Chapter 3

Finite elements for mixed and saddle points problems

3.1 Galerkin method for mixed problems

Notations and setting. The notation is the same as in the abstract framework introduced in the previous chapter. Let $(X, \|\cdot\|_X)$ and $(M, \|\cdot\|_M)$ be two real Hilbert spaces and $a : X \times X \to \mathbb{R}$ and $b : X \times M \to \mathbb{R}$ be two continuous bilinear forms. For $f \in X'$ and $g \in M'$, the following problem is considered

Find $(u, p) \in X \times M$ *such that, for all* $(v, q) \in X \times M$

$$\left\{ \begin{array}{rcl} a(u,v) + b(v,p) & = & \langle f, v\rangle, \\ b(u,q) & = & \langle g, q\rangle. \end{array} \right. \tag{3.1}$$

This chapter focuses on the approximation of problem (3.1) by a Galerkin method. To simplify the presentation, it is assumed that $a(\cdot,\cdot)$ is coercive on $V \times V$ — that is,

$$\exists\, \alpha > 0, \quad a(v,v) \geq \alpha \|v\|^2, \quad \forall\, v \in V, \tag{3.2}$$

which implies assumption (i) of Theorem 2.5 (by the Lax-Milgram theorem). In addition, we assume that the inf-sup condition (2.9) holds true on $X \times M$. Thus, assumptions (i) and (ii) of Theorem 2.5 are fulfilled. This ensures that problem (3.1) is well-posed.

As in any Galerkin type method (see Section 1.2), *finite dimensional subspaces* X_h of X and M_h of M are considered, and the following discrete problem, which is the *mixed Galerkin approximation* of problem (3.1), is introduced

Find $(u_h, p_h) \in X_h \times M_h$*, such that for all* $(v_h, q_h) \in X_h \times M_h$*,*

$$\left\{ \begin{array}{rcl} a(u_h, v_h) + b(v_h, p_h) & = & \langle f, v_h \rangle, \\ b(u_h, q_h) & = & \langle g, q_h \rangle. \end{array} \right. \tag{3.3}$$

Operators $A_h : X_h \to X_h'$ and $B_h : X_h \to M_h'$ can be defined as in the continuous case by

$$\langle A_h u_h, v_h \rangle = a(u_h, v_h), \quad \forall\, (u_h, v_h) \in X_h \times X_h, \tag{3.4}$$

$$\langle B_h v_h, q_h \rangle = b(v_h, q_h), \quad \forall\, (v_h, q_h) \in X_h \times M_h. \tag{3.5}$$

The dual operator of B_h is denoted by B_h^T and is defined by $B_h^T : M_h \to X_h'$, $\langle B_h^T q_h, v_h \rangle = b(v_h, q_h) = \langle B_h v_h, q_h \rangle$, for all $(v_h, q_h) \in X_h \times M_h$.

Differences with coercive problems. When a problem can be studied using the Lax-Milgram theorem on a space X, any of its finite dimensional internal approximation on $X_h \subset X$ can also be treated by the Lax-Milgram theorem. The well-posedness of the discrete problem is therefore a straightforward consequence of the well-posedness of the continuous problem. In contrast, the well-posedness of problem (3.1) does not imply in general that its discrete counterpart (3.3) is also well-posed. The reason for this is twofold:

- First, if one defines

  $$V_h = \mathrm{Ker}\, B_h = \{u_h \in X_h, \forall\, q_h \in M_h, b(u_h, q_h) = 0\}, \tag{3.6}$$

  then V_h is not necessarily included in V (for example, for the Stokes problem, $u_h \in V_h$ does not imply that u_h is divergence free). Thus, the fact that the continuous problem satisfies property (i) of Theorem 2.1 on V does not imply that the discrete problem satisfies the analogous property on V_h.

- Second, the inf-sup condition on $X \times M$,

  $$\exists\, \beta > 0, \inf_{q \in M} \sup_{v \in X} \frac{b(v, q)}{\|v\|_X \|q\|_M} \geq \beta,$$

  only implies

  $$\exists\, \beta > 0, \inf_{q_h \in M_h} \sup_{v \in X} \frac{b(v, q_h)}{\|v\|_X \|q_h\|_M} \geq \beta,$$

  which is an inf-sup condition on $X \times M_h$. The latter does not imply in general an inf-sup condition on $X_h \times M_h$, since $X \supset X_h$.

Therefore, it is necessary to assume that assumptions (i) and (ii) of Theorem 2.5 are satisfied by the discrete problem on $X_h \times M_h$ itself. This assumption is very strong, since in many practical cases the inf-sup condition is not satisfied if spaces X_h and M_h are not chosen adequately. This difficulty will be illustrated later for the Stokes problem. For now, it will be assumed that one is able to build X_h and M_h that satisfy the inf-sup condition. Then, the following theorem gives an estimate of the error of the Galerkin method by the interpolation error. In other words, the following theorem is the analogous of the Céa Lemma (Theorem 1.3, p. 10) for mixed problems.

Theorem 3.1
Assume that the coercivity hypothesis (3.2) on V and the inf-sup condition (2.9) on $X \times M$ hold and let (u,p) be the solution of (3.1). Assume in addition that[1]

(i) $\exists\, \alpha_h > 0$ *such that* $\forall\, v_h \in V_h$, $a(v_h, v_h) \geq \alpha_h \|v_h\|_X^2$.

(ii) $\exists\, \beta_h > 0$, *such that* $\displaystyle\inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X \|q_h\|_M} \geq \beta_h$.

Then problem (3.3) admits a unique solution and this solution satisfies

$$\begin{aligned} \|u - u_h\|_X \quad &\leq \quad \left(1 + \frac{\|a\|}{\alpha_h}\right)\left(1 + \frac{\|b\|}{\beta_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_X \\ &\qquad + \frac{\|b\|}{\alpha_h} \inf_{q_h \in M_h} \|p - q_h\|_M, \end{aligned} \tag{3.7}$$

and

$$\begin{aligned} \|p - p_h\|_X \quad &\leq \quad \frac{\|a\|}{\beta_h}\left(1 + \frac{\|a\|}{\alpha_h}\right)\left(1 + \frac{\|b\|}{\beta_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_X \\ &\qquad \left(1 + \frac{\|b\|}{\beta_h} + \frac{\|a\|\|b\|}{\alpha_h \beta_h}\right) \inf_{q_h \in M_h} \|p - q_h\|_M. \end{aligned} \tag{3.8}$$

Considering the standard interpolation errors estimates (like (1.26) and (1.27) p. 13), this result readily gives the convergence of the mixed Galerkin method. To achieve an optimal convergence rates, α_h and β_h are typically required to be independent of h.

The proof of Theorem 3.1 is much more involved than its counterpart in the coercive framework. It is based on the two following lemmas concerning the following set

$$\begin{aligned} V_h(g) \quad &= \quad \{v_h \in X_h, B_h v_h = g\} \\ &= \quad \{v_h \in X_h, b(v_h, q_h) = \langle g, q_h\rangle, \forall q_h \in M_h\}. \end{aligned} \tag{3.9}$$

The reader can observe that the space V_h introduced in (3.6) is nothing but $V_h(0)$.

Lemma 3.1 *Assume that*

[1] Space V_h is defined in (3.6).

(i) $\exists\,\alpha_h > 0$ *such that* $\forall\, v_h \in V_h$, $a(v_h, v_h) \geq \alpha_h \|v_h\|_X^2$.

(ii) The set $V_h(g)$ *is non-empty.*

Then, the problem of finding $u_h \in V_h(g)$ *such that for all* $v_h \in V_h$

$$a(u_h, v_h) = \langle f, v_h \rangle$$

has a unique solution. In addition, this solution satisfies

$$\|u - u_h\|_X \leq \left(1 + \frac{\|a\|}{\alpha}\right) \inf_{z_h \in V_h(g)} \|u - z_h\|_X + \frac{\|b\|}{\alpha} \inf_{q_h \in M_h} \|p - q_h\|_M.$$

The following Lemma shows the role played by the discrete inf-sup condition in the approximation property of the set $V_h(g)$.

Lemma 3.2 *Assume that the inf-sup condition is satisfied on* $X_h \times M_h$ *— that is,*

$$\exists\,\beta_h > 0, \text{ such that } \inf_{q_h \in M_h} \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X \|q_h\|_M} \geq \beta_h.$$

Let $u \in X$ *such that* $b(u, q) = \langle g, q \rangle, \forall q \in M$. *Then*

$$\inf_{z_h \in V_h(g)} \|u - z_h\| \leq \left(1 + \frac{\|b\|}{\beta_h}\right) \inf_{v_h \in X_h} \|u - v_h\|_X. \tag{3.10}$$

Remark 3.1 *Lemma 3.2 is useful for understanding the so-called* locking *phenomena (see Section 3.4).*

3.2 Algebraic aspects

Consider a basis $(\varphi_i)_{i=1..N_u}$ (respectively, $(\psi_i)_{i=1..N_p}$) of X_h (respectively of M_h). Any element $u_h \in X_h$ and $p_h \in M_h$ can be decomposed on these bases as follows

$$u_h = \sum_{i=1}^{N_u} U_i \varphi_i, \qquad \text{and} \qquad p_h = \sum_{i=1}^{N_p} P_i \psi_i.$$

Denote by U the vector $(U_1, .., U_{N_u})^T \in \mathbb{R}^{N_u}$, and by P the vector $(P_1, .., P_{N_p})^T \in \mathbb{R}^{N_p}$. In the sequel, this notation is systematically used (for example, $\mathrm{V} = (V_1, \ldots, V_{N_u}) \in \mathbb{R}^{N_u}$ represents the coordinates of v_h on $(\varphi_i)_{i=1..N_u}$).

Although all norms are equivalent in a finite dimensional space, it is useful to introduce the following specific ones. For any $\mathrm{V} \in \mathbb{R}^{N_u}$, define

$$\|\mathrm{V}\|_X = \|v_h\|_X, \qquad \|\mathrm{V}\|_* = \sup_{\mathrm{W} \in \mathbb{R}^{N_u}} \frac{(\mathrm{V}, \mathrm{W})}{\|\mathrm{W}\|_X},$$

where $v_h = \sum_{i=1..N_u} V_i \varphi_i$. In the same spirit, for any $\mathrm{Q} \in \mathbb{R}^{N_p}$, define

$$\|\mathrm{Q}\|_M = \|q_h\|_M,$$

where $q_h = \sum_{i=1..N_p} Q_i \psi_i$.

Denote by F the vector $(\langle f, \varphi_1 \rangle, .., \langle f, \varphi_{N_u} \rangle)^T \in \mathbb{R}^{N_u}$ and define the following matrices

$$\mathrm{A} = [a(\varphi_j, \varphi_i)]_{i,j=1..N_u}, \qquad \mathrm{B} = [b(\varphi_j, \psi_i)]_{i=1..N_p, j=1..N_u}, \tag{3.11}$$

where the index i indicates the rows and j the columns. Assuming for simplicity that $g = 0$, problem (3.3) takes the following algebraic form

Find $(\mathrm{U}, \mathrm{P}) \in \mathbb{R}^{\mathrm{N_u}} \times \mathbb{R}^{\mathrm{N_p}}$ *such that*

$$\begin{bmatrix} \mathrm{A} & \mathrm{B^T} \\ \mathrm{B} & 0 \end{bmatrix} \begin{bmatrix} \mathrm{U} \\ \mathrm{P} \end{bmatrix} = \begin{bmatrix} \mathrm{F} \\ 0 \end{bmatrix}. \tag{3.12}$$

This problem is sometimes referred to as the *primal problem*. If the unknown U is eliminated from system (3.12), we obtain the so-called *dual problem*[2]

$$\left(\mathrm{BA^{-1}B^T}\right)\mathrm{P} = \left(\mathrm{BA^{-1}}\right)\mathrm{F}. \tag{3.13}$$

Once the dual problem is solved, U can be recovered by solving

$$\mathrm{AU} = \mathrm{F} - \mathrm{B^T P}.$$

Let

$$\mathrm{S} = \begin{bmatrix} \mathrm{A} & \mathrm{B^T} \\ \mathrm{B} & 0 \end{bmatrix}.$$

The following proposition summarizes important properties of the matrices introduced above.

Proposition 3.1
Assume that $a(\cdot,\cdot)$ is symmetric and coercive on $X_h \times X_h$ and that an inf-sup condition holds on $X_h \times M_h$. Then,

(i) Matrix A is symmetric positive definite.

(ii) Matrix B^T is injective (thus B has full column rank).

(iii) Matrix $\mathrm{BA^{-1}B^T}$ is symmetric positive definite.

(iv) Matrix S is symmetric, invertible, non-definite. More precisely it has N_u positive and N_p negative eigenvalues.

[2] The matrix $\mathrm{BA^{-1}B^T}$ is the *Schur complement* with respect to P.

Proof.

(i) This is an immediate consequence of the symmetry and coercivity of $a(\cdot,\cdot)$.

(ii) From the satisfaction of the inf-sup condition, it follows that

$$\forall q_h \in M_h, \sup_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X} \geq \beta_h \|q_h\|_M,$$

which after introducing the vectors $\mathrm{V} \in \mathbb{R}^{N_u}$ and $\mathrm{Q} \in \mathbb{R}^{N_p}$ representing v_h and q_h gives

$$\forall \mathrm{Q} \in \mathbb{R}^{N_p}, \sup_{\mathrm{V} \in \mathbb{R}^{N_u}} \frac{(\mathrm{BV}, \mathrm{Q})}{\|\mathrm{V}\|_*} \geq \beta_h \|\mathrm{Q}\|_M.$$

Thus, using the above definition of $\|\cdot\|_*$, one can write

$$\forall \mathrm{Q} \in \mathbb{R}^{N_p}, \|\mathrm{B}^T\mathrm{Q}\|_* \geq \beta_h \|\mathrm{Q}\|_M,$$

which shows that B^T is indeed injective (and therefore B has full column rank).

(iii) This property is the immediate consequence of the two previous ones.

(iv) Notice that

$$\mathrm{S} = \begin{bmatrix} \mathrm{A} & \mathrm{B}^{\mathrm{T}} \\ \mathrm{B} & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} \mathrm{A} & 0 \\ \mathrm{B} & \mathrm{I} \end{bmatrix}}_{\mathrm{P}} \underbrace{\begin{bmatrix} \mathrm{A}^{-1} & 0 \\ 0 & -\mathrm{BA}^{-1}\mathrm{B}^{\mathrm{T}} \end{bmatrix}}_{\widetilde{\mathrm{S}}} \underbrace{\begin{bmatrix} \mathrm{A} & \mathrm{B}^{\mathrm{T}} \\ 0 & \mathrm{I} \end{bmatrix}}_{\mathrm{P}^{\mathrm{T}}}$$

Since $\det \mathrm{P} = \det \mathrm{A} \neq 0$, it follows that P is non singular. Thus S and $\widetilde{\mathrm{S}}$ are the representation of the same quadratic form in two different basis. From the Sylvester inertia theorem, one knows that the signature of a matrix (*i.e.* the number of positive, negative and zero eigenvalues) is independent of the basis in which it is written. Since $\widetilde{\mathrm{S}}$ has N_u positive and N_p negative eigenvalues, it follows that S has N_u positive and N_p negative eigenvalues too. □

Remark 3.2 (Algebraic inf-sup condition.) *In point (ii) of Proposition 3.1, it was shown that the inf-sup condition on $X_h \times M_h$ yields*

$$Ker\,\mathrm{B}^{\mathrm{T}} = \{0\}.$$

Conversely, if $Ker\,\mathrm{B}^{\mathrm{T}} = \{0\}$, for all $\mathrm{Q} \in \mathbb{R}^{\mathrm{N_p}}$ of norm 1, $\mathrm{B}^{\mathrm{T}}\mathrm{Q} \neq 0$. The application $\mathrm{Q} \rightarrow \mathrm{B}^{\mathrm{T}}\mathrm{Q}$ is continuous and does not vanish on the unit sphere. In finite dimension, the unit sphere being compact, there exists β_h such that for all $\mathrm{Q} \in \mathbb{R}^{N_p}$ with $\|Q\|_M = 1$,

$$\|\mathrm{B}^{\mathrm{T}}\mathrm{Q}\|_* \geq \beta_h,$$

and hence the inf-sup condition. It is very convenient to realize that, at the discrete level, stating the inf-sup condition is equivalent to stating that matrix B^T is injective (or matrix B has full column rank). Nevertheless, note that because the constant β_h depends a priori *on h, the convergence rate may be non optimal.*

Remark 3.3 (Spurious mode.) *If the inf-sup condition does not hold, Ker* $\mathrm{B}^{\mathrm{T}} \neq \{0\}$*, i.e. there exists* $\mathrm{P}^* \in \mathbb{R}^{\mathrm{N_p}} - \{0\}$ *such that* $\mathrm{B}^{\mathrm{T}}\mathrm{P}^* = 0$*. Thus, the linear systems (3.12) and (3.13) are ill-posed. Indeed, if* (U, P) *is a solution of (3.12), one can build an infinite number of solutions* $(\mathrm{U}, \mathrm{P} + k\mathrm{P}^*)$*,* $k \in \mathbb{R}$*. Such a vector* P^* *is often called a* spurious mode *(or a* spurious pressure *in the context of the hydrodynamic equations). In such a case, the approximation by* X_h *and* M_h *is said to be* unstable.

Dual problem. Typically, it is more convenient to solve a symmetric positive definite problem than a non-definite one. This motivates the solution of the dual problem instead of the primal one. Since $\mathrm{BA}^{-1}\mathrm{B}^{\mathrm{T}}$ is symmetric positive definite, the dual problem is often solved by a gradient based iterative method such as the *Uzawa algorithm*. The convergence properties of this iterative method are related to the condition number of the matrix $\mathrm{B}\ \mathrm{A}^{-1}\ \mathrm{B}^T$ which can be evaluated using the following result.

Lemma 3.3 *For all* $\mathrm{Q} \in \mathbb{R}^{N_p}$*,*

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q}, \mathrm{Q}) = \sup_{\mathrm{V} \in \mathbb{R}^{N_u}} \frac{(\mathrm{BV}, \mathrm{Q})^2}{(\mathrm{AV}, \mathrm{V})}.$$

Proof. Indeed,

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q}, \mathrm{Q}) = (\mathrm{B}^T\mathrm{Q}, \mathrm{A}^{-1}\mathrm{B}^T\mathrm{Q}) = (\mathrm{AV}, \mathrm{V}), \quad \text{where} \quad \mathrm{V} = \mathrm{A}^{-1}\mathrm{B}^T\mathrm{Q}.$$

Furthermore, the reader can easily verify that

$$(\mathrm{AV}, \mathrm{V}) = \sup_{\mathrm{W} \in \mathbb{R}^{N_u}} \frac{(\mathrm{AV}, \mathrm{W})^2}{(\mathrm{AW}, \mathrm{W})},$$

for example, by applying the Cauchy-Schwarz inequality using the scalar product $(\mathrm{A}\cdot, \cdot)$. Thus,

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q}, \mathrm{Q}) = \sup_{\mathrm{W} \in \mathbb{R}^{N_u}} \frac{(\mathrm{AV}, \mathrm{W})^2}{(\mathrm{AW}, \mathrm{W})} = \sup_{\mathrm{W} \in \mathbb{R}^{N_u}} \frac{(\mathrm{B}^T\mathrm{Q}, \mathrm{W})^2}{(\mathrm{AW}, \mathrm{W})}.$$

□

Proposition 3.2
Denote by α *the coercivity constant of* $a(\cdot, \cdot)$ *on* $X \times X$ *and by* β *the inf-sup constant of* $b(\cdot, \cdot)$ *on* $X \times M$*. Let* M_p *denote the mass matrice on the space* M *— that is,* $\mathrm{M}_p = [(\psi_j, \psi_i)]_{i,j=1..N_p}$*. Then,*

$$\kappa(\mathrm{M}_p^{-1}\mathrm{BA}^{-1}\mathrm{B}^T) \leq \frac{\|a\|\|b\|^2}{\alpha\beta^2}.$$

Proof. From Lemma 3.3, it follows that

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q},\mathrm{Q}) = \sup_{\mathrm{V}\in\mathbb{R}^{N_u}} \frac{(\mathrm{BV},\mathrm{Q})^2}{(\mathrm{AV},\mathrm{V})} = \sup_{\mathrm{V}\in\mathbb{R}^{N_u}} \frac{(\mathrm{BV},\mathrm{Q})^2}{\|v_h\|_X^2}\frac{\|v_h\|_X^2}{(\mathrm{AV},\mathrm{V})}.$$

From the continuity of $a(\cdot,\cdot)$, it follows that

$$(\mathrm{AV},\mathrm{V}) = a(v_h,v_h) \leq \|a\|\|v_h\|_X^2,$$

and from the inf-sup condition, one concludes that

$$\sup_{\mathrm{V}\in\mathbb{R}^{N_u}} \frac{(\mathrm{BV},\mathrm{Q})^2}{\|v_h\|_X^2} = \sup_{v_h\in X} \frac{b(v_h,q_h)^2}{\|v_h\|_X^2} \geq \beta^2\|q_h\|_M^2 = \beta^2(\mathrm{M}_p\mathrm{Q},\mathrm{Q}).$$

Hence,

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q},\mathrm{Q}) \geq \frac{\beta^2}{\|a\|}(\mathrm{M}_p\mathrm{Q},\mathrm{Q}).$$

Furthermore, from the coercivity of $a(\cdot,\cdot)$ and the continuity of $b(\cdot,\cdot)$, it follows that

$$(\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q},\mathrm{Q}) = \sup_{\mathrm{V}\in\mathbb{R}^{N_u}} \frac{(\mathrm{BV},\mathrm{Q})^2}{\|v_h\|_X^2}\frac{\|v_h\|_X^2}{(\mathrm{AV},\mathrm{V})} \leq \frac{\|b\|^2}{\alpha}\|q_h\|_M^2 = \frac{\|b\|^2}{\alpha}(\mathrm{M}_p\mathrm{Q},\mathrm{Q}).$$

Therefore,

$$\frac{\beta^2}{\|a\|}(\mathrm{M}_p\mathrm{Q},\mathrm{Q}) \leq (\mathrm{BA}^{-1}\mathrm{B}^T\mathrm{Q},\mathrm{Q}) \leq \frac{\|b\|^2}{\alpha}(\mathrm{M}_p\mathrm{Q},\mathrm{Q}),$$

which, given Proposition 1.3, p. 14, concludes the proof. □

It is interesting to compare the result of Proposition 3.2 with the results of Section 1.2.3, p. 13: whereas the condition number of the stiffness matrix typically grows as $1/h^2$, Proposition 3.2 shows that the condition number of the dual problem is essentially independent of h (as long as α and β are independent of h). This explains why the Uzawa method is popular for solving the algebraic systems of equations arising from mixed finite element methods. Nevertheless, it may happen that the ratio "continuity over coercivity constants" be very large (for example, when using a small time step for solving a transient problem). In such a case, better preconditioners than the mass matrix need be developed.

3.3 Finite Element for the Stokes problem

In this section, a few example of stable finite element spaces are discussed for the solution of the Stokes problem.

3.3.1 A result to prove the inf-sup condition

The following result is sometimes useful for proving that a proposed finite element pair of spaces satisfies the inf-sup condition.

Theorem 3.2 (Fortin's Lemma.)
Assume the inf-sup condition holds on $X \times M$ — that is,

$$\inf_{q \in M} \sup_{v \in X} \frac{b(v,q)}{\|v\|_X \|q\|_M} \geq \beta.$$

Let $X_h \subset X$ and $M_h \subset M$. An inf-sup condition holds on $X_h \times M_h$ with a constant β^* independent of h if and only if there exists a restriction operator $\Pi_h \in \mathcal{L}(X, X_h)$ and a constant $C > 0$ independent of h such that

(i) $\forall v \in X, \quad b(\Pi_h v - v, q_h) = 0, \forall q_h \in M_h.$

(ii) $\forall v \in X, \quad \|\Pi_h v\|_X \leq C\|v\|_X.$

Proof.

$\boxed{\Leftarrow}$ For $q_h \in M_h$,

$$\begin{aligned} \sup\nolimits_{v_h \in X_h} \frac{b(v_h, q_h)}{\|v_h\|_X} &\geq \sup\nolimits_{v \in X} \frac{b(\Pi_h v, q_h)}{\|\Pi_h v\|_X} \\ &\geq \frac{1}{C} \sup_{v \in X} \frac{b(v, q_h)}{\|v\|_X} \\ &\geq \frac{\beta}{C} \|q_h\|_M. \end{aligned}$$

Hence, $\beta^* = \beta/C$.

$\boxed{\Rightarrow}$ Suppose that the inf-sup condition holds in $X_h \times M_h$ with a constant β^*. In this case, the operator $B_h : X_h \to M_h'$ is an isomorphism from $V_h^\perp$ onto M_h' and for the *same constant* β^*, $\|B_h v_h\|_M \geq \beta^* \|v_h\|_X$. Let $v \in X$. Since the application $q_h \to b(v, q_h)$ is in M_h', there exists a unique element in X_h, denoted by $\Pi_h v$, such that

$$b(\Pi_h v, q_h) = b(v, q_h),$$

and

$$\begin{aligned} \|\Pi_h v\|_X &\leq \frac{1}{\beta^*}\|B_h \Pi_h v\|_{M'} = \frac{1}{\beta^*}\|B_h v\|_{M'} = \frac{1}{\beta^*} \sup_{q_h \in M_h} \frac{b(v, q_h)}{\|q_h\|_M} \\ &\leq \frac{\|b\|}{\beta^*}\|v\|_X, \end{aligned}$$

hence the result, with $C = \|b\|/\beta^*$.□

Remark 3.4 *The operator Π_h is often searched for as the sum of two operators Π_1 and Π_2 where Π_1 satisfies*

$$\|\Pi_1 v - v\|_X \leq Ch^s \|v\|_X$$

and Π_2 is an operator which is "locally" built (for example, with a bubble function) in order to satisfy point (i) of Proposition 3.2. See [4], p. 60.

When $X = H_0^1(\Omega)^d$ (for example, for the Stokes problem), it may be difficult to build the operator Π_1 mentioned in the previous Remark. Indeed, the classical interpolation operator cannot be used since it typically supposes that v is continuous. A projection operator could be used instead, but a convexity hypothesis must be made in order to apply the Aubin-Nitsche theorem. The best solution is to consider the Clement operator (see [6], Lemma 1.127, p. 60).

Theorem 3.3 (Clément.)
If the mesh family is shape-regular, there exists $C > 0$ such that $\forall v \in H_0^1(\Omega)^d$, $\exists R_h(v) \in X_h^1$ such that $\forall K \in \mathcal{T}_h$, $0 \leq l \leq 1$,

$$\|R_h(v) - v\|_{l,K} \leq Ch_K^{1-l} \|v\|_{1,\triangle_K} \tag{3.14}$$

where $\triangle_K = \cup_{\bar{K}' \cap \bar{K} \neq \emptyset} K'$.

3.3.2 Mixed finite element for the Stokes problem

Section 3.1 has provided the theoretical results needed for analyzing the convergence of a Galerkin discretization of Problem (3.1). In this section, several finite element spaces adapted to the discretization of the Stokes equations (1.37) are presented. The reader is referred to Section 1.2.2 for the notation and the basic results on finite elements.

Consider a shape-regular family of simplicial meshes $(\mathcal{T}_h)_{h>0}$. Denote by X_h^k the Lagrange finite element space of degree k built on $\mathcal{T}_h$.

For example, the spaces X_h and M_h can be defined as $X_h = (X_h^k)^3 \cap H_0^1(\Omega)^3$ and $M_h = X_h^s \cap L_0^2(\Omega)$. In such a case, one says that the problem is solved with the $\mathbb{P}_r/\mathbb{P}_s$ finite element, meaning that the velocity (respectively, the pressure) is approximated in a space of continuous piecewise polynomials of degree r (respectively, s). As mentioned above, the main difficulty is to find spaces X_h and M_h that satisfy the inf-sup condition

$$\exists\, \beta_h > 0, \inf_{q_h \in M_h} \sup_{\boldsymbol{v}_h \in X_h} \frac{\displaystyle\int_\Omega q_h \operatorname{div} \boldsymbol{v}_h}{\|\boldsymbol{v}_h\|_1 \|q_h\|_0} \geq \beta_h. \tag{3.15}$$

For example, the pairs $\mathbb{P}_r/\mathbb{P}_r$, $r \geq 1$, or $\mathbb{P}_1/\mathbb{P}_0$ are known to be *unstable*. This means that the spaces X_h and M_h built on these finite elements do not satisfy the inf-sup condition (3.15), and therefore the resulting linear system is singular.

A simple example of a stable pair using the standard spaces X_h^k is the $\mathbb{P}_2/\mathbb{P}_1$ known as the Taylor-Hood finite element (Figure 3.1). The inf-sup constant β_h of this element is independent of h which ensures an optimal convergence rate. The following result can be proved (see for example [6], section 4.2.5).

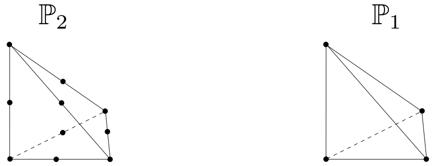

Figure 3.1: The Taylor-Hood ($\mathbb{P}_2/\mathbb{P}_1$) finite element.

Proposition 3.3
Assume that the solution of the Stokes problem satisfies $\boldsymbol{u} \in \left(H^3(\Omega) \cap H_0^1(\Omega)\right)^3$ and $p \in H^2(\Omega) \cap L_0^2(\Omega)$. Assume moreover that each tetrahedron of the mesh has at least three edges within Ω. Then, there exists $c > 0$ such that the solution $(\boldsymbol{u}_h, p_h)$ computed with the $\mathbb{P}_2/\mathbb{P}_1$ finite element satisfies for all $h > 0$

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \leq ch^2(\|\boldsymbol{u}\|_{3,\Omega} + \|p\|_{2,\Omega}).$$

Another popular element is the so-called $\mathbb{P}_1$-bubble/$\mathbb{P}_1$ pair, also known as the *mini-element*. It consists of adding to the $\mathbb{P}_1/\mathbb{P}_1$ element one degree of freedom for each component of the velocity on the barycenters of the tetrahedra (Figure 3.2). Let $\hat{b} \in H^1(\hat{K})$ denote a function which takes the value 1 at the barycenter of the reference $\hat{K}$, vanishes on its boundary $\partial\hat{K}$ and verifies $0 \leq \hat{b} \leq 1$. Such a function is known as a “bubble function”. Define then the space

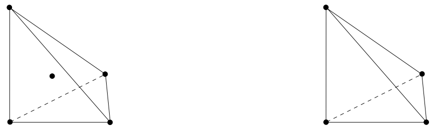

Figure 3.2: The mini ($\mathbb{P}_1$-bubble/$\mathbb{P}_1$) finite element.

$$\mathcal{P}_{1,h}^b = \left\{ v_h \in \mathcal{C}^0(\bar{\Omega}), v_h \circ \mathcal{F}_K \in \mathbb{P}_1(\hat{K}) \oplus \text{span}\{\hat{b}\}, \forall\, K \in \mathcal{T}_h \right\},$$

where $\mathcal{F}_K$ is the application that maps the reference element $\hat{K}$ on an element K of $\mathcal{T}_h$ (see Section 1.2.2). The Stokes problem is said to be solved with the mini-element, or the $\mathbb{P}^1$-bubble/$\mathbb{P}^1$ finite element, when one uses the spaces $X_h = (\mathcal{P}_{1,h}^b)^3 \cap H_0^1(\Omega)^3$ to approximate the velocity and $M_h = X_h^1 \cap L_0^2(\Omega)$ to approximate the pressure. As for the Taylor-Hood element, the inf-sup constant β_h of the mini-element is independent of h which ensures an optimal convergence rate. Then, the following result holds (see for example [6], section 4.2.4).

Proposition 3.4
Assume that the solution of the Stokes problem satisfies $\boldsymbol{u} \in \left(H^2(\Omega) \cap H^1_0(\Omega)\right)^3$ *and* $p \in H^1(\Omega) \cap L^2_0(\Omega)$. *Then, there exists* $c > 0$ *such that the solution* $(\boldsymbol{u}_h, p_h)$ *computed with the* $\mathbb{P}_1$*-bubble/* $\mathbb{P}_1$ *finite element satisfies for all* $h > 0$

$$\|\boldsymbol{u} - \boldsymbol{u}_h\|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \leq ch(\|\boldsymbol{u}\|_{2,\Omega} + \|p\|_{1,\Omega}).$$

The proof of this Proposition is based on Proposition 3.2. Many other examples of stable pairs can be found in the literature. The interested reader is referred to V. Girault and P.A. Raviart [7, Chapter 2], F. Brezzi and M. Fortin [4, Chapter 4] or A. Ern and J.-L. Guermond [6, Chapter 4].

3.4 Locking phenomena

Consider the elasticity problem (1.39) introduced in Section 1.3.3, a finite element space $X_h \subset (H^1_0(\Omega))^d$, and the search for $\boldsymbol{u}_h \in X_h$ such that

$$2G\int_\Omega \boldsymbol{\epsilon}^D(\boldsymbol{u}_h) : \boldsymbol{\epsilon}^D(\boldsymbol{v}_h)\,dx + \kappa\int_\Omega \operatorname{div}\boldsymbol{u}_h \operatorname{div}\boldsymbol{v}_h\,dx = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h\,dx. \tag{3.16}$$

When κ is very large (which corresponds to an almost incompressible material), results of poor quality are obtained when solving this equation. More specifically, it is observed that the material deforms as if it were much stiffer. In other words, it appears to “lock” (and hence the name of *locking* for describing this phenomenon). Here, it is explained why a mixed method satisfying the inf-sup is a good remedy for this phenomenon.

First, a heuristic explanation of locking is presented. To this effect, the following space is introduced

$$\widetilde{V}_h(q_h) = \{\boldsymbol{v}_h \in X_h, \operatorname{div}\boldsymbol{v}_h = q_h\} \qquad \text{and} \qquad \widetilde{V}_h = \widetilde{V}_h(0).$$

Formally, one sees in (3.16) that in the limit $\kappa = \infty$, $\operatorname{div}\boldsymbol{u}_h = 0$ (divide the equation by κ and let κ goes to ∞). Thus, the solution $\boldsymbol{u}_h$ is constrained to lie in space $\widetilde{V}_h$. Therefore, instead of being controlled by

$$\inf_{\boldsymbol{v}_h \in X_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1,$$

the error is actually controlled by

$$\inf_{\boldsymbol{v}_h \in \widetilde{V}_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1.$$

Whereas the approximation properties of X_h are usually well-known (standard finite element space, see for example (1.26) and (1.27)), the approximation properties of

$\widetilde{V}_h$ are less clear and can be very poor. The extreme case is when $\tilde{V}_h$ is reduced to $\{0\}$: the elastic solid is then completely stuck! This explains the locking problem. This phenomenon would not occur in the presence of an inequality such as

$$\inf_{\boldsymbol{v}_h \in \widetilde{V}_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1 \leq C \inf_{\boldsymbol{v}_h \in X_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1, \tag{3.17}$$

where C is a constant independent of h. Indeed in such a case, the approximation properties of $\widetilde{V}_h$ would be the same as those of X_h. But inequality (3.17) is not true in general.

To avoid locking, it has been proposed in the engineering literature to slightly modify the energy of the problem as follows

$$J_h(\boldsymbol{v}_h) = G\int_\Omega |\boldsymbol{\epsilon}^D(\boldsymbol{v}_h)|^2\,dx + \frac{\kappa}{2}\int_\Omega (P_h(\operatorname{div}\boldsymbol{v}_h))^2\,dx - \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h\,dx, \tag{3.18}$$

where P_h is the L^2 projector onto another finite dimensional space M_h (to be determined). Let $p \in L^2(\Omega)$. The reader is reminded that $P_h(p)$ is characterized by $P_h(p) \in M_h$ and

$$\int_\Omega q_h P_h(p)\,dx = \int_\Omega q_h p\,dx, \qquad \forall q_h \in M_h.$$

Minimizing J_h over X_h is equivalent to solving

$$2G\int_\Omega \boldsymbol{\epsilon}^D(\boldsymbol{u}_h) : \boldsymbol{\epsilon}^D(\boldsymbol{v}_h)\,dx + \kappa\int_\Omega P_h(\operatorname{div}\boldsymbol{u}_h)P_h(\operatorname{div}\boldsymbol{v}_h)\,dx = \int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h\,dx.$$

Introducing $p_h = -\kappa P_h(\operatorname{div}\boldsymbol{u}_h) \in M_h$ leads to

$$\kappa\int_\Omega P_h(\operatorname{div}\boldsymbol{u}_h)P_h(\operatorname{div}\boldsymbol{v}_h)\,dx = -\int_\Omega p_h P_h(\operatorname{div}\boldsymbol{v}_h)\,dx = -\int_\Omega p_h \operatorname{div}\boldsymbol{v}_h\,dx.$$

Thus, minimizing the modified energy J_h is equivalent to searching for $(\boldsymbol{u}_h, q_h) \in X_h \times M_h$ such that for all $(\boldsymbol{v}_h, q_h) \in X_h \times M_h$,

$$\left\{\begin{array}{rcl} 2G\displaystyle\int_\Omega \boldsymbol{\epsilon}^D(\boldsymbol{u}_h) : \boldsymbol{\epsilon}^D(\boldsymbol{v}_h)\,dx - \int_\Omega p_h \operatorname{div}\boldsymbol{v}_h\,dx & = & \displaystyle\int_\Omega \boldsymbol{f}\cdot\boldsymbol{v}_h\,dx, \\ \displaystyle\int_\Omega q_h \operatorname{div}\boldsymbol{u}_h\,dx + \frac{1}{\kappa}\int_\Omega q_h p_h\,dx & = & 0. \end{array}\right. \tag{3.19}$$

Define the operator B_h as for the Stokes equation by $\langle B_h v_h, q_h\rangle = b(v_h, q_h)$ with

$$b(\boldsymbol{v}_h, q_h) = -\int_\Omega q_h \operatorname{div}\boldsymbol{v}_h.$$

Define also the set $V_h(g)$ as in (3.9) and the space $V_h = V_h(0) = \operatorname{Ker} B_h$ as in (3.6). Since $\int_\Omega q_h \operatorname{div}\boldsymbol{v}_h\,dx = \int_\Omega q_h P_h(\operatorname{div}\boldsymbol{v}_h)\,dx$, note that

$$B_h = P_h(\operatorname{div}\boldsymbol{v}_h).$$

Thus, the space V_h can also be defined as

$$V_h = \{\boldsymbol{v}_h \in X_h / P_h(\operatorname{div} \boldsymbol{v}_h) = 0\}.$$

Under this form, the difference with the space $\widetilde{V}_h$ is now clear. Going back to the heuristic argument, one sees that when κ goes to infinity, the solution $\boldsymbol{u}_h$ to problem (3.19) is constrained to lie in V_h (instead of $\widetilde{V}_h$). Whereas $\widetilde{V}_h$ was a "hidden" and not very convenient space, the space V_h is linked to the choice of space M_h. To choose M_h, a trade-off has to be found between locking and accuracy. On one-hand, a smaller M_h makes V_h larger and thus, an inequality like (3.17) easier to obtain, which avoids locking but enforces poorly the incompressibility constraint. On the other-hand, a larger M_h enforces better the incompressibility constraint but leads to a smaller V_h and therefore is likely to introduce locking $\big($since inequality (3.17) is more difficult to achieve with a small $V_h\big)$. The heuristic is now clear. However, it does not help choosing M_h. The inf-sup condition does. Indeed, it has been proved in Lemma 3.2 that if M_h is chosen so that the inf-sup condition holds with a constant independent of h, then inequality (3.10) holds, *i.e.*

$$\inf_{\boldsymbol{v}_h \in V_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1 \leq C \inf_{\boldsymbol{v}_h \in X_h} \|\boldsymbol{u} - \boldsymbol{v}_h\|_1.$$

In other words, whereas inequality (3.17) is not true in general, it is actually verified as soon as the energy J is replaced by J_h $\big($defined in (3.18)$\big)$ with a projector P_h on a space M_h such that the inf-sup condition holds on $X_h \times M_h$. This explains why the introduction of a well-chosen projector in the energy can indeed be a good remedy to the locking phenomenon.

Remark 3.5 *Locking is not restricted to quasi-incompressible materials. It is also encountered, for example, in thin plates and shell problems.*