COMPUTER METHODS IN APPLIED MECHANICS AND ENGINEERING 82 (1990) 27–57
NORTH-HOLLAND

A DISCOURSE ON THE STABILITY CONDITIONS FOR MIXED FINITE ELEMENT FORMULATIONS

Franco BREZZI
Dipartimento di Meccanica Strutturale and Istituto di Analisi, Numerica del C.N.R., 27100 Pavia, Italy

Klaus-Jürgen BATHE
Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA 02139, U.S.A.

Received 13 February 1990

We discuss the general mathematical conditions for solvability, stability and optimal error bounds of mixed finite element discretizations. Our objective is to present these conditions with relatively simple arguments. We present the conditions for solvability and stability by considering the general coefficient matrix of mixed finite element discretizations, and then deduce the conditions for optimal error bounds for the distance between the finite element solutions and the exact solution of the mathematical problem. To exemplify our presentation we consider the solutions of various example problems. Finally, we also present a numerical test that is useful to identify numerically whether, for the solution of the general Stokes flow problem, a given finite element discretization satisfies the stability and optimal error bound conditions.

1. Introduction

During the recent years it has been recognized to an increasing extent that the use of mixed finite elements can be of great benefit and may even be necessary to obtain reliable and accurate solutions in certain fields of engineering analysis. Mixed finite elements are currently used with much success in the solution of incompressible fluid flows, and continue to provide great promise for the analysis of solids and structures [1, 2].

Of course, the largest area of finite element applications is still structural analysis and mixed finite elements are, in principle, much suited for use in the analysis of almost incompressible media (for example, for the analysis of rubber-like materials, elasto-plasticity and creep) and the analysis of plates and shells. However, although many mixed finite elements have been proposed over the last two decades in the research literature, it is apparent that mixed finite elements are hardly used in practical structural analysis.

The reason why mixed finite elements are not used abundantly in engineering practice is that their predictive behavior is much more difficult to assess than for the conventional and commonly used displacement-based elements. Whereas displacement-based elements, once formulated and shown to work well on certain sets of examples (including the patch tests), can be generally employed, mixed finite elements cannot be recommended for general use unless a deeper analysis and understanding is available. Namely, considering a certain category of

problems, a mixed finite element may work well in the solution of certain problems but perform very poorly on other problems. Therefore, a mathematical analysis (even a limited one) for the stability and convergence of a proposed formulation is an important requirement. Such mathematical analysis should give sufficient insight as to the general applicability of the finite element under consideration, and is in general no easy task.

Some researchers have proposed some easily applied 'counting rules' to assess whether a mixed finite element can be recommended [3, 4]. However, such rules can at best give some guide-lines and do not give the necessary information to assess whether an element is stable and accurate.

Considering mixed finite element discretizations, we recognize that they are governed by a system of equations with a coefficient matrix C, that we may write as

$$C = \begin{bmatrix} A & B^{\mathrm{t}} \\ B & 0 \end{bmatrix}. \tag{1.1}$$

We quote as a main example the analysis of incompressible fluid flow, the Stokes problem, when using the velocity–pressure formulation. Other important examples of interest are the analysis of incompressible solids and the analysis of plates and shells. In principle, many solutions can be formulated using a mixed or hybrid method that results into the coefficient matrix (1.1), because this matrix is reached by minimizing a functional under linear constraints [1].

The general mathematical theory for the solution of problems that are governed by the coefficient matrix in (1.1) is now quite well established and the detailed applications of this theory to a number of important problem categories is available. We know necessary and sufficient conditions for the existence and uniqueness of the solution, both for the continuous and the discretized problems. We also know necessary and sufficient conditions on the choice of the discretizations in order to have optimal error bounds [1, 5]. This information is most valuable for the design *and* analysis of mixed finite elements because the basic mathematical results are quite generally applicable (while the detailed application to problem areas may of course not be straight-forward).

Our objective in this paper is two-fold. The first aim is to present the general mathematical results quoted above with relatively simple arguments. For this purpose we consider the general coefficient matrix of mixed finite element formulations and deduce the conditions of solvability and stability. In proceeding this way, we refer to the continuous problem only when necessary (since the treatment of the continuous problem requires a background in functional analysis) and we concentrate on the discretized (finite-dimensional) problem. However, we succeed in pointing out the basic mathematical conditions on the discretization and in showing that they are necessary to have stability and optimal error estimates.

Our second aim in this paper is to propose a simple numerical procedure for checking whether the above mathematical conditions are satisfied for a given mixed finite element formulation. Such a procedure is useful because it may be employed to check a formulation and its computer program implementation (much like the patch test is used for incompatible displacement-based finite element formulations). We consider in this discussion the analysis of incompressible fluid flow and our test is closely related to 'Fortin's trick' to identify whether the mathematical conditions of stability and optimal error bounds are satisfied.

The paper is organized into the following sections. In Section 2 we recall some basic properties of square matrix systems and introduce the basic concepts of stability and optimality. In Section 3 we then deal with the special case of systems of the form (1.1); hence here we focus onto the analysis of mixed finite element formulations in detail. Finally, in Section 4 we discuss two applications and introduce our test for checking the good quality of a given discretization, using as an example the case of an incompressible fluid. We then conclude our presentation in Section 5.

2. Some preliminaries and the general problem of solvability and stability

Let us consider the general case of an $N \times N$ matrix M and the associated system

$$\text{Given } b \in \mathbb{R}^N \quad \text{find } x \in \mathbb{R}^N \quad \text{such that} \quad Mx = b\,. \tag{2.1}$$

The following theorem is a well-known cornerstone in linear algebra.

THEOREM 2.1. Problem (2.1) has a unique solution for every given right-hand side b, if and only if the associated homogeneous system $Mx = 0$ has only the solution $x = 0$.

In other words, in order to have a solvable problem in (2.1) for every possible $b \in \mathbb{R}^N$ we need the following condition to hold:

$$\text{if } Mx = 0\,, \quad \text{then } x = 0\,. \tag{2.2}$$

Condition (2.2) answers the problem of the solvability of (2.1) but not of its stability. Roughly speaking, we would like that a small change in b determines only a small change in x. However, in order to measure the magnitude of such change we have to introduce norms. Assume that we choose a norm $\| \ \|_L$ for measuring the size of solutions and a norm $\| \ \|_R$ for right-hand sides. In principle, we are allowed to choose the same norm for both, but we shall see that this, in general, is not the most convenient choice. We also point out explicitly that, in finite-dimensional spaces, all norms are equivalent in the sense that for any two norms $\| \ \|_{s_1}$ and $\| \ \|_{s_2}$ in $\mathbb{R}^N$ there exist two positive constants s_1 and s_2 such that

$$\|v\|_{s_1} \leq s_1 \|v\|_{s_2}\,, \tag{2.3}$$

$$\|v\|_{s_2} \leq s_2 \|v\|_{s_1} \tag{2.4}$$

for every vector v in $\mathbb{R}^N$. However, these constants s_1 and s_2 will, in general, depend on the dimension N.

EXAMPLE. This is a very simple example, only used to fix our ideas [2]. Let

$$\|v\|_{s_1} := \max_i |v_i| = \|v\|_{l_\infty}\,, \tag{2.5}$$

$$\|v\|_{s_2} := \sum_i |v_i| = \|v\|_{l_1}\,, \tag{2.6}$$

then it is easy to see that

$$\max_i |v_i| \leq \sum_i |v_i| , \tag{2.7}$$

$$\sum_i |v_i| \leq N \max_i |v_i| , \tag{2.8}$$

so that $s_1 = 1$ and $s_2 = N$. Similarly we have for the Euclidean norm,

$$\|v\|_E := \left(\sum_i |v_i|^2\right)^{1/2} = \|v\|_{l_2} , \tag{2.9}$$

that

$$\|v\|_{l_1} \leq \sqrt{N}\|v\|_{l_2} , \qquad \|v\|_{l_2} \leq \sqrt{N}\|v\|_{l_\infty} . \tag{2.10}$$

We have seen that the choice of one norm or another can change, asymptotically, the dependence on N of the various constants. We shall come back to this point with useful guidelines for the most convenient choices. For the moment, we assume that the choice of $\| \ \|_L$ and $\| \ \|_R$ has been performed and define stability in terms of these norms.

DEFINITION. Let M be a non-singular $N \times N$ matrix. We define the stability constant of M with respect to the norms $\| \ \|_L$ and $\| \ \|_R$ as the smallest possible constant S_{LR} such that

$$\frac{\|\delta x\|_L}{\|x\|_L} \leq S_{LR} \frac{\|\delta b\|_R}{\|b\|_R} \tag{2.11}$$

for all vectors x and δx in $\mathbb{R}^N$ with $Mx =: b$ and $M\,\delta x =: \delta b$. □

In other words, (2.11) bounds the relative change in x (in the norm L) by means of the relative change in the right-hand side b (in the norm R). We point out that such a constant S_{LR} always exists. However, if we consider a sequence of problems of type (2.1) with increasing dimension N (corresponding, in general, to a finer and finer finite element mesh) we might find that the corresponding constants S_{LR} depend on N and become infinitely large when $N \to +\infty$. Thus we might say that a sequence of problems of the type (2.1) is stable with respect to the norms $\| \ \|_L$ and $\| \ \|_R$ if the stability constant S_{LR} is uniformly bounded.

We would like now to present stability from a slightly different point of view. For this, let us introduce the matrix norms

$$\|M\|_{LR} = \sup_y \frac{\|My\|_R}{\|y\|_L} \tag{2.12}$$

and

$$\|M^{-1}\|_{RL} = \sup_z \frac{\|M^{-1}z\|_L}{\|z\|_R} . \tag{2.13}$$

From (2.13) for $z = \delta b$ (so that $M^{-1}z = \delta x$) we easily obtain

$$\|M^{-1}\|_{RL} \geq \frac{\|\delta x\|_L}{\|\delta b\|_R} , \tag{2.14}$$

while (2.12), for $y = x$ (and $My = b$), gives

$$\|M\|_{LR} \geq \frac{\|b\|_R}{\|x\|_L} . \tag{2.15}$$

From (2.14) and (2.15) we then have

$$\frac{\|\delta x\|_L}{\|x\|_L} \leq \|M\|_{LR} \cdot \|M^{-1}\|_{RL} \frac{\|\delta b\|_R}{\|b\|_R} , \tag{2.16}$$

from which

$$S_{LR} = \|M\|_{LR} \|M^{-1}\|_{RL} . \tag{2.17}$$

REMARK 2.1. Noting that for every x one has $x = M^{-1}Mx$, we obtain

$$\|x\|_L \leq \|M^{-1}\|_{RL} \|M\|_{LR} \|x\|_L , \tag{2.18}$$

which easily implies

$$S_{LR} = \|M^{-1}\|_{RL} \|M\|_{LR} \geq 1 . \tag{2.19}$$

REMARK 2.2. If we choose $\| \ \|_L = \| \ \|_R = \| \ \|_E$ (Euclidean norm) and if M is symmetric and positive definite, then

$$\|M\|_{LR} = \lambda_{max} , \qquad \|M^{-1}\|_{LR} = 1/\lambda_{min} , \tag{2.20}$$

where λ_{max} and λ_{min} are the maximum and (respectively) minimum eigenvalues of M.

Hence, for the case of M being symmetric and positive definite, we have that

$$S_{LR} = S_{EE} = \frac{\lambda_{max}}{\lambda_{min}} \tag{2.21}$$

coincides with the usual condition number. Note however, that a different choice of norms will (obviously) produce different stability constants. For instance, by taking

$$M = \begin{bmatrix} 2 & 1 \\ 1 & 1 \end{bmatrix} , \qquad \| \ \|_L = \| \ \|_{l_\infty} , \qquad \| \ \|_R = \| \ \|_{l_1} \tag{2.22}$$

(see (2.5) and (2.6) for the definition of the norms $\| \ \|_{l_\infty}$ and $\| \ \|_{l_1}$), we have

$$\lambda_{max} = \tfrac{1}{2}(3 + \sqrt{5}) , \quad \lambda_{min} = \tfrac{1}{2}(3 - \sqrt{5}) , \quad \|M\|_{LR} = 5 , \quad \|M^{-1}\|_{RL} = 1 . \tag{2.23}$$

so that $S_{EE} = (3 + \sqrt{5})/(3 - \sqrt{5})$ (= condition number) while $S_{LR} = 5$. We shall see in the following that for practical problems we have, in a natural way, choices for the norms $\| \ \|_L$ and $\| \ \|_R$ for which S_{LR} will be uniformly bounded while S_{EE} is not.

From (2.17) we see that a sequence of problems will be stable with respect to the norms $\| \ \|_L$ and $\| \ \|_R$ if both $\|M\|_{LR}$ and $\|M^{-1}\|_{RL}$ are uniformly bounded. In the applications it is very easy to find norms $\| \ \|_L$ such that

$$y^t M x \leq k_M \|y\|_L \|x\|_L \quad \forall x, y , \tag{2.24}$$

with k_M uniformly bounded from above and from below. From (2.24) we have a natural choice for $\| \ \|_R$ that produces a uniform bound for $\|M\|_{LR}$. Indeed, if we define the dual norm of $\| \ \|_L$ by

$$\|z\|_{DL} := \sup_y \frac{y^t z}{\|y\|_L} , \tag{2.25}$$

we have the following proposition.

PROPOSITION 2.1. Let M be an $N \times N$ matrix, let $\| \ \|_L$ be a norm in $\mathbb{R}^N$ and let k_M be the smallest possible constant for which (2.24) holds true, that is,

$$k_M = \sup_{x,y} \frac{y^t M x}{\|y\|_L \|x\|_L} . \tag{2.26}$$

If we choose $\| \ \|_R = \| \ \|_{DL}$ (dual norm of $\| \ \|_L$ as defined in (2.25)), then

$$\|M\|_{LR} = k_M . \tag{2.27}$$

PROOF. We have

$$\begin{aligned} \|M\|_{LR} &= \sup_x \frac{\|Mx\|_R}{\|x\|_L} \quad \text{(use (2.12))} \\ &= \sup_x \frac{\|Mx\|_{DL}}{\|x\|_L} \quad (\text{use } \| \ \|_R = \| \ \|_{DL}) \\ &= \sup_x \left\{ \frac{1}{\|x\|_L} \sup_y \frac{y^t M x}{\|y\|_L} \right\} \quad \text{(use (2.25))} \\ &= \sup_{x,y} \frac{y^t M x}{\|x\|_L \|y\|_L} = k_M \quad \text{(use (2.26))} . \ \square \end{aligned} \tag{2.28}$$

If we assume now that we are given a sequence of problems such that

$$y^t M x \leq k_M \|y\|_L \|x\|_L \quad \forall x, y ,$$

with k_M uniformly bounded from above and from below, and if we choose $\| \ \|_R = \| \ \|_{DL}$, then the sequence of problems will be stable with respect to the norms $\| \ \|_L$ and $\| \ \|_R$ if and only if $\|M^{-1}\|_{RL}$ is uniformly bounded. In the following proposition we express $\|M^{-1}\|_{RL}$ in terms of the norm $\| \ \|_L$ alone.

PROPOSITION 2.2. Let M be a non-singular $N \times N$ matrix. Let $\| \ \|_L$ be a norm in $\mathbb{R}^N$ and let $\| \ \|_R$ be the dual norm of $\| \ \|_L$ as defined in (2.25). Then

$$(\|M^{-1}\|_{RL})^{-1} = \inf_x \sup_y \frac{y^t M x}{\|y\|_L \|x\|_L} . \tag{2.29}$$

PROOF. We have

$$\begin{aligned}
(\|M^{-1}\|_{RL})^{-1} &= \left(\sup_z \frac{\|M^{-1}z\|_L}{\|z\|_R}\right)^{-1} = \inf_z \frac{\|z\|_R}{\|M^{-1}z\|_L} \quad \text{(use (2.13))} \\
&= \inf_x \frac{\|Mx\|_R}{\|x\|_L} \quad (\text{set } z = Mx) \\
&= \inf_x \frac{\|Mx\|_{DL}}{\|x\|_L} \quad (\text{use } \| \ \|_R = \| \ \|_{DL}) \\
&= \inf_x \left\{ \frac{1}{\|x\|_L} \sup_y \frac{y^t M x}{\|y\|_L} \right\} \quad \text{(use (2.25))} \\
&= \inf_x \sup_y \frac{y^t M x}{\|x\|_L \|y\|_L} . \quad \square
\end{aligned} \tag{2.30}$$

The following proposition summarizes Propositions 2.1 and 2.2.

PROPOSITION 2.3. Let M be an $N \times N$ non-singular matrix, let $\| \ \|_L$ be a norm in $\mathbb{R}^N$ and let $\| \ \|_R$ be its dual norm as defined in (2.25). Setting

$$k_M = \sup_{x,y} \frac{y^t M x}{\|y\|_L \|x\|_L} , \tag{2.31}$$

$$\gamma_M = \inf_x \sup_y \frac{y^t M x}{\|x\|_L \|y\|_L} , \tag{2.32}$$

the stability constant S_{LR} of M is given by

$$S_{LR} = k_M / \gamma_M . \tag{2.33}$$

The proof is obvious from (2.17), (2.27), (2.29), (2.31) and (2.32).

REMARK 2.3. If we assume to be dealing with a sequence of problems where

$$y^t M x \leq k \|y\|_L \|x\|_L \quad \forall x, y , \tag{2.34}$$

with k uniformly bounded from above, then $k_M \leq k$ and in order to have a uniform bound for S_{LR} we only need γ_M to be uniformly bounded from below, that is, we need a constant $\gamma > 0$

such that

$$\inf_x \sup_y \frac{y^t M x}{\|x\|_L \|y\|_L} \geq \gamma > 0 \tag{2.35}$$

for every problem of the sequence.

REMARK 2.4. We remember that the solvability of (2.1) was expressed in (2.2) by

$$Mx = 0 \quad \Rightarrow \quad x = 0 . \tag{2.36}$$

Under the assumption (2.34) we have now that the stability can be expressed by (2.35) which, in turn, can be written as

$$\exists \gamma > 0 \quad \text{such that} \quad \|Mx\|_{DL} \geq \gamma \|x\|_L \quad \forall x . \tag{2.37}$$

Indeed (2.35) can be written as

$$\exists \gamma > 0 \quad \text{such that} \quad \sup_y \frac{y^t M x}{\|y\|_L} \geq \gamma \|x\|_L \quad \forall x , \tag{2.38}$$

which becomes (2.37) by using the definition of the dual norm (2.25).

We end this section by analyzing the connection of the above results with the use of Galerkin methods for the discretization of variational problems. Let us consider a general linear elasticity problem characterized by a given Hilbert space W and a bilinear form $m(\phi, \psi)$ defined on $W \times W$. Given a linear functional $\beta(\phi)$ from W to $\mathbb{R}$, we want to approximate the solution of the continuous problem

$$\text{Find } \phi \in W \quad \text{such that} \quad m(\phi, \psi) = \beta(\psi) \quad \forall \psi \in W , \tag{2.39}$$

by means of the sequence of finite dimensional problems

$$\text{Find } \phi_h \in W_h \quad \text{such that} \quad m(\phi_h, \psi_h) = \beta(\psi_h) \quad \forall \psi_h \in W_h , \tag{2.40}$$

where W_h is a sequence of finite dimensional subspaces of W_h. Let us note that (2.40) is a very general mixed formulation. However, it may help the intuition of the reader to think of a displacement-based finite element discretization, which is the easiest case.

By choosing a basis $\phi^{(1)}, \ldots, \phi^{(N)}$ in W_h we can associate with every vector $\xi \in \mathbb{R}^N$ the element

$$\sum_i \xi_i \phi^{(i)} \in W_h \tag{2.41}$$

(in the usual way). Every problem (2.40) has now the form (2.1) with

$$M_{ij} := m(\phi^{(j)}, \phi^{(i)}) , \qquad b_i := \beta(\phi^{(i)}) . \tag{2.42}$$

If the bilinear form $m(\phi, \psi)$ satisfies

$$m(\phi, \psi) \leq k_m \|\phi\|_W \|\psi\|_W \quad \forall \phi, \psi \in W , \tag{2.43}$$

then (2.34) will easily hold with $k = k_m$ (with k_m independent of h) if we choose

$$\|\xi\|_L := \left\| \sum_i \xi_i \phi^{(i)} \right\|_W \tag{2.44}$$

as a norm in $\mathbb{R}^N$. The stability condition (2.35) can now be written in terms of the bilinear form $m(\phi, \psi)$ and of the space W_h as

$$\inf_{\phi_h \in W_h} \sup_{\psi_h \in W_h} \frac{m(\phi_h, \psi_h)}{\|\phi_h\|_W \|\psi_h\|_W} \geq \gamma > 0 \tag{2.45}$$

with γ independent of h.

We point out that (2.45) on one hand implies (as we have seen) the solvability of every discrete problem (2.40). On the other hand, if (2.45) holds with γ independent of h, then one can deduce optimal error bounds for the distance between the solution ϕ of (2.39) and the solution ϕ_h of (2.40). Incidentally, we point out that (2.45), together with

$$\lim_{h \to 0} \{ \inf_{\psi_h \in W_h} \|\psi - \psi_h\|_W \} = 0 \quad \forall \psi \in W \tag{2.49}$$

implies that (2.39) has a unique solution. We shall not report here the proof of this fact (which has basically little bearing upon our discussion), and shall instead report the proof of the optimal error bounds.

THEOREM 2.2 [6]. *Assume that the bilinear form $m(\phi, \psi)$ and the sequence of subspaces $W_h \subset W$ satisfy (2.43), (2.45) and (2.49). Let ϕ be the solution of (2.39) and ϕ_h the solution of (2.40). Then*

$$\|\phi - \phi_h\|_W \leq (1 + k_m/\gamma) \inf_{\psi_h \in W_h} \|\phi - \psi_h\|_W . \tag{2.50}$$

PROOF. For every $\psi_h \in W_h$ we have

$$\begin{aligned}
\gamma \|\psi_h - \phi_h\|_W &\leq \sup_{\chi_h \in W_h} \frac{m(\psi_h - \phi_h, \chi_h)}{\|\chi_h\|_W} \quad \text{(use (2.45))} \\
&= \sup_{\chi_h \in W_h} \frac{\{m(\psi_h - \phi, \chi_h) - m(\phi - \phi_h, \chi_h)\}}{\|\chi_h\|_W} \quad \text{(add and subtract } \phi\text{)} \\
&= \sup_{\chi_h \in W_h} \frac{m(\psi_h - \phi, \chi_h)}{\|\chi_h\|_W} \quad \text{(use (2.39) and (2.40))} \\
&\leq \sup_{\chi_h \in W_h} \frac{k_m \|\psi_h - \phi\|_W \|\chi_h\|_W}{\|\chi_h\|_W} \quad \text{(use (2.43))} \\
&= k_m \|\psi_h - \phi\|_W .
\end{aligned} \tag{2.51}$$

From (2.51) we have, using the triangle inequality,

$$\begin{aligned}\|\phi_h - \phi\|_W &\leq \|\phi_h - \psi_h\|_W + \|\psi_h - \phi\|_W \\ &\leq (k_m/\gamma)\|\psi_h - \phi\|_W + \|\psi_h - \phi\|_W \\ &= (1 + k_m/\gamma)\|\psi_h - \phi\|_W \end{aligned} \tag{2.52}$$

and (2.50) follows since (2.52) holds for every $\psi_h \in W_h$. □

3. Solvability and stability of mixed finite element formulations

We consider now a special case of (2.1). Namely we assume that the matrix M has the typical form arrived at when using a mixed finite element formulation,

$$M = \begin{bmatrix} A & B^t \\ B & 0 \end{bmatrix}, \tag{3.1}$$

where A is a square $NA \times NA$ matrix and B a rectangular $NB \times NA$ matrix with (obviously) $NA + NB = N$. Accordingly we split the unknown $x = (u, p)$ with $u \in \mathbb{R}^{NA}$ and $p \in \mathbb{R}^{NB}$ and the right-hand side $b = (f, g)$ with $f \in \mathbb{R}^{NA}$, $g \in \mathbb{R}^{NB}$. With this notation, the linear system under examination can be written as

$$Au + B^t p = f\,, \qquad Bu = g\,. \tag{3.2}$$

The analysis of the solvability, stability and optimality of mixed formulations has been performed in [5]. However, we shall follow here the more elegant presentation of Arnold [7]. In any case, the following space is of crucial importance. We set

$$K = \{v \in \mathbb{R}^{NA} \mid Bv = 0\} \tag{3.3}$$

(in other words $K = \mathrm{Ker}(B)$). Let NK be the dimension of K; we can split $\mathbb{R}^{NA}$ as

$$\mathbb{R}^{NA} = T \oplus K\,, \tag{3.4}$$

where T is the orthogonal of K in $\mathbb{R}^{NA}$. As a consequence of (3.4) every $v \in \mathbb{R}^{NA}$ can be split, in a unique way, as a sum

$$v = v_T + v_K \quad \text{with } v_T \in T,\ v_K \in K \text{ and } v_T^t v_K = 0\,. \tag{3.5}$$

If NT is the dimension of T, we will obviously have $NT + NK = NA$.

Let us now assume that the system of equations with the matrix M has been established in a suitable basis so that we can write the matrix A as

$$A = \begin{bmatrix} A_{TT} & A_{TK} \\ A_{KT} & A_{KK} \end{bmatrix}. \tag{3.6}$$

The notation (3.6) implies that the choice of the basis and the ordering of the unknowns in $\mathbb{R}^{NA}$ has been done in such a way that every $v_T \in T$ has only the first NT components which are a priori different from zero, while every $v_K \in K$ has only the last NK components (a priori) different from zero. With a (quite natural) abuse of notation we shall therefore, when convenient, treat v_T as an element of $\mathbb{R}^{NT}$ (discarding the last NK components which are identically zero). Similarly we shall treat, when convenient, v_K as an element of $\mathbb{R}^{NK}$ (discarding the first NT components) so that, for $v = v_T + v_K$, we can write

$$Av = (A_{TT}v_T + A_{TK}v_K) + (A_{KT}v_T + A_{KK}v_K)\,. \tag{3.7}$$

In (3.7) the first term of the right-hand side belongs to T and the second one belongs to K. Similarly, the matrix B will have the form

$$B = [B_T \quad B_K]\,, \tag{3.8}$$

with

$$Bv = B_T v_T + B_K v_K\,, \tag{3.9}$$

always with the notation (3.5). Note now that from (3.3) and the definition of T we will have

$$Bv_K = B_K v_K = 0 \quad \forall v_K \in K \tag{3.10}$$

and

$$Bv_T = B_T v_T = 0 \quad \text{iff } v_T = 0\,, \tag{3.11}$$

so that (3.9) can actually be written as

$$Bv = B_T v_T\,. \tag{3.12}$$

We also have

$$B^t q = B_T^t q \in T \quad \forall q \in \mathbb{R}^{NB}\,. \tag{3.13}$$

With a similar splitting for the right-hand side $f = f_T + f_K$ the original system (3.2) can now be written as

$$\begin{aligned} &A_{TT}u_T + A_{TK}u_K + B_T^t p = f_T\,,\\ &A_{KT}u_T + A_{KK}u_K = f_K\,,\\ &B_T u_T = g\,. \end{aligned} \tag{3.14}$$

The conditions for the solvability of (3.14) (and hence of (3.2)) are now clear: we need that (i) the equation $B_T u_T = g$ is solvable for every $g \in \mathbb{R}^{NB}$, (ii) the equation $A_{KK}u_K = f_K$ is solvable for every $f_K \in K$ and (iii) the equation $B_T^t p = f_T$ is solvable for every f_T in T. Condition (i) is equivalent to

$$B_T, \text{ as a mapping: } T \to \mathbb{R}^{NB}, \text{ is invertible}\,. \tag{3.15}$$

On the other hand, condition (ii) is equivalent to

$$A_{KK}, \text{ as a mapping: } K \to K, \text{ is invertible .} \tag{3.16}$$

Note now that, from the definition of B_T and in particular from (3.11) we have that B_T is always injective, so that (3.15) implies $NT = NB$. Now we conclude that the matrix B_T, as a mapping: $T \to \mathbb{R}^{NB}$, is a non-singular square matrix, and therefore its transposed matrix B_T^t is also non-singular and (iii) is automatically satisfied.

We now want to express (3.15) and (3.16) in terms of the matrices A and B, and of the kernel K (defined in (3.3)). Condition (3.15) is clearly equivalent to

$$B^t p = 0 \quad \Rightarrow \quad p = 0 , \tag{3.17}$$

while (3.16) can be written as

$$(u \in K \quad \text{and} \quad v^t A u = 0 \quad \forall v \in K) \quad \Rightarrow \quad u = 0 . \tag{3.18}$$

Conditions (3.17) and (3.18) are necessary and sufficient for the solvability of (3.2) for every right-hand side $f \in \mathbb{R}^{NA}$ and $g \in \mathbb{R}^{NB}$. We can summarize the above results in the following proposition.

PROPOSITION 3.1. Let A be an $NA \times NA$ square matrix and let B be an $NB \times NA$ matrix, and let K (the kernel of B) be defined as in (3.3). The linear system (3.2) is uniquely solvable for every $f \in \mathbb{R}^{NA}$ and for every $g \in \mathbb{R}^{NB}$ if and only if conditions (3.17) and (3.18) are satisfied. □

Note that, in particular, condition (3.17) implies

$$NA = NK + NT = NK + NB \geq NB , \tag{3.19}$$

which is (obviously) a necessary condition for the solvability of (3.2). However, we now recognize that the use of (3.19) as a test for solvability (or, worse, for stability) is too simplistic and hence misleading. Note also that, if A is symmetric and positive semi-definite, then (3.18) can be expressed by the easier form

$$v^t A v > 0 \quad \forall v \in K . \tag{3.20}$$

We address now the problem of stability of (3.2). In agreement with the approach of the previous section we might decide now from the very beginning to use dual norms for measuring the right-hand sides. Hence we assume that we have chosen a norm $\| \ \|_V$ in $\mathbb{R}^{NA}$ and a norm $\| \ \|_Q$ in $\mathbb{R}^{NB}$ and define the stability constant S as the smallest constant such that

$$\frac{\|\delta u\|_V + \|\delta p\|_Q}{\|u\|_V + \|p\|_Q} \leq S \, \frac{\|\delta f\|_{DV} + \|\delta g\|_{DQ}}{\|f\|_{DV} + \|g\|_{DQ}} \tag{3.21}$$

for all u, p, δu, δp and f, g, δf, δg with $Au + B^t p = f$; $Bu = g$; $A\,\delta u + B^t\,\delta p = \delta f$ and $B\,\delta u = \delta g$. From the previous section we have again $S \geq 1$.

REMARK 3.1. Definition (3.21) coincides with (2.11) if we take

$$\|(u, p)\|_{\mathrm{L}} = \|u\|_V + \|p\|_Q , \tag{3.22}$$

$$\|(f, g)\|_{\mathrm{R}} = \|f\|_{\mathrm{DV}} + \|g\|_{\mathrm{DQ}} . \tag{3.23}$$

We notice that in this case the norm $\| \ \|_{\mathrm{R}}$ is not the dual norm of $\| \ \|_{\mathrm{L}}$. Actually we have

$$\|(f, g)\|_{\mathrm{DL}} = \max (\|f\|_{\mathrm{DV}}, \|g\|_{\mathrm{DQ}}) . \tag{3.24}$$

However one can easily check that

$$\|(f, g)\|_{\mathrm{DL}} \leq \|(f, g)\|_{\mathrm{R}} \leq 2\|(f, g)\|_{\mathrm{DL}} , \tag{3.25}$$

so that the conditions for the uniform stability are still as discussed in the previous section.

Our aim is now to give conditions on a sequence of problems (3.2) in order to have S uniformly bounded. We might of course use, for instance, (2.34) and (2.35) as in the previous section (since we are dealing here with a particular case of the previous discussion). However, we prefer to have separate conditions on the (sequence of) matrices A and B, as we did for the solvability problem. This, actually, is much more convenient in actual applications.

We assume, for the sake of simplicity, that there exist two constants k_A and k_B such that

$$v^{\mathrm{t}} A u \leq k_A \|v\|_V \, \|u\|_V \quad \forall v, u , \tag{3.26}$$

$$v^{\mathrm{t}} B^{\mathrm{t}} q \leq k_B \|v\|_V \, \|q\|_Q \quad \forall v, q , \tag{3.27}$$

with k_A and k_B uniformly bounded from above and from below. In actual applications, (3.26) and (3.27), are easily fulfilled with the 'natural choice' for the norms $\| \ \|_V$ and $\| \ \|_Q$. Notice that (3.26) and (3.27) immediately imply that A has norm $\leq k_A$ from $\| \ \|_V$ into $\| \ \|_{\mathrm{DV}}$ (as in the previous section, Proposition 2.1). Similarly B_T has norm $\leq k_B$ from $\| \ \|_V$ into $\| \ \|_{\mathrm{DQ}}$. On the other hand (3.26) and (3.27) also imply

$$\|M\|_{\mathrm{LR}} \leq k_A + 2k_B \tag{3.28}$$

for M given in (3.1) and the norms $\| \ \|_{\mathrm{L}}$ and $\| \ \|_{\mathrm{R}}$ as in (3.22), (3.23). Hence, in view of (2.17) we only have to control $\|M^{-1}\|_{\mathrm{RL}}$. Assuming that M is invertible (and hence, by (3.14) A_{KK} and B_T are also invertible) we easily have from (3.14) that

$$\|u_T\|_V \leq \|B_T^{-1}\| \, \|g\|_{\mathrm{DQ}} , \tag{3.29}$$

$$\begin{aligned} \|u_K\|_V &\leq \|A_{KK}^{-1}\|(\|f_K\|_{\mathrm{DV}} + \|A_{KT} u_T\|_{\mathrm{DV}}) \\ &\leq \|A_{KK}^{-1}\|(\|f_K\|_{\mathrm{DV}} + k_A \|u_T\|_V) , \end{aligned} \tag{3.30}$$

$$\|p\|_Q \leq \|(B_T^t)^{-1}\|(\|f_T\|_{DV} + \|A_{TT}u_T\|_{DV} + \|A_{TK}u_K\|_{DV})$$
$$\leq \|(B_T^t)^{-1}\|(\|f_T\|_{DV} + k_A(\|u_T\|_V + \|u_K\|_V)), \tag{3.31}$$

where

$$\|B_T^{-1}\| = \sup_g \frac{\|B_T^{-1}g\|_V}{\|g\|_{DQ}}, \tag{3.32}$$

$$\|A_{KK}^{-1}\| = \sup_{v \in K} \frac{\|A_{KK}^{-1}v\|_V}{\|v\|_{DV}} \tag{3.33}$$

and

$$\|(B_T^t)^{-1}\| = \sup_{f_T} \frac{\|(B_T^t)^{-1}f_T\|_Q}{\|f_T\|_{DV}}. \tag{3.34}$$

Substituting (3.29) into (3.30) and then (3.29) and (3.30) into (3.31), one obtains

$$\|(u, p)\| \leq \mathscr{C}(\|B_T^{-1}\|, \|A_{KK}^{-1}\|, \|(B_T^t)^{-1}\|, k_A) \cdot \|(f, g)\|. \tag{3.35}$$

Since, as is easy to check,

$$\|B_T^{-1}\| = \|(B_T^t)^{-1}\|, \tag{3.36}$$

it follows from (3.35) that, in order to have a uniform bound for S (in (3.21)), we only need that $\|B_T^{-1}\|$ (or $(B_T^t)^{-1}\|$) and $\|A_{KK}^{-1}\|$ are uniformly bounded from above. In order to express this condition in terms of the matrices A, B (and of the kernel K as defined in (3.3)) we shall rather write that $\|(B_T^t)^{-1}\|^{-1}$ and $\|A_{KK}^{-1}\|^{-1}$ are uniformly bounded from below by some positive constant. Actually we have

$$\|(B_T^t)^{-1}\|^{-1} = \left(\sup_{f_T} \frac{\|(B_T^t)^{-1}f_T\|_Q}{\|f_T\|_{DV}}\right)^{-1} \quad \text{(use (3.35))}$$
$$= \inf_{f_T} \frac{\|f_T\|_{DV}}{\|(B_T^t)^{-1}f_T\|_Q}$$
$$= \inf_q \frac{\|B_T^t q\|_{DV}}{\|q\|_Q} \quad (\text{use } f_T = B_T^t q)$$
$$= \inf_q \frac{\|B^t q\|_{DV}}{\|q\|_Q} \quad \text{(use (3.13))}$$
$$= \inf_q \sup_v \frac{v^t B^t q}{\|q\|_Q \|v\|_V} \quad \text{(use (2.25))}$$
$$= \inf_q \sup_v \frac{q^t B v}{\|v\|_V \|q\|_Q} \tag{3.37}$$

and

$$\|A_{KK}^{-1}\|^{-1} = \left(\sup_{v\in K} \frac{\|A_{KK}^{-1}v\|_V}{\|v\|_{DV}}\right)^{-1} \quad \text{(use (3.33))}$$

$$= \inf_{v\in K} \frac{\|v\|_{DV}}{\|A_{KK}^{-1}v\|_V}$$

$$= \inf_{u\in K} \frac{\|A_{KK}u\|_{DV}}{\|u\|_V} \quad (\text{use } v = A_{KK}u)$$

$$= \inf_{u\in K} \sup_{z\in K} \frac{z^t A_{KK} u}{\|u\|_V \|z\|_V} \quad \text{(use (2.25))}$$

$$= \inf_{u\in K} \sup_{z\in K} \frac{z^t A u}{\|u\|_V \|z\|_V} \quad \text{(use (3.7))}. \tag{3.38}$$

From (3.37), (3.38) and the previous discussion we now have the following proposition.

PROPOSITION 3.2. Assume that we are given a sequence of problems of type (3.2). Assume that the matrices A and B satisfy (3.26) and (3.27) with k_A and k_B uniformly bounded. The stability constant S in (3.21) will then be uniformly bounded if and only if there exist two positive constants α and β such that

$$\inf_{u\in K} \sup_{v\in K} \frac{v^t A u}{\|v\|_V \|u\|_V} \geq \alpha > 0 \tag{3.39}$$

and

$$\inf_{q} \sup_{v} \frac{q^t B v}{\|v\|_V \|q\|_Q} \geq \beta > 0\,. \tag{3.40}$$

for every problem of the sequence.

REMARK 3.2. If every matrix A is symmetric and positive semi-definite, then (3.39) takes the simpler form

$$\exists \alpha > 0 \quad \text{such that} \quad v^t A v \geq \alpha \|v\|_V^2 \quad \forall v \in K\,, \tag{3.41}$$

with K (as in (3.39)) always given by (3.3). In some applications (typically in the solution of Stokes fluid flow problems) the matrices A will be positive definite and satisfy (3.41) for all v in $\mathbb{R}^{NA}$. This led some authors to consider (3.40) as *the* condition for stability and convergence of mixed methods, which obviously is not the case. For instance, in the analysis of the mixed (σ, u) formulation of elasticity problems in the nearly incompressible case, condition (3.39) is more delicate to enforce than (3.40). On the same erroneous trend, some authors seem incapable of distinguishing between (2.35) (which is a condition on the whole matrix M) and (3.40) (which is a condition on the rectangular submatrix B of a special case of the matrix M, namely (3.1)).

Let us consider now, as we did in the previous section, an abstract continuous problem and its Galerkin approximation. Assume that we are given two Hilbert spaces V and Q and two bilinear forms $a(u, v)$ (on $V \times V$) and $b(v, p)$ (on $V \times Q$). We assume from the beginning that the two forms are continuous in the sense that there exist two positive constants k_a and k_b such that

$$a(u, v) \leq k_a \|u\|_V \|v\|_V \quad \forall u, v \in V \tag{3.42}$$

and

$$b(v, p) \leq k_b \|v\|_V \|p\|_Q \quad \forall v \in V, p \in Q . \tag{3.43}$$

We can also introduce a kernel

$$\mathcal{K} = \{v \in V \mid b(v, q) = 0 \ \forall q \in Q\} , \tag{3.44}$$

which is the continuous version of the kernel K defined by (3.3). For the sake of simplicity we shall also assume that $a(u, v)$ is symmetric and positive semi-definite, that is,

$$a(u, v) = a(v, u) \quad \forall u, v \in V , \tag{3.45}$$

$$a(v, v) \geq 0 \quad \forall v \in V . \tag{3.46}$$

Finally, in analogy with (3.40) and (3.41) we make the following assumptions:

$$\exists \bar{\alpha} > 0 \quad \text{such that} \quad a(v, v) \geq \bar{\alpha} \|v\|_V^2 \quad \forall v \in \mathcal{K} , \tag{3.47}$$

$$\exists \bar{\beta} > 0 \quad \text{such that} \quad \inf_{q \in Q} \sup_{v \in V} \frac{b(v, q)}{\|v\|_V \|q\|_Q} \geq \bar{\beta} . \tag{3.48}$$

We have the following existence and uniqueness theorem.

THEOREM 3.1. [5] *Assume (3.42)–(3.48). For every $f \in V'$ and for every $g \in Q'$, where V' and Q' are the dual spaces of V and Q, respectively, there exists a unique pair (u, p) in $V \times Q$ such that*

$$\begin{aligned} &a(u, v) + b(v, p) = f(v) \quad \forall v \in V , \\ &b(u, q) = g(q) \quad \forall q \in Q . \end{aligned} \tag{3.49}$$

Assume now that we are given a sequence (V_h, Q_h) of finite dimensional subspaces of V and Q, respectively, and consider the finite dimensional approximations of (3.49):

$$\begin{aligned} &\text{Find } u_h \in V_h \text{ and } p_h \in Q_h \text{ such that} \\ &a(u_h, v_h) + b(v_h, p_h) = f(v_h) \quad \forall v_h \in V_h , \\ &b(u_h, q_h) = g(q_h) \quad \forall q_h \in Q_h . \end{aligned} \tag{3.50}$$

It will also be convenient to introduce the finite dimensional kernels

$$\mathcal{K}_h = \{v_h \in V_h \mid b(v_h, q_h) = 0 \ \forall q_h \in Q_h\} . \tag{3.51}$$

It is clear that, by choosing bases in V_h and Q_h, (3.50) can be written in the form (3.2). As a consequence, the solvability conditions for (3.50) will be

$$a(v_h, v_h) > 0 \quad \forall v_h \in \mathcal{K}_h , \tag{3.52}$$

$$\{b(v_h, q_h) = 0 \quad \forall v_h \in V_h\} \quad \Rightarrow \quad q_h = 0 , \tag{3.53}$$

as it can easily be deduced from (3.20) and (3.17). The uniform stability conditions now become

$$\exists \alpha^* > 0 \quad \text{such that} \quad a(v_h, v_h) \geq \alpha^* \|v_h\|_V^2 \quad \forall v_h \in \mathcal{K}_h , \tag{3.54}$$

$$\exists \beta^* > 0 \quad \text{such that} \quad \inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V \, \|q_h\|_Q} \geq \beta^* , \tag{3.55}$$

with α^* and β^* independent of h. It is clear that (3.54) and (3.55) are just a different way of writing (3.41) and (3.40). It is also clear that (3.54) implies (3.52), and (3.55) implies (3.53), so that stability implies solvability.

As far as error estimates are concerned we have the following theorem.

THEOREM 3.2. [5] *Assume that the sequence of subspaces* (V_h, Q_h) *satisfies* (3.54) *and* (3.55). *Then problem* (3.50) *has a unique solution* (u_h, p_h) *for every* $h > 0$. *Moreover there exists a constant* $c > 0$, *depending only on* k_a (3.42), k_b (3.43), α^* (3.54) *and* β^* (3.55) *such that*

$$\|u - u_h\|_V + \|p - p_h\|_Q \leq C\{\inf_{v_h \in V_h} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q\} , \tag{3.56}$$

where (u, p) *is the solution of* (3.49).

PROOF. We shall only sketch the proof, which is based on a classical 'stability–consistency' argument. Let u_h^* and p_h^* be the best approximation one can have for u and p, respectively, in the subspaces, that is,

$$\|u - u_h^*\|_V = \inf_{v_h \in V_h} \|u - v_h\|_V , \tag{3.57}$$

$$\|p - p_h^*\|_Q = \inf_{q_h \in Q_h} \|p - q_h\|_Q . \tag{3.58}$$

Let now

$$\tilde{f}(v_h) := a(u_h^*, v_h) + b(v_h, p_h^*) , \tag{3.59}$$

$$\tilde{g}(q_h) := b(u_h^*, q_h) , \tag{3.60}$$

and notice that

$$(f-\tilde{f})(v_h) = a(u-u_h^*, v_h) + b(v_h, p-p_h^*)\,, \tag{3.61}$$

$$(g-\tilde{g})(q_h) = b(u-u_h^*, q_h)\,. \tag{3.62}$$

Notice finally that $(u_h - u_h^*, p_h - p_h^*)$ solves a problem of type (3.50) with right-hand side given by $(f-\tilde{f}, g-\tilde{g})$. The stability of (3.50) implies that

$$\begin{aligned}\|u_h - u_h^*\|_V + \|p_h - p_h^*\|_Q &\leq C_1(\|f-\tilde{f}\|_{DV} + \|g-\tilde{g}\|_{DQ}) \\ &= C_1\left\{\sup_{v_h}\frac{(f-\tilde{f})(v_h)}{\|v_h\|_V} + \sup_{q_h}\frac{(g-\tilde{g})(q_h)}{\|q_h\|_Q}\right\} \leq C_2\{\|u-u_h^*\|_V + \|p-p_h^*\|_Q\}\,,\end{aligned} \tag{3.63}$$

with C_1 and C_2 depending only on α^*, β^*, k_a, k_b. From (3.62) and the triangle inequality we now have

$$\|u-u_h\|_V + \|p-p_h\|_Q \leq (1+C_2)(\|u-u_h^*\|_V + \|p-p_h^*\|_Q)\,, \tag{3.64}$$

and (3.64) with (3.57) and (3.58) gives (3.56). □

We end this section with some observations regarding penalty methods applied to systems of the form (3.2). For the sake of simplicity, assume that $NA = 3$, $NB = 2$ and that M has the form

$$M = \begin{bmatrix} \alpha_1 & 0 & 0 & \beta_1 & 0 \\ 0 & \alpha_2 & 0 & 0 & \beta_2 \\ 0 & 0 & \alpha_3 & 0 & 0 \\ \beta_1 & 0 & 0 & 0 & 0 \\ 0 & \beta_2 & 0 & 0 & 0 \end{bmatrix}. \tag{3.65}$$

For a more realistic situation we have to think of (3.65) as a block partitioning of M. It is clear that the system (3.2) splits now into

$$\alpha_i u_i + \beta_i p_i = f_i\,, \quad \beta_i u_i = g_i\,, \quad i = 1,2 \tag{3.66}$$

and

$$\alpha_3 u_3 = f_3\,. \tag{3.67}$$

If one of the β_i vanishes then (3.17) is violated and M is singular. If instead $\beta_i \neq 0$ $(i = 1,2)$, then

$$K = \{(0,0,u_3) \mid u_3 \in \mathbb{R}\} \tag{3.68}$$

and $\alpha_3 \neq 0$ satisfies (3.18). Assuming that all the α_i $(i = 1,2,3)$ are bounded away from zero (for the sake of simplicity) we only have to consider systems of type (3.66) that we consider

through their typical representative:

$$\alpha u + \beta p = f\,, \qquad \beta u = g\,. \tag{3.69}$$

A penalty approach to (3.69) consists in finding, for $\varepsilon > 0$ (and 'small'), the solution of

$$\alpha u_\varepsilon + \beta p_\varepsilon = f\,, \qquad \beta u_\varepsilon - \varepsilon p_\varepsilon = g\,, \tag{3.70}$$

which is given by

$$u_\varepsilon = \frac{\varepsilon f + \beta g}{\varepsilon\alpha + \beta^2}\,, \qquad p_\varepsilon = \frac{\beta f - \alpha g}{\varepsilon\alpha + \beta^2}\,. \tag{3.71}$$

It is clear that (for $\alpha \neq 0$) the solution (3.70) always exists, even for $\beta = 0$. However, for $\beta = 0$, $p_\varepsilon \to \infty$ as $\varepsilon \to 0$ when $g \neq 0$. On the other hand, in some applications (as, for instance, incompressibility conditions with zero Dirichlet boundary conditions), we have $g = 0$, and the situation improves. For $g = 0$ (3.71) becomes

$$u_\varepsilon = \frac{\varepsilon f}{\varepsilon\alpha + \beta^2}\,, \qquad p_\varepsilon = \frac{\beta f}{\varepsilon\alpha + \beta^2}\,. \tag{3.72}$$

For $\beta = 0$ (3.72) gives

$$u_\varepsilon = \frac{f}{\alpha}\,, \qquad p_\varepsilon = 0\,, \tag{3.73}$$

which is a nice result. Actually a closer look at (3.69) for $\beta = g = 0$ shows a singular but compatible system with solution $u = f/\alpha$, $p =$ undetermined. Clearly (3.73) gives the solution of minimum norm.

Let us now consider the case $g = 0$ and β very small. The system (3.69) will have a very large stability constant. However, if we look only at the u_ε component of (3.72), we have

$$u_\varepsilon \to 0 \quad \text{for } \varepsilon \to 0 \ (\beta \text{ fixed}) \tag{3.74}$$

and u_ε is uniformly bounded (in ε) as ε goes to zero. On the other hand

$$u_\varepsilon \to \frac{f}{\alpha} \quad \text{for } \beta \to 0 \ (\varepsilon \text{ fixed})\,, \tag{3.75}$$

which shows that we have difficulties to interpret the results even if u_ε is computed as a number of reasonable size. A look at the p_ε part of the solution shows that

$$p_\varepsilon \to \frac{f}{\beta} \quad \text{for } \varepsilon \to 0 \ (\beta \text{ fixed})\,. \tag{3.76}$$

If β is very small, p_ε will be very large and this indicates that a change in discretization may be required.

In a practical analysis, there will generally be only a limited number of the β_i's that are small. Hence, only the corresponding p_i components will be large, and this can explain the appearance of the so-called checker-board modes that appear, even when u_ε behaves nicely. Note also that, when solving with the penalty approach (for $g = 0$) a small β_i can be more dangerous than a $\beta_i = 0$, as shown by (3.73) compared with (3.74) and (3.76).

4. Examples of applications

In this section we present two examples that demonstrate the theory we have presented in the earlier part of the paper, and we also present a numerical procedure to test whether the inf–sup condition is satisfied for a finite element formulation to solve Stokes fluid flow problems.

4.1. Mixed methods for linear second-order elliptic problems

We start here with a very simple example to show the importance of the condition (3.54) ($\mathcal{K}_h$-ellipticity). Consider the mixed formulation of the model problem

$$\psi'' = 1 \quad = \text{in }]-1, 1[\,, \qquad \psi(-1) = \psi(1) = 0\,. \tag{4.1}$$

The solution is clearly

$$\psi(x) = \tfrac{1}{2}(x^2 - 1)\,. \tag{4.2}$$

Introducing the additional variable

$$\sigma = \psi'\,, \tag{4.3}$$

the mixed formulation of (4.1) now reads

$$\int_{-1}^{1} \sigma\tau \,\mathrm{d}x + \int_{-1}^{1} \psi\tau' \,\mathrm{d}x = 0 \quad \forall\tau\,, \tag{4.4}$$

$$\int_{-1}^{1} \sigma'\phi \,\mathrm{d}x = \int_{-1}^{1} \phi \,\mathrm{d}x \quad \forall\phi\,, \tag{4.5}$$

which is clearly of the form (3.49) with

$$V = \{\tau \in L^2(]-1, 1[) \,|\, \tau' \in L^2(]-1, 1[\}\,, \qquad \|\tau\|_V^2 = \|\tau\|_0^2 + \|\tau'\|_0^2\,, \tag{4.6}$$

$$Q = L^2(]-1, 1[)\,, \qquad \|\phi\|_Q = \|\phi\|_0\,, \tag{4.7}$$

$$a(\sigma, \tau) = \int_{-1}^{1} \sigma\tau \,\mathrm{d}x\,, \qquad b(\tau, \psi) = \int_{-1}^{1} \psi\tau' \,\mathrm{d}x\,, \tag{4.8}$$

where in (4.6) and (4.7) we used

$$\|v\|_0^2 := \int_{-1}^{1} v^2(x)\,dx\,. \tag{4.9}$$

Note how the form of a and b in (4.8) easily determines the norms (4.6) and (4.7) which are needed to have (3.42) and (3.43).

Let us check, as an exercise, that our problem satisfies (3.47) and (3.48). We first have to find what $\mathcal{K}$ is, as defined by (3.44). We have

$$\left\{\int_{-1}^{1} \phi\tau'\,dx = 0 \quad \forall\phi\right\} \quad \Leftrightarrow \quad \tau' = 0 \quad \Leftrightarrow \quad \tau = \text{constant}\,, \tag{4.10}$$

so that $\mathcal{K}$ contains only the constant functions. For $\tau \in \mathcal{K}$ we have

$$a(\tau,\tau) = \|\tau\|_0^2 = \|\tau\|_V^2 \quad (\text{since } \tau' = 0) \tag{4.11}$$

and therefore (3.47) holds with $\bar{\alpha} = 1$.

Let us now turn to (3.48). For every $\bar{\phi} \in L^2(]-1,1[)$ we can set

$$\bar{\tau}(x) = \int_0^x \bar{\phi}(t)\,dt\,. \tag{4.12}$$

We then obviously have

$$b(\bar{\tau},\bar{\phi}) = \int_{-1}^{1} \bar{\phi}^2 = \|\bar{\phi}\|_0^2\,, \tag{4.13}$$

$$\|\bar{\tau}'\|_0^2 = \|\bar{\phi}\|_0^2\,. \tag{4.14}$$

Furthermore

$$\|\bar{\tau}\|_0^2 \leq \|\bar{\phi}\|_0^2 \tag{4.15}$$

and hence we have

$$\begin{aligned}\sup_{\tau} \frac{b(\tau,\bar{\phi})}{\|\tau\|_V} &\geqslant \frac{b(\bar{\tau},\bar{\phi})}{\|\bar{\tau}\|_V} = \frac{\|\bar{\phi}\|_0^2}{(\|\bar{\tau}\|_0^2 + \|\bar{\tau}'\|_0^2)^{1/2}} \\ &\geqslant \frac{\|\bar{\phi}\|_0^2}{(\|\bar{\phi}\|_0^2 + \|\bar{\phi}\|_0^2)^{1/2}} = \frac{1}{\sqrt{2}}\,\|\bar{\phi}\|_0 \quad (\text{use (4.14) and (4.15)})\,.\end{aligned} \tag{4.16}$$

Since (4.16) holds for every $\bar{\phi}$ we obtain (3.48) with $\bar{\beta} = 1/\sqrt{2}$.

Let us now consider the discretization of (4.4), (4.5). We take a decomposition of $]-1,1[$ into N equal intervals and set

$$Q_h = \{\text{piecewise constants } (= \mathcal{L}_0^0 \text{ with the notation of [1]})\}\,. \tag{4.17}$$

It would now be reasonable to take

$$V_h = \{\text{piecewise linear continuous functions } (= \mathscr{L}_1^1)\} . \tag{4.18}$$

In this case it is easy to check that $\mathscr{K}_h$ is also reduced to the constant functions, and therefore (3.54) holds with $\alpha^* = 1$ (always by (4.11)). On the other hand the construction (4.12) still works, since $\bar{\phi} \in \mathscr{L}_0^0$ implies $\bar{\tau} \in \mathscr{L}_1^1$. Hence (3.55) also holds, with $\beta^* = 1/\sqrt{2}$, and the assumptions of Theorem 3.2 are fulfilled. In our particular case ($g \equiv 1$) it is also easy to prove that, for every decomposition, we have

$$\sigma_h(x) = x \quad \text{in }]-1, 1[, \tag{4.19}$$

which is the exact solution.

Assume now, for our discussion, that we take a larger V_h, namely

$$V_h = \{\text{piecewise quadratic continuous functions } (= \mathscr{L}_2^1)\} . \tag{4.20}$$

Setting

$$B_2 = \{\text{piecewise quadratic functions vanishing at the subdivision nodes}\} , \tag{4.21}$$

we easily have

$$V_h = \mathscr{L}_2^1 = \mathscr{L}_1^1 \oplus B_2 . \tag{4.22}$$

We can now make an observation which is of general validity: the choice of a larger V_h (with the same Q_h) makes the inf–sup condition (3.55) easier to satisfy and the $\mathscr{K}_h$-ellipticity condition (3.54) more difficult to satisfy (unless, obviously, the bilinear form a is V-elliptic: in such a case (3.54) is always satisfied for all choices of V_h and Q_h). In our case

$$\mathscr{K}_h = \left\{ \tau \in \mathscr{L}_2^1 \,\middle|\, \int_{-1}^{1} \phi \tau' \, dx = 0 \ \forall \phi \in \mathscr{L}_0^0 \right\} = \mathscr{K} \oplus B_2 , \tag{4.23}$$

where $\mathscr{K}$ is the space of global constants as in (4.10). Let now, for every subinterval $I_k (k = 1, \ldots, N)$, b_k be the second order polynomial vanishing at the endpoints of I_k and normalized in such a way that

$$\int_{I_k} b_k^2(x) \, dx = 1 . \tag{4.24}$$

A simple computation shows that

$$\|b_k'\|_0^2 = 10/h , \tag{4.25}$$

so that

$$a(b_k, b_k) = \|b_k\|_0^2 = 1 \tag{4.26}$$

and

$$\|b_k\|_V^2 = \|b_k\|_0^2 + \|b_k'\|_0^2 = 1 + 10/h\,, \tag{4.27}$$

and condition (3.54) can only hold for

$$\alpha^* = h/(h+10)\,, \tag{4.28}$$

which is not bounded uniformly from below. On the other hand, it is obvious that for every $\phi \in \mathscr{L}_0^0$ we have

$$\sup_{\tau\in\mathscr{L}_2^1} \frac{b(\tau,\phi)}{\|\tau\|_V} \geqslant \sup_{\tau\in\mathscr{L}_1^1} \frac{b(\tau,\phi)}{\|\tau\|_V} \geqslant \frac{1}{\sqrt{2}}\,\|\phi\|_Q\,, \tag{4.29}$$

and (3.55) as predicted is easier satisfied with a larger V_h. We are therefore facing a case where the inf–sup condition (3.55) is easily satisfied, but the $\mathscr{K}_h$-ellipticity condition (3.54) holds only with $\alpha^* \sim h$. Notice that (3.52) is still satisfied so that the discrete problem will be uniquely solvable. Notice as well that we started with an effective discretization ($V_h = \mathscr{L}_1^1$, $Q_h = \mathscr{L}_0^0$), that gave the exact value for σ_h, and that we enlarged V_h (which, as we have seen, does not affect the ability to satisfy the inf–sup condition). The key question must now of course be 'how is our solution accuracy affected by the enlargement by V_h?'

The solution of the discrete problem:

Find $\sigma_h \in \mathscr{L}_2^1$ and $\psi_h \in \mathscr{L}_0^0$ such that

$$\int_{-1}^{1} \sigma_h \tau\,\mathrm{d}x + \int_{-1}^{1} \psi_h \tau'\,\mathrm{d}x = 0 \quad \forall \tau \in \mathscr{L}_2^1\,, \tag{4.30}$$

$$\int_{-1}^{1} \phi \sigma_h'\,\mathrm{d}x - \int_{-1}^{1} \phi\,\mathrm{d}x = 0 \quad \forall \phi \in \mathscr{L}_0^0\,, \tag{4.31}$$

can again be computed by hand. Namely, from (4.31) we obtain

$$\sigma_h(x) = x + c + \sum_{k=1}^{N} c_k b_k(x)\,,$$

with c and c_k to be determined. Now $c = 0$ for symmetry reasons (the solution is unique) and choosing $\tau = b_k(x)$ in (4.30) yields

$$-c_k = \int_{I_k} x b_k(x)\,\mathrm{d}x =: (x, b_k)\,, \tag{4.32}$$

so that

$$\sigma_h(x) = x - \sum_{k=1}^{N} b_k(x)(x, b_k)\,, \tag{4.33}$$

and the L^2 norm of the error $\sigma_h - \sigma = \sigma_h - x$ is given by

$$\|\sigma_h - \sigma\|_0^2 = \sum_{k=1}^{N} (x, b_k)^2 , \tag{4.34}$$

which does not tend to zero. Hence our solution scheme with the enlarged V_h is not acceptable.

4.2. *Analysis of stationary incompressible fluids (the Stokes problem)*

We consider now the following model problem in a smooth domain $\Omega \subset \mathbb{R}^2$:

$$\begin{aligned} &\text{Find } u \in (H_0^1(\Omega))^2 \text{ and } p \in L^2(\Omega)/\mathbb{R} \quad \text{such that} \\ &\int_\Omega \operatorname{grad} u : \operatorname{grad} v \, d\Omega - \int_\Omega p \operatorname{div} v \, d\Omega = \int_\Omega f \cdot v \, d\Omega \quad \forall v , \\ &\int_\Omega q \operatorname{div} u \, d\Omega = 0 \quad \forall q , \end{aligned} \tag{4.35}$$

which is again of type (3.49) for

$$V = (H_0^1(\Omega))^2 = \{ v \in (L^2(\Omega))^2 \,|\, \operatorname{grad} v \in (L^2(\Omega))^4 \text{ and } v|_{\partial\Omega} = 0 \} , \tag{4.36}$$

$$Q = L^2(\Omega)/\mathbb{R} = \left\{ q \in L^2(\Omega) \,\middle|\, \int_\Omega q \, d\Omega = 0 \right\} \tag{4.37}$$

and

$$a(u, v) = \int_\Omega \operatorname{grad} u : \operatorname{grad} v \, d\Omega , \qquad b(v, q) = \int_\Omega q \operatorname{div} v \, d\Omega . \tag{4.38}$$

Note again how the form of a and b in (4.38) easily determines the norms (4.36) and (4.37) which are needed for having (3.42) and (3.43).

Problem (4.35) is the variational formulation of the problem

$$-\Delta u + \nabla p = f \quad \text{in } \Omega , \qquad \operatorname{div} u = 0 \quad \text{in } \Omega , \qquad u = 0 \quad \text{on } \partial\Omega , \tag{4.39}$$

which are the governing equations of an incompressible fluid. The well-known Poincaré inequality,

$$\exists c = c(\Omega) \quad \text{with} \quad \int_\Omega |v|^2 \, d\Omega \leq c(\Omega) \int_\Omega |\operatorname{grad} v|^2 \, d\Omega \quad \forall v \in V \tag{4.40}$$

now ensures that

$$a(v, v) \geq \left(\frac{1}{c+1} \right) \|v\|_V^2 \quad \forall v \in V , \tag{4.41}$$

so that (3.47) holds with $\bar{\alpha} = 1/(c+1)$ for all $v \in V$ and we do not need to be concerned with kernel $\mathcal{K}$. In particular, we have that (3.54) will also hold (with $\alpha^* = 1/(c+1)$) for every

choice of $V_h \subset V$ and $Q_h \subset Q$. Hence we can concentrate our attention on (3.48) and (3.55), that is, on the inf–sup condition. As far as (3.48) is concerned we remark that we actually have

$$\exists \beta(\Omega) > 0 \quad \text{such that} \quad \inf_{q \in Q} \sup_{v \in V} \frac{\int_\Omega q \operatorname{div} v \, d\Omega}{\|q\|_Q \|v\|_V} \geq \beta(\Omega) , \tag{4.42}$$

which is a nontrivial result in functional analysis (see, e.g., [8, 9]). We also notice that the following result obviously holds as an immediate consequence of (4.42): for every set $\mathcal{V}$ with

$$(H_0^1(\Omega))^2 \subseteq \mathcal{V} \subseteq (H^1(\Omega))^2 , \tag{4.43}$$

we have

$$\inf_{q \in Q} \sup_{v \in \mathcal{V}} \frac{\int_\Omega q \operatorname{div} v \, d\Omega}{\|q\|_Q \|v\|_V} \geq \beta(\Omega) , \tag{4.44}$$

since the supremum over $\mathcal{V}$ is obviously larger than the supremum over $V \equiv (H_0^1(\Omega))^2$. In a sense we can therefore say that the case of homogeneous Dirichlet boundary conditions is the most difficult to treat. This is the reason why we shall mainly concentrate on this case only.

Assume now that we are given a sequence of finite dimensional subspaces $V_h \in V$ and $Q_h \in Q$ and consider the discrete problem:

$$\begin{aligned} &\text{Find } u_h \in V_h \text{ and } p_h \in Q_h \quad \text{such that} \\ &a(u_h, v_h) - b(v_h, p_h) = \int_\Omega f \cdot v_h \, d\Omega \quad \forall v_h \in V_h , \\ &b(u_h, q_h) = 0 \quad \forall q_h \in Q_h . \end{aligned} \tag{4.45}$$

As we have observed above, we only have to check condition (3.55) in order to have solvability, stability and optimal error bounds. The following theorem, known as 'Fortin's trick' (see [10]) is often useful in order to prove (3.55) (see e.g. [11]).

THEOREM 4.1. Assume that (3.48) holds and that, for every h, we can build a linear operator $\Pi_h : V \to V_h$ with the following properties:

$$b(v - \Pi_h v, q_h) = 0 \quad \forall v \in V \quad \forall q_h \in Q_h , \tag{4.46}$$

$$\exists \gamma > 0 \quad \text{such that} \quad \|\Pi_h v\|_V \leq \gamma \|v\|_V \quad \forall v \in V , \tag{4.47}$$

where γ is independent of h. Then (3.55) holds with $\beta^* = \bar{\beta}/\gamma$.

PROOF. We have for every $q_h \in Q_h$

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq \sup_{v \in V} \frac{b(\Pi_h v, q_h)}{\|\Pi_h v\|_V}$$

$$= \sup_{v \in V} \frac{b(v, q_h)}{\|\Pi_h v\|_V} \quad \text{(use (4.46))}$$

$$\geq \sup_{v \in V} \frac{b(v, q_h)}{\gamma \|v\|_V} \quad \text{(use (4.47))}$$

$$\geq \bar{\beta}/\gamma \|q_h\|_Q \quad \text{(use (3.48))}\,, \tag{4.48}$$

where the first inequality holds since the image $\Pi_h(V)$ is contained in V_h. □

In many cases, it will actually be sufficient to prove that for every $q_h \in Q_h$ we have

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq \kappa \|q_h - \bar{q}_h\|_0\,, \tag{4.49}$$

where κ is independent of h and

$$\bar{q}_h = \{L^2\text{-projection of } q_h \text{ onto the space } \mathscr{L}_0^0 \text{ of piecewise constants}\}\,. \tag{4.50}$$

Here for instance we present two classes of discretizations for which (4.49) implies (3.55).

PROPOSITION 4.1. Assume that $Q_h \subset C^0(\Omega)$ (and piecewise polynomial) and that V_h is locally first order accurate in the sense that for every $v \in V$ there exists a $v^I \in V_h$ with

$$\|v - v^I\|_{L^2} \leq c_1 h \|v\|_V\,, \tag{4.51}$$

$$\|v^I\|_V \leq c_2 \|v\|_V\,. \tag{4.52}$$

Assume finally that the decomposition is quasi-uniform, in the sense that the maximum diameter h is bounded by $c_3 h_{\min}$ ($h_{\min}$ is the minimum diameter of the element). Then (4.49) implies (3.55).

PROOF. We note that for every $q_h \in Q_h$ there exists a $\bar{v} \in V$ such that

$$\frac{b(\bar{v}, q_h)}{\|\bar{v}\|_V} \geq \bar{\beta} \|q_h\|_Q\,. \tag{4.53}$$

Hence

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq \frac{b(\bar{v}^I, q_h)}{\|\bar{v}^I\|_V}$$

$$= \frac{b(\bar{v}^I - \bar{v}, q_h)}{\|\bar{v}^I\|_V} + \frac{b(\bar{v}, q_h)}{\|\bar{v}^I\|_V} \quad (\pm \bar{v})$$

$$\geq \frac{b(\bar{v}^I - \bar{v}, q_h)}{c_2 \|\bar{v}\|_V} + \frac{b(\bar{v}, q_h)}{c_2 \|\bar{v}\|_V} \quad \text{(use (4.52))}$$

$$\geqslant \frac{b(v^I - \bar{v}, q_h)}{c_2 \|\bar{v}\|_V} + \bar{\beta}\|q_h\|_Q \quad \text{(use (4.53))}$$

$$= \frac{\int_\Omega q_h \operatorname{div}(\bar{v}^I - \bar{v})\, d\Omega}{c_2\|\bar{v}\|_V} + \bar{\beta}\|q_h\|_Q \quad \text{(definition of } b\text{)}$$

$$= -\frac{\int_\Omega (\bar{v} - \bar{v}^I)\cdot \operatorname{grad} q_h\, d\Omega}{c_2\|\bar{v}\|_V} + \bar{\beta}\|q_h\|_Q \quad \text{(integrate by parts)}$$

$$\geqslant \bar{\beta}\|q_h\|_Q - \frac{\|\bar{v} - \bar{v}^I\|_0}{c_2\|\bar{v}\|_V}\|\operatorname{grad} q_h\|_0$$

$$\geqslant \bar{\beta}\|q_h\|_Q - \frac{c_1}{c_2}\|\operatorname{grad} q_h\|_0 h \quad \text{(use (4.51))}. \tag{4.54}$$

Now a simple scaling argument (using the quasi-uniformity of the mesh) shows that

$$h\|\operatorname{grad} q_h\|_0 \leqslant c_4\|q_h - \bar{q}_h\|_0 \tag{4.55}$$

and from (4.54) and (4.55) we have

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geqslant \bar{\beta}\|q_h\|_0 - c_5\|q_h - \bar{q}_h\|_0 . \tag{4.56}$$

It is now clear that (4.49) and (4.56) imply

$$\left(1 + \frac{\kappa}{c_5}\right)\left(\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V}\right) \geqslant \frac{\bar{\beta}\cdot\kappa}{c_5}\|q_h\|_0 \tag{4.57}$$

and (3.55) holds with $\beta^* = \bar{\beta}\kappa/c_5/(1 + \kappa/c_5)$. □

REMARK 4.1. The quasi-uniformity assumption is actually not necessary. We used it only in order to simplify the argument. See [1, 12] for the general case.

PROPOSITION 4.2. Assume that we know that for all $\bar{q}_h \in \mathcal{L}_0^0$

$$\sup_{v_h \in V_h} \frac{b(v_h, \bar{q}_h)}{\|v_h\|_V} \geqslant \gamma_1\|\bar{q}_h\|_0 . \tag{4.58}$$

Then (4.49) implies (3.55).

PROOF. We have, for every $q_h \in Q_h$,

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} = \sup_{v_h \in V_h}\left\{\frac{b(v_h, q_h - \bar{q}_h)}{\|v_h\|_V} + \frac{b(v_h, \bar{q}_h)}{\|v_h\|_V}\right\} \quad (\pm\bar{q}_h)$$

$$\geqslant \sup_{v_h \in V_h} \frac{b(v_h, \bar{q}_h)}{\|v_h\|_V} - \sup_{v_h \in V_h} \frac{b(v_h, q_h - \bar{q}_h)}{\|v_h\|_V}$$

$$\geqslant \gamma_1\|\bar{q}_h\|_0 - \|q_h - \bar{q}_h\|_0 . \tag{4.59}$$

Now from (4.59) and (4.49) we deduce (as in (4.57)) that

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq \frac{\kappa\gamma_1}{1+\kappa} \|\bar{q}_h\|_0 . \tag{4.60}$$

Finally from (4.49) and (4.60) the result follows. □

REMARK 4.2. As we can see, (4.49) implies the inf–sup condition in an impressive number of cases: using (4.49) basically all the approximations with continuous pressure can be considered, as well as all the choices of V_h that give a stable pair when used with a piecewise constant pressure. This second case for instance holds true, if for triangular elements V_h contains the space of piecewise quadratic functions, and if for quadrilateral elements, V_h contains the reduced (8 nodes) biquadratic functions. If we succeed in giving an easy test for (4.49) then we can treat a very wide number of cases. Such a test will be a consequence of Proposition 4.3 below.

REMARK 4.3. More generally the condition we need on V_h (in order to have (4.58)) is the following: we can use as degrees of freedom the values $v \cdot n$ (normal component of velocity) at midpoints of edges (respectively, faces in $\mathbb{R}^3$). Indeed, if this is the case, we can (roughly) consider a 'Fortin interpolator' Π_h such that

$$\int_e (v - \Pi_h v) \cdot n \, de = 0 \quad \text{for all edges (faces) } e \tag{4.61}$$

and

$$\text{'}v = \Pi_h v\text{'} \quad \text{for the other degrees of freedom .} \tag{4.62}$$

From (4.61) and (4.62) we now have, for every $\bar{q}_h \in \mathcal{L}_0^0$,

$$\begin{aligned} \int_K \operatorname{div}(v - \Pi_h v)\bar{q}_h \, dK &= \int_{\partial K} (v - \Pi_h v) \cdot n \bar{q}_h \, de \quad \text{(use Gauss' theorem)} \\ &= 0 \quad \text{(use (4.61) on each } e\text{)} \end{aligned} \tag{4.63}$$

for every element K. This gives (4.46) for $Q_h = \mathcal{L}_0^0$, which is the essential step (through Theorem 4.1) in order to have (4.58). Note however that the actual proof of this fact has some more technicalities (see [1, 13] for similar arguments). It is clear by now that, in designing a new element, to satisfy (4.49) is the essential step in almost every case.

PROPOSITION 4.3. Assume that $\mathcal{P}_1$ and $\mathcal{P}_2$ are unions of elements, with $\mathcal{P}_1 \cap \mathcal{P}_2 = \emptyset$. Assume that, for $i = 1, 2$, we have for all $q_h \in Q_h$

$$\sup_{\substack{v_h \in V_h \\ v_h = 0 \text{ in } \Omega \setminus \mathcal{P}_i}} \frac{\int_{\mathcal{P}_i} q_h \operatorname{div} v_h}{\|v_h\|_V} \geq \kappa_i \|q_h - \bar{q}_h\|_{L^2(\mathcal{P}_i)} , \tag{4.64}$$

then

$$\sup_{\substack{v_h \in V_h \\ v_h = 0 \text{ in } \Omega \setminus (\mathscr{P}_1 \cup \mathscr{P}_2)}} \frac{\int_{\mathscr{P}_1 \cup \mathscr{P}_2} q_h \operatorname{div} v_h}{\|v_h\|_V} \geq \kappa \|q_h - \bar{q}_h\|_{L^2(\mathscr{P}_1 \cup \mathscr{P}_2)} \tag{4.65}$$

for all $q_h \in Q_h$ and with

$$\kappa \geq \min(\kappa_1, \kappa_2) . \tag{4.66}$$

PROOF. The proof is elementary. From (4.64) we have, for all $q_h \in Q_h$, two elements $v^i \in (H_0^1(\mathscr{P}_i))^2 \cap V_h$ $(i = 1, 2)$, such that

$$b(v^i, q_h) = \kappa_i \|q_h - \bar{q}_h\|^2_{L^2(\mathscr{P}_i)} \tag{4.67}$$

and

$$\|v^i\|_V \leq \|q_h - \bar{q}_h\|_{L^2(\mathscr{P}_i)} . \tag{4.68}$$

Taking $v = v^1 + v^2$ we have

$$b(v, q_h) = \sum_{i=1}^{2} \kappa_i \|q_h - \bar{q}_h\|^2_{L^2(\mathscr{P}_i)} \geq \kappa \|q_h - \bar{q}_h\|^2_{L^2(\mathscr{P}_1 \cup \mathscr{P}_2)} \quad \text{(use (4.66))} \tag{4.69}$$

and

$$\|v\|^2_V = \|v^1\|^2_V + \|v^2\|^2_V \leq \|q_h - \bar{q}_h\|^2_{L^2(\mathscr{P}_1 \cup \mathscr{P}_2)} \tag{4.70}$$

and (4.69) with (4.70) implies (4.65). □

REMARK 4.4. The condition $\mathscr{P}_1 \cap \mathscr{P}_2 = \emptyset$, as we can see from the proof, is not crucial. Its only purpose is to avoid a factor 2 in (4.69) and (4.70). However, we always think of using the result in the 'disjoint' case.

REMARK 4.5. In (4.64), (4.65) the choice of $\bar{q}_h$ as an element by element projection onto the space $\mathscr{L}_0^0$ of piecewise constants is unnecessary. We might as well use a 'patch by patch' projection, that is, we might assume that $\bar{q}_h$ is constant in every patch (and equal to the mean value of q_h). On the other hand, the choice $\bar{q}_h = 0$ (which would give directly the inf–sup condition without passing through Propositions 4.1 and 4.2) is not allowed. If q_h has zero mean value on $\mathscr{P}_1 \cup \mathscr{P}_2$ it does not necessarily have zero mean value separately on $\mathscr{P}_1$ and on $\mathscr{P}_2$. But (4.64) is unrealistic if the right-hand side does not have zero mean value. Finally let us note that, if $\bar{q}_h$ is the element by element projection and $\tilde{q}_h$ is the patch by patch projection, then

$$\|q - \tilde{q}_h\|_{L^2} \geq \|q - \bar{q}_h\|_{L^2} .$$

REMARK 4.6. Proposition 4.3 deals with two patches. It is clear that the argument applies as well to any finite number of patches: the smallest κ_i gives the global κ.

REMARK 4.7. It is very important to point out that conditions (4.64) do not depend on the

size of the patches. Assume that we have, for a given patch, say, of size one

$$\sup_{v_h \in V_h(\mathcal{P})} \frac{\int_{\mathcal{P}} q_h \operatorname{div} v_h}{\|v_h\|_{H^1(\mathcal{P})}} \geq \|q_h - \tilde{q}_h\|_{L^2(\mathcal{P})} \tag{4.71}$$

for all $q_h \in Q_h(\mathcal{P})$, where $V_h(\mathcal{P})$ is a finite element subspace of $(H_0^1(\mathcal{P}))^2$ and $Q_h(\mathcal{P})$ is a finite element subspace of $L^2(\mathcal{P})$ and finally $\tilde{q}_h$ is the mean value of q_h on $\mathcal{P}$. If we shrink $\mathcal{P}$ to a small patch $\mathcal{P}^s$ of size h by the change of variable,

$$\mathcal{P} \ni x = \xi/h\,, \quad \xi \in \mathcal{P}^s\,, \tag{4.72}$$

and if we change the finite element spaces accordingly, we have

$$\sup_{v_h^s \in V_h^s} \frac{\int_{\mathcal{P}^s} q_h^s \operatorname{div} v_h^s}{\|v_h^s\|_{H^1(\mathcal{P}^s)}} \geq \kappa \|q_h^s - \tilde{q}_h^s\|_{L^2(\mathcal{P}^s)} \tag{4.73}$$

exactly with the same κ as in (4.71).

The same is true if, instead of the change of variable (4.72), we apply any other change of variable which is 'affine', that is with constant Jacobian: actually a distortion in the shape might change κ to some $\delta \cdot \kappa$, where δ depends on the amount of distortion (essentially if J is the Jacobian matrix, δ will depend on $|J^{-1}| \cdot \|J\|^2$ and $|J| \cdot \|J^{-1}\|^2$), and for reasonable distortions κ will remain greater than zero.

Proposition 4.3 and all the remarks after it suggest now a strategy for checking the inf–sup condition, if we have a continuous pressure field or if we know already that the velocity space V_h under consideration can be used with piecewise constant pressures. Assume that we can find a finite number of 'representative patches', $\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_R$, such that one can cover the decomposition with affine images of the $\mathcal{P}_i$'s. For each patch $\mathcal{P}_i$ we then check if the discrete problem is uniquely solvable on the patch (for both velocity and pressure) with homogeneous Dirichlet boundary conditions for velocities and obviously discarding the constant value of the pressure on the patch. If this is true, then a constant κ_i must exist; we do not need to compute it, we just want to know that it exists. Finally, the smallest κ_i will give the constant κ in (4.49) and the inf–sup condition will hold.

5. Concluding remarks

Our objective in this paper was to review and discuss conditions for the stability of mixed finite element formulations. We also presented a numerical test that can be employed to check whether a given mixed finite element formulation for the general Stokes problem satisfies the mathematical conditions of stability and optimal error bounds.

While the general mathematical theory for mixed formulations is quite well established, both for the continuous and discretized problems, the actual detailed use of that theory for the design and analysis of mixed finite element formulations can be a very difficult task. We note that quite effective mixed finite elements that satisfy the mathematical conditions of stability

and optimal error bounds are available for the solution of incompressible fluid flow [1] and the analysis of incompressible or almost incompressible solid media [1, 14, 15]. However, the situation is quite different, for example, in the field of analysis of plate and shell structures [16]. Here numerous mixed finite elements have been proposed but mathematical analyses are hardly available. Indeed the construction of effective mixed plate and shell elements that can be analysed and satisfy the mathematical conditions of stability and optimal error bounds is very difficult, and such elements are now under active research, see for example [16–18].

References

[1] F. Brezzi and M. Fortin, Mixed and Hybrid Finite Element Methods, to appear.
[2] K.J. Bathe, Finite Element Procedures in Engineering Analysis (Prentice-Hall, Englewood Cliffs, NJ, 1982).
[3] O.C. Zienkiewicz, S. Qu, R.L. Taylor and S. Nakazawa, The patch test for mixed formulations, Internat. J. Numer. Methods Engrg. 23 (1986) 1873–1883.
[4] O.C. Zienkiewicz and D. Lefebvre, Three-field mixed approximation and the plate bending problem, Comm. Appl. Numer. Methods 3 (1987) 301–309.
[5] F. Brezzi, On the existence, uniqueness, and approximation of saddle-point problems arising from Lagrange multipliers, RAIRO B-R2 (1974) 129–151.
[6] I. Babuška, The finite element method with Lagrangian multipliers, Numer. Math. 20 (1973) 179–192.
[7] D.N. Arnold, Discretization by finite elements of a model parameter dependent problem, Numer. Math. 37 (1981) 405–421.
[8] O. Ladyzhenskaya, The Mathematical Theory of Viscous Incompressible Flows (Gordon and Breach, London, 1969).
[9] R. Temam, Navier–Stokes Equations (North-Holland, Amsterdam, 1977).
[10] M. Fortin, An analysis of the convergence of mixed finite element method, RAIRO Anal. Numér. 11 (1977) 341–354.
[11] F. Brezzi and K.J. Bathe, Studies of finite element procedures—The inf–sup condition, equivalent forms and applications, in: K.J. Bathe and D.R.J. Owen, Eds., Reliability of Methods for Engineering Analysis (Pineridge, Swansea, 1986).
[12] R. Verfürth, Error estimates for a mixed finite element approximation of the Stokes equations, RAIRO 18 (1984) 175–182.
[13] M. Fortin, Old and new finite elements for incompressible flows, Internat. J. Numer. Methods Fluids 1 (1981) 347–364.
[14] J.T. Oden and N. Kikuchi, Finite element methods for constrained problems in elasticity, Internat. J. Numer. Methods Engrg. 18 (1982) 701–725.
[15] T. Sussman and K.J. Bathe, A finite element formulation for nonlinear incompressible elastic and inelastic analysis, Comput. & Structures 26 (1/2) (1987) 357–409.
[16] A.K. Noor, T. Belytschko and J.C. Simo, eds. Analytical and Computational Models of Shells (ASME, New York, 1989).
[17] F. Brezzi, K.J. Bathe and M. Fortin, Mixed-interpolated elements for Reissner/Mindlin plates, Internat. J. Numer. Methods Engrg. 28 (1989) 1787–1801.
[18] K.J. Bathe, S.W. Cho, M.L. Bucalem and F. Brezzi, On our MITC plate bending/shell elements, in: A.K. Noor, T. Belytschko and J.C. Simo, eds., Analytical and Computational Models of Shells (ASME, New York, 1989).