JOHANNES KEPLER
UNIVERSITÄT LINZ
Netzwerk für Forschung, Lehre und Praxis

Uzawa-Type Methods For The Obstacle Problem

MASTERARBEIT

zur Erlangung des akademischen Grades

DIPLOMINGENIEUR

im Masterstudium

INDUSTRIEMATHEMATIK

Angefertigt am *Institut für Numerische Mathematik*

Betreuung:

A. Univ.-Prof. Dipl.-Ing. Dr. Walter Zulehner

Eingereicht von:

Henry Kasumba

Linz, July 2007

To my family

Abstract

In this thesis we study the obstacle problem. It is a free boundary problem and the computation of approximate solution can be difficult and expensive. This thesis addresses some aspects of this issue.

We discretize the gradient with lowest order Raviart-Thomas elements and functional values by piecewise constant elements. Existence and uniqueness of solutions to the discrete problems is studied and error estimates are obtained.

We develop the Uzawa-type algorithms for the discrete system of linear equations and inequalities that results from the discretization of the mixed formulation of the obstacle problem.

The convergence of classical Uzawa method is analyzed and we display numerical results that agree with theoretical results.

Acknowledgements

The Writing of this thesis has been one of the most significant challenges i have ever had to face. Without support, patience and guidance of the following people, this study would not have been completed. it is to them that i owe my deepest gratitude.

• Professor Walter Zulehner , who undertook to act as my supervisor, despite his many other academic and professional commitment. His wisdom, knowledge and commitment to highest standards inspired and motivated me.

• My Friends Vincent Ssemaganda, Kho sinatra, Yayun Zhou, Edward Bbaale and Philip Ngale who inspired my final effort despite the enormous work pressures we were facing together.

• My parents Noel and Christopher Kiyemba, who have always supported , encouraged and believed in me, in all my endeavors and who so lovingly and unselfishly supported all my plans during my stay away from home.

• My Lecturers both in Eindhoven University of Technology (The Netherlands) and Johannes Kepler University Linz (Austria) for giving me all the necessary knowledge during the course of my study in these two patterner universities.

Last I would like to acknowledge the support of European union for supporting my study as well as my stay in Europe for a period of two years.

Johannes Kepler University Henry Kasumba
July , 2007

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

One important type of contact problem is that which involves the contact between an elastic solid and a rigid obstacle. Problems in this category are sometimes called 'obstacle problems'. Obstacle problems are a type of free boundary problem. They are of interest both for their intrinsic beauty and for the wide range of applications they describe in subjects ranging from physics to finance. Many important problems can be formulated by transformation to an obstacle problem, for example, the filtration dam problem [21], the Stefan problem [21], the subsonic flow problem [12], American options pricing model [15],etc. Since obstacle problems are highly nonlinear, the computation of approximate solutions can be a challenge. The main existing numerical methods for the solution of contact problems in general and obstacle problems in particular, are 'gap element' scheme, mathematical programming approach, and schemes based on penalty formulations and Lagrange multiplier formulations.

In this thesis we concentrate on the latter. In the Lagrange multiplier formulation, we reformulate the boundary value problem formulation of the obstacle problem in a weak form. It has been shown, see e.g Noor and Whiteman [18], that in the presence of a constraint, such an approach leads to a variational inequality. This variational inequality is equivalent to the minimization problem. There are classical iterative methods like point projection methods and point over-relaxation methods [6] for solving this minimization problem but these suffer from slow convergence rates on finer meshes due to the fact that the L^2 -norm is used as a measure for the projection onto the convex set instead of the actual dual norm . In these approaches

H^1 conforming finite elements are used as an approximating space for the primal variables. For this approximating space, one needs to choose basis functions such that they are continuous across the inter element boundaries.

On the other hand, it has been noticed that one could weaken the requirement of inter element continuity for the functions in the finite element subspace of $H^1(\Omega)$ and still obtain a convergent finite element method. This finite element approach is called a mixed finite element which is founded on a variational principle expressing an equilibrium (saddle point) condition and not a minimization principle. With this approach, we approximate both the scalar variable and the vector variable and here comes the name mixed. For the approach that we consider in this thesis, the primal variable is in $H(div, \Omega)$, and the dual variable is in $L^2(\Omega)$ and this thus makes the definition of the projection easier. The resulting system of equations and inequalities are solved using Uzawa's method. We investigate the performance of the classical Uzawa algorithm and its variants. Our ultimate goal will be to obtain a robust Uzawa-type algorithm for solving the mixed variational inequality. We further present numerical examples to verify the error estimates. Moreover, we give numerical examples which include problems involving an elastic membrane, encountering a flat or a non-flat rigid obstacle, and investigate the finite element convergence to the location of the free boundary

The outline of the remainder of this thesis is as follows:
In Chapter 2, a mathematical model for the obstacle problem is derived. Different equivalent formulations are derived. Existence and uniqueness of solutions to these formulations is also discussed.

In Chapter 3, the new primal problem and its equivalent mixed formulations are discretized using the mixed finite element method of Raviart-Thomas [20]. Error estimates for the approximation of the solution are provided.

In Chapter 4, we look at how to solve the discretized problem. The classical Uzawa algorithm and its variants for solving the system of equations and inequalities is presented. We analyze it and show its convergence.

In Chapter 5, several cases involving an elastic membrane encountering a flat or non-flat rigid obstacle are considered and numerical solutions are displayed. In order

to actually compute the free boundary, the case where the exact solution is known is chosen to demonstrate our findings. We present the performance of the classical Uzawa algorithm and its variants.

In Chapter 6 , we give some conclusive remarks and some ideas for the extension of this work.

Chapter 2

Obstacle Problem Formulation

In this chapter we study the obstacle problem. We give a physical example of this obstacle problem and further derive the different mathematical formulations for this problem with reference to [16]. We show the equivalence of these formulations and further analyze the existence and uniqueness of the solution to these formulations. Motivated by the mixed method of Raviart-Thomas [20], we focus on the mixed formulation of the obstacle Problem. In Chapter 3 we shall introduce finite dimensional spaces and approximate the mixed problem.

2.1 Physical Example

Let us consider a horizontal circular wire and a membrane hanging on this wire. We assume that this membrane is horizontal and above a plate. When we load the membrane with a force f in the vertical direction, it undergoes deflection and we get a contact area between the membrane and the obstacle which is the plate. This contact area is called the coincidence set (fig 2.1(b)).

The boundary of the coincidence set is called the free boundary for the obstacle problem. The location of this boundary is not known apriori and its part of our problem.

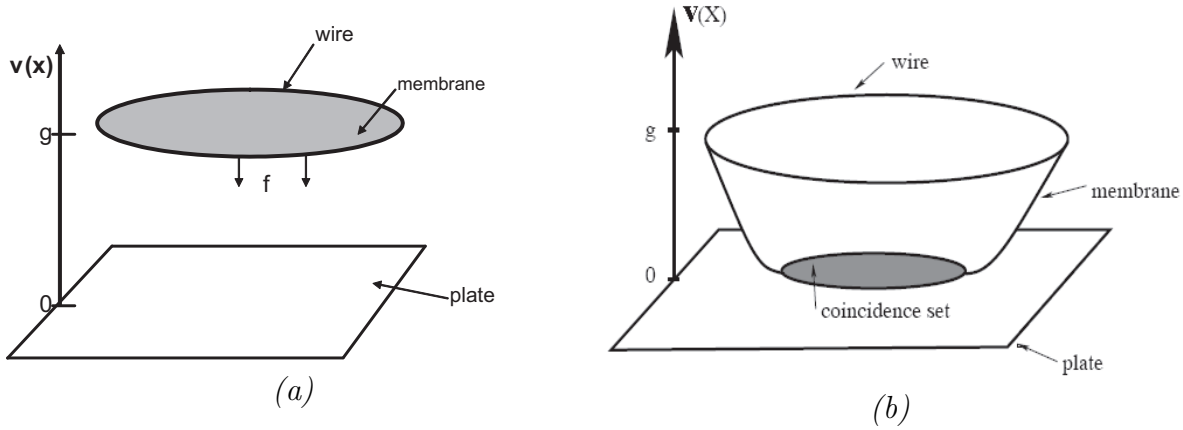


Figure 2.1: Membrane over a plate

2.1.1 Mathematical Formulation

We now give a simple mathematical model for this problem. We assume a homogeneous membrane represented by a domain $\Omega \subset \mathbb{R}^2$ a distance g from the plate (obstacle) (fig 2.1(a)). When this membrane is loaded with a force f in the vertical direction, it undergoes a deflection (fig 2.1(b)). Let v describe the new position of this membrane at a point $(x, y) \in \Omega$. The membrane is restricted from below by a horizontal plate,i.e.,

$$v \geq 0 \text{ on } \Omega. \tag{2.1.1}$$

In addition the wire is described by

$$v \;=\; constant \;=\; g \text{ on } \partial\Omega. \tag{2.1.2}$$

From calculus of variations, the surface area of the deformed membrane is given by

$$\text{Surface Area} = \int_\Omega \sqrt{1 + v_x^2 + v_y^2}\, dxdy.$$

We assume that the potential energy of the deflected membrane is proportional to the change of area of its surface[8], such that

$$P(v) = \int_\Omega \sqrt{1 + v_x^2 + v_y^2}\, dxdy - \text{meas}(\Omega),$$

where meas(Ω) is the surface area of the undeflected membrane (in figure 2.1(a)). Assuming small displacements ($(v-g) \ll 1$), higher order terms are neglected. Hence

we obtain

$$P(v) = \frac{1}{2}\int_{\Omega} |\nabla v|^2.$$

The work of external forces, corresponding to v is given by

$$E(v) = \int_{\Omega} fv dxdy,$$

and total energy $J(v) = P(v) - E(v)$. i.e

$$J(v) = \frac{1}{2}\int_{\Omega} |\nabla v|^2 - \int_{\Omega} fv dxdy. \tag{2.1.3}$$

2.1.2 Minimization and Variational Formulation

From the Lagrange principle of minimizing the total energy, the equilibrium state of the membrane is realized by a function u minimizing J over a class of functions v with finite energy, and with a prescribed value g on the boundary $\partial\Omega$. More precisely

$$u \in \tilde{K}^* \text{ such that } J(u) \leq J(v)\ \forall v \in \tilde{K}^*, \tag{2.1.4}$$

where

$$\tilde{K}^* = \{v \in H^1(\Omega) \ \ | v = g \text{ on } \partial\Omega, v \geq 0\}. \tag{2.1.5}$$

More generally, a rigid obstacle is represented by a body occupying the set

$$Q = \{[x, y, z] \in \mathbb{R}^3 | z \leq \psi(x, y)\}.$$

In addition, we also allow non-constant boundary conditions such that the set of admissible deflections is now given as,

$$\tilde{K} = \{v \in H^1(\Omega) : \quad v = g \text{ on } \partial\Omega,\ v \geq \psi \ a.e\ in\ \ \Omega\}. \tag{2.1.6}$$

Let ψ be such that $\tilde{K}$ is non-empty. Then $\tilde{K}$ is convex and closed (see[6] for the proof).

Therefore the obstacle problem can be posed as a problem in the calculus of variations. It is solved by the solution of the minimization problem:

$$\text{Find u} \in \tilde{K} : \ \ J(u) \leq J(v) \ \text{ for any } \ v \in \tilde{K}. \tag{2.1.7}$$

By putting J in the abstract form:

$$J(v) = \frac{1}{2}\tilde{a}(v,v) - (\tilde{F}, v),$$

where $(\cdot,\cdot)$ denotes the $L^2(\Omega)$ inner product, and $\tilde{a}(\cdot,\cdot)$ is of the form

$$\tilde{a}(u,v) \;=\; (\nabla u, \nabla v) \quad for\ all\ u,\ v \in H^1(\Omega), \tag{2.1.8}$$

and $\tilde{F} = f$. We assume that the bilinear form $\tilde{a}$ is positive, symmetric and that the convex set $\tilde{K}$ is non empty, convex and closed, then the following proposition holds.

Proposition 2.1.1. *$u \in \tilde{K}$ solves (2.1.7) if and only if it solves*

$$\tilde{a}(u, v-u) \geq (\tilde{F}, v-u) \quad for\ all\ v \in \tilde{K}. \tag{2.1.9}$$

Proof. Suppose 2.1.7 holds, then for $0 < \epsilon < 1$, for any $v \in \tilde{K}$, $u + \epsilon(v-u) \in \tilde{K}$ as $\tilde{K}$ is convex.
Let $J(u) = \inf_{v\in\tilde{K}} J(v)$, this implies that

$$\begin{aligned} J(u) &\leq J(u + \epsilon(v-u)) \\ &= \frac{1}{2}\tilde{a}(u+\epsilon(v-u), u+\epsilon(v-u)) - (\tilde{F}, u+\epsilon(v-u)) \\ &= \frac{1}{2}\tilde{a}(u,u) - (f,u) + \epsilon\left(\tilde{a}(u, v-u) - (\tilde{F}, v-u)\right) + \frac{\epsilon^2}{2}\tilde{a}(v-u, v-u) \\ &= J(u) + \epsilon\left(\tilde{a}(u, v-u) - (\tilde{F}, v-u)\right) + \frac{\epsilon^2}{2}\tilde{a}(v-u, v-u). \end{aligned}$$

By subtracting $J(u)$ and dividing by ϵ we obtain

$$0 \;\leq\; \tilde{a}(u, v-u) - (\tilde{F}, v-u) + \frac{\epsilon}{2}\tilde{a}(v-u, v-u).$$

In the limit as $\epsilon \to 0$, we see $\tilde{a}(u, v-u) - (\tilde{F}, v-u) \geq 0$, the variational inequality. Conversely suppose u solves (2.1.9). Let $v \in \tilde{K}$ with $v \neq u$, and let $f(\epsilon) = J(u + \epsilon(v-u))$. For $0 \leq \epsilon \leq 1$. Note $f(0) = J(u),\ f(1) = J(v)$ and f is continuous .
Now we calculate $f'(\epsilon)$ for $0 \leq \epsilon \leq 1$

$$f'(\epsilon) = \lim_{h\to 0} \frac{J\left(u + (\epsilon+h)(v-u)\right) - J\left(u+\epsilon(v-u)\right)}{h}$$

Take $w = u + \epsilon(v - u)$

$$\begin{aligned}
\Rightarrow f'(\epsilon) &= \lim_{h\to 0} \frac{J\left(w + h(v-u)\right) - J\left(w\right)}{h} \\
&= \lim_{h\to 0} \frac{h\tilde{a}(w,(v-u)) + \frac{h^2}{2}\tilde{a}(v-u,v-u) - h(\tilde{F},(v-u))}{h} \\
&= \tilde{a}(w,v-u) - (\tilde{F},v-u) \\
&= \tilde{a}(u,v-u) - (\tilde{F},v-u) + \epsilon\tilde{a}(v-u,v-u) \\
&> \tilde{a}(u,v-u) - (\tilde{F},v-u) \geq 0 \quad (by \;\; 2.1.9).
\end{aligned}$$

Thus $f(1) > f(0)$ and u minimizes $J(u)$ that is (2.1.7). □

Condition (2.1.9) is called the variational inequality formulation. This formulation sometimes called the primal formulation of the obstacle problem. Having established the equivalence of the minimization and variational formulation, it remains to show the existence and uniqueness of solution to the primal formulation (2.1.9), which we discuss in the following section (2.2)

2.2 Existence of the solution

As we have seen in the previous section, the obstacle problem can be described as a minimization problem. This minimization is equivalent to the variational inequality which we call a primal formulation of the obstacle problem. We will now present some general theory about the existence and uniqueness of the solution to these problems. The idea of proofs to the general theorems stated in this section were obtained from [16], [14]. We then adapt these general ideas to the obstacle problem.

2.2.1 Existence and Uniqueness for the Primal Problem

Throughout this thesis, unless specified, the following parenthesis $\langle\cdot,\cdot\rangle_R$ shall denote the $L^2(R)$ inner product on Ω. For example,

$$\langle f,g\rangle_\Omega = \int_\Omega f(x,y)g(x,y)dxdy \;\; \text{and} \;\; \langle f,g\rangle_{\partial\Omega} = \int_{\partial\Omega} f(x,y)g(x,y)d\sigma.$$

If the subscript R is omitted, we assume that $R = \Omega$. The scalar product and the norm in the Sobolev space $H^k(\Omega)$ are denoted by $(\cdot,\cdot)_k$ and $\|\cdot\|_k$ respectively. Let

V be a real Hilbert space equipped with scalar product $(\cdot,\cdot)_V$ and the norm $\|\cdot\|_V$. Let V^* denote its dual and $(\cdot,\cdot)$ the duality pairing. Additionally, we need V^0 to be a closed subspace of V, then $V_g := u_g + V_0$ denote a linear manifold containing all functions satisfying essential boundary conditions i.e ($u = g$ on $\partial\Omega$). We assume that the set of admissible solutions $\tilde{K}$ is non-empty, closed and convex. we further assume that the bilinear form $\tilde{a} : V \times V \to \mathbb{R}$ is symmetric, bounded and elliptic on V_0 such that,

$$|\tilde{a}(u,v)| \leq \alpha_1 ||u||_V ||v||_V \quad \forall \ u, v \in V \tag{2.2.1}$$

$$\tilde{a}(v,v) \geq \alpha_2 ||v||_V^2 \quad \forall \ v \in V_0 \tag{2.2.2}$$

Then the following theorem holds.

Theorem 2.2.1. *Let V be the Hilbert space, and $\tilde{K}$ non-empty , closed and convex subset of V_g. Let $\tilde{a}$ be symmetric and satisfy (2.2.1) and (2.2.2) and $\tilde{F} \in V^*$ such that*

$$\tilde{a}(u, v-u) \geq (\tilde{F}, v-u) \quad for\ all\ v \in \tilde{K}. \tag{2.2.3}$$

Then there exists a unique solution $u \in \tilde{K}$ of the variational problem (2.2.3). If u_1, u_2 are solutions to problem (2.2.3) with corresponding right hand sides $\tilde{F}_1,\ \tilde{F}_2 \in V^*$, then the following stability estimate holds

$$\|u_1 - u_2\| \leq \frac{1}{\alpha_2} \|\tilde{F}_2 - \tilde{F}_1\| \ . \tag{2.2.4}$$

Proof. We begin with the proof of the existence of solution u to (2.2.3), which we present in several steps. First we define the functional $J(u)$ and use our assumption that a is symmetric

$$J(u) = \frac{1}{2}\tilde{a}(u,u) - (\tilde{F}, u). \quad u \in V \tag{2.2.5}$$

Let $d = \inf_{\tilde{K}} \mathrm{J(u)}$. Since

$$\begin{aligned} J(u) &\geq \frac{\alpha_2}{2}\|u\|_V^2 - \|\tilde{F}\|_{V^*}\|u\|_V \\ &\geq \frac{\alpha_2}{2}\|u\|_V^2 - \frac{1}{2\alpha_2}\|\tilde{F}\|_{V^*}^2 - \frac{\alpha_2}{2}\|u\|_V^2 \\ &\geq -\frac{1}{2\alpha_2}\|\tilde{F}\|_{V^*}^2 \ , \end{aligned}$$

we see that $d \geq -\frac{1}{2\alpha_2}\|\tilde{F}\|^2_{V^*} \geq -\infty$. Let u_n be a minimizing sequence of J in $\tilde{K}$ such that

$$\left\{u_n \in \tilde{K} : d \leq J(u_n) \leq d + (1/n)\right\}$$

Applying the parallelogram law, and keeping in mind that $\tilde{K}$ is convex, we see that

$$\begin{aligned}
\alpha_2\|u_n - u_m\|^2_V &\leq \tilde{a}(u_n - u_m, u_n - u_m) \\
&= 2\tilde{a}(u_n, u_n) + 2\tilde{a}(u_m, u_m) - 4\tilde{a}(\frac{1}{2}(u_n + u_m), \frac{1}{2}(u_n + u_m)) \\
&= 4J(u_n) + 4J(u_m) - 8J(\frac{u_n + u_m}{2}) \\
&\leq 4[(1/n) + (1/m)],
\end{aligned}$$

where we have used

$$4(\tilde{F}, u_n) + 4(\tilde{F}, u_m) - 8(\tilde{F}, \frac{1}{2}(u_n + u_m)) = 0.$$

Hence the sequence $\{u_n\}$ is a cauchy sequence and the closed set $\tilde{K}$ contains an element u such that $u_n \to u$ in V and $J(u_n) \to J(u)$. So $J(u) = d$.
Now that we have seen the existence of the solution. It remains to show the uniqueness of this solution and consequently the stability estimate (2.2.4). First, suppose there exists $u_1, u_2 \in V$ be solutions to the variational inequalities

$$u \in \tilde{K} : \tilde{a}(u_i, v - u_i) \geq (\tilde{F}_i, v - u_i) \quad \forall\ v \in \tilde{K}, i = 1, 2. \tag{2.2.6}$$

Setting $v = u_2$ in the variational inequality for u_1 and $v = u_1$ in that of u_2 we obtain, upon adding

$$\tilde{a}(u_1 - u_2, u_1 - u_2) \leq (\tilde{F}_1 - \tilde{F}_2, u_1 - u_2).$$

Hence by coerciveness of $\tilde{a}$

$$\alpha_2\|u_1 - u_2\|^2_V \leq \|\tilde{F}_2 - \tilde{F}_1\|_{V^*}\|u_1 - u_2\|_V.$$

and thus there holds the stability estimate (2.2.4). For $\tilde{F}_1 = \tilde{F}_2$, we can also see uniqueness of the solution. □

Application to the obstacle problem

We have seen a general theorem on variational inequalities, that assures existence and uniqueness of a solution to a primal problem, provided all conditions are satisfied. So

our next aim is to prove that the primal formulation for the obstacle problem satisfy the required conditions.

Primal variational formulation

We formulated the following form of the obstacle problem:
Find $u \in \tilde{K} = \{v \in H^1(\Omega) : v \geq \psi \text{ in } \Omega,\ (v-g)|_{\partial\Omega} = 0\}$ such that

$$\tilde{a}(u, v-u) \geq \langle \tilde{F}, v-u \rangle \quad for\ all\ v \in \tilde{K}, \tag{2.2.7}$$

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

$$\begin{aligned} \tilde{a}(u,v) &= \int_\Omega \nabla u \nabla v \ dx \ \text{ for all } \ u, v \in H^1(\Omega)\ , \\ \langle \tilde{F}, v \rangle &= \int_\Omega f.v \ \ dx \ \ . \end{aligned}$$

We can immediately see that the bilinear form $\tilde{a} : H^1(\Omega) \times H^1(\Omega) \to \mathbb{R}$ is symmetric. Boundedness of $\tilde{a}$ follows of cauchy's inequality i.e.

$$\begin{aligned} |\tilde{a}(u,v)| &= |\int_\Omega \nabla u \nabla v \ dx| \\ &\leq \|\nabla u\|_{L^2(\Omega)} \|\nabla v\|_{L^2(\Omega)} \\ &\leq \|u\|_1 \|v\|_1 \ . \end{aligned}$$

In addition, $\tilde{a}$ is coercive on $H_0^1(\Omega)$ i.e.

$$\tilde{a}(u,u) = \int_\Omega |\nabla u|^2 \ dx \ = \ \frac{1}{2}\int_\Omega |\nabla u|^2 + \frac{1}{2}\int_\Omega |\nabla u|^2 \ .$$

Using Friedrichs' Inequality [9], we have $\|u\|_0^2 \leq \gamma \|\nabla u\|_0^2 \ \forall u \in H_0^1(\Omega)$ hence

$$\begin{aligned} \tilde{a}(u,u) &\geq \frac{1}{2}\int_\Omega |\nabla u|^2 + c\|u\|_0^2 \\ &\geq c'\|u\|_1^2 \quad \text{for all } u \in H_0^1(\Omega)\ , \end{aligned}$$

where $c = \frac{1}{\gamma}, c' = \min(1/2, 1/2c)$ are positive constants.

Clearly $\tilde{F}$ is bounded and linear. Therefore we get existence and uniqueness of a solution $u \in \tilde{K}$ by Theorem (2.2.1).

Linear Complementarity Problem Formulation

As a motivation towards deriving a new primal formulation that is well suited for the mixed method of Raviart-Thomas, we recall the obstacle problem: Given $f \in L^2(\Omega)$ and $(g, \psi) \in H^1(\Omega)$ with $\psi \geq g$ almost everywhere on $\partial\Omega$, find $u \in \tilde{K}$ such that

$$\langle \nabla u, \nabla(v-u) \rangle \geq \langle f, v-u \rangle \quad \text{for all } v \in \tilde{K}, \tag{2.2.8}$$

where

$$\tilde{K} = \{v \in H^1(\Omega) | v \geq \psi \ \ a.e \text{ in } \ \Omega, \ \ (v-g)|_{\partial\Omega} = 0\} \ . \tag{2.2.9}$$

We suppose that we have a high regularity of the solution ,i.e., $u \in H^2(\Omega) \cap \tilde{K}$. Then, by applying Greens formula to the left hand side of (2.2.8), we obtain

$$\int_\Omega -\Delta u(v-u)dxdy + \int_{\partial\Omega} \frac{\partial u}{\partial n}(v-u)ds \geq \int_\Omega f(v-u)dxdy \ \ \forall v \in \tilde{K} \ . \tag{2.2.10}$$

As $v - u = 0$ on $\partial\Omega$, the integral along the boundary $\partial\Omega$ vanishes. Let $\phi \in C_0^\infty(\Omega)$ be a nonnegative function in Ω. Then $v = u + t\phi \in \tilde{K}$ for any $0 < t \leq 1$. Substituting this element into (2.2.10) we get

$$\int_\Omega (-\Delta u - f)\phi dxdy \geq 0,$$

for all functions $\phi \in C_0^\infty(\Omega)$, with $\phi \geq 0$ in Ω. Hence

$$-\Delta u \geq f \ \ \text{a.e in } \Omega \ .$$

The domain Ω can now be divided as follows:

$$\Omega = \Omega_0 \cup \Omega_+,$$

where

$$\Omega_0 = \{A \in \Omega | u(A) = \psi(A)\},$$
$$\Omega_+ = \{A \in \Omega | u(A) > \psi(A)\}.$$

Let us assume that $\psi \in C(\bar{\Omega})$. As $H^2(\Omega) \hookrightarrow C(\bar{\Omega})$, the set Ω_+ is open. Let $\tilde{A} \in \Omega_+$ be given. Then there exists a neighborhood $U_\delta(\tilde{A}) \subset \Omega_+$. If $\phi \in C_0^\infty(U_\delta(\tilde{A}))$, the

function $v = u \pm t\phi$ belongs to $\tilde{K}$ i.e $v = u \pm t\phi \geq \psi$ provided t is sufficiently small, both positive and negative. Substituting v into (2.2.10), we obtain

$$(\pm t) \int_{U_\delta(\tilde{A})} -\Delta u\phi \; dxdy \geq (\pm t) \int_{U_\delta(\tilde{A})} f\phi \; dxdy,$$

which holds for $\phi \in C_0^\infty(U_\delta(\tilde{A}))$. Therefore

$$-\Delta u = f \quad \text{a.e in} \;\; U_\delta(\tilde{A})).$$

Summing up, we have proved that for $u \in H^2(\Omega) \cap \tilde{K}$, being the solution of (2.1.9) satisfies the following set of relations

$$\begin{aligned} -\Delta u &\geq f \;\; \text{a.e in } \Omega, \\ u &\geq \psi \;\; \text{a.e in } \Omega, \\ \text{if} \qquad u(A) &> \psi(A) \text{ then } \; -\Delta u(A) = f(A). \end{aligned} \tag{2.2.11}$$

Instead of (2.2.11) we can write also

$$-\Delta u \geq f, \;\; u \geq \psi \in \Omega, \tag{2.2.12}$$

$$(u - \psi)(-\Delta u - f) = 0 \;\; \text{a.e } in \; \Omega. \tag{2.2.13}$$

This formulation (2.2.12 - 2.2.13) is the **linear complementarity** problem(LCP) formulation of the obstacle problem.

New Primal Formulation

Now we have seen that, provided everything is smooth enough, then the LCP formulation (2.2.12) holds. Motivated by this formulation, we would like to further give a new primal formulation of (2.2.12 - 2.2.12) that is well suited for the mixed method of Raviart and Thomas [20]. Let us define the space

$$H(div, \Omega) = \{q \in L^2(\Omega) \times L^2(\Omega) : \nabla.q \in L^2(\Omega)\},$$

with associated scalar product form

$$[q, p]_\Omega = \langle q, p \rangle_\Omega + \langle \nabla.q, \nabla.p \rangle_\Omega, \qquad ||q||_{H,\Omega} = [q, q]_\Omega^{1/2}.$$

The first step towards deriving a new primal formulation is to let

$$p = \nabla u \ . \tag{2.2.14}$$

We also define a new convex set K such that

$$K = \{q \in H(div, \Omega) : \nabla.q + f \leq 0 \ \ a.e \text{ in } \ \Omega\}.$$

If we multiply equation (2.2.14) by a test function q from K and integrate over Ω, by using Greens formula, and noting that $u = g$ on the boundary $\partial\Omega$ we obtain

$$\langle p, q\rangle = \langle \nabla u, q\rangle = -\langle u, \nabla.q\rangle + \langle u, q.n\rangle_{\partial\Omega} = -\langle u, \nabla.q\rangle + \langle g, q.n\rangle_{\partial\Omega} \ . \tag{2.2.15}$$

We can replace q by $(q - p)$ in (2.2.15) and obtain

$$\langle p, q - p\rangle = -\langle u, \nabla.(q - p)\rangle + \langle g, (q - p).n\rangle_{\partial\Omega} \ . \tag{2.2.16}$$

From conditions (2.2.12) and (2.2.13) one has $\psi - u \leq 0$ and $\nabla.q + f \leq 0, \ \ \forall \ q \in K$, so combining these two conditions, we obtain

$$\langle \psi - u, (\nabla.q + f)\rangle \geq 0 \ . \tag{2.2.17}$$

From (2.2.13) one has

$$\langle \psi - u, \nabla.p + f\rangle = 0 \quad \text{ a.e in } \Omega \ . \tag{2.2.18}$$

So we can re-write (2.2.17) as

$$\langle \psi - u, (\nabla.q + f)\rangle - \langle \psi - u, \nabla.p + f\rangle \geq 0 \quad \text{ a.e in } \Omega \ . \tag{2.2.19}$$

From this equation we see that

$$\langle \psi, \nabla.(q - p)\rangle - \langle u, \nabla.(q - p)\rangle \geq 0 \ . \tag{2.2.20}$$

Combining (2.2.20) and (2.2.16) we obtain the new primal formulation

$$\langle p, q - p\rangle \geq -\langle \psi, \nabla.(q - p)\rangle + \langle g, (q - p).n\rangle_{\partial\Omega} \quad \text{ for all } q \in K. \tag{2.2.21}$$

Putting together all the ideas we have gathered so far, we have derived a new primal formulation of (2.2.12) that is well suited for the mixed method i.e:
Find $p \in K$ such that

$$\langle p, q-p\rangle \geq -\langle \psi, \nabla.(q-p)\rangle + \langle g, (q-p).n\rangle_{\partial\Omega} \quad \text{for all } q \in K, \tag{2.2.22}$$

where

$$K = \{q \in H(div,\Omega) : \nabla.q + f \leq 0 \ \ a.e \text{ in } \ \Omega\}. \tag{2.2.23}$$

For all $q \in H(div,\Omega)$, Greens formula implies that $q.n \in W^{-1/2}(\partial\Omega)$. Since $g \in W^1(\Omega)$, we also have $g \in W^{1/2}(\partial\Omega)$ by the trace theorem. Hence the last term in (2.2.22) is well defined. Let us define the new closed convex cone Λ in $L^2(\Omega)$:

$$\Lambda = \{\mu \in L^2(\Omega) : \mu \geq 0 \text{ a.e. } \text{ in } \ \Omega\} \ . \tag{2.2.24}$$

The new primal formulation (2.2.22-2.2.23) above can then be written in abstract form:
Find $p \in K$ such that

$$a(p, q-p) \ \ \geq \ \ (F, q-p) \quad \text{for all } q \in K, \tag{2.2.25}$$

where

$$K = \{q \in V : b(q,\mu) \leq (\phi,\mu) \ \text{ for all } \mu \in \Lambda\}. \tag{2.2.26}$$

and

$$a(p,q) = \langle p, q\rangle, \quad b(q,\mu) = \langle \nabla.q, \mu\rangle \ \ , \tag{2.2.27}$$

$$(\phi,\mu) = -\langle f, \mu\rangle, \quad V = H(div,\Omega) \ , \tag{2.2.28}$$

$$(F,q) = -\langle \psi, \nabla.q\rangle + \langle g, q.n\rangle_{\partial\Omega} \ . \tag{2.2.29}$$

We have seen the existence of the solution $p = \nabla u$ to (2.2.25) provided $u \in \tilde{K} \cap H^2(\Omega)$. It remains to show the uniqueness of this solution. We note that the bilinear form a is not coercive and hence Lax-Milgram Theorem 2.2.1 can not be used to show the uniqueness of the solution. However it is important to note that the bilinear form

$$a(q,q) = \int q^2 > 0 \ \text{ for all } q \in V \text{ with } q \neq 0$$

We now state a theorem from [14] that assures us about the uniqueness of the solution p

Theorem 2.2.2. *If $f \in L^2(\Omega)$, $(g, \psi) \in H^2(\Omega)$, and u is the unique solution to (2.2.8), then $p = \nabla u$ is the unique solution to (2.2.22).*

Proof. Now we have seen that $p = \nabla u$ is the solution to (2.2.25), it remains to show the uniqueness of this solution.

First, suppose there exists $p_1, p_2 \in V$ solutions to the variational inequality (2.2.25) then for all $q \in K$, we have:

$$a(p_1, q - p_1) \geq (F, q - p_1), \tag{2.2.30}$$
$$a(p_2, q - p_2) \geq (F, q - p_2) \ . \tag{2.2.31}$$

setting $q = p_2$ in the variational inequality for p_1 and $q = p_1$ in that of p_2 we obtain, upon adding

$$a(p_1 - p_2, p_1 - p_2) \leq 0 \ .$$

But since a is positive, then we have that $p_1 = p_2$. □

Mixed formulation

Having obtained the new formulation that is well suited for the mixed method of Raviart-Thomas, we now want to give a mixed formulation of (2.2.25). In addition to the assumptions on V, V_0, and V_g we made in the previous section, as stated earlier, we again stress the assumption here that the bilinear form

$$a(q, q) > 0 \ \text{ for all } q \in V \text{ with } q \neq 0.$$

and that the convex cone is given by (2.2.23). In addition the bilinear form $b(\cdot, \cdot)$ is assumed to be continuous on $V \times Q$, Q a Hilbert space and Λ is a closed convex cone in Q with the vertex at the origin, and $\phi \in Q^*$.

We consider the following mixed formulation of (2.2.25):

Find a pair $(p, \lambda) \in V \times \Lambda$ such that

$$a(p, q) + b(q, \lambda) = \langle F, q \rangle \qquad \forall q \in V, \tag{2.2.32}$$
$$b(p, \mu - \lambda) \leq \langle \phi, \mu - \lambda \rangle_Q \qquad \forall \mu \in \Lambda \ . \tag{2.2.33}$$

2.2.2 Existence and Uniqueness of a solution to the Mixed Problem

Provided that all the assumptions stated above hold, following [14], we develop general results about the existence and uniqueness of a solution $(p, \lambda) \in V \times \Lambda$ to the mixed problem (2.2.32-2.2.33). To this end we state the following theorem.

Theorem 2.2.3. *Suppose that there exists a constant $\beta > 0$ such that*

$$\inf_{\mu \in Q} \sup_{q \in V} \frac{b(q, \mu)}{\|q\|_V \|\mu\|_Q} \geq \beta, \ \ \mu, p \neq 0. \tag{2.2.34}$$

Then problems (2.2.25) and (2.2.32)-(2.2.33) have at most one solution. If either problem has a solution, then they both have solutions. Furthermore if (p, λ) solves (2.2.32)-(2.2.33), then p solves (2.2.25)

Proof. Let us begin by establishing the uniqueness of (2.2.32)-(2.2.33). The uniqueness of (2.2.25) has already been established from the proof to Theorem (2.2.2). If (p_1, λ_1) and (p_2, λ_2) are solutions of (2.2.32)-(2.2.33), then

$$\begin{aligned} a(p_1, q) + b(q, \lambda_1) &= \langle f, q \rangle, & (2.2.35)\\ b(p_1, \mu - \lambda_1) &\leq \langle \phi, \mu - \lambda_1 \rangle_Q, & (2.2.36)\\ a(p_2, q) + b(q, \lambda_2) &= \langle f, q \rangle, & (2.2.37)\\ b(p_2, \mu - \lambda_2) &\leq \langle \phi, \mu - \lambda_2 \rangle_Q. & (2.2.38) \end{aligned}$$

Choosing $q = p_1 - p_2$ in (2.2.35), $\mu = \lambda_2$ in (2.2.36) and $q = p_2 - p_1$ in (2.2.37), $\mu = \lambda_1$ in (2.2.38) we get

$$\begin{aligned} a(p_1, p_1 - p_2) + b(p_1 - p_2, \lambda_1) &= \langle f, p_1 - p_2 \rangle, & (2.2.39)\\ b(p_1, \lambda_2 - \lambda_1) &\leq \langle \phi, \lambda_2 - \lambda_1 \rangle_Q, & (2.2.40)\\ a(p_2, p_2 - p_1) + b(p_2 - p_1, \lambda_2) &= \langle f, p_2 - p_1 \rangle, & (2.2.41)\\ b(p_2, \lambda_1 - \lambda_2) &\leq \langle \phi, \lambda_1 - \lambda_2 \rangle_Q. & (2.2.42) \end{aligned}$$

Adding these relations gives $a(p_2 - p_1, p_2 - p_1) \leq 0$, but since a is positive, we then have $p_1 = p_2$. Now assume that (p, λ_1) and (p, λ_2) are solutions to (2.2.32) then:

$$\begin{aligned} a(p, q) + b(q, \lambda_1) &= \langle f, q \rangle, & (2.2.43)\\ a(p, q) + b(q, \lambda_2) &= \langle f, q \rangle. & (2.2.44) \end{aligned}$$

Subtracting, we obtain $b(q, \lambda_2 - \lambda_1) = 0 \ \ \forall \ \ q \in V$. Using the inf-sup condition, it follows that

$$0 = b(q, \lambda_2 - \lambda_1) \geq \beta \|q\|_V \|\lambda_1 - \lambda_2\|_Q \quad \forall \ \ q \in V, \tag{2.2.45}$$

which implies $\lambda_1 = \lambda_2$.
Does p belong to K? To answer this question we observe that $p \in V, \lambda \in Q$ satisfy (2.2.33) if and only if $p \in K$ and

$$b(p, \lambda) = \langle \phi, \lambda \rangle, \tag{2.2.46}$$

i.e , if $p \in K$, then $b(p, \mu) \leq \langle \phi, \mu \rangle$ and using (2.2.46) implies $b(p, \mu - \lambda) \leq \langle \phi, \mu - \lambda \rangle$ hence (2.2.33) is satisfied.
Conversely, if (2.2.33) holds, we choose $\mu = 0$ and $\mu = 2\lambda$ to obtain $b(p, \lambda) = \langle \phi, \lambda \rangle$ and from (2.2.33), we have then that $b(q, \mu) \leq \langle \phi, \mu \rangle \ \ \forall \mu \in \Lambda$, hence $p \in K$.

Now we consider the last part of the theorem. If $\langle p, \lambda \rangle$ satisfies (2.2.32-2.2.33), we just observed that $p \in K$. Moreover (2.2.32) and (2.2.46) gives us

$$\begin{aligned} a(p, q - p) &= \langle f, q - p \rangle - b(q - p, \lambda) \\ &= \langle f, q - p \rangle - b(q, \lambda) + b(p, \lambda) \\ &= \langle f, q - p \rangle + \langle \phi, \lambda \rangle - b(q, \lambda) \\ &\geq \langle f, q - p \rangle \text{ for all } q \in K \ . \end{aligned} \tag{2.2.47}$$

Hence $p \in K$ satisfies (2.2.3) with respect to variable p.
Conversely, let p be the solution to (2.2.3), and let the bounded linear operators $A : V \to V^*$ and $B : V \to Q^*$ be defined by:

$$\begin{aligned} \langle Au, q \rangle &= a(u, q) \text{ for all } u, q \in V, \\ \langle Bq, \mu \rangle &= b(q, \mu) \text{ for all } q \in V, \ \ \mu \in Q \ . \end{aligned} \tag{2.2.48}$$

If $z \in Z = Ker\{B\}$, i.e., $b(z, \mu) = 0 \ \forall \ z \in Z$, then $p \pm z$ belongs to K and (2.2.3) implies

$$\begin{aligned} a(p, p \pm z - p) &\geq (f, p \pm z - p), \\ a(p, z) &= (f, z) \text{ for all } z \in Z \ . \end{aligned} \tag{2.2.49}$$

This implies that $\langle Ap - f, z \rangle = 0$ for all $z \in Z$ hence $Ap - f$ belongs to the polar set Z^0 of Z (see[19]). Using the hypothesis (2.2.34), the result from Brezzi asserts that the adjoint operator B^* is an isomorphism from Q to Z^0. Therefore there exists $\lambda \in Q$ such that $B^*\lambda = f - Ap$, i.e.,

$$\begin{aligned} (B^*\lambda, q) &= \langle f - Ap, q \rangle, \\ (\lambda, Bq) &= (f, q) - (Ap, q) \ . \end{aligned}$$

and (2.2.32) is established.
To conclude we prove (2.2.33). It is important to note that $(Z^0)^*$ can be isometrically identified with the orthogonal complement $Z^\perp$ of Z. The operator B is an isomorphism from $Z^\perp$ to Q^*. Hence there exists $\bar{p} \in Z^\perp$ such that $B\bar{p} = \phi$. By definition of K, both $\bar{p}$ and $2p - \bar{p}$ are in K and by (2.2.25) we get

$$a(p, \bar{p} - p) = (f, \bar{p} - p) \ . \tag{2.2.50}$$

Inserting $q = \bar{p} - p$ into (2.2.32) gives us

$$0 = b(\bar{p} - p, \lambda) = b(\bar{p}, \lambda) - b(p, \lambda) \ . \tag{2.2.51}$$

Since $p \in K$, (2.2.51) and the equivalence (2.2.46) imply that (2.2.33) holds. Finally the next task is to show if $\lambda \in \Lambda$. Relations (2.2.25), (2.2.32) and (2.2.51) imply that for all $p \in K$,

$$0 \leq b(q - p, \lambda) = b(q, \lambda) - (\phi, \lambda).$$

Since the range of B is Q^*, we have that $\lambda \in \Lambda$. □

Application to the mixed problem

We have seen that the general theorem , assures us about existence and uniqueness of a solution to the primal problem (2.2.25) and the abstract mixed problem (2.2.32-2.2.33). As we saw earlier, the new primal formulation (2.2.22) has at most one solution. Thus under the hypothesis of Theorem (2.2.3), there exists an equivalent mixed problem
Find $(p, \lambda) = (\nabla u, u - \psi) \in H(div,\ \Omega) \times \Lambda$ such that

$$\begin{aligned} \langle p, q\rangle + \langle \lambda, \nabla.q\rangle &= -\langle \psi, \nabla.q\rangle + \langle g, q.n\rangle_{\partial\Omega} \quad for\ all\ q \in H(div\ \Omega), \quad (2.2.52) \\ \langle \nabla.p, \mu - \lambda\rangle &\leq -\langle f, \mu - \lambda\rangle \quad for\ all\ \mu \in \Lambda \ . \quad (2.2.53) \end{aligned}$$

Moreover if all the conditions of Theorem 2.2.3 are satisfied, then

Theorem 2.2.4. $(p, \lambda) = (\nabla u, u - \psi) \in H(div,\ \Omega) \times \Lambda$ *is the unique solution to the mixed problem (2.2.52-2.2.53)*

Proof. Theorem 2.2.3 is applied now with $V = H(div, \Omega), Q = L^2(\Omega),\ a(p, q) = \langle p, q\rangle$ and $b(q, \mu) = \langle \mu, \nabla.q\rangle$. As we saw earlier, the bilinear form $a(q, q) = \int q^2 > 0\ \ \forall q \neq 0$. Moreover the bilinear form b is bounded since $\mu \in L^2(\Omega)$ and $\nabla.q \in L^2(\Omega)\ \ \forall\ \ q \in H(div, \Omega)$. Now we verify the condition (2.2.34): Given $\mu \in Q$, let α satisfy:

$$\Delta\alpha = \mu \text{ in } \Omega, \qquad \alpha|_{\partial\Omega} = 0,$$

and set $q = \nabla\alpha$. Since α vanishes on $\partial\Omega$, there exists a constant γ (independent of μ) such that $\|\nabla\alpha\|_Q \leq \gamma\|\mu\|_Q$ (Fredrich's inequality). Hence we have

$$\begin{aligned} ||q||_{H,\Omega} &= \|\mu\|_Q^2 + \|\nabla\alpha\|_Q^2 \leq (1+\gamma^2)\|\mu\|_Q^2, \\ \frac{b(q,\mu)}{\|q\|_{H,\,\Omega}} &= \frac{\langle\mu, \nabla.q\rangle}{\|q\|_{H,\,\Omega}} = \frac{\|\mu\|_Q^2}{\|q\|_{H,\,\Omega}} \geq \beta\|\mu\|_Q, \end{aligned}$$

where $\beta = (1+\gamma^2)^{-1/2}$.
Since relation (2.2.34) holds, by Theorem (2.2.3), the mixed formulation, (2.2.52)-(2.2.53) has a unique solution (p, λ) and $p = \nabla u$ implies (using Greens formula) that

$$\langle p, q\rangle = -\langle u, \nabla.q\rangle + \langle g, q.n\rangle_{\partial\Omega} \quad \forall q \in H(div, \Omega).$$

Similarly

$$\langle p, q\rangle + \langle \lambda + \psi, \nabla.q\rangle = \langle g, q.n\rangle_{\partial\Omega}.$$

So combining both implies

$$\langle \lambda + \psi - u, \nabla.q\rangle = 0 \quad \text{for all } q \in H(div, \Omega). \tag{2.2.54}$$

Choosing $q = \nabla\alpha$, where $\nabla.q = \Delta\alpha = \mu = \lambda + \psi - u$,(2.2.54) implies $\lambda = u - \psi$. □

So far we have seen that the obstacle problem can be formulated as a minimization problem. This minimization problem is equivalent to the variational inequality. Further we have established that if $u \in H^2(\Omega) \cap \tilde{K}$, then the variational inequality can be expressed as a linear complementarity problem. We developed a new primal formulation (2.2.22) of this problem (LCP) that is suitable for the mixed method of Raviart-Thomas . We have further shown that a unique solution to (2.2.22) is also a unique solution to the mixed problem ([2.2.52]- [2.2.53]) provided, the `inf-sup` condition is satisfied and the bilinear form $a(q, q)$ is positive. In order to solve this mixed problem, we approximate it using finite dimensional subspaces which we introduce in the next chapter.

Chapter 3

Approximation of the Obstacle problem

In this chapter, we are interested in finding an approximative solution to the obstacle problem. We shall prove existence and uniqueness of a solution to this finite dimensional problem. Furthermore, we shall study the behavior of the approximative solutions, as the parameter of discretization tends to zero. In the next chapter, we shall see how we can solve the discrete problem that we are going to develop.

3.1 Introduction

Let us recall from the previous chapter the new primal formulation of the obstacle problem :
Find $p \in K$ such that

$$\langle p, q-p\rangle \geq -\langle \psi, \nabla.(q-p)\rangle + \langle g, (q-p).n\rangle_{\partial\Omega} \quad \text{for all } q \in K, \tag{3.1.1}$$

where

$$K = \{q \in H(div, \Omega) : \nabla.q + f \leq 0 \ \ a.e \text{ in } \ \Omega\}. \tag{3.1.2}$$

This formulation was found to be equivalent to the mixed problem :
Find $(p, \lambda) = (\nabla u, u - \psi) \in H(div,\ \Omega) \times \Lambda$ such that

$$\begin{aligned} \langle p, q\rangle + \langle \lambda, \nabla.q\rangle &= -\langle \psi, \nabla.q\rangle + \langle g, q.n\rangle_\Gamma \quad for\ all\ q \in H(div\ \Omega), & (3.1.3)\\ \langle \nabla.p, \mu - \lambda\rangle &\leq -\langle f, \mu - \lambda\rangle \quad for\ all\ \mu \in \Lambda. & (3.1.4) \end{aligned}$$

The existence and uniqueness of the solution (p, λ) to (3.1.3-3.1.4) and its equivalence with (3.1.1-3.1.2) were established in the previous chapter. We now want to discretize our domain Ω in order to solve these problems. We remark here that the accuracy of a finite element discretization is determined by the approximability of the exact solution by the finite element subspace and the stability of the discretization [1]. These two properties, together with implementational issues, furnish the major factors for the construction and evaluation of the finite element spaces to be used. Stability is automatic for coercive methods so that the finite element space can be chosen on the basis of approximation and ease of implementation alone.
It is important to remark here that the bilinear form $a(p, q) = \langle p, q \rangle$ is not coercive and stability is by no means automatic. In fact elements chosen without due regard to stability will prove to be unstable[1]. This then leads us to use elements that satisfy the Brezzi condition ([5], [14],[1]). Various techniques have been developed for design of stable mixed finite elements [1]. Polynomials whose normal components are continuous across the inter-element boundaries is the family of many mixed finite elements satisfying this property. An example among this family is the lowest order Raviart-Thomas elements which we shall use in the discretization of our space $V(\equiv H(div, \Omega))$.

3.2 Discretization of the obstacle problem

3.2.1 Definition of spaces

Let $\mathcal{T}_h$ be a regular family of decomposition of Ω into triangles [7]. The parameter of discretization h corresponds to the largest length of the edge of such a triangle. We define our finite dimensional spaces $V^h \subset V$ and $Q^h \subset L^2(\Omega)$ as

$$V^h = \{\underline{\phi} \in V : \underline{\phi}|_T = (a_1 + bx, a_2 + by),\ a_1,\ a_2,\ b \in \mathbb{R}\}, \tag{3.2.1}$$

$$Q^h = \{\tau : \tau \text{ is piecewise constant on each triangle}\}. \tag{3.2.2}$$

We further set $\Lambda^h = \Lambda \cap Q^h$:

$$\Lambda^h = P_0(\mathcal{T}) := \{\mu^h \in L^2(\Omega) : T \in \mathcal{T}, \mu^h|_T \in P_0(T), \mu^h \geq 0\}. \tag{3.2.3}$$

Remark 3.2.1. Notice that a vector in V^h has a normal component which is constant on each edge. Moreover if E is an edge in the triangulation (see figure 3.1) shared by two triangles T_+ and T_-, then

$$\forall E \in \mathcal{E}_\Omega, \quad [\underline{\phi}]_E.\nu_E = 0, \tag{3.2.4}$$

where $\mathcal{T}$ is a regular triangulation , $\mathcal{E}_\Omega$ is the set of all interior edges, and $[\underline{\phi}]_E.\nu_E := \underline{\phi}|_{T_+} - \underline{\phi}|_{T_-}$ along E denotes the jump of $\underline{\phi}$ across the edge $E = T_+ \cap T_-$ shared by the two neighboring elements $T+$ and $T-$ in $\mathcal{T}$. ν_E is the outer normal of E, whose orientation is such that it equals the outer normal of T_+ (and hence points into T_-).
This space V^h is the lowest order Raviart-Thomas $RT_0(\mathcal{T})$[20]. The continuity of the

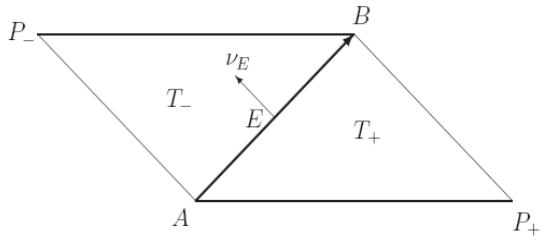


Figure 3.1: Two neighboring triangles T_+ and T_- that share the edge $E = \partial T+ \cap \partial\, T-$ with initial node A and end node B and unit normal ν_E. The orientation of ν_E is such that it equals the outer normal of T_+ (and hence points into T_-).

normal components on the boundaries reflects the conformity $RT_0(\mathcal{T}) \subset H(div, \Omega)$.

3.2.2 Approximation of the new primal and mixed formulations

We now approximate the new primal variational inequality (3.1.1 - 3.1.2) and the mixed system (3.1.3 - 3.1.4) using the above mentioned finite dimensional subspaces. The discrete new primal problem then reads :
Find $p^h \in K^h$ such that

$$\langle p^h, q^h - p^h\rangle \geq -\langle \psi, \nabla.(q^h - p^h)\rangle + \langle g, (q^h - p^h).n\rangle_{\partial\Omega} \quad \text{for all } q^h \in K^h, \tag{3.2.5}$$

where

$$K^h = \{q^h \in V^h : \nabla.q^h + f \leq 0 \ \ a.e \text{ in } \ \Omega\}. \tag{3.2.6}$$

Following the general results discussed in [20], we now state and prove the following Theorem.

Theorem 3.2.1. *There exits a unique solution p^h to the problem (3.2.5). Moreover if (p, λ) satisfies (3.1.3)-(3.1.4), then for all $q^h \in K^h$ and $\mu^h \in \Lambda^h$ we have*

$$\|p-p^h\|_0^2 \leq \langle p-p^h, p-q^h\rangle + \langle \lambda-\mu^h, \nabla.p+f\rangle + \langle \lambda-\mu^h, \nabla.(p^h-q^h)\rangle + \langle \mu^h, \nabla.(p-q^h)\rangle. \tag{3.2.7}$$

Proof. Since V^h is finite dimensional , $\langle \cdot, \cdot \rangle$ is coercive on V^h. Therefore there exists a unique solution to (3.2.5) by classical arguments. Now we focus our attentional to the error bound.

$$\begin{aligned} \langle p-p^h, p-p^h\rangle &= \langle p-p^h, p-q^h+q^h-p^h\rangle \\ &= \langle p-p^h, p-q^h\rangle + \langle p-p^h, q^h-p^h\rangle \\ &= \langle p-p^h, p-q^h\rangle + \langle p, q^h-p^h\rangle - \langle p^h, q^h-p^h\rangle. \end{aligned} \tag{3.2.8}$$

But $\langle p, q^h-q^h\rangle = \langle g, (q^h-p^h).n\rangle - \langle \lambda+\psi, \nabla.(p^h-q^h)\rangle$. If we substitute this in 3.2.8, we obtain

$$\begin{aligned} \langle p-p^h, p-p^h\rangle &= \langle p-p^h, p-q^h\rangle + \langle g, (q^h-p^h).n\rangle + \langle \lambda+\psi, \nabla.(p^h-q^h)\rangle - \langle p^h, q^h-p^h\rangle \\ &= \langle p-p^h, p-q^h\rangle + \langle \lambda, \nabla.(p^h-q^h)\rangle + \\ &\quad \underbrace{\langle g, (q^h-p^h).n\rangle - \langle \psi, \nabla.(q^h-p^h)\rangle - \langle p^h, q^h-p^h\rangle}_{\leq 0} \\ &\leq \langle p-p^h, p-q^h\rangle + \langle \lambda, \nabla.(p^h-q^h)\rangle. \end{aligned}$$

Note: By using the fact that $\langle \nabla.p, \lambda\rangle = \langle -f, \lambda\rangle$ and that $\langle \nabla.p, \mu^h\rangle \leq \langle -f, \mu^h\rangle$ we have:

$$\begin{aligned} \langle \lambda, \nabla.(p^h-q^h)\rangle &= \langle \nabla.(p^h-q^h), \lambda-\mu^h\rangle + \langle \nabla.(p-q^h), \mu^h\rangle \\ &\quad + \langle \nabla.p, \lambda-\mu^h\rangle + \langle f, \lambda-\mu^h\rangle + \langle \nabla.p^h, \mu^h\rangle + \langle f, \mu^h\rangle \\ &\leq \langle \nabla.(p^h-q^h), \lambda-\mu^h\rangle + \langle \nabla.(p-q^h), \mu^h\rangle \\ &\quad + \langle \nabla.p, \lambda-\mu^h\rangle + \langle f, \lambda-\mu^h\rangle. \end{aligned}$$

Hence (3.2.7) is clear from the above arguments ∎

Now we have seen the existence of a unique solution to the discrete new primal formulation (3.2.5-3.2.6). Theorem [2.2.3] in section [2.2.2] assures us that there exists

an equivalent discrete mixed formulation to the problem (3.2.5-3.2.6) which reads: Find $(p^h, \lambda^h) \in V^h \times \Lambda^h$:

$$\begin{aligned} \langle p^h, q^h\rangle + \langle \lambda^h, \nabla.q^h\rangle &= -\langle \psi, \nabla.q^h\rangle + \langle g, q^h.n\rangle_\Gamma \quad \text{for all } q^h \in V^h, \quad (3.2.9)\\ \langle \nabla.p^h, \mu^h - \lambda^h\rangle &\leq -\langle f, \mu^h - \lambda^h\rangle \qquad \text{for all } \mu^h \in \Lambda^h, \quad (3.2.10) \end{aligned}$$

provided the discrete inf-sup condition is satisfied. Moreover this discrete mixed formulation also has a unique solution. To this end we state the following theorem

Theorem 3.2.2. *Suppose that there exits a constant $\beta > 0$ (independent of h) such that*

$$\inf_{\mu^h \in Q^h} \sup_{q^h \in V^h} \frac{b(\mu^h, \nabla.q^h)}{\|q^h\|_{H,\Omega}\|\mu^h\|_0} \geq \beta. \tag{3.2.11}$$

Then there exists a unique pair (p^h, λ^h) solving the problem (3.2.9)-(3.2.10). If (p^h, λ^h) solves (3.2.9)-(3.2.10), then p^h solves (3.2.5). Moreover if (p, λ) satisfies (3.1.3)-(3.1.4), there exists a constant C (independent of h) such that

$$\|\lambda - \lambda^h\|_0 \leq C\{\|p - p^h\|_0 + \inf_{\mu^h \in Q^h} \|\lambda - \mu^h\|_0\}. \tag{3.2.12}$$

Proof. Uniqueness is an immediate consequence of Theorem(2.2.3). Now we prove the error estimate. We note that

$$\|\lambda - \lambda^h\|_0 \leq \|\lambda - \mu^h\|_0 + \|\mu^h - \lambda^h\|_0. \tag{3.2.13}$$

The task now is to estimate each of these terms in this equation. Let us begin by estimating the term $\|\mu^h - \lambda^h\|_0$.

From (3.2.9) and (3.1.3), one obtains

$$\begin{aligned} \langle \lambda^h - \mu^h, \nabla.q^h\rangle &= \langle g, q^h.n\rangle - \langle \psi, \nabla.q^h\rangle - \langle p^h, q^h\rangle - \langle \mu^h, \nabla.q^h\rangle, \quad (3.2.14)\\ \langle g, q^h.n\rangle_\Gamma &= \langle p, q^h\rangle + \langle \lambda + \psi, \nabla.q^h\rangle. \quad (3.2.15) \end{aligned}$$

Substituting (3.2.15) in (3.2.14) we obtain

$$\langle \lambda^h - \mu^h, \nabla.q^h\rangle = \langle p - p^h, q^h\rangle + \langle \lambda - \mu^h, \nabla.q^h\rangle. \tag{3.2.16}$$

We divide this equation (3.2.16) by $\|q^h\|_H$,and take the supremum we obtain

$$\sup_{q^h \in V^h} \frac{\langle \lambda^h - \mu^h, \nabla.q^h\rangle}{\|q^h\|_H} = \sup_{q^h \in V^h} \frac{\langle p - p^h, q^h\rangle}{\|q^h\|_H} + \sup_{q^h \in V^h} \frac{\langle \lambda - \mu^h, \nabla.q^h\rangle}{\|q^h\|_H}. \tag{3.2.17}$$

But (3.2.11) implies that

$$\beta\|\lambda^h - \mu^h\|_0 \leq \sup_{q^h \in V^h} \frac{\langle \lambda^h - \mu^h, \nabla.q^h \rangle}{\|q^h\|_H}. \tag{3.2.18}$$

So we have

$$\beta\|\lambda^h - \mu^h\|_0 \leq \sup_{q^h \in V^h} \frac{\langle p - p^h, q^h \rangle}{\|q^h\|_H} + \sup_{q^h \in V^h} \frac{\langle \lambda - \mu^h, \nabla.q^h \rangle}{\|q^h\|_H}, \tag{3.2.19}$$

$$\beta\|\lambda^h - \mu^h\|_0 \leq \sup_{q^h \in V^h} \frac{\langle p - p^h, q^h \rangle}{\|q^h\|_H} + \|\lambda - \mu^h\|_0, \tag{3.2.20}$$

Now we remark that $H(div, \Omega)$ is continuously imbedded in $L^2(\Omega)$, that is, $H(div, \Omega) \subset L^2(\Omega)$ and there exists a constant γ such that $\|q\|_0 \leq \|q\|_H$ for all $q \in H(div, \Omega)$. $a(\cdot, \cdot) = \langle p, q \rangle$ is continuous on $H(div, \Omega)$, which then implies

$$\frac{\langle p - p^h, q^h \rangle}{\|q^h\|_H} \leq \|p - p^h\|_0. \tag{3.2.21}$$

We obtain

$$\beta\|\lambda^h - \mu^h\|_0 \leq \|p - p^h\|_0 + \|\lambda - \mu^h\|_0. \tag{3.2.22}$$

We take the infimum over all $\mu^h \in Q^h$. Combining (3.2.22) and (3.2.13) will then yield the required result. □

Having obtained the above estimates, we now embark on the task of estimating $\|\lambda - \lambda^h\|_0$ and $\|p - p^h\|_0$. If one uses (3.2.12) to estimate $\|\lambda - \lambda^h\|_0$, we find that we need information about $\|p - p^h\|_0$ and $\inf_{\mu^h \in Q^h} \|\lambda - \mu^h\|_0$ to estimate $\|p - p^h\|_0$. We will use (3.2.7). To this end, we state the following theorem from [14] that hold in case of a polygonal domain.

Theorem 3.2.3. *If Ω is polygonal , $(\psi, g) \in H^2(\Omega)$, $f \in L^2(\Omega)$ and (p, λ) is a solution to (3.1.3-3.1.4) then the error in piecewise constant approximation (p^h, λ^h) generated by (3.2.9)-(3.2.10) satisfies*

$$\|p - p^h\|_0 = O(h) \tag{3.2.23}$$

$$\|\lambda - \lambda^h\|_0 = O(h) \tag{3.2.24}$$

Proof. Raviart and Thomas [20] constructed $p^I \in V^h$ such that

$$\langle 1, (p - p^I).n \rangle_E = 0 \tag{3.2.25}$$

for edges E in the triangulation where n is the unit normal along E. Moreover the following estimates hold

$$\begin{aligned} \|p - p^I\|_0 &= O(h) \text{ if } p \in H^1(\Omega), && (3.2.26)\\ \|\nabla.(p - p^I)\|_0 &= O(h) \text{ if } \nabla.p \in H^1(\Omega), && (3.2.27)\\ \langle \mu^h, \nabla.(p - p^I)\rangle &= 0 \ forall\ \mu^h \in Q^h. && (3.2.28) \end{aligned}$$

Combining the relations $\Lambda^h \subset \Lambda$, (3.2.28) and $\nabla.p + f \leq 0$, we see that $p^I \in K^h$ i.e

$$\langle \mu^h, \nabla.p^I + f\rangle = \langle \mu^h, \nabla.p + f\rangle \leq 0,$$

for all $\mu^h \in \Lambda^h$. Letting λ^I denote the L^2 projection of $\lambda = u - \psi$ onto Q^h, observe that $\lambda^I \geq 0$. Therefore we can apply error bound (3.2.7) using $(q^h, \mu^h) = (p^I, \lambda^I)$ Now we estimate term by term in (3.2.7). Let us start with

$\langle \lambda - \mu^h, \nabla.(p^h - q^h)\rangle$ and $\langle \mu^h, \nabla.(p - q^h)\rangle$

By (3.2.28) and the orthogonality relation $(\lambda - \lambda^I) \perp Q^h$, we have

$$\begin{aligned} \langle \lambda - \lambda^I, \nabla.(p^h - p^I)\rangle = 0, && (3.2.29)\\ \langle \lambda^I, \nabla.(p - p^I)\rangle = 0. && (3.2.30) \end{aligned}$$

Looking at the second term $\langle p - p^h, p - q^h\rangle$, since $p = \nabla u \in H^1(\Omega)$, then using (3.2.26) we have

$$\begin{aligned} \langle p - p^h, p - p^I\rangle &\leq \|p - p^h\|_0 \|p - p^I\|_0 && (3.2.31)\\ &= ch\|p - p^h\|_0. && (3.2.32) \end{aligned}$$

To conclude we estimate the term $\langle \lambda - \mu^h, \nabla.p + f\rangle = \langle \lambda - \lambda^I, \nabla.p + f\rangle$.
We consider two cases, i.e., $u > \psi$ and $u = \psi$:
We first observe that if $u > \psi$ on triangle T, then $\nabla.p + f = 0$ on T and in this case $\langle \lambda - \lambda^I, \nabla.p + f\rangle = 0$.
Now if we consider the case $u = \psi$, i.e.,
If $\lambda = 0$ on a subset of T with positive measure, then by the contraction property [1] of projections and by classical interpolation theory, one obtains

$$\|\lambda^I\|_{0,T} \leq \|\lambda\|_{0,T} \leq ch^2\|\lambda\|_{2,T}. \quad (3.2.33)$$

If Z is the region of triangles where λ vanishes on a subset of positive measure, then (3.2.33) gives us:

$$\begin{aligned} \langle \lambda - \lambda^I, \nabla.p + f\rangle = \langle \lambda - \lambda^I, \nabla.p + f\rangle_Z &\leq \|\nabla.p + f\|_{0,Z}\|\lambda - \lambda^I\|_{0,Z}.\\ &\leq 2\|\nabla.p + f\|_{0,Z}\|\lambda\|_{0,Z} = O(h^2). \end{aligned} \quad (3.2.34)$$

[1] $\|p_\Lambda(\lambda) - p_\Lambda(0)\|_0 \leq \|\lambda - 0\|_0$

Combining all the results from (3.2.29), (3.2.31) and (3.2.34), Theorem (3.2.1) yields

$$\|p - p^h\|_0^2 \leq ch\|p - p^h\|_0 + O(h^2). \tag{3.2.35}$$

Hence $\|p - p^h\|_0 = O(h)$. Since Raviart and Thomas established the existence of a constant $\beta > 0$ satisfying (3.2.11) uniformly in h, then if one replaces μ^h by λ^I in equation (3.2.12) we obtain the required estimate $\|\lambda - \lambda^h\|_0 = O(h)$ □

Brezzi, Hager and Raviart [20] assert that to obtain an optimal error estimate with linear elements, we need an assumption about the behavior of λ near the free boundary. Given a constant $c > 0$, one defines the sets

$$F^h = \{(x, y) : 0 < \lambda(x, y) < ch^2\} \tag{3.2.36}$$

and Ω^h=the union of all triangles $T \subset \Omega^h$ such that $T \cap F^h \neq \emptyset$. They assume that the measure $(\Omega_F^h) = O(h)$. Moreover in applications, this condition is always satisfied since the free boundary usually has finite length and the normal derivative $\partial^2\lambda/\partial n^2$ along the free boundary is positive.
To conclude this subsection, we have seen that applying the finite elements of Raviart-Thomas to our models, gives a $O(h)$ L^2 convergence of the function values and gradients for piecewise constant elements. Moreover for the obstacle problem, Brezzi et.al. [14] further established that:

Theorem 3.2.4. *If Ω is polygonal , $(u, \psi) \in W^{2,\infty}(\Omega)$, $f \in L^\infty(\Omega)$, $u \in W^{s,p}(\Omega)$ for all $1 < p < \infty$ and $s < 2 + 1/p$, and (p, λ) is a solution to (3.1.3-3.1.4) then the error in piecewise linear approximation (p^h, λ^h) generated by (3.2.9)-(3.2.10) satisfies*

$$\|p - p^h\|_0 = O(h^{\frac{3}{2}-\epsilon}) \tag{3.2.37}$$
$$\|\lambda - \lambda^h\|_0 = O(h^{\frac{3}{2}-\epsilon}) \tag{3.2.38}$$

for any $\epsilon > 0$.

The proof to this theorem is can be found in [14]. Its important to remark here that although this result was proved for a polygonal domain, it still holds if $\partial\Omega$ is C^2. See ([14] Theorem 4.6) for the proof.

3.2.3 Discretization of the Mixed Formulation

In this subsection we shall focus on the computational details of the mixed formulation. In the next chapter, we shall then see how to solve the resulting systems of equations. Following our approximation approach in the previous subsection, we choose finite dimensional subspaces $V^h \subset H(div, \Omega)$ and $Q^h \subset L^2(\Omega)$ and consider the approximate problem:
Find $(p^h, \lambda^h) \in V^h \times \Lambda^h$ such that

$$\begin{aligned} \langle p^h, q^h \rangle + \langle \lambda^h, \nabla . q^h \rangle &= -\langle \psi, \nabla . q^h \rangle + \langle g, q^h . n \rangle_\Gamma \quad \text{for all } q^h \in V^h, \quad (3.2.39) \\ \langle \nabla . p^h, \mu^h - \lambda^h \rangle &\leq -\langle f, \mu^h - \lambda^h \rangle \quad \text{for all } \mu^h \in \Lambda^h. \quad (3.2.40) \end{aligned}$$

Choice of basis for V^h

If we assume that $\phi_1, \phi_2, \cdots, \phi_N$ be the edge oriented basis for the N-dimensional space $V^h \subseteq RT_0(\mathcal{T})$. Here N denotes the number of edges in the triangulation $\mathcal{T}(dim(V^h))$.

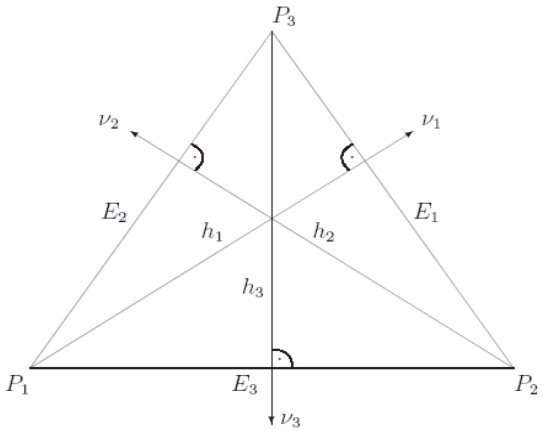


Figure 3.2: Triangle T with vertices (P_1, P_2, P_3) (ordered counterclockwise) and opposite edges E_1, E_2, E_3 of lengths $|E_1|, |E_2|, |E_3|$, respectively. The heights h_1, h_2, h_3 depicted

Following [2], we give the following definitions

Definition 3.2.1. (Local definition of ϕ_E). Let E_1, E_2, E_3 be the edges of a triangle T opposite to its vertices P_1, P_2, P_3, respectively(see figure 3.2), and let ν_{E_j} denote the unit normal vector of E_j chosen with a global fixed orientation while ν_j denotes the outer unit normal of T along E_j. Define

$$\phi_{E_j}(x) = \sigma_j \frac{|E_j|}{2|T|}(x - P_j) \quad \text{for } j = 1, 2, 3 \text{ and } x \in T, \tag{3.2.41}$$

where $\sigma_j = \nu_j.\nu_{E_j}$ is +1 if ν_{E_j} points outward and otherwise -1. $|E_j|$ is the length of E_j, and $|T|$ is the area of T,

$$2|T| = det(P_2 - P_1, P_3 - P_1) = det \begin{pmatrix} P_1 & P_2 & P_3 \\ 1 & 1 & 1 \end{pmatrix} \tag{3.2.42}$$

(with the 3×3-matrix that consists of the 2×3 matrix of the three vectors $P_1, P_2, P_3 \in \mathbb{R}^2$ plus three ones in the last row).

Suppose now, we have more than one element, then there holds the following definition

Definition 3.2.2. (Global definition of ϕ_E). Given an edge $E \in \mathcal{E}$ there are either two elements T_+ and T_- in $\mathcal{T}$ with the joint edge $E = \partial T_+ \cap \partial T_-$ (see Fig: 3.1) or there is exactly one element $T_+ \in \mathcal{T}$ with $E \subset \partial T_+$. Then if $T_\pm = conv(E \cup \{P_\pm\})$ for the vertex $P_\pm$ opposite to E set

$$\phi_E(x) := \begin{cases} \pm\frac{|E|}{2|T_\pm|}(x - P_\pm) & \text{for } x \in T_\pm \\ 0 & \text{elsewhere.} \end{cases} \tag{3.2.43}$$

We now state the following Lemma without proof. The details of the proof can be found in [2]

Lemma 3.2.5. *There hold*

(a) $\phi_E.\nu_E = \begin{cases} 0 & along\ (\cup\mathcal{E}) \setminus E, \\ 1 & alongE; \end{cases}$

(b) $\phi_E \in H(div, \Omega)$*;*

(c) $(\phi_E : E \in \mathcal{E})$ *is a basis of* $RT_0(\mathcal{T})$

(d) $div\phi_E = \begin{cases} \pm\frac{|E|}{|T_\pm|} & on\ T_\pm, \\ 0 & elsewhere. \end{cases}$

From now, unless specified, we shall abbreviate $\phi_j := \phi_{E_j}$.

Choice of basis for Q^h

If we assume that $\tau_1, \tau_2, \cdots, \tau_L$ is a basis of the L-dimensional space Q^h. In our case we choose τ_k as a characteristic function of the element $\mathcal{T}_l$ in $\mathcal{T}$(piecewise constant), where $L = card(\mathcal{T})$.

System of equations and inequalities

Having set up the basis for our spaces, then equation (3.2.39) can be written as

$$\langle p^h, \phi_i \rangle + \langle \lambda^h + \psi, \nabla.\phi_i \rangle = \langle g, \phi_i.n \rangle_\Gamma \qquad \text{for all } \phi_i \in V^h.$$

Each $p^h \in V^h$ and $\mu^h \in Q^h$ can be uniquely represented by vectors $\underline{p} = (p_k),\ p_k \in \mathbb{R}$ and $\underline{\lambda} = (\lambda_i),\ \lambda_i\ \in \mathbb{R}$ respectively. Thus we can make a series expansion of p_h and λ_h in terms of ϕ_k and τ_j such that

$$p^h = \sum_{k=1}^{N} p_k \phi_k \quad (\phi_k \in V^h) \qquad \text{and} \qquad \lambda^h = \sum_{j=1}^{L} \lambda_j \tau_j.$$

We obtain

$$\int_\Omega \sum_{k=1}^{N} p_k \phi_k \phi_i dx + \int_\Omega \sum_{j=1}^{L} \lambda_j \tau_j \nabla.\phi_i dx = \int_\Gamma g\phi_i.nds - \int_\Omega \psi \nabla.\phi_i dx, \qquad (3.2.44)$$

which can be rewritten as

$$\sum_{k=1}^{N} p_k \int_\Omega \phi_k \phi_i dx + \sum_{j=1}^{L} \lambda_j \int_\Omega \tau_j \nabla.\phi_i dx = \int_\Gamma g\phi_i.nds - \int_\Omega \psi \nabla.\phi_i dx.$$

If we further denote by $\underline{g}$, the vector induced by the right hand side (3.2.44) such that

$$\underline{g} = (g_i), \quad (g_i) = \int_\Gamma g\phi_i.nds - \int_\Omega \psi \nabla.\phi_i dx,$$

and the matrices A, B^T such that

$$A = (A_{ik}), \qquad A_{ik} = \int_\Omega \phi_k \phi_i dx,$$
$$B^T = (B^T_{ij}), \qquad B^T_{ij} = \int_\Omega \tau_j \nabla.\phi_i dx.$$

We can finally cast (3.2.44) into a system of linear equations

$$A\underline{p} + B^T\underline{\lambda} = \underline{g}. \tag{3.2.45}$$

Similarly for (3.2.40) we have

$$\begin{aligned}
\int_\Omega \left(\nabla . \sum_{k=1}^{N} p_k \phi_k + f \right) (\mu^h - \lambda^h)\, dx &\leq 0 \\
\sum_{j=1}^{L} (\mu_j - \lambda_j) \int_\Omega \left(\nabla . \sum_{k=1}^{N} p_k \phi_k + f \right) \tau_j \, dx &\leq 0 \\
\sum_{j=1}^{L} \sum_{k=1}^{N} (\mu_j - \lambda_j) p_k \int_\Omega \nabla . \phi_k \tau_j dx &\leq -\sum_{j=1}^{L} (\mu_j - \lambda_j) \int_\Omega f \tau_j dx.
\end{aligned} \tag{3.2.46}$$

If we again denote by $\underline{f}$, the vector induced by the right hand side (3.2.46) such that

$$\underline{f} = (f_i), \qquad (f_i) = \int_\Omega f \tau_j dx,$$

and $\mu^h \in Q^h$ such that it has a vector representation $\underline{\mu}$ such that

$$\underline{\mu} = (\mu_k), \qquad (\mu_k) \geq 0,$$

then (3.2.46) is equivalent to

$$(\underline{\mu} - \underline{\lambda})^T B \underline{p} \leq (\underline{\mu} - \underline{\lambda})^T \underline{f}. \tag{3.2.47}$$

Putting together all the ideas that we have gathered so far in this subsection, we can see that the discrete mixed formulation can be rewritten as a system of equations and inequalities for $\underline{p} \in \mathbb{R}^N, \underline{\lambda} \in \mathbb{R}^L, \underline{\lambda} \in \underline{\Lambda}$

$$A\underline{p} + B^T\underline{\lambda} = \underline{g}, \tag{3.2.48}$$

$$(\underline{\mu} - \underline{\lambda})^T B \underline{p} \leq (\underline{\mu} - \underline{\lambda})^T \underline{f} \quad \forall \underline{\mu} \in \underline{\Lambda}, \tag{3.2.49}$$

where

$$\underline{\Lambda} = \{\underline{\mu} : \quad \mu_k \geq 0, \quad k \leq L\}.$$

3.2.4 Discretization Matrices and Right Hand Sides

In the previous subsection, we have developed a system of equations and inequalities. We now want to see how we can calculate each of these terms involved in these systems. It is important to note that each ϕ_i defined in the previous subsection has support over at most two elements and each τ_j has support over one element, thus when a regular triangulation $\mathcal{T}$ has been generated for the domain Ω, one can calculate the stiffness matrix A, the matrix B^T and the right hand sides f and g as a sum over all elements i.e

$$A_{ik} = \sum_{T\in\mathcal{T}} \int_T \phi_k \phi_i dx, \qquad B^T_{ij} = \sum_{T\in\mathcal{T}} \int_T \nabla.\phi_i dx \quad (\tau_j = 1), \tag{3.2.50}$$

and the right hand sides

$$g_i = \int_\Gamma g\phi_i.nds - \sum_{T\in\mathcal{T}} \int_T \psi\nabla.\phi_i dx, \tag{3.2.51}$$

$$f_j = \sum_{T\in\mathcal{T}} \int_T f dx. \tag{3.2.52}$$

Definition 3.2.3. Let the local stiffness matrices $A_T, D_T \in \mathbb{R}^{3\times 3}$ and $B^T \in \mathbb{R}^{3\times 1}$ be defined by

$$(A_T)_{ik} : = \int_T \phi_k \phi_i dx \quad for\ i,k = 1,2,3\ , \tag{3.2.53}$$

$$(D_T) : = diag\left(\int_T div\phi_1 dx, \int_T, div\phi_2 dx, \int_T div\phi_3 dx\right), \tag{3.2.54}$$

$$(B^T)_i = \int_T div\phi_i dx. \tag{3.2.55}$$

Given (3.2.53)-(3.2.54) and the matrices

$$M := \begin{pmatrix} 2 & 0 & 1 & 0 & 1 & 0 \\ 0 & 2 & 0 & 1 & 0 & 1 \\ 1 & 0 & 2 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 & 0 \\ 0 & 1 & 0 & 1 & 0 & 2 \end{pmatrix} \in \mathbb{R}^{6\times 6} \text{ and } N := \begin{pmatrix} 0 & P_1 - P_2 & P_1 - P_3 \\ P_2 - P_1 & 0 & P_2 - P_3 \\ P_3 - P_1 & P_3 - P_2 & 0 \end{pmatrix} \in \mathbb{R}^{6\times 3},$$

Then there holds

$$A_T = \frac{1}{48|T|} D_T^T N^T M N D_T. \tag{3.2.56}$$

We follow the proof in [2] and try to adapt it to our problem.

Proof. Let $\lambda_1, \lambda_2, \lambda_3$ denote the barycentric coordinates in the triangle T of Figure (3.2), with properties

$$\begin{aligned} 1 &= \lambda_1 + \lambda_2 + \lambda_3, & (3.2.57)\\ x &= \lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3. & (3.2.58) \end{aligned}$$

Then an affine function (3.2.41) reads

$$\begin{aligned} \phi_i(x) &= \frac{\sigma_{E_i}|E_i|}{2|T|}\left(\lambda_1 P_1 + \lambda_2 P_2 + \lambda_3 P_3 - P_i(\lambda_1 + \lambda_2 + \lambda_3)\right) & (3.2.59)\\ &= \frac{\sigma_{E_i}|E_i|}{2|T|}\sum_{l=1}^{3}\lambda_l(P_l - P_i). & (3.2.60) \end{aligned}$$

Similarly, we obtain

$$\phi_k(x) = \frac{\sigma_{E_k}|E_k|}{2|T|}\sum_{m=1}^{3}\lambda_m(P_m - P_k). \tag{3.2.61}$$

Then

$$\begin{aligned} (A_T)_{ik} = \int_T \phi_k\phi_i dx &= \frac{\sigma_{E_i}|E_i|\sigma_{E_k}|E_k|}{4|T|^2}\int_T \sum_{l=1}^{3}\lambda_l(P_l - P_i).\sum_{m=1}^{3}\lambda_m(P_m - P_k)dx\\ &= \frac{\sigma_{E_i}|E_i|\sigma_{E_k}|E_k|}{4|T|^2}\sum_{l=1}^{3}\sum_{m=1}^{3}(P_l - P_i)P_m - P_k)\int_T \lambda_l\lambda_m dx. \end{aligned}$$

Since $\int_T \lambda_l\lambda_m = \frac{|T|}{12}(1+\delta_{lm})$, this yields

$$\begin{aligned} (A_T)_{ik} &= \frac{\sigma_{E_i}|E_i|\sigma_{E_k}|E_k|}{4|T|^2}\sum_{l=1}^{3}\sum_{m=1}^{3}(P_l - P_i)(P_m - P_k)\left(\frac{|T|}{12}(1+\delta_{lm})\right)\\ &= \frac{\sigma_{E_i}|E_i|}{48|T|}\left(\left(\sum_{l=1}^{3}(P_l - P_i)\right).\left(\sum_{m=1}^{3}(P_m - P_k)\right) + \sum_{l=1}^{3}(P_l - P_i)(P_l - P_k)\right)\sigma_{E_k}|E_k|. \end{aligned} \tag{3.2.62}$$

Direct calculations of $(A_T)_{ik}$ for each $i, k = 1, 2, 3$ and $P_{i,k} = (P_{i,k,x},\ P_{i,k,y})^T$ verify the relation (3.2.56). □

Handling of the Right Hand Sides

We now analyze each term for the expression

$$g_i = \int_\Gamma g\phi_i.nds - \int_\Omega \psi\nabla.\phi_i dx, \tag{3.2.63}$$

and later the term $f_j = \int_\Omega f\tau_j dx = \int_{T_l} f dx$, where the integral is over each element T_l with center of gravity z_{T_l}, $l = 1, \cdots, L$.
Let us start with $\int_\Gamma g\phi_i.n$.
Since the normal components of the test functions ϕ are zero or equal to one along the edge $E \in \mathcal{E}_D$ of number j with mid-point (x_m, y_m) (See lemma 3.2.5), a simple one point integration reads

$$\int_\Gamma g\phi_i.nds = \int_E g ds \approx g(x_m, y_m)|E|, \tag{3.2.64}$$

where $|E|$ is the length of the edge, and $\mathcal{E}_D$ is a set of all edges on the boundary.
Now let us turn our attention to the second term

$$\int_\Omega \psi\nabla.\phi_i dx. \tag{3.2.65}$$

This term can be re-written as a sum over all elements in the triangulation $\mathcal{T}$ i.e

$$\int_\Omega \psi\nabla.\phi_i dx = \sum_j \int_{T_l} \psi\nabla.\phi_i dx. \tag{3.2.66}$$

Next approximate the integral $\int_{T_l} \psi\nabla.\phi_i dx$ using the mid-point rule, such that if z_{T_l} is the value in the centroid of each triangular element, we can rewrite equation (3.2.66) as

$$\sum_{j=1}^{N} \psi(z_{T_j}) \int_{T_l} \nabla.\phi_i dx = \sum_{j=1}^{N} \psi(z_{T_j}) B_{ij}^T, \tag{3.2.67}$$

where B_{ij}^T is the global matrix defined in (3.2.50). Finally we approximate the term $f_j = \int_\Omega f\tau_j dx = \int_{T_l} f dx$. Numerical realization of this term also in the simplest case involves one point numerical quadrature. Based on results from [3], one can evaluate the following integral as

$$\int_{T_l} f dx = \frac{|T_l|}{3} f(z_l), \tag{3.2.68}$$

where z_{T_l} is the centroid of triangle T_l. It is important to remark here that the system matrices that we have developed so far are sparse [17], and thus easy to implement them. In the next chapter we shall see how the system (3.2.48 - 3.2.49) can be solved.

Chapter 4

Solution of discretized Problem

In this chapter we want to develop an algorithm for solving the system of equations and inequalities that we obtained in the previous chapter. We shall also analyze the convergence of this algorithm to the exact solution of the discrete problem
Find $\underline{p} \in \mathbb{R}^N, \underline{\lambda} \in \mathbb{R}^L, \underline{\lambda} \in \underline{\Lambda}$ such that

$$\begin{aligned} A\underline{p} + B^T\underline{\lambda} &= g, && (4.0.1)\\ (\underline{\mu} - \underline{\lambda})^T B\underline{p} &\leq (\underline{\mu} - \underline{\lambda})^T f \quad \forall \underline{\mu} \in \underline{\Lambda}, && (4.0.2) \end{aligned}$$

where

$$\underline{\Lambda} = \{\underline{\mu} : \quad \mu_k \geq 0, \quad k \leq L\}.$$

From now onwards, for the sake of simplicity of the notation, we shall omit the underlines and denote the vector induced by some $\lambda_h \in \Lambda_h$ by λ, similarly $\underline{p}$ as p . Using equation (4.0.1), we have

$$p = -A^{-1}(B^T\lambda - g). \tag{4.0.3}$$

Inserting this into our inequality system (4.0.2) , we get

$$\begin{aligned} -(\mu - \lambda)^T B(A^{-1}(B^T\lambda - g)) &\leq (\mu - \lambda)^T f,\\ -(\mu - \lambda)^T BA^{-1}B^T\lambda &\leq (\mu - \lambda)^T(f - BA^{-1}g). \end{aligned}$$

This leaves us with a problem for λ only. Setting the **Schur complement** $S = BA^{-1}B^T$ and right hand side $h = BA^{-1}g - f$, we obtain

$$(\mu - \lambda)^T(S\lambda - h) \geq 0 \quad \forall \ \ \mu \in \Lambda. \tag{4.0.4}$$

The Schur complement S is symmetric as A is symmetric. We also have that A is positive definite and the rank of $B = L$, therefore also S is positive definite and there exists numbers s_1, s_2 such that

$$\langle S\mu, \mu\rangle \geq s_1\|\mu\|^2, \tag{4.0.5}$$

$$(S\mu, \lambda) \leq s_2\|\mu\|\|\lambda\|, \tag{4.0.6}$$

where $(\cdot,\cdot)$ and $\|\cdot\|$ denote the Euclidean norm and scalar product in $\mathbb{R}^L$. Due to this, equation [4.0.4] is equivalent to minimization problem

$$\begin{aligned}\mathcal{F}(\lambda) &= \min_{\mu\in\Lambda}\mathcal{F}(\mu) \quad \text{where}\\ \mathcal{F}(\mu) &= \frac{1}{2}\mu^T S\mu - h^T\mu.\end{aligned} \tag{4.0.7}$$

We denote the gradient of the energy $\mathcal{F}$ by r such that

$$r(\mu) = \nabla(\mathcal{F}(\mu)) = S\mu - h.$$

In the following we present the algorithm for minimization of $\mathcal{F}$ over Λ. Once we have computed λ by such a method, we can then determine u from $u = \lambda + \psi$. In the following, we will need the projection operator P_Λ, which acts from R^L onto Λ. For $\mu \in \mathbb{R}^L$, $P_\Lambda(\mu)$ is the closest on the convex cone Λ in the L^2 norm defined by

$$(\mu, \lambda)_M = (M\mu, \lambda), \tag{4.0.8}$$

where

$$M = (M_{ij}), \qquad M_{ij} = \int_\Omega \tau_i\tau_j dx. \tag{4.0.9}$$

M is a diagonal mass matrix with each of the entries on the diagonal equivalent to the area of triangle say T_l in the mesh. Consequently the associated norm in vector notation is then given by

$$\|\lambda\|_M = \sqrt{(M\mu, \lambda)}. \tag{4.0.10}$$

Lemma 4.0.6. *The components of $P_\Lambda(\mu)$ are given by:*

$$P_\Lambda(\mu)_k = \max(\mu_k, 0).$$

Proof. $P_\Lambda(\mu)$ is the closest on the convex cone Λ to some μ , thus we have

$$||P_\Lambda(\mu) - \mu||_M \leq ||\mu_h - \mu||_M \qquad \forall \mu_h \in \Lambda. \tag{4.0.11}$$

If we let

$$\mu = \sum \mu_i \tau_j,$$

we set

$$P_\Lambda(\mu) = \sum \max(\mu_i, 0)\tau_j.$$

If we square both sides of (4.0.11) and divide by 2 both sides, we obtain

$$\frac{1}{2}||P_\Lambda(\mu) - \mu||_M^2 \leq \frac{1}{2}||\mu_h - \mu||_M^2 \qquad \forall \mu_h \in \Lambda. \tag{4.0.12}$$

Now we define the functional

$$\begin{aligned} J(\mu_h) &= \frac{1}{2}||\mu_h - \mu||_M^2 \qquad \forall \mu_h \in \Lambda \\ &= \frac{1}{2}(\mu_h - \mu, \mu_h - \mu) \\ &= \frac{1}{2}(\mu_h, \mu_h) - (\mu, \mu_h) + \frac{1}{2}(\mu, \mu). \end{aligned} \tag{4.0.13}$$

We want to minimize the functional (4.0.13). Since μ is constant, it does not affect the minimizer of (4.0.13), thus one can minimize instead

$$\min_{\mu \in \Lambda} \tilde{J}(\mu_h) = \frac{1}{2}(\mu_h, \mu_h) - (\mu, \mu_h). \tag{4.0.14}$$

This minimization is equivalent to the variational inequality (see subsection (2.1.2) proposition [2.1.1])

$$(\bar{\mu}^h, \mu_h - \bar{\mu}^h) \geq (\mu, \mu_h - \bar{\mu}^h), \tag{4.0.15}$$

where $\bar{\mu}^h$ is the solution to (4.0.14). (4.0.15) can be written as

$$(\mu - \bar{\mu}^h, \mu_h - \bar{\mu}^h) \leq 0. \tag{4.0.16}$$

By our claim, we set

$$\bar{\mu}^h = \sum \max(\mu_i, 0)\tau_j,$$

and show that with this choice, (4.0.16) holds. Consider

$$\left(\sum(\mu_i - \max(\mu_i, 0))\tau_j, \sum(\tilde{\mu}_i - \max(\mu_i, 0))\tau_j\right). \tag{4.0.17}$$

If $\mu_i \geq 0$, $\max(\mu_i, 0) = \mu_i$ and (4.0.17) will be zero. On the other hand if $\mu_i \leq 0$, then (4.0.17) becomes $(\mu_i, \tilde{\mu}_i) \leq 0$ (since $\tilde{\mu}_i \in \Lambda$) which completes the proof. □

4.1 Uzawa's method

Uzawa's method is an iterative method for solving systems of equations and inequalities of the form (4.0.1-4.0.2). Starting with some initial guess, λ can be computed from the constrained minimization problem (4.0.7). Then equation (4.0.3) allows the computation of p. A classical method of this type is the Uzawa algorithm [11], which relies on an exact solver for (4.0.1) and a Jacobi-like iteration for the constrained minimization problem (4.0.7). Below we outline the algorithm for this method.

Algorithm for Uzawa's method

Algorithm 1 Uzawa

1: give some initial value $\lambda^{(0)}$
2: k=0
3: **repeat**
4: compute $p^{(k+1)}$ from the equation:
$Ap^{(k+1)} + B^T\lambda^{(k)} := g$
5: $\lambda_*^{(k+1)} := \lambda^{(k)} + \alpha M^{-1}(Bp^{(k+1)} - f)$
$\{\alpha$ is some given, small positive constant$\}$
6: take $\lambda^{(k+1)}$ as the projection of $\lambda_*^{(k+1)}$ onto Λ :
$\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)})$
7: k=k+1
8: **until** $\max\{\|p^{(k+1)} - p^{(k)}\|/\|p^{(k+1)}\|, \quad \|\lambda^{(k+1)} - \lambda^{(k)}\|/\|\lambda^{(k+1)}\|\} \leq \varepsilon$

For an appropriate choice of α one can show that the sequence $(p^{(k)}, \lambda^{(k)})$ converges to (p, λ) of the reduced obstacle problem (3.2.48-3.2.49). By using the Schur compliment S and the right hand side h we introduced before, we can eliminate $p^{(k+1)}$ from the construction of $\lambda^{(k+1)}$. This means that we compute $\lambda^{(k+1)}$ directly from $\lambda^{(k)}$ without finding $p^{(k+1)}$. From the first step of the k^{th} iteration , we obtain

$$p^{(k+1)} = -A^{-1}(B^T\lambda^{(k)} - g).$$

Inserting this into the formula for $\lambda_*^{(k+1)}$ we get

$$\begin{aligned} \lambda_*^{(k+1)} &= \lambda^{(k)} + \alpha(-BA^{-1}B^T\lambda^{(k)} + BA^{-1}g - f) \\ &= \lambda^{(k)} - \alpha(S\lambda^{(k)} - h). \end{aligned} \tag{4.1.1}$$

This is a fixed-parameter first order Richardson iteration [24] applied to the system

$$S\lambda = h. \tag{4.1.2}$$

Therefore we can reformulate the algorithm 1. In each iteration , we compute $\lambda^{(k+1)}$ from $\lambda^{(k)}$ in two steps.

Algorithm 2 Reformulation of the Uzawa

1: give some initial value $\lambda^{(0)}$
2: k=0
3: **repeat**
4: $\lambda_*^{(k+1)} := \lambda^{(k)} - \alpha M^{-1}(S\lambda^{(k)} - h)$
$\{\alpha$ is some given, small positive constant$\}$
5: take $\lambda^{(k+1)}$ as the projection of $\lambda_*^{(k+1)}$ on Λ :
$\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)})$
6: k=k+1
7: **until** $\|\lambda^{(k+1)} - \lambda^{(k)}\| \leq \varepsilon\|\lambda^{(k+1)}\|$
8: $p^{(k+1)} = -A^{-1}(B^T\lambda^{(k+1)} - f)$

Convergence of Uzawa's Method

The choice of the parameter α determines the convergence of the sequence $(p^{(k)}, \lambda^{(k)})$ to (p, λ) of the reduced obstacle problem (3.2.48-3.2.49). Good choices of the scalar α are determined from observation of equation (4.1.2). Our goal now is to analyze the convergence of Uzawa's Method. Before we analyze the convergence of the Uzawa method, we shall first briefly recall some of the properties of the projection operator for our discussion. Recall that for $z \in Q$, we define $P_\Lambda(z)$, called the projection of z onto Λ as the closest point in Λ to z.
The projection function has several interesting properties, which we define in the following theorem:

Theorem 4.1.1. *For Λ a convex, closed and non-empty set, we have*

1. *$\forall z \in Q$ there is a unique optimum of (4.0.7) which we call $x^* = P_\Lambda(z)$*
2. *x^* is the unique point in Λ such that $\forall x \in \Lambda, (z - x^*)(x - x^*) \leq 0$*

3. Projection is non-expansive:

$$\forall x, y \in Q, ||P_\Lambda(x) - P_\Lambda(x)|| \leq ||x - y||$$

in other words P_Λ is Lipschitz with constant 1.

We give the figures to illustrate these properties. The proofs of these theorems

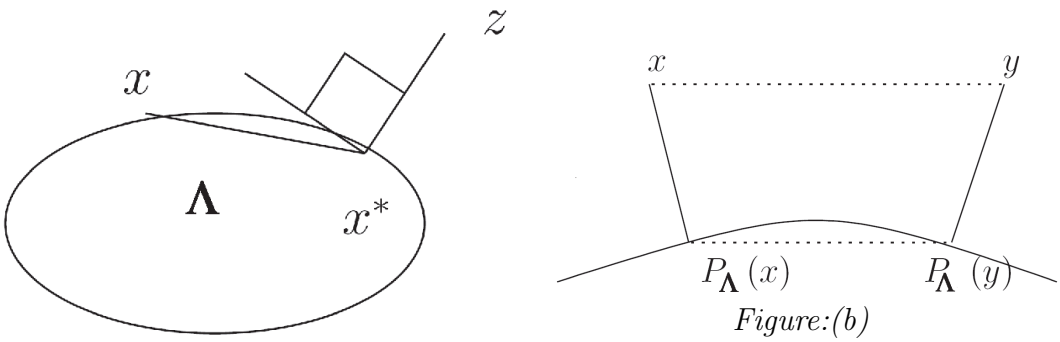


Figure:(a) *Figure:(b)*

Figure 4.1: Explanatory pictures for the projection theorem. In general, $z - x^*$ must form an obtuse angle with $x - x^*$ for any $x \in \Lambda$ (left), The projection also has the property that projections x^* and y^* of points x and y are at least as close together as x and y are (right).

can be found in [25]. We shall make use of the last property in order to discuss the convergence of Uzawa method.

We now state some theorem that assures us about the convergence of Uzawa's Method.

Theorem 4.1.2. *Let A, B be defined as above, and (p, λ) be a solution to the system (4.0.1-4.0.2). Let s_1, s_2 denote the smallest and the largest eigenvalues of $M^{-1}S$, and let $(p^{(k)}, \lambda^{(k)})$ be defined by Uzawa's method. Then there exists a positive constant $\bar{\alpha} > 0$ such that for each choice $\alpha \in (0, \bar{\alpha})$ there holds*

$$p^{(k)} \to p, \quad \lambda^{(k)} \to \lambda.$$

Proof. Recall from Algorithm 1 that

$$\lambda_*^{(k+1)} : = \lambda^{(k)} - \alpha M^{-1}(S\lambda^{(k)} - h) \tag{4.1.3}$$

$$= (I - \alpha S)\lambda^{(k)} + \alpha h. \tag{4.1.4}$$

One then defines

$$\lambda^{(k+1)} := P_\Lambda \left((I - \alpha S)\lambda^{(k)} + \alpha M^{-1} h \right). \tag{4.1.5}$$

Similarly , if $\lambda \in \Lambda$ is a solution to the mixed problem, then one has

$$\lambda = P_\Lambda \left((I - \alpha M^{-1}S)\lambda + \alpha M^{-1}h\right). \tag{4.1.6}$$

Then the errors satisfy

$$\lambda^{(k+1)} - \lambda = P_\Lambda \left((I - \alpha M^{-1}S)\lambda^{(k)} + \alpha M^{-1}h\right) - P_\Lambda \left((I - \alpha M^{-1}S)\lambda + \alpha M^{-1}h\right). \tag{4.1.7}$$

By the contraction property of projections, we have

$$\|\lambda^{(k+1)} - \lambda\|_M \leq \|(I - \alpha M^{-1}S)\|_M \|\lambda^{(k)} - \lambda\|_M. \tag{4.1.8}$$

Let us denote $\|\lambda^{(k+1)} - \lambda\|_M$ by e^{k+1} and $\|\lambda^{(k)} - \lambda\|_M$ by e^{k+1} , then the errors satisfy

$$e^{k+1} \leq \|(I - \alpha M^{-1}S)\|_M e^k. \tag{4.1.9}$$

Therefore we have,

$$(e^{k+1}, e^{k+1}) \leq ((I - \alpha M^{-1}S)e^k, (I - \alpha M^{-1}S)e^k). \tag{4.1.10}$$

Since S is symmetric, it follows that $\rho(I - \alpha M^{-1}S) = ||I - \alpha M^{-1}S||_M$, so that the error norm satisfies

$$||e^{k+1}||_M \leq \rho((I - \alpha M^{-1}S)||e^k||_M. \tag{4.1.11}$$

If let s_1, s_2 denote the smallest and the largest eigenvalues of $M^{-1}S$, respectively, and μ_i denote the eigen values of $(I - \alpha M^{-1}S)$. Then

$$1 - \alpha s_2 \leq \mu_i \leq 1 - \alpha s_1,$$

and the Uzawa's Method is convergent provided $\rho(I - \alpha M^{-1}S) < 1$ i.e $0 < \alpha < 2/s_2$. As p^k depends continuously on λ^k, we can conclude that $p^k \to p$. □

We will take $\rho(I - \alpha M^{-1}S)$ as the measure of effectiveness of the algorithm and refer to it as the convergent factor.

4.2 Inexact Uzawa Method

One problem with the algorithm (1) is that the bulk of computational effort is spent in the computation of A^{-1} at each step of the iteration. For many application this is an expensive operation. The inexact Uzawa replaces the action of A^{-1} by a preconditioner $\hat{A}$. This preconditioner is a linear operator $\hat{A} : V \to V$ which is symmetric

and positive definite. Moreover $\hat{A}$ should be relatively cheap to invert i.e the computational cost of $\hat{A}^{-1}$ should be comparable to that of A and not A^{-1}. The inexact algorithm then reads

Algorithm 3 Inexact Uzawa Algorithm

1: given some initial value $\lambda^{(0)} \in \Lambda$, $p^{(0)} \in V$
2: k=0
3: **repeat**
4: compute $p^{(k+1)}$,$\lambda^{(k+1)}$ from the equations:
$p^{k+1} = p^k + \hat{A}^{-1}(g - Ap^k - B^T\lambda^k)$
$\lambda_*^{(k+1)} = \lambda^k + \alpha M^{-1}(Bp^{k+1} - f)$
$\{\alpha$ is some given, small positive constant$\}$
5: take $\lambda^{(k+1)}$ as the projection of $\lambda_*^{(k+1)}$ onto Λ :
$\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)})$
6: k=k+1
7: **until** $\max\{\|p^{(k+1)} - p^{(k)}\|/\|p^{(k+1)}\|, \quad \|\lambda^{(k+1)} - \lambda^{(k)}\|/\|\lambda^{(k+1)}\|\} \leq \varepsilon$

One step of the inexact Uzawa Algorithm involves an evaluation of each of the operators $A,\ B,\ B^T,\ \hat{A}^{-1}$ and M^{-1}. The complete analysis of convergence of algorithm3 for the case of equations can be found in [11], [4], [26].

Preconditioning

The inexact Uzawa method above can be seen as preconditioned Richardson method

$$\hat{\mathcal{K}_1}\begin{pmatrix} p_{k+1} - p_k \\ \lambda_*^{(k+1)} - \lambda_k \end{pmatrix} = \begin{pmatrix} g \\ f \end{pmatrix} - \mathcal{K}\begin{pmatrix} p_k \\ \lambda_k \end{pmatrix}, \tag{4.2.1}$$

which an additional projection step for $\lambda^{(k+1)}$ defined as

$$\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)}).$$

With a preconditioner $\hat{\mathcal{K}_1} = \begin{pmatrix} \hat{A} & 0 \\ B & -\hat{S} \end{pmatrix}$, and the matrix $\mathcal{K}$ is given by $\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}$.

A class of symmetric preconditioners

We now want to discuss a class of symmetric preconditioners. These preconditioners were discussed in detail in [26] for a situation where we have a saddle point system

with equations. It was found that one can factorize matrix $\mathcal{K}$ in the following form

$$\mathcal{K} = \begin{pmatrix} A & 0 \\ B & I \end{pmatrix} \begin{pmatrix} A^{-1} & 0 \\ 0 & -C \end{pmatrix} \begin{pmatrix} A & B^T \\ 0 & I \end{pmatrix}, \tag{4.2.2}$$

which motivates the use of the preconditioner

$$\mathcal{K} = \begin{pmatrix} \hat{A} & 0 \\ B & I \end{pmatrix} \begin{pmatrix} \hat{A}^{-1} & 0 \\ 0 & -\hat{C} \end{pmatrix} \begin{pmatrix} \hat{A} & B^T \\ 0 & I \end{pmatrix}. \tag{4.2.3}$$

This is equivalent to the following procedure for the case of equations

$$\hat{p}^{k+1} = p^k + \hat{A}^{-1}(g - Ap^k - B^T\lambda^k), \tag{4.2.4}$$

$$\lambda^{k+1} = \lambda^k + \hat{S}^{-1}(B\hat{p}^{k+1} - f), \tag{4.2.5}$$

$$p^{k+1} = \hat{p}^{k+1} - \hat{A}^{-1}B^T(\lambda^{k+1} - \lambda^k). \tag{4.2.6}$$

This method can be viewed as inexact Uzawa algorithm with additional correction step for p. Actually we can call it a symmetric inexact Uzawa algorithm. We can then modify this algorithm so that it can be used to solve the mixed system (4.0.1-4.0.2). The modified algorithm then reads

Algorithm 4 Symmetric Inexact Uzawa Algorithm

1: given some initial value $\lambda^{(0)} \in \Lambda$, $p^{(0)} \in V$, $\mathcal{A}^{(0)}$, $\hat{S} = M$
2: k=0
3: **repeat**
4: compute $p^{(k+1)}$,$\lambda^{(k+1)}$ from the equations:
$\hat{p}^{k+1} = p^k + \hat{A}^{-1}(g - Ap^k - B^T\lambda^k)$
$\lambda_*^{(k+1)} = \lambda^k + \alpha\hat{S}^{-1}(B\hat{p}^{k+1} - f)$
take $\lambda^{(k+1)}$ as the projection of $\lambda_*^{(k+1)}$ onto Λ :
$\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)})$
$p^{k+1} = \hat{p}^{k+1} - \hat{A}^{-1}(B^T\lambda^{(k+1)} - \lambda^k)$
5: k=k+1
6: **until** $\max\{\|p^{(k+1)} - p^{(k)}\|/\|p^{(k+1)}\|, \quad \|\lambda^{(k+1)} - \lambda^{(k)}\|/\|\lambda^{(k+1)}\|\} \leq \varepsilon$

If we choose $\hat{A} = A$ and α is some properly chosen constant, the above algorithm can then be seen as the classical Uzawa algorithm. However, when A is singular, it cannot be inverted and the Schur complement does not exist. In this case one possible way

of dealing with the system is by argumentation, for example by replacing A by $\hat{A} = A + B^T M^{-T} B$ where M is a mass matrix. Such kind of preconditioner has been found to be effective in simulation of incompressible flow problems, Maxiwell equations in mixed form (see [13]and references there in for details). For this particular choice of $\hat{A}$, we shall study a mixed problem from a Poisson equation which after discretisation can be written in matrix form.
Find $p \in \mathbb{R}^N, \lambda \in \mathbb{R}^L$ such that

$$Ap + B^T\lambda = g, \tag{4.2.7}$$

$$Bp = f, \tag{4.2.8}$$

with $g_i = \int_{\partial\Omega} g\phi_i.n$ and A, B and f are as defined in chapter (3) subsection (3.2.3). Due to the presence of inequality constraints in the mixed system of the obstacle problem, we shall need to modify algorithm (4) such that a variable preconditioner can be used. This preconditioner depends of the set of active indices $\mathcal{A}$:

$$\mathcal{A} = \{i \quad : \quad \lambda_*^{k+1}(i) > 0\}. \tag{4.2.9}$$

To this end, we introduce the following algorithm.

Algorithm 5 Symmetric inexact Uzawa Algorithm with a variable preconditioner

given some initial value $\lambda^{(0)} \in \Lambda$, $p^{(0)} \in V$, $\mathcal{A}^{(0)}$, $\hat{S} = M$
k=0
repeat
  compute $p^{(k+1)}$,$\lambda^{(k+1)}$ from the equations:
  $\hat{p}^{k+1} = p^k + (\hat{A}^k)^{-1}(g - Ap^k - B^T\lambda^k)$
  $\lambda_*^{(k+1)} = \lambda^k + \alpha\hat{S}^{-1}(B\hat{p}^{k+1} - f)$
  take $\lambda^{(k+1)}$ as the projection of $\lambda_*^{(k+1)}$ onto Λ :
  $\lambda^{(k+1)} := P_\Lambda(\lambda_*^{(k+1)})$
  $p^{k+1} = \hat{p}^{k+1} - (\hat{A}^k)^{-1}B^T(\lambda^{(k+1)} - \lambda^k)$
  Compute $\mathcal{A}^{(k+1)} = \{i \quad : \quad \lambda^{k+1}(i) > 0\}$.
  $\hat{A}^{(k+1)} = A + \sum_{i\in\mathcal{A}^{(k+1)}} b_i^T M_i^{-1} b_i$.
  k=k+1
until $\max\{\|p^{(k+1)} - p^{(k)}\|/\|p^{(k+1)}\|, \quad \|\lambda^{(k+1)} - \lambda^{(k)}\|/\|\lambda^{(k+1)}\|\} \leq \varepsilon$

Based on this algorithm, we shall thus study three interesting cases,

- The case when $\hat{A} = A$, i.e. $\mathcal{A}^{(k+1)} = \emptyset$.
- The case when $\hat{A} = A + B^T M^{-1} B$. In this case we assume $\mathcal{A}^{(k+1)}$ is fully indexed such that $\sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_i^{-1} b_i = B^T M^{-1} B$. This represents a situation where we have only equations and no inequalities and thus we shall investigate the performance of this case.
- The case where we choose $\hat{A} = A + \sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_i^{-1} b_i$, such that the preconditioner $\hat{A}$ is dynamically changing during each step of iteration due to the change of the index set $\mathcal{A}$.

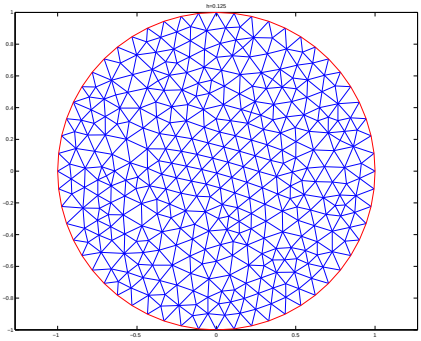
Chapter 5

Numerical Experiments

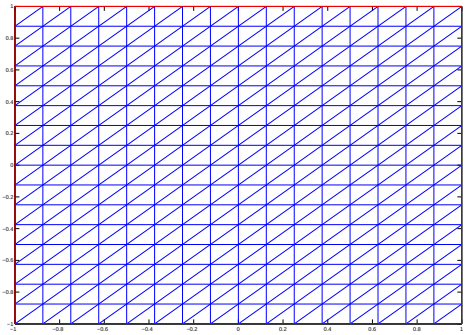
In this chapter, we solve the obstacle problem using the Uzawa algorithm presented in the previous chapter. We prescribe the obstacle using Matlab inline functions. We take different kinds of surfaces representing our obstacle and we ensure that each of these surfaces satisfy the required conditions that we discussed in the introductory chapter. For this problem, we approximate Ω by a triangular mesh and use the preexisting Matlab pdetool for mesh generation and refinement. The data describing the triangulation consists of a list of coordinates for the R node locations p (an $R \times 2$ array of real coordinates), a list of L triangles (an $L \times 3$ array of indices into p), and a list of all edges on the boundary on which we prescribe the Dirichlet boundary conditions. All this data is stored in 3 matrices namely , nodes2element, which is a sparse matrix describing the number of an element as a function of its 2 vertices, nodes2edge which is a symmetric sparse matrix of dimension $card(N)$ (N total number of edges), that describes number of edges, and element2edge which stores information of an initial and end node of an edge E shared by any two elements in the triangulation. For example, from this matrix, we can obtain information about the edges in the interior and on the boundary of the domain Ω. (We remark here that these names are not standard in literature but were choosen for the sake of experiments).

Choice of mesh

We use both structured and unstructured mesh as shown in figure (5.1). These meshes are generated by a Matlab pdetool. The unstructured mesh is constructed using a Delaunay[1] type algorithm. The structured triangular mesh consist of a triangulation in some particular direction as shown in Figure(5.1 (structured mesh)) .



Unstructured mesh



structured mesh

Figure 5.1: meshing of domain Ω

We consider the unstructured mesh on a circular domain to analyze the order of convergence of the mixed method of Raviart-Thomas because it is then easy to construct the analytic solution to this domain which we discuss in section (5.3). In addition, due to the availability of the exact solution on this domain, we also identify the free boundary of the obstacle problem on this domain Ω. To study the convergence of Uzawa's method as the paramenter of discretization becomes small, we use the structured mesh on a square domain $(-1,1) \times (-1,1)$ because of a good sparsity pattern of matrix A on this mesh.

[1]Delaunay triangulation or Delone triangularization for a set P of points in the plane is a triangulation DT(P) such that no point in P is inside the circumcircle of any triangle in DT(P). Delaunay triangulations maximize the minimum angle of all the angles of the triangles in the triangulation; they tend to avoid "sliver" triangles. The triangulation was invented by Boris Delaunay in 1934. [wikipedia]

5.1 Identification of free boundary

In this section we want to identify the boundary of the coincidence set where the obstacle and the membrane touch each other.

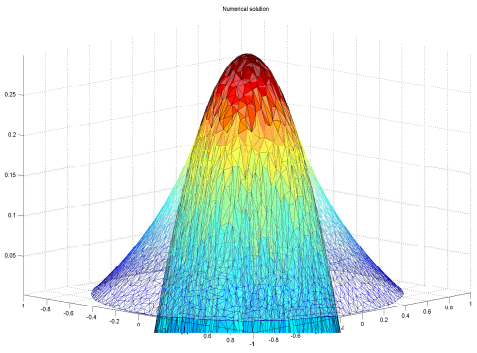


Figure:(a)

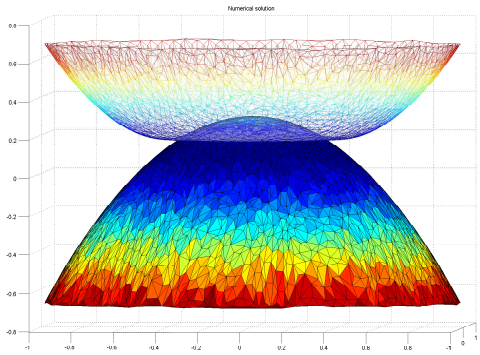


Figure:(b)

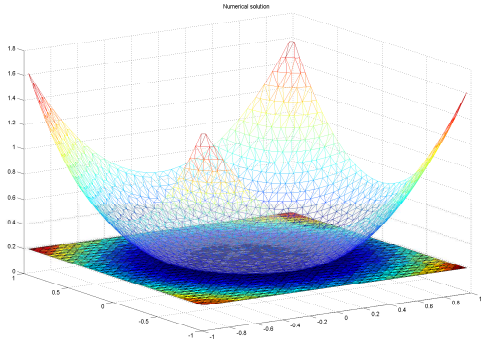


Figure:(c)

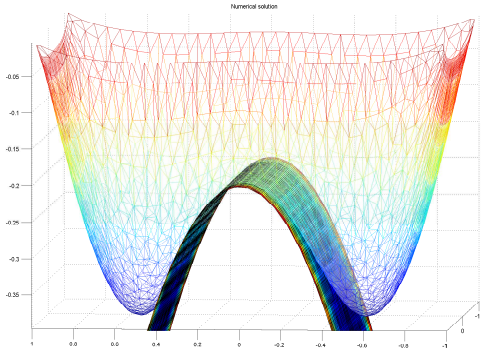


Figure:(d)

Figure 5.2: Membrane above an Obstacle (different cases)

Several cases of the membrane above an obstacle are illustrated in figure (5.2) . In table below we prescribe the conditions that each of these cases satisfy

Figure	Obstacle	f	g
a	$\psi = -x^2 - y^2 + 0.3$	0	0
b	$\psi = -x^2 - y^2 + 0.3$	-4	0.7
c	$\psi = 0.2$	-4	$x^2 + y^2 + 0.3$
d	$\psi = -x^2$	-4	0

In each of these cases, the membrane is constrained to lie above an obstacle ψ.
As a simple case, we consider a case where the exact solution is known. Consider the problem

$$-\Delta u \geq -4 \quad \text{in} \quad \Omega \tag{5.1.1}$$

$$u \geq -x^2 - y^2 + 0.3, \quad \text{in} \quad \Omega \tag{5.1.2}$$

$$(-\Delta u + 4)(u + x^2 + y^2 - 0.3) = 0 \quad \text{in} \quad \Omega \tag{5.1.3}$$

$$u = 0.7 \quad \text{on} \quad \partial\Omega \tag{5.1.4}$$

on the disc of radius 1 centered at the origin i.e

$$\Omega = \{(x, y) | x^2 + y^2 < 1\} \tag{5.1.5}$$

subject to the constraint $u \geq \psi$ where

$$\psi(x, y) = -x^2 - y^2 + 0.3. \tag{5.1.6}$$

That is we suppose that the membrane is attached at a point y=0.7 and is loaded by a force $f = -4$. In this case the problem is fully radial and $u = u(r)$. Thus

$$\Delta u = u_{rr} + \frac{1}{r} u_r.$$

And if $u > \psi$ then u solves

$$-\Delta u = -4.$$

Hence we have that

$$u_{rr} + \frac{1}{r} u_r = 4, \tag{5.1.7}$$

from which we obtain

$$u(r) = r^2 + C_1 \ln(r) + C_2. \tag{5.1.8}$$

If the free boundary is at position a then we seek a, C_1 and C_2 satisfying

$$u(a) = \psi(a), \quad u'(a) = \psi'(a), \quad u(1) = 0.7. \tag{5.1.9}$$

It is then clear that $0 < a < 1$. Using relations (5.1.9), we obtain

$$a^2 + C_1 \ln a + C_2 = -a^2 + 0.3, \tag{5.1.10}$$

$$2a + \frac{C_1}{a} = -2a. \tag{5.1.11}$$

From which we obtain $C_1 = -4a^2$ and $C_2 = -2a^2 + 0.3 + 4a^2 \ln a$. Hence

$$u(r) = r^2 - 4a^2 \ln r - 2a^2 + 0.3 + 4a^2 \ln a. \tag{5.1.12}$$

Now using the fact that $u(1) = 0.7$, this implies that $C_2 = -0.3$ and hence we obtain a scalar nonlinear equation for a

$$-2a^2 + 0.3 + 4a^2 \ln a = -0.3. \tag{5.1.13}$$

A quick plot[2] of (5.1.13) gives the solution is near 0.3 (see figure [5.3]). Now when we solve (5.1.13) analytically, we obtain $a = 0.2953$. Thus to see the convergence of

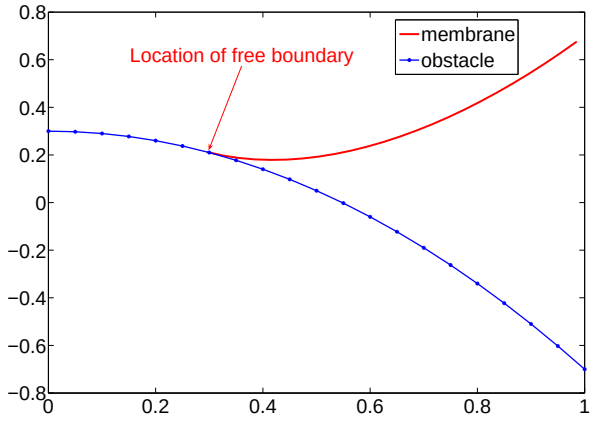


Figure 5.3: Location of free boundary

the finite element solution to the exact location of the boundary, we run the program for computing the obstacle problem for triangulations with decreasing values of step size h_k starting from 0.5. By using the assumption the free boundary is circular,

[2] a=0:.01:1;plot(a,0.3+4*a.*a.*log(a),a,2*a.*a-0.3);

we estimate it by obtaining all indices which we denote by "nogap"[3] where $u = \psi$. The x and y coordinates corresponding to each of these indices is obtained and the maximum value r_h:

$$r_h = \max \sqrt{x^2 + y^2},$$

is obtained. In Table 5.1 we report the results obtained.

step size h_k	0.5	0.25	0.125	0.0625	0.0313
$max(rh(rh < 1))$	0.3702	0.3236	0.3130	0.3049	0.3001
error (abs(r-rh))	0.0749	0.0283	0.0177	0.0096	0.0048

Table 5.1: Location of free boundary and error

From this table, we see that as step size h_k is made smaller, we have convergence to the exact location of the free boundary. In figure 5.4, we display the analytic and numerical solution for step size $h_k = \frac{1}{32}$.

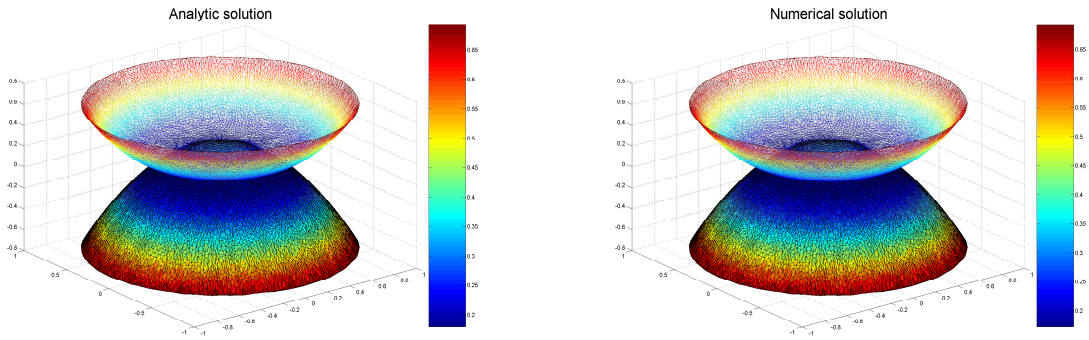


Figure:(a) *Figure:(b)*

Figure 5.4: Analytic and numerical solution

Further to support our results, we plot in figure (5.5), the free boundary estimate with iteration for several values of h_k. This graph shows that for each value of h_k, the sequence converge to a constant value after some fixed number of iterations. Similar treatment for the computation of free boundary can also be seen in [6].

[3] $nogap = (uh == obs(W)); rh = sqrt(W(nogap, 1)^2 + W(nogap, 2)^2); max(rh(rh < 1))$

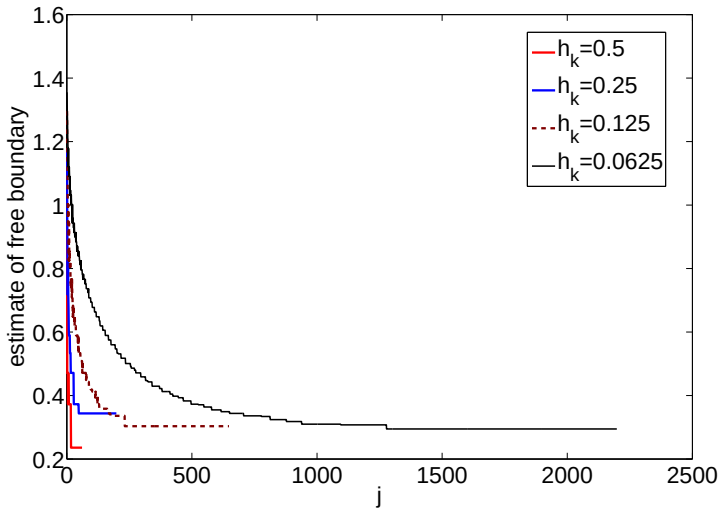


Figure 5.5: Convergence to the exact location of the free boundary

Remark 5.1.1. The mixed finite element solution is piecewise constant (figure 5.6(a)) and thus to obtain its continuous representation , we considered the solution at the centroid of each of the triangular elements and consequently made an interpolation to obtain a smooth solution (figure 5.6 (b)).

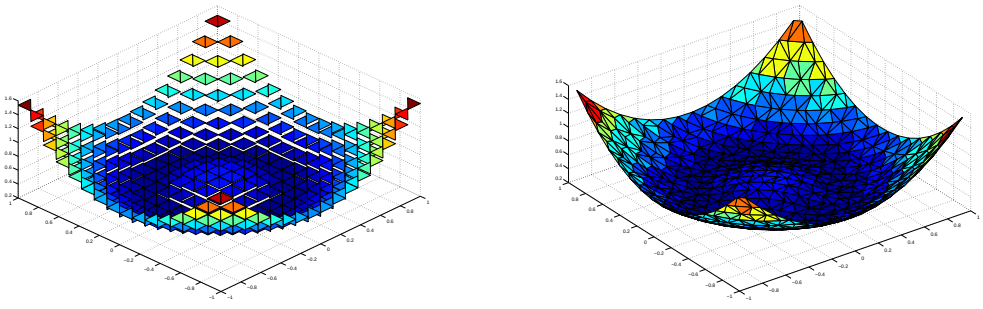


Figure:(a) *Figure:(b)*

Figure 5.6: smooth and piecewise constant numerical solution for obstacle problem

5.2 Convergence of Uzawa's method

5.2.1 Convergence of Classical Uzawa for Obstacle problem

In this section, we study the dependency of the convergence of Uzawa method on the mesh size h_k, we use four different refinement levels corresponding to $h_k = \frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$. The respective meshes consist of 25, 81, 289, 1089 and 4225 nodes.
For the dual variable λ, we get 32, 128, 512, 2048, 8192 degrees of freedom, while for the variable p we obtain 56, 208, 800, 3136 and 12416 (no of edges). Thus the dimension of the matrix S for a finer mesh is 8192×8192 and it takes on average 0.6 seconds in Matlab to compute this matrix. We compute $\alpha_{max} := \frac{2}{\lambda_{max}(M^{-1}S)}$, where $\lambda_{max}(M^{-1}S)$ is the maximum eigenvalue of $M^{-1}S$. Moreover this value is constant if a uniform mesh is used. We further compute the condition number $k(S) = \frac{\lambda_{max}(S)}{\lambda_{min}(S)}$ for different values of step size h_k. In Table 5.2 we put the number of iterations required

	$\alpha_{max} = 0.1111$, λ_0=0			
step size h_k	α	Uzawa steps	No of nodes	$k(S)$
1/2	0.103	62	25	28.7160
1/4	0.103	199	81	116.2310
1/8	0.11	709	289	466.3903
1/16	0.11	2199	1089	$1.8671e+003$
1/32	0.11	6989	4225	$\approx 2.9032e+003$

Table 5.2: Convergence of Uzawa method

for the convergence of λ^k to λ to a degree of relative tolerance 1×10^{-6} using classical Uzawa's method.
From Table 5.2, we can see that the condition number $k(S)$ of $S = BA^{-1}BT$ depends on step size h_k, in fact $k(S)$ grows like h_k^{-2}, this thus leads to the number of iterations required for computation of p, λ by Uzawa's method to increase and this can be observed in Table 5.2. Thus we can see that the operator S is not uniformly well conditioned and thus should be preconditioned to get an efficient algorithm. In Figure (5.7) we display the convergence history of the iteration error.

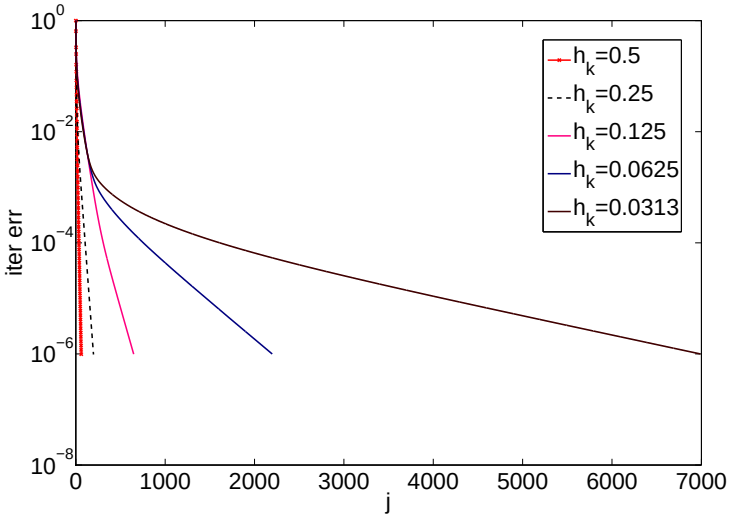


Figure 5.7: Convergence history for different values of step size h_k

To explain the behavior, we introduce the experimental convergence factor E_k computed as follows

$$E_k = \left(\frac{\|h - S\lambda^k\|}{\|h - S\lambda^0\|} \right)^{\frac{1}{k}},$$

with corresponding convergence rate (CR) defined by $CR = -ln(E_k)$ (see [22]). In

step size h_k	E_k	CR
1/2	0.9714	0.0290
1/4	0.9885	0.0116
1/8	0.9957	0.0043
1/16	0.9985	0.0015
1/32	0.9995	4.6767e-004

Table 5.3: Convergence factor

the table (5.3) we display this experimental convergence factor. From this table, we can see that the convergence factor of the classical Uzawa without preconditioning is close to 1, so the algorithm converges very slowly and this gets worse as the mesh

becomes finer.

5.2.2 Preconditioning of Classical Uzawa's method

Convergence results for the mixed system for Poisson equation

In the previous subsection, we have seen that Uzawa's methods takes long to convergence, moreover the number of iterations increase as the mesh becomes fine. In this section we want to introduce a preconditioner $\hat{A}$ for the matrix A and we study the convergence of preconditioned Uzawa's method. As a motivation towards achieving a robust numerical precondtioner for the Uzawa method for the obstacle problem, we start by looking at the Uzawa method for the mixed system for the Poisson equation, since theory regarding effective precondtioners for mixed systems of equations exist(see [13]). As discussed earlier in the previous section, Algorithm (4) is used, with $\hat{A} = A + B^T M^{-1} B$ as a suitable preconditioner. In Table 5.4 we report the results obtained. If we compare with results in Table (5.5) obtained using classical Uzawa's

	$\alpha = 1$				
step size h_k	0.5	0.25	0.125	0.0625	0.0313
iter	16	16	16	16	16

Table 5.4: Convergence for poisson equation with preconditioner $\hat{A} = A + B^T M^{-1} B$

method for the mixed system of poisson equation, we can see that with this choice

	$\alpha = 0.11$				
step size h_k	0.5	0.25	0.125	0.0625	0.0313
iter	226	461	1489	4636	> 5000

Table 5.5: Convergence for poisson equation with preconditioner $\hat{A} = A$

of preconditioner, we get a robust method, where the condition number of the new Schur complement $M^{-1} B^T \hat{A}^{-1} B$ is independent of h_k, and consequently the number

of iterations counts also does not grow with step size which is the ultimate goal of any numerical experiment. These results further agree with literature (see [13]) where the robustness of the preconditioner of the form $\hat{A} = A + B^T W^{-1} B$ is claimed, where W is the weight matrix. Motivated by these results, as discussed earlier, we now want to see if such a preconditioner can be applied to the inequality system.

Convergence results for the mixed system for obstacle problem

We now want to study the convergence of preconditioned Uzawa's method for the mixed system for the obstacle problem. As discussed earlier in the previous section, Algorithm (5) is used, with three possible alternative cases for the preconditioner $\hat{A}$. We discuss the three cases i.e $\mathcal{A}^{(k+1)} = \emptyset \left(\hat{A} = A\right)$, the case where $\mathcal{A}^{(k+1)}$ is fully indexed such that $\sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_{ii}^{-1} b_i = B^T M^{-1} B$, and the case when the index matrix is dynamically changing such that

$$\hat{A}^{(k+1)} = A + \sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_{ii}^{-1} b_i.$$

- The case $\mathcal{A}^{(k+1)} = \emptyset \left(\hat{A} = A\right)$.
  For this, convergence results the same as those obtained for the classical Uzawa Method are obtained (see Table(5.2)) .

- The case $\hat{A} = A + B^T M^{-1} B$.
  In Table (5.6) we put the results obtained for this case.

	$\alpha = 0.11$				
step size h_k	0.5	0.25	0.125	0.0625	0.0313
iter	268	495	677	2562	8000

Table 5.6: Convergence for obstacle problem with preconditioner $\hat{A} = A + B^T M^{-1} B$

If we compare these results with those from classical Uzawa method, we see that they are of the same order. This thus means that although this preconditioner is

good for mixed system equations, it is not optimal for variational Inequalities. It is comparable to the classical Uzawa method. In other words by presuming that the index set comprise of only equations is as equally bad as assuming that it comprise of only inequalities.

We now study the intermediate case where we take into consideration of having some knowledge about the index set $\mathcal{A}$ where $\lambda_i > 0$. We use algorithm (5) and take as initial conditions, $\hat{A} = A$, $\alpha = \alpha_0$ (α_0 optimal parameter for convergence of classical Uzawa method). After some iter* determined experimentally, we update the matrix $\hat{A} = A$ to $\hat{A} = A + \sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_{ii}^{-1} b_i$ and α_0 to 1. On each consecutive iteration, we compute the index and consequently update $\hat{A}$ while we keep α fixed to 1. In Table 5.7 we report the results obtained.

step size h_k	iter*	iter(classical Uzawa)	iter	Extra iteration needed
1/2	16	62	29	13
1/4	128	199	140	12
1/8	330	709	342	12
1/16	1750	2199	1762	12
1/32	4500	6989	4512	12

Table 5.7: Dynamically changing index set

Here we can see that provided we have identified clearly a better starting value, we just need an extra constant 12 iterations to achieve convergence, and this is independent of the mesh size. Thus this means that provided we have a better way of computing the initial guess rather than using the classical uzawa method, the update by the preconditioner $\hat{A} = A + \sum_{i \in \mathcal{A}} b_i^T M_{ii}^{-1} b_i$ seems to be robust. It remains to be investigated how well one can handle this initial computational step before we carry out the update.

5.3 Discretization Error Estimation

In this section, we present the results from numerical experiments that illustrate the theory of error estimates developed in chapter[3]. Our objective is then to establish

experimentally the conclusions from the theoretical analysis of the approximation of the obstacle problem. To compute the L^2 error, we approximate element wise both the numerical solution u_h and the exact solution u at the centroid $(\bar{x}, \bar{y})$ of each of the triangles $\mathcal{T}_j$. Then the L^2 error is computed according to

$$||u - u_h|| = \sqrt{\sum Area(T_j) * (u_j(\bar{x}) - u_{hj}(\bar{x}))}. \tag{5.3.1}$$

The maximum edge length in the triangulation is equivalent to h_{kmax} and is computed from matrix B^T. The radial solution $u(r)$ which was computed in section (5.1) is converted into $u(x, y)$ and used as the analytic solution such that

$$u(x,y) = \begin{cases} 0.3 - x^2 - y^2 & \text{on } \ 0 \leq \sqrt{x^2+y^2} \leq 0.2953 \\ x^2 + y^2 - 0.3 - 0.3489166809 \ln(\sqrt{x^2+y^2}) & \text{on } \ 0.2953 \leq \sqrt{x^2+y^2} \leq 1 \end{cases}$$

Below in Table (5.8) we report the error computed for different values of mesh size h_{kmax}. If we denote the step size h_k for triangulation $\mathcal{T}_k$ and h_{k+1} for triangulation $\mathcal{T}_{k+1}$ with respective errors obtained at these steps denoted by e_k and e_{k+1}, then the value of p is computed according to

$$p = \frac{\log\left(\frac{h_{k+1}}{h_k}\right)}{\log\left(\frac{e_{k+1}}{e_k}\right)}$$

	λ_0=0		
step size h_k	$h_{\max}$	$e = \|\|u - u_h\|\|_2$	p
1/2	0.6019	0.02230297	-
1/4	0.3108	0.003368309	2.8600
1/8	0.1500	0.001218882	1.3953
1/16	0.0769	0.0003664056	1.7990
1/32	0.0398	7.637119e-005	2.3809

Table 5.8: Error estimates for the obstacle problem

Figure (5.8) indicates the loglog plot of the L^2-error against h_{max}. The solid line indicates the error plot while the dotted line indicates order 2 convergence . Thus we conclude from these observations of p in Table 5.8 that the error $\|u - u_h\|$ is $O(h_k^{2\pm\epsilon})$ which still agrees with the theory we presented, since h_k is less than one.

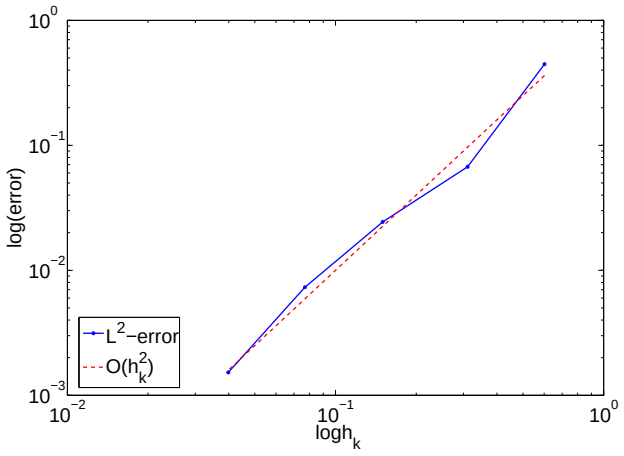


Figure 5.8: Convergence order for obstacle problem

Chapter 6

Conclusions and Future Work

6.1 Conclusions

Bearing in mind that many important problems ranging from physics to finance can be formulated by transformation to an obstacle problem, an efficient and robust way of solving this free boundary problem was a major motivation of our study.
Starting from a simple physical example of this problem, we derived a simple mathematical model to this problem. We showed that a solution to this problem is equivalent to solving a constrained minimization problem over a non-empty closed convex set. We further showed that the minimization problem is equivalent to a variational problem (primal problem). Motivated by the mixed method of Raviart-Thomas, under the assumptions of high regularity of our solution u, we further derived a new primal variational formulation and an equivalent mixed formulation both of which included variational inequalities. The existence and uniqueness of these formulations were further established. In order to actually compute the deflection of the membrane, we discretized the problem. In the discrete setting, we showed that there exists a unique solution to the discrete problem since the inf-sup condition is satisfied. We proposed Uzawa algorithm to solve the variational inequalities induced by the discrete problem. We proved convergence of the Uzawa algorithm, and for this method, we obtained linear convergence. Finally, we implemented the problem numerically using Matlab. The classical Uzawa's method was used to solve the system of equations and inequalities and a solution to the obstacle problem was obtained.

We observed that as the parameter of discretization was made smaller, the mixed finite element method converged to the exact location of the free boundary which was part of our problem. We further established numerically the validity of the theoretical error estimates which were found to be sharper. On the other hand, as we expected, we found that the Uzawa method does not converge faster, since it uses directions of steepest descent. In addition the convergence depended on the parameter of discretization, as the condition number of the system matrices gets worse for finer meshes.

Thus in order to improve on the convergence of Uzawa method, a symmetric three step Uzawa type algorithm was introduced. Studies involving mixed systems for a Poisson equation with a preconditioner $\hat{A} = A + B^T M^{-1} B$ for the block matrix A were carried out and proved to be robust, since the condition number of the associated Schur complement was found to be independent of the parameter of the discretization h_k. As a motivation from this approach, we modified the symmetric Uzawa algorithm by introducing an index set which comprised of all indices where λ is positive. A preconditioner $\hat{A}$ was computed based on this set. We investigated two extreme cases where the index set is empty. Results same as the classical Uzawa method were obtained. Further we tested the case where the index set is full, i.e $\hat{A} = A + B^T M^{-1} B$. Experiments revealed that although this preconditioner is robust for the mixed system of a poisson equation, it seems to be not any better than the classical Uzawa. In fact the number of iterations obtained are of the same order as the classical Uzawa. We investigated the intermediate case where an index set of positive λ was considered at each iteration step. Results revealed that if we switch to the preconditioner $\hat{A}$, and the parameter α to 1 after some number of iterations, we obtain a robust numerical method independent of h_k.

6.2 Future Work

Based on conclusions that we have discussed in the previous section, the following are possible suggestions for the extension of this work,

- We propose a thorough study of the case where we choose only indices where

$\lambda > 0$ such that the preconditioner

$$\hat{A} = A + \sum_{i \in \mathcal{A}^{(k+1)}} b_i^T M_{ii}^{-1} b_i$$

where

$$\mathcal{A}^{(k+1)} = \{i \ \mid \ \lambda_i > 0\}$$

- Provided we have a better way of computing the initial guess rather than the classical uzawa method, the update by the preconditioner $\hat{A} = A + \sum_{i \in \mathcal{A}} b_i^T M_{ii}^{-1} b_i$ seems to be robust. It remains to be investigated how well one can handle this initial computational step before we carry out the update.

- In algorithm 5, we needed to update the parameter α from α_0 to 1. It remains to be investigated on how well one can continuously update this parameter as we simultaneously update the index set.

- Further, by assuming that we have no prior knowledge of the the indices, one can modify the symmetric Uzawa algorithm such that the set $\mathcal{A}^{(k+1)}$ is dynamically changing.

- The Uzawa type methods we studied replaced the exact inverse of A by an "approximate" evaluation of A^{-1} or a preconditioner for its symmetric part. However several other efficient algorithms such as: (i) a linear one-step method, where the action of the inverse is replaced by a linear preconditioner such as one sweep of a multigrid procedure; (ii) a multistep method, where a sufficiently accurate approximation to A^{-1} is provided by some iterative method, e.g., preconditioned GMRES or preconditioned Lanczos , can be used.

Thus in short summary, we hope that this approach can yield a robust numerical method for computing the numerical solution of the obstacle problem using the Uzawa type methods provided the above points are taken into considerations.

Bibliography

[1] D. N. Arnold, *Mixed finite element methods for elliptic problems*, Comput. Methods Appl. Mech. Eng. **82** (1990), no. 1-3, 281–300.

[2] C. Bahriawati and C. Carstensen, *Three matlab implementations of the lowest-order raviart- thomas mfem with a posteriori error control.*, Computational Methods In Applied Mathematics **5(4)** (2005), 333361.

[3] Gerald E. Bartholomew, *Numerical integration over the triangle*, Mathematical Tables and Other Aids to Computation **13** (1959/10), 295–298.

[4] J.H. Bramble, J E Pasciak and A T Vassilev, *Analysis of the Inexact Uzawa Algorithm for saddle point problem*

[5] F. Brezzi D. N. Arnold and M. Fortin., *A stable finite element for the stokes equations*, Calcolo **21** (1984), 337–344.

[6] ED. Bueler ,*An easy Finite element Implementation of the obstacle problem for Poisson equation* www.math.uaf.edu/ bueler/poissoncont.pdf, 2004.

[7] P. G. Ciarlet and J.L. Lions (eds.), *The finite element method for elliptic problems Vol. I*, Handbook of Numerical Analysis, I, North-Holland, Amsterdam, 1978.

[8] P. G. Ciarlet and J.L. Lions (eds.), *Handbook of numerical analysis. Vol. I*, Handbook of Numerical Analysis, I, North-Holland, Amsterdam, 1990.

[9] W. Drenth , *On Friedrichs' inequalities* Eindhoven Univeristity of Technolgy. January 24, 2002.

[10] LC. Evans, *An introduction to variational inequalities and their applications (d. kinderlehrer and g. stampacchia)*, SIAM Review **23** (1981), no. 4, 539–543.

[11] Howard C. Elman and Gene H. Golub, *Inexact and preconditioned uzawa algorithms for saddle point problems*, SIAM J. Numer. Anal. **31** (1994), no. 6, 1645–1661.

[12] R. Glowinski, *Numerical methods for nonlinear variational problems*, Springer-Verlag, Berlin-Heidelberg-New York:, 1984.

[13] C. Greif and D. Schötzau, *Preconditioners for saddle point linear systems with highly singular (1,1) blocks*, Num Analysis **22** (2006), 114–121.

[14] W. W. Hager, F. Brezzi and P.A. Raviart, *Error estimates for the finite element solution of variational part ii. mixed methods inequalities*, Springer-Verlag **2** (1978).

[15] S. Howison , F. Wilmott and J. Dewynne, *The mathematics of financial derivative*, Cambridge University Press, Cambridge, 1995.

[16] D. Kinderlehrer and G. Stampacchia, *An introduction to Variational Inequalities and Their Applications*. Academic Press New York London Toronto Sydney San Francisco, 1980

[17] R.C Kirby, *Arbitrary order mixed finite elements for second order elliptic problems*, Math. Comput. (2002).

[18] M.A. Noor and J.R. Whiteman, *Error bounds for finite elementsolutions of mildly nonlinear ellitic boundary value problems*, Num.Math. **26** (1976), 107–116.

[19] F.A. Radu, *Lecture notes on Mixed and Mixed Hybrid Finite Element Methods: Theory, Implementation and Applications* Max-Planck Institute for Mathematics in the Sciences Leipzig, Germany.

[20] P.A. Raviart, J.M Thomas *A mixed finite element method for second order elliptic problems. in: Mathematical aspects of finite element methods.*, Springer-Verlag, Berlin-Heidelberg-New York:, 1977.

[21] J. Rodrigues, *Obstacle problems in mathematical physics*, Elsevier Science (1987).

[22] Y. Saad (2003), *Iterative Methods for Sparse Linear Systems*, second edn, SIAM, Philadelphia, PA.

[23] A. Sinwel, *Numerical simulation of contact problems with coulomb friction*, Master's thesis, Johannes Kepler University Linz Austria, 2006.

[24] R.S. Varga, *Matrix iterative analysis*, Prentice-Hall, Engewood Cliffs, NJ, 1962.

[25] M. Wainwright, *Lecture notes EE 227A / STAT 260: Nonlinear and Convex Optimization, Lecture 7 September Fall 2004.* URL: http://www.eecs.berkeley.edu/w̃ainwrig/ee227a/# lecture

[26] W. Zulehner , *Analysis of iterative Methods For Saddle Point Problems: A Unified Approach* Math. Comp. 71, 479505.

Eidesstattliche Erklärung

Ich, Henry Kasumba, erkläre an Eides statt, dass ich die vorliegende Masterarbeit selbstständig und ohne fremde Hilfe verfasst, andere als die angegebenen Quellen und Hilfsmittel nicht benutzt bzw. die wörtlich oder sinngemäß entnommenen Stellen als solche kenntlich gemacht habe.

Author: _____

Henry Kasumba

Vitae

Surname:	Kasumba
Given Names:	Henry
Place of Birth:	Masaka, Uganda
Date of Birth:	January 1st, 1981

Educational Institutions Attended

Johannes Kepler University Linz (Austria)	2006-2007
Techinische Universiteit Eindhoven (The Netherlands)	2005-2006
Makerere University kampala (Uganda)	2001-2004

Degrees Awarded

B. Sc/Education (Mathematics/Physics)	
Makerere University Kampala (Uganda)	2004
M. Sc-Industrial Mathematics (Computational Science and Engineering)	
Techinische Universiteit Eindhoven (The Netherlands)	2006
M. Sc-Industrial Mathematics (Computational Science and Engineering)	
Johannes Kepler University Linz (Austria)	2007

Special Activities

Participant modeling week Lyngby-Denmark	2006
Participant modeling week Mathematics for Industry-Eindhoven Netherlands	2007