Mixed Finite Element Methods applied to Saddle Point Problems

Álvaro Pé de la Riva

Trabajo de fin de Máster en Matemáticas
Universidad de Zaragoza

Directores del trabajo: Francisco Gaspar Lorenz,
Carmen Rodrigo Cardiel.

Summary

In the context of PDEs approximation, saddle point problems take an important place. This type of problems appear when the weak formulation of a PDE system shows a special structure. Moreover, these problems can be ill-posed: There might not be unique solvability for every right hand side. They can be formulated in Hilbert spaces as a model variational problem, that we describe in Chapter 2. At this level, their existence and uniqueness of solution are not always ensured. That is the case when the so-called inf-sup conditions do not hold. However, when these conditions are satisfied, one is even able to bound the solution using constants related to them. This topic will be developed in Chapter 3, where some estimates for the solution of a saddle point problem are given.

Besides, Chapter 4 is devoted to the approximation of saddle point problems in finite dimensional spaces. One important issue is the fact that the fulfillment of the inf-sup conditions on Hilbert spaces does not imply the fulfillment of their discrete version on finite dimensional subspaces. Due to this, one has to be careful when choosing those subspaces. Furthermore, this choice may yield better or worse estimates depending on the size of the discrete inf-sup constants.

The discretisation of these problems using a numerical method, like the finite element method, yields a linear system whose matrix is called saddle point matrix. Again, we find particular properties on them. As we will see along Chapter 5, although one is interested on invertibility of these matrices, they are indefinite. Thus, one has to make a further study about solvability conditions.

Among the PDEs whose variational formulation yields a saddle point problem, we find the Stokes equations. In order to discretize the Stokes equations, it is usual to apply mixed finite element methods. This fact is due to the non fulfillment of the discrete inf-sup conditions if one applies the same finite element spaces for both variables. In addition, not every pair of finite element spaces guarantees the stability of the method. For instance, some troubles may take place like the spurious pressure modes and the locking phenomenon. This will be studied in Chapter 6, where we will comment on the stability of several finite element pairs for the Stokes equations.

Finally, the choice of a good finite element pair implies satisfactory numerical results and expected convergence rates. That is the case of the Minielement and the Taylor-Hood finite element method for the Stokes equations. In order to prove this statement, we show the obtained numerical results in Chapter 7. Furthermore, the interested reader can find the implementation of these mixed finite element methods in the appendices.

Resumen

Los problemas de tipo punto silla son sistemas lineales cuya matriz de coeficientes tiene una estructura específica. Su nombre se debe a que la solución a dicho sistema es precisamente un punto silla para un problema de minimización cuadrática asociado. Antes de estudiar las propiedades de estas matrices, que reciben el nombre de matrices de punto silla, consideramos una generalización de este tipo de problemas en espacios de Hilbert. Es en estos espacios donde introducimos un problema variacional que nos sirve como modelo para iniciar nuestro estudio.

En primer lugar, denotamos con V y Q a los dos espacios de Hilbert con los que vamos a desarrollar dicho modelo. Así, considerando dos formas bilineales $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ y $b(\cdot,\cdot) : V \times Q \longrightarrow \mathbb{R}$, formulamos el siguiente problema variacional:

$$\left\{ \begin{array}{l} \text{Dadas } f \in V' \text{ y } g \in Q', \text{ hallar } (u,p) \in V \times Q \text{ tal que:} \\ \begin{array}{rcll} a(u,v) & + & b(v,p) = \langle f, v \rangle_{V' \times V}, & \forall v \in V, \\ & & b(u,q) = \langle g, q \rangle_{Q' \times Q}, & \forall q \in Q. \end{array} \end{array} \right. \tag{1}$$

A su vez, este problema modelo puede ser reformulado empleando operadores lineales construidos a partir de las formas bilineales $a(\cdot,\cdot)$ y $b(\cdot,\cdot)$. Dichos operadores vienen dados respectivamente por:

$$A : V \longrightarrow V', \qquad B : V \longrightarrow Q'.$$

$$\begin{array}{l} A : V \longrightarrow V', \ \text{ con } \ \langle Au, v \rangle_{V' \times V} := a(u,v), \ \forall u, v \in V, \\ B : V \longrightarrow Q', \ \text{ con } \ \langle Bv, q \rangle_{Q' \times Q} := b(v,q), \ \forall v \in V, \ \forall q \in Q. \\ B^t : Q \longrightarrow V', \ \text{ con } \ \langle B^t q, v \rangle_{V' \times V} := \langle B^t q, v \rangle_{V' \times V} = \langle Bv, q \rangle_{Q' \times Q} = b(v,q), \ \forall v \in V. \end{array}$$

Así, el problema variacional (1) es equivalente al siguiente problema:

$$\left\{ \begin{array}{l} \text{Dadas } f \in V' \text{ y } g \in Q', \text{ hallar } (u,p) \in V \times Q \text{ tal que:} \\ \begin{array}{rcll} Au & + & B^t p = f, & \\ & & Bu = g. & \end{array} \end{array} \right. \tag{2}$$

Lo primero que nos planteamos es si el problema (1) tiene solución y si es única. Para abordar esta cuestión, llamamos $K := \operatorname{Ker} B$ y denotamos con $A_{KK'} := \Pi_{K'} A E_K$, donde $E_K : K \longrightarrow V$ es el operador extensión y $\Pi_{K'} : V' \longrightarrow K'$ es el operador proyección entre espacios duales. Con esta notación presentamos el siguiente resultado:

Teorema 1. *Para cada $(f,g) \in V' \times Q'$, el problema variacional dado en (1) tiene una única solución $(u,p) \in V \times Q$ sí y sólo si $A_{KK'}$ es un isomorfismo de K en K' y el operador B es suprayectivo.*

No obstante, se puede demostrar que las anteriores condiciones impuestas sobre los operadores lineales $A_{KK'}$ y B son equivalentes a las conocidas como condiciones inf-sup. En concreto:

$$B \text{ es suprayectivo} \quad \Longleftrightarrow \quad \inf_{q \in Q} \sup_{v \in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta.$$

$$A_{KK'} \text{ es un isomorfismo} \iff \exists \alpha_1 > 0 \text{ tal que} \begin{cases} \inf\limits_{v_0 \in K} \sup\limits_{w_0 \in K} \dfrac{a(v_0,w_0)}{\|v_0\|_V \|w_0\|_V} \geq \alpha_1, \\ \\ \inf\limits_{w_0 \in K} \sup\limits_{v_0 \in K} \dfrac{a(v_0,w_0)}{\|v_0\|_V \|w_0\|_V} \geq \alpha_1. \end{cases}$$

A parte de asegurar la existencia y unicidad de la solución del problema modelo, con las condiciones inf-sup nos es posible acotar la solución en la norma del espacio correspondiente. El siguiente resultado determina dichas cotas:

Teorema 2. *Sean β, α_1 dos constantes positivas con las que se cumplen las condiciones inf-sup de la forma bilineal $b(\cdot,\cdot)$ y de la forma bilineal $a(\cdot,\cdot)$ restringida a K. Entonces, para cada $(f,g) \in V' \times Q'$, el problema modelo (1) tiene una única solución $(u,p) \in V \times Q$ acotada por:*

$$\|u\|_V \leq \frac{1}{\alpha_1}\|f\|_{V'} + \frac{2\|a\|}{\alpha_1 \beta}\|g\|_{Q'},$$
$$\|p\|_Q \leq \frac{2\|a\|}{\alpha_1 \beta}\|f\|_{V'} + \frac{2\|a\|^2}{\alpha_1 \beta^2}\|g\|_{Q'}.$$

Nuestro siguiente objetivo es aproximar la solución del problema modelo (1) definido en espacios de Hilbert. Para ello, consideramos los subespacios $V_h \subset V$, $Q_h \in Q$ de dimensión finita y buscamos en ellos una aproximación que denotamos con $(u_h, q_h) \in V_h \times Q_h$. De este modo obtenemos el siguiente problema discreto:

$$\begin{cases} \text{Dadas } f \in V' \text{ y } g \in Q', \text{ hallar } (u_h, p_h) \in V_h \times Q_h \text{ tal que:} \\ a(u_h, v_h) + b(v_h, p_h) = \langle f, v_h \rangle_{V' \times V}, \quad \forall v_h \in V_h, \\ \qquad\qquad b(u_h, q_h) = \langle g, q_h \rangle_{Q' \times Q}, \quad \forall q_h \in Q_h. \end{cases} \tag{3}$$

A partir de este problema, podemos definir los respectivos operadores lineales y condiciones inf-sup discretas. De este modo es posible obtener un resultado similar al anterior teorema para acotar el error entre la aproximación $(u_h, p_h) \in V_h \times Q_h$ y $(u,p) \in V \times Q$. Sin embargo, hay que tener en cuenta que contamos con nuevas dificultades añadidas: Los núcleos de los operadores lineales discretos no tienen por qué estar contenidos en los núcleos de los operadores A y B. Esto último puede introducir aproximaciones defectuosas y nos obliga a ser cuidadosos con la elección de los espacios V_h y Q_h. A lo anterior se le suma que, el hecho de que se cumplan las condiciones inf-sup, no implica que se cumplan sus versiones discretas.

Cuando el problema (1) se corresponde con la formulación variacional de un sistema de ecuaciones en derivadas parciales, la discretización produce un sistema lineal. Además, las matrices de estos sistemas tienen una estructura especial y reciben el nombre de matrices de punto silla:

Definición 1. *Un sistema lineal describe un problema de tipo punto silla si se corresponde con un sistema estructurado en 2×2 bloques tal que*

$$\underbrace{\begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix}}_{\mathscr{A}} \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{pmatrix} f \\ g \end{pmatrix}}_{b}, \quad \textit{donde:} \quad \begin{cases} A \in \mathbb{R}^{n \times n}, \\ B_1, B_2 \in \mathbb{R}^{m \times n}, \quad n \geq m. \\ C \in \mathbb{R}^{m \times m}, \end{cases} \tag{4}$$

y dichos bloques satisfacen al menos una de las siguientes condiciones:

i) A es simétrica.

ii) La parte simétrica de A dada por $H = \frac{1}{2}\left(A + A^T\right)$ es semidefinida positiva.

iii) $B_1 = B_2 = B$.

iv) *C es simétrica y definida positiva.*

v) $C = 0_{m\times m}$.

Dada la anterior definición, nuestro interés se centra en estudiar las propiedades espectrales de la matriz para determinar cuándo es invertible. Así, bajo ciertas condiciones sobre los bloques constituyentes, sabremos en qué casos nos es posible asegurar que el sistema tiene solución única. Para profundizar en este estudio, empleando el complemento de Schur $S = -(C + B_2A^{-1}B_1^T)$ de la matriz A, obtenemos que:

$$\mathscr{A} = \begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B_2A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & A^{-1}B_1^T \\ 0 & I \end{pmatrix}.$$

Luego con esta factorización en matrices triangulares por bloques vemos que $\mathscr{A}$ será invertible sí y sólo sí los bloques A y S son matrices no singulares. En el capítulo 5 se incluyen varios resultados relacionados con esta idea, alguno destinado a mejorar una posible resolución mediante métodos iterativos.

Tras desarrollar la teoría sobre problemas de tipo punto silla, el propósito del trabajo es resolver las ecuaciones de Stokes, puesto que su formulación variacional adopta la estructura de un problema de punto silla. Las ecuaciones de Stokes describen el flujo de un fluido incompresible cuando el flujo es lento. Considerando un dominio acotado $\Omega \subset \mathbb{R}^n$, $n \geq 2$, la formulación fuerte de las ecuaciones de Stokes considerando condiciones de contorno de tipo Dirichlet viene dada por:

$$\begin{cases} -\boldsymbol{\Delta}\mathbf{u} + \nabla p = \mathbf{f}, & \text{en } \Omega, \\ \nabla \cdot \mathbf{u} = 0, & \text{en } \Omega, \\ \mathbf{u} = \mathbf{g}, & \text{en } \partial\Omega, \end{cases} \tag{5}$$

donde $\mathbf{u}$ es el vector de las velocidades y p es la presión. Además, diremos que $(\mathbf{u}, p)$ es una solución clasica de las ecuaciones de Stokes si resuelven (5) y $(\mathbf{u}, p) \in \left(\mathscr{C}^2(\Omega) \cap \mathscr{C}(\bar{\Omega})\right) \times \mathscr{C}^1(\Omega)$. Por otro lado, para llegar a la formulación débil de las ecuaciones de Stokes, definimos como espacios test $\mathbf{V} = \mathbf{H}_0^1$ y $Q = L_0^2(\Omega)$. De este modo, obtenemos que la formulación débil es la siguiente:

$$\begin{cases} \displaystyle\int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, dx - \int_\Omega (\nabla \cdot \mathbf{v})\, p\, dx = \int_\Omega \mathbf{f} \cdot \mathbf{v}\, dx, & \forall \mathbf{v} \in \mathbf{V}, \\ \displaystyle\int_\Omega (\nabla \cdot \mathbf{u}) \cdot q = 0, & \forall q \in Q. \end{cases} \tag{6}$$

Tras comprobar que el anterior problema variacional presenta una estructura de problema de tipo punto silla, nuestro siguiente paso es obtener una aproximación de la solución empleando elementos finitos mixtos. Es en este punto donde de nuevo las condiciones inf-sup discretas toman importancia, puesto que si elegimos de manera inadecuada los subespacios V_h y Q_h, el método empleado será inestable. Para evidenciar esto, entre los pares de elementos finitos inestables para las ecuaciones de Stokes mencionamos el par $\mathbf{P}_1/P_0$ y la aproximación $\mathbf{P}_1/P_1$: El primer par no cumple la condición inf-sup discreta mientras que el segundo es inestable debido a que el núcleo de $B_h := \Pi_{Q_h'} B E_{V_h}$ no está contenido en Ker B.

Finalmente, el método del Minielemento y los elementos de Taylor-Hood convergen y producen buenas aproximaciones de la solución. Para comprobar que las cotas del error se cumplen, en el último capítulo del trabajo presentamos los resultados númericos obtenidos al implementar dichos métodos. Además incluimos los resultados obtenidos tras implementar el par $\mathbf{P}_1/P_1$, de este modo el lector puede comprobar que este último par de elementos finitos es claramente inestable.

Contents

Chapter 1

Previous results

In order to study the uniqueness and existence of solutions for saddle point problems, first some functional analysis results are needed. This work focuses on analyzing the characteristics of this type of problems and furthermore to solve some examples numerically, so we will treat the previous results briefly giving some references to the reader for detailing proofs. Moreover, the notation given along this chapter will be used in order to state the variational formulation that we will deal with.

Definition. A normed linear space $(H, \|\cdot\|_H)$ is said to be complete if for every Cauchy sequence $\{v_n\}$ in H there exists an element $v \in H$ such that $v_n \longrightarrow v$, that is $\|v - v_n\|_H \xrightarrow{n} 0$.

Definition. A pre-Hilbert space $(H, \|\cdot\|_H)$ is a space provided with a scalar product $\langle\cdot,\cdot\rangle_H$ and a norm $\|\cdot\|_H$ satisfying the usual properties of every norm including the so called parallelogram identity:

$$\|v+u\|_H^2 + \|v-u\|_H^2 = 2\left(\|u\|_H^2 + \|v\|_H^2\right).$$

Definition. A Hilbert space is a pre-Hilbert space that is complete. (i.e.: A Hilbert space is a Banach space whose norm satisfies the parallelogram identity)

Definition. Let V and W be Hilbert spaces and let $f : V \longrightarrow W$ be a linear mapping. We say that f is bounded or that it is continuous if there exists a constant $\lambda \in \mathbb{R}$ such that

$$\|fv\|_W \leq \lambda \|v\|_V, \qquad \forall v \in V.$$

We say that f is bounding if there exists a constant $\mu \in \mathbb{R}$ such that

$$\|fv\|_W \geq \mu \|v\|_V, \qquad \forall v \in V.$$

We will assume V and W are always Hilbert spaces. Then, we will denote by $\mathscr{L}(V,W)$ the linear space composed by the set of all $f : V \longrightarrow W$ linear continuous operators. Let's also introduce a norm for this space:

$$\|f\|_{\mathscr{L}(V,W)} := \sup_{v \in V} \frac{\|fv\|_W}{\|v\|_V}.$$

We define the dual space of a Hilbert space V as the space of all linear continuous functionals $f : V \longrightarrow \mathbb{R}$. The dual space of V will be denoted with V'. For the dual space a dual norm is also given:

$$\|f\|_{V'} := \sup_{v \in V} \frac{|f(v)|}{\|v\|_V} = \sup_{v \in V} \frac{\langle f, v\rangle_{V' \times V}}{\|v\|_V}.$$

The next one is the first important result:

Theorem 1.1 (Banach Theorem). *Let V and W be Hilbert spaces and let $M \in \mathscr{L}(V,W)$ be a one to one mapping. Then, its inverse operator $M^{-1} : W \longrightarrow V$ is also continuous.*

Proof. [2] □

Theorem 1.2 (**Riesz Theorem**). *Let V be a Hilbert space. For every $l \in V'$ continuous, there exists an unique $z \in V$ such that*

$$\langle l, v\rangle_{V'\times V} = \langle z, v\rangle_V, \qquad \forall v \in V.$$

From another point of view, if we define the operator $R_V : V \longrightarrow V'$ that to each $z \in V$ associates the functional $f_z = R_V z \in V'$ defined as

$$\langle f_z, v\rangle_{V'\times V} = \langle z, v\rangle_V,$$

then R_V is one to one and $\|R_V\|_{\mathscr{L}(H,H')} = \|R_V^{-1}\|_{\mathscr{L}(H',H)} = 1$.

Proof. See [9], pp. 97. □

Definition. Let Z be a subspace of a Hilbert space V. We call Z^0 the polar space of Z defined as

$$Z^0 := \{f \in V' \text{ such that } f(z) = 0 \ \forall z \in Z\} = R_V(Z^\perp).$$

Bilinear Forms and transposed operators: Let V and Q be Hilbert spaces. A bilinear form $b : V \times Q \longrightarrow \mathbb{R}$ is continuous if $\exists \mu_b \in \mathbb{R}$ such that $b(v,q) \leq \mu_b \|v\|_V \|q\|_Q$, $\forall v \in V$, $\forall q \in Q$. We denote with $\mathscr{B}(V \times Q, \mathbb{R})$ the set of all bilinear continuous operators from $V \times Q$ to $\mathbb{R}$. The norm of the continuous bilinear form b is defined as follows:

$$\|b\|_{\mathscr{B}(V\times Q,\mathbb{R})} := \sup_{v\in V,\, q\in Q} \frac{b(v,q)}{\|v\|_V \|q\|_Q}.$$

Let B be a linear operator from V to Q' defined as:

$$\langle Bv, q\rangle_{Q'\times Q} := b(v,q), \qquad \forall v \in V, \forall q \in Q.$$

Furthermore it is possible to prove that the operator B is continuous if and only if the associated bilinear form b is continuous. We can associate to the linear operator $B : V \longrightarrow Q'$ the so called transposed operator $B^t : Q \longrightarrow V'$ given by

$$\langle v, B^t q\rangle_{V\times V'} := \langle Bv, q\rangle_{Q'\times Q} = b(v,q).$$

These operators are deeply related, even their norm is the same in the corresponding spaces:

$$\|B\|_{\mathscr{L}(V,Q')} = \|B^t\|_{\mathscr{L}(Q,V')} = \|b\|_{\mathscr{B}(V\times Q,\mathbb{R})}.$$

It might be useful for future results to add the definition of the next subspaces:

$$\text{Ker } B := \big\{v \in V \big| b(v,q) = 0, \ \forall q \in Q\big\} = \big\{v \in V \big| \langle v, B^t q\rangle_{V\times V'} = 0, \ \forall q \in Q\big\},$$
$$\text{Ker } B^t := \big\{q \in Q \big| b(v,q) = 0, \ \forall v \in V\big\} = \big\{q \in Q \big| \langle Bv, q\rangle_{Q'\times Q} = 0, \ \forall v \in V\big\}.$$

Theorem 1.3 (**Banach Closed Range Theorem**). *Let V and Q be Hilbert spaces and let B be a linear continuous operator from V to Q'. Set:*

$$K := Ker\ B \subset V, \qquad H := Ker\ B^t \subset Q.$$

Then, the following statements are equivalent:

- *Im B is closed in Q'.*
- *Im B^t is closed in V'.*
- *$K^0 = Im\ B^t$.*

- $H^0 = Im\ B$.
- *There exists* $L_B \in \mathscr{L}(Im\ B, K^{\perp})$ *and* $\exists \beta > 0$ *such that* $B(L_B(g)) = g \quad \forall g \in Im\ B$ *and moreover* $\beta \|L_B g\|_V \leq \|g\|_{Q'} \quad \forall g \in Im\ B$.
- *There exists* $L_{B^t} \in \mathscr{L}(Im\ B^t, H^{\perp})$ *and* $\exists \beta > 0$ *such that* $B^t(L_{B^t}(f)) = f \quad \forall f \in Im\ B^t$ *and moreover* $\beta \|L_{B^t} f\|_Q \leq \|f\|_{V'} \quad \forall f \in Im\ B^t$.

The assumption of the special case in which B is surjective leads us to the next corollary:

Corollary 1.1. *Let V and Q be Hilbert spaces, and let B be a linear continuous operator from V to Q'. Then the following statements are equivalent:*

- $Im\ B = Q'$.
- $Im\ B^t$ *is closed and* B^t *is injective.*
- B^t *is bounding:* $\exists \beta \in \mathbb{R}$ *such that* $\|B^t q\|_{V'} \geq \beta \|q\|_Q, \quad \forall q \in Q$.
- *There exists* $L_B \in \mathscr{L}(Q', V)$ *such that* $B(L_B(g)) = g \quad \forall g \in Q'$ *with* $\|L_B\| = \dfrac{1}{\beta}$.

Finally, as a consequence of this corollary we reach a first result about uniqueness and existence of solution for variational problems: The Lax-Milgram lemma. In the next chapter, some proofs will be supported by this result.

Lax-Milgram Lemma. Let V be a Hilbert space and let $a(\cdot,\cdot)$ be a bilinear continuous form on V. Assume that a is coercive ($\exists \alpha > 0$ such that $a(v,v) \geq \alpha \|v\|_V^2$, $\forall v \in V$). Then for every $f \in V'$, the problem:

$$\text{Find } u \in V \text{ such that } a(u,v) = \langle f, v \rangle_{V' \times V}, \ \forall v \in V,$$

has an unique solution.

Proof. See [9], pp. 140. □

Chapter 2

Saddle point problems on Hilbert spaces

The previous section allows us to obtain results about existence and uniqueness of solutions for saddle point problems on Hilbert spaces adding some conditions due to the form of this kind of problems. We will assume along this section that V and Q are Hilbert spaces and we are given two bilinear forms $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ and $b(\cdot,\cdot) : V \times Q \longrightarrow \mathbb{R}$ with their corresponding linear continuous operators $A : V \longrightarrow V'$ and $B : V \longrightarrow Q'$. At this point we build the following model problem:

$$\left\{ \begin{array}{l} \text{Given } f \in V' \text{ and } g \in Q', \text{ find } (u,p) \in V \times Q \text{ such that:} \\ a(u,v) \quad + \quad b(v,p) = \langle f, v \rangle_{V' \times V}, \quad \forall v \in V, \\ \qquad\qquad\quad b(u,q) = \langle g, q \rangle_{Q' \times Q}, \quad \forall q \in Q. \end{array} \right. \tag{2.1}$$

This model problem can be rewritten. The following problem is an equivalent formulation of (2.1):

$$\left\{ \begin{array}{lll} Au & + & B^t p = f \quad \text{in } V', \\ & & Bu = g \quad \text{in } Q', \end{array} \right. \tag{2.2}$$

where we are dealing with the corresponding linear continuous operators. Assuming that $a(\cdot,\cdot)$ is symmetric, our model problem can be understood as the optimality conditions for the following minimisation problem:

$$\inf_{Bv=g} \frac{1}{2} a(v,v) - \langle f, v \rangle_{V' \times V}.$$

Then, variable p takes the role of the Lagrange multiplier associated with the constraint $Bu = g$ and the saddle point problem concerning us is:

$$\inf_{v \in V} \sup_{q \in Q} \left\{ \frac{1}{2} a(v,v) + b(v,q) - \langle f, v \rangle_{V' \times V} - \langle g, q \rangle_{Q' \times Q} \right\}.$$

Now, it is possible to introduce a first result about existence and uniqueness of solution for our model problem, noting that the second equation of (2.1) requires the surjectivity of the linear operator B in order to ensure the existence of solution:

Theorem 2.1. *Assume that Im $B = Q'$ and $a(\cdot,\cdot)$ is coercive on $K := Ker\, B$. Then, for every $(f,g) \in V' \times Q'$ problem (2.1) has a unique solution.*

Proof. As far as B is surjective, Corollary 1.1. ensures us that there exists $L_B \in \mathscr{L}(Q',V)$ such that $B(L_B g) = g,\ \forall g \in Q'$. Let us call $u_g := L_B g$, then $Bu_g = g$ and considering $u_0 := u - u_g$ such that $Bu = g$ it implies that $u_0 \in K$. Furthermore, testing the first equation of (2.1) with every $v_0 \in K$ and using that $a(\cdot,\cdot)$ is a bilinear form, one obtains:

$$a(u_0, v_0) = \langle f, v_0 \rangle_{V' \times V} - a(u_g, v_0), \quad \forall v_0 \in K.$$

At this point, the Lax-Milgram Lemma ensures that there is an unique solution $u_0 \in K$ to this equation. Now we define the functional $l : V \longrightarrow \mathbb{R}$ given by $l(v) = \langle f, v\rangle_{V'\times V} - a(u,v)$ for every $v \in V$. This functional vanishes identically for every $v \in K$, so $l \in K^0$ and then the Banach Closed Range Theorem provides us $l \in K^0 = \text{Im } B^t$. Then, there exists an elemenent $p \in Q$ such that $B^t p = l$ and yields:

$$\langle l, v\rangle_{V'\times V} = \langle B^t p, v\rangle_{V'\times V} = \langle f, v\rangle_{V'\times V} - a(u,v),\ \forall v \in V \iff a(u,v) + b(v,p) = \langle f, v\rangle_{V'\times V},\ \forall v \in V.$$

And hence the first equation is satisfied, but thanks to the definition of $u := u_g + u_0$ one also gets the second one:

$$B(u) = B(u_g + u_0) \underset{u_0\in K}{=} Bu_g + 0 \underset{u_g := L_B g}{=} g.$$

In addition to existence, we now prove uniqueness of solution for the model problem: Using linearity, it is sufficient to check that the case with $f = 0$ and $g = 0$ has an unique solution. Testing the first equation on this problem with $v = u$ we get $a(u,u) = 0$. On one hand, the coercivity of the bilinear form $a(\cdot,\cdot)$ provides $u = 0$. On another hand we get $B^t p = 0$, which also implies $p = 0$ because as it was seen in Corollary 1.1, Im $B = Q' \iff B$ is bounding.

□

A generalisation of the previous theorem provides us a necessary and sufficient condition for existence and uniqueness of solution. Let us call again $K := \text{Ker } B \subset V$ and let us define the operator $A_{KK'}$ as the composition $A_{KK'} := \Pi_{K'} A E_K$, where $E_K : K \longrightarrow V$ is the extension operator and $\Pi_{K'}$ is the projection between the corresponding dual spaces $\Pi_{K'} : V' \longrightarrow K'$.

Theorem 2.2. *Let us consider the model problem and the operator $A_{KK'}$ defined as before. Then, for every $(f,g) \in V' \times Q'$:*

There exists a unique solution $(u,p) \in V \times Q \iff$ $\begin{cases} A_{KK'} \text{ is an isomorphism from K to K'} \\ Im\ B = Q' \end{cases}$

Proof. We start assuming $A_{KK'}$ is an isomorphism and B is surjective, then we get the uniqueness and existence of solution for our model problem by following the same steps given in the proof of the last theorem, but this time we use $A_{KK'}$ is an isomorphism in order to get the existence of a solution $u_0 \in K$, instead of Lax-Milgram theorem.
Conversely, let us assume the model problem has a unique solution for every $(f,g) \in V' \times Q'$. Using as example the case $f = 0$ and any $g \in Q'$, our assumption yields that for every $g \in Q'$ there exists $u \in V$ such that:

$$\begin{cases} Au + & B^t p & = 0, & \\ & Bu & = g, & \text{for every } g \in Q'. \end{cases}$$

That means Im $B = Q'$, so this proves that the operator B is surjective. At this point, we can prove $A_{KK'}$ is an isomorphism: For every $\varphi \in K'$, we can build a model problem with $f_\varphi = E_{K'}\varphi$ and $g = 0$, each one owns a unique solution (u_φ, p_φ) and $u_\varphi \in K$ since $g = 0$. Now, if we test the first equation of the model problem with $v_0 \in K$:

$$a(u_\varphi, v_0) = \langle f_\varphi, v_0\rangle_{K'\times K} = \langle \varphi, v_0\rangle_{V'\times V}, \qquad \forall v_0 \in K, \varphi \in K'.$$

Then, we get $A_{KK'} u_\varphi = \varphi$ and hence $A_{KK'}$ is surjective. Finally, in order to prove that $A_{KK'}$ is also injective, let us assume $A_{KK'} w = 0$ for some $w \in K,\ w \neq 0$. This would imply $a(w, v_0) = 0,\ \forall v_0 \in K$ and hence $Aw \in K^o$. Now we apply the Banach Closed Range Theorem and then one gets $Aw \in$ Im B^t since $K^o =$ Im B^t. Moreover, it leads us to ensure the existence of a $p_w \in Q,\ p_w \neq 0$ such that $B^t p_w = A_w$ and then $(w, -p_w)$ would be a solution for the model homogeneous problem, making uniqueness to be lost. Therefore such a w $\neq 0$ cannot exist and it shows that $A_{KK'}$ must be injective. □

Chapter 3

Stability Constants and inf-sup Conditions

In this section, we point out an important topic: The Banach Closed Range Theorem ensures us that the operator B is surjective if and only if its transpose operator is bounding. In a practical point of view, it is easier to prove that one operator is bounding than its surjectiveness, for this reason our next goal is to define the best possible constant $\beta > 0$ that fits in the bounding condition for the linear operator B^t. Note that using the definition of norm in a dual space:

$$\|B^t q\|_{V'} \geq \beta \|q\|_Q \quad \forall q \in Q \quad \Longleftrightarrow \quad \inf_{q\in Q} \frac{\|B^t q\|_{V'}}{\|q\|_Q} \geq \beta \quad \Longleftrightarrow \quad \inf_{q\in Q} \sup_{v\in V} \frac{b(v,q)}{\|v\|_V \|q\|_Q} \geq \beta.$$

With a similar argument one can also gets that:

$$A_{KK'} \text{ is an isomorphism} \quad \Longleftrightarrow \quad \exists \alpha_1 > 0 \text{ such that } \begin{cases} \inf\limits_{v_0\in K} \sup\limits_{w_0\in K} \dfrac{a(v_0,w_0)}{\|v_0\|_V \|w_0\|_V} \geq \alpha_1, \\ \\ \inf\limits_{w_0\in K} \sup\limits_{v_0\in K} \dfrac{a(v_0,w_0)}{\|v_0\|_V \|w_0\|_V} \geq \alpha_1. \end{cases}$$

Thus, we get an equivalence for the second condition given in the if and only if theorem for existence and uniqueness. Now, we are also interested in obtaining a bound for the solution of the model problem. Furthermore, these inf-sup constants α_1 and β will appear in these bounds. Hence, the inf-sup constants take an important role in this chapter. Before showing the main result about this topic, first we need two useful lemmas for its proof:

Lemma 3.1. *Let V be a Hilbert space and let $a(\cdot,\cdot)$ be a symmetric bilinear continuous form on V. Assume that $a(v,v) \geq 0, \ \forall v \in V$. Then,*

$$(a(v,w))^2 \leq a(v,v)a(w,w), \qquad \forall v,w \in V,$$

and for the associated operator A:

$$\|Av\|_{V'}^2 \leq \|a\| a(v,v) \equiv \|A\| \langle Av, v\rangle_{V'\times V}.$$

Proof. Let us consider $v,w \in V$. Applying that $a(\cdot,\cdot)$ is a bilinear form, one gets that:

$$a(v+\lambda w, v+\lambda w) = \lambda^2 a(w,w) + 2\lambda a(v,w) + a(v,v).$$

Thanks to our hypothesis, $a(v+\lambda w, v+\lambda w) \geq 0, \ \forall v,w \in V, \ \forall \lambda \in \mathbb{R}$ so the equation $\lambda^2 a(w,w) + 2\lambda a(v,w) + a(v,v) = 0$ can not have two different real roots, that is:

$$\Delta \equiv 4\left(a(v,w)\right)^2 - 4a(v,v)a(w,w) \leq 0.$$

And finally dividing it by four, one gets the expected result. □

Lemma 3.2. *Let V be a Hilbert space and let $a(\cdot,\cdot)$ be a symmetric bilinear continuous form on V. Assume that $a(v,v) \geq 0,\ \forall v \in V$. Then, the double inf-sup condition on the restriction of $a(\cdot,\cdot)$ to K implies coercivity of $a(\cdot,\cdot)$ on the kernel K.*

Proof. From the inf-sup condition, using the previous lemma one gets:

$$\alpha_1 \leq \inf_{v\in K} \sup_{w\in K} \frac{a(v,w)}{\|v\|_V\|w\|_V} \implies \alpha_1\|v\|_V \leq \sup_{w\in K} \frac{a(v,w)}{\|w\|_V} \implies \alpha_1^2\|v\|_V^2 \leq \left(\sup_{w\in K} \frac{a(v,w)}{\|w\|_V}\right)^2 \implies$$
$$\alpha_1^2\|v\|_V^2 \leq \sup_{w\in K} \frac{a(v,w)^2}{\|w\|_V^2} \leq \sup_{\substack{w\in K \\ \|w\|_V=1}} \frac{a(v,v)a(w,w)}{\|w\|_V^2} \leq \|a\|\, a(v,v).$$

Thus, the result is reached with $\alpha_0 = \alpha_1^2/\|a\|$. □

Now we are ready to present a first bound result for solutions of our model problem. The next theorem is only valid for the case of Im $B = Q'$, further results can be found on the literature, but won't be needed for the problems treated in this work.

Theorem 3.1. *Assume that there exists two positive constants β, α_1 such that the inf-sup condition on $b(\cdot,\cdot)$ and the double inf-sup condition on the restriction of $a(\cdot,\cdot)$ to the kernel K are satisfied. Then, for every $(f,g) \in V' \times Q'$, the model problem has a unique solution $(u,p) \in V \times Q$ satisfying the following bounds:*

$$\|u\|_V \leq \frac{1}{\alpha_1}\|f\|_{V'} + \frac{2\|a\|}{\alpha_1\beta}\|g\|_{Q'},$$
$$\|p\|_Q \leq \frac{2\|a\|}{\alpha_1\beta}\|f\|_{V'} + \frac{2\|a\|^2}{\alpha_1\beta^2}\|g\|_{Q'}.$$

If, moreover, $a(\cdot,\cdot)$ is symmetric and satisfies $a(v,v) \geq 0\ \forall v \in V$, then we have the improved estimates:

$$\|u\|_V \leq \frac{1}{\alpha_0}\|f\|_{V'} + \frac{2\|a\|^{1/2}}{\alpha_0^{1/2}\beta}\|g\|_{Q'},$$
$$\|p\|_Q \leq \frac{2\|a\|^{1/2}}{\alpha_0^{1/2}\beta}\|f\|_{V'} + \frac{\|a\|}{\beta^2}\|g\|_{Q'},$$

where $\alpha_0 > 0$ satisfies the coercivity condition for the bilinear form $a(\cdot,\cdot)$ restricted to K (also known as the elker condition).

Proof. The Closed Range Theorem ensures us the existence of a continuous lifting operator $L \in \mathscr{L}(Q',V)$ such that we can split $u = u_0 + u_g$ where $u_0 \in \text{Ker } B$ and $u_g := Lg$. Continuity of the lifting operator provides:

$$\|u_g\|_V = \|Lg\|_V \leq \|L\|\|g\|_{Q'} \overset{\|L\|=\frac{1}{\beta}}{\implies} \|u_g\|_V \leq \frac{\|g\|_{Q'}}{\beta}.$$

Then we reach the following estimate: $\|Au_g\|_{V'} \leq \|a\|\|u_g\|_V \leq \dfrac{\|a\|}{\beta}\|g\|_{Q'}$.

The solution for our model problem with $g \equiv 0$ is $(u_0,p_0),\ u_0,p_0 \in \text{Ker } B$. Considering this problem, one easily gets that $\|Au_0\|_{V'} = \|f - Au_g\|_{V'}$, but using the inf-sup condition for the bilinear form $a(\cdot,\cdot)$:

$$\frac{\|Au_0\|_{V'}}{\|u_0\|_V} = \sup_{v\in V} \frac{\langle Au_0, v\rangle_{V'\times V}}{\|u_0\|_V\|v\|_V} \geq \sup_{v\in K} \frac{a(u_0,v)}{\|u_0\|_V\|v\|_V} \geq \inf_{u_0\in K} \sup_{v\in K} \frac{a(u_0,v)}{\|u_0\|_V\|v\|_V} \geq \alpha_1,\ \forall u_0 \in \text{Ker } B \implies$$
$$\implies \|u_0\|_V \leq \frac{1}{\alpha_1}\|f - Au_g\|_{V'} \leq \frac{1}{\alpha_1}\left(\|f\|_{V'} + \frac{\|a\|}{\beta}\|g\|_{Q'}\right).$$

And then the estimate for $\|u\|_V$ is:

$$\|u\|_V = \|u_0+u_g\|_V \leq \frac{1}{\alpha_1}\|f\|_{V'} + \left(\frac{\|a\|}{\alpha_1\beta}+\frac{1}{\beta}\right)\|g\|_{Q'} \leq \frac{1}{\alpha_1}\|f\|_{V'} + \frac{2\|a\|}{\alpha_1\beta}\|g\|_{Q'}.$$

Note that the last inequality is trivial because $1 \leq \|a\|/\alpha_1$. Now we are looking for an estimate of p. In order to get it, we recall the inf-sup condition for the bilinear form $b(\cdot,\cdot)$, that is equivalent to:

$$\beta\|p\|_Q \leq \|B^T p\|_{V'}, \qquad \forall p \in Q.$$

We return again to the model problem. This time we consider its general version with a solution $(u,p) \in V \times Q$. Then the first equation and the previous estimates yield:

$$\|p\|_Q \leq \frac{1}{\beta}\|f-Au\|_{V'} \leq \frac{1}{\beta}\left(\|f\|_{V'}+\|a\|\|u\|_V\right) \leq \left(\frac{1}{\beta}+\frac{\|a\|}{\alpha_1\beta}\right)\|f\|_{V'} + \frac{2\|a\|^2}{\alpha_1\beta^2}\|g\|_{Q'} \leq$$
$$\leq \frac{2\|a\|}{\alpha_1\beta}\|f\|_{V'} + \frac{2\|a\|^2}{\alpha_1\beta^2}\|g\|_{Q'}.$$

Furthermore, assuming $a(\cdot,\cdot)$ is symmetric and $a(v,v) \geq 0, \ \forall v \in V$, Lemma 3.1 and Lemma 3.2 lead us easily to the improved estimates. In order to get this last result, we build the following estimates and collecting them the proof shall be ended. Those estimates require the statement of two cases derived from our model problem, where we recall $u := u_0 + u_g$ with $u_0 \in$ Ker B, u_g such that $Bu_g = g$, and $p = p_0 + p_g$ satisfying:

$$(1)\begin{cases} Au_0 + & B^t p_0 & = f, \\ & Bu_0 & = 0, \end{cases} \qquad (2)\begin{cases} Au_g + & B^t p_g & = 0, \\ & Bu_g & = g. \end{cases}$$

Let us begin with the first problem in order to get estimates of u_0 and Au_0. Taking the first equation multiplied to the right by u_0:

$$\left.\begin{array}{c} \langle Au_0,u_0\rangle_{V'\times V} + \langle B^t p_0,u_0\rangle_{V'\times V} = \langle f,u_0\rangle_{V'\times V} \\ \langle B^t p_0,u_0\rangle_{V'\times V} = \langle Bu_0,p_0\rangle_{V'\times V} = 0 \end{array}\right\} \implies \langle Au_0,u_0\rangle_{V'\times V} = \langle f,u_0\rangle_{V'\times V}.$$

Now applying the ellipticity condition on the kernel one gets easily:

$$\alpha_0\|u_0\|_V^2 \leq \langle Au_0,u_0\rangle = \langle f,u_0\rangle \leq \|f\|_{V'}\|u_0\|_V \implies \|u_0\|_V \leq \frac{\|f\|_{V'}}{\alpha_0} \quad , \quad \langle Au_0,u_0\rangle \leq \frac{\|f\|_{V'}^2}{\alpha_0} \implies$$
$$\implies \|Au_0\|_{V'} \leq \|a\|^{1/2}\left(\langle Au_0,u_0\rangle\right)^{1/2} \leq \frac{\|a\|^{1/2}}{\alpha_0^{1/2}}\|f\|_{V'}.$$

We are interested on an estimate for $\|p\|_{Q'}$, then we recall the given inf-sup condition on the B^t operator and, thanks to the previous estimates:

$$\beta\|p_0\|_Q \leq \|B^t p_0\|_{V'} = \|f-Au_0\|_{V'} \quad \implies \quad \|p_0\|_Q \leq \frac{1}{\beta}\|f-Au_0\|_{V'} \leq \frac{1}{\beta}\|f\|_{V'} + \frac{1}{\beta}\frac{\|a\|^{1/2}}{\alpha_0^{1/2}}\|f\|_{V'} \leq$$
$$\leq \frac{2\|a\|^{1/2}}{\beta\alpha_0^{1/2}}\|f\|_{V'}.$$

Here, we come back to the problem (2) derived from the model problem. Multiplying the first equation by u_g:

$$\langle Au_g,u_g\rangle + \langle B^t p_g,u_g\rangle = 0 \quad \implies \quad \langle Au_g,u_g\rangle = -\langle B^t p_g,u_g\rangle = -\langle g,u_g\rangle \leq \|p_g\|_Q\|g\|_{Q'}.$$

The next step in our proof is to get an estimate of $\|p_g\|$ by $\left(\langle Au_g,u_g\rangle\right)^{1/2}$. The inf-sup condition provides the existence of a $\tilde{u}$ such that $\beta\|\tilde{u}\|\|p_g\| \leq b(\tilde{u},p_g)$ and then combining it with the result of Lemma 3.1:

$$\beta\|\tilde{u}\|_V\|p_g\|_Q \le \langle B\tilde{u}, p_g\rangle = -\langle A\tilde{u}, u_g\rangle \le \|a\|^{1/2}\|\tilde{u}\|_V\left(\langle Au_g, u_g\rangle\right)^{1/2} \implies \|p_g\|_Q \le \frac{\|a\|^{1/2}}{\beta}\left(\langle Au_g, u_g\rangle\right)^{1/2}.$$

At this point, we combine the last inequality with $\langle Au_g, u_g\rangle \le \|p\|_Q\|g\|_{Q'}$, so inmediately:

$$\langle Au_g, u_g\rangle \le \frac{\|a\|^{1/2}}{\beta}\left(\langle Au_g, u_g\rangle\right)^{1/2}\|g\|_{Q'} \implies \left(\langle Au_g, u_g\rangle\right)^{1/2} \le \frac{\|a\|^{1/2}}{\beta}\|g\|_{Q'} \implies \|p_g\|_Q \le \frac{\|a\|}{\beta^2}\|g\|_{Q'}.$$

Finally, we recall again Lemma 3.1 in order to get the new estimate for $\|u_g\|_V$:

$$\alpha_0\|u_g\|_V^2 \le \langle Au_g, u_g\rangle \le \|p_g\|_V\|g\|_{Q'} \le \frac{\|a\|}{\beta^2}\|g\|_{Q'}^2 \quad \implies \quad \|u_g\|_V \le \frac{\|a\|}{\alpha_0^{1/2}\beta}\|g\|_{Q'}.$$

Hence, collecting the estimates above, one obtains the improved estimates for $\|u\|_V$ and $\|p\|_Q$.

□

Chapter 4

Approximation of Saddle Point Problems

In order to ensure the existence and uniqueness of solutions for saddle point problems in the context of Hilbert spaces, we have already seen that the inf-sup conditions for the bilinear form $a(\cdot,\cdot)$ and the linear form $b(\cdot,\cdot)$ are needed. However, when we build an approximation problem such that its solution belongs to a finite dimensional subspace, some additional issues might appear. Firstly, we start defining the corresponding discrete operators and finite dimensional subspaces:

Let $V_h \subset V$ and $Q_h \subset Q$ be finite dimensional subspaces. We define the extension operators $E_{V_h} : V_h \longrightarrow V$, $E_{Q_h} : Q_h \longrightarrow Q$ and the projection operators $\Pi_{V_h} : V \longrightarrow V_h$, $\Pi_{Q_h} : Q \longrightarrow Q_h$. Along this chapter, we will use the notation $a(u_h, v_h) := a(E_{V_h}u_h, E_{V_h}v_h)$ and $b(v_h, q_h) := b(E_{V_h}v_h, E_{Q_h}q_h)$. We also define their associated linear operators as follows:

$$\begin{aligned} B_h v_h &= \Pi_{Q_h'} B E_{V_h} v_h, \ \ \forall v_h \in V_h, & \qquad B_h^t q_h &= \Pi_{V_h'} B^t E_{Q_h} q_h, \ \ \forall q_h \in Q_h, \\ A_h v_h &= \Pi_{V_h'} A E_{V_h} v_h, \ \ \forall v_h \in V_h, & \qquad A_h^t v_h &= \Pi_{V_h'} A^t E_{V_h} v_h, \ \ \forall v_h \in V_h. \end{aligned}$$

Finally we consider the discrete kernels:

$$\begin{aligned} K_h \equiv \ \text{Ker}\, B_h &:= \left\{ v_h \in V_h \;\middle|\; b(v_h, q_h) = 0, \ \ \forall q_h \in Q_h \right\}, \\ H_h \equiv \ \text{Ker}\, B_h^t &:= \left\{ q_h \in Q_h \;\middle|\; b(v_h, q_h) = 0, \ \ \forall v_h \in V_h \right\}. \end{aligned}$$

Hence, we are ready to state the approximation problem:

$$\begin{cases} \text{Given } f \in V' \text{ and } g \in Q', \text{ find } (u_h, p_h) \in V_h \times Q_h \text{ such that:} \\ \begin{aligned} a(u_h, v_h) \ \ + \ \ b(v_h, p_h) &= \langle f, v_h \rangle_{V' \times V}, \quad \forall v_h \in V_h, \\ b(u_h, q_h) &= \langle g, q_h \rangle_{Q' \times Q}, \quad \forall q_h \in Q_h. \end{aligned} \end{cases}$$

At this point, it is natural to assume that the necessary and sufficient condition for getting the existence and uniqueness of solution in this case might be the following pair of inf-sup conditions:

$$\begin{aligned} &\forall h > 0, \ \exists \alpha_1^h > 0 : \qquad \inf_{v_0^h \in K_h} \sup_{w_0^h \in K_h} \frac{a(v_0^h, w_0^h)}{\|v_0^h\|_{K_h} \|w_0^h\|_{K_h}} = \inf_{w_0^h \in K_h} \sup_{v_0^h \in K_h} \frac{a(v_0^h, w_0^h)}{\|v_0^h\|_{K_h} \|w_0^h\|_{K_h}} \geq \alpha_1^h, \\ &\forall h > 0, \ \exists \beta_h > 0 : \qquad \inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_Q} \geq \beta_h. \end{aligned}$$

The cases where the last two parameters do not depend on h are really interesting, but it is not always possible to find them. For this reason, in general, we will assume their dependence on h, which will refer to the mesh size chosen.

Here there are some important considerations to make: At the approximation problem, everything seems to work fine, but it is not like that due to the choice of our finite dimensional subspaces V_h and Q_h. Hence, we remark that:

- The kernel K_h is not in general a subspace of K.
- The kernel H_h is not in general a subspace of $H := \text{ Ker } B^t$.
- The existence of $\alpha_1 > 0$ satisfying the inf-sup condition for the operator $a(\cdot,\cdot)$ at K does not imply the existence of $\alpha_1^h > 0$ for the same inf-sup condition at K_h.
- The existence of $\beta > 0$ satisfying the inf-sup condition for the operator $b(\cdot,\cdot)$ at (Q,V) does not imply the existence of $\beta_h > 0$ for the same inf-sup condition at (Q_h,V_h).

It worths also to remark that the property $K_h \subseteq K$ is a nice one but we can dispense with it. Nevertheless, the property $H_h \subseteq H$ is absolutely important, because when that inclusion does not hold then the solution of the approximation problem will be determined up to elements of H_h. In fact, those elements use to be called spurious numerical artefacts because they might yield false solutions. For example, let (u_h, p_h) be the solution of our approximation problem, $p_0^h \in H_h \setminus H$. Then the pair $(u_h, p_h + p_0^h)$ is a solution of the approximation problem, but not at all an approximation of the original one. Therefore sometimes we restrict the space Q to Q/H. This is the case of the Stokes equations, where we take $L_0^2(\Omega)$ in the place of $L^2(\Omega)$.

As we have seen, it is important to hold the inclusion $H_h \subseteq H$ in order to control the solution of our discrete problem. Due to this, we are now interested in the next proposition:

Proposition 4.1. *Let $b(\cdot,\cdot) : V \times Q \longrightarrow \mathbb{R}$ be a bilinear operator, $V_h \subset V$, $Q_h \subset Q$ finite dimensional subspaces. Let also B_h, B_h^t be the discrete operators as we defined before, $H_h = \text{ Ker } B_h^t$. Suppose that there exists a linear operator $\Pi_{V_h} : V \longrightarrow V_h$ such that*

$$b(v - \Pi_{V_h} v, q_h) = 0 \qquad \forall v \in V,\ q_h \in Q_h.$$

Then $\Pi_{Q_h'}(\text{Im } B) \subseteq \text{Im } B_h$ and equivalently $H_h \subseteq H$.

Proof. We have that $b(v - \Pi_{V_h} v, q_h) = \langle Bv, q_h\rangle - \langle BE_{V_h}\Pi_{V_h} v, q_h\rangle = 0,\ \ \forall q_h \in Q_h$. And applying the $\Pi_{Q_h'}$ operator to the last equality we get:

$$\langle \Pi_{Q_h'} Bv, q_h\rangle - \langle B_h \Pi_{V_h} v, q_h\rangle = 0,\ \ \forall q_h \in Q_h.$$

Then we get the following equality between linear operators:

$$\Pi_{Q_h'} Bv = B_h \Pi_{V_h} v,\ \ \forall v \in V \Longrightarrow \Pi_{Q_h'}(\text{ Im } B) \subseteq \text{ Im } B_h.$$

Our next point is to prove that the last inclusion $\Pi_{Q_h'}(\text{ Im } B) \subseteq \text{ Im } B_h$ implies that $H_h \subseteq H$. Let us consider $q_0^h \in H_h$, that is:

$$\langle v_h, B_h^t q_0^h\rangle = 0 \iff \langle \Pi_{Q_h'} Bv_h, q_0^h\rangle = \langle Bv_h, q_0^h\rangle = \langle v_h, B^t q_0^h\rangle = \langle v_h, \Pi_{Q_h'} B^t q_0^h\rangle = 0, \quad \forall v_h \in V_h.$$

We recover the definition of polar space and realise that, then, $q_0^h \in \big(\text{ Im } \Pi_{Q_h'} BE_{V_h}\big)^o = \big(\text{ Im } B_h\big)^o$. The inclusion $\Pi_{Q_h'}(\text{ Im } B) \subseteq \text{ Im } B_h$ clearly implies the inclusion between the corresponding polar spaces $\big(\text{ Im } B_h\big)^o \subseteq \big(\Pi_{Q_h'}(\text{ Im } B)\big)^o$, due to their closeness, and, then, we get $q_0^h \in \big(\Pi_{Q_h'}(\text{ Im } B)\big)^o$. Finally we have $\langle v, B^t q_0^h\rangle = \langle \Pi_{Q_h'} Bv, q_0^h\rangle = 0,\ \forall v \in V$, and hence $q_0^h \in H_h$. □

Operators holding the proposition above are called B-compatible operators, which are really important to find inf-sup conditions. Here, we should remark that the converse for last proposition is also true, even one could go further and find an equivalence between them and the condition $K_h \subseteq K$.

Another important topic is the stability and error estimates of our finite dimensional solutions, so we are interested in finding bounds on $\|u - u_h\|_V$ and $\|p - p_h\|_Q$. In order to get them, first we will consider the

best approximations $\tilde{u} \in V_h$, $\tilde{p} \in Q_h$ for the solution $(u,p) \in V \times Q$ of our model problem. Secondly we take $(u_h, p_h) \in V_h \times Q_h$ as the solution of the discretised problem and, then, we will bound the distance between them in terms of the distance of $(\tilde{u}, \tilde{p})$ from (u,p). Finally, using the triangle inequality we will be able to obtain an expression of $\|u - u_h\|_V$ and $\|p - p_h\|_Q$ in terms of the distance between the original solution and its best approximation.

Hence, let us define the so-called approximation errors:

$$E_u := \inf_{v_h \in V_h} \|u - v_h\|_V, \qquad E_p := \inf_{q_h \in Q_h} \|p - q_h\|_Q.$$

Now using the linearity of operators $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ we substract the continuous problem to the discretised problem and then one easily introduces $\tilde{u}$ and $\tilde{p}$ getting the variational problem:

$$\begin{cases} \text{Find } (u_h - \tilde{u}, p_h - \tilde{p}) \in V_h \times Q_h \text{ such that} \\ a(u_h - \tilde{u}, v_h) + b(v_h, p_h - \tilde{p}) = a(u - \tilde{u}, v_h) + b(v_h, p - \tilde{p}) \quad \forall v_h \in V_h, \\ b(u_h - \tilde{u}, q_h) = b(u - \tilde{u}, q_h) \quad \forall q_h \in Q_h. \end{cases}$$

Therefore, for every $(\tilde{u}, \tilde{p}) \in V_h \times Q_h$ we can ensure that $(u_h - \tilde{u}, p_h - \tilde{p})$ is the unique solution of the variational problem above in $V_h \times Q_h$. For brevity, it is useful to define the operators $\tilde{f}$ and $\tilde{g}$ as follows:

$$\begin{aligned} \langle \tilde{f}, v_h \rangle_{V_h' \times V_h} &= a(u - \tilde{u}, v_h) + b(v_h, p - \tilde{p}) \quad \forall v_h \in V_h, \\ \langle \tilde{g}, q_h \rangle_{Q_h' \times Q_h} &= b(u - \tilde{u}, q_h) \quad \forall q_h \in Q_h. \end{aligned}$$

Now we are ready to state a first basic estimate theorem just combining the estimates given in chapter 3:

Theorem 4.1. *(The basic estimate) Let $a(\cdot,\cdot) : V \times V \longrightarrow \mathbb{R}$ and $b(\cdot,\cdot) : V \times Q \longrightarrow \mathbb{R}$ be two bilinear operators, $V_h \subset V$, $Q_h \subset Q$ finite dimensional subspaces. Let us consider the restrictions of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to these subspaces and assume that the corresponding inf-sup conditions hold for $\alpha_1^h > 0$ and $\beta_h > 0$. Let $f \in V'$, $g \in Q'$ and $\tilde{f} \in V_h'$, $\tilde{g} \in Q_h'$ be as we have defined. Then, for every $(\tilde{u}, \tilde{p}) \in V_h \times Q_h$ combining the estimates for the variational problem above, the continuous variational problem with unique solution $(u,p) \in V \times Q$ and the discretised problem with its unique solution $(u_h, p_h) \in V_h \times Q_h$, one gets the estimates:*

$$\begin{aligned} \|u_h - \tilde{u}\|_V &\leq \frac{1}{\alpha_1^h} \|\tilde{f}\|_{V_h'} + \frac{2\|a\|}{\alpha_1^h \beta_h} \|\tilde{g}\|_{Q_h'}, \\ \|p_h - \tilde{p}\|_Q &\leq \frac{2\|a\|}{\alpha_1^h \beta_h} \|\tilde{f}\|_{V_h'} + \frac{2\|a\|^2}{\alpha_1^h \beta_h^2} \|\tilde{g}\|_{Q_h'}. \end{aligned}$$

If moreover $a(\cdot,\cdot)$ is symmetric and semidefinite positive in V_h, then we have the improved estimates:

$$\begin{aligned} \|u_h - \tilde{u}\|_V &\leq \frac{1}{\alpha_0^h} \|\tilde{f}\|_{V_h'} + \frac{2\|a\|^{1/2}}{(\alpha_0^h)^{1/2} \beta_h} \|\tilde{g}\|_{Q_h'}, \\ \|p_h - \tilde{p}\|_Q &\leq \frac{2\|a\|^{1/2}}{(\alpha_0^h)^{1/2} \beta_h} \|\tilde{f}\|_{V_h'} + \frac{\|a\|}{\beta_h^2} \|\tilde{g}\|_{Q_h'}, \end{aligned}$$

with $\alpha_0^h = \dfrac{(\alpha_1^h)^2}{M_a^h}$, where $M_a^h = \sup_{v_h \in V_h} \sup_{w_h \in V_h} \dfrac{a(v_h, w_h)}{\|v_h\|_V \|w_h\|_V}$.

Furthermore, if we apply to the previous result the inequalities below

$$\|\tilde{f}\|_{V_h'} \leq \|a\| \|u - \tilde{u}\|_V + \|b\| \|p - \tilde{p}\|_Q, \qquad \|\tilde{g}\|_{Q_h'} \leq \|b\| \|u - \tilde{u}\|_V,$$

and the triangular inequalities for $\|u_h - u\|_V$ and $\|p_h - p\|_Q$:

$$\|u_h - u\| \leq \|u_h - \tilde{u}\| + \|\tilde{u} - u\| \leq \|u_h - \tilde{u}\| + \frac{\|a\|\|b\|}{\alpha_1^h \beta_h}\|\tilde{u} - u\|,$$
$$\|p_h - p\| \leq \|p_h - \tilde{p}\| + \|\tilde{p} - p\| \leq \|p_h - \tilde{p}\| + \frac{\|a\|\|b\|}{\alpha_1^h \beta_h}\|\tilde{p} - p\|.$$

Then we obtain as a corollary the so-called basic error estimate:

Corollary 4.1. *(The basic error estimate) Under the same assumptions, we have the following error estimates:*

$$\|u_h - u\|_V \leq \frac{4\|a\|\|b\|}{\alpha_1^h \beta_h} E_u + \frac{\|b\|}{\alpha_1^h} E_p,$$
$$\|p_h - p\|_Q \leq \left(\frac{2\|a\|^2}{\alpha_1^h \beta_h} + \frac{2\|a\|\|b\|}{\beta_h^2}\right) E_u + \frac{3\|a\|\|b\|}{\alpha_1^h \beta_h} E_p.$$

If moreover $a(\cdot,\cdot)$ is symmetric and semidefinite positive in V_h, then we have the improved estimates:

$$\|u_h - u\|_V \leq \left(\frac{2\|a\|}{\alpha_0^h} + \frac{2\|a\|^{1/2}\|b\|}{(\alpha_0^h)^{1/2}\beta_h}\right) E_u + \frac{\|b\|}{\alpha_0^h} E_p,$$
$$\|p_h - p\|_Q \leq \left(\frac{2\|a\|^{3/2}}{(\alpha_0^h)^{1/2}\beta_h} + \frac{\|a\|\|b\|}{\beta_h^2}\right) E_u + \frac{3\|a\|^{1/2}\|b\|}{(\alpha_0^h)^{1/2}\beta_h} E_p.$$

Finally, we close this chapter with the statement of a commonly used estimate avaiable for a particular case, when the inf-sup condition on the bilinear operator $b(\cdot,\cdot)$ is satisfied with a constant $\beta > 0$ independent of h and the bilinear operator $a(\cdot,\cdot)$ satisfies uniformly the discretised elker condition:

Theorem 4.2. *Let $(u,p) \in V \times Q$ and $(u_h, p_h) \in V_h \times Q_h$ be respectively solutions of the continuous and discretised problems. Assume that the inf-sup condition*

$$\inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_Q} \geq \beta > 0,$$

is satisfied and let $a(\cdot,\cdot)$ be uniformly coercive on $K_h := Ker\ B_h$:

$$a(v_0^h, v_0^h) \geq \alpha_0 \|v_0^h\|_V^2, \quad \forall v_0^h \in K_h.$$

Then, one has the following estimate with C a constant depending on $\|a\|$, $\|b\|$, β and α_0 but independent of h:

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq C \left(\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - p_h\|_Q \right).$$

Moreover, when we have the inclusion $K_h \subset K$, we have the better estimate

$$\|u - u_h\|_V \leq C \inf_{v_h \in V_h} \|u - v_h\|_V.$$

Remark: In the previous results, when $H := \text{Ker } B^t$ is not zero, then the constant β_h goes to zero when h tends to zero. Instead of Q, one must apply these results with $Q_{/H}$, that is the case of Stokes equations, for example.

Chapter 5

Properties of Saddle Point Matrices

This section is devoted to introduce the basic properties of saddle point matrices and how we could make a good use of them in order to get numerical solutions through adaptive methods. Due to its block structure, it will be appropiate to handle with a Schur complement reduction for the appearing saddle point matrices. Thus, our management of the linear system will be easier and we will improve our understanding of the problem we are dealing with. In the previous chapters, we could deduce how the structure of a saddle point problem is, but now we state its general structure as follows:

Definition. A linear system is said to describe a saddle point problem if its form corresponds to a block 2×2 linear system such that

$$\underbrace{\begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix}}_{\mathscr{A}} \begin{pmatrix} x \\ y \end{pmatrix} = \underbrace{\begin{pmatrix} f \\ g \end{pmatrix}}_{b} \quad \text{where:} \quad \begin{cases} A \in \mathbb{R}^{n\times n}, \\ B_1, B_2 \in \mathbb{R}^{m\times n}, \quad n \geq m. \\ C \in \mathbb{R}^{m\times m}, \end{cases} \tag{5.1}$$

and the constitutive blocks satisfy at least one of the following conditions:

i) A is symmetric.

ii) The symmetric part of A, $H = \frac{1}{2}\left(A + A^T\right)$, is positive semidefinite.

iii) $B_1 = B_2 = B$.

iv) C is symmetric and positive semidefinite.

v) $C = 0_{m\times m}$.

On the other hand, the saddle point problem receives its name due to the case where all the conditions above are hold: Then, one can prove that the solution of our problem is also the solution of a quadratic programming problem whose matrix $\mathscr{A}$ contains exactly n positive eigenvalues and m negative eigenvalues. However, we need still to develop a Schur complement reduction in order to check this.

At this point, we introduce a Schur complement based block factorization of $\mathscr{A}$ valid if A is nonsingular. Although other factorizations could be found by means of C Schur complement when A is singular, in general A will be nonsingular for us. More detailed information about this topic can be found in [8]. Then, under the last assumption, our saddle point matrix admits the following block triangular factorization:

$$\mathscr{A} = \begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix} = \begin{pmatrix} I & 0 \\ B_2 A^{-1} & I \end{pmatrix} \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix} \begin{pmatrix} I & A^{-1}B_1^T \\ 0 & I \end{pmatrix},$$

where $S = -(C + B_2 A^{-1} B_1^T)$ is the Schur complement of A in $\mathscr{A}$. The reader might note this factorization yields a lot of information about the spectral properties of our saddle point matrix because the eigenvalues of $\mathscr{A}$ are then given by the corresponding eigenvalues of A and its Schur complement S. Moreover, we are able to deal directly with solvability conditions for saddle point problems due to $\mathscr{A}$ is nonsingular if and only if S is (assuming A is nonsingular). Therefore, our next point is to find out under which assumptions S is nonsingular and which restrictions on the blocks A, B_1, B_2 and C are necessary/sufficient to get the desired result.

By this way, our first approach to avoid 0 to be an eigenvalue of $\mathscr{A}$ is to require A and S to be definite. Hence, now we consider the so-called symmetric case: Let us assume A to be symmetric positive definite, $B_1 = B_2 = B$ and $C = 0$. Clearly, the Schur complement of A takes the form $S = -BA^{-1}B^T$ and this matrix is symmetric negative semidefinite matrix, so we can hold that $\mathscr{A}$ is invertible if and only if rank $(B) = m$. That would imply S was symmetric negative definite and then we would back to the problem

$$\begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix},$$

whose solution solves the following quadratic programming problem:

$$\text{Find } x \in \mathbb{R}^n \text{minimizing} \begin{cases} \mathscr{J}(x) = \dfrac{1}{2} x^T A x - f^T x, \\ \text{s.t. } Bx = g, \end{cases}$$

where $y \in \mathbb{R}^m$ takes the role of Lagrange multiplier. Now we compute the associated Lagrangian for the problem above, its gradient and hessian matrix:

$$\begin{cases} \mathscr{L}_{x,y} = \dfrac{1}{2} x^T A x - f^T x + (Bx - g)^T y, \\ \nabla \mathscr{L}_{x,y} = \left(Ax - f^T + By, (Bx - g)^T \right) = 0, \\ H\mathscr{L}_{x,y} = \begin{pmatrix} A & B^T \\ B & 0 \end{pmatrix}. \end{cases}$$

Finally, the Hessian matrix of the Lagrangian provides us the answer to why our problem receives the saddle point problem name: As far as $H\mathscr{L}_{x,y}$ has the eigenvalues of A and S, it has n positive eigenvalues and m negative eigenvalues. Thus, $H\mathscr{L}_{x,y}$ is indefinite and therefore the solution of the quadratic programming problem we are looking for is a saddle point.

On the other hand, the last assumption we did about the constitutive blocks can be relaxed for C, if we just assume $C \neq 0$ is symmetric positive definite, we obtain $S = -(C + BA^{-1}B^T)$ is symmetric negative semidefinite also. However, for this discussion we need an additional condition in order to ensure the invertibility of $\mathscr{A}$ and it is exposed in the following theorem:

Theorem 5.1. *Assume A is symmetric positive definite, $B_1 = B_2 = B$, and C is symmetric positive semidefinite. If Ker $(C) \cap$ Ker $(B^T) = \{0\}$, then the saddle point matrix $\mathscr{A}$ is nonsingular. In particular $\mathscr{A}$ is invertible if B has full rank.*

The assumption taken for A can be relaxed. There are many results with A positive semidefinite, but easy counterexamples appear when A is considered indefinite. In that case $\mathscr{A}$ might be singular as it is in the following example:

$$\mathscr{A} = \begin{pmatrix} 1 & 0 & -1 \\ 0 & -1 & 1 \\ -1 & 1 & 0 \end{pmatrix}, \qquad \text{with } A = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \; B^T = \begin{pmatrix} -1 \\ 1 \end{pmatrix} \text{ and } C = 0.$$

As we have already checked, the requirement for A can be relaxed only to be positive semidefinite. Hence the next result that we expose with its respective proof provides an if and only if condition for nonsingularity of $\mathscr{A}$:

Theorem 5.2. *Assume A is symmetric positive semidefinite, $B_1 = B_2 = B$ has full rank and $C = 0$. Then the saddle point matrix $\mathscr{A}$ holds:*

$$\mathscr{A} \text{ is nonsingular } \Leftrightarrow \text{ Ker } (A) \cap \text{ Ker } (B) = \{0\}.$$

Proof. First, let $(x,y) \in \mathbb{R}^{n+m}$ be such that $\mathscr{A}(x,y)^T = 0$. In order to prove the condition above is a sufficient condition, let us see that (x,y) must be zero using Ker $(A) \cap$ Ker $(B) = \{0\}$. We note that the pair (x,y) satisfies

$$Ax + B^T y = 0, \qquad Bx = 0.$$

Then, one easily deduces multiplying on the left by x^T that $x^T Ax = -x^T B^T y = -(Bx)^T y = 0$. Moreover, $Ax = 0$ since A is symmetric positive semidefinite and then it follows that $x \in$ Ker $(A) \cap$ Ker (B). Consequently we have $B^T y = 0$ and rank $(B) = m$ implies $y \in \mathbb{R}^m$ must be zero. At this point, it follows that $(x,y) = 0$ and, thus, $\mathscr{A}$ is nonsingular.

In order to prove the converse, let us assume there exists $x \in$ Ker $(A) \cap$ Ker (B) with $x \neq 0$. Inmediately one notes that $\mathscr{A}(x,0)^T = 0$ implying that $\mathscr{A}$ is singular. This proves the given condition Ker $(A) \cap$ Ker $(B) = \{0\}$ is also necessary. □

More results can be found holding positive definiteness over the symmetric part of A instead of A making possible even to relax the $C = 0$ condition. Then we will be able to talk about solvability of saddle point problems where matrix A does not have to be symmetric. The proof of the following theorem follows the same steps for proving Theorem 5.1 and we refer the interested reader to [12] for finding more details about these results.

Theorem 5.3. *Assume $H = \frac{1}{2}(A + A^T)$ the symmetric part of A is positive semidefinite, $B_1 = B_2 = B$ has full rank and C is symmetric positive semidefinite. Then the following holds:*

i) Ker $(H) \cap$ Ker $(B) = \{0\} \Longrightarrow \mathscr{A}$ *is invertible.*

ii) $\mathscr{A}$ *is invertible* $\Longrightarrow$ Ker $(A) \cap$ Ker $(B) = \{0\}$.

All these results serve us for ensuring that $\mathscr{A}$ is invertible in some cases thanks to the nonsingularity of A and its Schur complement S. For those cases, an explicit expression for the inverse of a general saddle point matrix can be given and it is:

$$\mathscr{A}^{-1} = \begin{pmatrix} A & B_1^T \\ B_2 & -C \end{pmatrix}^{-1} = \begin{pmatrix} A^{-1} + A^{-1}B_1^T S^{-1} B_2 A^{-1} & -A^{-1}B_1^T S^{-1} \\ -S^{-1}B_2 A^{-1} & S^{-1} \end{pmatrix}.$$

However, a more interesting case appears when we assume again A to be symmetric positive definite, $B_1 = B_2 = B$, $C = 0$, S nonsingular and in addition we consider $g = 0$. Then, the explicit expression for $\mathscr{A}^{-1}$ yields the following explicit expression for the solution of the respective saddle point problem in words of the constitutive blocks:

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} (I + A^{-1}B^T S^{-1} B)A^{-1} f \\ S^{-1} B A^{-1} f \end{pmatrix}.$$

This case would correspond to the discretisation of the Stokes equations with a Dirichlet boundary condition using for instance Taylor-Hood finite elements.

Our next point is to study the spectral properties of the saddle point matrices. These properties are really important when we solve the system by an iterative method and they determine for each case the methods which are usable. As we have seen before, the symmetric case has an important role in the saddle point problems. Therefore, we start our study assuming that A is symmetric positive definite, $B_1 = B_2 = B$ has full rank and C is symmetric positive semidefinite. Then, the matrix $\mathscr{A}$ is congruent with the following block diagonal matrix $\begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}$, where the Schur complement of A given by $S = -(C + BA^{-1}B^T)$ is consequently symmetric negative definite. We can ensure this last since the factorization of $\mathscr{A}$ holds for this case as follows:

$$\begin{pmatrix} I & 0 \\ -BA^{-1} & I \end{pmatrix} \begin{pmatrix} A & B^T \\ B & -C \end{pmatrix} \begin{pmatrix} I & -A^{-1}B^T \\ 0 & I \end{pmatrix} = \begin{pmatrix} A & 0 \\ 0 & S \end{pmatrix}.$$

With this argument, it is clear that with the last assumption the matrix $\mathscr{A}$ has exactly n positive and m negative eigenvalues. Hence, $\mathscr{A}$ is indefinite as we pointed out before. Further results ensure that even if we just assume that A is symmetric positive semidefinite and the condition $\operatorname{Ker} A \cap \operatorname{Ker} B = \{0\}$ is given, then $\mathscr{A}$ is congruent with a block diagonal matrix where the S block has $m - r$ negative eigenvalues, with $r = \operatorname{rank}(S)$.

Now, we focus in the A symmetric positive definite case in order to get eigenvalue bounds. The reason why we are interested in these bounds lies in obtaining estimates for the condition number of $\mathscr{A}$, that provides us light about convergence of iterative methods on each case. Furthermore, these bounds will be useful for checking the inf-sup condition and thus the stability of mixed finite element discretisations. Then we open up the following theorem:

Theorem 5.4. *Assume A is symmetric positive definite, $B_1 = B_2 = B$ with full rank, and $C = 0$. Let μ_1 and μ_n denote the largest and smallest eigenvalues of A, and let σ_1 and σ_m denote the largest and smallest singular values of B. Let $\sigma(\mathscr{A})$ denote the spectrum of $\mathscr{A}$. Then, $\sigma(\mathscr{A}) \subset I^- \cup I^+$ where*

$$I^- = \left[\frac{1}{2}\left(\mu_n - \sqrt{\mu_n^2 + 4\sigma_1^2} \right), \frac{1}{2}\left(\mu_1 - \sqrt{\mu_1^2 + 4\sigma_m^2} \right) \right],$$
$$I^+ = \left[\mu_n, \frac{1}{2}\left(\mu_1 + \sqrt{\mu_1^2 + 4\sigma_1^2} \right) \right].$$

Let us note that clearly $I^- \subset (-\infty, 0)$ and $I^+ \subset (0, +\infty)$. Although our matrix $\mathscr{A}$ is indefinite, one is usually interested in a symmetric positive (semi-)definite matrix $\tilde{\mathscr{A}}$ in order to define an inner product on $\mathbb{R}^{n+m}$ which suits better to iterative methods like Krylov subspaces methods, more if we want to use certain preconditioners. For this reason, we can rewrite our saddle point problem as follows:

$$\begin{pmatrix} A & B^T \\ -B & C \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} f \\ -g \end{pmatrix} \qquad \text{or} \qquad \tilde{\mathscr{A}} u = \tilde{b}.$$

Note that $\tilde{\mathscr{A}} = \mathscr{J}\mathscr{A} = \begin{pmatrix} I_n & 0 \\ 0 & -I_m \end{pmatrix} \mathscr{A}$ is nonsymmetric positive (semi-)definite, but here one can define a $\mathscr{J}$-symmetric matrix as a matrix $\mathscr{M}$ that satisfies $\mathscr{J}\mathscr{M} = \mathscr{M}^T \mathscr{J}$. Then, we note that $\tilde{\mathscr{A}}$ holds this property and it is symmetric respect to the indefinite inner product defined on $\mathbb{R}^{n+m}$ given by $\langle u, v \rangle_{\mathscr{J}} = v^T \mathscr{J} u$. Saving some algebraic issues that we will not treat along this work, this nonstandard inner product with respect to which $\tilde{\mathscr{A}}$ is symmetric positive (semi-)definite lead us to a favourable situation for applying some iterative methods. Thus, we show the following result for the alternative formulation $\tilde{\mathscr{A}} u = \tilde{b}$:

Theorem 5.5. *Assume that A is symmetric positive definite, $B_1 = B_2 = B$ has full rank, and $C = 0$. Let μ_n denote the smallest eigenvalue of A. If $\mu_n \geq 4\|S\|_2$, then, all the eigenvalues of $\tilde{\mathscr{A}}$ are real and positive.*

Proof. See [13]. □

Another important topic about saddle point matrices is the condition number of $\mathscr{A}$. The system matrix of these problems uses to be poorly conditioned. In particular, for mixed finite element formulations of elliptic partial differential equations μ_n and σ_m tends to zero as h goes to zero. Hence, the condition number of $\mathscr{A}$ is given by

$$\mathscr{K}(\mathscr{A}) = \frac{\max |\lambda(\mathscr{A})|}{\min |\lambda(\mathscr{A})|}, \quad \text{where } \lambda(\mathscr{A}) \text{ denotes the set of eigenvalues of } \mathscr{A},$$

and it grows like $\mathscr{O}(h^{-p})$, with $p > 0$. This means that a refinement of our mesh will imply a deteriorated convergence rate of iterative methods for solving the system. However, a possible way to save this issue is to use preconditioners.

Finally, we finish this chapter adding another strategy for solving a saddle point system: A Schur complement reduction. This way only requires to assume that A and $\mathscr{A}$ are nonsingular, but on the other hand, only transforms our original system to a block upper triangular system. With this last assumption, we recall S the nonsingular Schur complement of the matrix A. Then, let us multiply both sides of the first equation of the saddle point system by B_2A^{-1}:

$$\begin{cases} Ax + B_1^T y &= f, \\ B_2 x - Cy &= g. \end{cases} \quad \Longrightarrow \quad \begin{cases} B_2 x + B_2 A^{-1} B_1^T y &= B_2 A^{-1} f, \\ B_2 x - Cy &= g. \end{cases}$$

Now, we substract the first equation to the second one and using $S = -(C + B_2 A^{-1} B_1^T)$ we obtain the following equation:

$$Sy = g - B_2 A^{-1} f.$$

At this point, we build a new system joining the first equation of the original system with this one:

$$\begin{cases} Ax + B_1^T y &= f, \\ Sy &= g - B_2 A^{-1} f. \end{cases} \tag{5.2}$$

In this system we are able to obtain x and y separately. We proceed solving first the system for y and then backsubstituting its solution on the system for x, so we solve two systems of m and n equations respectively. Nevertheless, those two systems might keep being costly to solve by iterative methods since we are not assuming positive definiteness of A and $-S$.

Chapter 6

The Stokes Equations

Introduction

The Stokes equations are a system of linear partial differential equations which describe incompressible fluid flows. This system represents a limiting case of the more general Navier-Stokes equations, but it is useful only when the flow is very slow. Thus, convection effects considered in Navier-Stokes equations can be neglected and one can obtain the stationary Stokes equations. On the other hand, a fluid is said to be incompressible if its density is constant along trajectories of a fluid element with a fixed temperature but changing pressure. The Stokes equations have many applications, for instance they can modtael how honey drops, flow of blood along veins and arteries...

In this chapter, we will see that the Stokes equations can be formulated as a saddle point problem. Another point is to use the results appearing in previous chapters in order to study existence and uniqueness of solution for this problem. Finally, the last aim of this chapter is to solve these equations through finite element approximations. However, the choice of discrete spaces can not be arbitrary: The inf-sup condition takes again an important role and our finite element methods will have to satisfy it.

The first step is to state the Stokes equations. Firstly we consider Dirichlet boundary conditions, but we will see that this problem also admits Neumann boundary conditions. Thus, let $\Omega \subset \mathbb{R}^n$ be a bounded, connected and open domain, with $n \geq 2$ and $\partial\Omega$-Lipschitz boundary. Then, the strong formulation of the Stokes equations is written as follows:

$$\begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{f}, & \text{in } \Omega, \\ \nabla \cdot \mathbf{u} = 0, & \text{in } \Omega, \\ \mathbf{u} = \mathbf{g}, & \text{on } \partial\Omega. \end{cases} \tag{6.1}$$

In the system above, variable $\mathbf{u}$ is a vector function that represents the velocity of the fluid meanwhile p is a scalar function that represents the pressure. The first equation models the conservation of the momentum of the fluid and, hence, it is known as the momentum equation. On the other hand, the second equation is the incompressibility constraint and enforces the conservation of mass. It is also known as the continuity equation. Given a force $\mathbf{f} \in \mathscr{C}(\Omega)$ and a boundary data $\mathbf{g} \in \mathscr{C}(\partial\Omega)$, we will say that a pair $(\mathbf{u}, p)$ is a classical solution of the Stokes problem if it fulfills (6.1) and $(\mathbf{u}, p) \in \left(\mathscr{C}^2(\Omega) \cap \mathscr{C}(\bar{\Omega})\right) \times \mathscr{C}^1(\Omega)$.

After this introduction to the Stokes equations with Dirichlet boundary conditions, it worths to remark that one has to be careful in the choice of the boundary conditions. Due to well-poseness of the problem, one condition needs to be satisfied:

Lemma 6.1. *The following compatibility constraint is a necessary condition for the existence of solutions for the Stokes equations:*

$$\int_{\partial\Omega} \boldsymbol{g}\cdot\boldsymbol{n}\, ds = 0.$$

That is, the volume of fluid entering the domain must be matched by the volume of fluid flowing out of the domain.

Proof. Let $\mathbf{u} = \mathbf{g}$ on $\partial\Omega$. First, integrating the incompressibility condition in the Stokes problem (6.1) one gets

$$\int_{\Omega} \nabla\cdot\mathbf{u}\, dx = 0.$$

And secondly using the Divergence theorem we reach the desired result:

$$0 = \int_{\Omega} \nabla\cdot\mathbf{u}\, dx = \int_{\partial\Omega} \mathbf{u}\cdot\mathbf{n}\, ds = \int_{\partial\Omega} \mathbf{g}\cdot\mathbf{n}\, ds.$$

□

This compatibility condition on the boundary data is very important: If it does not hold, then, there is no solution for the Stokes equations. For the Dirichlet problem given in (6.1), pressure p is unique up to a constant. This is due to the appearance of ∇p instead of p in the system of equations: Note that given a constant $c \in \mathbb{R}$, one has $\nabla(p+c) = \nabla p$. On the other hand, Neumann boundary conditions can be given by making a boundary partition $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$ and stating

$$\begin{aligned} &\text{Dirichlet boundary condition:} && \mathbf{u} = \mathbf{g} \quad \text{on } \partial\Omega_D. \\ &\text{Neumann boundary condition:} && \frac{\partial\mathbf{u}}{\partial\mathbf{n}} - \mathbf{n}\cdot p = \mathbf{s} \quad \text{on } \partial\Omega_N. \end{aligned}$$

where $\mathbf{s} \in \mathscr{C}(\partial\Omega_N)$, $\mathbf{n}$ is the outward unit normal to the boundary and $\dfrac{\partial\mathbf{u}}{\partial\mathbf{n}}$ denotes the directional derivative in the normal direction.

In this case, we need that $\partial\Omega_D \neq \emptyset$ in order to ensure uniqueness of the velocity solution $\mathbf{u}$. Fortunately, the condition appearing in the lemma given above holds when Neumann boundary conditions are taken into account too. That is caused by a self-acting adjustment of the product $\mathbf{u}\cdot\mathbf{n}$ in order to satisfy the incompressibility constraint.

Moreover, adding Neumann boundary conditions we can ensure the uniqueness of solution p to (6.1): Although p is unique up to a constant, known as hydrostatic pressure level, a Neumann outflow condition will fix it. One is able to check this statement with a simple example: Let us consider a two-dimensional domain $\Omega = [0,L]\times[0,H]$ with a Neumann outflow condition given in $\partial\Omega_N = \{(L,y)|0 < y < H\}$, and a homogeneous Dirichlet boundary condition in $\partial\Omega_D = \partial\Omega \setminus \partial\Omega_N$. Then, the natural outflow condition is

$$\begin{cases} \dfrac{\partial u_x}{\partial x} - p = s_x, \\ \dfrac{\partial u_y}{\partial x} = s_y. \end{cases}$$

Now, let us integrate the normal component over $\partial\Omega_N$. Thus, one gets,

$$\int_0^H \frac{\partial u_x}{\partial x} dy - \int_0^H p dy = \int_0^H s_x dy.$$

Then, using the incompressibility condition $\dfrac{\partial u_x}{\partial x} = -\dfrac{\partial u_y}{\partial y}$ and the Dirichlet boundary condition:

$$\int_0^H p dy = -\int_0^H \frac{\partial u_y}{\partial y} dy - \int_0^H s_x dy = -u_y(H) + u_y(0) - \int_0^H s_x dy = -\int_0^H s_x dy.$$

Finally, the average pressure value at $\partial\Omega_N$ gets fixed. One can even set it to zero by taking $\mathbf{s} = \mathbf{0}$. However, for sake of symplicity, we will study the Dirichlet problem (6.1).

Weak Formulation of the Stokes Equations

We are interested in approximating the solution of the Stokes problem given in (6.1) using mixed finite element methods. In order to do that, we need first to build the variational formulation of the Stokes problem and then apply the results given along this work. As one usually does, we start defining our test spaces:

$$\mathbf{V} = \mathbf{H}_0^1(\Omega) := \left\{ \mathbf{v} \in \mathbf{H}^1(\Omega) \ : \ \mathbf{v}_{|_{\partial\Omega}} = 0 \right\},$$
$$Q = L_0^2(\Omega) := \left\{ q \in L^2(\Omega) \ : \ \int_\Omega q \, dx = 0 \right\}.$$

The next step in order to obtain the weak formulation is to multiply the Stokes equations by test functions. Thus, we test the first equation by a function $\mathbf{v} \in \mathbf{V}$ and the second equation by a function $q \in Q$:

$$\begin{cases} -\boldsymbol{\Delta}\mathbf{u} \, : \, \mathbf{v} + \nabla p \cdot \mathbf{v} = \mathbf{f} \cdot \mathbf{v}, & \forall \mathbf{v} \in \mathbf{V}, \\ (\nabla \cdot \mathbf{u}) \, q = 0, & \forall q \in Q. \end{cases}$$

Finally, we integrate both equations in Ω and then apply integration by parts using again the Divergence theorem. Hence one gets the weak formulation of the Stokes equations: Given $\mathbf{f} \in \mathbf{H}^{-1}(\Omega) = \mathbf{V}'$, find $(\mathbf{u}, p) \in \mathbf{V} \times Q$ such that

$$\begin{cases} \displaystyle\int_\Omega \nabla\mathbf{u} \, : \, \nabla\mathbf{v} \, dx - \int_\Omega (\nabla \cdot \mathbf{v}) \, p \, dx = \int_\Omega \mathbf{f} \cdot \mathbf{v} \, dx, & \forall \mathbf{v} \in \mathbf{V}, \\ \displaystyle\int_\Omega (\nabla \cdot \mathbf{u}) \cdot q = 0, & \forall q \in Q. \end{cases} \tag{6.2}$$

A pair $(\mathbf{u}, p)$ is said to be weak solution of the problem (6.1) if it fulfills (6.2). Moreover, a weak solution is also a classical solution if it fulfills the smoothness requirements given in the strong formulation.

The variational problem (6.2) can be written using the same notation given in chapter 2. Thus, let us introduce the following bilinear and linear forms:

$$\begin{aligned} &a(\cdot,\cdot) : \mathbf{V} \times \mathbf{V} \longrightarrow \mathbb{R} \\ &\quad a(\mathbf{u}, \mathbf{v}) := \int_\Omega \nabla\mathbf{u} \, : \, \nabla\mathbf{v} \, dx, \qquad \forall \mathbf{u}, \mathbf{v} \in \mathbf{V}, \\ &b(\cdot,\cdot) : \mathbf{V} \times Q \longrightarrow \mathbb{R} \\ &\quad b(\mathbf{v}, p) := -\int_\Omega (\nabla \cdot \mathbf{v}) \, p \, dx, \qquad \forall \, \mathbf{v} \in \mathbf{V}, \, \forall p \in Q, \\ &f : \mathbf{V} \longrightarrow \mathbb{R} \\ &\quad f(\mathbf{v}) := \int_\Omega \mathbf{f} \cdot \mathbf{v} \, dx, \qquad \text{with } \mathbf{f} \in \mathbf{V}', \, \forall \mathbf{v} \in \mathbf{V}. \end{aligned}$$

Since our test spaces $\mathbf{V}$ and Q are Hilbert spaces which own a defined scalar product, we are interested in proving that the operators given above satisfy some linearity and continuity conditions. Thus, we will be able to state that the weak formulation of the Stokes problem and the model variational problem given in (2.1) with $\mathbf{g} = \mathbf{0}$ are equivalent. The following result lets us to ensure that:

Lemma 6.2. *It holds*

i) $a(\cdot,\cdot)$ *is symmetric, bounded, positive definite and coercive bilinear form.*

ii) $b(\cdot,\cdot)$ *is a bounded bilinear form.*

iii) $f(\cdot)$ *is a bounded linear functional.*

Proof. We proof *i*) here, the interested reader can find the proof of *ii*) and *iii*) in [9], pp. 23. First, the operator $a(\cdot,\cdot)$ is symmetric since the tensor product is symmetric too:

$$\begin{aligned} a(\mathbf{u},\mathbf{v}) &= \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, dx = \int_\Omega \sum_{i,j=1}^{n} \frac{\partial \mathbf{u}_i}{\partial x_j}\frac{\partial \mathbf{v}_i}{\partial x_j}\, dx = \int_\Omega \sum_{i,j=1}^{n} \frac{\partial \mathbf{v}_i}{\partial x_j}\frac{\partial \mathbf{u}_i}{\partial x_j}\, dx = \\ &= \int_\Omega \nabla\mathbf{v} : \nabla\mathbf{u}\, dx = a(\mathbf{v},\mathbf{u}), \quad \forall \mathbf{u},\mathbf{v} \in \mathbf{V}. \end{aligned}$$

Secondly, we prove the boundedness of $a(\cdot,\cdot)$ using the Cauchy-Schwartz inequality.

$$\begin{aligned} |a(\mathbf{u},\mathbf{v})| &= \left| \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{v}\, dx \right| \leq \int_\Omega |\nabla\mathbf{u} : \nabla\mathbf{v}|\, dx = \|\nabla\mathbf{u} : \nabla\mathbf{v}\|_{\mathbf{L}^1(\Omega)} \leq \|\nabla\mathbf{u}\|_{\mathbf{L}^2(\Omega)} \cdot \|\nabla\mathbf{v}\|_{\mathbf{L}^2(\Omega)} = \\ &= |\mathbf{u}|_{\mathbf{V}}|\mathbf{v}|_{\mathbf{V}} \leq \|\mathbf{u}\|_{\mathbf{V}}\|\mathbf{v}\|_{\mathbf{V}}, \quad \forall \mathbf{u},\mathbf{v} \in \mathbf{V}. \end{aligned}$$

Note that a linear operator between normed Hilbert spaces is continuous if and only if it is bounded. Therefore the operator $a(\cdot,\cdot)$ is continuous. Moreover, one realises that the seminorm defined in $\mathbf{V}$ and its norm are equivalent by using the Poincaré inequality:

$$\|\mathbf{u}\|_{L^2(\Omega)} \leq C_\Omega \|\nabla\mathbf{u}\|_{L^2(\Omega)} \quad \text{with } C_\Omega > 0,\ \forall \mathbf{u} \in \mathbf{V}.$$

Then, $a(\cdot,\cdot)$ is clearly positive definite:

$$a(\mathbf{u},\mathbf{u}) = \int_\Omega \nabla\mathbf{u} : \nabla\mathbf{u}\, dx = \|\nabla\mathbf{u}\|^2_{\mathbf{L}^2(\Omega)} = |\mathbf{u}|^2_{\mathbf{V}} \geq \frac{1}{(1+C_\Omega)^2}\|\mathbf{u}\|^2_{\mathbf{V}} > 0, \quad \forall \mathbf{u} \in \mathbf{V}\setminus\{0\}.$$

And consequently $a(\cdot,\cdot)$ is coercive too. Finally, bilinearity of $a(\cdot,\cdot)$ is due to its already proved symmetry and the fact that to integrate and to derive are linear operators. □

With this lemma, the weak formulation for the Stokes equations takes the form given in (2.1):

$$\left\{ \begin{array}{l} \text{Given } \mathbf{f} \in \mathbf{V}',\ \text{find } (\mathbf{u},p) \in \mathbf{V}\times Q \text{ such that:} \\ \begin{array}{lll} a(\mathbf{u},\mathbf{v}) & + & b(\mathbf{v},p) = \langle \mathbf{f},\mathbf{v}\rangle_{\mathbf{V}'\times\mathbf{V}}, \quad \forall \mathbf{v} \in \mathbf{V}. \\ & & b(\mathbf{u},q) = 0 \qquad\qquad\quad \forall q \in Q \end{array} \end{array} \right. \tag{6.3}$$

Moreover, we can define the associated linear operators corresponding to $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ such that

$$\begin{aligned} &A = -\boldsymbol{\Delta} : \mathbf{V} \longrightarrow \mathbf{V}' \\ &\quad \langle A\mathbf{u},\mathbf{v}\rangle_{\mathbf{V}'\times\mathbf{V}} := a(\mathbf{u},\mathbf{v}), \qquad \forall \mathbf{u},\mathbf{v} \in \mathbf{V}, \\ &B = \text{ div } : \mathbf{V} \longrightarrow Q' \\ &\quad \langle B\mathbf{v},q\rangle_{Q'\times Q} := b(\mathbf{v},q), \qquad \forall\, \mathbf{v} \in \mathbf{V},\ \forall q \in Q, \\ &B^t = \text{ grad } : Q \longrightarrow \mathbf{V}' \\ &\quad \langle B^t q,\mathbf{v}\rangle_{\mathbf{V}'\times\mathbf{V}} := \langle B^t q,\mathbf{v}\rangle_{\mathbf{V}'\times\mathbf{V}} = \langle B\mathbf{v},q\rangle_{Q'\times Q} = b(\mathbf{v},q), \qquad \forall\, \mathbf{v} \in \mathbf{V},\ \forall q \in Q. \end{aligned}$$

Then, we obtain an operator formulation for the Stokes problem that is equivalent to a saddle point problem on the Hilbert spaces $\mathbf{V}$ and Q: Given $\mathbf{f} \in \mathbf{V}'$, find $(\mathbf{u},p) \in \mathbf{V}\times Q$ such that

$$\left\{ \begin{array}{ll} A\mathbf{u} + B^t p & = \mathbf{f}, \\ B\mathbf{u} & = \mathbf{0}. \end{array} \right. \tag{6.4}$$

At this point, let us call $\mathbf{K} := \text{Ker } B \subset \mathbf{V}$ and define $A_{KK'} := \Pi_{K'} A E_K$. Hence, one can consider theorem 2.2 for talking about existence and uniqueness of solutions. As we saw in chapter 3, the conditions required in that theorem can be replaced by the following inf-sup conditions

$$\begin{aligned} \exists \beta > 0 : \inf_{q \in Q} \sup_{\mathbf{v} \in V} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_V \|q\|_Q} \geq \beta, \\ \exists \alpha_1 > 0 : \inf_{\mathbf{v}_0 \in K} \sup_{\mathbf{w}_0 \in K} \frac{a(\mathbf{v}_0,\mathbf{w}_0)}{\|\mathbf{v}_0\|_V \|\mathbf{w}_0\|_V} = \inf_{\mathbf{w}_0 \in K} \sup_{\mathbf{v}_0 \in K} \frac{a(\mathbf{v}_0,\mathbf{w}_0)}{\|\mathbf{v}_0\|_V \|\mathbf{w}_0\|_V} \geq \alpha_1. \end{aligned} \tag{6.5}$$

Since $a(\cdot,\cdot)$ is symmetric and $a(\mathbf{v},\mathbf{v}) > 0$, it is possible to prove that the converse statement of Lemma 3.2 is also true. Then, the inf-sup condition for the bilinear form $a(\cdot,\cdot)$ can be replaced by the elker condition. Note that we checked this condition in the proof of Lemma 6.2. Hence, we are almost ready to show a result about existence and uniqueness of solutions for the variational formulation of the Stokes equations. In order to prove it, we will need the following lemma:

Lemma 6.3. *Let $q \in Q$. Then*

$$\exists | \boldsymbol{v} \in \text{Ker } B : \nabla \cdot \boldsymbol{v} = q \text{ and } \|\boldsymbol{v}\|_V \leq C \|q\|_Q,$$

for a constant $C > 0$.

Proof. See [14], pp. 40. □

Finally, we finish this section with the mentioned result about existence and uniqueness.

Theorem 6.1. *Let Ω be a bounded domain in $\mathbb{R}^n$ with $\partial\Omega$ lipschitz boundary. Then, for every $\boldsymbol{f} \in \boldsymbol{H}^{-1}$, the weak Stokes problem (6.3) has a unique solution $(\boldsymbol{u}, p) \in \boldsymbol{V} \times Q$.*

Proof. As we have said before, the bilinear form $a(\cdot,\cdot)$ holds the ellipticity condition on Ker B and then we need only to prove that $b(\cdot,\cdot)$ satisfies the inf-sup condition given in (6.6). Using the result provided in lemma 6.3 one gets

$$\sup_{\mathbf{v} \in \mathbf{V}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{\mathbf{V}} \|q\|_Q} = \sup_{\mathbf{v} \in \mathbf{H}_0^1(\Omega) \setminus \mathbf{0}} \frac{\langle \nabla \cdot \mathbf{v}, q \rangle_{L_0^2(\Omega)}}{\|\mathbf{v}\|_{\mathbf{V}} \|q\|_Q} \geq \frac{\langle \nabla \cdot \mathbf{v}, q \rangle_{L_0^2(\Omega)}}{\|\mathbf{v}\|_{\mathbf{V}} \|q\|_Q} = \frac{\langle q, q \rangle_{L_0^2(\Omega)}}{\|\mathbf{v}\|_{\mathbf{V}} \|q\|_Q} = \frac{\|q\|^2_{L_0^2(\Omega)}}{\|\mathbf{v}\|_{\mathbf{V}} \|q\|_Q} \geq \frac{1}{C} \|q\|_{L_0^2(\Omega)}$$

by taking $\mathbf{v}$ such that $\nabla \cdot \mathbf{v} = q$ and $\|\mathbf{v}\|_{\mathbf{V}} \leq C \|q\|_Q$. Finally, we can choose $q \in L_0^2(\Omega)$ arbitrarily and then we obtain

$$\inf_{q \in L_0^2(\Omega) \setminus 0} \sup_{\mathbf{v} \in \mathbf{H}_0^1(\Omega) \setminus \mathbf{0}} \frac{b(\mathbf{v},q)}{\|\mathbf{v}\|_{\mathbf{H}_0^1(\Omega) \setminus \mathbf{0}} \|q\|_{L_0^2(\Omega) \setminus 0}} \geq \frac{1}{C} =: \beta.$$

□

Discrete Stokes Problem

A good-working discretisation is needed when we consider the variational formulation of the Stokes equations and our aim is to find an approximation of the problem (6.2). As far as we find two variables in the Stokes equations, a pair of finite element spaces will be required. Furthermore, when we use a different test space for each variable our numerical method receives the name of mixed finite element method.

In regard to the discrete Stokes problem, finite dimensional subspaces $\mathbf{V}_h \subset \mathbf{H}_0^1(\Omega)$ and $Q_h \subset L_0^2(\Omega)$ have to be chosen carefully. As we pointed out in chapter 4, firstly we have to take care about whether $\mathbf{K}_h :=$ Ker B_h and $H_h :=$ Ker B_h^t are actually subspaces of the original kernels $\mathbf{K}$ and H. Secondly, the finite element spaces have to verify the corresponding discrete inf-sup conditions, that are not implied by

the fulfillment of the inf-sup conditions on the general spaces. These remarks will make the difference between the choice of stable or unstable finite element spaces. Thus, we start this section stating the discrete Stokes problem: Let $\mathbf{V}_h \subset \mathbf{H}_0^1(\Omega)$ and $Q_h \subset L_0^2(\Omega)$ be finite dimensional subspaces,

$$\left\{ \begin{array}{l} \text{Given } \mathbf{f} \in \mathbf{V}', \text{ find } (\mathbf{u}_h, p_h) \in \mathbf{V}_h \times Q_h \text{ such that:} \\ a(\mathbf{u}_h, \mathbf{v}_h) \quad + \quad b(\mathbf{v}_h, p_h) = \langle \mathbf{f}, \mathbf{v}_h \rangle_{\mathbf{V}' \times \mathbf{V}}, \quad \forall \mathbf{v}_h \in \mathbf{V}_h. \\ \qquad\qquad\quad b(\mathbf{u}_h, q_h) = 0 \qquad\qquad\quad \forall q_h \in Q_h. \end{array} \right. \tag{6.6}$$

Since $a(\cdot,\cdot)$ is coercive on $\mathbf{V}$, we don't have to check whether the inf-sup condition for $a(\cdot,\cdot)$ holds when we consider homogeneous Dirichlet boundary conditions. However, there might be troubles with the uniqueness of p_h and we will be interested in verifying the following discrete inf-sup condition for $b(\cdot,\cdot)$

$$\exists \beta_h > 0 \text{ such that : } \inf_{q_h \in Q_h} \sup_{\mathbf{v}_h \in \mathbf{V}_h} \frac{b(\mathbf{v}_h, q_h)}{\|\mathbf{v}_h\|_{\mathbf{V}} \|q_h\|_Q} \geq \beta_h.$$

If this condition is verified, we are able to use corollary (4.1) since $a(\cdot,\cdot)$ is symmetric, definite positive in $\mathbf{V}_h$ and coercive in $\mathbf{V}$. Then, we ensure that there exists an unique solution $(\mathbf{u}_h, p_h)$ for (6.6) and we obtain the following improved error estimates

$$\|\mathbf{u}_h - \mathbf{u}\|_{\mathbf{V}} \leq \left(\frac{2\|a\|}{\alpha} + \frac{2\|a\|^{1/2}\|b\|}{(\alpha)^{1/2}\beta_h} \right) E_u + \frac{\|b\|}{\alpha} E_p, \tag{6.7}$$

$$\|p_h - p\|_Q \leq \left(\frac{2\|a\|^{3/2}}{(\alpha)^{1/2}\beta_h} + \frac{\|a\|\|b\|}{\beta_h^2} \right) E_u + \frac{3\|a\|^{1/2}\|b\|}{(\alpha)^{1/2}\beta_h} E_p, \tag{6.8}$$

with $\alpha > 0$ such that the coercivity condition $a(\mathbf{u},\mathbf{u}) \geq \alpha \|\mathbf{u}\|_V$, $\forall \mathbf{u} \in \mathbf{V}$ holds. Moreover, when the discrete inf-sup condition is satisfied with a constant $\beta_0 > 0$ which does not depend on the mesh size h, the same assumptions lead us to the following estimate

$$\|\mathbf{u}_h - \mathbf{u}\|_{\mathbf{V}} + \|p_h - p\|_Q \leq C\,(E_u + E_p)\,. \tag{6.9}$$

Mixed Finite Element Methods for the Stokes Equations

In this section we verify the importance of the inf-sup conditions for building an stable pair of finite element spaces. Some combinations violate the discrete inf-sup condition and consequently they do not yield a good approximation at all. In order to illustrate this, we will show some unstable pairs of finite element spaces explaining the reason why they do not work. Before that, let us introduce some common notation for a given decomposition of our domain Ω:

Definition. A triangulation $\mathscr{T}_h$ is a decomposition of a domain $\overline{\Omega} \subset \mathbb{R}^n$ into polyhedrons $T \in \mathscr{T}_h$. The union of all these polyhedrons is called grid or mesh and the grid size is defined as follows

$$h := \max_{T \in \mathscr{T}_h} \text{diam}\,(T).$$

We will consider that the polyhedrons for our triangulation are triangles. At using a finite element method, variables are approximated elementwise by polynomial functions. However, the general finite element theory is developed for a reference triangle.

Definition. The reference triangle is the convex hull of the points $\hat{x}_0 = \mathbf{0}$, $\hat{x}_i = e_i$, $\forall i = 1, \ldots n$, where $e_i, i = 1, \ldots n$ are the cartesian unit vectors in $\mathbb{R}^n$. In addition, we will denote the reference triangle with $\hat{T}$.

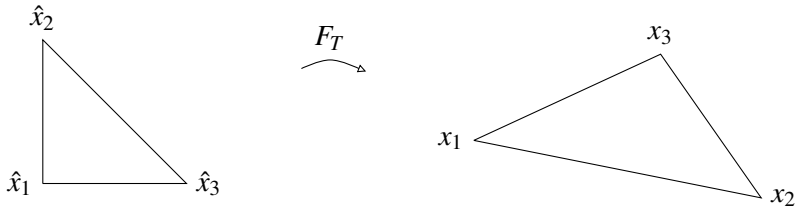

Figure 6.1: There exists an affine map between the reference triangle and each mesh cell.

Although the cells might have different size and shape in every triangulation, it is possible to find an affine map between the reference triangle and every cell. Let $F_T : \hat{T} \longrightarrow T$ be this map such that $F(\hat{x}_i) = x_i,\ i = 0,\dots,n$, where $x_i, i = 0,\dots,n$ are the vertices of a mesh triangle.

Another point of this grid introduction is to define the polynomial spaces on a cell mesh T as follows

$$\mathbf{P}_k(T) := \left\{ s : T \longrightarrow \mathbb{R} \;\middle|\; s(\mathbf{x}) = \sum_{|\alpha| \leq k} c_\alpha \mathbf{x}^\alpha,\ c_\alpha \in \mathbb{R} \right\}.$$

Now, let us realise that since $\mathbf{V}_h$ and Q_h are finite dimensional spaces, we can find a finite basis for each one. Let $\{\varphi_i\}_{i=1}^N$ be a basis of $\mathbf{V}_h$, $\{\psi_i\}_{i=1}^M$ a basis of Q_h. Then, every $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in Q_h$ can be expressed as

$$\mathbf{u}_h = \sum_{i=1}^{N} \alpha_i \varphi_i, \ \text{ with } \alpha_i \in \mathbb{R}, \quad i = 1,\dots,N; \qquad p_h = \sum_{i=1}^{M} \beta_i \psi_i, \ \text{ with } \beta_i \in \mathbb{R}, \quad i = 1,\dots,M.$$

In addition, note that it suffices to know the coefficients α_i and β_i in order to compute our approximations $\mathbf{u}_h$ and p_h. Also, both equations of the discrete Stokes problem are tested with functions $\mathbf{v}_h \in \mathbf{V}_h$ and $p_h \in Q_h$. Hence, using bilinearity of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ one notes that it is enough to solve the following system:

$$\begin{aligned} &\sum_{l=1}^{N} a(\varphi_l, \varphi_i)\alpha_l + \sum_{k=1}^{M} b(\varphi_i, \psi_k)\beta_k = f(\varphi_i), && \forall i = 1,\dots,N, \\ &\sum_{l=1}^{N} b(\varphi_l, \psi_j)\alpha_l = 0, && \forall j = 1,\dots,M, \end{aligned} \tag{6.10}$$

where $N = \dim \mathbf{V}_h$, $M = \dim Q_h$. Thus, we are able to obtain a linear system describing a saddle point problem by defining

$$\begin{aligned} &(A_h)_{i,j} := a(\varphi_j, \varphi_i), \ A \in \mathbb{R}^{N\times N}, \quad (\mathbf{u}_h)_j := \alpha_j, \quad \mathbf{u}_h \in \mathbb{R}^N, \\ &(B_h)_{i,j} := b(\varphi_j, \psi_i), \ B \in \mathbb{R}^{M\times N}, \quad (p_h)_j := \beta_j, \quad p_h \in \mathbb{R}^M, \\ &(f_h)_j := f(\varphi_j), \quad f \in \mathbb{R}^N. \end{aligned}$$

Thereby, one gets finally the following saddle point problem:

$$\begin{pmatrix} A_h & B_h^t \\ B_h & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u}_h \\ p_h \end{pmatrix} = \begin{pmatrix} \mathbf{f}_h \\ 0 \end{pmatrix}. \tag{6.11}$$

Note again that this system is uniquely solvable if and only if rank $(A_h) = N$ and rank $(B_h) = M$. However, A_h is positive definite due to $a(\cdot,\cdot)$ is positive definite too. Therefore, A_h has full rank and we will have to deal only with the condition: rank $(B_h) = M$. On the other hand, $M < N$ is a necessary condition because the opposite case means that the system is overconstrained and then we have linearly dependent rows. Let us introduce an important result that relates the full rank of B_h with the discrete inf-sup condition of $b(\cdot,\cdot)$:

Theorem 6.2. *Let $B_h \in \mathbb{R}^{M\times N}$, with $M \leq N$. Then we have,*

$$rank_{row}(B_h) = M \iff \inf_{q\in\mathbb{R}^M\setminus\{0\}} \sup_{\boldsymbol{v}\in\mathbb{R}^N\setminus\{0\}} \frac{q^T B_h \boldsymbol{v}}{\|\boldsymbol{v}\|_2\|q\|_2} \geq \beta > 0. \tag{6.12}$$

Proof. We prove here that $\text{rank}_{\text{row}}(B_h)$ is a necessary condition for holding the discrete inf-sup condition. The interested reader can find the converse ($\Rightarrow$) proven in [10], pp. 45. Hence, let us prove the $\Leftarrow$ direction by contradiction: Assume that there exists $\beta > 0$ such that the discrete inf-sup condition above holds. In addition, let us suppose $\text{rank}_{\text{row}}(B_h) < M$. Then,

$$\begin{aligned}
\text{rank}_{\text{row}}(B_h) < M \;&\Leftrightarrow\; \dim\left(\text{Ker}\left(B_h^T\right)\right) \geq 1 \;\Leftrightarrow\; \exists q \in \mathbb{R}^M\setminus\{0\}:\; q^T B = B^T q = 0 \;\Rightarrow \\
&\Rightarrow \exists q \in \mathbb{R}^M\setminus\{0\}:\; q^T B\mathbf{v} = 0,\ \forall \mathbf{v}\in\mathbb{R}^N \;\Rightarrow \exists q \in \mathbb{R}^M\setminus\{0\}:\; \sup_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{q^T B_h\mathbf{v}}{\|\mathbf{v}\|_2} = 0 \;\Rightarrow \\
&\Rightarrow \inf_{q\in\mathbb{R}^M\setminus\{0\}} \sup_{\mathbf{v}\in\mathbb{R}^N\setminus\{0\}} \frac{q^T B_h\mathbf{v}}{\|\mathbf{v}\|_2\|q\|_2} \leq 0,
\end{aligned}$$

and then, we obtain a contradiction. Thus, it has to be necessarily $\text{rank}_{\text{row}}(B_h) = M$. □

Our next aim is to show some pairs of finite element spaces for the Stokes equations. In order to explain why some of them do not work, we have to introduce the main troubles that might appear when one applies a mixed finite element method for these equations. Hence, we might face with:

- **Spurious Pressure Modes**: The choice of a too large finite dimensional space Q_h might cause the existence of a $\tilde{p}_h \in Q_h$ with $\tilde{p}_h \neq 0$ such that

  $$b(\mathbf{v}_h, \tilde{p}_h) = 0, \quad \forall \mathbf{v}_h \in \mathbf{V}_h.$$

  Then, the discrete inf-sup condition for $b(\cdot,\cdot)$ is violated. Taking such a $\tilde{p}_h$ one gets

  $$\sup_{\mathbf{v}\in\mathbf{V}_h\setminus 0} \frac{b(\mathbf{v}_h,\tilde{p}_h)}{\|\mathbf{v}_h\|_{\mathbf{V}_h}} = 0 \Longrightarrow \inf_{q_h\in Q_h\setminus 0} \sup_{\mathbf{v}\in\mathbf{V}_h\setminus 0} \frac{b(\mathbf{v}_h,q_h)}{\|\mathbf{v}_h\|_{\mathbf{V}_h}\|q_h\|_{Q_h}} \leq 0.$$

  Thus, the appearance of a spurious pressure mode $\tilde{p}_h \in Q_h$ makes the finite element pair $\mathbf{V}_h/Q_h$ inf-sup unstable: If we assume that the pair $(\mathbf{u}_h, p_h)$ solves the discrete Stokes equations, then $(\mathbf{u}_h, p_h + \tilde{p}_h)$ solves them too.

- **The Locking Phenomenon**: This phenomenon appears when the space $\mathbf{V}_h$ has not been chosen large enough. In that case, $\mathbf{V}_h$ might not contain nontrivial discretely divergence-free functions. In other words, when we have

  $$b(\mathbf{u}_h, q_h) = 0, \quad \forall q_h \in Q_h \iff \mathbf{u}_h = \mathbf{0}.$$

  Consequently, the discete velocity field approximation will be $\mathbf{u}_h = \mathbf{0}$.

At this point, we are ready to show some pairs of finite element spaces for the Stokes equations. When a pair of elements is inf-sup unstable, one of the two main troubles given above appears. On the other hand, when a stable pair is given, one can prove this fact by checking that the discrete inf-sup condition holds. There are some practical ways to do this, a common one is the so-called Fortin's trick. This one consists in building a B-compatible operator Π_h as we described in Proposition 4.1. However, this task might be sometimes pretty toilsome and we will reference the reader to detailed proofs if it is the case.

- **Linear-Constant Element: The $\mathbf{P}_1/P_0$ Approximation**: Velocity and pressure are respectively approximated by elementwise polynomials of degree one and constant funcions. Considering thus homogeneous Dirichlet boundary condition, the choice of our finite dimensional spaces is

$$\mathbf{V}_h := \left\{\mathbf{v}_h \in \mathscr{C}^0(\bar{\Omega}) : \mathbf{v}_{h|_T} \in \mathbf{P}_1(T), \forall T \in \mathscr{T}_h, \ \ \mathbf{v}_{h|_{\partial\Omega}} = 0\right\},$$
$$Q_h := \left\{q_h \in L^2(\Omega) : q_{h|_T} \in P_0(T), \forall T \in \mathscr{T}_h, \ \ \int_\Omega q_h \, dx = 0\right\}.$$

  Let us see an example for which this choice does not fulfill the inf-sup condition. Assume that Ω is a square domain subdivided into $2N^2$ triangles. In addition, let us consider Dirichlet boundary conditions. Thus, for this example the dimension of $\mathbf{V}_h$ is given by:

$$\dim(\mathbf{V}_h) = 2 \cdot \#\{\text{Inner nodes}\} = 2(N-1)^2.$$

  On the other hand since we are approximating the pressure by constant functions, one gets one degree of freedom per element. Note that the pressure mean value has to be balanced in Ω since $q_h \in L^2_0(\Omega)$. Due to this balance, the value in one mesh cell is determined by the rest of mesh cells values. Therefore,

$$\dim(Q_h) = \#\{\text{Elements}\} - 1 = 2N^2 - 1.$$

  Hence the locking phenomenon appears because $\dim Q_h > \dim \mathbf{V}_h$, for $N \geq 1$ and then $\mathbf{u}_h \in \mathbf{V}_h$ is overconstrained. Finally, the unique discrete divergence-free velocity field might be $\mathbf{u}_h = 0$ making the pair $\mathbf{P}_1/P_0$ unstable.

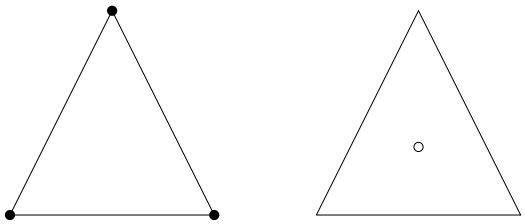

Figure 6.2: Local degrees of freedom for the velocity (left) and pressure (right) for the inf-sup unstable pair P_1/P_0.

- **Linear-Linear Element: The $\mathbf{P}_1/P_1$ Approximation**: The velocity field and the pressure are both approximated by elementwise linear functions. This finite element pair yield bad approximations since the appearance spurious pressure modes might take place. Our finite dimensional spaces are:

$$\mathbf{V}_h := \left\{\mathbf{v}_h \in \mathscr{C}^0(\bar{\Omega}) : \mathbf{v}_{h|_T} \in \mathbf{P}_1(T), \forall T \in \mathscr{T}_h, \ \ \mathbf{v}_{h|_{\partial\Omega}} = 0\right\},$$
$$Q_h := \left\{q_h \in \mathscr{C}^0(\Omega) : q_{h|_T} \in P_1(T), \forall T \in \mathscr{T}_h, \ \ \int_\Omega q_h \, dx = 0\right\}.$$

  However, this choice makes the $\mathbf{P}_1/P_1$ element is inf-sup unstable. For instance, let us consider our bidimensional domain $\Omega = (0,1)^2$ with a given triangulation $\mathscr{T}_h$. Note that every element is determined by the convex hull of its nodes $\{x_1, x_2, x_3\}$. At this point, let $\tilde{p}_h$ be a nonzero elementwise linear function such that

$$\sum_{i=1}^{3} \tilde{p}_h(x_{i,T}) = 0, \quad \forall T \in \mathscr{T}_h.$$

Since we are using a linear approximation for the velocity field, it holds that $(\nabla \cdot \mathbf{v}_h)_{|_T} = c_T$ constant on each triangle $T \in \mathscr{T}_h$. Thus, applying a Gaussian quadrature rule exact for linear functions one gets:

$$\int_\Omega \tilde{p}_h \nabla \cdot \mathbf{v}_h dx = \sum_{T\in\mathscr{T}_h} (\nabla \cdot \mathbf{v}_h)_{|_T} \int_T \tilde{p}_h dx = \sum_{T\in\mathscr{T}_h} (\nabla \cdot \mathbf{v}_h)_{|_T} \int_T \tilde{p}_h dx =$$
$$= \sum_{T\in\mathscr{T}_h} (\nabla \cdot \mathbf{v}_h)_{|_T} \frac{|T|}{3} \sum_{i=1}^{3} \tilde{p}_h(x_{i,T}) = 0, \quad \forall \mathbf{v}_h \in V_h.$$

Hence, this spurious pressure mode $\tilde{p}_h$ can yield any approximation of the pressure for the discrete Stokes problem. Note that calling $(\mathbf{u}_h, p_h)$ to its solution, then $(\mathbf{u}_h, p_h + c \cdot \tilde{p}_h)$ with $c \in \mathbb{R}$ solves the discrete problem too.

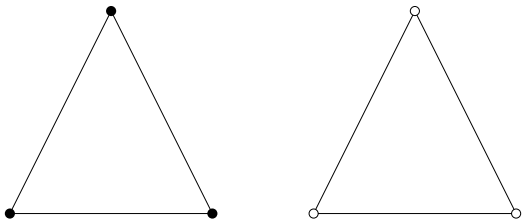

Figure 6.3: Local degrees of freedom for the velocity (left) and pressure (right) for the inf-sup unstable pair P_1/P_1.

- **The Minielement:** $(\mathbf{P}_1 + \mathbf{B}_3)/P_1$: As we have seen in the example above, $\mathbf{P}_1/P_1$ is an inf-sup unstable pair, but if we enrich enough the space $\mathbf{V}_h$ the spurious pressure modes might disappear. In order to do this, we introduce the so-called bubble functions: Given an element $T \in \mathscr{T}_h$, a function is said to be a bubble function in T if it vanishes on ∂T. In addition, one uses to find polynomial bubble functions that also belong to the infinite dimensional test space restricted to the element T, that is: $\mathbf{V} = \mathbf{H}_0^1(T)$. Furthermore, the bubble funcions for an element T are chosen such that their value on the barycenter is equal to 1. Thus, the following bubble space is used for the Minielement:

$$\mathbf{B}_3 := \{\mathbf{v}_h \in \mathscr{C}^0(\bar{\Omega}) : \mathbf{v}_{h|_T} \in \mathbf{P}_3(T) \cap \mathbf{H}_0^1(T), \ \forall T \in \mathscr{T}_h\}$$

Thereby, the finite dimensional spaces given for this pair are

$$\mathbf{V}_h := \left\{\mathbf{v}_h \in \mathscr{C}^0(\bar{\Omega}) : \mathbf{v}_{h|_T} \in (\mathbf{P}_1(T) \oplus \mathbf{B}_3(T)), \forall T \in \mathscr{T}_h, \ \mathbf{v}_{h|_{\partial\Omega}} = 0\right\},$$
$$Q_h := \left\{q_h \in \mathscr{C}^0(\Omega) : q_{h|_T} \in P_1(T), \forall T \in \mathscr{T}_h, \ \int_\Omega q_h \, dx = 0\right\}.$$

The Minielement yields an inf-sup stable pair of finite element spaces. It is possible to prove that the discrete inf-sup condition holds by building a proper B-compatible operator known as Clément operator. The detailed proof can be found in [4], pp. 470. Moreover, one finds out that the inf-sup constant does not depend on the mesh size h. It worths to say that the Minielement is considered the most economic element for the Stokes equations because the additional degrees of freedom do not increase significantly the computational cost.

Finally, assuming that the solution $(\mathbf{u}, p) \in \left(\left(H^2(\Omega) \cap H_0^1(\Omega)\right) \times \left(H^1(\Omega) \cap L_0^2(\Omega)\right)\right)$, one can obtain the following bound:

$$|\mathbf{u} - \mathbf{u}_h|_{\mathbf{H}^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq Ch \left(\|\mathbf{u}\|_{H^2(\Omega)} + \|p\|_{H^1(\Omega)}\right), \quad \text{with } C > 0. \tag{6.13}$$

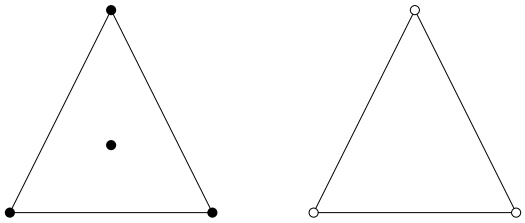

Figure 6.4: Local degrees of freedom for the velocity (left) and pressure (right) for the Minielement.

- **Taylor-Hood Element: $\mathbf{P}_k/P_{k-1}$, with $k \geq 1$:** The Taylor-Hood element also holds the discrete inf-sup condition and then it yields stable pairs of finite elements spaces. Again, the search of an uniformly stable Fortin operator becomes tedious and we reference the interested reader to [6], pp. 252. On the other hand, the approximation of velocity field and pressure is given by elementwise polynomials of degree k and $k-1$ respectively. Thus, we have

$$\mathbf{V}_h := \left\{ \mathbf{v}_h \in \mathscr{C}^0(\bar{\Omega}) : \mathbf{v}_{h|_T} \in \mathbf{P}_k(T), \forall T \in \mathscr{T}_h, \ \mathbf{v}_{h|_{\partial\Omega}} = 0 \right\},$$
$$Q_h := \left\{ q_h \in \mathscr{C}^0(\Omega) : q_{h|_T} \in P_{k-1}(T), \forall T \in \mathscr{T}_h, \ \int_\Omega q_h \, dx = 0 \right\}.$$

  This time, if we assume that the solution $(\mathbf{u}, p) \in \left(\left(H^{k+1}(\Omega) \cap H_0^1(\Omega) \right) \times \left(H^k(\Omega) \cap L_0^2(\Omega) \right) \right)$, then the following inequality holds:

$$|\mathbf{u} - \mathbf{u}_h|_{\mathbf{H}^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq Ch^k \left(\|\mathbf{u}\|_{H^{k+1}(\Omega)} + \|p\|_{H^k(\Omega)} \right), \quad \text{with } C > 0. \tag{6.14}$$

  In addition, the Taylor-Hood element has optimal convergence rate and the lowest order is given by $\mathbf{P}_2/P_1$. To end, we show the local degrees of freedom for this pair:

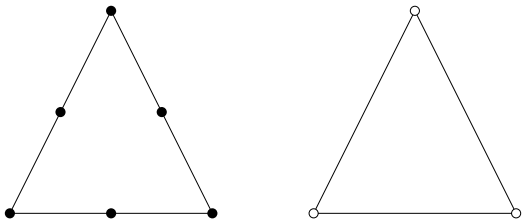

Figure 6.5: Local degrees of freedom for the velocity (left) and pressure (right) for the stable Taylor-Hood element $\mathbf{P}_2/P_1$.

Chapter 7

Numerical Results for the Stokes Equations

The previous chapter introduced us to the strong and weak formulation of the Stokes equations. Also, the discrete problem and some mixed finite element methods were included in order to open this chapter. Now, we are about to check that the theory and practice agree by approximating numerically the Stokes equations. Our aim is to show the results obtained using some finite element methods and compare them with the exact solution on each case. First, we consider a square domain $\Omega = (0,1) \times (0,1)$ with homogeneus Dirichlet boundary conditions on $\partial\Omega = \partial\Omega_D$. Secondly, we consider the following velocity field and pressure:

$$\mathbf{u} = \left(\sin(\pi x)\cos(\pi y) \ , \ -\cos(\pi x)\sin(\pi y) \right), \qquad p = \frac{1}{2} - x^2.$$

Thereby, the right-hand side is given by

$$\mathbf{f} = \left(2\pi^2 \sin(\pi x)\cos(\pi y) - 2x \ , \ -2\pi^2 \cos(\pi x)\sin(\pi y) \right).$$

Note that the force field $\mathbf{f}$ has been computed in order to make $(\mathbf{u}, p)$ a classical solution for the strong formulation of the Stokes equations. These solutions are depicted in Figure 7.1.

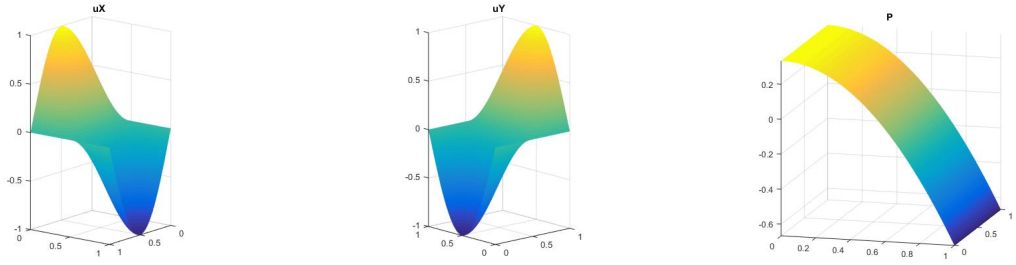

Figure 7.1: The exact solutions $(\mathbf{u}, p)$ for our example to the Stokes equations.

To discretise the problem we also consider a conforming triangulation. It worths to remark that a triangulation is said to be compatible or conforming if the intersection of any mesh cells $T, T' \in \mathscr{T}_h$ is either empty, a vertex or a whole edge. Figure 7.2 shows the triangulation that has been used in our numerical implementation.

As we saw, the approximation of this problem by a mixed finite element method yields a saddle point problem. Moreover, in this type of problems the discrete inf-sup conditions are required in order to obtain stability. If they are not fulfilled, some troubles may appear yielding a bad approximation. That was the case of the $\mathbf{P}_1/P_1$ pair, where the spurious pressure modes take place. Let us see the numerical

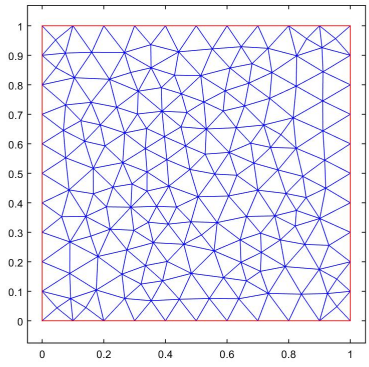

Figure 7.2: The triangulation $\mathscr{T}_h$ of domain Ω.

results obtained with Matlab using this mixed finite element method. The error computed in $H^1(\Omega)$ for each component of the velocity field $\mathbf{u}$ and the error for p in L^2 are showed in Table 7.1.

H^1-norm	u_X	u_Y	L^2-norm	p
$i=0$	6.427 e-02	7.195e-02	$i=0$	1.948e+00
$i=1$	2.084e-02	2.018e-02	$i=1$	1.517e+02
$i=2$	9.622e-03	9.357e-03	$i=2$	2.694e+02
$i=3$	4.516e-03	4.398e-03	$i=3$	9.111e+02
$i=4$	2.146e-03	2.093e-03	$i=4$	3.135e+03
$i=5$	1.024e-03	1.002e-03	$i=5$	6.379e+03

Table 7.1: H^1 errors for velocities and L^2 errors for pressure by using the inf-sup unstable $\mathbf{P}_1/P_1$ pair of finite element spaces.

It worths to remark that in these tables, i denotes the number of refinements applied to the triangulation $\mathscr{T}_h$. Thus, we see that the approximation of p obtained with $\mathbf{P}_1/P_1$ and the exact solution are totally different. As it was expected, the spurious pressure modes make this method unstable and the errors obtained are very large. In Figure 7.3, we show a plot of the pressure error in order to see this effect

On the other hand, the results obtained by using inf-sup stable pairs are pretty good. Applied to the same example, the Minielement yields satisfactory results and the convergence order agrees with the bound for the error given in (6.13). This can be see in Table 7.2.

H^1-norm	u_X	u_Y	L^2-norm	p
$i=0$	3.289e-01	3.264e-01	$i=0$	1.201e-01
$i=1$	1.668e-01	1.654e-01	$i=1$	5.545e-02
$i=2$	8.384e-02	8.314e-02	$i=2$	2.514e-02
$i=3$	4.198e-02	4.162e-02	$i=3$	1.009e-02
$i=4$	2.099e-02	2.082e-02	$i=4$	3.762e-03
$i=5$	1.050e-02	1.359e-02	$i=5$	1.359e-03

Table 7.2: H^1 errors for velocities and L^2 errors for pressure by using the Minielement method.

With every refinement of our mesh, h is divided by two approximately. Since the sum given by

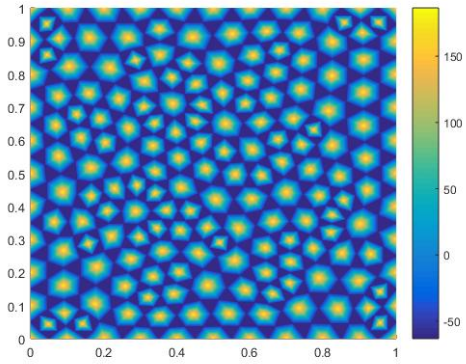

Figure 7.3: Pressure approximation by using the pair $\mathbf{P}_1/P_1$ with mesh refinement $i = 1$.

$|\mathbf{u} - \mathbf{u}_h|_{\mathbf{H}^1(\Omega)} + \|p - p_h\|_{L^2(\Omega)}$ is also divided by two, we deduce that the Minielement method has convergence rate $\mathscr{O}(h)$.

In regard to the Taylor-Hood finite element method, the bound given in (6.14) points out that its convergence rate is $\mathscr{O}(h^k)$. Hence the inf-sup stable pair $\mathbf{P}_2/P_1$ should have convergence rate $\mathscr{O}(h^2)$. In Table 7.3 we show the corresponding errors and we observe the expected results.

H^1-norm	u_X	u_Y	L^2-norm	p
$i=0$	1.768e-03	1.732e-03	$i=0$	1.908e-03
$i=1$	3.223e-04	3.153e-04	$i=1$	2.260e-04
$i=2$	5.726e-05	5.592e-05	$i=2$	3.662e-05
$i=3$	1.013e-05	9.892e-06	$i=3$	7.524e-06
$i=4$	1.791e-06	1.748e-06	$i=4$	1.766e-06
$i=5$	3.167e-07	3.091e-07	$i=5$	4.343e-07

Table 7.3: H^1 errors for velocities and L^2 errors for pressure by using the inf-sup stable $\mathbf{P}_2/P_1$ pair of finite element spaces.

In this case, with high refinements, the sum of the velocity field error measured in $H^1(\Omega)$ and the pressure error in $L^2(\Omega)$ is divided approximately by 4. Thus, we check that the expected convergence rate is given.

Bibliography

[1] A. LÓPEZ NIETO, *Métodos numéricos aplicados a la mecánica de fluidos*, Trabajo de Fin de Grado, Universidad de Zaragoza.

[2] A. N. KOLMOGOROV, S. V. FOMIN, *Introductory Real Analysis*, Dover Publications, New York.

[3] CIARLET PG, *The finite element method for elliptic problems*, Studies in mathematics and its applications, North-Holland, Amsterdam, 1978.

[4] D. BOFFI, F. BREZZI, M. FORTIN, *Mixed Finite Element Methods and Applications*, vol. 44 of Springer Series in Computational Mathematics, Springer, Heidelberg, 2013.

[5] ENDRE SÜLI, *A Brief Excursion into the Mathematical Theory of Mixed Finite Element Methods*, Mixed FEM lectures, 2014.

[6] F. AURICCHIO, F. BREZZI AND C. LOVADINA, *Mixed Finite Element Methods. Encyclopedia of Computational Mechanics. 1:9.*, Università di Pavia and IMATI-C.N.R, Pavia, Italy, 2004.

[7] F. BREZZI, M. FORTIN, *Mixed and Hybrid Finite Element Methods*, Springer, New York, 1991.

[8] GENE H. GOLUB, JÖRG LIESEN, MICHELE BENZI, *Numerical Solution of saddle point problems*, Cambridge University Press, 2005.

[9] HAIM BREZIS, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Springer, 2010.

[10] LAURA BLANK, *On Divergence-Free Finite Element Methods for the Stokes Equations*, Master Thesis, Berlin, 2014.

[11] LONG CHEN, *Finite Element Methods for Stokes Equations*.

[12] M. BENZI, G. H. GOLUB, *A preconditioner for generalized saddle point problems*, SIAM, 2004.

[13] M. BREZZI, V. SIMONCINI, *On the eigenvalues of a class of saddle point matrices*, Springer, 2005.

[14] V. JOHN, *Numerical methods for incompressible flow problems I. Lecture Notes*, Freie Universität Berlin, 2014.

Appendix A. The Minielement Method

scriptStokesP1bubbleP1.m

```
% Exact solution, rhs and Dirichlet boundary conditions
uExactX = @(x,y) (sin(pi*x)).*cos(pi*y);
uExactY = @(x,y) -(cos(pi*x)).*sin(pi*y);
pExact  = @(x,y) 1/2 - x.^2;
fX      = @(x,y) (2*pi*pi*(sin(pi*x)).*cos(pi*y) - 2*x).*x.^0;
fY      = @(x,y) (-2*pi*pi*(cos(pi*x)).*sin(pi*y) - 0).*y.^0;
gN      = @(x,y) [0.*x, 0.*y];

T= cuadrado;
%T= cuadradoref1;
%T= cuadradoref2;
%T= cuadradoref3;
%T= cuadradoref4;
%T= cuadradoref5;

if length(T.neumann)>0
disp('Neumann condition detected')
disp('Domain is not suitable for Stokes equation')
disp('Try with another one')
return
end

chr = uint16('.');
DispText    = zeros(1,40);
DispText(:) = chr;

disp('Stokes FEM experiment')
disp('======================')
disp('Full vectorized P1bubble-P1 FEM implementation')
disp(' ')
disp(['Mesh with ' num2str(length(T.elements)) ' triangles and '...
num2str(max(max(T.elementsbubble))) ' nodes'])

tic

[S,~]    = FEMmatricesbubble_v4(T);
[~,M]    = FEMmatrices_v4(T);
[B1,B2] = FEMmatricesP1bubbleP1_v4(T);
ele = sum(M,2);
```

```
clear M

mess='Assembly of the matrix: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

tic

DispText(:)=chr;
mess='Assembly of the rhs: ';
DispText(1:length(mess))= mess;

[LoadX,~]=FEMrhsbubble_v4(T,fX,gN);
[LoadY,~]=FEMrhsbubble_v4(T,fY,gN);

disp([DispText num2str(toc) ' seconds '] )

% Solver

nNodes = max(max(T.elements(:,1:3)));
nNodesbubble = max(max(T.elementsbubble(:,1:4)));
iD     = unique(T.dirichlet(:));
iND   = 1:nNodesbubble; iND(iD)=[];
uX    = zeros(nNodesbubble,1);
uY    = zeros(nNodesbubble,1);

% Direct method for solving the linear system

mess='Size of the matrix:..........';
DispText(1:length(mess))= mess;

disp([DispText num2str(2*length(iND)+nNodes) ' x ' ...
      num2str(2*length(iND)+nNodes) ] )
disp('Direct method')

uX(iD) = uExactX(T.coordinates(iD,1),T.coordinates(iD,2));
uY(iD) = uExactY(T.coordinates(iD,1),T.coordinates(iD,2));

LoadX = LoadX-S(:,iD)*uX(iD);
LoadY = LoadY-S(:,iD)*uY(iD);
LoadP = B1(:,iD)*uX(iD)+B2(:,iD)*uY(iD);

matrix = kron(eye(2),S(iND,iND));
matrix = [matrix                          -[B1(:,iND)'; B2(:,iND)'];...
-[B1(:,iND)    B2(:,iND)]    sparse(nNodes,nNodes)   ];
matrix = [matrix [sparse(2*length(iND),1); ele];...
sparse(1,2*length(iND)) ele' 0 ];
rhs    = [LoadX(iND); LoadY(iND); LoadP; 0];
tic
sol    = matrix\rhs;
toc
```

```
niND=length(iND);
uX(iND) = sol(1:niND);
uY(iND) = sol((1:niND) + niND);
P = sol(2*niND+1:(end-1));

DispText(:)=chr;
mess='Solution of the linear system: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

disp(' ' )
DispText(1:length(mess))= mess;
disp('Error computed in H1 and L2')

DispText(:)=chr;

% Error computed in H1 and L2

ErrorH1_uX= [(uExactX(T.coordinates(:,1),T.coordinates(:,2)) ...
-uX(1:length(T.coordinates))) ; ...
(uExactX(T.baryc(:,1),T.baryc(:,2))-(uX(T.elements(:,1)) ...
+uX(T.elements(:,2))+uX(T.elements(:,3)))/3 ...
 -uX(length(T.coordinates)+1:(length(T.coordinates)+length(T.baryc))))];
ErrorH1_uX=(ErrorH1_uX' * S)*ErrorH1_uX;
ErrorH1_uX=sqrt(ErrorH1_uX);
mess=' uX.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uX,'%8.3e')] )

ErrorH1_uY= [(uExactY(T.coordinates(:,1),T.coordinates(:,2)) ...
-uY(1:length(T.coordinates))) ; ...
(uExactY(T.baryc(:,1),T.baryc(:,2))-(uY(T.elements(:,1)) ...
+uY(T.elements(:,2))+uY(T.elements(:,3)))/3 ...
-uY(length(T.coordinates)+1:(length(T.coordinates)+length(T.baryc))))];
ErrorH1_uY=(ErrorH1_uY' * S)*ErrorH1_uY;
ErrorH1_uY=sqrt(ErrorH1_uY);
mess=' uY.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uY,'%8.3e')] )

errorP = (pExact(T.coordinates(1:nNodes,1), ...
                 T.coordinates(1:nNodes,2))-P) ;
meanP=sum(errorP)/nNodes;
errorP2 = (pExact(T.coordinates(1:nNodes,1), ...
                  T.coordinates(1:nNodes,2))-P-meanP).^2 ;

for i=1:length(T.elements)
aux(i)=errorP2(T.elements(i,1))+errorP2(T.elements(i,2)) ...
       +errorP2(T.elements(i,3));

ErrorL2_P(i)=(T.detB(i)/3)*aux(i);
```

```
end
ErrorL2_P=sqrt(sum(ErrorL2_P));

mess=' P.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorL2_P,'%8.3e')] )
```

FEMmatricesbubble_v4.m

```
function [S,M,Load,Tr]=FEMmatricesbubble_v4(T,f,gN)
% Matrices in the reference element
Mxx =  0.5 * [1,-1,0,0;
              -1,1,0,0;
               0,0,0,0;
               0,0,0,81/10];
Mxy =  0.5 * [1,-1,0,0;
               0,0,0,0;
              -1,1,0,0;
               0,0,0,81/20];
Myy =  0.5 * [1,0,-1,0;
               0,0,0,0;
              -1,0,1,0;
               0,0,0,81/10];
M0  = [1/12  1/24  1/24  3/40;
       1/24  1/12  1/24  3/40;
       1/24  1/24  1/12  3/40;
       3/40  3/40  3/40  81/560];

nTr = length(T.elements);
nbubbles = nTr;
nNodes  = max(T.elements(:)) + nbubbles;
S       = sparse(nNodes,nNodes);
M       = sparse(nNodes,nNodes);
Load    = zeros(nNodes,1);

[j,i] = meshgrid([1 2 3 4],[1 2 3 4]);

indi = zeros(4,4*nTr);
indj = indi;
indi(:) = T.elementsbubble(:,i)';
indj(:) = T.elementsbubble(:,j)';
M  = kron(T.detB',M0);
M  = sparse(indi,indj,M);
S  = kron(T.c11',Mxx)+kron(T.c22',Myy)+kron(T.c12',Mxy+Mxy');
S  = sparse(indi,indj,S);

return
```

FEMmatrices_v4.m

```
function [S,M,Load,Tr]=FEMmatrices_v4(T,f,gN)
% Matrices in the reference element

K11 =  0.5 * [1 -1  0; -1  1  0;  0  0  0];
K12 =  0.5 * [1  0 -1; -1  0  1;  0  0  0];
K22 =  0.5 * [1  0 -1;  0  0  0; -1  0  1];
Mk  = 1/24 * [2  1  1;  1  2  1;  1  1  2];

nTr = length(T.elements);
nNodes  = max(T.elements(:)); % other choices: nTr = max(T.coord);
S       = sparse(nNodes,nNodes);
M       = sparse(nNodes,nNodes);
Load    = zeros(nNodes,1);

[j,i] = meshgrid([1 2 3],[1 2 3]); %%

indi = zeros(3,3*nTr);
indj = indi;
indi(:) = T.elements(:,i)';
indj(:) = T.elements(:,j)';
M  = kron(T.detB',Mk);
M  = sparse(indi,indj,M);
S  = kron(T.c11',K11)+kron(T.c22',K22)+kron(T.c12',K12+K12');
S  = sparse(indi,indj,S);

return
```

FEMmatricesP1bubbleP1_v4.m

```
function [B1,B2]=FEMmatricesP1bubbleP1_v4(T)

% Matrices in the reference element

Kx0 = [-1/6  1/6  0  9/40;
       -1/6  1/6  0  -9/40;
       -1/6  1/6  0  0;
       -9/40  9/40  0  0];
Ky0 = [-1/6  0  1/6  9/40;
        -1/6   0  1/6  0;
        -1/6  0  1/6  -9/40;
        -9/40  0  9/40  0];%

if ~isfield(T,'b11')
px = T.coordinates(:,1);
py = T.coordinates(:,2);

T.b11 = px(T.elements(:,2))-px(T.elements(:,1));
T.b12 = px(T.elements(:,3))-px(T.elements(:,1));
T.b21 = py(T.elements(:,2))-py(T.elements(:,1));
T.b22 = py(T.elements(:,3))-py(T.elements(:,1));
clear px py
end
nTr = length(T.elements);
nNodesP1  = max(max(T.elements(:,1:3)));
nNodesP1bubble  = max(max(T.elementsbubble(:,1:4)));
[j,i] = meshgrid([1 2 3 4 ],[1 2 3]);
indi = zeros(3,4*nTr)';
indj = zeros(3,4*nTr)';
indi(:) = T.elementsbubble(:,i)';
indj(:) = T.elementsbubble(:,j)';
B1  = kron(T.b22',Kx0(1:3,:))-kron(T.b21',Ky0(1:3,:));
B1  = sparse(indi,indj,B1);
B2  = -kron(T.b12',Kx0(1:3,:))+kron(T.b11',Ky0(1:3,:));
B2  = sparse(indi,indj,B2);

return
```

FEMrhsbubble_v4.m

```
function [Load,Tr]=FEMrhsbubble_v4(T,f,gN)

nodes    = [1/2   0    1/2   ;...
            1/2   1/2    0   ];
nodesB   = [1-nodes(1,:)-nodes(2,:);
            nodes];
P1bubbleValues = [nodesB ; ...
                  27.*nodes(1,:).*nodes(2,:).*(1-nodes(1,:)-nodes(2,:))];
weights  = [1/6; 1/6; 1/6];

nTr    = length(T.elements);
nNodes = max(T.elements(:)); % other choices: nTr = max(T.coord);

px = T.coordinates(:,1);
py = T.coordinates(:,2);

val = f(px(T.elements(:,1:3))*nodesB, py(T.elements(:,1:3))*nodesB);
indNodes = kron((1:nTr)',[1 1 1 1]');
aux = kron(T.detB(:),P1bubbleValues);
aux = aux.*val(indNodes,:);
clear val
indT =T.elementsbubble'; indT=indT(:);
aux3=aux*weights ; aux3 = aux3(:);
Load = accumarray(indT, aux3);

Tr = zeros(nNodes,1);


return
```

Appendix B. Taylor-Hood Finite Element Method

scriptStokesP2P1.m

```
% Exact solution, rhs and Dirichlet boundary conditions
uExactX = @(x,y) (sin(pi*x)).*cos(pi*y);
uExactY = @(x,y) -(cos(pi*x)).*sin(pi*y);
pExact  = @(x,y) 1/2 - x.^2;
fX      = @(x,y) (2*pi*pi*(sin(pi*x)).*cos(pi*y) - 2*x).*x.^0;
fY      = @(x,y) (-2*pi*pi*(cos(pi*x)).*sin(pi*y) - 0).*y.^0;
gN      = @(x,y) [0.*x, 0.*y];

T= cuadrado;
%T= cuadradoref1;
%T= cuadradoref2;
%T= cuadradoref3;
%T= cuadradoref4;
%T= cuadradoref5;

if length(T.neumann)>0
disp('Neumann condition detected')
disp('Domain is not suitable for Stokes equation')
disp('Try with another one')
return
end

T2 = prepareGridP2(T);

chr = uint16('.');
DispText    = zeros(1,40);
DispText(:) = chr;

disp('Stokes FEM experiment')
disp('======================')

disp(' ')
disp('Full vectorized P2 FEM implementation')
disp(' ')
disp(['Mesh with ' num2str(length(T2.elements)) ' triangles and '...
num2str(length(T2.coordinates)) ' nodes'])
```

```
tic

[S,~]    = FEMmatricesP2_v4(T2);
[~,M]    = FEMmatrices_v4(T);
[B1,B2] = FEMmatricesP2P1_v4(T2);

% ele(j) =\int_\Omega \varphi_j
ele = sum(M,2);
clear M

mess='Assembly of the matrix: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

tic

DispText(:)=chr;
mess='Assembly of the rhs: ';
DispText(1:length(mess))= mess;

[LoadX,~]=FEMrhsP2_v4(T2,fX,gN);
[LoadY,~]=FEMrhsP2_v4(T2,fY,gN);

disp([DispText num2str(toc) ' seconds '] )

% Solver
nNodesP1 = max(max(T2.elements(:,1:3)));
nNodesP2 = max(max(T2.elements(:,1:6)));
iD     = unique(T2.dirichlet(:));
iND   = 1:nNodesP2; iND(iD)=[];
uX    = zeros(length(T2.coordinates),1);
uY    = zeros(length(T2.coordinates),1);

% Direct method for solving the linear system

mess='Size of the matrix:.....';
DispText(1:length(mess))= mess;

disp([DispText num2str(2*length(iND)+nNodesP1) ' x ' ...
num2str(2*length(iND)+nNodesP1) ] )
disp('Direct method')

uX(iD) = uExactX(T2.coordinates(iD,1),T2.coordinates(iD,2));
uY(iD) = uExactY(T2.coordinates(iD,1),T2.coordinates(iD,2));

LoadX = LoadX-S(:,iD)*uX(iD);
LoadY = LoadY-S(:,iD)*uY(iD);
LoadP = B1(:,iD)*uX(iD)+B2(:,iD)*uY(iD);

matrix = kron(eye(2),S(iND,iND));
matrix = [matrix                          -[B1(:,iND)'; B2(:,iND)'];...
```

```
-[B1(:,iND)    B2(:,iND)]    sparse(nNodesP1,nNodesP1)   ];
matrix = [matrix [sparse(2*length(iND),1); ele];...
sparse(1,2*length(iND)) ele' 0 ];
rhs    = [LoadX(iND); LoadY(iND); LoadP; 0];
tic
sol    = matrix\rhs;
toc
niND=length(iND);
uX(iND) = sol(1:niND);
uY(iND) = sol((1:niND) + niND);
P = sol(2*niND+1:(end-1));

DispText(:)=chr;
mess='Solution of the linear system: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

disp(' ' )
DispText(1:length(mess))= mess;
disp('Error computed in H1 and L2')

DispText(:)=chr;

% Error computed in H1 and L2

ErrorH1_uX= [(uExactX(T2.coordinates(:,1),T2.coordinates(:,2)) ...
-uX(1:length(T2.coordinates)))];
ErrorH1_uX=(ErrorH1_uX' * S)*ErrorH1_uX;
ErrorH1_uX=sqrt(ErrorH1_uX);
mess=' uX.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uX,'%8.3e')] )

ErrorH1_uY= [(uExactY(T2.coordinates(:,1),T2.coordinates(:,2)) ...
-uY(1:length(T2.coordinates)))];
ErrorH1_uY=(ErrorH1_uY' * S)*ErrorH1_uY;
ErrorH1_uY=sqrt(ErrorH1_uY);
mess=' uY.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uY,'%8.3e')] )

errorP = (pExact(T.coordinates(1:nNodesP1,1), ...
                 T.coordinates(1:nNodesP1,2))-P) ;
meanP=sum(errorP)/nNodesP1;
errorP2 =((pExact(T.coordinates(1:nNodesP1,1), ...
                  T.coordinates(1:nNodesP1,2))-P-meanP)).^2 ;

for i=1:length(T.elements)
aux(i)=errorP2(T.elements(i,1))+errorP2(T.elements(i,2)) ...
       +errorP2(T.elements(i,3));
ErrorL2_P(i)=(T.detB(i)/3)*aux(i);
```

```
end
ErrorL2_P=sqrt(sum(ErrorL2_P));

mess=' P.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorL2_P,'%8.3e')] )
```

prepareGridP2.m

```
function T2 = prepareGridP2(T)
T2 = T;
nNodesOld = max(T.elements(:));
px = T.coordinates(:,1);
py = T.coordinates(:,2);

edgesLocal =[2 3; 3 1; 1 2];

edges=[];
for j=1: length(edgesLocal)
edges = [edges; T.elements(:,edgesLocal(j,:))];
end
edges = sort(edges,2);
edges = unique(edges,'rows');
newNodes = (1:length(edges))+nNodesOld;
mCon = sparse([edges(:,1); edges(:,2)],[edges(:,2); edges(:,1)],...
[newNodes newNodes]);
T2.coordinates=[T.coordinates;...
(px(edges(:,1))+px(edges(:,2)))/2 ...
(py(edges(:,1))+py(edges(:,2)))/2];
for j=1: length(edgesLocal)

ind = sub2ind(size(mCon),...
T.elements(:,edgesLocal(j,1)),T.elements(:,edgesLocal(j,2)));
T2.elements(:,end+1)=mCon(ind);
end

% Dirichlet
ind = sub2ind(size(mCon),T.dirichlet(:,1),T.dirichlet(:,2));
T2.dirichlet(:,end+1)=mCon(ind);
```

FEMmatricesP2P1_v4.m

```
function [B1,B2]=FEMmatricesP2P1_v4(T)
% Matrices in the reference element

B1hat = [-1/6    0    0   1/6    -1/6    1/6;
          0     1/6   0   1/6    -1/6    -1/6;
          0      0    0   1/3    -1/3     0;];

B2hat = [-1/6    0    0    1/6    1/6    -1/6;
          0      0    0    1/3     0     -1/3;
          0      0   1/6   1/6   -1/6    -1/6;];

if ~isfield(T,'b11')
px = T.coordinates(:,1);
py = T.coordinates(:,2);

T.b11 = px(T.elements(:,2))-px(T.elements(:,1));
T.b12 = px(T.elements(:,3))-px(T.elements(:,1));
T.b21 = py(T.elements(:,2))-py(T.elements(:,1));
T.b22 = py(T.elements(:,3))-py(T.elements(:,1));
clear px py
end
nTr = length(T.elements);
nNodesP1  = max(max(T.elements(:,1:3)));
nNodesP2  = max(max(T.elements(:,1:6)));

[j,i] = meshgrid([1 2 3 4 5 6],[1 2 3]);
indi = zeros(3,6*nTr)';
indj = zeros(3,6*nTr)';
indi(:) = T.elements(:,i)';
indj(:) = T.elements(:,j)';
B1  = kron(T.b22',B1hat)-kron(T.b21',B2hat);
B1  = sparse(indi,indj,B1);
B2  = -kron(T.b12',B1hat)+kron(T.b11',B2hat);
B2  = sparse(indi,indj,B2);

return
```

FEMmatricesP2_v4.m

```
function [S,M]=FEMmatricesP2_v4(T)

% Matrices in the reference element

S11 = [ 3     1     0     0     0    -4 ; ...
        1     3     0     0     0    -4 ; ...
        0     0     0     0     0     0 ; ...
        0     0     0     8    -8     0 ; ...
        0     0     0    -8     8     0 ; ...
       -4    -4     0     0     0     8]/6;

S12 =[  3     0     1     0    -4     0 ; ...
        1     0    -1     4     0    -4 ; ...
        0     0     0     0     0     0 ; ...
        0     0     4     4    -4    -4 ; ...
        0     0    -4    -4     4     4 ; ...
       -4     0     0    -4     4     4 ]/6;

S22 = [ 3     0     1     0    -4     0;...
        0     0     0     0     0     0;...
        1     0     3     0    -4     0;...
        0     0     0     8     0    -8;...
       -4     0    -4     0     8     0;...
        0     0     0    -8     0     8]/6;


Mk = [  6    -1    -1    -4     0     0;...
       -1     6    -1     0    -4     0;...
       -1    -1     6     0     0    -4;...
       -4     0     0    32    16    16;...
        0    -4     0    16    32    16;...
        0     0    -4    16    16    32]/360;


nTr = length(T.elements);
nNodes  = max(T.elements(:));
S       = sparse(nNodes,nNodes);
M       = sparse(nNodes,nNodes);

[j,i] = meshgrid([1 2 3 4 5 6],[1 2 3 4 5 6]);
indi = zeros(6,6*nTr);
indj = indi;
indi(:) = T.elements(:,i)';
indj(:) = T.elements(:,j)';
M  = kron(T.detB',Mk);
M  = sparse(indi,indj,M);
S  = kron(T.c11',S11)+kron(T.c22',S22)+kron(T.c12',S12+S12');
S  = sparse(indi,indj,S);

return
```

FEMrhsP2_v4.m

```
function [Load,Tr]=FEMrhsP2_v4(T,f,gN)

nodes    = [1/2   0    1/2  ;...
            1/2   1/2    0  ];

nodesB   = [1-nodes(1,:)-nodes(2,:); nodes];
P2Values = [2*(1-nodes(1,:)-nodes(2,:)).*(0.5-nodes(1,:)-nodes(2,:));...
             2*nodes(1,:).*(nodes(1,:)-0.5);...
             2*nodes(2,:).*(nodes(2,:)-0.5);...
             4*nodes(1,:).*nodes(2,:);...
             4*(1-nodes(1,:)-nodes(2,:)).*nodes(2,:);...
             4*(1-nodes(1,:)-nodes(2,:)).*nodes(1,:)];
weights  = [1/6; 1/6; 1/6];

nTr    = length(T.elements);
nNodes = max(T.elements(:));

px = T.coordinates(:,1);
py = T.coordinates(:,2);

val = f(px(T.elements(:,1:3))*nodesB,py(T.elements(:,1:3))*nodesB);

indNodes = kron((1:nTr)',[1 1 1 1 1 1]');
aux = kron(T.detB(:),P2Values);
aux = aux.*val(indNodes,:);
clear val
indT =T.elements'; indT=indT(:);
Load = accumarray(indT, aux*weights);

clear indT aux

nNeumann = length(T.neumann);
Tr = zeros(nNodes,1);

return
```

Appendix C. The Linear-Linear Approximation

scriptStokesP1P1.m

```
% Exact solution, rhs and Dirichlet boundary conditions
uExactX = @(x,y) (sin(pi*x)).*cos(pi*y);
uExactY = @(x,y) -(cos(pi*x)).*sin(pi*y);
pExact  = @(x,y) 1/2 - x.^2;
fX      = @(x,y) (2*pi*pi*(sin(pi*x)).*cos(pi*y) - 2*x).*x.^0;
fY      = @(x,y) (-2*pi*pi*(cos(pi*x)).*sin(pi*y) - 0).*y.^0;
gN      = @(x,y) [0.*x, 0.*y];

T= cuadrado;
%T= cuadradoref1;
%T= cuadradoref2;
%T= cuadradoref3;
%T= cuadradoref4;
%T= cuadradoref5;

if length(T.neumann)>0
disp('Neumann condition detected')
disp('Domain is not suitable for Stokes equation')
disp('Try with another one')
return
end

chr = uint16('.');
DispText    = zeros(1,40);
DispText(:) = chr;

disp('Stokes FEM experiment')
disp('====================')
disp('Full vectorized P1-P1 FEM implementation')
disp(' ')
disp(['Mesh with ' num2str(length(T.elements)) ' triangles and '...
num2str(length(T.coordinates)) ' nodes'])

tic

[S,~]   = FEMmatrices_v4(T);
[~,M]   = FEMmatrices_v4(T);
```

```
[B1,B2] = FEMmatricesP1P1_v4(T);

% ele(j) =\int_\Omega \varphi_j
ele = sum(M,2);
clear M

mess='Assembly of the matrix: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

tic

DispText(:)=chr;
mess='Assembly of the rhs: ';
DispText(1:length(mess))= mess;

[LoadX,~]=FEMrhs_v4(T,fX,gN);
[LoadY,~]=FEMrhs_v4(T,fY,gN);

disp([DispText num2str(toc) ' seconds '] )

% Solver
nNodes = max(max(T.elements(:,1:3)));
iD     = unique(T.dirichlet(:));
iND    = 1:nNodes; iND(iD)=[];
uX     = zeros(nNodes,1);
uY     = zeros(nNodes,1);

% Direct method for solving the linear system

mess='Size of the matrix:..........';
DispText(1:length(mess))= mess;

disp([DispText num2str(2*length(iND)+nNodes) ' x ' ...
num2str(2*length(iND)+nNodes) ] )
disp('Direct method')

uX(iD) = uExactX(T.coordinates(iD,1),T.coordinates(iD,2));
uY(iD) = uExactY(T.coordinates(iD,1),T.coordinates(iD,2));

LoadX = LoadX-S(:,iD)*uX(iD);
LoadY = LoadY-S(:,iD)*uY(iD);
LoadP = B1(:,iD)*uX(iD)+B2(:,iD)*uY(iD);

matrix = kron(eye(2),S(iND,iND));
matrix = [matrix                          -[B1(:,iND)'; B2(:,iND)'];...
-[B1(:,iND)    B2(:,iND)]    sparse(nNodes,nNodes)   ];
matrix = [matrix [sparse(2*length(iND),1); ele ];...
sparse(1,2*length(iND)) ele' 0 ];
rhs    = [LoadX(iND); LoadY(iND); LoadP; 0];
tic
```

```
sol     = matrix\rhs;
toc
niND=length(iND);
uX(iND) = sol(1:niND);
uY(iND) = sol((1:niND) + niND);
P = sol(2*niND+1:(end-1));

DispText(:)=chr;
mess='Solution of the linear system: ';
DispText(1:length(mess))= mess;
disp([DispText num2str(toc) ' seconds '] )

disp(' ' )
DispText(1:length(mess))= mess;
disp('Error computed in H1 and L2')
DispText(:)=chr;

ErrorH1_uX= [(uExactX(T.coordinates(:,1),T.coordinates(:,2)) ...
-uX(1:length(T.coordinates)))];
ErrorH1_uX=(ErrorH1_uX' * S)*ErrorH1_uX;
ErrorH1_uX=sqrt(ErrorH1_uX);
mess=' uX.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uX,'%8.3e')] )

ErrorH1_uY= [(uExactY(T.coordinates(:,1),T.coordinates(:,2)) ...
-uY(1:length(T.coordinates))) ];
ErrorH1_uY=(ErrorH1_uY' * S)*ErrorH1_uY;
ErrorH1_uY=sqrt(ErrorH1_uY);
mess=' uY.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorH1_uY,'%8.3e')] )
errorP = (pExact(T.coordinates(1:nNodes,1), ...
                 T.coordinates(1:nNodes,2))-P) ;
meanP=sum(errorP)/nNodes;
errorP2 = ((pExact(T.coordinates(1:nNodes,1), ...
                   T.coordinates(1:nNodes,2))-P-meanP)).^2 ;

for i=1:length(T.elements)
aux(i)=errorP2(T.elements(i,1))+errorP2(T.elements(i,2))...
       +errorP2(T.elements(i,3));
ErrorL2_P(i)=(T.detB(i)/3)*aux(i);
end
ErrorL2_P=sqrt(sum(ErrorL2_P));

mess=' P.';
DispText(1:length(mess))= mess;
disp([DispText num2str(ErrorL2_P,'%8.3e')] )
```

FEMmatricesP1P1_v4.m

```
function [B1,B2]=FEMmatricesP1P1_v4(T)

Kx0 = [-1/6  1/6  0 ;
       -1/6  1/6  0 ;
       -1/6  1/6  0 ];
Ky0 = [-1/6  0  1/6 ;
       -1/6   0  1/6 ;
       -1/6  0  1/6 ];

if ~isfield(T,'b11')
px = T.coordinates(:,1);
py = T.coordinates(:,2);
T.b11 = px(T.elements(:,2))-px(T.elements(:,1));
T.b12 = px(T.elements(:,3))-px(T.elements(:,1));
T.b21 = py(T.elements(:,2))-py(T.elements(:,1));
T.b22 = py(T.elements(:,3))-py(T.elements(:,1));
clear px py
end
nTr = length(T.elements);
nNodesP1   = max(max(T.elements(:,1:3)));

[j,i] = meshgrid([1 2 3],[1 2 3]);
indi = zeros(3,3*nTr)';
indj = zeros(3,3*nTr)';
indi(:) = T.elementsbubble(:,i)';
indj(:) = T.elementsbubble(:,j)';
B1  = kron(T.b22',Kx0(1:3,:))-kron(T.b21',Ky0(1:3,:));
B1  = sparse(indi,indj,B1);
B2  = -kron(T.b12',Kx0(1:3,:))+kron(T.b11',Ky0(1:3,:));
B2  = sparse(indi,indj,B2);

return
```

FEMrhs_v4.m

```
function [Load,Tr]=FEMrhs_v4(T,f,gN)

nodes    = [1/2   0    1/2   ;...
            1/2   1/2    0   ];
nodesB   = [1-nodes(1,:)-nodes(2,:); nodes];
P1Values = nodesB;
weights  = [1/6; 1/6; 1/6 ];

nTr    = length(T.elements);
nNodes = max(T.elements(:));

px = T.coordinates(:,1);
py = T.coordinates(:,2);

val = f(px(T.elements(:,1:3))*nodesB,py(T.elements(:,1:3))*nodesB);

indNodes = kron((1:nTr)',[1 1 1]');
aux = kron(T.detB(:),P1Values);
aux = aux.*val(indNodes,:);
clear val
indT =T.elements'; indT=indT(:);
Load = accumarray(indT, aux*weights);

clear indT aux
Tr = zeros(nNodes,1);
return
```