Mixed FE-models for variational inequalities

DISSERTATION

zur Erlangung des Grades eines Doktors

der Naturwissenschaften

vorgelegt von

Andrej Garanza, M.Sc.

eingereicht bei der Naturwissenschaftlich-Technischen Fakultät

der Universität Siegen

Siegen 2020

gedruckt auf alterungsbeständigem holz- und säurefreiem Papier

1. Gutachter: Prof. Dr. Franz-Theo Suttmeier, Universität Siegen

2. Gutachter: Prof. Dr. Thomas Wick, Universität Hannover

Tag der mündlichen Prüfung: 20. Oktober 2020

Acknowledgements

First, I would like to thank Prof. Dr. Franz-Theo Suttmeier for giving me the opportunity to write this thesis and gain experience in a broad range of mathematical questions during this time, for his mentoring and for creating space to explore different aspects of a problem and an atmosphere of open exchange.

Furthermore, I thank Prof. Dr. Thomas Wick for his interest in my work, for the opportunity to present it to his team and of course for his input.

Moreover, I would like to gratefully acknowledge the contributions of all former members of the "Wissenschaftliches Rechnen AG" (Scientific Computing Group) at the University of Siegen, which lead to the ideas, that are further explored in this thesis. Special thanks go to colleagues and "office neighbors" Prof. Dr. Robert Plato, Prof Dr. Volker Michel, Dr. Max Kontak and Dr. Sarah Leweke, as well as Bianca Kretz and Naomi Schneider for many helpful and inspiring discussions, motivation and support over the years.

From the bottom of my heart I am grateful for those who have been with me through it all:

to my parents Irina and Sergej for always putting the family first,

for your sacrifices along the way,

for your encouragement, help and support,

to my sister Anna and her husband Valentin for always being there for me,

to my niece Lisa and nephew Niklas for making me forget my worries and putting a smile on my face every time we spend time together.

Abstract

In this work we deepen our studies on the numerical FE-treatment of systems of partial differential equations, where the solution is subjected to inequality constraints. Especially we focus on Lagrange-settings, which can be employed to handle the given constraints. In this way additional auxiliary variables are introduced which are determined simultaneously to the original primal solution within a so-called mixed system.
On this basis efficient solution processes for the mixed systems are constructed by eliminating inequality constraints yielding nonlinear equation systems. These can easily be solved by (non-smooth) Newton-type schemes. Furthermore concepts for a posteriori error control are reviewed and refined.

Zusammenfassung

In dieser Arbeit werden Systeme partieller Differentialgleichungen mit Ungleichungsnebenbedingungen behandelt.

Genauer geht es um die numerische Analyse mit Finite-Element-Methoden (FEM). Besonderes Augenmerk liegt hierbei auf dem Einsatz von Lagrange-Techniken. Die dadurch eingeführten Hilfsvariablen werden simultan zur primalen Lösung im Rahmen eines sogenannten gemischten Systems bestimmt.

Auf der Basis von Projektionstechniken können die Ungleichungsnebenbedingungen eliminiert werden. Die dann entstehenden nicht-linearen Probleme werden dann mit nichtglatten Verfahren vom Newton-Typ effizient gelöst.

Darüber hinaus werden Techniken zur a posteriori Fehlerkontrolle verfeinert und auf die vorliegende neue Situation erweitert.

Contents

List of Figures

List of Tables

1. Introduction

In many fields, like continum mechanics and fluid dynamics, problems occur, that are a result of the functional minimization on a restricted subset. This can be e. g. some kind of an energy functional $\mathcal{J}$, that is derived from the, appropriate for the field, principle of conservation of energy. The system strives towards the lowest possible energy level. So if the energy functional $\mathcal{J}$ depends on the value of some parameter $u \in \mathbf{K}$, we are interested in the solution of the minimization problem

$$\mathcal{J}(u) \leq \mathcal{J}(v) \qquad \forall v \in \mathbf{K},$$

where $\mathbf{K}$ is the permitted subset of a, for the physical variable u appropriate, space $\mathbf{V}$. Usually $\mathbf{V}$ is a Sobolev space and the subset $\mathbf{K}$ is defined by an inequality, which in turn results in, as one of the possible interpretation of the problem, a search for u that solves a variational inequality.

The main aim of this work is to introduce an efficient solution strategy for this kind of problems. To be more precise, let us assume, that $\mathbf{V}$ is a Sobolev space defined on the domain $\Omega \subset \mathbb{R}^d$ with $d \in \mathbb{N}$. Now, the subset $\mathbf{K}$ can be defined by using an operator $G : \mathbf{V} \to \mathbf{L}^2(\Omega)$. Examples of such spaces can be $\{u \in \mathbf{V} \,|\, \mathrm{div}(u) \geq 0 \text{ a.e. on } \Omega\}$ or $\{u \in \mathbf{V} \,|\, u \geq \psi \text{ a.e. on } \Omega\}$, $\psi \in \mathbf{L}^2(\Omega)$ denoting an obstacle. In the framework of this dissertation, we will discuss these examples in greater detail as well as a more general case.

In order to find a more efficient solution strategy, we have to identify the main difficulty in this type of problems. Since we cannot directly solve variational differential inequalities, the usual approach is an iteration process, where in each step we calculate a new approximation of the the physical variable u, examine it and, if the new approximation violates the restriction, make a correction. The more efficient way, to calculate this approximation combines two ideas. First we can find an equivalent variational formulation of the minimization problem, that does not contain any inequalities, but instead can be expressed as a non-linear variational equation. The second part is to find an appropriate Newton-type method for a fast iteration process. The Newton method for solving non-linear equations requests a derivative, a Gateaux derivative in this case, but as we will see, for this type of problems, the Gateaux derivative has to be approximated. Here we have to balance between complexity of the estimated derivative, that results in a longer computation time, and the stability of the solution strategy. For the numerical tests we use the finite element method to provide discrete space,s, but the general approach is not restricted to these spaces.

So to summarize the above in two concrete tasks for this work:

- We want to present two equivalent formulations of the problems introduced above in the form of a variational equation or a system of variational equations, instead of the usual inequality approach,
- and we want to apply the Newton-type methods and examine the efficiency of resulting algorithms compared to the classical projection methods.

Before dealing with the actual problems, that we would like to solve, we establish a notation scheme for the further course of this work. The *second* chapter is dedicated to this task.

Following up, in the *third* chapter we start with a simple one dimensional obstacle problem. This allows us to start the discussion with a comparatively simple representative of the targeted type of problems. After a brief introduction we describe possible equivalent formulations, where one of those can be used in a projected gradient type iteration approach. Thereafter we introduce the alternative formulation of the problem, that rests upon projection operator. This should be considered as a motivation for the solution strategy and not a formal proof, since it requires unnecessary assumptions.

The *fourth* chapter contains extensive numerical tests. Here, first of all, we consider another example for the set of problems described above. The problem at hand comes from the field of fluid dynamic and describes the flow of a viscous liquid, that can turn into a gas under appropriate conditions. It is a Stokes problem, but instead of usual incompressibility condition, we use the inequality $\operatorname{div}(u) > 0$ a.e. on Ω. In the similar fashion to the third chapter, we introduce the problem as well as the equivalent formulations and the usual solving approach, underlining this way the similarities of both cases as well as the incised complexity.

After we sufficiently discussed the different formulations for the flow problem, the next step is to compare different solution strategies and the first case is a simple fixed-point iteration. This represents one of the slowest ways to obtain an approximation for the solution of the new variational formulation, that was introduced earlier, but, as shown in the first numerical tests, this may be already enough to improve upon the calculation time. As a comparable classical approach for solving this type of problems we use conjugate gradient projection method.

Next we consider ways to improve upon the iteration process and apply the Newton-type method mentioned earlier. The main problem in calculating the Gateaux derivative poses a projection operator, that the new variational formulation contains. We test three Newton-type methods: in the first we ignore the part with the projection operator completely, in the second we partially include it in the calculations and in the third we regularize it. For the tests we simulate the flow through a T-pipe section. As we will see, even so the setup remains the same, changes in the viscosity constant of the fluid impact the solution massively. Combined, we consider three cases, which result in different calculation times for the compared methods. The other main factor for determining the shortest calculation time is the starting value. Here we compaire a zero starting value, which are far from good as starting approximation, and an interpolation of the solution calculated for the discrete space with less degrees of freedom. Further more we have to

consider how different meshes and, as a result, different discrete spaces might impact on stability of the iteration process. Therefore we use an error estimator to identify parts of the mesh to refine, which on one hand allows us to assign more degrees of freedom to the parts of the domain Ω, where they are needed, but the handling of the so called hanging nodes can produce a mesh vulnerable to the oscillations. During the numerical tests we calculate and compare the calculation time for the different cases, where a combination of those factors is considered. Another part of the numerical analysis is the question of behavior of the solution with rising number of degrees of freedom. Usually we calculate the dependence on the cell size and, since the continuous solution is not available, first we consider the different parts of the error estimator and then we compare the differences between two solution approximations on two consecutive finite element spaces.

As for the rest of the fourth chapter, here we proof the existence and the uniqueness of the solution for the regularized problem and show how it relates to the solution of the original problem. This is also the part, where the conditions, that discrete spaces have to fulfill, are introduced and used in the proof. Also we derive the error estimator, that was mentioned before. We consider two possible cases: a conform ($\mathbf{V}_h \subset \mathbf{V}$) and a non-conform ($\mathbf{V}_h \not\subset \mathbf{V}$) discretization, where $\mathbf{V}_h$ is the finite element space used for the approximation.

In the *fifth* chapter we revisit the obstacle problem, but we consider a two dimensional case. After modifying the formulation and applying some stabilization techniques, we calculate a membrane test example and compare the results, in particular the numbers of iterations, with different methods, applied to the same example. Also we derive a similar error estimator as we did for the Stokes problem.

The *sixth* chapter is devoted to the general case. First of all we have to specify the condition that the functional $\mathcal{J}$ and the subset defining operator G must fulfill. On one hand those are necessary to ensure the existence and uniqueness of the solution of the minimization problem. On the other hand we need this conditions to proof convergence of the proposed solution strategy. This describes the next parts of the chapter: after a short excursus into the minimization theory for functional and the derivation of the varitional formulation with a projection operator, we look into the iteration process itself. Here we proof the existence of a new estimation in each step as well as convergence of the method.

Finally in the last chapter is devoted to the summary and outlook.

2. Notation

Before we speak about the differential inequality and their solution strategies, let us clarify notations, that are used in this work.

In the Euclidean space $\mathbb{R}^n$ with $n \in \mathbb{N}$ the inner product or the dot product of two vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ is defined as $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_{k=1}^{n} x_k y_k$. Analogously, we define the double-dot product for space of $\mathbb{R}^{n \times m}$ type as $X : Y = \sum_{j=1}^{n} \sum_{k=1}^{m} x_{jk} y_{jk}$ with $X, Y \in \mathbb{R}^{n \times m}$ and $n, m \in \mathbb{N}$.

When studying the differential inequalities we consider different functions. Typical real-valued function f can be written as $f : \Omega \longrightarrow \mathbb{R},\ x \mapsto f(x)$, where the domain Ω is an open subset of $\mathbb{R}^d$ with $d \in \mathbb{N}$ a number of the problems dimensions. The symbol ∂ used with a domain designates the boundary of this domain, so $\partial\Omega$ is the boundary of Ω. The notation $\bar{\Omega}$ stands for the closure of the domain Ω, i.e. $\bar{\Omega} = \Omega \cup \partial\Omega$. Since we are not only interested in the scalar function, we can also define, using the notation above, vector-valued function $\boldsymbol{f}$ as $\boldsymbol{f} : \Omega \longrightarrow \mathbb{R}^n,\ x \mapsto \boldsymbol{f}(x)$ and matrix-valued function F as $F : \Omega \longrightarrow \mathbb{R}^{n \times m},\ x \mapsto F(x)$ with $n, m \in \mathbb{N}$.

Using vectors and matrices we compactly write inequality system, like $\boldsymbol{v} > 0$, which means that all entries of the vector $\boldsymbol{v}$ are positive.

The set of the eigenvalues of a matrix M can be written as $\mathrm{Eig}\,(M)$ and smallest of them is $\mathrm{Eig}_{min}\,(M) = \min\{\lambda \mid \lambda \in \mathrm{Eig}\,(M)\}$.

Together with the functions, we use a number of different operators:

- the operator $\partial^\alpha q = \frac{\partial^{|\alpha|} q}{\partial^{\alpha_1} x_1 \ldots \partial^{\alpha_n} x_n}$ describe a partial derivative of function q for a multi-index $\alpha \in \mathbb{N}_0^n$ with $\alpha = (\alpha_1, \ldots, \alpha_n)$,
- $\frac{\partial q}{\partial \boldsymbol{n}}$ is a directional derivative along a vector $\boldsymbol{n}$ (usually, vector $\boldsymbol{n}$ is a vector-valued function, which represents the normal vector of the domain boundary),
- $\nabla q = \begin{pmatrix} \frac{\partial q}{\partial x_1} \\ \vdots \\ \frac{\partial q}{\partial x_n} \end{pmatrix}$ stands for the gradient of the real-valued function q and
- $\nabla \boldsymbol{\varphi} = \begin{pmatrix} \frac{\partial \varphi_1}{\partial x_1} & \cdots & \frac{\partial \varphi_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \varphi_n}{\partial x_1} & \cdots & \frac{\partial \varphi_n}{\partial x_n} \end{pmatrix}$ is the gradient of the vector-valued function $\boldsymbol{\varphi}$,

- $\nabla \cdot \boldsymbol{\varphi} = \sum\limits_{j=1}^{n} \frac{\partial \varphi_j}{\partial x_j}$ and $\nabla \cdot A = \begin{pmatrix} \sum\limits_{j=1}^{n} \frac{\partial A_{1,j}}{\partial x_j} \\ \vdots \\ \sum\limits_{j=1}^{n} \frac{\partial A_{n,j}}{\partial x_j} \end{pmatrix}$ are two variants of divergence operator, depending on the art of function it is applied to (in this case $\boldsymbol{\varphi} : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ and $A : \mathbb{R}^n \longrightarrow \mathbb{R}^{n \times n}$ with $n \in \mathbb{N}$),

- $\Delta q = \nabla \cdot (\nabla q) = \sum\limits_{j=1}^{n} \frac{\partial^2 q}{\partial x_j^2}$ and $\Delta \boldsymbol{\varphi} = \nabla \cdot (\nabla \boldsymbol{\varphi}) = \begin{pmatrix} \sum\limits_{j=1}^{n} \frac{\partial^2 \varphi_1}{\partial x_j^2} \\ \vdots \\ \sum\limits_{j=1}^{n} \frac{\partial^2 \varphi_n}{\partial x_j^2} \end{pmatrix}$ describe the Laplace operator, which also depends on the art of function it is applied to,

- since an operator is mapping from one vector space to another, we can use the notation e.g. $\breve{G} : \mathbf{V} \longrightarrow \mathbf{W}$ to introduce a general not further specified operator $\breve{G}$,

- $\mathit{\Pi}_\Lambda$ is a projection operator on the subset $\boldsymbol{\Lambda}$. If the subset $\boldsymbol{\Lambda}$ is part of the space $\mathbf{Q}$, then for all the elements $x \in \mathbf{Q}$ following applies $\|x - \mathit{\Pi}_\Lambda x\|_{\mathbf{Q}} = \inf\limits_{y \in \boldsymbol{\Lambda}} \|x - y\|_{\mathbf{Q}}$.

- I_h is a Clément's interpolation operator (see f.e. Braess [4, p. 80]).

- $\mathbf{1}_{\tilde{\Omega}}$ is an indicator function, defined by $\mathbf{1}_{\tilde{\Omega}}(x) = \begin{cases} 1 & \text{if } x \in \tilde{\Omega} \\ 0 & \text{else} \end{cases}$.

If we want to describe some properties of a function, we usually write it as an element of the corresponding space, for example real-valued function f which defined on Ω is an element of $\mathbf{C}^0(\Omega)$. In this work we use following function spaces:

- $\mathbf{C}^m(\Omega)$, with $m \in \mathbb{N}$, is the space of functions which, together with their derivatives of order less or equal to m, are continuous on Ω.

- $\mathbf{L}^2(\Omega)$ is the space of measurable functions $v : \Omega \longrightarrow \mathbb{R}$ such that

$$\|v\|_{0,\Omega} = \sqrt{\int_\Omega (v(x))^2 \, \mathrm{d}x} < \infty.$$

  The norm, that is introduced above, is induced by the scalar product

$$(v, \varphi)_{0,\Omega} = \int_\Omega v(x)\varphi(x)\mathrm{d}x \qquad \forall v, \varphi \in \mathbf{L}^2(\Omega).$$

- $\mathbf{H}^m(\Omega)$, with $m \in \mathbb{N}$, is the Sobolev space of functions $v \in \mathbf{L}^2(\Omega)$ such that for each multi-index α with $|\alpha| \leq m$, the α^{th} weak derivative $\partial^\alpha v$ exist and $\partial^\alpha v \in \mathbf{L}^2(\Omega)$.

The scalar product and the norm of this space can be written as

$$(v,\varphi)_{m,\Omega} = \int_\Omega \sum_{|\alpha|\le m} (\partial^\alpha v(x)\partial^\alpha \varphi(x))\,\mathrm{d}x = \sum_{|\alpha|\le m} (\partial^\alpha v, \partial^\alpha \varphi)_{0,\Omega} \quad \forall v,\varphi \in \mathbf{H}^m(\Omega)$$

$$\text{and} \quad \|v\|_{m,\Omega} = \sqrt{(v,v)_{m,\Omega}} \qquad \forall v \in \mathbf{H}^m(\Omega).$$

- $\mathbf{C}_0^\infty(\Omega) = \{v \in \mathbf{C}^\infty(\Omega) \,|\, \operatorname{supp}(v) \text{ is a proper subset of } \Omega\}$, with a support of the function v defined as $\operatorname{supp}(v) = \overline{\{\boldsymbol{x} \in \Omega |\, v(\boldsymbol{x}) \neq 0\}}$.
- $\mathbf{H}_0^m(\Omega)$ is the closure of $\mathbf{C}_0^\infty(\Omega)$ in $\mathbf{H}^m(\Omega)$. As suggested in Dobrowoski [7, p. 120] we have an equivalent scalar product for such spaces:

$$(v,\varphi)_{m*,\Omega} = \sum_{|\alpha|=m} (\partial^\alpha v, \partial^\alpha \varphi)_{0,\Omega} \qquad \forall v,\varphi \in \mathbf{H}_0^m(\Omega).$$

- $\mathbf{H}^{-m}(\Omega)$ is the dual space of $\mathbf{H}^m(\Omega)$ with

$$\langle v,\varphi\rangle_{\mathbf{H}^m(\Omega)} = (v,\varphi)_{0,\Omega} \qquad \forall v \in \mathbf{H}^m(\Omega), \varphi \in \mathbf{H}^{-m}(\Omega)$$

$$\text{and} \quad \|\varphi\|_{-m,\Omega} = \sup_{v\in\mathbf{H}^m(\Omega)} \frac{\left|(v,\varphi)_{0,\Omega}\right|}{\|v\|_{m,\Omega}} \qquad \forall \varphi \in \mathbf{H}^{-m}(\Omega).$$

- In some cases we need space, which are combinations of the spaces above. This spaces will be defined during examination of the associated problem. The scalar products and the norms of such spaces can be often written using the some differential operators, i.e. $\mathbf{V} = \left(\mathbf{H}_0^1(\Omega)\right)^n$ and $\mathbf{Q} = \left(\mathbf{L}^2(\Omega)\right)^n$, where $\left((\boldsymbol{v},\boldsymbol{p})^{\mathrm{T}}, (\boldsymbol{\varphi},\boldsymbol{q})^{\mathrm{T}}\right)_{\mathbf{V}} = (\nabla\boldsymbol{v}, \nabla\boldsymbol{\varphi})_{0,\Omega} + (\boldsymbol{v},\boldsymbol{\varphi})_{0,\Omega} + (\boldsymbol{p},\boldsymbol{q})_{0,\Omega}$.

We use functionals to formulate the alternative forms of the problems we consider. The suitable functional can be defined as $\mathcal{L} : \mathbf{V} \longrightarrow \mathbb{R}$, where $\mathbf{V}$ is a function space. If the limit

$$\left\langle \mathcal{L}'(v), \varphi \right\rangle_{\mathbf{V}} := \lim_{\varepsilon\to 0} \frac{1}{\varepsilon} \left(\mathcal{L}(v+\varepsilon\varphi) - \mathcal{L}(v)\right)$$

exists, then we call the functional $\mathcal{L}$ differentiable at $v \in \mathbf{V}$ in direction $\varphi \in \mathbf{V}$. $\langle \mathcal{L}'(v), \varphi\rangle_{\mathbf{V}}$ is the so called Gateaux derivative of $\mathcal{L}$ at v in direction φ and $\mathcal{L}' : \mathbf{V} \longrightarrow \mathbf{V}^*, v \mapsto \mathcal{L}'(v)$ is a mapping into the dual space. (Note: The so called dual pair $\langle .\,,.\rangle_{\mathbf{V}}$ is a bilinear form, which allows us to evaluate, in case of $\langle \mathcal{L}'(v), \varphi\rangle_{\mathbf{V}}$, the functional $\mathcal{L}'(v) \in \mathbf{V}^*$ for a argument $\varphi \in \mathbf{V}$.)

We split the domain Ω into decomposition $\mathbb{T}_h = \{T_j | 1 \le j \le N_h\}$, consisting of N_h element or cells T_j. The decomposition $\mathbb{T}_h$ is usually called mesh or triangulation. The width of each element T_j can be described with with h_j =diam(T_j). We call $h = \min\{h_j | 1 \le j \le N_h\}$ the width of the mesh. Based on the mesh, we define standard finite element spaces, i.e. $\mathbf{V_h} = \operatorname{span}\{\varphi_1, \ldots, \varphi_N\} \subseteq \mathbf{V}$, where $\{\varphi_1, \ldots, \varphi_N\}$ is the basis of the finite element space $\mathbf{V_h}$ with $N \in \mathbb{N}$, which depends on N_h and art of chosen finite element space.

3. Basic idea applied to 1-D obstacle problem

In order to introduce the basic concept of the solution strategy, that we propose, let us consider a simple example from a set of problems, to which this solution strategy can be applied. The figure 3.1 depicts an elastic rope or cable, that is held in place on the right and the left sides. The cable is affected by the gravitational force f and the Hooke's law applies resulting in the deformation of the cable. Depending on the elasticity of the cable, it can still hang above the ground or in parts touch the ground.

In mathematical terms we define an external force in each point of the cable and the deviation from straight line as $f : \Omega \to \mathbb{R}$ and $u : \Omega \to \mathbb{R}$ respectively, where Ω is the interval between the two points, where the rope is fixed. The obstacle describes the function $\psi : \Omega \to \mathbb{R}$. The equilibrium of the forces can then be formulated as a formula:

$$-u'' = f.$$

Assuming zero displacement at the boundary, then $u \in \mathbf{C}_0^2(\Omega) \cap \mathbf{C}^0(\bar{\Omega})$. In case $u > \psi$ every where on the interval Ω, this is the one dimensional Poisson's problem and it can be solved easily. If on the other side this condition is not automatically satisfied, then we need to examine this problem more closely. If the cable can actually reach the obstacle, then two distinct effects will accrue on the subsets of the interval Ω. The first one is the equilibrium of forces, that is described above. The second possibility is $u = \psi$, in case the Hook's force is not sufficient and the reaction force from the ground prevents the

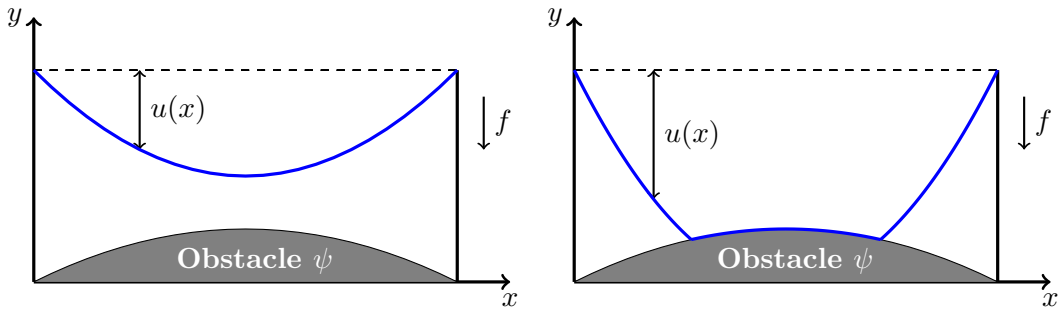

Figure 3.1.: Obstacle problem in one dimension.

further decent below the obstacle. This results in the classical formulation:

$$\begin{aligned} -u'' - f &\geq 0\,, \\ u - \psi &\geq 0\,, \\ (u-\psi)(-u''-f) &= 0\,. \end{aligned}$$

3.1. Possible variational formulations

Since we typically interested in the variational formulations for the numerical approximations, let us consider the initial minimization problem. From the point of view of physics, instead of searching for the force equilibrium, we look for the lowest energy level of the system. In the case of our 1-D obstacle problem, this would be equivalent to:

Find $u \in \mathbf{K}$, that satisfy

$$\mathcal{J}(u) \leq \mathcal{J}(\varphi) \qquad\qquad \forall \varphi \in \mathbf{K}\,,$$

where $\mathcal{J}(\varphi) = \frac{1}{2}(\varphi', \varphi')_{0,\Omega} - (f, \varphi)_{0,\Omega}$ for all $\varphi \in \mathbf{K}$ and the subset $\mathbf{K}$ is defined as $\mathbf{K} = \left\{\varphi \in \mathbf{H}_0^1(\Omega) \;\middle|\; \varphi \geq \psi \text{ o.e. on } \Omega\right\}$.

The first variational formulation of this problem can be derived by calculating the Gateaux derivative, resulting in the variational inequality

$$\left\langle \mathcal{J}'(u), \varphi - u \right\rangle \geq 0 \qquad\qquad \forall \varphi \in \mathbf{K}.$$

According to the definition of the functional, this can be written with two scalar products:

$$\left(u', \varphi' - u'\right)_{0,\Omega} - (f, \varphi - u)_{0,\Omega} \geq 0 \qquad\qquad \forall \varphi \in \mathbf{K}.$$

The constraint $u - \psi \geq 0$ is, in this case, a part of the subset $\mathbf{K}$ and the typical solution strategy (f.e. the gradient projection method) involves an iteration, where we calculate the update for the solution u without the constraint, and in a post-process correct it locally, if the new solution would violate the constraint.

In order to make the constraint a part of the calculation, we use the Lagrangian mechanics, which leads to the equivalent saddle point problem:

Find $u \in \mathbf{H}_0^1(\Omega)$ and $p \in \mathbf{L}^2(\Omega)$, that satisfy

$$\mathcal{L}(u, p) = \inf_{\varphi \in \mathbf{V}} \sup_{q \in \mathbf{Q}} \mathcal{L}(\varphi, q)$$

where $\mathbf{V} = \mathbf{H}_0^1(\Omega)$ and $\mathbf{Q} = \mathbf{L}^2(\Omega)$, as well as

$$\mathcal{L}(\varphi, q) = \frac{1}{2}\left(\varphi', \varphi'\right)_{0,\Omega} - (f, \varphi)_{0,\Omega} - (q, \varphi - \psi)_{0,\Omega}\,.$$

Next step is to formulate this saddle point problem in the similar fashion to the Karush-Kuhn-Tucker-conditions, that are used for finite dimensional minimization problems (see

f.e. Geiger and Kanzow [8]). Because of the similarities, we will call this new formulation hereafter the Karush-Kuhn-Tucker-conditions:

$$\begin{aligned} \left(u', \varphi'\right)_{0,\Omega} - (f, \varphi)_{0,\Omega} - (p, \varphi)_{0,\Omega} &= 0 && \forall \varphi \in \mathbf{V} \\ (u - \psi, q - p)_{0,\Omega} &\geq 0 && \forall q \in \mathit{\Pi}^{+}\mathbf{Q} \\ (u - \psi, p)_{0,\Omega} &= 0 \\ p &\in \mathit{\Pi}^{+}\mathbf{Q} \end{aligned}$$

where $\mathit{\Pi}^{+}\mathbf{Q} = \left\{\omega \in \mathbf{L}^2(\Omega) |\ \omega \geq 0 \text{ o.e. on } \Omega\right\}$. In this formulation the constraints on the function u are no longer a part of the allowed subset and we can use the entire space $\mathbf{V}$ for the test functions. But now a restriction applies on the test space for the Lagrange multiplier p ($q \in \mathit{\Pi}^{+}\mathbf{Q}$) and we still have to deal with a variational inequality.

3.2. Transformation into the system of equations

For the further transformation of the problem, we define the bilinear form $\mathcal{A} : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ with $\mathcal{A}(u, \varphi) = (u', \varphi')_{0,\Omega}$. According to the Lax-Milgram-Theorem, there is exactly one invertible operator $\breve{A} : \mathbf{V} \to \mathbf{V}$, such that

$$\mathcal{A}(v, \varphi) = \left(v, \breve{A}\varphi\right)_{0,\Omega} \qquad \forall v, \varphi \in \mathbf{V}.$$

Since the operator $\breve{A}$ is invertible, there is a inverse operator $\breve{A}^{-1}$, which allows us to write the following equation:

$$(u, \varphi)_{0,\Omega} = \mathcal{A}(u, A^{-1}\varphi) \qquad \forall v, \varphi \in \mathbf{V}.$$

Let assume for a moment, that $\mathbf{Q} = \mathbf{H}^1(\Omega)$. Then with the first variation equation of the previous Karush-Kuhn-Tucker-conditions and $\varphi = A^{-1}(q - p)$, we get

$$(u, (q - p))_{0,\Omega} = \left(f, A^{-1}(q - p)\right)_{0,\Omega} + \left(p, A^{-1}(q - p)\right)_{0,\Omega} \qquad \forall q \in \mathit{\Pi}^{+}\mathbf{Q}. \tag{3.1}$$

Now, with the second of the previous Karush-Kuhn-Tucker-conditions and the equation above we get

$$\begin{aligned} \left(p, A^{-1}(q - p)\right)_{0,\Omega} + \left(f, A^{-1}(q - p)\right)_{0,\Omega} - (\psi, (q - p))_{0,\Omega} &\geq 0 && \forall q \in \mathit{\Pi}^{+}\mathbf{Q} \\ \text{or} \qquad \left\langle \tilde{\mathcal{J}}'(p), q - p \right\rangle &\geq 0 && \forall q \in \mathit{\Pi}^{+}\mathbf{Q}, \end{aligned}$$

$$\text{with} \qquad \tilde{\mathcal{J}}(q) = \frac{1}{2}\left(q, A^{-1}q\right)_{0,\Omega} + \left(f, A^{-1}q\right)_{0,\Omega} - (\psi, q)_{0,\Omega}\ .$$

This is an another minimization problem with q as a variable, p as a solution and the restriction $p, q \in \mathit{\Pi}^{+}\mathbf{Q}$:

$$\tilde{\mathcal{J}}(p) \leq \tilde{\mathcal{J}}(q) \qquad \forall q \in \mathit{\Pi}^{+}\mathbf{Q}\,.$$

Analogous to the previous minimization problem we can introduce another Lagrange multiplier λ alongside a saddle point problem:

$$\tilde{\mathcal{L}}(p,\lambda) = \inf_{q\in\mathbf{Q}} \sup_{\omega\in\Pi^+\mathbf{Q}} \tilde{\mathcal{L}}(q,\omega) \qquad \text{with} \qquad \tilde{\mathcal{L}}(q,\omega) = \tilde{\mathcal{J}}(q) - (\omega,q)_{0,\Omega}\,.$$

In the similar fashion we derive yet an another set of Karush-Kuhn-Tucker-conditions:

$$\begin{aligned} \left(p, A^{-1}q\right)_{0,\Omega} + \left(f, A^{-1}q\right)_{0,\Omega} - (\psi,q)_{0,\Omega} - (\lambda,q)_{0,\Omega} = 0 \qquad & \forall q\in\mathbf{Q}\,, \\ (p,\omega-\lambda)_{0,\Omega} \geq 0 \qquad & \forall\omega\in\Pi^+\mathbf{Q}\,, \\ (\lambda,p)_{0,\Omega} = 0, & \\ \lambda\in\Pi^+\mathbf{Q}. & \end{aligned}$$

Here, we can use the equation (3.1) once again to not only simplify the first of the conditions, but also to eliminate the operator $\breve{A}^{-1}$ from the conditions entirely, which makes the assumption $\mathbf{Q} = \mathbf{H}^1(\Omega)$ unnecessary. The first of the conditions now reads as

$$(u-\psi,q)_{0,\Omega} - (\lambda,q)_{0,\Omega} = 0 \qquad \forall q\in\mathbf{Q}\,.$$

Next we replace the test function with $q = \omega-\lambda$, multiply the second inequality condition with a positive constant ϑ and subtract it from the the equation above, resulting in

$$((u-\psi-\vartheta p)-\lambda,\omega-\lambda)_{0,\Omega} \leq 0 \qquad \forall\omega\in\Pi^+\mathbf{Q}.$$

This is a so called projection inequality (see f.e. Alt [1, p. 96]), which means that the Lagrange multiplier λ can also be interpreted as a projection on the subset $\Pi^+\mathbf{Q}$. Combining all of the above, we can derive a new variational formulation, that is equivalent to the original minimization problem:

Find $u\in\mathbf{H}_0^1(\Omega)$ and $p\in\mathbf{L}^2(\Omega)$, that satisfy for $\vartheta>0$

$$\begin{aligned} \left(u',\varphi'\right)_{0,\Omega} - (p,\varphi)_{0,\Omega} = (f,\varphi)_{0,\Omega} \qquad & \forall\varphi\in\mathbf{V} \\ (u-\psi,q)_{0,\Omega} = (\lambda,q)_{0,\Omega} \qquad & \forall q\in\mathbf{Q}\,, \\ \text{where} \qquad \lambda = \Pi^+\left(u-\psi-\vartheta p\right). & \end{aligned}$$

Neither has this formulation variational inequalities in it, nor are the test spaces in any form restricted to subsets, which means no post-processing is necessary. The projection operator Π^+ make the problem non-linear, but it can be handled relatively easy, since it means

$$\lambda = \begin{cases} u-\psi-\vartheta p \text{ a.e. on } \tilde{\Omega}\subset\Omega & \text{if } u-\psi\geq\vartheta p \text{ a.e. on } \tilde{\Omega}\subset\Omega \\ 0 \text{ a.e. on } \tilde{\Omega}\subset\Omega & \text{else} \end{cases}.$$

This concludes the introduction of the basic idea. The possible numerical algorithms will be disused in the next chapter and applied to an another problem. After that we revisit the obstacle problem, but with $\Omega\subset\mathbb{R}^2$ and compare the results of numerical test with other algorithms.

4. Stokes problem with cavitation effects

The Stokes problem or the Stokes flow problem is a system of the differential equation, where the solution describe the flow of a viscous fluid. In the classical variant of the Stokes problem without cavitation we search a velocity function $\boldsymbol{u} \in \left(\mathbf{C}_0^2(\Omega) \cap \mathbf{C}^0(\bar{\Omega})\right)^n$ and a pressure function $p \in \left(\mathbf{C}^1(\Omega) \cap \mathbf{C}^0(\bar{\Omega})\right)$, that satisfy the differential equations

$$\begin{aligned} -\mu\Delta\boldsymbol{u} + \nabla p &= \boldsymbol{f} \\ \text{and} \qquad \nabla \cdot \boldsymbol{u} &= 0 \end{aligned}$$

for a constant viscosity factor $\mu > 0$. The second equation is a so called incompressibility condition, based on the assumption, that a fluid cannot be compressed. On the other side the same condition does not allow us to take into account the possible cavitation effect.

The cavitation can take place, if the fluid is kept in a liquid state only by the external pressure. So when the pressure drops below a certain level, the fluid switches into the gas state and expands. This is a violation of the incompressibility condition. If we want to allow the cavitation to take place in the simulation we must replace it with inequality condition

$$\nabla \cdot \boldsymbol{u} \geq 0 \,.$$

Modelling phenomena arising in the context of cavitation (see e.g. Nilsson and Hansbo [[13], [14]]) the additional condition $p \geq 0$ on Ω has to be incorporated. Also the pressure and the velocity must now satisfy the complementary slackness

$$p\,(\nabla \cdot \boldsymbol{u}) = 0 \,.$$

As we will see this condition derives from the initial minimization problem.

4.1. Minimization problem and equivalent variational formulation

In this section we go back to the initial minimization problem, which is the result of the modulation of the flow problem according to the laws of the continuous mechanic, from which the other formulations can be derived. In the context of the physics this can be considered as minimization of the energy functional $\mathcal{J} : \left(\mathbf{H}_0^1(\Omega)\right)^n \to \mathbb{R}$, with

$$\mathcal{J}(\boldsymbol{\varphi}) = \frac{\mu}{2}\,(\nabla\boldsymbol{\varphi}, \nabla\boldsymbol{\varphi})_{0,\Omega} - (f, \boldsymbol{\varphi})_{0,\Omega} \qquad\qquad \forall \boldsymbol{\varphi} \in \left(\mathbf{H}_0^1(\Omega)\right)^n \;,$$

on the subset $\mathbf{K}$, where

$$\mathbf{K} = \left\{ \boldsymbol{\varphi} \in \left(\mathbf{H}_0^1(\Omega)\right)^n \middle| \ \nabla \cdot \boldsymbol{\varphi} \geq 0 \text{ a.e. in } \Omega \right\} .$$

Summarising the above the Stokes problem can be written as:

Find $\boldsymbol{u} \in \mathbf{K}$ that satisfied the inequality

$$\mathcal{J}(\boldsymbol{u}) \leq \mathcal{J}(\boldsymbol{\varphi}) \qquad \forall \boldsymbol{\varphi} \in \mathbf{K}.$$

Since functional $\mathcal{J}$ is convex the inequality above can be replaced with a dual pair inequality (see section 6.2.2)

$$\left\langle \mathcal{J}'(\boldsymbol{u}), \boldsymbol{\varphi} - \boldsymbol{u} \right\rangle_{\mathbf{K}} \geq 0 \qquad \forall \boldsymbol{\varphi} \in \mathbf{K}$$

or an equivalent variational inequality

$$\mu \left(\nabla \boldsymbol{u}, \nabla(\boldsymbol{\varphi} - \boldsymbol{u})\right)_{0,\Omega} - (\boldsymbol{f}, \boldsymbol{\varphi} - \boldsymbol{u})_{0,\Omega} \geq 0 \qquad \forall \boldsymbol{\varphi} \in \mathbf{K} .$$

In this formulation of the problem the inequality condition, which allows the cavitation to take place, is a part of the definition of the subset $\mathbf{K}$ and there is no pressure function in it. Only when we apply the method of Lagrange multipliers, we will need an another function in our setting, which corresponds to the physical pressure in the system. So the minimization problem also has an equivalent Lagrange formulation:

Find a pair $(\boldsymbol{u}, p)^{\mathrm{T}} \in \mathbf{V} \times \boldsymbol{\Lambda}$ with

$$\mathcal{L}(\boldsymbol{u}, p) = \inf_{\boldsymbol{\varphi} \in \mathbf{V}} \sup_{q \in \boldsymbol{\Lambda}} \mathcal{L}(\boldsymbol{\varphi}, q) ,$$

where $\boldsymbol{\Lambda} = \left\{ q \in \mathbf{L}^2(\Omega) \middle| q \geq 0 \text{ a.e. in } \Omega \right\}$ and $\mathbf{V} = \left(\mathbf{H}_0^1(\Omega)\right)^n$ as well as

$$\mathcal{L}(\boldsymbol{\varphi}, q) = \frac{\mu}{2} \left(\nabla \boldsymbol{\varphi}, \nabla \boldsymbol{\varphi}\right)_{0,\Omega} - (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} - (q, \nabla \cdot \boldsymbol{\varphi})_{0,\Omega} .$$

Using the stationarity condition for a saddle point we can derive mixed variational formulation (similar to Karush-Kuhn-Tucker-conditions): Find a pair $(\boldsymbol{u}, p)^{\mathrm{T}} \in \mathbf{V} \times \boldsymbol{\Lambda}$ fulfilling the mixed formulation

$$\mu \left(\nabla \boldsymbol{u}, \nabla \boldsymbol{\varphi}\right)_{0,\Omega} - (p, \nabla \cdot \boldsymbol{\varphi})_{0,\Omega} = (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} \qquad \forall \boldsymbol{\varphi} \in \mathbf{V} , \tag{4.1}$$

$$(\nabla \cdot \boldsymbol{u}, q)_{0,\Omega} \geq 0 \qquad \forall q \in \boldsymbol{\Lambda}. \tag{4.2}$$

$$p(\nabla \cdot \boldsymbol{u}) = 0 \quad \text{ a.e. in } \Omega . \tag{4.3}$$

Next we define a bilinear form $\mathcal{A} : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ with $\mathcal{A}(\boldsymbol{v}, \boldsymbol{\varphi}) = \mu \left(\nabla \boldsymbol{v}, \nabla \boldsymbol{\varphi}\right)_{0,\Omega}$. It is obvious, that this form is V-elliptic and continuous. This means, that, according to the Lax–Milgram theorem, there is a invertible linear operator $\breve{A}$, that allows us to write

$$\mathcal{A}(\boldsymbol{v}, \boldsymbol{\varphi}) = \left(\boldsymbol{v}, \breve{A}\boldsymbol{\varphi}\right)_{0,\Omega} .$$

Also, if $\breve{G}^*$ the adjoint operator of divergence exists, meaning

$$(\nabla \cdot \boldsymbol{\varphi}, q)_{0,\Omega} = -\left(\boldsymbol{\varphi}, \breve{G}^* q\right)_{0,\Omega} ,$$

then we can write the equation (4.1) and the inequality (4.2) as

$$\begin{aligned} \left(\boldsymbol{u}, \breve{A}\boldsymbol{\varphi}\right)_{0,\Omega} + \left(\boldsymbol{\varphi}, \breve{G}^* p\right)_{0,\Omega} &= (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} && \forall \boldsymbol{\varphi} \in \mathbf{V} \\ \text{and} \qquad -\left(\boldsymbol{u}, \breve{G}^* q\right)_{0,\Omega} &\geq 0 && \forall q \in \boldsymbol{\Lambda}. \end{aligned} \tag{4.4}$$

We gain as well an implication of the complementarity condition (4.3)

$$\left(\boldsymbol{u}, \breve{G}^* p\right)_{0,\Omega} = 0 .$$

By using the test function $\boldsymbol{\varphi} = \breve{A}^{-1}\breve{G}^*(q-p)$ we obtain the inequality

$$\left(\breve{A}^{-1}\breve{G}^*(q-p), \breve{G}^* p\right)_{0,\Omega} - \left(\boldsymbol{f}, \breve{A}^{-1}\breve{G}^*(q-p)\right)_{0,\Omega} \geq 0 \qquad \forall q \in \boldsymbol{\Lambda},$$

which is equivalent to

$$\left\langle \mathcal{J}'(\breve{A}^{-1}\breve{G}^* p), \breve{A}^{-1}\breve{G}^*(q-p) \right\rangle \geq 0 \qquad \forall q \in \boldsymbol{\Lambda}.$$

This leads to an another minimization problem with a convex functional: Find $p \in \boldsymbol{\Lambda}$, such that

$$\mathcal{J}(\breve{A}^{-1}\breve{G}^* p) \leq \mathcal{J}(\breve{A}^{-1}\breve{G}^* q) \qquad \forall q \in \boldsymbol{\Lambda}. \tag{4.5}$$

That is the classical way of solving this problem: we consider the obstacle problem in variable p (because $p \in \boldsymbol{\Lambda}$ means essential that $p \geq 0$ a.e. on Ω). This can be done (in a discrete space) by conjugate gradient algorithm, see for example Blum, Braess and Suttmeier [3], or by conjugate gradient projection method, see Dembo and Tulowitzky [6]. After that the velocity $\boldsymbol{u}$ can be calculated using the equation (4.1). More on this can be found in the section 4.3, where we compare the calculation times of the the second method mentioned above with our first proposal for solution strategy from the next section.

4.2. First proposal for solution strategy

In last section we derived an other minimization problem, that can be solved as an obstacle problem in the variable p. In this section we go further and derive an equivalent variational problem with two Lagrange multipliers. The second multiplier λ is, in a way, artificial and can be calculated, if we know $\boldsymbol{u}$ and p.

In order to do that, we use the Lagrange multiplier methods the same way as above and obtain an another set of KKT-conditions for a pair $(p,\lambda)^{\mathrm{T}} \in \mathbf{Q}\times\mathbf{\Lambda}$ with $\mathbf{Q}=\mathbf{L}^2(\Omega)$:

$$\begin{aligned}
\mathcal{A}\left(\breve{A}^{-1}\breve{G}^*q,\breve{A}^{-1}\breve{G}^*p\right)-\left(\boldsymbol{f},\breve{A}^{-1}\breve{G}^*q\right)_{0,\Omega}-(\lambda,q)_{0,\Omega}=0 \qquad & \forall q\in\mathbf{Q}\,,\\
(\omega,p)_{0,\Omega}\geq 0 \qquad & \forall\omega\in\mathbf{\Lambda}\,,\\
p\lambda=0 \quad \text{a.e. in } \Omega\,. &
\end{aligned}$$

The equation (4.4) makes it possible to simplify the first of the new KKT-conditions:

$$\begin{aligned}
0=\left(\breve{A}^{-1}\breve{G}^*q,\breve{G}^*p\right)_{0,\Omega}-\left(\boldsymbol{u},\breve{G}^*q\right)_{0,\Omega}-\left(\breve{A}^{-1}\breve{G}^*q,\breve{G}^*p\right)_{0,\Omega}-(\lambda,q)_{0,\Omega}\\
\text{or}\qquad (\nabla\cdot\boldsymbol{u},q)_{0,\Omega}-(\lambda,q)_{0,\Omega}=0 \qquad \forall q\in\mathbf{Q}.
\end{aligned}$$

In summary we can combine the two set of the KKT-conditions to a new equivalent problem: Find a triple $(\boldsymbol{u},p,\lambda)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}\times\mathbf{\Lambda}$, that satisfy

$$\mu\,(\nabla\boldsymbol{u},\nabla\boldsymbol{\varphi})_{0,\Omega}-(p,\nabla\cdot\boldsymbol{\varphi})_{0,\Omega}=(\boldsymbol{f},\boldsymbol{\varphi})_{0,\Omega} \qquad \forall\boldsymbol{\varphi}\in\mathbf{V}\,, \tag{4.6}$$

$$(\nabla\cdot\boldsymbol{u},q)_{0,\Omega}-(\lambda,q)_{0,\Omega}=0 \qquad \forall q\in\mathbf{Q} \tag{4.7}$$

$$(\omega,p)_{0,\Omega}\geq 0 \qquad \forall\omega\in\mathbf{\Lambda}\,, \tag{4.8}$$

$$p\lambda=0 \quad \text{a.e. in } \Omega\,. \tag{4.9}$$

Unfortunately the operator $\breve{G}^*$ is not well defined for all elements of $\mathbf{Q}$. This is why this introduction of the the second Lagrange-multiplier should be considered a motivation for the introduction of the variational problem above. The more formal proof of equivalence of this and minimization problems can be found in the theorem 6.1.9.

Next, we consider modified versions of (4.7) and (4.8):

$$(\nabla\cdot\boldsymbol{u}-\lambda,\omega-\lambda)_{0,\Omega}=0 \qquad \forall\omega\in\mathbf{\Lambda} \tag{4.10}$$

$$(\omega-\lambda,p)_{0,\Omega}\geq 0 \qquad \forall\omega\in\mathbf{\Lambda}\,. \tag{4.11}$$

By dividing the inequality (4.11) by a constant $\delta>0$ and subtract it from the equality (4.10), we obtain

$$\left(\nabla\cdot\boldsymbol{u}-\frac{1}{\delta}p-\lambda,\omega-\lambda\right)_{0,\Omega}\leq 0 \qquad \forall\omega\in\mathbf{\Lambda}.$$

The lemma of projection operator A.0.3 allows us to interpret the inequality above as a projection

$$\lambda=\mathit{\Pi}_{\mathbf{\Lambda}}\left(\nabla\cdot\boldsymbol{u}-\frac{1}{\delta}p\right).$$

By applying the definition of the space $\mathbf{\Lambda}$ we obtain following variational problem: Find $(\boldsymbol{u},p)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}$ fulfilling the mixed formulation

$$\begin{aligned}
\mu\,(\nabla\boldsymbol{u},\nabla\boldsymbol{\varphi})_{0,\Omega}-(p,\nabla\cdot\boldsymbol{u})_{0,\Omega}=(\boldsymbol{f},\boldsymbol{\varphi})_{0,\Omega} \qquad & \forall\boldsymbol{\varphi}\in\mathbf{V}\,,\\
(\nabla\cdot\boldsymbol{u},q)_{0,\Omega}=(\lambda,q)_{0,\Omega} \qquad & \forall q\in\mathbf{Q}\\
\text{with}\qquad \lambda=\max\left\{0,\nabla\cdot\boldsymbol{u}-\frac{1}{\delta}p\right\}. &
\end{aligned}$$

We won't be using neither complementary condition $p\lambda = 0$ a.e. in Ω nor the fact that $p \in \mathbf{\Lambda}$ in our solution strategy explicitly, but this are important properties of the solution, which we evaluate in the chapter on error estimate.

4.3. First numerical results

In this section we compare results of two different approaches discussed in the previous section. The problem we consider is the cross section of a T-pipe with a viscous fluid, that can change in the gas state without the external pressure. Our research of the stokes problem with cavitation was initiated in cooperation with industry (see f.e. Gimbel et al [9]). In order to develop basic concepts for simulation we choose the T-pipe as prototype example. We set a non-zero flow profile, as the boundary condition the bottom end. The x-component of the flow profile is set to zero and the y-component can be calculated with the mapping $f_{flow} : [-1, 1] \to \mathbb{R}$ with $x \mapsto 10(1 - x^2)$. On the left and right side of the pipe we prescribe Neumann boundary condition as free flow out of the system. The external force is equals ten and oriented towards bottom, which represents the gravitation. The viscosity coefficient is set to one.

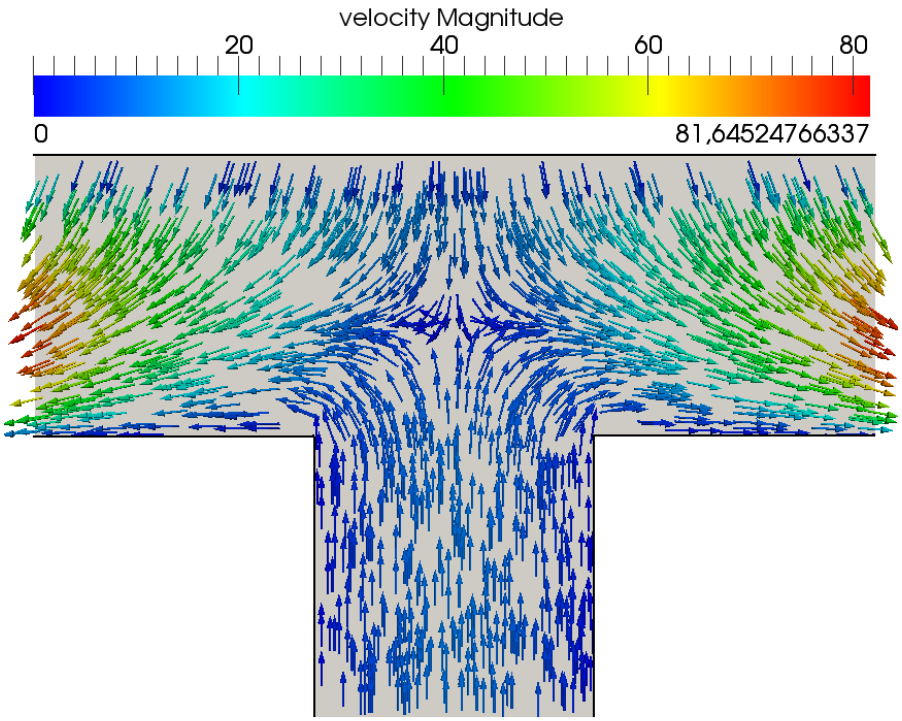

Figure 4.1.: The velocity vectors for Stokes problem with cavitation in the cross section of a T-pipe for the viscosity factor $\mu = 0.01$.

First we introduce the discrete finite-element-spaces. We divide Ω in triangular mesh $\mathbb{T}_h$ and, based on that mesh, we use the Croizeix-Raviart-elements or the non-conform

P_1-elements (see Braess [4, p. 103]) for $\mathbf{V}_h$ and the constant elements for $\mathbf{Q}_h$. This leads to the discrete problem

$$\mu\,(\nabla \boldsymbol{u}_h, \nabla \boldsymbol{\varphi}_h)_{0,\Omega} - (p_h, \nabla\cdot\boldsymbol{\varphi}_h)_{0,\Omega} = (\boldsymbol{f}, \boldsymbol{\varphi}_h)_{0,\Omega} \qquad \forall \boldsymbol{\varphi}_h \in \mathbf{V}_h\,, \tag{4.12}$$

$$(\nabla\cdot\boldsymbol{u}_h, q_h)_{0,\Omega} \geq 0 \qquad \forall q_h \in \mathit{\Pi}^{+}\mathbf{Q}_h, \tag{4.13}$$

$$(p_h, \nabla\cdot\boldsymbol{u}_h)_{0,T} = 0 \qquad \forall T \in \mathbb{T}_h\,. \tag{4.14}$$

Since all elements of the corresponding discrete space can be written as linear combinations, we define three appropriate vectors $\hat{\boldsymbol{u}}$, $\hat{\boldsymbol{u}}_0$ and $\hat{\boldsymbol{p}}$, such that $\boldsymbol{u}_h = \sum_j (\hat{u}_j + \hat{u}_{0,j})\boldsymbol{\varphi}_{h,j}$ and $p_h = \sum_j \hat{p}_j q_{h,j}$, where $\boldsymbol{\varphi}_{h,j}$ and $q_{h,j}$ are the basis functions of discrete spaces $\mathbf{V}_h$ and $\mathbf{Q}_h$ respectively. During the calculation process, we determine a vector $\hat{\boldsymbol{u}}_0$ in such a way, that the function $\sum_j \hat{u}_{0,j}\boldsymbol{\varphi}_{h,j}$ satisfy the non-zero boundary conditions and $\sum_j \hat{u}_j \boldsymbol{\varphi}_{h,j}\Big|_{\partial\Omega} = 0$. The operators $\breve{A}$ and $\breve{G}^*$ receive corresponding matrices A and G, defined as follows:

$$\begin{aligned} A_{i,j} &= \mu\,(\nabla\boldsymbol{\varphi}_{h,j}, \nabla\boldsymbol{\varphi}_{h,i})_{0,\Omega} \\ \text{and} \qquad G_{i,j} &= -\,(\nabla\cdot\boldsymbol{\varphi}_{h,i}, q_{h,j})_{0,\Omega}\,. \end{aligned}$$

At the same time we introduce two right hand side vectors $\hat{\boldsymbol{f}}$ and $\hat{\boldsymbol{g}}$ defined as

$$\begin{aligned} \hat{f}_j &= (\boldsymbol{f}, \boldsymbol{\varphi}_{h,i})_{0,\Omega} - (A\hat{\boldsymbol{u}}_0)_j \\ \text{and} \qquad \hat{g}_j &= (G\hat{\boldsymbol{u}}_0)_j \end{aligned}$$

respectively. This is a common way of transforming a problem into one with zero-boundary condition.

Next we consider the minimization problem (4.5) and get a discrete version of it: Find $\hat{\boldsymbol{p}} \geq 0$ such that

$$\frac{\mu}{2}\hat{\boldsymbol{p}}^{\mathrm{T}} G^{\mathrm{T}} A^{-1} G\hat{\boldsymbol{p}} - \hat{\boldsymbol{f}}^{\mathrm{T}} A^{-1} G\hat{\boldsymbol{p}} \leq \frac{\mu}{2}\hat{\boldsymbol{q}}^{\mathrm{T}} G^{\mathrm{T}} A^{-1} G\hat{\boldsymbol{q}} - \hat{\boldsymbol{f}}^{\mathrm{T}} A^{-1} G\hat{\boldsymbol{q}} \qquad \forall \hat{\boldsymbol{q}} \geq 0\,.$$

As mentioned before, we use conjugate gradient projection method, see Dembo and Tulowitzky [6], to solve this obstacle problem with the Schur complement $\mu G^{\mathrm{T}} A^{-1} G$ instead of the usual matrix and the vector $G^{\mathrm{T}} A^{-1}\hat{\boldsymbol{f}}$ as the right hand side. After the appropriate tolerance is achieved, we finally calculate $\hat{\boldsymbol{u}}$ using an equation derived from (4.12):

$$\hat{\boldsymbol{u}} = A^{-1}\left(\hat{\boldsymbol{f}} - G\hat{p}\right).$$

For our first solution strategy, in order to find the vectors $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{p}}$, we need to iterate the system

$$\begin{aligned} \mu A\hat{\boldsymbol{u}} + G\hat{\boldsymbol{p}} &= \hat{\boldsymbol{f}}, \\ G^{\mathrm{T}}\hat{\boldsymbol{u}} &= \hat{\boldsymbol{g}} + \hat{\boldsymbol{\lambda}}, \end{aligned}$$

cells	DoFs	iterations	tolerance	Calculation time in seconds setup	inverting A	solving
512	2,128	43	$4.51 \cdot 10^{-7}$	0.00526	0.00193	0.0365
2,048	8,352	54	$3.73 \cdot 10^{-7}$	0.01321	0.00718	0.3283
8,192	33,088	66	$5.51 \cdot 10^{-7}$	0.04162	0.08631	2.7729
32,768	131,712	74	$5.78 \cdot 10^{-7}$	0.16461	1.21509	21.862
131,072	525,568	108	$7.34 \cdot 10^{-7}$	0.74542	17.3248	214.9

Table 4.1.: Stokes problem with cavitation for the T-pipe example **(solved with cg projection method)**, until relative tolerance $\frac{\|p_{new}-p_{old}\|}{\|p_{new}+p_{old}\|} < 10^{-6}$, and with global refinement of cells

cells	DoFs	iterations	tolerance	Calculation time in seconds setup	inverting A	solving
512	2,128	10	$9.64 \cdot 10^{-7}$	0.00434	0.00077	0.0141
2,048	8,352	13	$9.15 \cdot 10^{-7}$	0.01144	0.00708	0.1046
8,192	33,088	22	$9.66 \cdot 10^{-7}$	0.03849	0.08584	1.0369
32,768	131,712	24	$9.78 \cdot 10^{-7}$	0.1629	1.21797	7.348
131,072	525,568	17	$9.97 \cdot 10^{-7}$	0.74385	17.3235	36.3318

Table 4.2.: Stokes problem with cavitation for the T-pipe example **(solved with Uzawa-Algorithm using projection operator as right hand side)**, until relative tolerance $\frac{\|p_{new}-p_{old}\|}{\|p_{new}+p_{old}\|} < 10^{-6}$, and with global refinement of cells

$$\text{with} \qquad \hat{\lambda}_j = \begin{cases} \left(G^{\mathrm{T}}(\hat{\boldsymbol{u}}+\hat{\boldsymbol{u}}_0)\right)_j - \frac{1}{\delta}\hat{p}_j & \text{if } \left(G^{\mathrm{T}}(\hat{\boldsymbol{u}}+\hat{\boldsymbol{u}}_0)\right)_j > \frac{1}{\delta}\hat{p}_j \\ 0 & \text{else} \end{cases} .$$

For this first numerical test we just use $\delta = 1$. In each iteration loop we "freeze" the vector $\hat{\boldsymbol{\lambda}}$ and search an approximation of the solution vectors $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{p}}$, using conjugate gradient Uzawa method (see Braess [4, p. 217]), until the quotient $\frac{\left\|\hat{\boldsymbol{p}}^{(k+1)}-\hat{\boldsymbol{p}}^{(k)}\right\|_2}{\left\|\hat{\boldsymbol{p}}^{(k)}\right\|_2}$ is below 10^{-3}. As the tests showed, the threshold of 10^{-3} is sufficient to proceed. Next we use this approximation to calculate the new vector $\hat{\boldsymbol{\lambda}}$ and repeat the process until the wanted precision is achieved.

In order to compare the algorithms above under fair condition we let them run until the relative tolerance $\frac{\left\|p^{(k+1)}-p^{(k)}\right\|}{\left\|p^{(k+1)}+p^{(k)}\right\|}$ is less then 10^{-6}. The tables 4.1 and 4.2 hold the results of the numerical tests conducted using the solving strategies, described above. The columns contain number of cells, number of degrees of freedom, that result from defining finite element space on each of this grids, number of iteration needed as well as actually achieved tolerance. We distinguish between inner and outer iteration cycles:

- the outer cycles are the number of actual fix-point-iterations,

- and the inner cycles are the number of steps for Uzawa-algorithms to calculate next $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{p}}$.

In case of our solution strategy we only count outer iteration cycles, since after a couple cycles, with between 3 to 13 Uzawa steps, we only need one Uzawa step pro cycle for the rest of the iteration. There are several calculation times, that are also included in those tables. They are split into three columns, two of which are more or less identical for the two strategies, since we use the same process to set up the problem and invert the resulting matrix A. The times in the last columns are quite different. According to this data, we can estimate that our first proposed solution strategy is about three to five times faster then the standard cg projection method for the given number of degrees of freedom.

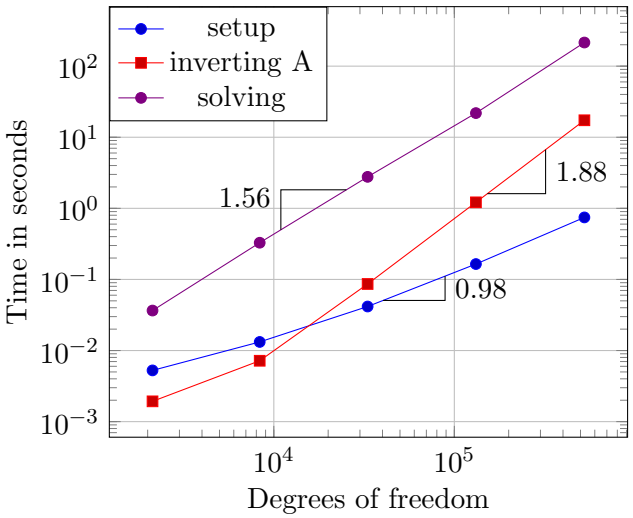

Figure 4.2.: A comparison of calculation times for the setup, the inverting matrix A and the actual iterations of cg projection method.

In the figures 4.2 and 4.3 we wanted to illustrate an increase of the calculation time relative to the growing number of degrees of freedom. Both the time needed for setup of the problem and the time for the inverting the matrix A increases dramatically with the number of degrees of freedom, but are only a fraction of the iteration time for the cg projection method. Furthermore on the graphic 4.3 we see that, not only the proposed strategy requires only a fraction of calculation time, the tests suggest, that the calculation time growth slop is decreasing with rising number of degrees of freedom.

4.4. Further analysis of the possible solution strategies

In this section we consider the possible Newton-type methods in order to increase convergence rate or, from the potential user point of view, in order to reduce computation time. Another aspect, that is worth addressing, is the number of degrees of freedom. Inevitably, the higher number of degree of freedom increases the calculation time. As a

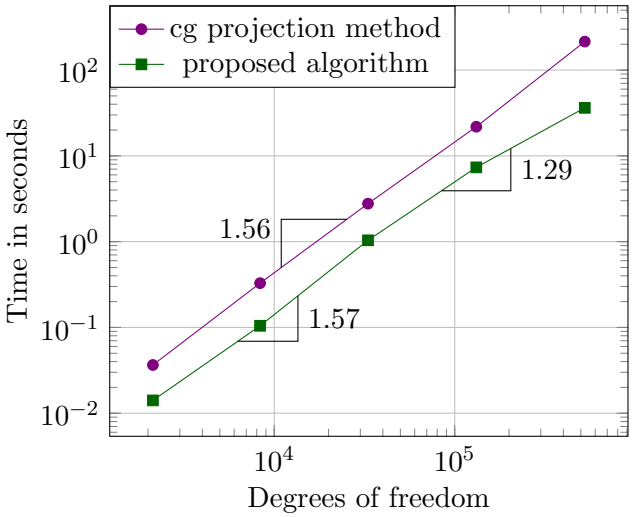

Figure 4.3.: A comparison of calculation times for the cg projection method and for the proposed algorithm with projection operator on the right hand side.

result there is a well know technique to refine only those parts of the mesh, that require more attention, see f. e. Braess [4, p. 173]. In order to be able to identify those parts we introduce our a posteriori error estimator. But a locally refined mesh, can also impose difficulties on the iteration process, that we will address in the section on further numerical results.

4.4.1. Newton-type methods

For any numerical algorithm we have to consider the start value and the iteration step.We start by looking on the possible iteration steps. For this purpose, we introduce a functional $\mathcal{F}_{\boldsymbol{\varphi} q} : \mathbf{V} \times \mathbf{Q} \longrightarrow \mathbb{R}$ with

$$\mathcal{F}_{\boldsymbol{\varphi} q}\left(\tilde{\boldsymbol{u}}, \tilde{p}\right) = \mathcal{A}(\tilde{\boldsymbol{u}}, \boldsymbol{\varphi}) - (\tilde{p}, \nabla \cdot \boldsymbol{\varphi})_{0,\Omega} - (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} + (\nabla \cdot \tilde{\boldsymbol{u}}, q)_{0,\Omega} - (\lambda(\tilde{\boldsymbol{u}}, \tilde{p}), q)_{0,\Omega} \ , \quad (4.15)$$

so that the functions $\boldsymbol{u}$ and p, which are the continuous solution of the variational formulation of the stokes problem with cavitation, are together a single zero spot of the functional $\mathcal{F}_{\boldsymbol{\varphi} q}$ for all $(\boldsymbol{\varphi}, q)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$. Since the functional $\mathcal{F}_{\boldsymbol{\varphi} q}$ is not Gateaux-differentiable, we cannot use the classical Newton method to find this zero spot. This is why, we introduce an element of the combined dual space $\mathcal{P}^*\left(\tilde{\boldsymbol{u}}, \tilde{p}\right) \in \mathbf{V}^* \times \mathbf{Q}^*$ as an approximation of the Gateaux-differential and the modified Newton step $(\boldsymbol{d}^u, d^p)^{\mathrm{T}}$, which can results from the equation

$$\left\langle \mathcal{P}^*\left(\tilde{\boldsymbol{u}}, \tilde{p}\right), (\boldsymbol{d}^u, d^p)^{\mathrm{T}} \right\rangle_{\mathbf{V}^* \times \mathbf{Q}^*} = \mathcal{F}_{\boldsymbol{\varphi} q}\left(\tilde{\boldsymbol{u}}, \tilde{p}\right).$$

Then the numerical solution can be obtained through the iteration process

$$\begin{pmatrix} \boldsymbol{u}^{(j+1)} \\ p^{(j+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}^{(j)} \\ p^{(j)} \end{pmatrix} - \begin{pmatrix} \boldsymbol{d}^u \\ d^p \end{pmatrix} \qquad \forall j \in \mathbb{N},$$

in which we use a vector $\boldsymbol{u}^{(0)}, p^{(0)}$ as start value. Of course, even though we define this $\mathcal{P}^*(\tilde{\boldsymbol{u}}, \tilde{p})$ as an element of the dual space to the combined continuous spaces $\mathbf{V}$ and $\mathbf{Q}$, we have an application on the Finite-Element-Spaces in mind. The purpose of this excise is to motivate the choices for the update calculation method and not to complicate the matter with choices of the appropriate Finite-Element-Spaces.

The first approximation of the Gateaux-differential, that come to mind, is the simplest one, such that ignores the dependency of λ on $\tilde{\boldsymbol{u}}$ and $\tilde{p}$, so that

$$\left\langle \mathcal{P}_1^*(\tilde{\boldsymbol{u}}, \tilde{p}), (\boldsymbol{d}^u, d^p)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} = \mathcal{A}(\boldsymbol{d}^u, \boldsymbol{\varphi}) - (d^p, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + (\nabla\cdot\boldsymbol{d}^u, q)_{0,\Omega}\,.$$

Next possibility to consider is the fact, that λ can be obtained using

$$\max\left\{0, \nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right\} = \frac{1}{2}\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right) + \frac{1}{2}\left|\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right|.$$

So for the next step we ignore absolute value function, which results in an another approximation of the Gateaux-differential

$$\left\langle \mathcal{P}_2^*(\tilde{\boldsymbol{u}}, \tilde{p}), (\boldsymbol{d}^u, d^p)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} = \mathcal{A}(\boldsymbol{d}^u, \boldsymbol{\varphi}) - (d^p, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + (\nabla\cdot\boldsymbol{d}^u, q)_{0,\Omega} + \frac{1}{\delta}(p, q)_{0,\Omega}\,.$$

The next step in order to obtain another approximation of the Gateaux-differential is to regularise the absolute value function, but before we move on, we modify the first equation of this new mixed formulation. First, we consider the mixed formulation with the inequality instead the projection operator

$$\begin{array}{lll} \mu(\nabla\boldsymbol{u}, \nabla\boldsymbol{\varphi})_{0,\Omega} & -(\nabla\cdot\boldsymbol{\varphi}, p)_{0,\Omega} & = (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega}\,, \\ (\nabla\cdot\boldsymbol{u}, q)_{0,\Omega} & -(\lambda, q)_{0,\Omega} & = 0\,, \\ -\delta(\nabla\cdot\boldsymbol{u}, \omega-\lambda)_{0,\Omega} & +(p, \omega-\lambda)_{0,\Omega} \quad +\delta(\lambda, \omega-\lambda)_{0,\Omega} & \geq 0\,. \end{array}$$

Written this way, it is obvious, that the problem has in a way almost a symmetric structure. We can obtain the missing part in the "right upper corner" by adding the second equation multiplied with factor $-\delta$ to the first one. Also we use $\nabla\cdot\boldsymbol{\varphi}$ instead of q as a test function to fit the variational formulation. This leads to

$$\begin{aligned} \mu(\nabla\boldsymbol{u}, \nabla\boldsymbol{\varphi})_{0,\Omega} - \delta(\nabla\cdot\boldsymbol{u}, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (p, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + \delta(\lambda, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} &= (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} \quad \forall\boldsymbol{\varphi}\in\mathbf{V}, \\ (\nabla\cdot\boldsymbol{u}, q)_{0,\Omega} - (\lambda, q)_{0,\Omega} &= 0 \qquad\qquad \forall q\in\mathbf{Q}, \\ \text{with} \qquad \lambda = \max\left\{0, \nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right\} &\text{ a.e. on } \Omega. \end{aligned}$$

This stabilizes the problem and makes the analysis much easier. The benefits of this alteration will be noticeable in the sections 4.6. Introduction of a bilinear form

$$\mathcal{A}^\delta(\boldsymbol{u}, \boldsymbol{\varphi}) = \mu(\nabla\boldsymbol{u}, \nabla\boldsymbol{\varphi})_{0,\Omega} - \delta(\nabla\cdot\boldsymbol{u}, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega}$$

enables more compact notation. Also the value of the Lagrange multiplier λ can be obtain throw square root function. In summary, we have the equivalent formulation of

the problem

$$\mathcal{A}^\delta(\boldsymbol{u},\boldsymbol{\varphi}) - (p, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + \delta\,(\lambda, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} = (\boldsymbol{f},\boldsymbol{\varphi})_{0,\Omega} \qquad \forall\boldsymbol{\varphi}\in\mathbf{V}, \tag{4.16}$$

$$(\nabla\cdot\boldsymbol{u}, q)_{0,\Omega} - (\lambda, q)_{0,\Omega} = 0 \qquad \forall q\in\mathbf{Q}, \tag{4.17}$$

$$\text{with}\quad \lambda = \frac{1}{2}\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right) + \frac{1}{2}\sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2}\ \text{a.e. on } \Omega. \tag{4.18}$$

Furthermore now regularized version of λ can be obtained using

$$\lambda^\xi = \frac{1}{2}\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right) + \frac{1}{2}\sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2 + \xi}\ \text{a.e. on } \Omega,$$

where ξ is a positive constant. With this in mind, we define a functional $\tilde{\mathcal{F}}_{\boldsymbol{\varphi}q} : \mathbf{V}\times\mathbf{Q} \longrightarrow \mathbb{R}$ so that

$$\begin{aligned}\tilde{\mathcal{F}}_{\boldsymbol{\varphi}q}(\tilde{\boldsymbol{u}},\tilde{p}) =& \mathcal{A}^\delta(\tilde{\boldsymbol{u}},\boldsymbol{\varphi}) - (\tilde{p}, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + \delta\left(\tilde{\lambda}^\xi, \nabla\cdot\boldsymbol{\varphi}\right)_{0,\Omega} - (\boldsymbol{f},\boldsymbol{\varphi})_{0,\Omega}\\ &+ (\nabla\cdot\tilde{\boldsymbol{u}}, q)_{0,\Omega} - \left(\tilde{\lambda}^\xi, q\right)_{0,\Omega}\end{aligned} \tag{4.19}$$

$$\text{as well as}\quad \tilde{\lambda}^\xi = \frac{1}{2}\left(\nabla\cdot\tilde{\boldsymbol{u}} - \frac{1}{\delta}\tilde{p}\right) + \frac{1}{2}\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{u}} - \frac{1}{\delta}\tilde{p}\right)^2 + \xi}\ \text{a.e. on } \Omega.$$

In the section 4.6 we discuss the conditions for the existence and the uniqueness of the zero spot of this functional, as well as its relation to the function, that we actually want to calculate. Since this stabilised regularise functional is Gateaux-differentiable, we can use its derivative to calculate modified Newton step, meaning

$$\begin{aligned}\left\langle \mathcal{P}_3^*(\tilde{\boldsymbol{u}},\tilde{p}), (\boldsymbol{d}^u, d^p)^{\mathrm{T}}\right\rangle_{\mathbf{V}\times\mathbf{Q}} &= \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}q}(\tilde{\boldsymbol{u}},\tilde{p}), (\boldsymbol{d}^u, d^p)^{\mathrm{T}}\right\rangle_{\mathbf{V}^*\times\mathbf{Q}^*}\\ &= \mathcal{A}^\delta(\boldsymbol{d}^u,\boldsymbol{\varphi}) - (d^p, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} + (\nabla\cdot\boldsymbol{d}^u, q)_{0,\Omega}\\ &\quad + \left(v^\xi\left(\nabla\cdot\boldsymbol{d}^u - \frac{1}{\delta}d^p\right), \delta\,(\nabla\cdot\boldsymbol{\varphi}) - q\right)_{0,\Omega},\end{aligned}$$

$$\text{with } v^\xi\in\mathbf{Q} \text{ and } v^\xi = \frac{1}{2}\left(1 + \frac{\nabla\cdot\tilde{\boldsymbol{u}} - \frac{1}{\delta}\tilde{p}}{\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{u}} - \frac{1}{\delta}\tilde{p}\right)^2 + \xi}}\right)\ \text{a.e. on } \Omega.$$

4.4.2. Possible start values

In the section 4.3 we already have seen that, since we impose boundary values on the velocity $\boldsymbol{u}_h$, we split the coefficient vector in $\hat{\boldsymbol{u}}_0$ and $\hat{\boldsymbol{u}}$. The vector $\hat{\boldsymbol{u}}_0$ satisfy the non-zero boundary conditions and one obvious start value for the vector $\hat{\boldsymbol{u}}$ would be $\mathbf{0}$.

On the other hand, it is possible for the algorithms to profit from a start value, that is already near the goal value. Since calculation for the grid with fewer degrees of freedom are much faster, it is interesting to assess the cascade approach. In this case, similar to the local refinement strategy, we conduct several calculation in a row, using ever further refined mesh in each step and the solution from the last mesh to calculate a start value for the new iteration.

4.4.3. A posteriori error estimator

The are several error estimators for the stokes problem without cavitation (see for example Verfürth [16] or Nobile [15]). Let $(\boldsymbol{u}, p)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$ be the continuous solution of the mixed variation formulation for the stokes problem with cavitation, given by (4.1) to (4.3), and $(\boldsymbol{u}_h, p_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ is the discrete solution, where $\mathbf{V_h} \subseteq \mathbf{V}$ and $\mathbf{Q_h} \subseteq \mathbf{Q}$, then the norm of the error $\left\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}$ can be estimated with the inequality

$$\left\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} \leq c\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) \ ,$$

where

$$\begin{aligned}\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T\in\mathbb{T}_h} \left(|T| \left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h + \nabla p_h\right\|^2_{0,T} + |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial n} - p_h\boldsymbol{n} \right\|^2_{0,\partial T\setminus\partial\Omega} \right) \\ + \left\|\nabla\cdot\boldsymbol{u}_h - \lambda_h\right\|^2_{0,\Omega} + (\mathit{\Pi}_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + \left\|\mathit{\Pi}_{\mathbf{\Lambda}} p_h - p_h\right\|^2_{0,\Omega}\end{aligned}$$

and

$$\lambda_h = \mathit{\Pi}_{\mathbf{\Lambda}_h}\left(\nabla\cdot\boldsymbol{u}_h - \frac{1}{\delta}p_h\right) \ .$$

Since the norm $\left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h + p_h\right\|_{0,T}$ is part of the error estimator, it can be classified as an residual type. The flow between the cell must satisfy the Neumann boundary condition on all edges of the cells, which a not part of $\partial\Omega$. This is measured by the norm $\left\| \frac{\partial \boldsymbol{u}_h}{\partial n} - p_h\boldsymbol{n} \right\|_{0,\partial T\setminus\partial\Omega}$. In addition to the residual on each cell, which is related with the equation (4.16), the estimator take account of the fulfilment of secondery condition (4.17), in the form of the norm $\left\|\nabla\cdot\boldsymbol{u}_h - \lambda_h\right\|_{0,\Omega}$. Since the function p was introduced as Lagrange multiplier, according to the KKT conditions the functions p and as result all the functions p_h must be greater or equal to zero almost everywhere on Ω. The norm $\left\|\mathit{\Pi}_{\mathbf{\Lambda}} p_h - p_h\right\|_{0,\Omega}$ measures the deviation of numerical solution from this condition. Finally, the scalar product $(\mathit{\Pi}_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega}$ reviews compliance of the complementary condition, which is also one of the KKT conditions. Especially the last to components of the error estimator are not really important, if you want to evaluate convergence rate of the solution on the entire subset Ω, but their local qualities, allow us to construct an efficient mesh. We will concentrate on the complete formal proof of the proposed a error estimator in the section 4.7. The rigorous proof is postponed to the section 4.7.

Further more, the discrete space $\mathbf{V}_h$, that we introduced in the section 4.3 is not conform, meaning $\mathbf{V}_h \not\subset \mathbf{V}$. In order to adapt this error estimator for the introduced Finite-Element-Space, we have to add another component into the mix. The norm $\left\|\nabla\left(\boldsymbol{v}_h - \boldsymbol{u}_h^{nc}\right)\right\|_{0,\Omega}$ measures the difference between the velocity approximation $\boldsymbol{u}_h^{nc}$ in the non-conform space and an element of one discrete subspace. More about the process to determinate appropriate $\boldsymbol{v}_h$ can be found in the section 4.5. Also the previous approach for deriving the error estimator would not work in the non-conform case. This is why, in the second part of the section 4.7 the alternative approach is introduced.

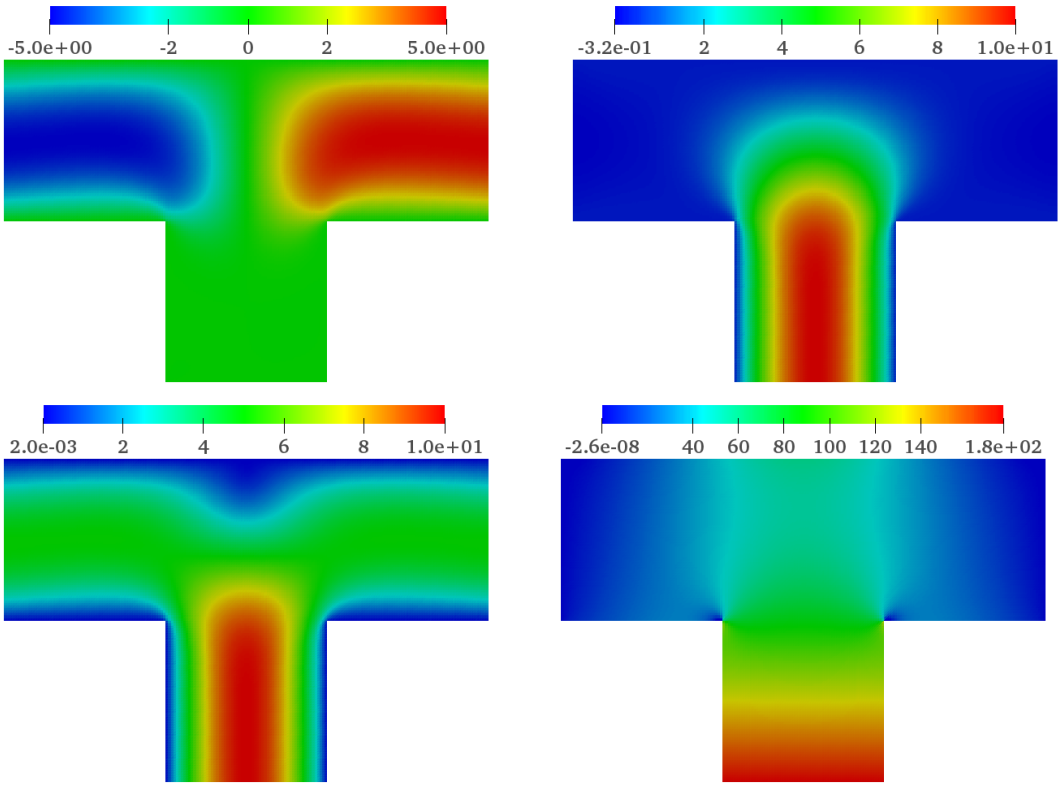

Figure 4.4.: The solution of the Stokes problem with cavitation for $\mu = 1$. Left above: x-component of the velocity. Right above: y-component of the velocity. Left below: absolute value of the velocity. Right below: pressure.

4.5. Numerical results

4.5.1. A comparison of proposed Newton-type iteration methods

For the further test we have to consider the effects of the viscosity constant μ on the simulation. As in the section 4.3, we calculate the flow in the T-pipe section, but different viscosity factors lead to completely different behavior (compare figures 4.4, 4.5 and 4.6).

The main reason for that is, that the differences in the viscosity lead to the different pressure distributions and, as a result of that, the different cavitation zones. If the pressure in a segment is too low, the cavitation effect takes place. As shown in the figure 4.7, in the example with $\mu = 1$ there are two small separate cavitation zones near the openings on the left and the right sides. In case of $\mu = 0.1$ the zones are still separate, but are much bigger. The simulation with the viscosity constant $\mu = 0.01$ shows one continuous cavitation zone in the upper part of the T-pipe segment. In the section 4.3 we only considered a fix point iteration approach for $\mu = 1$, and achieved superior result, compared with the cg projection method. In case of $\mu = 0.1$ this approach is still faster method for the larger number of degrees of freedom, but does not outperform

the cg projection method that drastically. That is why, we focus in this chapter on the numerical test for the case $\mu = 0.1$.

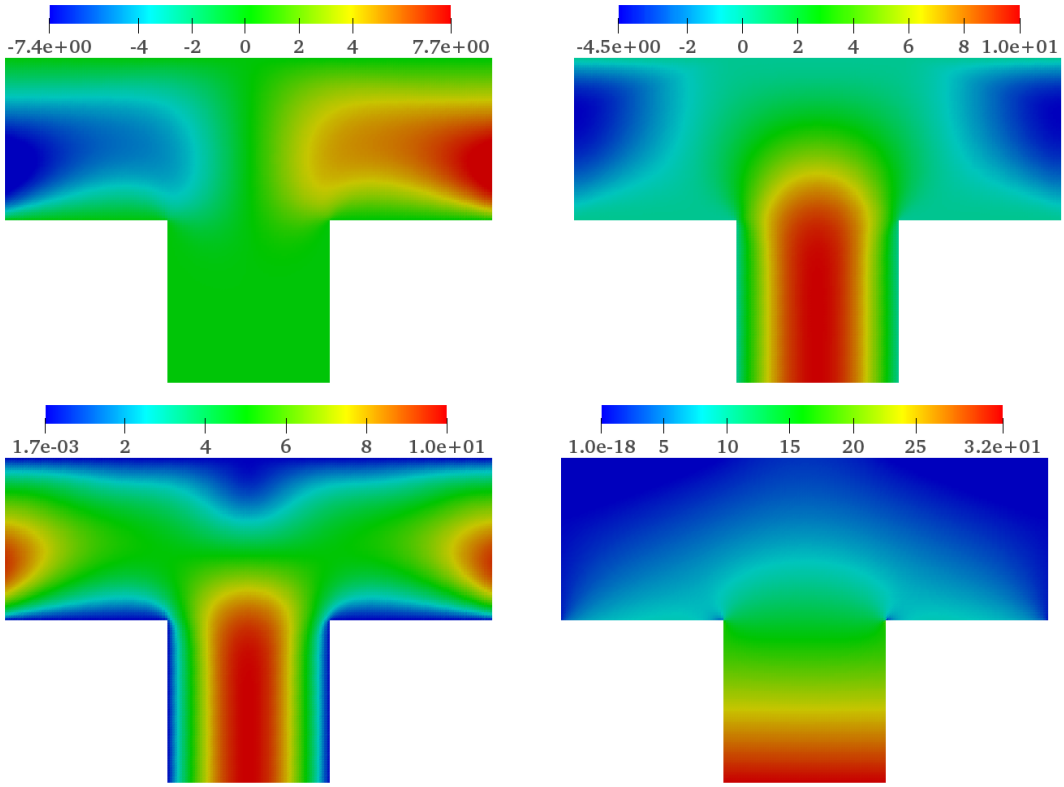

Figure 4.5.: The solution of the Stokes problem with cavitation for $\mu = 0.1$. Left above: x-component of the velocity. Right above: y-component of the velocity. Left below: absolute value of the velocity. Right below: pressure.

The other important point to discuss is the matrix A. In the section 4.3 we used $A_{i,j} = \mu\,(\nabla\boldsymbol{\varphi}_{h,j}, \nabla\boldsymbol{\varphi}_{h,i})_{0,\Omega}$. Our velocity solution vectors $\hat{\boldsymbol{u}}_0$ and $\hat{\boldsymbol{u}}$ consist of to blocks each (for example first half for x-components and the second half for the y-components). This result in a block structure of the matrix, such that for all $\boldsymbol{w}_h, \boldsymbol{v}_h \in \mathbf{V}_h$

$$\mathcal{A}(\boldsymbol{w}_h, \boldsymbol{v}_h) = \hat{\boldsymbol{w}}^{\mathrm{T}} \begin{pmatrix} A_{xx} & 0 \\ 0 & A_{xx} \end{pmatrix} \hat{\boldsymbol{v}} \quad \text{with } A_{xx} = \mu \begin{pmatrix} & \vdots & \\ \cdots & (\nabla\varphi_{h,j}, \nabla\varphi_{h,i})_{0,\Omega} & \cdots \\ & \vdots & \end{pmatrix}.$$

So instead of calculate the inverse of the whole matrix, we can take advantage of the structure, only invert one block after the assembling step and use it in combination with specially written matrix-vector-product routine in each iteration step. This works for the Newton-type methods P_1 and P_2 as well, but in the Newton-type methods P_3 we added the stabilization, which resulted in the bilinear form

$$\mathcal{A}^{\delta}(\boldsymbol{u}, \boldsymbol{\varphi}) = \mu\,(\nabla\boldsymbol{u}, \nabla\boldsymbol{\varphi})_{0,\Omega} - \delta\,(\nabla\cdot\boldsymbol{u}, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega}\,.$$

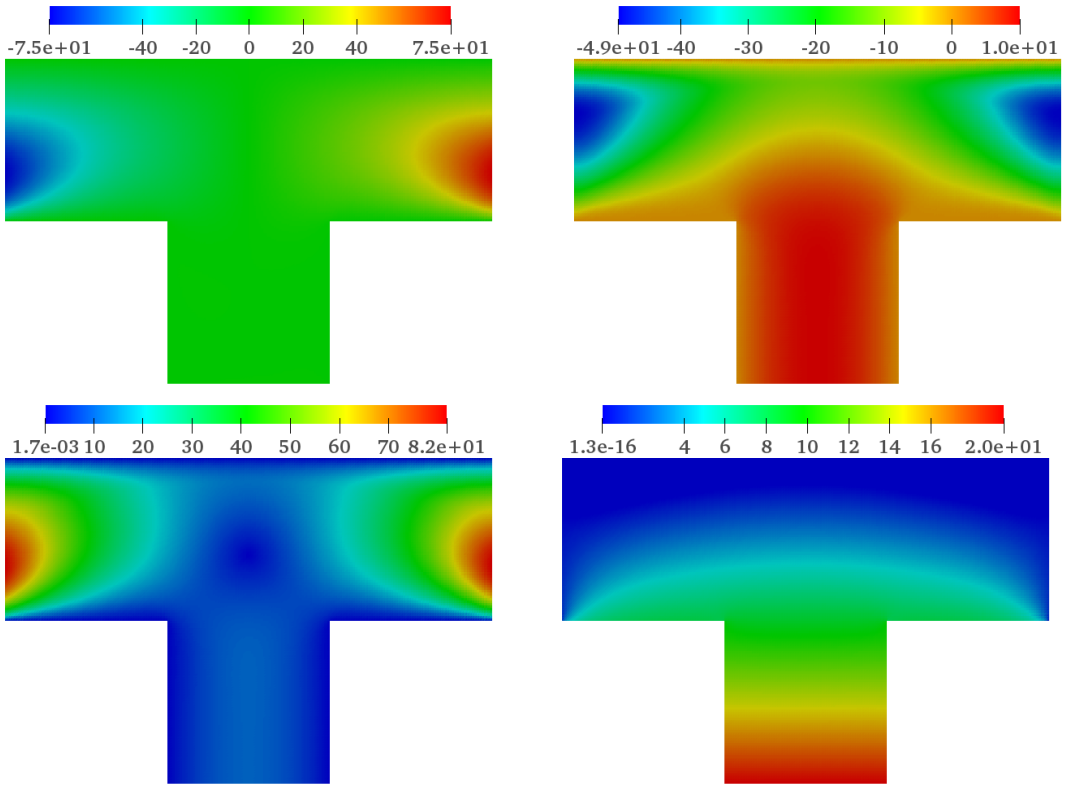

Figure 4.6.: The solution of the Stokes problem with cavitation for $\mu = 0.01$. Left above: x-component of the velocity. Right above: y-component of the velocity. Left below: absolute value of the velocity. Right below: pressure.

Figure 4.7.: The cavitation zone in the T-pipe segment for different viscosity factors: left to right $\mu = 1$, $\mu = 0.1$ and $\mu = 0.01$.

The resulting matrix is still very sparse and we can benefit from the block structure, but it is not possible to invert the whole matrix at the same "low cost" as previously. Further more, since the focus of this work was not on efficient algebraic computation, we decided to make quick numerical calculation of A^{-1} in each step and subtract the time from the overall calculation time. This way we didn't have to program an inverting routine for the block matrix and concentrate on the algorithm itself. That is why, the solving time for the P_3 method must be adjusted. In the subsequent comparison of the results we will be using this adjusted time, when referring to the P_3 method.

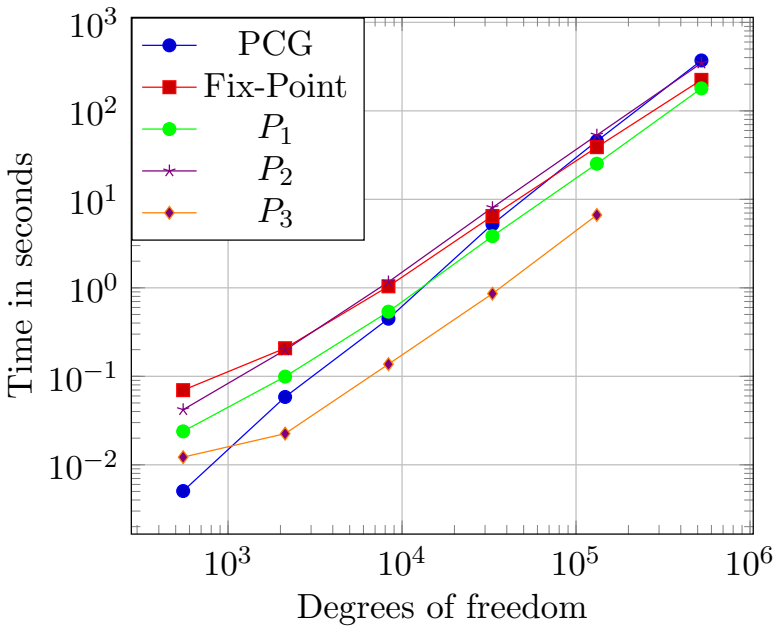

Figure 4.8.: A comparison of solving strategies of Stokes problem with cavitation for the T-pipe example with $\mu = 0.1$ **(zero start value and with global refinement of cells)**.

For the first numerical test we compare the calculation times of the different solution strategies for zero start value and with global refinement of cells. As illustrated in the figure 4.8, the P_2 update calculation had, as suspected, some oscillation problems, which resulted in the worst calculation time. That is why, in the comparison table 4.3 we left it out. The Fix-Point-Approach was, as stated before, better as the projected cg method. What is more important, it has shown a lower exponential growth of the calculation time in dependence on the growing number of degrees of freedom. The P_1 solution update strategy has shown similar results. Even though the calculation times were low than in the Fix-Point-Approach, they grow a little bit faster with the rising number of degrees of freedom. The solution update strategy P_3 required about the third of the calculation time of P_1 and less then a fifth part of the calculation time of the projected cg method.

Next we try the cascade approach and keep the pressure from the previous mesh to calculate the start value for the new one. The reason to use the pressure is, that we can easily interpolate it. In order to generate a new mesh we split the triangles of the old mesh into a number of new triangles. Since in our Finite-Element-Space pressure is constant on each triangular mesh element, we can just assign the pressure value from the "parent" cell to the all "child" cell, that are the result of refinement of the old "parent" cell in question. As already stated, we calculate the inverse of the matrix A, so the start value for velocity can be easily obtain using the equation (4.12).

As shown in the figure 4.9, Fix-Point-Iteration and P_2 approach profit the least from this new starting value. The projected cg method in the cascade calculation shows

DoFs	PCG		Fix-Point		P_1		P_3	
n	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$
552	0.005		0.07		0.024		0.012	
2,128	0.058	1.81	0.208	0.81	0.099	1.06	0.022	0.45
8,352	0.448	1.49	1.04	1.18	0.538	1.24	0.137	1.32
33,088	5.24	1.79	6.485	1.33	3.832	1.43	0.859	1.33
131,712	45.624	1.57	38.891	1.3	25.266	1.37	6.659	1.48
525,568	371.271	1.51	223.979	1.27	179.465	1.42		

Table 4.3.: The calculation times and their dependence on the number of degrees of freedom for Stokes problem with cavitation in the T-pipe with $\mu = 0.1$ **(zero start value and with global refinement of cells)**.

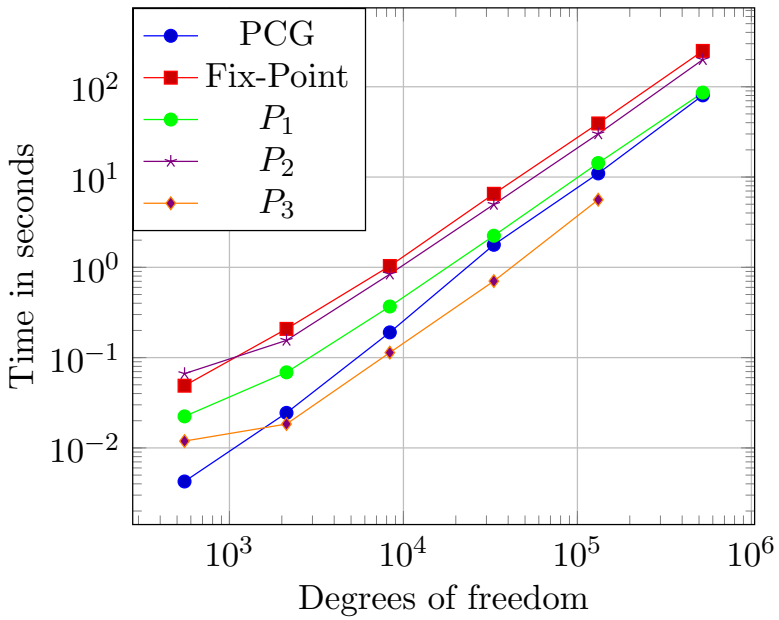

Figure 4.9.: A comparison of solving strategies of Stokes problem with cavitation for the T-pipe example with $\mu = 0.1$ **(interpolation of previous solution as start value and with global refinement of cells)**.

DoFs	PCG		Fix-Point		P_1		P_3	
n	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$
552	0.004		0.049		0.022		0.012	
2,128	0.024	1.3	0.208	1.07	0.069	0.83	0.018	0.32
8,352	0.19	1.5	1.035	1.17	0.368	1.23	0.114	1.33
33,088	1.777	1.62	6.522	1.34	2.239	1.31	0.704	1.32
131,712	10.969	1.32	39.239	1.3	14.288	1.34	5.605	1.5
525,568	80.308	1.44	249.394	1.34	86.138	1.3		

Table 4.4.: The calculation times and their dependence on the number of degrees of freedom for Stokes problem with cavitation in the T-pipe with $\mu = 0.1$ **(interpolation of previous solution as start value and with global refinement of cells)**.

better result then all proposed strategies except for the P_3 Newton-type method. The calculation times in the table 4.4 show, that all methods can profit from a better choice of the start value, but we have to take the calculation time for this start value into account two. So under this condition the cascade calculation is valuable strategy for the projected cg method and P_1 Newton-type method, but not for P_3 approach, which still has shown the best results.

The in the third test we introduce the local cell refinement. In the base problem with the T-pipe segment we are in particular interested in value in two point. Those are the points, where the pipes meet. The 90 degree angle leads to the artificial singularity in the flow profile. Our error estimator recognizes the problem and the cell in the proximity are refined in each step. The other interesting zone is the border to the area, where the cavitation take place. Refining those cells makes more efficient use of degrees of freedom, but we also have to take care of the hanging nodes. This results in less optimal triangles in the mesh. Also since we use the non-conform Finite-Element-Space $\mathbf{V}_h$, we need to calculate a conform approximation of the velocity. In the conform Finite-Element-Space with linear triangle elements (see f.e. Braess [4, p. 62 ff]) we need to know the function value in the nodes and in the Croizeix-Raviart-elements the value is attributed to the sides of the cell. So one of the simplest ways to obtain conform approximation from the non-conform solution with the Croizeix-Raviart-elements is to calculate in every node the average between the corresponding values of the cell sides, that have the node in common.

As the stability becomes more of an issue as before all solution strategies show an increase in the calculation time due to the oscillations. The Fix-Point-Iteration calculations were terminated early, because it was obvious, that they are not comparable with the rest. The methods P_1 and P_2 started slower as projected cg method, but P_1 showed better results for the higher number of degrees of freedom. And again was the solution strategy P_3 exceptionally fast due to the still much lower number of iteration circles, even so those iteration were more complex.

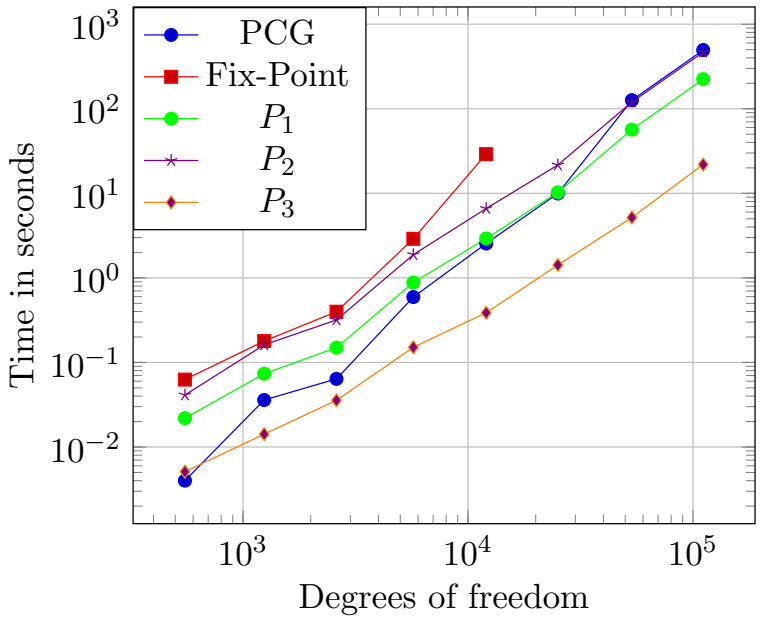

Figure 4.10.: A comparison of solving strategies of Stokes problem with cavitation for the T-pipe example with $\mu = 0.1$ **(zero start value and with local refinement of cells)**.

DoFs	PCG		P_1		P_2		P_3	
n	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$
552	0.004		0.022		0.041		0.005	
1,242	0.036	2.71	0.074	1.49	0.161	1.68	0.014	1.26
2,603	0.064	0.78	0.15	0.96	0.318	0.92	0.036	1.24
5,717	0.593	2.83	0.88	2.25	1.883	2.26	0.151	1.83
12,050	2.549	1.95	2.923	1.61	6.613	1.68	0.387	1.26
25,094	9.948	1.86	10.265	1.71	21.638	1.62	1.42	1.77
53,532	126.294	3.35	56.69	2.26	120.23	2.26	5.18	1.71
111,057	494.824	1.87	223.872	1.88	466.614	1.86	21.904	1.98

Table 4.5.: The calculation times and their dependence on the number of degrees of freedom for Stokes problem with cavitation in the T-pipe with $\mu = 0.1$ **(zero start value and with local refinement of cells)**.

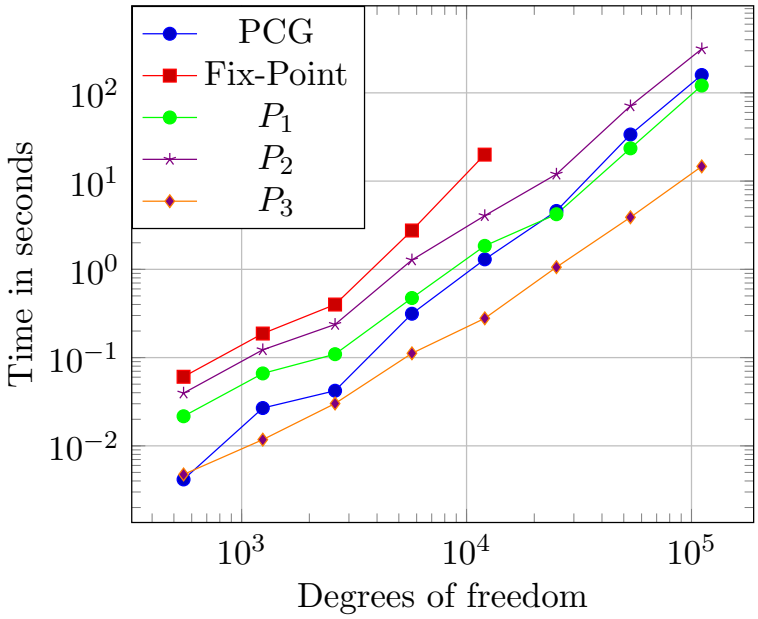

Figure 4.11.: A comparison of solving strategies of Stokes problem with cavitation for the T-pipe example with $\mu = 0.1$ **(interpolation of previous solution as start value and with local refinement of cells)**.

The last test with $\mu = 0.1$, were we use interpolation of previous solution as start value and an error estimator for the local refinement of cells, reviles the similar picture as the test with the global refinement of cells. All solution strategies profit from the better choice of the start value. The least improvement of the calculation time shows the method P_3, but it is still far more efficient then the rest of the competition.

The numbers for the case with viscosity factor $\mu = 0.01$ are similar. We limit our self to report the result of only one test as an example of the overall trend. The projection cg method can outperform all other solution strategies, except for the P_3 algorithm, but the results also suggest, that calculation time will increase faster with the growing number of degrees of freedom. Even so the solution strategy P_3 is superior in both calculation time and its growth due to the rising number of degrees of freedom. Since the data indicates the overall success of this method, we will analyze it deeper in the following section.

4.5.2. An examination of the convergence rates depending on the mesh size

In this part we compare, how does the refinement of the mesh affects the accuracy of the calculated solution. For this purpose, we consider the results of the tests, in which interpolation of previous solution was used as start value and with local refinement of cells. Since we already have an error estimator, it is interesting to look at this data first. The tables 4.8 and 4.9 contain the raw values for different parts of the error estimator,

DoFs	PCG		P_1		P_2		P_3	
n	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$
552	0.004		0.022		0.04		0.005	
1,242	0.027	2.3	0.066	1.38	0.123	1.39	0.012	1.12
2,603	0.042	0.61	0.109	0.68	0.237	0.89	0.03	1.27
5,717	0.314	2.55	0.473	1.86	1.277	2.14	0.112	1.66
12,050	1.298	1.9	1.845	1.82	4.058	1.55	0.279	1.22
25,094	4.587	1.72	4.215	1.13	11.967	1.47	1.058	1.82
53,532	33.719	2.63	23.51	2.27	71.145	2.35	3.891	1.72
111,057	159.344	2.13	121.012	2.25	315.004	2.04	14.678	1.82

Table 4.6.: The calculation times and their dependence on the number of degrees of freedom for Stokes problem with cavitation in the T-pipe with $\mu = 0.1$ **(interpolation of previous solution as start value and with local refinement of cells)**.

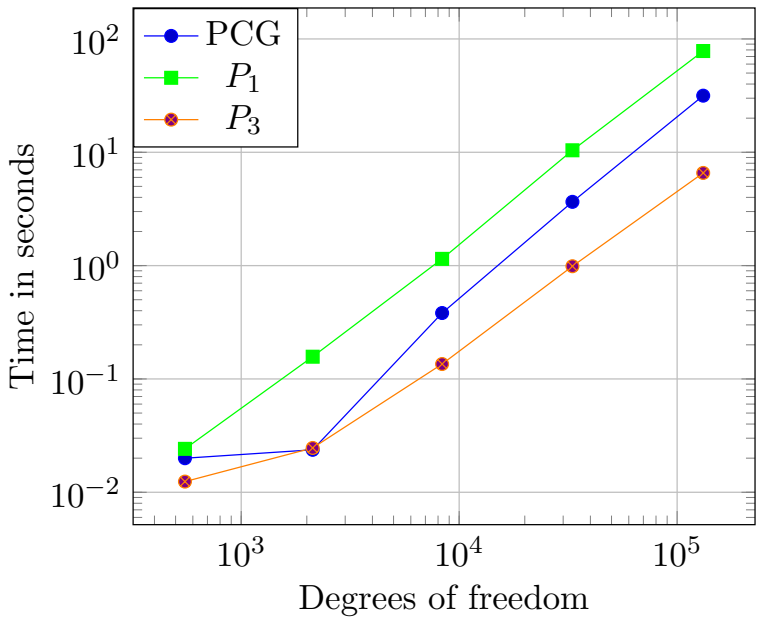

Figure 4.12.: A comparison of solving strategies of Stokes problem with cavitation for the T-pipe example with $\mu = 0.01$ **(zero start value and with global refinement of cells)**.

DoFs	PCG		P_1		P_3	
n	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$	Time	$\mathcal{O}(n)$
552	0.02		0.024		0.012	
2,128	0.024	0.13	0.157	1.39	0.025	0.51
8,352	0.382	2.03	1.147	1.45	0.135	1.25
33,088	3.654	1.64	10.412	1.6	0.987	1.44
131,712	31.588	1.56	78.254	1.46	6.574	1.37

Table 4.7.: The calculation times and their dependence on the number of degrees of freedom for Stokes problem with cavitation in the T-pipe with $\mu = 0.01$ **(zero start value and with global refinement of cells)**.

	Different errors					
cells	e^2_{res}	e^2_{jump}	e^2_{cond}	e^2_{compl}	e^2_{p+}	e^2_{nc}
128	12.5	194.5	$1.62 \cdot 10^{-9}$	0	0	344.02
512	3.13	15.9	$4.94 \cdot 10^{-9}$	0	0	117.22
2,048	0.78	1.19	$3.65 \cdot 10^{-9}$	0	0	38.68
8,192	0.2	$8.92 \cdot 10^{-2}$	$3.95 \cdot 10^{-9}$	0	0	13.64
32,768	$4.88 \cdot 10^{-2}$	$7.1 \cdot 10^{-3}$	$5.15 \cdot 10^{-9}$	0	0	5.3
$1.31 \cdot 10^5$	$1.22 \cdot 10^{-2}$	$6.03 \cdot 10^{-4}$	$4.46 \cdot 10^{-8}$	0	0	2.24

Table 4.8.: The error estimation data for **the projected cg method with global refinement**, applied on Stokes problem with cavitation in the T-pipe with $\mu = 0.1$.

that were introduced in the section 4.4:

$$e^2_{res} = \sum_{T\in\mathbb{T}_h} \left(|T| \, \|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h + \nabla p_h\|^2_{0,T} \right), \qquad e^2_{compl} = (\Pi_{\boldsymbol{\Lambda}} p_h, \lambda_h)_{0,\Omega},$$
$$e^2_{jump} = \sum_{T\in\mathbb{T}_h} \left(|\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial n} - p_h \boldsymbol{n} \right\|^2_{0,\partial T\setminus\partial\Omega} \right), \qquad e^2_{p+} = \|\Pi_{\boldsymbol{\Lambda}} p_h - p_h\|^2_{0,\Omega},$$
$$e^2_{cond} = \|\nabla\cdot\boldsymbol{u}_h - \lambda_h\|^2_{0,\Omega}, \qquad e^2_{nc} = \|\nabla(\boldsymbol{v}_h - \boldsymbol{u}^{nc}_h)\|^2_{0,\Omega}.$$

As already mentioned, the parts e_{cond}, e_{compl} and e_{p+} contribute relatively little to the global error estimation and are more interesting in the selecting process for the local cell refinement. In the case of the projected cg method e_{compl} and e_{p+} are zero, since this is essential part of the post-processing. The error e_{nc}, that is the result of the use of the non-conform elements, is dominant overall, but it is not the best candidate for evaluating convergence rates, since we use relatively unsophisticated method for calculating $\boldsymbol{u}^{nc}_h$. Different methods might be more valuable in this regard, but we have other data available. The residual part e_{res} is usually very important would serve the tusk quite well, but since we used the Croizeix-Raviart-elements for $\mathbf{V}_h$ and the constant elements for $\mathbf{Q}_h$, it is reduced to $e^2_{res} = \sum_{T\in\mathbb{T}_h} \left(|T| \, \|\boldsymbol{f}\|^2_{0,T} \right)$. This leaves us with another impor-

	Different errors					
cells	e^2_{res}	e^2_{jump}	e^2_{cond}	e^2_{compl}	e^2_{p+}	e^2_{nc}
128	12.5	194.5	$1.52 \cdot 10^{-15}$	0	$5.31 \cdot 10^{-20}$	344.02
512	3.13	15.9	$1.51 \cdot 10^{-16}$	$1.88 \cdot 10^{-21}$	$5.42 \cdot 10^{-24}$	117.22
2,048	0.78	1.19	$4.49 \cdot 10^{-17}$	$5.2 \cdot 10^{-17}$	$1.48 \cdot 10^{-25}$	38.68
8,192	0.2	$8.92 \cdot 10^{-2}$	$6.43 \cdot 10^{-15}$	$8.44 \cdot 10^{-15}$	$2.1 \cdot 10^{-20}$	13.64
32,768	$4.88 \cdot 10^{-2}$	$7.13 \cdot 10^{-3}$	$1.69 \cdot 10^{-15}$	$1.08 \cdot 10^{-12}$	$1.94 \cdot 10^{-29}$	5.34

Table 4.9.: The error estimation data for **the Newton-type methods** P_3, applied on Stokes problem with cavitation in the T-pipe with $\mu = 0.1$.

	PCG			P_3		
cells	$\kappa(e_{jump})$	$\kappa(e_{nc})$	$\kappa(\sqrt{\eta_{\text{Res}}})$	$\kappa(e_{jump})$	$\kappa(e_{nc})$	$\kappa(\sqrt{\eta_{\text{Res}}})$
512	1.81	0.78	1.72	1.81	0.78	1.72
2,048	1.87	0.8	1.64	1.87	0.8	1.64
8,192	1.87	0.75	1.4	1.87	0.75	1.4
32,768	1.83	0.68	1.17	1.83	0.68	1.17

Table 4.10.: The convergence rate κ for the different parts of the error estimator (calculated with the data from the tables 4.8 and 4.9).

tant part of the estimator e_{jump}, which provides information about interdependence of the discrete solution and the cell size.

In the tables 4.8 and 4.9 the number of cells quadruples from row to row, and since the mesh consists of triangular cells, the cell size h is halved in every step. This allows us to calculate the convergence rate κ as follows: if e_h is a value of an error in the current mesh and e_{2h} was the value in the previous mesh, then

$$\kappa(e_h) = -\frac{1}{\ln(2)} \ln\left(\frac{e_h}{e_{2h}}\right) .$$

In the table 4.10 we compare the convergence rates for projected cg method and Newton-type method P_3. For this purpose $\kappa(e_{jump})$, $\kappa(e_{nc})$ and $\kappa(\eta_{\text{Res}})$ were calculated, where η_{Res} is a combined residual estimator with

$$\eta_{\text{Res}}(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T \in \mathbb{T}_h} \left(c^2_{|T|} |T| \, \|\boldsymbol{f} + \mu \Delta \boldsymbol{u}_h - \nabla p_h\|^2_{0,T} + c^2_{|\partial T|} |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|^2_{0, \partial T \setminus \partial\Omega} \right),$$

that is used in the further course of this section.

We do not have a continuous solution for this problem to directly calculate the difference between it an our discrete solutions. Instead we can compare the discrete solution for the different meshes with each other. The table 4.11 provide such a comparison. It is convenient to calculate the norm $\|p_h - p_{2h}\|_{0,\Omega}$, since p_h is constant on each cell and the

	PCG		P_3	
cells	$\|p_h - p_{2h}\|^2_{0,\Omega}$	$\kappa(\|p_h - p_{2h}\|_{0,\Omega})$	$\|p_h - p_{2h}\|^2_{0,\Omega}$	$\kappa(\|p_h - p_{2h}\|_{0,\Omega})$
128	18.4344		18.4331	
512	5.60869	0.86	5.6071	0.86
2,048	1.10101	1.17	1.10143	1.17
8,192	0.2267	1.13	0.22667	1.14
32,768	0.05785	0.99	0.05627	1

Table 4.11.: A comparison of the discrete solutions of the Stokes problem with cavitation for the T-pipe example for the different meshes (solved until relative tolerance $\frac{\|p_{new}-p_{old}\|}{\|p_{new}+p_{old}\|} < 10^{-8}$ and with global refinement of cells).

value from the previous mesh is stored for test, in which it is need for calculating better start value. The norm $\|\nabla(u_h - u_{2h})\|_{0,\Omega}$ can be estimated using the similar trick as in lemma 4.7.8, which leads to an inequality

$$\|\nabla(u_h - u_{2h})\|^2_{0,\Omega} \leq c_\eta \eta_{\mathrm{Res}}(\boldsymbol{u}_{2h}, p_{2h}, \boldsymbol{f}) + c_{ph} \|p_h - p_{2h}\|^2_{0,\Omega}$$

with positive constants c_η and c_{ph}, for which we proved proof at the end of this section. To conclude, it can therefore be said that the accuracy of the discrete solution is approximately linearly dependent upon cell size h.

Lemma 4.5.1. *Let* $\mathbf{V_{2h}}$, $\mathbf{V_h}$, $\mathbf{Q_{2h}}$ *and* $\mathbf{Q_h}$ *be discrete spaces, such that* $\mathbf{V}_{2h} \subset \mathbf{V}_h$ *and* $\mathbf{Q}_{2h} \subset \mathbf{Q}_h$. *Furthermore let* $(\boldsymbol{u}_h, p_h)^T \in \mathbf{V_h} \times \mathbf{Q_h}$ *and* $(\boldsymbol{u}_{2h}, p_{2h})^T \in \mathbf{V_{2h}} \times \mathbf{Q_{2h}}$ *be the discrete solutions of for the stokes problem with cavitation, given by (4.1) to (4.3), on the respective discrete spaces. Then the following inequality applies*

$$\|\nabla(u_h - u_{2h})\|^2_{0,\Omega} \leq c_\eta \eta_{Res}(\boldsymbol{u}_{2h}, p_{2h}, \boldsymbol{f}) + c_{ph} \|p_h - p_{2h}\|^2_{0,\Omega}$$

with positive constants c_η *and* c_{ph}.

Proof. First we define the difference $\boldsymbol{e}_{2h} = \boldsymbol{u}_h - \boldsymbol{u}_{2h}$ and its interpolation on the discrete space with less degrees of freedom $I_h\boldsymbol{e}_{2h} \in \mathbf{V_h}$. Using the equation (4.12) we obtain $\mu\left(\nabla \boldsymbol{e}_{2h}, I_h\boldsymbol{e}_{2h}\right)_{0,\Omega} = \left(p_h - p_{2h}, \nabla \cdot (I_h\boldsymbol{e}_{2h})\right)_{0,\Omega}$ and this leads to

$$\begin{aligned}
\mu \left\|\nabla\left(\boldsymbol{u}_h - \boldsymbol{u}_{2h}\right)\right\|^2_{0,\Omega} =& \mu\left(\nabla\left(\boldsymbol{u}_h - \boldsymbol{u}_{2h}\right), \nabla\left(\boldsymbol{e}_{2h} - I_h\boldsymbol{e}_{2h}\right)\right)_{0,\Omega} \\
&- \left(p_h - p_{2h}, \nabla \cdot \left(\boldsymbol{e}_{2h} - I_h\boldsymbol{e}_{2h}\right)\right)_{0,\Omega} + \left(p_h - p_{2h}, \nabla \cdot \boldsymbol{e}_{2h}\right)_{0,\Omega} \\
=& \left(\boldsymbol{f}, \boldsymbol{e}_{2h} - I_h\boldsymbol{e}_{2h}\right)_{0,\Omega} - \mu\left(\nabla \boldsymbol{u}_{2h}, \nabla\left(\boldsymbol{e}_{2h} - I_h\boldsymbol{e}_{2h}\right)\right)_{0,\Omega} \\
&+ \left(p_{2h}, \nabla \cdot \left(\boldsymbol{e}_{2h} - I_h\boldsymbol{e}_{2h}\right)\right)_{0,\Omega} + \left(p_h - p_{2h}, \nabla \cdot \boldsymbol{e}_{2h}\right)_{0,\Omega} \ .
\end{aligned}$$

By applying the lemma 4.7.2 and Cauchy-Schwarz inequality we can estimate

$$\mu \left\|\nabla\left(\boldsymbol{u}_h - \boldsymbol{u}_{2h}\right)\right\|^2_{0,\Omega} \leq \|\nabla \boldsymbol{e}_{2h}\|_{0,\Omega} \sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_{2h}, p_{2h}, \boldsymbol{f})} + \|\nabla \cdot \boldsymbol{e}_{2h}\|_{0,\Omega} \|p_h - p_{2h}\|^2_{0,\Omega} \ .$$

Since $\|\nabla \cdot \boldsymbol{e}_{2h}\|_{0,\Omega} \leq \|\nabla \boldsymbol{e}_{2h}\|_{0,\Omega}$, the only step that is left to do, in order to obtain target estimation, is to apply Young's inequality and calculate the constants c_η and c_{ph}. □

4.6. Existence and uniqueness of the continuous and the FEM solutions

In this section we want to proof the existence an uniqueness of the solution of the continuous stabilized and regularised problem, as well as the existence an uniqueness of the solution of the discrete stabilized problem. Then we show, that the stabilized solution converges to the original solution for $\xi \to 0$. It can be argued that, if we would use Newton algorithms for the regularised problem with an appropriate $\xi > 0$, the solution would approximate the solution of the not regularised problem. The process should highlight the few criteria that are meet in continuous spaces and are required from the discrete spaces, for the solution strategy to work. The other aspect of this discussion is the preparation for the more abstract problem, in which the same properties will be required. Finally, from the discussion of the existence an the uniqueness of the solution we can derive the condition $0 < \delta < \mu$. First we describe the ideas, that we use, and the formal proof as well as the lemmas, that are used in the process, will come at the end of this section.

Theorem 4.6.1. *There is an unique solution* $(\boldsymbol{u}^\xi, p^\xi)^T \in \mathbf{V} \times \mathbf{Q}$ *of the continuous stabilized problem*

$$\tilde{\mathcal{F}}_{\boldsymbol{\varphi} q}(\boldsymbol{u}^\xi, p^\xi) = 0 \qquad\qquad \forall (\boldsymbol{\varphi}, q)^T \in \mathbf{V} \times \mathbf{Q}\,,$$

with $\tilde{\mathcal{F}}$ *defined by equation (4.19).*

To proof this theorem, we consider a mapping in to a dual space $\breve{E} : \mathbf{V} \times \mathbf{Q} \to \mathbf{V}^* \times \mathbf{Q}^*$ with

$$\left\langle \breve{E}(\tilde{\boldsymbol{u}}, \tilde{p})^{\mathrm{T}}, (\boldsymbol{\varphi}, q)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} = \tilde{\mathcal{F}}_{\boldsymbol{\varphi} q}(\tilde{\boldsymbol{u}}, \tilde{p}) \qquad\qquad \forall (\tilde{\boldsymbol{u}}, \tilde{p})^{\mathrm{T}}, (\boldsymbol{\varphi}, q)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$$

and proof, that the operator $\breve{E}$ strong monotone and Lipschitz continuous is. According to the lemma A.0.7, this qualities secure he existence an uniqueness of the solution.

The proof is relative technical (see last part of this section). The basis of it are following qualities:

i) There are positive constant c and δ, such that

$$\mathcal{A}^\delta(\boldsymbol{\varphi}, \boldsymbol{\varphi}) = \mu \left\|\nabla \boldsymbol{\varphi}\right\|_{0,\Omega}^2 - \delta \left\|\nabla \cdot \boldsymbol{\varphi}\right\|_{0,\Omega}^2 > c \left\|\nabla \boldsymbol{\varphi}\right\|_{0,\Omega}^2 \qquad\qquad \forall \boldsymbol{\varphi} \in \mathbf{V}\,,$$

where $\mathcal{A}^\delta(\boldsymbol{u}, \boldsymbol{\varphi}) = \mu \left(\nabla \boldsymbol{u}, \nabla \boldsymbol{\varphi}\right)_{0,\Omega} - \delta \left(\nabla \cdot \boldsymbol{u}, \nabla \cdot \boldsymbol{\varphi}\right)_{0,\Omega}$.

ii) The bilinear form $\mathcal{B}(.\,,.)$ fulfil the inf-sup-condition

$$\inf_{q \in \mathbf{Q}} \sup_{\boldsymbol{\varphi} \in \mathbf{V}} \frac{\mathcal{B}(\boldsymbol{\varphi}, q)}{\left\|\boldsymbol{\varphi}\right\|_{\mathbf{V}} \left\|q\right\|_{\mathbf{Q}}} = \inf_{q \in \mathbf{Q}} \sup_{\boldsymbol{\varphi} \in \mathbf{V}} \frac{-\left(\nabla \cdot \boldsymbol{\varphi}, q\right)_{0,\Omega}}{\left\|\boldsymbol{\varphi}\right\|_{\mathbf{V}} \left\|q\right\|_{\mathbf{Q}}} \geq \tilde{c}\,,$$

where $\tilde{c}$ is a positive constant and $\mathcal{B}(\boldsymbol{\varphi}, q) = -\left(\nabla \cdot \boldsymbol{\varphi}, q\right)_{0,\Omega}$.

This leads to the similar theorem for the discrete case.

Theorem 4.6.2. *There is an unique solution* $(\boldsymbol{u}_h^\xi, p_h^\xi)^T \in \mathbf{V_h} \times \mathbf{Q_h}$ *of the discrete stabilized and regularised problem*

$$\tilde{\mathcal{F}}_{\boldsymbol{\varphi}_h q_h}(\boldsymbol{u}_h^\xi, p_h^\xi) = 0 \qquad \forall (\boldsymbol{\varphi}_h, q_h)^T \in \mathbf{V_h} \times \mathbf{Q_h} \,,$$

with $\tilde{\mathcal{F}}$ *defined by equation (4.19), the following criteria:*

i) There are positive constant c *and* δ*, such that*

$$\mathcal{A}^\delta(\boldsymbol{\varphi}_h, \boldsymbol{\varphi}_h) = \mu \left\|\nabla \boldsymbol{\varphi}_h\right\|_{0,\Omega}^2 - \delta \left\|\nabla \cdot \boldsymbol{\varphi}_h\right\|_{0,\Omega}^2 > c \left\|\nabla \boldsymbol{\varphi}_h\right\|_{0,\Omega}^2 \qquad \forall \boldsymbol{\varphi}_h \in \mathbf{V}_h \,.$$

ii) The bilinear form $\mathcal{B}(.\,,.)$ *fulfil the inf-sup-condition*

$$\inf_{q_h \in \mathbf{Q_h}} \sup_{\boldsymbol{\varphi}_h \in \mathbf{V_h}} \frac{\mathcal{B}(\boldsymbol{\varphi}_h, q_h)}{\left\|\boldsymbol{\varphi}_h\right\|_{\mathbf{V}} \left\|q_h\right\|_{\mathbf{Q}}} = \inf_{q_h \in \mathbf{Q_h}} \sup_{\boldsymbol{\varphi}_h \in \mathbf{V_h}} \frac{-\left(\nabla \cdot \boldsymbol{\varphi}_h, q_h\right)_{0,\Omega}}{\left\|\boldsymbol{\varphi}_h\right\|_{\mathbf{V}} \left\|q_h\right\|_{\mathbf{Q}}} \geq \tilde{c} \,,$$

where $\tilde{c}$ *is a positive constant.*

Proof. Analogue to the proof of the theorem 4.6.1. □

Its important to know how does the stabilisation of the square root affect the solution. In the next theorem we consider limit of difference between the actual solution and the solution of the stabilized problem.

Theorem 4.6.3. *Let* $(\boldsymbol{u}, p)^T$ *be the solution of the continuous problem*

$$\mathcal{F}_{\boldsymbol{\varphi} q}(\boldsymbol{u}, p) = 0 \qquad \forall (\boldsymbol{\varphi}, q)^T \in \mathbf{V} \times \mathbf{Q} \,,$$

with $\mathcal{F}$ *defined by equation (4.15). Let* $(\boldsymbol{u}^\xi, p^\xi)^T$ *be the solution of the continuous stabilized problem*

$$\tilde{\mathcal{F}}_{\boldsymbol{\varphi} q}(\boldsymbol{u}^\xi, p^\xi) = 0 \qquad \forall (\boldsymbol{\varphi}, q)^T \in \mathbf{V} \times \mathbf{Q} \,,$$

with $\tilde{\mathcal{F}}$ *defined by equation (4.19). Then*

$$\lim_{\xi \to 0} \left\| (\boldsymbol{u}, p)^T - (\boldsymbol{u}^\xi, p^\xi)^T \right\|_{\mathbf{V} \times \mathbf{Q}} = 0 \,.$$

Besides the convergence for $\xi \to 0$, the proof show how the convergence rate is affected by value of $|\Omega|$. This lead to the conclusion, to see ξ not as a constant, but more as a piecewise constant function $\xi : \Omega \to \mathbb{R}$. The rest of the chapter are the proof of the theorems 4.6.1 and 4.6.3 as well as some lemma, that are used in this proofs.

Proof. (**Theorem 4.6.1**) As shown in the proof of the lemma 4.6.4, there is a positive constant c, such that

$$\mathcal{A}^\delta(\boldsymbol{\varphi}, \boldsymbol{\varphi}) > c \left\|\nabla \boldsymbol{\varphi}\right\|_{0,\Omega}^2 \qquad \forall \boldsymbol{\varphi} \in \mathbf{V} \,.$$

i) (**Strong monotony**) We consider the dualpair

$$\left\langle \breve{E}(\boldsymbol{\varphi},q)^{\mathrm{T}} - \breve{E}(\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}}, (\boldsymbol{\varphi},q)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} = \mathcal{A}^{\delta}(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}},\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) + \left(\boldsymbol{\lambda}^{\xi} - \tilde{\boldsymbol{\lambda}}^{\xi}, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega} \qquad \forall (\boldsymbol{\varphi},q)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}} \in \mathbf{V}\times\mathbf{Q}\,.$$

According to inequality (4.22)

$$\begin{aligned}\left(\boldsymbol{\lambda}^{\xi} - \tilde{\boldsymbol{\lambda}}^{\xi}, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega} &> \frac{\delta c}{2}\left\|\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - \frac{1}{\delta}(q-\tilde{q})\right)\right\|^2_{0,\Omega} \\ &\geq \frac{\delta c}{2}\left\|\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})\right\|^2_{0,\Omega} + \frac{c}{2\delta}\left\|q-\tilde{q}\right\|^2_{0,\Omega} - c\,(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}), q-\tilde{q})_{0,\Omega}\,.\end{aligned}$$

By using $(\delta+1)$ as a constant in the Young inequality we obtain

$$-c\,(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}), q-\tilde{q})_{0,\Omega} \geq -\frac{c(\delta+1)}{2}\left\|\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})\right\|^2_{0,\Omega} - \frac{c}{2(\delta+1)}\left\|q-\tilde{q}\right\|^2_{0,\Omega}\,.$$

Putting all those estimations together result in the following inequality:

$$\begin{aligned}\left\langle \breve{E}(\boldsymbol{\varphi},q)^{\mathrm{T}} - \breve{E}(\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}}, (\boldsymbol{\varphi},q)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} &> c\left\|\nabla(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})\right\|^2_{0,\Omega} \\ &\quad -\frac{c}{2}\left\|\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})\right\|^2_{0,\Omega} + \frac{c}{2\delta(\delta+1)}\left\|q-\tilde{q}\right\|^2_{0,\Omega} \\ &> \tilde{\alpha}\left\|(\boldsymbol{\varphi},q)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} \qquad \forall (\boldsymbol{\varphi},q)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}} \in \mathbf{V}\times\mathbf{Q}\,,\end{aligned}$$

with a constant $\tilde{\alpha} = \dfrac{c}{2\max\{2,\delta(\delta+1)\}}$, which means that the operator $\breve{E}$ is strong monotone.

(ii) (**Lipschitz continuity**) Using the inequalities (4.23) and (4.20), we can demonstrate, that the mapping $\breve{E}$ is Lipschitz continuous:

$$\begin{aligned}\left\|\breve{E}(\boldsymbol{\varphi},q)^{\mathrm{T}} - \breve{E}(\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}}\right\|_{\mathbf{V}^*\times\mathbf{Q}^*} &= \sup_{(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}} \frac{\left\langle \breve{E}(\boldsymbol{\varphi},q)^{\mathrm{T}} - \breve{E}(\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}}, (\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}}}{\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}} \\ \leq \sup_{(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}} &\frac{\mathcal{A}^{\delta}(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}},\tilde{\boldsymbol{u}}) - (q-\tilde{q},\nabla\cdot\tilde{\boldsymbol{u}})_{0,\Omega} + (\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}),\tilde{p})_{0,\Omega}}{\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}} \\ &+ \sup_{(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}} \frac{\left(\boldsymbol{\lambda}^{\xi} - \tilde{\boldsymbol{\lambda}}^{\xi}, \delta\nabla\cdot\tilde{\boldsymbol{u}} - \tilde{p}\right)_{0,\Omega}}{\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}} \\ \leq \left(\alpha + \frac{\delta^2+1}{\delta}\right)&\left\|(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}, q-\tilde{q})^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} \qquad \forall (\boldsymbol{\varphi},q)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}},\tilde{q})^{\mathrm{T}} \in \mathbf{V}\times\mathbf{Q}\,.\end{aligned}$$

iii) (**Conclusion**) According to the theorem A.0.7 the equation

$$\breve{E}(\boldsymbol{u}^{\xi}, p^{\xi})^{\mathrm{T}} = \mathbf{0}^*$$

has exactly one solution $(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$, where $\mathbf{0}^* \in \mathbf{V}^*$ is additive neutral element of the dual space. This means, that the equation

$$\tilde{\mathcal{F}}_{\boldsymbol{\varphi} q}(\boldsymbol{u}^\xi, p^\xi) = 0 \qquad\qquad \forall (\boldsymbol{\varphi}, q)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}\,.$$

has a unique solution $(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$. □

Proof. **(Theorem 4.6.3)** Since the operator $\breve{E}$ is strong monotone we obtain the following equality :

$$\left\|(\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} < \frac{1}{\tilde{\alpha}} \left\langle \breve{E}(\boldsymbol{u}, p)^{\mathrm{T}} - \breve{E}(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}, (\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}}\,.$$

According to the definitions of $(\boldsymbol{u}, p)^{\mathrm{T}}$ and $(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}$

$$\left\langle \breve{E}(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}, (\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}} = 0$$

and
$$\left\langle \breve{E}(\boldsymbol{u}, p)^{\mathrm{T}}, (\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}}$$
$$= \frac{1}{2\tilde{\alpha}} \left(\sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2 + \xi} - \sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2}, \delta\nabla\cdot\left(\boldsymbol{u} - \boldsymbol{u}^\xi\right) - \left(p - p^\xi\right) \right)_{0,\Omega}\,.$$

Finally, using the Cauchy-Schwarz, the Young's and Poincare's inequalities, we receive

$$\left\|(\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} < \frac{1}{\tilde{\alpha}} \left\langle \breve{E}(\boldsymbol{u}, p)^{\mathrm{T}} - \breve{E}(\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}, (\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}} \right\rangle_{\mathbf{V}\times\mathbf{Q}}$$
$$= \frac{1}{2\tilde{\alpha}} \left(\sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2 + \xi} - \sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2}, \delta\nabla\cdot\left(\boldsymbol{u} - \boldsymbol{u}^\xi\right) - \left(p - p^\xi\right) \right)_{0,\Omega}$$
$$\leq \frac{1}{2\tilde{\alpha}} \underbrace{\left\| \frac{\xi}{\sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2 + \xi} + \sqrt{\left(\nabla\cdot\boldsymbol{u} - \frac{1}{\delta}p\right)^2}} \right\|_{0,\Omega}}_{\leq \|\sqrt{\xi}\|_{0,\Omega}} \left\|\delta\nabla\cdot\left(\boldsymbol{u} - \boldsymbol{u}^\xi\right) - \left(p - p^\xi\right)\right\|_{0,\Omega}$$
$$\leq |\Omega|\sqrt{\xi}\frac{\sqrt{\delta^2+1}}{2\tilde{\alpha}} \left\|(\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\,.$$

This means, that $\left\|(\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} \leq |\Omega|\sqrt{\xi}\frac{\sqrt{\delta^2+1}}{2\tilde{\alpha}}$ and as a result,

$$\lim_{\xi\to 0} \left\|(\boldsymbol{u}, p)^{\mathrm{T}} - (\boldsymbol{u}^\xi, p^\xi)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} = 0\,.$$ □

Lemma 4.6.4 (Linear problem)**.** *Let $\breve{L} : \mathbf{V} \times \mathbf{Q} \to \mathbf{V}^* \times \mathbf{Q}^*$, $(\boldsymbol{u}, p)^T \mapsto (\boldsymbol{f}^*, \boldsymbol{g}^*)^T$ be a linear mapping into the dual space, defined by the saddle point problem*

$$\begin{aligned} \mathcal{A}^\delta(\boldsymbol{u}, \boldsymbol{\varphi}) + \mathcal{B}(\boldsymbol{\varphi}, p) &= \langle \boldsymbol{f}^*, \boldsymbol{\varphi}\rangle_{\mathbf{V}} && \forall \boldsymbol{\varphi} \in \mathbf{V}\,, \\ \mathcal{B}(\boldsymbol{u}, q) &= \langle g^*, q\rangle_{\mathbf{Q}} && \forall q \in \mathbf{Q}\,, \end{aligned}$$

where $\mathcal{A}^\delta(\boldsymbol{u},\boldsymbol{\varphi}) = \mu\,(\nabla\boldsymbol{u},\nabla\boldsymbol{\varphi})_{0,\Omega} - \delta\,(\nabla\cdot\boldsymbol{u},\nabla\cdot\boldsymbol{\varphi})_{0,\Omega}$ *and* $\mathcal{B}(\boldsymbol{\varphi},q) = -\,(\nabla\cdot\boldsymbol{\varphi},q)_{0,\Omega}$. *Then for* $0<\delta<\mu$ *the operator* $\breve{L}$ *is isomorph and the following inequalities apply*

$$\begin{aligned}\left|\left\langle \breve{L}(\tilde{\boldsymbol{u}},\tilde{p})^T,(\boldsymbol{\varphi},q)^T\right\rangle_{\mathbf{V}\times\mathbf{Q}}\right| &= \left|\mathcal{A}^\delta(\tilde{\boldsymbol{u}},\boldsymbol{\varphi}) - (\nabla\cdot\boldsymbol{\varphi},\tilde{p})_{0,\Omega} - (q,\nabla\cdot\tilde{\boldsymbol{u}})_{0,\Omega}\right| \\ &\le \alpha\left\|(\tilde{\boldsymbol{u}},\tilde{p})^T\right\|_{\mathbf{V}\times\mathbf{Q}}\left\|(\boldsymbol{\varphi},q)^T\right\|_{\mathbf{V}\times\mathbf{Q}} \qquad \forall(\boldsymbol{\varphi},q)^T,(\tilde{\boldsymbol{u}},\tilde{p})^T\in\mathbf{V}\times\mathbf{Q} \quad (4.20)\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\beta}\left\|(\tilde{\boldsymbol{u}},\tilde{p})^T\right\|_{\mathbf{V}\times\mathbf{Q}} &\le \sup_{(\boldsymbol{\varphi},q)^T\in\mathbf{V}\times\mathbf{Q}}\frac{\left\langle \breve{L}(\tilde{\boldsymbol{u}},\tilde{p})^T,(\boldsymbol{\varphi},q)^T\right\rangle_{\mathbf{V}\times\mathbf{Q}}}{\|(\boldsymbol{\varphi},q)^T\|_{\mathbf{V}\times\mathbf{Q}}} \\ &= \sup_{(\boldsymbol{\varphi},q)^T\in\mathbf{V}\times\mathbf{Q}}\frac{\mathcal{A}^\delta(\tilde{\boldsymbol{u}},\boldsymbol{\varphi}) - (\tilde{p},\nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (q,\nabla\cdot\tilde{\boldsymbol{u}})_{0,\Omega}}{\|(\boldsymbol{\varphi},q)^T\|_{\mathbf{V}\times\mathbf{Q}}} \quad \forall(\tilde{\boldsymbol{u}},\tilde{p})^T\in\mathbf{V}\times\mathbf{Q}\,. \quad (4.21)\end{aligned}$$

Proof. Since $\|\nabla\boldsymbol{\varphi}\|_{0,\Omega} > \|\nabla\cdot\boldsymbol{\varphi}\|_{0,\Omega}$, for $0<\delta<\mu$ there is a positive constant c, such that

$$\begin{aligned}\mathcal{A}^\delta(\boldsymbol{\varphi},\boldsymbol{\varphi}) =& \mu\,(\nabla\boldsymbol{\varphi},\nabla\boldsymbol{\varphi})_{0,\Omega} - \delta\,(\nabla\cdot\boldsymbol{\varphi},\nabla\cdot\boldsymbol{\varphi})_{0,\Omega} \\ =& \mu\,\|\nabla\boldsymbol{\varphi}\|^2_{0,\Omega} - \delta\,\|\nabla\cdot\boldsymbol{\varphi}\|^2_{0,\Omega} \\ >& c\,\|\nabla\boldsymbol{\varphi}\|^2_{0,\Omega} \\ \ge& \frac{c}{2}\,\|\boldsymbol{\varphi}\|^2_{1,\Omega} \qquad\qquad \forall\boldsymbol{\varphi}\in\mathbf{V}\,,\end{aligned}$$

This means the bilinear form $\mathcal{A}^\delta(.\,,.)$ is V-elliptic. The bilinear form $\mathcal{B}(.\,,.)$ fulfil the inf-sup-condition

$$\inf_{q\in\mathbf{Q}}\sup_{\boldsymbol{\varphi}\in\mathbf{V}}\frac{\mathcal{B}(\boldsymbol{\varphi},\boldsymbol{q})}{\|\boldsymbol{\varphi}\|_{\mathbf{V}}\,\|\boldsymbol{q}\|_{\mathbf{Q}}} = \inf_{q\in\mathbf{Q}}\sup_{\boldsymbol{\varphi}\in\mathbf{V}}\frac{-\,(\nabla\cdot\boldsymbol{\varphi},q)_{0,\Omega}}{\|\boldsymbol{\varphi}\|_{\mathbf{V}}\,\|q\|_{\mathbf{Q}}} \ge \tilde{c}\,,$$

where $\tilde{c}$ is a positive constant (see f.e. Girault & Raviart [10]). According to the Brezzi's splitting theorem A.0.5, the linear mapping $\breve{L}$ is an isomorphism and, as per the abstract existence theorem A.0.4, we obtain following inequalities, with constants $\alpha,\beta\in\mathbb{R}^+$:

$$\begin{aligned}\left|\left\langle \breve{L}(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}},(\boldsymbol{\varphi},q)^{\mathrm{T}}\right\rangle_{\mathbf{V}\times\mathbf{Q}}\right| &= \left|\mathcal{A}^\delta(\tilde{\boldsymbol{u}},\boldsymbol{\varphi}) - (\nabla\cdot\boldsymbol{\varphi},\tilde{p})_{0,\Omega} - (q,\nabla\cdot\tilde{\boldsymbol{u}})_{0,\Omega}\right| \\ &\le \alpha\left\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\left\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} \qquad \forall(\boldsymbol{\varphi},q)^{\mathrm{T}},(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}\end{aligned}$$

and

$$\begin{aligned}\frac{1}{\beta}\left\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} &\le \sup_{(\boldsymbol{\varphi},q)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}}\frac{\left\langle \breve{L}(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}},(\boldsymbol{\varphi},q)^{\mathrm{T}}\right\rangle_{\mathbf{V}\times\mathbf{Q}}}{\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}} \\ &= \sup_{(\boldsymbol{\varphi},q)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}}\frac{\mathcal{A}^\delta(\tilde{\boldsymbol{u}},\boldsymbol{\varphi}) - (\tilde{p},\nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (q,\nabla\cdot\tilde{\boldsymbol{u}})_{0,\Omega}}{\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}} \qquad \forall(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}\,. \quad \square\end{aligned}$$

Lemma 4.6.5. *We define* $\boldsymbol{\lambda}^\xi, \tilde{\boldsymbol{\lambda}}^\xi \in \mathbf{Q}$ *as*

$$\boldsymbol{\lambda}^\xi = \frac{1}{2}\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right) + \frac{1}{2}\sqrt{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right)^2 + \xi}$$
$$\textit{and}\quad \tilde{\boldsymbol{\lambda}}^\xi = \frac{1}{2}\left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right) + \frac{1}{2}\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)^2 + \xi}\,.$$

Then for the constants $0 < c < 1$ *and* $0 < \delta$ *the following inequalities apply:*

$$\left(\boldsymbol{\lambda}^\xi - \tilde{\boldsymbol{\lambda}}^\xi, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega} > \frac{\delta c}{2}\left\|\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - \frac{1}{\delta}(q-\tilde{q})\right)\right\|^2_{0,\Omega} \tag{4.22}$$
$$\forall(\boldsymbol{\varphi},q)^T, (\tilde{\boldsymbol{\varphi}},\tilde{q})^T \in \mathbf{V}\times\mathbf{Q}$$
$$\textit{and}\quad \left(\boldsymbol{\lambda}^\xi - \tilde{\boldsymbol{\lambda}}^\xi, \delta\tilde{\boldsymbol{u}} - \tilde{p}\right)_{0,\Omega} \le \frac{\delta^2+1}{\delta}\left\|(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}, q-\tilde{q})^T\right\|_{\mathbf{V}\times\mathbf{Q}}\left\|(\tilde{\boldsymbol{u}},\tilde{p})^T\right\|_{\mathbf{V}\times\mathbf{Q}} \tag{4.23}$$
$$\forall(\boldsymbol{\varphi},q)^T, (\tilde{\boldsymbol{\varphi}},\tilde{q})^T, (\tilde{\boldsymbol{u}},\tilde{p})^T \in \mathbf{V}\times\mathbf{Q}\,.$$

Proof. We consider the scalar product

$$\begin{aligned}
&\left(\lambda^\xi - \tilde{\lambda}^\xi, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega}\\
&= \frac{1}{2}\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - \frac{1}{\delta}(q-\tilde{q}), \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega}\\
&\quad + \frac{1}{2}\left(\sqrt{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right)^2 + \xi} - \sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)^2 + \xi}, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega}\\
&= \frac{1}{2}\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - \frac{1}{\delta}(q-\tilde{q}), \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega}\\
&\quad + \frac{1}{2}\left(\frac{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right)^2 - \left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)^2}{\sqrt{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right)^2 + \xi} + \sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)^2 + \xi}}, \delta\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - (q-\tilde{q})\right)_{0,\Omega}\\
&= \frac{\delta}{2}\left(1 + \frac{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right) + \left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)}{\sqrt{\left(\nabla\cdot\boldsymbol{\varphi} - \frac{1}{\delta}q\right)^2 + \xi} + \sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}} - \frac{1}{\delta}\tilde{q}\right)^2 + \xi}}, \left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}) - \frac{1}{\delta}(q-\tilde{q})\right)^2\right)_{0,\Omega}
\end{aligned}$$

Using the Hölder's inequality we can find a constant $0 < c < 1$, such that

$$\begin{aligned}
&\frac{\delta}{2}\left(1+\frac{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)+\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)}{\sqrt{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)^2+\xi}+\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)^2+\xi}},\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})-\frac{1}{\delta}(q-\tilde{q})\right)^2\right)_{0,\Omega}\\
&\quad>\frac{\delta}{2}\left(1-\left\|\frac{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)+\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)}{\sqrt{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)^2+\xi}+\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)^2+\xi}}\right\|_{\mathbf{L}^\infty(\Omega)}\right)\\
&\qquad\qquad\times\left\|\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})-\frac{1}{\delta}(q-\tilde{q})\right)\right\|_{0,\Omega}^2\\
&\quad>\frac{\delta c}{2}\left\|\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})-\frac{1}{\delta}(q-\tilde{q})\right)\right\|_{0,\Omega}^2.
\end{aligned}$$

On the other hand, using the Cauchy-Schwarz, the Young's and Poincare's inequalities, we receive

$$\begin{aligned}
\left(\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}},\delta\tilde{\boldsymbol{u}}-\tilde{p}\right)_{0,\Omega}=\frac{1}{2}\int_\Omega\underbrace{\left(1+\frac{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)+\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)}{\sqrt{\left(\nabla\cdot\boldsymbol{\varphi}-\frac{1}{\delta}q\right)^2+\xi}+\sqrt{\left(\nabla\cdot\tilde{\boldsymbol{\varphi}}-\frac{1}{\delta}\tilde{q}\right)^2+\xi}}\right)}_{<2}\times\\
\left(\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})-\frac{1}{\delta}(q-\tilde{q})\right)(\delta\tilde{\boldsymbol{u}}-\tilde{p})\,\mathrm{d}\boldsymbol{x}\,,
\end{aligned}$$

which leads to

$$\begin{aligned}
\left(\boldsymbol{\lambda}-\tilde{\boldsymbol{\lambda}},\delta\tilde{\boldsymbol{u}}-\tilde{p}\right)_{0,\Omega}&<\left(\left|\nabla\cdot(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}})-\frac{1}{\delta}(q-\tilde{q})\right|,|\delta\tilde{\boldsymbol{u}}-\tilde{p}|\right)_{0,\Omega}\\
&\le\left(\|\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}\|_{0,\Omega}+\frac{1}{\delta}\|q-\tilde{q}\|_{0,\Omega}\right)\left(\delta\|\tilde{\boldsymbol{u}}\|_{0,\Omega}+\|\tilde{p}\|_{0,\Omega}\right)\\
&=\sqrt{\left(\|\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}\|_{0,\Omega}+\frac{1}{\delta}\|q-\tilde{q}\|_{0,\Omega}\right)^2}\sqrt{\left(\delta\|\tilde{\boldsymbol{u}}\|_{0,\Omega}+\|\tilde{p}\|_{0,\Omega}\right)^2}\\
&\le\sqrt{\left(1+\frac{1}{\delta^2}\right)\left(\|\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}}\|_{0,\Omega}^2+\|q-\tilde{q}\|_{0,\Omega}^2\right)}\sqrt{(\delta^2+1)\left(\|\tilde{\boldsymbol{u}}\|_{0,\Omega}^2+\|\tilde{p}\|_{0,\Omega}^2\right)}\\
&\le\frac{\delta^2+1}{\delta}\left\|(\boldsymbol{\varphi}-\tilde{\boldsymbol{\varphi}},q-\tilde{q})^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\left\|(\tilde{\boldsymbol{u}},\tilde{p})^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}.
\end{aligned}$$

□

4.7. Error estimate

In the next sections we derive an "a posteriori" error estimator for the stokes problem with cavitation, that was introduced in the section 4.4.3, using the projection operator for the second Lagrange multiplier. As a base we use the method of error analysis introduced by Verfürth [16].

Summarised we will proof that, if $(\boldsymbol{u}, p)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$ is the continuous solution of the mixed variation formulation for the stokes problem with cavitation, given by (4.1) to (4.3), and $(\boldsymbol{u}_h, p_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ is the discrete solution, where $\mathbf{V_h} \subseteq \mathbf{V}$ and $\mathbf{Q_h} \subseteq \mathbf{Q}$, then the norm of the error $\left\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}$ can be estimated with the inequality

$$\left\|(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} \leq c\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) \ ,$$

where

$$\begin{aligned}\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T\in\mathbb{T}_h} \left(|T| \left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h + \nabla p_h\right\|^2_{0,T} + |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial n} - p_h \boldsymbol{n} \right\|^2_{0,\partial T\setminus\partial\Omega} \right) \\ + \left\|\nabla\cdot\boldsymbol{u}_h - \lambda_h\right\|^2_{0,\Omega} + \left(\Pi_{\mathbf{\Lambda}} p_h, \lambda_h\right)_{0,\Omega} + \left\|\Pi_{\mathbf{\Lambda}} p_h - p_h\right\|^2_{0,\Omega}\end{aligned}$$

and

$$\lambda_h = \Pi_{\mathbf{\Lambda}_h} \left(\nabla\cdot\boldsymbol{u}_h - \frac{1}{\delta} p_h \right) \ .$$

We proof this hypothesis in several steps:

- First of all we collect and prove, if necessary, some inequalities, that are used in the further course of the chapter.
- The next step is to estimate the difference between the second Lagrange parameter λ and its approximation λ_h.
- Using the results of this sections we finally derive the a posteriori error estimator.

4.7.1. Helpful estimates

The first two useful inequalities, that should be introduced are

$$\left\|\boldsymbol{v} - I_h\boldsymbol{v}\right\|_{0,T} \leq c_{|T|}|T|^{\frac{1}{2}} \left\|\nabla\boldsymbol{v}\right\|_{0,T} \tag{4.24}$$

$$\text{and} \qquad \left\|\boldsymbol{v} - I_h\boldsymbol{v}\right\|_{0,\partial T} \leq c_{|\partial T|}|\partial T|^{\frac{1}{2}} \left\|\nabla\boldsymbol{v}\right\|_{0,T} \ , \tag{4.25}$$

where $\boldsymbol{v} \in \left(\mathbf{H}^1(\Omega)\right)^n$ with $n \in \mathbb{N}$ and $I_h\boldsymbol{v}$ is its interpolation (see Verfürth [16, p. 313]). The next two lemmas are leading to inequality, which contains the typical norms of local residual of the classical formulation and the norms of the directional derivative of $\boldsymbol{u}_h$ in the direction of the outward pointing normal $\boldsymbol{n}$ on each cell.

Lemma 4.7.1. *For all $\theta : \mathbb{T}_h \longrightarrow \mathbb{R}^+_0$ and $\boldsymbol{\varphi} \in \left(\mathbf{L^2}(\Omega)\right)^n$ with $n \in \mathbb{N}$ the following inequality applies*

$$\sum_{T\in\mathbb{T}_h} \theta(T) \left\|\boldsymbol{\varphi}\right\|_{0,T} \leq \left\|\boldsymbol{\varphi}\right\|_{0,\Omega} \left(\sum_{T\in\mathbb{T}_h} \theta^2(T) \right)^{\frac{1}{2}} \ . \tag{4.26}$$

Proof. The sum $\sum_{T\in\mathbb{T}_h} \theta(T)\| \|\boldsymbol{\varphi}\|_{0,T} = \sum_{j=1}^{N_h} \theta(T_j) \|\boldsymbol{\varphi}\|_{0,T_j}$ can be interpreted as a scalar product in the vector space $\mathbb{R}^{N_h}$. By applying Cauchy-Schwarz inequality we get

$$\begin{aligned}\sum_{j=1}^{N_h} \theta(T_j) \|\boldsymbol{\varphi}\|_{0,T_j} &\leq \left(\sum_{j=1}^{N_h} \theta^2(T_j)\right)^{\frac{1}{2}} \left(\sum_{j=1}^{N_h} \|\boldsymbol{\varphi}\|_{0,T_j}^2\right)^{\frac{1}{2}} \\ &= \left(\sum_{j=1}^{N_h} \theta^2(T_j)\right)^{\frac{1}{2}} \|\boldsymbol{\varphi}\|_{0,\Omega} \ .\end{aligned}$$

□

Lemma 4.7.2. *For all $\boldsymbol{u}_h \in \mathbf{V_h}$, $p_h \in \mathbf{Q_h}$ and $\boldsymbol{f}, \boldsymbol{\varphi} \in \mathbf{V}$ the following inequality applies*

$$\begin{aligned}(\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - \mu \left(\nabla \boldsymbol{u}_h, \nabla \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} + \left(p_h, \nabla \cdot \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} \\ \leq 2 \|\nabla\boldsymbol{\varphi}\|_{0,\Omega} \sqrt{\eta_{Res}(\boldsymbol{u}_h, p_h, \boldsymbol{f})},\end{aligned} \tag{4.27}$$

where

$$\eta_{Res}(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T\in\mathbb{T}_h} \left(c_{|T|}^2 |T| \, \|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T\setminus\partial\Omega}^2 \right).$$

Proof. First of all we split the scalar products and consider them cell wise. Using Green's first identity and Cauchy-Schwarz inequality we receive fist estimate:

$$\begin{aligned}&(\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - \mu \left(\nabla \boldsymbol{u}_h, \nabla \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} + \left(p_h, \nabla \cdot \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} \\ &\leq \sum_{T\in\mathbb{T}_h} \Big((\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,T} - \mu \left(\nabla \boldsymbol{u}_h, \nabla \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,T} + \left(p_h, \nabla \cdot \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,T} \Big) \\ &\leq \sum_{T\in\mathbb{T}_h} \left((\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,T} + \left(\frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi} \right)_{0,\partial T\setminus\partial\Omega} \right) \\ &\leq \sum_{T\in\mathbb{T}_h} \Big(\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T} \, \|\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\|_{0,T} \\ &\quad + \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n} \right\|_{0,\partial T\setminus\partial\Omega} \|\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\|_{0,\partial T\setminus\partial\Omega} \Big) \ .\end{aligned}$$

Application of the inequalities (4.24) and (4.25) transforms the estimate into

$$\begin{aligned}&(\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - \mu \left(\nabla \boldsymbol{u}_h, \nabla \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} + \left(p_h, \nabla \cdot \left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} \\ &\leq \sum_{T\in\mathbb{T}_h} \|\nabla\boldsymbol{\varphi}\|_{0,T} \left(c_{|T|} |T|^{\frac{1}{2}} \, \|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T} + c_{|\partial T|} |\partial T|^{\frac{1}{2}} \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n} \right\|_{0,\partial T\setminus\partial\Omega} \right) \ .\end{aligned}$$

At this point we just need to apply the inequality (4.26) and the Young's inequality to receive the assertion from above:

$$\begin{aligned}(\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - &\mu\left(\nabla \boldsymbol{u}_h, \nabla\left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} + \left(p_h, \nabla\cdot\left(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}\right)\right)_{0,\Omega} \\ &\leq \|\nabla\boldsymbol{\varphi}\|_{0,\Omega}\left(\sum_{T\in\mathbb{T}_h}\left(c_{|T|}|T|^{\frac{1}{2}}\left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\right\|_{0,T}\right.\right. \\ &\quad \left.\left. + \; c_{|\partial T|}|\partial T|^{\frac{1}{2}}\left\|\frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n}\right\|_{0,\partial T\setminus\partial\Omega}\right)^2\right)^{\frac{1}{2}}\end{aligned}$$

$$\begin{aligned}\leq &2\|\nabla\boldsymbol{\varphi}\|_{0,\Omega}\left(\sum_{T\in\mathbb{T}_h}\left(c^2_{|T|}|T|\left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\right\|^2_{0,T}\right.\right. \\ &\left.\left. + \; c^2_{|\partial T|}|\partial T|\left\|\frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right)\right)^{\frac{1}{2}}. \qquad \square\end{aligned}$$

4.7.2. Errror estimator for second Lagrange multiplier

An important part of the development of the error estimator is finding proper estimate for the difference between the second Lagrange multiplier and its approximation. In this section we introduce such inequality.

Theorem 4.7.3. *Let $(\boldsymbol{u}, p)^T \in \mathbf{V}\times\mathbf{Q}$ be the continuous solution of the mixed variation formulation for the stokes problem with cavitation, given by (4.1) to (4.3), and $(\boldsymbol{u}_h, p_h)^T \in \mathbf{V_h}\times\mathbf{Q_h}$ is the discrete solution. Then the norm of $\|\lambda - \lambda_h\|_{0,\Omega}$ can be estimated with the inequality*

$$\begin{aligned}\epsilon\|\lambda - \lambda_h\|^2_{0,\Omega} \leq &\frac{1}{2}\left\|\left(\boldsymbol{u} - \boldsymbol{u}_h, p - p_h\right)^T\right\|^2_{\mathbf{V}\times\mathbf{Q}} \\ &+ c_\lambda\sum_{T\in\mathbb{T}_h}\left(c^2_{|T|}|T|\left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\right\|^2_{0,T} + c^2_{|\partial T|}|\partial T|\left\|\frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h\boldsymbol{n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right) \\ &+ c_\lambda\|\nabla\cdot\boldsymbol{u}_h - \lambda_h\|^2_{0,\Omega} + c_\lambda\left(\Pi_{\mathbf{\Lambda}}p_h, \lambda_h\right)_{0,\Omega} + c_\lambda\|p_h - \Pi_{\mathbf{\Lambda}}p_h\|^2_{0,\Omega}\ ,\end{aligned} \tag{4.28}$$

where $\epsilon > 0$ and $c_\lambda > 0$ are constants.

Proof. Since $\mathcal{A}^\delta(\boldsymbol{v}, \boldsymbol{v}) \geq 0$ is V-elliptic we consider our start inequality

$$\delta\|\lambda - \lambda_h\|^2_{0,\Omega} \leq \mathcal{A}^\delta(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{u}_h) + \delta\left(\lambda - \lambda_h, \lambda - \lambda_h\right)_{0,\Omega}\ .$$

On the right side of the inequality we add and subtract a number of scalar products:

$$\begin{aligned}
&\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)+\delta\,(\lambda-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\quad=\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)-(p-p_h,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}+\delta\,(\lambda-\lambda_h,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}\\
&\qquad+(p-p_h,\lambda-\lambda_h)_{0,\Omega}-\delta\,(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h),\lambda-\lambda_h)_{0,\Omega}+\delta\,(\lambda-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\qquad-(\lambda-\lambda_h,p-p_h)_{0,\Omega}+(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h),p-p_h)_{0,\Omega}\ .
\end{aligned}$$

Applying (4.16) and (4.17), we obtain the equality

$$\begin{aligned}
&\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)+\delta\,(\lambda-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\quad=\underbrace{\mathcal{A}^\delta(\boldsymbol{u},\boldsymbol{u}-\boldsymbol{u}_h)-(p,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}+\delta\,(\lambda,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}}_{=(\boldsymbol{f},\boldsymbol{u}-\boldsymbol{u}_h)_{0,\Omega}}\\
&\qquad-\mathcal{A}^\delta(\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)+(p_h,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}-\delta\,(\lambda_h,\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h))_{0,\Omega}\\
&\qquad+(p-p_h,\lambda-\lambda_h)_{0,\Omega}+\delta\,(\nabla\cdot\boldsymbol{u}_h-\lambda_h,\lambda-\lambda_h)_{0,\Omega}+(\lambda_h-\nabla\cdot\boldsymbol{u}_h,p-p_h)_{0,\Omega}\\
&\qquad-\delta\underbrace{(\nabla\cdot\boldsymbol{u}-\lambda,\lambda-\lambda_h)_{0,\Omega}}_{=0}+\underbrace{(\nabla\cdot\boldsymbol{u}-\lambda,p-p_h)_{0,\Omega}}_{=0}\ .
\end{aligned}$$

Using the notation

$$\begin{aligned}
&\boldsymbol{e}_u=\boldsymbol{u}-\boldsymbol{u}_h && \boldsymbol{e}_u\in\mathbf{V}\\
\text{and}\qquad &I_h\boldsymbol{e}_u=I_h(\boldsymbol{u}-\boldsymbol{u}_h) && I_h\boldsymbol{e}_u\in\mathbf{V_h}\ ,
\end{aligned}$$

we can write the equation above as

$$\begin{aligned}
&\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)+\delta\,(\lambda-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\quad=(\boldsymbol{f},\boldsymbol{e}_u)_{0,\Omega}-\mathcal{A}^\delta(\boldsymbol{u}_h,\boldsymbol{e}_u)+(p_h,\nabla\cdot\boldsymbol{e}_u)_{0,\Omega}-\delta\,(\lambda_h,\nabla\cdot\boldsymbol{e}_u)_{0,\Omega}\\
&\qquad+(p-p_h,\lambda-\lambda_h)_{0,\Omega}+\delta\,(\nabla\cdot\boldsymbol{u}_h-\lambda_h,\lambda-\lambda_h)_{0,\Omega}+(\lambda_h-\nabla\cdot\boldsymbol{u}_h,p-p_h)_{0,\Omega}\ .
\end{aligned}$$

Next, we apply the discrete version of the equation (4.16), where we use $I_h\boldsymbol{e}_u$ as a test function, and insert a projection $\Pi_{\mathbf{\Lambda}}p_h$ into one of the scalar product. Eventually we obtain

$$\begin{aligned}
&\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h,\boldsymbol{u}-\boldsymbol{u}_h)+\delta\,(\lambda-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\quad=(\boldsymbol{f},\boldsymbol{e}_u-I_h\boldsymbol{e}_u)_{0,\Omega}-\mu\,(\nabla\boldsymbol{u}_h,\nabla\,(\boldsymbol{e}_u-I_h\boldsymbol{e}_u))_{0,\Omega}+(p_h,\nabla\cdot(\boldsymbol{e}_u-I_h\boldsymbol{e}_u))_{0,\Omega}\\
&\qquad+\delta\,(\nabla\cdot\boldsymbol{u}_h-\lambda_h,\nabla\cdot(\boldsymbol{e}_u-I_h\boldsymbol{e}_u))_{0,\Omega}+\delta\,(\nabla\cdot\boldsymbol{u}_h-\lambda_h,\lambda-\lambda_h)_{0,\Omega}\\
&\qquad+(\lambda_h-\nabla\cdot\boldsymbol{u}_h,p-p_h)_{0,\Omega}+(p-\Pi_{\mathbf{\Lambda}}p_h+\Pi_{\mathbf{\Lambda}}p_h-p_h,\lambda-\lambda_h)_{0,\Omega}\ .
\end{aligned}$$

The first three summands on the right side are estimatable by (4.27). We apply the inequality (4.24) on the fourth and Cachy-Schwarz inequality on fifth an sixth scalar products respectively. The last summand can be estimated by using (4.9), $p\geq 0$ a. e. in

Ω and Cachy-Schwarz inequality as follows

$$\begin{aligned}(p - \Pi_{\mathbf{\Lambda}} p_h + \Pi_{\mathbf{\Lambda}} p_h - p_h, \lambda - \lambda_h)_{0,\Omega} = &\underbrace{(p, \lambda)_{0,\Omega}}_{=0} - \underbrace{(\Pi_{\mathbf{\Lambda}} p_h, \lambda)_{0,\Omega}}_{\geq 0} - \underbrace{(p, \lambda_h)_{0,\Omega}}_{\geq 0} \\ &+ (\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + (\Pi_{\mathbf{\Lambda}} p_h - p_h, \lambda - \lambda_h)_{0,\Omega} \\ \leq &(\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + (\Pi_{\mathbf{\Lambda}} p_h - p_h, \lambda - \lambda_h)_{0,\Omega} \\ \leq &(\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + \|\Pi_{\mathbf{\Lambda}} p_h - p_h\|_{0,\Omega} \|\lambda - \lambda_h\|_{0,\Omega} \ .\end{aligned}$$

Summarising the above one obtains

$$\begin{aligned}\mathcal{A}^{\delta}(\boldsymbol{u} - \boldsymbol{u}_h, \boldsymbol{u} - \boldsymbol{u}_h) + \delta\,(\lambda - \lambda_h, \lambda - \lambda_h)_{0,\Omega} \leq\ & 2\,\|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{0,T} \\ \times &\left(\sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T|\,\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2\right)\right)^{\frac{1}{2}} \\ &+ \delta\,\|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}\,\|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{0,\Omega} + \delta\,\|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}\,\|\lambda - \lambda_h\|_{0,\Omega} \\ &+ \|\lambda_h - \nabla \cdot \boldsymbol{u}_h\|_{0,\Omega}\,\|p - p_h\|_{0,\Omega} + (\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} \\ &+ \|\Pi_{\mathbf{\Lambda}} p_h - p_h\|_{0,\Omega}\,\|\lambda - \lambda_h\|_{0,\Omega} \ .\end{aligned}$$

Next we apply the Young's inequality with positive constants $c_{y,1}, c_{y,2}, c_{y,3}, c_{y,4}, c_{y,5} \in \mathbb{R}^+$ and receive

$$\begin{aligned}\delta\,\|\lambda - \lambda_h\|_{0,\Omega}^2 \leq &\left(c_{y,1} + \frac{\delta}{2} c_{y,2}\right) \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{0,T}^2 + \frac{c_{y,4}}{2}\,\|p - p_h\|_{0,\Omega}^2 \\ &+ \frac{1}{c_{y,1}} \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T|\,\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2\right) \\ &+ \left(\frac{\delta}{2c_{y,2}} + \frac{\delta}{2c_{y,3}} + \frac{1}{2c_{y,4}}\right) \|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 \\ &+ (\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + \frac{1}{2c_{y,5}}\,\|p_h - \Pi_{\mathbf{\Lambda}} p_h\|_{0,\Omega}^2 + \frac{\delta c_{y,3} + c_{y,5}}{2}\,\|\lambda - \lambda_h\|_{0,\Omega}^2 \ .\end{aligned}$$

We set $c_{y,3} = \frac{1}{2}$ and $c_{y,5} = \frac{\delta}{2}$, in order to absorb the norm $\|\lambda - \lambda_h\|_{0,\Omega}^2$ on the right side by the left side, multiply subsequently both sides with $\frac{2\epsilon}{\delta}$, where $\epsilon \in \mathbb{R}^+$ is a positive constant, and obtain

$$\begin{aligned}\epsilon\,\|\lambda - \lambda_h\|_{0,\Omega}^2 \leq &\frac{\epsilon(c_{y,1} + \delta c_{y,2})}{\delta}\,\|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{0,T}^2 + \frac{\epsilon c_{y,4}}{\delta}\,\|p - p_h\|_{0,\Omega}^2 \\ &+ \frac{2\epsilon}{\delta c_{y,1}} \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T|\,\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2\right) \\ &+ \frac{\epsilon\,(\delta c_{y,4}(1 + 2c_{y,2}) + c_{y,2})}{\delta c_{y,2} c_{y,4}}\,\|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 \\ &+ \frac{2\epsilon}{\delta}\,(\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + \frac{2\epsilon}{\delta^2}\,\|p_h - \Pi_{\mathbf{\Lambda}} p_h\|_{0,\Omega}^2 \ .\end{aligned}$$

Setting $c_{y,1} = \frac{\delta}{4\epsilon}$, $c_{y,2} = \frac{1}{4\epsilon}$ and $c_{y,4} = \frac{\delta}{2\epsilon}$ allow us to simplify the right side:

$$\begin{aligned}
\epsilon \|\lambda - \lambda_h\|_{0,\Omega}^2 \leq{}& \frac{1}{2} \|\nabla(\boldsymbol{u} - \boldsymbol{u}_h)\|_{0,T}^2 + \frac{1}{2} \|p - p_h\|_{0,\Omega}^2 \\
&+ \frac{8\epsilon^2}{\delta^2} \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T| \, \|\boldsymbol{f} + \mu \Delta \boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2 \right) \\
&+ \frac{\epsilon \left(\delta^2 (4\epsilon + 2) + 4\epsilon \right)}{\delta^2} \|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 \\
&+ \frac{2\epsilon}{\delta} \left(\Pi_{\boldsymbol{\Lambda}} p_h, \lambda_h \right)_{0,\Omega} + \frac{2\epsilon}{\delta^2} \|p_h - \Pi_{\boldsymbol{\Lambda}} p_h\|_{0,\Omega}^2 \, .
\end{aligned}$$

Finally, we define the constant

$$c_\lambda = \max \left\{ \frac{8\epsilon^2}{\delta^2}, \frac{\epsilon \left(\delta^2 (4\epsilon + 2) + 4\epsilon \right)}{\delta^2}, \frac{2\epsilon}{\delta}, \frac{2\epsilon}{\delta^2} \right\}$$

and obtain the hypothesis as a estimation result

$$\begin{aligned}
\epsilon \|\lambda - \lambda_h\|_{0,\Omega}^2 \leq{}& \frac{1}{2} \left\| (\boldsymbol{u} - \boldsymbol{u}_h, p - p_h)^{\mathrm{T}} \right\|_{\mathbf{V} \times \mathbf{Q}}^2 \\
&+ c_\lambda \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T| \, \|\boldsymbol{f} + \mu \Delta \boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2 \right) \\
&+ c_\lambda \|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 + c_\lambda \left(\Pi_{\boldsymbol{\Lambda}} p_h, \lambda_h \right)_{0,\Omega} + c_\lambda \|\Pi_{\boldsymbol{\Lambda}} p_h - p_h\|_{0,\Omega}^2 \, .
\end{aligned}$$

□

Remark 4.7.4. *By using another values for the constants $c_{y,1}, c_{y,2}, c_{y,3}, c_{y,4}$ and $c_{y,5}$ we can also obtain the estimation*

$$\begin{aligned}
\epsilon \|\lambda - \lambda_h\|_{0,\Omega}^2 \leq{}& \frac{1}{2} \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathbf{V}}^2 + \|p - p_h\|_{\mathbf{Q}}^2 \\
&+ c_{\lambda,u} \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T| \, \|\boldsymbol{f} + \mu \Delta \boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2 \right) \\
&+ c_{\lambda,u} \|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 + c_{\lambda,u} \left(\Pi_{\boldsymbol{\Lambda}} p_h, \lambda_h \right)_{0,\Omega} + c_{\lambda,u} \|\Pi_{\boldsymbol{\Lambda}} p_h - p_h\|_{0,\Omega}^2 \, ,
\end{aligned} \tag{4.29}$$

as well as the estimation

$$\begin{aligned}
\epsilon \|\lambda - \lambda_h\|_{0,\Omega}^2 \leq{}& \|\boldsymbol{u} - \boldsymbol{u}_h\|_{\mathbf{V}}^2 + \frac{1}{2} \|p - p_h\|_{\mathbf{Q}}^2 \\
&+ c_{\lambda,p} \sum_{T \in \mathbb{T}_h} \left(c_{|T|}^2 |T| \, \|\boldsymbol{f} + \mu \Delta \boldsymbol{u}_h - \nabla p_h\|_{0,T}^2 + c_{|\partial T|}^2 |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial \boldsymbol{n}} - p_h \boldsymbol{n} \right\|_{0,\partial T \setminus \partial\Omega}^2 \right) \\
&+ c_{\lambda,p} \|\nabla \cdot \boldsymbol{u}_h - \lambda_h\|_{0,\Omega}^2 + c_{\lambda,p} \left(\Pi_{\boldsymbol{\Lambda}} p_h, \lambda_h \right)_{0,\Omega} + c_{\lambda,p} \|\Pi_{\boldsymbol{\Lambda}} p_h - p_h\|_{0,\Omega}^2 \, .
\end{aligned} \tag{4.30}$$

4.7.3. Complete errror estimator

Theorem 4.7.5. *Let $(\boldsymbol{u}, p)^T \in \mathbf{V} \times \mathbf{Q}$ be the continuous solution of the mixed variation formulation for the stokes problem with cavitation, given by (4.1) to (4.3). Let $(\boldsymbol{u}_h, p_h)^T \in$*

$\mathbf{V_h}\times\mathbf{Q_h}$ *be the discrete solution. Then the norm* $\left\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)^T\right\|_{\mathbf{V}\times\mathbf{Q}}$ *can be estimated with the inequality*

$$\left\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)^T\right\|^2_{\mathbf{V}\times\mathbf{Q}} \leq c\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) \ ,$$

where

$$\begin{aligned}\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T\in\mathbb{T}_h} \left(|T| \left\|\boldsymbol{f} + \mu\Delta\boldsymbol{u}_h + \nabla p_h\right\|^2_{0,T} + |\partial T| \left\| \frac{\partial \boldsymbol{u}_h}{\partial n} - p_h\boldsymbol{n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right) \\ + \|\nabla\cdot\boldsymbol{u}_h - \lambda_h\|^2_{0,\Omega} + (\Pi_{\mathbf{\Lambda}} p_h, \lambda_h)_{0,\Omega} + \|\Pi_{\mathbf{\Lambda}} p_h - p_h\|^2_{0,\Omega}\end{aligned}$$

and

$$\lambda_h = \Pi_{\mathbf{\Lambda}_h}\left(\nabla\cdot\boldsymbol{u}_h - \frac{1}{\delta}p_h\right) \ .$$

Proof. We start with an inequality (4.21), where $\tilde{\boldsymbol{u}} = \boldsymbol{u}-\boldsymbol{u}_h$ and $\tilde{p} = p - p_h$,

$$\begin{aligned}&\left\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}} \\ &\quad\leq \beta \sup_{(\boldsymbol{\varphi},q)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}} \frac{\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{\varphi}) - (p-p_h, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega}}{\|(\boldsymbol{\varphi}, q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}}\end{aligned} \tag{4.31}$$

with a constant $\beta > 0$. First of all we estimate the numerator. We use (4.16) to obtain a start equality

$$\begin{aligned}\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{\varphi}) - (p-p_h, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} = (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} - \delta\,(\lambda, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} \\ - \mathcal{A}^\delta(\boldsymbol{u}_h, \boldsymbol{\varphi}) + (p_h, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} \qquad \forall\boldsymbol{\varphi}\in\mathbf{V}\end{aligned}$$

Subtracting the scalar product $(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega}$ from both sides gives us the numerator of (4.31) on the left side. On the right side we apply the discrete version of (4.16) and obtain

$$\begin{aligned}&\mathcal{A}^\delta(\boldsymbol{u}-\boldsymbol{u}_h, \boldsymbol{\varphi}) - (p-p_h, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega} \\ &\quad= (\boldsymbol{f}, \boldsymbol{\varphi})_{0,\Omega} - \delta\,(\lambda, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - \mathcal{A}^\delta(\boldsymbol{u}_h, \boldsymbol{\varphi}) + (p_h, \nabla\cdot\boldsymbol{\varphi})_{0,\Omega} - (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega} \\ &\quad= (\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - \mathcal{A}^\delta(\boldsymbol{u}_h, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi}) + (p_h, \nabla\cdot(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} \\ &\qquad - \underbrace{\left(\mathcal{A}^\delta(\boldsymbol{u}_h, I_h\boldsymbol{\varphi}) - (p_h, \nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega} + \delta\,(\lambda_h, \nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega} - (\boldsymbol{f}, I_h\boldsymbol{\varphi})_{0,\Omega}\right)}_{=0} \\ &\qquad - \delta\,(\lambda, \nabla\cdot(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} - (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega} + \delta\,(\lambda_h - \lambda, \nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega} \\ &\quad= (\boldsymbol{f}, \boldsymbol{\varphi} - I_h\boldsymbol{\varphi})_{0,\Omega} - \mu\,(\nabla\boldsymbol{u}_h, \nabla(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} + (p_h, \nabla\cdot(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} \\ &\qquad + \delta\,(\nabla\cdot\boldsymbol{u}_h, \nabla\cdot(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} - \delta\,(\lambda, \nabla\cdot(\boldsymbol{\varphi} - I_h\boldsymbol{\varphi}))_{0,\Omega} \\ &\qquad - (\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h), q)_{0,\Omega} + \delta\,(\lambda_h - \lambda, \nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega} \ .\end{aligned}$$

The first part of the right side of the equation can be estimated with (4.27). We consider the second part separately, using the equality (4.17) as well as the Cauchy-Schwarz and the Young's inequalities:

$$
\begin{aligned}
&\delta\left(\nabla\cdot\boldsymbol{u}_h-\lambda,\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})\right)_{0,\Omega}-(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h),q)_{0,\Omega}+\delta\left(\lambda_h-\lambda,\nabla\cdot(I_h\boldsymbol{\varphi})\right)_{0,\Omega}\\
&\quad=\delta\left(\nabla\cdot\boldsymbol{u}_h-\lambda,\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})\right)_{0,\Omega}-(\lambda-\nabla\cdot\boldsymbol{u}_h,q)_{0,\Omega}+\delta\left(\lambda_h-\lambda,\nabla\cdot(I_h\boldsymbol{\varphi})\right)_{0,\Omega}\\
&\quad=\delta\left(\nabla\cdot\boldsymbol{u}_h-\underbrace{\lambda_h+\lambda_h}_{=0}-\lambda,\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})\right)_{0,\Omega}-\left(\lambda-\underbrace{\lambda_h+\lambda_h}_{=0}-\nabla\cdot\boldsymbol{u}_h,q\right)_{0,\Omega}\\
&\quad+\delta\left(\lambda_h-\lambda,\nabla\cdot(I_h\boldsymbol{\varphi})\right)_{0,\Omega}\\
&\quad=(\lambda_h-\lambda,\delta\nabla\cdot\boldsymbol{\varphi}+q)_{0,\Omega}+(\nabla\cdot\boldsymbol{u}_h-\lambda_h,\delta\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})+q)_{0,\Omega}\\
&\quad\leq\|\lambda_h-\lambda\|_{0,\Omega}\left(\delta\|\nabla\cdot\boldsymbol{\varphi}\|_{0,\Omega}+\|q\|_{0,\Omega}\right)\\
&\quad+\|\nabla\cdot\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}\left(\delta\|\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})\|_{0,\Omega}+\|q\|_{0,\Omega}\right)\\
&\quad\leq\left(\|\lambda_h-\lambda\|_{0,\Omega}+\|\nabla\cdot\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}\right)\left(2\delta^2\|\nabla\boldsymbol{\varphi}\|_{0,\Omega}^2+2\|q\|_{0,\Omega}^2\right)^{\frac{1}{2}}\\
&\quad\leq c_\delta\left\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\left(\|\lambda_h-\lambda\|_{0,\Omega}+\|\nabla\cdot\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}\right)
\end{aligned}
$$

with $c_\delta=\sqrt{2}\max\{1,\delta\}$. Combining the results from above and the inequalities (4.27) and (4.31) we receive:

$$
\begin{aligned}
&\left\|(\boldsymbol{u}-\boldsymbol{u}_h,p-p_h)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\\
&\leq\beta\sup_{(\boldsymbol{\varphi},q)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}}\frac{\mathcal{A}^\delta(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h),\boldsymbol{\varphi})-(p-p_h,\nabla\cdot\boldsymbol{\varphi})_{0,\Omega}-(\nabla\cdot(\boldsymbol{u}-\boldsymbol{u}_h),q)_{0,\Omega}}{\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}}\\
&\leq\beta\sup_{(\boldsymbol{\varphi},q)^{\mathrm{T}}\in\mathbf{V}\times\mathbf{Q}}\left(\frac{c_\delta\left\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}\left(\|\lambda_h-\lambda\|_{0,\Omega}+\|\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}\right)}{\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}}+\frac{2\|\nabla\boldsymbol{\varphi}\|_{0,T}}{\|(\boldsymbol{\varphi},q)^{\mathrm{T}}\|_{\mathbf{V}\times\mathbf{Q}}}\right.\\
&\quad\left.\times\left(\sum_{T\in\mathbb{T}_h}\left(c_{|T|}^2|T|\,\|\boldsymbol{f}+\mu\Delta\boldsymbol{u}_h+\nabla p_h\|_{0,T}^2+c_{|\partial T|}^2|\partial T|\left\|\frac{\partial\boldsymbol{u}_h}{\partial n}-p_h\boldsymbol{n}\right\|_{0,\partial T\setminus\partial\Omega}^2\right)\right)^{\frac{1}{2}}\right)\\
&\leq\beta c_\delta\left(\|\lambda_h-\lambda\|_{0,\Omega}+\|\nabla\cdot\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}\right)\\
&\quad+2\beta\left(\sum_{T\in\mathbb{T}_h}\left(c_{|T|}^2|T|\,\|\boldsymbol{f}+\mu\Delta\boldsymbol{u}_h+\nabla p_h\|_{0,T}^2+c_{|\partial T|}^2|\partial T|\left\|\frac{\partial\boldsymbol{u}_h}{\partial n}-p_h\boldsymbol{n}\right\|_{0,\partial T\setminus\partial\Omega}^2\right)\right)^{\frac{1}{2}}.
\end{aligned}
$$

Now, by squaring and employing the Young's inequality ones again, the estimation

$$
\begin{aligned}
\left\|(\boldsymbol{u}-\boldsymbol{u}_h,p-p_h)^{\mathrm{T}}\right\|_{\mathbf{V}\times\mathbf{Q}}^2&\leq3\beta^2c_\delta^2\|\lambda_h-\lambda\|_{0,\Omega}^2+3\beta^2c_\delta^2\|\nabla\cdot\boldsymbol{u}_h-\lambda_h\|_{0,\Omega}^2\\
&\quad+6\beta^2\sum_{T\in\mathbb{T}_h}\left(c_{|T|}^2|T|\,\|\boldsymbol{f}+\mu\Delta\boldsymbol{u}_h+\nabla p_h\|_{0,T}^2+c_{|\partial T|}^2|\partial T|\left\|\frac{\partial\boldsymbol{u}_h}{\partial n}-p_h\boldsymbol{n}\right\|_{0,\partial T\setminus\partial\Omega}^2\right)
\end{aligned}
$$

can be obtained. Next, we apply (4.28), setting $\epsilon = 3\beta^2 c_\delta^2$,

$$\begin{aligned}\frac{1}{2}\left\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} &\leq \left(c_\lambda + 3\beta^2 c_\delta^2\right)\|\nabla\cdot\boldsymbol{u}_h - \lambda_h\|^2_{0,\Omega} \\ &+ \left(c_\lambda + 6\beta^2\right)\sum_{T\in\mathbb{T}_h}\left(c^2_{|T|}|T|\,\|\boldsymbol{f}+\mu\Delta\boldsymbol{u}_h+\nabla p_h\|^2_{0,T} + c^2_{|\partial T|}|\partial T|\left\|\frac{\partial\boldsymbol{u}_h}{\partial n} - p_h\boldsymbol{n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right) \\ &+ c_\lambda\left(\Pi_{\mathbf{\Lambda}}p_h, \lambda_h\right)_{0,\Omega} + c_\lambda\|\Pi_{\mathbf{\Lambda}}p_h - p_h\|^2_{0,\Omega}\ .\end{aligned}$$

From defining a constant c as

$$c = 2\max\left\{c_\lambda, c_\lambda + 3\beta^2c_\delta^2, \left(c_\lambda+6\beta^2\right)c^2_{|T|}, \left(c_\lambda+6\beta^2\right)c^2_{|\partial T|}\right\}$$

follows the hypothesis of the theorem, which is

$$\left\|(\boldsymbol{u}-\boldsymbol{u}_h, p-p_h)^{\mathrm{T}}\right\|^2_{\mathbf{V}\times\mathbf{Q}} \leq c\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f})\ ,$$

where

$$\begin{aligned}\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f}) = \sum_{T\in\mathbb{T}_h}&\left(|T|\,\|\boldsymbol{f}+\mu\Delta\boldsymbol{u}_h+\nabla p_h\|^2_{0,T} + |\partial T|\left\|\frac{\partial\boldsymbol{u}_h}{\partial n} - p_h\boldsymbol{n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right) \\ &+ \|\nabla\cdot\boldsymbol{u}_h - \lambda_h\|^2_{0,\Omega} + \left(\Pi_{\mathbf{\Lambda}}p_h, \lambda_h\right)_{0,\Omega} + \|\Pi_{\mathbf{\Lambda}}p_h - p_h\|^2_{0,\Omega}\end{aligned}$$

and

$$\lambda_h = \Pi_{\mathbf{\Lambda}_h}\left(\nabla\cdot\boldsymbol{u}_h - \frac{1}{\delta}p_h\right)\ . \qquad \square$$

Remark 4.7.6. *By using the inequality (4.29) instead of (4.28) we can also obtain the estimation*

$$\|\boldsymbol{u}-\boldsymbol{u}_h\|^2_{1,\Omega} \leq c_u\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f})\ ,$$

as well as the estimation

$$\|p-p_h\|^2_{0,\Omega} \leq c_p\eta(\boldsymbol{u}_h, p_h, \boldsymbol{f})\ ,$$

if we use the inequality (4.30) instead of (4.28), where $c_u, c_p \in \mathbb{R}$ are constants.

4.7.4. Errror estimator for non-conform case

The error estimator above is designed for the conform discretisation of spaces $\mathbf{V}$ and $\mathbf{Q}$, meaning $\mathbf{V}_h \subset \mathbf{V}$ and $\mathbf{Q}_h \subset \mathbf{Q}$. In the chapters 4.3 and 4.5 we used the Croizeix-Raviart-elements or the non-conform P_1-elements (see Braess [4, p. 103]) for the space $\mathbf{V}_h$,which leads to a minor problem. The error estimator for the non-conform case is almost identical, but, as already mentioned, the proof varies quoit a bit. Also, an introduction of a function from a conform discrete space is necessary.

Theorem 4.7.7. *Let $(\boldsymbol{u},p)^T \in \mathbf{V}\times\mathbf{Q}$ be the continuous solution of the mixed variation formulation for the stokes problem with cavitation, given by (4.1) to (4.3). Let $(\boldsymbol{u}_h,p_h)^T \in \mathbf{V}_h^{nc}\times\mathbf{Q}_h$ be the discrete solution with a non-conform discrete space $\mathbf{V}_h^{nc}$. Furthermore let the space $\mathbf{V}_h \subset \mathbf{V}$ be a conform discrete space. Then the norm of the error $\left\|(\boldsymbol{u}-\boldsymbol{u}_h,p-p_h)^T\right\|_{\mathbf{V}\times\mathbf{Q}}$ can be estimated with the inequality*

$$\left\|(\boldsymbol{u}-\boldsymbol{u}_h^{nc},p-p_h)^T\right\|_{\mathbf{V}\times\mathbf{Q}}^2 \leq c\eta(\boldsymbol{u}_h^{nc},p_h,\boldsymbol{f}) + c_{nc}\left\|\nabla(\boldsymbol{v}_h-\boldsymbol{u}_h^{nc})\right\|_{0,\Omega} \qquad \forall \boldsymbol{v}_h\in\mathbf{V}_h\ ,$$

where c and c_{nc} are positive constants and the functional η is the same error estimator, as in conform case

The proof of this theorem can be found after the following lemma

Lemma 4.7.8. *Let the assumptions of the theorem 4.7.7 hold, then*

$$\beta\left\|p-p_h\right\|_{0,\Omega} \leq \left\|\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc}\right\|_{0,\Omega} + 2\sqrt{\eta_{Res}(\boldsymbol{u}_h^{nc},p_h,\boldsymbol{f})}\,. \tag{4.32}$$

Proof. Since $I_h\boldsymbol{\varphi}\in\mathbf{V}_h$ and $\mathbf{V}_h\subset\mathbf{V}_h^{nc}$ as well as $\mathbf{V}_h\subset\mathbf{V}$, we take difference between (4.1) and (4.12):

$$(\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc},\nabla I_h\boldsymbol{\varphi})_{0,\Omega} = (p-p_h,\nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega}\ . \tag{4.33}$$

In the lemma 4.7.2 we did not explicitly defined $\mathbf{V}_h$ as an non-conform space, this lemma applies for $\mathbf{V}_h^{nc}$ too. Using this result and the inequality (4.27) we receive following estimation:

$$\begin{aligned}
(\nabla\cdot\boldsymbol{\varphi},p_h-p)_{0,\Omega} &= (\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}),p_h-p)_{0,\Omega} + (p_h-p,\nabla\cdot(I_h\boldsymbol{\varphi}))_{0,\Omega}\\
&= (\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}),p_h-p)_{0,\Omega} + (\nabla\boldsymbol{u}_h^{nc}-\nabla\boldsymbol{u},\nabla I_h\boldsymbol{\varphi})_{0,\Omega}\\
&= (\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}),p_h-p)_{0,\Omega} - (\nabla\boldsymbol{u}_h^{nc}-\nabla\boldsymbol{u},\nabla(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}))_{0,\Omega}\\
&\quad + (\nabla\boldsymbol{u}_h^{nc}-\nabla\boldsymbol{u},\nabla\boldsymbol{\varphi})_{0,\Omega}\\
&= (\boldsymbol{f},\boldsymbol{\varphi}-I_h\boldsymbol{\varphi})_{0,\Omega} - (\nabla\boldsymbol{u}_h^{nc},\nabla(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}))_{0,\Omega} + (p_h,\nabla\cdot(\boldsymbol{\varphi}-I_h\boldsymbol{\varphi}))_{0,\Omega}\\
&\quad + (\nabla\boldsymbol{u}_h^{nc}-\nabla\boldsymbol{u},\nabla\boldsymbol{\varphi})_{0,\Omega}\\
&\leq \left\|\nabla\boldsymbol{\varphi}\right\|_{0,\Omega}\left(\left\|\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc}\right\|_{0,\Omega} + 2\sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_h^{nc},p_h,\boldsymbol{f})}\right)
\end{aligned}$$

According to the inf-sup-condition, we can finally obtain following estimation:

$$\begin{aligned}
\beta\left\|p-p_h\right\|_{0,\Omega} &\leq \sup_{\boldsymbol{\varphi}\in\mathbf{V}}\frac{(\nabla\cdot\boldsymbol{\varphi},p-p_h)_{0,\Omega}}{\left\|\boldsymbol{\varphi}\right\|_{1,\Omega}}\\
&\leq \left\|\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc}\right\|_{0,\Omega} + 2\sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_h^{nc},p_h,\boldsymbol{f})}
\end{aligned}$$

□

Proof. **(Theorem 4.7.7)** We start by introducing $\boldsymbol{v}_h\in\mathbf{V}_h$. According to the Schwarz-inequality

$$\begin{aligned}
\left\|\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc}\right\|_{0,\Omega}^2 &= (\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc},\nabla(\boldsymbol{u}-\boldsymbol{v}_h)+\nabla(\boldsymbol{v}_h-\boldsymbol{u}_h^{nc}))_{0,\Omega}\\
&\leq (\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc},\nabla(\boldsymbol{u}-\boldsymbol{v}_h))_{0,\Omega} + \left\|\nabla\boldsymbol{u}-\nabla\boldsymbol{u}_h^{nc}\right\|_{0,\Omega}\left\|\nabla(\boldsymbol{v}_h-\boldsymbol{u}_h^{nc})\right\|_{0,\Omega}
\end{aligned}$$

Next we define $\boldsymbol{e} \in \mathbf{V}$ as $\boldsymbol{e} = \boldsymbol{u} - \boldsymbol{v}_h$ use it as test function in the equation (4.33):

$$(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}_h^{nc}, \nabla I_h \boldsymbol{e})_{0,\Omega} = (p - p_h, \nabla \cdot (I_h \boldsymbol{e}))_{0,\Omega} \ .$$

This leads, according to the lemma 4.7.8, to

$$\begin{aligned}
(\nabla \boldsymbol{u} - \nabla \boldsymbol{u}_h^{nc}, \nabla(\boldsymbol{u} - \boldsymbol{v}_h))_{0,\Omega} =& (\nabla \boldsymbol{u} - \nabla \boldsymbol{u}_h^{nc}, \nabla(\boldsymbol{e} - I_h \boldsymbol{e}))_{0,\Omega} \\
&- (p - p_h, \nabla \cdot (\boldsymbol{e} - I_h \boldsymbol{e}))_{0,\Omega} + (p - p_h, \nabla \cdot \boldsymbol{e})_{0,\Omega} \\
=& (\boldsymbol{f}, \boldsymbol{e} - I_h \boldsymbol{e})_{0,\Omega} - (\nabla \boldsymbol{u}_h^{nc}, \nabla (\boldsymbol{e} - I_h \boldsymbol{e}))_{0,\Omega} \\
&+ (p_h, \nabla \cdot (\boldsymbol{e} - I_h \boldsymbol{e}))_{0,\Omega} + (p - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{v}_h))_{0,\Omega} \\
=& 2 \left\| \nabla(\boldsymbol{u} - \boldsymbol{v}_h) \right\|_{0,\Omega} \sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_h^{nc}, p_h, \boldsymbol{f})} \\
&+ (p - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{v}_h))_{0,\Omega} \\
=& 2 \left\| \nabla(\boldsymbol{u} - \boldsymbol{v}_h) \right\|_{0,\Omega} \sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_h^{nc}, p_h, \boldsymbol{f})} \\
&+ \left\| p - p_h \right\|_{0,\Omega} \left\| \nabla \cdot (\boldsymbol{u}_h^{nc} - \boldsymbol{v}_h) \right\|_{0,\Omega} \\
&+ (p - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc}))_{0,\Omega} \, .
\end{aligned}$$

Next, we consider the term $(p - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc}))_{0,\Omega}$ using the positive projection $\Pi^+ p$

$$\begin{aligned}
(p - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc}))_{0,\Omega} =& \left(\Pi^+ p_h - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc})\right)_{0,\Omega} + \left(p - \Pi^+ p_h, \nabla \cdot \boldsymbol{u} - \lambda_h\right)_{0,\Omega} \\
&+ \left(p - \Pi^+ p_h, \lambda_h - \nabla \cdot \boldsymbol{u}_h^{nc}\right)_{0,\Omega} \\
=& \left(\Pi^+ p_h - p_h, \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc})\right)_{0,\Omega} + \underbrace{(p, \nabla \cdot \boldsymbol{u})_{0,\Omega}}_{=0} \\
&- \underbrace{\left(\Pi^+ p_h, \nabla \cdot \boldsymbol{u}\right)_{0,\Omega}}_{\geq 0} - \underbrace{(p, \lambda_h)_{0,\Omega}}_{\geq 0} + \left(\Pi^+ p_h, \lambda_h\right)_{0,\Omega} \\
&+ (p - p_h, \lambda_h - \nabla \cdot \boldsymbol{u}_h^{nc})_{0,\Omega} \\
\leq& \left\| \Pi^+ p_h - p_h \right\|_{0,\Omega} \left\| \nabla \cdot (\boldsymbol{u} - \boldsymbol{u}_h^{nc}) \right\|_{0,\Omega} + \left(\Pi^+ p_h, \lambda_h\right)_{0,\Omega} \\
&+ \left\| p - p_h \right\|_{0,\Omega} \left\| \nabla \cdot \boldsymbol{u}_h^{nc} - \lambda_h \right\|_{0,\Omega}
\end{aligned}$$

Combining the results above and applying the Young's inequality with positive constants c_{y1}, c_{y2}, c_{y3} and c_{y4} leads to the estimation

$$\begin{aligned}
\left\| \nabla \boldsymbol{u} - \nabla \boldsymbol{u}_h^{nc} \right\|_{0,\Omega}^2 \leq& \left\| \nabla \boldsymbol{u} - \nabla \boldsymbol{u}_h^{nc} \right\|_{0,\Omega} \left\| \nabla(\boldsymbol{v}_h - \boldsymbol{u}_h^{nc}) \right\|_{0,\Omega} + \left\| p - p_h \right\|_{0,\Omega} \left\| \nabla \left(\boldsymbol{u}_h^{nc} - \boldsymbol{v}_h\right) \right\|_{0,\Omega} \\
&+ 2 \left(\left\| \nabla(\boldsymbol{u} - \boldsymbol{u}_h^{nc}) \right\|_{0,\Omega} + \left\| \nabla(\boldsymbol{u}_h^{nc} - \boldsymbol{v}_h) \right\|_{0,\Omega} \right) \sqrt{\eta_{\mathrm{Res}}(\boldsymbol{u}_h^{nc}, p_h, \boldsymbol{f})} \\
&+ \left\| p - p_h \right\|_{0,\Omega} \left\| \nabla \cdot \boldsymbol{u}_h^{nc} - \lambda_h \right\|_{0,\Omega} \\
&+ \left\| \Pi^+ p_h - p_h \right\|_{0,\Omega} \left\| \nabla \left(\boldsymbol{u} - \boldsymbol{u}_h^{nc}\right) \right\|_{0,\Omega} + \left(\Pi^+ p_h, \lambda_h\right)_{0,\Omega}
\end{aligned}$$

Together with the results of the lemma 4.7.8, the appropriate set of the constants for this inequality leads to the error estimator:

$$\left\| (\boldsymbol{u} - \boldsymbol{u}_h^{nc}, p - p_h)^{\mathrm{T}} \right\|_{\mathbf{V} \times \mathbf{Q}}^2 \leq c \eta(\boldsymbol{u}_h^{nc}, p_h, \boldsymbol{f}) + c_{nc} \left\| \nabla(\boldsymbol{v}_h - \boldsymbol{u}_h^{nc}) \right\|_{0,\Omega} \ . \qquad \square$$

5. Revisiting Obstacle problem in 2-D

In this section we return to the obstacle problem (see f.e. Biermann et al [2]), but now on a domain $\Omega \subset \mathbb{R}^2$, which in classical notation reads

$$\begin{aligned} -\Delta u - f &\geq 0\,, \\ u - \psi &\geq 0\,, \\ (u-\psi)(-\Delta u - f) &= 0\,, \end{aligned} \tag{5.1}$$

with $u : \Omega \to \mathbb{R}$ sufficiently smooth and the Dirichlet boundary condition $u = 0$ on $\partial\Omega$. Here, $f : \Omega \to \mathbb{R}$ is a so called right-hand-side function, which describes the effect of external forces. The function $\psi : \Omega \to \mathbb{R}$ represents the obstacle. The purpose of this section is to extend the problem by considering the 2-D case, introduce stabilization technique and compare numerical results.

5.1. Introduction of the mixed formulation

Analogous to the chapter 3.1, the system (5.1) can be rewritten as a variational inequality. A new obstacle type variational problem can be described as a search for $u \in \mathbf{K}$ that satisfied the inequality

$$(\nabla u, \nabla(\varphi - u))_{0,\Omega} \geq (f, \varphi - u)_{0,\Omega} \qquad \forall \varphi \in \mathbf{K}\,,$$

where we set $\mathbf{V} := \mathbf{H}_0^1(\Omega)$ and $\mathbf{K} := \{v \in \mathbf{V} \mid v \geq \psi \text{ a.e. in } \Omega\}$ and assume $\psi \in \mathbf{H}^1(\Omega)$ as well as $f \in \mathbf{L}^2(\Omega)$.

This variational problem has also an equivalent minimization problem: Find $u \in \mathbf{K}$ with

$$\mathcal{J}(u) = \inf_{\varphi \in \mathbf{K}} \mathcal{J}(\varphi)\,,$$

where $\mathcal{J}(\varphi) = \frac{1}{2}(\nabla\varphi, \nabla\varphi)_{0,\Omega} - (f, \varphi)_{0,\Omega}$. In the chapter 3.2 we already offered the motivation for the introduction of the Lagrange-multipliers, that are used in the corresponding solution strategy. At this point we would like to refer to the section 6.2.2, where the generalized case is discussed, and just check the condition of the theorem 6.1.9.

Fist of all the subset $\mathbf{K}$ is described using continuous linear operator $\breve{Id} : \mathbf{L}^2(\Omega) \to \mathbf{L}^2(\Omega)$ with $\varphi \mapsto \varphi$. Also according to the definition

$$\sup_{\varphi \in \mathbf{H}^1(\Omega)} \frac{\left(q, \breve{Id}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{1,\Omega}} = \sup_{\varphi \in \mathbf{H}^1(\Omega)} \frac{(q,\varphi)_{0,\Omega}}{\|\varphi\|_{1,\Omega}} = \|q\|_{0,\Omega} \qquad \forall q \in \mathbf{L}^2(\Omega)\,.$$

As show in the theorem 6.1.8 the subset $\mathbf{K}$ is convex. Next we check the following points

- $\mathbf{V}$ is a reflexive Banach space,
- $\mathcal{J}$ has a gradient $\mathcal{J}'(u) \in \mathbf{V}^*$ everywhere in $\mathbf{K}$, such that

$$\left\langle \mathcal{J}'(u), \varphi \right\rangle_{\mathbf{V}} = (\nabla\varphi, \nabla u)_{0,\Omega} - (f, \varphi)_{0,\Omega}$$

- $\mathcal{J}$ is twice Gateaux-differentiable in all directions and satisfies the condition

$$\left\langle \mathcal{J}''(u)\varphi, \varphi \right\rangle_{\mathbf{V}} = (\nabla\varphi, \nabla\varphi)_{0,\Omega} \geq \frac{1}{2} \|\varphi\|^2_{1,\Omega} \ .$$

Combined, we can conclude that all the conditions of the theorem 6.1.9, which means, that there is an equivalent problem: Find a point $(u, p, \lambda)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q} \times \mathbf{\Lambda}$ such that

$$(\nabla\varphi, \nabla u)_{0,\Omega} - (p, \varphi)_{0,\Omega} = (f, \varphi)_{0,\Omega} \qquad \forall\varphi \in \mathbf{V} \tag{5.2}$$

$$(q, u)_{0,\Omega} - (q, \lambda)_{0,\Omega} = (q, \psi)_{0,\Omega} \qquad \forall q \in \mathbf{Q} \tag{5.3}$$

$$(p, \omega)_{0,\Omega} \geq 0 \qquad \forall\omega \in \mathbf{\Lambda} \tag{5.4}$$

$$p\lambda = 0 \qquad \text{a.e. on } \Omega\,, \tag{5.5}$$

with $\mathbf{Q} = \mathbf{L}^2(\Omega)$ and $\mathbf{\Lambda} = \mathit{\Pi}^+\mathbf{Q} = \{\omega \in \mathbf{Q} |\ \omega \geq 0 \text{ a.e. on } \Omega\}$. Also according to the theorems 6.1.5 and 6.1.9 both the initial variational obstacle problem and the equivalent problem above have a unique solution. Furthermore the function $u \in \mathbf{V}$, which is the first part of the solution of the equivalent problem above, is the solution of the initial variational obstacle problem.

5.2. Numerical treatment of the mixed obstacle problem

We apply, like in Birmann et al [2], the finite element method to compute an approximate solution on the triangulation $\mathbb{T}_h$ of Ω. Based on the mesh, we introduce standard finite element spaces $\mathbf{V_h} = \mathcal{Q}_1(\mathbb{T}_h)$ and $\mathbf{Q_h} = \mathcal{Q}_0(\mathbb{T}_h)$, which means, that we intend to deal with quadrangular mesh elements and bilinear base functions in case of the space $\mathbf{V_h}$, as well as piecewise constant base functions in case of the space $\mathbf{Q_h}$. This is a conform discretisation since $\mathbf{V_h} \subset \mathbf{V}$ and $\mathbf{Q_h} \subset \mathbf{L}^2(\Omega)$ The new goal is computing a approximate, or discrete, solution $(u_h, p_h, \lambda_h) \in \mathbf{V_h} \times \mathbf{Q_h} \times \mathbf{\Lambda_h}$ corresponding discrete version of the equations (5.2), (5.3) and (5.5) as well as the inequality (5.4) reads

$$(\nabla\varphi_h, \nabla u_h)_{0,\Omega} - (\varphi_h, p_h)_{0,\Omega} = (f, \varphi_h)_{0,\Omega} \qquad \forall\varphi_h \in \mathbf{V_h} \tag{5.6}$$

$$(u_h, q_h)_{0,\Omega} - (\lambda_h, q_h)_{0,\Omega} = (\psi, q_h)_{0,\Omega} \qquad \forall q_h \in \mathbf{Q_h} \tag{5.7}$$

$$(\omega_h, p_h)_{0,\Omega} \geq 0 \qquad \forall\omega_h \in \mathbf{\Lambda_h} \tag{5.8}$$

$$p_h\lambda_h = 0 \qquad \text{a.e. on } \Omega\,, \tag{5.9}$$

respectively, where $\mathbf{\Lambda_h} = \mathit{\Pi}^+\mathbf{Q_h} = \{\omega_h \in \mathbf{Q_h} |\ \omega_h \geq 0 \text{ a.e. on } \Omega\}$. Like in continuous case, according to the theorem 6.1.9 there is a unique solution $(u_h, p_h, \lambda_h) \in \mathbf{V_h} \times \mathbf{Q_h} \times \mathbf{\Lambda_h}$,

such that is equivalent to the discrete minimization problem, if the inf sup condition is satisfied: There is a constant $\beta \in \mathbb{R}^+$, such that

$$\sup_{\varphi \in \mathbf{V_h}} \frac{(\varphi_h, q_h)_{0,\Omega}}{\|\varphi_h\|_{1,\Omega}} \geq \beta \|q_h\|_{0,\Omega} \qquad \forall q_h \in \mathbf{Q_h} \,. \tag{5.10}$$

By applying the theorem 6.1.10 we obtain following variational problem: Find $(u_h, p_h) \in \mathbf{V_h} \times \mathbf{Q_h}$ fulfilling the mixed formulation

$$\begin{aligned} (\nabla u_h, \nabla \varphi_h)_{0,\Omega} - (p_h, \varphi_h)_{0,\Omega} &= (f, \varphi_h)_{0,\Omega} && \forall \varphi_h \in \mathbf{V_h} \\ (u_h, q_h)_{0,\Omega} &= (\psi + \lambda_h, q_h)_{0,\Omega} && \forall q_h \in \mathbf{Q_h} \\ \text{with} \qquad \lambda_h &= \max\left\{0, u_h - \psi - \frac{1}{\delta} p_h\right\}. \end{aligned}$$

As in case of the stokes problem we want to apply the Newton-type method to solve this problem, which means we approximate the zero spot $(u_h, p_h) \in \mathbf{V_h} \times \mathbf{Q_h}$ of the functional $\mathcal{F}_{\varphi_h q_h} : \mathbf{V_h} \times \mathbf{Q_h} \longrightarrow \mathbb{R}$ with

$$\begin{aligned} \mathcal{F}_{\varphi_h q_h}(\tilde{u}_h, \tilde{p}_h) = &(\nabla \tilde{u}_h, \nabla \varphi_h)_{0,\Omega} - (\tilde{p}_h, \varphi_h)_{0,\Omega} - (f, \varphi_h)_{0,\Omega} \\ &+ (\tilde{u}_h, q_h)_{0,\Omega} - \left(\psi + \tilde{\lambda}_h, q_h\right)_{0,\Omega} \\ \text{and} \quad \tilde{\lambda}_h = &\frac{1}{2}\left(\tilde{u}_h - \psi - \frac{1}{\delta}\tilde{p}_h\right) + \frac{1}{2}\left|\tilde{u}_h - \psi - \frac{1}{\delta}\tilde{p}_h\right| \,. \end{aligned} \tag{5.11}$$

The Newton-type method itself is similar to $\mathcal{P}_3$ update calculation algorithm from chapter 4.4, but with two differences. First we multiply the update with a relaxation factor $\gamma_r \in \mathbb{R}$. Then the iteration process can be written as

$$\left(u_h^{(k+1)}, p_h^{(k+1)}\right)^{\mathrm{T}} = \left(u_h^{(k)}, p_h^{(k)}\right)^{\mathrm{T}} - \gamma_r \left(d_h^u, d_h^p\right)^{\mathrm{T}} \qquad \forall k \in \mathbb{N}_0 \,,$$

where $\left(u_h^{(0)}, p_h^{(0)}\right)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ is a chosen start point and the update $\left(d_h^u, d_h^p\right)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ solves the equation

$$\left\langle \tilde{\mathcal{F}}'_{\varphi_h q_h}(u_h^{(k)}, p_h^{(k)}), \left(d_h^u, d_h^p\right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h} \times \mathbf{Q_h}} = \mathcal{F}_{\varphi_h q_h}(u_h^{(k)}, p_h^{(k)}) \qquad \forall \left(\varphi_h, q_h\right)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h} \,.$$

In chapter 4.4 we used consistent stabilization for the first equation by using the second equation with appropriate test function. Here the condition number of the resulting matrix can be improved by non-consistent stabilization. More specific, in order to stabilize the the equation (5.7), we modify this discrete formulation by adding mesh-dependent terms to the original problem. Eventually we consider substitutional equation

$$(u_h, q_h)_{0,\Omega} + c_{\mathrm{Q}} \sum_{T \in \mathbb{T}_h} |T| \, (p_h, q_h)_{0,T} = (\psi + \lambda_h, q_h)_{0,\Omega} \qquad \forall q_h \in \mathbf{Q_h}, \tag{5.12}$$

where $c_{\mathrm{Q}} > 0$ is a constant. This is stabilization method based on the ideas discussed in Biermann et al. [2].

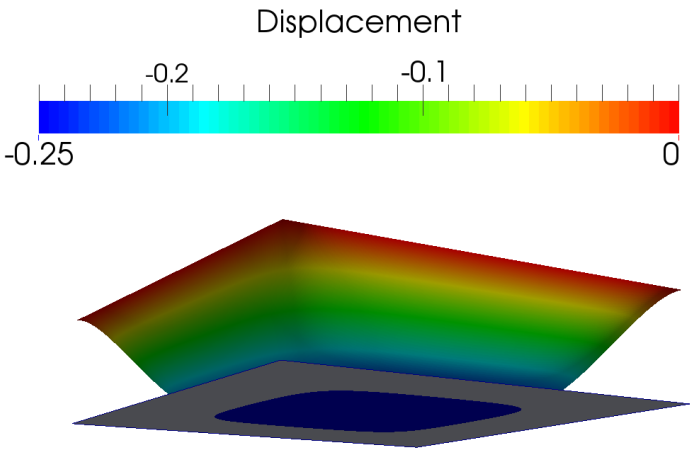

Figure 5.1.: The test example for the obstacle problem with $\psi = -0.25$.

The main difference between the functional $\mathcal{F}$ and it's regularized version $\tilde{\mathcal{F}}$ is the use of $\tilde{\lambda}_h^\xi$ instead of $\tilde{\lambda}_h$, where

$$\tilde{\lambda}_h^\xi = \frac{1}{2}\left(\tilde{u}_h - \psi - \frac{1}{\delta}\tilde{p}_h\right) + \frac{1}{2}\sqrt{\left(\tilde{u}_h - \psi - \frac{1}{\delta}\tilde{p}_h\right)^2 + \xi}\,, \tag{5.13}$$

So with regard to the changed equation (5.12), we can define the functional $\tilde{\mathcal{F}}$ as

$$\begin{aligned}\tilde{\mathcal{F}}_{\varphi_h q_h}(\tilde{u}_h, \tilde{p}_h) = &(\nabla\tilde{u}_h, \nabla\varphi_h)_{0,\Omega} - (\tilde{p}_h, \varphi_h)_{0,\Omega} - (f, \varphi_h)_{0,\Omega} \\ &+ (\tilde{u}_h, q_h)_{0,\Omega} - \left(\psi + \lambda_h^\xi, q_h\right)_{0,\Omega} + c_{\mathrm{Q}} \sum_{T\in\mathbb{T}_h} |T|\,(\tilde{p}_h, q_h)_{0,T}\,. \end{aligned} \tag{5.14}$$

5.3. Numerical results

The figure 5.1 depicts the results of the numerical test for the obstacle problem. Here we have a membrane displacement on $\Omega = [0,1] \times [0,1]$ with zero boundary conditions $u|_{\partial\Omega} = 0$. The right-hand-side function and the obstacle are constant on Ω, so that $f|_\Omega = 10$ and $\psi|_\Omega = -0.25$.

The constant, we use in the the calculations, are $c_{\mathrm{Q}} = 1$ and $\delta = 0.5$. The relaxation factor γ_r is different in each of the iteration steps. For $r \in \{0,1,2,3,4,5\}$ we try the relaxation factors $\gamma_r = \left(\frac{1}{2}\right)^r$ in that order and calculate therefor $\mathcal{F}_{\varphi_h q_h}\left(u_h^{(k+1)}, p_h^{(k+1)}\right)$, where

cells	Newton-type method	SQOPT	cGPSSOR	precond. Uzawa
64	3	4	11	8
256	6	18	20	14
1024	10	86	46	30
4096	13	304	140	58
16384	21	1188	444	263
65536	28	4632	1534	428

Table 5.1.: A comparison of needed iterations for proposed Newton-type algorithm and the other methods, found in Biermann et al. [2].

$\left(u_h^{(k+1)}, p_h^{(k+1)}\right)^{\mathrm{T}}$ is the new approximation of the zero spot. If the current relaxation factor improves the approximation, we test the next one, if not the previous relaxation factor is used.

The reason for choosing this specific problem is, that in Biermann et al. [2] the exact same test example was used. So instead of implementing all of the algorithms we can compare the iteration numbers. Even so the calculation time for each of the methods can vary significantly, the table 5.1 shows, that the rate, which the number of iteration rises with growing the number of cells, is much low that by other algorithms.

5.4. Error estimate

In the next sections we derive a posteriori error estimator, based on the solution strategy, introduced above. If $(u,p)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q}$ is the continuous solution of the mixed variation formulation for the obstacle problemand $(u_h, p_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ is the discrete solution, then the norm of the error $\|u - u_h\|_{0,\Omega}$ can be estimated with the inequality

$$\|u - u_h\|_{0,\Omega}^2 \leq c\,\eta(u_h, p_h, f, \psi)\,, \tag{5.15}$$

where c is a positive constant and

$$\begin{aligned}\eta(u_h, p_h, f, \psi) = \sum_{T \in \mathbb{T}_h} \left(|T|\,\|f + \Delta u_h + p_h\|_{0,T}^2 + |\partial T| \left\| \frac{\partial u_h}{\partial n} \right\|_{0,\partial T \setminus \partial\Omega}^2 \right) \\ + \left(\Pi^+ p_h, \lambda_h\right)_{0,\Omega} + \|u_h - \psi - \lambda_h\|_{0,\Omega}^2 + \left\|\Pi^+ p_h - p_h\right\|_{0,\Omega}^2\,.\end{aligned}$$

It has similar structure as the error estimator in for the stokes problem (see chapter 4.7). Four of the summands can be identified with the KKT-condition (5.2) to (5.5) and the norm $\left\| \frac{\partial u_h}{\partial n} \right\|_{0,\partial T \setminus \partial\Omega}$ measures the differentiability of the numerical solution u_h on the edges of the cells.

We proof this hypothesis, using a technique similar to strategy e.g. found in the proof of lemma 3.1 in Duran e. a. [5]. The proof consists of three steps:

- First we define residual error estimator $\eta_{\text{Res}}(u_h, p_h, f)$. (Lemma 5.4.1)
- Then we use it to get an estimate of the error $\|p - p_h\|_{0,\Omega}$. (Lemma 5.4.2)
- Finally, the steps above result in the proof of the inequality (5.15). (Theorem 5.4.3)

Lemma 5.4.1. *For all $u_h \in \mathbf{V_h}$, $p_h \in \mathbf{Q_h}$ and $f, \varphi \in \mathbf{V}$ the following inequality applies*

$$\begin{aligned}(f, \varphi - I_h\varphi)_{0,\Omega} - (\nabla u_h, \nabla(\varphi - I_h\varphi))_{0,\Omega} + (p_h, \varphi - I_h\varphi)_{0,\Omega} \\ \leq 2\|\nabla\varphi\|_{0,\Omega}\sqrt{\eta_{Res}(u_h, p_h, f)} \end{aligned} \tag{5.16}$$

where $$\eta_{Res}(u_h, p_h, f) = \sum_{T\in\mathbb{T}_h}\left(c^2_{|T|}|T|\,\|f + \Delta u_h + p_h\|^2_{0,T} + c^2_{|\partial T|}|\partial T|\left\|\frac{\partial u_h}{\partial n}\right\|^2_{0,\partial T\setminus\partial\Omega}\right).$$

Proof. Same as in 4.7.2. □

Lemma 5.4.2. *Let $(u, p)^T \in \mathbf{V}\times\mathbf{Q}$ be the continuous solution of the mixed variation formulation for the obstacle problem, given by (5.2) to (5.5). Let $(u_h, p_h)^T \in \mathbf{V_h}\times\mathbf{Q_h}$ be the discrete solution, obtained with method describe in section 5.2. Then the norm of the error $\|p - p_h\|_{0,\Omega}$ can be estimated with the inequality*

$$\beta\|p - p_h\|_{0,\Omega} \leq \|\nabla u - \nabla u_h\|_{0,\Omega} + 2\sqrt{\eta_{Res}(u_h, p_h, f)}\,,$$

where $\eta_{Res}(u_h, p_h, f)$ defined in the lemma 5.4.1.

Proof. First we take difference between (5.2) and (5.6):

$$(\nabla u - \nabla u_h, \nabla I_h\varphi)_{0,\Omega} = (p - p_h, I_h\varphi)_{0,\Omega}\ .$$

Using this result and the inequality (5.16) we receive following estimation:

$$\begin{aligned}(\varphi, p_h - p)_{0,\Omega} &= (\varphi - I_h\varphi, p_h - p)_{0,\Omega} + (p_h - p, I_h\varphi)_{0,\Omega} \\ &= (\varphi - I_h\varphi, p_h - p)_{0,\Omega} + (\nabla u_h - \nabla u, \nabla I_h\varphi)_{0,\Omega} \\ &= (\varphi - I_h\varphi, p_h - p)_{0,\Omega} - (\nabla u_h - \nabla u, \nabla(\varphi - I_h\varphi))_{0,\Omega} \\ &\quad + (\nabla u_h - \nabla u, \nabla\varphi)_{0,\Omega} \\ &= (f, \varphi - I_h\varphi)_{0,\Omega} - (\nabla u_h, \nabla(\varphi - I_h\varphi))_{0,\Omega} + (p_h, \varphi - I_h\varphi)_{0,\Omega} \\ &\quad + (\nabla u_h - \nabla u, \nabla\varphi)_{0,\Omega} \\ &\leq \|\nabla\varphi\|_{0,\Omega}\left(\|\nabla u - \nabla u_h\|_{0,\Omega} + 2\sqrt{\eta_{\text{Res}}(u_h, p_h, f)}\right)\end{aligned}$$

According to the inf-sup-condition, we can finally obtain following estimation:

$$\begin{aligned}\beta\|p - p_h\|_{0,\Omega} &\leq \sup_{\varphi\in\mathbf{V}}\frac{(\varphi, p - p_h)_{0,\Omega}}{\|\varphi\|_{1,\Omega}} \\ &\leq \|\nabla u - \nabla u_h\|_{0,\Omega} + 2\sqrt{\eta_{\text{Res}}(u_h, p_h, f)}\end{aligned}$$

□

Theorem 5.4.3. *Let $(u,p)^T \in \mathbf{V}\times\mathbf{Q}$ be the continuous solution of the mixed variation formulation for the obstacle problem, given by (5.2) to (5.5). Let $(u_h,p_h)^T \in \mathbf{V_h}\times\mathbf{Q_h}$ be the discrete solution, obtained with method describe in section 5.2. Then the norm of the error $\|u-u_h\|_{0,\Omega}$ can be estimated with the inequality*

$$\|u-u_h\|_{0,\Omega}^2 \le c\,\eta(u_h,p_h,f,\psi)\,,$$

where c is a positive constant and

$$\begin{aligned}\eta(u_h,p_h,f,\psi) = \sum_{T\in\mathbb{T}_h}\left(|T|\,\|f+\Delta u_h+p_h\|_{0,T}^2 + |\partial T|\left\|\frac{\partial u_h}{\partial n}\right\|_{0,\partial T\setminus\partial\Omega}^2\right)\\ + \left(\Pi^+p_h,\lambda_h\right)_{0,\Omega} + \|u_h-\psi-\lambda_h\|_{0,\Omega}^2 + \left\|\Pi^+p_h-p_h\right\|_{0,\Omega}^2\,.\end{aligned}$$

Proof. Let $e\in\mathbf{V}$ be defined as $e=u-u_h$, then ,as stated in proof of the lemma 5.4.2,

$$(\nabla u-\nabla u_h,\nabla I_h e)_{0,\Omega} = (p-p_h,I_h e)_{0,\Omega}\,,$$

which leads, according to the lemma 4.7.2, to

$$\begin{aligned}\|\nabla u-\nabla u_h\|_{0,\Omega}^2 =& (\nabla u-\nabla u_h,\nabla e-I_h e)_{0,\Omega} - (p-p_h,e-I_h e)_{0,\Omega} + (p-p_h,e)_{0,\Omega}\\ =& (f,e-I_h e)_{0,\Omega} - (\nabla u_h,\nabla(e-I_h e))_{0,\Omega} + (p_h,e-I_h e)_{0,\Omega}\\ &+ (p-p_h,u-u_h)_{0,\Omega}\\ =& 2\,\|\nabla(u-u_h)\|_{0,\Omega}\sqrt{\eta_{\mathrm{Res}}(u_h,p_h,f)} + (p-p_h,u-u_h)_{0,\Omega}\,.\end{aligned}$$

Next, we consider the term $(p-p_h,u-u_h)_{0,\Omega}$ using the positive projection Π^+p

$$\begin{aligned}(p-p_h,u-u_h)_{0,\Omega} =& \left(\Pi^+p_h-p_h,u-u_h\right)_{0,\Omega} + \left(p-\Pi^+p_h,u-\psi-\lambda_h\right)_{0,\Omega}\\ &+ \left(p-\Pi^+p_h,\lambda_h+\psi-u_h\right)_{0,\Omega}\\ =& \left(\Pi^+p_h-p_h,u-u_h\right)_{0,\Omega} + \underbrace{(p,u-\psi)_{0,\Omega}}_{=0} - \underbrace{\left(\Pi^+p_h,u-\psi\right)_{0,\Omega}}_{\ge 0}\\ &- \underbrace{(p,\lambda_h)_{0,\Omega}}_{\ge 0} + \left(\Pi^+p_h,\lambda_h\right)_{0,\Omega} + (p-p_h,\lambda_h+\psi-u_h)_{0,\Omega}\\ \le& \left\|\Pi^+p_h-p_h\right\|_{0,\Omega}\|u-u_h\|_{0,\Omega} + \left(\Pi^+p_h,\lambda_h\right)_{0,\Omega}\\ &+ \|p-p_h\|_{0,\Omega}\,\|u_h-\psi-\lambda_h\|_{0,\Omega}\\ \le& \left\|\Pi^+p_h-p_h\right\|_{0,\Omega}\|u-u_h\|_{0,\Omega} + \left(\Pi^+p_h,\lambda_h\right)_{0,\Omega}\\ &+ \frac{1}{\beta}\,\|u_h-\psi-\lambda_h\|_{0,\Omega}\left(\|\nabla u-\nabla u_h\|_{0,\Omega} + 2\sqrt{\eta_{\mathrm{Res}}(u_h,p_h,f)}\right)\end{aligned}$$

Combining the results above and applying the Young's inequality with positive constants

c_{y1}, c_{y2}, c_{y3} and c_{y4} leads to the estimation

$$\begin{aligned}\|\nabla u - \nabla u_h\|_{0,\Omega}^2 \leq & 2\,\|\nabla(u-u_h)\|_{0,\Omega}\,\sqrt{\eta_{\mathrm{Res}}(u_h,p_h,f)} \\ & + \left\|\Pi^+ p_h - p_h\right\|_{0,\Omega}\|\nabla(u-u_h)\|_{0,\Omega} + \left(\Pi^+ p_h,\lambda_h\right)_{0,\Omega} \\ & + \frac{1}{\beta}\,\|u_h-\psi-\lambda_h\|_{0,\Omega}\left(\|\nabla u-\nabla u_h\|_{0,\Omega} + 2\sqrt{\eta_{\mathrm{Res}}(u_h,p_h,f)}\right) \\ \leq & \left(c_{y1}+\frac{c_{y2}}{2}+\frac{c_{y3}}{2}\right)\|\nabla(u-u_h)\|_{0,\Omega}^2 + \left(\frac{1}{c_{y1}}+c_{y4}\right)\eta_{\mathrm{Res}}(u_h,p_h,f) \\ & + \frac{1}{\beta^2}\left(\frac{1}{2c_{y3}}+\frac{1}{c_{y4}}\right)\|u_h-\psi-\lambda_h\|_{0,\Omega}^2 \\ & + \frac{1}{2c_{y2}}\left\|\Pi^+ p_h - p_h\right\|_{0,\Omega}^2 + \left(\Pi^+ p_h,\lambda_h\right)_{0,\Omega}\end{aligned}$$

Finally, we set $c_{y1} = c_{y2} = c_{y3} = \frac{1}{4}$ and $c_{y4} = 1$, so that

$$\begin{aligned}\|\nabla u - \nabla u_h\|_{0,\Omega}^2 \leq & \frac{1}{2}\,\|\nabla(u-u_h)\|_{0,\Omega}^2 + 5\,\eta_{\mathrm{Res}}(u_h,p_h,f) + \left(\Pi^+ p_h,\lambda_h\right)_{0,\Omega} \\ & + \frac{3}{\beta^2}\,\|u_h-\psi-\lambda_h\|_{0,\Omega}^2 + 2\left\|\Pi^+ p_h - p_h\right\|_{0,\Omega}^2\,,\end{aligned}$$

which leads for $c = \max\left\{\frac{12}{\beta^2}, 20c_{|T|}^2, 20c_{|\partial T|}^2, 8\right\}$ to the error estimation

$$\begin{aligned}\frac{1}{c}\,\|u-u_h\|_{1,\Omega}^2 \leq & \sum_{T\in\mathbb{T}_h}\left(|T|\,\|f+\Delta u_h+p_h\|_{0,T}^2 + |\partial T|\left\|\frac{\partial u_h}{\partial n}\right\|_{0,\partial T\setminus\partial\Omega}^2\right) \\ & + \left(\Pi^+ p_h,\lambda_h\right)_{0,\Omega} + \|u_h-\psi-\lambda_h\|_{0,\Omega}^2 + \left\|\Pi^+ p_h - p_h\right\|_{0,\Omega}^2\,. \qquad \square\end{aligned}$$

6. Framework for first kind problems with linear inequality conditions

In the following, in order to provide general framework for broader class of problems, we apply the ideas sketched examples above (Stokes and obstacle problems) to an abstract setting. We start with a minimization problem on a subset, that can be derived, for example, from applied laws of continuous mechanics. This abstract problem can be written as: Find $\boldsymbol{u} \in \mathbf{K}$ such that

$$\mathcal{J}(\boldsymbol{u}) = \inf_{\boldsymbol{\varphi} \in \mathbf{K}} \mathcal{J}(\boldsymbol{\varphi}) \,,$$

where $\mathbf{K}$ is subset of a Sobolev-space $\mathbf{V} = (\mathbf{H}^s(\Omega))^n$ with $s \in \mathbb{N}$. We consider the subset $\mathbf{K}$, that is described by a number of linear condition, which can be combined into an continuous linear operator $\breve{G} : \mathbf{V} \longrightarrow \mathbf{Q} = (\mathbf{L}(\Omega))^m$, where $m \in \mathbb{N}$ depends on the number of condition. Using $\boldsymbol{\psi} \in \mathbf{Q}$ we define the subset $\mathbf{K}$ as

$$\mathbf{K} = \left\{ \boldsymbol{\varphi} \in \mathbf{V} \,\middle|\, \breve{G}(\boldsymbol{\varphi}) \leq \boldsymbol{\psi} \text{ a.e. on } \Omega \right\} .$$

In order to have an unique solution for this problem, we assume that the functional $\mathcal{J}$ has a gradient $\mathcal{J}'(\boldsymbol{u}) \in \mathbf{V}'$ everywhere in $\mathbf{K}$, as well as $\mathcal{J}$ is twice Gateaux-differentiable in all directions and satisfies the condition

$$\left\langle \mathcal{J}''(\boldsymbol{u})\boldsymbol{\varphi}, \boldsymbol{\varphi} \right\rangle_{\mathbf{V}} \geq \alpha \left\| \boldsymbol{\varphi} \right\|_{\mathbf{V}}^2 \qquad\qquad \forall \boldsymbol{\varphi} \in \mathbf{V}$$

for a positive constant α. Furthermore, we assume that, there is a constant $\beta \in \mathbb{R}^+$, such that

$$\sup_{\boldsymbol{\varphi} \in \mathbf{V}} \frac{\left(\boldsymbol{q}, \breve{G}(\boldsymbol{\varphi}) \right)_{0,\Omega}}{\left\| \boldsymbol{\varphi} \right\|_{\mathbf{V}}} \geq \beta \left\| \boldsymbol{q} \right\|_{\mathbf{Q}} \qquad\qquad \forall \boldsymbol{q} \in \mathbf{Q} := \mathbf{L}^2(\Omega) \,.$$

According to the lemma 6.1.9 this conditions not only secure the existence and uniqueness of the solution, but also allow us to formulate an equivalent variational problem: To find a point $(\boldsymbol{u}, \boldsymbol{p}, \boldsymbol{\lambda})^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q} \times \boldsymbol{\Lambda}$ such that

$$\mathcal{A}(\boldsymbol{\varphi}, \boldsymbol{u}) + \mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{p}) = \mathcal{L}(\boldsymbol{\varphi}) \qquad \forall \boldsymbol{\varphi} \in \mathbf{V} \tag{6.1}$$

$$\mathcal{G}(\boldsymbol{u}, \boldsymbol{q}) = (\boldsymbol{q}, \boldsymbol{\psi} - \boldsymbol{\lambda})_{0,\Omega} \qquad \forall \boldsymbol{q} \in \mathbf{Q} \tag{6.2}$$

$$(\boldsymbol{p}, \boldsymbol{\omega})_{0,\Omega} \geq 0 \qquad \forall \boldsymbol{\omega} \in \boldsymbol{\Lambda} \tag{6.3}$$

$$\boldsymbol{p} \cdot \boldsymbol{\lambda} = 0 \qquad \text{a.e. on } \Omega \,, \tag{6.4}$$

with $\mathbf{\Lambda} = \mathit{\Pi}^{+}\mathbf{Q} = \{\tilde{\boldsymbol{q}} \in \mathbf{Q} |\ \tilde{\boldsymbol{q}} \geq 0 \text{ a.e. on } \Omega\}$, the functional $\mathcal{A} : \mathbf{V} \times \mathbf{V} \to \mathbb{R}$ and the linear form $\mathcal{L} : \mathbf{V} \to \mathbb{R}$, such that

$$\left\langle \mathcal{J}'(\boldsymbol{v}), \boldsymbol{\varphi} \right\rangle_{\mathbf{V}} = \mathcal{A}(\boldsymbol{\varphi}, \boldsymbol{v}) - \mathcal{L}(\boldsymbol{\varphi}) \qquad \forall \boldsymbol{\varphi}, \boldsymbol{v} \in \mathbf{V}\,,$$

as well as the functional $\mathcal{G} : \mathbf{V} \times \mathbf{Q} \to \mathbb{R}$, such that

$$\left(\boldsymbol{q}, \breve{G}(\boldsymbol{\varphi})\right)_{0,\Omega} = \mathcal{G}(\boldsymbol{\varphi}, \boldsymbol{q}) \qquad \forall \boldsymbol{\varphi} \in \mathbf{V} \text{ and } \forall \boldsymbol{q} \in \mathbf{Q}\,.$$

In the section 6.2.2 we further restrict the proof of the convergence to the cases, in which, similar to the obstacle and Stokes problems, the functional $\mathcal{A}$ is a bilinearform. Before we examine the Newton-type algorithm, let us first make a small excursion into the different but equivalent formulations of the minimization problems.

6.1. Optimization theory for functionals

As it was shown on the examples in the previous chapters, weak formulations of the differential inequalities and minimization problems are closely related. That is why in this part we would like to discuss some aspects of the optimization theory for the functionals. As a basis we use the work of Cea and Murthy [12]. First of all we need some criteria for the general existence of minima of certain funktionals.

Theorem 6.1.1. *Suppose* $\mathbf{V}$, $\mathbf{K}$ *and* $\mathcal{J} : \mathbf{K} \to \mathbb{R}$ *satisfy the following hypothesis:*

(H1) $\mathbf{V}$ *is a reflexive Banach space,*

(H2) $\mathbf{K}$ *is weakly closed.*

(H3) $\mathbf{K}$ *is bounded and*

(H4) $\mathcal{J} : \mathbf{K} \subset \mathbf{V} \to \mathbb{R}$ *is weakly lower semi-continuous.*

Then $\mathcal{J}$ *has a global minimum in* $\mathbf{K}$.

Proof. See Cea and Murthy [12, p. 22] □

If $\mathbf{K}$ is not weakly closed, then we still can apply another theorem to secure the existence of a minimum.

Theorem 6.1.2. *If* $\mathbf{V}$, $\mathbf{K}$ *and* $\mathcal{J}$ *satisfy the Hypothesis (H1), (H2), (H4) and* $\mathcal{J}$ *satisfies*

(H3)' $\lim\limits_{\|v\|_{\mathbf{V}} \to \infty} \mathcal{J}(v) = \infty$

then $\mathcal{J}$ *admits a global minimum in* $\mathbf{K}$.

Proof. See Cea and Murthy [12, p. 22] □

In case of the Gateaux-differentiable functional we have the following necessary condition for the existence of a local minimum.

Theorem 6.1.3. *Suppose a functional $\mathcal{J} : \mathbf{K} \subset \mathbf{V} \to \mathbb{R}$ has a local minimum at a point $u \in \mathbf{K}$ and is Gateaux-differentiable at u in all directions then $\langle \mathcal{J}'(u), v - u \rangle_{\mathbf{V}'} \geq 0$ for every $v \in \mathbf{V}$ such that $v - u$ is a strongly admissible direction. Furthermore, if $\mathbf{K}$ is an open set then $\langle \mathcal{J}'(u), \varphi \rangle_{\mathbf{V}'} = 0$ for all $\varphi \in \mathbf{V}$.*

Proof. See Cea and Murthy [12, p. 24] □

The convexity is in in context of optimization an extreme powerful tool. It allows us to make statements about the existence and uniqueness of the minimum.

Theorem 6.1.4. *If $\mathbf{K}$ is a convex subset of a normed vector space and $\mathcal{J} : \mathbf{K} \subset \mathbf{V} \to \mathbb{R}$ is strictly convex then there exists a unique minimum $u \in \mathbf{K}$ for $\mathcal{J}$.*

Proof. See Cea and Murthy [12, p. 26] □

The convexity correspond to the differentiability of the functional. In case of the twice Gateaux-differentiable functional there is a special existence and uniqueness theorem.

Theorem 6.1.5. *Let $\mathcal{J} : \mathbf{K} \to \mathbb{R}$ be a functional on $\mathbf{V}$, $\mathbf{K}$ a subset of $\mathbf{V}$ satisfying the following hypothesis:*

(H1) *$\mathbf{V}$ is a reflexive Banach space,*

(H2) *$\mathcal{J}$ has a gradient $\mathcal{J}'(u) \in \mathbf{V}'$ everywhere in $\mathbf{K}$;*

(H3) *$\mathcal{J}$ is twice Gateaux-differentiable in all directions $\varphi, \psi \in \mathbf{V}$ and satisfies the condition*

$$\left\langle \mathcal{J}''(u)\varphi, \varphi \right\rangle_{\mathbf{V}'} \geq \|\varphi\|_{\mathbf{K}} \chi\left(\|\varphi\|_{\mathbf{K}}\right)$$

where $t \mapsto \chi(t)$ is a function on $\{t \in \mathbb{R} | t \geq 0\}$ such that $\chi(t) \geq 0$ and $\lim_{t \to \infty} \chi(t) = \infty$;

(H4) *$\mathbf{K}$ is a closed convex set.*

Then there exists at least one minimum $u \in \mathbf{K}$ of $\mathcal{J}$. Furthermore, if in (H3)

(H5) $$\chi(t) > 0 \quad \text{for} \quad t > 0$$

is satisfied by χ then there exists a unique minimum of $\mathcal{J}$ in $\mathbf{K}$.

Proof. See Cea and Murthy [12, p. 27] □

Again using the convexity we can formulate a minimisation problem as an equivalent differential inequality.

Theorem 6.1.6. *Suppose* $\mathbf{K}$ *is a convex subset of a Banach space* $\mathbf{V}$ *and* $\mathcal{J} : \mathbf{K} \subset \mathbf{V} \to \mathbb{R}$ *is a Gateaux-differentiable (in all directions) convex functional. Then* $u \in \mathbf{K}$ *is a minimum for* $\mathcal{J}$ *(i.e.* $\mathcal{J}(u) \leq \mathcal{J}(v)$ *for all* $v \in \mathbf{V}$ *if and only if* $u \in \mathbf{K}$ *and* $\langle \mathcal{J}'(u), v - u \rangle_{\mathbf{V}'} \geq 0$ *for all* $v \in \mathbf{K}$*).*

Proof. See Cea and Murthy [12, p. 28] □

Next, we consider the case, where the subset $\mathbf{K}$ can be described using a mapping Φ. In context of minimization this is a constraint on the solution u and allows us to use the Lagrange-multiplier method. The next theorem contains the conditions, should be imposed on the mapping Φ, in order to obtain the equivalent mixed problem.

Theorem 6.1.7. *Let* $\mathbf{V}$ *be a normed space and* $\mathbf{K}$ *be a subset of* $\mathbf{V}$ *such that we can find a cone* $\mathbf{\Lambda}$ *with vertex at 0 (in a suitable vector space) and a function* $\Phi : \mathbf{V} \times \mathbf{\Lambda} \to \mathbb{R}$ *satisfying the conditions:*

i) The mapping $\mathbf{\Lambda} \ni q \mapsto \Phi(\varphi, q) \in \mathbb{R}$ *is homogeneous of degree one i.e.*

$$\Phi(\varphi, \rho q) = \rho \Phi(\varphi, q) \qquad \forall \rho \geq 0 \, .$$

ii) A point $\varphi \in \mathbf{V}$ *belongs to* $\mathbf{K}$ *if and only if*

$$\Phi(\varphi, q) \leq 0 \qquad \forall q \in \mathbf{\Lambda} \, .$$

Then the following two problems are equivalent: Let $\mathcal{J} : \mathbf{K} \to \mathbb{R}$ *be a given functional*

Primal problem: *To find* $u \in \mathbf{K}$ *such that* $\mathcal{J}(u) = \inf_{\varphi \in \mathbf{K}} \mathcal{J}(\varphi)$.

Minimax problem: *To find a point* $(u, p)^T \in \mathbf{V} \times \mathbf{\Lambda}$ *such that*

$$\mathcal{J}(u) + \Phi(u, p) = \inf_{\varphi \in \mathbf{V}} \sup_{q \in \mathbf{\Lambda}} \left(\mathcal{J}(\varphi) + \Phi(\varphi, q) \right) \, .$$

Proof. See Cea and Murthy [12, p. 28] □

In order to obtain an appropriate mapping Φ we need to take a close look at the space $\mathbf{V}$ and the subset $\mathbf{K}$. Let $\mathbf{V}$ be the Sobolev space $\mathbf{H}_0^n(\Omega)$ with $n \in \mathbb{N}_0$ and $\breve{G} : \mathbf{V} \to \mathbf{H}^s(\Omega)$, with $s \in \mathbb{N}_0$ and $s \leq n$, be an operator that defines the subset $\mathbf{K}$ as

$$\mathbf{K} = \left\{ \varphi \in \mathbf{V} \middle| \, \breve{G}(\varphi) \leq \psi \text{ a.e. on } \Omega \right\} \, ,$$

where $\psi \in \mathbf{L}^2(\Omega)$. Next we define a space $\mathbf{Q} = \mathit{\Pi}^+ \mathbf{L}^2(\Omega)$ and a cone $\mathbf{\Lambda} = \mathit{\Pi}^+ \mathbf{Q} = \left\{ \omega \in \mathbf{L}^2(\Omega) \middle| \, \omega \geq 0 \text{ a.e. on } \Omega \right\}$. Obviously for all $\varphi \in \mathbf{K}$ and all $q \in \mathbf{\Lambda}$ the inequality $\left(\breve{G}(\varphi) - \psi, q \right)_{0,\Omega} \leq 0$ applies. On the other hand, if $\varphi \in \mathbf{V}$, but $\varphi \notin \mathbf{K}$, then there is $\tilde{\Omega} \subset \Omega$, with $|\tilde{\Omega}| \neq 0$, such that $\breve{G}(\varphi) > \psi$ a.e. on $\tilde{\Omega}$. Then there is also an indicator

function $1_{\tilde{\Omega}} \in \mathbf{\Lambda}$ with $\left(\breve{G}(\varphi) - \psi, 1_{\tilde{\Omega}}\right)_{0,\Omega} > 0$. This means we can describe the subset $\mathbf{K}$ as

$$\mathbf{K} = \left\{ \varphi \in \mathbf{V} \middle| \left(q, \breve{G}(\varphi) - \psi\right)_{0,\Omega} \leq 0 \quad \forall q \in \mathbf{\Lambda} \right\} .$$

This way we obtain an appropriate functional $\Phi(\varphi, q) = \left(q, \breve{G}(\varphi) - \psi\right)_{0,\Omega}$, that satisfy the conditions of the theorem 6.1.7. Next we want to find some criteria for the operator $\breve{G}$, that describes the subset $\mathbf{K}$, which secure the convexity of $\mathbf{K}$. This way we will be able to use existence and uniqueness theorems above.

Theorem 6.1.8. *Let $\mathbf{K}$ be a subset of $\mathbf{V}$, with spaces and the subset defined as above. Let $\breve{G} : \mathbf{V} \to \mathbf{H}^s(\Omega)$ be a sublinear operator, that satisfies the condition:*

i) There is a constant $\alpha \in \mathbb{R}^+$, such that

$$\left\| \breve{G}(\varphi) \right\|_{0,\Omega} \leq \alpha \, \|\varphi\|_{\mathbf{V}} \qquad \forall \varphi \in \mathbf{V} .$$

ii) There is a constant $\beta \in \mathbb{R}^+$, such that

$$\sup_{\varphi \in \mathbf{V}} \frac{\left(q, \breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}} \geq \beta \, \|q\|_{\mathbf{Q}} \qquad \forall q \in \mathbf{Q} .$$

In this case, the subset $\mathbf{K}$ is closed convex.

Proof. The proof can be found at the end of the section. □

In the previous chapters we motivated the further transformation of the minimax problem from the theorem 6.1.7. So, the next step is to prove those transformations in two steps:

- the set of variational conditions equivalent to the minimax problem (theorem 6.1.9),
- the replacement of the variational inequalities in the those conditions with a projection operator equation (theorem 6.1.10).

Theorem 6.1.9. *Let $\mathbf{K}$ be a subset of $\mathbf{V}$, with $\mathbf{K} = \left\{ \varphi \in \mathbf{V} \middle| \breve{G}(\varphi) \leq \psi \text{ a.e. on } \Omega \right\}$, with spaces as defined above. Let $\breve{G} : \mathbf{V} \to \mathbf{H}^s$ be a continuous linear operator, that satisfies the condition: There is a constant $\beta \in \mathbb{R}^+$, such that*

$$\sup_{\varphi \in \mathbf{V}} \frac{\left(q, \breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}} \geq \beta \, \|q\|_{\mathbf{Q}} \qquad \forall q \in \mathbf{Q} .$$

If $\mathcal{J} : \mathbf{K} \to \mathbb{R}$ is a functional on $\mathbf{V}$ that satisfies the hypothesis (H1), (H2) and (H3) of the theorem 6.1.5, then there exists at least one minimum $u \in \mathbf{K}$ of $\mathcal{J}$. Furthermore,

if the condition (H5) of the theorem 6.1.5 is satisfied by χ then there exists a unique minimum of $\mathcal{J}$ in $\mathbf{K}$. Also there is an ***equivalent variational problem:*** *To find a point $(u,p,\lambda)^T \in \mathbf{V}\times\mathbf{Q}\times\mathbf{\Lambda}$ such that*

$$\begin{aligned}
\left\langle \mathcal{J}'(u),\varphi\right\rangle_{\mathbf{V}'} + \left(p,\breve{G}(\varphi)\right)_{0,\Omega} &= 0 && \forall\varphi\in\mathbf{V}\\
\left(q,\breve{G}(u)-\psi\right)_{0,\Omega} + (q,\lambda)_{0,\Omega} &= 0 && \forall q\in\mathbf{Q}\\
(p,\omega)_{0,\Omega} &\geq 0 && \forall\omega\in\mathbf{\Lambda}\\
(q-p,\lambda)_{0,\Omega} &\geq 0 && \forall q\in \mathit{\Pi}^{+}\mathbf{Q}\,.
\end{aligned}$$

Proof. Since $\mathbf{K}$ is a closed convex set (see theorem 6.1.8), the condition (H4) of the theorem 6.1.5 is fulfilled. This means, that according to the theorem 6.1.5, there exists at leas one or even a unique minimum of $\mathcal{J}$ in $\mathbf{K}$, depending on the (H5).

As mentioned earlier, using the cone $\mathit{\Pi}^{+}\mathbf{Q}$ with vertex at 0, we can define a functional $\Phi : \mathbf{V}\times\mathit{\Pi}^{+}\mathbf{Q}\to\mathbb{R}$ with $\Phi(\varphi,q) = \left(q,\breve{G}(\varphi)-\psi\right)_{0,\Omega}$ and formulate, according to the theorem 6.1.7, the equivalent dual minimax problem

$$\mathcal{J}(u)+\Phi(u,q)\leq\mathcal{J}(u)+\Phi(u,p)\leq\mathcal{J}(\varphi)+\Phi(\varphi,p)\qquad\forall\varphi\in\mathbf{V}\text{ and }\forall q\in\mathit{\Pi}^{+}\mathbf{Q}\,.$$

The right inequality is a minimization problem on the open set with a convex Gateaux-differentiable functional, according to the (H2) and (H3), and leads to the equation

$$\left\langle \mathcal{J}'(u),\varphi\right\rangle_{\mathbf{V}'} + \left(p,\breve{G}(\varphi)\right)_{0,\Omega} = 0 \qquad\forall\varphi\in\mathbf{V}\,.$$

Since $u\in\mathbf{K}$, there is $\lambda\in\mathbf{\Lambda}$ such that

$$\left(q,\breve{G}(u)-\psi\right)_{0,\Omega} + (q,\lambda)_{0,\Omega} = 0$$

The left inequality of the minimax problem leads to

$$-\left(q-p,\breve{G}(u)-\psi\right)_{0,\Omega}\geq 0\qquad\text{or}\qquad\langle q-p,\lambda\rangle_{\mathbf{Q}}\geq 0\qquad\forall q\in\mathit{\Pi}^{+}\mathbf{Q}\,.$$

Since $p\in\mathit{\Pi}^{+}\mathbf{Q}$, according to the definition of the cone $\mathbf{\Lambda}$, we obtain

$$(p,\omega)_{0,\Omega}\geq 0\qquad\forall\omega\in\mathbf{\Lambda}$$

On the other hand, if we start with the equivalent variational problem the convexity of the functional J leads to

$$\begin{aligned}
0 &= \underbrace{\left\langle \mathcal{J}'(u),\varphi-u\right\rangle_{\mathbf{V}'}}_{=-\Phi(\varphi,p)+\Phi(u,p)} + \Phi(\varphi,p)-\Phi(u,p)\\
&\leq (\mathcal{J}(\varphi)+\Phi(\varphi,p))-(\mathcal{J}(u)+\Phi(u,p)) \qquad\forall\varphi\in\mathbf{V}\,.
\end{aligned}$$

Also

$$\begin{aligned} 0 \leq & (q-p,\lambda)_{0,\Omega} \\ = & \left(p, \breve{G}(u)-\psi\right)_{0,\Omega} - \left(q, \breve{G}(u)-\psi\right)_{0,\Omega} \\ = & \Phi(u,p) - \Phi(u,q) \qquad \forall q \in \mathit{\Pi}^{+}\mathbf{Q}\,, \end{aligned}$$

which is equivalent to

$$\mathcal{J}(u) + \Phi(u,q) \leq \mathcal{J}(u) + \Phi(u,p) \qquad \forall q \in \mathit{\Pi}^{+}\mathbf{Q}\,.$$

That means, that u and p of the equivalent variational problem also solve the minimax problem. □

The variational inequalities in the equivalent problem above makes the solution process relatively complex. We can substitute it with an equation by using the lemma of projection operator A.0.3.

Theorem 6.1.10. *Let* $\mathbf{V}$, $\mathbf{Q}$, $\mathbf{\Lambda}$ *and* $\breve{G}$ *be defined as above for the subset* $\mathbf{K}$. *Furthermore, let* $u \in \mathbf{K} \subset \mathbf{V}$ *and* $\lambda \in \mathbf{\Lambda}$ *satisfy the equation*

$$\left(q, \breve{G}(u)-\psi\right)_{0,\Omega} + (q,\lambda)_{0,\Omega} = 0 \qquad \forall q \in \mathbf{Q}\,.$$

Then the following problems are equivalent:

- *To find a point* $p \in \mathbf{Q}$, *such that*

$$\begin{aligned} & (p,\omega)_{0,\Omega} \geq 0 && \forall \omega \in \mathbf{\Lambda} \\ \textit{and} \quad & (q-p,\lambda)_{0,\Omega} \geq 0 && \forall q \in \mathit{\Pi}^{+}\mathbf{Q}\,. \end{aligned}$$

- *To find a point* $p \in \mathbf{Q}$, *such that*

$$\lambda = \mathit{\Pi}_{\mathbf{\Lambda}}\left(\psi - \breve{G}(u) - \varepsilon p\right) \qquad \forall \varepsilon > 0\,.$$

Proof. We start with first statement:

$$\begin{aligned} & (p,\omega)_{0,\Omega} \geq 0 && \forall \omega \in \mathbf{\Lambda} \\ \text{and} \; & (q-p,\lambda)_{0,\Omega} \geq 0 && \forall q \in \mathit{\Pi}^{+}\mathbf{Q}\,. \end{aligned}$$

Since a constant zero function is valid choice for $q \in \mathbf{Q}$, we can use it an add two inequalities, obtaining the inequality

$$(p,\omega-\lambda)_{0,\Omega} \geq 0 \qquad \forall \omega \in \mathbf{\Lambda}$$

The other valid choice is $q = \omega - \lambda$ a.e. on Ω for all $\omega \in \mathbf{\Lambda}$, which leads to the equation

$$\left(\omega-\lambda, \psi - \breve{G}(u) - \lambda\right)_{0,\Omega} = 0 \qquad \forall \omega \in \mathbf{\Lambda}\,.$$

By combing this equation with the inequality above, that we multiply with the positive constant ε, it leads to another inequality

$$\Big(\psi - \breve{G}(u) - \varepsilon p - \lambda, \omega - \lambda\Big)_{0,\Omega} \leq 0 \qquad\qquad \forall \omega \in \mathbf{\Lambda}\,.$$

The lemma A.0.3 allows us to interpret this inequality as a condition for a $\mathbf{L}^2(\Omega)$-projection on the subset $\mathbf{\Lambda}$:

$$\lambda = \mathit{\Pi}_{\mathbf{\Lambda}} \Big(\psi - \breve{G}(u) - \varepsilon p\Big)\,.$$

On the other hand if $\lambda = \mathit{\Pi}_{\mathbf{\Lambda}} \Big(\psi - \breve{G}(u) - \varepsilon p\Big)$, then we obtain the inequality

$$\underbrace{\Big(\psi - \breve{G}(u) - \lambda, \omega - \lambda\Big)_{0,\Omega}}_{=0} - (\varepsilon p, \omega - \lambda)_{0,\Omega} \leq 0 \qquad\qquad \forall \omega \in \mathbf{\Lambda}\,,$$

$$\text{which leads to } (p, \omega - \lambda)_{0,\Omega} \geq 0 \qquad \Leftrightarrow \qquad (p,\omega)_{0,\Omega} \geq (p,\lambda)_{0,\Omega} \qquad \forall \omega \in \mathbf{\Lambda}\,.$$

Since we allowed to use all $\omega \in \mathbf{\Lambda}$, we suggest $\omega = 2p$ a.e. on Ω. This leads to the inequality $(p,\lambda)_{0,\Omega} \geq 0$, which in turn leads to the inequality

$$(p,\omega)_{0,\Omega} \geq 0 \qquad \forall \omega \in \mathbf{\Lambda}\,.$$

This also means, that $p \in \mathit{\Pi}^+\mathbf{Q}$. Now if use zero function as ω, we obtain

$$0 \geq (p,\lambda)_{0,\Omega} \geq 0 \qquad \Leftrightarrow \qquad (p,\lambda)_{0,\Omega} = 0\,.$$

Since we already established, that $(q,\lambda)_{0,\Omega} \geq 0$ for all $q \in \mathit{\Pi}^+\mathbf{Q}$, this results in the inequality

$$\langle q - p, \lambda\rangle_{\mathbf{Q}} \geq 0 \qquad\qquad \forall q \in \mathit{\Pi}^+\mathbf{Q}\,.$$

□

Proof. **(Theorem 6.1.8)**

First of all the subset $\mathbf{K}$ is convex, because for all $v_1, \ldots, v_k \in \mathbf{K}$ and $\alpha_1, \ldots, \alpha_k \in \mathbb{R}^+$, with $k \in \mathbb{N}$ and $\sum\limits_{j=1}^{k} \alpha_j = 1$,

$$\begin{aligned}
\left(q, \psi - \breve{G}\left(\sum_{j=1}^{k} \alpha_j v_j\right)\right)_{0,\Omega} &\geq \left(q, \psi - \sum_{j=1}^{k} \alpha_j \breve{G}\left(v_j\right)\right)_{0,\Omega} \\
&\geq \inf_{1\leq i\leq k} \left(q, \psi - \breve{G}(v_i) \sum_{j=1}^{k} \alpha_j\right)_{0,\Omega} \\
&= \inf_{1\leq i\leq k} \Big(q, \psi - \breve{G}(v_i)\Big)_{0,\Omega} \geq 0 \qquad\qquad \forall q \in \mathit{\Pi}^+\mathbf{Q}\,.
\end{aligned}$$

According to the first condition

$$\sup_{\varphi\in\mathbf{V}} \frac{\left(q,\breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}} \leq \|q\|_{\mathbf{Q}} \sup_{\varphi\in\mathbf{V}} \frac{\left\|\breve{G}(\varphi)\right\|_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}} \leq \alpha \|q\|_{\mathbf{Q}} \qquad \forall q \in \mathbf{Q} .$$

Next consider a convergent sequence $(v_k)_{k\in\mathbb{N}} \subset \mathbf{K}$ with $\lim_{k\to\infty} v_k = v \in \mathbf{V}$. For all $j \in \mathbb{N}$ and $q \in \mathbf{Q}$, with $\|q\|_{\mathbf{Q}} \neq 0$, there is $\tilde{k} \in \mathbb{N}$, such that for all $k \in \mathbb{N}$ with $k \geq \tilde{k}$

$$\begin{aligned}\|v_k - v\|_{\mathbf{V}} <& \frac{\beta}{2^{j-2}\|q\|_{\mathbf{Q}}(\alpha+\beta)^2} \\ &\leq \frac{1}{2^{j-2}\|q\|_{\mathbf{Q}}^2(\alpha+\beta)^2}\left((\alpha+\beta)\|q\|_{\mathbf{Q}} - \sup_{\varphi\in\mathbf{V}} \frac{\left(q,\breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}}\right) .\end{aligned}$$

If for $q \in \mathit{\Pi}^+\mathbf{Q}$, with $\|q\|_{\mathbf{Q}} \neq 0$, there is an index $k \in \mathbb{N}$, with $k \geq \tilde{k}$, such that $\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega} \leq -\varepsilon < 0$, then there is $j \in \mathbb{N}$, such that

$$\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega} \leq -\frac{1}{2^j} .$$

Together with the estimation above, this leads to

$$\begin{aligned}&\left(q,\breve{G}(v) - \psi\right)_{0,\Omega} \leq \left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega} + \left(q,\breve{G}(v - v_k)\right)_{0,\Omega} \\ &\leq -\frac{1}{2^j} + \|v - v_k\|_{\mathbf{V}} \sup_{\varphi\in\mathbf{V}} \frac{\left(q,\breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}} \\ &\leq -\frac{1}{2^j} + \frac{1}{2^{j-2}\|q\|_{\mathbf{Q}}^2(\alpha+\beta)^2} \underbrace{\left((\alpha+\beta)\|q\|_{\mathbf{Q}} - \sup_{\varphi\in\mathbf{V}} \frac{\left(q,\breve{G}(\varphi)\right)_{0,\Omega}}{\|\varphi\|_{\mathbf{V}}}\right) \sup_{\varphi\in\mathbf{V}} \frac{\left(q,\breve{G}(\varphi)\right)_{\mathbf{Q}}}{\|\varphi\|_{0,\Omega}}}_{\leq \frac{(\alpha+\beta)^2\|q\|_{\mathbf{Q}}^2}{4}} \\ &\leq 0 .\end{aligned}$$

On the other hand, if for all $q \in \mathit{\Pi}^+\mathbf{Q}$, with $\|q\|_{\mathbf{Q}} \neq 0$, and all $k \in \mathbb{N}$, with $k \geq \tilde{k}$, $\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega} = 0$, then we have a constant zero sequence $\left(\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega}\right)_{k\in\mathbb{N}}$. This sequence converges to $\left(q,\breve{G}(v) - \psi\right)_{0,\Omega}$, because

$$\begin{aligned}\left|\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega} - \left(q,\breve{G}(v) - \psi\right)_{0,\Omega}\right| &\leq \left|\left(q,\breve{G}(v - v_k)\right)_{0,\Omega}\right| + \underbrace{2\left|\left(q,\breve{G}(v_k) - \psi\right)_{0,\Omega}\right|}_{=0} \\ &\leq \|q\|_{\mathbf{Q}} \left\|\breve{G}(v - v_k)\right\|_{0,\Omega} \\ &\leq \alpha \|q\|_{\mathbf{Q}} \|v - v_k\|_{\mathbf{V}}\end{aligned}$$

Since we have a constant zero sequence for $k > \tilde{k}$ this leads to the conclusion

$$\left(q, \breve{G}(v) - \psi\right)_{0,\Omega} = \lim_{k\to\infty} \left(q, \breve{G}(v_k) - \psi\right)_{0,\Omega} = 0\,.$$

Combining both cases we can conclude that $\left(q, \breve{G}(v) - \psi\right)_{\mathbf{Q}} \leq 0$ for all $q \in \Pi^+\mathbf{Q}$, which means that $v \in \mathbf{K}$ and that the subset $\mathbf{K}$ is closed. □

6.2. Newton-type method for the generalized problem

In order to numerically solve the variational problem, introduced above, we replace the continuous space with appropriated conform finite dimensional spaces $\mathbf{V_h}$ and $\mathbf{Q_h}$, that satisfy the conditions

$$\left\langle \mathcal{J}''(\boldsymbol{u}_h)\boldsymbol{\varphi}_h, \boldsymbol{\varphi}_h\right\rangle_{\mathbf{V_h}} = \mathcal{A}(\boldsymbol{\varphi}_h, \boldsymbol{\varphi}_h) \geq \alpha \left\|\boldsymbol{\varphi}_h\right\|^2_{\mathbf{V_h}} \qquad \forall \boldsymbol{\varphi} \in \mathbf{V_h} \tag{6.5}$$

$$\text{and} \qquad \sup_{\boldsymbol{\varphi}_h \in \mathbf{V_h}} \frac{\left(\boldsymbol{q}_h, \breve{G}(\boldsymbol{\varphi}_h)\right)_{0,\Omega}}{\left\|\boldsymbol{\varphi}_h\right\|_{\mathbf{V}}} \geq \beta \left\|\boldsymbol{q}_h\right\|_{\mathbf{Q_h}} \qquad \forall \boldsymbol{q}_h \in \mathbf{Q_h} \tag{6.6}$$

for positive constants α and β. This spaces $\mathbf{V_h}$ and $\mathbf{Q_h} \subseteq \mathbf{L}^2(\Omega)$ are usually referred to as discreet space. Furthermore it is important, that we don't use any properties of the FE-spaces. The only requirement we impose on the spaces $\mathbf{V_h}$ and $\mathbf{Q_h}$ is, that linear a problem, like the one we use to calculate the update vector, are solvable, if the unique solution exist. As in the case of the Stokes and the obstacle problems before, we consider the discreet problem: To find a point $(\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ such that

$$\mathcal{A}(\boldsymbol{\varphi}_h, \boldsymbol{u}_h) + \left(\breve{G}(\boldsymbol{\varphi}_h), \boldsymbol{p}_h\right)_{0,\Omega} = \mathcal{L}(\boldsymbol{\varphi}_h) \qquad \forall \boldsymbol{\varphi_h} \in \mathbf{V_h}, \tag{6.7}$$

$$\left(\breve{G}(\boldsymbol{u}_h), \boldsymbol{q}_h\right)_{0,\Omega} = (\boldsymbol{q}_h, \boldsymbol{\psi} - \boldsymbol{\lambda}_h)_{0,\Omega} \qquad \forall \boldsymbol{q}_h \in \mathbf{Q_h} \tag{6.8}$$

$$\text{and} \quad \boldsymbol{\lambda}_h = \Pi^+\left(\boldsymbol{\psi} - \breve{G}(\boldsymbol{u}_h) - \frac{1}{\delta}\boldsymbol{p}_h\right) \text{ a.e. on } \Omega. \tag{6.9}$$

Since the problem above is non-linear, we want an iteration instruction

$$\begin{pmatrix} \boldsymbol{u}_h^{(k+1)} \\ \boldsymbol{p}_h^{(k+1)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{u}_h^{(k)} \\ \boldsymbol{p}_h^{(k)} \end{pmatrix} - \gamma_r \begin{pmatrix} \boldsymbol{d}_h^u \\ \boldsymbol{d}_h^p \end{pmatrix}$$

$$\text{where } \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}), \left(\boldsymbol{d}_h^u, \boldsymbol{d}_h^p\right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} = \mathcal{F}_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}) \quad \forall\, (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}\,,$$

with a positive constant γ_r. In order to calculate the update vector $\left(\boldsymbol{d}_h^u, \boldsymbol{d}_h^p\right)^{\mathrm{T}}$ we need to define the functional $\mathcal{F}_{\boldsymbol{\varphi}_h\boldsymbol{q}_h} : \mathbf{V_h} \times \mathbf{Q_h} \to \mathbb{R}$ be a functional with the parameters $\boldsymbol{\varphi}_h \in \mathbf{V_h}$ and $\boldsymbol{q}_h \in \mathbf{Q_h}$, where

$$\begin{aligned}\mathcal{F}_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h) =& \mathcal{A}^\delta(\boldsymbol{\varphi}_h, \tilde{\boldsymbol{u}}_h) + \left(\breve{G}(\boldsymbol{\varphi}_h), \boldsymbol{p}_h\right)_{0,\Omega} - \mathcal{L}(\boldsymbol{\varphi}_h) + \delta\left(\breve{G}(\boldsymbol{\varphi}_h), \boldsymbol{\psi} - \boldsymbol{\lambda}_h\right)_{0,\Omega} \\ & - \left(\breve{G}(\tilde{\boldsymbol{u}}_h), \boldsymbol{q}_h\right)_{0,\Omega} + (\boldsymbol{q}_h, \boldsymbol{\psi} - \boldsymbol{\lambda}_h)_{0,\Omega}\,,\end{aligned}$$

$$\text{with} \qquad \mathcal{A}^{\delta}(\boldsymbol{\varphi}_h, \tilde{\boldsymbol{u}}_h) = \mathcal{A}(\boldsymbol{\varphi}_h, \tilde{\boldsymbol{u}}_h) - \delta \left(\breve{G}(\boldsymbol{\varphi}_h), \breve{G}(\tilde{\boldsymbol{u}}_h) \right)_{\mathbf{Q}}$$

$$\text{and} \qquad \boldsymbol{\lambda}_h = \frac{1}{2} \left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right) + \frac{1}{2} \begin{pmatrix} \left| \left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right)_1 \right| \\ \vdots \\ \left| \left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right)_m \right| \end{pmatrix} .$$

As well, we define a more stabilized and regularised Gateaux-differentiable functional $\tilde{\mathcal{F}}_{\boldsymbol{\varphi q}} : \mathbf{V} \times \mathbf{Q} \to \mathbb{R}$ with the parameters $\boldsymbol{\varphi} \in \mathbf{V}$ and $\boldsymbol{q} \in \mathbf{Q}$, where

$$\begin{aligned} \tilde{\mathcal{F}}_{\boldsymbol{\varphi}_h \boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h) =& \mathcal{A}^{\delta}(\boldsymbol{\varphi}_h, \tilde{\boldsymbol{u}}_h) + \left(\breve{G}(\boldsymbol{\varphi}_h), \boldsymbol{p}_h \right)_{0,\Omega} - \mathcal{L}(\boldsymbol{\varphi}_h) + \delta \left(\breve{G}(\boldsymbol{\varphi}_h), \boldsymbol{\psi} - \boldsymbol{\lambda}_h \right)_{0,\Omega} \\ & - \left(\breve{G}(\tilde{\boldsymbol{u}}_h), \boldsymbol{q}_h \right)_{0,\Omega} + (\boldsymbol{q}_h, \boldsymbol{\psi} - \boldsymbol{\lambda}_h)_{0,\Omega} + c_{\mathrm{Q}} \sum_{T \in \mathbb{T}_h} |T| \, (\tilde{\boldsymbol{p}}_h, \boldsymbol{q}_h)_{0,T} \ , \end{aligned}$$

$$\text{with } \boldsymbol{\lambda}^{\xi} = \frac{1}{2} \left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right) + \frac{1}{2} \begin{pmatrix} \sqrt{\left(\left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right)_1 \right)^2 + \xi} \\ \vdots \\ \sqrt{\left(\left(\boldsymbol{\psi} - \breve{G}(\tilde{\boldsymbol{u}}_h) - \frac{1}{\delta} \breve{\Psi}_s(\tilde{\boldsymbol{p}}_h) \right)_m \right)^2 + \xi} \end{pmatrix} .$$

The hypothesis is, that on the discrete spaces $\mathbf{V}_h$ and $\mathbf{Q}_h$ the iteration sequence converges towards the best possible approximation $(\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}$. To proof it, we define two operators $\breve{U} : \mathbf{V_h} \times \mathbf{Q_h} \to \mathbf{V_h} \times \mathbf{Q_h}$ and $\breve{T} : \mathbf{V_h} \times \mathbf{Q_h} \to \mathbf{V_h} \times \mathbf{Q_h}$, where

$$\begin{aligned} & \breve{T}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} = (\tilde{\boldsymbol{u}}_{\boldsymbol{h}}, \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} - \gamma_r \breve{U}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} \\ \text{and} \quad & \breve{U}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} = (\breve{U}^u \tilde{\boldsymbol{u}}_h, \breve{U}^p \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} = \left(\boldsymbol{d}_h^u, \boldsymbol{d}_h^p \right)^{\mathrm{T}} \ , \\ \text{with} \quad & \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h \boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h), \left(\boldsymbol{d}_h^u, \boldsymbol{d}_h^p \right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h} \times \mathbf{Q_h}} = \mathcal{F}_{\boldsymbol{\varphi}_h \boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h) \quad \forall \, (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h} \, . \end{aligned}$$

We have to make sure, that the update vector $\left(\boldsymbol{d}_h^u, \boldsymbol{d}_h^p \right)^{\mathrm{T}}$ can be calculated, as well as that the operator $\breve{T}$ defines a fixed point iteration. Last but not least we have to verify, that the limit of this sequence is $(\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} \in \mathbf{V}_h \times \mathbf{Q}_h$, the solution of the problem defined by equations 6.7 to 6.9. We use the theorem A.0.7 from Großmann and Roos [11, p. 109 ff] and the scheme of the poof looks as follows:

In the **subsection 6.2.1**, we consider $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h \tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) \in (\mathbf{V} \times \mathbf{Q})^*$ with $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h \tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) = \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h \boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h)$ and show, that for all $\tilde{\boldsymbol{f}}^* \in \mathbf{V_h}^* \times \mathbf{Q}^*$ there exist exactly one solution $(\boldsymbol{\varphi}_f, \boldsymbol{q}_f)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$, such that

$$\mathcal{Z}^*_{\tilde{\boldsymbol{u}} \tilde{\boldsymbol{p}}}(\boldsymbol{\varphi}_f, \boldsymbol{q}_f) = \tilde{\boldsymbol{f}}^* \qquad \forall (\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h)^{\mathrm{T}} \in \mathbf{V} \times \mathbf{Q} \, .$$

To do so, we proof, that $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h \tilde{\boldsymbol{p}}_h}$ has following properties:

a) $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is strong monotone, meaning that meaning that there is a constant $\alpha_{\mathcal{Z}} \in \mathbb{R}^+$, such that for all $(\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$

$$\begin{aligned}\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) - \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h), (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} \\ \geq \alpha_{\mathcal{Z}} \left\| (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}.\end{aligned}$$

b) $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is Lipschitz continuous, meaning that there is a constant $\beta_{\mathcal{Z}} \in \mathbb{R}^+$, such that for all $(\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$

$$\left\| \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) - \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h) \right\|_{(\mathbf{V_h}\times\mathbf{Q_h})^*} \leq \beta_{\mathcal{Z}} \left\| (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}}.$$

The properties above allow us to show in the **subsection 6.2.2** that, in each step the unique update vector $\left(\boldsymbol{d}^u_h, \boldsymbol{d}^p_h\right)^{\mathrm{T}}$ exist. For further steps we will need the properties:

a) There is an lower estimation, with a positive constant $\alpha_{\breve{U}}$, for the scalar product of the update operator in each iteration step:

$$\begin{aligned}&\left(\left(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h\right)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}, \breve{U}\left(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h\right)^{\mathrm{T}} \right)_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &\geq \alpha_{\breve{U}} \left\| \left(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h\right)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} \qquad \forall k \in \mathbb{N}.\end{aligned}$$

b) There is an upper estimation, with a positive constant $\beta_{\breve{U}}$, for the norm of the update operator in each iteration step:

$$\left\| \breve{U}\left(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \leq \beta_{\breve{U}} \left\| \left(\boldsymbol{u}^{(k)}_h - \boldsymbol{u}_h, \boldsymbol{p}^{(k)}_h - \boldsymbol{p}_h\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \qquad \forall k \in \mathbb{N}.$$

Finally we can estimate how much near does the iteration process brigs us to the solution $(\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}$ in each step. For this we derive the inequality

$$\begin{aligned}&\left\| (\boldsymbol{u}^{(k+1)}_h - \boldsymbol{u}_h, \boldsymbol{p}^{(k+1)}_h - \boldsymbol{p}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &= \left\| (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} - (\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} + \gamma_r \breve{U}(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &= \left\| (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} - (\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &\quad - \gamma_r \left((\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}, \breve{U}(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} \right)_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &\quad + \gamma_r^2 \left\| \breve{U}(\boldsymbol{u}^{(k)}_h, \boldsymbol{p}^{(k)}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}.\end{aligned}$$

If the update operator $\breve{U}$ has the properties described above, this means that

$$\begin{aligned}&\left\|(\boldsymbol{u}_h^{(k+1)}-\boldsymbol{u}_h,\boldsymbol{p}_h^{(k+1)}-\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}\\&\quad\leq\left(1-2\gamma_r\alpha_{\breve{U}}+\gamma_r^2\beta_{\breve{U}}^2\right)\left\|(\boldsymbol{u}_h^{(k)}-\boldsymbol{u}_h,\boldsymbol{p}_h^{(k)}-\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}.\end{aligned}$$

As a result we can conclude, that for $0<\gamma_r<\frac{2\alpha_{\breve{U}}}{\beta^2_{\breve{U}}}$ the iteration algorithm has a contraction property. Furthermore the inequality above proves, that for the appropriate positive constants γ_r, δ and c_{Q} we receive a convergent sequence, that has the solution of the problem, defined by equations 6.7 to 6.9, as it's limit.

6.2.1. Properties of the Gateaux-derivative $\mathcal{F}'_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)$

Theorem 6.2.1. *Let the derivative $\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}$ and the functional $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ be defined as above, then $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is* ***strong monotone****.*

Proof. Using the definition above we get the equation

$$\begin{aligned}&\left\langle\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)-\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_{\boldsymbol{h}},\tilde{\boldsymbol{q}}_{\boldsymbol{h}}),(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)^{\mathrm{T}}-(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}\right\rangle_{\mathbf{V}_h\times\mathbf{Q}_h}\\&=\left\langle\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h),(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h,\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}\\&-\left\langle\tilde{\mathcal{F}}'_{\tilde{\boldsymbol{\varphi}}_h\tilde{\boldsymbol{q}}_h}(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h),(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h,\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}\\&=\mathcal{A}^{\delta}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h,\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)+\left(\breve{G}(\boldsymbol{\varphi}_h)-\breve{G}(\tilde{\boldsymbol{\varphi}}_h),\boldsymbol{q}_h-\tilde{\boldsymbol{q}}\right)_{0,\Omega}\\&-\left(\breve{G}(\boldsymbol{\varphi}_h)-\breve{G}(\tilde{\boldsymbol{\varphi}}_h),\boldsymbol{q}_h-\tilde{\boldsymbol{q}}\right)_{0,\Omega}+c_{\mathrm{Q}}\sum_{T\in\mathbb{T}_h}|T|\,\|\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h\|_{0,T}\\&+\frac{1}{2}\sum_{j=1}^{m}\left(\delta\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)_j+(q_j-\tilde{q}_j),\left(1+\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)\right)\left(\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)_j+\frac{1}{\delta}(q_j-\tilde{q}_j)\right)\right)_{0,\Omega},\end{aligned}$$

where $\breve{V}_j:\mathbf{V_h}\times\mathbf{Q_h}\to\mathbf{L}^{\infty}(\Omega)$ with

$$\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)=\frac{\left(\boldsymbol{\psi}-\breve{G}(\tilde{\boldsymbol{u}}_h)-\frac{1}{\delta}\tilde{\boldsymbol{p}}_h\right)_j}{\sqrt{\left(\left(\boldsymbol{\psi}-\breve{G}(\tilde{\boldsymbol{u}}_h)-\frac{1}{\delta}\tilde{\boldsymbol{p}}_h\right)_j\right)^2+\xi}}$$

and therefor $0<\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)<1$ a.e. on Ω. This helps us to estimate using Hölder-inequality, that

$$\begin{aligned}&\sum_{j=1}^{m}\left(\delta\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)_j+(q_j-\tilde{q}_j),\left(1+\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)\right)\left(\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)_j+\frac{1}{\delta}(q_j-\tilde{q}_j)\right)\right)_{0,\Omega}\\&\geq\frac{1}{\delta}\left\|\delta\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)+(\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h)\right\|^2_{0,\Omega}\max_{1\leq j\leq m}\left\|1-\left|\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)\right|\right\|_{\mathbf{L}^{\infty}(\Omega)}\end{aligned}$$

In the case of mapping $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h} : \mathbf{V_h} \times \mathbf{Q_h} \to (\mathbf{V_h} \times \mathbf{Q_h})^*$ we treat $\tilde{\boldsymbol{u}}_h$ and $\tilde{\boldsymbol{p}}_h$ just as parameter and there for can define a constant $c_H = \max\limits_{1\le j\le m} \left\| 1 - \left| \breve{V}_j(\tilde{\boldsymbol{u}}_h, \tilde{\boldsymbol{p}}_h) \right| \right\|_{\mathbf{L}^\infty(\Omega)}$ with $0 < c_H < 1$. According to the ellipticity condition 6.5, we can estimate, that

$$
\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) - \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h), (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\rangle_{\mathbf{V}_h \times \mathbf{Q}_h} \\
\geq & \alpha \left\| \boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h \right\|^2_{\mathbf{V}_h} - \delta \left\| \breve{G}(\boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h) \right\|^2_{0,\Omega} + \frac{c_H}{2\delta} \left\| \delta \breve{G}(\boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h) - (\boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h) \right\|^2_{0,\Omega} \\
& + c_{\mathrm{Q}} |T_{min}| \left\| \boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h \right\|^2_{0,\Omega} \\
\geq & \alpha \left\| \boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h \right\|^2_{\mathbf{V}_h} + \delta \left(\frac{c_H}{2} - 1 \right) \left\| \breve{G}(\boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h) \right\|^2_{0,\Omega} + \frac{c_H}{2\delta} \left\| \boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h \right\|^2_{0,\Omega} \\
& - c_H \left\| \breve{G}(\boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h) \right\|_{0,\Omega} \left\| \boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h \right\|_{0,\Omega} + c_{\mathrm{Q}} |T_{min}| \left\| \boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h \right\|^2_{0,\Omega} .
\end{aligned}
$$

Using the continuity of the operator $\breve{G}$, the inequality

$$
\left\| \breve{G}(\boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h) \right\|_{0,\Omega} \leq c_{con} \left\| \boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h \right\|_{\mathbf{V}} , \tag{6.10}
$$

with a constant $c_{con} > 0$, can be introduced. Lastly, we can apply the Young's inequality with a constant $c_Y > 0$:

$$
\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) - \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h), (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\rangle_{\mathbf{V_h} \times \mathbf{Q_h}} \\
&\geq \left(\alpha + \delta c^2_{con} \left(\frac{c_H}{2} (1 - c_Y) - 1 \right) \right) \left\| \boldsymbol{\varphi}_h - \tilde{\boldsymbol{\varphi}}_h \right\|^2_{\mathbf{V_h}} \\
&+ \left(\frac{c_H}{2\delta} \left(1 - \frac{1}{2c_Y} \right) + c_{\mathrm{Q}} |T_{min}| \right) \left\| \boldsymbol{q}_h - \tilde{\boldsymbol{q}}_h \right\|^2_{0,\Omega} .
\end{aligned}
$$

In order for the constants $\left(\alpha + \delta c^2_{con} \left(\frac{c_H}{2} (1 - c_Y) - 1 \right) \right)$ and $\frac{c_H}{2\delta} \left(1 - \frac{1}{2c_Y} \right)$ to be positive, the constant c_Y must satisfy the condition

$$
\frac{1}{2} < c_Y < 2 \frac{\alpha - \delta c^2_{con}}{\delta c^2_{con} c_H} + 1 .
$$

This is possible, if $\delta < \frac{\alpha}{c^2_{con}} \leq \frac{4\alpha}{c^2_{con}(4 - c_H)}$. Since we now that an appropriate contant c_Y exist, by defining

$$
\alpha_{\mathcal{Z}} = \min \left\{ \left(\alpha + \delta c^2_{con} \left(\frac{c_H}{2} (1 - c_Y) - 1 \right) \right), \frac{c_H}{2\delta} \left(1 - \frac{1}{2c_Y} \right) + c_{\mathrm{Q}} |T_{min}| \right\}
$$

we obtain the estimation

$$
\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h, \boldsymbol{q}_h) - \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h), (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\rangle_{\mathbf{V_h} \times \mathbf{Q_h}} \\
&\geq \alpha_{\mathcal{Z}} \left\| (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}} - (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \right\|^2_{\mathbf{V_h} \times \mathbf{Q_h}} \qquad \forall (\boldsymbol{\varphi}_h, \boldsymbol{q}_h)^{\mathrm{T}}, (\tilde{\boldsymbol{\varphi}}_h, \tilde{\boldsymbol{q}}_h)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h} . \square
\end{aligned}
$$

Theorem 6.2.2. *Let the derivative $\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_h\boldsymbol{q}_h}$ and the functional $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ be defined as above, then $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is **Lipschitz continuous**.*

Proof. Using the definition of $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)$ and other notation, used previously, we get for all $(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)^{\mathrm{T}}$, $(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}$ and $\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}$ the equation

$$\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)-\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h),\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}\right\rangle_{\mathbf{V}_h\times\mathbf{Q}_h}\\
=&\mathcal{A}^\delta(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h,\boldsymbol{d}_h^u)+\left(\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h),\boldsymbol{d}_h^p\right)_{0,\Omega}\\
&-\left(\breve{G}(\boldsymbol{d}_h^u),\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h\right)_{0,\Omega}+c_{\mathrm{Q}}\sum_{T\in\mathbb{T}_h}|T|\left(\boldsymbol{d}_h^p,\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h\right)_{0,T}\\
&+\frac{1}{2}\sum_{j=1}^{m}\left(\delta\breve{G}(\varphi_j-\tilde{\varphi}_j)+(q_j-\tilde{q}_j),\left(1+\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)\right)\left(\breve{G}(\boldsymbol{d}_h^u)_j+\frac{1}{\delta}(d^p_{h,j})\right)\right)_{0,\Omega}.
\end{aligned}$$

By applying the Cauchy-Schwarz inequality we receive the first estimation:

$$\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)-\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h),\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}\right\rangle_{\mathbf{V}_h\times\mathbf{Q}_h}\\
&\le\mathcal{A}^\delta(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h,\boldsymbol{d}_h^u)+\left\|\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)\right\|_{0,\Omega}\left\|\boldsymbol{d}_h^p\right\|_{0,\Omega}\\
&+\left\|\breve{G}(\boldsymbol{d}_h^u)\right\|_{0,\Omega}\|\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h\|_{0,\Omega}+c_{\mathrm{Q}}|T_{max}|\left\|\boldsymbol{d}_h^p\right\|_{0,\Omega}\|\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h\|_{0,\Omega}\\
&+\frac{1}{2\delta}\left(1+\sup_{1\le j\le m}\left\|\breve{V}_j(\tilde{\boldsymbol{u}}_h,\tilde{\boldsymbol{p}}_h)\right\|_{\mathbf{L}^\infty(\Omega)}\right)\left\|\delta\breve{G}(\boldsymbol{\varphi}_h-\tilde{\boldsymbol{\varphi}}_h)+(\boldsymbol{q}_h-\tilde{\boldsymbol{q}}_h)\right\|_{0,\Omega}\\
&\quad\times\left\|\left(\delta\breve{G}(\boldsymbol{d}_h^u)+\boldsymbol{d}_h^p\right)\right\|_{0,\Omega}.
\end{aligned}$$

Since the bilinear form $\mathcal{A}^\delta$ is continuous, last step is to apply the Young's inequality and choose the appropriate constants, which results in

$$\begin{aligned}
&\left\langle \mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)-\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h),\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}\right\rangle_{\mathbf{V}_h\times\mathbf{Q}_h}\\
&\le\beta_{\mathcal{Z}}\left\|(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)^{\mathrm{T}}-(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\left\|\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}},
\end{aligned}$$

with a positive constant $\beta_{\mathcal{Z}}$, and therefor

$$\left\|\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)-\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h)\right\|_{(\mathbf{V_h}\times\mathbf{Q_h})^*}\le\beta_{\mathcal{Z}}\left\|(\boldsymbol{\varphi}_h,\boldsymbol{q}_h)^{\mathrm{T}}-(\tilde{\boldsymbol{\varphi}}_h,\tilde{\boldsymbol{q}}_h)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\square$$

6.2.2. Properties of the update operator $\breve{U}$

The properties, derived in the previous subsection allow us two conclusions. First according to the abstract existence theorem A.0.4 there is a unique update vector $\left(\boldsymbol{d}_h^u,\boldsymbol{d}_h^p\right)^{\mathrm{T}}$ for each $\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}$, and therefor the update operator $\breve{U}$ is well defined. On the other hand, for all $\tilde{\boldsymbol{f}}^*\in(\mathbf{V_h}\times\mathbf{Q_h})^*$ there is $(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\in\mathbf{V_h}\times\mathbf{Q_h}$ that satisfies the equation

$$\mathcal{Z}^*_{\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}}(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)=\tilde{\boldsymbol{f}}^*.$$

Theorem 6.2.3. *Let the update operator $\breve{U}$ be defined as above, then there is an lower estimation, with a positive constant $\alpha_{\breve{U}}$, for the scalar product of the update operator in each iteration step:*

$$\begin{aligned}&\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^T-\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^T,\breve{U}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^T\right)_{\mathbf{V_h}\times\mathbf{Q_h}}\\&\quad\geq\alpha_{\breve{U}}\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^T-\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^T\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}\qquad\forall k\in\mathbb{N}.\end{aligned}$$

Proof. Let $\tilde{\boldsymbol{f}}^*\in(\mathbf{V_h}\times\mathbf{Q_h})^*$ be defined as

$$\left\langle\tilde{\boldsymbol{f}}^*,\,.\,\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}=\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^{\mathrm{T}},\,.\,\right)_{\mathbf{V_h}\times\mathbf{Q_h}},$$

where $\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\in\mathbf{V_h}\times\mathbf{Q_h}$ is the k-th step of the iteration and $\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^{\mathrm{T}}\in(\mathbf{V_h}\times\mathbf{Q_h})$ is the solution of the problem defined by equations 6.7 to 6.9. Furthermore let $(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\in\mathbf{V_h}\times\mathbf{Q_h}$ be the solution of the equation

$$\mathcal{Z}^*_{\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}}\left(\boldsymbol{\varphi}_f,\boldsymbol{q}_f\right)=\tilde{\boldsymbol{f}}^*$$

$$\text{and therefor}\quad\left\langle\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f\boldsymbol{q}_f}(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}),\,.\,\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}=\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^{\mathrm{T}},\,.\,\right)_{\mathbf{V_h}\times\mathbf{Q_h}}.$$

According to the definitions of the functional $\mathcal{F}_{\boldsymbol{\varphi_1},\boldsymbol{q_1}}$, the dual pair $\left\langle\tilde{\mathcal{F}}'_{\boldsymbol{\varphi_1},\boldsymbol{q_1}},\,.\,\right\rangle_{\mathbf{V}\times\mathbf{Q}}$ and the operator $\breve{U}$, the equation above is equivalent to

$$\begin{aligned}&\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^{\mathrm{T}},\breve{U}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\right)_{\mathbf{V_h}\times\mathbf{Q_h}}\\&=\left\langle\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f\boldsymbol{q}_f}(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}),\breve{U}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}=\mathcal{F}_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right).\end{aligned}$$

Since $\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)^{\mathrm{T}}\in(\mathbf{V_h}\times\mathbf{Q_h})$ is the solution of the problem defined by equations 6.7 to 6.9, it means that $\mathcal{F}_{\boldsymbol{\varphi}_h,\boldsymbol{q}_h}\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)=0$ for all $\left(\boldsymbol{\varphi}_h,\boldsymbol{q}_h\right)^{\mathrm{T}}\in(\mathbf{V_h}\times\mathbf{Q_h})$. By applying the Taylor's theorem we get

$$\begin{aligned}\underbrace{\tilde{\mathcal{F}}_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)}_{=\mathcal{F}_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)+c_{\mathrm{Q}}\sum\limits_{T\in\mathbb{T}_h}|T|\left(\boldsymbol{p}_h^{(k)},\boldsymbol{q}_f\right)_{0,T}}&=\underbrace{\tilde{\mathcal{F}}_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)}_{=\mathcal{F}_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}\left(\boldsymbol{u}_h,\boldsymbol{p}_h\right)+c_{\mathrm{Q}}\sum\limits_{T\in\mathbb{T}_h}|T|\left(\boldsymbol{p}_h,\boldsymbol{q}_f\right)_{0,T}}\\&\quad+\left\langle\tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f\boldsymbol{q}_f}(\tilde{\boldsymbol{u}}_\theta,\tilde{\boldsymbol{q}}_\theta),\left(\boldsymbol{u}_h^{(k)}-\boldsymbol{u}_h,\boldsymbol{p}_h^{(k)}-\boldsymbol{p}_h\right)^{\mathrm{T}}\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}},\end{aligned}$$

where $(\tilde{\boldsymbol{u}}_\theta,\tilde{\boldsymbol{q}}_\theta)^{\mathrm{T}}\in(\mathbf{V_h}\times\mathbf{Q_h})$ is a vector, such that for all components of the vector there is a real number $0<\theta_j<1$ with $\tilde{u}_{\theta,j}=\theta_j u_{h,j}+(1-\theta_j)u_{h,j}^{(k)}$ or $\tilde{p}_{\theta,j}=\theta_j p_{h,j}+(1-\theta_j)p_{h,j}^{(k)}$.

Again, according to the definitions of the dual pair $\left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f,\boldsymbol{q}_f}, \, . \right\rangle_{\mathbf{V}\times\mathbf{Q}}$, the operator $\breve{V}$ and the previous equations

$$
\begin{aligned}
&\left(\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}, \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right)_{\mathbf{V_h}\times\mathbf{Q_h}} \\
&= \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f\boldsymbol{q}_f}(\tilde{\boldsymbol{u}}_\theta, \tilde{\boldsymbol{q}}_\theta), \left(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h, \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} - c_{\mathrm{Q}} \sum_{T\in\mathbb{T}_h} |T| \left(\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h, \boldsymbol{q}_f\right)_{0,T} \\
&= \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_f\boldsymbol{q}_f}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}), \left(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h, \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} - c_{\mathrm{Q}} \sum_{T\in\mathbb{T}_h} |T| \left(\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h, \boldsymbol{q}_f\right)_{0,T} \\
&+ \frac{1}{2}\sum_{j=1}^{m} \Big(\delta\breve{G}(\boldsymbol{\varphi}_f)_j + q_{f,j}, \left(\breve{V}_j(\tilde{\boldsymbol{u}}_\theta, \tilde{\boldsymbol{p}}_\theta) - \breve{V}_j(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)})\right) \\
&\qquad\qquad \times \left(\breve{G}(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h)_j + \frac{1}{\delta}(p_{h,j}^{(k)} - p_{h,j}) \right) \Big)_{0,\Omega} \\
&= \left(\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}}, \left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} - (\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} \right)_{\mathbf{V_h}\times\mathbf{Q_h}} \\
&\quad - c_{\mathrm{Q}} \sum_{T\in\mathbb{T}_h} |T| \left(\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h, \boldsymbol{q}_f\right)_{0,T} \\
&\quad + \frac{1}{2}\sum_{j=1}^{m} \Big(\delta\breve{G}(\boldsymbol{\varphi}_f)_j + q_{f,j}, \left(\breve{V}_j(\tilde{\boldsymbol{u}}_\theta, \tilde{\boldsymbol{p}}_\theta) - \breve{V}_j(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)})\right) \\
&\qquad\qquad \times \left(\breve{G}(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h)_j + \frac{1}{\delta}(p_{h,j}^{(k)} - p_{h,j}) \right) \Big)_{0,\Omega}.
\end{aligned}
$$

Next we apply Hölder's and Cauchy-Schwarz's theorems, as well as the property of the operator $\breve{V}$, which for all $(\boldsymbol{\varphi}_f, \boldsymbol{q}_f)^{\mathrm{T}} \in (\mathbf{V}\times\mathbf{Q})$ holds $0 \leq \breve{V}_j(\boldsymbol{\varphi}_f, \boldsymbol{q}_f) \leq 1$ a.e. on Ω. Using the continuity of the operator $\breve{G}$ (see the inequality (6.10)) and Young's inequality with

appropriate constants we can further refine our estimation to the form

$$
\begin{aligned}
&\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}},\breve{U}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\right)_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\geq\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}-c_{\mathrm{Q}}|T_{\max}|\left\|\boldsymbol{p}_h^{(k)}-\boldsymbol{p}_h\right\|_{0,\Omega}\|\boldsymbol{q}_f\|_{0,\Omega}\\
&\quad-\frac{1}{2}\underbrace{\left\|\max_{1\leq j\leq m}\left|\breve{V}_j(\tilde{\boldsymbol{u}}_\theta,\tilde{\boldsymbol{p}}_\theta)-\breve{V}_j(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)})\right|\right\|_{\mathbf{L}^\infty(\Omega)}}_{\leq 1}\\
&\qquad\qquad\times\left\|\left|\delta\breve{G}(\boldsymbol{\varphi}_f)+\boldsymbol{q}_f\right|\left|\breve{G}(\boldsymbol{u}_h^{(k)}-\boldsymbol{u}_h)+\frac{1}{\delta}(\boldsymbol{p}_h^{(k)}-\boldsymbol{p}_h)\right|\right\|_{\mathbf{L}^1(\Omega)}\\
&\geq\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\quad-c_{\mathrm{Q}}|T_{\max}|\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\left\|(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\quad-\frac{1}{2}\left\|\delta\breve{G}(\boldsymbol{\varphi}_f)+\boldsymbol{q}_f\right\|_{0,\Omega}\left\|\breve{G}(\boldsymbol{u}_h^{(k)}-\boldsymbol{u}_h)+\frac{1}{\delta}(\boldsymbol{p}_h^{(k)}-\boldsymbol{p}_h)\right\|_{0,\Omega}\\
&\geq\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\quad-\left(c_{\mathrm{Q}}|T_{\max}|+\frac{1+\delta^2c_{con}^2}{2\delta}\right)\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\left\|(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}},
\end{aligned}
$$

where $|T_{\max}|=\max\limits_{T\in\mathbb{T}_h}|T|$. In the previous subsection we demonstrated, that $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is strong monotone, which particularly means that

$$
\begin{aligned}
\alpha_{\mathcal{Z}}\left\|(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}&\leq\left\langle\mathcal{Z}^*_{\boldsymbol{u}_h^{(k)}\boldsymbol{p}_h^{(k)}}(\boldsymbol{\varphi}_f,\boldsymbol{q}_f),(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&=\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}},(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right)_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\leq\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}\left\|(\boldsymbol{\varphi}_f,\boldsymbol{q}_f)^{\mathrm{T}}\right\|_{\mathbf{V_h}\times\mathbf{Q_h}}.
\end{aligned}
$$

In combination with the previous results we finally get

$$
\begin{aligned}
&\left(\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}},\breve{U}\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\right)_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\quad\geq\left(1-\frac{1}{\alpha_{\mathcal{Z}}}\left(c_{\mathrm{Q}}|T_{\max}|+\frac{1+\delta^2c_{con}^2}{2\delta}\right)\right)\left\|\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}-(\boldsymbol{u}_h,\boldsymbol{p}_h)^{\mathrm{T}}\right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}}\\
&\qquad\qquad\forall\left(\boldsymbol{u}_h^{(k)},\boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}\in\mathbf{V_h}\times\mathbf{Q_h}.
\end{aligned}
$$

By choosing f.e. $c_{\mathrm{Q}}|T_{\max}|<\alpha_{\mathcal{Z}}-\frac{1+\delta^2c_{con}^2}{2\delta}$ and $\frac{\alpha_{\mathcal{Z}}-\sqrt{\alpha_{\mathcal{Z}}^2-c_{con}^2}}{c_{con}^2}<\delta<\frac{\alpha_{\mathcal{Z}}+\sqrt{\alpha_{\mathcal{Z}}^2-c_{con}^2}}{c_{con}^2}$ we can define a constant $\alpha_{\breve{U}}:=\left(1-\frac{1}{\alpha_{\mathcal{Z}}}\left(c_{\mathrm{Q}}|T_{\max}|+\frac{1+\delta^2c_{con}^2}{2\delta}\right)\right)>0$ which means, that, for an

appropriate choice of the stabilization constants δ and c_{Q}, the scalar product with the operator $\breve{U}$ have a lower estimate with a positive constant. □

Theorem 6.2.4. *Let the update operator $\breve{U}$ be defined as above, then there is an upper estimation, with a positive constant $\beta_{\breve{U}}$, for the norm of the update operator in each iteration step:*

$$\left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^T \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \leq \beta_{\breve{U}} \left\| \left(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h, \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)^T \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \qquad \forall k \in \mathbb{N}.$$

Proof. Let $\tilde{\boldsymbol{q}}^* \in (\mathbf{V_h} \times \mathbf{Q_h})^*$ be defined as

$$\langle \tilde{\boldsymbol{g}}^*, \, . \, \rangle_{\mathbf{V_h}\times\mathbf{Q_h}} = \left(\breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}, \, . \, \right)_{\mathbf{V_h}\times\mathbf{Q_h}},$$

where $\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ is the k-th step of the iteration. Furthermore let the vector $(\boldsymbol{\varphi}_g, \boldsymbol{q}_g)^{\mathrm{T}} \in \mathbf{V_h} \times \mathbf{Q_h}$ be the solution of the equation

$$\mathcal{Z}^*_{\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}}(\boldsymbol{\varphi}_g, \boldsymbol{q}_g) = \tilde{\boldsymbol{g}}^*$$

$$\text{and therefor} \quad \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}), \, . \, \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} = \left(\breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}, \, . \, \right)_{\mathbf{V_h}\times\mathbf{Q_h}}.$$

This results in the equation

$$\begin{aligned} \left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} &= \left\langle \tilde{\mathcal{F}}'_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}), \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} \\ &= \mathcal{F}_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}). \end{aligned}$$

Let $(\boldsymbol{u}_h, \boldsymbol{p}_h)^{\mathrm{T}} \in (\mathbf{V_h} \times \mathbf{Q_h})$ is the solution of the problem defined by equations 6.7 to 6.9, then using the definition of the functional $\mathcal{F}_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}$ and the Cauchy-Schwarz inequality we can estimate that

$$\begin{aligned} \left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\|^2_{\mathbf{V_h}\times\mathbf{Q_h}} &= \mathcal{F}_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}) - \underbrace{\mathcal{F}_{\boldsymbol{\varphi}_g \boldsymbol{q}_g}(\boldsymbol{u}_h, \boldsymbol{p}_h)}_{=0} \\ &\leq \mathcal{A}(\boldsymbol{\varphi}_g, \boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h) + \left(\breve{G}(\boldsymbol{\varphi}_g), \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)_{0,\Omega} - \frac{1}{2}\left(\delta\breve{G}(\boldsymbol{\varphi}_g) + \boldsymbol{q}_g, \breve{G}(\boldsymbol{u}_h^{(k)}) - \breve{G}(\boldsymbol{u}_h)\right)_{0,\Omega} \\ &\quad + \frac{1}{2}\left(\delta\breve{G}(\boldsymbol{\varphi}_g) + \boldsymbol{q}_g, \frac{1}{\delta}\left(\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)\right)_{0,\Omega} \\ &\quad - \frac{1}{2}\sum_{j=1}^{m}\left(\delta\breve{G}(\boldsymbol{\varphi}_g)_j + \boldsymbol{q}_{g,j}, \left|\psi_j - \breve{G}(\boldsymbol{u}_h^{(k)})_j - \frac{1}{\delta}p_j^{(k)}\right| - \left|\psi_j - \breve{G}(\boldsymbol{u}_h)_j - \frac{1}{\delta}p_j\right|\right)_{0,\Omega} \end{aligned}$$

$$\begin{aligned} &\leq \mathcal{A}(\boldsymbol{\varphi}_g, \boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h) + \left\|\breve{G}(\boldsymbol{\varphi}_g)\right\|_{0,\Omega}\left\|\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right\|_{0,\Omega} \\ &\quad + \left\|\delta\breve{G}(\boldsymbol{\varphi}_g) + \boldsymbol{q}_g\right\|_{0,\Omega}\left\|\breve{G}(\boldsymbol{u}_h^{(k)}) - \breve{G}(\boldsymbol{u}_h)\right\|_{0,\Omega} + \left\|\delta\breve{G}(\boldsymbol{\varphi}_g) + \boldsymbol{q}_g\right\|_{0,\Omega}\left\|\frac{1}{\delta}\left(\boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)\right\|_{0,\Omega}. \end{aligned}$$

Since both the biliear form $\mathcal{A}$ and the operator $\breve{G}$ we can obtain the following estimation by applying once again the Young's inequality with appropriate constants:

$$\left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}}^2 \leq \tilde{\beta}_{\breve{U}} \left\| \left(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h, \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \left\| (\boldsymbol{\varphi}_g, \boldsymbol{p}_g)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}},$$

where $tilde\beta_{\breve{U}}$ is a positive constant.In the previous subsection we demonstrated, that $\mathcal{Z}^*_{\tilde{\boldsymbol{u}}_h\tilde{\boldsymbol{p}}_h}$ is strong monotone, which particularly means that

$$\begin{aligned}
\alpha_{\mathcal{Z}} \left\| (\boldsymbol{\varphi}_g, \boldsymbol{q}_g)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}}^2 &\leq \left\langle \mathcal{Z}^*_{\boldsymbol{u}_h^{(k)}\boldsymbol{p}_h^{(k)}} (\boldsymbol{\varphi}_g, \boldsymbol{q}_g), (\boldsymbol{\varphi}_g, \boldsymbol{q}_g)^{\mathrm{T}} \right\rangle_{\mathbf{V_h}\times\mathbf{Q_h}} \\
&= \left(\breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}}, (\boldsymbol{\varphi}_g, \boldsymbol{q}_g)^{\mathrm{T}} \right)_{\mathbf{V_h}\times\mathbf{Q_h}} \\
&\leq \left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \left\| (\boldsymbol{\varphi}_g, \boldsymbol{q}_g)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}}.
\end{aligned}$$

In combination with the previous results we finally get for a constant $\beta_{\breve{U}} = \sqrt{\frac{\tilde{\beta}_{\breve{U}}}{\alpha_{\mathcal{Z}}}}$ the upper estimation for the update operator:

$$\left\| \breve{U}\left(\boldsymbol{u}_h^{(k)}, \boldsymbol{p}_h^{(k)}\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}} \leq \beta_{\breve{U}} \left\| \left(\boldsymbol{u}_h^{(k)} - \boldsymbol{u}_h, \boldsymbol{p}_h^{(k)} - \boldsymbol{p}_h\right)^{\mathrm{T}} \right\|_{\mathbf{V_h}\times\mathbf{Q_h}}. \qquad \square$$

7. Summary and outlook

In summary, it can be said that, the scope of this work consists of three parts. First of all we used Lagrange-settings to handle the given constraints and derived a new formulation of the a given problem, a so-called mixed system. The second aspect were the Newton-type schemes, which lead to the tested iteration algorithms. The third part were the refined version of the a posteriori error estimator, that is corresponding to the appropriate mixed system.

By deriving the equivalent mixed formulations, we discussed the obstacle problem, the Stokes problem and the problem in abstract setting in the order of the rising complexity. First we introduced the one dimensional obstacle problem, where the solution function is directly the subject to the restriction. In other words, here we have the simplest example of the constrain operator $\breve{G}$, namely the identity operator. This way it was relatively easy to introduce the usual solution approach and the basics of the proposed mixed system. The Stokes problem provides an example with the more complicated constrain operator $\breve{G}$. This is why, this example of the typical problem was investigated more closely (for example the prove of the existence and the uniqueness of the the solution for stabilized and regularized problem in the sectin 4.6). The imposed the condition on the involved functionals were the basis for the further development in the abstract setting. In both cases, namely the obstacle and the Stockes problems, we merely motivated the proposed mixed formulation. The proper proof of the equivalence of the several formulations were derived in the curse of the section 6.1. Here we combined the a small excursion in the theory of functional minimization with the so far accumulated experience in the mixed variational formulations.

Regarding the second aspect of the work, the Newton-type schemes, the most important question is how to regularize the functional. It a balance act, because on the one hand the more simple approach might have a worse update in each iteration step, an therefor will need more iteration steps to achieve the same precision, when compare to a more complex regularization. On the other hand, the higher calculation "costs" should not be neglected, since even if the specific Newton-type scheme can decrease the number of iteration steps, the overall calculation time might still be worse. Therefor we used an exaple for a Stokes problem with allowed cavitation effects test several Newton-type schemes. The results suggest that in all scenarios, that we considered in the section 4.5, the best approach was the method with the most precise approximation of the Gateau-derivative. The more complex calculation of the update not only paid of, with rising number of degrees of freedom the calculation time was the fraction of the time needed, when solving the problem with the usual method. In the section 4.5 we also examined, how the rising number of degrees of freedom, or the size of the mesh, affects the solution.

We were able to estimate, that the norm of the difference between estimated and the exact solution goes linear down with the mesh size value h. Last but not least, in the section 6.2.2 we transferred the solution strategy to abstract setting and could proof the convergence of the iteration algorithm.

The third part, the a posteriori error estimator is strongly related to the proposed mixed formulation. In the section 4.4 we described, how all the part of the estimator relate to the corresponding conditions of the mixed formulation. The difficulty here was, that we had to derive it twice. The first approach work only for the conform discretization of the continuous spaces. Since we used the Croizeix-Raviart-elements in the numerical test the non-conform discretization needed to be taken into account. This resulted in an additional term in the error estimator and a different concept to derive the estimator.

The possibilities for extensions and future tasks can be found for all the thematic scopes summarized above. For example, we restricted the constrains in the abstract setting to the linear operator $\breve{G}$, but in the theorem 6.1.8 we considered the larger family of the sublinear operators. The next step can be to examine this constrains using f.e. the torsion problem and in similar fashion derive equivalent mixed formulations with the target to solve the problem with Newton-type schemes.

On the subject of the Newton-type schemes, we examined the case, where the the second derivative of the energy functional $\mathcal{J}$ is a bilinearform, that is independent of the solution u. Making it u-dependent would not only result in a wide verity of Newton-type schemes, but also may require additional properties to ensure the convergence.

As for the error estimator, we saw that the worse part of it, in terms of the convergence rates, was the norm of the difference between the non-conform approximation calculate by the solution strategy and the conform approximation estimated using the letter. There we mentioned, that basically "taking the average" is easy to calculate, but not a good estimation. There for it would be interesting to see what better ways can be used and how this would affect the convergence rate and the calculation time.

Appendices

A. Useful theorems and lemmas

Lemma A.0.1 (Friedrichs' inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Then there is a constant $c \in \mathbb{R}^+$ so that*

$$\|v\|_{0,\Omega} \leq c \, \|\nabla v\|_{0,\Omega} \qquad \forall v \in \mathbf{H}_0^1(\Omega) \, .$$

More specific for a two dimensional convex domain Ω

$$\|v\|_{0,\Omega} \leq \frac{d_\Omega^2}{\pi^2} \, \|\nabla v\|_{0,\Omega} \qquad \forall v \in \mathbf{H}_0^1(\Omega) \, ,$$

where d_Ω is the diameter of the domain Ω and π is the ratio of a circle's circumference to its diameter.

Proof. See, for example [11, p. 84] □

Lemma A.0.2 (Poincare's inequality)**.** *Let $\Omega \subset \mathbb{R}^n$ be a bounded domain. Then the inequality*

$$\|v\|_{0,\Omega} \leq \|v\|_{1,\Omega} \qquad \forall v \in \mathbf{H}_0^1(\Omega)$$

results directly from lemma above.

Proof.

$$\begin{aligned} \|v\|_{0,\Omega}^2 =& \frac{1}{1+c^2} \, \|v\|_{0,\Omega}^2 + \frac{c^2}{1+c^2} \, \|v\|_{0,\Omega}^2 \\ \leq& \frac{c^2}{1+c^2} \, \|\nabla v\|_{0,\Omega}^2 + \frac{c^2}{1+c^2} \, \|v\|_{0,\Omega}^2 \\ =& \frac{c^2}{1+c^2} \, \|v\|_{1,\Omega}^2 \\ \leq& \|v\|_{1,\Omega}^2 \, . \end{aligned}$$

□

Lemma A.0.3 (Lemma of projection operator)**.** *Let $\mathbf{Q}$ be a Hilbert space and let $\mathbf{\Lambda} \subset \mathbf{Q}$ be non-empty, closed and convex. Then there is exactly one mapping $\mathit{\Pi}_{\mathbf{\Lambda}} : \mathbf{Q} \longrightarrow \mathbf{\Lambda}$ with*

$$\|p - \mathit{\Pi}_{\mathbf{\Lambda}} p\|_{\mathbf{Q}} = \inf_{q \in \mathbf{Q}} \|p - q\| \qquad \forall p \in \mathbf{Q} \, .$$

For all $p \in \mathbf{Q}$ there is an equivalent characterisation of $\mathit{\Pi}_{\mathbf{\Lambda}} p$ as

$$Re\,(p - \mathit{\Pi}_{\mathbf{\Lambda}} p, q - \mathit{\Pi}_{\mathbf{\Lambda}} p)_{\mathbf{Q}} \leq 0 \qquad q \in \mathbf{\Lambda} \, .$$

The mapping $\mathit{\Pi}_{\mathbf{\Lambda}} : \mathbf{Q} \longrightarrow \mathbf{\Lambda}$ is called (orthogonal) projection of $\mathbf{Q}$ on $\mathbf{\Lambda}$, and $\mathit{\Pi}_{\mathbf{\Lambda}}$ is referred to as projection operator.

Proof. See, for example [1, p. 100] □

Theorem A.0.4 (Abstract existence theorem)**.** *Let* $\mathbf{U}$ *and* $\mathbf{V}$ *be Hilbert spaces. A linear mapping* $\breve{L} : \mathbf{U} \to \mathbf{V}^*$ *is an isomorphism exactly when the bilinear form* $\mathcal{A} : \mathbf{U} \times \mathbf{V} \to \mathbb{R}$*, with* $\mathcal{A}(\boldsymbol{u}, \boldsymbol{v}) = \left\langle \breve{L}\boldsymbol{u}, \boldsymbol{v} \right\rangle_{\mathbf{V}}$ *for* $\boldsymbol{v} \in \mathbf{V}$*, fulfil following conditions:*

(i) (Continuity) There is $\alpha \in \mathbb{R}^+$ *such that*

$$|\mathcal{A}(\boldsymbol{u}, \boldsymbol{v})| \leq \alpha \, \|\boldsymbol{u}\|_{\mathbf{U}} \, \|\boldsymbol{v}\|_{\mathbf{V}} \qquad \forall \boldsymbol{u} \in \mathbf{U}, \forall \boldsymbol{v} \in \mathbf{V} \, . \tag{A.1}$$

(ii) (inf-sup-condition) There is $\beta \in \mathbb{R}^+$ *such that*

$$\sup_{\boldsymbol{v} \in \mathbf{V}} \frac{\mathcal{A}(\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{v}\|_{\mathbf{V}}} \geq \beta \, \|\boldsymbol{u}\|_{\mathbf{U}} \qquad \forall \boldsymbol{u} \in \mathbf{U} \, . \tag{A.2}$$

(iii) For all $\boldsymbol{v} \in \mathbf{V}$*,* $v \neq 0$*, there is* $\boldsymbol{u} \in \mathbf{U}$ *such that*

$$\mathcal{A}(\boldsymbol{u}, \boldsymbol{v}) \neq 0 \, . \tag{A.3}$$

Proof. See, for example [4, p. 119] □

Theorem A.0.5 (Brezzi's splitting theorem)**.** *We consider a mapping into the dual space* $\breve{L} : \mathbf{V} \times \mathbf{Q} \to \mathbf{V}^* \times \mathbf{Q}^*$*,* $(\boldsymbol{u}, \boldsymbol{p})^T \mapsto (\boldsymbol{f}^*, \boldsymbol{g}^*)^T$*, defined by the saddle point problem*

$$\begin{aligned} \mathcal{A}(\boldsymbol{u}, \boldsymbol{\varphi}) + \mathcal{B}(\boldsymbol{\varphi}, \boldsymbol{p}) &= \langle \boldsymbol{f}^*, \boldsymbol{\varphi} \rangle_{\mathbf{V}} && \text{for } \boldsymbol{\varphi} \in \mathbf{V} \, , \\ \mathcal{B}(\boldsymbol{u}, \boldsymbol{q}) &= \langle \boldsymbol{g}^*, \boldsymbol{q} \rangle_{\mathbf{Q}} && \text{for } \boldsymbol{q} \in \mathbf{Q} \, , \end{aligned}$$

where $\mathcal{A}(.,.)$ *and* $\mathcal{B}(.,.)$ *are continuous bilinear forms. The mapping* $\breve{L}$ *is an isomorphism exactly when following condition are met:*

(i) The bilinear form $\mathcal{A}(.\,,.)$ *is V-elliptic, which means there is* $\alpha \in \mathbb{R}^+$ *such that*

$$\mathcal{A}(\boldsymbol{v}, \boldsymbol{v}) \geq \alpha \, \|\boldsymbol{v}\|_{\mathbf{V}}^2 \qquad \forall \boldsymbol{v} \in \{\boldsymbol{\varphi} \in \mathbf{V} |\ \mathcal{B}(\boldsymbol{\varphi}, \boldsymbol{q}) = 0 \ \forall \boldsymbol{q} \in \mathbf{Q}\} \, . \tag{A.4}$$

(ii) The bilinear form $\mathcal{B}(.\,,.)$ *fulfils the inf-sup-condition:*

$$\text{There is } \beta \in \mathbb{R}^+ \text{ such that } \inf_{\boldsymbol{q} \in \mathbf{Q}} \sup_{\boldsymbol{\varphi} \in \mathbf{V}} \frac{\mathcal{B}(\boldsymbol{\varphi}, \boldsymbol{q})}{\|\boldsymbol{\varphi}\|_{\mathbf{V}} \, \|\boldsymbol{q}\|_{\mathbf{Q}}} \geq \beta \, . \tag{A.5}$$

Proof. See, for example [4, p. 126] □

Lemma A.0.6 (Equivalent inf-sup-conditions)**.** *Let $\mathcal{B} : \mathbf{V} \times \mathbf{Q} \to \mathbb{R}$ be a continuous bilinear form. We define*

$$\mathbf{V}_B = \{\boldsymbol{v} \in \mathbf{V} |\, \mathcal{B}(\boldsymbol{v}, \boldsymbol{q}) = 0 \quad \forall \boldsymbol{q} \in \mathbf{Q}\}$$

$\mathbf{V}_B^\circ$ is its polar set and $\mathbf{V}_B^\perp$ is its orthogonal compliment. There are two operators corresponding to the bilinear form $\mathcal{B}$:

$$\breve{B} : \mathbf{V} \to \mathbf{Q}^* \qquad \breve{B}^* : \mathbf{Q} \to \mathbf{V}^*$$
$$\left\langle \breve{B}\boldsymbol{\varphi}, \boldsymbol{q} \right\rangle_{\mathbf{Q}} = \mathcal{B}(\boldsymbol{v}, \boldsymbol{q}) \qquad \left\langle \boldsymbol{\varphi}, \breve{B}^*\boldsymbol{q} \right\rangle_{\mathbf{V}} = \mathcal{B}(\boldsymbol{v}, \boldsymbol{q})\,.$$

Then the following statements are equivalent:

(i) There is a constant $\beta \in \mathbb{R}^+$ such that

$$\inf_{\boldsymbol{q} \in \mathbf{Q}} \sup_{\boldsymbol{\varphi} \in \mathbf{V}} \frac{\mathcal{B}(\boldsymbol{\varphi}, \boldsymbol{q})}{\|\boldsymbol{\varphi}\|_{\mathbf{V}} \, \|\boldsymbol{q}\|_{\mathbf{Q}}} \geq \beta\,.$$

(ii) The mapping $\breve{B} : \mathbf{V}_B^\perp \to \mathbf{Q}^*$ is an isomorphism and

$$\left\| \breve{B}\boldsymbol{v} \right\|_{\mathbf{Q}^*} \geq \beta \, \|\boldsymbol{v}\|_{\mathbf{V}} \qquad \forall \boldsymbol{v} \in \mathbf{V}_B^\perp\,.$$

(iii) The mapping $\breve{B}^* : \mathbf{Q} \to \mathbf{V}_B^\circ \subset \mathbf{V}^*$ is an isomorphism and

$$\left\| \breve{B}^*\boldsymbol{q} \right\|_{\mathbf{V}^*} \geq \beta \, \|\boldsymbol{q}\|_{\mathbf{Q}} \qquad \forall \boldsymbol{q} \in \mathbf{Q}\,.$$

Proof. See, for example [4, p. 125] □

Theorem A.0.7 (Abstract existence theorem for non-linear case)**.** *We consider a mapping in to a dual space $\breve{E} : \mathbf{V} \to \mathbf{V}^*$. Let $\breve{E}$ fulfil following conditions:*

(i) (Strong monotony) There is $\gamma \in \mathbb{R}^+$ such that

$$\left\langle \breve{E}\boldsymbol{\varphi} - \breve{E}\tilde{\boldsymbol{\varphi}}, \boldsymbol{\varphi} - \tilde{\boldsymbol{\varphi}} \right\rangle_{\mathbf{V}} \geq \gamma \, \|\boldsymbol{\varphi} - \tilde{\boldsymbol{\varphi}}\|_{\mathbf{V}}^2 \qquad \forall \boldsymbol{\varphi}, \tilde{\boldsymbol{\varphi}} \in \mathbf{V}\,. \tag{A.6}$$

(ii) (Lipschitz continuity) There is $c \in \mathbb{R}^+$ such that

$$\left\| \breve{E}\boldsymbol{\varphi} - \breve{E}\tilde{\boldsymbol{\varphi}} \right\|_{\mathbf{V}^*} = \sup_{\boldsymbol{v} \in \mathbf{V}} \frac{\left\langle \breve{E}\boldsymbol{\varphi} - \breve{E}\tilde{\boldsymbol{\varphi}}, \boldsymbol{v} \right\rangle_{\mathbf{V}}}{\|\boldsymbol{v}\|_{\mathbf{V}}} \leq c \, \|\boldsymbol{\varphi} - \tilde{\boldsymbol{\varphi}}\|_{\mathbf{V}} \qquad \forall \boldsymbol{\varphi}, \tilde{\boldsymbol{\varphi}} \in \mathbf{V}\,. \tag{A.7}$$

Then the equation

$$\breve{E}\boldsymbol{u} = \boldsymbol{f}^*$$

has exactly one solution $\boldsymbol{u} \in \mathbf{V}$ for any $\boldsymbol{f}^* \in \mathbf{V}^*$.

Proof. See, for example [11, p. 109 ff.] □

Theorem A.0.8 (Abstract existence theorem of a numerical solution for non-linear case)**.** *Let* $\breve{E} : \mathbf{V} \to \mathbf{V}^*$ *be strong monotone and Lipschitz continues mapping and* $\boldsymbol{f} \in \mathbf{V}^*$*. Then there is exactly one* $\boldsymbol{u}_h \in \mathbf{V}_h$ *for each linear subset* $\mathbf{V}_h \subset \mathbf{V}$ *with* $\dim \mathbf{V}_h < \infty$ *that solves the equation*

$$\left\langle \breve{E}\boldsymbol{u}_h, \boldsymbol{\varphi}_h \right\rangle_{\mathbf{V}} = \langle \boldsymbol{f}^*, \boldsymbol{\varphi}_h \rangle_{\mathbf{V}} \qquad \forall \boldsymbol{\varphi}_h \in \mathbf{V}_h \, .$$

Proof. See, for example [11, p. 112 ff.] □

Bibliography

[1] H. W. Alt. *Lineare Fuktionanalysis*. Springer-Verlag, Berlin Heidelberg, 2012.

[2] D. Biermann, H. Blum, I. Iovkov, N. Klein, A. Rademacher, and F.-T. Suttmeier. Stabilization techniques and a posteriori error estimates for the obstacle problem. *Applied Mathematical Science*, 7(127):6329–6346, 2013.

[3] H. Blum, D. Braess, and F.T. Suttmeier. A cascadic multigrid algorthm for variational inequalities. *Computing and Visualization in Science*, 2004.

[4] D. Braess. *Finite Elemente - Theorie, schnelle Löser und Anwendungen in der Elastizitätstheorie*. Springer-Verlag, Berlin Heidelberg, 2013.

[5] E. Dari, R. Duran, and C. Padra. Error Estimators for Nonconforming Finite Element Approximations of the Stokes Problem. *Math.Comp.*, 64(211), 1995.

[6] R. S. Dembo and U. Tulowitzky. On the minimization of quadratic functions subject to box constrints. *Yale Research Center for Scientific Computation*, 1984.

[7] M. Dobrowolski. *Angewandte Funktionalanalysis*. Springer, 2006.

[8] C. Geiger and Kanzow Chr. *Theorie und Numerik restringierter Optimierungsaufgaben*. Springer-Verlag, Berlin Heidelberg, 2002.

[9] F. Gimbel, P. Hansbo, and F.T. Suttmeier. An adaptive low-order FE-scheme for stokes flow with cavitation. *J. Numer. Math.*, 18(3):177–186, 2010.

[10] V. Girault and P.-A. Raviart. *Finite Element Approximation of the Navier-Stokes Equations*. Springer Verlag, Heidelberg New York, 1979.

[11] Ch. Großmann and H.-G. Roos. *Numerik partieller Differentialgleichungen*. B.G. Teubner, Stuttgart, 1994.

[12] M.K.V. Murthy J. Cea. *Lectures on Optimization - Theory and Algorithms (Tata Institute Lectures on Mathematics and Physics)*. Springer, 1979.

[13] B. Nilsson and P. Hansbo. A stokes model with cavitation for the numerical simulation of hydrodynamic lubrication. Technical Report 17, Chalmers University of Gothenburg, 2008.

[14] B. Nilsson and P. Hansbo. Weak coupling of a reynolds model and a stokes model for hydrodynamic lubrication. *Internat. J. Numer. Methods Fluids*, 2010. DOI: 10.1002/fld.2281, to appear.

[15] Fabio Nobile. A posteriori error estimates for the finite element approximation of the stokes problem. Technical report, 2003.

[16] R. Verfürth. A posteriori error estimators for the stokes equation. *Numerische Mathematik*, 55:309–325, 1989.