text{for all } k \in \mathbb{N}\,.$$

Lemma 4.9. Let A, D, L, and $(\underline{u}_k)_k$ be defined as in Lemma 4.6. Assume that A has a unique critical point $\underline{u}$ in L. Assume, finally, that the sequence $(\underline{s}_k) \in \mathbb{R}^N$ of nonzero vectors is gradient-related to $\underline{u}_k$, then

$$\lim_{k\to\infty} \underline{u}_k = \underline{u}\,.$$

Proof. See [OR70, Theorem 14.3.2]. □

The next lemma states a F-function φ such that the sequence $(\underline{s}_k)_k$ is gradient related to $(\underline{u}_k)_k$, both sequences produced by Algorithm 4.2.

Lemma 4.10. Let A, D, L, and $(\underline{u}_k)_k$, $(\underline{s}_k)_k$ be defined as in Lemma 4.6. Again, for ease of notation, let $\underline{g}(\underline{v}) := \nabla A(\underline{v})$ and $H(\underline{v}) := \nabla^2 A(\underline{v})$ and assume that the Hessians $H(\underline{v})$ are bounded by $\|H(\underline{v})\| \leq \mu_2$ and are positive definite with $\mu_1\|\underline{v}\|^2 \leq \underline{v}^T H(\underline{v})\,\underline{v}$ for all $\underline{v} \in D$, where $0 < \mu_1, \mu_2$ are constants. Finally, assume $\eta_k \leq \frac{1}{2}\mu_1^2\,\mu_2^{-2}$. Then we have

$$-\frac{\underline{g}^T(\underline{u}_k)\underline{s}_k}{\|\underline{s}_k\|} \geq \varphi(\|\underline{g}(\underline{u}_k)\|) \qquad \text{with } \varphi(t) := \frac{\mu_1^2}{2\mu_2^2(1+\eta_k)}\,t, \quad t \in [0,\infty)\,.$$

Here, φ is a F-function and $(\underline{g}(\underline{v}_k))_k$ is gradient related to $(\underline{s})_k$.

Proof. Since $k \in \mathbb{N}$ can be considered fixed, we abbreviate $\underline{g} = \underline{g}(\underline{u}_k)$, $\underline{s} = \underline{s}_k$, and $H = H(\underline{u}_k)$. Due to the strictly positive definiteness of H there exists a $\underline{d} \in \mathbb{R}^N$ such that

$$-\underline{s} = H^{-1}(\underline{g} + \underline{d}) = \hat{\underline{g}} + \hat{\underline{d}}\,. \tag{4.6}$$

Using the *While* condition of Algorithm 4.2 *Step 2* we estimate

$$\|\underline{g} + H\underline{s}\| = \|\underline{d}\| \leq \eta_k\|\underline{g}\|\,. \tag{4.7}$$

The Cauchy Schwarz inequality and, again, the strictly positive definiteness of H yield

$$\|\underline{w}\|\,\|H\underline{w}\| \geq \mu_1\|\underline{w}\|^2\,, \qquad\qquad \|H\underline{w}\| \geq \mu_1\|\underline{w}\|\,,$$

and

$$\|\underline{w}\| \geq \|HH^{-1}\underline{w}\| \geq \mu_1\|H^{-1}\underline{w}\|\,. \tag{4.8}$$

The boundedness of H provides

$$\|\underline{w}\| \leq \|HH^{-1}\underline{w}\| \leq \|H\|\|H^{-1}\underline{w}\| \leq \mu_2\|H^{-1}\underline{w}\|\,.$$

Using (4.6) and the above inequalities we estimate

$$\|\hat{\underline{g}}\| \geq \mu_2^{-1}\|\underline{g}\| \qquad \text{and} \qquad \|\hat{\underline{d}}\| \leq \mu_1^{-1}\|\underline{d}\|\,. \tag{4.9}$$

Combining the inequalities (4.6), (4.8), and (4.7) we obtain

$$\|\underline{s}\| = \| - H^{-1}\,(\underline{g} + \underline{d})\| \leq \mu_1^{-1}\|g + d\| \leq \mu_1^{-1}(1 + \eta_k)\|g\|\,. \tag{4.10}$$

Combining (4.9), (4.8) and (4.7) yields the following chain of inequalities

$$\begin{aligned} -\underline{g}^T\underline{s} \quad &= \underline{g}^T H^{-1}(\underline{g} + \underline{d}) = \hat{\underline{g}}^T H^T\,(\hat{\underline{g}} + \hat{\underline{d}}) \\ &\geq \|\hat{\underline{g}}^T H^T\,\hat{\underline{g}}\| - \|\hat{\underline{g}}^T H^T\,\hat{\underline{d}}\| \\ &\geq \mu_1\|\hat{\underline{g}}\|^2 - \|\underline{g}\,\hat{\underline{d}}\| \\ &\geq \mu_1\mu_2^{-2}\|\underline{g}\|^2 - \eta_k\mu_1^{-1}\|\underline{g}\|^2\,. \end{aligned}$$

Using (4.10) and the assumption $\eta_k \leq \frac{1}{2}\mu_1^2\mu_2^{-2}$ we conclude the proof with

$$-\underline{g}^T\underline{s} \;\geq\; (\mu_1^2\mu_2^{-2} - \eta_k)(1 + \eta_k)^{-1}\|\underline{g}\|\|\underline{s}\| \;\geq\; \varphi(\|\underline{g}\|)\,\|\underline{s}\|\,.$$

□

For the inexact backtracking method we summarize the convergence properties in the next theorem.

Theorem 4.11. Let A be given by (4.3) and let $\underline{u} \in \mathbb{R}^N$ be the unique solution of the minimization problem $A(\underline{u}) \leq A(\underline{v})$ for all $\underline{v} \in \mathbb{R}^N$. Furthermore, let $\underline{u}_0 \in \mathbb{R}^N$ and the sequences $(\underline{u}_k)_{k\to\infty}$, $(\underline{s}_k)_{k\to\infty}$, and $(\sigma_k)_{k\to\infty}$ be generated by Algorithm 4.2. Then

$$\lim_{k\to\infty} \frac{\underline{g}^T(\underline{u}_k)s_k}{\|s_k\|} = 0\,, \tag{4.11}$$

$$\lim_{k\to\infty} \sigma_k\|s_k\| = 0\,, \tag{4.12}$$

and we have the r-linear convergence of $\underline{u}_k$ against $\underline{u}$.

Proof. We know that there is a unique $\underline{u}$ with $\nabla\underline{u} = 0$ and that it satisfies $A(\underline{u}) \leq A(\underline{v})$ for all $v \in \mathbb{R}^N$. We may take $D := B_\delta(\underline{u})$ as the open convex set of Lemmata 4.6, 4.9, and 4.10 with δ such that $\underline{u}_0 \in D$. Therefore, Lemma 4.6 yields (4.11).

By construction of Algorithm 4.2 there is $\sigma_k \leq \sigma_0$. Using Lemma 4.10 and (4.10) we obtain

$$|\underline{g}^T(\underline{u}_k)\underline{s}_k| \geq \rho\|\underline{s}_k\|^2 \qquad \text{with} \quad \rho := \frac{\mu_1^3}{2\mu_2^2(1 + \eta_k)^2}\,.$$

We estimate for $\sigma_0 = 1$

$$\sigma_k\|\underline{s}_k\| \;\leq\; \sigma_0\|\underline{s}_k\| \;\leq\; \rho^{-1}\,\frac{|\underline{g}^T(\underline{u}_k)\underline{s}_k|}{\|\underline{s}_k\|}$$

which implies (4.12) due to (4.11).
For $\sigma_0 < 1$ we substitute $\sigma_0 = -q'_{\underline{s}_k}(0)/q''_{\underline{s}_k}(0)$ (see Algorithm 4.2 *Steps 4, 5*) and get

$$\sigma_k\|\underline{s}_k\| \;\leq\; \sigma_0\|\underline{s}_k\| \;=\; \frac{-\underline{g}^T(\underline{u}_k)\,\underline{s}_k}{\underline{s}_k^T H(\underline{u}_k)\,\underline{s}_k}\,\|\underline{s}_k\| \;=\; \frac{-\underline{g}^T(\underline{u}_k)\,\underline{s}_k}{\|\underline{s}_k\|}\,\frac{\|\underline{s}_k\|^2}{\underline{s}_k^T H(\underline{u}_k)\,\underline{s}_k} \leq \mu_1^{-1}\,\frac{|\underline{g}^T(\underline{u}_k)\,\underline{s}_k|}{\|\underline{s}_k\|}\,.$$

Here, the last inequality follows with μ_1 from the assumption on the positive definiteness of $H(\underline{v})$ for all $\underline{v} \in D$ in Lemma 4.10. The convergence of $\underline{u}_k$ towards $\underline{u}$ follows from Lemma (4.9), the r-linear convergence is obtained by [OR70, 14.3.6]. □

We conclude the convergence properties of Algorithm 4.2 stated in Theorem 4.11 and Theorem 4.2, respectively:

(i) We have r-linear convergence of $\underline{u}_k$ towards $\underline{u}$.

(ii) Setting $\alpha = 2$ in the algorithm, we have q-quadratic convergence of $\underline{u}_k$ towards $\underline{u}$, when $\underline{u}_0$ is near to $\underline{u}$.

4.3 Bounded constrained nonlinear problems

It is the intention of this section to introduce a solver for the *large-scale minimization problem with equality and inequality constraints* (4.4), i.e.,

$$\begin{aligned} &\min \quad A(\underline{u}) \\ &\text{subject to } u_i(\underline{u}) = g_D(x_i), \qquad x_i \in \Gamma_{D,p}, \\ &\text{subject to } u_i(\underline{u}) \geq \psi(x_i), \qquad x_i \in G_p. \end{aligned}$$

For an overview on the subject of nonlinear minimization with side conditions we refer to [NW99, Chapter 16] here. The *generalized conjugated gradient algorithm* proposed by [O'L80] represents an efficient method for *quadratic programming functionals.* Considering our large-scale nonlinear convex programming problem the optimization literature shows that the *projected gradient method* should be chosen because it is designed to make rapid changes to the active set.

The following algorithm combines the unconstrained minimizer of the previous section and the projected gradient method. Thereby, the gradient projection steps suggest a set of *active components* i which fulfill the constraint condition with $v_i = \psi_i$. Then the unconstrained minimizer of Algorithm 4.2 is applied to the *non active* or *free components* to explore the affine subspace

$$\{\underline{w} = \underline{\psi} + \underline{v} \,|\, \underline{v} \in \mathbb{R}^N \text{ with } v_i = 0 \text{ for all active components } i\}$$

where $\underline{\psi}$ is given by $\psi_i = \psi(x_i)$, $x_i \in G_p$. The affine subspace can be viewed as a face of the half hyper-plane $\mathbb{R}_{\geq \underline{\psi}}$.

Firstly, we describe the *gradient projection method* which realizes steepest descent steps for the constrained problem. Therefore, we introduce the *projection* $P_{\underline{\psi}}$ and the *projected gradient* $\nabla_{\underline{\psi}}$. We recall necessary conditions of a bounded constrained minimum (Lemma 4.12) and sufficient conditions (Remark 4.13).

Let $P_{\underline{\psi}} : \mathbb{R}^N \to \mathbb{R}^N_{\geq \underline{\psi}}$ denote *the projection onto the feasible region* $\mathbb{R}^N_{\geq \underline{\psi}}$ defined component wise by

$$(P_{\underline{\psi}}(\underline{v}))_i := \begin{cases} \psi_i, & \text{if } v_i \leq \psi_i, \\ v_i, & \text{if } v_i > \psi_i \,. \end{cases} \tag{4.13}$$

Let $\underline{s} \in \mathbb{R}^N$ and let $\underline{v}$ be a feasible point, i.e., $\underline{v} \in \mathbb{R}^N_{\geq \underline{\psi}}$, with at least one non active component. Then $P_{\underline{\psi}}(\underline{v} + t\underline{s})$, $t \geq 0$, defines a polygon. Now, it is the idea of the gradient projection method to set $s = -\nabla A$ and to seek for the minimum of the constrained problem by an approximate minimum search on this polygon. This means that we start our search in the direction of steepest descent.

Considering

$$\frac{A(P_{\underline{\psi}}(\underline{v} + t\underline{s})) - A(P_{\underline{\psi}}(\underline{v}))}{t}$$

leads us to the definition of the *projected gradient on* $\mathbb{R}_{\geq \underline{\psi}}$

$$(\nabla_{\underline{\psi}} A(\underline{v}))_i := \begin{cases} \min\{0, (\nabla A(\underline{v}))_i\}\,, & \text{if } v_i \leq \psi_i, \\ (\nabla A(\underline{v}))_i\,, & \text{if } v_i > \psi_i, \end{cases} \tag{4.14}$$

if we insert the coordinate vectors for s. With the notations of the projection and the projected gradient we can write the following equivalence which gives a necessary condition for a minimum.

Lemma 4.12. Let $\underline{u} \in \mathbb{R}^N_{\geq \underline{\psi}}$ and A be continuously differentiable on $\mathbb{R}^N_{\geq \underline{\psi}}$. Then, the following statements are equivalent.

$$\nabla^T A(\underline{u})\,(\underline{v} - \underline{u}) \geq 0 \quad \text{for all } \underline{v} \in \mathbb{R}^N_{\geq \underline{\psi}}\,, \tag{4.15a}$$

$$\nabla_{\underline{\psi}} A(\underline{u}) = 0\,, \tag{4.15b}$$

$$\underline{u} - P_{\underline{\psi}}(\underline{u} - \nabla A(\underline{u})) = 0\,. \tag{4.15c}$$

Proof. We consider the three equations for active and non active components. Let i be an active component, i.e., $u_i = \psi_i$. Let $\underline{v}$ in (4.15a) be given by $v_j = u_j$ for $j \neq i$ and $v_j = u_j + 1$ for $j = i$. Then $\nabla^T A(\underline{u})\,(\underline{v} - \underline{u}) = (\nabla A(\underline{u}))_i \geq 0$ which implies $(\nabla_{\underline{\psi}} A(\underline{u}))_i = \min\{0, \nabla A(\underline{u}))_i\} = 0$ and $(P_{\underline{\psi}}(\underline{u} - \nabla A(\underline{u})))_i = \psi_i = u_i$. Thus, we obtain (4.15b) and (4.15c). Reversely, (4.15c) implies $(\nabla A(\underline{u}))_i \geq 0$ and with that (4.15b).

Now, let i be a non active component, i.e., $u_i > \psi_i$. Taking $\underline{v}$ as above we get again $(\nabla A(\underline{u}))_i \geq 0$. If we take $\underline{v}$ with $v_i = u_i$ for $j \neq i$ and $\underline{v}_j = \psi_i$ for $j = i$, we get $(\nabla A(\underline{u}))_i \leq 0$, which implies $(\nabla A(\underline{u}))_i = 0$ and henceforth, (4.15b), (4.15c). Reversely, (4.15c) implies $(\nabla A(\underline{u}))_i = 0$ and with that (4.15b).

For $\underline{v} \in \mathbb{R}^N_{\geq \underline{\psi}}$ we have $v_i - u_i \geq 0$ for active components i. Combining the above implications of (4.15c) for active and non active components yields $(\nabla A(\underline{u}))_i\,(v_i - u_i) \geq 0$. Adding up these component products, we obtain (4.15a) . □

Remark 4.13. In case of the elliptic obstacle problem as defined above, we know that the solution $\underline{u}$ of (4.15a) is the unique minimizer of the bounded constrained problem. Thus, we know that the solution $\underline{u}$ of (4.15c) solves the bounded constrained problem. To verify the condition $\|\underline{u} - P_{\underline{\psi}}(\underline{u} - \nabla A(\underline{u}))\| \leq \varepsilon$ instead of $\|\nabla_{\underline{\psi}} A(\underline{u})\| \leq \varepsilon$ in Algorithm 4.3 has numerical advantages, because $(\underline{u} - P_{\underline{\psi}}(\underline{u} - \nabla A(\underline{u})))_i$ is continuous in ψ_i. In a more general context of constrained minimization the statements of Lemma 4.12 are necessary, but not sufficient for a minimizer.

In [CM87, Theorem 3.4], it is proved that the successive application of the minimization rule

$$\underline{u}_{k+1} = P_{\underline{\psi}}(\underline{u}_k - \sigma_k \nabla A(\underline{u}_k))$$

where σ_k satisfies the Goldstein-Armijo line-search of Algorithm 4.1 generates a sequence $(\underline{u}_k)_k$ with $\lim_{k\to\infty}\|\nabla_{\underline{\psi}} A(\underline{u}_k)\| = 0$ when $\underline{u}_k$ is bounded. This minimization rule is realized by *Step 3* of Algorithm 4.3. As in Algorithm 4.2, the choice of the Goldstein-Armijo line-search initial $\sigma_{P,0}$ results from the assumption that $q(t) := h(t, \underline{s}_k) = A(P_{\underline{\psi}}(\underline{u}_k + t\underline{s}_k))$ can be approximated by the quadratic model $\tilde{q}(t) = \alpha t^2 + \beta t + \gamma$ defined by $\tilde{q}(0) = q(0)$, $\tilde{q}'(0) = q'(0)$, and $\tilde{q}''(0) = q''(0)$. If $h_{tt}(0, \underline{s}_k) > 0$, the model $\tilde{q}(t)$ is minimal in $t = \sigma_{P,0} = -h_t(0, \underline{s}_k)/h_{tt}(0, \underline{s}_k)$.

Secondly, we try to enhance the performance of the projected gradient method by minimizing A on the free variables under the assumption that these remain free. Using this assumption we can perform unconstrained minimization steps in the sense of the inexact Newton backtracking method (see Algorithm 4.2). We use the following notation in the algorithm.

Let $\underline{v} \in \mathbb{R}^N$ and $\mathcal{I} \subset \{1, \ldots, N\}$ an index set. We define the $(\underline{v}, \mathcal{I})$-*reduced function* $A_{\underline{v},\mathcal{I}} : \mathbb{R}^{\operatorname{card}\mathcal{I}} \to \mathbb{R}$ by

$$A_{\underline{v},\mathcal{I}}(\underline{w}) := A(\underline{v}_{\mathcal{I}}(\underline{w})) \quad \text{with} \quad \underline{v}_{\mathcal{I}}(\underline{w}) := \begin{cases} w_i & \text{if } i \in \mathcal{I}, \\ v_i & \text{otherwise.} \end{cases} \tag{4.16}$$

Noting the *reduced identity matrix*

$$I_{\mathcal{I}} := \big(e_j\big)_{j\in\mathcal{I}} \in \mathbb{R}^{N\times \operatorname{card}\mathcal{I}} \quad \text{with the column unit vectors } \underline{e}_j \in \mathbb{R}^N,$$

we obtain the *gradient* and *Hessian* of the $(v, \mathcal{I})$-reduced function,

$$\nabla A_{\underline{v},\mathcal{I}}(\underline{w}) = I_{\mathcal{I}}^T \, \nabla A(\underline{v}_{\mathcal{I}}(\underline{w})) \qquad \text{and} \qquad \nabla^2 A_{\mathcal{I}}(\underline{w}) = I_{\mathcal{I}}^T \, \nabla^2 A(\underline{v}_{\mathcal{I}}(\underline{w})) \, I_{\mathcal{I}} \,, \tag{4.17}$$

respectively. We name

$$\mathcal{F}(\underline{v}) := \{i \in \{1, \ldots, N\} \,|\, v_i > \psi_i\} \,.$$

the *set of free variables (non-active set)* and its complement

$$\mathcal{A}(\underline{v}) := \{1, \ldots, N\} \backslash \mathcal{F}(\underline{v}) \; = \{i \in \{1, \ldots, N\} \,|\, v_i \le \psi_i\} \,.$$

the *active set*. Using the notation $\underline{v}_{\mathcal{I}}(\underline{w})$ of (4.16), the projection $P_{\underline{\psi}}$ from (4.13) can be rewritten as

$$P_{\underline{\psi}} = \underline{\psi}_{\mathcal{F}(\underline{v})}(\underline{v}) \,.$$

It is the idea of Algorithm 4.3 to start the Newton iteration for the actually free variables $\mathcal{F}(\underline{u}_k)$ and to go on with this iterations until the iteration $\underline{u}_{k+i} = \underline{\psi}_{\mathcal{F}(\underline{u}_k)}(\underline{w}_i)$, $i \in \mathbb{N}$, either violates a constraint condition or satisfies the stopping criterion $\|\nabla A_{\underline{u}_k,\mathcal{F}(\underline{u}_k)}(\underline{w}_i)\| \le \epsilon$. The Newton iterations are performed by *Steps 6–9* of Algorithm 4.3.

If the set of free variables $\mathcal{F}(\underline{u})$ of the minimum $\underline{u}$ is already identified correctly at iteration k, i.e., $\mathcal{F}(\underline{u}) = \mathcal{F}(\underline{u}_k)$, then the Newton steps yield the minimum $\underline{u}$ due to Theorem 4.2, Theorem 4.11, and the iteration loop stops due to the stopping criterion (see *Step 8,* Lemma 4.12).

If the iteration $\underline{u}_{k+i}$ violates a constraint condition, i.e., $\underline{u}_{k+i} \notin \mathbb{R}^N_{\geq \underline{\psi}}$ and $\mathcal{F}(\underline{u}_{k+i}) \not\subseteq \mathcal{F}(\underline{u}_k)$, then we switch to the projected gradient search.

But how can we be sure that the set of free variables $\mathcal{F}(\underline{u}_k)$ is big enough and is not reduces to an empty set due to violations of constraints too early? Here, we demand a sufficient decrease of A with respect to the free variables when the Newton iterations are applied. If the Newton iteration does not yield a sufficient decrease in comparison to its predecessors or already yielded the minimum $\min\{A_{\underline{u}_k,\mathcal{F}(\underline{u}_k)}(\underline{v}) \,|\, \underline{v} \in \mathbb{R}^N\}$ with respect to the free variables, we switch to the projected gradient method.

In Algorithm 4.3 the switching to the projected gradient method due to constraint violation, insufficient decrease, or a reached minimum with respect to the free variables is controlled by *Steps 9, 10*. To take advantages of the excellent convergence properties of Newton's method as much as possible, it is useful to choose the initial $\underline{u}_0$ such that all variables are free, i.e., $\mathcal{F}(\underline{u}_0) = \{1, \ldots, N\}$.

It remains to give a criterion for switching from projected gradient iterations to Newton iterations. Here, we assume that projected gradient iterations have found the correct set of free variables $\mathcal{F}(\underline{u})$ when $\mathcal{F}(\underline{u}_k)$ has not been changed for the last l iterations, i.e., $\mathcal{F}(\underline{u}_{k+1}) = \mathcal{F}(\underline{u}_k) = \cdots = \mathcal{F}(\underline{u}_{k+1-l})$. If the assumption $\mathcal{F}(\underline{u}) = \mathcal{F}(\underline{u}_k)$ is true, Newton's method will find the minimum quickly. If the assumption is not fulfilled, this will be detected by the criterion realized in *Step 9* and projected gradient iterations will be used further. The criterion $\mathcal{F}(\underline{u}_{k+1}) = \mathcal{F}(\underline{u}_k) = \cdots = \mathcal{F}(\underline{u}_{k+1-l})$ is controlled by *Step 5*. As a second criterion we use a sufficient decrease of A in comparison to the preceding decreases (see *Step 5*).

Algorithm 4.3 can be viewed more generally as a scheme of the kind given in [CM87, Algorithm 5.3]. Mutually, starting with an initial $\underline{u}_0 \in \mathbb{R}^N_{\geq \underline{\psi}}$, $\underline{u}_{k+1}$ follows from $\underline{u}_k$ recursively by the application of one of the two minimization rules:

1. Let $\underline{u}_{k+1} = P_{\underline{\psi}}(\underline{u}_k - \sigma_k \nabla A(\underline{u}_k))$ where σ_k satisfies the Goldstein-Armijo line-search of Algorithm 4.1.

2. Determine $\underline{u}_{k+1} \in \mathbb{R}^N_{\geq \underline{\psi}}$ such that $A(\underline{u}_{k+1}) < A(\underline{u}_k)$ and $\mathcal{A}(\underline{u}_{k+1}) \subset \mathcal{A}(\underline{u}_k)$.

The first minimization rule is realized in *Step 3* of Algorithm 4.3, the second in *Steps 6–9*. In [CM87, Theorem 3.4] it is proved that the exclusive application of the first minimization rule generates a sequence $(\underline{u}_k)_k$ with $\lim_{k\to\infty} \|\nabla_{\underline{\psi}} A(\underline{u}_k)\| = 0$ when $\underline{u}_k$ is bounded. It is the goal of the second rule to accelerate this convergence by solving unconstrained subproblems over the free variables.

[CM87, Theorem 5.4] shows that a bounded sequence generated by a combination of the above minimization rules identifies the active set of the minimum $\mathcal{A}(\underline{u})$ after a finite number of iterations whenever the *strict complementary condition* holds in each stationary point $\underline{u}$, i.e.,

$$(\nabla A(\underline{u}))_i > 0 \quad \text{if } u_i = \psi_i \,. \tag{SCC}$$

When the correct active set $\mathcal{A}$ is detected, i.e., $\mathcal{A}(\underline{u}_{k+1}) = \mathcal{A}(\underline{u}_k)$ for all following k (condition of *Step 9*), the minimization remains as an unconstrained minimization problem

Algorithm 4.3 $\underline{u}_k = \text{pginbm}(\underline{u}_0, \underline{\psi}, \|\cdot\|, \epsilon, \mu_1, \mu_2, \eta_0, \gamma, \alpha, M, l)$;
Projected gradient inexact Newton backtracking method

Let an initial $\underline{u}_0 \in \mathbb{R}^N_{\geq \underline{\psi}}$ and a norm $\|\cdot\|$ be given. We have a termination parameter $\epsilon > 0$, decrease thresholds $0 < \mu_1, \mu_2 < 1$, and a number $2 \leq l \in \mathbb{N}$ of active sets to compare. Furthermore, we hand over a preconditioner M, and the parameters $0 \leq \gamma < 1$, $1 < \alpha \leq 2$, for the calculation of the threshold η_i needed by the inner inbm-iterations with Algorithm 4.2.
To ease notation, we write $\underline{g}(\underline{v}) := \nabla A(\underline{v})$, $H(\underline{v}) := \nabla^2 A(\underline{v})$ for the Hessian, and $\underline{g}_{\underline{\psi}} := \nabla_{\underline{\psi}} A$ for the projected gradient.

1. Set $k = 0$.

2. If $\|\underline{g}_{\underline{\psi}}(\underline{u}_k)\| < \epsilon$
   Exit with $\underline{u}_k$.
   Else
   Continue with *Step 6.*

3. Set $k_P = k$.
   Set the search direction $\underline{s}_k = -\underline{g}(\underline{u}_k)$.
   Set $h(t, \underline{s}_k) := A(P_{\underline{\psi}}(\underline{u}_k + t\,\underline{s}_k))$, i.e., restrict A onto a polygonal search path given by $\underline{s}_k$ and $P_{\underline{\psi}}$.

   Compute derivative $h_t(0, \underline{s}_k) = \underline{s}_k^T\, \underline{g}_{\underline{\psi}}(\underline{u}_k) = -\|\underline{g}_{\underline{\psi}}(\underline{u}_k)\|_2^2$.

   Compute derivative $h_{tt}(0, \underline{s}_k) = \underline{g}_{\underline{\psi}}(\underline{u}_k)^T H(\underline{u}_k)\, \underline{g}_{\underline{\psi}}(\underline{u}_k)$.

   Set
   $$\sigma_{P,0} = \begin{cases} \min\left\{1, -\frac{h_t(0,\underline{s}_k)}{h_{tt}(0,\underline{s}_k)}\right\} & \text{if } h_{tt}(0, \underline{s}_k) > 0 \\ 1 & \text{if } h_{tt}(0, \underline{s}_k) = 0. \end{cases}$$
   Determine step length $\sigma_{P,k} = \sigma_P = \text{linesearch}(h(\cdot, \underline{s}_k), \sigma_{P,0}, \delta, \beta)$ with Algorithm 4.1.
   Set $\underline{u}_{k+1} = P_{\underline{\psi}}(\underline{u}_k + \sigma_P \underline{s}_k)$.

4. If $\|\underline{g}_{\underline{\psi}}(\underline{u}_{k+1})\| < \epsilon$
   Set $k = k + 1$.
   Exit with $\underline{u}_k$.

5. If $\mathcal{A}(\underline{u}_{k+1}) = \mathcal{A}(\underline{u}_k) = \cdots = \mathcal{A}(\underline{u}_{k+1-l})$
   or $A(\underline{u}_k) - A(\underline{u}_{k+1}) \leq \mu_1 \max\{A(\underline{u}_j) - A(\underline{u}_{j+1}) \,|\, k_P \leq j < k\}$
   Set $k = k + 1$. Continue with *Step 6.*
   Else
   Set $k = k + 1$. Continue with *Step 3.*

6. Set $\mathcal{F} = \mathcal{F}(\underline{u}_k)$. Set $i = 0$. Define the $\mathcal{F}$-reduced function $A_{\underline{\psi},\mathcal{F}}$.
   Set $\underline{w}_0 = I_{\mathcal{F}}^T \underline{u}_k$.

7. Compute one iteration $\underline{w}_{i+1} = \text{inbm}(\underline{w}_i, \|\cdot\|, \epsilon, \eta_i, \gamma, \alpha, M)$ with Algorithm 4.2 using $A_{\underline{\psi},\mathcal{F}}$ instead of A, $k = i$ instead of $k = 0$.
   Set $\underline{u}_{k+1} = P_{\underline{\psi}}(\underline{\psi}_{\mathcal{F}}(I_{\mathcal{F}}\, \underline{w}_{i+1}))$.

continued on next page

continued from previous page

1. If $\|\underline{g}_{\underline{\psi}}(\underline{u}_{k+1})\| < \epsilon$

   Set $k = k + 1$.
   Exit with $\underline{u}_k$.

2. If $\mathcal{A}(\underline{u}_{k+1}) = \mathcal{A}(\underline{u}_k)$
   and $A_{\underline{u}_k,\mathcal{F}}(\underline{w}_i) - A_{\underline{u}_k,\mathcal{F}}(\underline{w}_{i+1}) > \mu_2 \max\{A_{\underline{u}_k,\mathcal{F}}(\underline{w}_j) - A_{\underline{u}_k,\mathcal{F}}(\underline{w}_{j+1}) \,|\, 0 \le j < i\}$
   and $\|\nabla A_{\underline{u}_k,\mathcal{F}}(\underline{w}_{i+1})\| \ge \epsilon$

   Set $k = k + 1$, $i = i + 1$.
   Continue with *Step 7*.

3. Set $k = k + 1$.
   Continue with *Step 3*.

over the free variables. As this is solved with Algorithm 4.2 by construction, then we have the convergence of Theorem 4.2 and Theorem 4.11.

Thus, taking into account that A is strongly elliptic and the discrete minimization problem of Theorem 2.7(i) has a unique critical point $\underline{u}$, we obtain the following theorem.

Theorem 4.14. Let the sequence $(\underline{u}_k)_{k\to\infty}$ be generated by Algorithm 4.3. If the (SCC) holds in the unique critical point $\underline{u}$ of the minimization problem (4.4), we have $\lim_{k\to\infty} \underline{u}_k = \underline{u}$. Further, there exists a $k_0 \in \mathbb{N}$ such that $\mathcal{A}(\underline{u}_k) = \mathcal{A}(\underline{u})$ for all $k \ge k_0$, i.e., the active set of $\underline{u}$ is detected in a finite number of iterations. $\underline{u}_k$ converges to $\underline{u}$ for $k \ge k_0$ as stated in Theorem 4.11.

Unfortunately, Example 4.15 shows that we can not assume the (SCC) to be always fulfilled. However, Remark 4.16 offers a work around to this problem.

Example 4.15. Let $\underline{u}$ be the minimizer of $A(\underline{v}) := v_1^2 + v_2^2$, $\underline{v} \in \mathbb{R}_{\underline{\psi}}$ with $\underline{\psi} := \left(\begin{smallmatrix} -1 \\ 0 \end{smallmatrix}\right)$. Then, A is strongly elliptic, but the unique minimizer $\underline{u} = \left(\begin{smallmatrix} 0 \\ 0 \end{smallmatrix}\right)$ does not fulfill the (SCC).

Remark 4.16. If a strongly elliptic A does not fulfill the (SCC) in the active component i, we know that $(\nabla A(\underline{u}))_i = 0$. If we replace the active set $\mathcal{A}(\underline{v})$ by the *binding set*

$$\tilde{\mathcal{A}}(\underline{v}) := \left\{ i \in \{1, \dots, N\} \,\middle|\, v_i = \psi_i \ \text{ and } \ (\nabla A(\underline{v}))_i > 0 \right\}$$

and the set of free variables $\mathcal{F}(\underline{v})$ by the *augmented set of free variables*

$$\tilde{\mathcal{F}}(\underline{v}) := \left\{ i \in \{1, \dots, N\} \,\middle|\, v_i > \psi_i \ \text{ or } \ \left(v_i \le \psi_i \ \text{ and } \ (\nabla A(\underline{v}))_i \le 0\right) \right\},$$

then the projected gradient steps consider only active variables which fulfill the (SCC). Thus, the binding set $\tilde{\mathcal{A}}(\underline{v})$ is identified in a finite number of steps (cf. [CM87, Theorem 4.2]).

Remark 4.17. A standard in mathematical programming software is the optimization package `LANCELOT A` from Conn, Gould and Toint [CGT92] and its successor `GALAHAD` [CGT02]. In the early 90th `LANCELOT A` was the first and only method to solve large-scale

problems. The `LANCELOT A` routines treat the convex bounded constrained nonlinear programming mutually equivalently to Algorithm 4.3. The main differences between the algorithm and the `LANCELOT A` implementation is that `LANCELOT A` handles non-convex problems additionally. The software checks the convexity of the problem by a trust region method and allows only small steepest descent steps when non-convexity is detected. Switching off the trust region control by tuning the respective parameters to a large trust region which will not be reduced in the outer iterations, gives an algorithm that behaves similar to Algorithm 4.3.

In [CGT02, §1], the authors conjecture that another approach, called *sequential quadratic programming (SQP)*, will be more successful in the long term, when high-quality quadratic programming codes for large-scale problems exist. This is the case for small- and medium-scaled problems ($N \leq 2000$). The results of comparative tests of other software packages (`SNOPT`, see Gill, Murray, and Saunders [GMS02], `LOCO`, see Vanderbei and Shanno [SV00], `KNITRO`, see Byrd, Hribar, and Nocedal [BHN99], and `FilterSQP`, see Fletcher and Leyffer [FL02]) against `LANCELOT A` showed `LANCELOT A` often (but not always) being significantly out performed, mainly in case of small- and medium-scaled problems.

Algorithm 5.2.1 presented in [Fel99] is similar to Algorithm 4.3, but allows the treatment of non-convex nonlinear programming problems. It uses a *truncated Lanczos decomposition* instead of the conjugate gradient method to compute search directions for the line-search of the minimum and a second line-search algorithm for the lines where the objective function owns a negative curvature.

4.4 Solving the linear systems of the unconstrained problem

The treatment of the unbounded and the bounded discrete nonlinear problems demands preconditioned gradient iterations to solve linear systems $H(\underline{u}_k)\,\underline{y} = -\underline{g}(\underline{u}_k)$ approximately until the relative residual of the i-th iteration satisfies

$$\|\underline{g}(\underline{u}_k) + H(\underline{u}_k)\underline{y}_i\| / \|\underline{g}(\underline{u}_k)\| \leq \eta := \eta_k. \tag{4.18}$$

(see *Step 2* of Algorithm 4.2 and *Step 9* of Algorithm 4.3). To estimate the total cost for solving a linear system $H(\underline{u}_k)\underline{y} = -\underline{g}(\underline{u}_k)$, we must estimate the number of preconditioned conjugate gradient iterations and the number of floating point operations needed for the computation of one iteration.

The order of costs for one iteration is determined by the matrix-vector product $H(\underline{u}_k)\underline{y}$ and by the application of the preconditioner M (cf. Algorithm A.1). In the following, we assume that the preconditioning is cheaper than the matrix-vector product. In Appendix B, we suggest an algorithm which needs $\mathcal{O}(p^4)$ floating point operations for one matrix-vector multiplication $H(\underline{u}_k)\underline{y}$.

In this section, we recall that the number of conjugate gradient iterations depends mutually on the condition number of $H(\underline{u}_k)$. Further, we cite the sharp bounds on the condition number given by Melenk in [Mel02]. Since the cg-iterations do not depend on $\underline{u}_k$ during the iteration process, we can neglect the dependency on $\underline{u}_k$ in the following.

Definition 4.18. Let X and Y be normed spaces and let $A : X \to Y$ be a bounded linear operator with a bounded inverse $A^{-1} : Y \to X$. Then

$$\operatorname{cond}(A) := \|A\| \, \|A^{-1}\|$$

is called the condition number of A.
The condition of a matrix $H \in \mathbb{R}^{N\times N}$ is defined as $\operatorname{cond}(H) = \|A\|_2 \, \|A^{-1}\|_2$ where $A : \mathbb{R} \to \mathbb{R}^N$ is given by $A(x) := Hx$ and $\|\cdot\|_2$ denotes the operator norm with respect to the Euclidian norm on $\mathbb{R}^N$.

Frequently, to develop the estimates for $\operatorname{cond}(H)$, we will use

Lemma 4.19. The condition number of a Hermitian matrix H is given by

$$\operatorname{cond}(H) = \frac{\lambda_N}{\lambda_1}$$

where λ_N and λ_1 denote the largest and the smallest eigenvalues of H, respectively.

Proof. We observe that $\|H\|_2 = \lambda_N$ and $\|H^{-1}\|_2 = \mu_N$ (cf. [Kre98, Theorem 3.31]) where μ_N is the largest eigenvalue of H^{-1}. But $\mu_N = \lambda_1^{-1}$. □

We cite the following result which allows us to estimate the relative error of the cg-scheme depending on the numbers of iterations.

Theorem 4.20. Let $\underline{r}_i := \underline{g} + H\underline{y}_i$ the residual of the i-th iteration. Taking the Euclidian norm $\|\underline{v}\| := \|\underline{v}\|_2 := (\underline{v}^T\underline{v})^{1/2}$ as the norm of Algorithm 4.2 and the H-norm $\|\underline{v}\|_H := (\underline{v}^T H\underline{v})^{1/2}$, the analysis of the conjugate gradient method shows that

$$\frac{\|\underline{r}_i\|}{\|\underline{g}\|} \;\le\; \sqrt{\operatorname{cond}(H)}\,\frac{\|\underline{y}_i - \underline{y}\|_H}{\|\underline{y}_0 - \underline{y}\|_H} \;\le\; 2\sqrt{\operatorname{cond}(H)}\left(\frac{\sqrt{\operatorname{cond}(H)}-1}{\sqrt{\operatorname{cond}(H)}+1}\right)^i$$

Proof. The conjugate gradient method is analyzed in [Kel95, §2.3] and [CGO76]. □

Using the previous theorem, we obtain that due to the stopping criterion (4.18) the cg-method will terminate at the latest in the iteration

$$i_{\max} = \left\lceil \log\left(\frac{\eta}{2\sqrt{\operatorname{cond}(H)}}\right) \Big/ \log\left(\frac{\sqrt{\operatorname{cond}(H)}-1}{\sqrt{\operatorname{cond}(H)}+1}\right) \right\rceil . \tag{4.19}$$

The following corollary allows us to replace the last equation by an estimate which is less complicated.

Corollary 4.21. Let $\delta > 0$. Then, there exists a constant $C > 0$ such that

$$i_{\max} \le C\big(\operatorname{cond}(H)\big)^{1/2+\delta} . \tag{4.20}$$

For condition numbers of numerical interest, e.g. $\operatorname{cond}(H) \le 10^{25}$, (4.20) holds also with $\delta = 0$.

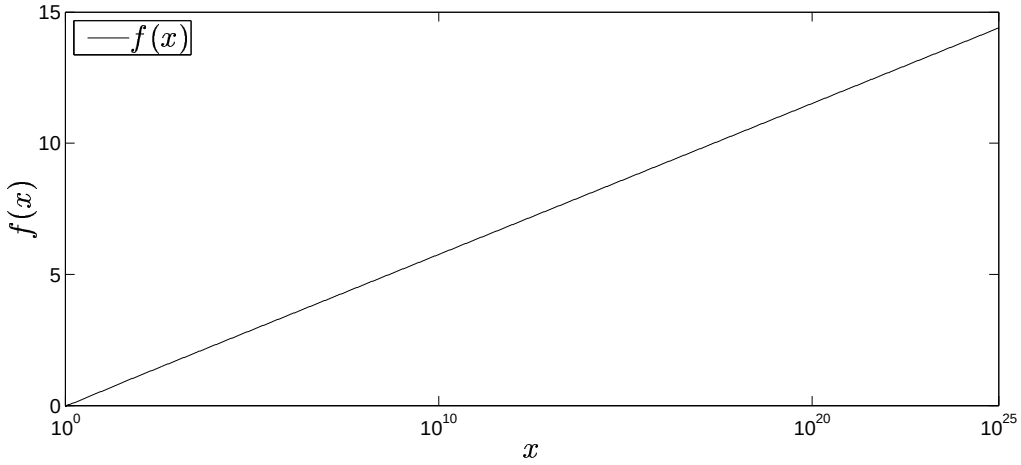

Figure 4.1: Semi-logarithmic plot of the quotient given in (4.23) with $\delta = 0$ and $\eta = 2$, i.e., $f(x) := \log\big(x^{-1/2}\big)\big(\sqrt{x}\big(\log(\sqrt{x}-1) - \log(\sqrt{x}+1)\big)\big)^{-1}$.

Proof. It suffices to proof that the right hand side of (4.19) divided by $(\mathrm{cond}(H))^{1/2+\delta}$ converges towards a finite real number for $\mathrm{cond}(H) \to \infty$. For brevity of notation, we substitute $\mathrm{cond}(H)$ by x. From $\lim_{y\to\infty}\big(1+\frac{1}{y}\big)^y = \exp(1)$, we obtain

$$\lim_{x\to\infty}\left(\frac{\sqrt{x}-1}{\sqrt{x}+1}\right)^{\sqrt{x}} = \exp(-2) \quad \text{and} \quad \lim_{x\to\infty}\sqrt{x}\,\log\left(\frac{\sqrt{x}-1}{\sqrt{x}+1}\right) = -2 \tag{4.21}$$

by the substitution $y = \frac{1}{2}(\sqrt{x}-1)$. Using l' Hospital rule in case of $\delta \neq 0$, we get

$$\lim_{x\to\infty}\frac{\log\frac{\eta}{2\sqrt{x}}}{x^{\delta}} = \begin{cases} 0 & \text{for } \delta > 0\,, \\ -\infty & \text{for } \delta \leq 0\,. \end{cases} \tag{4.22}$$

Combining (4.21) and (4.22), yields

$$\lim_{x\to\infty}\frac{\log\frac{\eta}{2\sqrt{x}}}{x^{\frac{1}{2}+\delta}\,\log\left(\frac{\sqrt{x}-1}{\sqrt{x}+1}\right)} = \begin{cases} 0 & \text{for } \delta > 0\,, \\ \infty & \text{for } \delta \leq 0\,, \end{cases} \tag{4.23}$$

which proves (4.20) for $\delta > 0$. The semi-logarithmic plot of Figure 4.1 shows that the quotient given in (4.23) grows very slowly for $\delta = 0$ and $\eta = 2$. For $\delta = 0$ and $\eta < 2$ we get a similar slow increase of the quotient because of $\log\frac{\eta}{2\sqrt{x}} = \log\frac{2}{2\sqrt{x}} + \log\frac{\eta}{2}$. Thus, (4.20) holds for $\delta = 0$ and $\mathrm{cond}(H) \leq 10^{25}$. □

As one cg-iteration costs $\mathcal{O}(p^4)$ floating point operations (cf. Algorithm B.3 and Proposition B.9), we obtain

Corollary 4.22. The linear system $H\underline{y} = -\underline{g}$ is solved approximately by the conjugate gradient method at a cost of $\mathcal{O}\big(p^4\sqrt{\mathrm{cond}(H)}\big)$ floating point operations, if we assume $\mathrm{cond}(H) \leq 10^{25}$.

4.4.1 Condition number estimates

The condition numbers in two-dimensional *hp*-FEM with locally refined meshes are analyzed by Melenk in [Mel02]. The standard technique to estimate the condition number splits the bilinear form

$$\langle v_1, v_2 \rangle_{\rho,u} := D^2 A(u; v_1, v_2) \quad \text{for all } v_1, v_2 \in V$$

into element contributions. Here, the parameters ρ and u refer to the function ρ in (1.10) and to u in $D^2 A(u; \cdot, \cdot)$). Then, the element contributions are analyzed using polynomial inverse estimates on the reference element and estimates on the geometric transformations.

In [BS89] Bank and Scott show that the combination of the element contributions results in the condition number bound

$$\operatorname{cond}(H) \leq C_{\rho,u} C_{\mathcal{T}} C_p$$

where $C_{\rho,u}$ is a function of the specific bilinear form $\langle v_1, v_2 \rangle_{\rho,u}$ and $C_{\mathcal{T}}$ reflects the dependence on the mesh. C_p depends only on the basis on the reference element $\tilde{Q}$ and on the polynomial degree. Thus, the influence of the mesh and of the polynomial degree can be studied separately.

From the bounds for the bilinear form $\langle v_1, v_2 \rangle_{\rho,u}$ given in (1.13), it follows that the $C_{\rho,u}$ can be replaced by a positive constant. In context of geometrically refined meshes $\mathcal{T}$ (cf. Section 2.4), we know from [BS89] that there exists a positive constant C such that

$$C_{\mathcal{T}} \leq C \cdot (\operatorname{card} \mathcal{T})^2 \,, \tag{4.24}$$

when we assume Dirichlet boundary conditions. Since adaptive refinement leads to a geometric refinement towards corners and edges with singularities, this estimate also holds for adaptively refined meshes. Thus, it remains to estimate C_p, i.e., to estimate $\operatorname{cond}(H)$ for a problem on the reference element $\tilde{Q}$ with a bilinear form equivalent to $D^2 A(u; \cdot, \cdot)$ for all ρ, u.

Melenk addresses this problem in [Mel02] for the basis $B_{\tilde{Q}}$ introduced in Section 4.1 and proves

Theorem 4.23. Using the coordinate representation with respect to the Lagrangian basis $B_{\tilde{Q}}$ on the reference square $\overline{\tilde{Q}} = [-1, 1]^2$ (see Section 4.1), there exists a constant $c > 0$ independent of p such that for all polynomials $v = \sum_{i,j=0}^{p} v_{ij} b_{\tilde{Q},ij} \in V(\tilde{Q})$ there holds

$$c^{-1} p^{-2} \sum_{i,j=0}^{p} v_{ij}^2 \leq \|v\|^2_{H^1(\tilde{Q})} \leq cp \sum_{i,j=0}^{p} v_{ij}^2 \,. \tag{4.25}$$

Proof. [Mel02, Proposition 2.8]. □

The optimality of (4.25) is not proved theoretically. Nevertheless, the convergence rates of Experiment 4.34 emphasize that (4.25) is optimal. As suggested by Melenk this stability result can be used to estimate the condition of the Hessian corresponding to the *p*-version on geometrically refined meshes.

Theorem 4.24. Let $r(x) := \operatorname{dist}(x, \partial\Omega)$ and let c_{geo}, h_0 be positive constants such that the elements Q of the mesh $\mathcal{T}$ satisfy $h_Q \geq c_{\text{geo}} h_0\, r(F_Q((0,0)))$ where h_Q are parameters proportional to the diameter of Q (see (2.1)). Assume Dirichlet boundary conditions, i.e., $\Gamma_D = \partial\Omega$ and $\Gamma_N = \emptyset$. Let $V_{p,g_D} := V_{p,g_D}(\mathcal{T})$, $g_D \equiv 0$, be given by Definition 2.4 for $p \geq \mathbb{N}$. Further, let $H := H(\underline{u}_k)$ be the Hessian in the k-th iteration of the inexact Newton backtracking method (see Algorithm 4.2) in the minimization process of A on V_{p,g_D}. Then, there exists a positive constant C depending only on the constants of (2.1), c_{geo}, and κ_l, κ_u of (1.13) such that the condensed matrix H^c satisfies

$$\|H\|_2 \leq Cp \qquad \text{and} \qquad \|\big(H\big)^{-1}\|_2 \leq C h_0^{-2} p^2\,. \tag{4.26}$$

There holds $\operatorname{cond}(H) = \mathcal{O}(h_0^{-2} p^3)$.

Proof. Let $\underline{v} = (v_k)_{k=1,\dots,N}$, $N := \dim V_p(\mathcal{T})$, be the coordinate representation of $v \in V_p(\mathcal{T})$ given by $v = \sum_{k=0}^{N} v_k b_k$ with b_k as in (4.2). With Theorem 4.23 we know that

$$c^{-1} p^{-2} \sum_{k=1}^{N} v_k^2 \leq \sum_{Q \in \mathcal{T}} \big(\|\nabla v\|_{L^2(Q)}^2 + h_Q^{-2} \|v\|_{L^2(Q)}^2 \big)\,.$$

By the two-dimensional specification of the embedding result in weighted Sobolev spaces [Gri85, Theorem 1.4.4.3], there exists a positive constant $\bar{c}$ with

$$\|r^{-1} v\|_{L^2(\Omega)} \leq \bar{c} \|\nabla v\|_{L^2(\Omega)} \qquad \text{for all } v \in H_0^1(\Omega)\,. \tag{4.27}$$

Thus, we get

$$c^{-1} p^{-2} \sum_{k=1}^{N} v_k^2 \leq \big(1 + c_{\text{geo}}^{-2} h_0^{-2}\, \bar{c}^{\,-2}\big) \|\nabla v\|_{L^2(\Omega)}^2 \leq \kappa_l^{-1} \big(1 + c_{\text{geo}}^{-2} h_0^{-2}\, \bar{c}^{\,-2}\big) D^2 A(u; v, v)$$

with (1.13). The correspondence $D^2 A(u; v, v) = \underline{v}^T H(\underline{u})\, \underline{v}$ for all $\underline{u}, \underline{v} \in V_p(\mathcal{T})$ yields

$$c_1 h_0^2 p^{-2} \leq \inf \left\{ \frac{\underline{v}^T H(\underline{u})\, \underline{v}}{\underline{v}^T\, \underline{v}} \middle| \underline{v} \in \mathbb{R}^N \setminus \{0\} \right\} = \lambda_1 = \|H^{-1}\|_2^{-1}$$

where λ_1 is the smallest eigenvalues of the positive definite matrix H and c_1 is a positive constant depending only on c_{geo}, h_0, $\bar{c}, \kappa_l$.

From the right inequality given in Theorem 4.23 and (1.13), we obtain

$$D^2 A(u; v, v) \leq \kappa_u \|v\|_{H^1(\Omega)}^2 \leq \kappa_u c p \sum_{k=1}^{N} v_k^2$$

which yields

$$\|H\|_2 = \lambda_N = \sup \left\{ \frac{\underline{v}^T H(\underline{u})\, \underline{v}}{\underline{v}^T\, \underline{v}} \middle| v \in \mathbb{R}^N \setminus \{0\} \right\} \leq c_2 p$$

with positive constant $c_2 = \kappa_u c$. Thus, we get (4.26) with $C := \max\{c_2, c_1^{-1}\}$. The statement on the condition number follows by Definition 4.18. □

From Corollary 4.22 we obtain

Corollary 4.25. With the assumption of Theorem 4.24 the linear system $H\underline{y} = -\underline{g}$ is solved approximately by the conjugate gradient method at a cost of $\mathcal{O}(p^{11/2})$ floating point operations.

Remark 4.26.

1. The assumption $h_Q \geq c_{\text{geo}}\, h_0\, r(F_Q((0,0)))$ allows meshes that are geometrically refined towards vertices and edges. Thus, adaptive refinements are also covered by Theorem 4.24.

2. The proof of Theorem 4.24 shows that the assumption of Dirichlet boundary conditions can be replaced by the assumption

$$\|r^{-1}v\|^2_{L^2(\Omega)} \leq D^2A(u;v,v) \quad \text{for all } u,v \in H^1_{g_D}(\Omega)$$

(cf. (4.27)). In particular, this is fulfilled, when we minimize the functional A with $\sigma > 0$ on a quasi-uniform mesh.

4.4.2 Solving the linear system by static condensation

Frequently, *static condensation* is suggested in FE literature for an efficient solution of the linear systems. Thus, we estimate the computational costs of this method for the basis defined in Section 4.1. *Static condensation* can be introduced as follows.

Let $O \cup I$ be a partition of the index set $\{1,\dots,N\}$, $N := \operatorname{card} G_p$, of the global basis such that O contains the indices of the node- and edge-associated degrees of freedom, and I contains the indices of the interior degrees of freedom, i.e., those with a local triple (Q,i,j), $1 \leq i,j \leq p-1$ (cf. (4.2)). Reordering the linear system $H\underline{v} = -\underline{g}$ according to this partition, we may write

$$\begin{pmatrix} H_{OO} & H_{OI} \\ H_{IO} & H_{II} \end{pmatrix} \begin{pmatrix} \underline{v}_O \\ \underline{v}_I \end{pmatrix} = - \begin{pmatrix} \underline{g}_O \\ \underline{g}_I \end{pmatrix}. \tag{4.28}$$

As $\operatorname{supp} b_{Q,i,j} = \overline{Q}$ for all $Q \in \mathfrak{T}$, $1 \leq i,j \leq p-1$, H_{II} has the block diagonal structure

$$H_{II} = \begin{pmatrix} H_{Q_1} & & \\ & \ddots & \\ & & H_{Q_{n_{\mathfrak{T}}}} \end{pmatrix}, \qquad n_{\mathfrak{T}} := \operatorname{card} \mathfrak{T},$$

with the local stiffness matrices $H_{Q_i} \in \mathbb{R}^{(p-1)^2 \times (p-1)^2}$, $Q_i \in \mathfrak{T}$, associated to the $(p-1)^2$ local basis functions $b_{Q,i,j}$, $1 \leq i,j \leq p-1$. Thus, we obtain the inverse of H simply by the inversion of the local blocks, i.e.,

$$H_{II}^{-1} = \begin{pmatrix} H_{Q_1}^{-1} & & \\ & \ddots & \\ & & H_{Q_{n_{\mathfrak{T}}}}^{-1} \end{pmatrix}.$$

This allows the reformulation

$$H^c\underline{v}_O = -\underline{g}^c \qquad \text{with } H^c := H_{OO} - H_{OI}H_{II}^{-1}H_{IO} \text{ and } \underline{g}^c := \underline{g}_O - H_{OI}H_{II}^{-1}\underline{g}_I \tag{4.29}$$

of (4.28) by applying the Schur complement, called *static condensation* in the context of finite elements. The condition number of the matrix H^c is estimated by

Theorem 4.27. With the assumption of Theorem 4.24, there holds

$$\|H^c\|_2 \le C\,, \quad \|\big(H^c\big)^{-1}\|_2 \le Ch_0^{-2}p, \qquad \text{and} \qquad \operatorname{cond}(H^c) = \mathcal{O}(h_0^{-2}p)\,.$$

Proof. The statements follow by applying [Mel02, Theorem 2.2] on the bilinear form given by $D^2A(u;\cdot,\cdot)$. □

Corollary 4.28. Let the assumptions of Theorem 4.24 be fulfilled. Then, using static condensation the linear system $H\underline{y} = -\underline{g}$ is solved approximately by the conjugate gradient method at a cost of $\mathcal{O}(p^6)$ floating point operations, when the local systems $H_{Q_i}\underline{v}_{Q_i} = -\underline{g}_{Q_i}$, $Q_i \in \mathcal{T}$, are solved by the cg-method.

Proof. With Corollary 4.25 all local systems H_{Q_i}, $Q_i \in \mathcal{T}$, are solved at a cost of $\mathcal{O}(p^{11/2})$ by the cg-method. Thus, the computation of the matrix-vector product $H^c\underline{v}_O$, and consequently one cg-iteration for the condensed system, costs $\mathcal{O}(p^{11/2})$ flops. The corollary follows with Corollary 4.21 due to $\operatorname{cond}(H^c) = \mathcal{O}(p)$. □

Remark 4.29. With the numerical experiments we can conjecture that $\operatorname{cond}(\tilde{H}_{Q_i}) = \mathcal{O}(p^2)$ for all $Q_i \in \mathcal{T}$ where $\tilde{H}_{Q_i}$ denotes the matrix corresponding to a diagonally preconditioned cg-scheme for H_{Q_i} (see Conjecture 4.35). This improves Corollary 4.28 to $\mathcal{O}(p^{11/2})$, when the local systems are solved with diagonal preconditioning.

Remark 4.30. Remark 4.26 holds also for Theorem 4.27 because we have the same assumptions.

4.4.3 Using a hierarchical basis

It is a standard in p-version FEM to employ a hierarchical basis. The analysis of the diagonally preconditioned system corresponding to this basis shows that the condition numbers grow mildly with a rate of $\mathcal{O}(p^2)$ (see Theorem 4.32, (4.33a)). In Section 4.5 we will show that we can use this advantage of the hierarchical basis over the Lagrangian basis for efficient preconditioning of the reduced linear systems raised by the nonlinear programming problem.

Let a basis $(\mathcal{L}_0, \ldots, \mathcal{L}_p)$ of $\mathbb{P}_p$ be given by

$$\mathcal{L}_0(\xi) := \tfrac{1}{2}(1-\xi), \qquad \mathcal{L}_1(\xi) := \tfrac{1}{2}(1+\xi),$$

and by the normed anti-derivatives of Legendre polynomials

$$\mathcal{L}_i(\xi) := \frac{1}{\|L_{i-1}\|_{L^2(-1,1)}} \int_{-1}^{\xi} L_{i-1}(t)\ \mathrm{d}t \quad \text{for all } \xi \in [-1,1],\ i = 2, \ldots, p.$$

Elementary integration of the Legendre polynomials yields $\|L_{i-1}\|_{L^2(-1,1)} = \sqrt{\frac{2}{2i+1}}$. We call $(\mathcal{L}_0, \ldots, \mathcal{L}_p)$ a hierarchical basis, since it can be expanded to a basis of $\mathbb{P}_{p+1}$ simply be adding $\mathcal{L}_{p+1}$.

We define the local basis

$$B^{\mathcal{L}}_{\tilde{Q},p} := (b^{\mathcal{L}}_{ij} \,|\, 0 \le i,j \le p) \tag{4.30}$$

on the reference square $\tilde{Q}$ by the tensor product polynomials $b_{ij}^{\mathcal{L}}((\xi_1,\xi_2)) := \mathcal{L}_i(\xi_1)\mathcal{L}_j(\xi_2)$. Again, as usual in finite elements, we get the local bases on the quadrilaterals with the transformations F_Q and the global basis $B_p^{\mathcal{L}} := (b_1^{\mathcal{L}},\dots,b_{N_p}^{\mathcal{L}})$ by assembling the local basis functions (cf. (4.1), (4.2)) and the introduction of a global counting.

Now, let $B_p := (b_1,\dots,b_{N_p})$ be the Lagrangian basis with respect to the Gauss-Lobatto nodes (cf. (4.2)) and let N be the matrix mapping the basis functions $\big(b_1,\dots,b_{N_p}\big)$ onto $\big(b_1^{\mathcal{L}},\dots,b_{N_p}^{\mathcal{L}}\big)$, i.e.,

$$\big(b_1^{\mathcal{L}},\dots,b_{N_p}^{\mathcal{L}}\big)^T = N\big(b_1,\dots,b_{N_p}\big)^T . \tag{4.31}$$

Lemma 4.31. Let $v \in \operatorname{span} B_p$, and let $\underline{v}^{\mathcal{L}}$, $\underline{v}$ be its vector representations with respect to $B_p^{\mathcal{L}}$ and B_p, respectively. Then, we have the coordinate transformation

$$\underline{v} = N^T\underline{v}^{\mathcal{L}} .$$

Proof. We write

$$v = \begin{pmatrix} v_1^{\mathcal{L}} \\ \vdots \\ v_N^{\mathcal{L}} \end{pmatrix}^T \begin{pmatrix} b_1^{\mathcal{L}} \\ \vdots \\ b_N^{\mathcal{L}} \end{pmatrix} = \begin{pmatrix} v_1 \\ \vdots \\ v_N \end{pmatrix}^T \begin{pmatrix} b_1 \\ \vdots \\ b_N \end{pmatrix} .$$

Using (4.31) to replace $\big(b_1^{\mathcal{L}},\dots,b_{N_p}^{\mathcal{L}}\big)^T$ yields the statement. □

We use the coordinate transformation of Lemma 4.31 to reformulate the linear system $H\underline{v} = -\underline{g}$ as

$$H^{\mathcal{L}}\underline{v}^{\mathcal{L}} = -\underline{g}^{\mathcal{L}} \qquad \text{with} \quad H^{\mathcal{L}} := NHN^T \quad \text{and} \quad g^{\mathcal{L}} := Ng . \tag{4.32}$$

Theorem 4.32. Let $H^{\mathcal{L}}$ be the Hessian of the discrete functional $A : \mathbb{R}^{N_p} \to \mathbb{R}$ given by the redefinition $A(\underline{w}^{\mathcal{L}}) := A\big(\sum_{i=1}^{N_p} w_i^{\mathcal{L}} b_i^{\mathcal{L}}\big)$ with respect to the basis $B_p^{\mathcal{L}}$ in the k-th iteration of the inexact Newton backtracking method.
Let $H_{II}^{\mathcal{L}}$ be the sub-matrix of $H^{\mathcal{L}}$ corresponding to the bubble basis functions $b_i^{\mathcal{L}}$, i.e., there exists a $Q \in \mathcal{T}$ such that $\operatorname{supp} b_i^{\mathcal{L}} \subset \overline{Q}$.
With the assumptions of Theorem 4.24, there holds

$$ch_0^{-2}p^4 \;\le\; \operatorname{cond}(H^{\mathcal{L}}) \;\le\; Ch_0^{-2}p^4(1+\log^2 p)\,, \tag{4.33a}$$

$$ch_0^{-2}p^4 \le \operatorname{cond}(H_{II}^{\mathcal{L}}) \;\le\; Ch_0^{-2}p^4\,, \tag{4.33b}$$

each with positive constants c, C independent of h_0 and p.

Further, let $\tilde{H}^{\mathcal{L}} := \Lambda^{-1/2}H^{\mathcal{L}}\Lambda^{-1/2}$ where Λ denotes the diagonal matrix yielded from $H^{\mathcal{L}}$ by setting all non-diagonal entries to zero. $\tilde{H}^{\mathcal{L}}$ corresponds to a diagonally preconditioned cg-scheme, i.e., the preconditioning step $M\underline{z}_i = \underline{r}_i$ in Algorithm A.1 is realized by the divisions $(\underline{z}_i)_k = (\underline{r}_i)_k/H_{kk}$, $k = 1,\dots,N_p$, where k denotes the component of the vectors and H_{kk} denotes the corresponding diagonal element of H.
Let $\tilde{H}_{II}^{\mathcal{L}}$ be defined analogously.
With the assumptions of Theorem 4.24, there holds

$$ch_0^{-2}p^2 \;\le\; \operatorname{cond}(\tilde{H}^{\mathcal{L}}) \;\le\; Ch_0^{-2}p^2(1+\log^2 p)\,, \tag{4.34a}$$

$$ch_0^{-2}p^2 \;\le\; \operatorname{cond}(\tilde{H}_{II}^{\mathcal{L}}) \;\le\; Ch_0^{-2}p^2\,, \tag{4.34b}$$

each with positive constants c, C independent of h_0 and p.

Proof. Analogously to the proof of Theorem 4.24, the estimates (4.33) and (4.34) follow from stability estimates on the reference element $\tilde{Q}$. The local estimates for (4.33a) and (4.33b) are stated in [MP96, Corollary 1 and Theorem 1, respectively], the local estimates for (4.34a) and (4.34b) are stated in [MP96, Corollary 2 and Theorem 3, respectively]. □

Remark 4.33.

1. The system $H^{\mathcal{L}}\underline{v}^{\mathcal{L}} = -\underline{g}^{\mathcal{L}}$ can be solved efficiently since the condition number of a diagonally preconditioned $H^{\mathcal{L}}$ is of $\mathcal{O}(p^2)$ due to Theorem 4.32. We also consider the static condensation of this system in Experiment 4.37 and conjecture that the condition of the condensed matrix grows at $\mathcal{O}(p^2)$ (Conjecture 4.38). Since the condensation process needs additionally the inversion of the $H^{\mathcal{L}}_{II}$-block corresponding to the interior basis functions, diagonal preconditioning should be preferred to static condensation, when the hierarchical basis is used.

2. $H^{\mathcal{L}}$ and $\underline{g}^{\mathcal{L}}$ can be computed directly as the stiffness matrix and the righthand side analogously to H and $\underline{g}$, when we replace the basis function b_k by $b_k^{\mathcal{L}}$. If we substitute (B.3) by the equation

$$\begin{pmatrix} \mathcal{L}_0(\xi) \\ \vdots \\ \mathcal{L}_p(\xi) \end{pmatrix} = C \begin{pmatrix} \lambda_0^q(\xi) \\ \vdots \\ \lambda_q^q(\xi) \end{pmatrix} \qquad \text{with } C := \begin{pmatrix} \mathcal{L}_0(\xi_0^{q+1}) & \cdots & \mathcal{L}_0(\xi_q^{q+1}) \\ \vdots & & \vdots \\ \mathcal{L}_p(\xi_0^{q+1}) & \cdots & \mathcal{L}_p(\xi_q^{q+1}) \end{pmatrix},$$

   then Algorithm B.3 performs the matrix-vector multiplication $H^{\mathcal{L}}(\underline{u})\,\underline{v}$ locally. Thus, the product costs $\mathcal{O}(p^4)$ floating point operations due to Proposition B.9.

3. The diagonal entries of $H^{\mathcal{L}}$ are not explicitly known, when only the matrix-vector product $H^{\mathcal{L}}\underline{v}^{\mathcal{L}}$ is implemented. Using C from 2. of this remark, Algorithm B.7 computes the diagonal entries at a cost of $\mathcal{O}(p^4)$ floating point operations.

4. We recall that the Dirichlet boundary condition yielded the definition of the affine subspace V_{p,g_D} and the equality constraints given by the boundary data g_D in the Gauss-Lobatto points on the boundary, when we used the coordinate representation with respect to the Lagrangian basis B_p. Using the hierarchical basis, we proceed as follows to get the same affine subspace V_{p,g_D}.

   Let $B_{p,\Gamma}$ and $B^{\mathcal{L}}_{p,\Gamma}$ be the subsets of $B_{p,\Gamma}$ and $B^{\mathcal{L}}_{p,\Gamma}$, respectively, which are nonzero on the Dirichlet boundary. Further, let $\underline{v}_\Gamma$ be the coordinates of the discrete boundary function with respect to $B_{p,\Gamma}$, i.e.,

$$v_{p|\Gamma} = \underline{v}_\Gamma B_{p,\Gamma}\,.$$

   The components $v_{\Gamma,k} = g_D(x_k)$, $k = 1, \ldots, \operatorname{card} B_{p,\Gamma}$, of $\underline{v}_\Gamma$ are given by the Dirichlet data in the Gauss-Lobatto points on Γ. As $\operatorname{span} B_{p,\Gamma} = \operatorname{span} B^{\mathcal{L}}_{p,\Gamma}$, there exists a sub-matrix N_Γ of the matrix N (see (4.31)), such that $B_{p,\Gamma} = N_\Gamma B^{\mathcal{L}}_{p,\Gamma}$. Thus, we can fulfill the Dirichlet condition in the Gauss-Lobatto points by computing the coordinate representation $\underline{v}^{\mathcal{L}}_\Gamma$ with respect to $B^{\mathcal{L}}_{p,\Gamma}$, i.e.,

$$\underline{v}^{\mathcal{L}}_\Gamma = N_\Gamma^{-T}\,\underline{v}_\Gamma\,.$$

   So, we can switch totally to hierarchical basis in case of variational equalities at a cost of the inversion of N_Γ^T. The linear systems corresponding to variational inequalities are considered in Section 4.5.

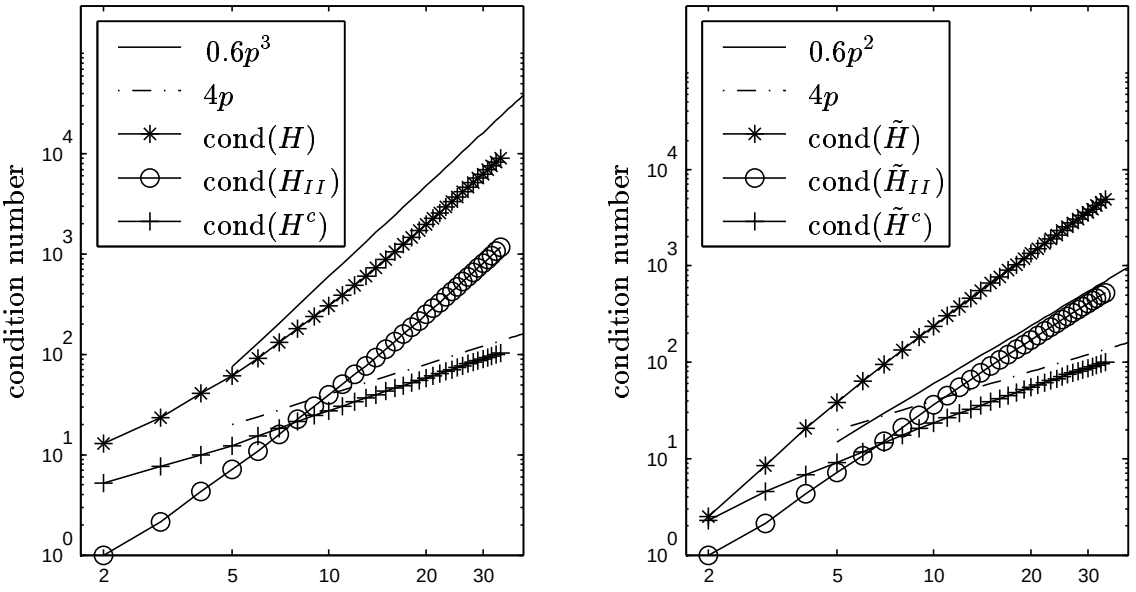

Figure 4.2: cond(H), cond(H_{II}), cond(H^c) (left) and cond$(\tilde{H})$, cond$(\tilde{H}_{II})$, cond$(\tilde{H}^c)$ (right).

4.4.4 Numerical experiments concerning the condition numbers

As the dependence of the condition numbers of H on the mesh geometry was explored already in [AMT99, BS89, Mel02], it is the intention of the following experiments to compare the influence of different p-version FE bases on the condition of the linear system. Further, we are interested in the condition numbers of the respective condensed matrix H^c and in cond(H_{II}) because H_{II} must be inverted to obtain H^c.

Experiment 4.34. We compute the stiffness matrix of the Laplacian on the reference element $\tilde{Q}$

$$H := \left(\int_{\tilde{Q}} (\nabla b_i \nabla b_j \, \mathrm{d}x \right)_{\substack{1 \le i \le N_p \\ 1 \le j \le N_p}}, \qquad N_p := \dim V_p(\tilde{Q}),$$

where $b_i \in B_{\tilde{Q}}$ are the Lagrangian basis functions with respect to the Gauss-Lobatto points of order p. Then H_{II} is regular, but H and H^c are not invertible since their null spaces are spanned by the coefficient vector corresponding to the function $u \equiv 1$. It would be a work around to fix one degree of freedom to 0, i.e., to reduce H, H^c by one column and one row.

Instead of this strategy we prefer to redefine the condition number for the semidefinite matrices as the quotient λ_N/λ_2. Here, λ_N and λ_2 denote the largest eigenvalue and the smallest eigenvalue greater than 0 of the positive semidefinite matrices H, H^c. Using this redefinition of cond(H), cond(H^c), we ensure that the influence of all basis functions on the condition is balanced in a way similar to that of FE-mesh with several quadrilaterals.

The left panel of Figure 4.2 presents a double logarithmic plot of cond(H), cond(H_{II}), cond(H^c) versus p. A comparison with the additionally plotted lines $0.6p^3$ and $4p$ confirms

the growth rates $\operatorname{cond}(H) = \mathcal{O}(p^3)$, and $\operatorname{cond}(H^c) = \mathcal{O}(p)$ stated in Theorem 4.24 and Theorem 4.27, respectively. As H_{II} is a sub-matrix of H, it follows that $\operatorname{cond}(H_{II}) = \mathcal{O}(p^3)$. This is also confirmed by the plot.

The right panel of Figure 4.2 presents the condition numbers of $\tilde{H} := \Lambda^{-1/2} H \Lambda^{-1/2}$ where Λ denotes the diagonal matrix yielded from H by setting all non-diagonal entries to zero. $\tilde{H}$ corresponds to a diagonally preconditioned cg-scheme, i.e., the preconditioning step $M\underline{z}_i = \underline{r}_i$ in Algorithm A.1 is realized by the divisions $(\underline{z}_i)_k = (\underline{r}_i)_k / H_{kk}$, $k = 1, \dots, N_p$, where k denotes the component of the vectors and H_{kk} is the corresponding diagonal entry of H. Further, $\operatorname{cond}(\tilde{H}_{II})$ and $\operatorname{cond}(\tilde{H}^c)$ are plotted in the right panel. Again, we get the $\sim$-expressions by diagonal preconditioning. In case of $\tilde{H}$ and $\tilde{H}^c$ we obtain the same growth rates as in the non-preconditioned schemes which means that these matrices are scaled well already. For $\tilde{H}_{II}$ we obtain the amazing result that the conditions numbers are natural numbers for $p \geq 8$.

Conjecture 4.35. There holds

$$\operatorname{cond}(\tilde{H}_{II}) = \tfrac{1}{2}(p-2)(p-1) \quad \text{for } p \geq 8\,.$$

Justification. The numerical experiment with double precision arithmetic yields

$$|\tfrac{1}{2}(p-2)(p-1) - \operatorname{cond}(\tilde{H}_{II})| \;\leq\; \begin{cases} 10^{-12} & \text{for } 8 \leq p \leq 17\,, \\ 10^{-11} & \text{for } 8 \leq p \leq 28\,, \\ 10^{-10} & \text{for } 8 \leq p \leq 34\,. \end{cases}$$

□

In addition to the graphical presentation, we list the condition numbers and the experimental convergence rates

$$\alpha_p = \frac{\log\big(\operatorname{cond}(H_p)/\operatorname{cond}(H_{p-2})\big)}{\log\big(p/(p-2)\big)} \quad \text{for } 3 \leq p \leq p_{\max} \tag{4.35}$$

in Table C.1 and Table C.2. Again, the experimental convergence rates confirm the theoretical results.

Experiment 4.36. We repeat Experiment 4.34 but replace the Lagrangian basis $B_{\tilde{Q}}$ by the Lagrangian basis B^{eq} with respect to the equidistant nodes $\big(-1 + 2i/p, -1 + 2j/p\big)$, $0 \leq i, j \leq p$.

Analogously to Figure 4.2, Figure 4.3 shows semilogarithmic plots of the condition of H^{eq}, H^{eq}_{II}, and $H^{c,\mathrm{eq}}$ (left panel), and of the respective diagonally preconditioned matrices (right panel), and compares these with the line $C\exp(bp)$, $C \approx 0.056$, $b \approx 2.73$. The growth rates $\operatorname{cond}(H^{\mathrm{eq}}) = \mathcal{O}(\exp(bx))$ and $\operatorname{cond}(H^{\mathrm{eq}}) = \mathcal{O}(\exp(bx))$ are asymptotically sharp and correspond to the theoretical result given in [OD95]. The exponential rates make it clear that the distribution of the nodes is essential for numerical stability, not only for variational inequalities but also for variational equalities. In addition to the graphical presentation, we list the condition numbers and the experimental exponents

$$b^{\mathrm{eq}}_p = \big(\operatorname{cond}(H_p)/\operatorname{cond}(H_{p-1})\big) \qquad \text{for } 2 \leq p \leq p_{\max}\,.$$

in Table C.3.

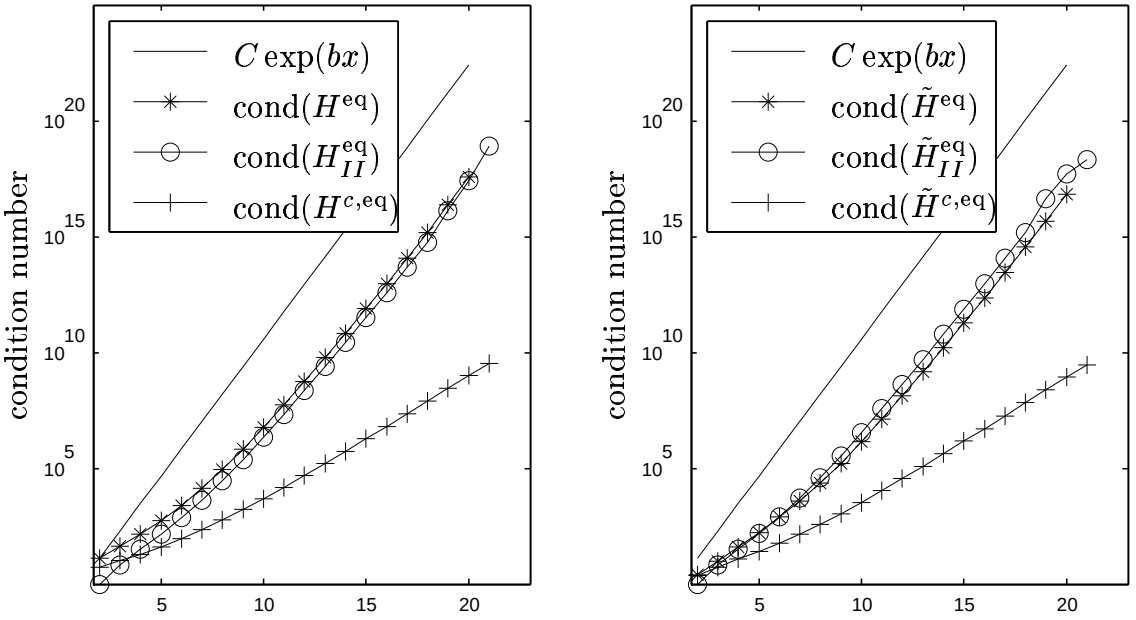

Figure 4.3: $\mathrm{cond}(H^{\mathrm{eq}})$, $\mathrm{cond}(H_{II}^{\mathrm{eq}})$, $\mathrm{cond}(H^{c,\mathrm{eq}})$ (left) and $\mathrm{cond}(\tilde{H}^{\mathrm{eq}})$, $\mathrm{cond}(\tilde{H}_{II}^{\mathrm{eq}})$, $\mathrm{cond}(\tilde{H}^{c,\mathrm{eq}})$ (right). The parameters of the line are $C \approx 0.056$, $b \approx 2.73$.

Experiment 4.37. We repeat Experiment 4.34 but replace the Lagrangian basis $B_{\tilde{Q}}$ by the hierarchical basis $B_{\tilde{Q},p}^{\mathcal{L}}$ defined by (4.30).

The left panel of Figure 4.4 presents $\mathrm{cond}(H^{\mathcal{L}})$, $\mathrm{cond}(H_{II}^{\mathcal{L}})$, $\mathrm{cond}(H^{c,\mathcal{L}})$. A comparison with the additionally plotted line $0.3p^4$ confirms the growth rates $\mathrm{cond}(H) = \mathcal{O}(p^4)$ and $\mathrm{cond}(H_{II}) = \mathcal{O}(p^4)$ stated in Theorem 4.32.

The right panel of Figure 4.4 presents the condition numbers of the corresponding diagonally preconditioned systems. A comparison with the line $4.8p^2$ confirms $\mathrm{cond}(\tilde{H}^{\mathcal{L}}) = \mathcal{O}(p^2)$ and $\mathrm{cond}(\tilde{H}_{II}^{\mathcal{L}}) = \mathcal{O}(p^2)$ stated in Theorem 4.32. Additionally, we list the condition numbers and the experimental convergence rates (cf. (4.35)) in Table C.4 and Table C.5. The experimental convergence rates agree with the abovely named convergence rates.

Conjecture 4.38. Let there hold the assumptions of Theorem 4.24. Further, let $H^{\mathcal{L}}$ be the Hessian of the discrete functional $A : \mathbb{R}^{N_p} \to \mathbb{R}$ given by the redefinition $A(\underline{w}^{\mathcal{L}}) := A\big(\sum_{i=1}^{N_p} w_i^{\mathcal{L}} b_i^{\mathcal{L}}\big)$ with respect to the basis $B_p^{\mathcal{L}}$ in the k-th iteration of the inexact Newton backtracking method. Let $H^{c,\mathcal{L}}$ be the condensed matrix of $H^{\mathcal{L}}$ analogously to (4.29) and $\tilde{H}^{c,\mathcal{L}}$ the corresponding diagonally preconditioned matrix. Then, there holds

$$ch_0^{-2}p^2 \leq \mathrm{cond}(H^{c,\mathcal{L}}) \leq Ch_0^{-2}p^2\,(1+\log^2 p)\,, \tag{4.36a}$$

$$ch_0^{-2}p \leq \mathrm{cond}(\tilde{H}^{c,\mathcal{L}}) \leq Ch_0^{-2}p\,(1+\log^2 p)\,. \tag{4.36b}$$

Justification. Due to Figure 4.4, Table C.4 and Table C.5 the stabilities (4.36) are proved for on the reference square $\tilde{Q}$ for $2 \leq p \leq 30$. We conjecture the stability for $p \in \mathbb{N}$. The dependency on h_0 is an analogy to Theorem 4.27 and its proof. □

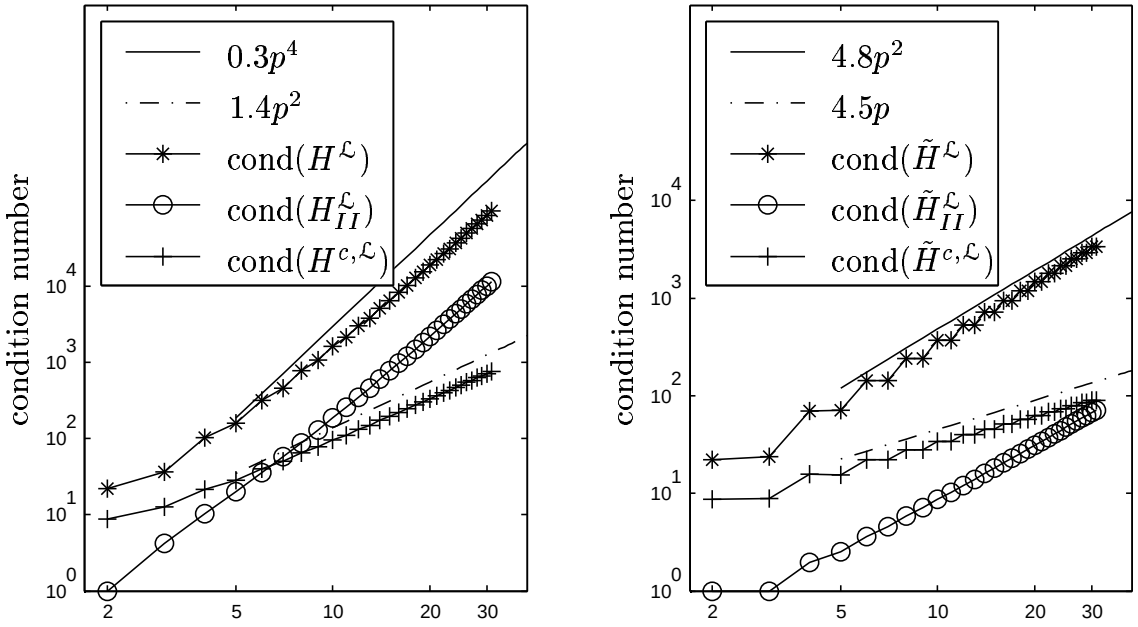

Figure 4.4: $\mathrm{cond}(H^{\mathcal{L}})$, $\mathrm{cond}(H^{\mathcal{L}}_{II})$, $\mathrm{cond}(H^{c,\mathcal{L}})$ (left) and $\mathrm{cond}(\tilde{H}^{\mathcal{L}})$, $\mathrm{cond}(\tilde{H}^{\mathcal{L}}_{II})$, $\mathrm{cond}(\tilde{H}^{c,\mathcal{L}})$ (right).

4.5 The linear system of the constrained problem

In Section 4.1, we use the Lagrangian basis with respect to Gauss-Lobatto points to define appropriate subsets which fulfill the Dirichlet boundary condition and the obstacle condition approximately (see V_{p,g_D} and K_{p,g_D}). *Step 7* of Algorithm 4.3 requires the unconstrained minimization of the $(\underline{\psi}, \mathcal{F}(\underline{u}_k))$-reduced function $A_{\underline{u}_k,\mathcal{F}(\underline{u}_k)}$ on $\mathbb{R}^{\operatorname{card}\mathcal{F}(\underline{u}_k)}$ by the inexact Newton backtracking method of Algorithm 4.2. Thus, we must solve the linear system $H_{\mathcal{F}}\, \underline{y}_{\mathcal{F}} = -\underline{g}_{\mathcal{F}}$ in *Step 2* of Algorithm 4.2 by a preconditioned conjugate gradient scheme approximately. Here,

$$H_{\mathcal{F}} = I_{\mathcal{F}}^T\, \nabla^2 A(\underline{\psi}_{\mathcal{F}}(\underline{u}_k))\; I_{\mathcal{F}} \qquad \text{and} \qquad \underline{g}_{\mathcal{F}} = I_{\mathcal{F}}^T \nabla A(\underline{\psi}_{\mathcal{F}}(\underline{u}_k))$$

denote the reduced Hessian and the reduced gradient, respectively, with $\mathcal{F} := \mathcal{F}(\underline{u}_k)$, see (4.17).

In this section, we want to demonstrate that the linear problems from the unbounded and the linear subsystems from the bounded discrete nonlinear problems both can be preconditioned with a cost of $\mathcal{O}(p^3)$ floating point operations such that the condition number of the respective system grows as the condition $\mathrm{cond}(\tilde{H}^{\mathcal{L}}) = \mathcal{O}(p^2)$ of the diagonally preconditioned system with respect to the p-hierarchical basis. For ease of notation, we only consider the p-version on the reference element $\tilde{Q}$. The treatment on more general meshes follows straightforwardly and any preconditioner given for the p-hierarchical basis can be used additionally.

However, it is in order here to recall that the *projected gradient inexact Newton backtracking method* of Algorithm 4.3 demands the solution of linear systems only in those steps which realize the adapted Newton method. Thus, good preconditioning of the reduced

system $H_{\mathcal{F}}\,\underline{y}_{\mathcal{F}} = -\underline{g}_{\mathcal{F}}$ enhances only the performance of the loop given by *Steps 7, 1, 2* of the algorithm.

4.5.1 Preconditioning the unconstrained problem

The convergence of the conjugate method can be improved significantly by solving $M^{-1}H\underline{y} = M^{-1}\underline{g}$ instead of the original problem. Of course, this makes only sense when M^{-1} can be applied cheaply. Using the preconditioned gradient scheme, it is not necessary to compute $M^{-1}H$ and $M^{-1}\underline{g}$. It suffices to calculate $\underline{z} = M^{-1}\underline{r}$ once an iteration (see Algorithm A.1). In the following we deduce a preconditioner M which yields $\mathrm{cond}(M^{-1}H) = \mathrm{cond}(\tilde{H}^{\mathcal{L}})$.

Again, let N be the matrix mapping the Lagrangian basis B_p onto the hierarchical basis $B_p^{\mathcal{L}}$ (see (4.31)). By linear algebra of coordinate transformation, we have

$$H^{\mathcal{L}} = NHN^T, \qquad g^{\mathcal{L}} = Ng, \quad \text{and} \quad N^T y^{\mathcal{L}} = y\,. \tag{4.37}$$

Now, let D be the diagonal matrix which contains the reciprocals H_{kk}^{-1} of the diagonal entries of H. We may write the following equivalences to the diagonally preconditioned system:

$$\begin{aligned} & DH^{\mathcal{L}}y^{\mathcal{L}} &&= Dg^{\mathcal{L}} \\ \iff \quad & DNHy &&= DNg \\ \iff \quad & (N^TDN)Hy &&= (N^TDN)g. \end{aligned}$$

Since $N^TDNH = N^TDH^{\mathcal{L}}N^{-T}$, the matrices N^TDNH and $DH^{\mathcal{L}}$ are spectrally equivalent. Further, $DH^{\mathcal{L}}$ is spectrally equivalent to $\tilde{H}^{\mathcal{L}}$ from Theorem 4.32.

Remark 4.39.

1. Let $M := (N^TDN)^{-1}$ be the preconditioner for the system $H\underline{y} = -\underline{g}$. The spectral equivalence of $M^{-1}H$ and $\tilde{H}^{\mathcal{L}}$ implies $\mathrm{cond}(M^{-1}H) = \mathcal{O}(p^2(1+\log^2 p))$ by Theorem 4.32.

2. In Lemma B.5 the matrix-vector products $N\underline{v}^q$ and $N^T\underline{v}$ (see (B.4)) can be computed efficiently due to the tensor product structure of the bases. Analogously, Lemma B.5 holds for N given in (4.31), if we replace $\underline{v}$ by $\underline{v}^{\mathcal{L}}$, $\underline{v}^q$ by $\underline{v}$, and the one dimensional coordinate transformation (B.3) by

$$\begin{pmatrix} \mathcal{L}_0(\xi) \\ \vdots \\ \mathcal{L}_p(\xi) \end{pmatrix} = C \begin{pmatrix} \lambda_0^p(\xi) \\ \vdots \\ \lambda_p^p(\xi) \end{pmatrix} \qquad \text{with } C := \begin{pmatrix} \mathcal{L}_0(\xi_0^{p+1}) & \cdots & \mathcal{L}_0(\xi_p^{p+1}) \\ \vdots & & \vdots \\ \mathcal{L}_p(\xi_0^{p+1}) & \cdots & \mathcal{L}_p(\xi_p^{p+1}) \end{pmatrix}.$$

3. Analogously to Remark B.6, the matrix-vector products $N\underline{v}$ and $N^T\underline{v}^{\mathcal{L}}$ cost $\mathcal{O}(p^3)$ floating point operations.

4. In actual computation matrices and vectors need not physically rearranged. With Lemma B.5 the preconditioning step $\underline{z} = M^{-1}\underline{r} = N^TDN\underline{r}$ can be computed by the following three steps:

(i) Compute $\underline{\underline{v}} = C\underline{\underline{r}}\,C^T$.

(ii) Set $\underline{v} = D\,\underline{v}$.

(iii) Compute $\underline{\underline{z}} = C^T\underline{\underline{v}}\,C$.

Thus, a preconditioning step costs $\mathcal{O}(p^3)$ floating point operations. When the matrix-vector products are performed by Algorithm B.3, D is not known and has to be calculated by Algorithm B.7 at a cost of $\mathcal{O}(p^4)$ once before the cg-iterations start.

5. One preconditioned cg-iteration costs $\mathcal{O}(p^4)$ floating points operations, when we use Algorithm B.3 and an implementation of the preconditioner according to 4. Thus, the approximate solution of the linear system costs $\mathcal{O}(p^5(1+\log p))$ floating point operations due to 1 and Corollary 4.19.

4.5.2 Preconditioning the constrained problem

Step 7 of Algorithm 4.3 requires the solution of the linear problem $H_{\mathcal{F}}\,y_{\mathcal{F}} = -\underline{g}_{\mathcal{F}}$. Generalizing an idea of O'Leary (cf. [O'L80]), we show that the preconditioners known for the full linear problem can be used for the restricted linear problem. The condition number of the preconditioned restricted linear system will be less than or equal that of the full system.

Suppose that P is the permutation matrix corresponding to the current partitioning $\mathcal{A} \cup \mathcal{F}$ into active and free variables given by

$$PHP^T = \begin{pmatrix} H_{II} & H_{IJ} \\ H_{IJ}^T & H_{JJ} \end{pmatrix} \qquad \text{where } I := \mathcal{A},\ J := \mathcal{F}.$$

Again, let M denote the positive definite preconditioner for H. As an effective preconditioning step $\underline{z}_J = \overline{M}^{-1}\underline{r}_J$ for the remaining system $H_{JJ}\,\underline{y}_J = -\underline{g}_J$ we give the following algorithm.

Algorithm 4.4 $\underline{z}_J = \overline{M}^{-1}\underline{r}_J$; Preconditioning step for $H_{JJ}\,\underline{y}_J = -\underline{g}_J$

---

1. Set $\underline{r} = P^T \begin{pmatrix} 0 \\ \underline{r}_J \end{pmatrix}$.

2. Compute $\underline{\underline{v}} = C\underline{\underline{r}}\,C^T$. Set $\underline{v} = D\underline{v}$. Compute $\underline{\underline{z}} = C^T\underline{\underline{v}}\,C$.
(The $\underline{\underline{\cdot}}$ notation is introduced in Lemma B.5, p. 146.)

3. Set $\begin{pmatrix} \underline{z}_I \\ \underline{z}_J \end{pmatrix} = P\underline{z}$.

Partitioning and rearranging the matrix M^{-1} in a manner corresponding to the current rearrangement of H we get

$$PM^{-1}P^T \;=\; \begin{pmatrix} (M^{-1})_{II} & (M^{-1})_{IJ} \\ (M^{-1})_{JI} & (M^{-1})_{JJ} \end{pmatrix}.$$

The effect of Algorithm 4.4 reads as $\underline{z}_J = (M^{-1})_{JJ}\,\underline{r}_J$. To estimate the condition number of $\overline{M}^{-1}H_{JJ}$, we consider the simple situation when there is only one active variable, i.e., $\operatorname{card} I = 1$. Then, the general situation with $1 \le \operatorname{card} I \le \operatorname{card}(I \cup J)$ can be deduced by induction.

Lemma 4.40. Let $\overline{M}^{-1}$ be obtained using Algorithm 4.4. Then it is symmetric positive definite. Suppose the dimension of $\overline{M}^{-1}$ is $N-1$ where N is the dimension of M, and let

$$\begin{aligned} \lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N > 0 \qquad & \text{be the eigenvalues of } M^{-1}H, \\ \overline{\lambda}_1 \ge \overline{\lambda}_2 \ge \cdots \ge \overline{\lambda}_{N-1} > 0 \qquad & \text{be the eigenvalues of } \overline{M}^{-1}H_{JJ}. \end{aligned}$$

Then $\quad \lambda_1 \ge \overline{\lambda}_1 \ge \lambda_2 \ge \overline{\lambda}_2 \ge \cdots \ge \lambda_{N-1} \ge \overline{\lambda}_{N-1} \ge \lambda_N$.

Proof. $\overline{M}^{-1}$ is symmetric positive definite, since it is a principal sub-matrix of a symmetric positive definite matrix. For ease of notation we introduce the notations $G := (D^{1/2}N)$ and $G_{JJ} := ((D^{1/2})_{JJ}\,N_{JJ})$. Note that

$$\begin{aligned} \det(M^{-1}H - \lambda I) \;&=\; \det(G^T G H - \lambda I) \;&&=\; \det(GHG^T - \lambda I) \\ \text{and} \quad \det(\overline{M}^{-1}H_{JJ} - \lambda I) \;&=\; \det(G_{JJ}^T G_{JJ} H - \lambda I) \;&&=\; \det(G_{JJ} H G_{JJ}^T - \lambda I)\,. \end{aligned}$$

By the Courant-Fischer characterization of eigenvalues (cf. [GVL96, Theorem 8.1.2]),

$$\begin{aligned} \lambda_k &= \max_{\dim(S)=k} \min_{0\ne\underline{v}\in S} \frac{\underline{v}^T G H G^T \underline{v}}{\underline{v}^T \underline{v}} \\ &= \max_{\dim(S)=k} \min_{0\ne\underline{w}\in S} \bigl\{ \underline{w}^T H \underline{w} \,|\, \|G^{-T}\underline{w}\| = 1 \bigr\} \qquad \text{for all } k = 1, \ldots, N, \end{aligned}$$

where S is any subspace of $\mathbb{R}$ with the indicated dimension. Without loss of generality, we suppose that H_{JJ} is obtained from H by deleting the first row and column. Then

$$\begin{aligned} \overline{\lambda}_k &= \max_{\dim(S_J)=k} \min_{0\ne\underline{v}_J\in S_J} \frac{\underline{v}_J^T G_{JJ} H_{JJ} G_{JJ}^T \underline{v}_J}{\underline{v}_J^T \underline{v}_J} \\ &= \max_{\dim(S_J)=k} \min_{0\ne\underline{w}\in S_J} \bigl\{ \underline{w}_J^T H_{JJ} \underline{w}_J \,|\, \|G_{JJ}^{-T}\underline{w}_J\| = 1 \bigr\} \\ &= \max_{\dim(S)=k+1} \min_{0\ne\underline{w}\in S} \bigl\{ \underline{w}^T H \underline{w} \,|\, w_1 = 0,\ (G^{-T}\underline{w})_1 = 0,\ \ \|G^{-T}\underline{w}\| = 1 \bigr\}. \end{aligned}$$

Here, S_J denotes any subspace of $\mathbb{R}^{N-1}$. Therefore, $\overline{\lambda}_k \ge \lambda_{k+1}$ for $k = 1, \ldots, N-1$. The corresponding min-max characterization of eigenvalues can be used in an analogous arrangement to prove that $\overline{\lambda}_k \le \lambda_k$ for $k = 1, \ldots, N-1$. □

Now, we can state the estimate for the condition number of the preconditioned reduced system $\overline{M}^{-1}H_{JJ}$. We stress that we only assumed M to be a positive definite preconditioner of the full system. This means that we can use other preconditioners $\breve{M}$ known for p-hierarchical bases for the preconditioning of reduced systems.

Theorem 4.41. The following relation holds for the conditions of the preconditioned sub-matrix H_{JJ} and the preconditioned matrix H.

$$\operatorname{cond}(\overline{M}^{-1}H_{JJ}) \le \operatorname{cond}(M^{-1}H)\,.$$

Proof. Let λ_1 and λ_N the largest and the smallest eigenvalues of $M^{-1}H$, let $\overline{\lambda}_1$ and $\overline{\lambda}_{N_J}$ the largest and the smallest eigenvalues of $\overline{M}^{-1}H_{JJ}$. Here, N_J denotes the dimension of H_{JJ}. It follows by Lemma 4.40 and induction that $\lambda_1 \geq \overline{\lambda}_1 \geq \overline{\lambda}_{N_J} \geq \lambda_N$. Noting

$$\operatorname{cond}(\overline{M}^{-1}H_{JJ}) = \frac{\overline{\lambda}_1}{\overline{\lambda}_{N_J}} \leq \frac{\lambda_1}{\lambda_N} = \operatorname{cond}(M^{-1}H)$$

finishes the proof. □

Corollary 4.42. Rewriting the reduced matrix-vector product

$$H_{JJ}\,\underline{y}_J \;=\; \left(H\begin{pmatrix}0\\ \underline{y}_J\end{pmatrix}\right)_J$$

where $(\cdot)_J$ means that we take the J components of the vector, we can use Algorithm B.3 to compute $H_{JJ}\,\underline{y}_J$. It follows analogously to Remark 4.39.5 that one $\overline{M}$-preconditioned cg-iteration costs $\mathcal{O}(p^4)$ floating points operations. Further, the approximate solution of the reduced linear system $H_{JJ}\underline{y}_J = -\underline{g}_J$ by the $\overline{M}$-preconditioned cg-scheme needs $\mathcal{O}(p^5(1+\log p))$ floating point operations.

4.5.3 Space decomposition methods for the constrained problem

In [BCMP91] a preconditioning technique based on space decomposition is developed for the p-version FEM by Babuška et al. There, the condition number of the preconditioned system grows at most as $\mathcal{O}(1+\log^2 p)$.

Ainsworth uses a similar approach with an additional domain decomposition for the hp-FEM in [Ain96] and gets a growth of $\mathcal{O}((1+\log^2 p)(1+\log^2(Hp/h))$ for the preconditioned system, where p is the polynomial degree, h is the size of the elements, and H is the size of the sub-domains. However, both methods demand a p-hierarchical basis given by the anti-derivatives of Legendre polynomials.

Nevertheless, we can use both preconditioners as preconditioners for the Lagrangian basis by a combination with the basis transformation N. Let $\check{M}$ denote one of the mentioned space decomposition preconditioners. Remark 4.39.4 presents a preconditioner for the Lagrangian basis which reduces the condition number to that of a diagonally preconditioned hierarchical basis. Replacing the diagonal matrix D by $\check{M}^{-1}$ reduces the condition number to that of the $\check{M}$ preconditioned system. This becomes clear when we write

$$\underline{z} = N^T\check{M}^{-1}N\underline{r}\,.$$

Analogously to Section 4.5.1, we obtain

$$\begin{aligned}\operatorname{cond}(N^T\check{M}^{-1}NH) &= \operatorname{cond}(\check{M}^{-1}H^{\mathcal{L}})\\ &= \begin{cases}\mathcal{O}((1+\log^2 p)) & \text{for } \check{M} \text{ from [BCMP91]},\\ \mathcal{O}((1+\log^2 p)(1+\log^2(Hp/h)) & \text{for } \check{M} \text{ from [Ain96]}.\end{cases}\end{aligned}$$

An inspection of the proof of Lemma 4.40 shows that we only use the symmetry and the positive definiteness of preconditioner M when going over to the preconditioner $\overline{M}$

of the reduced system. Thus, we can define the preconditioner $\overline{\check{M}}$ by replacing *Step 2* of Algorithm 4.4 by *Step*

2'. Compute $\underline{\underline{v}} = C\underline{\underline{r}}\,C^T$. Set $\underline{v} = \check{M}^{-1}\underline{v}$. Compute $\underline{\underline{z}} = C^T\underline{\underline{v}}\,C$.

Analogously to Lemma 4.40, we obtain $\mathrm{cond}(\overline{\check{M}}) \leq \mathrm{cond}(\check{M})$.

Nevertheless, it is in order here to recall that the costs for the approximate solution of the linear systems are dominated by the $\mathcal{O}(p^4)$ floating point operations needed for the matrix-vector products. The implementation of a space decomposition preconditioner will reduce only the number of these products. Due to Corollary 4.21 and Theorem 4.32 (4.34a) the diagonal preconditioned cg-scheme introduced in Section 4.5.2 leads to $\mathcal{O}(p(1+\log p))$ cg-iterations. The $\overline{\check{M}}$ preconditioned cg-scheme leads to $\mathcal{O}(1+\log p)$ cg-iterations in case of $\check{M}$ from [BCMP91], to $\mathcal{O}((1+\log p)(1+\log Hp/h))$ cg-iterations in case of $\check{M}$ from [Ain96].

4.6 Non uniform *hp*-meshes

In Section 2.4.1, we developed non uniform *hp*-FE spaces and *hp*-FE subsets. To achieve the continuity across inter-element boundaries between quadrilaterals with hanging nodes and with different polynomial degrees, we demanded that the restrictions of the discrete functions onto the edges are polynomials of degree $p_e = \min\{p_Q \,|\, Q \in \mathcal{T},\, e \cap \overline{Q}_i \neq \emptyset\}$, i.e., the minimum of polynomial degrees of the adjacent quadrilaterals. The obstacle condition was controlled on the set $G_{\vec{p}}$ given by (2.42) which contained the interior Gauss-Lobatto points of the quadrilaterals and the edge associated Gauss-Lobatto points $G_{\mathcal{E},\vec{p}}$.

In Section 4.1 (4.4), we reformulated the discrete minimization problem in coordinate notation for the quasi-uniform *p*-version. There, the continuity of the FE functions was guaranteed by the choice of continuous global basis functions. Now, it would be possible to define appropriate nodal global basis functions for the non-uniform *p*-version. But it would be a disadvantage that both, the assembly of the global stiffness matrix H and the quadrature would become complicated. One would need an extra quadrature routine to integrate the edge associated nodal basis functions. As an alternative we maintain the continuity by means of linear algebra.

Let

$$B := \{b_{Q,i,j} \,|\, Q \in \mathcal{T}, 0 \leq i,j \leq p_Q\}$$

where $b_{Q,i,j}$ are the tensor product basis functions given by (4.1) and p_Q denotes the polynomial degree on Q given by the degree vector $\vec{p}$. Of course, B contains discontinuous basis functions. We split up B into the basis functions associated to the interior nodes on the quadrilaterals Q and to those associated to the nodes on the boundaries of the quadrilaterals,

$$\underline{b}_I := \{b_{Q,i,j} \,|\, Q \in \mathcal{T}, 1 \leq i,j \leq p_Q - 1\} \quad \text{and} \quad \underline{b}_O := \{b_{Q,i,j} \,|\, Q \in \mathcal{T}, i,j \in \{0, p_Q\}\}\,.$$

Introducing a global numbering, we can represent all linear combinations $v \in \operatorname{span} B$ using

coordinate notation as

$$v = \begin{pmatrix} \underline{v}_I \\ \underline{v}_O \end{pmatrix}^T \begin{pmatrix} \underline{b}_I \\ \underline{b}_O \end{pmatrix} \qquad \text{for a unique} \qquad \begin{pmatrix} \underline{v}_I \\ \underline{v}_O \end{pmatrix} \in \mathbb{R}^{\operatorname{card} B}.$$

In the following, we take up the notation of Section 2.4.1, pp. 63, 64. It is the goal to express $V_{\vec{p}}$, $V_{\vec{p},g_D}$, and $K_{\vec{p},g_D}$ from Definition 2.27 by appropriate vectors $(\underline{v}_I^T, \underline{v}_O^T)$. Therefore, we define an auxiliary basis associated to the Gauss-Lobatto points on the edges

$$\hat{G}_{\mathcal{E},\vec{p}} := G_{\mathcal{E},\vec{p}} \cup \{x_h \,|\, x_h \text{ is a ficticious hanging node}\}$$

(see (2.41)). Let $k = 1, \ldots, N_e$, $N_e := \operatorname{card} \hat{G}_{\mathcal{E},\vec{p}}$ a global counting for the $x_k \in \hat{G}_{\mathcal{E},\vec{p}}$. By construction of $\hat{G}_{\mathcal{E},\vec{p}}$, there exists an edge $e \in \mathcal{E}$ and a local edge number $i \in \{0, \ldots, p_e\}$ such that

$$x_k = x_i^{p+1,e}\,. \tag{4.38}$$

Here, the pair (e, i) is not determined uniquely by k, since vertices and non ficticious hanging nodes are elements of multiple edges. Thus, we define the set

$$Y_k := \{(e,i) \in \mathcal{E} \times \{0, \ldots, p_e\} \,|\, (4.38) \text{ holds for } (e,i)\} \quad \text{for all } k = 1, \ldots, N_e\,.$$

Let $Q \in \mathcal{T}$ be the quadrilateral with the lowest ordering number and $j \in \{1, \ldots, 4\}$ the local edge number which fulfill $e = F_Q(e^{(j)})$ (see (2.38)). Q and j are determined uniquely by e. For every $x \in e$ there exists a unique $\tilde{\xi} = (\tilde{\xi}_1, \tilde{\xi}_2) \in e^{(j)}$ such that $x = F_Q(\tilde{\xi})$. Now, we can define the local edge basis by

$$c_{e,i}(x) := \begin{cases} \lambda_i^{p_e}(\tilde{\xi}_1) & \text{for } x \in e \text{ and } j \in \{1, 3, \} \\ \lambda_i^{p_e}(\tilde{\xi}_2) & \text{for } x \in e \text{ and } j \in \{2, 4\} \\ 0 & \text{for } x \notin e \end{cases}, \tag{4.39}$$

and assemble the global nodal edge basis function c_k associated to ξ_k,

$$c_k := \sum_{(e,i) \in Y_k} c_{e,i} \qquad \text{for all } k = 1, \ldots, N_e\,. \tag{4.40}$$

Taking off the basis functions c_k corresponding to ficticious hanging nodes, we obtain a partition of the basis functions which we note by the column vector $\left(\begin{smallmatrix} \underline{c}_E \\ \underline{c}_F \end{smallmatrix}\right)$. Now, any $w \in \operatorname{span}\{c_k \,|\, k = 1, \ldots, N_e\}$ can be represented using the coordinate representation

$$w = \begin{pmatrix} \underline{w}_E \\ \underline{w}_F \end{pmatrix}^T \begin{pmatrix} \underline{c}_E \\ \underline{c}_F \end{pmatrix} \qquad \text{for a unique} \qquad \begin{pmatrix} \underline{w}_E \\ \underline{w}_F \end{pmatrix} \in \mathbb{R}^{N_e}\,.$$

Here, $\underline{w}_F$ is already determined, when we require the continuity in the ficticious nodes. Let $w \in \operatorname{span}\{c_k \,|\, k = 1, \ldots, N_e\}$ be continuous, let w_f be the coefficient scaling the basis function c_f associated to the ficticious node x_f, and let $\hat{e}$ be the edge with $x_f \in \hat{e}$, but x_f is not an end point of $\hat{e}$. Then,

$$w_f = w(x_f) = \sum_{i=0}^{p_{\hat{e}}} w_{\hat{e},i} c_{\hat{e},i}(x_f)\,, \tag{4.41}$$

i.e., all ficticious degrees of freedom can be resolved locally. Switching back to global numbering, there exists a matrix $C_{E,F}$ with

$$C_{E,F} \underline{w}_E = \underline{w}_F\,. \tag{4.42}$$

Similarly, we get the continuity of $v \in \operatorname{span} B$ by local operations. Let e be the edge of Q with the local edge number $j = 1$. The restriction $v_{|e}$ can be represented for all $x \in e$ by the linear combinations

$$v(x) = \begin{pmatrix} v_{e,0} \\ \vdots \\ v_{e,p_e} \end{pmatrix}^T \begin{pmatrix} c_{e,0}(x) \\ \vdots \\ c_{e,p_e}(x) \end{pmatrix} \qquad \text{using the local basis on the edge } e,$$

$$\text{and} \quad v(x) = \begin{pmatrix} v_{Q,0,0} \\ \vdots \\ v_{Q,p_Q,0} \end{pmatrix}^T \begin{pmatrix} b_{Q,0,0}(x) \\ \vdots \\ b_{Q,p_Q,0}(x) \end{pmatrix} \qquad \text{using the local basis on the quadrilateral } Q.$$

Inserting $x_i = F_Q(\xi_i^{p_Q+1}, -1)$, $i = 0, \ldots, p_Q$, into both representations yields

$$v_{Q,i,0} = \begin{pmatrix} v_{e,0} \\ \vdots \\ v_{e,p_e} \end{pmatrix}^T \begin{pmatrix} \lambda_0^{p_e}(\xi_i^{p_Q+1}) \\ \vdots \\ \lambda_{p_e}^{p_e}(\xi_i^{p_Q+1}) \end{pmatrix} \qquad \text{for } 0 \le i \le p_Q \,,$$

and further

$$\begin{pmatrix} v_{Q,0,0} \\ \vdots \\ v_{Q,p_Q,0} \end{pmatrix} = C_{e,Q} \begin{pmatrix} v_{e,0} \\ \vdots \\ v_{e,p_e} \end{pmatrix} \qquad \text{with } \; C_{e,Q} := \begin{pmatrix} \lambda_0^{p_e}(\xi_0^{p_Q+1}) & \cdots & \lambda_{p_e}^{p_e}(\xi_0^{p_Q+1}) \\ \vdots & & \vdots \\ \lambda_0^{p_e}(\xi_{p_Q}^{p_Q+1}) & \cdots & \lambda_{p_e}^{p_e}(\xi_{p_Q}^{p_Q+1}) \end{pmatrix} . \tag{4.43}$$

In case of a local edge number $j \in \{2, 3, 4\}$, we get analog expressions for

$$\begin{pmatrix} v_{Q,p_Q,0} \\ \vdots \\ v_{Q,p_Q,p_Q} \end{pmatrix} , \quad \begin{pmatrix} v_{Q,0,p_Q} \\ \vdots \\ v_{Q,p_Q,p_Q} \end{pmatrix} , \quad \text{and} \quad \begin{pmatrix} v_{Q,0,0} \\ \vdots \\ v_{Q,0,p_Q} \end{pmatrix} , \quad \text{respectively.}$$

Combining the local linear mappings $C_{e,Q}$ for all $Q \in \mathcal{T}$ and turning back to global numbering yields a matrix $C_{EF,O}$ which determines the coefficient vector $\underline{v}_O$ by

$$\underline{v}_O = C_{EF,O} \begin{pmatrix} \underline{v}_E \\ \underline{v}_F \end{pmatrix} .$$

Replacing $\underline{v}_F$ by the right hand side of (4.42), we get a matrix $C_{E,O}$ such that

$$\underline{v}_O = C_{E,O}\, \underline{v}_E \,.$$

Collecting the above arguments, we can conclude with the following proposition which gives the coordinate expressions of $V_{\vec{p}}$, $V_{\vec{p},g_D}$, and $K_{\vec{p},g_D}$.

Proposition 4.43. Let $\{1, \ldots, N\}$, $N := \operatorname{card} G_{\vec{p}}$ be the index set of $G_{\vec{p}}$, let $I \cup E = \{1, \ldots, N\}$ be the partition of the index set obtained by taking off the Gauss-Lobatto points $G_{\mathcal{E},\vec{p}}$, and let $N_I := \operatorname{card} I$, $N_E := \operatorname{card} E$. Using the above coordinate notation $(v_I, v_O, v_E, v_F)^T$, $V_{\vec{p}}$, $V_{\vec{p},g_D}$, and $K_{\vec{p},g_D}$ from Definition (2.27) can be rewritten

as

$$V_{\vec{p}} = V_{\vec{p}}(\mathcal{T}) = \left\{ \begin{pmatrix} v_I \\ v_O \end{pmatrix}^T \begin{pmatrix} b_I \\ b_O \end{pmatrix} \;\middle|\; v_I \in \mathbb{R}^{N_I} \text{ and } v_O = C_{E,O} v_E,\ v_E \in \mathbb{R}^{N_E} \right\},$$

$$V_{\vec{p},g_D} = \left\{ \begin{pmatrix} v_I \\ v_O \end{pmatrix}^T \begin{pmatrix} b_I \\ b_O \end{pmatrix} \;\middle|\; v_I \in \mathbb{R}^{N_I} \text{ and } v_O = C_{E,O} v_E,\ v_E \in \mathbb{R}^{N_E}_{=g_D} \right\},$$

$$K_{\vec{p},g_D} = \left\{ \begin{pmatrix} v_I \\ v_O \end{pmatrix}^T \begin{pmatrix} b_I \\ b_O \end{pmatrix} \;\middle|\; v_I \in \mathbb{R}^{N_I}_{\geq \underline{\psi}_I} \text{ and } v_O = C_{E,O} v_E,\ v_E \in \mathbb{R}^{N_E}_{=g_D} \cap \mathbb{R}^{N_E}_{\geq \underline{\psi}_E} \right\}.$$

Here, $\underline{\psi} = (\psi(x_k))_{x_k \in G_{\vec{p}}}$ is the obstacle vector corresponding to $G_{\vec{p}}$ and $\underline{\psi}^T = (\underline{\psi}_I^T, \underline{\psi}_E^T)$ denotes its partition with respect to the index sets I, E. Further,

$$\mathbb{R}^{N_E}_{=\underline{g}_D} := \{\underline{w} \in \mathbb{R}^{N_E} \,|\, w_k = g_D(x_k) \text{ for all } x_k \in \Gamma_{D,\vec{p}}\}$$

and $\mathbb{R}^{N_I}_{\geq \underline{\psi}_I}$, $\mathbb{R}^{N_E}_{\geq \underline{\psi}_E}$ are given by introducing $\underline{\psi}_I$, N_I, and $\underline{\psi}_E$, N_E for $\underline{\psi}$, N into

$$\mathbb{R}^{N}_{\geq \underline{\psi}} := \{\underline{w} \in \mathbb{R}^{N} \,|\, \underline{w} \geq \underline{\psi}\}.$$

Here, the relation $\geq$ has to be understood component-wise.

Corollary 4.44. Analogously to Section 4.1 (4.3), we define

$$A\left(\begin{pmatrix} \underline{w}_I \\ \underline{w}_E \end{pmatrix}\right) := A\left(\begin{pmatrix} \underline{w}_I \\ C_{E,O}\, \underline{w}_E \end{pmatrix}^T \begin{pmatrix} \underline{b}_I \\ \underline{b}_O \end{pmatrix}\right) \tag{4.44}$$

for $(w_I, w_E)^T \in \mathbb{R}^{N_I}_{\geq \underline{\psi}_I} \times (\mathbb{R}^{N_E}_{=g_D} \cap \mathbb{R}^{N_E}_{\geq \underline{\psi}_E})$. Again, we use the arguments $\underline{w} \in \mathbb{R}^N$ and $w \in V_{\vec{p}}$ to differentiate between the coordinate and the vector notation of A. We rewrite the discrete obstacle problem from Theorem 2.7(i) equivalently as

$$\underline{u} \text{ minimizes } A \text{ on } \mathbb{R}^{N_I}_{\geq \underline{\psi}_I} \times (\mathbb{R}^{N_E}_{=g_D} \cap \mathbb{R}^{N_E}_{\geq \underline{\psi}_E}).$$

Remark 4.45. In practical computations, there is no obligation to perform global operations on matrices, or to assemble the global matrix H. Particularly, there is no need to store the global matrix $C_{E,O}$. Using an iterative solver such as Algorithm 4.3, it suffices to store the local matrices H_Q and the local right hand sides g_Q yielded by quadrature routines, and the vectors $(\underline{v}_I, \underline{v}_E)^T$, $(\underline{\psi}_I, \underline{\psi}_E)^T$. Taking a local counting of the components of v_F and v_O on the quadrilaterals Q, the auxiliary vectors $v_{F,Q}$ and $v_{O,Q}$ are generated by exclusively local calculations of (4.41) and (4.43). These can be done simultaneously on the quadrilaterals.

Remark 4.46. The preconditioners given in Remark 4.39 for the linear system of the unconstrained problem and in Algorithm 4.4 for the linear system of the constrained problem can be generalized straightforwardly to the non uniform *hp*-version, because they are given as local operations and $(v_I, v_O)^T$ offers a local representation of v.

Remark 4.47. In Section 2.2 we defined p-FE on the reference triangle $\tilde{T}$. Global nodal edge basis functions c_k (see ((4.40)) can be introduced, when Ω is divided into triangles. Of course, the local edge number j has to be in $\{1, 2, 3\}$, and the definition of the local edge basis functions $c_{e,i}$ (see (4.39)) has to be accommodated. The continuity on the inter-element boundaries and the obstacle condition are obtained completely analogously as in Proposition 4.43 by means of linear algebra. Even hanging nodes are allowed.

4.7 Numerical experiments

A practical way to design structures that combine minimum weight with maximum strength is to experiment with soap bubbles. To construct surface structures, for example, architects would bend a wire frame, to the outline of a vaulted alcove and dip the wire frame into a soap sud. The shape formed by the soap film on the outline gives the most economical form in the real structure. The book *Natürliche Konstruktionen* from Otto presents examples of roofs, vaults, and tents which shapes were designed by soap film models (cf. p. 72 in [Ott85]). The link between mathematics and architectural construction is discussed by Emmer in [Emm96].

Because of the tension within the liquid, a soap film will form the smallest surface area between edges. The effect of gravity is negligible for small areas of soap films because their weight is minimal. Thus, a film stretched out across a hoop, for example, will form a flat disc, and a film around a volume of air will form a sphere. We refer to the books *Soap bubbles and forces which mould them* from C. V. Boys [Boy90] and *Demonstrating science with soap films* from D. Lovett [Lov94] for the physical backgrounds of soap films and bubbles. The shape of a soap film on a wire frame can be modeled mathematically according to the following minimal surface problem.

The minimal surface model problem with inhomogeneous boundary data.
Let $\Omega \subset \mathbb{R}^2$ be a bounded Lipschitz domain. Let the wire frame $(x, g_D(x))$ be given by the Dirichlet data $g_D \in H^{1/2}(\partial\Omega)$ on the Lipschitz boundary $\partial\Omega$. We ask for the element $u \in H^1_{g_D}(\Omega)$ of least area, namely,

$$A(u) \le A(v) \quad \text{for all } v \in H^1_{g_D}(\Omega) \quad \text{with} \quad A(v) := \int_\Omega \sqrt{1 + |\nabla v|^2}\, \mathrm{d}x\,. \tag{4.45}$$

Further, we introduce an obstacle given by the $\psi \in C^0(\overline{\Omega}) \cap H^1(\Omega)$ with $\psi \le g_D$ almost everywhere in a neighborhood of $\partial\Omega$ into the minimal surface problem and demand that the surface u fulfills the condition $u(x) \ge \psi(x)$ almost everywhere on Ω. Physically speaking, the soap films must lie above the obstacle. Now, we ask for the element

$$u \in K := \{v \in H^1_{g_D}(\Omega) \,|\, v \ge \psi \text{ a.e. on } \Omega\}$$

of least area, namely

$$A(u) \le A(v) \quad \text{for all } v \in K\,.$$

The functional defined by $A(v)$ is a particular case of the functional defined by (1.8) in Chapter 1. This becomes clear, when we take $\sigma = 0$, $f \equiv 0$, $\Gamma_N = \emptyset$, and

$$p(t) := \sqrt{1 + t^2} \quad \text{in (1.8).}$$

Unfortunately, the corresponding $\rho(t)$ does not fulfill the condition (1.6), here,

$$\rho_0 \le \rho(t) = (1+t^2)^{-1/2} \le \rho_1\,, \qquad \rho_2 \le \rho(t) + t\rho'(t) = (1+t^2)^{-3/2} \le \rho_3 \quad \text{for all } t \in \mathbb{R}_{\ge 0}\,,$$

demanded for existence and uniqueness of a minimum by Theorem 1.22 and Theorem 1.23. We can take $\rho_1 = \rho_3 = 1$ as upper bounds, but the lower bounds $\rho_0 > 0$ and $\rho_2 > 0$ do not exist because t can become arbitrarily large. To achieve the existence and the regularity

of the minimal surface, we need an a priori estimate for the gradient of the solution. With the boundedness of $t = |\nabla u|$ it suffices to show that there exist positive constants ρ_0 and ρ_2, when t is bounded arbitrarily.

An extensive analysis of the contact minimal surface problem is given by Kinderlehrer and Stampacchia in terms of continuously differentiable locally coercive vector fields $a(x) := \rho(|x|)x,\ x \in \mathbb{R}^d,\ d \geq 1$. Due to [KS80, Theorem IV.4.3], there exists a unique minimal surface $u \in H^2(\Omega)$, when we assume homogeneous boundary data, a convex domain Ω with smooth boundary $\partial\Omega$, and an obstacle $\psi \in C^2(\Omega)$.

Solving the unconstrained discrete minimum problems. The discretization of the minimum surface problem yields discrete nonlinear unconstrained and constrained minimum problems. The unconstrained minimum problems are solved using the inexact Newton backtracking method of Algorithm 4.2 with the initial $\underline{u}_0 \in \mathbb{R}^N$ with the components given by $u_{0,i} = 0$ for the interior degrees of freedom and $u_{0,i} = g_D(x_i)$ for the boundary degrees of freedom where x_i is a Gauss-Lobatto points on the boundary $\partial\Omega$. The Newton iterations are performed until the Euclidian norm of the gradient $\|\underline{g}(\underline{u}_{k+1})\|_2$ is smaller or equal $\epsilon = 10^{-12}$ (see *Step 7*). The linear systems are solved using diagonally preconditioned conjugate gradient iterations until the stopping criterion of *Step 2* is fulfilled for the forcing term η_k computed by *Step 7* (4.5) with the parameters $\gamma = 0.9$ and $\alpha = 2$. As initial forcing term, we take $\eta_0 = 1.0$. The line-search with Algorithm 4.1 is done with the "Armijo"-parameter $\delta = 10^{-4}$ and the backtracking parameter $\beta = 0.5$.

Solving the constrained discrete minimum problems. The constrained discrete minimum problems are solved using Algorithm 4.3 with the initial $\underline{u}_0$ given by the vector components

$$u_{0,i} = \begin{cases} \max\{\psi(x_i), 0\}, & \text{if } x_i \in \Omega\,, \\ g_D(x_i), & \text{if } x_i \in \partial\Omega\,. \end{cases}$$

where $x_i \in G_p$ denote the Gauss-Lobatto points corresponding to the respective degrees of freedom. The projected gradient and the Newton iterations of Algorithm 4.3 are performed until the Euclidian norm condition $\|\underline{u} - P_{\underline{\psi}}(\underline{u} - \nabla A(\underline{u}))\|_2 \leq \epsilon = 10^{-10}$ holds. This ensures $\|g_{\underline{\psi}}(\underline{u}_k)\|_2 \leq \epsilon$ due to Remark 4.13. We use $l = 3$ as the maximum number of projected gradient steps without changing the active set in *Step 5* and set $\eta_1 = 0.1$ as the decrease threshold for the projected gradient steps in *Step 5*, $\eta_2 = 0.1$ as the decrease threshold for the Newton steps in *Step 2*. The remaining parameters α, β, γ, δ, η_0, are taken as above in Algorithm 4.2. The linear problems are solved by a diagonally preconditioned cg-scheme.

Estimating the error of the discrete solution. The exact solutions u of the following numerical experiments are not known. Thus, the following lemma proves to be useful estimating the error of an approximate solution.

Lemma 4.48. Let the functional $A : H^1(\Omega) \to \mathbb{R}$ be given by (1.8) and let there exist positive constants ρ_i, $i = 0, 1, \ldots, 5$, such that the assumptions of Lemma 1.21 and Theorem 1.23 are satisfied. Further, let u be the unique minimizer of A according to Theorem 1.23 and let $\xi_0 := \frac{\rho_2}{2\rho_4+\rho_5}$. Then, for $v \in K$ with $|v - u|_{H^1(\Omega)} \leq \xi_0$, there holds

$$|v - u|^2_{H^1(\Omega)} \leq \tfrac{3}{\rho_2} \left|A(v) - A(u)\right|. \tag{4.46}$$

Defining the sequence $(\xi_k)_{k\in\mathbb{N}}$ recursively by ξ_0 as above and

$$\xi_{k+1} := \min\left\{\xi_0\ ,\ \sqrt{\frac{6|A(u)-A(v)|}{3\rho_2-(2\rho_4+\rho_5)\xi_k}}\right\},$$

we may estimate more precisely

$$|v-u|_{H^1(\Omega)} \le \inf\{\xi_k \,|\, k\in\mathbb{N}\}\,. \tag{4.47}$$

Proof. By setting

$$\varphi_w(t) := A(u+tw)\ , \qquad t\in\mathbb{R}_{\ge 0} \tag{4.48}$$

where $w\in H^1(\Omega)$, we know from the classical Taylor theorem that there exists a $\theta\in[0,1]$ such that

$$\begin{aligned}\varphi_w(1) &= \varphi_w(0)+\varphi_w'(0)+\varphi_w''(0)+\varphi_w'''(\theta)\\ &= A(u)+DA(u;w)+\tfrac{1}{2}D^2A(u;w,w)+\tfrac{1}{6}\theta D^3A(u;w,w,w)\,.\end{aligned}$$

Inserting $w=v-u$, $v\in K$, we estimate

$$DA(u;w)\ge 0 \qquad \text{and} \qquad D^2A(u;w,w)\ge \rho_2|v-u|^2_{H^1(\Omega)}$$

with Theorem 1.23(ii) and Lemma 1.21 (1.13), respectively, and

$$|D^3A(u;w,w,w)|\le(2\rho_4+\rho_5)|v-u|^3_{H^1(\Omega)}$$

with (1.16) from the proof of Lemma 1.21. Combining the last three inequalities and taking $t=1$ in (4.48), we get

$$2A(v)-2A(u)\ge(\rho_2-\tfrac{1}{3}(2\rho_4+\rho_5)|v-u|_{H^1(\Omega)})|v-u|^2_{H^1(\Omega)}\,. \tag{4.49}$$

Assuming $v\in K$ with $|v-u|_{H^1(\Omega)}\le\xi_0$ yields (4.46) because of

$$2|A(v)-A(u)|\ge\tfrac{2}{3}\rho_2|v-u|^2_{H^1(\Omega)}\,.$$

Moreover, assuming v as above, we obtain

$$2|A(v)-A(u)|\ge(\rho_2-\tfrac{1}{3}(2\rho_4+\rho_5)\xi_0)|v-u|^2_{H^1(\Omega)}$$

and consequently by the recursive definition of ξ_{k+1}

$$\xi_1^2\ge|v-u|^2_{H^1(\Omega)}\,.$$

Inserting ξ_{k+1} for ξ_1 and ξ_k for ξ_0 yields (4.47) by induction. The infimum exists since

$$\xi_k\ge\sqrt{\frac{2|A(v)-A(u)|}{\rho_2}}\,.$$

□

For the error estimation, we assume that the assumptions of Lemma 4.48 hold and neglect the consistency error, i.e., we ignore the fact that the discrete solutions $u_p\in K_{p,g_D}$ are not in K.

For the extrapolation of the minimum $\bar{A} := A(u)$ on the subset K we use a generalization of the Richardson extrapolation. For ease of notation we write $A_k := A(u_k)$ where $u_k := u_{p_k} \in K_{p_k,g_D}$ is the solution of the p-version FE computation with respect to Theorem 2.7. Let us assume that $(A_k)_{k\in\mathbb{N}}$ fulfills

$$A_k = \bar{A} + Cp_k^{-r}$$

with a real constants $\bar{A}$, C, r independent of k. Inserting $k+1$, $k+2$ instead of k into this equation gives us a system of three nonlinear equations and leads to

$$\left|\frac{A_k - A_{k+1}}{A_k - A_{k+2}}\right| = \frac{1 - \left(\frac{p_k}{p_{k+1}}\right)^r}{1 - \left(\frac{p_k}{p_{k+2}}\right)^r} .$$

Thus, finding a real $r > 0$ which fulfills the last equation allows the computation of $\bar{A}$ and C. In case of the h-version we extrapolate $\bar{A}$ analogously using $A_k = \bar{A} + Ch_k^r$. In [CP89] Christiansen and Petersen propose a more general extrapolation method starting from

$$A_k = \bar{A} + C_1 h_k^{r_1} + C_2 h_k^{r_2} + \dots$$

with positive real constants C_i, r_i, $i = 1, 2, \dots$. Nevertheless, the simple approach presented above is sufficient to extrapolate significant digits of $\bar{A}$ for the following experiment from the highest dimensional h- and p-version problems.

Then, the extrapolation $\bar{A}$ can be used to compute the experimental convergence rates with respect to the number of unknowns N_k assuming $|A(u_k) - \bar{A}| \approx CN_k^{\alpha}$ where C, α are constants independent of k. This leads to the experimental convergence rates

$$\alpha_{k+1} = \log\left|\frac{A_k - \bar{A}}{A_{k+1} - \bar{A}}\right| \Big/ \log\left(\frac{N_k}{N_{k+1}}\right) .$$

Beneath the disregarding of the inconsistency error, the main problem with the extrapolation of $\bar{A}$ and the experimental convergence rates is that we do not have a monotone decreasing sequence A_k in case of inhomogeneous boundary conditions or a real obstacle problem. In contrast to the FEM where we have $V_p \subset V_q$ for $p < q$, we do not have $V_{p,g_D} \subset V_{q,g_D}$ or $K_{p,g_D} \subset K_{q,g_D}$ for $p < q$. As a work around the extrapolation and the convergence rates are computed with respect to $p_{k+1} = p_k + 2$ instead of $p_{k+1} = p_k + 1$. It is a further advantage of taking $p_{k+1} = p_k + 2$ that oscillations of the error with respect to p caused by the symmetry of the solution do not influence the extrapolation and the convergence rates.

Experiment 4.49 (*Minimal surface with inhomogeneous boundary data***).** We take the unit square $\Omega := [-1, 1]^2$ and the minimal surface functional A as defined by (4.45). The inhomogeneous boundary data is given by $g_D := (1 - x_1^8) - (1 - x_2^8)$ for $x \in \partial\Omega$. We compare the h- and the p-version as follows:
For the h-version we start with a uniform square grid $\mathcal{T}_0$ on which we take the FE space $V_1(\mathcal{T}_0)$. The squares of the grid have the length h_0. Then, the elements $Q \in \mathcal{T}_0$ are refined into 4 elements by bisection of the edges. This yields the grid $\mathcal{T}_1$ and the grids $\mathcal{T}_k$, $k \geq 2$ by recursive continuation. The discrete minima $A(\underline{u}_k)$ on $V_{1,g_D}(\mathcal{T}_k)$ for different initial mesh widths are listed in Table 4.1.

For the p-version, again, we start with a uniform square grid $\mathcal{T}_0$ with the initial mesh parameter h_0 on which we take the FE space $V_p(\mathcal{T}_0)$, $p = 1$. Then, we increment the

polynomial degree p by 1 recursively and ask for the discrete minima on $V_{p,g_D}(\mathcal{T}_0)$. The minima are listed in Table 4.2.

The minimal surface is visualized in Figure 4.5. The p-version with different mesh parameters h_0 and the h-version with different initial values h_0 give the same values for the extrapolation $\bar{A}$ in the first 6 digits. For the computation of the convergence rate, we used $\bar{A} = 9.014014$. In case of the h-version the experimental convergence rates confirm $|A_k - \bar{A}| = \mathcal{O}(N_k^{-1})$ (see Table 4.1). Assuming (4.46) of Lemma 4.48 to hold, this gives $|u_h - u|_{H^1(\Omega)} = \mathcal{O}(h)$ because of $N = \mathcal{O}(h^{-2})$.

The p-version experiment confirms $|A_k - \bar{A}| = \mathcal{O}(N_k^{-2})$ (see Table 4.2). Thus, again assuming (4.46), yields $|u_h - u|_{H^1(\Omega)} = \mathcal{O}(p^{-2})$ because of $N = \mathcal{O}(p^2)$. The best h-version result yielded by solving a nonlinear system with 123201 unknowns can be calculated much more efficiently by the p-version by solving a nonlinear system with less than 2000 unknowns for all tested mesh width h_0.

Experiment 4.50 (*Minimal surface with inhomogeneous boundary data over an obstacle***).** We repeat the h-version and the p-version computations of Experiment 4.49, but demand that the minimal surface fulfills the condition $u(x_i) \geq \psi(x_i)$ for all Gauss-Lobatto points $x_i \in G_p$. Here, the obstacle ψ is defined by

$$\psi(x) := \begin{cases} \sqrt{\frac{1}{16} - x_1^2} - 0.3 & \text{for } |x_1| < \frac{1}{4}, \\ -\infty & \text{for } |x_1| \geq \frac{1}{4}\,. \end{cases} \tag{4.50}$$

This can be illustrated by moving a cylinder parallel to the x_2-axis from underneath against the minimal surface of Experiment 4.49 (see Figure 4.6). Using a wire frame for the cylinder allows us to visualize the coincidence set Ψ. The lower plot of Figure 4.6 shows that for those $x \in \Omega$ where the obstacle condition is violated, we have $|\psi(x) - u_p(x)| \leq 2 \cdot 10^{-3}$ for $h = \frac{1}{8}$ and $p = 4$.

The numerical results are listed in Table 4.3 for the h-version and in Table 4.4 for the p-version. The p-version with the tested mesh widths h_0 and the h-version with the tested mesh widths h_0 lead to the extrapolation $\bar{A} = 9.5545$. The h-version experiment confirms $|A_k - \bar{A}| = \mathcal{O}(N_k^{-1})$. Thus, assuming analogously to the interpretation of Experiment 4.49, we obtain $|u_h - u|_{H^1(\Omega)} = \mathcal{O}(h)$. The p-version experiment confirms a rate of $|A_k - \bar{A}| = \mathcal{O}(N_k^{-3/2})$. Under the assumption that (4.46) holds, this leads to $|u_h - u|_{H^1(\Omega)} = \mathcal{O}(p^{-3/2})$. The best h-version solution calculated from 123201 unknowns is achieved nearly by the p-version with less than 8000 unknowns.

Experiment 4.51 (*Minimal surface with homogeneous boundary data over an obstacle***).** We take Ω, the h-discretization, and the p-discretization as in Experiment 4.49. Now, we demand homogeneous Dirichlet conditions, i.e., $g_D \equiv 0$ and define an obstacle ψ by

$$\psi(x) := \begin{cases} \sqrt{\frac{1}{16} - (x - x_m)^2} + \frac{1}{4} & \text{for } |x - x_M| < \frac{1}{4}\,, \\ -\infty & \text{for } |x - x_M| \geq \frac{1}{4}\,, \end{cases} \tag{4.51}$$

where $x_M = \binom{0.3}{0.1}$. Thus, we ask for the minimal surface on the square $[-1, 1]^2$ over a ball with radius $\frac{1}{4}$ which touches the square at x_M (see Figure 4.7).

The surface plots of Figure 4.8 visualize for $h = \frac{1}{8}$ and $p = 4$ where the obstacle condition is violated, i.e., $\psi > u_p$. There, we have $|\psi(x) - u_p(x)| \leq 5 \cdot 10^{-3}$. The numerical results

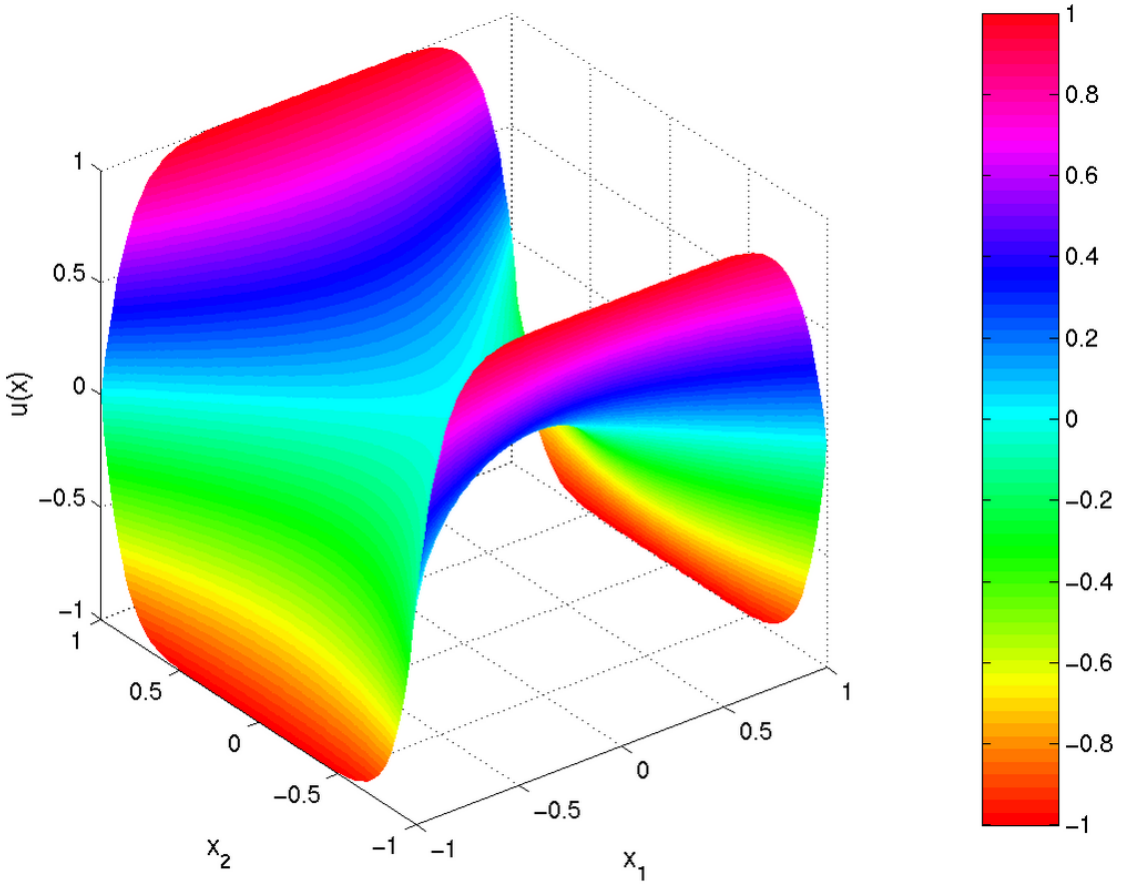

Figure 4.5: Surface plot of u from Experiment 4.49: The approximation is computed due to the mesh parameters $h = \frac{1}{8}$ and $p = 4$. This yields 3969 degrees of freedom.

are listed in Table 4.5 for the h-version and in Table 4.6 for the p-version. The p-version with the tested mesh widths h_0 and the h-version with the tested mesh widths h_0 yield the extrapolation $\bar{A} = 4.3118$. The h-version experiment confirms $|A_k - \bar{A}| = \mathcal{O}(N_k^{-3/4})$ which corresponds to $|u_h - u|_{H^1(\Omega)} = \mathcal{O}(h^{3/4})$ (cf. interpretation of Experiment 4.49). The p-version experiment confirms convergence and numerical stability of the method for different mesh widths h_0. For the width $h_0 = \frac{2}{5}$ the experimental convergence rates for $p \geq 6$ do not indicate further convergence. We explain this by the exactness of the solution for $p = 5$, $p = 6$. As the approximation can hardly be improved, a further convergence can not be shown. In addition, we note that the experimental convergence rates depend highly on the extrapolated value $\bar{A}$. Nevertheless, the p-version performs stable and achieves the most exact approximation of the h-version calculated from 123201 unknowns already with less than 5000 unknowns for the mesh widths $h_0 = \frac{2}{7}, \frac{2}{9}, \frac{2}{11}$.

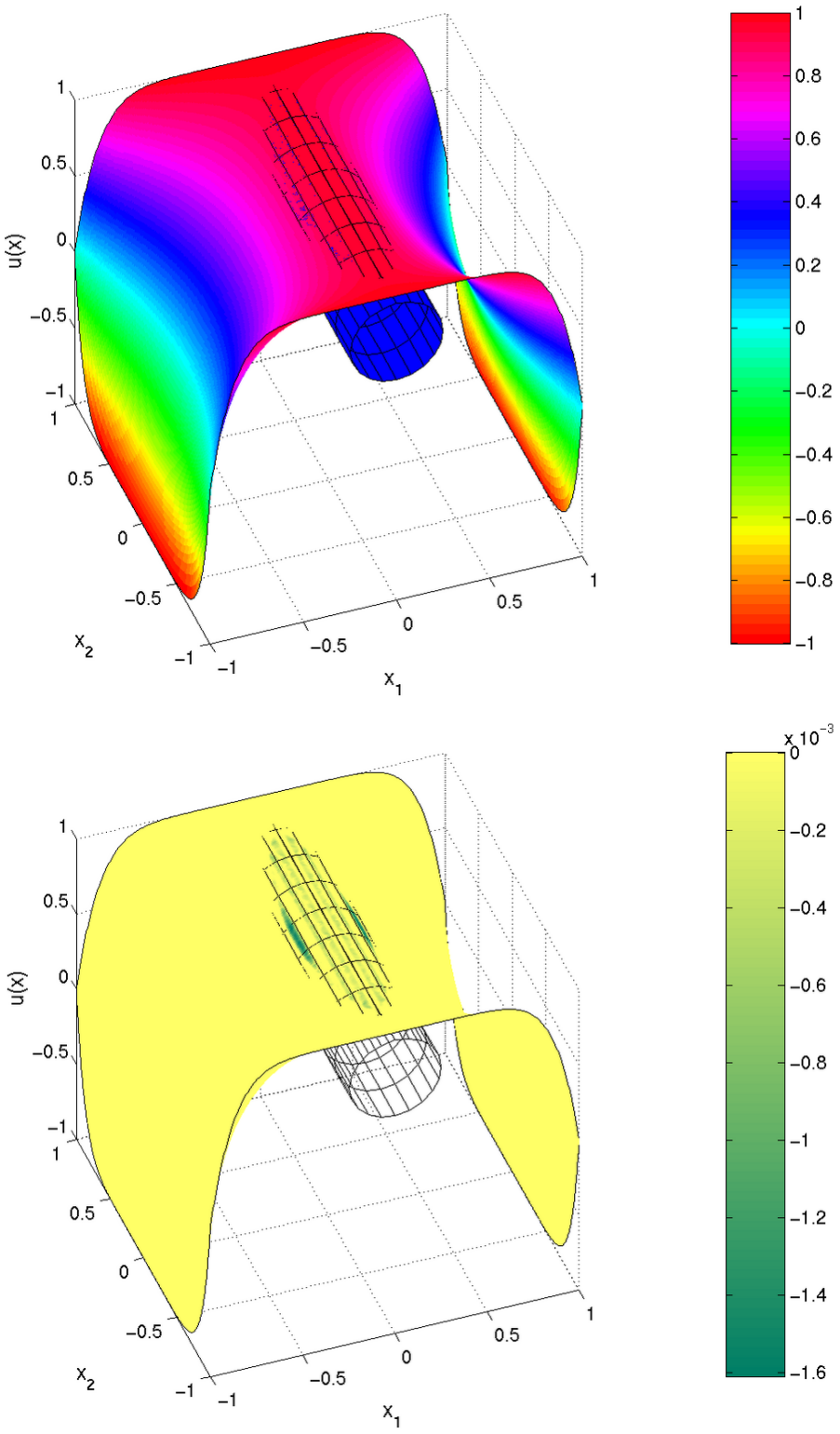

Figure 4.6: Surface plots of u from Experiment 4.50: The approximation is computed according to the mesh parameters $h = \frac{1}{8}$ and $p = 4$. 435 of the 3969 degrees of freedom are active. The wire frame model of the cylinder in the upper plot shows the coincidence set Ψ. The lower plot visualizes the area where the obstacle condition is violated, i.e., $\psi(x) > u_p(x)$.

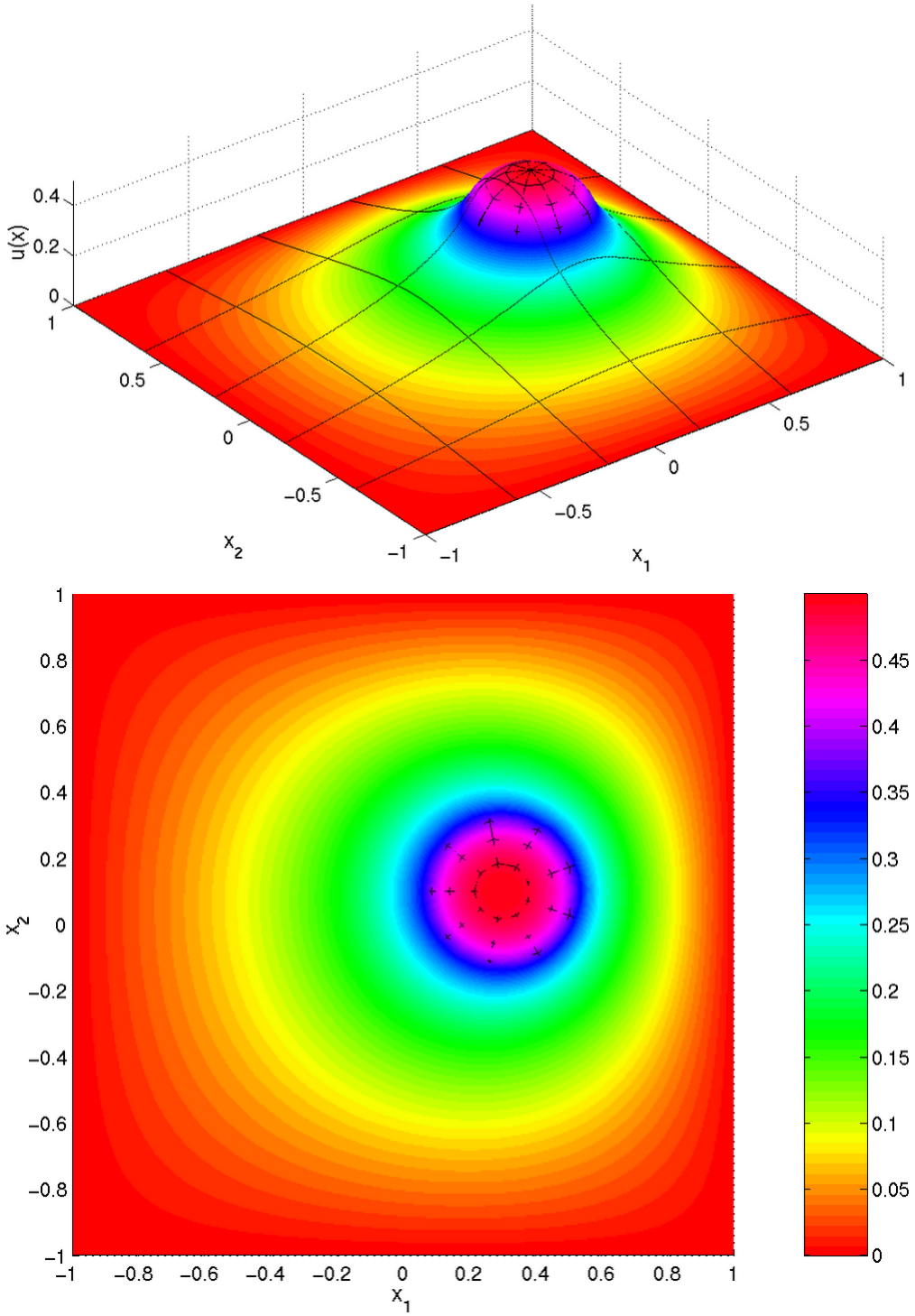

Figure 4.7: Side view and top view of the solution u from Experiment 4.51: The approximation is computed according to the mesh parameters $h = \frac{1}{8}$ and $p = 4$. 168 of the 3969 degrees of freedom are active. The wire frame model of the ball shows the coincidence set Ψ.

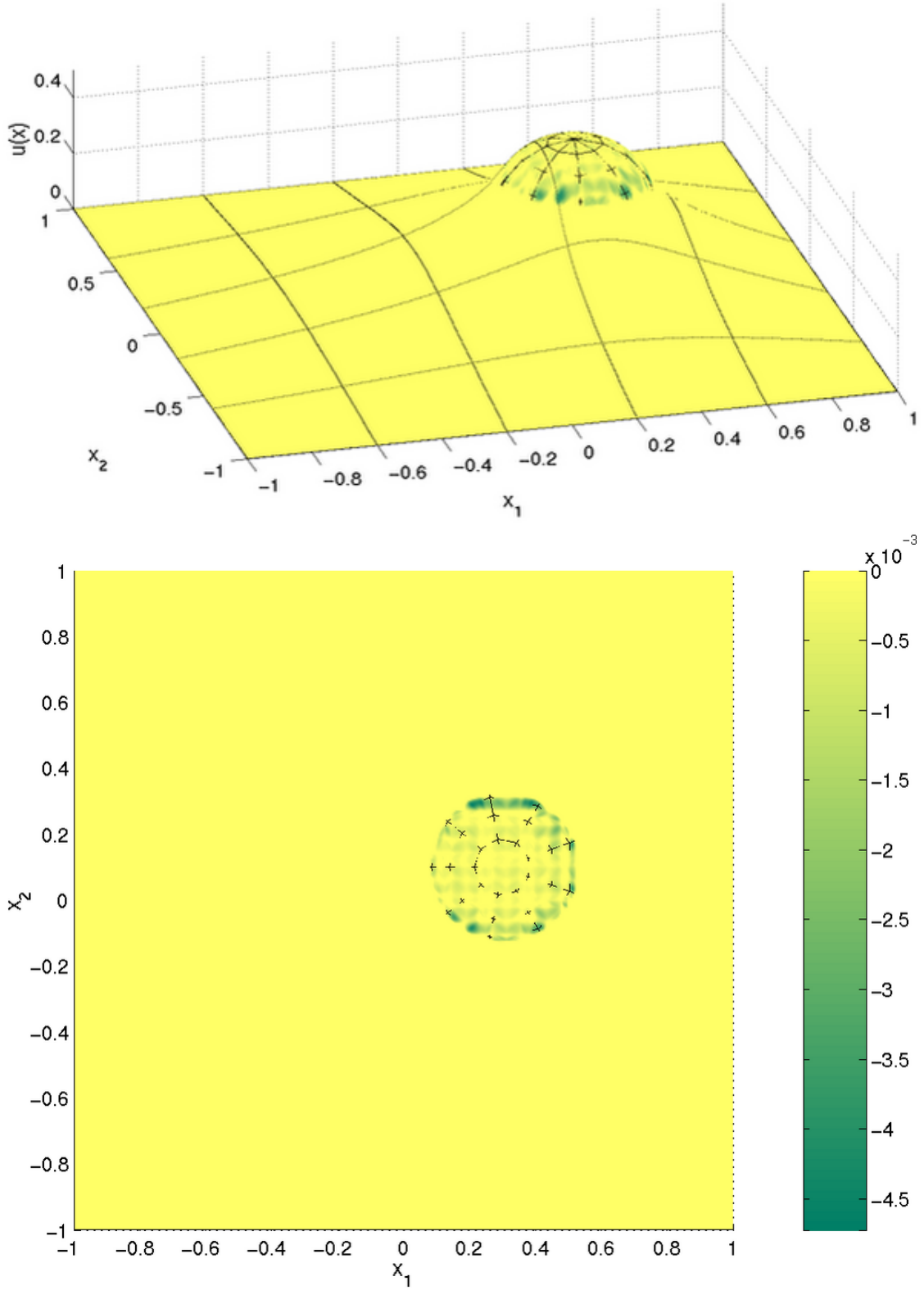

Figure 4.8: Side and top view of the consistency error from Experiment 4.51: The plots visualize the area where the obstacle condition is violated, i.e., $\psi(x) > u_p(x)$.

h_0	k	N_k	$A(u_k)$	$\lvert A(u_k) - \bar{A}\rvert$	α
$\frac{2}{5}$	1	16	8.53396	4.80e-01	—
	2	81	8.90022	1.14e-01	-0.89
	3	361	8.98726	2.68e-02	-0.97
	4	1521	9.00761	6.40e-03	-0.99
	5	6241	9.01245	1.56e-03	-1.00
	6	25281	9.01363	3.87e-04	-1.00
$\frac{2}{7}$	1	36	8.77347	2.41e-01	—
	2	169	8.95776	5.63e-02	-0.94
	3	729	9.00069	1.33e-02	-0.99
	4	3025	9.01079	3.22e-03	-1.00
	5	12321	9.01322	7.93e-04	-1.00
	6	49729	9.01382	1.97e-04	-1.00
$\frac{2}{9}$	1	64	8.87204	1.42e-01	—
	2	289	8.98070	3.33e-02	-0.96
	3	1225	9.00607	7.94e-03	-0.99
	4	5041	9.01208	1.93e-03	-1.00
	5	20449	9.01354	4.78e-04	-1.00
	6	82369	9.01389	1.19e-04	-1.00
$\frac{2}{11}$	1	100	8.92084	9.32e-02	—
	2	441	8.99206	2.20e-02	-0.97
	3	1849	9.00875	5.27e-03	-1.00
	4	7569	9.01273	1.29e-03	-1.00
	5	30625	9.01369	3.20e-04	-1.00
	6	123201	9.01393	7.98e-05	-1.00

Table 4.1: h-version for Experiment 4.49

h_0	p	N_p	$A(u_p)$	$\lvert A(u_p) - \bar{A}\rvert$	α
$\frac{2}{5}$	1	16	8.53853	4.75e-01	—
	2	81	9.02307	9.06e-03	—
	3	196	9.01802	4.01e-03	-1.9
	4	361	9.01492	9.08e-04	-1.5
	5	576	9.01428	2.62e-04	-2.5
	6	841	9.01409	7.98e-05	-2.9
	7	1156	9.01404	2.48e-05	-3.4
	8	1521	9.01402	7.86e-06	-3.9
	9	1936	9.01402	2.54e-06	-4.4
	10	2401	9.01401	8.44e-07	-4.9
	11	2916	9.01401	2.79e-07	-5.4
	12	3481	9.01401	8.00e-08	-6.3
$\frac{2}{7}$	1	36	8.77347	2.41e-01	—
	2	169	9.02096	6.95e-03	—
	3	400	9.01560	1.58e-03	-2.1
	4	729	9.01435	3.37e-04	-2.1
	5	1156	9.01410	8.25e-05	-2.8
	6	1681	9.01403	2.09e-05	-3.3
	7	2304	9.01402	5.43e-06	-3.9
	8	3025	9.01402	1.45e-06	-4.5
	9	3844	9.01401	3.93e-07	-5.1
$\frac{2}{9}$	1	64	8.87204	1.42e-01	—
	2	289	9.01837	4.35e-03	—
	3	676	9.01479	7.79e-04	-2.2
	4	1225	9.01416	1.49e-04	-2.3
	5	1936	9.01405	3.13e-05	-3.1
	6	2809	9.01402	6.81e-06	-3.7
	7	3844	9.01402	1.52e-06	-4.4
	8	5041	9.01401	3.44e-07	-5.1
$\frac{2}{11}$	1	100	8.92084	9.32e-02	—
	2	441	9.01683	2.82e-03	—
	3	1024	9.01445	4.31e-04	-2.3
	4	1849	9.01409	7.34e-05	-2.5
	5	2916	9.01403	1.35e-05	-3.3
	6	4225	9.01402	2.56e-06	-4.1
	7	5776	9.01401	4.96e-07	-4.8

Table 4.2: p-version for Experiment 4.49

h_0	k	N_k	$A(u_k)$	$\lvert A(u_k) - \bar{A}\rvert$	α_k
$\frac{2}{5}$	1	16	9.04436	5.10e-01	—
	2	81	9.44237	1.12e-01	-0.93
	3	361	9.53080	2.37e-02	-1.04
	4	1521	9.54963	4.87e-03	-1.10
	5	6241	9.55350	1.00e-03	-1.12
	6	25281	9.55431	1.91e-04	-1.19
$\frac{2}{7}$	1	36	9.31889	2.36e-01	—
	2	169	9.50355	5.09e-02	-0.99
	3	729	9.54362	1.09e-02	-1.06
	4	3025	9.55222	2.28e-03	-1.10
	5	12321	9.55405	4.55e-04	-1.15
	6	49729	9.55442	7.77e-05	-1.27
$\frac{2}{9}$	1	64	9.42125	1.33e-01	—
	2	289	9.52499	2.95e-02	-1.00
	3	1225	9.54828	6.22e-03	-1.08
	4	5041	9.55323	1.27e-03	-1.12
	5	20449	9.55425	2.47e-04	-1.17
	6	82369	9.55446	3.51e-05	-1.40
$\frac{2}{11}$	1	100	9.46819	8.63e-02	—
	2	441	9.53542	1.91e-02	-1.02
	3	1849	9.55053	3.97e-03	-1.10
	4	7569	9.55370	8.01e-04	-1.14
	5	30625	9.55435	1.49e-04	-1.20
	6	123201	9.55449	1.49e-05	-1.66

Table 4.3: h-version for Experiment 4.50

h_0	p	N_p	$A(u_p)$	$\lvert A(u_p) - \bar{A}\rvert$	α_p
$\frac{2}{5}$	1	16	9.04436	5.10e-01	—
	2	81	9.58605	3.16e-02	—
	3	196	9.56674	1.22e-02	-1.49
	4	361	9.55889	4.39e-03	-1.32
	5	576	9.55635	1.85e-03	-1.75
	6	841	9.55540	9.01e-04	-1.87
	7	1156	9.55497	4.70e-04	-1.97
	8	1521	9.55476	2.58e-04	-2.11
	9	1936	9.55466	1.64e-04	-2.04
	10	2401	9.55460	1.01e-04	-2.06
	11	2916	9.55457	6.89e-05	-2.12
	12	3481	9.55455	4.66e-05	-2.08
$\frac{2}{7}$	1	36	9.31889	2.36e-01	—
	2	169	9.57405	1.95e-02	—
	3	400	9.56025	5.75e-03	-1.54
	4	729	9.55655	2.05e-03	-1.54
	5	1156	9.55539	8.89e-04	-1.76
	6	1681	9.55494	4.38e-04	-1.85
	7	2304	9.55473	2.30e-04	-1.96
	8	3025	9.55463	1.27e-04	-2.10
	9	3844	9.55457	7.28e-05	-2.25
$\frac{2}{9}$	1	64	9.42125	1.33e-01	—
	2	289	9.56568	1.12e-02	—
	3	676	9.55767	3.17e-03	-1.59
	4	1225	9.55580	1.30e-03	-1.49
	5	1936	9.55498	4.77e-04	-1.80
	6	2809	9.55471	2.12e-04	-2.19
	7	3844	9.55463	1.33e-04	-1.86
	8	5041	9.55456	6.08e-05	-2.13
	9	6400	9.55454	3.57e-05	-2.58
$\frac{2}{11}$	1	100	9.42125	1.33e-01	—
	2	441	9.56568	1.12e-02	—
	3	1024	9.55767	3.17e-03	-1.61
	4	1849	9.55580	1.30e-03	-1.50
	5	2916	9.55498	4.77e-04	-1.81
	6	4225	9.55471	2.12e-04	-2.20
	7	5776	9.55458	7.89e-05	-2.63
	8	7569	9.55456	6.08e-05	-2.14
	9	9604	9.55454	3.57e-05	-1.56

Table 4.4: p-version for Experiment 4.50

h_0	k	N_k	$A(u_k)$	$\lvert A(u_k) - \bar{A}\rvert$	α_k
$\frac{2}{5}$	1	16	4.18788	1.24e-01	—
	2	81	4.28220	2.96e-02	-0.88
	3	361	4.30337	8.40e-03	-0.84
	4	1521	4.30949	2.28e-03	-0.91
	5	6241	4.31111	6.61e-04	-0.88
	6	25281	4.31160	1.71e-04	-0.97
$\frac{2}{7}$	1	36	4.24838	6.34e-02	—
	2	169	4.29280	1.90e-02	-0.78
	3	729	4.30692	4.85e-03	-0.93
	4	3025	4.31046	1.31e-03	-0.92
	5	12321	4.31143	3.39e-04	-0.96
	6	49729	4.31168	9.09e-05	-0.94
$\frac{2}{9}$	1	64	4.26195	4.98e-02	—
	2	289	4.29912	1.27e-02	-0.91
	3	1225	4.30876	3.01e-03	-0.99
	4	5041	4.31096	8.08e-04	-0.93
	5	20449	4.31156	2.08e-04	-0.97
	6	82369	4.31171	5.81e-05	-0.92
$\frac{2}{11}$	1	100	4.28571	2.61e-02	—
	2	441	4.30348	8.29e-03	-0.77
	3	1849	4.30969	2.08e-03	-0.97
	4	7569	4.31124	5.31e-04	-0.97
	5	30625	4.31163	1.44e-04	-0.93
	6	123201	4.31173	4.13e-05	-0.90

Table 4.5: h-version for Experiment 4.51

h_0	p	N_p	$A(u_p)$	$\lvert A(u_p) - \bar{A}\rvert$	α_p
$\frac{2}{5}$	1	16	4.18788	1.24e-01	—
	2	81	4.30886	2.91e-03	—
	3	196	4.30934	2.43e-03	-1.57
	4	361	4.31169	8.08e-05	-2.40
	5	576	4.31183	6.11e-05	-3.41
	6	841	4.31184	6.61e-05	-0.24
	7	1156	4.31196	1.93e-04	1.65
	8	1521	4.31167	9.75e-05	0.66
	9	1936	4.31202	2.54e-04	0.53
	10	2401	4.31191	1.38e-04	0.75
	11	2916	4.31170	6.95e-05	-3.17
	12	3481	4.31190	1.29e-04	-0.17
$\frac{2}{7}$	1	36	4.24838	6.34e-02	—
	2	169	4.30706	4.71e-03	—
	3	400	4.31097	7.99e-04	-1.82
	4	729	4.31243	6.56e-04	-1.35
	5	1156	4.31151	2.59e-04	-1.06
	6	1681	4.31202	2.55e-04	-1.13
	7	2304	4.31174	2.90e-05	-3.17
	8	3025	4.31187	9.76e-05	-1.63
	9	3844	4.31187	1.05e-04	2.51
	10	4761	4.31175	2.49e-05	-3.01
$\frac{2}{9}$	1	64	4.26195	4.98e-02	—
	2	289	4.31294	1.17e-03	—
	3	676	4.31202	2.47e-04	-2.25
	4	1225	4.31166	1.08e-04	-1.65
	5	1936	4.31169	8.34e-05	-1.03
	6	2809	4.31179	2.08e-05	-1.99
	7	3844	4.31180	2.59e-05	-1.71
	8	5041	4.31178	5.39e-06	-2.31
	9	6400	4.31177	3.78e-06	-3.78
$\frac{2}{11}$	1	100	4.28571	2.61e-02	—
	2	441	4.30938	2.39e-03	—
	3	1024	4.31170	7.18e-05	-2.53
	4	1849	4.31182	5.28e-05	-2.66
	5	2916	4.31179	2.34e-05	-1.07
	6	4225	4.31180	2.96e-05	-0.70
	7	5776	4.31177	4.96e-07	-5.64
	8	7569	4.31178	1.06e-05	-1.76
	9	9604	4.31175	1.57e-05	6.80

Table 4.6: p-version for Experiment 4.51

Chapter 5

Prolongation and space decomposition methods for nonlinear PDE and PDI

It is the purpose of this chapter to discuss some techniques which try to speed up the solving of nonlinear minimization problems with and without constraints by reducing it to nonlinear problems of lower dimension. In Section 4.4 and Section 4.5, we considered efficient solvers for the linear subproblems of the nonlinear minimization algorithms and showed that *hp*-preconditioners known for linear problems can be adapted to the treatment of the linear subproblems originated by minimization problems with box constraints (see Subsections 4.5.2 and 4.5.3). The nonlinear problem treated by the Newton like outer iterations of Algorithm 4.2 and Algorithm 4.3 remained as a global problem. In contrast to efficient linear solvers, we propose two nonlinear methods in this chapter, one based on the prolongation of a coarse space solution, the other based on a space decomposition.

Section 5.1 introduces two algorithms which prolongate a p-FE solution u_p to a higher dimensional p-FE space with polynomial degree $q > p$ by one Newton iteration. Algorithm 5.1 is suited to problems without inequality constraints. Algorithm 5.2 extends Algorithm 5.1 to the treatment of lower bounds on the vector components. Both algorithms should be understood as a first attempt of a numerical analysis for two or multi space discretization techniques tailored to solve the p-version discretization of a PDI efficiently. We give a posteriori estimates for the results of Algorithm 5.1 and Algorithm 5.2 in Proposition 5.1 and Proposition 5.3, respectively.

Section 5.2 describes a nonlinear solver which decomposes the minimization space into a direct sum of subspaces. Then, the original minimization problem is solved in parallel or sequentially over each of the low dimensional subspaces. Algorithm 5.3 and Algorithm 5.4, both generalize the known additive and multiplicative methods from linear problems to the class of unconstrained minimization problems given by the functional A defined by (1.8). This techniques is motivated by the publications *Rate of convergence of some space decomposition methods for linear and nonlinear problems* and *Applications of a space decomposition method to linear and nonlinear elliptic problems* [TE98a, TE98b]. The authors use standard h-version multi-grid and domain decompositions from efficient

solvers of linear problems to develop fast solvers for nonlinear problems and present a uniform linear rate of convergence for the additive and multiplicative nonlinear Schwarz method.

We transfer this approach to a space decomposition proposed by Babuška et al. in [BCMP91]. There, a p-version FE space on a quasi-uniform triangle or quadrilateral mesh is decomposed into the sum of non-overlapping subspaces given by nodal, edge, and interior associated global basis functions. A uniform rate of convergence is proven for the additive Algorithm 5.3 and for the multiplicative Algorithm 5.4. Unfortunately, the contraction factors of both methods converge to 1, when p becomes big (see (5.12), (5.15)).

In comparison to the global Newton approach to an unconstrained nonlinear problem the algorithms proposed in Section 5.1 and Section 5.2 show a disappointing performance in the numerical experiments documented in Section 5.3. We put this down to the excellent convergence properties of the global Newton's method (quadratic convergence, when the initial is near by the solution) and to the fact that the number of Newton iterations does not depend on the dimension of the problem as much as the costs of iterative linear solvers do, see Remark 5.10.

Nevertheless, the algorithms of this chapter might represent good starting points for further developments in numerical analysis and numerical experiments. The prolongation idea of Section 5.1 may help to find good initials for constrained minimization problems originated from high polynomial degrees. The space decomposition method of Section 5.2 may perform better, when overlapping subspaces are used.

5.1 Prolongation of a discrete solution into a higher dimensional space

The nonlinear minimizers presented in Section 4.2 and Section 4.3 for the unconstrained and the constrained problem, respectively, both demand an initial vector (see Algorithm 4.2, Algorithm 4.3). It is the idea of the following algorithms to take the solution $u_p \in V_p$ from a low dimensional problem as the initial for the search of the solution $u_q \in V_q$, $q > p$, of a high dimensional problem. We call this process *prolongation of* u_p *into* V_q.

For the numerical analysis we consider the quality of this prolongation after an application of a Newton iteration. Firstly, we consider the case of an unconstrained problem corresponding to a variational equality. Secondly, we generalize the algorithm and its analysis to a variational inequality.

Now, we describe Algorithm 5.1. *Step 1* takes care of the Dirichlet boundary conditions. $u_{p|\Gamma}$ can be used as an initial for the solver of the nonlinear variational problem of *Step 2*. *Step 3* realizes one Newton iteration. The terms $-u_{p|\Gamma} + u_{q|\Gamma}$ of the last assignment of the algorithm ensure that $\tilde{u}_q$ fulfills the Dirichlet boundary condition in the discrete sense.

The following proposition gives an a posteriori estimate for the difference between the fine space solution u_q and the result $\tilde{u}_q$ of Algorithm 5.1.

Algorithm 5.1 Prolongation of the PDE solution u_p into V_{q,g_D}, $q > p$

Let V_p and V_q be two p-version FE spaces with $1 \leq p < q$ as defined in (2.2). Let $V_{p,g_D} \subset V_p$, $V_{q,g_D} \subset V_q$ be the FE subsets given in Definition 2.4 which discretize the Dirichlet boundary condition. Further, let $V_{q,0_D} \subset V_q$ be the FE subset which discretizes the homogeneous Dirichlet boundary condition, i.e., we set $g_D \equiv 0$ in Definition 2.4.
Let G_p be the set of Gauss-Lobatto points on the mesh $\mathcal{T}$ (see Definition 2.2), let $\Gamma_{D,p} \subset G_p$ be the subset of Gauss-Lobatto points on the Dirichlet boundary, and let $b_{p,x}$ be the global p-Lagrangian basis functions associated to each $x \in G_p$ (cf. (4.2)). Let $\Gamma_{D,q}$ and $b_{q,x}$ be defined analogously by taking q instead of p.

1. Set
$$u_{p,\Gamma} = \sum_{x \in \Gamma_{D,p}} g_D(x)\, b_{p,x} \qquad \text{and} \qquad u_{q,\Gamma} = \sum_{x \in \Gamma_{D,q}} g_D(x)\, b_{q,x}\,.$$
2. Find $u_p \in V_{p,g_D}$ such that $DA(u_p; v - u_p) = 0$ for all $v \in V_{p,g_D}$.
3. Find $d_q \in V_{q,0_D}$ such that $D^2A(u_p; d_q, v) = -DA(u_p; v)$ for all $v \in V_{q,0_D}$.
Set $\tilde{u}_q = u_p + d_q - u_{p,\Gamma} + u_{q,\Gamma}$.

Proposition 5.1. Let the functional A be given by (1.8) and let ρ_i, $i = 1, \ldots, 5$, be the constants of the assumptions on $\rho(t)$ in Lemma 1.21.
Let $u_q \in V_{q,g_D}$ be the solution of the variational equation $DA(u_q; v - u_q) = 0$ for all $v \in V_{q,g_D}$ and let $\tilde{u}_q$ be computed by Algorithm 5.1. Then, we have the a posteriori error estimate

$$|u_q - \tilde{u}_q|_{H^1(\Omega)} \leq \kappa_l^{-1}\big(\kappa_u |u_{q|\Gamma} - u_{p|\Gamma}|_{H^1(\Omega)} + \sqrt{2}(5\rho_4 + \rho_5)\|\nabla(u_p - \tilde{u}_q)\|^2_{L^4(\Omega)}\big)$$

in case of $\sigma = 0$ (see (1.8)). In case of $\sigma > 0$ the semi-norm $|\cdot|_{H^1(\Omega)}$ has to be replaced by the norm $\|\cdot\|_{H^1(\Omega)}$. Here, κ_l and κ_u are the ellipticity constants defined in Lemma 1.21.

To prove the proposition, we need the following lemma.

Lemma 5.2. For any $w, u, v \in H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$ a bounded Lipschitz domain, there holds

$$DA(w; v) = DA(u; v) + D^2A(u; w - u, v) + R(u; w - u, v) \tag{5.1}$$

with the estimate for the remainder R

$$|R(u; w - u, v)| \leq \sqrt{2}(5\rho_4 + \rho_5)\, \|\nabla(w - u)\|^2_{L^4(\Omega)}\, |v|_{H^1(\Omega)}\,.$$

Proof. The lemma is an application of Taylor's formula and standard calculus. The proof is given in Appendix D. □

Proof of Proposition 5.1. By Lemma 5.2 we have the identity

$$DA(\tilde{u}_q; v) = DA(u_p; v) + D^2A(u_p; \tilde{u}_q - u_p, v) + R(u_p; \tilde{u}_q - u_p, v) \text{ for all } v \in V_{q,0_D}$$

with the estimate

$$|R(u_p; \tilde{u}_q - u_p, v)| \leq \sqrt{2}(5\rho_4 + \rho_5)\, \|\nabla(\tilde{u}_q - u_p)\|^2_{L^4(\Omega)}\, |v|_{H^1(\Omega)}$$

for the remainder R. Due to *Step 3* of Algorithm 5.1, we can substitute

$$DA(u_p; v) + D^2A(u_p; \tilde{u}_q - u_p, v) = D^2A(u_p; u_{q,\Gamma} - u_{p,\Gamma}, v).$$

Inspecting the proof of Lemma 1.21 (see (1.14)), we get

$$D^2A(u_p; u_{q,\Gamma} - u_{p,\Gamma}, v) \le \kappa_u |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} |v|_{H^1(\Omega)} ,$$

if $\sigma = 0$. For $\sigma > 0$ we obtain the same estimate with the norm $\|\cdot\|_{H^1(\Omega)}$ instead of the semi-norm $|\cdot|_{H^1(\Omega)}$.

Using the uniform monotonicity (1.12) of DA stated in Lemma 1.21 and noting that $DA(u_q; v) = 0$ for all $v \in V_{q,0}$, i.e., in particular $DA(u_q; \tilde{u}_q - u_q) = 0$, we combine

$$\begin{aligned}
\kappa_l |\tilde{u}_q - u_q|^2_{H^1(\Omega)} &\le |DA(\tilde{u}_q; \tilde{u}_q - u_q) - DA(u_q; \tilde{u}_q - u_q)| \\
&\le D^2A(u_p; u_{q,\Gamma} - u_{p,\Gamma}, \tilde{u}_q - u_q) \qquad + R(u_p; \tilde{u}_q - u_p, \tilde{u}_q - u_p) \\
&\le |\tilde{u}_q - u_q|_{H^1(\Omega)} \big(\kappa_u |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} + \sqrt{2}(5\rho_4 + \rho_5) \|\nabla(\tilde{u}_q - u_p)\|^2_{L^4(\Omega)}\big) ,
\end{aligned}$$

if $\sigma = 0$. For $\sigma > 0$ the proposition follows by taking the norm $\|\cdot\|_{H^1(\Omega)}$ instead of the semi-norm $|\cdot|_{H^1(\Omega)}$ (cf. Lemma 1.21). □

Now, we extend Algorithm 5.1 to the treatment of variational inequalities as follows. In *Step 2* of Algorithm 5.2, we look for the coarse space solution of the variational inequality in K_{p,g_D}. *Step 3* changes *Step 3* of Algorithm 5.1. Here, (5.2) ensures $\tilde{u}_q \in K_{q,g_D}$.

Algorithm 5.2 Prolongation of the PDI solution u_p into K_{q,g_D}, $q > p$

Let V_p and V_q be two p-version FE spaces with $1 \le p < q$ as defined in (2.2) and $K_{p,g_D} \subset V_p$, $K_{q,g_D} \subset V_q$ be the FE subsets given in Definition 2.4 which discretize the Dirichlet boundary condition and the obstacle condition. Further, let $V_{q,0_D} \subset V_q$ the FE subset which discretizes the homogeneous Dirichlet boundary condition, i.e., we set $g_D \equiv 0$ in Definition 2.4.
Let G_p be the set of Gauss-Lobatto points on the mesh $\mathcal{T}$ (see Definition 2.2) and $\Gamma_{D,p} \subset G_p$ the subset of points on the Dirichlet boundary. We denote the global p-Lagrangian basis functions associated to each $x \in G_p$ (cf. (4.2)) by $b_{p,x}$.
Additionally, G_q, $\Gamma_{D,q}$, and $b_{q,x}$ are defined analogously by taking q instead of p.

1. Set
$$u_{p,\Gamma} = \sum_{x \in \Gamma_{D,p}} g_D(x)\, b_{p,x} \qquad \text{and} \qquad u_{q,\Gamma} = \sum_{x \in \Gamma_{D,q}} g_D(x)\, b_{q,x} .$$
2. Find $u_p \in K_{p,g_D}$ such that $DA(u_p; v - u_p) \ge 0$ for all $v \in K_{p,g_D}$.
3. Find $d_q \in V_{q,0_D}$ such that $D^2A(u_p; d_q, v) = -DA(u_p; v)$ for all $v \in V_{q,0_D}$.
   Set $\hat{u}_q = u_p + d_q - u_{p,\Gamma} + u_{q,\Gamma}$.
   Set
$$\tilde{u}_q = \sum_{x \in G_p} \max\{\hat{u}_q(x), \psi(x)\} b_{q,x} . \tag{5.2}$$

We can estimate the error $|u_q - \tilde{u}_q|_{H^1(\Omega)}$ between the fine space solution u_q and the result $\tilde{u}_q$ of Algorithm 5.2 a posteriori from $\|u_p - \tilde{u}_q\|_{H^1(\Omega)}$. To increase the readability of the

following proposition and its proof, we demand that the problem has no Neumann boundary conditions. The proposition can be extended to Neumann conditions by introducing the respective boundary terms used in the proof of Theorem 2.11 into the proposition's proof.

Proposition 5.3. Let the functional A be given by (1.8) without the Neumann boundary condition, i.e., $\Gamma_N = \emptyset$. Let ρ_i, $i = 1, \dots, 5$, be the constants of the assumptions on $\rho(t)$ in Lemma 1.21.
Further, let $u_q \in K_{q,g_D}$ be the solution of the variational inequality $DA(u_q; v - u_q) \geq 0$ for all $v \in K_{q,g_D}$ and let $\tilde{u}_q$ be computed by Algorithm 5.2. Then, there exist constants C_i, $i = 1, 2, 3, 4$, independent of p and q, such that the a posteriori error estimate

$$\begin{aligned}|\tilde{u}_q - u_q|^2_{H^1(\Omega)} \leq C_2\|\tilde{u}_q - u_p\|_{L^2(\Psi)} \quad &+ C_3|u_q - u|^2_{H^1(\Omega)} \quad + C_4 p^{-1} \\ &+ C_1\big(|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)} + |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} + \sqrt{2}(5\rho_4 + \rho_5)\|\nabla(\tilde{u}_q - u_p)\|^2_{L^4(\Omega)}\big)^2\end{aligned}$$

holds in case of $\sigma = 0$. In case of $\sigma > 0$ the estimate holds, when the semi-norm $|\cdot|_{H^1(\Omega)}$ is replaced by the norm $\|\cdot\|_{H^1(\Omega)}$.

Proof. In the following κ_l and κ_u are the ellipticity constants given in Lemma 1.21. We start from the uniform monotonicity estimate

$$\begin{aligned}\kappa_l|\tilde{u}_q - u_q|^2_{H^1(\Omega)} &\leq |DA(\tilde{u}_q; \tilde{u}_q - u_q) - DA(u_q; \tilde{u}_q - u_q)| \\ &\leq |DA(\tilde{u}_q; \tilde{u}_q - u_q)| \; + \; |DA(u_q; \tilde{u}_q - u_q)|\,. \end{aligned} \tag{5.3}$$

Firstly, we consider $|DA(\tilde{u}_q; \tilde{u}_q - u_q)|$. As in the proof of Proposition 5.1, we have the identity

$$DA(\tilde{u}_q; v) = DA(u_p; v) + D^2A(u_p; \tilde{u}_q - u_p, v) + R(u_p; \tilde{u}_q - u_p, v)$$

for all $v \in V_{q,0}$ with

$$|R(u_p, \tilde{u}_q - u_p, v)| \leq \sqrt{2}(5\rho_4 + \rho_5)\,\|\nabla(\tilde{u}_q - u_p\|^2_{L^4(\Omega)}\,|v|_{H^1(\Omega)}$$

by Lemma 5.2. Due to *Step 3* of Algorithm 5.2, we can replace

$$DA(u_p; v) + D^2A(u_p; \tilde{u}_q - u_p, v) = D^2A(u_p; \tilde{u}_q - \hat{u}_q, v) + D^2A(u_p; u_{q,\Gamma} - u_{p,\Gamma}, v)\,.$$

Inspecting the proof of Lemma 1.21 (see (1.14)), yields

$$\begin{aligned}|D^2A(u_p; \tilde{u}_q - \hat{u}_q, v) + D^2A(u_p; u_{q,\Gamma} - u_{p,\Gamma}, v)| \\ \leq \kappa_u\left(|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)} + |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)}\right)|v|_{H^1(\Omega)}\,.\end{aligned}$$

Combining these estimates we get

$$\begin{aligned}DA(\tilde{u}_q; \tilde{u}_q - u_q) \;\leq\; &\kappa_u\big(|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)} + |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} \\ &\qquad + \sqrt{2}(5\rho_4 + \rho_5)\|\nabla(\tilde{u}_q - u_p)\|^2_{L^4(\Omega)}\big)\,|\tilde{u}_q - u_q|_{H^1(\Omega)} \\ \leq\; &\tfrac{\kappa_u^2}{\kappa_l}\big(|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)} + |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} + \sqrt{2}(5\rho_4 + \rho_5)\|\nabla(\tilde{u}_q - u_p)\|^2_{L^4(\Omega)}\big)^2 \\ &\qquad + \tfrac{\kappa_l}{4}\,|\tilde{u}_q - u_q|^2_{H^1(\Omega)}\,.\end{aligned}$$

Here, the last inequality follows from $ab \leq \frac{\mu}{2}a^2 + \frac{1}{2\mu}b^2$, $a, b \in \mathbb{R}$, with $\mu = \frac{2}{\kappa_l}$.

Secondly, we consider $|DA(u_q, \tilde{u}_q - u_q)|$. By Taylor's formula there holds,

$$DA(u_q; \tilde{u}_q - u_q) = DA(u; \tilde{u}_q - u_q) + D^2A(u + \theta(u_q - u); u_q - u, \tilde{u}_q - u_q) \tag{5.4}$$

for a $\theta \in [0, 1]$. Analogously to (2.22), we have

$$D^2A(u + \theta(u_q - u); u_q - u, \tilde{u}_q - u_q) \leq \tfrac{\kappa_u^2}{2\kappa_l}|u_q - u|^2_{H^1(\Omega)} + \tfrac{\kappa_l}{2}|\tilde{u}_q - u_q|^2_{H^1(\Omega)}\,. \tag{5.5}$$

Using the notations $\mathcal{P}(u)$ from (1.18) and Ψ from (1.22) for the coincidence set, p. 28, partial integration of $DA(u; \tilde{u}_q - u_q)$ yields

$$\begin{aligned} DA(u; \tilde{u}_q - u_q) &\leq \|\mathcal{P}(u)\|_{L^2(\Psi)}\, \|\tilde{u}_q - u_q\|_{L^2(\Psi)} \\ &\leq \|\mathcal{P}(u)\|_{L^2(\Psi)} \left(\|\tilde{u}_q - u_p\|_{L^2(\Psi)} + \|u_p - \psi\|_{L^2(\Psi)} + \|\psi - u_q\|_{L^2(\Psi)}\right). \end{aligned} \tag{5.6}$$

Now, let

$$\bar{\psi}_p := \begin{cases} u_p \text{ on } \Omega \setminus \Psi, \\ \psi \text{ on } \Psi. \end{cases}$$

Due to [KS80, Theorem II.A.1], we know that $\|\bar{\psi}_p\|_{H^1(\Omega)} \leq \|\psi\|_{H^1(\Omega)} + \|u_p\|_{H^1(\Omega)}$. Further, $\bar{\psi}_p$ is interpolated by u_p in the Gauss-Lobatto points, i.e., $i_p\, \bar{\psi}_p = u_p$. Thus, there exists a constant c independent of p such that

$$\|u_p - \psi\|_{L^2(\Psi)} \leq \|u_p - \bar{\psi}_p\|_{L^2(\Omega)} \leq cp^{-1}\|\bar{\psi}_p\|_{H^1(\Omega)}$$

due to the the interpolation result Theorem 2.3. Analogously, we define the extension $\bar{\psi}_q$ of ψ and estimate

$$\|u_q - \psi\|_{L^2(\Psi)} \leq cq^{-1}\|\bar{\psi}_q\|_{H^1(\Omega)}\,.$$

It follows by the convergence of u_p and u_q towards u with respect to the $\|\cdot\|_{H^1(\Omega)}$-norm that $\|\bar{\psi}_p\|_{H^1(\Omega)}$ and $\|\bar{\psi}_q\|_{H^1(\Omega)}$ are bounded independently of p and q by $2\|u\|_{H^1(\Omega)} + \|\psi\|_{H^1(\Omega)}$. Combining the estimates (5.4), (5.5) and (5.6) , we obtain that there holds

$$|DA(u_q; \tilde{u}_q - u_q)| \leq \|\mathcal{P}(u)\|_{L^2(\Psi)}\, \|\tilde{u}_q - u_p\|_{L^2(\Psi)} + \tilde{C}_3|u_q - u|^2_{H^1(\Omega)} + \tfrac{\kappa_l}{2}|\tilde{u}_q - u_q|^2_{H^1(\Omega)} + \tilde{C}_4 p^{-1}$$

with $\tilde{C}_3 := \frac{\kappa_u^2}{2\kappa_l}$ and $\tilde{C}_4 := 4c\|u\|_{H^1(\Omega)} + 2c\|\psi\|_{H^1(\Omega)}$.
Setting $\tilde{C}_1 := \kappa_l^{-1}\kappa_u^2$, $\tilde{C}_2 := \|\mathcal{P}(u)\|_{L^2(\Psi)}$ and using (5.3), we summarize

$$\begin{aligned} \kappa_l|\tilde{u}_q - u_q|^2_{H^1(\Omega)} \;\leq\; & \tfrac{\kappa_l}{4}\,|\tilde{u}_q - u_q|^2_{H^1(\Omega)} \\ & + \tilde{C}_1\big(|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)} + |u_{q,\Gamma} - u_{p,\Gamma}|_{H^1(\Omega)} + \sqrt{2}(5\rho_4 + \rho_5)\|\nabla(\tilde{u}_q - u_p)|^2_{L^4(\Omega)}\big)^2 \\ & + \tilde{C}_2\|\tilde{u}_q - u_p\|_{L^2(\Psi)} \;+\; \tilde{C}_3|u_q - u|^2_{H^1(\Omega)} \;+\; \tfrac{\kappa_l}{2}|\tilde{u}_q - u_q|^2_{H^1(\Omega)} + \tilde{C}_4 p^{-1}\,. \end{aligned}$$

For $\sigma = 0$ arithmetic transformations yield the proposition with the constants $C_i = \frac{4}{\kappa_l}\tilde{C}_i$, $i = 1, 2, 3, 4$. As the semi-norm $|\cdot|_{H^1(\Omega)}$ in the proof can be replaced by the norm $\|\cdot\|_{H^1(\Omega)}$ in case of $\sigma > 0$, the statement for $\sigma > 0$ follows analogously. □

In comparison to the a posteriori estimate for Algorithm 5.1 the treatment of variational inequalities causes four additional terms,

$$C_1|\tilde{u}_q - \hat{u}_q|_{H^1(\Omega)}\,, \quad C_2\|\tilde{u}_q - u_p\|_{L^2(\Psi)}\,, \quad C_3|u_q - u|^2_{H^1(\Omega)}\,, \quad \text{and} \quad + C_4 p^{-1}\,.$$

Here, the first and the last term can be calculated easily. The second term can not be computed exactly because the coincidence set Ψ is not known. A simple work around is to calculate the norm $\|\cdot\|_{L^2(\Omega)}$ instead of $\|\cdot\|_{L^2(\Psi)}$. The third term may be estimated based on a priori knowledge (cf. Theorem 2.8 and Theorem 2.11).

5.2 Space decomposition methods for nonlinear problems

Two algorithms are described in this section. Both are devoted to unconstrained nonlinear minimization problems originated by the p-version FEM on quasi-uniform meshes in two dimensions and generalize an approach of Babuška, Craig, Mandel, and Pitkäranta in [BCMP91] to nonlinear solvers. There, the authors suggest a space decomposition of p-version FE spaces on curvilinear triangles and quadrilaterals for the efficient preconditioning of linear systems.

In the following, we present their space decomposition for a p-version space on quadrilaterals and the stability estimate given in [BCMP91] (see Theorem 5.4). Then, the stability estimate will be used to prove linear convergence rates of the nonlinear decomposition solvers defined by Algorithm 5.3 and Algorithm 5.4. Both algorithms were described by Tai and Espedal in *Rate of convergence of some space decomposition methods for linear and nonlinear problems* [TE98a]. There, the authors applied the algorithms to an overlapping two-grid decomposition of a h-version FE space.

To start with, we recall the hierarchical p-version FE basis given in [BCMP91] which allows the direct sum decomposition of V_p into nodal, edge, and interior associated spaces.

(i) As usual, we take the bilinear nodal functions with the property $N_i(\xi_j) = 0$ for $i \neq j$ and $N_i(\xi_j) = 1$ for $i = j$. Here, $\xi_j = (\xi_{j,1}, \xi_{j,2})$ denotes the node j of the reference square $\tilde{Q}$, $j = 1, 2, 3, 4$. These read as

$$\begin{aligned} N_1(\xi) &= \tfrac{1}{4}(1-\xi_{\cdot,1})(1-\xi_{\cdot,2}), \qquad & N_2(\xi) &= \tfrac{1}{4}(1+\xi_{\cdot,1})(1-\xi_{\cdot,2}), \\ N_3(\xi) &= \tfrac{1}{4}(1+\xi_{\cdot,1})(1+\xi_{\cdot,2}), \qquad & N_4(\xi) &= \tfrac{1}{4}(1-\xi_{\cdot,1})(1-\xi_{\cdot,2}), \end{aligned} \tag{5.7}$$

and give the one dimensional spaces $\mathcal{N}_j$, $j = 1, 2, 3, 4$.

(ii) The edges $\hat{\Gamma}_j$, $j = 1, 2, 3, 4$, of $\tilde{Q}$ are associated with

$$\begin{aligned} N_i^{[1]}(\xi) &= \tfrac{1}{2}(1-\xi_{\cdot,2})\Phi_i(\xi_{\cdot,1}), \qquad & i &= 1, 2, \ldots, p-1, \\ N_i^{[2]}(\xi) &= \tfrac{1}{2}(1+\xi_{\cdot,1})\Phi_i(\xi_{\cdot,2}), \qquad & i &= 1, 2, \ldots, p-1, \\ N_i^{[3]}(\xi) &= \tfrac{(-1)^i}{2}(1+\xi_{\cdot,2})\Phi_i(\xi_{\cdot,1}), \qquad & i &= 1, 2, \ldots, p-1, \\ N_i^{[4]}(\xi) &= \tfrac{(-1)^i}{2}(1-\xi_{\cdot,1})\Phi_i(\xi_{\cdot,2}), \qquad & i &= 1, 2, \ldots, p-1, \end{aligned}$$

where Φ_i denotes the scaled anti-derivative of the Legendre polynomial P_i of degree i

$$\Phi_i(t) := \sqrt{\frac{2i-1}{2}} \int_{-1}^{t} P_i(\tau)\ \mathrm{d}\tau\,.$$

(iii) The set $\hat{\mathcal{J}}$ of the internal shape functions. For $p \geq 4$ there are $(p-1)^2$ internal shape functions defined as

$$N_{i,j}^0(\xi) = (1-\xi_{\cdot,1}^2)(1-\xi_{\cdot,2}^2)\, P_i(\xi_{\cdot,1})\, P_j(\xi_{\cdot,2}), \quad 0 \leq i, j \leq p-2\,,$$

where P_i, P_j denote Legendre polynomials of degree i, j, respectively.

As usual, we obtain the nodal, edge and interior associated global basis functions b_N, b_e^i, and b_Q^{ij}, respectively, by the mappings F_Q from the above defined basis functions on $\bar{Q}$. We denote their sets $\mathcal{N} := \{b_N : N \in \text{Nodes}\}$, $\mathcal{E}_e := \{b_e^i : \ i \text{ as in (ii)}\}$, $e \in \text{Edges}$, and $\mathcal{I}_Q := \{b_Q^{ij} : \ i,j \text{ as in (iii)}\}$, $Q \in \mathcal{T}$, and write the direct sum decomposition

$$V_p = \operatorname{span}\mathcal{N} \oplus \bigoplus_{e\in\text{Edges}} \operatorname{span}\mathcal{E}_e \oplus \bigoplus_{Q\in\mathcal{T}} \operatorname{span}\mathcal{I}_Q\,. \tag{5.8}$$

Here, Nodes and Edges are the sets of nodes and edges of the partition $\mathcal{T}$. Babuška et. al proved the following stability estimate for this decomposition.

Theorem 5.4. Let $\mathcal{T}$ be a quasi-uniform mesh of curvilinear quadrilaterals and let $v \in V_p = V_p(\mathcal{T})$ be decomposed as above, i.e.,

$$v = v_N + \sum_{e\in\text{Edges}} v_e + \sum_{Q\in\mathcal{T}} v_Q$$

with $v_N \in \operatorname{span}\mathcal{N}$, $v_e \in \operatorname{span}\mathcal{E}_e$, $v_Q \in \operatorname{span}\mathcal{I}_Q$. Then, there exists a constant $C_d > 0$ such that

$$|v_N|^2_{1,\Omega} + \sum_{e\in\text{Edges}} |v_e|^2_{1,\Omega} + \sum_{Q\in\mathcal{T}} |v_Q|^2_{1,\Omega} \ \le\ C_d^2(1+\log^2 p)\,|v|^2_{1,\Omega}\,. \tag{5.9}$$

Proof. We have (5.9) according to [BCMP91, Theorem 3.5]. □

Now, we use the space decomposition of V_p to calculate the minimizer $u_p \in V_{p,g_D}$ of the functional A. To generalize and to simplify the notation, we introduce a global numbering $V^{(i)}$, $i = 1,2,\ldots,m$, of the m subspaces, i.e., we write the decomposition of $v \in V_p$ given by Theorem 5.4 as

$$v = \sum_{i=1}^{n} v_i \quad \text{where } v_i \in V^{(i)}$$

and formulate the following two algorithms. In both the constant ϵ_0 controls how accurately the subproblems in *Step 2* will be solved. Algorithm 5.3 employs parallel subspace corrections to yield the minimum of the global problem, whereas Algorithm 5.4 uses successive subspace corrections.

Algorithm 5.3 Additive space decomposition method

Let initial values $u_i^0 \in V^{(i)}$ and relaxation parameters $\alpha_i > 0$ be given such that $\sum_{i=1}^m \alpha_i \leq 1$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^N$. Further, we have the termination parameters $\epsilon_0 > 0$ for the local nonlinear minimizations and $\epsilon_{glob} > 0$ for the outer iterations.

1. For $n \geq 0$, let $\hat{u}_i^{n+\frac{1}{2}} \in V^{(i)},\ i = 1, \ldots, m$, satisfy

$$A\Big(\sum_{k=1,k\neq i}^{m} u_k^n + \hat{u}_i^{n+\frac{1}{2}}\Big) \leq A\Big(\sum_{k=1,k\neq i}^{m} u_k^n + v_i\Big) \quad \text{for all } v_i \in V^{(i)}\,.$$

   Use an approximate solver to find $u_i^{n+\frac{1}{2}}$ in parallel for $i = 1, \ldots, m$, such that

$$\|u_i^{n+\frac{1}{2}} - \hat{u}_i^{n+\frac{1}{2}}\|_{H^1(\Omega)} \leq \epsilon_0 \|u_i^n - \hat{u}_i^{n+\frac{1}{2}}\|_{H^1(\Omega)}\,. \tag{5.10}$$

2. For $i = 1, \ldots, m$, set

$$u_i^{n+1} = u_i^n + \alpha_i (u_i^{n+\frac{1}{2}} - u_i^n)\,.$$

   If $\|\nabla A\big(\sum_{k=1}^m u_k^{n+1}\big)\| > \epsilon_{glob}$, set $n = n+1$ and continue with *Step 1,* else exit.

Algorithm 5.4 Multiplicative space decomposition method

Let initial values $u_i^0 \in V^{(i)}$ and relaxation parameters $\alpha_i > 0$ be given such that $\sum_{i=1}^m \alpha_i \leq 1$. Let $\|\cdot\|$ be a norm on $\mathbb{R}^N$. Further, we have the termination parameters $\epsilon_0 > 0$ for the local nonlinear minimizations and $\epsilon_{glob} > 0$ for the outer iterations.

1. For $n \geq 0$, let $\hat{u}_i^{n+1} \in V^{(i)}$, $i = 1, \ldots, m$, satisfy

$$A\Big(\sum_{1\leq k<i} u_k^{n+1} + \hat{u}_i^{n+1} + \sum_{i<k\leq m} u_k^n\Big) \leq A\Big(\sum_{1\leq k<i} u_k^{n+1} + v_i + \sum_{i<k\leq m} u_k^n\Big) \quad \text{for all } v_i \in V^{(i)}\,.$$

   Use an approximate solver to find u_i^{n+1} sequentially for $i = 1, \ldots, m$, such that

$$\|u_i^{n+1} - \hat{u}_i^{n+1}\|_{H^1(\Omega)} \leq \epsilon_0 \|u_i^n - \hat{u}_i^{n+1}\|_{H^1(\Omega)}\,.$$

2. If $\|\nabla A\big(\sum_{k=1}^m u_k^{n+1}\big)\| > \epsilon_{glob}$, set $n = n+1$ and continue with *Step 1,* else exit.

Due to the nonlinearity of A we can not state an energy norm to analyze the convergence of u^n of Algorithms 5.3 and 5.4, respectively, against the minimizer $u_p \in V_{p,g_D}$. Instead we introduce the A-energy form $d_A(\cdot,\cdot)$ defined by

$$d_A(u,v) := |DA(u; u-v) - DA(v; u-v)| \qquad \text{for all } u, v \in H^1(\Omega)\,. \tag{5.11}$$

The following two theorems state linear convergence rates for both algorithms.

Theorem 5.5. Let the functional A be defined by (1.8) with ρ satisfying (1.6). Then

Algorithm 5.3 converges against the minimizer $u_p \in V_{p,g_D}$ with

$$d_A(u^{n+1}, u_p) \leq \sqrt{1 - \frac{1}{C \log^2 p}}\, d_A(u^n, u_p) \tag{5.12}$$

with a constant C independent of p.

Proof. Algorithm 5.3 is a special case of [TE98a, Algorithm 2.1]. With [TE98a, Theorem 3.1] and an inspection of equation (3.7) of its proof it suffices to prove the following hypotheses.
H1 There are constants $\kappa_l > 0$, $\kappa_u < \infty$ such that

$$\kappa_l |u - v|^2_{1,\Omega} \leq DA(u; u - v) - DA(v; u - v) \leq \kappa_u |u - v|^2_{1,\Omega} \quad \text{for all } u, v \in H^1(\Omega)\,.$$

H2 There exists a $C_{H2} > 0$, independently of v, such that for any $v \in V_p$ decomposed as in Theorem 5.4 we have

$$\sum_{i=1}^{m} |v_i|^2 \leq C^2_{H2} |v|^2\,.$$

H3 There exists a $C_{H3} > 0$ such that

$$\sum_{i=1}^{m} D^2 A(w_i; u_i, \sum\nolimits_{j=1}^{m} v_j) \leq C_{H3} \left(\sum\nolimits_{i=1}^{m} |u_i|^2_{H^1(\Omega)} \right)^{1/2} \left(\sum\nolimits_{i=1}^{m} |v_i|^2_{H^1(\Omega)} \right)^{1/2}$$
$$\text{for all } w_i \in V_p, \quad u_i \in V^{(i)}, \quad v_j \in V^{(j)}\,.$$

H4 The subproblems of *Step 1* are solved accurately enough, i.e., there holds

$$(1 + \epsilon_0)\epsilon_0 \leq \frac{\kappa_l}{4\kappa_u} \leq \frac{1}{4}$$

where κ_l, κ_u are the constants from **H1**.

Verification of H1. Since A satisfies the assumptions of Lemma 1.21, we can take κ_l and κ_u as in the proof of the lemma and obtain the hypothesis with (1.12).
Verification of H2. This hypothesis is satisfied due to Theorem 5.4, if we take $C_{H2} = C_d(1 + \log^2 p)^{1/2}$.
Verification of H3. For all $w_i \in V$ and for all $u_i \in V^{(i)}$, $v_j \in V^{(j)}$ we get with $t_i := |\nabla w_i|$

and the constants ρ_1, ρ_3 from Lemma 1.21

$$\begin{aligned}
\sum_{i=1}^m D^2A(w_i;u_i,\textstyle\sum_{j=1}^m v_j) &= \sum_{i=1}^m \int_\Omega \rho(t_i)(\nabla u_i)^T\big(\textstyle\sum_{j=1}^m \nabla v_j\big)\ \mathrm{d}x \\
&\qquad + \sum_{i=1}^m \int_\Omega t_i\rho'(t_i)\Big(\tfrac{(\nabla w_i)^T}{t_i}\nabla u_i\Big)\Big(\tfrac{(\nabla w_i)^T}{t_i}\big(\textstyle\sum_{j=1}^m \nabla v_j\big)\Big)\Big)\ \mathrm{d}x \\
&\le \int_\Omega \big(\textstyle\sum_{i=1}^m |\rho(t_i)\nabla u_i|\big)\,\big|\sum_{j=1}^m \nabla v_j\big|\ \mathrm{d}x + \int_\Omega \big(\sum_{i=1}^m |t_i\rho'(t_i)|\,|\nabla u_i|\big)\,\big|\sum_{j=1}^m \nabla v_j\big|\ \mathrm{d}x \\
&\le \rho_1 \int_\Omega \big(\textstyle\sum_{i=1}^m |\nabla u_i|\big)\,\big|\sum_{j=1}^m \nabla v_j\big|\ \mathrm{d}x + (\rho_1+\rho_3)\int_\Omega \big(\sum_{i=1}^m |\nabla u_i|\big)\,\big|\sum_{j=1}^m \nabla v_j\big|\ \mathrm{d}x \\
&\le (2\rho_1+\rho_3)\big\|\textstyle\sum_{i=1}^m |\nabla u_i|\big\|_{L^2(\Omega)}\,\big\|\,\big|\sum_{i=1}^m \nabla v_i|\,\big\|_{L^2(\Omega)} \\
&\le (2\rho_1+\rho_3)\Big(\sum_{i=1}^m \||\nabla u_i|\|^2_{L^2(\Omega)}\Big)^{1/2}\Big(\sum_{i=1}^m \||\nabla v_i|\|^2_{L^2(\Omega)}\Big)^{1/2} \\
&= (2\rho_1+\rho_3)\Big(\sum_{i=1}^m |u_i|^2_{H^1(\Omega)}\Big)^{1/2}\Big(\sum_{i=1}^m |v_i|^2_{H^1(\Omega)}\Big)^{1/2}.
\end{aligned} \tag{5.13}$$

Thus, **H3** holds with $C_{H3} = 2\rho_1 + \rho_3$.

Verification of H4. H4 holds if the subproblems are solved accurately enough. If a subproblem is solved by an iteration procedure the relative change against the preceding iteration gives an estimate how accurate the subproblem is solved already. The norm on the right of (5.10) can be estimated by comparing u_i^n with its predecessor u_{i-1}^n. The iteration process continues as long as H4 is not satisfied.

Introducing the notation

$$e^n := d_A(u^n, u_p) = |DA(u^n; u^n - u_p) - DA(u_p; u^n - u_p)| \tag{5.14}$$

we know again by [TE98a, Theorem 3.1(b)]

$$e^{n+1} \le \beta_n^{1/2}\, e^n \quad \text{for all } n \ge 0$$

with

$$\lim_{n\to\infty} \beta_n = \frac{C_\beta}{1+C_\beta} \qquad \text{and} \qquad C_\beta := \frac{2}{\kappa_l^2} C_{H3}^2\, C_{H2}^2 \left((1+\epsilon_0)\alpha_{\min}^{-1/2} + \alpha_{\max}^{1/2}\right)^2.$$

Here, $\alpha_{\min}$ and $\alpha_{\max}$ denote the minimum and maximum of the relaxation parameters α_i. Substituting κ_l, C_{H2}, C_{H3}, $\alpha_{\min}$, $\alpha_{\max}$, and ϵ_0 according to the verification of H1–H4 yields

$$e_{n+1} \le \left(1 - \frac{1}{C\log^2 p}\right)^{1/2} e_n$$

with a constant C independent of p for n sufficiently high. (5.12) follows by definition of e_n in (5.14). □

Theorem 5.6. Let the functional A be defined by (1.8) with ρ satisfying (1.6). Then Algorithm 5.4 converges against the minimizer $u_p \in V_{p,g_D}$ with

$$d_A(u^{n+1}, u_p) \le \sqrt{1 - \frac{1}{C\log p}}\; d_A(u^n, u_p) \tag{5.15}$$

with a constant C independent of p.

Proof. Algorithm 5.4 is a special case of [TE98a, Algorithm 2.2]. With [TE98a, Theorem 4.1] and an inspection of equation (4.4) of its proof it suffices to prove the same four hypotheses as in the proof of Theorem 5.5 to get the convergence rate

$$e^{n+1} \leq \beta_n^{1/2} e^n$$

with

$$\lim_{n\to\infty} \beta_n = \frac{C_\beta}{1+C_\beta} \qquad \text{and} \qquad C_\beta := \frac{2}{\kappa_l^2} C_{H2} C_{H3}\,.$$

Substituting κ_l, C_{H2}, C_{H3} according to the verification of H1–H4 in the proof of Theorem 5.5 gives

$$e^{n+1} \leq \left(1 - \frac{1}{C \log p}\right)^{1/2} e^n$$

with a constant C independent of p for n sufficiently high. (5.15) follows by the definition of e^n in (5.14). □

Theorem 5.5 and Theorem 5.6 also imply the convergence of the additive and the multiplicative space decomposition method against u_p in the $\|\cdot\|_{H^1(\Omega)}$-norm due to inequality (1.12) and the Poincaré-Friedrichs inequality. Using Banach's fixed point theorem gives the following a priori error estimates.

Corollary 5.7. The iterative u^n of Algorithms 5.3 converges against $u_p \in V_{p,g_D}$ in the norm $\|\cdot\|_{H^1(\Omega)}$ with

$$\|u^n - u_p\|_{H^1(\Omega)} \leq C_{asm1} \left(1 - \frac{1}{C_{asm2} \log^2 p}\right)^{\frac{n}{2}} \|u^1 - u^0\|_{H^1(\Omega)}\,.$$

The iterative u^n of Algorithms 5.4 converges against $u_p \in V_{p,g_D}$ in the norm $\|\cdot\|_{H^1(\Omega)}$ with

$$\|u^n - u_p\|_{H^1(\Omega)} \leq C_{msm1} \left(1 - \frac{1}{C_{msm2} \log p}\right)^{\frac{n}{2}} \|u^1 - u^0\|_{H^1(\Omega)}\,.$$

Here, C_{asm1}, C_{asm2}, C_{msm1}, and C_{msm2} denote positive constants independent of p.

Unfortunately, the convergence rate of both algorithms depends on p. Both algorithm converged slowly in the numerical experiments. We document and comment the performance of Algorithm 5.4 in the next section (see Experiment 5.9).

5.3 Numerical experiments

The model problem. We consider the model potential problem

$$-\operatorname{div}(\rho(|\nabla u|)\nabla u) + f = 0 \quad \text{with } \rho(t) = \tfrac{1}{6}\big(1 + \tfrac{5}{1+5t}\big)$$

and the homogeneous Dirichlet condition $u_{|\partial\Omega} = 0$ on the unit square $\Omega := [-1,1]^2$. The function f is determined such that the boundary value problem is solved by

$$u(x) = \sin(\tfrac{3}{2}\pi(x_1 - 1)) \cdot \sin(\pi x_2)\,.$$

Discretization. We discretize the model problem by the p-version on a uniform grid $\mathcal{T}$ of 16 squares, each with a side length of 0.5 units. Using the notation of Section 2.1, we look for the unconstrained minimum of the functional A as defined in (1.8) on $V_{p,0}$. The integrals of $A(v)$, the gradient $\nabla A(v)$, and the Hessian $\nabla^2 A(v)$, $v \in V_p$, are calculated with $4 + p$ quadrature points.

Solving the discrete nonlinear problems. The nonlinear global systems of Experiment 5.8 and the nonlinear local systems of the multiplicative Schwarz method used in Experiment 5.9 are solved by a modified Newton backtracking method. In the modified version of Algorithm 4.2, the Hessian $H(\underline{u}_k)$ used in *Step 2* and *Step 4* are updated only for the Newton iterations $k = 0, 5, 10, \ldots$. The computations of the other Newton iterations were done using the Hessian of the previous iterations. This modification is named *Shamanskii method* in [Kel95]. It converges q-superlinearly due to [Kel95, Theorem 5.4.5]. It is a drawback of this nonlinear solver that it increases the number of Newton like iterations. Nevertheless, the Shamanskii method performs faster as the strict Newton approach because it reduces the computation time needed for the calculations of the Hessian $H(\underline{u}_k)$. In Corollary B.10, it is stated that the calculation of one Hessian costs $\mathcal{O}(p^6)$ floating point operations. Thus, the Hessians are the bottle-neck of the nonlinear solvers for the p-version. In the experiments, a re-computation of the Hessian in every fifth iteration showed a quite good performance for different p. The Newton iterations are repeated until the gradient $\nabla A(\underline{u}_k)$ fulfills the exactness criterion $\|\nabla A(\underline{u}_k)\|_2 \leq 10^{-9}$ (see *Step 7* of Algorithm 4.2). As initial $\underline{u}_0$ we choose the zero vector.

Implementation. The following two experiments are implemented in FORTRAN 90. All real variables are defined as double precision. The program is compiled with the SUN compiler `f90` using the `-fast` option for the cpu-timings . The computations are performed on a SUN Ultra-5.

Experiment 5.8 (*Prolongation cascade***).** To get an impression of the practical performance of the prolongation idea presented in Algorithm 5.1, we test a heuristic extension, a cascade of prolongations, here. To give an example, we consider the field

50
10 13 16
7 7 5

of Column $j = 3$ and Row $p = 16$ in Table 5.1. The row and the column give us the information that we search the minimum of A on $V_{p,0}$, $p = 16$, with the prolongation increment $j = 3$. Firstly, we minimize A on $V_{1,0}$. The solution is prolongated into $V_{4,0}$ and A is minimized on $V_{4,0}$. Prolongating and solving is repeated for $p = 7, 10, 13, 16$. Speaking generally, the prolongation cascade searches the minimizer of A on V_p by solving the minimization problems successively on V_{p_i} with $p_0 := p \mod j$, $p_i := p_{i-1} + j$, $i = 1, \ldots, \lceil \frac{p}{j} \rceil - 1$.

From the first line of the field, we know that the sum of Newton iterations needed on the levels $p = 1, 4, \ldots, 13, 16$ is 50. The second line of the field lists the polynomial degrees of the subspaces used in the last three prolongation steps, here, $p = 10, 13, 16$. The last line lists the number of Newton iterations needed on each of the last three levels $p = 10, 13, 16$, here, $7, 7, 5$.

p	j=1	j=3	j=5	j=7	j=9	j=11	j=13	j=p
2	20 1 2 12 8							12
4	38 2 3 4 8 10 8	24 1 4 12 12						12
6	49 4 5 6 8 7 4	21 3 6 12 9	24 1 6 12 12					12
8	63 6 7 8 4 7 7	31 2 5 8 12 12 7	20 3 8 12 8	27 1 8 12 15				12
10	76 8 9 10 7 6 7	38 4 7 10 12 7 7	19 5 10 12 7	20 3 10 12 8	31 1 10 12 19			12
12	90 10 11 12 7 7 7	32 6 9 12 9 5 6	30 2 7 12 12 12 6	19 5 12 12 7	20 3 12 12 8	34 1 12 12 22		12
14	103 10 11 12 7 7 7	42 8 11 14 7 5 6	24 4 9 14 12 7 5	19 7 14 12 7	18 5 14 12 6	20 3 14 12 8	36 1 14 12 24	12
16	117 7 8 9 7 7 6	50 10 13 16 7 7 5	35 6 11 16 12 5 6	34 2 9 16 12 15 7	19 7 16 12 7	19 5 16 12 7	22 3 16 12 10	12

Legend:

Number of iterations Last 3 prolongation levels Number of iterations of the last 3 levels

Table 5.1: Number of iterations of the prolongation scheme (see Experiment 5.8)

When the prolongation increment j is big or p is small, there exist less than three prolongation levels. In case of $j = p$, no prolongation is employed to solve the problem.

As the most important information, we obtain from the last column of Table 5.1 that the number of Newton iterations does not depend on p.

Furthermore, we note that the number of iterations needed for the full problem can be reduced by the prolongation method, when the increments $j = 1, 3, 5$ are used. But prolongation increments $j \geq 7$ can even increase the number of iterations on the last level, e.g. $p = 8$, $j = 7$.

The total number of iterations and the number of iterations on different levels are not very interessting from the practical point of view, since the costs of one iteration depend on p. The cpu-timings documented in Table 5.2 are more important because they give an estimate on the proportions of the sum of floating point operations needed for the different tasks, namely, the computation of the Hessians and the gradients, the solving of the linear systems by the diagonally preconditioned cg-method. Only in a few cases the prolongation is superior to the global solving with respect to the total timing. The prolongation increments $j = 3, 5$ yielded total timings which are still comparable to the

p	j=1	j=3	j=5	j=7	j=9	j=11	j=13	j=p
2	20 0.045 0.117 0.007 0.169							12 0.043 0.112 0.039 0.195
4	38 0.262 0.379 0.105 0.748	24 0.227 0.264 0.103 0.595						12 0.211 0.208 0.056 0.475
6	49 0.747 0.663 0.475 1.892	21 0.571 0.423 0.473 1.470	24 0.713 0.425 0.593 1.735					12 0.693 0.368 0.273 1.335
8	63 3.397 1.475 2.958 7.854	31 2.139 0.855 2.085 5.091	20 1.798 0.692 2.051 4.552	27 2.596 1.082 3.684 7.372				12 2.545 0.811 1.236 4.593
10	76 7.424 3.064 10.071 20.618	38 7.757 2.743 12.635 23.163	19 6.081 2.814 9.820 18.738	20 5.392 2.604 10.888 19.065	31 10.461 5.616 25.327 41.466			12 3.894 1.627 3.155 8.678
12	90 38.563 12.872 74.703 126.590	32 17.771 4.758 36.082 58.670	30 17.901 4.708 32.202 54.858	19 15.966 3.817 36.072 55.898	20 15.275 4.044 42.419 61.819	34 37.440 10.751 112.493 160.805		12 11.209 2.943 11.444 25.599
14	103 41.212 10.813 93.994 146.268	42 17.758 4.298 40.672 62.838	24 8.315 2.984 30.269 41.653	19 14.609 3.494 40.404 58.583	18 13.756 2.785 32.780 49.394	20 13.408 3.485 43.696 60.660	36 33.122 10.213 133.853 177.258	12 39.889 9.920 45.831 95.644
16	117 92.980 19.534 224.260 337.229	50 27.572 7.187 82.899 117.854	35 33.719 5.338 69.911 109.115	34 32.844 6.191 78.746 117.910	19 31.985 5.205 71.520 108.832	19 31.019 4.910 75.098 111.144	22 30.725 6.757 108.092 145.688	12 46.091 7.796 43.313 97.204

Legend:

Number of iterations
Hessian time
Gradient time
Conjugate gradient time
Total time

Table 5.2: Number of iterations and cpu-timings in seconds on a SUN Ultra-5 (see Experiment 5.8)

non-prolongated scheme. Estimating roughly the calculation of the Hessians and the solution of the linear systems cause the same costs for $10 \leq p \leq 16$, when prolongation is omitted.

Experiment 5.9 (*Nonlinear multiplicative Schwarz method***).** We decompose the global space $V_{p,0}$ according to (5.8) into the 9 dimensional subspace $V^{(1)}$ associated to the interior nodes, the $(p-1)$ dimensional subspaces $V^{(i)}$, $i = 2, \ldots, 25$, associated to the 24 interior edges, and the $(p-1)^2$ dimensional subspaces $V^{(i)}$, $i = 26, 41$, associated to the interior degrees of freedom on the 16 squares.

The minimum $\hat{u}_i^{n+1}$ of the local problem $A^{(i)}$ on $V^{(i)}$ defined by *Step 2* of Algortihm 5.4 is determined as pointed out in the paragraph *Solving the discrete nonlinear problems* above such that the gradient fulfills $\|\nabla A^{(i)}(\hat{u}_i^{n+1})\|_2 \leq 10^{-11}$. This means that the stopping criterion

$$\|u_i^{n+1} - \hat{u}_i^{n+1}\|_{H^1(\Omega)} \leq \epsilon_0 \|u_i^n - \hat{u}_i^{n+1}\|_{H^1(\Omega)}$$

of Algorithm 5.4 which demands only a relative improvement u_i^{n+1} in the local minimization process, is fulfilled because of $u_i^{n+1} = \hat{u}_i^{n+1}$. Of course, this exactness leads to a higher number of inner iterations i needed for the local minimizations. But it ensures that the number ouf outer iterations n needed to solve the global problem by the multiplicative Schwarz method until $\|\nabla A(\hat{u}_{n+1})\|_2 \leq \epsilon_{glob} = 10^{-9}$ will be minimal. Nevertheless, Table 5.3 shows big numbers for the outer iterations.

Corresponding to Theorem 5.6, the multiplicative method converges and needs an increasing number of outer iterations to reach the asked exactness $\epsilon_{glob} = 10^{-9}$, when p increases. The cpu-timings make it clear that the multiplicative Schwarz method of Algorithm 5.4 can not be used practically.

We also used the decomposition of the experiment for the additive Schwarz method defined by Algorithm 5.3. As the results were worse than those of multiplicative Schwarz method, we have not documented them.

We give an interpretation of the disappointing performance of the prolongation and the multiplicative Schwarz method for nonlinear PDE in the following remark.

Remark 5.10. The main reason to employ multi-level and space decomposition methods in FEM for the solution of linear systems is the reduction of the condition number to ensure numerical stability and the efficiency of iterative solvers. The numerical analysis of the h- and the p-version shows that the condition number of global linear systems can grow drastically, if we choose a wrong basis (cf. Experiment 4.36) or do not apply a preconditioning technique. An appropriate preconditioning is necessary to guarantee the solvability of large-scale linear systems.

Due to the last column of Table 5.1 the number of Newton iterations needed to solve the problem is independent of p, i.e., of the dimension of the system. This observation corresponds to the numerical analysis of Newton's method which shows that the neighborhood of the minimizer u, where we have quadratic convergence, depends mainly on $\|(\nabla^2 A(u))^{-1}\|$ and $\|\nabla A(u)\|$. Here, $\|\cdot\|$ is a norm on $\mathbb{R}^n$ (see [Kel95, Theorem 5.1.1 and Theorem 5.1.2]). Therefore, the main reason for a successful application of multi-level and space decomposition methods to linear problems, the reduction of the iteration number needed by the iterative linear solver, has been dropped. The number of Newton or

p=2	p=4	p=6	p=8	p=10	p=12
25	75	130	210	310	429
0.369	6.096	39.904	233.220	1040.191	1568.639
0.735	9.433	50.454	165.965	505.209	867.635
0.178	0.673	3.535	21.505	126.418	275.273
1.285	16.216	93.934	420.925	1672.142	2712.144

Legend:

Number of outer iterations
Stiffness time
Residual time
Conjugate gradient time
Total time

Table 5.3: Iterations counts and cpu-timings of the multiplicative Schwarz method (see Experiment 5.9)

Shamanskii iterations remains already low, even for large scale problems, and the methods converge quadratically or q-superlinearly, respectively.

In comparison to this convergences, the linear and p-dependent convergence rates stated for the additive and the multiplicative Schwarz method in Theorem 5.5 and Theorem 5.6, respectively, look bad.

Appendix A

Preconditioned conjugate gradient method

Algorithm A.1 $\underline{u} = \text{pcg}(H, \underline{g}, M, \underline{u}_0, \varepsilon, i_{\max})$; Preconditioned conjugate gradient method

Given a symmetric positive definite $H \in \mathbb{R}^{N\times N}$, $\underline{g} \in \mathbb{R}^N$, a symmetric positive preconditioner M, an initial guess $\underline{u}_0$, a termination parameter ε, and a maximum number of iterations $i_{\max}$, the following algorithm solves the linear system $H\underline{u} = \underline{g}$ approximately until the residual $\|\underline{g} - H\underline{u}\| < \varepsilon$.

1. Set $i = 0$.
   Compute $\underline{r}_0 = \underline{g} - H\underline{u}_0$.
2. Do while $\|\underline{r}_i\| \geq \varepsilon$ and $i < i_{\max}$

   Solve $M\underline{z}_i = \underline{r}_i$.
   Set $i = i + 1$.
   If $i = 1$

   Set $\underline{p}_1 = \underline{z}_0$.
   Else

   Set $\beta_i = \underline{r}_{i-1}^T \underline{z}_{i-1} \,/\, \underline{r}_{i-2}^T \underline{z}_{i-2}$.
   Set $\underline{p}_i = \underline{z}_{i-1} + \beta_i, \underline{p}_{i-1}$.
   End if

   Set $\alpha_i = \underline{r}_{i-1}^T \underline{z}_{i-1} \,/\, \underline{p}_i^T H \underline{p}_i$.
   Set $\underline{u}_i = \underline{u}_{i-1} + \alpha_i \underline{p}_i$.
   Set $\underline{r}_i = \underline{r}_{i-1} - \alpha_i H \underline{p}_i$.
   End while
3. Set $\underline{u} = \underline{u}_i$.

Appendix B

Efficient implementation of the matrix-vector multiplication $H(\underline{u})\,\underline{v}$

The minimization processes of Algorithm 4.2 and Algorithm 4.3, both demand the approximate solutions of linear system $H(\underline{u}_k)\underline{y} = -g(\underline{u}_k)$. In Section 4.4, it was shown that the number of iterations grows as $\mathcal{O}(p^{3/2})$, when the preconditioned conjugate gradient method is applied straightforwardly, and as $\mathcal{O}(p^{1/2})$, when static condensation is used to eliminate the interior degrees of freedom locally. The first implementations of the numerical examples from Section 2.3 made it clear soon, that most of the costs for computing $\underline{y}$ were not caused by the total of matrix-vector multiplications in the preconditioned conjugate gradient scheme, but by the computation of the matrix $H(\underline{u})$. The evaluation of the integrands for the numerical quadrature became the most expensive part of the whole computation. It is the purpose of this section to discuss the expense of a full matrix computation and to give an efficient alternative. Remark B.4 and Corollary B.10 state that the number of floating point operations (flops) grows with $\mathcal{O}(p^6)$, when the full matrix is computed.

Fortunately, there is no need to calculate all entries of the matrix $H(\underline{u})$. Due to the use of an iterative solver, it suffices to calculate the matrix-vector product $H(\underline{u})\,\underline{v}$ for particular $\underline{v}$. It will be shown in the following that this product can be calculated efficiently with $\mathcal{O}(p^4)$ flops. Noting, that we need $\mathcal{O}(p^{3/2})$ diagonally preconditioned conjugate gradient iterations, this yields a total cost of $\mathcal{O}(p^{11/2})$ flops for the solution of a linear system. Additionally, we save a lot of memory because we do not need to store $H(\underline{u})$. This should improve the performance further since the information needed for the local computations can remain in the CPU's cache.

For a start, we recall the costs for the basic linear operations matrix-vector multiplication, matrix-matrix multiplication, and solution of a linear system $M\underline{v} = \underline{w}$ by Gaussian elimination. As usual the Gaussian elimination is computed in two steps. Firstly, the matrix is factorized into a unit lower triangular matrix L and an upper triangular matrix U. Secondly, the factorized system $LU\underline{v} = \underline{w}$ is solved. We give operation counts for the double precision LAPACK routines (cf. [ABB+95]) needed for these operations in Table B.1.

Remark B.1. We know that $\underline{y}$ has $\mathcal{O}(p^2)$ components due to the definition of V_p and

Let $k, m, n \in \mathbb{N}$, let $\underline{w} \in \mathbb{R}^m$, $\underline{v} \in \mathbb{R}^n$, let $M \in \mathbb{R}^{m\times n}$, $M_1, M_2 \in \mathbb{R}^{m\times k}$, $M_3 \in \mathbb{R}^{k\times n}$. Let L be a unit lower triangular matrix, U an upper triangular matrix with $L, U \in \mathbb{R}^{m\times m}$, let $M_4 \in \mathbb{R}^{m\times m}$. Then, the number of flops needed to perform the named linear operations is counted as follows:

linear operation	routine	multiplications	additions	total flops
$\underline{w} = M\underline{v}$	`dgemv`	mn	mn	2mn
$M = M_2M_3$	`dgemm`	mkn	mkn	2mkn
$LU = M_4$	`dgertf`	$\frac{1}{3}m^3 + \frac{2}{3}m$	$\frac{1}{3}m^3 - \frac{1}{2}m^2 + \frac{1}{6}m$	$\frac{2}{3}m^3 - \frac{1}{2}m^2 + \frac{5}{6}m$
$M_1 = U^{-1}L^{-1}M_2$	`dgerts`	km^2	$k(m^2 - m)$	$k(2m^2 - m)$

Table B.1: Operations counts for BLAS and LAPACK routines

that each local Galerkin matrices has $(p+1)^4$ entries. Assuming that $H(\underline{u})$ was already calculated, it follows with Table B.1 that the matrix-vector multiplication $H(\underline{u})\,\underline{v}$ costs $\mathcal{O}(p^4)$ flops.

Now, we develop an algorithm which computes the matrix-vector product $H(\underline{u})\,\underline{v}$ with $\mathcal{O}(p^4)$ flops and does not require the explicit computation of the matrix $H(\underline{u})$. Switching back from coordinate to vector notation $v = \sum_{l=1}^{N} v_l b_l$, and using the definition of $H(\underline{u})$,

$$H(\underline{u}) := \nabla^2 A(\underline{u}) = \big(D^2 A(\underline{u}; b_k, b_l)\big)_{0\le k,l\le N}\,,$$

the components of the product $(H(\underline{u})\,\underline{v})_k$, $k = 1, \dots, N := \operatorname{card} G_p$, can be rewritten as

$$(H(\underline{u})\underline{v})_k = D^2 A(u; b_k, v) = \sum_{Q\in\mathcal{T}} \int_Q f_{u;b_k,v}(x)\,\mathrm{d}x$$

where

$$f_{u;b_k,v}(x) := \Big(\rho(t)\nabla^T b_k \nabla v + t\rho'(t) s_1 s_2 + \sigma b_k v\Big)(x) \quad \text{with } t := |\nabla u|\,, \tag{B.1}$$

and

$$s_1 := \begin{cases} \frac{1}{t}\nabla^T u \nabla b_k, & \text{if } t > 0,\\ 0, & \text{if } t = 0,\end{cases} \qquad s_2 := \begin{cases} \frac{1}{t}\nabla^T u \nabla v, & \text{if } t > 0,\\ 0 & \text{if } t = 0.\end{cases}$$

Assuming that the integrands are polynomials of degree $\le 2q-1$, $q \in \mathbb{N}$, we have

$$\int_Q f_{u;b_k,v}(x)\,\mathrm{d}x = \int_{\tilde{Q}} \big(f_{u;b_k,v}\circ F_Q\big)(\eta)\,|\det DF_Q|\;\mathrm{d}\eta = \sum_{\xi\in G_{\tilde{Q},q}} \big(f_{u;b_k,v}\circ F_Q\big)(\xi)\,w(\xi)$$

where the set of the Gauss-Lobatto points $G_{\tilde{Q},q}$ is defined as in Definition 2.2 and the weights $w(\xi)$ are given by

$$w(\xi) := |\det DF_Q(\xi)|\,\rho_{r_1}^{q+1}\rho_{r_2}^{q+1} \quad \text{for } \xi = (\xi_{r_1}^{q+1}, \xi_{r_2}^{q+1}) \in G_{\tilde{Q},q}\,.$$

Here, ρ_r^{q+1}, $r = 0, \dots, q$, denote the weights of the Gauss-Lobatto quadrature (see (2.3)).

Remark B.2. The integrand $f_{u;b_k,v} \circ F_Q$ is not a polynomial in general because $\rho(t)$ can be non-polynomial. Assuming that $\rho(t)$ is a polynomial, the polynomial degree of the integrand $f_{u;b_k,v} \circ F_Q$ depends on the transformation F_Q, the function $\rho(t)$, the constant σ, and the polynomial degree p of the basis functions. If F_Q is an affine linear transformation, $\rho(t) = 1$, and $\sigma = 0$, then $f_{u;b_k,v} \circ F_Q$ has degree $2p - 2$. When an iso-parametric mesh is used, i.e., $Q = F_Q(\tilde{Q})$ with an iso-parametric transformation given by

$$F_Q(\eta) = \sum_{0 \le i_1, i_2 \le p} \lambda_{i_1}^p(\eta_1)\lambda_{i_2}^p(\eta_2) x_{i_1,i_2} \quad \text{for all } \eta \in \tilde{Q}$$

for the distinct points $x_{i_1,i_2} \in Q$, then the polynomial degree of the integrand caused by the term $\sigma b_k v$ increases to $4p$.

In our numerical experiments we used affine linear transformations F_Q and a nonlinear function $\rho(t)$. The numerical experiments yielded quite similar results for different $q \ge p + 2$.

Assumption B.3. In the remaining part of this section, we assume that the quadrature parameter q is given by $q = p + \tilde{q}$ where $\tilde{q}$ is a constant surplus which guarantees a sufficient approximation of the integrals $\int_Q \cdot \, \mathrm{d}x$ independently of p. Further, we assume that the costs for the evaluation of F_Q do not depend on p.

It is a drawback of Assumption B.3 that F_Q can not be an iso-parametric mapping. The assumption on F_Q is not essential for the following algorithms and their cost analysis. It serves an ease of notation, and the algorithms can be modified appropriately without increasing the order of costs.

Remark B.4. The numerical computation of the Hessian $H(\underline{u})$ using Gaussian quadrature or Gauss-Lobatto quadrature costs $\mathcal{O}(p^6)$ floating point operations.

Proof. We obtain the $(p+1)^4$ entries of the local Galerkin matrix $\big(H(\underline{v})\big)_{k_1,k_2} = \big(H(\underline{v})b_{k_2}\big)_{k_1}$, $1 \le k_1, k_2 \le (p+1)^2$, by evaluating $\big(f_{u;b_k,v} \circ F_Q\big)(\xi)w(\xi)$ for all $\xi \in G_{\tilde{Q},q}$. Since the number of flops needed for this evaluations does not depend on p, the remark follows because of $\operatorname{card} G_{\tilde{Q},q} = (q+1)^2 = \mathcal{O}(p^2)$. □

Using the local counting (Q, j_1, j_2) from (4.1), (4.2), we may write the basis functions as

$$b_k = b_{Q,j_1,j_2} = b_{j_1,j_2} \circ F_Q^{-1} \qquad \text{with } b_{j_1,j_2} \text{ defined by } b_{j_1,j_2}(\eta) := \lambda_{j_1}^p(\eta_1)\lambda_{j_2}^p(\eta_2).$$

To characterize v on Q it suffices to consider the local coordinates $v_{j_1,j_2} = v_{Q,j_1,j_2} = v_k$. In the following, we note these coordinates as column vector

$$\underline{v} := \big(v_{0,0}, \dots, v_{p,0} \mid \dots \mid v_{p,p}, \dots, v_{p,p}\big)^T \quad \text{and as matrix} \quad \underline{\underline{v}} := \begin{pmatrix} v_{0,0} & \dots & v_{0,p} \\ \vdots & & \vdots \\ v_{p,0} & \dots & v_{p,p} \end{pmatrix}.$$

The algorithms of this section are motivated by the following idea: Let f be a polynomial given by the Lagrangian polynomials with respect to the quadrature points, i.e., $f(t) = \sum_{i=0}^q f_i \lambda_i^q(t)$. Then, we have $\int_{-1}^1 f \, \mathrm{d}t = \sum_{i=0}^q f_i \rho_i^{q+1}$. This means that the integrand f does not need to be evaluated when coordinates with respect to the Lagrangian

basis $\left(\lambda_i^q\right)_{0\le i\le q}$ are already known. Since the coordinate transformation of $\underline{v}$ into the coordinates $\underline{v}^q$ with respect to the basis $\tilde{B}_q$ can be performed efficiently with $\mathcal{O}(p^3)$ flops (see Lemma B.5), a consequent generalization of this concept saves evaluations of the basis functions at the quadrature points.

Let N be the matrix mapping the basis functions $\tilde{B}_q = (b^q_{i_1,i_2})_{0\le i_1,i_2\le q}$ onto $\tilde{B}_p = (b_{i_1,i_2})_{0\le i_1,i_2\le p}$, i.e.,

$$\left(b_{0,0}\ \cdots\ b_{p,0}\,\middle|\,\cdots\,\middle|\,b_{0,p}\ \cdots\ b_{p,p}\right)^T = N\,\left(b^q_{0,0}\ \cdots\ b^q_{q,0}\,\middle|\,\cdots\,\middle|\,b^q_{0,q}\ \cdots\ b^q_{q,q}\right)^T. \tag{B.2}$$

It is shown in the following lemma that N and N^T can be applied efficiently due to the tensor product structure of the bases $\tilde{B}_p$ and $\tilde{B}_q$.

Lemma B.5. Suppose that C is the matrix mapping the Lagrangian polynomials of degree q to the Lagrangian polynomials of degree p, both with respect to the Gauss-Lobatto points of the respective degrees (see Section4.1) so that

$$\begin{pmatrix}\lambda_0^p(\xi)\\ \vdots\\ \lambda_p^p(\xi)\end{pmatrix} = C\begin{pmatrix}\lambda_0^q(\xi)\\ \vdots\\ \lambda_q^q(\xi)\end{pmatrix} \qquad \text{with } C := \begin{pmatrix}\lambda_0^p(\xi_0^{q+1}) & \cdots & \lambda_0^p(\xi_q^{q+1})\\ \vdots & & \vdots\\ \lambda_p^p(\xi_0^{q+1}) & \cdots & \lambda_p^p(\xi_q^{q+1})\end{pmatrix}. \tag{B.3}$$

Let $\underline{v}^q = \left(v^q_{0,0}\ \cdots\ v^q_{q,0}\,\middle|\,\cdots\,\middle|\,v^q_{0,q}\ \cdots\ v^q_{q,q}\right)^T \in \mathbb{R}^{(q+1)^2}$ be a column vector and

$$\underline{\underline{v}}^q = \begin{pmatrix}v^q_{0,0} & \cdots & v^q_{0,q}\\ \vdots & & \vdots\\ v^q_{q,0} & \cdots & v^q_{q,q}\end{pmatrix} \in \mathbb{R}^{(q+1)\times(q+1)}$$

the corresponding matrix obtained by taking $\underline{v}^q$ to $q+1$ columns of length $q+1$ one after another. Further, let $\underline{v} \in \mathbb{R}^{(p+1)^2}$ and $\underline{\underline{v}} \in \mathbb{R}^{(p+1)\times(p+1)}$ be defined correspondingly. We have the following correspondence of linear mappings on $\underline{v}$ and $\underline{\underline{v}}$.

$$\begin{array}{rcl}\underline{v} &=& N\underline{v}^q\\ \underline{v}^q &=& N^T\underline{v}\end{array} \qquad \text{corresponds to} \qquad \begin{array}{rcl}\underline{\underline{v}} &=& C\,\underline{\underline{v}}^q\,C^T,\\ \underline{\underline{v}}^q &=& C^T\,\underline{\underline{v}}\,C.\end{array} \tag{B.4}$$

Proof. We note the tensor product basis $\tilde{B}_p$ in matrix form,

$$\begin{aligned}\tilde{B}_p((\eta_1,\eta_2)) &= \begin{pmatrix}\lambda_0^p(\eta_1)\\ \vdots\\ \lambda_p^p(\eta_1)\end{pmatrix}\begin{pmatrix}\lambda_0^p(\eta_2)\\ \vdots\\ \lambda_p^p(\eta_2)\end{pmatrix}^T\\ &= C\begin{pmatrix}\lambda_0^q(\eta_1)\\ \vdots\\ \lambda_q^q(\eta_1)\end{pmatrix}\begin{pmatrix}\lambda_0^q(\eta_2)\\ \vdots\\ \lambda_q^q(\eta_2)\end{pmatrix}^T C^T &= C\,\tilde{B}_q((\eta_1,\eta_2))\,C^T\end{aligned}$$

for all $(\eta_1,\eta_2) \in \tilde{Q}$. Inserting the Gauss-Lobatto points $(\xi_{i_1}^{q+1}, \xi_{i_2}^{q+1}) \in G_{\tilde{Q},q}$, $0 \le i_1, i_2 \le q$, yields the mapping of all unit coordinate vectors of $\mathbb{R}^{(q+1)^2}$ and $\mathbb{R}^{(q+1)\times(q+1)}$. Thus, the first line of (B.4) is obtained due to linearity of N and C.

The second line of (B.4) is proved by comparing vector and matrix notation of a coordinate transformation. Let $v \in \operatorname{span} \tilde{B}_p$, and let $\underline{v}$ be its coordinate representations with respect to $\tilde{B}_p$. There exists a coordinate representation $\underline{v}^q$ of v with respect to the basis $\tilde{B}_q$ because of $\operatorname{span} \tilde{B}_p \subset \operatorname{span} \tilde{B}_q$. Thus, we can write

$$v = \sum_{0 \le i_1, i_2 \le p} v_{i_1,i_2} b_{i_1,i_2} = \left(\begin{array}{c} v_{0,0} \\ \vdots \\ v_{p,0} \\ \hline \vdots \\ \hline v_{0,p} \\ \vdots \\ v_{p,p} \end{array}\right)^T \left(\begin{array}{c} b_{0,0} \\ \vdots \\ b_{p,0} \\ \hline \vdots \\ \hline b_{0,p} \\ \vdots \\ b_{p,p} \end{array}\right) = \sum_{0 \le i_1, i_2 \le q} v^q_{i_1,i_2} b^q_{i_1,i_2} = \left(\begin{array}{c} v^q_{0,0} \\ \vdots \\ v^q_{q,0} \\ \hline \vdots \\ \hline v^q_{0,q} \\ \vdots \\ v^q_{q,q} \end{array}\right)^T \left(\begin{array}{c} b^q_{0,0} \\ \vdots \\ b^q_{q,0} \\ \hline \vdots \\ \hline b^q_{0,q} \\ \vdots \\ b^q_{q,q} \end{array}\right).$$

Using (B.2) to substitute $(b_{00}, \ldots, b_{pp})^T$ leads to $N^T \underline{v} = \underline{v}^q$. Now, let $\underline{\underline{v}}$, $\underline{\underline{v}}^q$ be the matrix representations with respect to $\tilde{B}_p$ and $\tilde{B}_q$, i.e.,

$$\begin{aligned} v((\eta_1,\eta_2)) = \sum_{0 \le i_1, i_2 \le p} v_{ij}\, \lambda^p_{i_1}(\eta_1)\lambda^p_{i_2}(\eta_2) \quad &= \begin{pmatrix} \lambda^p_0(\eta_1) \\ \vdots \\ \lambda^p_p(\eta_1) \end{pmatrix}^T \underline{\underline{v}} \begin{pmatrix} \lambda^p_0(\eta_2) \\ \vdots \\ \lambda^p_p(\eta_2) \end{pmatrix} \\ = \sum_{0 \le i_1, i_2 \le q} v^q_{i_1,i_2}\, \lambda^q_{i_1}(\eta_1)\lambda^q_{i_2}(\eta_2) \quad &= \begin{pmatrix} \lambda^q_0(\eta_1) \\ \vdots \\ \lambda^q_q(\eta_1) \end{pmatrix}^T \underline{\underline{v}}^q \begin{pmatrix} \lambda^q_0(\eta_2) \\ \vdots \\ \lambda^q_q(\eta_2) \end{pmatrix} \end{aligned}$$

for all $(\eta_1, \eta_2) \in \tilde{Q}$. Using (B.3) to replace $(\lambda^p_0, \ldots, \lambda^p_p)$ and inserting the Gauss-Lobatto points $(\xi^{p+1}_{i_1}, \xi^{p+1}_{i_2}) \in G_{\tilde{Q},q}$, $0 \le i_1, i_2 \le q$, yields $C^T \underline{\underline{v}} C = \underline{\underline{v}}^q$. This proves the second statement of (B.4). □

Remark B.6. In actual computation matrices and vectors do not need to be rearranged physically. With Lemma B.5 the coordinate representation of v with respect to the basis $\tilde{B}_q$ can be computed from $\underline{v}$ by $\underline{\underline{v}}^q = C \underline{\underline{v}} C^T$. This costs $\mathcal{O}(p^3)$ floating point operations (cf. Table B.1).

Using Remark B.6, we can represent v on Q by

$$v = \sum_{i_1,i_2=0}^{q} v^q_{i_1,i_2} b^q_{i_1,i_2} \circ F_Q^{-1} \quad \text{with } b^q_{i_1,i_2}(\eta) = \lambda^q_{i_1}(\eta_1)\, \lambda^q_{i_2}(\eta_2)\,.$$

This yields the gradient $\nabla^T v = \sum_{i_1,i_2=0}^q v^q_{i_1,i_2} \big(\nabla^T b^q_{i_1,i_2} \circ F_Q^{-1}\big) \cdot DF_Q^{-1}$ with

$$\nabla^T b^q_{i_1,i_2}(\eta) = \begin{pmatrix} \big(\lambda^q_{i_1}\big)'(\eta_1)\, \lambda^q_{i_2}(\eta_2) \\ \lambda^q_{i_1}(\eta_1)\, \big(\lambda^q_{i_2}\big)'(\eta_2) \end{pmatrix}^T.$$

Inserting $\xi_{r_1,r_2} := (\xi^{q+1}_{r_1}, \xi^{q+1}_{r_2}) \in G_{\tilde{Q},q}$, $0 \le r_1, r_2 \le q$, we obtain

$$b^q_{i_1,i_2}(\xi_{r_1,r_2}) = \delta_{i_1,r_1}\delta_{i_2,r_2} \qquad \text{and} \qquad \nabla b^q_{i_1,i_2}(\xi_{r_1,r_2}) = \begin{pmatrix} \big(\lambda^q_{i_1}\big)'(\xi^{q+1}_{r_1})\, \delta_{i_2,r_2} \\ \big(\lambda^q_{i_2}\big)'(\xi^{q+1}_{r_2})\, \delta_{i_1,r_1} \end{pmatrix}^T. \tag{B.5a}$$

With these expressions, it follows that

$$v(F_Q(\xi_{r_1,r_2})) = v^q_{r_1,r_2} \tag{B.5b}$$

$$\text{and} \quad \nabla^T v(F_Q(\xi_{r_1,r_2})) = \begin{pmatrix} \sum_{i_1=0}^q v_{i_1,r_2} \big(\lambda^q_{i_1}\big)'(\xi^{q+1}_{r_1}) \\ \sum_{i_2=0}^q v_{r_1,i_2} \big(\lambda^q_{i_2}\big)'(\xi^{q+1}_{r_2}) \end{pmatrix}^T \big(DF_Q(\xi_{r_1,r_2})\big)^{-1}. \tag{B.5c}$$

We use the equations (B.5) to formulate three algorithms. Algorithm B.1 computes the values $\big(\lambda^q_i\big)'(\xi^{q+1}_r)$ for all $0 \le i,r \le q$. Algorithm B.2 computes the products $\big(\nabla^T v\, \nabla b_{Q,j_1,j_2}\big) \circ F_Q$ for all $0 \le j_1, j_2 \le q$ at the quadrature point ξ_{r_1,r_2}. Using these auxiliary algorithms, Algorithm B.3 realizes the numerical quadrature of the integrands $f_{u;b_k,v} \circ F_Q$ for all k corresponding to the local index triples (Q, j_1, j_2), $0 \le j_1, j_2 \le p$.

Algorithm B.1 Computation of $\underline{\underline{\mu}} := \big(\mu_{i,r}\big)_{0\le i,r\le q}$, $\mu_{i,r} := \big(\lambda^q_i\big)'(\xi^{q+1}_r)$

---

1. For $r = 0, \dots, q$

   (a) Set $f'_{0,r} = 0$, $f'_{1,r} = L_0(\xi^{q+1}_r) = 1$, $f'_{2,r} = L_1(\xi^{q+1}_r) = \xi^{q+1}_r$.
   For $i = 2, \dots q-1$
   Compute $f'_{i+1,r} = L_i(\xi^{q+1}_r) = \frac{1}{i}\big((2i-1)\xi^{q+1}_r L_{i-1}(\xi^{q+1}_r) - (i-1)L_{i-2}(\xi^{q+1}_r)\big)$.
   End for i

   (b) Set $f_{0,r} = \mathcal{L}_0(\xi^{q+1}_r) = 1$, $f_{1,r} = \mathcal{L}_1(\xi^{q+1}_r) = \xi^{q+1}_r$.
   For $i = 2, \dots, q$
   $f_{i,r} = \mathcal{L}_i(\xi^{q+1}_r) = \frac{1}{2i-1}\big(L_i(\xi^{q+1}_r) - L_{i-2}(\xi^{q+1}_r)\big)$
   End for i

   End for r

2. Set $\underline{\underline{\mu}} = \underline{\underline{f}}^{-1}\underline{\underline{f'}}$ where $\underline{\underline{f}} = \big(f_{i,r}\big)_{0\le i,r\le q}$, $\underline{\underline{f'}} = \big(f'_{i,r}\big)_{0\le i,r\le q}$.
   Exit.

Lemma B.7. Algorithm B.1 computes the matrix $\underline{\underline{\mu}} := \big(\mu_{i,r}\big)_{0\le i,r\le q}$, $\mu_{i,r} := \big(\lambda^q_i\big)'(\xi^{q+1}_r)$, at a cost of $\mathcal{O}(q^3)$ flops.

Proof. From the theory of orthogonal polynomials (cf. [Sze75]) we know that the recursive scheme $L_0(t) := 1$, $L_1(t) := t$,

$$L_i(t) := \tfrac{1}{i}\big((2i-1)tL_{i-1}(t) - (i-1)L_{i-2}(t)\big) \qquad \text{for all } i \ge 2,$$

yields polynomials with the orthogonal property

$$\int_{-1}^{1} L_i(t)L_j(t)\,dt = \tfrac{2}{2i+1}\delta_{i,j} \qquad \text{for all } i,j \ge 0\,.$$

Further, it can be shown by induction and elementary integration that the anti-derivatives $\mathcal{L}_i(t) := \int_{-1}^t L_{i-1}(\tau)\,\mathrm{d}\tau$, $i \ge 1$, satisfy the equation

$$\mathcal{L}_i(t) = \tfrac{1}{2i-1}\big(L_i(t) - L_{i-2}(t)\big) \qquad \text{for all } i \ge 2\,.$$

For ease of notation we set $\mathcal{L}_0(t) = 1$ and $L_{-1}(t) := 0$. The algorithm computes the square matrices

$$\underline{\underline{f}} = \big(\mathcal{L}_i(\xi_r^{q+1})\big)_{0\le i,r\le q} \text{ in } \textit{Step 1b)} \qquad \text{and } \underline{\underline{f'}} = \big(L_{i-1}(\xi_r^{q+1})\big)_{0\le i,r\le q} \text{ in } \textit{Step 1a).}$$

Now, let $\lambda_i^q = \sum_{j=0}^q c_{i,j}\mathcal{L}_j$, $0 \le i \le q$, be the coordinate representation of the Lagrangian polynomials with respect to $\mathcal{L}_j$, $0 \le j \le q$. Then, it follows that $\big(\lambda_i^q\big)' = \sum_{j=0}^q c_{i,j}L_{j-1}$. Noting that the coefficients c_{ij} are given by $\underline{\underline{f}}^{-1}$, it follows that the algorithm computes $\underline{\underline{\mu}}$. *Step 1* of the algorithm needs $\mathcal{O}(q^2)$ flops. *Step 2* can be implemented as the solution of the linear system $\underline{\underline{f}}\,\underline{\underline{\mu}} = \underline{\underline{f'}}$. Using *LU* factorization of $\underline{\underline{f}}$ causes an expense of $\mathcal{O}(q^3)$ flops (see Table B.1, `dgetrf`, `dgetrs`). □

Algorithm B.2 Compute $\beta_{j_1,j_2} = (z_1,z_2)\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$.

---

1. Set $(z_1,z_2) = (z_1,z_2)\big(DF_Q\big)^{-T}$.
   Set $\underline{\underline{\beta}} = \big(\beta_{j_1,j_2}\big)_{0\le j_1,j_2\le q} = 0$.
2. For $j_1 = 0,\dots,q$
   Set $\beta_{j_1,r_2} = \beta_{j_1,r_2} + z_1\,\mu_{j_1,r_1}$.
   End for j_1
3. For $j_2 = 0,\dots,q$
   Set $\beta_{r_1,j_2} = \beta_{r_1,j_2} + z_2\,\mu_{j_2,r_2}$.
   End for j_2
   Exit.

Lemma B.8. Algorithm B.2 computes $\beta_{j_1,j_2} = (z_1,z_2)\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$, at a cost of $\mathcal{O}(q)$ flops.

Proof. The statement follows by the right equation of (B.5a) and by noting that the executions of *Step 2* and *Step 3* totals up to $2(q+1)$ multiplications and $2(q+1)$ additions. □

Proposition B.9 (*Costs of a matrix-vector product***).** Algorithm B.3 needs $6(p+\tilde{q}+1)^4 + \mathcal{O}(p^3)$ floating point operations to compute $H(\underline{u})\,\underline{v}$.

Proof. Firstly, we check that the algorithm computes the matrix-vector product. After the evaluation of $\big(\lambda_i^q\big)'$ at the quadrature points ξ_r in *Step 1* and the coordinate transformation of $\underline{u}$ and $\underline{v}$ in *Step 2* (see Remark B.6), $\underline{\underline{s}}^q$ is initialized to 0 for the numerical quadrature in *Step 3*. In *Step 4* the integrands $f_{u;b^q_{Q,j_1,j_2},v} \circ F_Q$ are computed for all points ξ_{r_1,r_2}. Here, $t = |\nabla u|$ is calculated according to (B.5c) in *Step 4d). Steps 4c), 4e),4f)* add the contributions corresponding to $\sigma b_k v$, $t\rho'(t)s_1s_2$ and $\rho(t)\nabla^T b_k \nabla v$ of (B.1), respectively.

With the termination of *Step 4f)* we know the values

$$\gamma_{j_1,j_2}(r_1,r_2) := \big(f_{u;b^q_{Q,j_1,j_2},v} \circ F_Q\big)(\xi_{r_1,r_2}) \qquad \text{for all } 0 \le, j_1, j_2 \le q\,.$$

Using column vector notation, we may write $b_{j_1,j_2} \in \tilde{B}_p$ as the linear combination

$$b_{j_1,j_2} = \vec{n}_{j_1,j_2}\big(b^q_{00},\ldots,b^q_{q,q}\big)^T \quad \text{where } \vec{n}_{j_1,j_2} \text{ denotes the row } (j_1,j_2) \text{ of } N \text{ from (B.2).}$$

Accordingly, writing $\gamma_{j_1,j_2}(r_1,r_2)$ as the column vector

$$\underline{\gamma}(r_1,r_2) := \big(\gamma_{0,0}(r_1,r_2)\,,\ \ldots\ ,\ \gamma_{q,q}(r_1,r_2)\,\big)^T,$$

we obtain

$$\big(f_{u;b_{Q,j_1,j_2},v} \circ F_Q\big)(\xi_{r_1,r_2}) = \vec{n}_{j_1,j_2}\underline{\gamma}(r_1,r_2)$$

and the integral

$$s_{j_1,j_2} = \int_{\tilde{Q}} \big(f_{u;b_{j_1,j_2},v} \circ F_Q\big)(\eta)\ \mathrm{d}\eta = \sum_{r_1,r_2=0}^{q} \vec{n}_{j_1,j_2} w(\xi_{r_1,r_2})\underline{\gamma}(r_1,r_2) = \vec{n}_{j_1,j_2}\underline{s}^q$$

where $\underline{s}^q = \big(s^q_{0,0}\,,\ \ldots\ ,\ s^q_{q,q}\,\big)^T$ is taken after the execution of the double loop *Step 4*. This yields $H(\underline{u})\,\underline{v} = \underline{s} = \big(s_{0,0}\,,\ \ldots\ ,\ s_{p,p}\,\big)^T = N\underline{s}^q$ or equivalently $\underline{\underline{s}} = C\underline{\underline{s}}^q\,C^T$ due to the Lemma B.5, when matrix notation is used.

Secondly, we count the flops. By Remark B.6 and Lemma B.7 we know that *Step 1* and *Step 2*, both cause an expense of $\mathcal{O}(q^3)$ flops. The cost of *Step 4a)* is independent of p due to Assumption B.3 on F_Q, *Step 4c)* costs one flop. The calculations of (z_1,z_2) and β_{j_1,j_2} for all j_1,j_2 in *Step 4d)*, *Step 4e)*, and *Step 4f)* cause an expense of $\mathcal{O}(q)$ flops. The summation processes for γ_{j_1,j_2} and $s^q_{j_1,j_2}$ in *Step 4e)*, *Step 4f)*, and *Step 4g)* cost $3(q+1)^2$ multiplications and $3(q+1)^2$ additions. Summing up the operations of *Step 4*, the complete execution of the double loop costs $6(q+1)^4 + \mathcal{O}(q^3)$ flops. The matrix-matrix multiplications of *Step 5* need $\mathcal{O}(q^3)$ flops. As we assumed $q = p + \tilde{q}$ with a surplus $\tilde{q}$ which is independent of p, it follows that the total expense of the algorithm is $6(p+\tilde{q}+1)^4 + \mathcal{O}(p^3)$ flops. □

Corollary B.10. Inserting the $(p+1)^2$ local unit vectors for $\underline{v}$ in Algorithm B.3 yields the $(p+1)^2$ columns of the local Galerkin matrix $H(\underline{u})$ at a cost of $\mathcal{O}(p^6)$ flops.

Remark B.11. Algorithm B.3 uses basis transformations for the efficient evaluation of the integrand $f_{u;b_k,v}(x)$ defined in (B.1) at the quadrature points. Of course, there is no need for this basis transformation and the integrand $f_{u;b_k,v}(x)$ can be evaluated directly. But this means that Algorithm B.2 must be replaced by a routine which computes $(z_1,z_2)\big(DF_Q\big)^{-T}\nabla b_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1,j_2 \le q$ at a cost of $\mathcal{O}(q^2)$ flops. This increases the $\mathcal{O}(p^4)$ costs of Algorithm B.3 drastically. It is a second advantage of the basis transformation approach that the basis transformations can be performed by highly optimized BLAS routines.

Remark B.12. We can generalize the idea of numerical quadrature by basis transformation to the computation of the whole Galerkin matrix $H(\underline{u})$. This idea is formulated in Algorithm B.4 and Algorithm B.5. As both algorithms are straightforward extensions of Algorithm B.2 and Algorithm B.3, respectively, we omit the details here. Counting the flops needed by Algorithm B.5, we obtain that the calculation with this algorithm needs $\mathcal{O}(p^6)$ flops, i.e., the order of costs is not improved in comparison to the simple approach of Corollary B.10. However, Algorithm B.5 saves basis transformations of the unit vectors to the representation with respect to the basis $\tilde{B}_q$ in *Step 2* and the calculation of

(z_1, z_2) in *Step 4f)* of Algorithm B.3 which both cost $\mathcal{O}(p^3)$ flops. It is a disadvantage of Algorithm B.5 that it needs more main memory for the multidimensional arrays β_{i_1,i_2,j_1,j_2}, γ_{i_1,i_2,j_1,j_2}, $s^{q,q}_{i_1,i_2,j_1,j_2}$, and the linear transformations of *Step 5*.

To finish this appendix, we consider diagonal preconditioning of the linear problem. When all entries of the Galerkin matrix $H(\underline{u})$ are computed diagonal preconditioning of the cg-scheme can be applied cheaply, because the diagonal entries are known in particular. Using only the matrix-vector multiplications in the cg-scheme, a diagonal preconditioning demands the explicit calculation of the diagonal entries

$$\big(H(\underline{u})\,\underline{b}_k\big)_k = D^2A(u; b_k, b_k) = \sum_{Q\in\mathcal{T}} \int_Q f_{u;b_k,b_k}(x)\,\mathrm{d}x$$

where

$$f_{u;b_k,b_k}(x) = \Big(\rho(t)\|\nabla b_k\|_2^2 + t\rho'(t)s^2 + \sigma b_k^2\Big)(x) \qquad \text{with } t := |\nabla u|\,,$$
$$\text{and} \qquad s := \begin{cases} \frac{1}{t}\nabla^T u\,\nabla b_k, & \text{if } t > 0,\\ 0, & \text{if } t = 0. \end{cases}$$

As the integrand $f_{u;b_k,b_k}$ is a specification of $f_{u;b_k,v}$ from (B.1) a straightforward modification of Algorithm B.2 and Algorithm B.3 yields

Proposition B.13 (*Diagonal preconditioning***).** Algorithm B.7 needs $\mathcal{O}(p^4)$ floating point operations to compute the diagonal entries of $H(\underline{u})$.

Proof. As Algorithm B.3 and Algorithm B.7 differ only in *Step 7f)* the proposition follows completely analogously to Proposition B.9. □

Since Algorithm B.3 and Algorithm B.7 are quite similar, they can be merged into a routine which computes the product $\Lambda^{-1}H(\underline{u})\,\underline{v}$ where Λ is the yielded from $H(\underline{u})$ by setting all non-diagonal entries to zero. Thus, a diagonally preconditioned cg-iteration costs $\mathcal{O}(p^4)$ floating point operations.

Algorithm B.3 Compute the matrix-vector product $H(\underline{u})\,\underline{v}$

1. Compute $\underline{\underline{\mu}} := \big(\mu_{i,r}\big)_{0\le i,r\le q}$, $\mu_{i,r} := \big(\lambda_i^q\big)'(\xi_r^{q+1})$ with Algorithm B.1.

2. Compute $\underline{\underline{u}}^q = C^T \underline{\underline{u}}\, C\,, \quad \underline{\underline{v}}^q = C^T \underline{\underline{v}}\, C\,.$

3. Set $\underline{\underline{s}}^q = \big(s^q_{j_1,j_2}\big)_{0\le j_1,j_2\le q} = 0.$

4. For $r_1 = 0,\dots,q$
   For $r_2 = 0,\dots,q$

   (a) Compute $\big(DF_Q\big)^{-1} = \big(DF_Q(\xi^q_{r_1},\xi^q_{r_2})\big)^{-1}.$

   (b) Set $\gamma_{j_1,j_2} = 0$ for all $0 \le j_1, j_2 \le q$.

   (c) Set $\gamma_{r_1,r_2} = \sigma v^q_{r_1,r_2}\,.$

   (d) Compute $(z_1,z_2) = \nabla^T u(\xi_{r_1,r_2})$
   $= \big(\sum_{i_1=0}^q u_{i_1,r_2}\mu_{i_1,r_1}\,,\ \sum_{i_2=0}^q u_{r_1,i_2}\mu_{i_2,r_2}\big)\,\big(DF_Q\big)^{-1}.$
   Compute $t = (z_1^2+z_2^2)^{1/2}, \quad \rho(t), \quad \text{and } \rho'(t)\,.$

   (e) If $t \neq 0$
   Compute $\beta_{j_1,j_2} = (z_1,z_2)\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$ with Algorithm B.2.
   Compute $\alpha = \sum_{i_1,i_2=0}^q v^q_{i_1,i_2}\,\beta_{i_1,i_2}\,.$
   Set $\gamma_{j_1,j_2} = \gamma_{j_1,j_2} + \frac{1}{t}\rho'(t)\alpha\beta_{j_1,j_2}$ for all $0 \le j_1, j_2 \le q$.
   End if

   (f) Compute $(z_1,z_2) = \nabla^T v(\xi_{r_1,r_2})$
   $= \big(\sum_{i_1=0}^q v_{i_1,r_2}\mu_{i_1,r_1}\,,\ \sum_{i_2=0}^q v_{r_1,i_2}\mu_{i_2,r_2}\big)\,\big(DF_Q\big)^{-1}.$
   Compute $\beta_{j_1,j_2} = (z_1,z_2)\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$ with Algorithm B.2.
   Set $\gamma_{j_1,j_2} = \gamma_{j_1,j_2} + \rho(t)\,\beta_{j_1,j_2}$ for all $0 \le j_1, j_2 \le q$.

   (g) Compute $w(\xi_{r_1,r_2}) = |\det DF_Q(\xi_{r_1,r_2})|\,\rho^{q+1}_{r_1}\rho^{q+1}_{r_2}.$
   Set $s^q_{j_1,j_2} = s^q_{j_1,j_2} + \gamma_{j_1,j_2}\cdot w(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$.

   End for r_2
   End for r_1

5. Compute $\underline{\underline{s}} = C\underline{\underline{s}}^q C^T$. Set $H(\underline{u})\,\underline{v} = \underline{s}$.
   Exit.

Algorithm B.4 Compute $\beta_{i_1,i_2,j_1,j_2} = \left(\nabla^T b^q_{i_1,i_2}(\xi_{r_1,r_2})\big(DF_Q\big)^{-1}\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}\right)(\xi_{r_1,r_2})$ for all $0 \le i_1, i_2, j_1, j_2 \le q$.

1. Compute $\tilde{F} = \big(DF_Q\big)^{-1}\big(DF_Q\big)^{-T}$.
   Set $\big(\beta_{i_1,i_2,j_1,j_2}\big)_{0\le i_1,i_2,j_1,j_2\le q} = 0$.
2. For $i_1 = 0, \ldots, q$
   For $i_2 = 0, \ldots, q$
   (a) Compute $(z_1, z_2) = (\mu_{i_1,r_1}, \mu_{i_2,r_2})\tilde{F}$.
   (b) For $j_1 = 0, \ldots, q$
   Set $\beta_{i_1,i_2,j_1,r_2} = \beta_{i_1,i_2,j_1,r_2} + z_1\,\mu_{j_1,r_1}$.
   End for j_1
   (c) For $j_2 = 0, \ldots, q$
   Set $\beta_{i_1,i_2,r_1,j_2} = \beta_{i_1,i_2,r_1,j_2} + z_2\,\mu_{j_2,r_2}$.
   End for j_2
   End for i_2
   End for i_1
   Exit.

Algorithm B.5 Compute the matrix $H(\underline{u})$

1. Compute $\underline{\underline{\mu}} := \left(\mu_{i,r}\right)_{0\le i,r\le q}$, $\mu_{i,r} := \left(\lambda_i^q\right)'(\xi_r^{q+1})$ with Algorithm B.1.

2. Compute $\underline{\underline{u}}^q = C^T \underline{\underline{u}}\, C$.

3. Set $\left(s^{q,q}_{i_1,i_2,j_1,j_2}\right)_{0\le i_1,i_2,j_1,j_2\le q} = 0$.

4. For $r_1 = 0,\dots,q$
   For $r_2 = 0,\dots,q$

   (a) Compute $DF_Q = DF_Q(\xi^q_{r_1},\xi^q_{r_2})$.

   (b) Set $\gamma_{i_1,i_2,j_1,j_2} = 0$ for all $0 \le i_1,i_2,j_1,j_2 \le q$.

   (c) Set $\gamma_{r_1,r_2,r_1,r_2} = \sigma$.

   (d) Compute $(z_1,z_2) = \nabla^T u(\xi_{r_1,r_2})$
   $= \left(\sum_{i_1=0}^q u_{i_1,r_2}\mu_{i_1,r_1}\,,\ \sum_{i_2=0}^q u_{r_1,i_2}\mu_{i_2,r_2}\right)\left(DF_Q\right)^{-1}$.
   Compute $t = (z_1^2+z_2^2)^{1/2}$, $\rho(t)$, and $\rho'(t)$.

   (e) If $t \ne 0$
   Compute $\alpha_{j_1,j_2} = (z_1,z_2)\left(DF_Q\right)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1,j_2 \le q$ with Algorithm B.2.
   Set $\gamma_{i_1,i_2,j_1,j_2} = \gamma_{i_1,i_2,j_1,j_2} + \frac{1}{t}\rho'(t)\alpha_{i_1,i_2}\alpha_{j_1,j_2}$ for all $0 \le i_1,i_2,j_1,j_2 \le q$.
   End if

   (f) Compute $\beta_{i_1,i_2,j_1,j_2} = \nabla^T b^q_{i_1,i_2}(\xi_{r_1,r_2})\left(DF_Q\right)^{-1}\left(DF_Q\right)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le i_1,i_2,j_1,j_2 \le q$ with Algorithm B.4.
   Set $\gamma_{i_1,i_2,j_1,j_2} = \gamma_{i_1,i_2,j_1,j_2} + \rho(t)\,\beta_{i_1,i_2,j_1,j_2}$ for all $0 \le i_1,i_2,j_1,j_2 \le q$.

   (g) Compute $w(\xi_{r_1,r_2}) = |\det DF_Q(\xi_{r_1,r_2})|\,\rho_{r_1}^{q+1}\rho_{r_2}^{q+1}$.
   Set $s^{q,q}_{i_1,i_2,j_1,j_2} = s^{q,q}_{i_1,i_2,j_1,j_2} + \gamma_{i_1,i_2,j_1,j_2}\cdot w(\xi_{r_1,r_2})$ for all $0 \le i_1,i_2,j_1,j_2 \le q$.

   End for r_2
   End for r_1

5. (a) For $i_1 = 0,\dots,q$
   For $i_2 = 0,\dots,q$
   Set $\underline{\underline{s}}^{q,q}_{i_1,i_2} = \left(s^{q,q}_{i_1,i_2,j_1,j_2}\right)_{0\le j_1,j_2\le q}$.
   Compute $\underline{\underline{s}}^{q}_{i_1,i_2} = C\,\underline{\underline{s}}^{q,q}_{i_1,i_2}\,C^T$.
   End for i_2
   End for i_1

   (b) For $j_1 = 0,\dots,p$
   For $j_2 = 0,\dots,p$
   Set $\underline{\underline{s}}^{q}_{j_1,j_2} = \left(s^{q}_{i_1,i_2,j_1,j_2}\right)_{0\le i_1,i_2\le q}$.
   Compute $\underline{\underline{s}}_{j_1,j_2} = C\,\underline{\underline{s}}^{q}_{j_1,j_2}\,C^T$.
   End for j_2
   End for j_1

   (c) Set $H(\underline{u}) = \left(s_{i_1,i_2,j_1,j_2}\right)_{\substack{0\le i_1,i_2\le p\\ 0\le j_1,j_2\le p}}$.
   Exit.

Algorithm B.6 Compute $\tilde{\beta}_{j_1,j_2} = \|\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})\|_2^2$ for all $0 \le j_1, j_2 \le q$.

1. Set $\begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix} = \big(DF_Q\big)^{-T}$.

   Set $\underline{\underline{\alpha}}_x = \big(\alpha_{x,j_1,j_2}\big)_{0\le j_1,j_2\le q} = 0, \quad \underline{\underline{\alpha}}_y = \big(\alpha_{y,j_1,j_2}\big)_{0\le j_1,j_2\le q} = 0$

2. For $j_1 = 0, \ldots, q$
   Set $\alpha_{x,j_1,r_2} = \alpha_{x,j_1,r_2} + f_{xx}\,\mu_{j_1,r_1}$.
   Set $\alpha_{y,j_1,r_2} = \alpha_{y,j_1,r_2} + f_{yx}\,\mu_{j_1,r_1}$.
   End for j_1

3. For $j_2 = 0, \ldots, q$
   Set $\alpha_{x,r_1,j_2} = \alpha_{x,r_1,j_2} + f_{xy}\,\mu_{j_2,r_2}$.
   Set $\alpha_{y,r_1,j_2} = \alpha_{y,r_1,j_2} + f_{yy}\,\mu_{j_2,r_2}$.
   End for j_2

4. For $j_1 = 0, \ldots, q$
   For $j_2 = 0, \ldots, q$
   Set $\tilde{\beta}_{j_1,j_2} = \alpha^2_{x,j_1,j_2} + \alpha^2_{y,j_1,j_2}$.
   End for j_2
   End for j_1
   Exit.

Algorithm B.7 Compute the diagonal entries $H_{i1,i2;i1,i2}$ of $H(\underline{u})$

1. Compute $\underline{\underline{\mu}} := \big(\mu_{i,r}\big)_{0\le i,r\le q}$, $\mu_{i,r} := \big(\lambda_i^q\big)'(\xi_r^{q+1})$ with Algorithm B.1.
2. Compute $\underline{\underline{u}}^q = C^T \underline{\underline{u}}\, C$.
3. Set $\underline{\underline{d}}^q = \big(d^q_{j_1,j_2}\big)_{0\le j_1,j_2\le q} = 0$.
4. For $r_1 = 0,\dots,q$
   For $r_2 = 0,\dots,q$
   (a) Compute $\big(DF_Q\big)^{-1} = \big(DF_Q(\xi^q_{r_1},\xi^q_{r_2})\big)^{-1}$.
   (b) Set $\tilde\gamma_{j_1,j_2} = 0$ for all $0 \le j_1, j_2 \le q$.
   (c) Set $\tilde\gamma_{r_1,r_2} = \sigma$.
   (d) Compute $(z_1,z_2) = \nabla^T u(\xi_{r_1,r_2})$
   $= \big(\sum_{i_1=0}^q u_{i_1,r_2}\mu_{i_1,r_1}\,,\ \sum_{i_2=0}^q u_{r_1,i_2}\mu_{i_2,r_2}\big)\,\big(DF_Q\big)^{-1}$.
   Compute $t = (z_1^2+z_2^2)^{1/2}$, $\rho(t)$, and $\rho'(t)$.
   (e) If $t \ne 0$
   Compute $\tilde\beta_{j_1,j_2} = (z_1,z_2)\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$ with Algorithm B.2.
   Set $\tilde\gamma_{j_1,j_2} = \tilde\gamma_{j_1,j_2} + \frac{1}{t}\rho'(t)\tilde\beta^2_{j_1,j_2}$ for all $0 \le j_1, j_2 \le q$.
   End if
   (f) Compute $\tilde\beta_{j_1,j_2} = \|\big(DF_Q\big)^{-T}\nabla b^q_{j_1,j_2}(\xi_{r_1,r_2})\|_2^2$ for all $0 \le j_1, j_2 \le q$ with Algorithm B.6.
   Set $\tilde\gamma_{j_1,j_2} = \tilde\gamma_{j_1,j_2} + \rho(t)\,\tilde\beta_{j_1,j_2}$ for all $0 \le j_1, j_2 \le q$.
   (g) Compute $w(\xi_{r_1,r_2}) = |\det DF_Q(\xi_{r_1,r_2})|\,\rho_{r_1}^{q+1}\rho_{r_2}^{q+1}$.
   Set $d^q_{j_1,j_2} = d^q_{j_1,j_2} + \tilde\gamma_{j_1,j_2}\cdot w(\xi_{r_1,r_2})$ for all $0 \le j_1, j_2 \le q$.
   End for r_2
   End for r_1
5. Compute $\underline{\underline{d}} = C\underline{\underline{d}}^q C^T$. Set $H_{i_1,i_2;i_1,i_2} = d_{i_1,i2}$ for all $0 \le i_1, i_2 \le p$.
   Exit.

Appendix C

Conditions numbers from Experiments 4.34, 4.36, 4.37

p	cond(H)	α	cond(H_{II})	α_{II}	cond(H^c)	α^c
2	12.94	—	1.00	—	5.22	—
3	23.39	—	2.14	—	7.60	—
4	40.68	1.65	4.28	2.10	9.87	0.92
5	61.49	1.89	7.09	2.34	12.36	0.95
6	91.60	2.00	10.79	2.28	15.41	1.10
7	131.73	2.26	15.94	2.41	18.44	1.19
8	179.90	2.35	22.59	2.57	21.48	1.15
9	238.09	2.36	30.50	2.58	24.51	1.13
10	307.03	2.40	39.74	2.53	27.55	1.12
11	390.36	2.46	50.46	2.51	30.59	1.10
12	487.95	2.54	62.84	2.51	33.64	1.10
13	603.66	2.61	77.14	2.54	36.70	1.09
14	737.08	2.68	93.64	2.59	39.76	1.08
15	888.59	2.70	112.62	2.64	42.83	1.08
16	1061.55	2.73	134.24	2.70	45.91	1.08
17	1254.26	2.75	158.77	2.74	48.99	1l.07
18	1471.30	2.77	186.32	2.78	52.08	1.07
19	1710.32	2.79	216.91	2.81	55.18	1.07
20	1976.28	2.80	250.84	2.82	58.29	1.07
21	2266.65	2.81	288.12	2.84	61.40	1.07
22	2586.50	2.82	329.08	2.85	64.52	1.07
23	2933.22	2.83	373.72	2.86	67.65	1.07
24	3311.93	2.84	422.38	2.87	70.78	1.06
25	3720.01	2.85	475.04	2.88	73.92	1.06
26	4162.59	2.86	532.07	2.88	77.07	1.06
27	4637.03	2.86	593.41	2.89	80.23	1.06
28	5148.47	2.87	659.48	2.90	83.39	1.06
29	5694.28	2.87	730.18	2.90	86.55	1.06
30	6279.59	2.88	805.94	2.91	89.73	1.06

Table C.1: Condition numbers of H, H_{II}, H^c.

p	cond($\tilde{H}$)	$\tilde{\alpha}$	cond($\tilde{H}_{II}$)	$\tilde{\alpha}_{II}$	cond($\tilde{H}^c$)	$\tilde{\alpha}^c$
2	2.51	—	1.00	—	2.29	—
3	8.53	—	2.14	—	4.54	—
4	20.48	3.03	4.29	2.10	6.80	1.57
5	38.47	2.95	7.14	2.35	9.08	1.36
6	63.06	2.77	10.79	2.27	11.76	1.35
7	94.77	2.68	15.21	2.25	14.55	1.40
8	134.09	2.62	21.00	2.32	17.47	1.38
9	181.51	2.59	28.00	2.43	20.43	1.35
10	237.43	2.56	36.00	2.42	23.44	1.32
11	302.28	2.54	45.00	2.36	26.49	1.29
12	376.44	2.53	55.00	2.32	29.57	1.27
13	460.26	2.52	66.00	2.29	32.68	1.26
14	554.07	2.51	78.00	2.27	35.81	1.24
15	658.19	2.50	91.00	2.24	38.96	1.23
16	772.87	2.49	105.00	2.23	42.12	1.22
17	898.39	2.49	120.00	2.21	45.31	1.21
18	1034.97	2.48	136.00	2.20	48.50	1.20
19	1182.82	2.47	153.00	2.18	51.70	1.19
20	1342.13	2.47	171.00	2.17	54.91	1.18
21	1513.09	2.46	190.00	2.16	58.12	1.17
22	1695.86	2.45	210.00	2.16	61.34	1.16
23	1890.59	2.45	231.00	2.15	64.57	1.16
24	2097.43	2.44	253.00	2.14	67.80	1.15
25	2316.50	2.44	276.00	2.13	71.04	1.14
26	2547.95	2.43	300.00	2.13	74.27	1.14
27	2791.89	2.43	325.00	2.12	77.51	1.13
28	3048.44	2.42	351.00	2.12	80.75	1.13
29	3317.71	2.41	378.00	2.11	84.00	1.12
30	3599.80	2.41	406.00	2.11	87.24	1.12

Table C.2: Condition numbers of the diagonally preconditioned matrices $\tilde{H}$, $\tilde{H}_{II}$, $\tilde{H}^c$.

p	H^{eq}	b^{eq}	$\tilde{H}^{\mathrm{eq}}$	$\tilde{b}^{\mathrm{eq}}$
2	13	—	2.5	—
3	43	—	9.6	—
4	1.4e+02	1.21	39	1.37
5	5.5e+02	1.27	1.7e+02	1.44
6	2.5e+03	1.41	7.8e+02	1.49
7	1.3e+04	1.60	4e+03	1.58
8	8.7e+04	1.79	2.4e+04	1.71
9	6.7e+05	1.96	1.7e+05	1.87
10	5.8e+06	2.10	1.4e+06	2.03
11	5.6e+07	2.21	1.3e+07	2.18
12	5.7e+08	2.29	1.3e+08	2.29
13	6.2e+09	2.35	1.5e+09	2.36
14	6.9e+10	2.40	1.7e+10	2.41
15	8e+11	2.43	1.9e+11	2.45
16	9.5e+12	2.46	2.3e+12	2.47
17	1.2e+14	2.49	2.9e+13	2.50
18	1.5e+15	2.52	3.6e+14	2.52
19	2.5e+16	2.68	4.8e+15	2.56
20	3.8e+17	2.78	7.1e+16	2.64

Table C.3: Condition numbers of H^{eq}, $\tilde{H}^{\mathrm{eq}}$

p	cond($H^{\mathcal{L}}$)	$\alpha^{\mathcal{L}}$	cond($H_{II}^{\mathcal{L}}$)	$\alpha_{II}^{\mathcal{L}}$	cond($H^{c,\mathcal{L}}$)	$\alpha^{c,\mathcal{L}}$
2	22.26	—	1.00	—	8.76	—
3	36.63	—	4.20	—	12.66	—
4	103.73	2.22	10.35	3.37	21.79	1.32
5	159.29	2.88	20.31	3.09	28.54	1.59
6	321.16	2.79	35.56	3.04	40.68	1.54
7	462.53	3.17	57.81	3.11	50.27	1.68
8	783.61	3.10	89.05	3.19	65.38	1.65
9	1072.30	3.35	131.52	3.27	77.84	1.74
10	1633.70	3.29	187.71	3.34	95.90	1.72
11	2148.02	3.46	260.31	3.40	111.23	1.78
12	3047.60	3.42	352.31	3.45	132.22	1.76
13	3882.61	3.54	466.90	3.50	150.44	1.81
14	5235.00	3.51	607.52	3.53	174.35	1.79
15	6502.55	3.60	777.89	3.57	195.47	1.83
16	8439.16	3.58	981.94	3.60	222.30	1.82
17	10267.85	3.65	1223.87	3.62	246.32	1.85
18	12936.86	3.63	1508.10	3.64	276.06	1.84
19	15472.08	3.69	1839.31	3.66	302.98	1.86
20	19038.45	3.67	2222.45	3.68	335.63	1.85
21	22442.38	3.72	2662.67	3.70	365.45	1.87
22	27087.84	3.70	3165.41	3.71	401.02	1.87
23	31539.39	3.74	3736.32	3.72	433.73	1.88
24	37462.45	3.73	4381.32	3.74	472.21	1.88
25	43157.36	3.76	5106.58	3.75	507.83	1.89
26	50573.30	3.75	5918.49	3.76	549.22	1.89
27	57724.04	3.78	6823.71	3.77	587.74	1.90
28	66864.93	3.77	7829.13	3.78	632.04	1.90
29	75700.77	3.79	8941.91	3.78	673.46	1.91
30	86815.43	3.78	10169.44	3.79	720.67	1.90

Table C.4: Condition numbers of $H^{\mathcal{L}}$, $H_{II}^{\mathcal{L}}$, $H^{c,\mathcal{L}}$.

p	$\mathrm{cond}(\tilde{H}^{\mathcal{L}})$	$\tilde{\alpha}^{\mathcal{L}}$	$\mathrm{cond}(\tilde{H}^{\mathcal{L}}_{II})$	$\tilde{\alpha}^{\mathcal{L}}_{II}$	$\mathrm{cond}(\tilde{H}^{c,\mathcal{L}})$	$\tilde{\alpha}^{c,\mathcal{L}}$
2	22.08	—	1.00	—	8.69	—
3	23.86	—	1.00	—	8.78	—
4	69.56	1.66	1.97	0.98	15.57	0.84
5	70.34	2.12	2.50	1.79	15.54	1.12
6	141.81	1.76	3.59	1.47	21.95	0.85
7	142.41	2.10	4.53	1.77	21.94	1.03
8	241.22	1.85	5.83	1.69	28.07	0.85
9	241.74	2.11	7.11	1.79	28.06	0.98
10	369.14	1.91	8.66	1.77	34.02	0.86
11	369.58	2.12	10.24	1.82	34.02	0.96
12	526.55	1.95	12.05	1.81	39.85	0.87
13	526.93	2.12	13.92	1.84	39.85	0.95
14	714.23	1.98	15.98	1.83	45.60	0.87
15	714.56	2.13	18.14	1.85	45.60	0.94
16	932.88	2.00	20.47	1.85	51.28	0.88
17	933.16	2.13	22.90	1.86	51.28	0.94
18	1183.08	2.02	25.51	1.87	56.91	0.88
19	1183.33	2.14	28.21	1.88	56.91	0.94
20	1465.38	2.03	31.09	1.88	62.51	0.89
21	1465.60	2.14	34.07	1.88	62.51	0.94
22	1780.25	2.04	37.21	1.89	68.07	0.89
23	1780.45	2.14	40.47	1.89	68.07	0.94
24	2128.12	2.05	43.88	1.90	73.61	0.90
25	2128.30	2.14	47.42	1.90	73.61	0.94
26	2509.39	2.06	51.10	1.90	79.13	0.90
27	2509.56	2.14	54.91	1.91	79.13	0.94
28	2924.43	2.07	58.87	1.91	84.62	0.91
29	2924.59	2.14	62.95	1.91	84.62	0.94
30	3373.59	2.07	67.17	1.91	90.11	0.91

Table C.5: Condition numbers of the diagonally preconditioned matrices $\tilde{H}^{\mathcal{L}}$, $\tilde{H}^{\mathcal{L}}_{II}$, $\tilde{H}^{c,\mathcal{L}}$.

Appendix D

Proof of Lemma 5.2

Lemma 5.2 For any $w, u, v \in H^1(\Omega)$, $\Omega \subset \mathbb{R}^2$ a bounded Lipschitz domain, there holds

$$DA(w; v) = DA(u; v) + D^2A(u; w - u, v) + R(u; w - u, v) \tag{D.1}$$

with the estimate for the remainder R

$$|R(u; w - u, v)| \leq \sqrt{2}(5\rho_4 + \rho_5) \, \|\nabla(w - u)\|^2_{L^4(\Omega)} \, |v|_{H^1(\Omega)} \, .$$

Proof. Let $e := w - u$ and $u(t) = u + te$. We consider

$$\begin{aligned} \eta(t) &= DA(u(t); v) \\ &= \int_\Omega \Big(\rho(|\nabla u(t)|)(\nabla u(t))^T \nabla v + \sigma \, u(t) \, v - fv \Big) \, \mathrm{d}x - \int_{\Gamma_N} g_N v_{|\Gamma} \quad \mathrm{d}s \, . \end{aligned} \tag{D.2}$$

By Taylor's formula we have that

$$\eta(1) = \eta(0) + \eta'(0) + \int_0^1 \eta''(t)(1 - t) \, \mathrm{d}t \, . \tag{D.3}$$

If the first and second derivative of the integrands of (D.2) are continuous and bounded for all $x \in \Omega$, we can write

$$\eta'(t) = \int_\Omega \tfrac{d}{dt} \big(\rho(|\nabla u(t)|)(\nabla u(t))^T \big) \, \nabla v \, \mathrm{d}x \; + \int_\Omega \tfrac{d}{dt} \big(\sigma u(t) \big) \, v \, \mathrm{d}x \tag{D.4a}$$

$$\eta''(t) = \int_\Omega \tfrac{d^2}{dt^2} \big(\rho(|\nabla u(t)|)(\nabla u(t))^T \big) \, \nabla v \, \mathrm{d}x + \int_\Omega \tfrac{d^2}{dt^2} \big(\sigma u(t) \big) \, v \, \mathrm{d}x \tag{D.4b}$$

due to Leibniz rule. Differentiating straightforwardly

$$\begin{aligned} \tfrac{d}{dt} \big(\sigma \, u(t) \big) &= \sigma e \, , & \tfrac{d^2}{dt^2} \big(\sigma \, u(t) \big) &= 0 \, , \\ \tfrac{d}{dt} (\nabla u(t)) &= \nabla e \, , & \tfrac{d}{dt} (|\nabla u(t)|) &= \frac{\big(\nabla u(t) \big)^T \nabla e}{|\nabla u(t)|} \, , \end{aligned}$$

$$\tfrac{d}{dt} \left(\frac{\nabla u(t)}{|\nabla u(t)|} \right) = \frac{1}{|\nabla u(t)|} \nabla e - \frac{\big(\nabla u(t) \big)^T \nabla e}{|\nabla u(t)|^3} \, \nabla u(t) \, ,$$

$$\tfrac{d^2}{dt^2} (|\nabla u(t)|) = \tfrac{d}{dt} \left(\frac{\big(\nabla u(t) \big)^T \nabla e}{|\nabla u(t)|} \right) = \frac{1}{|\nabla u(t)|^3} \Big(|\nabla u(t)|^2 |\nabla e|^2 - \big(\big(\nabla u(t) \big)^T \nabla e \big)^2 \Big) \, ,$$

we obtain the derivatives

$$\tfrac{d}{dt}\big(\rho(|\nabla u(t)|)\big(\nabla u(t)\big)^T\big) = \rho(|\nabla u(t)|)\big(\nabla e\big)^T + \rho'(|\nabla u(t)|)\frac{\big(\nabla u(t)\big)^T\nabla e}{|\nabla u(t)|}\big(\nabla u(t)\big)^T$$

and

$$\begin{aligned}\tfrac{d^2}{dt^2}\big(\rho(|\nabla u(t)|)\big(\big(\nabla u(t)\big)^T\big) = 2\rho'(|\nabla u(t)|)\frac{\big(\nabla u(t)\big)^T\nabla e}{|\nabla u(t)|}\big(\nabla e\big)^T + \bigg(\frac{\rho'(|\nabla u(t)|)}{|\nabla u(t)|^3}\big(|\nabla u(t)|^2|\nabla e|^2 \\ - \big(\big(\nabla u(t)\big)^T\nabla e\big)^2\big) + \rho''(|\nabla u(t)|)\frac{\big(\big(\nabla u(t)\big)^T\nabla e\big)^2}{|\nabla u(t)|^2}\bigg)\big(\nabla u(t)\big)^T.\end{aligned}$$

Using the Cauchy-Schwarz inequality $|\big(\nabla u(t)\big)^T\nabla e| \le |\nabla u(t)|\,|\nabla e|$ yields

$$\begin{aligned}\big|\tfrac{d^2}{dt^2}\big(\rho(|\nabla u(t)|)\nabla u(t)\big)\big| &\le \big(4\,|\rho'(|\nabla u(t)|)| + |\rho''(|\nabla u(t)|)|\,|\nabla u(t)|\big)\,|\nabla e|^2 \\ &\le (5\rho_4+\rho_5)\,|\nabla e|^2\end{aligned}$$

where ρ_4, ρ_5 are the positive constants from Lemma 1.21. Applying Cauchy-Schwarz to (D.4b), we estimate

$$\max\big\{|\eta''(t)|\,\big|\,t\in[0,1]\big\} \le \sqrt{2}(5\rho_4+\rho_5)\,\|\nabla e\|^2_{L^4(\Omega)}\,\|\nabla v\|_{L^2(\Omega)}\,. \tag{D.5}$$

Defining the remainder R by

$$R(u;e,v) := \int_0^1 \eta''(t)(1-t)\;\mathrm{d}t\,,$$

we obtain (D.1) by rewriting (D.3) and the estimate for $|R|$ from (D.5). □

Bibliography

[ABB+95] E. Anderson, Z. Bai, C. Bischof, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, S. Ostrouchov, and D. Sorensen. *LAPACK user's guide. 2nd ed.* Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1995.

[Ada75] Robert A. Adams. *Sobolev spaces.* Academic Press [A subsidiary of Harcourt Brace Jovanovich, Publishers], New York-London, 1975. Pure and Applied Mathematics, Vol. 65.

[Ain96] Mark Ainsworth. A preconditioner based on domain decomposition for *hp*-finite-element approximation on quasi-uniform meshes. *SIAM J. Numer. Anal.*, 33(4):1358–1376, 1996.

[Alt91] Hans Wilhelm Alt. *Lineare Funktionalanalysis. Eine anwendungsorientierte Einführung. (Linear functional analysis. An application oriented introduction). 2. verbesserte Auflage.* Springer-Lehrbuch. Berlin: Springer, 1991.

[AMT99] Mark Ainsworth, William McLean, and Thanh Tran. The conditioning of boundary element equations on locally refined meshes and preconditioning by diagonal scaling. *SIAM J. Numer. Anal.*, 36(6):1901–1932, 1999.

[AO00] Mark Ainsworth and J.Tinsley Oden. *A posteriori error estimation in finite element analysis.* Pure and Applied Mathematics. A Wiley-Interscience Series of Texts, Monographs, and Tracts. Chichester: Wiley, 2000.

[AOL93] Mark Ainsworth, J.Tinsley Oden, and C.Y. Lee. Local a posteriori error estimators for variational inequalities. *Numer. Methods Partial Differ. Equations*, 9(1):23–33, 1993.

[AS97] Mark Ainsworth and Bill Senior. Aspects of an adaptive *hp*-finite element method: Adaptive strategy, conforming approximation and efficient solvers. *Comput. Methods Appl. Mech. Eng.*, 150(1-4):65–87, 1997.

[AS98] Mark Ainsworth and Bill Senior. An adaptive refinement strategy for *hp*-finite element computations. In *Proceedings of the International Centre for Mathematical Sciences Conference on Grid Adaptation in Computational PDEs: Theory and Applications (Edinburgh, 1996)*, volume 26, pages 165–178, 1998.

[BA72] Ivo Babuška and A. K. Aziz. Survey lectures on the mathematical foundations of the finite element method. In *The mathematical foundations of the finite*

element method with applications to partial differential equations (Proc. Sympos., Univ. Maryland, Baltimore, Md., 1972), pages 1–359. Academic Press, New York, 1972. With the collaboration of G. Fix and R. B. Kellogg.

[BCMP91] I. Babuška, A. Craig, J. Mandel, and J. Pitkäranta. Efficient preconditioning for the p-version finite element method in two dimensions. *SIAM J. Numer. Anal.*, 28(3):624–661, 1991.

[Ber96] Christine Bernardi. Indicateurs d'erreur en h-N version des éléments spectraux. *RAIRO Modél. Math. Anal. Numér.*, 30(1):1–38, 1996.

[BG88] I. Babuška and B. Q. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. I. Boundary value problems for linear elliptic equation of second order. *SIAM J. Math. Anal.*, 19(1):172–203, 1988.

[BG89] I. Babuška and B. Q. Guo. Regularity of the solution of elliptic problems with piecewise analytic data. II. The trace spaces and application to the boundary value problems with nonhomogeneous boundary conditions. *SIAM J. Math. Anal.*, 20(4):763–781, 1989.

[BHN99] Richard H. Byrd, Mary E. Hribar, and Jorge Nocedal. An interior point algorithm for large-scale nonlinear programming. *SIAM J. Optim.*, 9(4):877–900, 1999.

[BM92] Christine Bernardi and Yvon Maday. Polynomial interpolation results in Sobolev spaces. *J. Comput. Appl. Math.*, 43(1-2):53–80, 1992.

[BM97] Christine Bernardi and Yvon Maday. Spectral methods. In *Handbook of numerical analysis, Vol. V*, Handb. Numer. Anal., V, pages 209–485. North-Holland, Amsterdam, 1997.

[Boy90] Charles. V. Boys. *Soap Bubbles and the Forces which Mould Them;* enlarged edition 1959: *Soap Bubbles, their Colours and the Forces which Mould Them,* Dover, New York, 1959. Society for the Promotion of Christian Knowledge, London, 1890.

[BS89] Randolph E. Bank and L.Ridgway Scott. On the conditioning of finite element equations with highly refined meshes. *SIAM J. Numer. Anal.*, 26(6):1383–1394, 1989.

[BS00] H. Blum and F.-T. Suttmeier. Weighted error estimates for finite element solutions of variational inequalities. *Computing*, 65(2):119–134, 2000.

[CGO76] Paul Concus, Gene H. Golub, and Dianne P. O'Leary. A generalized conjugate gradient method for the numerical solution of elliptic partial differential equations. In *Sparse matrix computations (Proc. Sympos., Argonne Nat. Lab., Lemont, Ill., 1975)*, pages 309–332. Academic Press, New York, 1976.

[CGT92] Andrew R. Conn, Nick I.M. Gould, and Philippe L. Toint. *LANCELOT. A Fortran package for large-scale nonlinear optimization (Release A).* Springer Series in Computational Mathematics. 17. Berlin etc.: Springer- Verlag, 1992.

[CGT96] A. R. Conn, Nick Gould, and Ph. L. Toint. Numerical experiments with the LANCELOT package (Release A) for large-scale nonlinear optimization. *Math. Programming*, 73(1, Ser. A):73–110, 1996.

[CGT02] Andrew R. Conn, Nick I.M. Gould, and Philippe L. Toint. GALAHAD, a library of thread-safe Fortran 90 packages for large-scale nonlinear optimization. *Preprint RAL-TR-2002-014,* `http://www.numerical.rl.ac.uk/reports/reports.html`*,* visited on 07/2002*, Computational Science and Engineering Department, Rutherford Appleton Laboratory*, 2002.

[CL96] Thomas F. Coleman and Yuying Li. A reflective Newton method for minimizing a quadratic function subject to bounds on some of the variables. *SIAM J. Optim.*, 6(4):1040–1058, 1996.

[CM87] Paul H. Calamai and Jorge J. Moré. Projected gradient methods for linearly constrained problems. *Math. Program.*, 39:93–116, 1987.

[CN00] Zhiming Chen and Ricardo H. Nochetto. Residual type a posteriori error estimates for elliptic obstacle problems. *Numer. Math.*, 84(4):527–548, 2000.

[CP89] Edmund Christiansen and Henrik Gordon Petersen. Estimation of convergence orders in repeated Richardson extrapolation. *BIT*, 29(1):48–59, 1989.

[DL93] Ronald A. DeVore and George G. Lorentz. *Constructive approximation.* Grundlehren der Mathematischen Wissenschaften. 303. Berlin: Springer-Verlag, 1993.

[DRD02] L. Demkowicz, W. Rachowicz, and Ph. Devloo. A fully automatic *hp*-adaptivity. *J. Sci. Comput.*, 17(1-4):117–142, 2002.

[Dub91] Moshe Dubiner. Spectral methods on triangles and other domains. *J. Sci. Comput.*, 6(4):345–390, 1991.

[Emm96] Michele Emmer. Architecture and mathematics: soap bubbles and soap films. In *Nexus: architecture and mathematics (Fucecchio, 1996)*, volume 2 of *Collana "Gli Studi"*, pages 53–65. Erba, Fucecchio, 1996.

[EW96] Stanley C. Eisenstat and Homer F. Walker. Choosing the forcing terms in an inexact Newton method. *SIAM J. Sci. Comput.*, 17(1):16–32, 1996.

[Fal74] Richard S. Falk. Error estimates for the approximation of a class of variational inequalities. *Math. Comput.*, 28:963–971, 1974.

[Fel99] Rolf Felkel. *On solving large scale nonlinear programming problems using iterative methods.* PhD thesis, Aachen: Shaker Verlag. Darmstadt: TU Darmstadt, Fachbereich Mathematik, 1999.

[Fic64] Gaetano Fichera. Problemi elastostatici con vincoli unilaterali: Il problema di Signorini con ambigue condizioni al contorno. *Atti Accad. Naz. Lincei Mem. Cl. Sci. Fis. Mat. Natur. Sez. I (8)*, 7:91–140, 1963/1964.

[FL02] Roger Fletcher and Sven Leyffer. Nonlinear programming without a penalty function. *Math. Program.*, 91(2, Ser. A):239–269, 2002.

[Glo84] Roland Glowinski. *Numerical methods for nonlinear variational problems.* Springer Series in Computational Physics. New York etc.: Springer-Verlag, 1984.

[GLT81] Roland Glowinski, Jacques-Louis Lions, and Raymond Tremolieres. *Numerical analysis of variational inequalities. Transl. and rev. ed.* Studies in Mathematics and its Applications, Vol. 8. Amsterdam, New York, Oxford: North-Holland Publishing Company, 1981.

[GMS02] Philip E. Gill, Walter Murray, and Michael A. Saunders. SNOPT: An SQP algorithm for large-scale constrained optimization. *SIAM J. Optim.*, 12(4):979–1006, 2002.

[Gri85] P. Grisvard. *Elliptic problems in nonsmooth domains.* Monographs and Studies in Mathematics, 24. Pitman Advanced Publishing Program. Boston-London-Melbourne: Pitman Publishing Inc., 1985.

[GS93] J. Gwinner and E.P. Stephan. Boundary element convergence for a variational inequality of the second kind. In *Guddat, Jürgen (ed.) et al., Parametric optimization and related topics. III. Proceedings of the 3rd conference held in Güstrow, Germany, August 30 - September 5, 1991*, Approximation Optimization. 3, pages 227–241. Peter Lang Verlag, Frankfurt am Main, 1993.

[GVL96] Gene H. Golub and Charles F. Van Loan. *Matrix computations.* Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[Hes98] J.S. Hesthaven. From electrostatics to almost optimal nodal sets for polynomial interpolation in a simplex. *SIAM J. Numer. Anal.*, 35(2):655–676, 1998.

[HHNL88] I. Hlaváček, J. Haslinger, J. Nečas, and J. Lovíšek. *Solution of variational inequalities in mechanics.* Applied Mathematical Sciences, 66. New York etc.: Springer-Verlag, 1988.

[HR03] V. Heuveline and R. Rannacher. Duality-based adaptivity in the *hp*-finite element method. *Preprint, Institut für Angewandte Mathematik, Universität Heidelberg*, 2003.

[HS96] N. Heuer and E.P. Stephan. The *hp*-version of the boundary element method on polygons. *J. Integral Equations Appl.*, 8(2):173–212, 1996.

[HS98] Norbert Heuer and Ernst P. Stephan. Boundary integral operators in countably normed spaces. *Math. Nachr.*, 191:123–151, 1998.

[Kel95] C.T. Kelley. *Iterative methods for linear and nonlinear equations.* Frontiers in Applied Mathematics. 16. Philadelphia, PA: SIAM, Society for Industrial and Applied Mathematics, 1995.

[Kre98] Rainer Kress. *Numerical analysis.* Graduate Texts in Mathematics. 181. New York, NY: Springer, 1998.

[KS80] David Kinderlehrer and Guido Stampacchia. *An introduction to variational inequalities and their applications.* Pure and Applied Mathematics, Vol. 88. New York etc.: Academic Press (A Subsidiary of Harcourt Brace Jovanovich, Publishers), 1980.

[LM72] J.-L. Lions and E. Magenes. *Non-homogeneous boundary value problems and applications. Vol. I.* Springer-Verlag, New York, 1972. Translated from the French by P. Kenneth, Die Grundlehren der mathematischen Wissenschaften, Band 181.

[Lov94] David R. Lovett. *Demonstrating science with soap films.* Institute of Physics Publishing, Bristol, Philadelphia, 1994.

[Mai01] Matthias Maischak. *FEM/BEM methods for Signorini-type problems, error analysis, adaptivity, preconditioners.* Habilitation thesis, University Hannover, 2001.

[Mel02] J.M. Melenk. On condition numbers in *hp*-FEM with Gauss-Lobatto-based shape functions. *J. Comput. Appl. Math.*, 139(1):21–48, 2002.

[Mor66] C.B.jun. Morrey. *Multiple integrals in the calculus of variations.* Die Grundlehren der mathematischen Wissenschaften. 130. Berlin-Heidelberg-New York: Springer-Verlag, 1966.

[MP96] Jean-Franois Maitre and Olivier Pourquier. Condition number and diagonal preconditioning: Comparison of the *p*-version and the spectral element methods. *Numer. Math.*, 74(1):69–84, 1996.

[MW01] J.M. Melenk and B.I. Wohlmuth. On residual-based a posteriori error estimation in hp-FEM. *Adv. Comput. Math.*, 15(1-4):311–331, 2001.

[NW99] Jorge Nocedal and Stephen J. Wright. *Numerical optimization.* Springer Series in Operations Research. New York, NY: Springer, 1999.

[OD95] Elwood T. Olsen and Jim jun. Douglas. Bounds on spectral condition numbers of matrices arising in the *p*-version of the finite element method. *Numer. Math.*, 69(3):333–352, 1995.

[O'L80] Dianne P. O'Leary. A generalized conjugate gradient algorithm for solving a class of quadratic programming problems. *Linear Algebra Appl.*, 34:371–399, 1980.

[OR70] J.M. Ortega and W.C. Rheinboldt. *Iterative solution of nonlinear equations in several variables.* Computer Science and Applied Mathematics. New York-London: Academic Press, 1970.

[Ott85] Frei Otto. *Natürliche Konstruktionen: Formen und Strukturen in Natur und Technik und Prozesse ihrer Entstehung.* Deutsche Verlags-Anstalt GmbH, Stuttgart, 1985.

[Sch98] Ch. 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.

[Sti85a] T. J. Stieltjes. Sur quelques théorèmes d'Algébre. *Comptes Rendus de l'Académie des Sciences, Paris*, 100:439–440, 1885.

[Sti85b] T. J. Stieltjes. Sur les polynômes de Jacobi. *Comptes Rendus de l'Académie des Sciences*, 100:620–622, 1885.

[SV00] David F. Shanno and Robert J. Vanderbei. Interior-point methods for nonconvex nonlinear programming: orderings and higher-order methods. *Math. Program.*, 87(2, Ser. B):303–316, 2000. Studies in algorithmic optimization.

[Sze75] Gabor Szegö. *Orthogonal polynomials*. American Mathematical Society, Providence, R.I., fourth edition, 1975. American Mathematical Society, Colloquium Publications, Vol. XXIII.

[TE98a] Xue-Cheng Tai and Magne Espedal. Rate of convergence of some space decomposition methods for linear and nonlinear problems. *SIAM J. Numer. Anal.*, 35(4):1558–1570 (electronic), 1998.

[TE98b] Xue-Cheng Tai and Magne Espedal. Applications of a space decomposition method to linear and nonlinear elliptic problems. *Numer. Methods Partial Differ. Equations*, 14(6):717–737, 1998.

[Yos94] Kosaku Yosida. *Functional analysis. Repr. of the 6th ed.* Berlin: Springer-Verlag, 1994.

[Zei85] Eberhard Zeidler. *Nonlinear functional analysis and its applications. III.* Springer-Verlag, New York, 1985. Variational methods and optimization, Translated from the German by Leo F. Boron.

Curriculum Vitae

General:

Name:	Andreas Krebs
Email:	krebs@math.tu-cottbus.de
Born:	April 30th, 1965 Bückeburg, Germany

Education:

09/2000 –	Ph.D. Student and Research Assistant at Department of Mathematics, Technical University of Cottbus, contract until 10/2010
11/1998 – 09/2000	Ph.D. Student and Research Assistant at Department of Applied Mathematics, University of Hanover
10/1993 – 10/1998	Student of Mathematics, University of Hanover
10/1986 – 07/1991	Student of Pedagogics for handicapped children, University of Hanover
07/1974 – 06/1984	Secondary School, Gymnasium Achim, A-levels

Examinations:

07/2004	Ph.D. in Mathematics (Dr. rer. nat)
10/1998	Diploma in Mathematics (Dipl.-Math.)
06/1991	1st State Examination as teacher for schools for mentally handicapped children

Practical Experience:

11/1998 –	Teaching mathematics and programming courses in `C` and `Fortran95` to students of mathematics, physics, computer sciences, and engineering at the universities of Cottbus and Hanover
10, 11/1999	DAAD sponsored research at the University of Conception, Chile
10/1993 – 03/1996	Unix system administration, Software development in `C` and `Fortran77` as Student Assistant at the Institute for Transport, Railway Construction and Operation, University of Hanover
10/1991 – 09/1993	Social Education Worker at an asylum for disturbed adolescents
07/1984 – 02/1986	Community Service at an asylum for mentally handicapped adults

Research Interests:

Nonlinear optimization, variational inequalities, orthogonal polynomials, p-version for finite elements and boundary elements and their coupling, a posteriori error estimating, software development with `C` and `Fortran 95`

Language Skills:

native language German, fluent English, some French