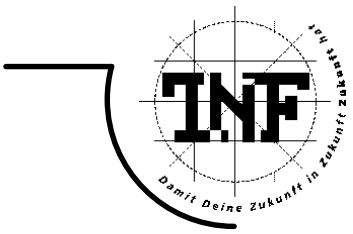
JOHANNES KEPLER
UNIVERSITÄT LINZ
Netzwerk für Forschung, Lehre und Praxis



Hermitian and skew-Hermitian Solvers and Preconditioners: Application to Symmetric and Indefinite Problems

Diplomarbeit

zur Erlangung des akademischen Grades

Diplomingenieurin

in der Studienrichtung

Technische Mathematik

Angefertigt am *Institut für Numerische Mathematik*

Betreuung:

O. Univ. Prof. Dipl. Ing. Dr. Ulrich Langer

Eingereicht von:

Lilya Ghazaryan

Linz, August 2008

Johannes Kepler Universität
A-4040 Linz · Altenbergerstraße 69 · Internet: http://www.jku.at · DVR 0093696

For my parents, Levon and Veta

Abstract

The aim of the master thesis is to apply the Hermitian and skew-Hermitian (HSS) iterative method and its inexact version to the solution of linear algebraic systems, arising in different applications, with symmetric, but indefinite system matrices. In particular, we consider saddle-point problems coming from a reformulation of the well-known domain decomposition Finite Element Tearing and Interconnecting method as a saddle-point problem with both primal and dual variables as unknowns. This is an alternative, hopefully better, approach to the existing methods, namely block-structured preconditioners combined with suitable Krylov subspace methods or Schur-complement conjugate gradient methods. The convergence of HSS method is studied numerically. The numerical experiments show that the use of the HSS method as a preconditioner in a Krylov subspace method is very efficient.

Acknowledgement

During my two-year study in Technical University of Eindhoven (TU/e) and Johannes Kepler Universität Linz (JKU) many people have influenced, directly or indirectly, on writing my master thesis. I consider writing this master thesis as one of the most important steps in my life. First of all, this is evaluation of accomplishing another stage in my education. Secondly, this is a culmination of my two-year journey in Europe.

I would like to thank first of all TU/e and JKU for hosting me for two years and giving me the chance to be a part of their academical life. I am very grateful to my supervisor Prof. Ulrich Langer for inspiring and encouraging me while working on the thesis. In general, I want to thank all professors in both universities who shared their knowledge and experience with me and my colleagues. These two years of my study would not be that fruitful without the discussions during the lectures, seminars and group-works. I am very happy that I can see the change in my knowledge and that the hopes I had before starting the master program are fulfilled. I found out for myself that the more I study applied mathematics, the more I like it!

I would not be able to reach this stage of my life without the help and support of my family. I am deeply grateful for my parents, Levon and Veta, and sisters, Sara and Tamara, for their believe in me. I am thankful to all my friends that I met in Europe for making me feel like at home and sharing the difficulties and happiness with me. These two years would not be as interesting as they were without all my friends.

Writing this thesis is another experience in my education, hopefully not the last one !

Contents

Chapter 1

Introduction

The main goal of this work is to consider the application of Hermitian and skew-Hermitian (HSS) splitting iteration [2] to the saddle point-system coming from Domain Decomposition (DD) Finite Element Tearing and Interconnecting (FETI) methods [10]. The choice of this particular domain decomposition method as an application is motivated by the fact that FETI and the more recent Dual-Primal Finite Element Tearing and Interconnecting (FETI-DP), Balancing Domain Decomposition by Constraints (BDDC) methods are the most widely used domain decomposition methods [12].

A large number of phenomena in nature are mathematically described by Partial Differential Equations (PDEs)[1]. Elliptic PDEs play the main role. The reason for this is that time integration methods for parabolic and hyperbolic PDEs finally lead to the solution of a sequence of elliptic PDEs. The theory of elliptic PDEs [3] is a well developed area in applied mathematics, and there are fundamental methods which are widely used for solving such problems. The machinery for solving linear elliptic PDEs can be roughly described in the following way:

- establishing the well posedness of the problem: namely, the existence, uniqueness and stability of the weak solution,

- discretization of the problem: transition from infinite dimensional space to finite dimensional space,

- solving the discrete system: usually formulated as a linear system of equations.

The first step of this description is more or less fixed, as the well posedness of the problem relies on the theorem of Lax-Milgram [1], or more generally, on Fredholm's theory [33]. The other two steps can be done in several ways. For instance, for discretization of the problem one can use different discretization techniques, like Finite Element Method (FEM) [9], Finite Difference Method (FDM) [4], Finite Volume Method (FVM) and Boundary Element Method (BEM) [5]. And, of course, nowadays there exist a large number of efficient ways for solving linear systems of equations [6],[7]. The choice of the method either for discretization or for solving the problem strongly depends on the nature of the problem. In order to find the most efficient way for solving a particular problem one has to take into account the properties of the problem. For example, if the resulting linear system is symmetric positive definite then a good choice can be the Conjugate Gradient (CG) method [6]. With this regard, certain methods can be quite efficient applied for certain class of problems.

So, the HSS method we are going to consider is a representative of iterative methods. Iterative methods for numerical computation of the solution of a linear system of equations (originally by Gauss in 1823, Liouville in 1837 and Jacobi in 1845) have popularity in several areas of scientific computing. The approach of iterative methods is quite different from the direct solution methods (such as Gaussian elimination) which until recent times were often preferred to iterative methods because of their predictable behavior. Though initially iterative methods had somehow special-purpose nature as they were developed for certain applications, nowadays iterative methods started to be superior to direct solvers for large scale problems. The most efficient direct solvers available nowadays might not be really efficient for solving for instance linear systems coming from the discretization of partial differential equation in three dimensional space because of the memory and computational requirements. In such cases iterative solvers are widely used as, compare to the direct methods, they are easier to implement efficiently on high-performance, and, in particular, on parallel computers.

The simplest iterative method (preconditioned Richardson) for solving linear system $Ax = b$ has the form:

$$x^{k+1} = x^k - \tau C^{-1}(Ax^k - b) \qquad k = 0, 1, 2.....$$

where x^0 is a given initial guess, τ is a suitably chosen relaxation parameter and C is some preconditioner. The classical preconditioners C are I (classical Richardson iteration) and diag$A$ (Jacobi)[8]. Under certain conditions the sequence $\{x^k\}$ converges to the solution of $Ax = b$ [6]. One of the advantages of the iterative

methods is that only matrix-vector multiplications and vector operations need to be performed. Also the computer storage is only required for the nonzero entries of A and the vector x in addition to one or two more vectors. Therefore, one can use the advantage of sparsity of A if that is the case. Sometimes fully populated matrices A are data sparse (for example circulant, block-circulant matrices) or can be approximated by data sparse matrices [32]. This allows to reduce the storage required for A and the cost of one matrix-vector multiplication from n^2 to $n(\log n)^\alpha$ for some $\alpha > 0$, usually $\alpha \in \{1, 2\}$, where n is the size of A. In that case, iterative solvers are even more efficient. While for direct methods one needs permutations in order to avoid a considerable amount of fill-in [32].

The disadvantage of the iterative methods is that the rate of convergence may be slow and a proper stopping criterium needs to be found. Nevertheless, for some important classes of matrices, convergence analysis results as well as practical implementation aspects are available nowadays, for instance, for Symmetric Positive Definite (SPD) problems [12], [9] and saddle-point problems [11]. Certainly, the choice of iterative method depends on the problem itself. More precisely, one has to take into account the properties of the problem which leads to the linear system. Nevertheless, one could combine somehow direct solvers and iterative solvers in a way, that the advantages of both iterative and direct methods are used. In this regard domain decomposition methods can be seen as a good combination of the two solution methods. The class of domain decomposition methods has gained enormous popularity during the last decade [12]. This methods follow the idea of "divide -and -conquer", which means that they divide the original problem into a number of smaller problems. This division is done for different reasons. Sometimes such a division arises from breaking up a domain with complicated geometry. In other cases the division is more artificial. The subproblems are easier to solve because of their smaller size and often parallel computation can be used. This is quite important for the efficiency of computations.

Domain decomposition methods can be seen from two different point of view. One is that they may arise from separation of a physical domain into regions (subdomains). In these regions the problem can be modeled by separate partial differential equations. On the interfaces between the subdomains various conditions, such as continuity, are imposed. The other approach is to see domain decomposition methods as methods for solving large algebraic linear systems arising from the discretization of partial differential equations. In that sense, a domain decomposition method can be seen as an algebraic method, where the large system is subdivided into smaller problems, whose solutions can be used to generate a preconditioner for the large system [13].

In proceeding chapters we will look at the HSS iterative method itself, its use for constructing preconditioners, discuss the domain decomposition method FETI and finally present the numerical result we have obtained. For the sake of completeness of the material throughout the paper we will repeat some of the proofs of basic results [2], [10], [14], [26], [29]. In more details, the paper is organized as follows. We will start with a small Chapter 2 where we will motivate the objective of this work. For this we will consider a model problem, a simple boundary value problem, solving which will suggest the consideration of FETI domain decomposition method and HSS iterative method. Then we will continue with Chapter 3 where we will introduce the Hermitian and skew-Hermitian iterative method itself. We will include the main convergence results for this method. Also we will discuss recently introduced Accelerated HSS method [14]. Afterwards in Chapter 4 the actual application of the HSS method in our paper, namely, the Finite Element Tearing and Interconnecting domain decomposition method will be considered. We will describe the method and give its saddle-point formulation in dual and primal variables. Later on in Chapter 5, after all the required methods are introduced, we will consider the application of HSS method to the FETI system for out model problem. The HSS method will be considered as an stationary iterative method and also as a preconditioner for GMRES, which is a widely used Krylov subspace method [6], [28]. We will present the numerical results we have obtained and interpret them. Since saddle-point systems arise in many scientific and engineering applications, including computational fluid dynamics [15], [16], [24], [17], mixed finite element approximation of elliptic PDE's [18], [19], [25], optimization [20], [23], [22], [21], in Chapter 6 we will present some more examples of problems leading to saddle-point systems. The paper will be concluded with a short summary and a brief discussion of possible continuation.

Chapter 2

Motivation

In this small chapter we will try to motivate the topic of our work by considering a model problem, a Dirichlet boundary value problem on a rectangular domain. We will present the main steps for solving such a problem and point out which steps are connected to our topic.

2.1 Some concepts

Before we introduce our model problem we will consider definitions of several basic concepts that we will need throughout the chapter.
We will denote by $\mathrm{L}^2(\Omega)$ the space of scalar functions which are defined and square integrable over Ω in Lebesgue sense, namely

$$\mathrm{L}^2(\Omega) = \{u : \int_\Omega u^2 \mathrm{d}x < \infty\}.$$

$\mathrm{L}^2(\Omega)$ is a Hilbert space with inner product $(u, v) = \int_\Omega uv\mathrm{d}x$ and the induced norm $||u||^2_{\mathrm{L}^2(\Omega)} = (u, u)$. In the same manner, for vector functions $\bar{v} = [v_1, v_2, ..., v_d]^T$ we define the Hilbert space $(\mathrm{L}^2(\Omega))^d$ in the following way:

$$(\mathrm{L}^2(\Omega))^d = \{\bar{v} = [v_1, v_2, ..., v_d]^T : v_i \in \mathrm{L}^2(\Omega) \quad \text{for} \quad i = 1, ..., d\}.$$

The inner product in $(\mathrm{L}^2(\Omega))^d$ is defined as $(u, v) = \int_\Omega \bar{u} \cdot \bar{v}\mathrm{d}x = \sum_{i=1}^d \int_\Omega u_i v_i \mathrm{d}x$ and the corresponding norm is $||\bar{v}||^2_{\left(\mathrm{L}^2(\Omega)\right)^d} = (\bar{v}, \bar{v})$.
We also introduce the notion of multi-index. A multi-index $\alpha = (\alpha_1, \alpha_2, ..., \alpha_d)$ is

a set of non-negative integers. We define $|\alpha| = \sum_{i=1}^{d} \alpha_i$. Using this, we define the partial derivative D^α as follows

$$\mathrm{D}^\alpha = \frac{\partial^{|\alpha|}}{\partial^{\alpha_1} x_1 \partial^{\alpha_1} x_1 \dots \partial^{\alpha_d} x_d}.$$

For fixed $m \geq 0$ we define the Sobolev spaces,

$$\mathrm{H}^m(\Omega) = \{v : v \in \mathrm{L}^2(\Omega) \text{ and } \mathrm{D}^\alpha v \in \mathrm{L}^2(\Omega) \text{ for } |\alpha| \leq m\},$$

and the associated seminorms and norms are defined as

$$|v|^2_{\mathrm{H}^m(\Omega)} = \sum_{|\alpha|=m} ||\mathrm{D}^\alpha v||_{\mathrm{L}^2(\Omega)} \text{and} ||v||^2_{\mathrm{H}^m(\Omega)} = \sum_{k \leq m} ||\mathrm{D}^k v||_{\mathrm{L}^2(\Omega)}.$$

We will mainly consider the space

$$\mathrm{H}^1(\Omega) = \{v : v \in \mathrm{L}^2(\Omega) \text{ and } \frac{\partial v}{\partial x_1}, ..., \frac{\partial v}{\partial x_d} \in \mathrm{L}^2(\Omega)\},$$

and subspace $\mathrm{H}^1_0(\Omega) \subset \mathrm{H}^1(\Omega)$, defined as

$$\mathrm{H}^1_0(\Omega) = \{v : v \in \mathrm{H}^1(\Omega) \text{ and } v = 0 \text{ on } \partial\Omega\}.$$

We will also come across with the space $\mathrm{H}(\mathrm{div}; \Omega)$, which is

$$\mathrm{H}(\mathrm{div}; \Omega) = \{v : v \in \mathrm{L}^2(\Omega)^d \text{ and } \mathrm{div} v \in \mathrm{L}^2(\Omega)\}.$$

If we assume that $\partial\Omega$ is smooth then we can define the *trace* $v/_{\partial\Omega}$ of any function $v \in \mathrm{H}^1(\Omega)$. The set of all traces of such functions gives rise to the Hilbert space $\mathrm{H}^{\frac{1}{2}}(\partial\Omega)$:

$$\mathrm{H}^{\frac{1}{2}}(\partial\Omega) = \{g : g = v/_{\partial\Omega} \text{ for some } v \in \mathrm{H}^1(\Omega)\}.$$

In the same manner for vector functions $\bar{v}$ in $\mathrm{H}(\mathrm{div}; \Omega)$, the set of *normal traces* $(\bar{v} \cdot \bar{n})/_{\partial\Omega}$, where $\bar{n}$ denotes the outward normal vector to $\partial\Omega$, gives rise to the dual space $\mathrm{H}^{-\frac{1}{2}}(\partial\Omega)$:

$$\mathrm{H}^{-\frac{1}{2}}(\partial\Omega) = \{q : q = (\bar{v} \cdot \bar{n})/_{\partial\Omega} \text{ for some } \bar{v} \in \mathrm{H}(\mathrm{div}; \Omega)\}.$$

Some fundamental inequalities in Sobolev spaces will be also needed in our analysis. In particular, the Poincare-Friedrich's inequality.

Lemma 2.1. *(Poincare-Friedrich's inequality). For all $v \in H^1_0(\Omega)$ the following inequality holds true:*

$$||u||_{L^2(\Omega)} \leq C_F(\Omega)|v|_{H^1(\Omega)}$$

where $C_F(\Omega)$ is a constant that depends on Ω.

Lemma 2.2. *For all $v \in H^1(\Omega)$ the following equality holds true:*

$$\int_\Omega \nabla \cdot \bar{v} u dx = -\int_\Omega \bar{v} \cdot \nabla u dx + \int_{\partial\Omega} \bar{v} \cdot \bar{n} u ds_x$$

2.2 Model problem

Let us consider a representative of elliptic equations, Laplace equation, which models electrostatic interaction and many other potential problems. So, our model problem will be the following boundary value problem in two dimensional space:

$$\begin{cases} -\triangle u(x,y) &= f(x,y) \quad \text{in} \quad \Omega = (0,2)\times(0,1), \\ \\ u(x,y) &= g(x,y) \quad \text{on} \quad \partial\Omega. \end{cases} \tag{2.1}$$

where $\triangle u(x,y) = \dfrac{\partial^2 u}{\partial^2 x}(x,y) + \dfrac{\partial^2 u}{\partial^2 y}(x,y)$ and $f \in \mathrm{L}^2(\Omega)$, $g \in \mathrm{L}^2(\partial\Omega)$ are given functions.
We denote by ∇v the gradient of v which is given by $\nabla v = [\dfrac{\partial v}{\partial x}, \dfrac{\partial v}{\partial y}]^T \in \mathbb{R}^2$. The derivative of v in the direction of the outward normal $\mathrm{n} = [n_x, n_y]^T \in \mathbb{R}^2$ is denoted by $\dfrac{\partial v}{\partial n}$ and defined as $\dfrac{\partial v}{\partial n} = \nabla v \cdot \mathrm{n}$.
Now let us multiply the first equation of (2.1) by a test function $v \in \mathrm{V} := \mathrm{H}^1(\Omega)$ and integrate over Ω. We will get:

$$-\int_\Omega \triangle u v \mathrm{d}x = \int_\Omega f v \mathrm{d}x.$$

According to Green's second formula [9] :

$$-\int_\Omega \triangle u v \mathrm{d}x = \int_\Omega \nabla u \nabla v \mathrm{d}x - \int_{\partial\Omega} \frac{\partial u}{\partial n} v \mathrm{d}s_x,$$

therefore, if we incorporate the essential boundary condition $v = 0$ on $\partial\Omega$, we will get:

$$-\int_\Omega \triangle u v \mathrm{d}x = \int_\Omega \nabla u \nabla v \mathrm{d}x.$$

Hence

$$\int_\Omega \nabla u \nabla v \mathrm{d}x = \int_\Omega f v \mathrm{d}x \quad \text{for all} \quad v \in \mathrm{V}_0,$$

where

$$\mathrm{V}_0 := \{v \in \mathrm{V} : v = 0 \quad \text{on} \quad \partial\Omega\}$$

Now, if we consider $\mathrm{V}_g = \{u \in \mathrm{V} : u = g \quad \text{on} \quad \partial\Omega\}$, then we can give the variational formulation of the boundary value problem (2.1).

For given $f \in \mathrm{L}^2(\Omega)$, find $u \in \mathrm{V}_g$ such that

$$\int_\Omega \nabla u \nabla v \mathrm{d}x = \int_\Omega f v \mathrm{d}x \quad \text{for all} \quad v \in \mathrm{V}_0. \tag{2.2}$$

Now, let us introduce bilinear form $a : \mathrm{V} \times \mathrm{V} \to \mathbb{R}$ and linear form $F : \mathrm{V} \to \mathbb{R}$ given by

$$a(u, v) = \int_\Omega \nabla u \nabla v \mathrm{d}x \quad \text{and} \quad \langle F, v \rangle = \int_\Omega f v \mathrm{d}x. \tag{2.3}$$

Then the variational problem (2.2) can be rewritten in the following way:

For given $f \in \mathrm{L}^2(\Omega)$, find $u \in \mathrm{V}_g$ such that

$$a(u, v) = \langle F, v \rangle \quad \text{for all} \quad v \in \mathrm{V}_0. \tag{2.4}$$

Next, we will formulate the fundamental theorem in the theory of variational problems, the well-known Lax-Milgram theorem [9], which will provide us with existence and uniqueness of the solution of (2.4). For this, let us first define some concepts.

Definition 2.3. *A bilinear form $a(\cdot, \cdot)$ on a normed linear space H is said to be bounded or continuous, if $\exists C < \infty$ such that*

$$|a(u, v)| \leq C||u||_H||v||_H \quad \textit{for all} \quad u, v \in H.$$

Definition 2.4. *A bilinear form $a(\cdot, \cdot)$ on a normed linear space H is said to be coercive on $V \subset H$, if $\exists \alpha > 0$ such that*

$$|a(u, u)| \geq \alpha||u||_H^2 \quad \textit{for all} \quad u \in V.$$

Definition 2.5. *A linear form $\langle F, \cdot \rangle$ on a normed linear space H is said to be bounded or continuous on H, if $\exists C < \infty$ such that*

$$|\langle F, v \rangle| \leq C||v||_H \quad \textit{for all} \quad v \in V.$$

Now, having these definitions we can state Lax-Milgram's theorem.

Theorem 2.6. *Let H be a Hilbert space and V a closed subspace of H. Assume that*

- *$a(\cdot, \cdot)$ is a bounded (continuous) bilinear form on H*
- *$a(\cdot, \cdot)$ is coercive on V*
- *$\langle F, \cdot \rangle$ is a bounded (continuous) functional on V*

then the the variational problem find $u \in V$ such that

$$a(u, v) = \langle F, v \rangle \quad \textit{for all} \quad v \in V \tag{2.5}$$

has a unique solution.

So, in order to show that our model problem is well posed it is sufficient to show that bilinear form $a(\cdot,\cdot)$ and functional $\langle F,\cdot\rangle$ satisfy the conditions of Lax-Milgram's theorem.

- boundedness of $a(\cdot,\cdot)$: from the definition of $a(\cdot,\cdot)$ for all $u,v\in \mathrm{V}_0$ we have

$$\begin{aligned} |a(u,v)| &= \left|\int_\Omega \nabla u\nabla v \mathrm{d}x\right| \\ \text{(triangle inequality)} &\leq \int_\Omega |\nabla u||\nabla v|\mathrm{d}x \\ \text{(Cauchy-Schwartz inequality)} &\leq (\int_\Omega |\nabla u|^2\mathrm{d}x)^{\frac{1}{2}}(\int_\Omega |\nabla v|^2\mathrm{d}x)^{\frac{1}{2}} \\ &\leq ||\nabla u||_{\mathrm{L}_2}||\nabla v||_{\mathrm{L}_2} \\ \text{(with } C_1=1) &\leq C_1||u||_{\mathrm{H}^1}||v||_{\mathrm{H}^1} \end{aligned}$$

- coercivity of $a(\cdot,\cdot)$: from the definition of $a(\cdot,\cdot)$ for all $u\in \mathrm{V}_0$ we have

$$\begin{aligned} a(u,u) &= \int_\Omega \nabla u\nabla u\mathrm{d}x \\ &= ||\nabla u||^2_{\mathrm{L}_2} \\ \text{(Friedrich's inequality)} &\geq \frac{1}{c_F^2}||u||^2_{\mathrm{H}^1} \end{aligned}$$

- boundedness of $\langle F,\cdot\rangle$: from the definition of $\langle F,\cdot\rangle$ for all $v\in \mathrm{V}_0$ we have

$$\begin{aligned} \langle F,v\rangle &= \left|\int_\Omega fv\mathrm{d}x\right| \\ \text{(triangle inequality)} &\leq \textstyle\int_\Omega |fv|\mathrm{d}x \\ \text{(Cauchy-Schwartz inequality)} &\leq (\int_\Omega |f|^2\mathrm{d}x)^{\frac{1}{2}}(\int_\Omega |v|^2\mathrm{d}x)^{\frac{1}{2}} \\ &\leq ||f||_{\mathrm{L}_2}||v||_{\mathrm{L}_2} \\ &\leq ||f||_{\mathrm{L}_2}||v||_{\mathrm{H}^1} \\ \text{(with } C_2=||f||_{\mathrm{L}_2}) &\leq C_2||v||_{\mathrm{H}^1} \end{aligned}$$

2.3 Finite Element discretization

So, now we know that our variational problem is well posed and the next step to solve such a problem is discretization. The precess of discretization will be based on a Galerkin approximation. In that case, an approximate solution in sought in a finite-dimensional subspace of the space in which the weak (variational) formulation is posed. With this from our continuous problem we will get a discrete problem, for which the conditions of Lax-Milgram are satisfied, which means that the discrete problem is well posed as well.
We choose a finite-dimensional subspace $\mathrm{V}_h \subset \mathrm{V}_0$ and, after the homogenization of problem (2.4), the Galerkin approximation is the solution of the following problem: Find $u_h \in \mathrm{V}_h$ such that

$$a(u_h, v_h) = \langle F, v_h \rangle \quad \text{for all} \quad v_h \in \mathrm{V}_h. \tag{2.6}$$

If we now choose a basis for V_h, the discrete problem (2.6) will lead to a system of linear equations. Let $\{\varphi\}_{i=1}^N$ be a basis of V_h. Then, assume

$$u_h = \sum_{i=1}^{N} u_i \varphi_i \tag{2.7}$$

and choose $v_h = \varphi_j \quad \text{for} \quad j = 1, 2, ..., N$. Since $a(\cdot, \cdot)$ is bilinear, if we substitute in (2.6) the expression for u_h given by (2.7) and consider the special choice of v_h we will get

$$\sum_{i=1}^{N} a(\varphi_i, \varphi_j) u_i = \langle F, \varphi_j \rangle \quad j = 1, 2, ..., N. \tag{2.8}$$

Hence, we need to solve the linear system

$$K\bar{u} = \bar{f} \tag{2.9}$$

Matrix $K_h = [a(\varphi_i, \varphi_j)]$ is called the stiffness matrix, and the right hand side vector $\bar{f} = [\langle F, \varphi_j \rangle]$ is called the load vector. So, what we need to evaluate is the vector of unknowns, also called the vector of degrees of freedom, $\bar{u} = [u_i]$.

Now, the next question is that the subspace V_h has to be chosen in such a way that the functions in V can be accurately approximated with functions in V_h. Solving the system (2.9) is the next important step. So, another advantage could be if the V_h, or more precisely, the basis $\{\varphi\}_{i=1}^N$, is chosen in such a way, that the stiffness matrix K is as sparse as possible. A good choice is the method of Finite

Elements. In this method the construction of suitable spaces V_h relies on triangulation, which splits the domain Ω into small disjoint regions of simple geometric shape, such as triangles or quadrangles in $\mathbb{R}^2$. In that sense the subscript h refers to the characteristic size of these regions. Under certain assumptions, which prevent the triangulation from being degenerating (for instance, the size of the angles in triangles is bounded from below so that they do not become sharp), the finer the triangulation, the closer a finite element Galerkin solution to the exact solution.

Usually the functions in a finite element spaces V_h arise from a polynomial on the elements on the triangulation. Each polynomial defined on a given region is uniquely determined by its values (and maybe also the values of its derivatives) at some nodal points, usually the vertices of the region. So, every function in V_h is then determined by a set of values at nodal points. A simple example of finite element spaces is the space formed by continuous functions linear on triangles in $\mathbb{R}^2$. This space is refereed to P_1 and the elements of this space are uniquely determined by their function values at the vertices of the triangulation.

The standard basis for a finite element space is the one in which each of the basis functions has exactly one degree of freedom equal to 1 and the rest 0. In that case the unknowns in the linear system arising from the discretization are directly the degrees of freedom of the Galerkin approximation. The overlaps of the supports of the basis functions are small, which causes the stiffness matrix to be sparse. On the other hand in our case, due to the properties of the bilinear form, stiffness matrix is also symmetric and positive definite.

So, the first thing we will do is triangulation of our rectangular domain $\Omega = [0,2] \times [0,1]$. We take $h = \frac{1}{N}$ and create a uniform meshes with grid points $x_1, x_2, ..., x_i, ..., x_N$ in x-direction and $y_1, y_2, ..., y_j, ..., y_N$ in y-direction. By doing this we generate N^2 nodal points $\mathbf{x} = (x_i, y_j) \in \mathbb{R}^2, \quad i, j = 1, 2, ..., N$. We number these nodal points $\mathbf{x}_k, \quad k = 1, ..., N^2$ in a convenient way so that the stiffness matrix will turn out to be as sparse as possible (this is linked to the fact that the supports of basis functions overlap for finite number of triangles).

For the simplicity of the analysis for our model problem we will take linear basis functions which are defined by the following property:

$$\varphi_i(\mathbf{x}_j) = \begin{cases} 0 & \text{if} \quad i \neq j \\ 1 & \text{if} \quad i = j \end{cases}$$

If we have the basis functions, then we can compute the stiffness matrix and the

corresponding load vector. More detailed description of finite element method and some issues of its implementation can be found in [27]. And the final step would be to solve the resulting system using some solver.

This is so to say the "direct" approach for solving the discrete problem (2.6). Another approach could be to use a domain decomposition method [12] for the same problem. A natural question would be why do we need to use a domain decomposition method? The point is that in real applications the domain Ω is not as regular as it is in our model problem or the resulting linear system is quite large. The approach of domain decomposition methods is to divide the original problem into a number of smaller subproblems. Sometimes such a division arises from breaking up a complicated geometry, and sometimes the division is done in artificial manner. The subproblems are much smaller and easier to solve, and very often a parallel computation can be exploited.

The choice of discretization method and the choice of a method to solve the resulting linear system have to be done taking into account the nature of the problem. Issues like, efficiency and accuracy, strongly depend on the choice of the methods and most of the time there is a trade-off involved. What we are interested in this paper is a particular discretization method with a combination of a particular iterative method. In the coming chapters first we will introduce the HSS iterative method for solving linear systems and then we will present the domain decomposition method, called FETI, which later will be applied to our model problem. Our aim is to analyze how good is the HSS Iterative method applied to the system which comes from the FETI method. Though our model problem is quite simple, but hopefully it will allow us to draw some conclusions about the HSS iterative method as a potential method to solve systems arising form FETI.

Chapter 3

Hermitian and skew-Hermitian splitting methods

In this chapter we will study an iterative method for solving large non-Hermitian positive definite system of equations based on the Hermitian and skew-Hermitian (HSS) splitting of the coefficient matrix. This method was introduced in [2]. Both exact and the inexact versions of the method will be discussed. We will state the main results of the general method and then consider the method for a special case when the coefficient matrix is given as a saddle-point matrix. A preconditioning strategy for Krylov subspace methods [6], [28] based on the Hermitian and skew-Hermitian splitting of the coefficient matrix will be presented. We will conclude the chapter by considering recently introduced Accelerated Hermitian and skew-Hermitian method [14] and the corresponding convergence analysis will be given.

3.1 HSS iteration method

Linear systems of equation of form

$$\mathcal{A}x = b \tag{3.1}$$

where $\mathcal{A} \in \mathbb{R}^{n\times n}$ is nonsingular and $x, b \in \mathbb{R}^n$, appear in many areas of scientific computing, for instance after discretization of partial differential equations. Many iterative methods for solving linear systems are based on an efficient splitting of the coefficient matrix $\mathcal{A}$ [8]. We will study an iterative method based on particular splitting of $\mathcal{A}$ into its Hermitian and skew-Hermitian parts. As we consider the

case of real coefficient matrix, let us split matrix $\mathcal{A}$ into its symmetric and skew-symmetric parts:

$$\mathcal{A} = \mathcal{H} + \mathcal{S} \tag{3.2}$$

where $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T)$ is the symmetric part of $\mathcal{A}$ and $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$ is the skew-symmetric part. Using this splitting we consider the following two-step iteration, which is called HSS iteration:

Algorithm 1. ***The HSS iteration method.*** *Given an initial guess $x^{(0)}$, for $k = 0, 1, 2, ...,$ until $x^{(k)}$ converges, compute*

$$\begin{cases} (\alpha I + \mathcal{H})x^{(k+\frac{1}{2})} &= (\alpha I - \mathcal{S})x^{(k)} + b, \\ (\alpha I + \mathcal{S})x^{(k+1)} &= (\alpha I - \mathcal{H})x^{(k+\frac{1}{2})} + b, \end{cases} \tag{3.3}$$

where α is a given positive constant.

As it can be seen each iterate of the HSS iteration alternates between the symmetric part $\mathcal{H}$ and the skew-symmetric part $\mathcal{S}$. It should be mentioned that we can also reverse the roles of the matrices $\mathcal{H}$ and $\mathcal{S}$ in the HSS iteration in the sense that we first have to solve the system of linear equations with coefficient matrix $\alpha I + \mathcal{S}$ and afterwards the system of linear equations with coefficient matrix $\alpha I + \mathcal{H}$.

Before we study the convergence properties of the HSS iteration let us note that at each half-step of the iteration we need to solve exactly two $n \times n$ systems with matrices $\alpha I + \mathcal{H}$ and $\alpha I + \mathcal{S}$. But solving exactly the two subsystems is not practical. For this reason in actual applications one can use CG method to solve the first system with the coefficient matrix $\alpha I + \mathcal{H}$ and for the second system with matrix $\alpha I + \mathcal{S}$ one can employ some Krylov subspace method with some prescribed accuracy. Of course, there are other possible choices for the inner iteration, for instance multigrid methods or multilevel methods. The idea of solving the two subsystems iteratively leads to the inexact version of the HSS algorithm, which is called Inexact Hermitian and skew-Hermitian (IHSS) iteration method. Now we will state the main theorem of the current section:

Theorem 3.1. *Let $\mathcal{A} \in \mathbb{R}^{n\times n}$ be a positive definite matrix, let $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T)$ and $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$ be its symmetric and skew-symmetric parts, and let $\alpha > 0$. Then the iteration matrix $\mathcal{M}(\alpha)$ of the HSS iteration is given by*

$$\mathcal{M}(\alpha) = (\alpha I + \mathcal{S})^{-1}(\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}(\alpha I - \mathcal{S})$$

and its spectral radius $\rho(\mathcal{M}(\alpha))$ *is bounded by*

$$\sigma(\alpha) \equiv \max_{\lambda_i \in \lambda(\mathcal{H})} \frac{|\alpha - \lambda_i|}{|\alpha + \lambda_i|}$$

where $\lambda(\mathcal{H})$ *is the spectral set of the matrix* $\mathcal{H}$*. Therefore, it holds that*

$$\rho(\mathcal{M}(\alpha)) \leq \sigma(\alpha) < 1 \quad \forall \alpha > 0$$

i.e., the HSS iteration converges to the unique solution $x^* \in \mathbb{R}^n$ *of the system of linear equations* $\mathcal{A}x = b$*.*

Proof[2]. Let us first rewrite the two-step HSS iteration method in a fixed point form. For this we eliminate $x^{(k+\frac{1}{2})}$ from the second equation of (3.3) by using the first one. We will get:

$$\mathbf{x}^{k+1} = \mathcal{T}_\alpha \mathbf{x}^k + c, \tag{3.4}$$

where

$$\mathcal{T}_\alpha := (\mathcal{S} + \alpha I)^{-1}(\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}(\alpha I - \mathcal{S}), \tag{3.5}$$

and

$$c := (\mathcal{S} + \alpha I)^{-1}[I + (\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}]\mathbf{b}. \tag{3.6}$$

Now, from the general theory of iterative methods we know that in order to show that the method converges it is enough to show that $\rho(\mathcal{M}(\alpha)) < 1$. Let us first note that, as $\mathcal{A}$ is nonsingular and $\alpha > 0$, the matrices $\alpha I + \mathcal{H}$ and $\alpha I + \mathcal{S}$ are nonsingular, therefore their inverse exists, and the matrix $\mathcal{M}(\alpha)$ given by (3.5) is well defined.
Observe that $\mathcal{M}(\alpha)$ is similiar to $\mathcal{B}(\alpha)$, which is given by

$$\mathcal{B}(\alpha) = (\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}$$

Indeed, it is ease to see that $\mathcal{M}(\alpha) = (\alpha I + \mathcal{S})^{-1}\mathcal{B}(\alpha)(\alpha I + \mathcal{S})$. Therefore, from the similarity invariance of the matrix spectrum, we have $\rho(\mathcal{M}(\alpha)) = \rho(\mathcal{B}(\alpha))$. Hence

$$\begin{aligned}
\rho(\mathcal{M}(\alpha)) &= \rho(\mathcal{B}(\alpha)) \\
&= \rho((\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}) \\
&\leq ||(\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}||_2 \\
&\leq ||(\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}||_2 ||(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}||_2
\end{aligned}$$

Now, consider $\mathcal{Q}(\alpha) = (\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}$. Having that $\mathcal{S}^T = -\mathcal{S}$ and matrices $(\alpha I + \mathcal{S})$ and $(\alpha I - \mathcal{S})$ commute, we can write that

$$\begin{aligned}
\mathcal{Q}(\alpha)^T\mathcal{Q}(\alpha) &= ((\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1})^T(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1} \\
&= ((\alpha I + \mathcal{S})^{-1})^T(\alpha I - \mathcal{S})^T(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1} \\
&= ((\alpha I + \mathcal{S})^T)^{-1}(\alpha I - \mathcal{S})^T(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1} \\
&= (\alpha I - \mathcal{S})^{-1}(\alpha I + \mathcal{S})(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1} \\
&= (\alpha I - \mathcal{S})^{-1}(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})(\alpha I + \mathcal{S})^{-1} \\
&= I.
\end{aligned}$$

Therefore, matrix $\mathcal{Q}(\alpha)$ is an orthogonal matrix ($\mathcal{Q}(\alpha)$ is also called Cayley transform of $\mathcal{S}$). This means that $||\mathcal{Q}(\alpha)||_2 = 1$. So

$$\begin{aligned}
\rho(\mathcal{M}(\alpha)) &\leq ||(\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}||_2||(\alpha I - \mathcal{S})(\alpha I + \mathcal{S})^{-1}||_2 \\
&\leq ||(\alpha I - \mathcal{H})(\alpha I + \mathcal{H})^{-1}||_2 \\
&= \max_{\lambda_i \in \lambda(\mathcal{H})} \frac{|\alpha - \lambda_i|}{|\alpha + \lambda_i|}
\end{aligned}$$

Since $\lambda_i > 0 (i = 1, 2, ..., n)$ and α is a positive constant, we can conclude that

$$\rho(\mathcal{M}(\alpha)) \leq \sigma(\alpha) < 1$$

which proves the theorem.
□

We should remark that from theorem 3.1 we can see that the convergence speed of the HSS iteration is bounded by $\sigma(\alpha)$, which depends only on the spectrum of the symmetric part $\mathcal{H}$, but does not depend on the spectrum of the skew-symmetric part, on the spectrum of $\mathcal{A}$ or on eigenvectors of $\mathcal{H}$,$\mathcal{S}$ and $\mathcal{A}$. Another remark is that if the maximum and minimum eigenvalues of $\mathcal{H}$ are known, then the optimal parameter α for $\sigma(\alpha)$, which is the upper bound of $\rho(\mathcal{M}(\alpha))$, can be obtained. This is given in the following corollary.

Corollary 3.2. *Let $\mathcal{A} \in \mathbb{R}^{n\times n}$ be positive definite matrix, let $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T)$ and $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$ be its symmetric and skew-symmetric parts, and let $\gamma_{\min}$ and $\gamma_{\max}$ be the minimum and maximum eigenvalues of the matrix $\mathcal{H}$ respectively, and let α be a positive constant. Then*

$$\alpha^* \equiv \arg\min_{\alpha} \left\{ \max_{\gamma_{\min} \leq \lambda \leq \gamma_{\max}} \frac{|\alpha - \lambda|}{|\alpha + \lambda|} \right\} = \sqrt{\gamma_{\min}\gamma_{\max}}$$

and

$$\sigma(\alpha^*) = \frac{\sqrt{\gamma_{\max}} - \sqrt{\gamma_{\min}}}{\sqrt{\gamma_{\max}} + \sqrt{\gamma_{\min}}} = \frac{\sqrt{\kappa(\mathcal{H})} - 1}{\sqrt{\kappa(\mathcal{H})} + 1}$$

where $\kappa(\mathcal{H})$ is the spectral condition number of $\mathcal{H}$.

Note that in the above corollary the optimal parameter α^* minimizes only the upper bound $\sigma(\alpha)$ of the spectral radius itself. It is clear that the asymptotic rate of the convergence of the alternating iteration heavily depends on the spectral radius of the iteration matrix $\mathcal{T}_\alpha$, so it makes sense to try to find α such that $\rho(\mathcal{T}_\alpha)$ is as small as possible. In general finding such α is a difficult problem.

3.2 HSS as a preconditioner for Krylov subspace methods

Even with the optimal choice of α the convergence of stationary iteration (3.3) is typically slow for the method to be competitive. In this small subsection we will show that the iterative method can be used in order to accelerate the convergence of other efficient iterative methods such as Krylov subspace methods [6], [28]. For this let us consider the following corollary.

Corollary 3.3. *There is a unique splitting $\mathcal{A} = \mathcal{M} - \mathcal{N}$ with $\mathcal{M}$ non singular such that the iteration matrix $\mathcal{T}_\alpha$ is the matrix induced by the splitting,namely, $\mathcal{T}_\alpha = \mathcal{M}^{-1}\mathcal{N} = \mathcal{I} - \mathcal{M}^{-1}\mathcal{A}$. An easy calculation shows that*

$$\mathcal{M} \equiv \mathcal{M}_\alpha = \frac{1}{2\alpha}(\alpha I + \mathcal{H})(\alpha I + \mathcal{S}) \tag{3.7}$$

It is therefore possible to rewrite the iteration (3.3) in the correction form:

$$x^{k+1} = x^k + \mathcal{M}_\alpha^{-1} r^k, \quad r^k = b - \mathcal{A}x^k$$

From the corollary above it follows that the linear system (3.1) is equivalent to (has the same solution as) the linear system

$$(I - \mathcal{T}_\alpha)x = \mathcal{M}_\alpha^{-1}\mathcal{A}x = c$$

where $c = \mathcal{M}_\alpha^{-1}b$. In other words, this equivalent system is a preconditioned system which can be solved with Krylov subspace method like GMRES to accelerate the convergence of the iteration. Hence, the matrix $\mathcal{M}_\alpha$ can be seen as a preconditioner

for GMRES. Equivalently, we can say that GMRES can be used to accelerate the convergence of the alternating iteration applied to $\mathcal{A}x = b$.

The factor $\frac{1}{2\alpha}$ in (3.7) has no effect on the preconditioned system and as a preconditioner we can use $\mathcal{M}_\alpha = (\alpha I + \mathcal{H})(\alpha I + \mathcal{S})$. Now, under the assumptions of Theorem 3.1 , since $\mathcal{M}_\alpha^{-1}\mathcal{A} = I - \mathcal{T}_\alpha$, we can see that for all $\alpha > 0$ the eigenvalues of the preconditioned matrix $\mathcal{M}_\alpha^{-1}\mathcal{A}$ (or the right preconditioned $\mathcal{A}\mathcal{M}_\alpha^{-1}$) are entirely contained in the open disc of radius 1 centered at (0,1). In particular, the preconditioned system is positive stable. Note that the smaller the spectral radius of $\mathcal{T}_\alpha$, the more clustered the eigenvalues of the preconditioned matrix (around 1). Very often a clustered spectrum leads to a rapid convergence of GMRES.

Choosing α in a way that the spectral radius of the iteration matrix is minimized does not necessarily imply that the same α is the best choice if the algorithm is used as a preconditioner for a Krylov subspace method. In certain problems in can be shown that when α is chosen sufficiently small the alternating iteration results in h-independent preconditioner for GMRES, and the spectral radius is very close to 1 [29].

Also, minimizing the spectral radius, or even the number of GMRES iterations, does not imply optimal performance in CPU time. It is clear that an efficient implementation of the method requires that two subsystems in (3.3) be solved inexactly. And the choice of α can influence the cost of the solves for corresponding subsystems. Large values of α will make the iterative solution of subsystems easy. But on the other hand if $\alpha \longrightarrow \infty$ as well as when $\alpha \longrightarrow 0$ then the nonzero eigenvalues of iteration matrix $\mathcal{T}_\alpha$ approach 1. In that case the convergence of the outer iteration will slow down. So we see that like in many cases one has to face a tradeoff approach. So if we define the "optimal" value of α as the one that minimizes the total amount of work required for computing an approximate solution, this will not necessary be the same as the value of α that minimizes the number of outer iterations.

3.3 IHSS iteration method

As mentioned before solving the two subsystems of (3.3) exactly is quite impractical and it is reasonable to solve the subsystems using some iterative methods. This results in the following IHSS iteration for solving linear equations (3.1).

Algorithm 2. ***The IHSS iteration method*** *Given an initial guess $x^{(0)}$, for $k = 0, 1, 2, ...$ and given tolerances η_k, ε_k until $\{x^{(k)}\}$ converges,*

1. *approximate the solution of $(\alpha I + \mathcal{H})z^{(k)} = r^{(k)}$ where $r^{(k)} = b - \mathcal{A}x^{(k)}$ by iterating until $z^{(k)}$ is such that the residual $p^{(k)} = r^{(k)} - (\alpha I + \mathcal{H})z^{(k)}$ satisfies*

$$||p^{(k)}|| \leq \varepsilon_k ||r^{(k)}||$$

and then compute $x^{(k+\frac{1}{2})} = x^{(k)} + z^{(k)}$

2. *approximate the solution of $(\alpha I + \mathcal{S})z^{(k+\frac{1}{2})} = r^{(k+\frac{1}{2})}$ where $r^{(k+\frac{1}{2})} = b - \mathcal{A}x^{(k+\frac{1}{2})}$ by iterating until $z^{(k)}$ is such that the residual $q^{(k+\frac{1}{2})} = r^{(k+\frac{1}{2})} - (\alpha I + \mathcal{H})z^{(k+\frac{1}{2})}$ satisfies*

$$||q^{(k+\frac{1}{2})}|| \leq \eta_k ||r^{(k+\frac{1}{2})}||$$

and then compute $z^{(k+1)} = x^{(k+\frac{1}{2})} + z^{(k+\frac{1}{2})}$

In [2] the analysis of IHSS in slightly general terms is given, and, as a main result, the following theorem is stated.

Theorem 3.4. *Let $\mathcal{A} \in \mathbb{R}^{n\times n}$ be a positive definite matrix, let $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T)$ and $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$ be its symmetric and skew-symmetric parts, and let $\alpha > 0$. If $\{x^k\}$ is an iterative sequence generated by the IHSS iteration method and if x^* is the exact solution of the system of linear equations (3.1), then it holds that*

$$|||x^{k+1} - x^*||| \leq (\sigma(\alpha) + \theta\rho\eta_k)(1 + \theta\epsilon_k)|||x^k - x^*|||, \qquad k = 0, 1, 2...,$$

where

$$\rho = ||(\alpha I + \mathcal{S})(\alpha I + \mathcal{H})^{-1}||_2, \qquad \theta = ||\mathcal{A}(\alpha I + \mathcal{S})^{-1}||_2$$

and

$$|||x||| \quad \textit{is defined as} \quad |||x||| := ||(\alpha I + \mathcal{S})x||_2 \quad \forall x \in \mathbb{C}^n.$$

In particular, if $(\sigma(\alpha) + \theta\rho\eta_{\max})(1 + \theta\epsilon_{\max}) < 1$, then the iterative sequence $\{x^k\}$ converges to x^*, where $\epsilon_{\max} = \max_k\{\epsilon_k\}$ and $\eta_{\max} = \max_k\{\eta_k\}$.

It should be mentioned that the tolerances ϵ_k and η_k are not required to approach zero as k increases in order to get the convergence of IHSS iteration, but are required to approach zero in order to asymptotically recover the original convergence rate of the HSS iteration. The following theorem presents one possible way of choosing the tolerances ϵ_k and η_k such that the original convergence rate of the two-step splitting iterative scheme can be asymptotically recovered [2].

Theorem 3.5. *Let the assumptions in Theorem 3.4 be satisfied. Suppose that both* $\{\tau_1(k)\}$ *and* $\{\tau_2(k)\}$ *are nondecreasing and positive sequences such that* $\tau_1(k) \geq 1$ *,* $\tau_2(k) \geq 1$ *and* $\lim_{k\longrightarrow\infty} \sup \tau_1(k) = \lim_{k\longrightarrow\infty} \sup \tau_2(k) = +\infty$ *, and that both* δ_1 *and* δ_2 *are real positive constants in the interval* $(0, 1)$ *satisfying*

$$\epsilon_k \leq c_1\delta_1^{\tau_1(k)} \quad \text{and} \quad \eta_k \leq c_2\delta_2^{\tau_1(k)} \quad k = 0, 1, 2, ...,$$

with c_1 *and* c_2 *nonnegative constants. Then it holds that*

$$|||x^{k+1} - x^*||| \leq (\sqrt{\sigma(\alpha)} + \omega\theta\delta^{\tau(k)})^2|||x^k - x^*|||, \qquad k = 0, 1, 2, ...,$$

where ρ *and* θ *are defined as in Theorem 3.4 and* $\tau(k)$ *,* δ *,* ω *are defined as*

$$\tau(k) = \min\{\tau_1(k), \tau_2(k)\}, \;\; \delta = \max\{\delta_1, \delta_2\}, \quad \omega = \max\left\{\sqrt{c_1c_2\rho}, \;\; \frac{1}{2\sqrt{\sigma(\alpha)}}(c_1\sigma(\alpha) + c_2\rho)\right\}.$$

In particular, we have

$$\lim_{k\longrightarrow\infty} \sup \frac{|||x^{k+1} - x^*|||}{|||x^k - x^*|||} = \sigma(\alpha)$$

which means that the convergence rate of the IHSS iteration method is asymptotically the same as that of the HSS iteration method.

3.4 HSS for generalized saddle-point problems

In the previous subsections we have considered the HSS/ IHSS method for solving general linear system of equations. In this subsection we will specify the linear system, namely we will consider the case when the system is given as a saddle-point system. Such systems appear quite often in applications, including computational fluid dynamics [15], [16], [24], [17], mixed finite element approximation of elliptic PDEs [18], [19], [25], electrical networks, optimization [20], [23], [22], [21], and solving them efficiently is an important issue. Nowadays there are many methods available for solving such systems. A very detailed and interesting overview on saddle-points problems and solvers for such systems is given in [11].
We consider the solution of system of linear equations with the following 2×2 block structure:

$$\begin{bmatrix} A & B^T \\ B & -C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \tag{3.8}$$

with $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{m\times n}$, $C \in \mathbb{R}^{m\times m}$, $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ and $m \leq n$. We will assume that:

- A has positive semidefinite symmetric part $H = \frac{1}{2}(A + A^T)$,
- $\text{rank}(B) = m$,
- $\ker(H) \cap \ker(B) = \{0\}$,
- C is symmetric positive semidefinite.

These assumptions guarantee existence and uniqueness of the solution. Very often A is symmetric positive definite but in some cases A is either symmetric and singular (i.e., only positive semidefinite), or nonsymmetric with positive definite symmetric part H. When A is symmetric positive (semi-)definite, the coefficient matrix in (3.8) is symmetric indefinite, and indefinite solvers can be used. Alternatively, it is possible instead of (3.8) to solve the equivalent nonsymmetric system

$$\begin{bmatrix} A & B^T \\ -B & C \end{bmatrix} \begin{bmatrix} u \\ p \end{bmatrix} = \begin{bmatrix} f \\ -g \end{bmatrix} \tag{3.9}$$

or

$$\mathcal{A}\mathbf{x} = \mathbf{b}$$

where $\mathcal{A} = \begin{bmatrix} A & B^T \\ -B & C \end{bmatrix}$, $\mathbf{x} = [u^T, p^T]^T$ and $\mathbf{b} = [f^T, -g^T]^T$.

In the following theorem we will summarize some properties of $\mathcal{A}$ which are quite useful and independent of the fact weather A is symmetric or not.

Theorem 3.6. *Let $\mathcal{A} \in \mathbb{R}^{(n+m)\times(n+m)}$ be the coefficient matrix in (3.9). Assume $H = \frac{1}{2}(A + A^T)$ is positive semidefinite, B has full rank, $C = C^T$ is positive semidefinite, and $\ker(H) \cap \ker(B) = \{0\}$. Let $\sigma(\mathcal{A})$ denote the spectrum of $\mathcal{A}$. Then*

- *(a) $\mathcal{A}$ is nonsingular,*
- *(b) $\mathcal{A}$ is semipositive real: $< \mathcal{A}\boldsymbol{v}, \boldsymbol{v} >= \boldsymbol{v}^T\mathcal{A}\boldsymbol{v} \geq 0$ for all $\boldsymbol{v} \in \mathbb{R}^{(n+m)}$,*
- *(c) $\mathcal{A}$ is positive semistable, that is , the eigenvalues of $\mathcal{A}$ nonnegative real part: $\Re(\lambda) \geq 0$ for all $\lambda \in \sigma(\mathcal{A})$,*
- *(d) If, in addition, $H = \frac{1}{2}(A + A^T)$ is positive definite, then $\mathcal{A}$ is positive stable: $\Re(\lambda) > 0$ for all $\lambda \in \sigma(\mathcal{A})$.*

Proof [29].

- (a)
  Let $\mathbf{x} = [u^T, p^T]^T$ be such that $\mathcal{A}\mathbf{x} = 0$. Then

$$Au + B^T p = 0 \quad \text{and} \quad -Bu + Cp = 0 \tag{3.10}$$

  Now, $\mathcal{A}\mathbf{x} = 0$ implies that $\mathbf{x}^T\mathcal{A}\mathbf{x} = u^T Au + p^T Cp = 0$. This and the fact that both $u^T Au$ and $p^T Cp$ are nonnegative imply $u^T Au = p^T Cp = 0$. On the other hand, $u^T Au = u^T Hu = 0$, which means that $u \in \ker(H)$ since H is symmetric positive semidefinite. In the same way, $p^T Cp = 0$ and C being positive semidefinite implies $Cp = 0$. The second equation in (3.10) and $Cp = 0$ gives $Bu = 0$. So, we get that $u \in \ker(B)$. Since $\ker(H) \cap \ker(B) = \{0\}$, we can conclude that $u = 0$. Then, from the first equation in (3.10) we can see that $B^T p = 0$. As B has full column rank, $B^T p = 0$ implies $p = 0$. Therefore, $\mathcal{A}\mathbf{x} = 0$ has the only solution $\mathbf{x} = 0$, which means that $\mathcal{A}$ is nonsingular.

- (b)
  Now, for any $\mathbf{v} \in \mathbb{R}^{(n+m)}$ we have $\mathbf{v}^T\mathcal{A}\mathbf{v} = \mathbf{v}^T\mathcal{H}\mathbf{v}$ where $\mathcal{H}$ is the symmetric part of $\mathcal{A}$, namely

$$\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T) = \begin{bmatrix} H & O \\ O & C \end{bmatrix}.$$

  As H and C are positive semidefinite ,then clearly $\mathcal{H} = \begin{bmatrix} H & O \\ O & C \end{bmatrix}$ is positive semidefinite, hence $\mathbf{v}^T\mathcal{A}\mathbf{v} \geq 0$.

- (c)
  Let $(\lambda, \mathbf{v})$ be an eigenpair of $\mathcal{A}$ with $||\mathbf{v}||_2 = 1$.Then $\mathbf{v}^*\mathcal{A}\mathbf{v} = \lambda$ and $(\mathbf{v}^*\mathcal{A}\mathbf{v})^* = \mathbf{v}^*\mathcal{A}\mathbf{v} = \bar{\lambda}$.Now, on the other hand

$$\mathbf{v}^*(\mathcal{A} + \mathcal{A}^T)\mathbf{v} = \Re(\mathbf{v})^T(\mathcal{A} + \mathcal{A}^T)\Re(\mathbf{v}) + \Im(\mathbf{v})^T(\mathcal{A} + \mathcal{A}^T)\Im(\mathbf{v}).$$

  We have already showed that $\mathcal{A}$ is semipositive real therefore the quantity $\Re(\mathbf{v})^T(\mathcal{A} + \mathcal{A}^T)\Re(\mathbf{v}) + \Im(\mathbf{v})^T(\mathcal{A} + \mathcal{A}^T)\Im(\mathbf{v})$ is nonnegative. Hence,

$$\Re(\lambda) = \frac{1}{2}\mathbf{v}^*(\mathcal{A} + \mathcal{A}^T)\mathbf{v} \geq 0.$$

- (d)
  Assume that $(\lambda, \mathbf{v})$ be an eigenpair of $\mathcal{A}$ with $\mathbf{v} = [u^T, p^T]^T$. Then

$$\Re(\lambda) = u^*Hu + p^*Cp = \Re(u)^T H\Re(u) + \Im(u)^T H\Im(u) + \Re(p)^T C\Re(p) + \Im(p)^T C\Im(p).$$

Since H is assumed to be positive definite, and the quantity above is nonnegative, then $\Re(\lambda)$ can be zero if and only if $u = 0$. But if $u = 0$ then, from the second equation in (3.10), we get $B^T p = 0$. Hence $p = 0$, as B has full column rank. So, we have that $u = p = 0$, which means that $\mathbf{v} = [u^T, p^T]^T = 0$. But $\mathbf{v}$ is assumed to be an eigenvector, therefore we get a contradiction. This implies that $\Re(\lambda) > 0$.

□

The theorem above shows that changing the sign of the last m equations in (3.8) might effect the fact that $\mathcal{A}$ is symmetric (in case if A is symmetric), but we gain positive (semi)-definitness. Having the coefficient matrix positive (semi)-definite can be an advantage when using Krylov subspace methods, for instance GMRES. So, under the conditions of the theorem the existence of the solution is guaranteed and some advantageous properties of the coefficient matrix are provided.
Now, let us solve the linear system (3.9) by using the HSS iterative method introduced in the first subsection of current chapter. According to the algorithm of the two-step iteration, what we need to solve at each iteration is the following system

$$\begin{cases} (\alpha I + \mathcal{H})x^{(k+\frac{1}{2})} &= (\alpha I - \mathcal{S})x^{(k)} + b, \\ (\alpha I + \mathcal{S})x^{(k+1)} &= (\alpha I - \mathcal{H})x^{(k+\frac{1}{2})} + b, \end{cases} \tag{3.11}$$

where $\mathcal{H}$ and $\mathcal{S}$ are the symmetric and skew-symmetric parts of $\mathcal{A}$ respectively, namely

$$\mathcal{A} = \begin{bmatrix} A & B^T \\ -B & C \end{bmatrix} = \begin{bmatrix} H & O \\ O & C \end{bmatrix} + \begin{bmatrix} S & B^T \\ -B & O \end{bmatrix} = \mathcal{H} + \mathcal{S},$$

where:

- $\mathcal{H} = \frac{1}{2}(\mathcal{A} + \mathcal{A}^T)$,
- $\mathcal{S} = \frac{1}{2}(\mathcal{A} - \mathcal{A}^T)$,
- $H = \frac{1}{2}(A + A^T)$,
- $S = \frac{1}{2}(A - A^T)$.

Now, as $\mathcal{H} = \begin{bmatrix} H & O \\ O & C \end{bmatrix}$, we can conclude that $\mathcal{A}$ will be positive real (i.e.,$\mathcal{H}$ is symmetric positive definite) if and only if both H and C are symmetric positive definite (SPD). Unfortunately, that is not the case in most of the applications. So the convergence theory discussed in first subsection is not applicable as in the context of

generalized saddle-point problems the matrix $\mathcal{H}$ is only positive semidefinite and, in general, singular. In this case more detailed analysis is required, since for matrices whose symmetric part is positive semidefinite and singular, the alternating iteration is not convergent in general. For the analysis we will consider the fixed point formulation of the HSS iteration:

$$\mathbf{x}^{k+1} = \mathcal{T}_\alpha \mathbf{x}^k + c, \tag{3.12}$$

where

$$\mathcal{T}_\alpha := (\mathcal{S} + \alpha I)^{-1}(\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}(\alpha I - \mathcal{S}) \tag{3.13}$$

and

$$c := (\mathcal{S} + \alpha I)^{-1}[I + (\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}]\mathbf{b}. \tag{3.14}$$

The following theorem shows that for a large class of generalized saddle-point problems the alternating iteration converges.

Theorem 3.7. *Consider problem (3.9) and assume that A is positive real, C symmetric positive semidefinite, and B has full rank. Then the iteration (3.11) is unconditionally convergent; that is, $\rho(\mathcal{T}_\alpha) < 1$ for all $\alpha > 0$.*

Proof [29]. Consider the matrix

$$\mathcal{K}_\alpha := (\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}(\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1} = \mathcal{R}\mathcal{U},$$

with $\mathcal{R} := (\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1}$ and $\mathcal{U} := (\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1}$. Let us first notice that $\mathcal{T}_\alpha$ is similar to $\mathcal{K}_\alpha$. Indeed, it is easy to see that

$$\mathcal{T}_\alpha = (\mathcal{S} + \alpha I)^{-1}\mathcal{K}_\alpha(\mathcal{S} + \alpha I).$$

Next, since $(\alpha I - \mathcal{H})$ and $(\mathcal{H} + \alpha I)^{-1}$ commute, and $\mathcal{H}$ is symmetric, it is easy to show that $\mathcal{R}$ is symmetric. Similarly, since $(\alpha I - \mathcal{S})$and $(\mathcal{S} + \alpha I)^{-1}$ commute and $\mathcal{S}^T = -\mathcal{S}$, one can show that $\mathcal{U}$ is orthogonal matrix. Indeed,

$$\begin{aligned}
\mathcal{R}^T &= ((\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1})^T \\
&= (\mathcal{H} + \alpha I)^{-T}(\alpha I - \mathcal{H})^T \\
&= (\mathcal{H} + \alpha I)^{-1}(\alpha I - \mathcal{H}) \\
&= (\alpha I - \mathcal{H})(\mathcal{H} + \alpha I)^{-1} \\
&= \mathcal{R},
\end{aligned}$$

therefore $\mathcal{R}^T = \mathcal{R}$. Now, to show that $\mathcal{U}$ is orthogonal we need to show that $\mathcal{U}\mathcal{U}^T = I$.

$$\begin{aligned}
\mathcal{U}\mathcal{U}^T &= (\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1}((\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1})^T \\
&= (\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1}(\mathcal{S} + \alpha I)^{-T}(\alpha I - \mathcal{S})^T \\
&= (\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1}(-\mathcal{S} + \alpha I)^{-1}(\alpha I + \mathcal{S}) \\
&= (\mathcal{S} + \alpha I)^{-1}(\alpha I - \mathcal{S})(-\mathcal{S} + \alpha I)^{-1}(\alpha I + \mathcal{S}) \\
&= (\mathcal{S} + \alpha I)^{-1}(\alpha I + \mathcal{S}) \\
&= I.
\end{aligned}$$

Now, since $\mathcal{R}$ is symmetric, then it is orthogonally similar to $(n+m) \times (n+m)$ diagonal matrix $\mathcal{D}$ given by:

$$\mathcal{D} = \begin{bmatrix}
\frac{\alpha-\mu_1}{\alpha+\mu_1} & & & & & & & & \\
& \frac{\alpha-\mu_2}{\alpha+\mu_2} & & & & & & & \\
& & \ddots & & & & & & \\
& & & \frac{\alpha-\mu_n}{\alpha+\mu_n} & & & & & \\
& & & & \frac{\alpha-\upsilon_1}{\alpha+\upsilon_1} & & & & \\
& & & & & \frac{\alpha-\upsilon_2}{\alpha+\upsilon_2} & & & \\
& & & & & & \ddots & & \\
& & & & & & & \frac{\alpha-\upsilon_m}{\alpha+\upsilon_m}
\end{bmatrix}$$

where $\mu_1, \mu_2, ..., \mu_n$ are the (positive) eigenvalues of H and $\upsilon_1, \upsilon_2, ..., \upsilon_m$ are the (nonnegative) eigenvalues of C. This means that there is an orthogonal matrix $\mathcal{V}$ of order $(n+m)$ such that

$$\mathcal{V}^T\mathcal{R}\mathcal{V} = \mathcal{D} = \begin{bmatrix} D_1 & O \\ O & D_2 \end{bmatrix},$$

where D_1 and D_2 are diagonal matrices of order n and m respectively. Observe that

$$\frac{|\alpha - \mu_i|}{|\alpha + \mu_i|} < 1 \quad \text{for} \quad 1 \leq i \leq n \quad \text{and} \quad \frac{|\alpha - \upsilon_i|}{|\alpha + \upsilon_i|} < 1 \quad \text{for} \quad 1 \leq i \leq m.$$

Now, since $\mathcal{V}$ is orthogonal we can see that $\mathcal{R}\mathcal{U}$ is orthogonally similar to $(\mathcal{V}^T\mathcal{R}\mathcal{V})(\mathcal{V}^T\mathcal{U}\mathcal{V})$. Indeed,

$$\mathcal{V}^T\mathcal{R}\mathcal{U}\mathcal{V} = (\mathcal{V}^T\mathcal{R}\mathcal{V})(\mathcal{V}^T\mathcal{U}\mathcal{V}) = \mathcal{D}\mathcal{Q},$$

where $\mathcal{Q} := \mathcal{V}^T\mathcal{U}\mathcal{V}$, as a product of orthogonal matrices, is an orthogonal matrix. We had that the iteration matrix $\mathcal{T}_\alpha$ is similar to $\mathcal{R}\mathcal{U}$ and, since $\mathcal{R}\mathcal{U}$ is similar to $\mathcal{D}\mathcal{Q}$, we can conclude that $\mathcal{T}_\alpha$ is similar to $\mathcal{D}\mathcal{Q}$. This means:

$$\rho(\mathcal{T}_\alpha) = \rho(\mathcal{D}\mathcal{Q}) = \rho(\mathcal{Q}\mathcal{D}).$$

If we manage to show that $\rho(\mathcal{QD}) < 1$ for all $\alpha > 0$ then the theorem will be proved. For this let us partition $\mathcal{Q}$ in the following way:

$$\mathcal{Q} = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}.$$

Then the matrix product $\mathcal{QD}$ can be written as:

$$\mathcal{QD} = \begin{bmatrix} Q_{11}D_1 & Q_{12}D_2 \\ Q_{21}D_1 & Q_{22}D_2 \end{bmatrix}.$$

Now, assume that $\lambda \in \mathbb{C}$ is an eigenvalue of $\mathcal{QD}$, and $\mathbf{x} \in \mathbb{C}^{n+m}$ is the corresponding eigenvector with $||\mathbf{x}||_2 = 1$. We will assume that $\lambda \neq 0$, otherwise the proof is finished. We want to show that $|\lambda| < 1$. Now,

$$\mathcal{QD}\mathbf{x} = \lambda\mathbf{x} \quad \text{implies} \quad \mathcal{D}\mathbf{x} = \lambda\mathcal{Q}^T\mathbf{x},$$

therefore, by taking the norms and recalling that $\mathcal{Q}$ is an orthogonal matrix, we will get:

$$||\mathcal{D}\mathbf{x}||_2 = |\lambda|||\mathcal{Q}^T\mathbf{x}||_2 = |\lambda|||\mathbf{x}||_2 = |\lambda|.$$

Hence

$$|\lambda|^2 = ||\mathcal{D}\mathbf{x}||_2^2 = \sum_{i=1}^{n}(\frac{\alpha-\mu_i}{\alpha+\mu_i})^2 x_i\bar{x}_i + \sum_{i=n+1}^{n+m}(\frac{\alpha-\upsilon_i}{\alpha+\upsilon_i})^2 x_i\bar{x}_i \leq ||\mathbf{x}||_2 = 1. \tag{3.15}$$

So, we showed that $|\lambda| \leq 1$. To prove that $|\lambda| < 1$ (strictly), we will show that there exists at least one $i(1 \leq i \leq n)$ such that $x_i \neq 0$. For this let us assume that $x_i = 0$ for all $i(1 \leq i \leq n)$. Then the eigenvector $\mathbf{x}$ is of the form $\mathbf{x} = \begin{bmatrix} 0 \\ \hat{x} \end{bmatrix}$ where $\hat{x} \in \mathbb{C}^m$. Therefore,

$$\mathcal{QD}\mathbf{x} = \begin{bmatrix} Q_{11}D_1 & Q_{12}D_2 \\ Q_{21}D_1 & Q_{22}D_2 \end{bmatrix}\begin{bmatrix} 0 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} Q_{12}D_2\hat{x} \\ Q_{22}D_2\hat{x} \end{bmatrix} = \begin{bmatrix} 0 \\ \lambda\hat{x} \end{bmatrix}, \tag{3.16}$$

which means that $Q_{12}D_2\hat{x} = 0$. Let us for a moment assume that Q_{12} has full column rank. Then $Q_{12}D_2\hat{x} = 0$ implies that $D_2\hat{x} = 0$. But according to (3.16) $\lambda\hat{x} = Q_{22}D_2\hat{x} = 0$. Since we have assumed that $\lambda \neq 0$ it must be $\hat{x} = 0$. Therefore $\mathbf{x} = \begin{bmatrix} 0 \\ \hat{x} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} = 0$, which is a contradiction, as $\mathbf{x}$ is an eigenvector. This means that if $x_i = 0$ for all $i(1 \leq i \leq n)$ then $\mathbf{x} = 0$, therefore there exists at least one $i(1 \leq i \leq n)$ such that $x_i \neq 0$.

To finalize the proof we only need to show that Q_{12} has full column rank. Now, let us recall that $\mathcal{Q} = \mathcal{V}^T\mathcal{U}\mathcal{V}$ with

$$\mathcal{V} = \begin{bmatrix} V_{11} & O \\ O & V_{22} \end{bmatrix},$$

where $V_{11} \in \mathbb{R}^{n\times n}$ is the orthogonal matrix that diagonalizes $(\alpha I_n - \mathcal{H})(\alpha I_n + \mathcal{H})^{-1}$ and $V_{22} \in \mathbb{R}^{m\times m}$ is the orthogonal matrix that diagonalizes $(\alpha I_n - \mathcal{C})(\alpha I_n + \mathcal{C})^{-1}$. Also let us recall that

$$\mathcal{U} = (\alpha I - \mathcal{S})(\mathcal{S} + \alpha I)^{-1} \begin{bmatrix} \alpha I_n - S & -B^T \\ B & \alpha I_m \end{bmatrix} \begin{bmatrix} \alpha I_n + S & B^T \\ -B & \alpha I_m \end{bmatrix} = \begin{bmatrix} U_{11} & U_{12} \\ U_{21} & U_{22} \end{bmatrix}.$$

One can show that

$$U_{12} = [(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n]B^T[\alpha I_m + B(\alpha I_n + S)^{-1}B^T]^{-1}.$$

Let us now show that -1 can not be an eigenvalue of an orthogonal matrix $(\alpha I_n - S)(\alpha I_n + S)^{-1}$. Indeed, if for some $(-1, \mathbf{x})$ with $\mathbf{x} \in \mathbb{C}^n$ is an eigenpair of $(\alpha I_n - S)(\alpha I_n + S)^{-1}$ then

$$\begin{array}{rcll} (\alpha I_n - S)(\alpha I_n + S)^{-1}\mathbf{x} & = & -\mathbf{x} & \Leftrightarrow \\ (\alpha I_n + S)^{-1}(\alpha I_n - S)\mathbf{x} & = & -\mathbf{x} & \Leftrightarrow \\ (\alpha I_n - S)\mathbf{x} & = & -(\alpha I_n + S)\mathbf{x} & \Leftrightarrow \\ \alpha I_n\mathbf{x} - S\mathbf{x} & = & -\alpha I_n\mathbf{x} - S\mathbf{x} & \Leftrightarrow \\ \mathbf{x} & = & -\mathbf{x} & \Leftrightarrow \\ \mathbf{x} & = & 0. & \end{array}$$

But $\mathbf{x} = 0$ contradicts with the assumption that $\mathbf{x}$ is an eigenvector, therefore we showed that -1 can not be an eigenvalue of $(\alpha I_n - S)(\alpha I_n + S)^{-1}$. This means that $(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n$ is nonsingular. Now, $(\alpha I_n + S)^{-1}$ is a positive real matrix as the inverse of a positive real matrix, B has full column rank, therefore $B(\alpha I_n + S)^{-1}B^T$ is a positive real matrix. And this implies that $\alpha I_m + B(\alpha I_n + S)^{-1}B^T$ is nonsingular. Now,

$$\mathcal{Q} = \mathcal{V}^T\mathcal{U}\mathcal{V} = \begin{bmatrix} V_{11}^T U_{11} V_{11} & V_{11}^T U_{12} V_{22} \\ \\ V_{22}^T U_{21} V_{11} & V_{22}^T U_{22} V_{22} \end{bmatrix},$$

which means that

$$Q_{12} = V_{11}^T U_{12} V_{22} = -V_{11}[(\alpha I_n - S)(\alpha I_n + S)^{-1} + I_n]B^T[\alpha I_m + B(\alpha I_n + S)^{-1}B^T]^{-1}V_{22}.$$

This shows that Q_{12} has a full column rank since V_{11}^T and V_{22} are orthogonal and B^T has full column rank. So, we have proved the theorem.
□

So, quite technical proof of the Theorem (3.7) insures that the HSS iteration method applied to a generalized saddle-point problem will converge, under certain conditions on matrices A, B and C. Luckily, in many applications these conditions are satisfied, so HSS iterative method can be seen as an potential solver for saddle-point problems.
Let us now have a closer look at the system (3.11), taking into account that $\mathcal{A}$ has a special block structure and make some general observations. The first half step of (3.11) requires to solve a system of two uncoupled linear systems:

$$\begin{cases} (H + \alpha I_n)u^{k+\frac{1}{2}} &= \alpha u^k - Su^k + f - B^T p^k, \\ (C + \alpha I_m)p^{k+\frac{1}{2}} &= \alpha p^k - g + Bu^k. \end{cases} \tag{3.17}$$

Since both systems in (3.17) are symmetric positive definite (SPD), any solver for SPD solvers can be applied, for instance preconditioned conjugate gradient method. It should be mentioned that addition of a positive term to the main diagonal elements of H and C considerably improves the condition numbers, which of course also results in better convergence rate of iterative methods applied to (3.17). If we normalize H in the sense that its largest eigenvalue is equal to 1, then for the spectral condition number of $H + \alpha I_n$ we will have:

$$\kappa(H + \alpha I_n) = \frac{1 + \alpha}{\lambda_{\min}(H) + \alpha} \leq 1 + \frac{1}{\alpha}.$$

Remark, that even if α is reasonably small, for instance $\alpha = 0.1$, the condition number is small as well ($\kappa(H + \alpha I_n) \leq 11$). Unless the value of α is very small, CG applied to method applied to (3.17) will converge rapidly, independent of the number n of unknows. Now, let us consider the second half-step of the algorithm given by (3.11). It requires a solution of two coupled linear systems of the form

$$\begin{cases} (\alpha I_n + S)u^{k+1} + B^T p^{k+1} &= (\alpha I_n - H)u^{k+\frac{1}{2}} + f \equiv f^k, \\ -Bu^{k+1} + \alpha p^{k+1} &= (\alpha I_m - C)p^{k+\frac{1}{2}} - g \equiv g^k. \end{cases} \tag{3.18}$$

Solving this system is less trivial and might be done in several ways. For example one can eliminate u_{k+1} from the second equation using the first one, which will result in smaller (order m) linear system of the form:

$$[B(I_n + \alpha^{-1}S)^{-1}B^T + \alpha^2 I_n m]p^{k+1} = B(I_n + \alpha^{-1}S)^{-1}f^k + \alpha g^k. \tag{3.19}$$

If the soultion of (3.19) is computed, then the vector u^{k+1} is can be computed by

$$u^{k+1} = (\alpha I_n + S)^{-1}(f^k - B^T p^{k+1}). \tag{3.20}$$

Note that if $S = O$, then (3.19) is simplified to:

$$[BB^T + \alpha^2 I_n m]p^{k+1} = B(I_n + \alpha^{-1}S)^{-1}f^k + \alpha g^k \tag{3.21}$$

and

$$u^{k+1} = \frac{1}{\alpha}(f^k - B^T p^{k+1}).$$

As we will see in proceeding chapter, this will be the case for our application of FETI domain decomposition method to the model problem. Moreover, in our case BB^T is sufficiently sparse, so system (3.21) can be solved using sparse Cholesky factorization. In general, if BB^T is not sparse enough, then other method, such as Preconditioned Conjugate Gradient (PCG), could be used.

As it has been already mentioned, the linear systems in (3.11) need not to be solved exactly. The inexact solves can considerably reduce the cost of each iteration though at the expense of slower convergence. But it should be mentioned that if alternating scheme is used as a preconditioner for Krylov subspace methods, then inexact solves are a good choice.

3.5 AHSS splitting iteration for saddle-point problems

In the previous section we considered the application of HSS iterative method to the generalized saddle-point problems and in this section we will discuss the Accelerated Hermitian and skew-Hermitian (AHSS) iteration method [14] for large sparse saddle-point problem by making use of the HSS iterative method. We will briefly present the AHSS methods, including the algorithmic description of the methods and unconditionally convergence property. Also we will point out the advantage of this class of methods over the HSS iterative methods.

AHSS iteration methods are two-parameter versions of HSS iterative methods. Theoretical and practical analysis shows that AHSS iteration methods algorithmically generalize the HSS methods, without any extra computational work, and the resulting iterative scheme turns out to be rapidly convergent. Moreover, the numerical sensitivity of the iterative schemes with respect to the iteration parameters is considerably decreased. As we will see, for all positive parameters the AHSS iteration methods converge unconditionally to the unique solution of the saddle-point problem.

We will consider a saddle-point problem:

$$\mathcal{A}\mathbf{x} \equiv \begin{bmatrix} A & B^T \\ -B & O \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \equiv \mathbf{b}, \tag{3.22}$$

with $A \in \mathbb{R}^{n\times n}$ symmetric positive definite , $B \in \mathbb{R}^{m\times n}$ full column rank, $f \in \mathbb{R}^n$, $g \in \mathbb{R}^m$ and $m \leq n$. As we have already seen, this assumptions guarantee the existence of the solution of (3.22). In order to simplify the analysis, without any loss of generality, we transform the saddle-point problem (3.22) into an equivalent form [14]. For this, we let $W \in \mathbb{R}^{n\times n}$ be a non-singular matrix such that $W^T A W = I_n$. For instance we can take $W = A^{-\frac{1}{2}}$. Consider also another non-singular matrix $Z \in \mathbb{R}^{m\times m}$ and construct $\bar{B}$, C , T, $\bar{\mathcal{A}}$, $\bar{\mathbf{x}}$ and $\bar{\mathbf{b}}$ in the following way:

$$\bar{B} = W^T B Z, \quad C = Z^{-T} Z^{-1}, \tag{3.23}$$

$$T = \begin{bmatrix} W & O \\ O & Z \end{bmatrix}, \quad \bar{\mathcal{A}} \equiv T^T \mathcal{A} T = \begin{bmatrix} I_n & \bar{B} \\ -\bar{B}^T & O \end{bmatrix}, \tag{3.24}$$

$$\bar{\mathbf{x}} \equiv \begin{bmatrix} \bar{y} \\ \bar{z} \end{bmatrix} = T^{-1}\mathbf{x} = \begin{bmatrix} W^{-1}y \\ Z^{-1}z \end{bmatrix}, \quad \bar{\mathbf{b}} \equiv \begin{bmatrix} \bar{f} \\ \bar{g} \end{bmatrix} = T^T \mathbf{b} = \begin{bmatrix} W^T f \\ Z^T g \end{bmatrix}. \tag{3.25}$$

With this one can show that the saddle-point problem (3.22) is equivalent to

$$\bar{\mathcal{A}}\bar{\mathbf{x}} = \bar{\mathbf{b}} \tag{3.26}$$

Now, let us split the coefficient matrix $\bar{\mathcal{A}}$ into its symmetric and skew-symmetric parts

$$\bar{\mathcal{A}} = \bar{\mathcal{H}} + \bar{\mathcal{S}},$$

with

$$\bar{\mathcal{H}} = \frac{1}{2}(\bar{\mathcal{A}} + \bar{\mathcal{A}}^T) = \begin{bmatrix} I_n & O \\ O & O \end{bmatrix} \text{ and } \bar{\mathcal{H}} = \frac{1}{2}(\bar{\mathcal{A}} - \bar{\mathcal{A}}^T) = \begin{bmatrix} O & \bar{B} \\ -\bar{B}^T & O \end{bmatrix}.$$

We apply the HSS iteration technique and obtain the following iteration scheme:

$$\begin{cases} (\Lambda + \bar{\mathcal{H}})\bar{\mathbf{x}}^{(k+\frac{1}{2})} &= (\Lambda - \bar{\mathcal{S}})\bar{\mathbf{x}}^{(k)} + \bar{\mathbf{b}}, \\ (\Lambda + \bar{\mathcal{S}})\bar{\mathbf{x}}^{(k+1)} &= (\Lambda - \bar{\mathcal{H}})\bar{\mathbf{x}}^{(k+\frac{1}{2})} + \bar{\mathbf{b}}, \end{cases} \tag{3.27}$$

where

$$\Lambda = \begin{bmatrix} \alpha I_n & O \\ O & \beta I_m \end{bmatrix}, \text{ with } \alpha \text{ and } \beta \text{ positive constants.}$$

As it can be seen the iterative scheme given by (3.27) is not the same as previously presented scheme (3.11), as it involves two arbitrary parameters α and β. When $\alpha \neq \beta$ the matrix $\mathcal{Q}(\alpha, \beta) = (\Lambda + \bar{\mathcal{S}})^{-1}(\Lambda + \bar{\mathcal{S}})$ is not unitary.
If we now write down $\bar{\mathcal{H}}$ and $\bar{\mathcal{S}}$ explicitly then (3.27) can be rewritten as follows:

$$\begin{bmatrix} \alpha A & B \\ -B^T & \beta C \end{bmatrix} \begin{bmatrix} y^{k+1} \\ z^{k+1} \end{bmatrix} = \begin{bmatrix} \frac{\alpha(\alpha-1)}{\alpha+1}A & -\frac{\alpha-1}{\alpha+1}B \\ B^T & \beta C \end{bmatrix} \begin{bmatrix} y^k \\ z^k \end{bmatrix} + \begin{bmatrix} \frac{2\alpha}{\alpha+1}f \\ 2g \end{bmatrix}, \tag{3.28}$$

or (3.28) can be equivalently written as

$$\begin{bmatrix} y^{k+1} \\ z^{k+1} \end{bmatrix} = \mathcal{T}(\alpha, \beta) \begin{bmatrix} y^k \\ z^k \end{bmatrix} + \mathcal{K}(\alpha, \beta) \begin{bmatrix} f \\ g \end{bmatrix}, \tag{3.29}$$

where

$$\mathcal{T}(\alpha, \beta) = \begin{bmatrix} \alpha A & B \\ -B^T & \beta C \end{bmatrix}^{-1} \begin{bmatrix} \frac{\alpha(\alpha-1)}{\alpha+1}A & -\frac{\alpha-1}{\alpha+1}B \\ B^T & \beta C \end{bmatrix} \tag{3.30}$$

and

$$\mathcal{K}(\alpha, \beta) = \begin{bmatrix} \alpha A & B \\ -B^T & \beta C \end{bmatrix}^{-1} \begin{bmatrix} \frac{2\alpha}{\alpha+1}I & O \\ O & 2I \end{bmatrix}. \tag{3.31}$$

One can show that there exists a splitting of the coefficient matrix

$$\mathcal{A} = \mathcal{M}(\alpha, \beta) - \mathcal{N}(\alpha, \beta)$$

such that the iteration matrix $\mathcal{T}(\alpha,\beta) = \mathcal{M}^{-1}(\alpha,\beta)\mathcal{N}(\alpha,\beta)$. It turns out that $\mathcal{M}(\alpha,\beta)$ and $\mathcal{N}(\alpha,\beta)$ that satisfy these conditions are given by

$$\mathcal{M}(\alpha,\beta) = \begin{bmatrix} \frac{\alpha+1}{2}A & \frac{\alpha+1}{2\alpha}B \\ -\frac{1}{2}B^T & \frac{\beta}{2}C \end{bmatrix}, \quad \mathcal{N}(\alpha,\beta) = \begin{bmatrix} \frac{\alpha-1}{2}A & -\frac{\alpha-1}{2\alpha}B \\ \frac{1}{2}B^T & \frac{\beta}{2}C \end{bmatrix}. \tag{3.32}$$

In fact, at each iteration we need to solve a linear system with coefficient matrix $\mathcal{M}'(\alpha,\beta)$ or, equivalently, with $\mathcal{M}(\alpha,\beta)$, where

$$\mathcal{M}'(\alpha,\beta) = \begin{bmatrix} \alpha A & B \\ -B^T & \beta C \end{bmatrix}.$$

If we consider the block-triangular factorization of the matrix $\mathcal{M}(\alpha,\beta)$, for an initial guess $\mathbf{x}^{(0)} = [{y^{(0)}}^T, {z^{(0)}}^T]^T \in \mathbb{R}^{n+m}$ and two positive constants α and β the algorithm of AHSS reads as:

Algorithm 3. ***The AHSS iteration method***
Given an initial guess $x^{(0)}$, for $k = 0, 1, 2, ...,$ until $\{x^{(k)}\}$ converges,

1. *Compute the current residual vector*

$$r^{(k)} = f - (Ay^{(k)} + Bz^{(k)}), \quad s^{(k)} = g + B^T y^{(k)},$$

2. *Compute the auxiliary vector*

$$u^{(k)} = \frac{2}{\alpha+1} r^{(k)}, \quad v^{(k)} = B^T A^{-1} u^{(k)} + 2s^{(k)},$$

3. *Compute the update vector*

$$(\beta C + \frac{1}{\alpha} B^T A^{-1} B) w^{(k)} = v^{(k)}, \quad A t^{(k)} = u^{(k)} - B w^{(k)},$$

4. *Form the next iterate*

$$y^{(k+1)} = y^{(k)} + t^{(k)}, \quad z^{(k+1)} = z^{(k)} + w^{(k)}.$$

According to the algorithm, at each step we have to solve two subsystems of linear equations with the coefficient matrix A and $(\beta C + \frac{1}{\alpha} B^T A^{-1} B)$ which is the Schur-complement of the matrix $\mathcal{M}'(\alpha,\beta)$. If we recall that C is an arbitrary

symmetric positive-definite matrix, we can choose it in a way that the matrix $(\beta C + \frac{1}{\alpha}B^TA^{-1}B)$ can be easily invertible in order to solve the subsystems as efficient as possible.
Now, we will shortly present the main results of convergence analysis. We will start with the following lemma which provides with explicit expressions for the eigenvalues of iteration matrix $\mathcal{T}(\alpha,\beta)$ [14].

Lemma 3.8. *Consider the saddle-point problem (3.22) and assume that $A \in \mathbb{R}^{n\times n}$ is symmetric positive definite, $B \in \mathbb{R}^{n\times m}$ has full column rank and $\alpha, \beta > 0$ are given. Let $C \in \mathbb{R}^{m\times m}$ be a symmetric positive definite matrix. If $\bar{\sigma}_k (k = 1, 2, ..., m)$ are the positive singular values of the matrix $\bar{B} \in \mathbb{R}^{n\times m}$, given by (3.23), then the eigenvalues of the iteration matrix of AHSS iteration method $\mathcal{T}(\alpha,\beta)$, defined in (3.30), are*

- $\dfrac{\alpha-1}{\alpha+1}$ *with multiplicity* $n-m$

and

- $\dfrac{1}{(\alpha+1)(\alpha\beta+\bar{\sigma}_k^2)}(\alpha(\alpha\beta-\bar{\sigma}_k^2) \pm \sqrt{(\alpha\beta+\bar{\sigma}_k^2)-4\alpha^3\beta\bar{\sigma}_k^2}), \quad k = 1, 2, ..., m.$

It should be mentioned that the singular values of $\bar{B}$ are exactly the square roots of the eigenvalues of the matrix $C^{-1}B^TA^{-1}B$.

Lemma 3.9. *Assume that the conditions of Lemma 3.8 are satisfied. If $\bar{\sigma}_k$ are the positive singular values of the matrix $\bar{B} \in \mathbb{R}^{n\times m}$ in (3.30), then the iteration matrix $\mathcal{T}(\alpha,\beta)$ of the AHSS iteration method has*

- $m-n$ *eigenvalues* λ *with absolute value* $|\lambda| = \dfrac{|\alpha-1|}{\alpha+1}$,

- $2m$ *eigenvalues* λ *such that for* $k = 1, 2, ..., m$
  - *if* $\alpha\beta + \bar{\sigma}_k^2 > 2\alpha\sqrt{\alpha\beta}\bar{\sigma}_k$ *, then there exist two correspoding eigenvalues* λ *such that*

$$|\lambda| = \frac{\alpha}{1+\alpha}\left(\frac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \sqrt{\frac{1}{\alpha^2} - \frac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right),$$

$$|\lambda| = \frac{\alpha}{1+\alpha}\left|\frac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} - \sqrt{\frac{1}{\alpha^2} - \frac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right|$$

– *if* $\alpha\beta + \bar{\sigma}_k^2 \leq 2\alpha\sqrt{\alpha\beta}\bar{\sigma}_k$ *, then there are two eigenvalues* λ *such that* $|\lambda| = \sqrt{\dfrac{\alpha-1}{\alpha+1}}$.

Using the lemmas above it is possible to show that the AHSS iterative method converges, namely the following theorem is hold true:

Theorem 3.10. *Consider the saddle-point problem (3.22) and assume that* $A \in \mathbb{R}^{n\times n}$ *is symmetric positive definite,* $B \in \mathbb{R}^{n\times m}$ *has full column rank and* $\alpha, \beta > 0$ *are given. Let* $C \in \mathbb{R}^{m\times m}$ *be a symmetric positive definite matrix. Then the spectral radius* $\rho(\mathcal{T}(\alpha,\beta)) < 1$ *for any* α *and* β *positive constants, where*

$$\rho(\mathcal{T}(\alpha,\beta)) = \begin{cases} \max\left\{\dfrac{1-\alpha}{1+\alpha}, \dfrac{\alpha}{1+\alpha}\left(\dfrac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \sqrt{\dfrac{1}{\alpha^2} - \dfrac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right)\right\} & \text{for} \quad \alpha \leq 1 \\ \max\left\{\sqrt{\dfrac{\alpha-1}{\alpha+1}}, \dfrac{\alpha}{1+\alpha}\left(\dfrac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \sqrt{\dfrac{1}{\alpha^2} - \dfrac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right)\right\} & \text{for} \quad \alpha > 1 \end{cases}$$

which means, that AHSS iteration converges to the exact solution of the saddle-point problem (3.22).

Proof [14]. First let us observe that

$$\frac{|1-\alpha|}{1+\alpha} < 1 \quad \forall\alpha > 0 \quad \text{and} \quad \sqrt{\frac{\alpha-1}{\alpha+1}} < 1 \quad \forall\alpha > 1.$$

Now, for $k = 1, 2, ..., m$ when $\alpha\beta + \bar{\sigma}_k^2 > 2\alpha\sqrt{\alpha\beta}\bar{\sigma}_k$, we have

$$\begin{aligned} \frac{\alpha}{1+\alpha}\left|\frac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \sqrt{\frac{1}{\alpha^2} - \frac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right| &\leq \frac{\alpha}{1+\alpha}\left(\frac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \sqrt{\frac{1}{\alpha^2} - \frac{4\alpha\beta\bar{\sigma}_k^2}{\alpha\beta+\bar{\sigma}_k^2}}\right) \\ &< \frac{\alpha}{1+\alpha}\left(\frac{|\alpha\beta-\bar{\sigma}_k^2|}{\alpha\beta+\bar{\sigma}_k^2} + \frac{1}{\alpha}\right) \\ &< \frac{\alpha}{1+\alpha}\left(1 + \frac{1}{\alpha}\right) \end{aligned}$$

So, if we take into account Lemma (3.9) then we have showed that $\rho(\mathcal{T}(\alpha,\beta)) < 1$ for any $\alpha > 0$ and $\beta > 0$.

□

Following theorem presents the results on optimal iteration parameters and the corresponding asymptotic convergence factor of the AHSS iteration.

Theorem 3.11. *Consider the saddle-point problem (3.22) and assume that* $A \in \mathbb{R}^{n\times n}$ *is symmetric positive definite,* $B \in \mathbb{R}^{n\times m}$ *has full column rank and* $\alpha, \beta > 0$ *are given. Let* $C \in \mathbb{R}^{m\times m}$ *be a symmetric positive definite matrix. If* $\sigma_k (k = 1, 2, ..., m)$ *are the positive singular values of the matrix* $W^T BZ \in \mathbb{R}^{n\times m}$*, and* $\sigma_{\min} = \min_{1\le k\le m}\{\sigma_k\}$ *and* $\sigma_{\max} = \max_{1\le k\le m}\{\sigma_k\}$*, then, for the AHSS iteration method applied to the saddle-point problem (3.22), the optimal values of the iteration parameters* α *and* β *are given by:*

$$\{\alpha^*, \beta^*\} = \arg\min_{\alpha,\beta>0} \rho(\mathcal{T}(\alpha, \beta)) = \left\{\tau, \frac{\sigma_{\min}\sigma_{\max}}{\tau}\right\},$$

and corresponding $\rho(\mathcal{T}(\alpha^*, \beta^*))$ *is*

$$\rho(\mathcal{T}(\alpha^*, \beta^*)) = \frac{\sqrt{\sigma_{\max}} - \sqrt{\sigma_{\min}}}{\sqrt{\sigma_{\max}} + \sqrt{\sigma_{\min}}} \equiv \frac{\sqrt[4]{\kappa} - 1}{\sqrt[4]{\kappa} + 1},$$

where $\kappa = \frac{\sigma_{\max}^2}{\sigma_{\min}^2}$ *is the condition number of the matrix* $C^{-1}B^T A^{-1}B$ *and*

$$\tau = \frac{\sigma_{\min} + \sigma_{\max}}{2\sqrt{\sigma_{\max}\sigma_{\min}}} \equiv \frac{1}{2}\left(\sqrt[4]{\kappa} + \frac{1}{\sqrt[4]{\kappa}}\right).$$

From Theorem (3.11) and Remark 4.3 in [30] it is known that the optimal convergence rate of AHSS iteration method is $\frac{2}{\sqrt[4]{\kappa}}$ and that of HSS iteration method is approximately $\frac{2}{\sqrt{\kappa}}$. This means that AHSS converges remarkably faster than HSS when the optimal iteration parameters are employed and $\kappa \gg 1$.

As in the case of HSS iteration method, AHSS iteration can be used as a preconditioner to accelerate Krylov subspace methods such as GMRES. Namely, matrix $\mathcal{M}(\alpha, \beta)$ can be seen as a preconditioner for the system (3.22).

Chapter 4

Finite Element Tearing and Interconnecting Method

In this chapter we will discuss the Finite Element Tearing and Interconnecting method. It is one of domain decomposition methods [12] for solving large systems of linear equations arising from finite element discretization of elliptic partial differential equations. We will present the derivation of the method which was first done by Farhat and Roux [10]. Similar to most of the domain decomposition methods, in this method the original problem is divided into number of subproblems which are easier to solve because of their smaller size. We will give the saddle-point formulation of the FETI method, and the main algebraic properties of the method will be presented as well [26].

4.1 About Domain Decomposition Methods

The class of domain decomposition methods has gained enormous popularity during the last decade. This methods follow the idea of "divide -and -conquer", which means that they divide the original problem into a number of smaller problems. This division is done for different reasons. Sometimes such a division arises from breaking up a domain with complicated geometry. In other cases, though, the division is more artificial. The subproblems are easier to solve because of their smaller size and often parallel computation can be used. This is quite important for the efficiency of computations.
Domain decomposition methods can be seen from two different point of view. One is that they may arise from separation of a physical domain into regions. In these regions the problem can be modeled by separate partial differential equations. On

the interfaces between the subdomains various conditions, such as continuity, are imposed. The other approach is to see domain decomposition methods as methods for solving large algebraic linear systems arising from the discretization of partial differential equations. In that sense, a domain decomposition method can be seen as an algebraic method, where the large system is subdivided into smaller problems, whose solutions can be used to generate a preconditioner for the large system.

4.2 Original FETI method

As we have mentioned before, Finite Element Tearing and Interconnecting (FETI) method is a domain decomposition method designed for solving systems arising from finite element discretization of elliptic partial differential equations. If we describe the method in general terms, it can be done as follows: a given domain is "torn " into non-overlapping subdomains where an incomplete solution of the primary field is first evaluated using a direct solver. Using Lagrangian multipliers intersubdomain field continuity is enforced. As a result of such "gluing" process a smaller size symmetric dual problem is generated where the unknowns are the Lagrange multipliers, and which is solved by a preconditioned conjugate gradient (PCG) method.

To describe the method in more details, let us first introduce some notations and assumptions needed in this chapter. For $u, v \in \mathbb{R}^n$, the inner product $\langle u, v\rangle = u^T v$ is also interpreted as duality product. For a symmetric positive semidefinite matrix A , we denote $||u||_A = \langle Au, v\rangle^{1/2}$, which is the seminorm induced by matrix A. Naturally, if A is positive definite then $\langle Au, v\rangle^{1/2}$ is norm. We will also need the notion of pseudoinverse of an operator which is given as follows:

Definition 4.1. *Let A be a linear operator. A pseudoinverse A^+ is any linear operator such that, if $a \in ImA$ then $AA^+a = a$.*

In general, a pseudoinverse is not unique. The algorithms discussed in this chapter will be invariant to a specific choice of the pseudoinverse. If A is a symmetric operator on a finite dimensional space, then the pseudoinverse A^+ can also be chosen symmetric. Indeed, if we consider the spectral decomposition of A,

$$A = \sum_{\sigma} \sigma v_\sigma v_\sigma^T, \qquad Av_\sigma = \sigma v_\sigma, \qquad v_\sigma^T v_\sigma = 1 \tag{4.1}$$

then as a pseudoinverse A^+ one can chose

$$A^+ = \sum_{\sigma \neq 0} \frac{1}{\sigma} v_\sigma v_\sigma^T$$

Now, let the domain Ω in $\mathbb{R}^2$ ($\mathbb{R}^3$) be decomposed into N_s non-overlapping subdomains $\Omega_1, \Omega_2, ..., \Omega_{N_s}$. Let u_i be the vector of degrees of freedom for subdomain Ω_i corresponding to a conforming finite element discretization of an elliptic problem (for instance linear elasticity, Stokes problem) defined on Ω, such that each subdomain is a union of some of the elements. Assume K_i and f_i are the local stiffness matrices and the load vectors respectively, associated with the subdomain Ω_i. Then we consider u, f and K defined as:

$$u = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{N_s} \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_{N_s} \end{bmatrix}, \quad K = \begin{bmatrix} K_1 & O & \dots & O \\ O & K_2 & \dots & O \\ \dots & \dots & \dots & \dots \\ O & O & \dots & K_{N_s} \end{bmatrix}. \tag{4.2}$$

Depending on the boundary conditions and the location of the subdomain, the local stiffness matrix K_i is positive definite or positive semidefinite. A subdomain without sufficient essential boundary conditions (which is done in order to prevent the local stiffness matrix K_i from being singular) is called a floating subdomain. Now, denote by Z_i the matrix with linearly independent columns that generate the kernel of K_i, which means that $\mathrm{Im} Z_i = \ker K_i$. If we consider

$$Z = \begin{bmatrix} Z_1 & O & \dots & O \\ O & Z_2 & \dots & O \\ \dots & \dots & \dots & \dots \\ O & O & \dots & Z_{N_s} \end{bmatrix}$$

then

$$\mathrm{Im} Z = \ker K \quad \text{and} \quad \ker Z = \{0\}.$$

For a single mesh point $x \in \Omega$ the algorithm assigns several degrees of freedom associated with it, if it lies on the intersection of boundaries, which is called interface.

Now, let matrix B be such that constraint $Bu = 0$ expresses the condition that for each mesh node shared by more than one subdomain the values of degrees of freedom associated with that node coincide. Consider the space of all vectors of degrees of freedom, which we denote by W and the space of the vectors of values of the continuity constraint, denoted by Λ . With this notations we can see that

$$K : \mathrm{W} \to \mathrm{W} \quad \text{and} \quad B : \mathrm{W} \to \Lambda.$$

In fact, the problem that we need to solve is the following minimization problem subject to intersubdomain continuity conditions:

$$E(u) = \frac{1}{2}u^T Ku - f^T u \to \min \quad \text{subject to} \quad Bu = 0, \quad u \in \mathrm{W}. \tag{4.3}$$

If we assume that

$$\ker B \cap \ker K = \{0\} \tag{4.4}$$

then the solution of (4.3) will be unique. For describing the FETI algorithm we need some more notations, which ar given in (4.5).

$$\begin{aligned} G &= BZ, \\ F &= BK^+B^T, \\ d &= BK^+f, \\ e &= Z^T f, \\ P &= I - G(G^TG)^{-1}G^T. \end{aligned} \tag{4.5}$$

Later on in this chapter we will show that P is well defined, namely the matrix G^TG is invertible.

As it was originally done by Farhat and Roux [10] Lagrange multipliers are introduced to enforce the continuity of the solution. With this, solving constraint minimization problem (4.3) leads to the following system of equations:

$$\begin{array}{lcl} Ku + B^t\lambda & = & f \\ Bu & = & 0 \end{array} \tag{4.6}$$

Now, let us note that a solution u of the first equation in (4.6) exists if and only if

$$f - B^T\lambda \in \mathrm{Im}K. \tag{4.7}$$

Then u should have a form

$$u = K^+(f - B^T\lambda) + Z\alpha, \tag{4.8}$$

where α still has to be specified. If we substitute the expression for u from (4.8) into the second equation in (4.6), we will get:

$$BK^+(f - B^T\lambda) + BZ\alpha = 0. \tag{4.9}$$

Multiplying (4.9) by P, which was defined in (4.5), and taking into account (4.7), we will obtain that λ satisfies the following system of equations:

$$\begin{cases} P(F\lambda - d) &= 0 \\ G^T\lambda &= e \end{cases} \tag{4.10}$$

with e, F, G defined in (4.5).
Let us now show that the orthogonal projection P is well defined.

Lemma 4.2. *The matrix G^TG is invertible, namely, $(G^TG)^{-1}$ exists.*

Proof[26]. Let $Gw = BZw = 0$. Then $Zw \in \ker B$. From definition of Z it follows that $Zw \in \ker K$. As we have assumed that $\ker K \cap \ker B = \{0\}$, then $Zw = 0$. Since we also assumed that Z is of full rank, then we immediately obtain that $w = 0$. This means that G is invertible, as well as G^T, and therefore, matrix G^TG, as product of two invertible matrices, is also invertible.
□
Next let us consider the system (4.10).

Theorem 4.3. *The solution λ of (4.10) is unique up to addition of a vector from $\ker B^T$. Any solution λ of (4.10) yields the same solution u of the minimization problem (4.3), using (4.8) with $\alpha = -(G^TG)^{-1}G^T(d - F\lambda)$.*

Proof. See [26]. □

The original FETI algorithm, which is an application of preconditioned CG for solving the equation $PF\lambda = d$, using a symmetric preconditioner D, can be written as follows.

Algorithm 4. ***(FETI)***
Given an initial $\bar{\lambda}_0$, compute the initial estimate

$$\lambda_0 = G(G^TG)^{-1}e + P\bar{\lambda}_0$$

and the initial residual

$$r_0 = P(F\lambda_0 - d).$$

Repeat for $k = 1, 2, ...$ *until convergence:*

$$\begin{aligned} z_{k-1} &= Dr_{k-1} \\ y_{k-1} &= Pz_{k-1} \\ \xi_k &= r_{k-1}^T y_{k-1} \\ p_k &= y_{k-1} + \frac{\xi_k}{xi_{k-1}} p_{k-1} \quad (p_1 = y_0) \\ \mu_k &= \frac{\xi_k}{p_k^T PF p_k} \\ \lambda_k &= \lambda_{k-1} + \mu_k p_k \\ r_k &= r_{k-1} + \mu_k PF p_k \end{aligned}$$

So, we see that in the original FETI algorithm the first equation of (4.10) is solved by a preconditioned conjugate gradient method using an initial approximation λ_0 such that it satisfies the second equation. For the conjugate gradient method it is required to evaluate the actions of PF. Since $F = BK^+B^T$, most of the computational work is concentrated in evaluation of K^+. On the other hand, K^+ is a block diagonal matrix, therefore its action can be computed in parallel which involves solving subdomain problems only. Application of P leads to solving a small coarse problem. For a scalar problem the size of the coarse problem corresponding to P is less than the number of subdomains N_s.

Another approach could be to solve the system (4.6) using HSS iterative method. For this let us rewrite the system (4.6) as a saddle-point system:

$$\begin{bmatrix} K & B^T \\ B & O \end{bmatrix} \begin{bmatrix} u \\ \lambda \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{4.11}$$

In fact, as it has been mentioned several times, this is the main interest of our work. So, in the next chapter we wil apply HSS method to the FETI system (4.11) for our model problem.

Chapter 5

HSS applied to FETI system

In the previous two chapters we introduced the HSS iterative method as a potential solver for saddle-point problems and also we looked at saddle-point formulation of the well known domain decomposition method FETI. In this chapter we will first apply FETI domain decomposition method to our model problem and then solve the resulting saddle-point system using HHS iterative method. We will use HSS method both as a stationary iterative method and also as a preconditioner for Krylov subspace method GMRES and the potential of this approach will be illustrated.

5.1 Numerical results

Let us recall that our model problem is (2.1):

$$\begin{cases} -\triangle u(x,y) &= f(x,y) \quad \text{in} \quad \Omega = (0,2) \times (0,1), \\ \\ u(x,y) &= g(x,y) \quad \text{on} \quad \partial\Omega. \end{cases} \tag{5.1}$$

Let us apply FETI method to our model problem. For this consider the simplest case, when the domain Ω is divided into two subdomains Ω_1 and Ω_2 such that $\Omega_1 \cap \Omega_2 = \emptyset$ and $\partial\Omega_1 \cap \partial\Omega_2 = \Gamma$. As it has been described in Chapter 3, for constructing FETI system, we need to compute the stiffness matrices K_i and the load vectors f_i for each subdomain Ω_i , $i \in \{1, 2\}$. In order to enforce the continuity condition we introduce the Lagrangian multiplier λ. Then the system to be solved

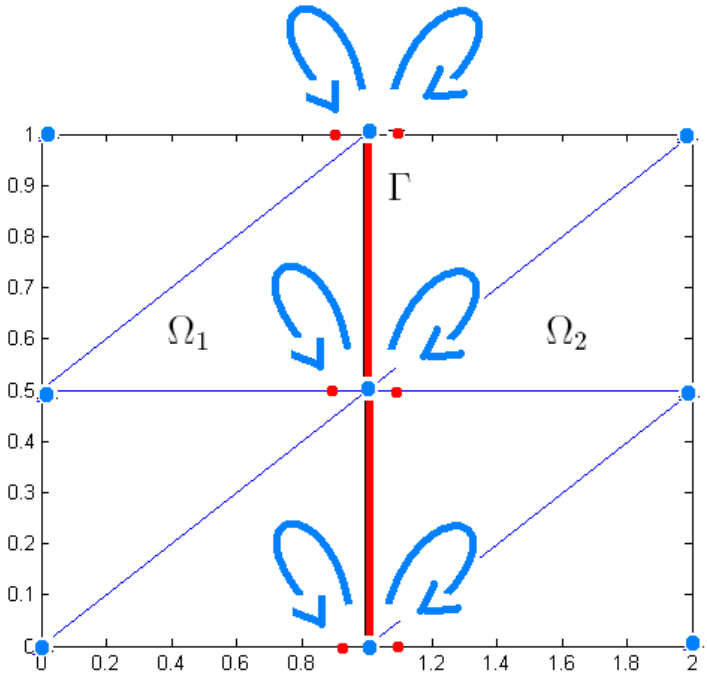


Figure 5.1: Tearing of unknowns on the interface

is:

$$\begin{bmatrix} K & B^T \\ B & O \end{bmatrix} \begin{bmatrix} u \\ \lambda \end{bmatrix} = \begin{bmatrix} f \\ 0 \end{bmatrix} \tag{5.2}$$

or

$$\mathcal{A}\mathbf{x} = \mathbf{b} \tag{5.3}$$

with

$$K = \begin{bmatrix} K_1 & O \\ O & K_2 \end{bmatrix}, \quad f = \begin{bmatrix} f_1 \\ f_2 \end{bmatrix}, \quad u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix},$$

$$\mathcal{A} = \begin{bmatrix} K & B^T \\ B & O \end{bmatrix}, \quad \mathbf{x} = \begin{bmatrix} u \\ \lambda \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} f \\ 0 \end{bmatrix}.$$

Recall that matrix B (constructed form $\{0, 1, -1\}$) is such that the constraint $Bu = 0$ provides the continuity of the solution u through the interface Γ. For the case of two subdomains one can show that, with certain numbering of nodal points, B is such that BB^T is a diagonal matrix, more precisely

$$BB^T = 2I.$$

If we recall the HSS algorithm for saddle-point problems, such a structure of the matrix BB^T is quite advantageous, which means that the two subsystems of HSS iteration can be solved quite accurate.
In our numerical experiments we used as a Dirichlet data and as a right hand side the functions

$$g(x, y) = x + y + 1 \quad \text{and} \quad f(x, y) = 0.$$

It can be easily seen that the exact solution of the model problem (2.1) is the harmonic function $u(x, y) = 1 + x + y$. We applied the HSS iterative scheme (3.17) for our model problem and also we used the HSS iterative method as a preconditioner for GMRES.
In Figure 5.2 we display the spectral radius of the iteration matrix $\mathcal{M}_\alpha$ in the case of $h = \frac{1}{9}$ for different values of α. So, if we define by α_{opt} the value of α that minimizes the spectral radius, then, as we can see from the figure, $\alpha_{\text{opt}} = 1.20$ and $\rho(\mathcal{M}_\alpha) = 0.77$.

Now, we have compared the number of HSS as an iterative solver, GMRES without preconditioner and GMRES with the HSS iterative scheme as a preconditioner. The results of the comparison are summarized in Table 5.1. In all our

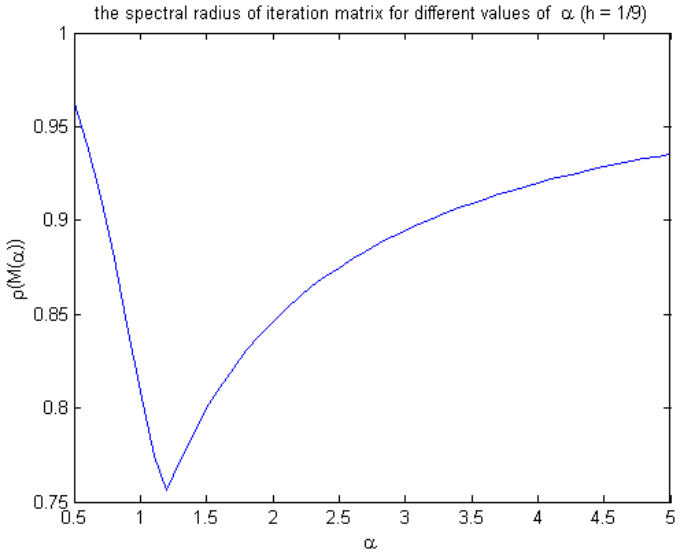


Figure 5.2: Spectral radius of iteration matrix $\rho(\mathcal{M}_\alpha)$ for different values of α ($h = \frac{1}{9}$)

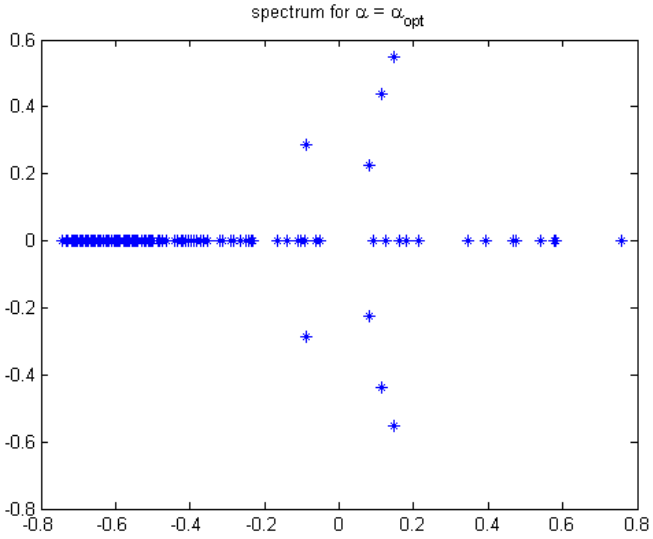


Figure 5.3: The spectrum of iteration matrix for optimal value of α

Table 5.1: Comparison of HSS iterative scheme, GMRES without preconditioning, GMRES with HSS iterative scheme as a preconditioner

h	Iterative	GMRES	Preconditioned GMRES
1	7	1	1 (inner 3)
1/2	13	7	4 (inner 3)
1/4	13	18	5 (inner 3)
1/8	34	30	7 (inner 3)
1/16	98	64	10(inner 3)

runs we used zero initial guess and the iteration is stopped when the relative residual has been reduced by at least four orders of magnitude, which means when $||\mathbf{b} - \mathcal{A}\mathbf{x}|| \leq 10^{-4}||\mathbf{b}||$. For the inner HSS iteration for preconditioned GMRES we used the accuracy $\epsilon = 10^{-1}$. As it can be seen form the table, the HSS iterative scheme as a preconditioner for GMRES remarkably reduces the number of iterations. For instance for $h = \frac{1}{17}$ the preconditioning halves the number of iterations. We have also tried to change the accuracy of the inner iteration for preconditioned GMRES. In the Table 5.2 we have presented the results.

Table 5.2: Number of iterations for preconditioned GMRES for different accuracies of inner HSS iteration

h	$\epsilon = 10^{-1}$	$\epsilon = 10^{-2}$	$\epsilon = 10^{-3}$
1	1 (inner 3)	1(inner 4)	1(inner 6)
1/2	4 (inner 3)	3(inner 6)	3(inner 9)
1/4	5 (inner 3)	3(inner 6)	3(inner 9)
1/8	7 (inner 3)	3(inner12)	3(inner 23)
1/16	10(inner 3)	5(inner 21)	3(inner 56)

We can observe that by making the inner iteration more accurate the outer iteration becomes independent of h. Nevertheless, it is obvious that the inner accuracy $\epsilon = 10^{-1}$ corresponds to the least number of total iterations.

Chapter 6

Other saddle-point problems

Saddle-point systems arise in many scientific and engineering applications, including computational fluid dynamics [15], [16], [24], [17], mixed finite element approximation of elliptic PDE's [18], [19], [25] and optimization [20], [23], [22], [21]. In this chapter we will present some more examples of problems leading to saddle-point systems. We will also consider the Boundary Element Method, which, as we will see, gives rise to saddle-point problem as well. We suggest HSS iterative method as a potential method for solving this problems.

6.1 Mixed Formulations of 2nd order elliptic problems

6.1.1 Linear elliptic problems

Let us consider the following problem

$$\begin{cases} \operatorname{div}(A(x)\nabla u) &= f \quad \text{in} \quad \Omega \subset \mathbb{R}^d, \\ u &= g_D \quad \text{on} \quad \Gamma_D, \\ (A(x)\nabla u)\cdot \bar{n} &= g_N \quad \text{on} \quad \Gamma_N, \end{cases} \tag{6.1}$$

where

- $\Gamma_D \cap \Gamma_N = \emptyset$ and $\Gamma_D \cup \Gamma_N = \partial\Omega$,

- A(x) is a smooth function on $\bar{\Omega}$ and $A(x) \geq a > 0$ for all $x \in \Omega$,
- $\bar{n}$ is the outward normal to $\partial\Omega$,
- f, g_D and g_N are given smooth functions in Ω , Γ_D and Γ_N respectively.

Consider the manifold $\mathrm{V}_g \subset \mathrm{H}^1(\Omega)$ and $\mathrm{V}_0 \subset \mathrm{H}^1(\Omega)$ defined as:

$$\mathrm{V}_g = \{v : v \in \mathrm{H}^1(\Omega) \ \text{ and } \ v = g_D \ \text{ on } \ \Gamma_D\},$$

$$\mathrm{V}_0 = \{v : v \in \mathrm{H}^1(\Omega) \ \text{ and } \ v = 0 \ \text{ on } \ \Gamma_D\}.$$

If we now multiply the first equation in (6.1) with a function $v \in \mathrm{V}_0$ and integrate over Ω we will get:

$$\int_\Omega \mathrm{div}(A(x)\nabla u)v\mathrm{d}x = \int_\Omega fv\mathrm{d}x \ \ \forall v \in \mathrm{V}_0.$$

Using the Green's formula and taking into account the third equation in (6.1), we will get

$$\int_\Omega A(x)\nabla u \cdot \nabla v\mathrm{d}x = -\int_\Omega fv\mathrm{d}x + \int_{\Gamma_N} g_N v\mathrm{d}s_x \ \ \forall v \in \mathrm{V}_0.$$

Note that for the formulation above we can relax the conditions on $A(x)$, requiring only $A(x) \in \mathrm{L}_\infty$ and $\bar{a} \geq A(x) \geq \underline{\mathrm{a}} > 0$ for almost all $x \in \Omega$. So, the primal variational problem can be formulated as:

Find $u \in \mathrm{V}_g$ such that

$$\int_\Omega A(x)\nabla u \cdot \nabla v\mathrm{d}x = -\int_\Omega fv\mathrm{d}x + \int_{\Gamma_N} g_N v\mathrm{d}s_x \ \ \forall v \in \mathrm{V}_0. \tag{6.2}$$

In order to get the mixed formulation of (6.1) we introduce the variable:

$$p = A\nabla u \ \text{ in } \ \Omega. \tag{6.3}$$

Then the first and third equations of (6.1) will become, respectively

$$\mathrm{div}p = f \ \text{ in } \ \Omega, \tag{6.4}$$

and

$$p \cdot \bar{n} = g_N \ \text{ on } \ \Gamma_N. \tag{6.5}$$

It is now possible to give two reasonable variational formulations for (6.2) - (6.5).

- first variational formulation

  Find $u \in \mathrm{V}_g$ and $p \in \left(\mathrm{L}^2(\Omega)\right)^d$ such that:

$$\begin{cases} \displaystyle\int_\Omega (A(x))^{-1} p \cdot q \mathrm{d}x - \int_\Omega q \cdot \nabla u \mathrm{d}x &= 0 & \forall q \in \left(\mathrm{L}^2(\Omega)\right)^d, \\ \displaystyle -\int_\Omega p \cdot \nabla v \mathrm{d}x &= \displaystyle\int_\Omega f v \mathrm{d}x - \int_{\Gamma_N} g_N v \mathrm{d}s_x & \forall v \in \mathrm{V}_0. \end{cases} \tag{6.6}$$

For introducing the second variational formulation we consider the space $\mathrm{H}_0(\mathrm{div};\Omega)$ and the manifold $\mathrm{H}_{g_N}(\mathrm{div};\Omega)$ defined by:

$$\mathrm{H}_0(\mathrm{div};\Omega) = \{q : q \in \left(\mathrm{L}^2(\Omega)\right)^d; \mathrm{div} q \in \mathrm{L}^2(\Omega); q \cdot \bar{n} = 0 \ \text{ on } \ \Gamma_N\}$$

and

$$\mathrm{H}_{g_N}(\mathrm{div};\Omega) = \{q : q \in \left(\mathrm{L}^2(\Omega)\right)^d; \mathrm{div} q \in \mathrm{L}^2(\Omega); q \cdot \bar{n} = g_N \ \text{ on } \ \Gamma_N\}.$$

Now, the second variational formulation can be formulated.

- second variational formulation

  Find $u \in \mathrm{L}^2(\Omega)$ and $p \in \mathrm{H}_{g_N}(\mathrm{div};\Omega)$ such that:

$$\begin{cases} \displaystyle\int_\Omega (A(x))^{-1} p \cdot q \mathrm{d}x + \int_\Omega u \mathrm{div} q \mathrm{d}x &= \displaystyle\int_{\Gamma_D} g_D q \cdot \bar{n} \mathrm{d}s_x & \forall q \in \mathrm{H}_0(\mathrm{div};\Omega), \\ \displaystyle\int_\Omega v \mathrm{div} p \mathrm{d}x &= \displaystyle\int_\Omega f v \mathrm{d}x & \forall v \in \mathrm{L}^2(\Omega). \end{cases} \tag{6.7}$$

The difference between the two variational formulations is simply in using the Green's formula (or integration by parts formula). Nevertheless, the regularity required for u and p is interchanged. So, for discretizing the first formulation one needs to use continuous finite elements for u and can use discontinuous finite elements for p. On the other hand, for discretizing the second formulation one can use discontinuous finite elements for u but the finite elements for p have to be such that $\mathrm{div} p \in \mathrm{L}^2(\Omega)$. Also we should mention that another difference between the two formulations is in the treatment of essential and natural boundary conditions. In

general the primal formulation is simpler, as it involves only one variable, and there are many robust methods based on this approximation. But in many applications the second variable p is the more relevant physical variable. In these cases the mixed formulation is preferred since very often it provides better accuracy for p.
If we now consider the second formulation and introduce the following notations:

- $a(p,q) = \int_\Omega (A(x))^{-1} p \cdot q \mathrm{d}x$ and $b(v,p) = \int_\Omega v \mathrm{div} p \mathrm{d}x$,
- $\langle F, v\rangle = \int_\Omega f v \mathrm{d}x$ and $\langle G, q\rangle = \int_{\Gamma_D} g_D q \cdot \bar{n} \mathrm{d}s_x$,

then the variational formulation can be reformulated in the following way:

Find $u \in \mathrm{L}^2(\Omega)$ and $p \in \mathrm{H}_{g_N}(\mathrm{div};\Omega)$ such that:

$$\begin{cases} a(p,q) + b(u,q) &= \langle F, v\rangle \quad \forall q \in \mathrm{H}_0(\mathrm{div};\Omega) \\ b(v,p) &= \langle G, q\rangle \quad \forall v \in \mathrm{L}^2(\Omega) \end{cases} \tag{6.8}$$

6.1.2 Linear elasticity problem

For vector valued function $v(x)$ we define by $\epsilon(v) = [\epsilon_{ij}]$ the following second order tensor:

$$\epsilon_{ij} = \frac{1}{2}(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}) \quad i,j = 1,...,d. \tag{6.9}$$

The linear elasticity equations are given by:

$$\begin{cases} \sigma &= E : \epsilon(u) \quad \text{in } \Omega \quad (\sigma_{ij} = \sum_{l=1}^{d}\sum_{m=1}^{d} E_{ijlm}\epsilon_{lm}(u)) \\ \mathrm{div}\sigma &= f \qquad \text{in } \Omega \end{cases} \tag{6.10}$$

If we substitute $\sigma = E : \epsilon(u)$ and $\epsilon_{ij} = \frac{1}{2}(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i})$ in the second equation of (6.10), we will get a second order elliptic system with the unknown u. In (6.10) the forth order tensor E is called elasticity tensor, which we will assume to be elliptic and have constant coefficients. We will denote the inverse tensor of E by C. This means

$$\tau = E : \epsilon(v) \iff \epsilon(v) = C : \tau.$$

Again, for simplicity, we will assume homogenous Dirichlet boundary condition:

$$u = 0 \ \text{ on } \ \partial\Omega.$$

Consider

$$\mathrm{H_S}(\mathrm{div};\Omega) = \{\tau : \tau \in (\mathrm{L}^2(\Omega))^{d^2}; \tau_{ij} = \tau_{ji} \ \ \forall i,j; \mathrm{div}\tau \in (\mathrm{L}^2(\Omega))^d\}.$$

Now we can formulate the mixed variational formulation.

Find $u \in (\mathrm{L}^2(\Omega))^{d^2}$ and $\sigma \in \mathrm{H_S}(\mathrm{div};\Omega)$ such that

$$\left\{ \begin{array}{rcll} \displaystyle\int_\Omega (C:\sigma) : \tau \mathrm{d}x + \int_\Omega u \cdot \mathrm{div}\tau \mathrm{d}x & = & 0 & \forall \tau \in \mathrm{H_S}(\mathrm{div};\Omega) \\ \\ \displaystyle\int_\Omega v \cdot \mathrm{div}\sigma \mathrm{d}x & = & \displaystyle\int_\Omega f \cdot v & \forall v \in (\mathrm{L}^2(\Omega))^{d^2} \end{array} \right. \tag{6.11}$$

By introducing the following notations

- $a(\sigma,\tau) = \displaystyle\int_\Omega (C:\sigma) : \tau \mathrm{d}x$ and $b(u,\tau) = \displaystyle\int_\Omega u \cdot \mathrm{div}\tau \mathrm{d}x$,
- $\langle F, v\rangle = \displaystyle\int_\Omega f v \mathrm{d}x$,

the mixed variational formulation can be rewritten as:

$$\left\{ \begin{array}{rcll} a(\sigma,\tau) + b(u,\tau) & = & 0 & \forall \tau \in \mathrm{H_S}(\mathrm{div};\Omega) \\ \\ b(v,\sigma) & = & \langle F, q\rangle & \forall v \in (\mathrm{L}^2(\Omega))^{d^2} \end{array} \right. \tag{6.12}$$

These mixed formulations are due to Hellinger and Reissner, and, sometimes, the procedure of derivation the mixed formulation is called Hellinger-Reissner principle.

6.1.3 Stokes problem

Next example we will consider is the Stokes equations for modeling the flow of incompressible fluids. The Stokes equations are:

$$\left\{ \begin{array}{rcll} -\nu\triangle u + \nabla p & = & f \ \text{ in } \ \Omega \\ \mathrm{div}u & = & 0 \ \text{ in } \ \Omega \end{array} \right. \tag{6.13}$$

Many kinds of boundary conditions can be used with these equations but for simplicity we will consider only the homogenous Dirichlet boundary condition, the so called no-slip condition:

$$u = 0 \quad \text{on} \quad \partial\Omega$$

As usual, in order to get the variational formulation we need to multiply the governing equations by the test functions and integrate over the domain Ω. By doing this and using the formula of integration by parts, we will get the following variational formulation.

Find $u \in (\mathrm{H}_0^1(\Omega))^d$ and $p \in \mathrm{L}^2(\Omega)$ such that:

$$\begin{cases} \displaystyle\int_\Omega \nu \nabla u : \nabla v \mathrm{d}x - \int_\Omega p \mathrm{div} v \mathrm{d}x &= \displaystyle\int_\Omega f \cdot v \mathrm{d}x \quad \forall v \in (\mathrm{H}_0^1(\Omega))^d \\ \displaystyle\int_\Omega p \mathrm{div} v \mathrm{d}x &= 0 \qquad\qquad \forall q \in \mathrm{L}^2(\Omega) \setminus \mathbb{R} \end{cases} \tag{6.14}$$

where $A : B$ is the direct matrix product defined as $[A : B]_{ij} = A_{ij} B_{ij}$. If we now introduce the following notations:

- $a(u, v) = \displaystyle\int_\Omega \nu \nabla u : \nabla v \mathrm{d}x$ and $b(q, v) = \displaystyle\int_\Omega q \mathrm{div} v \mathrm{d}x$,
- $\langle F, v \rangle = \displaystyle\int_\Omega f v \mathrm{d}x$,

then the variation formulation can be rewritten as:

Find $u \in (\mathrm{H}_0^1(\Omega))^d$ and $p \in \mathrm{L}^2(\Omega) \setminus \mathbb{R}$ such that:

$$\begin{cases} a(u, v) - b(p, v) &= \langle F, v \rangle \quad \forall v \in (\mathrm{H}_0^1(\Omega))^d \\ b(q, u) &= 0 \qquad \forall q \in \mathrm{L}^2(\Omega) \end{cases} \tag{6.15}$$

6.2 Boundary Element Method

Boundary Element Method (BEM) is a numerical computational method of solving linear partial differential equations which have been formulated as integral equations. BEM attempts to use the given boundary conditions to fit boundary values into the integral equation, rather than values throughout the space defined by a

PDE. Once this is done, the integral representation can then be used again to calculate numerically the solution directly at any desired point in the interior of the solution domain. Very often BEM is more efficient than other methods, for instance finite elements, in terms of computational resources. Conceptually, it works by constructing a "mesh" over the surface of the solution domain. However, for many problems BEM is significantly less efficient than volume-discretisation methods (FEM, FDM, FVM). Boundary element formulations typically result in fully populated matrices. This means that the storage requirements and computational time will grow according to the square of the number of unknowns. By contrast, finite element matrices are sparse (since the elements are only locally connected) and the storage requirements for the system matrices typically grow linearly with the problem size. Another requirement for BEM is that it can be applied to problems for which the fundamental solution can be calculated. But once the fundamental solution is known, and the boundary values are also approximated the solution in the interior of the domain is calculated very accurately, due to the representation formula. For introducing the technique of boundary element methods we will need some definitions.

Consider $\Omega \subset \mathbb{R}^d$. Define $C_0^\infty(\Omega)$ as:

$$C_0^\infty(\Omega) := \{\phi \in C^\infty(\mathbb{R}^d) : \phi \ \text{ has compact support}\}.$$

We will say that the sequence $\phi_n \in C_0^\infty(\Omega)$ converges to 0 if there exists a compact subset $K \subset \Omega$ such that

- $\phi_n(x) = 0 \ \ \forall x \in K, \forall n \in \mathbb{N}$,
- $\partial^\alpha \phi_n \rightrightarrows 0$ (uniformly converges to 0) on K.

We denote by $\mathcal{D}$ the space $C_0^\infty(\Omega)$ equipped with the topology described above. The space of continuous linear functionals on $\mathcal{D}$ is the dual space $\mathcal{D}'$, which is called distribution space. Next, we need the concept of the Dirac delta function, which is defined as a functional from $\mathcal{D}'$, such that

$$\langle \delta(x), \phi(x) \rangle_{\mathcal{D}' \times \mathcal{D}} = \phi(0) \ \ \forall \phi \in \mathcal{D}.$$

Definition 6.1. *The fundamental solution of some scalar elliptic partial differential operator L_x is denoted by $E(x, y)$ and defined by the following relation:*

$$L_x E(x, y) = \delta(x - y) \ \ \textit{in} \ \ \mathcal{D}'(\Omega), \quad \Omega \subset (\mathbb{R}^d)$$

or

$$\langle L_x E(x, y), \phi(x) \rangle_{D' \times D} = \langle \delta(x - y), \phi(x) \rangle := \phi(y) \ \ \textit{for every} \ \ \phi \in \mathcal{D}(\Omega),$$

where y is a parameter.

For instance the fundamental solution for he Laplace operator in d dimensional space ($d = 1, 2, 3$) is given by [5]:

$$E(x,y) = \begin{cases} \frac{1}{2}(1 - |x-y|) & \text{for} \quad d = 1 \\ -\frac{1}{2\pi}\log|x-y| & \text{for} \quad d = 2 \\ \frac{1}{4\pi}\dfrac{1}{|x-y|} & \text{for} \quad d = 3 \end{cases}$$

Lemma 6.2. *Let us assume that $\Omega \subset (\mathbb{R}^d)$ is bounded and $\partial\Omega := \Gamma \in C^{0,1} \cap PC^1$ (piecewise smooth). Then the following Green's formulas are valid:*

- *First Green's formula. $\forall u \in W_p^1(\Omega), \forall v \in W_q^2(\bar{\Omega})$ with $\frac{1}{p} + \frac{1}{q} = 1$. it holds:*

$$\int_\Omega u(x)\triangle v(x)dx = \int_\Gamma u(x)\frac{\partial v}{\partial n_x}(x)ds_x - \int_\Omega \nabla u(x)\nabla v(x)dx \tag{6.16}$$

- *Second Green's formula. $\forall u, v \in W_2^2 = H^2$ it holds:*

$$\int_\Omega (u(x)\triangle v(x) - v(x)\triangle u(x))dx = \int_\Gamma (u(x)\frac{\partial v}{\partial n_x}(x) - v(x)\frac{\partial u}{\partial n_x}(x))ds_x \tag{6.17}$$

- *Third Green's formula or **Representation formula**: if we take $v(\cdot) = E(\cdot, y)$ in the second Green's formula and if we take the traces on Γ, then we obtain [5]:*

$$\sigma(y)u(y) = -\int_\Gamma \frac{\partial E}{\partial n_x}(x,y)u(x)ds_x + \int_\Gamma E(x,y)\frac{\partial u}{\partial n_x}(x)ds_x + \int_\Omega E(x,y)(-\triangle u(x))dx \tag{6.18}$$

with

$$\sigma(y) := \begin{cases} 0 & \text{if} \quad y \in \mathbb{R}^d \setminus \bar{\Omega} \\ \frac{1}{2} & \text{almost everywhere on} \quad \Gamma \\ 1 & \text{if} \quad y \in \Omega \end{cases}$$

Note that once we know the Cauchy data $u(x)/_\Gamma$, $\dfrac{\partial u}{\partial n_x}(x)/_\Gamma$ and $-\triangle u$ on Ω, then using the representation formula we can calculate the value of the function u at any point in Ω. This gives the motivation to the Boundary Element Method.

In order to see how it works, let us consider the following mixed boundary value problem:

$$\begin{cases} -\triangle u(x) &= 0 \qquad \text{in} \quad \Omega, \\ u(x) &= g_D(x) \quad \text{on} \quad \Gamma_D(\text{Dirichlet boundary}), \\ \dfrac{\partial u}{\partial n_x}(x) &= g_N(x) \quad \text{on} \quad \Gamma_N(\text{Neumann boundary}), \end{cases} \tag{6.19}$$

where $\Omega \subset \mathbb{R}^2$ and $\Gamma_D \cup \Gamma_N = \Gamma := \partial\Omega$, $\Gamma_D \cap \Gamma_N = \emptyset$. We assume that the boundary Γ has 1-periodic parameter representation. Namely, we assume that there exists 1-periodic function $\mathrm{x}(t)$ such that $\mathrm{x}(0) = \mathrm{x}(1), |\dot{\mathrm{x}}(t)| \geq \kappa > 0$, and

$$\Gamma = \{x = \mathrm{x}(t) = [x_1(t), x_2(t)]^T \in \mathbb{R}^2 : 0 < t \leq 1\}.$$

Now, if we rewrite the Green's third formula and take into account that for our problem $-\triangle u(x) = 0$, then we will get the following identity:

$$\int_\Gamma E(x,y)v(x)\mathrm{d}s_x = \frac{1}{2}u(y) + \int_\Gamma \frac{\partial E}{\partial n_x}(x,y)u(x)\mathrm{d}s_x \quad \forall y \in \Gamma \tag{6.20}$$

where we denote by $v(x) := \dfrac{\partial u(x)}{\partial n_x}$ the Neumann data. Let us now 'formally' take the derivative of the (6.20) in the direction of n_y. We will get

$$\frac{1}{2}\frac{\partial u(y)}{\partial n_y} = -\frac{\partial}{\partial n_y}\int_\Gamma \frac{\partial E(x,y)}{\partial n_x}u(x)\mathrm{d}s_x + \int_\Gamma \frac{\partial E}{\partial n_y}(x,y)v(x)\mathrm{d}s_x. \tag{6.21}$$

Next, we introduce the following integral operators (BIO's):

- Single layer potential operator V: $Vv(y) := \displaystyle\int_\Gamma E(x,y)v(x)\mathrm{d}s_x$,
- Double layer potential operator K: $Ku(y) := \displaystyle\int_\Gamma \frac{\partial E}{\partial n_x}(x,y)u(x)\mathrm{d}s_x$,
- Adjoint double later potential operator K^* : $K^*v(y) := \displaystyle\int_\Gamma \frac{\partial E}{\partial n_y}(x,y)v(x)\mathrm{d}s_x$,
- Hyper singular potential operator D: $Du(y) := -\dfrac{\partial}{\partial n_y}\displaystyle\int_\Gamma \frac{\partial E(x,y)}{\partial n_x}u(x)\mathrm{d}s_x$.

Both in (6.20) and (6.21) the integrals over Γ we can split up into the sum of two integrals:

$$\int_{\Gamma} = \int_{\Gamma_D} + \int_{\Gamma_n}.$$

This will result in:

$$\begin{aligned} \frac{1}{2}u(y) &= \int_{\Gamma_D} E(x,y)v(x)\mathrm{d}s_x + \int_{\Gamma_N} E(x,y)v(x)\mathrm{d}s_x \\ &\quad - \int_{\Gamma_D} \frac{\partial E}{\partial n_x}(x,y)u(x)\mathrm{d}s_x - \int_{\Gamma_N} \frac{\partial E}{\partial n_x}(x,y)u(x)\mathrm{d}s_x \end{aligned} \tag{6.22}$$

and

$$\begin{aligned} \frac{1}{2}v(y) &= -\frac{\partial}{\partial n_y}\int_{\Gamma_D} \frac{\partial E(x,y)}{\partial n_x}u(x)\mathrm{d}s_x - \frac{\partial}{\partial n_y}\int_{\Gamma_N} \frac{\partial E(x,y)}{\partial n_x}u(x)\mathrm{d}s_x \\ &\quad + \int_{\Gamma_D} \frac{\partial E}{\partial n_y}(x,y)v(x)\mathrm{d}s_x + \int_{\Gamma_N} \frac{\partial E}{\partial n_y}(x,y)v(x)\mathrm{d}s_x. \end{aligned} \tag{6.23}$$

One can show that the boundary operators we have presented have the following properties [5]:

- $V = V^*$ is self adjoint in $\mathrm{H}^{-\frac{1}{2}}(\Gamma)$,
- $D = D^*$ is self adjoint in $\mathrm{H}^{\frac{1}{2}}(\Gamma)$,
- K^* is adjoint to K in $\mathrm{L}^2(\Gamma)$,
- D is positive semidefinite on the space $\mathrm{H}^{\frac{1}{2}}(\Gamma)$ and positive definite on $\mathrm{H}^{\frac{1}{2}}(\Gamma)/_{\ker D}$, means there exists a constant $\mu_{\mathrm{D}} > 0$ such that

$$\langle Du, u\rangle_{\mathrm{H}^{-\frac{1}{2}}\times\mathrm{H}^{\frac{1}{2}}} \geq \mu_{\mathrm{D}}||u||^2_{\mathrm{H}^{\frac{1}{2}}} \quad \forall u \in \mathrm{H}^{\frac{1}{2}}(\Gamma)/_{\ker D}$$

- if the diameter of Ω is small enough, namely, $\mathrm{diam}(\Omega) < 1$, then V is positive definite on the space $\mathrm{H}^{-\frac{1}{2}}(\Gamma)$, means there exists a constant $\mu_{\mathrm{V}} > 0$ such that:

$$\langle v, Vv\rangle_{\mathrm{H}^{-\frac{1}{2}}\times\mathrm{H}^{-\frac{1}{2}}(\Gamma)} \geq \mu_{\mathrm{V}}||v||^2_{\mathrm{H}^{\frac{1}{2}}} \quad \forall v \in \mathrm{H}^{\frac{1}{2}}(\Gamma)$$

We skip the technical details which can be found in [5] and present only the final result of Galerkin approximation of (6.22)- (6.23). So, taking into account the properties of boundary integral operators we have introduced and following the general procedure of Galerkin approximation scheme, we will get a saddle-point system of the following form:

$$\begin{bmatrix} V_h & -K_h \\ K_h^T & D_h \end{bmatrix} \begin{bmatrix} \bar{v} \\ \bar{u} \end{bmatrix} = \begin{bmatrix} \bar{f} \\ \bar{g} \end{bmatrix} \tag{6.24}$$

Since our purpose is to stress that the resulting system is a saddle-point system, we will not detail in the derivation of this system and the exact expressions for block matrices V_h, K_h, D_h as well as the right hand side vectors $\bar{f}$ and $\bar{g}$. What is more relevant for us in here is that the Boundary Method Technique is one of the method which lead to saddle-point problems.

6.3 Analysis and numerics for Mixed Variational Problems

In the previous section we have considered some examples of mixed variational formulations. As we have mentioned, in some cases mixed formulations are preferred to primal formulation. For some cases this comes from physical reasons. For instance, in case of linear elasticity, the stresses are more relevant unknowns than the displacements. In some other cases the reason for using the mixed formulation of a particular problem lies on algebraic level. For instance, in the same linear elasticity problem, for isotropic materials, the finite element disretization leads to a ill-conditioned matrix. In this section we will consider the mixed variational problems in quite general framework. Let us first introduce some notations. Let X and Λ be Hilbert spaces with inner products $(\cdot,\cdot)_X$, $(\cdot,\cdot)_\Lambda$ and induced norms $||\cdot||_X$, $||\cdot||_\Lambda$ respectively. The dual spaces of X and Λ will be denoted by X^* and Λ^* with duality products $\langle\cdot,\cdot\rangle_{X^*\times X}$ and $\langle\cdot,\cdot\rangle_{\Lambda^*\times\Lambda}$. All the examples we have considered so far are problems of this type:

Find $(u,\lambda) \in X \times \Lambda$ such that

$$\begin{cases} a(u,v) + b(v,\lambda) &= \langle F, v\rangle \quad \forall v \in X \\ b(u,\mu) &= \langle G, v\rangle \quad \forall \mu \in \Lambda \end{cases} \tag{6.25}$$

with $F \in X^*$ and $G \in \Lambda^*$ given linear continuous functionals. We also define the following operators:

- $A : X \to X^* : \quad \langle Au, v\rangle_{X^*\times X} := a(u,v) \quad \forall u, v \in X$

- $B : X \to \Lambda^* : \quad \langle Bu, \lambda\rangle_{\Lambda^*\times\Lambda} := b(u,\lambda) \quad \forall u \in X, \forall \lambda \in \Lambda$

Then in the operator notations (6.25) can be reformulated as:

Find $(u, \lambda) \in X \times \Lambda$ such that:

$$\left\{ \begin{array}{lcll} Au + B^*\lambda & = & f & \text{in } X^* \\ Bu & = & g & \text{in } \Lambda^* \end{array} \right. \tag{6.26}$$

where the operator $B^* : \Lambda \to X^*$ is the adjoint operator of B, uniquely defined by the relation

$$\langle B^*\lambda, u\rangle := \langle Bu, \lambda\rangle = b(u,\lambda) \quad \forall u \in X, \forall \lambda \in \Lambda.$$

Let

$$\mathrm{V}_0 := \{v \in X : b(v,\mu) = 0 \quad \forall \mu \in \Lambda\}$$

and

$$\mathrm{V}_g := \{v \in X : b(v,\mu) = \langle G, \mu\rangle \quad \forall \mu \in \Lambda\}.$$

In here the notations V_0 and V_g should not be confused with the same notations in previous chapters.

6.3.1 Brezzi's theorem

Now, we will state the fundamental theorem of mixed variational problems, which is called Brezzi's theorem [18].

Theorem 6.3. *Let the following assumptions are fulfilled*

- $F \in X^*$ *and* $G \in \Lambda^*$ *are given*

- *there exists* $\alpha_2 > 0$ *such that*

$$|a(u,v)| \le \alpha_2 ||u||_X ||v||_X, \quad \forall u, v \in X$$

- *there exists* $\beta_2 > 0$ *such that*

$$|b(u,\mu)| \le \beta_2 ||u||_X ||\mu||_\Lambda \quad \forall u \in X, \forall \mu \in \Lambda$$

- *there exists* $\beta_1 > 0$ *such that the so called LBB hold true:*

$$\sup_{v \in X} \frac{b(v,\mu)}{||v||_X} \geq \beta_1 ||\mu|| \quad \forall \mu \in \Lambda$$

- $a(\cdot,\cdot)$ *is elliptic on* $\ker B$*, namely, there exists* α_1 *such that*

$$a(u,v) = \langle Au, v \rangle \geq \alpha_1 ||v||_X^2 \quad \forall v \in \ker B = V_0$$

Then there exists unique solution (u, λ) *such that*

$$\begin{cases} a(u,v) + b(v,\lambda) &= \langle F, v \rangle \quad \forall v \in X \\ b(u,\mu) &= \langle G, v \rangle \quad \forall \mu \in \Lambda \end{cases}$$

or, equivalently

$$\begin{cases} Au + B^*\lambda &= f \quad in \quad X^* \\ Bu &= g \quad in \quad \Lambda^* \end{cases}$$

Moreover, the following apriori estimates for the solution are known:

$$\begin{cases} ||u||_X &\leq \frac{1}{\alpha_1}||F||_{X^*} + \frac{1}{\beta_1}(1 + \frac{\alpha_2}{\alpha_1})||G||_{\Lambda^*} \\ ||\lambda||_\Lambda &\leq \frac{1}{\beta_1}(1 + \frac{\alpha_2}{\alpha_1})||F||_{X^*} + \frac{\alpha_2}{\beta_1^2}(1 + \frac{\alpha_2}{\alpha_1})||G||_{\Lambda^*} \end{cases} \tag{6.27}$$

One can show that for all the examples we have considered the conditions of Brezzi's theorem hold true. So, the continuous mixed variational problems for these examples are well posed.

6.3.2 Mixed Finite Element Approximation

Let $X_h := \text{span}\{p^{(i)}; i = 1, 2, ..., n_h\} \subset X$ and $\Lambda_h := \text{span}\{q^{(i)}; i = 1, 2, ..., m_h\} \subset \Lambda$ be finite dimensional subspaces of X and Λ respectively. Then the Galerkin approximation to (6.25) reads as follows:

Find $(u_h, \lambda_h) \in X_h \times \Lambda_h$ such that

$$\begin{cases} a(u_h, v_h) + b(v_h, \lambda_h) &= \langle F, v_h \rangle \quad \forall v_h \in X_h \\ b(u_h, \mu_h) &= \langle G, v_h \rangle \quad \forall \mu_h \in \Lambda_h \end{cases} \tag{6.28}$$

Using the representation of u_h by the basis functions in X_h and the representation of λ by the basis functions in Λ_h :

$$u_h = \sum_{i=1}^{n_h} u^{(i)} p^{(i)} \quad \text{and} \quad \lambda_h = \sum_{i=1}^{m_h} \lambda^{(i)} q^{(i)}$$

we will get:

Find $\bar{u}_h = [u^{(i)}]_{i=1}^{n_h} \in \mathbb{R}^{n_h}$ and $\bar{\lambda}_h = [\lambda^{(i)}]_{i=1}^{m_h} \in \mathbb{R}^{m_h}$ such that

$$\begin{bmatrix} A_h & B_h^T \\ B_h & O \end{bmatrix} \begin{bmatrix} \bar{u}_h \\ \bar{\lambda}_h \end{bmatrix} = \begin{bmatrix} \bar{f}_h \\ \bar{g}_h \end{bmatrix} \tag{6.29}$$

where

$$\bar{f}_h = [\langle F, p^{(k)} \rangle]_k \in \mathbb{R}^{n_h} \quad \text{and} \quad \bar{g}_h = [\langle G, q^{(k)} \rangle]_k \in \mathbb{R}^{m_h}$$

$$A_h = [a(p^{(i)}, p^{(k)})]_{ik} \in \mathbb{R}^{n_h \times n_h}$$

$$B_h = [b(p^{(i)}, p^{(j)})]_{ij} \in \mathbb{R}^{n_h \times m_h}$$

Let

$$\mathrm{V}_{0h} := \{v_h \in X_h : b(v_h, \mu_h) = 0 \ \ \forall \mu_h \in \Lambda_h\}$$

and

$$\mathrm{V}_{gh} := \{v_h \in X_h : b(v_h, \mu_h) = \langle G, \mu_h \rangle \ \ \forall \mu_h \in \Lambda_h\}.$$

In general if the continuous problem satisfies the conditions of Brezzi's theorem the discrete problem does not necessarily satisfy the same conditions. The reasons for this is that in general

$$X_h \subset X \ \text{ and } \ \Lambda_h \in \Lambda \ \ \not\Rightarrow \ \ \mathrm{V}_{0h} \subset \mathrm{V}_h \ \text{ and } \ \mathrm{V}_{gh} \subset \mathrm{V}_g.$$

So, if $\mathrm{V}_{0h} \subset \mathrm{V}_h$ and $\mathrm{V}_{gh} \subset \mathrm{V}_g$ (which guarantees that the discrete LBB condition and the V_{0h}-ellipticity of $a(\cdot, \cdot)$ are satisfied), then the system (6.29) has unique solution. Moreover, for the solution we have the following result:

Theorem 6.4. *Let*

- $F \in X^*$ *and* $G \in \Lambda^*$ *are given*

- *there exists* $\alpha_2 > 0$ *such that*

$$|a(u, v)| \leq \alpha_2 ||u||_X ||v||_X, \quad \forall u, v \in X$$

- *there exists* $\beta_2 > 0$ *such that*

$$|b(u,\mu)| \leq \beta_2 ||u||_X ||\mu||_\Lambda \quad \forall u \in X, \forall \mu in \Lambda$$

- *there exists* $\beta_1 > 0$ *such that the so called LBB condition holds true:*

$$\sup_{v \in X} \frac{b(v,\mu)}{||v||_X} \geq ||\mu|| \quad \forall \mu \in \Lambda$$

- $a(\cdot,\cdot)$ *is elliptic on* $\ker B$*, namely, there exists* α_1 *such that*

$$a(u,v) = \langle Au, v \rangle \geq \alpha_1 ||v||_X^2 \quad \forall v \in \ker B = V_0$$

- *there exists* $\beta_{1h} > 0$ *such that the discrete LBB condition holds true:*

$$\sup_{v_h \in X_h} \frac{b(v_h,\mu_h)}{||v_h||_X} \geq ||\mu_h|| \quad \forall \mu_h \in \Lambda_h$$

- $a(\cdot,\cdot)$ *is elliptic on* V_{0h}*, namely, there exists* α_{1h} *such that*

$$a(u_h, v_h) = \langle Au_h, v_h \rangle \geq \alpha_{1h} ||v_h||_X^2 \quad \forall v_h \in V_{0h}$$

Then there exists a constant $C > 0$ *such that* $C \neq C(h)$ *(does not depend on* h*) and*

$$||u - u_h||_X + ||\lambda - \lambda_h||_\Lambda \leq C(\inf_{w_h \in X_h})||u - w_h||_X + \inf_{\gamma_h \in \Lambda_h})||\lambda - \gamma_h||_\Lambda$$

Now, the next step to be done is to solve the system (6.29). We can see that the system is a saddle-point system and one can use the HSS iterative method to solve such a system.

Chapter 7

Conclusion

We started this thesis by looking at the main steps of solving elliptic PDEs. We motivated our topic by considering a model problem, which is a representative of elliptic PDEs. We showed the well possedness of our model problem, namely, the existence of solution, its uniqueness and stability. This gave us the background for solving the problem.
From infinite dimensional space we transferred our problem to finite dimensional space. So, the continuous problem was then replaced by the discrete problem. Due to the choice of finite dimensional subspace of our original space, the solution of discrete problem approximates the solution of continuous problem quite accurate. As a result of discretization, by using Finite Element Tearing and Interconnecting method we arrived at a saddle-point problem. And, as a last step in solving our model problem, we applied the Hermitian and skew-Hermitian iterative method for solving the resulting saddle-point problem.
We used the HSS method first as a stationary iterative method and we saw that the number of iterations depends on the discretization parameter h. Then we applied HSS method as a preconditioner for Krylov subspace method GMRES. We also applied to our saddle-point system GMRES without preconditioner and we compared the results. What we observed was that the HSS preconditioning remarkably reduces the number of GMRES iterations. So, though the HSS iterative method as a iterative scheme itself does not converge rapidly, as a preconditioner for GMRES it worked quite well. Of course, due to some limitations we have restrained ourselves, the investigation we have started can be further continued. For instance, as a first further consideration we can increase the number of subdomains for FETI method (H) and look at the behavior of the HSS iteration for increasing number of subdomains of constant size. This is done, for instance, in [31] for linear elasticity problem. Another step in our investigation can be considering a 3D model problem

instead of 2D problem. A very important consideration can be a comparison of HSS iterative method with other solvers, for example block-structured preconditioners combined with Krylov subspace methods.

Bibliography

[1] L.C. Evans, *Partial Differential Equation,* American Mathematical Society, 1998.

[2] Z.Z. Bai, G.H. Golub, M.K. Ng, *Hermitian and skew-Hermitian splitting methods for non-Hermitian positive definite linear system,* SIAM J. Matrix Anal.Appl., 24(2003), pp.603-626.

[3] D. Gilbarg, N. S. Trudinger, *Elliptic Partial Differential Equations of Second Order,* Springer (2001).

[4] G. D. Smith, *Numerical Solution of Partial Differential Equations: Finite Difference Method,* Clarendon Press (1985).

[5] S. Rjasanow, O. Steinbach, *The Fast Solution of Boundary Integral Equations,* Springer (2007).

[6] Y. Saad, *Iterative methods for sparse linear systems,* SIAM (2003).

[7] T.A. Davis, *Direct Methods for Sparse Linear Systems (Fundamentals of Algorithms),* Society for Industrial and Applied Mathematic.

[8] G.H. Golub, A.J. Wathen, *Matrix computations*, 3rd ed., The Johns Hopkins Uiversity Press, Baltimore, 1996.

[9] D. Braess, *Theory, Fast solvers and Application in Solid Mechanics*, Cambridge University Press (2002).

[10] C. Fahrat, F.-X. Roux, *A method of finite element tearing and interconnecting and its parallel solution algorithm,* Internat.J.Numer.Methods Engrg., 32(1991), pp.1205 -1227.

[11] M. Benzi, G.H. Golub, J. Liesen, *Numerical solution of saddle point problems,* Acta Numerica (2005), pp. 1137.

[12] A. Toselli, O.B. Widlund, *Domain Decomposition Methods - Algorithms and Theory,* Springer (2004).

[13] K.C. Park, M.R. Justiono, Jr. and C.A. Felippa, *An algebraically partitioned FETI method for parallel structural analysis: Algorithm discription,* Internat. J. for Numerical methods in Engineering, vol. 40 , pp. 2717-2737 (1997).

[14] Z.Z. Bai, G.H. Golub, *Accelerated Hermitian and skew-Hermitian splitting iteration methods for saddle-point problems,* IMA Journal of Numerical Analysis.

[15] H.C. Elman, *Preconditioners for saddle point problems arising in computational fluid dynamics,* Appl.Numer.Math.,43(2002), pp.75-89.

[16] H.C. Elman, D.J. Silvester, A.J. Wathen, *Performance and analysis of saddle point preconditioners for the discrete steady-state Navier-Stokes equations,* Numer.Math.,90 (2002), pp.665-688.

[17] M. Fortin, R. Glowinski, *Augmented Lagrangian Methods: Application to the Numerical Solution of Boundary-Value Problems,* Stud.Math.Appl.15, North-Holland, Amsterdam, 1983.

[18] F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods*, Springer-Verlag, New-York, 1991.

[19] I. Perugia, V. Simoncini, *Bolck-diagonal and indefinite symmetric preconditioners for mixed finite element formulations,* Numer. Linear Algebra Appl.,7(2000), pp. 585- 616.

[20] M. Benzi, *Solution of equality constrained quadratic programming problems by a projection iterative method,* Rend. mat. Appl., 13 (1993), pp.275-296.

[21] P.E. Gill, W. Murray, M.H. Wright, *Practical Optimization,* Academic Press, New-York, 1981.

[22] P.E. Gill, W. Murray, D.B. Poneceleon, M.A. Saunders, *Preconditioners for indefinite systems arising in optimization,* SIAM J. Matrix Anal. Appl., 13 (1992), pp.292-311.

[23] N.I.M. Gould, M.E. Hribar, J. Nocedal, *On the solution of equality constrained quadratic programming problems arising in optimization,* SIAM J. Sci.Comput., 23(2001), pp.1376-1395.

[24] R. Glowinski, *Numerical Methods for Nonlinear Variational Problems,* Springer-Verlag, New-York, 1984.

[25] F. Brezzi, *A survey of mixed finite element methods,* NASA Contractor Report, 1987.

[26] R. Tezaur, *Analysis of Lagrange multiplier based domain decomposition,* PhD Thesis, University Karlova, Chech Republic, 1998.

[27] J. Alberty, C. Carstensen, S.A. Funken, *Remarks around 50 lines of Matlab: short finite element implementation,* Numerical Algorithms 20 (1999) 117137.

[28] A. Greenbaum, *Iterative methods for solving linear systems,* SIAM Frontiers Appl. MAth. 17, Philadelphia 1997.

[29] M. Benzi, G.H. Golub, *A preconditioner for generalized saddle point problems,* SIAM J.Matrix Anal. Appl., Vol. 26 No. 1, pp. 20 -41.

[30] Z.Z. Bai, G.H. Golub, C.K. Li, *Optimal parameter in Hermitian and skew-Hermitian splitting method for certain two-by-two block matrices,* SIAM J.Sci.Comput.,28,pp.583-603 (2006b).

[31] A. Klawonn, O.B. Widlund, *A domain decomposition method with lagrange multipliers and inexact solvers for linear elasticity,* SIAM J.Sci.Comput. vol.22, No.2, pp.1199-1219.

[32] M. Bebendorf, *Hierarchical Matrices,* Springer-Verlag, 2008.

[33] O.A. Ladyschenskaja, *Boundary Value problems of Mathematical Physiscs,* Hauka, Moscow 1973 (in Russian).

[34] O. Steinbach, *Numerical Approximation Methods for Elliptic Boundary Value Problems: Finite and Boundary Elements,* Springer-Verlag, 2008.

List of Figures

List of Tables