Brezzi • Fortin
Mixed and Hybrid Finite Element Methods

Mixed and hybrid finite element methods are widely employed in many areas of computational engineering. Their mathematical properties have been thoroughly studied and a general theory has emerged. This book gives a unified presentation of this mathematical theory. It presents all the standard results, in particular the construction of elements adapted to the mixed formulations and also the more recent techniques of stabilization. It also introduces the treatment of continuity conditions at interfaces through multipliers, which leads to better properties to the discrete problem.

Besides the general theory, many applications are considered. In particular, elasticity problems and moderately thick plates problems are treated extensively. An entire chapter is devoted to the approximation of incompressible flows or incompressible elasticity, providing a state-of-the-art access to the field.

ISBN 0-387-97582-9
ISBN 3-540-97582-9

SPRINGER SERIES IN COMPUTATIONAL MATHEMATICS

Franco Brezzi
Michel Fortin

Mixed and Hybrid Finite Element Methods

Springer-Verlag

Springer Series in
Computational
Mathematics

15

Editorial Board

R. L. Graham, Murray Hill (NJ)
J. Stoer, Würzburg
R. Varga, Kent (Ohio)

Franco Brezzi Michel Fortin

Mixed and Hybrid Finite Element Methods

With 65 Illustrations

Springer-Verlag
New York Berlin Heidelberg London
Paris Tokyo Hong Kong Barcelona

Franco Brezzi
University of Pavia
Institute of Numerical Analysis
5 Corso Carlo Alberto
I-27100 Pavia
Italy

Michel Fortin
Département de Mathématiques
et de Statistique
Université Laval
Quebec G1K 7P4
Canada

Mathematics Subject Classification: 73XX, 76XX

Library of Congress Cataloging-in-Publication Data
Brezzi, Franco.
Mixed and hybrid finite element methods / Franco Brezzi, Michel Fortin.
p. cm -- (Springer series in computational mathematics ; 15)
Includes bibliographical references and index.
ISBN 0-387-97582-9 (alk. paper)
1. Finite element method. I. Brezzi, F. (Franco), 1945– .
II Title. III. Series.
TA347.F5F68 1991
620'.001'51535--dc20 91-10909

Printed on acid-free paper.

© 1991 Springer-Verlag New York Inc.
All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher (Springer-Verlag New York, Inc., 175 Fifth Avenue, New York, NY 10010, USA), except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.
The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Photocomposed copy prepared from the author's TEX file.
Printed and bound by Edwards Brothers, Inc., Ann Arbor, Michigan.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-97582-9 Springer-Verlag New York Berlin Heidelberg
ISBN 3-540-97582-9 Springer-Verlag Berlin Heidelberg New York

Preface

When, a few years ago, we began the redaction of this book, we had the naive thought that the theory of mixed and hybrid finite element methods was ripe enough for a unified presentation. We soon realized that things were not so simple and that, if basic facts were known, many obscure zones remained in many applications. Indeed the literature about nonstandard finite element method is still evolving rapidly and this book cannot pretend to be complete. We would rather like to lead the reader through the general framework in which development is taking place.

We have therefore built our presentation around a few classical examples: Dirichlet's problem, Stokes problem, linear elasticity, ... They are sketched in Chapter I and basic methods to approximate them are presented in Chapter IV, following the general theory of Chapter II and using finite element spaces of Chapter III. Those four chapters are therefore the essential part of the book. They are complemented by the following three chapters which present a more detailed analysis of some problems.

Chapter V comes back to mixed approximations of Dirichlet's problem and analyses, in particular, the (lambda)-trick that enables to make the link between mixed methods and more classical non-conforming methods. Chapter VI deals with Stokes problem and Chapter VII with linear elasticity and the Mindlin-Reissner plate model.

The reader should not look here for practical implementation tricks. Our goal was to provide an analysis of the methods in order to understand their properties as thoroughly as possible. We refer, among others, to the recent work of BATHE [A] or HUGHES [A] or to the classical and indispensable book

of ZIENKIEWICZ [A] for practical considerations. We are of course strongly indebted to CIARLET [A] which remains the essential reference for the classical theory of finite element methods. Finally we also refer to ROBERTS–THOMAS [A] for another presentation of mixed methods.

This book would never have come to its end without the help, encouragement and criticisms of our friends and colleagues. We must also thank all those who took the time reading the first draft of our manuscript and proposed significant improvements. We hope that the final result will better than what one might expect, according to the quotations thereafter, of the hybrid resulting of a collaboration between Pavia and Québec.

Apris atque sui setosus nascitur hybris. (C. Plinius Caecilius Secondus.)

Mixtumque genus prolesque biformis. (Publius Virgilius Maro.)

Contents

Preface v

Chapter I: Variational Formulations and Finite Element Methods 1
§1. Classical Methods 1
§2. Model Problems and Elementary Properties of Some Functional Spaces 4
§3. Duality Methods 11
3.1. Generalities 11
3.2. Examples for symmetric problems 13
3.3. Duality methods for nonsymmetric bilinear forms 23
§4. Domain Decomposition Methods, Hybrid Methods 24
§5. Augmented Variational Formulations 30
§6. Transposition Methods 33
§7. Bibliographical remarks 35

Chapter II: Approximation of Saddle Point Problems 36
§1. Existence and Uniqueness of Solutions 36
1.1. Quadratic problems under linear constraints 37
1.2. Extensions of existence and uniqueness results 44
§2. Approximation of the Problem 51
2.1. Basic results 51
2.2. Error estimates for the basic problem 54
2.3. The inf-sup condition: criteria 57
2.4. Extensions of error estimates 61
2.5. Various generalizations of error estimates 63

2.6. Perturbations of the problem, nonconforming methods 65
2.7. Dual error estimates 70
§3. Numerical Properties of the Discrete Problem 73
3.1. The matrix form of the discrete problem 73
3.2. Eigenvalue problem associated with the inf–sup condition ... 75
3.3. Is the inf–sup condition so important? 78
§4. Solution by Penalty Methods, Convergence of Regularized Problems 80
§5. Iterative Solution Methods. Uzawa's Algorithm 87
5.1. Standard Uzawa's algorithm 87
5.2. Augmented Lagrangian algorithm 87
§6. Concluding Remarks 88

Chapter III: Function Spaces and Finite Element Approximations 89
§1. Properties of the spaces $H^s(\Omega)$ and $H(\mathrm{div};\Omega)$ 89
1.1. Basic results 89
1.2. Properties relative to a partition of Ω 93
1.3. Properties relative to a change of variables 95
§2. Finite Element Approximations of $H^1(\Omega)$ and $H^2(\Omega)$ 99
2.1. Conforming methods 99
2.2. Nonconforming methods 107
2.3. Nonpolynomial approximations: Spaces $\mathfrak{L}_k^s(\mathfrak{E}_h)$ 111
2.4. Scaling arguments 111
§3. Approximations of $H(\mathrm{div};\Omega)$ 113
3.1. Simplicial approximations of $H(\mathrm{div};K)$ 113
3.2. Rectangular approximations of $H(\mathrm{div};K)$ 119
3.3. Interpolation operator and error estimates 124
3.4. Approximation spaces for $H(\mathrm{div};\Omega)$ 131
§4. Concluding Remarks 132

Chapter IV: Various Examples 133
§1. Nonstandard Methods for Dirichlet's Problem 134
1.1. Description of the problem 134
1.2. Mixed finite element methods for Dirichlet's problem 135
1.3. Primal hybrid methods 140
1.4. Dual hybrid methods 149
§2. Stokes Problem 155
§3. Elasticity Problems 159
§4. A Mixed Fourth-Order Problem 164
4.1. The ψ–ω biharmonic problem 164
§5. Dual Hybrid Methods for Plate Bending Problems 167

Chapter V: Complements on Mixed Methods for Elliptic Problems 177
§1. Numerical Solutions 177
1.1. Preliminaries 177
1.2. Interelement multipliers 178

§2. A Brief Analysis of the Computational Effort 182
§3. Error Analysis for the Multiplier 186
§4. Error Estimates in Other Norms 191
§5. Application to an Equation Arising from Semiconductor Theory 193
§6. How Things Can Go Wrong 195
§7. Augmented Formulations 198

Chapter VI: Incompressible Materials and Flow Problems 200
§1. Introduction 201
§2. The Stokes Problem as a Mixed Problem 202
2.1. Mixed Formulation 202
§3. Examples of Elements for Incompressible Materials 206
3.1. Simple examples 207
§4. Standard Techniques of Proof for the inf–sup Condition 219
4.1. General results 224
4.2. Higher order methods 227
§5. Macroelement Techniques and Spurious Pressure Modes 228
5.1. Some remarks about spurious pressure modes 228
5.2. An abstract convergence result 231
5.3. Macroelement techniques 235
5.4. The bilinear velocity-constant pressure (Q_1–P_0) element 240
5.5. Other stabilization procedures, (Augmented Formulations) 246
§6. An Alternative Technique of Proof and Generalized Taylor–Hood Element 252
§7. Nearly Incompressible Elasticity, Reduced Integration Methods and Relation with Penalty Methods 258
7.1. Variational formulations and admissible discretizations 258
7.2. Reduced integration methods 259
7.3. Effects of inexact integration 263
§8. Divergence-Free Basis, Discrete Stream Functions 267
§9. Other Mixed and Hybrid Methods for Incompressible Flows 272

Chapter VII: Other Applications 274
§1. Mixed Methods for Linear Thin Plates 274
§2. Mixed Methods for Linear Elasticity Problems 282
§3. Moderately Thick Plates 295
3.1. Generalities 295
3.2. Discretization of the problem 303
3.3. Continuous Pressure Approximations 322
3.4. Discontinuous Pressure Approximations 322

References 324

Index 344

I

Variational Formulations and Finite Element Methods

Although we shall not define in this chapter mixed and hybrid (or other non-standard) finite element methods in a very precise way, we would like to situate them in a sufficiently clear setting. As we shall see, boundaries between different methods are sometimes rather fuzzy. This will not be a real drawback if we nevertheless know how to apply correctly the principles underlying their analysis.

After having briefly recalled some basic facts about classical methods, we shall present a few model problems. The study of these problems will be the kernel of this book. We shall thereafter rapidly recall basic principles of *duality theory* as this will be our starting point to introduce mixed methods. *Domain decomposition* methods (allied to duality) will lead us to hybrid methods. Finally, we shall present a few ideas about transposed formulations as they can help to understand some of the weak problems generated by the previous methods.

I.1 Classical Methods

We recall here in a very simplified way some results about optimization methods and the classical finite element method. Such an introduction cannot be complete and does not want to be. We refer the reader to CIARLET [A] or RAVIART–THOMAS [D], among others, where standard finite element methods are clearly exposed. We also refer to DAUTRAY–LIONS [A] where an exhaustive analysis of many of our model problems can be found.

Let us then consider a very common situation where the solution of a physical problem minimizes some functional (usually an "energy functional"),

in a "well chosen" space of admissible functions V that we take for the moment as a Hilbert space,

$$\inf_{v \in V} J(v). \tag{1.1}$$

If the functional $J(\cdot)$ is differentiable (cf. EKELAND–TEMAM [A] for instance), the minimum (whenever it exists) will be characterized by *a variational equation*

$$\langle J'(u), v\rangle_{V' \times V} = 0, \qquad \forall v \in V, \tag{1.2}$$

where $\langle \cdot, \cdot \rangle_{V' \times V}$ denotes duality between V and its topological dual V', the derivative $J'(u)$ at point u being considered as a linear form on V.

A classical method, Ritz's method, to approximate the solution of (1.1) consists in looking for $u_m \in V_m$, where V_m is a finite-dimensional subspace of V, which is solution of the problem

$$\inf_{v_m \in V_m} J(v_m) \tag{1.3}$$

or, differentiating,

$$\langle J'(u_m), v_m\rangle_{V' \times V} = 0, \qquad \forall v_m \in V_m. \tag{1.4}$$

Let us consider to fix ideas the quadratic functional

$$J(v) = \tfrac{1}{2} a(v, v) - L(v), \tag{1.5}$$

where $a(\cdot, \cdot)$ is a bilinear form on V, which we suppose to be continuous and symmetric, and $L(\cdot)$ is a linear form on V. The variational equation (1.4) can then be written as: $u_m \in V_m$ and

$$a(u_m, v_m) = L(v_m), \qquad \forall v_m \in V_m. \tag{1.6}$$

If a basis $w_1, w_2, \ldots, w_m$ of V_m is chosen and if one writes

$$u_m = \sum_{i=1}^{m} \alpha_i w_i, \tag{1.7}$$

problem (1.6) is reduced to the solution of the linear system

$$\sum_{i=1}^{m} a_{ij} \alpha_i = b_j, \qquad 1 \le j \le m, \tag{1.8}$$

where one defines

$$a_{ij} = a(w_i, w_j), \qquad b_j = L(w_j). \tag{1.9}$$

This formulation can be extended to the case where the bilinear form $a(\cdot,\cdot)$ *is not symmetric* and where problem (1.6) no longer corresponds to a minimization problem. This is then usually called a Galerkin's method. Let us recall that problems of type (1.6) will have a unique solution if, in particular, the bilinear form $a(\cdot,\cdot)$ is *coercive*, that is, if there exists a positive real number α such that for all v in V

$$a(v,v) \geq \alpha \|v\|_V^2 . \tag{1.10}$$

The above-described methodology is very general and classical. We can consider the finite element method to be a special case of it in the following sense.

The finite element method is a general technique to build finite-dimensional subspaces of a Hilbert space V in order to apply the Ritz–Galerkin method to a variational problem.

This technique is based on a few simple ideas. The fundamental one is the *partition of the domain* Ω in which the problem is posed into a set of "simple" subdomains, called elements. These elements are usually triangles, quadrilaterals, tetrahedra, etc. A space V *of functions defined on* Ω *is then approximated* by "simple" functions, defined on each subdomain with suitable matching conditions at interfaces. Simple functions are usually polynomials or functions obtained from polynomials by a change of variables.

This is of course a very summarized way of defining finite elements and this is for sure not the best way to understand it from the computational point of view. We shall come back to this in Chapter III with a much more workable approach.

The point that we want to emphasize here is the following. *A finite element method can only be considered in relation with a variational principle and a functional space. Changing the variational principle and the space in which it is posed leads to a different finite element approximation (even if the solution for the continuous problem can remain the same).*

In the remainder of this chapter, we shall see how different variational formulations can be built for the same physical problem. Each of these formulations will lead to a new setting for finite element approximations. *The common point of the methods analyzed in this book is that they are founded on a variational principle expressing an equilibrium* (saddle point) *condition* rather than on a minimization principle. We shall now try to see, on some examples, how such equilibrium principles can be built.

I.2 Model Problems and Elementary Properties of Some Functional Spaces

The aim of this section is to introduce notation and to present four model problems that will underlie almost all cases analyzed in the book. They will be the Dirichlet problem for the Laplace equation, linear elasticity, the Stokes problem, and finally a fourth-order problem modeling the deflection of a thin clamped plate. These problems are closely interrelated and methods to analyze them will also be.

We shall present, in this section, the most classical variational formulation of these problems. The following sections will lead us to less standard forms.

We shall assume, in our exposition, that the problems are posed in a domain Ω of $\mathbb{R}^n$, with a sufficiently smooth boundary $\partial\Omega = \Gamma$ (for instance a Lipschitz continuous boundary). In practice $n = 2$ or 3, and we shall present most of our examples in a two-dimensional setting for the sake of simplicity. In the problems considered here, working in $\mathbb{R}^2$ rather than in $\mathbb{R}^3$ is not really restrictive and extensions are generally straightforward. (This is however not always the case for numerical methods).

Let us first set a few notation. We shall constantly use the Sobolev spaces (ADAMS [A], LIONS–MAGENES [A], NEČAS [A]). They are based on

$$L^2(\Omega) = \left\{ v \middle| \int_\Omega |v|^2 dx = \|v\|^2_{L^2(\Omega)} < +\infty \right\}, \tag{2.1}$$

the space of square integrable functions on Ω. (These functions **must of course** be measurable). We then define in general, for m integer ≥ 0,

$$H^m(\Omega) = \left\{ v \middle| \, D^\alpha v \in L^2(\Omega), \, \forall |\alpha| \leq m \right\}, \tag{2.2}$$

where

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}, \; |\alpha| = \alpha_1 + \cdots + \alpha_n,$$

these derivatives being taken in the sense of distributions. On this space, we shall use the semi-norm

$$|v|^2_{m,\Omega} = \sum_{|\alpha|=m} |D^\alpha v|^2_{L^2(\Omega)}, \tag{2.3}$$

and the norm

$$\|v\|^2_{m,\Omega} = \sum_{k \leq m} |v|^2_{k,\Omega}. \tag{2.4}$$

The space $L^2(\Omega)$ is then $H^0(\Omega)$ and we shall usually write $|v|_{0,\Omega}$ to denote its norm $\|v\|_{L^2(\Omega)}$. Let us denote, as usual, by $\mathcal{D}(\Omega)$ the space of indefinitely

differentiable functions having a compact support in Ω, and by $H_0^m(\Omega)$ the completion of $\mathfrak{D}(\Omega)$ for the topology defined by the norm (2.4). If the boundary is smooth enough (e.g., Lipschitz continuous boundary), this simple definition will coincide, without troublesome pathologies, with more sophisticated ones.

Among the spaces introduced, the most commonly used, apart from $L^2(\Omega)$, will be $H^1(\Omega)$, $H_0^1(\Omega)$, $H^2(\Omega)$, and $H_0^2(\Omega)$.

If the boundary $\partial\Omega$ is sufficiently smooth, (we consider only Lipschitz continuous boundaries), one can show that there exists an operator $\gamma_0 : H^1(\Omega) \mapsto L^2(\Gamma)$, linear and continuous, such that $\gamma_0 v$ = trace of v on Γ for every v smooth [say, to fix the ideas, for every $v \in C^1(\bar{\Omega})$]. It then seems natural to call $\gamma_0 v$ "the trace of v on Γ", and denote it by $v|_\Gamma$ even if v is a general function in $H^1(\Omega)$. A deeper analysis shows that by taking all the traces of all the functions of $H^1(\Omega)$ one does not obtain the whole space $L^2(\Omega)$ but only a subspace of it. Further investigations show that such a subspace contains $H^1(\Gamma)$ as a proper subset. Hence we have,

$$H^1(\Gamma) \subset \gamma_0(H^1(\Omega)) \subset L^2(\Gamma) \equiv H^0(\Gamma),$$

where every inclusion is strict. It is finally recognized that the space $\gamma_0(H^1(\Omega))$ belongs to a family of spaces $H^s(\Gamma)$ (that we are not going to define here) and corresponds exactly to the value $s = 1/2$. Hence we have

$$H^{1/2}(\Gamma) = \gamma_0(H^1(\Omega))$$

with

$$\|g\|_{H^{1/2}(\Gamma)} = \inf_{\substack{v \in H^1(\Omega) \\ \gamma_0 v = g}} \|v\|_{H^1(\Omega)}.$$

In a similar way, one can see that the traces of functions in $H^2(\Omega)$ belong to a space $H^s(\Gamma)$ for $s = 3/2$. We may therefore set

$$H^{3/2}(\Gamma) = \gamma_0(H^2(\Omega)),$$

$$\|g\|_{H^{3/2}(\Gamma)} = \inf_{\substack{v \in H^2(\Omega) \\ \gamma_0 v = g}} \|v\|_{H^2(\Omega)}.$$

This can be generalized to the traces of higher-order derivatives. For instance, if the boundary Γ is smooth enough, one can define $\partial v/\partial n\big|_\Gamma \in H^{1/2}(\Gamma)$ for $v \in H^2(\Omega)$. We shall not discuss in a more precise way trace theorems on Sobolev spaces of fractional order. (The reader may refer to the authors quoted above.) Intuitively, Sobolev spaces of fractional order can be considered as having regularity properties that are intermediate between the properties of the neighboring integer-order spaces and they can indeed be defined as *interpolation spaces*. Taking this as granted, we then have

$$H_0^1(\Omega) = \left\{ v \mid v \in H^1(\Omega),\ v\big|_\Gamma = 0 \right\}, \tag{2.5}$$

$$H_0^2(\Omega) = \left\{ v \mid v \in H^2(\Omega),\ v\big|_\Gamma = 0,\ \frac{\partial v}{\partial n}\Big|_\Gamma = 0 \right\}. \tag{2.6}$$

For $v \in H_0^1(\Omega)$, we have the *Poincaré inequality*,

$$|v|_{0,\Omega} \le C(\Omega)|v|_{1,\Omega} \tag{2.7}$$

and the seminorm $|\cdot|_{1,\Omega}$ is therefore a norm on $H_0^1(\Omega)$, equivalent to $\|\cdot\|_{1,\Omega}$. We shall also need to consider functions that vanish on a part of the boundary; suppose $\Gamma = D \cup N$, a partition of Γ into disjoint parts, one then defines

$$H_{0,D}^1(\Omega) = \{v \mid v \in H^1(\Omega),\ v\big|_D = 0\} \tag{2.8}$$

and one has $H_0^1(\Omega) \subset H_{0,D}^1(\Omega) \subset H^1(\Omega)$. In Chapter III, we will discuss the properties of these spaces; the above definitions are sufficient to allow us to present some examples.

Example 2.1: *Boundary value problems for the Laplace equation.*
This is a very classical case that in fact led to the definition of Sobolev spaces. Let us consider, on $H_0^1(\Omega)$, $f \in L^2(\Omega)$ being given, the following minimization problem

$$\inf_{v \in H_0^1(\Omega)} \frac{1}{2} \int_\Omega |\underline{\text{grad}}\, v|^2 dx - \int_\Omega f v\, dx, \tag{2.9}$$

where $|\underline{\text{grad}}\, v|^2 = \left|\partial v/\partial x_1\right|^2 + \left|\partial v/\partial x_2\right|^2 = \underline{\text{grad}}\, v \cdot \underline{\text{grad}}\, v$. One shows easily (cf. CEA [A], LIONS–MAGENES [A], NEČAS [A] for instance) that this problem has a unique solution u, characterized by: $u \in H_0^1(\Omega)$ and

$$\int_\Omega \underline{\text{grad}}\, u \cdot \underline{\text{grad}}\, v\, dx = \int_\Omega f v\, dx, \qquad \forall v \in H_0^1(\Omega). \tag{2.10}$$

This solution u then satisfies, in the sense of distributions,

$$\begin{cases} -\Delta u = f \text{ in } \Omega, \\ u\big|_\Gamma = 0, \end{cases} \tag{2.11}$$

which is a standard Dirichlet problem. If $H_0^1(\Omega)$ were replaced by $H_{0,D}^1(\Omega)$, one would get instead of (2.11), a mixed type problem

$$\begin{cases} -\Delta u = f \text{ in } \Omega, \\ u\big|_D = 0, \\ \dfrac{\partial u}{\partial n}\Big|_N = 0. \end{cases} \tag{2.12}$$

We thus have Dirichlet boundary conditions on D and Neumann conditions on N. In particular, for $N = \Gamma$, we get a Neumann problem. It must be noted that minimizing (2.9) on $H^1(\Omega)$ instead of $H_0^1(\Omega)$ will define u up to an additive constant, and requires the compatibility condition $\int_\Omega f dx = 0$, which can be seen to be necessary from (2.10), taking $v \equiv 1$ in Ω.

If we denote by $H^{-1/2}(\Gamma)$ the dual space of $H^{1/2}(\Gamma)$, and we take $g \in H^{-1/2}(\Gamma)$, we can consider the functional

$$\frac{1}{2}\int_\Omega |\underline{\text{grad}}\, v|^2 dx - \int_\Omega fv\, dx - \langle g, v\rangle, \tag{2.13}$$

where the bracket $\langle\cdot,\cdot\rangle$ denotes duality between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. We shall sometimes write *formally* $\int_\Gamma gv\, ds$ instead of $\langle g, v\rangle$. Minimizing (2.13) on $H^1_{0,D}(\Omega)$ leads to the problem

$$\begin{cases} -\Delta u = f \text{ in } \Omega, \\ u|_D = 0, \\ \dfrac{\partial u}{\partial n}\Big|_N = g. \end{cases} \tag{2.14}$$

When $D = \emptyset$ the solution is defined up to an additive constant and we must choose f and g such that $\int_\Omega f\, dx - \int_\Gamma g\, ds = 0$.

These problems are among the most classical of mathematical physics and we do not have to emphasize their importance. In the following chapters we shall need to use *regularity results* for the problems introduced above. We have supposed up to now $f \in L^2(\Omega)$. For the Dirichlet problem (2.11) we could have assumed f to belong to a weaker space, namely, $f \in H^{-1}(\Omega) = (H_0^1(\Omega))'$, and nevertheless obtained $u \in H_0^1(\Omega)$. Indeed, if f is taken in $L^2(\Omega)$ and the boundary Γ is Lipschitzian and convex, one can prove (NEČAS [A]) that $u \in H^2(\Omega)$ and that

$$||u||_{2,\Omega} \le c|f|_{0,\Omega}. \tag{2.15}$$

Regularity results are essential to many approximation results and are fundamental to obtain error estimates. We refer the reader to GRISVARD [B] for the delicate questions of the regularity of the general problem (2.12) in a domain with corners. □

Example 2.2: *Linear elasticity.*
We shall try to determine the displacement $\underline{u} = \{u_1, u_2\}$ of an elastic material under the action of some external forces. We suppose the displacement to be small and the material to be isotropic and homogeneous. (CIARLET [A], MARSDEN–HUGHES [A]). The domain Ω is the initial configuration of the

body. To set our problem, we must introduce some notation from continuum mechanics. First we define the linearized strain tensor $\underline{\underline{\varepsilon}}(\underline{v})$ by

$$\varepsilon_{ij}(\underline{v}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right). \tag{2.16}$$

The trace $\mathrm{tr}(\underline{\underline{\varepsilon}})$ of this tensor is nothing but the divergence of the displacement field

$$\mathrm{tr}(\underline{\underline{\varepsilon}})(\underline{v}) = \mathrm{div}\,\underline{v}. \tag{2.17}$$

We shall also use the *deviatoric* $\underline{\underline{\varepsilon}}^D$ of the tensor $\underline{\underline{\varepsilon}}$, that is,

$$\underline{\underline{\varepsilon}}^D = \underline{\underline{\varepsilon}} - \tfrac{1}{2}\,\mathrm{tr}(\underline{\underline{\varepsilon}})\underline{\underline{\delta}}, \tag{2.18}$$

where $\underline{\underline{\delta}}$ is the standard Kronecker tensor. The deviatoric is evidently built to have $\mathrm{tr}(\underline{\underline{\varepsilon}}^D) = 0$. Let then Γ_0 be a part of Γ on which we assume $\underline{u} = 0$. We also assume the existence in Ω of a distributed force $\underline{f}$ (e.g., gravity) and on Γ_1 of a traction $\underline{g}$ that is decomposed into a normal part g_n and a tangential part g_t (Figure I.1). We denote by $\underline{n}$ and $\underline{t}$ the normal and tangential unit vectors to Γ.

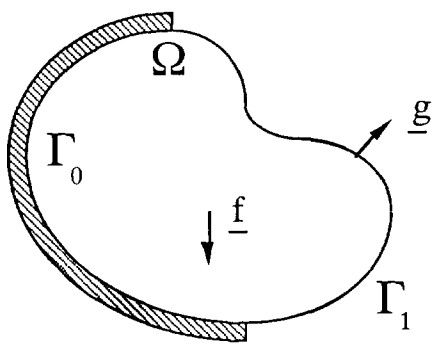


Figure I.1

Let us denote

$$|\underline{\underline{\varepsilon}}|^2 = \sum_{i,j} \varepsilon_{ij}^2 = \underline{\underline{\varepsilon}} : \underline{\underline{\varepsilon}}, \tag{2.19}$$

and let us consider in $(H^1_{0,\Gamma_0}(\Omega))^2 = V$ the minimization problem

$$\begin{aligned} \inf_{\underline{v}\in V}\Big\{ \int_\Omega \frac{1}{2}\big(\lambda|\mathrm{div}\,\underline{v}|^2 + 2\mu|\underline{\underline{\varepsilon}}(v)|^2\big)\,dx - \int_\Omega \underline{f}\cdot\underline{v}\,dx \\ - \int_{\Gamma_1} g_n \underline{v}\cdot\underline{n}\,ds - \int_{\Gamma_1} g_t\,\underline{v}\cdot\underline{t}\,ds \Big\}. \end{aligned} \tag{2.20}$$

Constants λ and μ, the Lamé coefficients, depend on the physical properties of the material considered. Solution $\underline{u}$ of this problem is then characterized by

$$
\begin{aligned}
2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx + \lambda \int_\Omega \operatorname{div} \underline{u} \operatorname{div} \underline{v}\, dx \\
= \int_\Omega \underline{f} \cdot \underline{v}\, dx + \int_{\Gamma_1} g_n\, \underline{v} \cdot \underline{n}\, ds + \int_{\Gamma_1} g_t\, \underline{v} \cdot \underline{t}\, ds.
\end{aligned} \tag{2.21}
$$

We now use the classical integration by parts formula, $\underline{\underline{m}}$ being some tensor,

$$
\int_\Omega \underline{\underline{m}} : \underline{\underline{\varepsilon}}(\underline{v})\, dx = - \int_\Omega (\underline{\operatorname{div}}\, \underline{\underline{m}}) \cdot \underline{v}\, dx + \int_\Gamma m_{nn} \underline{v} \cdot \underline{n}\, ds + \int_\Gamma m_{nt} \underline{v} \cdot \underline{t}\, ds, \tag{2.22}
$$

where m_{nn} and m_{nt} denote the normal and tangential parts of the traction vector m_n, i.e.,

$$
\begin{cases}
m_{nn} = \sum_{i,j} m_{ij}\, n_i\, n_j = \sum_i \{ \sum_j m_{ij} n_j \} n_i = \sum_i (\underline{m}_n)_i\, n_i, \\
m_{nt} = \sum_{i,j} m_{ij}\, t_i\, n_j = \sum_i \{ \sum_j m_{ij} n_j \} t_i = \sum_i (\underline{m}_n)_i\, t_i.
\end{cases} \tag{2.23}
$$

Equation (2.21) can now be interpreted as

$$
\begin{cases}
-(2\mu\, \underline{\operatorname{div}}\, \underline{\underline{\varepsilon}}(\underline{u}) + \lambda\, \underline{\operatorname{grad}} \operatorname{div} \underline{u}) = \underline{f} \text{ in } \Omega, \\
\underline{u}|_{\Gamma_0} = 0, \\
2\mu\varepsilon_{nn} + \lambda \operatorname{div} \underline{u} = g_n \text{ on } \Gamma_1, \\
2\mu\varepsilon_{nt} = g_t \text{ on } \Gamma_1.
\end{cases} \tag{2.24}
$$

Let us now introduce the stress tensor $\underline{\underline{s}} = \underline{\underline{s}}^D + p\underline{\underline{\delta}}$ and the constitutive law

$$
\begin{cases}
\underline{\underline{s}}^D = 2\mu\, \underline{\underline{\varepsilon}}^D(\underline{u}), \\
p = 2(\lambda + \mu) \operatorname{div} \underline{u},
\end{cases} \tag{2.25}
$$

relating stresses to displacements. It is now clear that the first equation of (2.24) expresses the equilibrium condition of continuum mechanics,

$$
\underline{\operatorname{div}}\, \underline{\underline{s}} + \underline{f} = 0. \tag{2.26}
$$

In applications, the constitutive law (2.25) will vary depending on the type of materials and will sometimes take very nonlinear forms. Moreover, large displacements will require a much more complex treatment. Nevertheless the problem described remains valuable as a model for more complicated situations.

The case of an incompressible material is specially important. It leads to the same equations as in the study of viscous incompressible flows. □

Example 2.3: *Stokes problem for viscous incompressible flow.*
We now consider a low velocity flow of a viscous incompressible fluid in a domain Ω. We denote by $\underline{u}$ the velocity field and by $\underline{\underline{\varepsilon}}(\underline{u})$ the (linearized) strain rate tensor defined in the same way as $\underline{\underline{\varepsilon}}$ in (2.16). We thus consider the mimimization problem with the same notation and the same space V as in Example 2.2, but now with the incompressibility condition $\operatorname{div} \underline{v} = 0$, that is,

$$\inf_{\substack{\underline{v} \in V \\ \operatorname{div} \underline{v} = 0}} \mu \int_\Omega |\underline{\underline{\varepsilon}}(\underline{v})|^2 - \int_\Omega \underline{f} \cdot \underline{v}\, dx + \int_{\Gamma_1} g_n\, \underline{v} \cdot \underline{n}\, ds + \int_{\Gamma_1} g_t\, \underline{v} \cdot \underline{t}\, ds. \tag{2.27}$$

As we shall see later, problem (2.20) can be considered, when λ is large, as an approximation (by a "penalty method") of problem (2.27). When λ is large, the second constitutive relation of (2.25) becomes meaningless. We shall see in Section I.3 that pressure can be introduced as a Lagrange multiplier associated with the constraint $\operatorname{div} \underline{u} = 0$. □

We finally present as a last example a fourth-order problem. It is again physically an elasticity problem but in a special modelization.

Example 2.4: *Deflection of a thin clamped plate.*
We consider here the problem of a thin clamped plate deflected under a distributed load f. The physical model will be described in Chapter VII. We also refer to CIARLET [B] and CIARLET–DESTUYNDER [A] for more details on plate problems. Under reasonable assumptions, one obtains that the vertical deflection u is a solution of the minimization problem

$$\inf_{v \in H_0^2(\Omega)} \frac{1}{2} \int_\Omega |\Delta v|^2 dx - \int_\Omega f v\, dx. \tag{2.28}$$

The unique solution u is characterized by

$$\int_\Omega \Delta u\, \Delta v\, dx = \int_\Omega f v\, dx, \qquad \forall v \in H_0^2(\Omega) \tag{2.29}$$

and is the solution of the boundary value problem

$$\begin{cases} \Delta^2 u = f, \\ u|_\Gamma = 0, \\ \dfrac{\partial u}{\partial n}|_\Gamma = 0. \end{cases} \tag{2.30}$$

For these boundary conditions (representing a clamped plate) one may use, instead of (2.28), the formulation

$$\inf_{v \in H_0^2(\Omega)} \frac{1}{2} \int_\Omega \left[\left\{ \frac{\partial^2 v}{\partial x_1^2} \right\}^2 + 2 \left\{ \frac{\partial^2 v}{\partial x_1 \partial x_2} \right\}^2 + \left\{ \frac{\partial^2 v}{\partial x_2^2} \right\}^2 \right] dx - \int_\Omega f v\, dx. \tag{2.31}$$

These two equivalent forms can lead to two different numerical methods. It must also be noted that *natural boundary conditions* (those arising from integration by parts) will not be the same if (2.28) and (2.31) are minimized on a space larger than $H_0^2(\Omega)$. Actually the true potential energy of the plate (that is, the true functional which has to be minimized) is given by

$$J(v) = \frac{Et^3}{12(1-\nu^2)} \int_\Omega \Big\{ \nu|\Delta v|^2 + (1-\nu)\Big[\Big(\frac{\partial^2 v}{\partial x_1^2}\Big)^2 + 2\Big(\frac{\partial^2 v}{\partial x_1 \partial x_2}\Big)^2 + \Big(\frac{\partial^2 v}{\partial x_2^2}\Big)^2\Big]\Big\}\, dx - \int_\Omega f v\, dx \tag{2.32}$$

where E is Young's modulus, ν the Poisson's coefficient, and t the thickness of the plate. In particular, E and ν can be expressed in terms of the Lamé coefficients λ, μ in the following way:

$$E = \frac{\mu(3\lambda + 2\mu)}{\lambda + \mu}, \qquad \nu = \frac{\lambda}{2(\lambda + \mu)}. \tag{2.33}$$

We also recall that the Stokes problem (2.27) can also be expressed as a biharmonic problem by the introduction of a stream function ψ such that

$$\underline{u} = \Big\{ \frac{\partial \psi}{\partial x_2}, -\frac{\partial \psi}{\partial x_1} \Big\} \tag{2.34}$$

We shall come back to this point in Section I.3. □

The examples presented are among the most fundamental of mathematical physics and engineering problems. A good understanding of their properties will enable one to extend the results obtained to more complex situations.

I.3 Duality Methods

I.3.1 Generalities

Up to now, we have introduced four problems that can be written as minimization problems of some functionals in properly chosen functional spaces. This is the most classical way of setting these problems. Finite element approximations, based on the formulations described earlier, are routinely used in commercial codes. Various reasons justified the introduction, for these same problems of different variational formulations and therefore different finite element approximations. This was done at the beginning by many engineers. The reader may refer, for example to REISNERR [A]–[B], PIAN–TONG [A].

The first reason may be the presence in the variational formulation of a constraint, such as the condition $\operatorname{div} \underline{u} = 0$ in problem (2.27). As we shall see,

it is difficult (and not necessary) to build finite element approximations satisfying exactly this constraint. It will be more efficient to modify the variational formulation and to introduce pressure.

A second reason may lie in the physical "importance" of the variables appearing in the problem. In elasticity problems, for example, it is often more useful to compute accurately stresses rather than displacements. In the standard formulation, stresses can be recovered from the displacements by (2.25) or some other analogue law. Their computation requires the derivatives of the displacement field $\underline{u}$. From a numerical point of view, differentiating implies a loss of precision. It is therefore appealing to look for a formulation in which constraints are readily accessible.

A third reason comes from difficulties arising in the discretization of spaces of regular functions such as $H_0^2(\Omega)$ appearing in Example 2.4. Approximating this space by a finite element method implies ensuring continuity of the derivatives at interfaces between elements. This is possible but more cumbersome than approximating, say, $H^1(\Omega)$. A variational formulation enabling to decompose a fourth-order problem into a system of second-order problems permits one to avoid building complicated elements at the price of introducing some other difficulties.

Finally, a last reason could be to look for *a weaker variational formulation* corresponding better in some cases to available data (e.g., punctual loads) for which standard formulations may become meaningless due to a lack of regularity of the solution.

We must also point out that the "nonstandard" formulations which we shall now describe have been initially introduced by engineers for one or some of the reasons discussed above. We quote in this respect, but in a totally non exhaustive way, FRAEIJS DE VEUBEKE [A], HELLAN [A], HERMANN [A], PIAN [A], TONG–PIAN [A]. On the other hand, very powerful tools for the transformation of variational problems can be found in convex analysis and duality theory (AUBIN [B], BARBU–PRECUPANU [A], EKELAND–TEMAM [A], ROCKAFELLAR [A]). It is neither possible nor desirable here to develop duality theory and we shall restrict ourselves to the most basic facts. The fundamental idea of duality theory is that one can represent a convex function by the family of its tangent affine functions. This is indeed the principle of the classical Legendre transformation. More precisely, let us define for a given convex function $f(v)$, defined on a space V, the conjugate function $f^*(v^*)$ on the dual space V' of V by

$$f^*(v^*) = \sup_{v \in V} \langle v, v^* \rangle_{V \times V'} - f(v). \tag{3.1}$$

Note that when $V = \mathbb{R}$, $f^*(v^*)$ is the intercept with the v axis of the tangent to f of slope v^*. The important point for what follows is that one can build

$f(v)$ from $f^*(v^*)$ by the following formula, symmetrical to (3.1),

$$f(v) = \sup_{v^* \in V'} \langle v^*, v \rangle_{V' \times V} - f^*(v^*). \tag{3.2}$$

Given then a problem of the form

$$\inf_{v \in V} g(v) + f(v), \tag{3.3}$$

we can use (3.2) to obtain

$$\inf_{v \in V} \left\{ g(v) + \sup_{v^* \in V'} \langle v^*, v \rangle_{V' \times V} - f^*(v^*) \right\}, \tag{3.4}$$

that is, the saddle point problem

$$\inf_{v \in V} \sup_{v^* \in V'} g(v) + \langle v^*, v \rangle_{V' \times V} - f^*(v^*). \tag{3.5}$$

Under simple regularity assumptions, one can then consider the *dual problem*

$$\sup_{v^* \in V'} \left\{ \inf_{v \in V} g(v) + \langle v^*, v \rangle_{V' \times V} - f^*(v^*) \right\}. \tag{3.6}$$

We now demonstrate on examples how this technique can be applied.

I.3.2 Examples for symmetric problems

Example 3.1: *Introduction of pressure in Stokes problem.*
Let us consider problem (2.27) where, to make the presentation easier, we take $\Gamma_0 = \Gamma$, that is, pure Dirichlet conditions on the boundary. This constrained problem can be written as an unconstrained problem, introducing the characteristic function $\delta(\cdot|\{0\})$ defined on $L^2(\Omega)$ by

$$\delta(v|\{0\}) = \begin{cases} 0 & \text{if } v = 0 \\ +\infty & \text{otherwise.} \end{cases} \tag{3.7}$$

It is thus a pure change of notation to write, instead of (2.27),

$$\inf_{v \in V} \mu \int_\Omega |\underline{\underline{\varepsilon}}(v)|^2 \, dx - \int_\Omega \underline{f} \cdot \underline{v} + \delta(\operatorname{div} \underline{v}|\{0\}), \tag{3.8}$$

where $V = (H_0^1(\Omega))^2$. On the other hand, one clearly has

$$\delta(\operatorname{div} \underline{u}|\{0\}) = \sup_{q \in L^2(\Omega)} \int_\Omega q \operatorname{div} \underline{u} \, dx, \qquad \forall \underline{u} \in (H_0^1(\Omega))^2, \tag{3.9}$$

and the minimization problem (3.8) can be transformed into the saddle point problem

$$\inf_{v\in V}\sup_{q\in L^2(\Omega)} \mu\int_\Omega |\underline{\underline{\varepsilon}}(v)|^2\,dx - \int_\Omega \underline{f}\cdot\underline{v}\,dx - \int_\Omega q\,\mathrm{div}\,\underline{v}\,dx. \tag{3.10}$$

This apparently simple trick has in reality completely changed the nature of the problem. We now have to find a pair $(\underline{u},p)$ solution of the variational system

$$\begin{cases} 2\mu\displaystyle\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\,dx - \int_\Omega \underline{f}\cdot\underline{v}\,dx - \int_\Omega p\,\mathrm{div}\,\underline{v}\,dx = 0, & \forall v\in V,\\ \displaystyle\int_\Omega q\,\mathrm{div}\,\underline{u}\,dx = 0, & \forall q\in L^2(\Omega). \end{cases} \tag{3.11}$$

The second equation of (3.11) evidently expresses the condition $\mathrm{div}\,\underline{u}=0$. In order to use (3.11), we shall have to show the existence of a saddle point $(\underline{u},p)$, in particular the existence of the Lagrange multiplier p. This will be done in Chapter II. The variational system (3.11) can be interpreted in the form

$$\begin{cases} -2\mu A\underline{u} + \underline{\mathrm{grad}}\,p = \underline{f},\\ \mathrm{div}\,\underline{u} = 0,\\ \underline{u}|_\Gamma = 0, \end{cases} \tag{3.12}$$

where one uses the operator $A\underline{u} = \underline{\mathrm{div}}\,\underline{\underline{\varepsilon}}(\underline{u})$

$$A\underline{u} = \begin{pmatrix} \dfrac{\partial^2 u_1}{\partial x_1^2} + \dfrac{1}{2}\dfrac{\partial}{\partial x_2}\left\{\dfrac{\partial u_1}{\partial x_2} + \dfrac{\partial u_2}{\partial x_1}\right\} \\ \dfrac{\partial^2 u_2}{\partial x_2^2} + \dfrac{1}{2}\dfrac{\partial}{\partial x_1}\left\{\dfrac{\partial u_1}{\partial x_2} + \dfrac{\partial u_2}{\partial x_1}\right\} \end{pmatrix}. \tag{3.13}$$

Under the *divergence-free condition* $\mathrm{div}\,\underline{u}=0$, this can also be written

$$\begin{cases} -\mu\Delta\underline{u} + \underline{\mathrm{grad}}\,p = \underline{f},\\ \mathrm{div}\,\underline{u} = 0, \end{cases} \tag{3.14}$$

which is the classical form of the Stokes problem. □

Problem (3.10) has the general form

$$\inf_{v\in V}\sup_{q\in Q} L(v,q), \tag{3.15}$$

where $L(v,q)$ is a convex-concave functional on $V\times Q$. If one first eliminates q by computing

$$J(v) = \sup_{q\in Q} L(v,q),$$

one falls back on the original problem, the *primal problem*. Reversing the order of operations (this cannot always be done, but no problems arise in the examples we present), and eliminating v from $L(v,q)$ by defining

$$J^*(q) = \inf_v L(v,q) \tag{3.16}$$

leads to the dual problem

$$\sup_{q\in Q} J^*(q). \tag{3.17}$$

We now apply this idea to the previous examples.

Example 3.2: *Dual problem for the Stokes problem.*
In the case of Stokes problem, the dual problem can be expressed, as we shall see, in many equivalent ways. In order to find it, we must, given q, find the minimum in v of $L(\underline{v},q) = \mu \int_\Omega |\underline{\underline{\varepsilon}}(v)|^2\,dx - \int_\Omega \underline{f}\cdot\underline{v}\,dx - \int_\Omega q \operatorname{div}\underline{v}\,dx$. This minimum is characterized by

$$2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_q) : \underline{\underline{\varepsilon}}(\underline{v})\,dx - \int_\Omega \underline{f}\cdot\underline{v} - \int_\Omega q \operatorname{div}\underline{v}\,dx = 0, \qquad \forall v \in V, \tag{3.18}$$

denoting by $\underline{u}_q$ the minimum point. Making $v = \underline{u}_q$ this gives

$$2\mu \int_\Omega |\underline{\underline{\varepsilon}}(\underline{u}_q)|^2\,dx - \int_\Omega \underline{f}\cdot\underline{u}_q\,dx - \int_\Omega q \operatorname{div}\underline{u}_q\,dx = 0. \tag{3.19}$$

Using (3.19) to evaluate $L(\underline{u}_q, q)$, the dual problem can be written as an **optimal** control problem,

$$\sup_{q\in L^2(\Omega)} -\mu \int_\Omega |\underline{\underline{\varepsilon}}(\underline{u}_q)|^2\,dx, \tag{3.20}$$

where $\underline{u}_q$ is the **solution** of

$$\begin{cases} -2\mu\, A\underline{u}_q + \underline{\operatorname{grad}}\, q = \underline{f}, \\ \underline{u}_q|_\Gamma = 0. \end{cases} \tag{3.21}$$

Denoting by G the Green operator defining the solution of (3.21), that is,

$$\underline{u}_q = G(\underline{f} - \underline{\operatorname{grad}}\, q), \tag{3.22}$$

and using (3.22) in (3.20), one can get from (3.19)

$$\inf_q \int_\Omega \underline{\operatorname{grad}}\, q \cdot G(\underline{\operatorname{grad}}\, q)\,dx - \int_\Omega G(f)\cdot \underline{\operatorname{grad}}\, q\,dx. \tag{3.23}$$

One notices that this dual problem is a problem in $\underline{\text{grad}}\, q$. It is well known that the solution p is defined (for Dirichlet conditions on $\underline{u}$) only up to an additive constant. One can interpret (3.23) as the equation

$$\text{div}(G\, \underline{\text{grad}}\, q) = \text{div}(Gf). \tag{3.24}$$

If one defines on $V' = (H^{-1}(\Omega))^2$ the norm,

$$\|f\|_G^2 = \langle Gf, f\rangle_{V\times V'}, \tag{3.25}$$

problem (3.23) can be written as a least-squares problem

$$\inf_{q\in L^2(\Omega)} \tfrac{1}{2}\|\, \underline{\text{grad}}\, q - f\|_G^2. \ \square \tag{3.26}$$

The presence of a Green operator makes this dual problem difficult to handle directly. It is however implicitly the basis of some numerical solution procedures (FORTIN–GLOWINSKI [B], THOMASSET [A]). We will present other dual problems that will have direct importance and that will be handled as such.

Example 3.3: *A duality method for nearly incompressible elasticity.*
We already noted in Example 2.2 and Example 2.3 that the linear elasticity problem and Stokes problem are very close when a nearly incompressible material is considered. We now develop this analogy in the framework of Example 3.1. The starting point will be the obvious result,

$$\frac{\lambda}{2}\int_\Omega |\,\text{div}\, \underline{v}|^2\, dx = \sup_{q\in L^2(\Omega)} \int_\Omega q\, \text{div}\, \underline{v}\, dx - \frac{1}{2\lambda}\int_\Omega |q|^2\, dx. \tag{3.27}$$

Substituting (3.27) into (2.20) we get, by the same methods as in the previous examples, the problem

$$\begin{aligned} \inf_{v\in V}\sup_{q\in L^2(\Omega)} \mu\int_\Omega |\underline{\underline{\varepsilon}}(\underline{v})|^2\, dx - \int_\Omega q\, \text{div}\, \underline{v}\, dx - \frac{1}{2\lambda}\int_\Omega |q|^2\, dx \\ - \int_\Omega \underline{f}\cdot\underline{v}\, dx - \int_{\Gamma_1} g_n\, \underline{v}\cdot\underline{n}\, ds - \int_{\Gamma_1} g_t\, \underline{v}\cdot\underline{t}\, ds. \end{aligned} \tag{3.28}$$

The solution $(\underline{u}, p)$ of problem (3.28) is characterized by the system

$$\left\{\begin{aligned} &2\mu\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx - \int_\Omega p\, \text{div}\, \underline{v}\, dx \\ &\qquad = \int_\Omega \underline{f}\cdot\underline{v}\, dx + \int_{\Gamma_1} (g_n\, \underline{v}\cdot\underline{n} + g_t\, \underline{v}\cdot\underline{t})\, ds, \qquad \forall v\in V, \\ &\int_\Omega q\, \text{div}\, \underline{u} + \frac{1}{\lambda}\int_\Omega pq\, dx = 0, \qquad \forall q\in L^2(\Omega). \end{aligned}\right. \tag{3.29}$$

This can be summarized by saying that we transformed our original problem into a system by introducing the auxiliary variable $p = \lambda\, \text{div}\, \underline{u}$. It must be noted that this also makes our minimization problem become a saddle point problem. We shall see in Chapter VI that this apparently *tautological change has implications in the building of numerical approximations to (2.20) that remain valid when λ is large.* $\square$

Example 3.4: *Dualization of the Dirichlet problem.*
The result that we shall get here can be obtained by many methods. Techniques of convex analysis permit one to extend what appears to be a trick to much more complex situations. However it will be sufficient for our purpose to use the simple development below. Let us then consider the Dirichlet problem,

$$\inf_{v \in H_0^1(\Omega)} \frac{1}{2} \int_\Omega |\,\underline{\text{grad}}\, v|^2 \, dx - \int_\Omega f v \, dx. \tag{3.30}$$

In many applications $\underline{\text{grad}}\, v$ rather than v is the interesting variable. For instance, in thermodiffusion problems, $\underline{\text{grad}}\, v$ will be the heat flux, which is very often more important to know than temperature v. What we now do is essentially to introduce the auxiliary variable $\underline{p} = \underline{\text{grad}}\, v$ to transform our Dirichlet problem into a system. To do so, we use the same trick as in Example 3.3 and write

$$\tfrac{1}{2} \int_\Omega |\,\underline{\text{grad}}\, v|^2 \, dx = \sup_{\underline{q} \in (L^2(\Omega))^2} \int_\Omega \underline{q} \cdot \underline{\text{grad}}\, v \, dx - \tfrac{1}{2} \int_\Omega |\underline{q}|^2 \, dx \tag{3.31}$$

which we use in (3.30) to get the saddle point problem

$$\inf_{v \in V} \sup_{\underline{q} \in Q} -\tfrac{1}{2} \int_\Omega |\underline{q}|^2 \, dx - \int_\Omega f v \, dx + \int_\Omega \underline{q} \cdot \underline{\text{grad}}\, v \, dx, \tag{3.32}$$

where $V = H_0^1(\Omega$ and $Q = (L^2(\Omega))^2$. The saddle point $(u, \underline{p})$ is characterized by

$$\begin{cases} \displaystyle\int_\Omega \underline{p} \cdot \underline{q} \, dx - \int_\Omega \underline{q} \cdot \underline{\text{grad}}\, u \, dx = 0, & \forall \underline{q} \in Q, \\ \displaystyle\int_\Omega \underline{p} \cdot \underline{\text{grad}}\, v \, dx = \int_\Omega f v \, dx, & \forall v \in V, \end{cases} \tag{3.33}$$

and this can be read as

$$\begin{cases} \underline{p} = \underline{\text{grad}}\, u, & u \in H_0^1(\Omega), \\ \operatorname{div} \underline{p} + f = 0, \end{cases} \tag{3.34}$$

which is evidently equivalent to a standard Dirichlet problem.

The *dual problem* is readily made explicit. Writing it as a minimization problem by changing the sign of the objective functional, we have

$$\inf \tfrac{1}{2} \int_\Omega |\underline{q}|^2 \, dx, \qquad \forall \underline{q} \in Z_f = \{\underline{q} \in (L^2(\Omega))^2 \mid \operatorname{div} \underline{q} + f = 0\}. \tag{3.35}$$

This is the classical *complementary energy principle.* □

We now want to get a weaker form of this problem. In order to do so, we must introduce a new functional space which will frequently occur in the following. We define

$$H(\mathrm{div};\Omega) = \{\underline{q} \mid \underline{q} \in (L^2(\Omega))^2,\ \mathrm{div}\,\underline{q} \in L^2(\Omega)\} \tag{3.36}$$

and its norm

$$\|\underline{q}\|^2_{H(\mathrm{div},\Omega)} = \|\underline{q}\|^2_{0,\Omega} + \|\mathrm{div}\,\underline{q}\|^2_{0,\Omega} \tag{3.37}$$

that makes it a Hilbert space. It can then be shown, (TEMAM [A]), by the methods of LIONS–MAGENES [A], that vectors of $H(\mathrm{div};\Omega)$ admit a well defined *normal trace on* $\Gamma = \partial\Omega$. This normal trace $\underline{q}\cdot\underline{n}$, lies in $H^{-1/2}(\Gamma)$ and one has the following "integration by parts" formula,

$$\int_\Omega \underline{q}\cdot \underline{\mathrm{grad}}\, v\, dx + \int_\Omega \mathrm{div}\,\underline{q}\, v\, dx = \langle v, \underline{q}\cdot\underline{n}\rangle_{H^{1/2}(\Gamma)\times H^{-1/2}(\Gamma)}, \tag{3.38}$$

for any $\underline{q} \in H(\mathrm{div};\Omega)$ and any $v \in H^1(\Omega)$. We shall often write **formally** $\int_\Gamma v\underline{q}\cdot\underline{n}\, ds$ instead of the duality product $\langle v,\ \underline{q}\cdot\underline{n}\rangle$.

Example 3.5: *Weak form of the dual Dirichlet problem.*
If we take $f \in L^2(\Omega)$, problem (3.35), which is a constrained problem, can be changed into a saddle point problem, as in Example 3.1, by introducing a Lagrange multiplier $u \in L^2(\Omega)$, that is, as $\underline{q}$ now belongs to $H(\mathrm{div};\Omega)$,

$$\inf_{\underline{q}\in H(\mathrm{div},\Omega)}\ \sup_{v\in L^2(\Omega)}\ \tfrac{1}{2}\int_\Omega |\underline{q}|^2\, dx + \int_\Omega fv\, dx + \int_\Omega v\,\mathrm{div}\,\underline{q}\, dx. \tag{3.39}$$

The functional spaces employed precisely enable us to write every term in (3.39) without ambiguity. We now look for a saddle point $(u,\underline{p})$ satisfying the variational system,

$$\begin{cases} \displaystyle\int_\Omega \underline{p}\cdot\underline{q}\, dx + \int_\Omega u\,\mathrm{div}\,\underline{q}\, dx = 0, & \forall \underline{q} \in H(\mathrm{div};\Omega), \\ \displaystyle\int_\Omega (\mathrm{div}\,\underline{p} + f)v\, dx = 0, & \forall v \in L^2(\Omega). \end{cases} \tag{3.40}$$

Using (3.38) with $\underline{q}\cdot\underline{n}|_\Gamma = 0$, we obtain from (3.40),

$$\underline{p} = \underline{\mathrm{grad}}\, u. \tag{3.41}$$

Now $u \in L^2(\Omega)$ and $\underline{\mathrm{grad}}\, u = \underline{p} \in (L^2(\Omega))^2$ imply that $u \in H^1(\Omega)$ and it is justified to consider its trace. Again using (3.38) with a general $\underline{q}$ shows that $u|_\Gamma = 0$. The solution of our "weaker" problem is then the solution of the standard problem. However the discretizations of problem (3.40) will be quite different from those used for the standard formulation. □

Remark 3.1: The previous formulation enables us to write directly in a variational form a nonhomogeneous Dirichlet problem. Indeed the solution $(u, \underline{p})$ of the saddle point problem with $g \in H^{1/2}(\Gamma)$,

$$\inf_{\underline{q}} \sup_{v} \tfrac{1}{2} \int_\Omega |\underline{q}|^2 \, dx + \int_\Omega (\operatorname{div} \underline{q} + f) v \, dx - \int_\Gamma g \, \underline{q} \cdot \underline{n} \, ds, \tag{3.42}$$

leads to $\underline{p} = \underline{\operatorname{grad}}\, u$, $\operatorname{div} \underline{p} + f = 0$, $u|_\Gamma = g$.

On the other hand, Neumann conditions become *essential* conditions that have to be incorporated into the construction of $\underline{p}$, that is, in the choice of the functional space. □

We now want to extend the previous results to the case of the linear elasticity problem. We shall thus get a second way to dualize problem (2.20). It is a general fact that there is no unique way to use duality techniques.

The lines of the development are the same as for Dirichlet problems and we shall avoid to write the details. Let us then define,

$$\underline{\underline{H}}(\underline{\operatorname{div}}; \Omega)_s = \{\underline{\underline{\sigma}} \mid \sigma_{ij} \in L^2(\Omega),\ \sigma_{ij} = \sigma_{ji},\ \underline{\operatorname{div}}\, \underline{\underline{\sigma}} \in (L^2(\Omega))^2\}, \tag{3.43}$$

where $\underline{\operatorname{div}}\, \underline{\underline{\sigma}}$ is the *vector* $\partial\sigma_{i2}/\partial x_1 + \partial\sigma_{i2}/\partial x_2$. On this space we use the norm,

$$\|\underline{\underline{\sigma}}\|^2_{\underline{\underline{H}}(\underline{\operatorname{div}};\Omega)_s} = \sum_{i,j} \int_\Omega |\sigma_{ij}|^2 \, dx + \|\underline{\operatorname{div}}\, \underline{\underline{\sigma}}\|^2_{(L^2(\Omega))^2}, \tag{3.44}$$

which makes it a Hilbert space.

One can then define [as for $H(\operatorname{div}; \Omega)$] the vector $\underline{\sigma}_n \in (H^{-1/2}(\Gamma))^2$

$$(\underline{\sigma}_n)_i = \sum_j \sigma_{ij} \, n_j \tag{3.45}$$

and we shall mostly use the normal and tangential components, σ_{nn} and σ_{nt}, of this vector, as defined in (2.23). We then have the following "integration by parts" formula:

$$\int_\Omega \underline{\underline{\sigma}} : \underline{\underline{\varepsilon}}(\underline{v}) \, dx + \int_\Omega \underline{\operatorname{div}}\, \underline{\underline{\sigma}} \cdot \underline{v} \, dx = \langle \underline{\sigma}_n, \underline{v} \rangle = \langle \sigma_{nn}, \underline{v} \cdot \underline{n} \rangle + \langle \sigma_{nt}, \underline{v} \cdot \underline{t} \rangle, \tag{3.46}$$

which is valid for any $\underline{\underline{\sigma}}$ and $\underline{v}$ smooth enough. We have denoted by $\langle \cdot, \cdot \rangle$ the duality between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$ and shall often write the formal expression $\int_\Gamma \sigma_{nn} \ \underline{v} \cdot \underline{n} \ ds + \int_\Gamma \sigma_{nt} \ \underline{v} \cdot \underline{t} \ ds$. We can now write our dual formulation for the linear elasticity problem.

Example 3.6: *Dualization of the linear elasticity problem.*
Following the same line as for the Dirichlet problem, we write

$$\begin{aligned}\mu \int_\Omega |\underline{\underline{\varepsilon}}(\underline{v})|^2 + \frac{\lambda}{2}\int_\Omega |\operatorname{div}\underline{v}|^2\,dx &= \sup_{\underline{\underline{\sigma}}\in \underline{\underline{H}}(\operatorname{div},\Omega)} \int_\Omega \underline{\underline{\sigma}}^D \cdot \underline{\underline{\varepsilon}}^D\,dx \\ &+ \int_\Omega \operatorname{tr}\underline{\underline{\sigma}}\ \operatorname{tr}\underline{\underline{\varepsilon}}\,dx - \frac{1}{4\mu}\int_\Omega |\underline{\underline{\sigma}}^D|^2\,dx - \frac{1}{2(\lambda+\mu)}\int_\Omega (\operatorname{tr}\underline{\underline{\sigma}})^2\,dx\,,\end{aligned} \tag{3.47}$$

which leads us to the saddle point problem in $\underline{\underline{H}}(\operatorname{div};\Omega)\times(L^2(\Omega))^2$,

$$\inf_{\underline{\underline{\sigma}}}\sup_{\underline{v}} \frac{1}{2(\lambda+\mu)}\int_\Omega |\operatorname{tr}\underline{\underline{\sigma}}|^2\,dx + \frac{1}{4\mu}\int_\Omega |\underline{\underline{\sigma}}^D|^2\,dx + \int_\Omega (\operatorname{div}\underline{\underline{\sigma}}+\underline{f})\,\underline{v}\,dx. \tag{3.48}$$

The solution $(\underline{\underline{\sigma}},\underline{u})$ of this saddle point problem is characterized by the system

$$\begin{cases} \operatorname{div}\underline{\underline{\sigma}} + \underline{f} = 0, \\ \operatorname{tr}\underline{\underline{\sigma}} = (\lambda+\mu)\ \operatorname{tr}\underline{\underline{\varepsilon}}(\underline{u}), \\ \underline{\underline{\sigma}}^D = 2\mu\,\underline{\underline{\varepsilon}}^D(\underline{u}), \end{cases} \tag{3.49}$$

which are the equilibrium condition (2.26) and the constitutive relations (2.25). The dual problem then consists in minimizing the *complementary energy*

$$\inf_{\underline{\underline{\sigma}}} \frac{1}{4\mu}\int_\Omega |\underline{\underline{\sigma}}^D|^2\,dx + \frac{1}{2(\lambda+\mu)}\int_\Omega |\operatorname{tr}\underline{\underline{\sigma}}|^2\,dx \tag{3.50}$$

under the constraint $\operatorname{div}\underline{\underline{\sigma}} + \underline{f} = 0$. Both the *mixed formulation* (3.48) and the dual formulation (3.50) are used in practice. They lead to different although similar approximations. □

To end this section we finally consider the thin plate problem of Example 2.4 to introduce a mixed formulation due to CIARLET–RAVIART [C] and MERCIER [A].

Example 3.7: *Decomposition of a biharmonic problem.*
Again using the same technique as in the dualization of the Dirichlet problem in Example 3.4, it is a simple exercise to transform problem (2.28) into the saddle point problem

$$\inf_{\mu\in L^2(\Omega)}\ \sup_{v\in H_0^2(\Omega)}\ \tfrac{1}{2}\int_\Omega |\mu|^2\,dx + \int_\Omega \mu\,\Delta v\,dx + \int_\Omega fv\,dx, \tag{3.51}$$

and to get the dual problem

$$\inf_{\mu\in M}\ \frac{1}{2}\int_\Omega |\mu|^2\,dx, \tag{3.52}$$

where $M = \{\mu \in L^2(\Omega),\ \Delta\mu + f = 0\}$. Integrating by parts the term $\int_\Omega \mu\, \Delta v\, dx$, we get, as in Example 3.5, a weaker formulation

$$(3.53)\qquad \inf_{\mu\in L^2(\Omega)} \sup_{v\in H_0^2(\Omega)} \frac{1}{2}\int_\Omega |\mu|^2\, dx - \int_\Omega \underline{\mathrm{grad}}\,\mu \cdot \underline{\mathrm{grad}}\, v\, dx + \int_\Omega fv\, dx.$$

Assume that (3.53) has a saddle point (ω, u) with $\omega \in H^1(\Omega)$. Then (ω, u) is characterized by the variational system

$$(3.54)\qquad \begin{cases} \displaystyle\int_\Omega \omega\mu - \int_\Omega \underline{\mathrm{grad}}\,\mu \cdot \underline{\mathrm{grad}}\, u\, dx = 0, & \forall \mu \in H^1(\Omega), \\ \displaystyle\int_\Omega \underline{\mathrm{grad}}\,\omega \cdot \underline{\mathrm{grad}}\, v\, dx = \int_\Omega fv\, dx, & \forall v \in H_0^1(\Omega). \end{cases}$$

If we use $\mu = \phi \in \mathfrak{D}(\Omega)$ in the first equation and $v = \phi \in \mathfrak{D}(\Omega)$ in the second one, we can interpret (3.54) as

$$(3.55)\qquad \begin{cases} -\Delta u = \omega, \\ u|_\Gamma = 0, \\ -\Delta\omega = f. \end{cases}$$

The first equation of (3.54) also yields $\partial u/\partial n = 0$, as we hope. In a sense, we thus have in (3.55) too many boundary conditions on u and none on ω. The system however has a solution (ω, u) (provided Ω and f are smooth enough) such that the solution of the Dirichlet problem in u also satisfies (through the choice of the right-hand side) the extra Neumann condition. □

Example 3.8: *Decomposition of the plate bending problem.*
We now consider the plate bending problem (2.32). In order to make the dual problem easier to introduce, we first write the energy functional in the form

$$(3.56)\qquad \frac{1}{2}\left(\frac{Et^3}{12(1-\nu^2)}\right)\int_\Omega \mathfrak{M}(\underline{\underline{D}}_2 v) : \underline{\underline{D}}_2 v\, dx - \int_\Omega fv\, dx,$$

where the operator $\underline{\underline{D}}_2$ is defined by

$$(3.57)\qquad (\underline{\underline{D}}_2 v)_{ij} = \frac{\partial^2 v}{\partial x_i \partial x_j}, \qquad 1 \le i,\ j \le 2,$$

and the operator $\mathfrak{M}$ by

$$(3.58)\qquad \mathfrak{M}(\underline{\underline{\tau}}) = \begin{pmatrix} \tau_{11} + \nu\tau_{22} & (1-\nu)\tau_{12} \\ (1-\nu)\tau_{12} & \nu\tau_{11} + \tau_{22} \end{pmatrix}$$

for any symmetric tensor $\underline{\underline{\tau}}$. Using the same kind of analysis as in the previous examples we then get the saddle point problem

$$(3.59)\qquad \begin{cases} \inf\limits_{\underline{\underline{\tau}}\in(L^2(\Omega))^4_s} \sup\limits_{v\in H^2_0(\Omega)} \frac{1}{2}\left(\frac{12(1-\nu^2)}{Et^3}\right)\int_\Omega \mathfrak{M}^{-1}(\underline{\underline{\tau}}) : \underline{\underline{\tau}}\, dx \\ \qquad\qquad - \int_\Omega \underline{\underline{\tau}} : \underline{\underline{D}}_2 v\, dx + \int_\Omega fv\, dx, \end{cases}$$

where $(L^2(\Omega))^4_s$ is the space of square integrable 2×2 symmetric tensors. We introduce, as dual variables, the bending moments, obtained from the second derivatives of the primal solution u by

$$(3.60)\qquad \underline{\underline{\sigma}} = \frac{Et^3}{12(1-\nu^2)}\mathfrak{M}(\underline{\underline{D}}_2 u),$$

or explicitly

$$(3.61)\qquad \begin{cases} \sigma_{11} = \frac{Et^3}{12(1-\nu^2)}\left(\frac{\partial^2 u}{\partial x_1^2} + \nu\frac{\partial^2 u}{\partial x_2^2}\right), \\ \sigma_{22} = \frac{Et^3}{12(1-\nu^2)}\left(\nu\frac{\partial^2 u}{\partial x_1^2} + \frac{\partial^2 u}{\partial x_2^2}\right), \\ \sigma_{12} = \frac{Et^3}{12(1+\nu)}\frac{\partial^2 u}{\partial x_1 \partial x_2}. \end{cases}$$

The dual problem can then be written as

$$(3.62)\qquad \inf_{\underline{\underline{\tau}}} \frac{1}{2}\left(\frac{12}{Et^3}\right)\int_\Omega [(\tau_{11}+\tau_{22})^2 + 2(1+\nu)(\tau_{12}^2 - \tau_{11}\tau_{22})]\, dx$$

under the constraint

$$(3.63)\qquad D_2^*\underline{\underline{\tau}} = f.$$

In (3.63) we denoted by D_2^* the transpose of the operator $\underline{\underline{D}}_2$ so that

$$(3.64)\qquad D_2^*\underline{\underline{\tau}} = \frac{\partial^2\tau_{11}}{\partial x_1^2} + 2\frac{\partial^2\tau_{12}}{\partial x_1\partial x_2} + \frac{\partial^2\tau_{22}}{\partial x_2^2}.$$

It is possible, as in the previous case, to integrate by parts expression (3.59) and to obtain formulations in different functional spaces. We shall see an example of such a procedure in Section IV.5. □

I.3.3 Duality methods for nonsymmetric bilinear forms

In all previous examples, our variational formulations were based on a minimization problem for a functional and we were led to introduce a genuine saddle point problem. Even if this classical framework is suitable for a first presentation, it is not the sole possibility and the techniques developed can also be applied to problems which are not optimization problems. Let us consider for instance in $H_0^1(\Omega)$ a *continuous* and *coercive* bilinear form $a(u,v)$. If we do not require $a(\cdot,\cdot)$ to be symmetric, the variational problem

$$a(u,v) = \int_\Omega fv\,dx, \qquad \forall v \in H_0^1(\Omega), \tag{3.65}$$

has for $f \in L^2(\Omega)$ a unique solution $u \in H_0^1(\Omega)$ but does not correspond to the minimization of any functional. To fix ideas, let us suppose that $a(u,v)$ can be written as

$$a(u,v) = m(\underline{\text{grad}}\, u,\ \underline{\text{grad}}\, v) = \int_\Omega M(\underline{\text{grad}}\, u)\cdot \underline{\text{grad}}\, v\,dx, \tag{3.66}$$

where $m(\cdot,\cdot)$ is a continuous bilinear form on $(L^2(\Omega))^2$, which, of course, is nonsymmetric, and M is the associated linear operator from $(L^2(\Omega))^2$ into $(L^2(\Omega))^2$. We can now introduce the auxiliary variable,

$$\underline{p} = M(\underline{\text{grad}}\, u) \tag{3.67}$$

and write problem (3.65) in the form

$$\begin{cases} \displaystyle\int_\Omega \underline{p}\cdot \underline{\text{grad}}\, v\,dx = \int_\Omega fv\,dx, \\ \displaystyle\int_\Omega M^{-1}\underline{p}\cdot \underline{q}\,dx = \int_\Omega q\cdot \underline{\text{grad}}\, u\,dx. \end{cases} \tag{3.68}$$

This can be integrated by parts to yield, as in Example 3.5: $\underline{p} \in H(\text{div};\Omega)$, $u \in L^2(\Omega)$ and

$$\begin{cases} \displaystyle\int_\Omega \text{div}\,\underline{p}\,v\,dx + \int_\Omega fv\,dx = 0, & \forall v \in L^2(\Omega), \\ \displaystyle\int_\Omega M^{-1}\underline{p}\cdot \underline{q}\,dx + \int_\Omega u\,\text{div}\,\underline{q}\,dx = 0, & \forall \underline{q} \in H(\text{div};\Omega). \end{cases} \tag{3.69}$$

We shall thus consider in Chapter II problems such as (3.69) without making reference to a saddle point problem. The same remark would apply to the methods of the following section. □

I.4 Domain Decomposition Methods, Hybrid Methods

We have shown in Section I.3 that duality techniques enable us to obtain alternate variational formulations for some problems. The method that we shall now describe will yield a new family of variational principles that can be more or less grouped under the name of hybrid methods. The common point between the examples that follow is that in all cases the variational principle will depend explicitly, independently of any discretization, on a partition of the domain Ω into subdomains. To make clearer some of the facts that will appear later, we first recall a very classical result.

Example 4.1: *A transmission problem.*
We consider the very classical case in which a domain Ω is split into two subdomains Ω_1 and Ω_2 by a smooth enough internal boundary S (Figure I.2). We consider the case of a Dirichlet problem with variable coefficient $a(x)$, $a(x)$ being discontinuous on S. This classically leads to the variational problem [where we want to find $u \in H_0^1(\Omega)$]

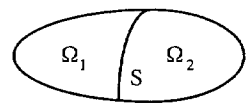


Figure I.2

$$\int_{\Omega_1} a_1(x)\ \underline{\text{grad}}\, u \cdot \underline{\text{grad}}\, v\ dx + \int_{\Omega_2} a_2(x)\ \underline{\text{grad}}\, u \cdot \underline{\text{grad}}\, v\ dx = \int_{\Omega} f v\ dx, \qquad \forall v \in H_0^1(\Omega). \tag{4.1}$$

Defining $u_1 = u|_{\Omega_1}$ and $u_2 = u|_{\Omega_2}$, it is standard to interpret (formally) problem (4.1) in the form

$$\begin{cases} -\operatorname{div}(a_1(x)\ \underline{\text{grad}}\, u_1) = f \text{ in } \Omega_1, \\ -\operatorname{div}(a_2(x)\ \underline{\text{grad}}\, u_2) = f \text{ in } \Omega_2, \\ u_1|_{\Gamma\cap\partial\Omega_1} = 0,\ u_2|_{\Gamma\cap\partial\Omega_2} = 0, \end{cases} \tag{4.2}$$

$$u_1 = u_2 \text{ on } S, \qquad a_1 \frac{\partial u_1}{\partial n_1} + a_2 \frac{\partial u_2}{\partial n_2} = 0 \text{ on } S, \tag{4.3}$$

where n_1 and n_2 are obviously the exterior normals to Ω_1 and Ω_2 (respectively) on S. Continuity conditions (4.3) are implicitly contained in the variational formulation. An important special case is $a_1(x) = a_2(x) = 1$. We then get the following result. □

Proposition 4.1: Let u be solution of the Dirichlet problem

$$\begin{cases} -\Delta u = f, \\ u|_\Gamma = 0. \end{cases} \tag{4.4}$$

Let the internal boundary S split Ω into Ω_1 and Ω_2. Then it is equivalent to say that u is solution of the problem

$$\begin{cases} -\Delta u_1 = f \text{ in } \Omega_1, \\ -\Delta u_2 = f \text{ in } \Omega_2, \\ u_1|_{\Gamma\cap\partial\Omega_1} = 0,\ u_2|_{\Gamma\cap\partial\Omega_2} = 0, \end{cases} \tag{4.5}$$

$$\begin{cases} u_1 = u_2 \text{ on } S, \\ \dfrac{\partial u_1}{\partial n_1} + \dfrac{\partial u_2}{\partial n_2} = 0 \text{ on } S. \end{cases} \tag{4.6}$$

To show this result we would have to define properly the normal derivatives $\partial u_1/\partial n_1$ and $\partial u_2/\partial n_2$ on S. This would require some regularity on f, for instance $f \in L^2(\Omega)$. □

What we really want to do is to consider a general partition of Ω

$$\bar{\Omega} = \bigcup_{i=1}^{N} \bar{K}_i. \tag{4.7}$$

We now write the classical Dirichlet functional of Example 2.1, in the following apparently strange way.

Example 4.2: *A domain decomposition method for Dirichlet problem.* Writing the Dirichlet functional as

$$J(v) = \sum_{i=1}^{N} \left\{ \frac{1}{2} \int_{K_i} |\ \underline{\text{grad}}\ v|^2\, dx - \int_{K_i} fv\, dx \right\} \tag{4.8}$$

and introducing now the functional space

$$X(\Omega) = \{v|\ v \in L^2(\Omega), v|_{K_i} \in H^1(K_i)\} \approx \prod_{i=1}^{N} H^1(K_i), \tag{4.9}$$

we can extend $J(v)$ on $X(\Omega)$. Moreover $H_0^1(\Omega)$ is a closed subspace of $X(\Omega)$ and we may consider "$v \in H_0^1(\Omega)$" as a linear constraint on $v \in X(\Omega)$. This constraint states that on $e_{ij} = \partial K_i \cap \partial K_j$ we must have, in $H^{1/2}(e_{ij})$,

$u_i = u_j$, where $u_\ell = u|_{K_\ell}$. We shall therefore, following a now familiar procedure, impose this constraint through a Lagrange multiplier properly chosen in $H^{-1/2}(e_{ij})$. As we shall see in Chapter III, it will be more convenient to introduce $\underline{q} \in H(\mathrm{div};\Omega)$ and to use as a multiplier the normal trace of $\underline{q}$ on ∂K_i. This leads us to the saddle point problem

$$(4.10)\quad \inf_{v\in X(\Omega)} \sup_{\underline{q}\in H(\mathrm{div},\Omega)} \sum_{i=1}^{N} \left\{ \frac{1}{2}\int_{K_i} |\,\underline{\mathrm{grad}}\, v|^2\, dx - \int_{\partial K_i} \underline{q}\cdot \underline{n}_i\, v\, ds - \int_{K_i} f v\, dx \right\}$$

for which we have the following optimality conditions: for $i = 1,\dots,N$,find $u_i \in H^1(K_i)$ such that,

$$(4.11)\quad \int_{K_i} \underline{\mathrm{grad}}\, u_i \cdot \underline{\mathrm{grad}}\, v_i\, dx = \int_{K_i} f v_i\, dx + \int_{\partial K_i} \underline{p}\cdot \underline{n}_i\, v_i\, ds,\ \forall v_i \in H^1(K_i),$$

$$(4.12)\qquad \sum_{i=1}^{N} \int_{\partial K_i} \underline{q}\cdot \underline{n}_i\, u_i\, ds = 0, \qquad \forall \underline{q} \in H(\mathrm{div};\Omega).$$

Condition (4.12) expresses continuity of u at interfaces e_{ij} and condition $u|_\Gamma = 0$. Condition (4.11) shows that u_i is solution in K_i of a Neumann problem

$$(4.13)\qquad \begin{cases} -\Delta u_i = f \text{ in } K_i, \\ \dfrac{\partial u_i}{\partial n_i} = \underline{p}\cdot \underline{n}_i \text{ on } \partial K_i. \end{cases}$$

Solving this problem obviously requires [make $v_i = 1$ in (4.11)] a compatibility condition

$$(4.14)\qquad \int_{\partial K_i} \underline{p}\cdot \underline{n}_i\, ds + \int_{K_i} f\, dx = 0,$$

on every subdomain K_i. This condition can also be written

$$(4.15)\qquad \int_{K_i} (\mathrm{div}\, \underline{p} + f)\, dx = 0.$$

From (4.13) we have that the multiplier $\underline{p}\cdot \underline{n}$ can be seen as the normal derivative of u. Indeed, when equilibrium is attained, we have on interfaces $\partial u_i/\partial n_i = \underline{p}\cdot \underline{n}_i = -\underline{p}\cdot \underline{n}_j = -\partial u_j/\partial n_j$ and $u_i = u_j$. A suitable lifting of $\underline{p}$ in each K_i in order to have $\mathrm{div}\, \underline{p} + f = 0$ can always be done because of (4.14) and (4.15). □

Example 4.3: *Dual problem of the domain decomposition method.*
We now consider the dual problem of the above saddle point formulations. It will be, as can be expected, very close to the dual problem introduced in Section I.3 for the Dirichlet problem. Let us first remark that taking the *infimum* on the constant part of $v \in X(\Omega)$ on each K_i leads to the *constraint* (4.15) on p. It is therefore possible to suppose $\operatorname{div} \underline{p} + f = 0$, as this can be attained by modifications to p that are internal to K_i (that is, not modifying $\underline{p} \cdot \underline{n}_i$) and are transparent to formulation (4.10). Writing

$$\int_{\partial K_i} \underline{q} \cdot \underline{n}_i \, v \, ds = \int_{K_i} \operatorname{div} \underline{q} \, v \, dx + \int_{K_i} \underline{q} \cdot \underline{\operatorname{grad}} \, v \, dx, \tag{4.16}$$

one gets from (4.10)

$$\sup_{\operatorname{div} \underline{q}+f=0} \inf_{v_i \in H^1(K_i)/\mathbb{R}} \sum_{i=1}^{N} \left\{ \frac{1}{2} \int_{K_i} |\underline{\operatorname{grad}} \, v_i|^2 \, dx - \int_{K_i} \underline{q} \cdot \underline{\operatorname{grad}} \, v_i \, dx \right\}. \tag{4.17}$$

From (4.17) we evidently get, setting $\underline{p}_i = \underline{p}|_{K_i}$,

$$\underline{\operatorname{grad}} \, u_i = P(\underline{p}_i), \tag{4.18}$$

where P is the projection operator in $(L^2(K_i))^2$ on $\underline{\operatorname{grad}}(H^1(K_i))$. We shall indeed prove in Chapter III that one has

$$(L^2(\Omega))^n = \{\underline{\operatorname{grad}} \, H^1(\Omega)\} \oplus \underline{\operatorname{rot}} \, H_0^1(\Omega)\}. \tag{4.19}$$

From this we can eliminate v_i and write the dual problem

$$\sup_{\substack{\underline{q} \in H(\operatorname{div},\Omega) \\ \operatorname{div} \underline{q}+f=0}} -\frac{1}{2} \sum_{i=1}^{N} \int_{K_i} |P(\underline{q}_i)|^2 \, dx. \tag{4.20}$$

We are therefore back to a variant of (3.35). Indeed, (3.35) shows that the projection operator P in (4.20) is unnecessary. □

Remark 4.1: One could obtain a variant of the above dual problem, without constraint (4.15) by using a "least-squares" solution of (4.13) whenever (4.14) does not hold. This could be done, for instance by solving on K_i, in a weak formulation that we shall not describe,

$$\begin{cases} \Delta^2 u_i = \Delta f \text{ in } K_i, \\ \dfrac{\partial}{\partial n_i} \Delta u_i = \dfrac{\partial f}{\partial n_i} \text{ on } \partial K_i, \\ \dfrac{\partial u_i}{\partial n_i} = \underline{q} \cdot \underline{n}_i \text{ on } \partial K_i, \end{cases} \tag{4.21}$$

for which a solution always exists, defined up to an additive constant. Such a procedure could be useful for algorithmic purposes since (4.21) is a local simple problem even if it is a fourth-order problem. □

Example 4.4: *Dual hybrid methods.*
We consider now the dual problem (3.35), that is, the complementary energy principle, that we now pose in $H(\mathrm{div};\Omega)$,

$$\inf_{\substack{\underline{q}\in H(\mathrm{div},\Omega)\\ \mathrm{div}\,\underline{q}+f=0}} \frac{1}{2}\int_\Omega |\underline{q}|^2\,dx. \tag{4.22}$$

We can apply the domain decomposition principle to such a problem by introducing

$$Y(\Omega) = \{\underline{q} \mid \underline{q}|_{K_i} \in H(\mathrm{div};K_i)\} \approx \prod_{i=1}^{N} H(\mathrm{div};K_i). \tag{4.23}$$

As we shall see in Chapter III, $H(\mathrm{div};\Omega)$ is now a closed subspace of $Y(\Omega)$ characterized by

$$\sum_{i=1}^{N}\int_{\partial K_i}(\underline{p}\cdot\underline{n}_i)\,v\,ds = 0, \qquad \forall v \in H_0^1(\Omega). \tag{4.24}$$

We can then transform (4.22) into the saddle point problem

$$\inf_{q\in Y(\Omega)}\ \sup_{v\in H_0^1(\Omega)} \sum_{i=1}^{N}\left\{\frac{1}{2}\int_{K_i}|\underline{q}_i|^2\,dx + \int_{\partial K_i}\underline{q}_i\cdot\underline{n}_i\,v\,ds\right\} \tag{4.25}$$

under the local constraint

$$\mathrm{div}\,\underline{q}_i + f = 0 \text{ on } K_i. \tag{4.26}$$

An advantage of this formulation is that it is easy to find $\underline{q}_i$ satisfying (4.26). We shall meet discretization methods, based on such a principle, under the name of *dual hybrid methods* for the treatment of almost any example considered in this book: Dirichlet problems, elasticity problems, fourth-order problems, etc. □

Example 4.5: *The Hellan–Hermann–Johnson method in elasticity.*
This is an example in which a domain decomposition is introduced, not by dualizing a continuity condition but by defining a variational formulation able to bypass this continuity by approximating weak derivatives. We shall first present formal results and delay a precise presentation of the functional framework. Our starting point will be the saddle point problem (3.48) and its optimality conditions (3.49) that we write, in variational form (with functional spaces to be defined),as

$$\frac{1}{\mu}\int_\Omega \underline{\underline{\sigma}}^D : \underline{\underline{\tau}}^D\,dx + \frac{1}{2(\lambda+\mu)}\int_\Omega \mathrm{tr}\,\underline{\underline{\sigma}}\ \mathrm{tr}\,\underline{\underline{\tau}}\,dx + \int_\Omega \underline{\underline{\varepsilon}}(u)\cdot\underline{\underline{\tau}}\,dx = 0, \qquad \forall \underline{\underline{\tau}} \in \underline{\underline{H}}(\mathrm{div};\Omega)_s, \tag{4.27}$$

$$(4.28)\qquad \int_\Omega \underline{\underline{\varepsilon}}(v) : \underline{\underline{\sigma}}\, dx + \int_\Omega \underline{f} \cdot \underline{v}\, dx = 0, \qquad \forall \underline{v} \in (H_0^1(\Omega))^2.$$

These conditions make sense for a space of $\underline{\underline{\sigma}}$ chosen so that $\underline{\mathrm{div}}\ \underline{\underline{\sigma}}$ is well defined, which implies, as we have seen, continuity of $\underline{\sigma}_n$ at interfaces. On the other hand $\underline{v}$ can be taken as completely discontinuous on these same interfaces. What we now try to do is to split continuity conditions between $\underline{s}_n$ and $\underline{v}$. Let us consider indeed the well-known integration by parts formula,

$$(4.29)\quad \int_\Omega \underline{\mathrm{div}}\ \underline{\underline{\sigma}} \cdot \underline{v}\, dx + \int_\Omega \underline{\underline{\sigma}} : \underline{\underline{\varepsilon}}(\underline{v})\, dx = \int_{\partial\Omega} \sigma_{nn}\, \underline{v} \cdot \underline{n}\, ds + \int_{\partial\Omega} \sigma_{nt}\, \underline{v} \cdot \underline{t}\, ds$$

Whenever $\underline{v}$ is a smooth [let us say $H^1(\Omega)$] vector, and σ_{nt} is continuous, we thus have

$$(4.30)\qquad \int_\Omega \underline{\underline{\varepsilon}}(\underline{v}) : SI\, dx = -\sum_{i=1}^N \left\{ \int_{K_i} \underline{\mathrm{div}}\ \underline{\underline{\sigma}} \cdot \underline{v}\, dx + \int_{\partial K_i} \sigma_{nn}\, \underline{v} \cdot \underline{n}\, ds \right\},$$

so that we can rewrite (4.27) and (4.28) in the following form:

$$(4.31)\quad \frac{1}{\mu} \int_\Omega \underline{\underline{\sigma}}^D : \underline{\underline{\tau}}^D\, dx + \frac{1}{2(\lambda+\mu)} \int_\Omega \mathrm{tr}\ \underline{\underline{\sigma}}\ \mathrm{tr}\ \underline{\underline{\tau}}\, dx$$
$$+ \sum_{i=1}^N \left\{ \int_{K_i} \underline{\mathrm{div}}\ \underline{\underline{\tau}} \cdot \underline{u}\, dx - \int_{\partial K_i} \tau_{nn}\, \underline{u} \cdot \underline{n}\, ds \right\} = 0, \qquad \forall \underline{\underline{\tau}},$$

$$(4.32)\quad \sum_{i=1}^N \left\{ \int_{K_i} \underline{\mathrm{div}}\ \underline{\underline{\sigma}} \cdot \underline{v}\, dx - \int_{\partial K_i} \sigma_{nn}\, \underline{v} \cdot \underline{n}\, ds \right\} + \int_\Omega \underline{f} \cdot \underline{v}\, dx = 0, \qquad \forall \underline{v}.$$

Formally this is well defined for $\underline{\underline{\sigma}}$ chosen with σ_{nt} continuous at interfaces while $\underline{u} \cdot \underline{n}$ is continuous. Then the term

$$(4.33)\qquad \sum_{i=1}^N \left\{ \int_{\partial K_i} \sigma_{nn}\, \underline{v} \cdot \underline{n}\, ds \right\},$$

therefore depends on the *jump* of σ_{nn} on ∂K_i and (4.32) can be read as $\underline{\mathrm{div}}\, \underline{\underline{\sigma}} + \underline{f} = 0$ in the sense of distributions. We shall consider in Chapter VI a discretization of problem (4.31) and (4.32) for $\lambda = +\infty$ (that is, $\mathrm{tr}\ \underline{\underline{\sigma}} = 0$), i.e., the case of an incompressible material. As we shall see, our main problem will then be to preserve symmetry in the discretized problem.

Up to now we considered a purely formal problem. Giving a good framework to (4.31) and (4.32) is a task that requires some care. The presence of traces, appearing explicitly, in the variational formulation leads one to deal with

spaces $H^{1/2}(\partial K_i)$ and $H^{-1/2}(\partial K_i)$ and to subtle considerations about the behavior of functions in these pathological spaces. Let us define

$$(4.34)\quad \Sigma = \prod_{K_i} (H^1(K_i))^4_s = \{\underline{\underline{\sigma}} \in (L^2(\Omega))^4_s, \quad \sigma_{ij}|_{K_i} \in H^1(K_i),\ \sigma_{ij} = \sigma_{ji}\}.$$

This is a space of smooth tensors and we can consider σ_{nt} on each interface $e_{ij} = \partial K_i \cap \partial K_j$ (cf. Chapter III). We have $\sigma_{nt} \in H^{1/2}(e_{ij})$ but we do not have $\sigma_{nt} \in H^{1/2}(\partial K_i)$; this would require some continuity at vertices which cannot, in general, take place due to the change of direction of $\underline{n}$ and $\underline{t}$. We can nevertheless consider in $\sum$, tensor functions $\underline{\underline{\sigma}}$ such that σ_{nt} is continuous on e_{ij}. To make (4.33) meaningful, we now have to choose $\underline{v}$ with $\underline{v} \cdot \underline{n}$ continuous on e_{ij}. We have already seen that for $\underline{v}$ in $H(\text{div};\ K_i)$ we can define $\underline{v} \cdot \underline{n}$ in $H^{-1/2}(\partial K_i)$. Unfortunately it is not possible to restrict $\underline{v} \cdot \underline{n}|_{e_{ij}}$ and get a result in $H^{-1/2}(e_{ij})$: something is lost in corners. In reality we only need an "infinitesimal" amount of extra smoothness and this will lead us to look for $\underline{v}$ in $(L^p(\Omega))^2 \cap H(\text{div};\Omega)$ for $p > 2$. This will cause some problems in applying the theory of Chapter II and existence of a solution will have to be deduced through special considerations. □

I.5 Augmented Variational Formulations

We shall present in this section other possible ways of defining variational principles associated with saddle point problems. The methods that we shall consider are known as Galerkin least-squares methods and were introduced by FRANCA–HUGHES [A]. To fix the ideas, let us consider the simplest cases of Examples 3.4 and 3.5, and in particular the saddle point formulations (3.32) and (3.39), respectively. In both cases the Euler equations are given by (3.34). It is clear that we can always add (or substract) the square of one of these equations to the functional without changing the min–max point. For instance, we can take (3.32) and add to it the square of the first equation of (3.34) to obtain

$$(5.1)\quad \inf_{v \in H^1_0(\Omega)} \sup_{\underline{q} \in (L^2(\Omega))^2} \left\{ -\tfrac{1}{2}\int_\Omega |\underline{q}|^2 dx - \int_\Omega f v\, dx + \int_\Omega \underline{q} \cdot \underline{\text{grad}}\, v\, dx + \tfrac{\alpha}{2}\int_\Omega |\underline{q} - \underline{\text{grad}}\, v|^2 dx \right\},$$

where α can be chosen arbitrarily provided $0 \le \alpha \le 1$. Similarly one can add, instead, the square of the second equation of (3.34) to the functional (3.39) to get

$$(5.2)\quad \inf_{\underline{q} \in H(\text{div},\Omega)} \sup_{v \in L^2(\Omega)} \left\{ \tfrac{1}{2}\int_\Omega |\underline{q}|^2 dx + \int_\Omega f v\, dx + \int_\Omega v\, \text{div}\, \underline{q}\, dx + \tfrac{\beta}{2}\int_\Omega (\text{div}\, \underline{q} + f)^2 dx \right\},$$

where $\beta \geq 0$ is arbitrary.

A third (reasonable) possibility is available: one might take (3.39), add to it the square of the second equation of (3.34) and substract the square of the first equation of (3.34):

$$\inf_{\underline{q}\in H(\mathrm{div},\Omega)} \sup_{v\in H_0^1(\Omega)} \left\{ \tfrac{1}{2}\int_\Omega |\underline{q}|^2 dx + \int_\Omega fv\,dx + \int_\Omega v \operatorname{div} \underline{q}\,dx \right.$$
$$\left. + \tfrac{\beta}{2}\int_\Omega (\operatorname{div}\underline{q}+f)^2 dx - \tfrac{\alpha}{2}\int_\Omega |\underline{q}-\underline{\mathrm{grad}}\,v|^2 dx \right\}. \tag{5.3}$$

Note that we had to change the regularity requirements on v in order to make the functional meaningful. Note as well that we could obtain (5.3) from (5.1) by substracting $(\beta/2)\,\|\operatorname{div}\underline{q}+f\|^2_{L^2(\Omega)}$ (and increasing the regularity requirements on $\underline{q}$) and by changing the sign. An easy computation shows that the Euler equations of (5.1) are

$$\int_\Omega \underline{p}\cdot\underline{q}\,dx + \int_\Omega fv\,dx - \int_\Omega \underline{p}\cdot\underline{\mathrm{grad}}\,v\,dx - \int_\Omega \underline{q}\cdot\underline{\mathrm{grad}}\,u\,dx$$
$$- \alpha\int_\Omega (\underline{p}-\underline{\mathrm{grad}}\,u)\cdot(\underline{q}-\underline{\mathrm{grad}}\,v)dx = 0,\ \forall(\underline{q},v)\in (L^2(\Omega))^2\times H_0^1(\Omega) \tag{5.4}$$

which are equivalent to (3.34) for $\alpha \neq 1$. For $\alpha = 1$, we just have

$$\int_\Omega \underline{\mathrm{grad}}\,u\cdot\underline{\mathrm{grad}}\,v\,dx = \int_\Omega fv\,dx, \qquad \forall v\in H_0^1(\Omega),$$

as should be expected since (5.1) reduces to (3.30) in this case. Similarly, (5.2) gives

$$\int_\Omega \underline{p}\cdot\underline{q}\,dx + \int_\Omega fv\,dx + \int_\Omega u\operatorname{div}\underline{q}\,dx + \int_\Omega v\operatorname{div}\underline{p}\,dx$$
$$+\beta\int_\Omega (\operatorname{div}\underline{p}+f)\operatorname{div}\underline{q}\,dx = 0, \qquad \forall(\underline{q},v)\in H(\mathrm{div};\Omega)\times L^2(\Omega), \tag{5.5}$$

which are again equivalent to (3.34). Finally, (5.3) gives

$$\int_\Omega \underline{p}\cdot\underline{q}\,dx + \int_\Omega fv\,dx + \int_\Omega u\operatorname{div}\underline{q}\,dx + \int_\Omega v\operatorname{div}\underline{p}\,dx$$
$$+\beta\int_\Omega (\operatorname{div}\underline{p}+f)\operatorname{div}\underline{q}\,dx - \alpha\int_\Omega (\underline{p}-\underline{\mathrm{grad}}\,u)\cdot(\underline{q}-\underline{\mathrm{grad}}\,v)dx = 0,$$
$$\forall(\underline{q},v)\in H(\mathrm{div};\Omega)\times H_0^1(\Omega). \tag{5.6}$$

Setting now $\psi = \operatorname{div}\underline{p}+f$ and $\underline{\phi} = \underline{p}-\underline{\mathrm{grad}}\,u$, we may rewrite (5.6) as

$$\begin{cases} (1-\alpha)\displaystyle\int_\Omega \underline{\phi}\cdot\underline{q}\,dx + \beta\int_\Omega \psi\operatorname{div}\underline{q}\,dx = 0, & \forall\underline{q}\in H(\underline{\mathrm{div}},\Omega), \\ \alpha\displaystyle\int_\Omega \underline{\phi}\cdot\underline{\mathrm{grad}}\,v\,dx + \int_\Omega \psi v\,dx = 0, & \forall v\in H_0^1(\Omega). \end{cases} \tag{5.7}$$

But taking $\underline{q} = \underline{\text{grad}}\, v$, with $v \in H^2(\Omega) \cap H_0^1(\Omega)$, we get from (5.7)

$$(5.8) \quad -(1-\alpha)\int_\Omega \psi v\, dx + \alpha\beta \int_\Omega \psi \triangle v\, dx = 0, \qquad \forall v \in H^2(\Omega) \cap H_0^1(\Omega).$$

(For $\alpha = 1$, we have $-\triangle u = f$ again. In conclusion, (5.6) is equivalent to (3.34).

Remark 5.1: We obtained in this way three more equivalent variational formulations for the original problem (3.30). Note however that the Euler equations (3.34) constitute a system of two first-order equations in the unknowns u and $\underline{p}$; on the contrary (5.4) is of second order in u and first order in $\underline{p}$, whereas (5.5) is of first order in u and second order in $\underline{p}$ and finally (5.6) is of second order in both variables. □

Remark 5.2: We presented here, on an example, a quite general idea which was presented in a general setting in FRANCA–HUGHES [A] and FRANCA [A]. Some examples of possible uses of these ideas will be developed in the following chapters. □

Remark 5.3: In the example presented above, Euler equations (3.34) were a system of first-order equations. This is not the case of Stokes problem (3.14), for which one of the equations is already second order. Applying the same procedure would lead to a fourth-order problem in the variable $\underline{u}$ which would lead to undesirable complications. Indeed the analogue of (5.1) would here be obtained from (3.10) as follows:

$$(5.9) \quad \begin{aligned} \inf_{\underline{v}\in(H_0^1(\Omega))^2} \sup_{q\in L^2(\Omega)} \mu \int_\Omega |\underline{\varepsilon}(v)|^2 dx - \int_\Omega \underline{f}\cdot\underline{v}\, dx - \int_\Omega q\, \text{div}\, \underline{v}\, dx \\ - \frac{\alpha}{2}\int_\Omega |-\triangle \underline{u} + \underline{\text{grad}}\, p - f|^2 dx, \end{aligned}$$

which would force us to use a very regular approximation for the variable $\underline{u}$. A possible one could be to employ this method in connection with a domain decomposition, Ω being partitioned into subdomains as in (4.7) and to change the last integral into

$$(5.10) \qquad -\frac{\hat{\alpha}}{2}\sum_{i=1}^{N}\int_{K_i} |-\triangle \underline{u} + \underline{\text{grad}}\, p - f|^2 dx$$

where $\hat{\alpha}$ will have to be suitably scaled. This is what has been done by HUGHES–FRANCA [A]. We shall come back to this in Chapter VI. □

Remark 5.4: Finally, another variant of the general idea developed above has been considered in DOUGLAS–WANG [A]. The formulation cannot in this case be written as a modified Lagrangian but must be introduced as a modification of Euler equations of (5.4). Indeed let us write instead of (5.4)

$$(5.11)\quad \int_\Omega \underline{p}\cdot\underline{q}\,dx + \int_\Omega fv\,dx - \int_\Omega \underline{p}\cdot\underline{\mathrm{grad}}\,v\,dx - \int_\Omega \underline{q}\cdot\underline{\mathrm{grad}}\,u\,dx$$
$$+\,\alpha\int_\Omega (\underline{p}-\underline{\mathrm{grad}}\,u)\cdot(\underline{q}+\underline{\mathrm{grad}}\,v)dx = 0, \quad \forall(\underline{q},v)\in (L^2(\Omega))^2\times H_0^1(\Omega).$$

This formulation cannot be obtained from a Lagrangian. It can easily be seen that it remains valid for $\alpha \geq 0$ arbitrary. Indeed (5.11) can also be writen as (5.12)

$$\begin{cases}(1+\alpha)\displaystyle\int_\Omega (\underline{p}-\underline{\mathrm{grad}}\,u)\,\underline{q}\,dx = 0, & \forall\underline{q}\in (L^2(\Omega))^2,\\ \alpha\displaystyle\int_\Omega \underline{\mathrm{grad}}\,u\cdot\underline{\mathrm{grad}}\,v\,dx+(1-\alpha)\int_\Omega \underline{p}\cdot\underline{\mathrm{grad}}\,v\,dx-\int_\Omega fv\,dx = 0, & \forall v\in H_0^1(\Omega)\end{cases}$$

and this is equivalent to (3.34) for any $\alpha \geq 0$. □

To end this chapter, we present a last type of variational formulation yielding weaker solutions than the formulations presented up to now.

I.6 Transposition Methods

Although we shall not consider in this book discretization methods directly based on transposition methods, some of the properties of these methods will be a good guide for understanding weak formulations. We present here the simplest possible case of a transposed problem and we refer to LIONS–MAGENES [A] for a complete discussion. Our starting point to obtain a weak formulation of a Dirichlet problem will be, paradoxically, a regularity result (AGMON–DOUGLIS–NIRENBERG [A], AGMON [A], NEČAS [A]). It is indeed well known that for $f \in L^2(\Omega)$ and when the boundary $\partial\Omega$ is smooth enough, the solution of the problem

$$(6.1)\qquad \begin{cases} -\Delta u = f,\\ u\in H_0^1(\Omega),\end{cases}$$

satisfies a regularity property

$$(6.2)\qquad u\in H^2(\Omega)\cap H_0^1(\Omega),$$

and then an a priori bound

$$(6.3)\qquad \|u\|_2 \leq c\,|f|_0.$$

One therefore has defined an isomorphism from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$ and it is then immediate that the transpose of this isomorphism is also an isomorphism from $L^2(\Omega)$ into the dual space $(H^2(\Omega) \cap H_0^1(\Omega))'$. Thus there exists a unique solution to problem

$$(6.4) \qquad -\int_\Omega u \, \Delta\phi \, dx = L(\phi), \qquad \forall \phi \in H^2(\Omega) \cap H_0^1, \; u \in L^2(\Omega),$$

for any continuous linear form $L(\cdot)$ on $H^2(\Omega) \cap H_0^1(\Omega)$. In particular, for $f \in L^2(\Omega)$ and $g \in H^{-1/2}$, the problem

(6.5)

$$-\int_\Omega u \, \Delta\phi \, dx = \int_\Omega f\phi \, dx - \int_\Gamma g \frac{\partial \phi}{\partial n} ds, \quad \forall \phi \in H^2(\Omega) \cap H_0^1(\Omega), \; u \in L^2(\Omega),$$

has a unique solution satisfying in the sense of distributions,

$$(6.6) \qquad -\Delta u = f$$

and in a weak sense (LIONS–MAGENES [A])

$$(6.7) \qquad u|_\Gamma = g.$$

One has solved a weak form of the Dirichlet problem with boundary values in $H^{-1/2}(\Gamma)$. In Example 3.5 we also had a weak form but boundary values had to be chosen in $H^{1/2}(\Gamma)$ and we were implicitly brought back to the strong problem.

It is also possible to define in $H^{-3/2}(\Gamma)$ the trace of the "normal derivative" of u. Indeed for every $\phi \in H^{3/2}(\Gamma)$ we can solve the problem:find $\Phi \in H^2(\Omega)$ such that,

$$(6.8) \qquad \begin{cases} -\Delta\Phi = 0 \text{ in } \Omega, \\ \Phi|_\Gamma = \phi \end{cases}$$

The normal derivative $P = \partial u / \partial n$ of u will then be the mapping $\phi \to \langle P, \phi \rangle$ defined as

$$(6.9) \qquad \langle P, \phi \rangle = -\int_\Omega f \, \Phi \, dx + \int_\Gamma g \frac{\partial \Phi}{\partial n} ds.$$

We shall have in chapter V to consider weak traces of the solution of a Dirichlet problem on interfaces between subdomains. Although we shall not make an explicit use of transposition, the presence of weak discrete norms in the error estimates indicates that we could then indeed get some insight from such a formulation.

I.7 Bibliographical Remarks

The purpose of this Chapter was to present examples which will be used later as a standing ground for our development. It was not possible in such a context to consider every case. We already referred the reader to DAUTRAY–LIONS [A] where the mathematical analysis of the problem selected, and many others, can be found in an unified setting. We also refer to more engineering oriented presentations such as BATHE [A], HUGHES [A], KIKUCHI–ODEN [A], and ZIENKIEWICZ [A]. In particular, nonlinear problems and their treatment are described in these references.

II

Approximation of Saddle Point Problems

This chapter is in a sense the kernel of the book. It sets a general framework in which mixed and hybrid finite element methods can be studied. Even if some applications will require variations of the general results, these could not be understood without the basic notions introduced here. Our first concern will be existence and uniqueness of solutions. We first consider in Section II.1.1 the simple case of a saddle point problem corresponding to the minimization of a linearly constrained quadratic functional. This case is extended in Section II.1.2 to a more general case. The matter of approximating the solution will then be considered under various (but classical) assumptions. Finally, we shall deal with numerical properties of the discretized problems and practical computational facts.

II.1 Existence and Uniqueness of Solutions

In the previous chapter, we introduced a large number of saddle point problems or generalizations of such problems. In most cases, the question of existence and uniqueness of solutions was left aside. We now introduce an abstract frame that is sufficiently general to cover all our needs. In order to make our presentation easier, we shall first consider the simpler case corresponding (under symmetry assumptions) to the minimization of a quadratic functional under linear constraints. We shall follow essentially the analysis of BREZZI [A] and FORTIN [C]. We also refer the reader to the paper of BABUŠKA [A] which was a fundamental step towards understanding mixed methods, and to the recent work of ROBERTS–THOMAS [A] for another general presentation of mixed methods.

II.1.1 Quadratic problems under linear constraints

Let V be some Hilbert space for the norm $\|\cdot\|_V$ and the scalar product $((\cdot,\cdot))_V$. We consider a continuous bilinear form on $V \times V$ (not being necessarily symmetric) and therefore satisfying

$$|a(u,v)| \leq \|a\| \, \|u\|_V \, \|v\|_V. \tag{1.1}$$

This bilinear form thus defines a linear continuous operator $A : V \to V'$ by

$$\langle Au, v\rangle_{V'\times V} = a(u,v), \qquad \forall v \in V, \ \forall u \in V. \tag{1.2}$$

Let us choose another Hilbert space Q, with norm $\|\cdot\|_Q$ and scalar product $((\cdot,\cdot))_Q$, and a continuous bilinear form $b(v,q)$ on $V \times Q$ with

$$|b(v,q)| \leq \|b\| \, \|v\|_V \, \|q\|_Q. \tag{1.3}$$

Again, we can introduce a linear operator $B : V \to Q'$, and its transpose $B^t : Q \to V'$, defined by,

$$\langle Bv, q\rangle_{Q'\times Q} = \langle v, B^t q\rangle_{V\times V'} = b(v,q), \ \forall v \in V, \qquad \forall q \in Q. \tag{1.4}$$

As we shall see, the properties of operator B are fundamental in the study of the problem; we consider in particular the range of B denoted $\operatorname{Im} B$ and its kernel $\operatorname{Ker} B$. *Let $f \in V'$, $g \in Q'$ be given; we want to find $u \in V$, $p \in Q$ solutions of*

$$\begin{cases} a(u,v) + b(v,p) = \langle f, v\rangle_{V'\times V}, & \forall v \in V, \\ b(u,q) = \langle g, q\rangle_{Q'\times Q}, & \forall q \in Q. \end{cases} \tag{1.5}$$

This can also be written as

$$\begin{cases} Au + B^t p = f \ \text{ in } V', \\ Bu = g \ \text{ in } Q'. \end{cases} \tag{1.6}$$

We now want to find conditions implying existence and possibly uniqueness of solutions to this problem. If the bilinear form $a(u,v)$ is symmetric, equations (1.5) are the optimality conditions of the saddle point problem

$$\inf_{v\in V} \sup_{q\in Q} \tfrac{1}{2} a(v,v) + b(v,q) - \langle f, v\rangle_{V'\times V} - \langle g, q\rangle_{Q'\times Q}. \tag{1.7}$$

This is the reason for the title of this chapter, in spite of the fact that we deal in fact with a more general case. We shall first give an existence and uniqueness result for a problem which is strongly related to (1.5). This result is a direct consequence of the classical Lax–Milgram theorem (CIARLET [A], LIONS [A]).

Proposition 1.1: Let $g \in \operatorname{Im} B$ and let the bilinear form $a(\cdot,\cdot)$ be coercive on $\operatorname{Ker} B$, that is, there exists α_0 such that

$$a(v_0, v_0) \geq \alpha_0 \, \|v_0\|_V^2, \qquad \forall v_0 \in \operatorname{Ker} B. \tag{1.8}$$

Then there exists a unique $u \in V$ solution of

$$a(u, v_0) = \langle f, v_0 \rangle_{V' \times V}, \qquad \forall v_0 \in \operatorname{Ker} B, \tag{1.9}$$

and

$$Bu = g. \tag{1.10}$$

Proof: The condition $g \in \operatorname{Im} B$ is of course necessary. Let us suppose it is satisfied; one can then find $u_g \in V$ with $Bu_g = g$. One then writes, in a classical way, the first equation of (1.5) in the form

$$a(u_0, v_0) = \langle f, v_0 \rangle_{V' \times V} - a(u_g, v_0), \qquad \forall v_0 \in \operatorname{Ker} B,\ u_0 \in \operatorname{Ker} B, \tag{1.11}$$

by setting $u = u_0 + u_g$ and taking $v = v_0 \in \operatorname{Ker} B$. A sufficient condition for the existence and uniqueness of u_0 is therefore the coercivity condition (1.8). There remains to check that $u = u_0 + u_g$ does not depend on the choice of u_g. Indeed if we had two solutions of (1.9) and (1.10) say u_1 and u_2 , we would have $u_1 - u_2 \in \operatorname{Ker} B$ and from (1.9)

$$a(u_1 - u_2, v_0) = 0, \qquad \forall v_0 \in \operatorname{Ker} B,$$

and this implies $u_1 - u_2 = 0$ by condition (1.8). □

Remark 1.1: It is clear that, if (1.5) has a solution (u, p), then u will be a solution of (1.9)-(1.10). Then Proposition 1.1 implies that the first component u of the solution (u, p) of (1.5) (if it exists) is unique. Moreover we note that by Proposition 1.1 we have

$$\|u\| \leq \|u_g\| + \frac{1}{\alpha_0}\Big\{\|f\|_{V'} + \|a\| \, \|u_g\|\Big\}. \ \square \tag{1.12}$$

We must therefore bound $\|u_g\|$ to get a proper a priori bound on u. □

Remark 1.2: The coercivity of $a(\cdot,\cdot)$ on $\operatorname{Ker} B$ may hold while there is no coercivity on V. We shall meet many examples of this situation. □

Remark 1.3: If the bilinear form $a(\cdot,\cdot)$ is symmetric, (1.9) and (1.10) are the optimality conditions of the minimization problem

$$\inf_{Bv=g} \frac{1}{2}a(v,v) - \langle f, v\rangle_{V'\times V}. \tag{1.13}$$

The variable p will be the Lagrange multiplier associated with the constraint $Bu = g$. □

We now turn to the problem of finding p. For this, we shall have to make an additional assumption on the operator B. Precisely the range of B, $\operatorname{Im} B$, will have to be closed in Q'. This will hold, in particular, in the frequently encountered cases where B is surjective or when $\operatorname{Im} B$ is of finite codimension.

This assumption (that $\operatorname{Im} B$ is closed) enables us to extend to the infinite-dimensional case, properties that are well-known to hold for matrices. We thus recall the following classical result of functional analysis (cf YOSIDA [A] for instance).

Proposition 1.2: The following statements are equivalent:

— $\operatorname{Im} B$ is closed in Q', that is for any sequence v_k such that Bv_k converges in Q', there exists $v \in V$ with $\lim_k Bv_k = Bv$.

— $\operatorname{Im} B^t$ is closed in V', that is for any sequence q_k such that $B^t q_k$ converges in V', there exists $q \in Q$ with $\lim_k B^t q_k = B^t q$.

— $(\operatorname{Ker} B)^0 = \{v' \in V' | \langle v', v\rangle_{V'\times V} = 0,\ \forall v \in \operatorname{Ker} B\} = \operatorname{Im} B^t$.

— $(\operatorname{Ker} B^t)^0 = \{q' \in Q' | \langle q', q\rangle_{Q'\times Q} = 0,\ \forall q \in \operatorname{Ker} B^t\} = \operatorname{Im} B$.

— There exists $k_0 > 0$ such that for any $g \in \operatorname{Im} B$, there exists $v_g \in V$ with $Bv_g = g$ and $\|v_g\|_V \le 1/k_0 \|g\|_{Q'}$.

— There exists $k_0 > 0$ such that for any $f \in \operatorname{Im} B^t$, there exists $q_f \in Q$ with $B^t q_f = f$ and $\|q_f\|_Q \le 1/k_0 \|f\|_{V'}$. □

If one of the above properties is satisfied, one can say that B admits *a continuous lifting* from Q' to V and B^t a continuous lifting from V' to Q. We have used the dual norms, $\|\cdot\|_{V'}$ and $\|\cdot\|_{Q'}$, defined by

$$\|f\|_{V'} = \sup_{v\in V} \frac{[g,q]}{\|v\|_V}, \qquad \|g\|_{Q'} = \sup_{q\in Q} \frac{\langle g, q\rangle}{\|q\|_Q}, \tag{1.14}$$

where, as in the rest of the book, we assumed implicitely that $\sup_x \phi(x)$ has to be taken for $\|x\| \neq 0$ if $\phi(x)$ contains $\|x\|$ in the denominator. The same will be true for $\inf_x \phi(x)$. The last two statements of Proposition 1.2 can then be written as

$$\sup_{q\in Q} \frac{b(v,q)}{\|q\|_Q} \ge k_0 \Big[\inf_{v_0 \in \operatorname{Ker} B} \|v + v_0\|_V\Big] = k_0 \|v\|_{V/\operatorname{Ker} B}, \qquad \forall v \in B, \tag{1.15}$$

and

$$(1.16)\quad \sup_{v\in V}\frac{b(v,q)}{\|v\|_V} \geq k_0\Big[\inf_{q_0\in \operatorname{Ker} B^t}\|q+q_0\|\Big] = k_0\|q\|_{Q/\operatorname{Ker} B^t}, \qquad \forall q\in Q,$$

by taking into account the fact that v_g can be chosen in $(\operatorname{Ker} B)^\perp$ and q_f in $(\operatorname{Ker} B^t)^\perp$, respectively. We can summarize by saying that operators with a closed range have the well-known properties of operators in finite-dimensional spaces (for which the range is always trivially closed). We can now proceed to study the full problem (1.5).

Proposition 1.3: Let $g\in \operatorname{Im} B$ and let u be the solution of problem (1.9) and (1.10). If $\operatorname{Im} B$ is closed in Q', there exists $p\in Q'$ such that (u,p) is solution of problem (1.5). Moreover, we have

$$(1.17)\qquad \|u\|_V \leq \frac{1}{\alpha_0}\|f\|_{V'} + \frac{1}{k_0}\Big(1+\frac{\|a\|}{\alpha_0}\Big)\|g\|_{Q'},$$

$$(1.18)\qquad \|p\|_{Q/\operatorname{Ker} B^t} \leq \frac{1}{k_0}\Big(1+\frac{\|a\|}{\alpha_0}\Big)\|f\|_{V'} + \frac{\|a\|}{k_0^2}\Big(1+\frac{\|a\|}{\alpha_0}\Big)\|g\|_{Q'}.$$

Proof: Indeed, let us consider in V', the linear form

$$(1.19)\qquad L(v) = \langle f,v\rangle_{V'\times V} - a(u,v).$$

By (1.9) we have $L(v_0)=0$, for any $v_0\in \operatorname{Ker} B$. By Proposition 1.2, we know that $L\in \operatorname{Im} B^t = (\operatorname{Ker} B)^0$, and there exists $p\in Q$ so that

$$(1.20)\qquad L(v) = b(v,p), \qquad \forall v\in V,$$

$$(1.21)\qquad \|p\|_{Q/\operatorname{Ker} B^t} \leq \frac{1}{k_0}\|L\|_{V'} \leq \frac{1}{k_0}\Big(\|a\|\,\|u\|_V + \|f\|_{V'}\Big).$$

From Proposition 1.2, we can also choose u_g in Proposition 1.1 to satisfy

$$(1.22)\qquad \|u_g\|_V \leq \frac{1}{k_0}\|g\|_{Q'}.$$

Thus (1.17) follows from (1.12) and (1.22) whereas (1.18) is readily deduced from (1.21) and (1.17). Finally, from (1.20) and (1.19) we have

$$(1.23)\qquad a(u,v)+b(v,p) = \langle f,v\rangle, \qquad \forall v\in V. \square$$

Remark 1.4: For *finite-dimensional* problems we shall always have existence of a solution (u, p) of (1.5) provided we have the existence of u in (1.9) and (1.10). □

Remark 1.5: If $\operatorname{Im} B$ is of finite codimension (and thus closed) it is possible to build explicitly the extension of $L(\cdot)$. □

Remark 1.6: It is clear (from the first equation of(1.6) for instance) that p is defined up to an arbitrary element of $\operatorname{Ker} B^t$. We then have *uniqueness* of p if and only if B is surjective. □

The proof of Proposition 1.1 shows that condition (1.8) is in fact too strong. What we really need is the condition that the restriction of A to $\operatorname{Ker} B$ be invertible on $\operatorname{Ker} B$. Equation (1.11) then defines u_0 in a unique way. This condition is equivalent to the condition that A_0, the restriction of A to $\operatorname{Ker} B$, is bijective and by Proposition 1.2 applied to A; this is equivalent to the conditions

$$\inf_{v_0 \in \operatorname{Ker} B} \sup_{u_0 \in \operatorname{Ker} B} \frac{a(u_0, v_0)}{\|u_0\| \, \|v_0\|} \geq \alpha_0, \tag{1.24}$$

which impliesKer $A_0^t = \{0\}$ that is, A_0 is surjective, and

$$\inf_{u_0 \in \operatorname{Ker} B} \sup_{v_0 \in \operatorname{Ker} B} \frac{a(u_0, v_0)}{\|u_0\| \, \|v_0\|} \geq \alpha_0, \tag{1.25}$$

which implies $\operatorname{Ker} A_0 = \{0\}$ that is, A_0 is injective. It must be pointed out that (1.25) for instance is not implied by a condition of the form

$$\inf_{u \in V} \sup_{v \in V} \frac{a(u, v)}{\|u\| \, \|v\|} \geq \alpha$$

which would mean A to be injective on V. The same remark holds for (1.24).

For example, we shall encounter in practice problems of the form

$$\begin{pmatrix} A & B_1^t & C_1^t \\ B_1 & 0 & C_2^t \\ C_1 & C_2 & 0 \end{pmatrix} \begin{pmatrix} U_1 \\ U_2 \\ P \end{pmatrix} = \begin{pmatrix} F_1 \\ F_2 \\ G \end{pmatrix}. \tag{1.26}$$

We would then have to check that

$$\begin{pmatrix} A & B_1^t \\ B_1 & O \end{pmatrix}$$

is invertible on $\operatorname{Ker} C = \operatorname{Ker}\{C_1, C_2\}$. It will not be sufficient to prove invertibility on V which is a problem of the type considered above and is in general much simpler.

We can now summarize the above results in the following.

Theorem 1.1: Let $a(\cdot,\cdot)$ be a continuous linear form on $V \times V$, let $b(\cdot,\cdot)$ be a continuous linear form on $V \times Q$. Let us suppose that the range of the operator B associated to $b(\cdot,\cdot)$ is closed in Q', that is, there exists $k_0 > 0$ such that

$$\sup_{v \in V} \frac{b(v,q)}{\|v\|_V} \geq k_0 \, \|q\|_{Q/\operatorname{Ker} B^t}. \tag{1.27}$$

If moreover $a(\cdot,\cdot)$ is invertible on $\operatorname{Ker} B$, that is, there exists $\alpha_0 > 0$, such that

$$\begin{cases} \displaystyle\inf_{u_0 \in \operatorname{Ker} B} \sup_{v_0 \in \operatorname{Ker} B} \frac{a(u_0,v_0)}{\|u_0\|_V \, \|v_0\|_V} \geq \alpha_0, \\ \displaystyle\inf_{v_0 \in \operatorname{Ker} B} \sup_{u_0 \in \operatorname{Ker} B} \frac{a(u_0,v_0)}{\|u_0\|_V \, \|v_0\|_V} \geq \alpha_0, \end{cases} \tag{1.28}$$

then there exists a solution (u,p) to problem (1.5) for any $f \in V'$ and for any $g \in \operatorname{Im} B$. The first component u is unique and p is defined up to an element of $\operatorname{Ker} B^t$. Moreover one has the bounds

$$\|u\|_V \leq \frac{1}{\alpha_0}\|f\|_{V'} + \left(\frac{\|a\|}{\alpha_0} + 1\right)\frac{1}{k_0}\|g\|_{Q'}, \tag{1.29}$$

$$\|p\|_{Q/\operatorname{Ker} B^t} \leq \frac{1}{k_0}\left(1 + \frac{\|a\|}{\alpha_0}\right)\|f\|_{V'} + \frac{\|a\|}{k_0^2}\left(1 + \frac{\|a\|}{\alpha_0}\right)\|g\|_{Q'}. \tag{1.30}$$ □

Remark 1.7: Conditions (1.27)-(1.28) of the theorem above are not only sufficient but necessary for the existence of a solution to the problem considered *for all* $g \in \operatorname{Im} B$ and *for all* $f \in V'$ (BREZZI [A]). □

To fix ideas, we shall apply the results just obtained to some of the examples introduced in Chapter I.

Example 1.1: *The Stokes problem.*

Let us go back to Example 3.1 of Chapter I. We look for $u \in (H_0^1(\Omega))^2 = V$, $p \in L^2(\Omega) = Q$, solutions of

$$\begin{cases} 2\mu \displaystyle\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx - \int_\Omega p \operatorname{div} \underline{v}\, dx = \int_\Omega \underline{v} \cdot \underline{f}\, dx, & \forall v \in V, \\ -\displaystyle\int_\Omega q \operatorname{div} \underline{u}\, dx = 0, & \forall q \in Q. \end{cases} \tag{1.31}$$

Here we have $g = 0$. Moreover the bilinear form $a(u,v) = 2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx$ is coercive on V (DUVAUT-LIONS [A], TEMAM [A]). The existence of u in (2.27) of chapter I is therefore established.

On the other hand, we have

$$b(\underline{v}, q) = -\int_\Omega q \operatorname{div} \underline{v}\, dx$$

and B is the divergence operator from $(H_0^1(\Omega))^2$ into $L^2(\Omega)$. One can show (LADYZHENSKAYA [A], TEMAM [A]) that

$$\operatorname{Im} B = \{q | \int_\Omega q\, dx = 0\},$$

and this subspace of $L^2(\Omega)$ is evidently closed (being of codimension one). We then have the existence of a pressure p defined up to an additive constant, and we can write

$$\operatorname{Ker} B^t = \operatorname{Ker}(-\operatorname{grad}) = \{q | \, q \text{ is constant on } \Omega\}. \square$$

Example 1.2: *Mixed formulation of the Dirichlet problem*

We consider here the case of Example 3.5 of Chapter I. We look for $\underline{p} \in H(\operatorname{div}; \Omega) = V$, $u \in L^2(\Omega) = Q$ such that, f being given in $L^2(\Omega)$,

$$(1.32) \qquad \begin{cases} \displaystyle\int_\Omega \underline{p} \cdot \underline{q}\, dx + \int_\Omega u \operatorname{div} \underline{q}\, dx = 0, & \forall \underline{q} \in H(\operatorname{div}; \Omega), \\ \displaystyle\int_\Omega (\operatorname{div} \underline{p} + f) v\, dx = 0, & \forall v \in L^2(\Omega). \end{cases}$$

There is a reversal of symbols with respect to the abstract result: this is probably the major difficulty of this example. Here B is the divergence operator from $H(\operatorname{div}; \Omega)$ into $L^2(\Omega)$ and it is surjective. We have

$$a(\underline{p}, \underline{q}) = \int_\Omega \underline{p} \cdot \underline{q}\, dx,$$

and this bilinear form is coercive on $\operatorname{Ker} B$, even if it is *not coercive* on $H(\operatorname{div}; \Omega)$, taking into account the definition (3.37) of chapter I of the norm in $H(\operatorname{div}; \Omega)$. $\square$

Example 1.3: *Domain decomposition for the Dirichlet problem.*

Refering to Example 4.2 of Chapter I, we have to solve the following problem: find $u \in X(\Omega) = V$, $\underline{p} \in H(\operatorname{div}; \Omega) = Q$, solutions of

$$(1.33) \qquad \begin{cases} \displaystyle\int_{K_i} \underline{\operatorname{grad}}\, u_i \cdot \underline{\operatorname{grad}}\, v_i\, dx - \int_{\partial K_i} \underline{p} \cdot \underline{n}_i v_i\, d\sigma = \int_{K_i} f v_i\, dx, \\ \qquad\qquad\qquad\qquad \forall v_i \in H^1(K_i), \forall K_i, \\ \displaystyle\sum_i \int_{\partial K_i} \underline{q} \cdot \underline{n}_i u_i\, d\sigma = 0, \qquad \forall \underline{q} \in H(\operatorname{div}; \Omega) \end{cases}$$

We thus have $b(v, \underline{q}) = \sum_i \int_{\partial K_i} \underline{q} \cdot \underline{n}_i \, v_i \, d\sigma$, and the operator B associates to $v \in X(\Omega)$ its jumps $v_i - v_j$ on the interfaces $e_{ij} = \partial K_i \cap \partial K_j$. The kernel of B is nothing but $H_0^1(\Omega)$ and the problem corresponding to (1.9) and (1.10) is the standard Dirichlet problem. To prove the existence of p we shall have to prove that $\operatorname{Im} B$ is closed in $(H(\operatorname{div};\Omega))'$ and we shall have to characterize $\operatorname{Ker} B^t$. This will be done in Chapter IV. □

We shall of course come back to these problems when studying more precisely mixed and hybrid methods. Checking the closedness of $\operatorname{Im} B$, even if existence proofs can be obtained through other considerations, is a crucial step ensuring that we have a well-posed problem and that we are working with the right functional spaces. This last fact is essential to obtain "natural" error estimates.

II.1.2 Extensions of existence and uniqueness results

Some applications, in particular nearly incompressible material (Section VI.6) will require a more general formulation than problem (1.5). Although the first generalization introduced will appear to be simple, we shall see that its analysis is rather more intricate.

Let us then introduce a new bilinear form $c(\cdot,\cdot)$ on $Q \times Q$ on which we suppose *continuity* and *positivity*,

$$|c(p,q)| \leq \|c\| \, \|p\|_Q \, \|q\|_Q, \qquad \forall p, q \in Q, \tag{1.34}$$

$$c(q,q) \geq 0, \qquad \forall q \in Q, \tag{1.35}$$

and let us denote by $C : Q \to Q'$ the operator associated with $c(\cdot,\cdot)$.

We now consider the following extension of problem (1.5): find $u \in V$ and $p \in Q$ such that

$$\begin{cases} a(u,v) + b(v,p) = \langle f, v\rangle_{V' \times V'}, & \forall v \in V, \\ b(u,q) - c(p,q) = \langle g, q\rangle_{Q' \times Q}, & \forall q \in Q. \end{cases} \tag{1.36}$$

Whenever $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are symmetric, this problem corresponds to the saddle point problem

$$\inf_{v \in V} \sup_{q \in Q} \tfrac{1}{2} a(v,v) + b(v,q) - \tfrac{1}{2} c(q,q) - \langle f, v\rangle + \langle g, q\rangle$$

and it is no longer equivalent to a minimization problem on u.

We now want to look for conditions on a, b, and c ensuring the existence and uniqueness of a solution to (1.36). We also want to find bounds on u and p and we would like these bounds to be independent of the properties of $c(\cdot,\cdot)$.

Let us first consider a special case. We assume that $c(\cdot,\cdot)$ is coercive on Q, that is,

$$\exists\gamma > 0 \text{ such that } c(q,q) \geq \gamma\,\|q\|_Q^2, \qquad \forall q \in Q \tag{1.37}$$

and that $a(\cdot,\cdot)$ is also coercive on V:

$$\exists\alpha > 0 \text{ such that } a(v,v) \geq \alpha\,\|v\|_V^2, \qquad \forall v \in V. \tag{1.38}$$

Then we have the following proposition.

Proposition 1.4: If (1.37) and (1.38) hold, then for every $f \in V'$ and $g \in Q'$ problem (1.36) has a unique solution (u,p). Moreover we have:

$$\frac{\alpha}{2}\|u\|_V^2 + \frac{\gamma}{2}\|p\|_Q^2 \leq \frac{1}{2\alpha}\|f\|_{V'}^2 + \frac{1}{2\gamma}\|g\|_{Q'}^2 \tag{1.39}$$

The proof is elementary. □

The estimate (1.39) is unsatisfactory. Actually, in many applications, we will deal with a bilinear form $c(\cdot,\cdot)$ defined by

$$c(p,q) = \lambda((p,q))_Q, \qquad \lambda \geq 0, \tag{1.40}$$

and we would like to get estimates that provide uniform bounds on the solution for λ small (say $0 \leq \lambda \leq 1$). Clearly if $c(\cdot,\cdot)$ has the form (1.40), one has $\gamma = \lambda$ in (1.37) and the bound (1.39) explodes for vanishing λ. This fact has practical implications, as we shall see, on the numerical approximations of some problems, for instance nearly incompressible materials. On the other hand Proposition 1.4 makes no assumptions on $b(\cdot,\cdot)$ [except the usual (1.3)] and it is then quite natural for the choice $c \equiv 0$ to be forbidden. However in Section II.1.2 we were able to get proper bounds for $c \equiv 0$ by using (1.27). Hence we now try to reduce the assumptions on $c \equiv 0$ and to add the assumption (1.27) on $b(\cdot,\cdot)$.

We therefore assume that $a(\cdot,\cdot)$ satisfies (1.1), (1.28), and

$$a(v,v) \geq 0, \qquad \forall v \in V. \tag{1.41}$$

We also assume that $b(\cdot,\cdot)$ satisfies (1.3) and (1.27) and $c(\cdot,\cdot)$ satisfies (1.34) and (1.35). For the sake of simplicity we also assume that c is symmetric, that is, $c(p,q) = c(q,p)$ for all p and q in Q. It is easy to check that (1.35), and the symmetry imply

$$(c(p,q))^2 \leq (c(p,p))(c(q,q)) \tag{1.42}$$

[consider the polynomial $P(t) = c(p+tq, p+tq)$ and remark that $P(t) \geq 0$, $\forall t \in \mathbb{R}$. This implies (1.42).]

Now consider, for every $\varepsilon > 0$, the "regularized" problem: find $u_\varepsilon \in V$ and $p_\varepsilon \in Q$ such that

$$\varepsilon((u_\varepsilon, v))_V + a(u_\varepsilon, v) + b(v, p_\varepsilon) = \langle f, v\rangle, \qquad \forall v \in V, \tag{1.43}$$

$$b(u_\varepsilon, q) - \varepsilon((p_\varepsilon, q))_Q - c(p_\varepsilon, q) = \langle g, q\rangle, \qquad \forall q \in Q. \tag{1.44}$$

Proposition 1.4 ensures existence and uniqueness of the solution of (1.43) and (1.44). If we can bound u_ε and p_ε independently of ε, a simple limiting argument will allow us to conclude. As a first step we take $v = u_\varepsilon$ in (1.43) and $q = p_\varepsilon$ in (1.44) and we subtract the two equations to get

$$\varepsilon\|u_\varepsilon\|_V^2 + \varepsilon\,\|p_\varepsilon\|_Q^2 + a(u_\varepsilon, u_\varepsilon) + c(p_\varepsilon, p_\varepsilon) = \langle f, u_\varepsilon\rangle - \langle g, p_\varepsilon\rangle. \tag{1.45}$$

From now on, we drop, for the sake of brevity, the subscript ε and we write

$$u = u_0 + \bar{u}, \qquad p = p_0 + \bar{p} \tag{1.46}$$

with $u_0 \in \operatorname{Ker} B$, $\bar{u} \in (\operatorname{Ker} B)^\perp$, $p_0 \in \operatorname{Ker} B^t$, $\bar{p} \in (\operatorname{Ker} B^t)^\perp$. The first step is to bound $\bar{u}$ and $\bar{p}$ by means of (1.27). We have

$$\begin{aligned} k_0\|\bar{p}\|_Q &\le \sup_{v\in V}\left\{\frac{a(u,v) + \varepsilon((u,v))_V - \langle f, v\rangle}{\|v\|_V}\right\} \\ &\le \{(\|a\| + \varepsilon)\|u\|_V + \|f\|_{V'}\}, \end{aligned} \tag{1.47}$$

and, using Proposition 1.2 (in particular (1.15)),

$$\begin{aligned} k_0\|\bar{u}\|_V &\le \sup_{q\in Q}\left\{\frac{c(p,q) + \varepsilon((p,q))_Q - \langle g, q\rangle}{\|q\|_Q}\right\} \\ &\le \left\{(\|c\|c(p,p))^{\frac{1}{2}} + \|g\|_{Q'} + \varepsilon\|p\|_Q\right\}, \end{aligned} \tag{1.48}$$

where (1.42) and (1.34) were also used. Next we bound u_0 in terms of $\bar{u}$, using (1.28):

$$\begin{aligned} \alpha_0\|u_0\|_V &\le \sup_{v_0\in \operatorname{Ker} B}\frac{a(u_0, v_0)}{\|v_0\|_V} = \sup_{v_0\in \operatorname{Ker} B}\left\{\frac{a(u, v_0) - a(\bar{u}, v_0)}{\|v_0\|_V}\right\} \\ &\le \|a\|\,\|\bar{u}\|_V + \|f\|_{V'} + \varepsilon\|\bar{u}\|_V + \varepsilon\|u_0\|_V. \end{aligned} \tag{1.49}$$

Collecting (1.48) and (1.49) we have

$$\|u\|_V \le \frac{1}{k_0}\{(\|c\|\,c(p,p))^{1/2} + \varepsilon\|p\|_Q + \|g\|_{Q'}\}\left(1 + \frac{\|a\| + \varepsilon}{\alpha_0 - \varepsilon}\right) + \frac{\|f\|_{V'}}{\alpha_0 - \varepsilon}. \tag{1.50}$$

On the other hand, if $g \in \operatorname{Im} B$ we have from (1.45)

$$\varepsilon\|p\|_Q^2 + a(u,u) + c(p,p) \le \|f\|_{V'}\,\|u\|_V + \|g\|_{Q'}\,\|\bar{p}\|_Q \tag{1.51}$$

and using (1.47)

$$(1.52)\quad \varepsilon\|p\|_Q^2 + a(u,u) + c(p,p) \le \left(\|f\|_{V'} + \|g\|_{Q'}\frac{(\|a\|+\varepsilon)}{k_0}\right)\|u\|_V + \frac{1}{k_0}\|f\|_{V'}\,\|g\|_{Q'}.$$

From (1.52) and (1.50) we have

$$(1.53)\qquad \varepsilon\|p\|_Q^2 + a(u,u) + c(p,p) \le K((\|c\|\ c(p,p))^{1/2} + 1 + \varepsilon\|p\|_Q),$$

where K is easily bounded by $\|f\|$, $\|g\|$, $\|a\|$, $1/k_0$, $1/\alpha_0$.

From (1.53) we deduce that $c(p,p)$ and $\sqrt{\varepsilon}\|p\|$ are uniformly bounded. Then (1.48) gives the bound on $\bar{u}$, (1.49) gives the bound on u_0 and (1.47) gives the bound on $\bar{p}$. We still have to bound p_0. Equation (1.44) implies

$$(1.54)\qquad \varepsilon((p_0,q))_Q + c(p_0,q) = -c(\bar{p},q), \qquad \forall q \in \operatorname{Ker} B^t.$$

In many particular cases, (1.54) provides a bound for p_0 in a very natural way. Let us consider the following general assumption.

(1.55) There exists a $\gamma_0 > 0$ such that for every $\bar{p} \in (\operatorname{Ker} B^t)^\perp$ and for every $\varepsilon > 0$ the solution $p_0 \in \operatorname{Ker} B^t$ of equation (1.54) is bounded by $\gamma_0\,\|p_0\|_Q \le \|\bar{p}\|_Q$.

If assumption (1.55) is satisfied, we obtain that both $\|u\|$ and $\|p\|$ and bounded uniformly in ε by a constant depending on $\|f\|$, $\|g\|$, $\|a\|$, $\|c\|$, $1/k_0$, $1/\alpha_0$, $\frac{1}{\gamma_0}$. This will imply existence and a priori bounds for the solution of (1.36). Before a deeper analysis of (1.55) let us state the result which has been obtained.

Theorem 1.2. Assume that $a(\cdot,\cdot)$, $b(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are continuous bilinear forms on $V \times V$ on $V \times Q$, and on $Q \times Q$ respectively. Assume further that $a(\cdot,\cdot)$ is positive semidefinite [i.e. (1.41)] and that $c(\cdot,\cdot)$ is positive semidefinite and symmetric. Finally assume that (1.27), (1.28), and (1.55) are satisfied. Then for every $f \in V'$ and every $g \in \operatorname{Im} B$ problem (1.36) has a solution (u,p), which is unique in $V \times Q/M$, where

$$(1.56)\qquad M = \operatorname{Ker} B^t \cap \operatorname{Ker} C.$$

Moreover we have the bound

$$(1.57)\qquad \|u\|_V + \|p\|_{Q/\operatorname{Ker} B^t} \le K(\|f\|_{V'} + \|g\|_{Q'})$$

with K a nonlinear function of $\|a\|$, $\|c\|$, $1/\alpha_0$, $1/k_0$, $1/\gamma_0$ which is bounded on bounded subsets. □

Remark 1.8: Theorem 1.2 assumes that $g \in \operatorname{Im} B$. Actually a careful look at its proof shows that the crucial steps (1.51)–(1.53) can be performed under the more general assumptions,

$$(1.58) \qquad \begin{cases} g = \bar{g} + g_0, \\ \bar{g} \in \operatorname{Im} B, \\ \exists \sigma > 0 \text{ such that for all } q \in Q,\ |\langle g_0, q\rangle| \le \sigma\,(c(q,q)^{1/2}). \end{cases}$$

This in particular will hold if g has the form

$$(1.59) \qquad \langle g, q\rangle = \langle \bar{g}, q\rangle + c(q_0, q)$$

with $\bar{g} \in \operatorname{Im} B$ and $q_0 \in Q$, provided $c(\cdot,\cdot)$ satisfies the assumptions of Theorem 1.2 and in particular (1.42). In this case, we clearly have $\sigma = (c(q_0, q_0)^{1/2})$.

If g has the form (1.59), equation (1.54) also keeps the same form, and condition (1.55) must now be written in the slightly stronger form

(1.60) There exists a $\gamma_0 > 0$ such that for every $\bar{p} \in Q$ and for every $\varepsilon > 0$ the solution $p_0 \in \operatorname{Ker} B^t$ of equation (1.54) is bounded by $\gamma_0 \|p_0\|_Q \le \|\bar{p}\|_Q$. □

Let us now discuss condition (1.55) in more detail, trying to find particular instances where it stands.

Case 1: $c(p,q) = \lambda((p,q))_Q$, $(\lambda \ge 0)$.

Then (1.54) reduces to

$$(1.61) \qquad (\varepsilon + \lambda)((p_0, q)) = 0, \qquad \forall q \in \operatorname{Ker} B^t;$$

hence $p_0 = 0$ and (1.55) is trivially satisfied. The result extends if g has the form (1.59), but p_0 is not bounded independently of λ for a general g. □

Case 2: $c(p,q) = \lambda\, \bar{c}(p,q)$, where $\bar{c}(\cdot,\cdot)$ is a bilinear form on $Q \times Q$ satisfying

$$(1.62) \qquad \begin{cases} \bar{c}(p,q) \le \|\bar{c}\|\, \|p\|_Q\, \|q\|_Q, & \forall p, q \in Q \\ \bar{c}(q,q) \ge \bar{\gamma} \|q\|_Q^2, & \forall q \in Q. \end{cases}$$

Then (1.55) is satisfied with $\gamma_0 = \bar{\gamma}/\|\bar{c}\|$ and hence independent of λ for $g \in \operatorname{Im} B$ or g satisfying (1.59). □

Case 3: $\operatorname{Ker} B^t = \{0\}$.

Then (1.55) is trivial. □

Case 4: $c(\cdot,\cdot)$ is coercive on $\operatorname{Ker} B^t$.

$$c(q_0, q_0) \geq \gamma_1 \, ||q_0||_Q^2, \qquad \forall q_0 \in \operatorname{Ker} B^t. \tag{1.63}$$

Then (1.55) holds with $\gamma_0 = \gamma_1/||c||$. In this case the assumption $g \in \operatorname{Im} B$ becomes unnecessary. □

An assumption weaker then (1.63), in the style of (1.28), still ensuring c to be invertible in $\operatorname{Ker} B^t$ would still be enough to get (1.55) and to avoid $g \in \operatorname{Im} B$.

Remark 1.9: One often has (like in Cases 1 and 2 above), $\operatorname{Ker} C = \{0\}$, so that the space M appearing in (1.56) actually reduces to $\{0\}$ and the solution (u,p) is unique. □

Remark 1.10: A closer look to the proof of Theorem 1.2 shows that, if one assumes $a(\cdot,\cdot)$ to be coercive in V, that is, (1.38), then $||u||_V$ is bounded directly by $(a(u,u))^{1/2}$ and (1.48) is useless. Hence the symmetry of $c(\cdot,\cdot)$ becomes unnecessary if (1.38) holds. □

Remark 1.11: The case $c = \lambda I$ was considered by ARNOLD [A]. □

Remark 1.12: Assuming that $\operatorname{Ker} B^t = \{0\}$ and (1.28), a continuity argument shows that (1.36) has a unique solution whenever $c(\cdot,\cdot)$ is small enough, without any further assumption. □

Remark 1.13: Another case strongly related to Theorem 1.2 will occur in applications. (cf. Section VII.3). Let us consider a bilinear form $c_\lambda(\cdot,\cdot)$ defined on a Hilbert space $W \hookrightarrow Q$ and satisfying

$$\begin{aligned} c_\lambda(p,q) &\leq c_0 \, \lambda \, ||p||_W \, ||q||_W, \qquad \forall p,q \in W \\ c_\lambda(p,p) &\geq \lambda \, \gamma \, ||p||_W^2, \qquad \forall q \in W. \end{aligned}$$

We now consider a problem of the form

$$\begin{cases} a(u,v) + b(v,p) = \langle f, v\rangle_{V'\times V}, & \forall v \in V, \\ b(u,q) - c_\lambda(p,q) = \langle g_1, q\rangle_{Q'\times Q} + \langle g_2, q\rangle_{W'\times W}, & \forall q \in W. \end{cases} \tag{1.64}$$

We now suppose, for simplicity, $a(\cdot,\cdot)$ to be coercive on V and $b(\cdot,\cdot)$ continuous on $V \times Q$ (hence on $V \times W$) with the range of B closed in Q'. We suppose in (1.64) that $g_1 \in \operatorname{Im} B$. It is then obvious that we get, instead of (1.39),

$$(1.65)\quad \alpha\|u\|_V^2 + \lambda\|p\|_W^2 \le \|f\|_{V'}\|u\|_V + \|g_1\|_{Q'}\|p\|_{Q/\operatorname{Ker} B^t} + \|g_2\|_{W'}\|p\|_W,$$

while one still has $k_0\|p\|_{Q/\operatorname{Ker} B^t} \le \|a\|\,\|u\|_V + \|f\|_{V'}$. Regrouping the terms one gets the estimate

$$(1.66)\quad \|u\|_V^2 + \|p\|_{Q/\operatorname{Ker} B^t}^2 + \lambda\|p\|_W^2 \le c\,(\|f\|_{V'}^2 + \|g_1\|_{Q'}^2 + \frac{1}{2\lambda}\|g_2\|_{W'}^2).$$

If we have $g_2 = g_2(\lambda)$ with $(\|g_2(\lambda)\|_{W'}^2/\lambda)$ bounded as $\lambda \to 0$, the solution will become unbounded in W but will remain bounded in $Q/\operatorname{Ker} B^t$ and will converge to the solution of problem (1.5). We shall come back to this convergence property in Section II.3. □

Finally, we must state here another type of generalization considered in NICOLAIDES [A] and BERNARDI–CANUTO–MADAY [A]. They consider a problem of type (1.36) but implying two bilinear forms $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ on $V \times Q$, that is,

$$(1.67)\quad \begin{cases} a(u,v) + b_1(v,p) = \langle f, v\rangle_{V'\times V}, & \forall v \in V,\\ b_2(u,q) - c(p,q) = \langle g, q\rangle_{Q'\times Q}, & \forall q \in Q. \end{cases}$$

In fact, the above references only deal with the case $C \equiv 0$. Conditions for existence of a solution are now that both $b_1(\cdot,\cdot)$ and $b_2(\cdot,\cdot)$ should satisfy an inf–sup condition of type (1.27) and that $a(u,v)$ should satisfy an invertibility condition from $\operatorname{Ker} B_2$ on $(\operatorname{Ker} B_1)'$, that is,

$$(1.68)\quad \inf_{u_0\in\operatorname{Ker} B_1}\ \sup_{v_0\in\operatorname{Ker} B_2} \frac{a(u_0,v_0)}{\|u_0\|_V\|v_0\|_V} \ge \alpha_0,$$

$$(1.69)\quad \inf_{v_0\in\operatorname{Ker} B_1}\ \sup_{u_0\in\operatorname{Ker} B_2} \frac{a(u_0,v_0)}{\|u_0\|_V\|v_0\|_V} \ge \alpha_0.$$

This condition is in general rather hard to check. We refer to BERNARDI–CANUTO–MADAY [A] for details and an application to a problem arising from spectral methods.

II.2 Approximation of the Problem

II.2.1 Basic results

We now turn to the approximation of problem (1.5). To make this presentation clearer, we shall first place ourselves in the simplest possible framework. Extensions of the theory to more complex cases will be introduced later. We again follow BREZZI [A] and FORTIN [C] while giving a more general presentation. We suppose known the standard approximation results such as can be found in CIARLET [A] or BABUŠKA [B].

Let then $V_h \hookrightarrow V$ and $Q_h \hookrightarrow Q$ be finite-dimensional subspaces of V and Q, respectively. The index h will eventually refer to a mesh from which these approximations are derived. We thus look for a couple $\{u_h, p_h\}$ in $V_h \times Q_h$ solution of

$$(2.1) \qquad \begin{cases} a(u_h, v_h) + b(v_h, p_h) = \langle f, v_h\rangle_{V'\times V}, & \forall v_h \in V_h, \\ b(u_h, q_h) = \langle g, q_h\rangle_{Q'\times Q}, & \forall q_h \in Q_h. \end{cases}$$

The problems we have to solve here concern the existence and uniqueness of $\{u_h, q_h\}$ and the estimation of $\|u - u_h\|_V$ and $\|p - p_h\|_Q$.

Remark 2.1: We can introduce, as in the continuous problem, operators A_h from V_h to V_h' and B_h from V_h to Q_h'. We identify Q_h' to a subspace of Q' extending bilinear forms on Q_h to bilinear forms on Q by the canonical "extension by zero" on $Q_h^\perp$, the orthogonal complement of Q_h in Q. We therefore set for $g_h' \in Q_h'$,

$$\langle g_h', q\rangle_{Q'\times Q} = \langle g_h', P_{Q_h} q\rangle.$$

It is also natural to define for $g \in Q'$ its projection onto $Q_h' \subset Q'$ by

$$\langle P_{Q_h'} g, q\rangle_{Q'\times Q} = \langle g, P_{Q_h} q\rangle_{Q'\times Q} = \langle P_{Q_h}^t g, q\rangle_{Q'\times Q}.$$

This being done, the operator B_h can then be interpreted as an operator from V_h into Q' which will not be, in general, the restriction to V_h of the operator B. In fact we shall have

$$(2.2) \qquad \begin{aligned} \langle B_h v_h, q\rangle_{Q'\times Q} = \langle B_h v_h, P_{Q_h} q\rangle_{Q_h'\times Q_h} &= b(v_h, P_{Q_h} q) \\ &= \langle B v_h, P_{Q_h} q\rangle_{Q'\times Q}. \end{aligned}$$

In other words, as $P_{Q_h'}$ can be seen as $(P_{Q_h})^t$ we can write

$$(2.3) \qquad B_h v_h = (P_{Q_h})^t\, B v_h = P_{Q_h'}\, B v_h, \qquad \forall v_h \in V_h.$$

B_h will therefore coincide with B only if $BV_h \subset Q_h'$. We shall meet cases where this inclusion holds but they are far from being the rule. □

Coming back to problem (2.1), we first introduce some notation. For $g \in Q'$, we set

$$Z_h(g) = \{v_h \in V_h \mid b(v_h, q_h) = \langle g, q_h \rangle,\ \forall q_h \in Q_h\}. \tag{2.4}$$

When $g = 0$, we have evidently,

$$Z_h(0) = \operatorname{Ker} B_h. \tag{2.5}$$

We shall also need constantly

$$\operatorname{Ker} B_h^t = \{q_h \in Q_h \mid b(v_h, q_h) = 0,\ \forall v_h \in V_h\}. \tag{2.6}$$

It is clear that a necessary condition in order to have existence **of a solution** of (2.1) is

$$Z_h(g) \neq \emptyset. \tag{2.7}$$

Some of the results of Section II.1 will apply directly to the present case, possibly with some simplifications due to the fact that we are dealing with finite-dimensional spaces. Note that, in particular, $\operatorname{Im} B_h$ will always be closed. Moreover, if we have the existence of a positive constant α_h^1 such that

$$\inf_{u_h \in \operatorname{Ker} B_h} \sup_{v_h \in \operatorname{Ker} B_h} \frac{a(u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \geq \alpha_h^1, \tag{2.8}$$

this will imply the existence of a positive α_h^2 such that

$$\inf_{v_h \in \operatorname{Ker} B_h} \sup_{u_h \in \operatorname{Ker} B_h} \frac{a(u_h, v_h)}{\|u_h\|_V \|v_h\|_V} \geq \alpha_h^2, \tag{2.9}$$

for in the finite-dimensional case, surjectivity and injectivity are equivalent. Hence we can collect the results of Theorem 1.1, applied to problem (2.1).

Proposition 2.1: Assume that (2.7) and (2.8) are satisfied. Then (2.1) has at least one solution (u_h, p_h). Moreover u_h is uniquely determined in V_h and p_h is uniquely determined in $Q_h / \operatorname{Ker} B_h^t$. □

Remark 2.2: In most applications, condition (2.8) will be a consequence of the ellipticity of $a(u, v)$ on $\operatorname{Ker} B_h$, that is,

$$\begin{aligned} &\text{There exists } \alpha_h^1 > 0 \text{ such that} \\ &a(v_h, v_h) \geq \alpha_h^1 \|v_h\|_V^2,\ \forall v_h \in \operatorname{Ker} B_h. \end{aligned} \tag{2.10}$$

Note however that (2.10) is not, in general, a consequence of (1.8) since the inclusion $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ is, in general, not true. □

Although Proposition 2.1 looks simple, its simplicity is hiding fundamental difficulties. Indeed, a problem may arise when we shall try to get error estimates in the Section II.3: p is defined up to an element of $\operatorname{Ker} B^t$ whereas p_h is defined up to an element of $\operatorname{Ker} B_h^t$. In practice, cases will be met where $\operatorname{Ker} B_h^t$ is larger than $\operatorname{Ker} B^t$ (In particular when B is surjective while B_h is not). The next result shows that this question is also related to condition (2.7).

Proposition 2.2: *The following statements are equivalent:*

— *For any* $g \in Im\, B$, $Z_h(g) \neq \emptyset$.

— *For any* $u \in V$, *there exists* $u_h = \Pi_h u \in V_h$, *such that* $b(u - \Pi_h u, q_h) = 0$, $\forall q_h \in Q_h$.

— $Ker\, B_h^t = Ker\, B^t \cap Q_h \subset Ker\, B^t$.

The first two statements are evidently synonymous. To check equivalence of the last one, take $q_{0h} \in \mathrm{Ker}\, B_h^t$. Then $b(v, q_{0h}) = 0$ for any $v \in V$ as v can be replaced by v_h by the second statement. The reciprocal is equally obvious. □

We evidently have by the same demonstration,

Proposition 2.3: *The following statements are equivalent:*

— *For any* $q \in Q$, *there exists* $q_h = \Phi_h q \in Q_h$, *such that* $b(v_h, q - \Phi_h q) = 0$, $\forall v_h \in V_h$.

— $Ker\, B_h = Ker\, B \cap V_h \hookrightarrow Ker\, B$ □

Proposition 2.2 can be summarized by the fact that the following diagram commutes.

$$\begin{array}{ccc} V & \xrightarrow{B} & Q' \\ {\scriptstyle \Pi_h}\downarrow & & \downarrow{\scriptstyle P_{Q'_h}} \\ V_h & \xrightarrow[B_h]{} & Q'_h \end{array}$$

Indeed the second statement can be written

$$B_h \Pi_h u = P_{Q'_h} B u.$$

Proposition 2.3 could also be summarized by a commuting diagram, that is, $B_h^t \Phi_h q = P_{V'_h} B^t q$.

Specially interesting is the case when B_h is the restriction of B to V_h; we then have

$$\langle v_h, B^t q\rangle = \langle B v_h, q\rangle = \langle B v_h, P_{Q_h} q\rangle = \langle v_h, B_h^t P_{Q_h} q\rangle$$

for $B v_h \in Q_h$ and Φ_h can be taken to be the projection from Q into Q_h.

Remark 2.3: A closer look to Proposition 2.2 may be worthwhile before going further. It tells us that when $\operatorname{Ker} B_h^t$ is larger than $\operatorname{Ker} B^t$, some discrete problems might not be well posed $[Z_h(g) = \emptyset]$ even if the continuous counterpart is $(g \in \operatorname{Im} B)$. In such cases, *additional compatibility conditions* have to be imposed and parasitic components from $\operatorname{Ker} B_h^t$ will pollute p_h. The inclusion $\operatorname{Ker} B_h^t \subset \operatorname{Ker} B^t$ is therefore "quasi-essential" even if approximations not satisfying it are still currently used. On the other hand, inclusion $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ will be exceptional and approximations where it holds will possess special properties. □

Remark 2.4: In practice, a very important case will be $\operatorname{Ker} B_h^t = \operatorname{Ker} B^t$. This will, in particular, be true if B and B_h are both surjective. □

II.2.2 Error estimates for the basic problem

We can now come to the essential part of this abstract theory: comparing the discrete solution $\{u_h, p_h\}$ of problem (2.1) to the exact solution of problem (1.5). Our first result will be:

Proposition 2.4: Let (u,p) be solution of problem (1.5). Assume that (2.7)-(2.8) are satisfied and let (u_h, p_h) be solution of problem (2.1). We then have

$$\|u - u_h\|_V \le \left(1 + \frac{\|a\|}{\alpha_h^1}\right) \inf_{w_h \in Z_h(g)} \|u - w_h\|_V + \frac{\|b\|}{\alpha_h^1} \inf_{q_h \in Q_h} \|q_h - p\|_Q. \tag{2.11}$$

Moreover if $\operatorname{Ker} B_h \subset \operatorname{Ker} B$, this can be reduced to

$$\|u - u_h\|_V \le \left(1 + \frac{\|a\|}{\alpha_h^1}\right) \inf_{w_h \in Z_h(g)} \|u - w_h\|_V \tag{2.12}$$

Proof: Let w_h be any element of $Z_h(g)$. Since $w_h - u_h \in \operatorname{Ker} B_h$, we have

$$\begin{aligned} \alpha_h^1 \|w_h - u_h\|_V &\le \sup_{v_h \in \operatorname{Ker} B_h} \frac{a(w_h - u_h, v_h)}{\|v_h\|_V} \\ &= \sup_{v_h \in \operatorname{Ker} B_h} \frac{a(w_h - u, v_h) + a(u - u_h, v_h)}{\|v_h\|_V} \\ &= \sup_{v_h \in \operatorname{Ker} B_h} \frac{a(w_h - u, v_h) - b(v_h, p - p_h)}{\|v_h\|_V}. \end{aligned} \tag{2.13}$$

If $\operatorname{Ker} B_h \subset \operatorname{Ker} B$, condition $v_h \in \operatorname{Ker} B_h$ implies $v_h \in \operatorname{Ker} B$ and (2.13) gives

$$\alpha_h^1 \|w_h - u_h\|_V \le \sup_{v_h \in \operatorname{Ker} B_h} \frac{a(w_h - u, v_h)}{\|v_h\|_V} \le \|a\| \, \|w_h - u\|_V, \tag{2.14}$$

and (2.12) follows from the triangle inequality. In the general case($\mathbf{Ker}\, B_h \not\subset \mathrm{Ker}\, B$), we still have as $v_h \in \mathrm{Ker}\, B_h$, for any $q_h \in Q_h$,

$$|b(v_h, p - p_h)| = |b(v_h, p - q_h)| \leq \|b\| \, \|v_h\|_V \|p - q_h\|_Q,$$

and (2.13) becomes

$$\text{(2.15)} \qquad \begin{cases} \alpha_h^1 \|u_h - w_h\|_V \leq \displaystyle\sup_{v_h \in \mathrm{Ker}\, B_h} \frac{a(w_h - u, v_h) - b(v_h, p - q_h)}{\|v_h\|} \\ \qquad\qquad \leq \|a\| \, \|u - w_h\|_V + \|b\| \, \|p - q_h\|_Q, \end{cases}$$

and (2.11) follows again using the triangle inequality. □

The result just proved is still very incomplete. Besides the fact that we still have to estimate $\|p - p_h\|_Q$, we also have to study the quantity

$$\text{(2.16)} \qquad \inf_{w_h \in Z_h(g)} \|u - w_h\|_V,$$

and eventually to relate it to the more standard quantity $\inf_{v_h \in V_h} \|v_h - u\|_V$, for which we can, using finite elements, get a mesh dependent bound.

Remark 2.5: We shall however meet cases where it will be possible, and even simpler to use directly a bound of (2.16). □

Proposition 2.5: Under the same assumptions as in Proposition 2.4, let k_h be the constant (in general dependent on h) such that

$$\text{(2.17)} \qquad \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq k_h \inf_{q_{0h} \in \mathrm{Ker}\, B_h^t} \|q_h + q_{0h}\|_Q = k_h \|q_h\|_{Q/\mathrm{Ker}\, B_h^t},$$

for every $q_h \in Q_h$. Then we have

$$\text{(2.18)} \qquad \inf_{w_h \in Z_h(g)} \|w_h - u\|_V \leq \left(1 + \frac{\|b\|}{k_h}\right) \inf_{v_h \in V_h} \|v_h - u\|_V.$$

Proof: Let v_h be any element of V_h. We look for $r_h \in V$ such that

$$\text{(2.19)} \qquad b(r_h, q_h) = b(u - v_h, q_h), \qquad \forall q_h \in Q_h.$$

As $Bu = g$, assumption (2.7) ensures that (2.19) has at least one solution. From Proposition 1.2 and (1.15), we can in fact find a solution satisfying

$$\text{(2.20)} \qquad \|r_h\|_V \leq \frac{1}{k_h} \sup_{q_h \in Q_h} \frac{b(u - v_h, q_h)}{\|q_h\|_Q} \leq \frac{1}{k_h} \|b\| \, \|u - v_h\|_V.$$

From (2.19), we also know that $w_h = r_h + v_h \in Z_h(g)$. Thus writing

$$\|u - w_h\|_V = \|u - v_h - r_h\|_V \leq \|u - v_h\|_V + \|r_h\|_V \leq \left(1 + \frac{\|b\|}{k_h}\right) \|u - v_h\|_V,$$

we get directly (2.18). □

Remark 2.6: Note that we had, as an intermediate result [using (2.20)],

$$\|u - w_h\|_V \le \|u - v_h\|_V + \|r_h\|_V \le \|u - v_h\|_V + \frac{1}{k_h} \sup_{q_h \in Q_h} \frac{b(u - v_h, q_h)}{\|q_h\|_Q},$$

which easily implies

$$\text{(2.21)} \quad \inf_{w_h \in Z_h(g)} \|u - w_h\|_V \le \inf_{v_h \in V_h} \Big(\|u - v_h\|_V + \frac{1}{k_h} \sup_{q_h \in Q_h} \frac{b(u - v_h, q_h)}{\|q_h\|_Q} \Big).$$

In some cases, (2.21) provides a sharper estimate than (2.18). □

We now marry Propositions 2.4 and 2.5 to get the following classical results.

Proposition 2.6: Assume that problem (1.5) has a solution (u, p). Assume that (2.7) and (2.8) are satisfied and let (u_h, p_h) be a solution of (2.1). Assume moreover that there exists two positive constants $\alpha_1 > 0$ and $k_0 > 0$ such that $\alpha_h^1 \ge \alpha_1$ in (2.8) and $k_h \ge k_0$ in (2.17). Then there exist two constants c_1 and c_2 independent of h such that

$$\text{(2.22)} \qquad \|u - u_h\|_V \le c_1 \inf_{v_h \in V_h} \|u - v_h\|_V + c_2 \inf_{q_h \in Q_h} \|p - q_h\|_Q.$$

If moreover $\operatorname{Ker} B_h \subset \operatorname{Ker} B$, we have

$$\text{(2.23)} \qquad \|u - u_h\|_V \le c_1 \inf_{v_h \in V_h} \|u - v_h\|_V.$$

The constants c_1 and c_2 satisfy

$$c_1 \le \Big(1 + \frac{\|a\|}{\alpha_1}\Big)\Big(1 + \frac{\|b\|}{k}\Big), \quad c_2 \le \frac{\|b\|}{\alpha_1}. \ \square$$

Remark 2.7: The condition $k_h \ge k_0 > 0$, will be known in the following as the **inf–sup condition**. It is often referred to in the literature as the **Babuška–Brezzi** condition. It is a sufficient but not a necessary condition for estimates (2.22) and (2.23). For additional comments on this point, see Remark 2.11 below. □

Having now obtained error bounds for u_h, there remains to study the error on the Lagrange multiplier p_h. Here again, the properties of B_h and of its lifting are an essential part of the discussion.

Proposition 2.7: Under the same hypotheses as in Proposition 2.4 and 2.5, we have the error estimate

$$\text{(2.24)} \quad \|p - p_h\|_{Q/\operatorname{Ker} B_h^t} \le \Big(1 + \frac{\|b\|}{k_h}\Big) \inf_{q_h \in Q_h} \|q_h - p\|_Q + \frac{\|a\|}{k_h} \|u - u_h\|_V.$$

Proof: Let us subtract the first equation of (2.1) from the first equation of (1.5). We get

$$a(u - u_h, v_h) + b(v_h, p - p_h) = 0, \qquad \forall v_h \in V_h,$$

so that for $q_h \in Q_h$ there comes

$$b(v_h, q_h - p_h) = -a(u - u_h, v_h) - b(v_h, p - q_h).$$

Using this and (2.17) we have,

$$\begin{aligned} \|q_h - p_h\|_{Q/\operatorname{Ker} B_h^t} &\leq \frac{1}{k_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h - p_h)}{\|v_h\|_V} \\ &= \frac{1}{k_h} \sup_{v_h \in V_h} \frac{b(v_h, p - q_h) + a(u - u_h, v_h)}{\|u_h\|_V}. \end{aligned} \tag{2.25}$$

One obtains therefore

$$\|q_h - p_h\|_{Q/\operatorname{Ker} B_h^t} \leq \frac{1}{k_h}(\|b\| \, \|p - q_h\|_Q + \|a\| \, \|u - u_h\|_V),$$

which implies (2.24) by the triangle inequality. □

Remark 2.8: Comparing estimates (2.24) and (2.22) one sees that the estimate for $\|p - p_h\|_Q$ depends on $1/k_h^2$ whereas the estimate for $\|u - u_h\|_V$ depends only on $1/k_h$. It can thus be expected, and this is verified in practice, that a small value of k_h will have more dramatic effects on p_h than on u_h. □

Remark 2.9: The inf–sup condition is again crucial to estimate $\|p - p_h\|_Q$. It must also be remarked that the error on p_h is estimated up to an element of $\operatorname{Ker} B_h^t$. If the kernel of B_h^t is larger than the kernel of B^t, this means that some components of p are not well approximated. However this possible imprecision on p_h might have no effect on the error $\|u_h - u\|_V$. □

We shall come back later to variants and extensions of the above results. Before doing so, we shall pay some special attention to the fundamental hypotheses of Proposition 2.6.

II.2.3 The inf–sup condition: criteria

We have met in Section II.1 the hypothesis,

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq k_0 \inf_{q_{0h} \in \operatorname{Ker} B_h^t} \|q_h + q_{0h}\|_Q, \tag{2.26}$$

which is nothing but (2.17) when $k_h \geq k_0 > 0$. Refering to Proposition 1.2, this hypothesis means that B_h has an *uniformly continuous lifting* (with respect to h). This is classically known as the inf–sup condition.

Going back to the estimates (2.22) or (2.24) we see that the case when $k_h \to 0$ will mean a loss of precision and even a lack of convergence. Checking the inf–sup condition is thus a very important point in the study of a saddle point problem. Condition (2.26) is rather abstract and is hard to check as such. We now give *criteria* that can be used in many important cases.

Proposition 2.8: Assume that we are given spaces $W \hookrightarrow V$ and S such that $S \cap Q_h \subset Q$. Let $|\cdot|$ be a seminorm on S and $\|\cdot\|_W$ be a norm on W. Suppose that

$$\sup_{w \in W} \frac{b(w, s_h)}{\|w\|_W} \geq \beta_W \, |s_h|_S, \qquad \forall s_h \in S \cap Q_h \tag{2.27}$$

and assume that there exists a family of uniformly continuous operators Π_h from W into V satisfying

$$\begin{cases} b(\Pi_h w - w, s_h) = 0, & \forall s_h \in S \cap Q_h, \\ \|\Pi_h w\|_V \leq c\|w\|_W. \end{cases} \tag{2.28}$$

with c independent of h. Then we have

$$\sup_{v_h \in V_h} \frac{b(v_h, s_h)}{\|v_h\|_V} \geq k_0 \, |s_h|_S, \qquad \forall s_h \in S \cap Q_h \tag{2.29}$$

with $k_0 = \beta_W / c$.

Proof: Indeed we have

$$\begin{aligned} \sup_{v_h \in V_h} \frac{b(v_h, s_h)}{\|v_h\|_V} &\geq \sup_{w \in W} \frac{b(\Pi_h w, s_h)}{\|\Pi_h w\|_V} = \sup_{w \in W} \frac{b(w, s_h)}{\|\Pi_h w\|_V} \\ &\geq \sup_{w \in W} \frac{1}{c} \frac{b(w, s_h)}{\|w\|_W} \geq \frac{\beta_W}{c} |s_h|_S. \ \square \end{aligned}$$

Remark 2.10: In most applications, we shall take $W = V$, $S = Q$ and $|\cdot|_S = \|\cdot\|_{Q/\operatorname{Ker} B^t}$. In this case, the first condition of (2.28) indeed implies from Proposition 2.2 that $\operatorname{Ker} B_h^t \subset \operatorname{Ker} B^t$ and we can summarize Proposition 2.8 by saying that if the continuous inf–sup condition (1.27) holds, and if we have (2.28), then the discrete inf–sup condition holds. $\square$

In some cases, it will be convenient to choose W to be a strict subspace of V. This will, for instance, be the case when V is not smooth enough to allow a simple construction of the operator Π_h. Obviously, we shall then have to check the inf–sup condition (2.27) on W, usually with $S = Q$ and $|\cdot|_S = \|\cdot\|_{Q/\operatorname{Ker} B^t}$.

The more general statement of Proposition 2.8 will also be useful for some special cases where $\operatorname{Ker} B_h^t$ is larger than $\operatorname{Ker} B^t$ and where we would like to use $|\cdot|_S = \|\cdot\|_{Q/\operatorname{Ker} B_h^t}$. In those cases, (2.28) will hold only for an ad hoc choice of W and the main trouble will be to obtain (2.27) for this W.

Finally, there will still be other cases in which a special choice of S is needed. We shall meet, for example, cases where $\operatorname{Ker} B_h^t = \operatorname{Ker} B^t = \{0\}$, where V is smooth enough to allow the construction of Π_h but where the continuous inf–sup condition holds only if one takes a space S which is larger than Q so that $|\cdot|_S \leq \|\cdot\|_Q$.

Remark 2.11: Assume that for every $w \in W \subset V$ one has $Z_h(Bw) \neq \emptyset$, so that the discrete problem (1.5) with $g = Bw$ and $f = Aw$ has a solution (u_h, p_h). If one requires that $\|u_h - w\|_V \leq c\|w\|_W$ (which is somehow a weaker condition than requiring the convergence of u_h to w), then one can set $u_h = \Pi_h w$ and (2.28) is satisfied, (hence also the inf–sup condition). This shows that the existence of an operator Π_h which satisfies (2.28) (hence the validity of the inf–sup condition) is in a sense *necessary* if we want a *reasonable behavior* of the discrete problem. However, the explicit construction of Π_h will be easy in some cases but very difficult in others. □

There will be cases in which Π_h will be constructed in two steps. Namely, we have the following proposition.

Proposition 2.9: Let $W \hookrightarrow V$ be a subspace of V for which (2.27) holds. Let $\Pi_1 \in \mathcal{L}(W, V_h)$ and $\Pi_2 \in \mathcal{L}(V, V_h)$ be such that

$$\text{(2.30)} \qquad \begin{cases} \|\Pi_1 w\|_V \leq c_1 \|w\|_W, \\ b(\Pi_2 v - v,\ q_h) = 0,\ \forall q_h \in Q_h, \\ \|\Pi_2(I - \Pi_1)w\|_V \leq c_2 \|w\|_W, \end{cases}$$

then (2.28) holds, hence the inf–sup condition follows.

Proof: We set $\Pi_h w = \Pi_2(w - \Pi_1 w) + \Pi_1 w$. It is easy to check that (2.28) holds. Indeed,

$$\begin{aligned} b(\Pi_h w, q_h) &= b(\Pi_2(w - \Pi_1 w), q_h) + b(\Pi_1 w, q_h) \\ &= b(w - \Pi_1 w, q_h) + b(\Pi_1 w, q_h) \\ &= b(w, q_h) \end{aligned}$$

and

$$\|\Pi_h w\|_V \leq \|\Pi_2(w - \Pi_1 w)\|_V + \|\Pi_1 w\|_V \leq (c_2 + c_1)\|w\|_W. \ \square$$

In applications, Π_1 will be a kind of "best approximation" operator. To fix ideas, it will verify an estimate of type $\|\Pi_1 w - w\|_V \leq ch^s \|w\|_W$. On the other hand, Π_2 will be a local adjustment (typically by bubble functions) in order to satisfy the first condition of (2.28).

A last remark about the operator Π_h. It is sometimes possible to build it so that an error bound is directly available on $\|\Pi_h u - u\|_V$, independently of the inf–sup condition. One then has

Proposition 2.10: Let u be solution of problem (1.5). If one can build $\Pi_h u \in Z_h(g)$, that is satisfying $b(u - \Pi_h u, q_h) = 0$, $\forall q_h \in Q_h$, then one has

$$\text{(2.31)} \qquad \begin{cases} \|u - u_h\|_V \leq c_1 \|u - \Pi_h u\|_V + c_2 \inf\limits_{q_h \in Q_h} \|p - q_h\|_Q, \\ c_1 = (1 + \|a\|/\alpha_h^1)\ , \qquad c_2 = \|b\|/\alpha_h^1. \end{cases}$$

Proof: This is obvious from Proposition 2.4. □

As the above results will play an essential role throughout this book, we shall now summarize its most usual form in the following.

Theorem 2.1: Let $(u, p) \in V \times Q$ and $(u_h, p_h) \in V_h \times Q_h$ be respectively solutions of problems,

$$\text{(2.32)} \qquad \begin{cases} a(u, v) + b(v, p) = \langle f, v \rangle, & \forall v \in V, \\ b(u, q) = \langle g, q \rangle, & \forall q \in Q, \end{cases}$$

and

$$\text{(2.33)} \qquad \begin{cases} a(u_h, v_h) + b(v_h, p_h) = \langle f, v_h \rangle, & \forall v_h \in V_h, \\ b(u_h, q_h) = \langle g, q_h \rangle, & \forall q_h \in Q_h. \end{cases}$$

Assume that the inf–sup condition

$$\text{(2.34)} \qquad \inf_{q_h \in Q_h} \sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_{Q/\operatorname{Ker} B^t}} \geq k_0 > 0$$

is satisfied and let $a(\cdot, \cdot)$ be uniformly coercive on $\operatorname{Ker} B_h$, that is, there exists $\alpha_0 > 0$ such that

$$\text{(2.35)} \qquad a(v_{0h}, v_{0h}) \geq \alpha_0 \|v_{0h}\|^2, \qquad \forall v_{0h} \in \operatorname{Ker} B_h$$

Then one has the following estimate, with a constant c depending on $\|a\|$, $\|b\|$, k_0, α_0 but independent of h:

$$\text{(2.36)} \quad \|u - u_h\|_V + \|p - p_h\|_{Q/\operatorname{Ker} B^t} \leq c\Big(\inf_{v_h \in V} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q\Big). \ \square$$

The reader may refer to the previous propositions to get more detailed forms of this result.

Remark 2.12: As we have seen, condition (2.35) can be replaced by the weaker condition (2.8). □

II.2.4 Extensions of error estimates

We have introduced in Section II.1.2 an extended problem in which a third bilinear form $c(\cdot,\cdot)$ appeared. We shall consider now the question of error estimation for this problem. We shall use a discrete analogue of Theorem 1.2 and in particular of its variant provided by Remark 1.8. We first introduce the discrete problem

$$(2.37)\qquad \begin{cases} \text{find } u_h \in V_h \text{ and } p_h \in Q_h \text{ such that} & \\ a(u_h, v_h) + b(v_h, q_h) = \langle f, v_h\rangle, & \forall v_h \in V_h, \\ b(u_h, q_h) - c(q_h, q_h) = \langle g, q_h\rangle, & \forall q_h \in Q_h, \end{cases}$$

where, as usual, $a(\cdot,\cdot)$, $b(\cdot,\cdot)$, and $c(\cdot,\cdot)$ are bilinear continuous forms on $V \times V$, on $V \times Q$, and $Q \times Q$ respectively, and $V_h \subset V$ and $Q_h \subset Q$ are finite-dimensional subspaces.

Proposition 2.11: Assume that $a(\cdot,\cdot)$ and $c(\cdot,\cdot)$ are positive semidefinite [that is (1.35) and (1.41), respectively]. Assume moreover that $c(\cdot,\cdot)$ is symmetric, that $a(\cdot,\cdot)$ satisfies (1.24) and (2.8), and that $\operatorname{Im} B$ is closed in Q' whereas $b(\cdot,\cdot)$ satisfies (2.34). Assume finally that condition (1.60) holds. Then for every $f \in V'$ and $g \in \operatorname{Im} B$ problems (1.36) and (2.37) have a unique solution. Moreover we have

$$(2.38)\qquad \|u - u_h\|_V + \|p - p_h\|_Q$$
$$\le K\Big(\|a\|, \|b\|, \|c\|, \frac{1}{k_0}, \frac{1}{\alpha_h}, \frac{1}{\gamma_0}\Big)\Big(\inf_{v_h \in V} \|u - v_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q\Big)$$

with K bounded on bounded subsets.

Proof. Existence and uniqueness for both problems follow from Theorem 1.2. In order to get the estimate (2.38) we note that from (1.36) and (2.37) we have, for all $\tilde{u}_h \in V_h$ and $\tilde{p}_h \in Q_h$,

$$(2.39)\qquad \begin{cases} a(\tilde{u}_h - u_h, v_h) + b(v_h, \tilde{p}_h - p_h) = a(\tilde{u}_h - u, v_h) + b(v_h, \tilde{p}_h - p), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall v_h \in V_h, \\ b(\tilde{u}_h - u_h, q_h) - c(\tilde{p}_h - p_h, q_h) = b(\tilde{u}_h - u, q_h) - c(\tilde{p}_h - p, q_h), \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall q_h \in Q_h. \end{cases}$$

Hence $(\tilde{u}_h - u_h, \tilde{p}_h - p_h)$ is the solution of a problem of type (2.37) with right-hand side $F \in V_h'$ and $G \in Q_h'$ defined by

$$(2.40)\qquad F : v_h \to a(\tilde{u}_h - u, v_h) + b(v_h, \tilde{p}_h - p_h),$$

$$G = \bar{G} - G_0, \tag{2.41}$$

$$\bar{G} : q_h \to b(\tilde{u}_h - u, q_h), \tag{2.42}$$

$$G_0 : q_h \to c(\tilde{p}_h - p, q_h) =: c(q_0, q_h). \tag{2.43}$$

It is clear that

$$\begin{aligned} &\|F\|_{V'} + \|\bar{G}\|_{Q'} + (c(q_0, q_0))^{1/2} \\ &\qquad \le (\|a\| + 2\|b\| + \|c\|)(\|\tilde{u}_h - u\|_V + \|\tilde{p}_h - p\|_Q). \end{aligned} \tag{2.44}$$

Applying Theorem 1.2 (or, rather, Remark 1.8) to (2.39) and using (2.44), we obtain

$$\begin{aligned} &\|\tilde{u}_h - u_h\|_V + \|\tilde{p}_h - p_h\|_Q \\ &\qquad \le K(\|a\|, \|b\|, \|c\|, \frac{1}{k_0}, \frac{1}{\alpha_h}, \frac{1}{\gamma_0})\,(\|\tilde{u}_h - u\|_V + \|\tilde{p}_h - p\|_Q). \end{aligned} \tag{2.45}$$

Since $\tilde{u}_h$ and $\tilde{p}_h$ are arbitrarily chosen in V_h and q_h, we obtain (2.38) from (2.45) and the triangle inequality. □

Remark 2.13: The above proof applies in particular to the case $c(\cdot,\cdot) = 0$. It is more general than Theorem 2.1 by allowing *coerciveness* to be replaced by (2.8). In practice such a condition may well be rather hard to check. □

We can also consider the case of Remark 1.13 in which $c(\cdot,\cdot)$ depends on a parameter λ and there exists constants c_0 and γ such that,

$$c_\lambda(p,q) \le c_0\lambda\, \|p\|_W\, \|q\|_W, \qquad \forall q \in W, \tag{2.46}$$

$$c_\lambda(q,q) \ge \gamma\lambda\, \|q\|_W^2, \qquad \forall q \in W, \tag{2.47}$$

with $W \hookrightarrow Q$. We have the following proposition:

Proposition 2.12: Let $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be as in Proposition 2.11 and assume that $a(\cdot,\cdot)$ is coercive on V and $c(\cdot,\cdot)$ satisfies (2.46), (2.47). Then for every $f \in V'$ and $g \in \operatorname{Im} B$ we have

$$\begin{aligned} &\|u_\lambda - u_h\|_V^2 + \|p_\lambda - p_h\|^2_{Q/\operatorname{Ker} B'} + \lambda\|p_\lambda - p_h\|_W^2 \\ &\qquad \le c\Big(\inf_{v_h \in V_h} \|u_\lambda - v_h\|_V^2 + \inf_{q_h \in Q_h} \{\|p_\lambda - q_h\|_Q^2 + \lambda\|p_\lambda - q_h\|_W^2\}\Big). \end{aligned} \tag{2.48}$$

The proof follows the lines of Proposition 2.11, using (1.66) with $\langle g_1, q_h\rangle = b(\tilde{u}_h - u_\lambda, q_h)$ and $\langle g_2, q_h\rangle = -c_\lambda\,(\tilde{p}_h - p_\lambda, q_h)$ (instead of using Remark 1.8). Estimate (2.48) indeed breaks down in W for λ small but still yields an optimal estimate in Q. □

II.2.5 Various generalizations of error estimates

We have considered up to now the most basic form of mixed problems. Numerous variations are however possible. Some of them are too special to merit an abstract treatment and will be presented on specific examples in subsequent chapters. We consider here some problems arising in quite a large number of practical situations.

The first pathology that we consider is the case where coerciveness on $\operatorname{Ker} B_h$ does not hold but can be replaced by a weaker condition. Let us suppose $V \hookrightarrow H$ where H is also a Hilbert space. We suppose that the bilinear form $a(\cdot,\cdot)$ satisfies the hypotheses

$$a(v,v) \geq \alpha \|v\|_H^2, \tag{2.49}$$

$$|a(u,v)| \leq \|a\| \, \|u\|_H \, \|v\|_H. \tag{2.50}$$

This situation arises in two kinds of examples.

— $\|\cdot\|_H$ is a norm on $\operatorname{Ker} B$ so that (2.49) implies coerciveness on $\operatorname{Ker} B$. It may happen however that for a discretization of the problem one does not have $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ and that the discrete problem is not coercive on $\operatorname{Ker} B_h$. Condition (2.49) nevertheless ensures existence of the discrete solution by the equivalence of norms in a finite-dimensional space (see below). Convergence properties are however likely to be altered. In the mixed formulation of elasticity introduced in Chapter I, we have $V = (H(\underline{\mathrm{div}};\Omega))_s^2$ whereas the bilinear form $a(\boldsymbol{u},v) = \int_\Omega \underline{\underline{\sigma}} : \underline{\underline{\tau}} \, dx$ is coercive only on $(L^2(\Omega))_s^4 = H$. This is enough to have coerciveness on $\operatorname{Ker} B$ but not in general on $\operatorname{Ker} B_h$ unless one is clever and builds V_h and Q_h in order to have $\operatorname{Ker} B_h \subset \operatorname{Ker} B$. In general the analysis of this problem is difficult as we shall see in Chapter VII.

— One considers an ill-posed problem in the sense that the existence of (u,p) cannot be obtained directly in $V \times Q$ but only for instance through a regularity argument. Existence of a discrete solution however holds and one would like to get error estimates. Such is the case in the $\psi - \omega$ mixed formulation of the biharmonic problem that we have seen in (3.54) of chapter I. For a more detailed analysis of this case, see chapter IV.

On the finite-dimensional space V_h, $\|\cdot\|_V$ and $\|\cdot\|_H$ are equivalent and we thus have

$$\|v_h\|_V \leq S(h) \, \|v_h\|_H. \tag{2.51}$$

In practice, $S(h)$ will be given, for finite element approximations, by the inverse inequality (CIARLET [B]).

We now consider an error estimate for the simple case.

Proposition 2.13: Let (u,p) be solution of problem (1.5) and (u_h, p_h) be solution of problem (2.1). Under hypotheses (2.49)–(2.50)–(2.51), we have the estimate

$$(2.52)\quad \|u-u_h\|_H \le \left(1+\frac{\|a\|}{\alpha}\right) \inf_{v_h\in Z_h(g)} \|v_h-u\|_H + \frac{\|b\|S(h)}{\alpha} \inf_{q_h\in Q_h} \|q_h-p\|_Q$$

The proof is the same as for Proposition 2.4, introducing the bound

$$b(u_h-v_h,\ q_h-p) \le \|b\|\ S(h)\ \|u_h-v_h\|_H\ \|q_h-p\|_Q.$$

The rest of the analysis can be continued from this point but with a loss of accuracy coming from the $S(h)$ factor. □

Remark 2.14: In fact, the general bound in this case would be

$$(2.53)\quad \|u-u_h\|_H \le \left(1+\frac{\|a\|}{\alpha}\right) \inf_{v_h\in Z_h(g)} \|v_h-u\|_H + \inf_{q_h} \sup_{v_h\in \operatorname{Ker} B_h} \frac{b(v_h,p-q_h)}{\alpha\|v_h\|_H}$$

for which (2.52) is a brute force bound. It is however reasonable to expect in some cases the term $b(v_h,p-q_h)$ to have for $v_h \in \operatorname{Ker} B_h$ some *superconvergence* property either in general or for special types of approximation (KIKUCHI–ANDO [A], SCAPOLLA [A]). If $\operatorname{Ker} B_h \subset \operatorname{Ker} B$, this term actually vanishes. It must also be noted that we do not have in general a bound of $\inf_{v_h\in Z_h(g)} \|v_h-u\|_H$ by $\inf_{v_h\in V_h} \|v_h-u\|_H$. □

Another variant that will be useful in the study of some hybrid methods is the following.

Let $|v|_V$ be a continuous *seminorm* on V and let M denote its kernel. Then $|\cdot|_V$ is a norm on the quotient space V/M. We suppose that we have

$$(2.54)\qquad a(v,v) \ge \alpha\ |v|_V^2, \qquad \forall v\in V,\ \alpha \text{ independent of } h,$$

and

$$(2.55)\qquad |a(u,v)| \le \|a\|\ |u|_V\ |v|_V.$$

Let us suppose $M \subset V_h$ and let us suppose that for $p\in Q$, one can build $q_h \in Q_h$ such that $b(v_h,p-q_h)=0$, $\forall v_h \in M$. We then have the bound

$$(2.56)\qquad |b(v_h,p-q_h)| \le \|b\|\ |v_h|_V\ \|p-q_h\|_Q,$$

and the following proposition holds.

Proposition 2.14: Let (u,p) be solution of problem (1.5) and $(\boldsymbol{u}_h, \boldsymbol{p}_h)$ be solution of problem (2.1). Let

$$\bar{Q}_h(p) = \{q_h \mid b(v, p - q_h) = 0.\ \forall v \in M\}, \tag{2.57}$$

If (2.54) and (2.55) are satisfied, then we have the estimate

$$|u - u_h|_V \leq \left[1 + \frac{\|a\|}{\alpha}\right] \inf_{v_h \in Z_h(g)} |u - v_h|_V + \|b\| \inf_{q_h \in \bar{Q}_h(p)} \|p - q_h\|_Q. \ \square \tag{2.58}$$

II.2.6 Perturbations of the problem, nonconforming methods

We shall now rapidly consider the effect on error estimates of changing problem (2.1) into a perturbed problem of the form

$$\begin{cases} a_h(u_h, v_h) + b_h(v_h, p_h) = \langle f, v_h \rangle_h, & \forall v_h \in V_h, \\ b_h(u_h, q_h) = \langle g, q_h \rangle_h, & \forall q_h \in Q_h, \end{cases} \tag{2.59}$$

where $a_h(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ are, in a sense to be made precise, approximations of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$, and where $\langle \cdot,\cdot \rangle_h$ denotes an approximation of the duality brackets $\langle \cdot,\cdot \rangle_{V' \times V}$ or $\langle \cdot,\cdot \rangle_{Q' \times Q}$

The reasons underlying such a study are twofold:

— Formulation of type (2.59) arise when *nonconforming* approximations are introduced. In this case we no longer have $V_h \subset V$ and $Q_h \subset Q$ so that the problem must be imbedded in larger spaces. We shall give an alternative treatment of nonconforming methods using domain decomposition methods in Chapter IV. However, their importance is worth their presence in our abstract discussion.

— Using numerical quadrature formulas also leads to problems of type (2.59). Numerical integration is a standard part of the finite element method and it is important to be acquainted with its consequences.

Remark 2.15: The concept of numerical integration is taken here in a very general sense and is not restricted to reduced integration methods (cf. Section VI.7). We also mean by this any procedure where $b(v_h, q_h)$ is replaced by a (weaker) expression of the form

$$b_h(v_h, q_h) = \langle IBv_h, q_h \rangle_{Q'_h \times Q_h}, \tag{2.60}$$

where I is a continuous operator from $B(V_h)$ into Q'_h. In practice this could mean as in Chapter VI an interpolation operator. The usual mixed formulation takes I to be the projection operator on Q'_h and we have

$$(IBv_h, q_h)_{Q_h} = \langle Bv_h, q_h \rangle.$$

In general we can consider I to project Bv_h with respect to a duality product $\langle \cdot,\cdot \rangle_h$ defined by $\langle IBv_h, q_h \rangle_{Q'_h \times Q_h} = \langle Bv_h, q_h) \rangle_h$. $\square$

To include the nonconforming case in our setting we suppose that there exist spaces X and Y such that V_h and V are closed subspaces of X. In the same way Q_h and Q should be closed subspaces of Y. We suppose that $a_h(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ satisfy

$$|a_h(u_h, v_h)| \leq c\, ||u_h||_X\, ||v_h||_X \tag{2.61}$$

$$|b_h(v_h, q_h)| \leq c\, ||v_h||_X\, ||q_h||_Y \tag{2.62}$$

We suppose that $a_h(\cdot,\cdot)$ is coercive on $\operatorname{Ker} B_h$ (where $\operatorname{Ker} B_h = \{v_h \in V_h \mid b_h(v_h, q_h) = 0, : \forall q_h \in Q_H)$, that is,

$$a_h(v_{0h}, v_{0h}) \geq \alpha\, ||v_{0h}||_X^2,\ \forall v_{0h} \in \operatorname{Ker} B_h. \tag{2.63}$$

We suppose that b_h satisfies, with k_0 independent of h,

$$\sup_{v_h \in V_h} \frac{b_h(v_h, q_h)}{||v_h||_X} \geq k_0\, ||q_h||_Y \tag{2.64}$$

that is, B_h is surjective (we consider the general case later). Finally we define

$$||f||_h = \sup_{v_h \in V_h} \frac{\langle f, v_h\rangle_h}{||v_h||_X}, \quad ||g||_h = \sup_{q_h \in Q_h} \frac{\langle g, q_h\rangle_h}{||q_h||_Y}. \tag{2.65}$$

We obviously have the following result.

Proposition 2.15: Under hypotheses (2.61) through (2.65), problem (2.59) has a unique solution and there exists a constant c independent of h such that

$$||u_h||_X + ||p_h||_Y \leq c\, (||f||_h + ||g||_h). \ \square \tag{2.66}$$

We now want, as in Proposition 2.12, to use this stability result to obtain an error estimate. Let then (u, p) be the solution of problem (1.5). After a few tedious manipulations, one gets from (2.59)

$$\begin{aligned} a_h(u - u_h, v_h) &+ b_h(v_h, p - p_h) \\ &= [a_h(u, u_h) + b_h(v_h, p) - \langle f, v_h\rangle] + [\langle f, v_h\rangle - \langle f, v_h\rangle_h], \end{aligned} \tag{2.67}$$

$$b_h(u - u_h, q_h) = [b_h(u, q_h) - \langle g, q_h\rangle] + [\langle g, q_h\rangle - \langle g, q_h\rangle_h], \tag{2.68}$$

provided we can give a meaning to all the terms in (2.67), (2.68). For numerical integration, this will require extra regularity whereas for nonconforming methods, $a_h(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ will be extensions of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to less regular spaces. It wil therefore be impossible to state a general precise result. However, using the same stability argument as in Proposition 2.12 we have formally

Proposition 2.16: Let (u,p) be the solution of problem (1.5) and (u_h,p_h) be the solution of problem (2.59). Assume that the hypotheses of Proposition 2.15 hold, then we have

$$(2.69)\quad \|u-u_h\|_X + \|p-p_h\|_Y \le c\left(\inf_{v_h\in V_h}\|u-v_h\|_X + \inf_{q_h\in Q_h}\|p-q_h\|_Y + M_{1h} + M_{2h} + M_{3h} + M_{4h}\right)$$

where we define the "consistency terms"

$$(2.70)\qquad M_{1h} = \sup_{v_h\in V_h}\frac{|a_h(u,v_h)+b_h(v_h,p)-\langle f,v_h\rangle|}{\|v_h\|_X}$$

$$(2.71)\qquad M_{2h} = \sup_{v_h\in V_h}\frac{|\langle f,v_h\rangle-\langle f,v_h\rangle_h|}{\|v_h\|_X},$$

$$(2.72)\qquad M_{3h} = \sup_{q_h\in Q_h}\frac{|b_h(u,q_h)-\langle g,q_h\rangle|}{\|q_h\|_Y},$$

$$(2.73)\qquad M_{4h} = \sup_{q_h\in Q_h}\frac{|\langle g,q_h\rangle-\langle g,q_h\rangle_h|}{\|q_h\|_Y}. \ \square$$

Using Proposition 2.16 in practice means giving **a sense to the extra terms** M_{1h}, M_{2h}, M_{3h}, M_{4h} and bounding them properly.

It is worth considering a few special cases. In many problems, it will be natural to use a nonconforming approximation of V but a conforming one on Q. For instance, in Stokes problem (Chapter VI) we have $Q = L^2(\Omega)$ and it is rather hard to think of a nonconforming approximation to this space. If we suppose then that $b_h(u,q_h) = b(u,q_h)$, which is usually the case when no numerical integration is used, then we have $M_{3h} = 0$.

The terms M_{2h} and M_{4h} normally come from the use of numerical quadrature formulas for the right-hand sides, and they can be handled by standard techniques (CIARLET [B]).

Finally an important case is the use of conforming approximations where $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ are computed by numerical quadrature. In this case, we have (if u is smooth enough to give a sense to $a_h(u,v_h)$)

$$(2.74)\qquad a(u,v_h)+b(v_h,p)-\langle f,v_h\rangle = 0$$

and we can transform M_{1h} to

$$(2.75)\qquad \hat{M}_{1h} = \sup_{v_h\in V_h}\frac{|a(u,v_h)-a_h(u,v_h)|}{\|v_h\|_V} + \sup_{v_h\in V_h}\frac{|b(v_h,p)-b_h(v_h,p)|}{\|v_h\|_V}$$

and M_{3h} to

$$(2.76)\qquad \hat{M}_{3h} = \sup_{q_h\in Q_h}\frac{|b(u,q_h)-b_h(u,q_h)|}{\|q_h\|_Q}. \ \square$$

We have assumed in the previous estimates that B_h was surjective. Let us now see how this condition can be checked and eventually relaxed for the case of nonconforming methods. We assume therefore that the bilinear forms $a_h(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ are now extensions of $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ to $X \times X$ and $X \times Y$ respectively. It is then natural to consider the extension $\tilde{B}$ of B with $\tilde{B} : X \to Y'$ and $\tilde{B}^t \cdot Y \to X'$. We thus have by the definitions of B and B_h

$$Bu = P_{Q'}(\tilde{B}u), \tag{2.77}$$

$$B_h u_h = P_{Q'_h}(\tilde{B}u_h). \tag{2.78}$$

Whenever $Q = Y$, this reduces to $Bu = \tilde{B}u$ and $B_h u_h = P_{Q'_h}(Bu_h)$ that is the standard conforming case.

Let us now make the assumption that $\tilde{B}$ has a closed range in Y', that is,

$$\sup_{x \in X} \frac{b_h(x,y)}{\|x\|_X} \geq k_0 \, \|y\|_{Y/\operatorname{Ker}\tilde{B}^t}. \tag{2.79}$$

In general checking (2.79) requires a good choice of X and Y. Whenever this holds we have:

Proposition 2.17: Let the bilinear form $b_h(\cdot,\cdot)$ satisfy (2.79) and let there exist a family of uniformly continuous operators $\Pi_h : X \to V_h$ such that one has

$$b_h(u, q_h) = b_h(\Pi_h u, q_h), \qquad \forall q_h \in Q_h, \tag{2.80}$$

$$\|\Pi_h u\|_X \leq c \, \|u\|_X, \tag{2.81}$$

with constant c independant of h. If, moreover, $\operatorname{Im} B_h \hookrightarrow \operatorname{Im} \tilde{B}$, **then** $\operatorname{Ker} B_h^t \hookrightarrow \operatorname{Ker} \tilde{B}^t$ and one has

$$\sup_{v_h \in V_h} \frac{b_h(v_h, q_h)}{\|v_h\|_X} \geq k_0' \, \|q_h\|_{Y/\operatorname{Ker}\tilde{B}^t}. \tag{2.82}$$

The proof is the same as for Proposition 2.8. □

The error estimate (2.69) can then be changed to bound $\|p - p_h\|_{Y/\operatorname{Ker}\tilde{B}^t}$ instead of $\|p - p_h\|_Y$. The case $Y = Q$ thus reduces to a standard $\|p - p_h\|_{Q/\operatorname{Ker}B^t}$ and using Proposition 2.17, requires a good choice of X. In the same way when $Y \neq Q$, it should be chosen in order to keep $\operatorname{Ker}\tilde{B}^t$ small enough, the best situation being $\operatorname{Ker}\tilde{B}^t = \operatorname{Ker}B^t$. These technical problems will of course have to be solved by different ways on each particular problem.

Let us turn now to the case of conforming approximations with numerical integration, in which $V_h \subset V$, $Q_h \subset Q$ but $b_h(\cdot,\cdot)$ is an approximation of $b(\cdot,\cdot)$. It is sometimes interesting to compare on $V_h \times Q_h$ the two operators B_h and $\hat{B}_h$ defined respectively by $b(v_h, q_h) = \langle B_h u, q_h \rangle$ and $b_h(v_h, q_h) = \langle \hat{B}_h v_h, q_h \rangle$. Knowing for instance that B_h satisfies the inf–sup condition, what can be said of $\hat{B}_h$. We have the following criterion.

Proposition 2.18: Let B_h and $\hat{B}_h$ be defined respectively by $b(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$. The following statements are then equivalent.

(2.83)
$$\begin{cases} \text{a) } \operatorname{Ker} \hat{B}_h^t \subset \operatorname{Ker} B_h^t, \\ \text{b) } \operatorname{Im} B_h \subset \operatorname{Im} \hat{B}_h, \\ \text{c) } \forall v_h,\ \exists w_h = \Pi_h v_h \text{ such that } b(v_h, q_h) = b_h(w_h, q_h),\ \ \forall q_h \in Q_h. \end{cases}$$

The proof is a simpler version of the proof of Proposition 2.2. □

The above result is symmetrical with respect to B_h and $\hat{B}_h$. It is then clear that we have the criterion given by the following

Proposition 2.19: Let B_h and $\hat{B}_h$ be defined respectively by $b(\cdot,\cdot)$ and $b_h(\cdot,\cdot)$ and suppose that there exists an invertible operator Π_h; $V_h \to V_h$ such that

$$\|\Pi_h v_h\|_V \le c_0 \|v_h\|_V,\ \|\Pi_h^{-1} v_h\|_V \le c_1 \qquad \|v_h\|_V \tag{2.84}$$

and such that

$$b(v_h, q_h) = b_h(\Pi_h v_h, q_h), \qquad \forall q_h \in Q_h \tag{2.85}$$

Then, if either of B_h or $\hat{B}_h$ satisfies the inf–sup condition, both do, and we have $\operatorname{Ker} B_h^t = \operatorname{Ker} \hat{B}_h^t$, $\operatorname{Im} B_h = \operatorname{Im} \hat{B}_h$. □

In practice this means that the numerical quadrature is not exact for the computation of $b(v_h, q_h)$ but rather integrates $b(\Pi_h v_h, q_h)$ with $\Pi_h v_h$ near enough to v_h.

It is also useful to consider the following result.

Proposition 2.20: Let us suppose that B_h^t and $\hat{B}_h$ have the same kernel, that B_h satisfies the inf–sup condition, and that there exists a constant $C(h)$, with $C(h) \to 0$ when $h \to 0$, such that

$$|b(v_h, q_h - b_h(v_h, q_h)| \le C(h)\, \|v_h\|_V\, \|q_h\|_{Q_h/\operatorname{Ker} B_h^t}. \tag{2.86}$$

Then for h small enough, $b_h(\cdot,\cdot)$ also satisfies the inf–sup condition.

Indeed one may write $b(v_h, q_h) = b_h(v_h, q_h) + (b(v_h, q_h) - b_h(v_h, q_h))$ and thus

$$\sup_{v_h \in V} \frac{b(v_h, q_h)}{\|v_h\|_V} \le \sup_{v_h \in V} \frac{b_h(v_h, q_h)}{\|v_h\|_V} + \sup_{v_h \in V} \frac{|b(v_h, q_h) - b_h(v_h, q_h)|}{\|v_h\|_V}$$

Using (2.85) and the inf–sup condition for $b(\cdot,\cdot)$ we get

$$\sup_{v_h \in V} \frac{b_h(v_h, q_h)}{\|v_h\|_V} \ge (k_0 - C(h))\|q_h\|_{Q_h/\operatorname{Ker} B_h^t}$$

that is, the desired result. □

II.2.7 Dual error estimates

We now present, to end this section on error estimates, an extension of the Aubin–Nitsche's duality technique (AUBIN [A], NITSCHE [A]) to the analysis of problem (1.5). We consider an abstract setting that will be general enough to include most cases where we will like to use such techniques for instance in Chapter VI for Stokes problem (to get $L^2(\Omega)$-estimates) or in Chapter V for Dirichlet's problem (to get H^{-1}-estimates). We refer to FALK–OSBORN [A] where similar, and in some cases more general, results are presented.

Let us then consider two spaces V_- and Q_- (the minus index intuitively meaning a "less regular" space) with the *dense* inclusions

$$V \hookrightarrow V_- \quad \text{and} \quad Q \hookrightarrow Q_-. \tag{2.87}$$

We would like to estimate $\|u - u_h\|_{V_-}$ and $\|p - p_h\|_{Q_-}$. Let us denote

$$V'_+ = (V_-)', \; Q'_+ = (Q_-)'. \tag{2.88}$$

We then have from (2.87)

$$V'_+ \hookrightarrow V', \; Q'_+ \hookrightarrow Q', \tag{2.89}$$

and we can thus make the following hypothesis.

Hypothesis H1: For any $f_+ \in V'_+$, $g_+ \in Q'_+ \cap \operatorname{Im} B$, **the solution (w, s) of the** problem

$$\begin{cases} a(v,w) + b(v,s) = \langle f_+, v\rangle, & \forall v \in V, \\ b(w,q) = \langle g_+, q\rangle, & \forall q \in Q, \end{cases} \tag{2.90}$$

belongs to $V_{++} \times Q_{++}$, where $V_{++} \hookrightarrow V$, $Q_{++} \hookrightarrow Q$ and we have the estimate

$$\|w\|_{V_{++}} + \|s\|_{Q_{++}/\operatorname{Ker} B^t} \le c\,(\|f\|_{V'_+} + \|g\|_{Q'_+}). \; \square \tag{2.91}$$

This hypothesis evidently means in practice that we have a regularity property and that $f \in V'_+$, $g \in Q'_+$ yield a more regular solution. Moreover we are implicitly assuming that $\operatorname{Ker} B^t \subset Q_{++}$. We then have

Theorem 2.2: Let hypothesis H1 hold and let (u, p) be the solution of problem (2.32) and (u_h, p_h) be the solution of problem (2.33). We then have under the hypotheses of Theorem 2.1

$$\begin{aligned} \|u-u_h\|_{V_-} + \|p-p_h\|_{Q_-/\operatorname{Ker} B^t} &\le c_1\,(\|u-u_h\|_V + \|p-p_h\|_{Q/\operatorname{Ker} B^t}) \\ &\times \Big(\inf_{w\in V_{++}} \sup_{w_h\in V_h} \frac{\|w-w_h\|_V}{\|w\|_{V_{++}}} + \inf_{s\in Q_{++}} \sup_{q_h\in Q_h} \frac{\|s-q_h\|_Q}{\|s\|_{Q_{++}}} \Big). \end{aligned} \tag{2.92}$$

Proof: Let us choose $f_+ \in V'_+$ and $g_+ \in Q'_+ \cap \operatorname{Im} B$ with $\|f_+\|_{V'_+} = 1$ and $\|g_+\|_{Q'_+} = 1$ such that one has

$$(2.93) \qquad \begin{cases} \langle f_+, u - u_h \rangle_{V'_+ \times V_-} = \|u - u_h\|_{V_-}, \\ \langle g_+, p - p_h \rangle_{Q'_+ \times Q_-} = \|p - p_h\|_{Q_- / \operatorname{Ker} B^t}, \end{cases}$$

and let $(w, s) \in V_{++} \times Q_{++}$ be the solution of (2.90) and therefore bounded by (2.91). We may thus write

$$(2.94) \qquad (\|w\|_{V_{++}} + \|s\|_{V_{++}}) \leq \hat{c}.$$

Making $v = u - u_h$ and $q = p - p_h$ in (2.90) we thus have from (2.93):

$$(2.95) \qquad \begin{aligned} &\|u - u_h\|_{V_-} + \|p - p_h\|_{Q_- / \operatorname{Ker} B^t} \\ &\qquad = a(u - u_h, w) + b(u - u_h, s) + b(w, p - p_h). \end{aligned}$$

But we know that one also has, substracting (2.32) and (2.33)

$$(2.96) \qquad \begin{cases} a(u - u_h, w_h) + b(w_h, p - p_h) = 0, & \forall w_h \in V_h, \\ b(u - u_h, q_h) = 0, & \forall q_h \in Q_h. \end{cases}$$

We may thus write in (2.95),

$$(2.97) \qquad \begin{aligned} &\|u - u_h\|_{V_-} + \|p - p_h\|_{Q_- / \operatorname{Ker} B^t} \\ &\qquad = a(u - u_h, w - w_h) + b(u - u_h, s - q_h) + b(w - w_h, p - p_h) \end{aligned}$$

and (2.92) follows. □

In practice we shall use the fact that $w \in V_{++}$ and $s \in Q_{++}$ are regular to obtain bounds

$$(2.98) \qquad \inf_{w_h \in V_h} \|w - w_h\|_V \leq m(h)\, \|w\|_{V_{++}} \leq m(h)\, \hat{c},$$

$$(2.99) \qquad \inf_{q_h \in Q_h} \|s - q_h\|_Q \leq n(h)\, \|s\|_{Q_{++}} \leq n(h)\, \hat{c},$$

so that (2.92) yields the estimate

$$(2.100) \qquad \begin{aligned} &\|u - u_h\|_{V_-} + \|p - p_h\|_{Q_- / \operatorname{Ker} B^t} \\ &\qquad \leq m(h)\, \hat{c}_1\, (\|u - u_h\|_V + \|p - p_h\|_Q) + n(h)\, \hat{c}_2\, (\|u - u_h\|_V), \end{aligned}$$

which will eventually be a better estimate if $m(h)$ and $n(h)$ are small. □

Remark 2.16: We shall also use in Chapter V a superconvergence result that can be extended to the abstract setting of Theorem 2.2.

Let us suppose that the approximations at hand satisfy the inclusion $\operatorname{Ker} B_h \hookrightarrow \operatorname{Ker} B$. From Proposition 2.3, we then know there exists $\bar{p}_h \in Q_h$ such that

$$b(v_h, \bar{p}_h - p) = 0, \qquad \forall v_h \in V_h, \tag{2.101}$$

and we now want to find an estimate on $\|\bar{p}_h - p_h\|_Q$. In order to do so, we consider (z, ϕ) the solution of the problem

$$\begin{cases} a(v, z) + b(v, \phi) = 0, & \forall v \in V, \\ b(z, q) = ((\bar{p}_h - p_h, q))_Q, & \forall q \in Q. \end{cases} \tag{2.102}$$

This is a well-posed problem in $V \times Q$. It may happen (that will be the case in the application of Chapter V) that (z, ϕ) is more regular and that (z, ϕ) belongs to $V_{++} \times Q_{++}$ for properly chosen spaces. We now show that it is then possible to estimate $\|\bar{p}_h - p_h\|_Q$. Indeed from (2.102) we have

$$\|\bar{p}_h - p_h\|_Q^2 = b(z, \bar{p}_h - p_h) = b(\Pi_h z, \bar{p}_h - p_h), \tag{2.103}$$

where $\Pi_h z$ is the special interpolate such that $b(z - \Pi_h z, q_h) = 0$ for all q_h as in (2.28). (The existence of $\Pi_h z$ is equivalent to say that the inf–sup condition holds, according to Proposition 2.8 and Remark 2.11.) Using (2.101) in (2.103) we then have

$$\begin{aligned} \|\bar{p}_h - p_h\|_Q^2 &= b(\Pi_h z, p - p_h) \\ &= a(u_h - u, \Pi_h z) \\ &= a(u_h - u, \Pi_h z - z) + a(u_h - u, z). \end{aligned} \tag{2.104}$$

Making $v = u_h - u$ in (2.102) this becomes, for all $q_h \in Q_h$,

$$\begin{aligned} \|\bar{p}_h - p_h\|_Q^2 &= a(u_h - u, \Pi_h z - z) - b(u_h - u, \phi) \\ &= a(u_h - u, \Pi_h z - z) - b(u_h - u, \phi - q_h). \end{aligned} \tag{2.105}$$

Finally from (2.105) we have, for all $q_h \in Q_h$,

$$\|\bar{p}_h - p_h\|^2 \le \|u_h - u\|_V (\|z - \Pi_h z\|_V + \|\phi - q_h\|_Q). \tag{2.106}$$

If $\Pi_h z$ approximates z with optimal order (for $z \in V_{++}$) and if we have an estimate $\|z\|_{V_{++}} + \|\phi\|_{Q_{++}} \le \hat{c}\, \|\bar{p}_h - p_h\|_Q$, then we get from (2.98), (2.99), and (2.106) the estimate:

$$\|\bar{p}_h - p_h\| \le \|u - u_h\|_V\, \hat{c}\, [m(h) + n(h)]. \tag{2.107}$$

This result uses the strong assumption $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ and its use is rather technical. Anyhow the above analysis shows when it can be expected to hold, besides the example of Chapter V. □

II.3 Numerical Properties of the Discrete Problem

This section will present a few general facts related to numerical computations with the previously described problem. As we are still in a rather abstract setting, we will not be able to obtain directly usable results. However some basic facts are common to a large number of methods and presenting them in a unified frame may help understand the relations existing between apparently different methods.

II.3.1 The matrix form of the discrete problem

We shall consider first problem (2.1) and develop a matrix form suited to numerical computation. We shall set, for the finite-dimensional spaces V_h and Q_h,

$$\begin{cases} N = \dim V_h, \\ M = \dim Q_h, \end{cases} \tag{3.1}$$

and we use a basis of theses spaces, namely, $\{v_{ih} \mid 1 \leq i \leq N\}$ for V_h and $\{q_{ih} \mid 1 \leq i \leq M\}$ for Q_h. We can now define

$$a_{ij} = a(v_{jh}, v_{ih}), \tag{3.2}$$

$$b_{ij} = b(v_{jh}, q_{ih}), \tag{3.3}$$

$$f_i = \langle f, v_{ih} \rangle, \tag{3.4}$$

$$g_i = \langle g, q_{ih} \rangle. \tag{3.5}$$

We denote $A_{N\times N} = (a_{ij})$, $B_{M\times N} = (b_{ij})$, $\mathbf{f}_N = (f_i)$, $\mathbf{g}_M = (g_i)$ and by $\mathbf{u} = \{\alpha_i\}$, $\mathbf{p} = \{\beta_i\}$ the vectors of $\mathbb{R}^N$ and $\mathbb{R}^M$ formed by the coeficients of u_h and p_h in the expressions

$$u_h = \sum_{i=1}^{N} \alpha_i \, v_{ih}, \tag{3.6}$$

$$p_h = \sum_{i=1}^{M} \beta_i \, q_{ih}, \tag{3.7}$$

Problem (2.1) can now be written in matrix form as

$$\begin{cases} A\mathbf{u} + B^t\mathbf{p} = \mathbf{f}, \\ B\mathbf{u} = \mathbf{g}, \end{cases} \tag{3.8}$$

or

$$\begin{pmatrix} A & B^t \\ B & O \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{p} \end{pmatrix} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{3.9}$$

In practice, the bases $\{v_{ih}\}$ and $\{q_{ih}\}$ will be built using a finite element technique. This will impose additional structure on problem (3.9). We can however see that for a symmetric bilinear form $a(\cdot,\cdot)$ we have to solve a symmetric but, in general, indefinite linear system. The fact that we have positive and negative eigenvalues is, of course, directly related to the fact that we discretize a saddle point problem.

It can be seen from (3.9) that the system will be singular if $\operatorname{Ker} B^t \neq \{0\}$. The right-hand side will then have to satisfy a compatibility condition: $\mathbf{g}$ must belong to $\operatorname{Im} B$. As we have already said, this can happen even if the continous problem is surjective.

Finally *one can eliminate the variable* $\mathbf{u}$ *from this linear system, at least if matrix* A *is invertible.* Indeed one gets from (3.8),

$$\mathbf{u} = A^{-1}\mathbf{f} - A^{-1}B^t\mathbf{p} \tag{3.10}$$

and thus denoting $\mathbf{u}_g$ any solution of $B\mathbf{u}_g = \mathbf{g}$

$$B\mathbf{u} = BA^{-1}\mathbf{f} - BA^{-1}B^t\mathbf{p} = \mathbf{g} = B\mathbf{u}_g. \tag{3.11}$$

We can then solve for $\mathbf{p}$ in the problem

$$BA^{-1}B^t\mathbf{p} = BA^{-1}\mathbf{f} - B\mathbf{u}_g. \tag{3.12}$$

This is a discrete form of the *dual problem* of Section I.3. Let us consider the matrix $BA^{-1}B^t$. If matrix A is positive definite, this matrix is also positive semidefinite. Indeed one has

$$\langle BA^{-1}B^t\mathbf{p}, \mathbf{p}\rangle_{\mathbb{R}^M} = \langle A^{-1}B^t\mathbf{p}, B^t\mathbf{p}\rangle_{\mathbb{R}^M} \geq \alpha \, \|B^t\,\mathbf{p}\|^2_{\mathbb{R}^M}. \tag{3.13}$$

It is positive definite if $\operatorname{Ker} B^t = \{0\}$. Problem (3.12) is therefore easier to solve than problem (3.9), as numerical methods for positive definite systems are more efficient and more stable.

Unfortunately this simplification of the problem cannot in general be done in practice. The trouble comes from A^{-1} which is likely to be a full matrix even if A is sparse. The system (3.12) is then too large to be stored and handled. *We shall however meet some cases where such a reduction of the problem can be done*, thus providing an efficient solution method.

II.3.2 Eigenvalue problem associated with the inf–sup condition

The convergence analysis of Section II.2 relied heavily on the inf–sup condition stating that the operator B_h must have an uniformly continuous lifting. In abstract form, this means checking that one has, with $k_h \geq k_0 > 0$,

$$\sup_{v_h \in V_h} \frac{b(v_h, q_h)}{\|v_h\|_V} \geq k_h \, \|q_h\|_{Q_h / \operatorname{Ker} B_h^t}, \qquad \forall q_h \in Q_h. \tag{3.14}$$

We shall now give an interpretation of constant k_h as a *generalized singular value* of the matrix representing the operator B_h in the discrete problem. This singular value can be identified with the square root of a generalized eigenvalue problem. Although the following development is very elementary, it cannot easily be found in elementary texts and we thought it was worth presenting it. We also refer to MALKUS [A] where similar discussions can be found.

We have already defined a matrix B such that if $v_h = \sum_{i=1}^N \alpha_i v_{ih}$ and $q_h = \sum_{j=1}^M \beta_i q_{ih}$, one has, denoting $\mathbf{v} = \{\alpha_i\}$, $\mathbf{q} = \{\beta_i\}$,

$$b(v_h, q_h) = \langle B\mathbf{v}, \mathbf{q}\rangle_{\mathbb{R}^M} = \langle \mathbf{v}, \ B^t\mathbf{q}\rangle_{\mathbb{R}^N}. \tag{3.15}$$

We shall also need the matrices S and T associated with the scalar product of V_h and Q_h. Let us define

$$s_{ij} = ((q_{ih}, q_{jh}))_{Q_h}, \qquad 1 \leq i, j \leq M, \tag{3.16}$$

$$t_{ij} = ((v_{ih}, v_{jh}))_{V_h}, \qquad 1 \leq i, j \leq N. \tag{3.17}$$

One then clearly has with the same notations as above

$$\|v_h\|^2_{V_h} = \langle T\mathbf{v}, \mathbf{v}\rangle_{\mathbb{R}^N} \tag{3.18}$$

and

$$\|q_h\|^2_{V_h} = \langle S\mathbf{q}, \mathbf{q}\rangle_{\mathbb{R}^M}. \tag{3.19}$$

The operators S and T can be considered as isomorphisms from the copy of $\mathbb{R}^N$ and $\mathbb{R}^M$ representing V_h and Q_h onto another copy representing the dual spaces V_h' and Q_h'. We can summarize the situation by the Diagram II.1.

$$\begin{tikzcd}
V_h \approx \mathbb{R}^N \arrow[r, "B"] & Q_h' \approx \mathbb{R}^M \arrow[d, "S"] \\
V_h' \approx \mathbb{R}^M \arrow[u, "T"] & Q_h \approx \mathbb{R}^M \arrow[l, "B^t"]
\end{tikzcd}$$

Diagram II.1

We now consider the following *generalized singular value problem.* Find a basis of T-orthonormal vectors of $\mathbb{R}^N$, $\{\mathbf{v}_i |\ 1 \le i \le N\}$, and a basis of S-orthonormal vectors of $\mathbb{R}^M$, $\{\mathbf{q}_i |\ 1 \le i \le M\}$, such that there exists $\mu_i > 0$ satisfying

$$B\mathbf{v}_i = \mu_i S\mathbf{q}_i, \qquad \forall \mathbf{v}_i \notin \operatorname{Ker} B, \tag{3.20}$$

$$B^t\mathbf{q}_i = \mu_i T\mathbf{v}_i, \qquad \forall \mathbf{q}_i \notin \operatorname{Ker} B^t, \tag{3.21}$$

The case $S = T = I$ is the standard singular value problem for matrix B. It is easily shown that such μ_i, $\mathbf{v}_i$ and $\mathbf{q}_i$ exist. Indeed one gets from (3.20) and (3.21), denoting by r the rank of B,

$$BT^{-1}B^t\mathbf{q}_i = \mu_i^2\, S\mathbf{q}_i, \qquad 1 \le i \le r, \tag{3.22}$$

$$B^tS^{-1}B\mathbf{v}_i = \mu_i^2\, T\mathbf{v}_i, \qquad 1 \le i \le r. \tag{3.23}$$

Both these problems are standard generalized eigenvalue problems. It is elementary to show that their solution yield a solution of (3.20) and (3.21). One then obtains the desired bases by completing them with vectors of the kernel subspaces. We then have

$$\begin{cases} B = SP\Sigma U^tT^t, \\ B^t = TU\Sigma P^tS^t, \end{cases} \tag{3.24}$$

where U and P are the matrices formed from the column vectors $\mathbf{v}_i$ and $\mathbf{q}_i$ and Σ is the $M \times N$, pseudo-diagonal matrix containing the μ_i on its main diagonal.

In problem (3.22), zero eigenvalues correspond to eigenvectors lying in $\operatorname{Ker} B^t$. We shall now see that the inf–sup condition is related to the behavior of the smallest nonzero eigenvalue. This eigenvalue is nothing but k_h and must remain bounded away from zero when the dimensions of the spaces increase. We shall in fact prove:

Proposition 3.1: Let k_h be defined by (3.14) and let $\mu_{\min} = \mu_r$ be the smallest nonzero singular value of B, as defined by (3.20) and (3.21). Then $k_h = \mu_{\min}$.

Proof: Let us first remark that (3.14) can be written as

$$\inf_{q_h \in (\operatorname{Ker} B_h^t)^\perp} \ \sup_{v_h \in (\operatorname{Ker} B_h)^\perp} \frac{b(v_h, q_h)}{\|v_h\|_V \|q_h\|_Q} = k_h. \tag{3.25}$$

Let us write as in (3.6) and (3.7), $v_h = \sum_i^r \alpha_i\, v_{ih}$ and $q_h = \sum_i^r \beta_i\, q_{ih}$, but taking now for v_{ih} and q_{ih} the elements of V_h and Q_h associated with the

vectors $\mathbf{v}_i$ and $\mathbf{q}_i$ solutions of the eigenvalue problems (3.22) and (3.23). Then $v_h \in (\mathbf{Ker}\, B_h^t)^\perp$ and $q_h \in (\mathrm{Ker}\, B_h)^\perp$. One has moreover

$$
\begin{aligned}
b(v_h, q_h) &= \langle B\mathbf{v}, \mathbf{q}\rangle_{\mathbb{R}^M} = \langle \sum_i \alpha_i B\mathbf{v}_i,\ \sum_j \beta_j\ \mathbf{q}_j\rangle \\
&= \sum_{k=1}^{r} \mu_k\ \alpha_k\ \beta_k
\end{aligned}
\tag{3.26}
$$

by the S orthonormality of the $\mathbf{q}_i$ and $B\mathbf{v}_i = \mu_i\ S\mathbf{q}_i$. Moreover we have

$$
\|v_h\|_{V_h}^2 = \sum_{i=1}^{r} \alpha_i^2, \tag{3.27}
$$

$$
\|q_h\|_{Q_h}^2 = \sum_{j=1}^{r} \beta_j^2. \tag{3.28}
$$

We want to evaluate

$$
\inf_{\beta_k} \sup_{\alpha_k} \frac{\sum_k \mu_k\ \alpha_k\ \beta_k}{\sqrt{\sum \alpha_k^2}\sqrt{\sum \beta_k^2}} = k_h. \tag{3.29}
$$

There is no loss in generality in taking $\|v_h\| = \|q_h\| = 1$. Then the supremum in α_k is then clearly attained for $\alpha_k = \mu_k\ \beta_k(\sum \mu_k^2\ \beta_k^2)^{-1/2}$ and its value is $\sqrt{\sum_k \mu_k^2\ \beta_k^2}$. This is clearly larger than μ_{min}. The minimum value is $\mu_{min} = \mu_r$, taking $\beta_r = 1$ and all other coefficients zero. □

Remark 3.1: We have in fact for the singular value a generalized Rayleigh's quotient with

$$
\mu_{min} = \inf_{\mathbf{v}} \sup_{\mathbf{q}} \frac{\langle B\mathbf{v}, \mathbf{q}\rangle}{\sqrt{\langle T\mathbf{v}, \mathbf{v}\rangle}\ \sqrt{\langle S\mathbf{q}, \mathbf{q}\rangle}} \tag{3.30}
$$

and other singular values corresponding to other extremal values.

In particular, we have

$$
\|b\| = \sup_{\mathbf{v}}\ \sup_{\mathbf{q}} \frac{\langle B\mathbf{v}, \mathbf{q}\rangle}{\sqrt{\langle T\mathbf{v}, \mathbf{v}\rangle}\sqrt{\langle S\mathbf{q}, \mathbf{q}\rangle}}. \ \square \tag{3.31}
$$

An interesting special case of the above result arises when one can identify the matrix T, associated with the scalar product on V_h and matrix A defined by (3.2). This is the case when $a(\cdot,\cdot)$ is continuous and coercive, thus defining on V_h a scalar product and hence a norm, equivalent to the standard norm. From

(3.22), the μ_i^2 are now the eigenvalues of the dual problem (3.12) and we have a result on the condition number of this problem (FORTIN–PIERRE [A]) which is now given by

$$\text{Cond}(BA^{-1}B^t) = \frac{\|b\|}{k_{\min}} \tag{3.32}$$

and will vary, following the dependence of $\|b\|$ or $k_{\min}$ on h.

The above analysis also allows us to give a closer look to the structure of the problem and the importance of the inf–sup condition for convergence.

II.3.3 Is the inf–sup condition so important?

One of the most frustrating things in the analysis of mixed finite element methods is often the apparent discrepancy between experience and theory. To quote FORTIN [D], "Computations were done (with success!) using theoretically dubious elements or at best, using elements on which theory remained silent." This is specially the case for Stokes problem of Chapter VI where velocity results are generally quite good, even with elements not satisfying the inf–sup condition, whereas reasonable pressure results can often be recovered after a filtering posttreatment of the raw results. The singular value decomposition introduced above allows us to get a better understanding of those disconcerting behaviors.

Let us go back to the matrix form (3.9) of our problem, which we shall now rewrite, using the fact that there exists a basis of V_h and a basis of Q_h such that matrix B takes the pseudo-diagonal form:

$$B = \begin{pmatrix} \mu_1 & \cdot & \cdot & \cdot & 0 & \cdot & 0 \\ \cdot & \mu_2 & \cdot & \cdot & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \mu_r & 0 & \cdot & 0 \\ \cdot & \cdot & 0 & \cdot & 0 & \cdot & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & 0 & \cdot & 0 \end{pmatrix}, \tag{3.33}$$

where we suppose that the singular values μ_i are written in decreasing order. The solution of our problem will depend directly on the behavior of those singular values in a way which we shall now try to describe. Let us first note that in (3.33), columns of zeros (i.e., $j > r$) , correspond to the kernel of B whereas rows of zeros correspond to the kernel of of B^t. Rows of zeros imply that it is possible to solve $B\mathbf{u} = \mathbf{g}$ only if $\mathbf{g}$ takes the form

$$\mathbf{g} = \begin{pmatrix} g_1 \\ g_2 \\ \cdot \\ g_r \\ 0 \\ \\ 0 \end{pmatrix}, \tag{3.34}$$

that is, $\mathbf{g}$ has no component in $\operatorname{Ker} B^t$. We have already discussed the importance of the dimension of $\operatorname{Ker} B^t$. If this dimension happens to be larger than the dimension of the kernel of the corresponding infinite-dimensional operator, we have spurious zero energy modes in $\mathbf{p}$ which imply artificial (nonphysical) constraints on $\mathbf{g}$.

Another important point is the dimension of $\operatorname{Ker} B$, that is, the number of zero columns. In order to get a good approximation of the infinite-dimensional kernel, this dimension should grow when the number of degrees of freedom increases. Whenever this growth is not occurring properly, we shall have a *locking phenomenon*, which may be *total*, that is,

$$B\mathbf{u} = \mathbf{0} \text{ implies } \mathbf{u} = \mathbf{0}, \tag{3.35}$$

or *partial*, $\mathbf{u}$ being restricted into too small a subspace. This will happen whenever the space Q_h is taken too large, thus overconstraining the solution. From Proposition 2.5, such a situation implies that some of the singular values μ_i will become vanishingly small when the mesh size decreases.

To complete our picture, we shall now divide the singular values of B into three sets, writing

$$B = \begin{pmatrix} \Sigma_1 & 0 & 0 \\ 0 & \Sigma_2 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{3.36}$$

where Σ_1 contains the "stable part of B" (i.e.,$\mu_i > k_0 \geq 0$), Σ_2 contains singular values vanishing when h gets small, and the zero singular values correspond to $\operatorname{Ker} B^t$. We can now write system (3.9) as

$$\begin{pmatrix} A_{11} & A_{12} & A_{13} & \Sigma_1 & 0 & 0 \\ A_{21} & A_{22} & A_{23} & 0 & \Sigma_2 & 0 \\ A_{31} & A_{32} & A_{33} & 0 & 0 & 0 \\ \Sigma_1 & 0 & 0 & 0 & 0 & 0 \\ 0 & \Sigma_2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ p_1 \\ p_2 \\ p_3 \end{pmatrix} = \begin{pmatrix} f_1 \\ f_2 \\ f_3 \\ g_1 \\ g_2 \\ g_3 \end{pmatrix}. \tag{3.37}$$

If we want to solve (3.37), we must first have $g_3 = 0$, leaving p_3, the component of $\mathbf{p}$ in $\operatorname{Ker} B^t$, indeterminate. As we have already discussed, this condition may imply artificial constraints if $\operatorname{Ker} B^t$ is too large: they could then eventually be verified by suitably modifying the data. The question is then whether this can be done without losing precision. Supposing that this point can be settled, we can now proceed in (3.37) to solve for u_1, u_2, and u_3,

$$\begin{cases} u_1 = \Sigma_1^{-1} g_1, \\ u_2 = \Sigma_2^{-1} g_2, \\ u_3 = A_{33}^{-1} - A_{31}^{-1} u_1 - A_{32}^{-1} u_2. \end{cases} \tag{3.38}$$

The potential trouble obviously lies in u_2 which depends on the inverse of the unstable part Σ_2. Again, if g_2 is null (or sufficiently small), u_2 will be null (or negligible) whereas u_1 and u_3 will behave correctly. This can happen because we can set g_2 to zero without loosing precision or because "normal data" contain only a small g_2 component corresponding, for example, to "high frequency components" which are small for regular functions. In such a case, one can expect reasonable results even if the inf–sup condition is not satisfied and B contains an unstable part Σ_2.

Finally, u_1, u_2, and u_3 being known, we get from (3.37)

$$\begin{cases} p_1 = -\Sigma_1^{-1}(A_{11}^{-1}u_1 + A_{12}^{-1}u^2 + A_{13}^{-1}u_3 - f_1), \\ p_2 = -\Sigma_2^{-1}(A_{21}^{-1}u_1 + A_{22}^{-1}u_2 + A_{23}^{-1}u_3 - f_2). \end{cases} \tag{3.39}$$

Here p_2 depends on the inverse of the unstable part Σ_2, in fact, from (3.38), on Σ_2^{-2} if g_2 is not zero. Even for $g_2 = 0$, p_2 cannot be expected to be correct but p_1 will then remain stable. If this stable part of p is rich enough to approximate the exact infinite-dimensional solution, filtering out p_2 will yield good results. This is indeed what happens in many situations. One may however think that relying on such borderline conditions is likely to lead to unreliable results at times.

The complete analysis of an approximation should therefore identify how well a "normal problem" can be approximated by the "good part" u_1, u_3, p_1 of the numerical solution. This would imply the knowledge of the singular decomposition, which is a rather strong requirement. We shall present in Chapter VI an example where this can partly be done. We also refer to Section V.6 where the role of "coerciveness on the kernel" and the inf–sup condition are discussed on a simple example.

II.4 Solution by Penalty Methods, Convergence of Regularized Problems

We now describe a procedure that is gaining more and more popularity for the numerical solution of some saddle point problems, specially the Stokes problem described in Chapter VI. The method is general and we believe it is worth presenting it in an abstract setting.

The main idea is very simple and is quite classical in the theory of mathematical programming. Let us suppose we want to solve a constrained problem

$$\inf_{v \in K} J(v), \tag{4.1}$$

where K is a closed convex subset of the Hilbert space V. Then we can approximate this problem by the unconstrained one

$$\inf_{v \in V} J(v) + \frac{1}{\varepsilon} d^2(v, K), \tag{4.2}$$

where $d(v, K)$ is the distance in V from v to K. It can be proved, under fairly general assumptions, that the solution u of (4.2) converges in V to the solution u of problem (4.1) when ε becomes small. This apparent simplification has to be paid: problem (4.2) is generally harder to solve when ε is small due to the ill conditioning of the functional

$$J_\varepsilon(v) = J(v) + \frac{1}{\varepsilon} d^2(v, K).$$

Let us apply this to problem (3.8). When A is a positive definite symmetric matrix, this system is equivalent to

$$\inf_{Bv=g} \left\{ \tfrac{1}{2}(Av, v) - (f, v) \right\}, \tag{4.3}$$

where $(\cdot, \cdot)$ denote the standard Euclidean scalar product in $\mathbb{R}^N$. Then let S be any positive definite matrix in $\mathbb{R}^N$. We can replace (4.3) by

$$\inf_v \left\{ \frac{1}{2}(Av, v) + \frac{1}{2\varepsilon}(S^{-1}(Bv - g, Bv - g) - (f, v) \right\} \tag{4.4}$$

or equivalently, writing $p_\varepsilon = (1/\varepsilon)S^{-1}(Bv - g)$

$$\begin{cases} Au_\varepsilon + B^t p_\varepsilon = f, \\ Bu_\varepsilon - g = \varepsilon S p, \end{cases} \tag{4.5}$$

which is nothing but

$$Au_\varepsilon + \frac{1}{\varepsilon} B^t S^{-1} B u_\varepsilon = f + \frac{1}{\varepsilon} B^t S^{-1} g. \tag{4.6}$$

If the matrix S is "easy to invert" (in particular if S^{-1} is a sparse matrix, by preference block diagonal), this provides a way to reduce our problem to a more standard quadratic unconstrained problem. This is a widely used technique and it is indeed quite efficient. One should however be aware that the penalty term $1/\varepsilon\ B^t S^{-1} B$ *has a strong negative impact on the condition number* of the linear system (4.6).

Using a penalty method is, for instance, almost impossible if an iterative method is used for the solution of the linear system (4.6), iterative methods being is general quite sensitive to the condition number of the matrix at hand.

We now consider the problem of estimating the effect of changing a problem with constrains to a penalized problem. We shall place ourselves in a general setting that can be applied as well to a finite-dimensional problem as to an infinite dimensional one. The result obtained will also show how the solution of a problem depending on a parameter may converge in some cases to a limit problem of mixed type. We therefore consider a problem of the form

$$\begin{cases} a(u_\varepsilon, v) + b(v, p_\varepsilon) = \langle f, v \rangle, \ \forall v \in V, \\ b(u_\varepsilon, q) - \varepsilon((p_\varepsilon, q))_Q = \langle g, q \rangle, \ \forall q \in Q. \end{cases} \tag{4.7}$$

If $S : Q \to Q'$, denotes the canonical isomorphism associated with the scalar product $((\cdot,\cdot))_Q$, this problem can be written as

$$(4.8) \qquad a(u_\varepsilon, v) + \frac{1}{\varepsilon}\langle Bv, S^{-1}Bu_\varepsilon\rangle_{Q'\times Q} = \langle f, v\rangle + \frac{1}{\varepsilon}\langle Bv, S^{-1}g\rangle_{Q'\times Q}.$$

If the space Q is identified to its dual space Q', we then have $S = I$ and (4.8) becomes

$$(4.9) \qquad a(u_\varepsilon, v) + \frac{1}{\varepsilon}((Bu_\varepsilon, Bv))_Q = \langle f, v\rangle + \frac{1}{\varepsilon}((g, Bv)).$$

A problem as (4.9) can thus arise from a penalty method applied to a constrained problem. We are then interested in knowing whether the solution of the penalized problem converges or not to the solution of the true problem. It is also possible that (4.9), usually with $g \equiv 0$, represents a physical system in which a parameter becomes small. This is the case for example in nearly incompressible materials (Section I.1 and Section VI.7). In this case we are interested in knowing the relation between the solution of (4.9) and the solution of the limit problem which has the form (1.5).

Both these questions are indeed solved by the same analysis. The assumption will be as usual coerciveness of $a(\cdot,\cdot)$, that is

$$(4.10) \qquad a(v, v) \geq \alpha \, \|v\|_V^2,$$

and the closedness of $\operatorname{Im} B$ in Q', that is,

$$(4.11) \qquad \sup_{v\in V} \frac{b(v, q)}{\|v\|_V} \geq k_0 \, \|q_0\|_{Q/\operatorname{Ker} B^t}.$$

We now prove (BERCOVIER [B], ODEN–KIKUCHI–SONG [A]), ODEN–JACQUOTTE [B].

Proposition 4.1: Let $g \in \operatorname{Im} B$, then solution of problem (4.7) converges strongly when $\varepsilon \to 0$, in $V \times Q$ to the solution $(u, \bar{p})$ of problem (1.5) with $\bar{p} \in (\operatorname{Ker} B^t)^\perp$ provided (4.10) and (4.11) hold. Moreovere there exists a constant depending only of f and g, k_0, α, and $\|a\|$ such that

$$(4.12) \qquad \|u - u_\varepsilon\|_V \leq c\,\varepsilon,$$

$$(4.13) \qquad \|\bar{p} - p_\varepsilon\|_Q \leq \frac{c}{k_0}\,\varepsilon.$$

Proof: Substracting (4.7) from (1.5) we have

$$(4.14)\qquad \begin{cases} a(u-u_\varepsilon,v)+b(v,p-p_\varepsilon)=0, & \forall v\in V,\\ b(u-u_\varepsilon,q)+\varepsilon((v,p-p_\varepsilon))_Q=\varepsilon\,((p,q))_Q, & \forall q\in V.\end{cases}$$

Then we apply Theorem 1.2 and we get the result. □

Remark 4.1: This result can of course be applied to a discretized problem. *The reader should notice that discretizing a penalized problem is not in general equivalent to penalize a discrete problem.* In this last case a choice of spaces $V_h \subset V$ and $Q_h \subset Q$ is explicitly done and the penalty method is to be considered as a solution procedure. Discretizing the penalized problem is in general equivalent to choosing $Q_h = BV_h$ which is in general a poor choice. Reduced integration penalty methods have been introduced to circumvent these difficulties and their equivalence with mixed method will be discussed in Chapter VI in the context of Stokes problem and in Chapter VII for moderately thick plates à la Mindlin in the slightly more general setting discussed below. In general, a discrete penalty method will take the form

$$(4.15)\qquad a(u_h,v_h)+\frac{1}{\varepsilon}((IBu_h,IBv_h))_{Q_h}=(f,v_h),$$

where I is an operator from Q into Q_h. In the context of Remark 2.15 this can be seen to be equivalent to the perturbed mixed problem

$$(4.16)\qquad \begin{cases} a(u_h,v_h)+b_h(v_h,p_h)=(f,v_h),\\ b_h(u_h,q_h)-\varepsilon((p_h,q_h))_{Q_h}=0,\end{cases}$$

where $b_h(v_h,q_h)=((IBv_h,q_h))=((Bv_h,q_h))_h$. Whenever I is not the projection operator on Q_h, a consistency error is introduced that has to be taken into account by the method of Section II.2.6. □

Remark 4.2: Proposition 4.1 can be extended to the case where the bilinear form $a(\cdot,\cdot)$ is coercive only on Ker B, that is,

$$a(v_0,v_0)\geq\alpha\,\|v_0\|^2,\qquad \forall v_0\in \operatorname{Ker} B.$$

Indeed one still has the estimate (4.12) and we must then show that p is bounded. But this will be true by applying Proposition 2.11 to problem (4.7). □

Let us now consider the case already discussed in Remark 1.13, that is a regularization of our problem by a scalar product $((\cdot,\cdot))_W$ in a dense subspace $W\hookrightarrow Q$. Depending on which space is identified to its dual space, we shall meet cases where $W\hookrightarrow Q=Q'\hookrightarrow W'$ or where $Q'\hookrightarrow W=W'\hookrightarrow Q$. In all cases the solution of a problem in Q is approximated by the smoother solution

of a problem in W. *We suppose as in Remark 1.13 that* $a(\cdot,\cdot)$ *is coercive on* V but this condition could probably be relaxed.

We thus want to compare $(u,p) \in V \times Q$ solution of problem (1.5) and $(u_\varepsilon, p_\varepsilon) \in V \times W$ solution of

$$(4.17) \qquad \begin{cases} a(u_\varepsilon, v) + b(v_\varepsilon, p) = \langle f, v\rangle_{V'\times V}, \\ b(u_\varepsilon, q) - c_\varepsilon(p_\varepsilon, q) = \langle g, q\rangle_{Q'\times Q}, \qquad \forall q \in W, \end{cases}$$

where c_ε is equivalent to ε times the scalar product on W; that is, it satisfies (1.44) with $\lambda = \varepsilon$ and when $g \in Q'$ is supposed to lie in $\operatorname{Im} B$, which is closed in Q' but not in W'.

Using the bound of Remark 1.13, one gets

$$(4.18) \qquad \|u_\varepsilon\|_V + \|p_\varepsilon\|_{Q/\operatorname{Ker} B^t} \le c\,(\|f\|_{V'} + \|g\|_{Q'}),$$

and we can prove the following proposition.

Proposition 4.2: The solution of problem (4.17) converges strongly in $V \times Q/\operatorname{Ker} B^t$, when ε goes to zero, to the solution of problem (1.5).

Proof: Weak convergence is obvious by the usual extraction of subsequences and by uniqueness of (u,p). To prove strong convergence we use the standard trick and write

$$(4.19) \qquad a(u-u_\varepsilon, u-u_\varepsilon) = a(u,u) - a(u_\varepsilon, u) - a(u, u_\varepsilon) + a(u_\varepsilon, u_\varepsilon).$$

Using (4.17) this can be written as

$$a(u-u_\varepsilon, u-u_\varepsilon) = a(u,u) - a(u_\varepsilon, u) - a(u, u_\varepsilon) + \langle f, u_\varepsilon\rangle - \langle g, p_\varepsilon\rangle - c_\varepsilon(p_\varepsilon, p_\varepsilon)$$

and thus by the positivity of $c_\varepsilon(\cdot,\cdot)$

$$(4.20) \quad a(u-u_\varepsilon, u-u_\varepsilon) \le a(u,u) - a(u_\varepsilon, u) - a(u, u_\varepsilon) + \langle f, u_\varepsilon\rangle - \langle g, p_\varepsilon\rangle.$$

But weak convergence and equations (1.5) show that the limit of the right-hand side is zero. Thus by coerciveness of $a(\cdot,\cdot)$, we have $\|u_\varepsilon - u\|_V \to 0$. Moreover one has

$$a(u-u_\varepsilon, v) + b(v, p-p_\varepsilon) = 0$$

and hence

$$k_0\,\|p-p_\varepsilon\|_Q \le \|a\|\,\|u-u_\varepsilon\|_V$$

and the proof is complete. □

Remark 4.3: This result applies, of course, to the case $W = Q$. It is then a special case of Proposition 4.1.

The problem that remains is to get an estimate on $\|u - u_\varepsilon\|_V$ and $\|p - p_\varepsilon\|_Q$. We now prove

Proposition 4.3: Let $(u_\varepsilon, p_\varepsilon) \in V \times W$ be the solution of problem (4.17) and $(u, p) \in V \times Q$ be the solution of problem (1.5). We then have

$$(4.21)\quad (\|u - u_\varepsilon\|_V + \|p - p_\varepsilon\|_{Q/\operatorname{Ker} B^t}) \leq c \inf_{p_w \in W} \left[\|p - p_w\|_Q + \sqrt{\varepsilon}\, \|p_w\|_W\right].$$

One easily sees by substracting (4.17) and (1.5) with $q \in W$ that one has

$$(4.22)\qquad \begin{cases} a(u - u_\varepsilon, v) + b(v, p - p_\varepsilon) = 0, & \forall v \in V, \\ b(u - u_\varepsilon, q) = c_\varepsilon(p_\varepsilon, q), & \forall q \in W. \end{cases}$$

The argument of Proposition 4.1 cannot be applied for it would require (in (4.22b)) $q \in Q$. However let p_w be any element of W. We rewrite (4.22) as

$$(4.23)\qquad \begin{cases} a(u - u_\varepsilon, v) + b(v, p_w - p_\varepsilon) = b(v, p_w - p), & \forall v \in V, \\ b(u - u_\varepsilon, q) - c_\varepsilon(p_w - p_\varepsilon, q) = -c_\varepsilon(p_w, q), & \forall q \in W. \end{cases}$$

We can now use estimate (1.44) with $\langle g_2, q\rangle = c_\varepsilon\ (p_w, q)$ to get

$$(4.24)\qquad \|u - u_\varepsilon\|_V^2 + \|p_w - p_\varepsilon\|_{Q/\operatorname{Ker} B^t}^2 \leq c\,(\|p_w - p\|_Q^2 + \varepsilon\, \|p_w\|_W^2).$$

From the triangle inequality and the arbitrariness of p_w, one deduces (4.21). □

Remark 4.4: The above result is not optimal. It does not for instance reduce to Proposition 4.2 when $W = Q$. Let us suppose however that there exists a space W_+ dense in W (and hence in Q) such that

$$(4.25)\qquad |c_\varepsilon(p_{w_+}, q)| \leq c_\varepsilon\ \|p_{w_+}\|_{W_+}\ \|q\|_{Q/\operatorname{Ker} B^t}, \qquad \forall q \in W.$$

W_+ is then a space of more regular functions. From (4.23) and (1.47) taking now $p_w = p_{w_+}$ and $g_2 = 0$, $\langle g_1, g\rangle = c_\varepsilon\ (p_{w_+}, q)$, one obtains

$$(4.26)\quad (\|u - u_\varepsilon\|_V + \|p - p_\varepsilon\|_{Q/\operatorname{Ker} B^t}) \leq c \left(\inf_{p_{w_+} \in W_+} \|p - p_{w_+}\|_Q + \varepsilon\, \|p_{w_+}\|_{W_+}\right)$$

and this is now optimal for $W_+ = Q$. □

Remark 4.5: We remark that the right-hand side of (4.26) can be bounded in term of ε whenever p is more regular. Precisely let us suppose that $p \in [W_+, Q]_{\theta,\frac{1}{2}}$ for $0 < \theta < 1$, then one has

$$\inf_{p_{w_+} \in W_+} (\|p - p_{w_+}\|_Q + \varepsilon \, \|p_{w_+}\|_{W_+}) \leq c_\theta \, \varepsilon^\theta \, \|p\|_\theta . \tag{4.27}$$

The space $[W_+, Q]_{\theta,\frac{1}{2}}$ used here is an interpolation space between W_+ and Q. We refer the reader to BERGH–LÖFSTROM [A] where inequality (4.27) is proved in Theorem 3.12. In particular if $W_+ = H^1(\Omega)$ and $Q = L^2(\Omega)$ we have $[W_+, Q]_{\theta,\frac{1}{2}} = H^\theta(\Omega)$. □

Remark 4.6: *Stabilization by a penalty method.*

In the estimate of penalty error presented above, we have supposed that the bilinear form $b(v, q)$ satisfied the inf–sup condition and we did not distinguish between discrete or continuous problems. It must however be said that penalty methods are sometimes used, on discrete problems that do not satisfy the inf–sup condition, as a stabilization procedure. We shall meet such a situation in Section VI.5. It is however worth to introduce the idea in a general setting. Suppose that problem (2.1) is replaced by

$$\begin{cases} a(u, v_h) + b(v_h, p_h) = \langle f, v_h \rangle, \\ b(u - u_h, q_h) - \varepsilon((p_h, q_h))_Q = \langle g, q_h \rangle. \end{cases} \tag{4.28}$$

Substracting from the continuous perturbed problem (4.7) we get

$$\begin{cases} a(u - u_h, v_h) + b(v_h, p - p_h) = 0, \\ b(u - u_h, q_h) - \varepsilon((p - p_h, q_h))_Q = 0. \end{cases} \tag{4.29}$$

By standard techniques and without using the inf–sup condition we can easily get the estimate

$$\begin{cases} \|u_h - u\|_V^2 + \varepsilon \, \|p - p_h\|_Q^2 \leq c_1 \inf\limits_{v_h \in V_h} [\|v_h - u\|_V^2 + \dfrac{1}{\varepsilon}\|v_h - u\|_V^2] \\ \qquad\qquad + c_2 \inf\limits_{q_h \in Q_h} [(1 + \varepsilon) \, \|q_h - p\|_Q^2]. \end{cases} \tag{4.30}$$

The presence of the $1/\varepsilon$ term on the right-hand side forbids ε going to zero. However taking ε to be a function of h leaves a downgraded but simple error estimate. We shall use in Section VI.5 a variant of this procedure that will not cause a loss of accuracy at least for low-order approximations. □

II.5 Iterative Solution Methods. UZAWA's algorithm

To conclude this section we present an iterative algorithm, namely Uzawa's algorithm that is quite efficient in the solution of problems of type (1.5) when the operator A associated with $a(\cdot,\cdot)$ is invertible. In this case, we have already seen that u can be eliminated and we get, in matrix form, problem (3.12). The matrix $BA^{-1}B^t$ that appears in this problem is positive semi-definite and is well suited to a solution by a descent method such as the gradient method or the conjugate gradient method. Moreover, in many important cases, *the condition number* of $BA^{-1}B^t$ will not grow as the discretization mesh is reduced so that convergence properties will be independent of the mesh, which is a very desirable feature. We refer to FORTIN–GLOWINSKI [B] for more details and convergence proofs of the algorithms described below that are nothing but a gradient method applied to (3.12) or a variant of (3.12). Multigrid versions of the method can also be found in VERFÜRTH [B].

II.5.1 Standard UZAWA's algorithm

A – Let p^0 be chosen arbitrarily,

B – p^n being given, find u^{n+1} solution of

$$a(u^{n+1},v) = b(v,p^n) = \langle f,v\rangle, \qquad \forall v \in V, \tag{5.1}$$

C – compute p^{n+1} using with ρ small enough

$$((p^{n+1}-p^n,q)) = \rho[b(u^{n+1},q) - \langle g,q\rangle], \qquad \forall q \in Q, \tag{5.2}$$

D – stop whenever $\|p^{n+1}-p^n\|_Q$ is small enough, otherwise go to step B. □

Although this algorithm behaves well by itself, it is specially well *adapted to being used in conjuction with a penaly method. In this case it can be considered as a way to eliminate the penalty error and to obtain the true solution of the underlying limit problem.* This extension of Uzawa's algorithm is called the *augmented Lagrangian algorithm* and it was introduce by HESTENES [A] and POWELL [A]. Its properties are discussed in details in FORTIN–GLOWINSKI [B].

II.5.2 Augmented Lagrangian algorithm

A – Let p^0 be chosen arbitrarily,

B – p^n being given, find u^{n+1}, the solution of

(5.3)

$$a(u^{n+1},v) + \frac{1}{\varepsilon}\langle S^{-1}Bu^{n+1},Bv\rangle_{Q'\times Q} = \langle f,v\rangle + \frac{1}{\varepsilon}\langle S^{-1}g,Bv\rangle, \quad \forall v\in V;$$

C – compute p^{n+1} using with $\rho \leq 2/\varepsilon$

$$(5.4) \qquad ((p^{n+1} - p^n, q)) = \rho \left[b(u^{n+1}, q) - \langle g, q \rangle \right], \qquad \forall q \in Q.$$

D – stop whenever $\|p^{n+1} - p^n\|_Q$ is small enough, otherwise go to step B. □

Remark 5.1: Using in (5.4) the value $\rho = 1/\varepsilon$ which is very close to the optimal value for ε small, one can rewrite (5.3) and (5.4) as

$$(5.5) \qquad \begin{cases} a(u^{n+1}, v) + b(v, p^{n+1}) = \langle f, v \rangle, & \forall v \in V, \\ b(u^{n+1}, q) - \varepsilon \left((p^{n+1} - p^n, q) \right)_Q = \langle g, q \rangle, & \forall q \in Q. \end{cases}$$

Taking $p^n = 0$, *this is the standard penalty method* and for ε small, p^{n+1} is a good approximation of p. In general two or three interations of (5.5) will be sufficient to complctely eliminate the error due to the penalty term. □

The augmented Lagrangian algorithm is a powerful tool for the numerical solution of Stokes problem (Chapter VI). It has also been applied to a large number of mixed problcms under names or forms that are somctimes hard to recognize but nevertheless equivalent. It is also worth noting that it may converge even for $a(\cdot,\cdot)$ being coercive only on $\operatorname{Ker} B$ as long as the matrix associated with (5.3) is invertible. This will be the case for instance if $a(v,v) \geq \alpha \|v\|_H^2$ as in Section II.2.5. □

II.6 Concluding Remarks

We tried to present in this chapter the basic facts that will serve throughout this book for the analysis of various applications. Many cases have not been treated. We however feel that it should enable the reader to master easily the different extensions that can be found in the literature and even to build by themselves the variants that would be necessary to cover new problems. Some important problems have not been treated in our presentation. This is the case in particular of eigenvalue problems for which mixed and hybrid methods can provide an alternate approach. We refer to the fundamental work of MERCIER–OSBORN–RAPPAZ–RAVIART [A] for a general analysis and also to CANUTO [A,B] where applications can be found. Other presentations of parts of the theory or variants are given in BABUŠKA–OSBORN [B], FALK [A], LEROUX [A], ODEN–REDDY [A]. Another general presentation, ROBERTS–THOMAS [A] can also be consulted for additional for references. We must also point to the analysis of some nonlinear problems given in PIRONNEAU–RAPPAZ [A] for isentropic compressible flows and KIKUCHI–ODEN [A] for contact problems in elasticity. Incompressible non-linear elasticity problems have been studied, for instance, by BERCOVIER–HASBANI–GIBON–BATHE [A], SUSSMAN–BATHE [A], and LE TALLEC–RUAS [A].

III

Function Spaces and Finite Element Approximations

This chapter will present some properties of function spaces that will be necessary for the application of the abstract theory of Chapter II to special problems. We also consider standard results about the finite element approximation of Sobolev spaces and finally we consider approximations of $H(\text{div};\Omega)$. The results of Section III.1 are technical and may be skipped by a reader interested mostly by numerical results.

III.1 Properties of the Spaces $H^m(\Omega)$ and $H(\text{div};\Omega)$

III.1.1 Basic results

We have already introduced, in Chapter I, the Sobolev spaces

$$H^m(\Omega) = \{v \mid v \in L^2(\Omega),\ D^\alpha v \in L^2(\Omega),\ |\alpha| \le m\} \tag{1.1}$$

where,

$$D^\alpha v = \frac{\partial^{|\alpha|} v}{\partial x_1^{\alpha_1} \partial x_2^{\alpha_2} \ldots \partial x_n^{\alpha_n}}, \qquad |\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

The most important of these spaces will be for us $H^1(\Omega)$ (and some of its subspaces) and for fourth-order problems $H^2(\Omega)$.

For a general study of Sobolev spaces, we refer the reader to LIONS–MAGENES [A], NEČAS [A], ADAMS [A]. It is well known that if $\Gamma = \partial\Omega$ is smooth enough (for instance Lipschitzian), it is possible to define the trace

$\gamma u = u|_\Gamma$ of $u \in H^1(\Omega)$ on the boundary Γ. The traces of functions in $H^1(\Omega)$ span a Hilbert space, denoted $H^{1/2}(\Gamma)$ that is a proper dense subspace of $L^2(\Omega)$. The mapping

$$\gamma : H^1(\Omega) \to H^{1/2}(\Gamma), \tag{1.2}$$

is surjective and possesses a continous lifting. The norm,

$$\|\gamma v\|_{1/2,\Gamma} = \inf_{\substack{w \in H^1(\Omega) \\ \gamma w = v}} \|w\|_{1,\Omega}, \tag{1.3}$$

is then equivalent to more standard norms on $H^{1/2}(\Omega)$ as defined in LIONS–MAGENES [A]. If we use on $H^1(\Omega)$ the standard norm $\|v\|^2_{1,\Omega} = |v|^2_{0,\Omega} + |v|^2_{1,\Omega}$ where we have, as defined in Chapter I,

$$|v|^2_{m,\Omega} = \sum_{|\alpha|=m} \int |D^\alpha v|^2 \, dx, \tag{1.4}$$

we can write

$$\|v\|_{1/2,\Gamma} = \|\bar{v}\|_{1,\Omega}, \tag{1.5}$$

where $\bar{v}$ is the unique solution in $H^1(\Omega)$ of the Dirichlet problem,

$$\begin{cases} -\Delta \bar{v} + \bar{v} = 0, \\ \bar{v}|_\Gamma = v. \end{cases} \tag{1.6}$$

We shall denote by $H^{-1/2}(\Gamma)$ the dual space of $H^{1/2}(\Gamma)$ with the dual norm

$$\|v^*\|_{-1/2,\Gamma} = \sup_{v \in H^{1/2}(\Gamma)} \frac{\langle v, v^* \rangle}{\|v\|_{1/2,\Gamma}}, \tag{1.7}$$

where the bracket $\langle \cdot, \cdot \rangle$ denotes duality between $H^{-1/2}(\Gamma)$ and $H^{1/2}(\Gamma)$. It is easily checked that one has,

$$\|v^*\|_{-1/2,\Gamma} = \|\bar{v}^*\|_{1,\Omega}, \tag{1.8}$$

where $\bar{v}^*$ is the solution of the variational Neumann problem,

$$\int_\Omega \underline{\text{grad}}\, \bar{v}^* \cdot \underline{\text{grad}}\, v \, dx + \int_\Omega \bar{v}^* v \, dx = \langle v^*, v \rangle, \qquad \forall v \in H^1(\Omega). \tag{1.9}$$

Remark 1.1: We shall sometimes write formally, $\int_\Gamma v^* v\, ds$ instead of $\langle v^*, v\rangle$, to denote duality between $H^{1/2}(\Gamma)$ and $H^{-1/2}(\Gamma)$. □

We can define in the same way a trace operator γ on $H^2(\Omega)$. It is now possible to define $v|_\Gamma$ in a space denoted $H^{3/2}(\Gamma)$ but also traces of $\underline{\text{grad}}\, v|_\Gamma \in (H^{1/2}(\Gamma))^n$ and thus the trace of the normal derivative $\partial v/\partial n$. We then define

$$H_0^1(\Omega) = \{v \mid v \in H^1(\Omega),\, v|_\Gamma = 0\}, \tag{1.10}$$

$$H_0^2(\Omega) = \{v \mid v \in H^2(\Omega),\, v|_\Gamma = 0,\ \left.\frac{\partial v}{\partial n}\right|_\Gamma = 0\}. \ \square \tag{1.11}$$

Remark 1.2: The reader should be aware that handling Sobolev spaces $H^s(\cdot)$ where $s =$ integer $1/2$ requires some caution (LIONS–MAGENES [A]). In the case of $H^{1/2}(\Gamma)$ it is important to recall some facts. Let Γ_0 be a part of Γ ; then $\phi \in H^{1/2}(\Gamma_0)$ cannot be extended by zero outside Γ_0 to a function in $H^{1/2}(\Gamma)$ [even if paradoxically $\mathfrak{D}(\Gamma_0)$ is dense in $H^{\frac{1}{2}}(\Gamma_0)$]. Dually if $\Gamma = \Gamma_0 \cup \Gamma_1$, one does not get the whole of $H^{-1/2}(\Gamma_0)$ by patching functions of $H^{-1/2}(\Gamma_0)$ and $H^{-1/2}(\Gamma_1)$. Unfortunately spaces $H^{1/2}(\partial K)$ and $H^{-1/2}(\partial K)$, with K an element of a partition of Ω, are met very often in the analysis of hybrid and mixed methods and one must be very careful in handling them. □

Having considered standard Sobolev spaces, we now present some properties of a space specially adapted to the study of mixed and hybrid methods.

The mathematical analysis of mixed methods will use constantly

$$H(\text{div};\Omega) = \{\underline{q} \mid \underline{q} \in (L^2(\Omega))^n,\ \text{div}\, \underline{q} \in L^2(\Omega)\}, \tag{1.12}$$

with the norm

$$\|\underline{q}\|^2_{\text{div},\Omega} = |\underline{q}|^2_{0,\Omega} + |\text{div}\, \underline{q}|^2_{0,\Omega}. \tag{1.13}$$

It is then possible to define $\underline{q} \cdot \underline{n}|_\Gamma$, the normal trace of $\underline{q}$ on Γ.

Lemma 1.1: For $\underline{q} \in H(\text{div},\Omega)$, we can define $\underline{q} \cdot \underline{n}|_\Gamma \in H^{-1/2}(\Gamma)$ and we have Green's formula,

$$\int_\Omega \text{div}\, \underline{q}\, v\, dx + \int_\Omega \underline{q} \cdot \underline{\text{grad}}\, v\, dx = \langle \underline{q} \cdot \underline{n}, v\rangle, \qquad \forall v \in H^1(\Omega). \tag{1.14}$$

Proof: For $\underline{q} \in (\mathfrak{D}(\bar{\Omega}))^n$ and $v \in \mathfrak{D}(\bar{\Omega})$, we have the standard Green's formula

$$\int_\Gamma \underline{q} \cdot \underline{n}\, v\, d\sigma = \int_\Omega \operatorname{div} \underline{q}\, v\, dx + \int_\Omega \underline{q} \cdot \underline{\operatorname{grad}}\, v\, dx$$

and therefore

$$\left| \int_\Gamma \underline{q} \cdot \underline{n}\, v\, d\sigma \right| \leq \|\underline{q}\|_{\operatorname{div},\Omega} \|v\|_{1,\Omega}.$$

Moreover the expression $\int_\Omega \operatorname{div} \underline{q}\, v\, dx + \int_\Omega \underline{q} \cdot \underline{\operatorname{grad}}\, v\, dx$ depends only on the trace $v|_\Gamma \in H^{1/2}(\Gamma)$. The result follows by density of $\mathfrak{D}(\bar{\Omega})$ and $(\mathfrak{D}(\bar{\Omega}))^n$ in $H^1(\Omega)$ and $H(\operatorname{div};\Omega)$, respectively. □

The trace operator defined above also satisfies a surjectivity property.

Lemma 1.2: The trace operator $\underline{q} \in H(\operatorname{div};\Omega) \to \underline{q} \cdot \underline{n}|_\Gamma \in H^{-1/2}(\Gamma)$ is surjective.

Proof: Let $g \in H^{-1/2}(\Gamma)$ be given. Then solving in $H^1(\Omega)$

$$\int_\Omega \underline{\operatorname{grad}}\, \phi \cdot \underline{\operatorname{grad}}\, v\, dx + \int_\Omega \phi v\, dx = \langle g, v \rangle, \qquad \forall v \in H^1(\Omega),$$

and making $\underline{q} = \underline{\operatorname{grad}}\, \phi$, implies $\underline{q} \cdot \underline{n}|_\Gamma = g$. □

Let us now suppose a partition $(\Gamma = D \cup N)$ of the boundary Γ. We define

$$H^1_{0,D}(\Omega) = \{v \mid v \in H^1(\Omega),\ v|_D = 0\}. \tag{1.15}$$

In particular, we have $H^1_{0,D}(\Omega) = H^1_0(\Omega)$ if $D = \Gamma$ and $H^1_0(\Omega) = H^1(\Omega)$ if $D = \emptyset$. We shall also need the space

$$H_{0,N}(\operatorname{div},\Omega) = \{\underline{q} \mid \underline{q} \in H(\operatorname{div}\Omega),\ \langle \underline{q} \cdot \underline{n}, v \rangle = 0,\ \forall v \in H^1_{0,D}(\Omega)\}. \tag{1.16}$$

Remark 1.3: This space contains functions of $H(\operatorname{div};\Omega)$ whose normal traces vanish on N. For reasons related to pathological properties of $H^{1/2}(D)$ and $H^{-1/2}(N)$, it is necessary to use definition (1.16) and not an expression such as $\underline{q} \cdot \underline{n}|_N = 0$ in $H^{-1/2}(N)$. □

In particular we denote $H_0(\operatorname{div};\Omega) = H_{0,N}(\operatorname{div};\Omega)$ when $N = \Gamma$. Finally another important subspace of $H(\operatorname{div};\Omega)$ will be

$$N^0(\operatorname{div};\Omega) = \{\underline{q} \mid \underline{q} \in H(\operatorname{div};\Omega),\ \operatorname{div} \underline{q} = 0\}. \tag{1.17}$$

We then have

Lemma 1.3: The normal trace operator $\underline{q} \to \underline{q} \cdot \underline{n}|_\Gamma$ is surjective from $N^0(\text{div};\Omega)$ onto $\{\mu^* \mid \mu^* \in H^{-1/2}(\Gamma),\ \langle \mu^*, 1\rangle = 0\}$.

Proof: By Green's formula (1.14) we have $\langle \underline{q} \cdot \underline{n}, 1\rangle = 0$ if $\underline{q} \in N^0(\text{div};\Omega)$. Reciprocally, if $g \in H^{-1/2}(\Gamma)$ is given with $\langle g, 1\rangle = 0$, we can solve in $H^1(\Omega)/\mathbb{R}$ the Neumann problem

$$\int_\Omega \underline{\text{grad}}\, \phi \cdot \underline{\text{grad}}\, v \, dx = \langle g, \phi\rangle, \qquad \forall \phi \in H^1(\Omega), \tag{1.18}$$

and making $\underline{q} = \underline{\text{grad}}\, \phi$ yields $\underline{q} \cdot \underline{n} = g$. □

Remark 1.4: In applications, D will be the part of Γ where Dirichlet's conditions are given, and N the part with Neumann's conditions. □

III.1.2 Properties relative to a partition of Ω

This section presents a short introduction to properties of some functional spaces. We refer to RAVIART–THOMAS [A], THOMAS [B] for more details.

Partitioning Ω into subdomains is an essential feature of both standard and non-standard methods. Continuity properties at interfaces between sub-domains are an essential part in the definition of a finite element approximation. Moreover we shall introduce here some notations that will be used throughout the book.

Let $\Omega = \bigcup_{r=1}^m K_r$ be partioned into a family of subdomains. In practice, the K_r will be triangles or quadrilaterals and we shall call them *elements. We shall denote by $\mathcal{T}_h$ a partition into triangles or into quadrilaterals* (or for three-dimensional domains, tetrahedra and hexahedra).

The edges of elements will be denoted e_i ($i = 1,2,3$ or $i = 1,2,3,4$) in the two-dimensional case. For three-dimensional elements we denote again the faces of the elements by e_i ($1 \le i \le 4$ or $1 \le i \le 6$). We also denote by

$$e_{ij} = \partial K_i \cap \partial K_j, \tag{1.19}$$

the interface between element K_i and K_j and

$$\mathfrak{E}_h = \bigcup_{ij} e_{ij} \bigcup \Gamma_h = \bigcup_K \partial K, \tag{1.20}$$

where Γ_h is the set of boundary edges or faces.

Remark 1.5: The index h will of course be related to mesh size, that is, to the size of elements. With an abuse of notation, we shall also use the symbol h for denoting the *maximum diameter* of the *elements of the decomposition.* □

We introduce the functional space

$$X(\Omega) = \{v \mid v \in L^2(\Omega),\ v|_{K_i} \in H^1(K_i),\ \forall i\} = \prod_r H^1(K_r), \tag{1.21}$$

with the norm

$$\|v\|^2_{X(\Omega)} = \sum_r \|v\|^2_{1,K_r} \tag{1.22}$$

and

$$Y(\Omega) = \{\underline{q} \mid \underline{q} \in (L^2(\Omega))^n,\ \underline{q}\,|_{K_i} \in H(\mathrm{div}; K_i),\ \forall i\} = \prod_r H(\mathrm{div}; K_r) \tag{1.23}$$

with the norm

$$\|\underline{q}\|^2_{Y(\Omega)} = \sum_r \|\underline{q}\|^2_{\mathrm{div},\Omega}. \tag{1.24}$$

We also consider the subspace of $Y(\Omega)$,

$$Y^0(\Omega) = \{\underline{q} \mid \underline{q} \in Y(\Omega),\ \underline{q}|_{K_i} \in N^0(\mathrm{div}; K_i),\ \forall i\}. \tag{1.25}$$

We shall now characterize $H^1_{0,D}(\Omega)$ and $H_{0,N}(\mathrm{div};\Omega)$ as subspaces of $X(\Omega)$ and $Y(\Omega)$, respectively. Let us first remark that for $v \in H^1(\Omega)$ and $\underline{q} \in H(\mathrm{div};\Omega)$, we have, denoting $\underline{n}_r$ the outward normal to $\Gamma_r = \partial K_r$,

$$\sum_r \langle v, \underline{q} \cdot \underline{n}_r \rangle_{\Gamma_r} = \langle v, \underline{q} \cdot \underline{n} \rangle_\Gamma \tag{1.26}$$

where $\langle \cdot, \cdot \rangle$ denotes duality between $H^{1/2}(\Gamma_r)$ and $H^{-1/2}(\Gamma_r)$. Indeed we can decompose the Green's formula as

$$\langle v, \underline{q} \cdot \underline{n} \rangle_\Gamma = \sum_r \Big\{ \int_{K_r} \mathrm{div}\ \underline{q}\ v\ dx + \int_{K_r} \underline{q} \cdot \underline{\mathrm{grad}}\ v\ dx \Big\}, \tag{1.27}$$

and apply it inside each element. We can now state

Proposition 1.1: $H^1_{0,D}(\Omega) = \{v \mid v \in X(\Omega),\ \sum_r \langle \underline{q} \cdot \underline{n}_r, v\rangle_{\Gamma_r} = 0,\ \forall \underline{q} \in H_{0,N}(\mathrm{div};\Omega)\}$.

Proof: It is clear by definition that if $v \in H^1_{0,D}(\Omega)$, we have by (1.26) that $\sum_r \langle \underline{q} \cdot \underline{n}_r, v\rangle = 0$, $\forall \underline{q} \in H_{0,N}(\mathrm{div};\Omega)$. Let us consider the reciprocal. Using Green's formula, we get

$$\int_\Omega v \operatorname{div} \underline{q}\, dx = -\sum_r \int_{K_r} \underline{\mathrm{grad}}\, v \cdot \underline{q}\, dx,\ \forall \underline{q} \in H_{0,N}(\mathrm{div};\Omega). \tag{1.28}$$

This implies for all $\underline{q}$, for instance $\underline{q} \in (\mathfrak{D}(\Omega))^n$,

$$\left|\int v \operatorname{div} \underline{q}\, dx\right| \le \Big(\sum_r |v|^2_{1,K_r}\Big)^{1/2} \|\underline{q}\|_{0,\Omega}, \tag{1.29}$$

and therefore $\underline{\mathrm{grad}}\, v \in (L^2(\Omega))^n$, thus $v \in H^1(\Omega)$. We then have $\langle \underline{q} \cdot \underline{n}, v\rangle = 0$, $\forall \underline{q} \in H_{0,N}(\mathrm{div};\Omega)$, so that $v \in H^1_{0,D}(\Omega)$. □

The same kind of proof would yield

Proposition 1.2: $H_{0,N}(\mathrm{div};\Omega) = \{\underline{q} \mid \underline{q} \in Y(\Omega),\ \sum_r \langle \underline{q} \cdot \underline{n}_r, v\rangle_{\Gamma_r} = 0,\ \forall v \in H^1_{0,D}(\Omega)\}$. □

This last result states that functions of $Y(\Omega)$ belong to $H(\mathrm{div};\Omega)$ if their normal traces are "continuous" at the interfaces. This will be an essential point for finite element approximations.

III.1.3 Properties relative to a change of variables

The use of a *reference element*, and therefore of coordinates changes is an essential ingredient of finite element methods, whether for convergence studies or for practical implementation. We must therefore study the effect of a change of variables on our function spaces. We refer to CIARLET [B] for a more complete presentation.

Let then $\hat{K} \subset \mathbb{R}^n$. We denote by $\partial\hat{K}$ its boundary, by $\hat{\underline{n}}$ the outward oriented normal, by $d\hat{x}$ the Lebesgue measure on $\hat{K}$ and, by $d\hat{s}$ the superficial measure induced by it on $\partial\hat{K}$.

Let now F be a smooth (at least C^1) mapping from $\mathbb{R}^n$ into $\mathbb{R}^n$. We define $K = F(\hat{K})$. We suppose that $DF(\hat{x})$, the Jacobian matrix is invertible for any $\hat{x}$ and that F is globally invertible on K. We then have

$$DF^{-1}(x) = (DF(\hat{x}))^{-1}. \tag{1.30}$$

An important case is $F(\hat{x}) = x_0 + B\hat{x}$, that is, F is an affine mapping. Then $DF(\hat{x}) = B$ is a constant matrix. We define

$$(1.31) \qquad \|DF\|_\infty = \sup_{\hat{x}\in K}\Big(\sup_{\xi\in\mathbb{R}^n} \frac{|DF(\hat{x})\xi|_{\mathbb{R}^n}}{|\xi|_{\mathbb{R}^n}}\Big)$$

the norm in $L^\infty(\hat{K})$ of the function $\hat{x} \to \|DF(\hat{x})\|$, that is, the matrix norm of $DF(\hat{x})$. In the same way we have,

$$(1.32) \qquad \|DF^{-1}\|_\infty = \sup_{x\in K}\Big(\sup_{\xi\in\mathbb{R}^n} \frac{|(DF^{-1}(x))\xi|_{\mathbb{R}^n}}{|\xi|_{\mathbb{R}^n}}\Big).$$

We write,

$$(1.33) \qquad J(\hat{x}) = |\det\, DF(\hat{x})|,$$

and for $\hat{x} \in \partial\hat{K}$,

$$(1.34) \qquad J_{\underline{\hat{n}}}(\hat{x}) = J(\hat{x})\, |(DF^{-1})^t\, \underline{\hat{n}}|_{\mathbb{R}^n}.$$

If $\hat{v}(\hat{x})$ is a function on $\hat{K}$, we define $v(x)$ on K by

$$(1.35) \qquad v = \hat{v} \circ F^{-1},$$

and we denote this by $v = \mathfrak{F}(\hat{v})$. We then have the classical formulas

$$(1.36) \qquad \underline{\text{grad}}\, v = (DF^{-1})^t\, \underline{\text{grad}}\, \hat{v} \circ F^{-1} = \mathfrak{F}((DF^{-1})^t\, \underline{\text{grad}}\, \hat{v})$$

and

$$(1.37) \qquad \int_K \mathfrak{F}(\hat{v})\, dx = \int_{\hat{K}} \hat{v}\, J\, d\hat{x},$$

$$(1.38) \qquad \int_{\partial K} \mathfrak{F}(\hat{v})\, d\sigma = \int_{\partial\hat{K}} \hat{v}\, J_{\underline{\hat{n}}}\, d\hat{s}.$$

From this, it is immediate to deduce

Lemma 1.4: The mapping $\mathfrak{F}$ is an isomorphism from $L^2(\hat{K})$ onto $L^2(K)$ and from $H^1(\hat{K})$ onto $H^1(K)$, satisfying,

$$(1.39) \qquad |v|_{0,K} \le \Big(\sup_{\hat{x}} J(\hat{x})\Big)^{1/2} \quad |\hat{v}|_{0,\hat{K}},$$

$$(1.40) \qquad |\hat{v}|_{0,\hat{K}} \le \Big(\inf_{\hat{x}} \; J(\hat{x})\Big)^{-1/2} \quad |v|_{0,K},$$

$$(1.41) \qquad |v|_{1,K} \le \Big(\sup_{\hat{x}} J(\hat{x})\Big)^{1/2} \quad \|DF^{-1}\|_\infty\, |\hat{v}|_{0,\hat{K}},$$

$$(1.42) \qquad |\hat{v}|_{1,\hat{K}} \le \Big(\inf_{\hat{x}} \; J(\hat{x})\Big)^{-1/2} \quad \|DF\|_\infty\, |v|_{1,K}. \;\square$$

Remark 1.6: If F is an affine mapping we also have (CIARLET [A])

$$|v|_{m,K} \leq c\,(\det B)^{1/2}\,\|B^{-1}\|^m\,|\hat{v}|_{m,\hat{K}} \tag{1.43}$$

and, similarly,

$$|\hat{v}|_{m,\hat{K}} \leq c\,(\det B)^{-1/2}\,\|B\|^m\,|v|_{m,K}, \tag{1.44}$$

where the constant c depends only on m and on the space dimension n. □

In the general case, one must use the Leibnitz's formula and the final result is much more complex. We refer to CIARLET-RAVIART [A,C] for this analysis which is beyond the scope of this presentation. □

When building approximations of $H(\mathrm{div};\Omega)$ in Section III.3, we shall be led to using the normal component of vectors as degrees of freedom. The above transformation obviously does not preserve normal components. It does neither map $H(\mathrm{div};\hat{K})$ into $H(\mathrm{div};K)$. To overcome this problem we have to introduce a special (contravariant) transformation known as the Piola's transformation.

Let then $DF(\hat{x})$ be the Jacobian matrix of the transformation $F(\hat{x})$. We consider, for $\underline{\hat{q}} \in (L^2(\hat{K}))^n$, the mapping

$$\mathfrak{G}(\underline{\hat{q}})(x) = \frac{1}{J(\hat{x})} DF(\hat{x})\,\underline{\hat{q}}(\hat{x}), \qquad x = F(\hat{x}). \tag{1.45}$$

It is then elementary to check that one has (in $\mathbb{R}^2$, but the result holds for $\mathbb{R}^n$)

$$\begin{pmatrix} \dfrac{\partial q_1}{\partial x} & \dfrac{\partial q_1}{\partial y} \\ \dfrac{\partial q_2}{\partial x} & \dfrac{\partial q_2}{\partial y} \end{pmatrix} = \frac{1}{J}(DF) \begin{pmatrix} \dfrac{\partial \hat{q}_1}{\partial \hat{x}} & \dfrac{\partial \hat{q}_1}{\partial \hat{y}} \\ \dfrac{\partial \hat{q}_2}{\partial \hat{x}} & \dfrac{\partial \hat{q}_2}{\partial \hat{y}} \end{pmatrix} (DF^{-1}). \tag{1.46}$$

As the trace of a matrix is invariant by a change of variables, we have

$$\mathrm{div}\,\underline{q} = \frac{1}{J}\mathrm{div}\,\underline{\hat{q}}. \tag{1.47}$$

More generally we have (THOMAS [B])

Lemma 1.5: Let $v = \mathfrak{F}(\hat{v})$ and $\underline{q} = \mathfrak{G}(\hat{\underline{q}})$, then

$$\int_K \underline{q} \cdot \underline{\text{grad}}\, v \, dx = \int_{\hat{K}} \hat{\underline{q}} \cdot \underline{\text{grad}}\, \hat{v} \, d\hat{x}, \tag{1.48}$$

$$\int_K v \cdot \text{div}\, \underline{q} \, dx = \int_{\hat{K}} \hat{v} \, \text{div}\, \hat{q} \, d\hat{x}, \tag{1.49}$$

$$\int_{\partial K} \underline{q} \cdot \underline{n} v \, d\sigma = \int_{\partial \hat{K}} \hat{\underline{q}} \cdot \hat{\underline{n}} \, \hat{v} \, d\hat{s}, \ \square \tag{1.50}$$

We refer to THOMAS [B] and RAVIART–THOMAS [A] for the proof of this result and most of the following ones.

From (1.50) we see that $\mathfrak{G}$ preserves the normal trace in $H^{-1/2}$ and enables us to define subspaces of $H(\text{div}; K)$ through the reference element $\hat{K}$. More precisely we have

Lemma 1.6: The mapping $\mathfrak{G}$ is an isomorphism of $H(\text{div}; \hat{K})$ onto $H(\text{div}; K)$ and of $H^0(\text{div}; \hat{K})$ onto $H^0(\text{div}; K)$. Moreover we have

$$|\underline{q}|_{0,K} \leq \left(\inf_{\hat{x}} J(\hat{x})\right)^{-1/2} \|DF\|_\infty \, |\hat{\underline{q}}|_{0,\hat{K}}, \tag{1.51}$$

$$|\hat{\underline{q}}|_{0,\hat{K}} \leq \left(\sup_{\hat{x}} J(\hat{x})\right)^{1/2} \|DF^{-1}\|_\infty \, |\underline{q}|_{0,K}, \tag{1.52}$$

$$|\text{div}\, \underline{q}|_{0,K} \leq \left(\inf_{\hat{x}} J(\hat{x})\right)^{-1/2} |\text{div}\, \hat{\underline{q}}|_{0,\hat{K}}, \tag{1.53}$$

$$|\text{div}\, \hat{\underline{q}}|_{0,\hat{K}} \leq \left(\sup_{\hat{x}} J(\hat{x})\right)^{1/2} |\text{div}\, \underline{q}|_{0,K}. \ \square \tag{1.54}$$

It is also possible to obtain relations between $|\underline{q}|_{m,K}$ and $|\hat{\underline{q}}|_{m,\hat{K}}$ or between $|\text{div}\, \underline{q}|_{m,K}$ and $|\text{div}\, \hat{\underline{q}}|_{m,\hat{K}}$. We refer to THOMAS [B] for details. For the sake of completeness we shall however present the result in the case where F is an affine transformation and $\underline{q} \in H^m(\text{div}; \Omega)$, where

$$H^m(\text{div}; \Omega) = \{\underline{q} \mid \underline{q} \in (H^m(\Omega))^n, \ \text{div}\, \underline{q} \in H^m(\Omega)\}. \tag{1.55}$$

We then have

Lemma 1.7: If the mapping F is affine and if $\underline{q} \in H^m(\text{div}; \Omega)$, the following estimates hold, with $B = DF$:

$$|\underline{q}|_{m,K} \leq (\det B)^{-1/2} \|B^{-1}\|^m \, \|B\| \, |\hat{\underline{q}}|_{m,\hat{K}}, \tag{1.56}$$

$$|\text{div}\, \underline{q}|_{m,K} \leq (\det B)^{-1/2} \|B^{-1}\|^m \, |\text{div}\, \hat{\underline{q}}|_{m,\hat{K}}. \ \square \tag{1.57}$$

The reverse inequalities also hold by a simple exchange of roles between K and $\hat{K}$. Such results are of course essential to the proofs of error estimates. The Piola transformation can be extended to tensor-valued functions with similar properties (cf., for instance, BREZZI–MARINI [A], MARSDEN–HUGHES [A] or CIARLET [A]).

III.2 Finite Element Approximations of $H^1(\Omega)$ and $H^2(\Omega)$

This section will be mainly devoted to the approximation of $H^1(\Omega)$ and its subspaces of the form $H^1_{0,D}(\Omega)$. We shall however sketch the results concerning the approximation of $H^2(\Omega)$. Standard approximations of Sobolev spaces can be subdivided into two classes: conforming and nonconforming methods. Even though nonconforming methods will be studied in the context of hybrid finite element methods, their importance makes it useful to introduce them here. We refer to CIARLET [A], BABUŠKA–AZIZ [A] or RAVIART–THOMAS [D] for a detailed presentation of the following results.

III.2.1 Conforming methods

Conforming methods are the most natural of finite element methods. They yield *internal approximations* in the sense that they enable us to build finite dimensional subspaces of the function space that we want to approximate.

Given a partition of the domain Ω, into triangles or quadrilaterals, a conforming approximation of $H^1(\Omega)$ is a space of *continuous functions* defined by a finite number of parameters (or degrees of freedom).

The last condition is *usually* met by using a space of piecewise polynomial functions or functions obtained from polynomials by a change of variables like (using the notations of Section III.1)

$$v_h|_K = \hat{v} \circ F^{-1}, \tag{2.1}$$

where $K = F(\hat{K})$ and $\hat{v}$ is a polynomial function on $\hat{K}$. Continuity is obtained by a clever choice of degrees of freedom.

Remark 2.1: For triangular elements it is usual and convenient to use piecewise polynomial functions on K. For quadrilaterals it is essential to use (2.1). It must then be noted that $v_h|_K$ is not in general a polynomial on K. This will be the case only for affine transformations. □

To give a more precise definition of our finite element approximations we shall need a few definitions. Let us define on an element K

$$P_k(K): \text{ the space of polynomials of degree } \leq k. \tag{2.2}$$

The dimension of $P_k(K)$ is $\frac{1}{2}(k+1)(k+2)$ for $n=2$, and for $n=3$, it is $\frac{1}{6}(k+1)(k+2)(k+3)$. It will sometimes be convenient to define (for $n=2$)

$$P_{k_1,k_2}(K) = \left\{ p(x_1,x_2) \;\middle|\; p(x_1,x_2) = \sum_{\substack{i\le k_1\\ j\le k_2}} a_{ij}\, x_1^i\, x_2^j \right\} \tag{2.3}$$

the space of polynomials of degree $\le k_1$ in x_1 and $\le k_2$ in x_2. In the same way we can define $P_{k_1,k_2,k_3}(K)$ for $n=3$. The dimensions of these spaces are respectively $(k_1+1)(k_2+1)$ and $(k_1+1)(k_2+1)(k_3+1)$. We then define

$$Q_k(K) = \begin{cases} P_{k,k}(K) & \text{for } n=2 \\ P_{k,k,k,}(K) & \text{for } n=3. \end{cases} \tag{2.4}$$

We shall also need polynomial spaces on the edges (or faces) of the elements. Using the notations of Section III.1.2, we define

$$R_k(\partial K) = \{\phi \mid \phi \in L^2(\partial K),\ \phi|_{e_i} \in P_k(e_i),\ \forall e_i\}, \tag{2.5}$$

$$T_k(\partial K) = \{\phi \mid \phi \in R_k(\partial K) \cap C^0(\partial K)\}. \tag{2.6}$$

Functions of $R_k(\partial K)$ are polynomials of degree $\le k$ on each side (or face) of K. They do not have to be continuous at vertices (or edges). The dimensions of $R_k(\partial K)$ and $T_k(\partial K)$ are respectively for $k \ge 1$:

— $3(k+1)$ and $3k$ for triangles,

— $4(k+1)$ and $4k$ for quadrilaterals,

— $2(k+1)(k+2)$ and $2(k^2+1)$ for tetrahedra.

For hexahedra, it will *usually* be more convenient to consider functions in $Q_k(e_i)$ in the definition of $R_k(\partial K)$ and $T_k(\partial K)$.

To define a finite element, we must, following CIARLET [A], specify three things.

— The geometry: we choose a reference element $\hat{K}$ and a change of variables $F(\hat{x})$, and we set $K = F(\hat{K})$.

— A set $\hat{P}$ of polynomials on $\hat{K}$. For $\hat{p} \in \hat{P}$ we define, on K, $p = \hat{p} \circ F^{-1}$.

— A set of degrees of freedom $\hat{\Sigma}$, that is, a set of linear forms $\{\hat{\ell}_i\}_{1\le i\le \dim \hat{P}}$ on $\hat{P}$. We say that this set is unisolvent when these linear forms are linearly independent, i.e., the knowledge of $\hat{\ell}_i(\hat{p})$ for all i completely defines $\hat{p}$.

A finite element is of *Lagrange type* if its degrees of freedom are *point values*, that is, one is given a set $\{\hat{a}_i\}_{1\le i\le \dim \hat{P}}$ of points in $\hat{K}$ and one defines

$$\hat{\ell}_i(\hat{p}) = \hat{p}(\hat{a}_i), \qquad 1 \le i \le \dim \hat{P}. \tag{2.7}$$

For the approximation of $H^1(\Omega)$, Lagrange type elements will be sufficient but approximating $H^2(\Omega)$ requires Hermite type elements, that is, degrees of freedom involving derivatives.

Remark 2.2: The reader should be aware that not any choice of points will yield an unisolvent set of degrees of freedom. Moreover the points have to be chosen in order to ensure interelement continuity. □

Example 2.1: *Affine finite elements*

This is the most classical family of finite elements. The reference element is the triangle $\hat{K}$ of Figure III.1 and we use the affine transformation

$$F(\hat{x}) = x_0 + B\hat{x}. \tag{2.8}$$

The element K is still a triangle and it is not degenerate provided det $B \neq 0$. We now take $\hat{P} = P_k(\hat{K})$ and choose an appropriate set of degrees of freedom. The standard choice for $k \leq 3$ are presented in Figure III.2.

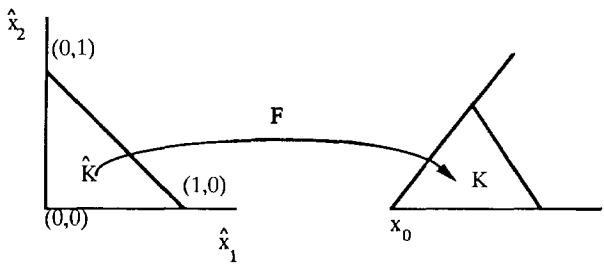


Figure III.1

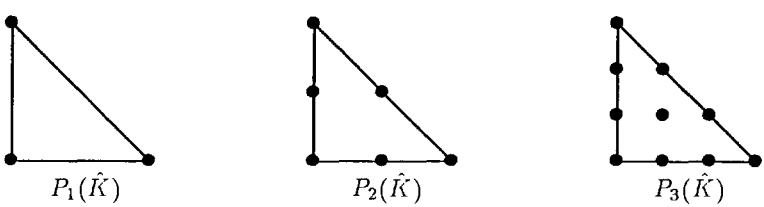


Figure III.2: Standard conforming elements

One notes that this choice of points ensures continuity at interfaces. □

Example 2.2: *Isoparametric triangular elements*

We use the same reference element and the same set $\hat{P}$ as in the previous example. We now take the transformation $F(\hat{x})$ such that each of its components F_i belongs to $P_k(\hat{K})$. For $k = 1$ nothing is changed but for $k \geq 2$, the element K now has curved boundaries. We present the case $k = 2$ in Figure III.3.

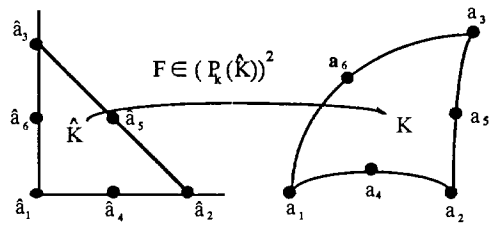
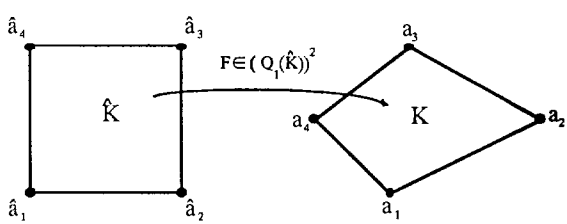


Figure III.3: Isoparametric triangle of degree two

Using such curved triangles enables us to obtain a better approximation of curved boundaries. It must be noted that the curvature of boundaries introduces additional terms in the approximation error and the curved elements should be used only when they are really necessary (CIARLET–RAVIART [C] or CIARLET [A]).

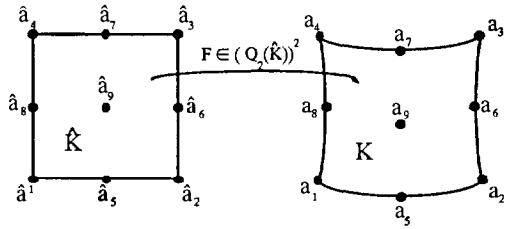
Example 2.3: *Isoparametric quadrilateral elements*

This is also a very classical family of finite elements. The reference element is the square $\hat{K} =]0,1[\times]0,1[$. We take $\hat{P} = Q_k(\hat{K})$ and a transformation F with each component in $Q_k(\hat{K})$. We present the standard choice of degrees of freedom for $k \leq 2$ in Figure III.4. It must be noted that we need $F \in (Q_1(\hat{K}))^2$ to define a general straight-sided quadrilateral.



a) The Q_1 isoparametric element

Figure III.4



b) The Q_2 isoparametric element

Figure III.4

Finally we recall that it is possible to eliminate internal nodes to get the so-called serendipity finite elements. For instance if we take

$$\hat{P} = Q_2'(\hat{K}) = \{\hat{p} \mid \hat{p} \in Q_2(\hat{K}),\ 4\hat{p}(\hat{a}_9) + \sum_{i=1}^{4} \hat{p}(\hat{a}_i) - 2\sum_{i=5}^{8} \hat{p}(\hat{a}_i) = 0\}$$
$$= P_3(\hat{K}) \cap Q_2(\hat{K})$$

we obtain the element of Figure III.5

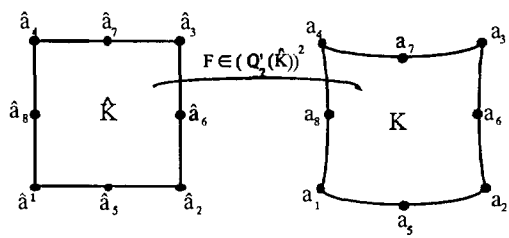


Figure III.5

One notices again that the degrees of freedom have been chosen in order to ensure continuity between elements. □

Example 2.4: *Hermite type elements*

Approximating $H^2(\Omega)$ will require continuity of derivatives at interelement boundaries and leads to the introduction of elements in which values of the derivatives are used as degrees of freedom. The simplest Hermite type element is the P_3 triangle of Figure III.6a.

a) P_3-triangle b) Argyris' triangle

- value of the function
- values of the function and its first derivative
- values of the function and its first and second derivative
- value of the normal derivative

Figure III.6

Here the degrees of freedom are values of the function and its derivative at vertices plus a point value at barycenter. This element *does not* enable us to build an approximation of $H^2(\Omega)$. To do so, one must use Argyris' triangle (Figure III.6b) where polynomials of degree 5 are used. (Composite elements may also be used.) For quadrilaterals the analogues are easily built. The difficulty of building approximations of $H^2(\Omega)$ by standard methods was one of the major reason for the introduction of various kinds of mixed or hybrid methods for plate problems (cf. Section IV.5 or Section VII.1).

We now have to say a few words about the approximation of a given function v by the finite element spaces just described or similar ones. We shall not give proofs, for which we refer to CIARLET [A], STRANG–FIX [A], CLEMENT [A].

We must first define *the interpolate of* v. For a general set of degrees of freedom $\{\hat{\ell}_i\}$ on $\hat{K}$, we define $\hat{r}_h\hat{v}$ by

$$\hat{\ell}_i(\hat{r}_h\hat{v}) = M_i(\hat{v}), \qquad 1 \le i \le \dim \hat{P}. \tag{2.9}$$

The operator M must be a well-defined continuous form. When the linear forms $\hat{\ell}_i$ are defined by (2.5), it is natural to set

$$(\hat{r}_h\hat{v})(\hat{a}_i) = \hat{v}(\hat{a}_i). \tag{2.10}$$

This definition makes sense only when $\hat{v}$ is a *continuous function* which is not the case when $v \in H^1(\Omega)$. For Lagrange type elements in $\mathbb{R}^2$ or $\mathbb{R}^3$, $\hat{v} \in H^2(\hat{K})$ is a sufficient condition for (2.10) to be justified and $\hat{r}_h\hat{v}$ is just the Lagrange interpolate, in the classical sense, of $\hat{v}$. For $v \in H^1(\Omega)$, CLEMENT [A] *has defined a continuous interpolate* $\hat{r}_h$ *using averages of* u *instead of point values*. This also implies a more elaborate use of reference elements. In particular, the operator $\hat{r}_h\hat{v}$ is no longer defined on an element. In fact the nodal

values of $\hat{r}_h\hat{v}$ depend on the value of $\hat{v}$ on the adjacent elements through an averaging process.

Once $\hat{r}_h\hat{v}$ is defined, we can define on K,

$$r_h v = (\hat{r}_h(v \circ F)) \circ F^{-1} = (\hat{r}_h \hat{v}) \circ F^{-1}. \tag{2.11}$$

We rapidly recall a few classical results. We refer the reader to CIARLET [A] for a detailed presentation. We first consider the case of *affine elements*, assuming first r_h to be defined by the usual interpolate (2.10).

Proposition 2.1: If the mapping F is affine, that is $F(\hat{x}) = x_0 + B\hat{x}$, and if $r_h p_k = p_k$ for any $p_k \in P_k(K)$, we have for $v \in H^s(\Omega)$, $m \leq s$, $1 < s \leq k+1$,

$$|v - r_h v|_{m,K} \leq c\, \|B^{-1}\|^m\, \|B\|^s\, |v|_{s,K}. \tag{2.12}$$

The proof uses (1.43), its reciprocal (1.44) and the classical results stated below. □

Lemma 2.1: $|\cdot|_{k+1,\Omega}$ is a norm on $H^{k+1}(\Omega)/P_k(\Omega)$, equivalent to the standard quotient norm. □

From this one deduces another classical result:

Lemma 2.2: (Bramble–Hilbert lemma) Let L be a continuous linear form on $H^{k+1}(\Omega)$ such that $L(p_k) = 0$ for any $p_k \in P_k(\Omega)$. Then there exists a constant c (depending on L and Ω) such that one has

$$|L(v)| \leq c\, |v|_{k+1,\Omega}. \; \square \tag{2.13}$$

Results similar to (2.12), although more complex, can be obtained for general isoparametric elements (CIARLET [A], CIARLET–RAVIART [A,C]). Let then h_K be the diameter of K. Provided some classical conditions on the *shape of elements* forbidding degeneracy (CIARLET [A]) are fulfilled, relation (2.12) can be converted into a relation involving a power of h_K. For affine elements one defines for instance

$$\sigma_K = \frac{h_K}{\rho_K}, \tag{2.14}$$

where ρ_K is the diameter of the largest inscribed disk (or sphere) in K.

We shall in the following always assume that the interpolation operator r_h is defined by the method of CLEMENT [A], that is, by a local projection instead

of a pointwise interpolate. This allows us to get rid of the condition $s > 1$ of Proposition 2.1. To state this result we define

$$\begin{aligned} \Delta K &= \{K' \mid \bar{K}' \cap \bar{K} \neq \emptyset\}, \\ h_{\Delta K} &= \sup_{K' \in \Delta K} h_{K'}, \\ \sigma_{\Delta K} &= \sup_{K' \in \Delta K} \sigma_{K'}. \end{aligned}$$

We then have (CLEMENT [A]):

Proposition 2.2: If the mapping F is affine and if $r_h p_k = p_k$ for any $p_k \in P_k(\Omega)$, then there is a constant, depending on k and m, such that for $0 \leq m \leq s$, $1 \leq s \leq k+1$,

$$|r_h v - v|_{m,K} \leq c\, \sigma_{\Delta K}\, h_{\Delta K}^{s-m}\, |v|_{s,\Delta K}. \ \square \tag{2.15}$$

We then say that *a family of triangulations* $(\mathcal{T}_h)_{h \geq 0}$ *is* ***regular*** **if**

$$\sigma_K < \sigma, \qquad \forall K \in \mathcal{T}_h, \qquad \forall h. \tag{2.16}$$

For the geometrical meaning of this condition, we refer to CIARLET [A]. We may recall however that (2.16) can be written as a condition on angles excluding degenerate elements. For general curved elements there is also a condition on the curvature of the sides.

We then have the approximation result,

Corollary 2.1: If $(\mathcal{T}_h)_{h \geq 0}$ is regular family of affine partitions, there exists a constant c depending on k and σ, such that

$$\begin{cases} |r_h v - v|_{m,K} \leq c\, h^{s-m}\, |v|_{s,\Delta K} \\ \displaystyle\sum_K h_K^{2m-2} |v - \Pi_1 v|_{m,K}^2 \leq c \|v\|_{1,\Omega}^2. \ \square \end{cases} \tag{2.17}$$

For more general partitions including general isoparametric elements, the result is qualitatively the same: we have an $O(h^k)$ approximation provided the family of partitions is regular in a sense to be precised.

We also refer the reader to JAMET [A] where some degenerate cases are analyzed.

From the elements described above, we can build approximations of $H^1(\Omega)$ and $H^2(\Omega)$. The idea is, of course, to use functions whose restriction to an element belongs to a set of polynomial (or image S of polynomial) functions. Let $S_k(K)$ be a subspace of $P_k(K)$. We define for a partition $\mathcal{T}_h$ of Ω

$$\mathfrak{L}^s(S_k, \mathcal{T}_h) = \{v \mid v \in H^s(\Omega),\ v|_K \in S_k(K)\}. \tag{2.18}$$

Remark 2.3: In the two-dimensional case, for $s = 1$ and $s = 2$, we have $\mathfrak{L}^s(S_k, \mathcal{T}_h) \subset C^{s-1}(\bar{\Omega})$. although, this is not true for $H^s(\Omega)$ □

We shall reduce this notation when no confusion is to be feared and write

$$\mathfrak{L}^s_k = \mathfrak{L}^s(P_k, \mathcal{T}_h) \tag{2.19}$$

when $\mathcal{T}_h$ is built from triangles, and $S_k = P_k$ (= the space of polynomials of degree $\leq k$). In the same way we shall write

$$\mathfrak{L}^s_{[k]} = \mathfrak{L}^s(Q_k, \mathcal{T}_h) \tag{2.20}$$

when $\mathcal{T}_h$ is built from quadrilaterals.

We shall often use in our constructions *bubble functions.* For an element K a bubble function is a function vanishing on ∂K. Thus we say that S_k is a set of bubble functions if $S_k \subset H^1_0(K)$. We then denote

$$B(S_k) = \mathfrak{L}^1(S_k, \mathcal{T}_h) = \mathfrak{L}^0(S_k, \mathcal{T}_k) \tag{2.21}$$

and we shall use the compact notation

$$\begin{cases} B_k = B(P_k \cap H^1_0(K)), \\ B_{[k]} = B(Q_k \cap H^1_0(K)), \end{cases} \tag{2.22}$$

when no ambiguity will be possible.

Spaces of bubble functions will be used to build enriched spaces. For instance the space $\mathfrak{L}^1_2 \oplus B_3$ will be useful in Chapter VI for the approximation of Stokes problem.

When approximating a standard elliptic problem, the finite element spaces introduced up to now can be used directly in the variational fomulation of the problem and error estimates follow from interpolation error estimates (CIARLET [A]). In many cases, however, nonconforming methods have proved to yield accurate (and sometimes easier to handle) approximations.

III.2.2 Nonconforming methods

We shall meet later nonconforming methods when studying hybrid finite element methods. In many cases, it will however be more convenient to see them in the frame of *external approximations*, which we now define.

Let us consider a variational problem (with $f \in V'$),

$$a(u, v) = \langle f, v \rangle_{V' \times V}, \qquad \forall v \in V,\ u \in V, \tag{2.23}$$

where V is some Hilbert space and $a(u, v)$ a bilinear (coercive) form on $V \times V$.

Suppose we can find a larger space $S \supset V$, such that there exists a canonical extension $\tilde{a}(\cdot,\cdot)$ to $S \times S$, satisfying

$$\tilde{a}(u,v) = a(u,v), \qquad \forall u,\ v \in V. \tag{2.24}$$

Moreover, let $V_h \subset S$ be a family of finite-dimensional subspaces of S such that

$$v = \lim_{h \to 0} v_h \Rightarrow v \in V. \tag{2.25}$$

V_h is said to be an external approximation to V. Assuming that f can be extended to an element $\tilde{f}$ in S', we can now approximate problem (2.23) by: find $u_h \in V_h$, the solution of

$$\tilde{a}(u_h, v_h) = \langle \tilde{f}, v_h \rangle_{S' \times S}, \qquad \forall v_h \in V_h. \tag{2.26}$$

Using standard coerciveness and continuity assumptions, one gets from (2.23) and (2.26) a result known as Strang's lemma (CIARLET [A], STRANG–FIX [A]).

$$\begin{aligned} \|u - u_h\|_S &\le c \inf_{v_h \in V_h} \|u - v_h\|_S + \sup_{v_h \in V_h} \frac{|\tilde{a}(u, v_h) - \langle \tilde{f}, v_h \rangle|}{\|v_h\|_S} \\ &\le c \inf_{v_h \in V_h} \|v - v_h\| + E_h(u, v_h). \end{aligned} \tag{2.27}$$

The last term can be seen as a *consistency* term: it measures how well the exact solution satisfies the discrete equation. This term vanishes when $V_h \subset V$ and we then get the standard result for the conforming case.

In classical situations we have $V = H^1(\Omega)$ or $V = H^2(\Omega)$ (or one of their subspaces). Introducing a partition of the domain into m subdomains K_r, and assuming $V = H^1(\Omega)$, we take $S = X(\Omega)$ as defined in Section III.1.3. Any bilinear form of the type

$$\int_\Omega a(x)\ \underline{\text{grad}}\ u \cdot \underline{\text{grad}}\ v\ dx \tag{2.28}$$

can immediately be extended to $X(\Omega)$ by writing

$$\tilde{a}(u,v) = \sum_{r=1}^{m} \int_{K_r} a(x)\ \underline{\text{grad}}\ u \cdot \underline{\text{grad}}\ v\ dx. \tag{2.29}$$

We now want to find a subspace of $X(\Omega)$, approximating $H^1(\Omega)$ and such that error estimates obtained from (2.27) be "optimal". Optimality is here relative to the degree of the polynomials from which the approximation is built:

we would like to keep $O(h^k)$ estimates when using polynomials of degree k. We are thus led to study the second term on the right-hand side of (2.27). We shall make this analysis later in the context of hybrid finite element methods; we shall therefore merely state the result which is quite classical (CEA [B], IRONS–RAZZAQUE [A], CROUZEIX–RAVIART [A], FRAEIJS DE VEUBEKE [B]) and was discovered on empirical grounds and is known as the *Céa-test: the moments up to degree $k-1$, of u_h on any interface of the partition must be continuous, that is,*

$$\int_S u_h \, p_{k-1} \, ds, \ \forall p_{k-1} \in P_{k-1}(S) \tag{2.30}$$

is continuous across S. □

A more general form was given by LASCAUX–LESAINT [A]. It states that the consistency term $E_h(u, v_h)$ must vanish whenever $u \in P_r(\Omega)$. For plate problems this implies a condition similar to (2.30) for u_h and its derivatives. To fix ideas we recall a few classical examples. □

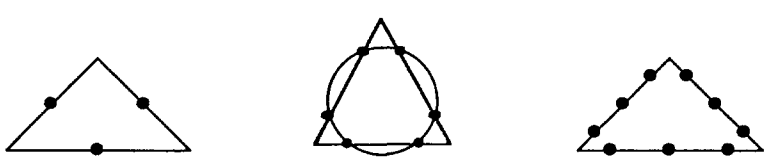
In conformity to notation (2.18) we denote by $\mathfrak{L}^{1,NC}(S_k, \mathcal{T}_h)$ a nonconforming approximation of $H^1(\Omega)$ built from functions of $S_k(K)$. We shall simplify this notation whenever possible as in the following example.

Example 2.4: *Nonconforming elements on the triangle*

Let us consider a partition of Ω into straight-sided triangles and an approximation

$$\mathfrak{L}_k^{1,NC} = \{v_h \mid v_h \in L^2(\Omega), \ v_h|_K \in P_k(K), \ \forall K \in \mathcal{T}_h, \quad \sum_K \int_{\partial K} u_h \phi \, ds = 0, \quad \forall \phi \in R_k(\partial K)\} \tag{2.31}$$

It is then easy to see that the patch-test implies that the functions of $\mathfrak{L}_k^{1,NC}$ should be continuous at the k Gauss–Legendre points on every side of the triangles (Figure III.7).



a) $k = 1$ b) $k = 2$ c) $k = 3$

Continuity points for nonconforming elements

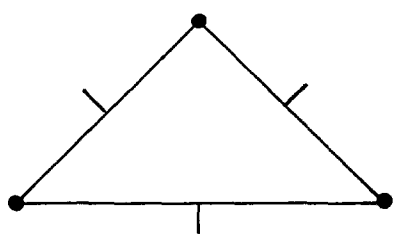
Figure III.7

For k odd, those points, with internal points for $k \geq 3$, can be used as degrees of freedom, but for k even it is not so. For instance the six Gaussian points of the $k = 2$ case lie on an ellipse and the values at these points are not independent. It was however shown in FORTIN–SOULIE [A] that this element can nevertheless be used in a very simple way. This was extended to the three-dimensional case in FORTIN [B]. It must be noted that in three-dimensional nonconforming elements, the patch-test implies in general no point continuity. □

Nonconforming approximations of $H^2(\Omega)$ (CIARLET [B]) have been widely used because of the difficulty to obtain conforming elements. We refer the reader to LASCAUX–LESAINT [A] where many examples are given. We shall however use in Chapters VI and VII the following nonconforming approximation of $H^2(\Omega)$.

Example 2.5: *Morley's triangle*

In plate problems, where an approximation of $H^2(\Omega)$ is needed, an important nonconforming element is Morley's triangle (Figure III.8).



• point value
— normal derivative value

Figure III.8

The functions v_h are supposed to be in $P_2(K)$ for every K. The degrees of freedom are point values at the vertices of the triangle and normal derivatives at mid-side points. It can be shown by the method of LASCAUX–LESAINT [A] that this provides a consistent approximation that will converge as $O(h)$ in a *discrete* $H^2(\Omega)$-norm. We shall denote by $\mathfrak{L}_2^{2,NC}$ the approximation of $H^2(\Omega)$ built from such elements. □

Example 3.6: *Nonconforming elements on the rectangle*

We could consider a partition of Ω into straight-side quadrilaterals, and an approximation of $H^1(\Omega)$ defined by

$$(2.32)\quad \mathfrak{L}_{[k]}^{1,NC} = \{u_h \in L^2(\Omega),\ u_h|_{K_r} = \hat{u}_h \circ F_r^{-1},\ \hat{u}_h \in Q_k(\hat{K}),\ \text{(2.30) holds}\}.$$

Here again the patch-test implies continuity at the Gauss–Legendre points of the interfaces. It is never possible to use these points as degrees of freedom. For $k = 1$, the functions $(\hat{x} - \frac{1}{2})(\hat{y} - \frac{1}{2}) \in Q(\hat{K})$ vanishes on the four Gauss–Legendre points of the sides that are indeed midpoints in this case. For $k = 2$, the points lie on an ellipse, and so on. This explains why nonconforming quadrilateral elements have been used so little. It is however possible to extend the method of FORTIN–SOULIE [A] to these cases. □

The above examples are in no way exhaustive: many other nonconforming approximations can be built and some are indeed effectively used (LESAINT [A], HENNART–JAFFRE–ROBERTS [A]). As we shall see in the sequel, nonconforming methods are strongly related, and often equivalent to hybrid methods or mixed methods. We think it is preferable to delay further examples until they are met in a proper context.

III.2.3 Nonpolynomial approximations: Spaces $\mathfrak{L}_k^s(\mathfrak{E}_h)$

In the applications involving hybrid methods it will be useful to consider approximation spaces built from functions that have a polynomial trace on ∂K but which are not necessarily polynomials inside K. These spaces will be useful whenever only the trace is computationally important: they can be thought of as defined only on $\mathfrak{E}_h = \bigcup_K \partial K$ (cf. (1.20)). We thus define for $s \geq 1$

$$\mathfrak{L}_k^s(\mathfrak{E}_h) = \{v \mid v \in H^s(\Omega),\ v|_{\partial K} \in T_k(\partial K),\ \forall K\}, \tag{2.33}$$

and for $s = 0$,

$$\mathfrak{L}_k^0(\mathfrak{E}_h) = \{v \mid v \in L^2(\mathfrak{E}_h),\ v|_{\partial K} \in R_k(\partial K),\ \forall K\}. \tag{2.34}$$

For $s \geq 1$, functions of $\mathfrak{L}_k^s(\mathfrak{E}_h)$ are evidently approximations of $H^1(\Omega)$ at optimal order with respect to k. It is also possible to get error estimates on the traces.

III.2.4 Scaling arguments

We shall briefly recall here the basic idea of the scaling arguments of DUPONT–SCOTT [A]. We shall do it on a very simple example, but it will be clear how the idea applies to more general cases. Assume that we want to prove the following *inverse inequality* for elements $v_h \in \mathfrak{L}_k^s$: there exists a constant c depending only on k and on the minimum angle θ_0 in $\mathcal{T}_h$ such that, on every element K, we have

$$|v_h|_{1,K} \leq c h_K^{-1}\, |v_h|_{0,K}. \tag{2.35}$$

We construct first a new element $\hat{K}$ such that the mapping $F : \hat{K} \to K$ is simply given by

$$\underline{x} = h_K \underline{\hat{x}} + \underline{b} \tag{2.36}$$

and K has a vertex in the origin. Formulas (1.43) and (1.44) then simply become (in two dimensions)

$$\hat{v}|_{m,\hat{K}} = h_K^{m-1}\, |v|_{m,K} \tag{2.37}$$

and we easily get

$$|v_h|_{1,K} = |\hat{v}|_{1,\hat{K}} \le c(k,\hat{K})\, |\hat{v}|_{0,\hat{K}} \le c(k,\hat{K})\, h_K^{-1}\, |v|_{0,K}. \tag{2.38}$$

Now we remark that $c(k,\hat{K})$ actually depends *continuously* on the shape of $\hat{K}$ (a similar argument was already used in BREZZI–MARINI [A]). In particular, if one considers the family K_{θ_0} of all the triangles having diameter $= 1$, one vertex in the origin and a minimum angle $\ge \theta_0$, one easily gets

$$\sup_{\hat{K}\in K_{\theta_0}} c(k,\hat{K}) \le c(k,\theta_0) \tag{2.39}$$

by compactness (DUPONT–SCOTT [A]). Hence from (2.38) and (2.39) we get

$$|v_h|_{1,K} \le c(k,\theta_0)\, h_K^{-1}\, |v|_{0,K}, \tag{2.40}$$

that is, (2.35).

Note that, in this particular case, it would have been equally easy (or even easier) to derive directly (2.35) by using (1.43) and (1.44) and a fixed $\hat{K}$ = unit triangle. However (2.37) is easier to use and the continuity argument (2.39) is always essentially the same in many different applications, so that using the scaling (2.36) actually results in a simplification. For instance one can easily get by this method the inequality

$$\int_{\partial K} |v_h|\, d\sigma = h_K \int_{\partial\hat{K}} |\hat{v}_h|\, d\hat{s} \le c(k,\theta_0)\, h_K\, |\hat{v}_h|_{o,\hat{K}} = c(k,\theta_o)\, |v_h|_{0,K}. \tag{2.41}$$

In the same way, one can guess, for instance, that one has

$$\|\partial v_h/\partial n\|_{L^\infty(\partial K)} \le c(k,\theta_0)\, h_K^{-2}\, |v_h|_{0,K}, \tag{2.42}$$

because both sides behave like h_K^{-1} in the transformation (2.36) and the inequality holds on a fixed element of size $= 1$. However, Note that an inequality of the type

$$\|v_h\|_{L^\infty(\partial K)} \le c(k,\theta_0)\, |v_h|_{1,K}$$

is still hopeless (take $v_h = 1$!) unless we specify, for instance, that v_h has zero mean value in K.

III.3 Approximations of $H(\text{div};\Omega)$

Although this section is important by itself, as we shall use $H(\text{div};\Omega)$ in many examples throughout this book, its importance also lies in its value as a model. The techniques introduced for the approximation of $H(\text{div};\Omega)$ can indeed be applied to other situations and we shall meet for instance similar constructions in the discretization of the Hermann–Johnson mixed formulations (Chapter VII). Our presentation does not follow the original work of RAVIART–THOMAS [A], and THOMAS [B] later generalized and extended to the three-dimensional case by NEDELEC [A]. We shall rather start from an approximation introduced by BREZZI–DOUGLAS–MARINI [B][C], BREZZI–DOUGLAS–DURAN–FORTIN [A], and NEDELEC [B] that contains (for triangular elements) the elements of NEDELEC [A] and RAVIART–THOMAS [A]. In the case of quadrilaterals, we introduce a general element containing the elements of RAVIART–THOMAS [A] the BREZZI–DOUGLAS–MARINI [B] elements and the ones of BREZZI–DOUGLAS–FORTIN–MARINI [A], thus clarifying the relation between those two. As the triangular case is simple and more intuitive, we shall first consider it into details. Quadrilateral elements will be treated afterwards.

III.3.1 Simplicial approximations of $H(\text{div}; K)$

In this section, the element K will be either a triangle ($n = 2$) or a tetrahedron ($n = 3$). We still denote e_i ($i = 1,2,3$ or $i = 1,2,3,4$) the sides (or the faces) of K.

Following BREZZI–DOUGLAS–MARINI [B] (for $n = 2$) and BREZZI–DOUGLAS–DURAN–FORTIN [A] (for $n = 3$) we now introduce, to approximate $H(\text{div}; K)$, the space

$$BDM_k(K) = (P_k(K))^n. \tag{3.1}$$

The dimension of BDM_k is thus

$$\dim BDM_k = \begin{cases} (k+1)(k+2) & \text{for } n = 2, \\ \frac{1}{2}(k+1)(k+2)(k+3) & \text{for } n = 3 \end{cases} \tag{3.2}$$

For $\underline{q} \in BDM_k(K)$, we evidently have div $\underline{q} \in P_{k-1}(\partial K)$. Moreover, the normal trace $\underline{q} \cdot \underline{n}$ on ∂K belongs to $R_k(\partial K)$ as defined by (2.5). In order to build from BDM_k an approximation of $H(\text{div};\Omega)$, it will be necessary to ensure continuity (up to the sign) of $\underline{q} \cdot \underline{n}$ at the interfaces. This will be made possible by the choice of appropriate degrees of freedom. Indeed we have

Proposition 3.1: For $k \geq 1$, and for any $\underline{q} \in BDM_k$ the following relations imply $\underline{q} = 0$.

$$\int_{\partial K} \underline{q} \cdot \underline{n}\, p_k \, ds = 0, \qquad \forall p_k \in R_k(\partial K), \tag{3.3}$$

$$(3.4) \qquad \int_K \underline{q} \cdot \underline{\text{grad}}\, p_{k-1} dx = 0, \qquad \forall p_{k-1} \in P_{k-1}(K),$$

$$(3.5) \quad \int_K \underline{q} \cdot \underline{\phi}_k \, dx = 0, \qquad \forall \underline{\phi}_k \in \{\underline{\phi}_k \in (P_k)^n \mid \text{div}\, \underline{\phi}_k = 0,\ \underline{\phi}_k \cdot \underline{n}|_{\partial K} = 0\} = \Phi_k.$$

Indeed, it is easy to check that (3.3) and (3.4) are equivalent to $\underline{q} \in \Phi_k$, since we may write

$$(3.6) \qquad \int_K \text{div}\, \underline{q}\ \text{div}\, \underline{q}\, dx = -\int_K \underline{q} \cdot \underline{\text{grad}}\ \text{div}\, \underline{q}\, dx + \int_{\partial K} \underline{q} \cdot \underline{n}\ \text{div}\, \underline{q}\, d\sigma$$

Thus (3.3) and (3.4) imply div $\underline{q} = 0$. Reciprocally, it is trivial that (3.3) and (3.4) hold for $\underline{\phi}_k \in \Phi_k$. □

To prove that (3.3),(3.4), and (3.5) can be used to define degrees of freedom for BDM_k by choosing bases for $R_k(\partial K)$, $P_{k-1}(K)$ and Φ_k, there remains to check that the set obtained from (3.3) and (3.4) is linearly independent, that is,

Lemma 3.1: Let $g \in R_k(\partial K)$ and $f \in P_{k-1}(K)$ such that

$$(3.7) \qquad \int_{\partial K} g\, \underline{q} \cdot \underline{n}\, ds + \int_K \underline{q} \cdot \underline{\text{grad}}\, f\, dx = 0, \qquad \forall \underline{q} \in BDM_k(K),$$

then $g = 0$ and $f =$ constant.

Proof: Using the change of variables (1.45) and Lemma 1.5, it is sufficient to prove the result on the reference element (Figure III.9). We give the construction for $n = 3$ as the case $n = 2$ is a simple restriction of it. One first uses in (3.7), λ_4 being the fourth barycentric coordinate (that is, $\lambda_4 = 1 - x - y - z$),

$$q_1 = x\frac{\partial f}{\partial x}\lambda_4, \quad q_2 = y\frac{\partial f}{\partial y}\lambda_4, \quad q_3 = z\frac{\partial f}{\partial z}\lambda_4.$$

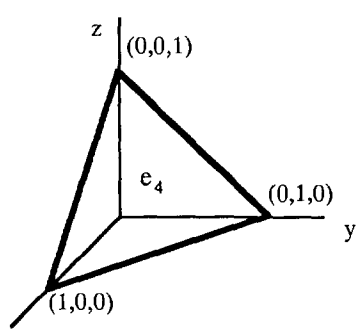


Figure III.9

Then $\underline{q} \cdot \underline{n}|_{\partial K} = 0$ and we get from (3.7)

$$\int_K \left[x\left(\frac{\partial f}{\partial x}\right)^2 + y\left(\frac{\partial f}{\partial y}\right)^2 + z\left(\frac{\partial f}{\partial z}\right)^2 \right] \lambda_4 \, dx = 0 \tag{3.8}$$

which implies $\underline{\text{grad}}\, f = 0$ as all terms in the integral are positive. We now take $q_1 = x p_{k-1}$, $q_2 = q_3 = 0$. From this comes $\int_{e_4} xg \; p_{k-1} \, ds = 0$. In the same way we get $\int_{e_4} yg \; p_{k-1} \, ds = \int_{e_4} zg \; p_{k-1} \, ds = 0$ and as $x + y + z = 1$ on e_4, $\int_{e_4} g \; p_{k-1} \, ds = 0$. All these conditions imply $g|_{e_4} = 0$. Finally we take $q_i = g|_{e_i} = g_i$ and (3.7) implies $\sum_{i=1}^{3} \int_{e_i} (g_i)^2 \, ds = 0$, hence $g = 0$. □

Let us now count the number of conditions thus induced:

(3.9)

$$\dim R_k(\partial K) + \dim P_{k-1}(K) - 1 = \begin{cases} \frac{1}{2}k^2 + 7k + 4 & \text{for } n = 2 \\ \frac{1}{6}k^3 + 15k^2 + 38k + 18 & \text{for } n = 3. \end{cases}$$

From this we can deduce, by standard arguments of linear algebra,

(3.10)

$$\dim \Phi_k = \begin{cases} \frac{1}{2}k(k-1) = \dim P_{k-2}(K) \text{ for } n = 2, \; k \geq 2, \\ \frac{1}{6}3k^3 - 3k - \frac{1}{6}(k-2)(k-1)k = \begin{cases} \dim[(P_{k-2})^3] - \dim(P_{k-3}), \\ \qquad n = 3, \; k \geq 3, \\ \dim[(P_{k-2})^3], \; n=3, \; k=2. \end{cases} \end{cases}$$

In the two-dimensional case, the space Φ_k can easily be characterized.

Lemma 3.2: For $n = 2$, we have,

$$\Phi_k = \{\underline{\phi}_k \mid \underline{\phi}_k = \underline{\text{curl}} \; b_K \; p_{k-2}, \; p_{k-2} \in P_{k-2}(K)\}, \tag{3.11}$$

where $b_K = \lambda_1 \lambda_2 \lambda_3 \in B_3(K)$ is the bubble function on K.

Proof: Any $\underline{\phi}_k \in \Phi_k$ is the curl of a polynomial of degree $k+1$. A simple count of degrees of freedom terminates the proof. □

In the three-dimensional case, the construction of Φ_k is less direct. It is still true that $\underline{\phi}_k \in \Phi_k$ implies that $\underline{\phi}_k$ is the $\underline{\text{curl}}$ of a vector function polynomial of degree $k+1$. However this representation is not, in general, unique and it is not so easy to explicit a basis (NEDELEC [B]).

One must however make two remarks. First it must be noted that building a basis for Φ_k is a simple problem of linear algebra, that is, building a basis for the kernel of a linear operator. If one really needs it, such a basis can be built by a simple procedure based on Gaussian elimination. But one must also say that the degrees of freedom described above have mainly *a theoretical importance* for instance in building an interpolation operator for proving the

inf–sup condition. In practice, as we shall see in the applications of Chapter IV and V, any basis of $(P_k)^n$ will be convenient and standard degrees of freedom can be used.

Although the spaces described above may seem quite natural, they were not, by far, the first approximations introduced to approximate $H(\text{div};\Omega)$. Other possibilities exist and we shall see later how they are related to the previous one. Therefore, we now introduce the approximations of RAVIART–THOMAS [A]. We first consider the case of affine triangular or tetrahedral elements and we use the definition introduced by NEDELEC [A].

Let K be an n-simplicial (triangular or tetrahedral) element. Then we define

$$RT_k(K) = (P_k(K))^n + \underline{x}P_k(K). \tag{3.12}$$

It can easily be checked that the dimension of $RT_k(K)$ is given by

$$\dim RT_k(K) = \begin{cases} (k+1)(k+3) & \text{for } n = 2 \\ \frac{1}{2}(k+1)(k+2)(k+4) & \text{for } n = 3, \end{cases} \tag{3.13}$$

and that only the part of $\underline{x}P_k(K)$ involving homogeneous polynomials of degree k is important. We now prove some basic result about RT_k spaces. These spaces have indeed been tailor designed in order to satisfy the properties we now state in the following proposition.

Remark 3.1: The original work of RAVIART–THOMAS [A] used an expression equivalent to (3.12) on the *reference element* $\hat{K}$ and defined $RT_k(K)$ by the change of variable $\mathfrak{G}$ of (1.45). It must be noted that this definition is not equivalent to the definition of $RT_k(K)$ given above: it depends on the orientation of space. For triangular elements, definition (3.12) is more natural and easier to handle. □

Proposition 3.2: For any n-simplicial element K we have for $\underline{q} \in RT_k(K)$

$$\begin{cases} \text{div } \underline{q} \in P_k(K), \\ \underline{q} \cdot \underline{n}|_{\partial K} \in R_k(\partial K). \end{cases} \tag{3.14}$$

Moreover, the divergence operator is surjective from $RT_k(K)$ onto $P_k(K)$.

Proof: $\underline{q} \in RT_k(K)$ can be written $\underline{q} = \underline{q}_0 + \underline{x}p_k$ with $\underline{q}_0 \in (P_k(K))^n$. It is then clear that div $\underline{q}$ is a polynomial of degree k. This proves the results about div $\underline{q}$. On the other hand let $\underline{n} = \{n_1, n_2\}$ be the normal to a side (we consider the two-dimensional case for simplicity)

$$\underline{q} \cdot \underline{n} = \underline{q}_0 \cdot \underline{n} + p_k(x_1 n_1 + x_2 n_2).$$

But on a side we have $x_1 n_1 + x_2 n_2$ constant so that $\underline{q} \cdot \underline{n}$ is a polynomial of degree k. The same argument holds in $\mathbb{R}^3$ (or in $\mathbb{R}^n$). □

We also have,

Proposition 3.3: For $k \geq 0$, and for any $\underline{q} \in RT_k$, the following relations imply $\underline{q} = 0$:

$$\int_{\partial K} \underline{q} \cdot \underline{n}\, p_k \, ds = 0, \qquad \forall p_k \in R_k(\partial K), \tag{3.15}$$

$$\int_K \underline{q} \cdot \underline{p}_{k-1} \, dx = 0, \qquad \forall \underline{p}_{k-1} \in (P_{k-1}(K))^n. \tag{3.16}$$

This is a variant of Proposition 3.1 and the proof is left as an exercise. □

Let us now define

$$RT_k^0(K) = \{\underline{q} \in RT_k(K) \mid \operatorname{div} \underline{q} = 0\}. \tag{3.17}$$

From (3.12), we can easily deduce

Corollary 3.1: $RT_k^0(K) \subset (P_k(K))^n$. □

Therefore $RT_k(K)$ and $BDM_k(K)$ contain the same divergence-free vectors. We then have

Corollary 3.2: For $n = 2$, any $\underline{q}_0 \in RT_k^0(K)$ is the curl of a stream-function $\psi \in P_{k+1}(K)$. The dimension of $RT_k^0(K)$ is equal to $(\dim P_{k+1}(K) - 1) = \frac{1}{2}(k+1)(k+4)$. □

The above result has been extended by NEDELEC [A] to the three-dimensional case using

$$H(\underline{\operatorname{curl}}\,;\Omega) = \{\underline{\phi} \mid \underline{\phi} \in (L^2(\Omega))^n,\ \underline{\operatorname{curl}}\ \underline{\phi} \in (L^2(\Omega))^n\}. \tag{3.18}$$

In the three-dimensional space, $H(\underline{\operatorname{curl}}\,;\Omega)$ will need special approximations, whereas in the two-dimensional case, approximations can be built from RT_k or BDM_k by a simple rotation of $\pi/2$, as we shall see in Chapter VII.

Propositions 3.2 and 3.3 imply that one can use as degrees of freedom for RT_k:

— the moments of order up to k of $\underline{q} \cdot \underline{n}$ on the sides or faces of K;

— the moments of order up to $k-1$ of $\underline{q}$ on K.

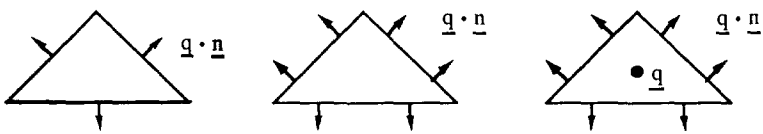
Before stating approximation results, we consider a few examples.

Example 3.1: *The spaces RT_0, RT_1, BDM_1*

From the results above, we know that RT_0 is a space of dimension 3 containing polynomials of the form

$$\begin{cases} q_1(x,y) = a + cx, \\ q_2(x,y) = b + cy. \end{cases} \tag{3.19}$$

We can specify it by the three normal components of $\underline{q}$ on ∂K as sketched in Figure III.10. Space BDM_1 is of dimension 6 and RT_1 is of dimension 8. It must be noted that div BDM_1 = div $RT_0 = P_0$. If one considers the subset of BDM_1 such that $\underline{q} \cdot \underline{n}|_{\partial K} \in R_0(\partial K)$, one easily sees that the resulting space is RT_0. In the same way, BDM_1 is the subset of RT_1 such that div $\underline{q} \in P_0$ instead of P_1.

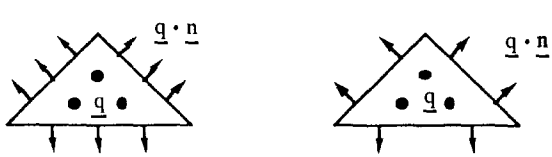


a) RT_0 b) BDM_1 c) RT_1

Figure III.10

Example 3.2: *Spaces BDM_2 and $BDFM_2$*

The space $BDM_2(K)$ is twelve dimensional. It is defined by nine boundary degrees of freedom and three internal ones (Figure III.11) derived from (3.3) through (3.5). It is then possible to restrict $\underline{q} \cdot \underline{n}$ to belong to $R_1(\partial K)$ instead of $R_2(\partial K)$. The reader may check that Proposition 3.1 would still be valid with the appropriate change of (3.3). This space is then closely related to the Raviart–Thomas space $RT_1(K)$, with, however, one more degree of freedom. We denote it $BDFM_2$, as it is the triangular analogue of the reduced element introduced in BREZZI–DOUGLAS–FORTIN–MARINI [A]. □



a) BDM_2 b) $BDFM_2$

Figure III.11

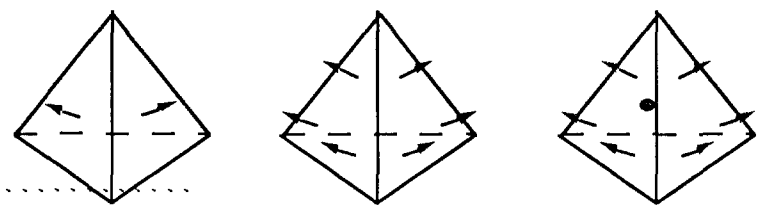
This last example was indeed quite general: restricting $\underline{q} \in BDM_k(K)$ to have a normal trace in $R_{k-1}(\partial K)$ yields a space larger than $RT_k(K)$, but having essentially the same properties, that we denote $BDFM_k(K)$. For the triangular case we thus have the following inclusions between the spaces just defined:

$$RT_0 \subset BDFM_1 \subset BDM_1 \subset RT_1 \subset BDFM_2 \subset BDM_2 \subset RT_2.$$

The spaces $BDM_k(K)$ are in a sense more natural as they use a full set of polynomials instead of (3.12). It must, however, be stated (cf. Chapter VII) that spaces $RT_k(K)$ are more suited to certain problems, specially in elasticity.

Example 3.3: *Three-dimensional elements:* RT_0, BDM_1, RT_1

The simplest cases of three-dimensional elements are depicted in Figure III.12. Their properties are exactly the same as in the two-dimensional case.



a) RT_0 b) BDM_1 c) RT_1

Figure III.12

III.3.2 Rectangular approximations of $H(\text{div}; K)$

We now consider the extension of the previous construction to rectangular elements and through the change of variables (1.45) to general quadrilaterals. In the present case, the use of a reference element is essential and *we shall build our spaces on* $\hat{K} =]-1,+1[^n$. Contrarily to the simplicial case, it will be simpler here to first introduce the approximations of Raviart and Thomas. The extension to the three-dimensional case is again due to NEDELEC [A].

Let us thus define, as in the previous section,

$$RT_{[k]} = (Q_k)^n + \underline{x}Q_k. \tag{3.20}$$

It can be checked that one has

$$RT_{[k]} = \begin{cases} P_{k+1,k} \times P_{k,k+1} & \text{for } n = 2 \\ P_{k+1,k,k} \times P_{k,k+1,k} \times P_{k,k,k+1} & \text{for } n = 3, \end{cases} \tag{3.21}$$

and that

$$\dim RT_{[k]} = \begin{cases} 2(k+1)(k+2) & \text{for } n = 2 \\ 3(k+1)^2(k+2) & \text{for } n = 3. \end{cases} \tag{3.22}$$

Moreover, these spaces have been defined in order to have

$$\text{div } \underline{q}_k \in Q_k \tag{3.23}$$

and

$$\begin{cases} \underline{q} \cdot \underline{n}|_{e_i} \in P_k(e_i) & \text{for } n = 2, \\ \underline{q} \cdot \underline{n}|_{e_i} \in Q_k(e_i) & \text{for } n = 3. \end{cases} \tag{3.24}$$

Defining, as in the simplicial case,

$$RT^0_{[k]} = \{\underline{q} \mid \underline{q} \in RT_{[k]},\ \text{div } \underline{q} = 0\}, \tag{3.25}$$

we have

Lemma 3.3: For n= 2, if $\hat{\underline{q}} \in RT^0_{[k]}(\hat{K})$, there exists $\psi \in Q_{k+1}(\hat{K})$ such that $\hat{\underline{q}} = \underline{\text{curl}}\ \psi$. The dimension of $RT^0_{[k]}(\hat{K})$ is $(k+1)(k+3)$. □

In order to choose an approximate set of degrees of freedom, we define

$$\Psi_k(K) = \begin{cases} P_{k-1,k}(K) \times P_{k,k-1}(K) & \text{for } n{=}2, \\ P_{k-1,k,k}(K) \times P_{k,k-1,k}(K) \times P_{k,k,k-1}(K) & \text{for } n{=}3. \end{cases} \tag{3.26}$$

We can now state

Proposition 3.4: For any $\hat{\underline{q}} \in RT_{[k]}(\hat{K})$, the relations

$$\int_{e_i} \phi_i\ \hat{\underline{q}} \cdot \hat{\underline{n}}\ d\hat{s} = 0, \quad \forall \phi_i \in Q_k(e_i) \text{ for } n = 3,\ \forall \phi_i \in P_k(e_i) \text{ for } n = 2, \tag{3.27}$$

$$\int_{\hat{K}} \hat{\underline{\phi}} \cdot \hat{\underline{q}}\ dx = 0, \qquad \forall \hat{\underline{\phi}} \in \Psi_k(\hat{K}) \tag{3.28}$$

imply $\hat{\underline{q}} = 0$.

For $n = 2$ the proof is analogous to the proof of Proposition 3.3. For $n = 3$ see NEDELEC [A]. Note that, for $n = 2$, the sides e_i are one-dimensional, so that actually $Q_k(e_i) = P_k(e_i)$ in (3.27). □

The $RT_{[k]}$ spaces just described are based on the idea that a finite element approximation of the rectangle should use a space of type Q_k. This is, however, by no means necessary in the present case. Let us define following BREZZI–DOUGLAS–MARINI [B] and BREZZI–DOUGLAS–DURAN–FORTIN [A], for $n = 2,\ k \geq 1$,

$$\begin{aligned} BDM_{[k]} = \{\underline{q} \mid \underline{q} = \underline{p}_k(x,y) + r\ \underline{\text{curl}}\ (x^{k+1}y) \\ + s\ \underline{\text{curl}}\ (xy^{k+1}),\ \underline{p}_k \in (P_k)^2\} \end{aligned} \tag{3.29}$$

and for $n = 3,\ k \geq 1$,

$$(3.30)\quad \begin{cases} BDM_{[k]} = \{\underline{q} \mid \underline{q} = \underline{p}_k(x,y,z) + \sum_{i=0}^{k}[r_i\ \underline{\mathrm{curl}}\ \{0,0,xy^{i+1}z^{k-i}\} \\ \qquad\qquad + s_i\ \underline{\mathrm{curl}}\ \{yz^{i+1}x^{k-i},0,0\} \\ \qquad\qquad + t_i\ \underline{\mathrm{curl}}\ \{0,zx^{i+1}y^{k-i},0\}];\ \underline{p}_k \in (P_k)^3\}. \end{cases}$$

Those spaces have been carefully defined in order to have

$$(3.31)\quad \begin{cases} \operatorname{div} \underline{q} \in P_{k-1}(K), \\ \underline{q}\cdot\underline{n}|_{e_i} \in P_k(e_i). \end{cases}$$

It must be remarked that these last conditions are rather unusual for a rectangular approximation. We have by a simple count,

$$(3.32)\quad \dim BDM_{[k]} = \begin{cases} (k+1)(k+2)+2 = k^2+3k+4 & \text{for } n=2 \\ \frac{(k+1)(k+2)(k+3)}{2} + 3(k+1) & \text{for } n=3. \end{cases}$$

For the choice of degrees of freedom, we have

Proposition 3.5: For $k \geq 1$, the following conditions imply $\underline{q} = 0$,

$$(3.33)\quad \int_{e_i} \underline{q}\cdot\underline{n}\, p_k\, d\sigma, \qquad \forall p_k \in P_k(e_i),$$

$$(3.34)\quad \int_K \underline{q}\cdot\underline{p}_{k-2}\, dx = 0, \qquad \forall \underline{p}_{k-2} \in (P_{k-2})^n.$$

We prove the proposition for $n = 3$. The case $n = 2$ is a simpler variant. It is sufficient to prove that (3.33) implies $\underline{q} \in (P_k)^n$, that is, all terms introduced through $\underline{\mathrm{curl}}$ vanish. Indeed, if $\underline{q} \in (P_k)^n$, then $\underline{q}\cdot\underline{n}|_{e_i} = 0$ implies $q_1 = (1-x^2)p_{k-2}$, $q_2 = (1-y^2)p_{k-2}$, $q_3 = (1-z^2)p_{k-2}$ and (3.34) eliminates $\underline{q}$. Let us consider the first component q_1. One has from (3.30)

$$(3.35)\quad \begin{aligned} q_1 =& p_k(x,y,z) + \sum_{i=0}^{k} r_i(i+1)xy^iz^{k-i} - \sum_{i=0}^{k} t_i x^{i+1}y^{k-i} \\ =& p_k(x,y,z) - \sum_{i=1}^{k} t_i x^2 x^{i-1}y^{k-i} + \sum_{i=0}^{k-1} r_i(i+1)xy^iz^{k-i} \\ &+ (r_k(k+1) - t_0)xy^k. \end{aligned}$$

In order to have $q_1 = 0$ for $x = \pm 1$, all terms that are for fixed x homogeneous polynomials of degree k in y, z must vanish. This implies

$$r_i = 0, \qquad 0 \leq i \leq k-1,$$
$$r_k(k+1) - t_0 = 0.$$

In the same way one has $s_i = 0$, $(0 \leq i \leq k-1)$, $s_k(k+1) - r_0 = 0$ from component q_2 considered at $y = \pm 1$ and $t_i = 0, (0 \leq i \leq k-1)$, $t_k(k+1) - s_0 = 0$ from q_3 considered at $z = \pm 1$. This completes the proof. □

Remark 3.2: Definitions (3.29) and (3.30) have been designed in order to keep $\operatorname{div} \underline{q}$ in P_{k-1} by adding divergence-free functions to $(P_k)^n$ while providing terms with a normal component in $P_k(e_i)$ on each side or face e_i. In the three-dimensional case, *there is no unique way* to give such a definition. For instance, one could have used, instead of (3.30),

$$
\begin{aligned}
BDM_{[k]} = \Big\{ \underline{q} \mid \underline{q} = p_k(x,y,z) &+ \sum_{i=0}^{k} r_i \ \underline{\operatorname{curl}} \ \{0,0,yx^{i+1}z^{k-i}\} \\
&+ \sum_{i=0}^{k} s_i \ \underline{\operatorname{curl}} \ \{zy^{i+1}x^{k-i},0,0\} \\
&+ \sum_{i=0}^{k} t_i \ \underline{\operatorname{curl}} \ \{0,xz^{i+1}y^{k-i},0\} \Big\}.
\end{aligned} \tag{3.36}
$$

We shall also show below on an example how other choices can be made. □

We would now like to see what are the relations between $BDM_{[k]}(K)$ and $RT_{[k]}(K)$. For the sake of simplicity, we restrict ourselves to the case $n = 2$ even though the result can easily be extended. First, one obviously has $BDM_{[k]} \subset RT_{[k]}$. However, the space obtained by restricting the normal component of $BDM_{[k]}$ to belong to $P_{k-1}(e_i)$ on each side has no direct relation to $RT_{[k-1]}$ and is a much smaller space (providing an approximation of the same accuracy). In order to get a pattern of inclusions, we define the space

$$
\begin{aligned}
S_{[k+1]} = RT_{[k]} + \operatorname{Span}\{&\underline{\operatorname{curl}} \ x^{k+2}y, \underline{\operatorname{curl}} \ yx^{k+2}, \\
&\underline{\operatorname{curl}} \ x^{k+2}, \underline{\operatorname{curl}} \ y^{k+2}\}.
\end{aligned} \tag{3.37}
$$

This space obviously contains $RT_{[k]}$ but also contains $BDM_{[k+1]}$.

We can also define the space $BDFM_{[k+1]}$ be restricting the normal component of $\underline{q} \in BDM_{[k+1]}$ to belong to $P_k(e_i)$ instead of $P_{k+1}(e_i)$ on each side (BREZZI–DOUGLAS–FORTIN–MARINI [A]).

It can easily be checked that $BDFM_{[1]} = RT_{[0]}$. To make things clear, let us consider a few diagrams (Figure III.13)

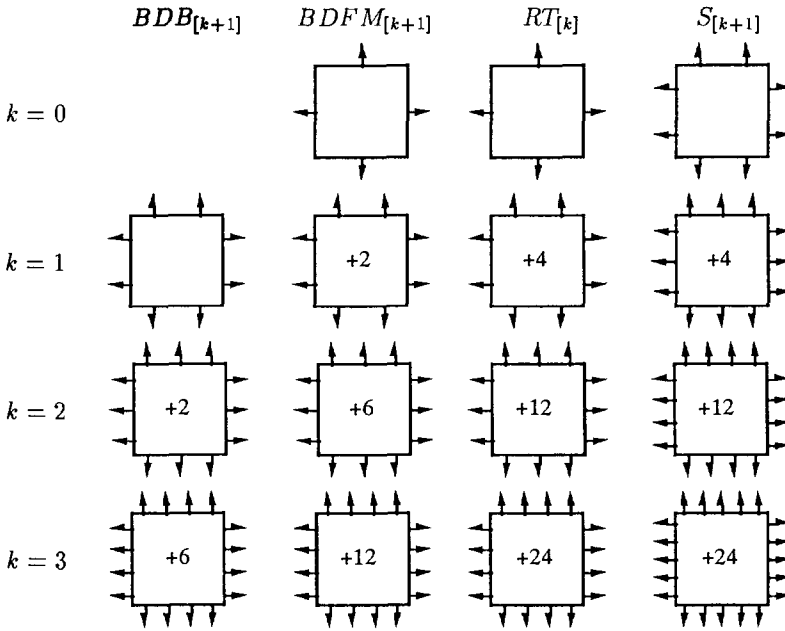


Figure III.13

We can then summarize the previous facts in Figure III.14 in which arrows indexed by b represent a reduction in boundary degrees of freedom and arrows indexed by i represent a reduction in internal degrees of freedom

$$\begin{tikzcd}
S_{[k+1]} \arrow[r, "b"] \arrow[d, "i"'] & RT_{[k]} \arrow[r, "i"] \arrow[d, "i"'] & S_{[k]} \arrow[r, "b"] \arrow[d, "i"'] & RT_{[k-1]} \arrow[d, "i"'] \\
BDM_{[k+1]} \arrow[r, "b"] & BDFM_{[k+1]} \arrow[r, "i"] & BDM_{[k]} \arrow[r, "b"] & BDFM_{[k]}
\end{tikzcd}$$

Figure III.14

Space $RT_{[0]}$ plays a special role in this set of spaces. It is the simplest possible space and it is related to the MAC scheme (HARLOW–WELSCH [A]) that has been extensively used in fluid mechanical computations. It is clear from Figure III.14 that both $RT_{[k]}$ and $BDFM_{[k+1]}$ are a generalization of this space with the same order of accuracy. One uses $RT_{[k]}$ whenever one wants div $\underline{q} \in Q_k$ and $BDFM_{[k+1]}$ if div $\underline{q} \in P_k$ is sufficient. It is thus worth considering $BDFM_{[k+1]}$ in more details. It is easy to check that one has in the two-dimensional case

$$BDFM_{[k+1]} = (P_{k+1})^2 \backslash \{0, x^{k+1}\} \backslash \{y^{k+1}, 0\}. \tag{3.38}$$

This shows that it is natural to move to $BDM_{[k+1]}$ and get an extra order of accuracy whenever one is ready to pay for extra boundary nodes.

To make our presentation complete we now consider a few three-dimensional elements.

Example 3.4: *Spaces $BDM_{[1]}$ and $BDFM_{[2]}$ for $n = 3$*

Space $BDM_{[1]}$ has 18 degrees of freedom. They are the moments of degree 1 on each face. Any function of $BDM_{[1]}$ can be written in the form (using (3.30))

$$q = \begin{cases} a_1 + b_1 x + c_1 y + d_1 z + r_0 zx + 2r_1 xy - t_0 xy - t_1 x^2, \\ a_2 + b_2 x + c_2 y + d_2 z - r_0 yz - r_1 y^2 + s_0 yx + 2s_1 yz, \\ a_3 + b_3 x + c_3 y + d_3 z - s_0 zx - s_1 z^2 + t_0 zy + 2t_1 zx. \end{cases} \tag{3.39}$$

The last terms have been generated (from (3.30)) by taking the curl of six vectors:

$$\{0,0,xyz\},\ \{0,0,xy^2\},\ \{yxz,0,0\},\ \{yz^2,0,0\}\ \{0,zxy,0\},\ \{0,zx^2,0\}.$$

A space with similar properties could be generated by taking the curl of

$$\{0,0,xyz^2\},\ \{0,0,x^2y\},\ \{xyz,0,0\},\ \{y^2z,0,0\},\ \{0,xyz,0\},\ \{0,xz^2,0\}$$

as in (3.36), or even the more symmetrical form

$$\{0,0,xy^2\},\ \{0,0,x^2y\},\ \{z^2y,0,0\},\ \{y^2z,0,0\},\ \{0,xz^2,0\},\ \{0,x^2z,0\}.$$

The last case can be written in a general form; however, it is more cumbersome than (3.30).

Space $BDM_{[2]}$ has 39 degrees of freedom. Boundary nodes account for 36 of them and 3 internal ones remain. The space $BDFM_{[2]}$ obtained by restricting the normal components to $P_1(e_i)$ on each face thus has 21 degrees of freedom, 18 of them on the boundary. By comparison, the space RT_1 has, for the same order of accuracy, 36 degrees of freedom, 24 of them on the boundary. Again, one has for $n = 3$, $BDFM_{[1]} = RT_{[0]}$. □

III.3.3 Interpolation operator and error estimates

Let now $\underline{q}$ be some function of $H(\text{div}; K)$. Using for each of the spaces the degrees of freedom previously described, it is possible to define an interpolation operator $\rho_K \underline{q}$, *provided $\underline{q}$ is slightly smoother than merely belonging* to $H(\text{div}; K)$. Indeed the degrees of freedom used always imply the moments of q on the faces (or sides) of an element. But the functions $p_k \in R_k(\partial K)$ do not

belong to $H^{1/2}(\partial K)$, and it is not possible in general to compute expressions like $\int_{\partial K} \underline{q} \cdot \underline{n}\, p_k\, ds$ as $\underline{q} \cdot \underline{n}$ is only defined in $H^{-1/2}(\partial K)$.

However it is easy to check that if $\underline{q}$ belongs to the space (3.40)

$$W(K) = \{\underline{q} \in (L^s(K))^n \mid \operatorname{div} \underline{q} \in L^2 \in (\Omega)\} \tag{3.40}$$

(for s fixed > 2), then such a construction is possible.

For the convenience of the reader we shall summarize now all the spaces introduced in this section, and, for each of them, we shall define the corresponding operator ρ_K that we will always assume to be defined in $W(K)$.

Case $n = 2$, triangular elements

(i) $BDM_k(K) = (P_k(K))^2$, $(k \geq 1)$.

$\rho_K\colon W(K) \to BDM_k(K)$ is defined by

$$\begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, \ \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\operatorname{grad}}\, p_{k-1}\, dx = 0, \ \forall p_{k-1} \in P_{k-1}(K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\operatorname{curl}}\, (b_K p_{k-2}) dx = 0, \ \forall p_{k-2} \in P_{k-2}(K)\ (k \geq 2). \end{cases} \tag{3.41}$$

(ii) $BDFM_k(K) = \{\underline{q} \in (P_k(K))^2 \mid \underline{q} \cdot \underline{n}|_{\partial K} \in R_{k-1}(\partial K)\}$, $(k \geq 1)$

$\rho_K\colon W(K) \to BDFM_k(K)$ is defined by

$$\begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_{k-1}\, ds = 0, \ \forall p_{k-1} \in R_{k-1}(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\operatorname{grad}}\, p_{k-1}\, dx = 0, \ \forall p_{k-1} \in P_{k-1}(K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\operatorname{curl}}\, (b_K p_{k-2}) dx = 0, \ \forall p_{k-2} \in P_{k-2}(K)\ (k \geq 2). \end{cases} \tag{3.42}$$

(iii) $RT_k(K) = (P_k(K))^2 \oplus \underline{x} P_k(K)$, $(k \geq 0)$.

$\rho_K\colon W(K) \to RT_k(K)$ is defined by

$$\begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, \quad \forall p_k \in R_k(K), \\ \displaystyle\int_k (\underline{q} - \rho_K \underline{q}) \cdot \underline{p}_{k-1}\, dx = 0, \quad \forall \underline{p}_{k-1} \in (P_{k-1}(K))^2\ (k \geq 1). \end{cases} \tag{3.43}$$

Case $n = 2$, $K =$ unit square

(i) $BDM_{[k]}(K) = (P_k(K))^2 \oplus \underline{\mathrm{curl}}\,(x^{k+1}y) \oplus \underline{\mathrm{curl}}\,(xy^{k+1})$ $(k \geq 1)$.

$\rho_K : W(K) \to BDM_{[k]}(K)$ is defined by

$$(3.44)\quad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{p}_{k-2}\, dx = 0, & \forall p_k \in (P_{k-2}(K))^2 \ \ (k \geq 2). \end{cases}$$

$$(ii)\quad \begin{cases} BDFM_{[k]}(K) & = \{\underline{q} \in BDM_{[k]}(K),\ \underline{q} \cdot \underline{n}|_{\partial K} \in R_{k-1}(\partial K)\} \\ & = (P_k(K) \backslash \{y^k\}) \times (P_k(K) \backslash \{x^k\})\ (k \geq 1). \end{cases}$$

$\rho_K \cdot W(K) \to BDFM_{[k]}(K)$ is defined by

$$(3.45)\quad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q})\ \ \underline{n}\, p_{k-1}\, ds = 0, & \forall p_{k-1} \in R_{k-1}(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{p}_{k-2}\, dx = 0, & \forall \underline{p}_{k-2} \in (P_{k-2}(K))^2\ (k \geq 2). \end{cases}$$

(iii) $RT_{[k]}(K) = P_{k+1,k}(K) \times P_{k,k+1}(K)$, $(k \geq 0)$.

$\rho_K \quad W(K) \to RT_{[k]}(K)$ is defined by

$$(3.46)\quad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\phi}_k\, dx = 0, & \forall \underline{\phi}_k \in P_{k-1,k}(K) \times P_{k,k-1}(K). \end{cases}$$

Before discussing the cases $n = 3$ we recall some additional notation. For $K =$ tetrahedron, we set

$$(3.47)\qquad \Phi_k(K) = \{\underline{\phi} \in (P_k(K))^3 \mid \mathrm{div}\, \underline{\phi} = 0,\ \underline{\phi} \cdot \underline{n}|_{\partial K} = 0\}.$$

For $K =$ cube, we set

$$(3.48)\qquad \Psi_k(K) = P_{k-1,k,k,}(K) \times P_{k,k-1,k}(K) \times P_{k,k,k-1}(K),$$

$$(3.49)\qquad \mathrm{Hom}_k(\xi, \eta) = \bigoplus_{i=0}^{k} \xi^i\ \eta^{k-i},$$

$$(3.50)\qquad R_{[k]}(\partial K) = \{v \in L^2(\partial K),\ v|_{e_i} \in Q_k(e_i),\ 1 \leq i \leq 6\}.$$

Case $n = 3$, $K =$ tetrahedra

(i) $BDM_k(K) = (P_k(K))^3$, $(k \geq 1)$.

$\rho_K : W(K) \to BDM_k(K)$ is defined by

$$(3.51)\qquad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\mathrm{grad}}\, \underline{p}_{k-1}\, dx = 0, & \forall p_{k-1} \in P_{k-1}(K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\phi}_k\, dx = 0, & \forall \underline{\phi}_k \in \underline{\Phi}_k(K), \end{cases}$$

(ii) $BDFM_k(K) = \{\underline{q} \in (P_k(K))^3,\ \underline{q} \cdot \underline{n}|_{\partial K} \in R_{k-1}(\partial K)\}$ $(k \geq 1)$

$\rho_K \quad W(K) \to BDFM_k(K)$ is defined by

$$(3.52)\qquad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_{k-1}\, ds = 0, & \forall p_{k-1} \in R_{k-1}(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\mathrm{grad}}\, p_{k-1}\, dx = 0, & \forall p_{k-1} \in P_{k-1}(K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\phi}_k\, dx = 0, & \forall \underline{\phi}_k \in \Phi_k(K), \end{cases}$$

(iii) $RT_k(K) = (P_k(K))^3 \oplus \underline{x} P_k(K)$ $(k \geq 0)$.

$\rho_K : W(K) \to RT_k(K)$ is defined by

$$(3.53)\qquad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{p}_k\, dx = 0, & \forall \underline{p}_{k-1} \in (P_{k-1}(K))^3. \end{cases}$$

Case $n = 3$, $K =$ unit cube

(i) $$BDM_{[k]}(K) = (P_k(K))^3 \bigoplus_{i=0}^{k} \underline{\mathrm{curl}}\ (0, 0, xy^{i+1}z^{k-i}) \bigoplus_{i=0}^{k} \underline{\mathrm{curl}}\ (0, x^{k-i}yz^{i+1}, 0) \bigoplus_{i=0}^{k} \underline{\mathrm{curl}}\ (x^{i+1}y^{k-i}z, 0, 0)\ \ (k \geq 1).$$

$\rho_K : W(K) \to BDM_{[k]}(K)$ is defined by

$$(3.54)\qquad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_k(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{p}_{k-2}\, dx = 0, & \forall \underline{p}_{k-2} \in (P_{k-2}(K))^2 \ \ (k \geq 2). \end{cases}$$

(ii) $$BDFM_{[k]}(K) = \{\underline{q} \in BDM_{[k]}(K),\ \underline{q}\cdot\underline{n}|_{\partial K} \in R_{k-1}(\partial K)\}$$
$$= (P_k \setminus \mathrm{Hom}_k(y,z)) \times (P_k \setminus \mathrm{Hom}_k(x,z)) \times (P_k \setminus \mathrm{Hom}_k(x,y)),\ (k \geq 1).$$

$\rho_K : W(K) \to BDFM_{[k]}(K)$ is defined by

$$(3.55)\quad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q})\cdot \underline{n}\, p_{k-1}\, ds = 0, & \forall p_{k-1} \in R_{k-1}(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q})\cdot \underline{p}_{k-2}\, dx = 0, & \forall \underline{p}_{k-2} \in (P_{k-2}(K))^2,\ (k \geq 2). \end{cases}$$

(iii) $RT_{[k]}(K) = P_{k+1,k,k,}(K) \times P_{k,k+1,k}(K) \times P_{k,k,k+1}(K)\ \ (k \geq 0).$

$\rho_K : W(K) \to RT_{[k]}(K)$ is defined by

$$(3.56)\quad \begin{cases} \displaystyle\int_{\partial K} (\underline{q} - \rho_K \underline{q})\cdot \underline{n}\, p_k\, ds = 0, & \forall p_k \in R_{[k]}(\partial K), \\ \displaystyle\int_K (\underline{q} - \rho_K \underline{q})\cdot \underline{\phi}_k\, dx = 0, & \forall \underline{\phi}_k \in \Psi_k(K). \end{cases}$$

Note that for rectangular elements we used the unit square for K (or the unit cube for $n = 3$). For a general K, the spaces and the interpolation operators ρ_K have to be modified by means of the contravariant mapping $\mathfrak{G}$ of (1.45). In particular, $\rho_K \underline{q} = \mathfrak{G}(\rho_{\hat{K}} \hat{\underline{q}})$, where $\hat{\underline{q}} = \mathfrak{G}^{-1}(\underline{q})$ and $\hat{K}$ is the unit square or the unit cube. As we have seen, everything works in the case of affine elements whereas some complications may arise for general elements.

In the following, whenever it may be convenient, we will denote by the symbol $M(K)$ anyone of the above approximations of $H(\mathrm{div}; K)$. Since, as we shall see, the accuracy of these approximations in the L^2-norm is especially relevant, we decided to indicate by $M_k(K)$ anyone of the above spaces such that $(P_k(K))^n \subseteq M_k(K)$ but $(P_{k+1}(K))^n \not\subseteq M_k(K)$. Hence, in the following, $M_k(K)$ will denote anyone of the following spaces: $BDM_k(K)$, $BDM_{[k]}(K)$, $RT_k(K)$, $RT_{[k]}(K)$, $BDFM_{k+1}(K)$, $BDFM_{[k+1]}(K)$.

Using Lemmas 1.6 and 1.7 and usual techniques (CIARLET [A]) we have immediately the following result.

Proposition 3.6: Let K be an affine element and ρ_K be the interpolation operator $W(K) \to M_k(K)$. There exists a constant c depending only on k and on the shape of K, such that, for $1 \leq m \leq k+1$, for $s = 0$ or 1 and for any $\underline{q}$ in $(H^m(K))^n$, we have

$$(3.57)\qquad \|\underline{q} - \rho_K \underline{q}\|_{s,K} \leq c h_K^{m-s}\, |\underline{q}|_{m,K}. \ \square$$

We now want to analyze the behavior of the error in $H(\text{div}; K)$. For this we need to characterize the space of the divergences of the vectors in $M_k(K)$. Let

$$D_k(K) := \text{div}(M_k(K)). \tag{3.58}$$

A simple analysis shows that, for affine elements, we have

$$\begin{aligned}
\text{div}(BDM_k(K)) &= \text{div}(BDM_{[k]}(K)) = P_{k-1}(K),\\
\text{div}(BDFM_{k+1}(K)) &= \text{div}(BDFM_{[k+1]}(K)) = P_k(K),\\
\text{div}(RT_k(K)) &= P_k(K),\\
\text{div}(RT_{[k]}(K)) &= \mathfrak{F}(Q_k(K)),
\end{aligned}$$

with $\mathfrak{F}$ defined in (1.35). (Note that Q_k is not invariant under affine transformations.) The following result is of paramount importance in the study of these approximations.

Proposition 3.7: Let K be an affine element and ρ_K the interpolation operator: $W(K) \to M_k(K)$. Let moreover π_K be the L^2-projection on $D_k(K) = \text{div}(M_k(K))$. Then we have, for all $\underline{q} \in W(K)$,

$$\text{div}(\rho_K \underline{q}) = \pi_K \text{ div } \underline{q}. \tag{3.59}$$

Proof: Since $\text{div}_K \underline{q} \in D_k(K)$ by definition, we only have to prove that

$$\int_K v \text{ div}(\rho_K \underline{q})\, dx = \int_K v \text{ div } \underline{q}\, dx, \qquad \forall v \in D_k(K). \tag{3.60}$$

Indeed

$$\begin{aligned}
\int_K v(\text{div } \rho_K \underline{q} - \text{div } \underline{q})\, dx &= \int_K (\underline{q} - \rho_K \underline{q}) \cdot \underline{\text{grad}}\, v\, dx\\
&\quad - \int_{\partial K} (\underline{q} - \rho_K \underline{q}) \cdot \underline{n}\, v\, ds,
\end{aligned} \tag{3.61}$$

and it is easy to check that, for all the possible choices of ρ_K , the right-hand side of (3.61) vanishes. □

Remark 3.3: The statement of Proposition 3.7 can also be expressed as

$$\begin{array}{ccc} W(K) & \xrightarrow{\text{div}} & L^2(K) \\ \rho_K \big\downarrow & & \pi_K \big\downarrow \\ M_k(K) & \xrightarrow{\text{div}} & D_k(K) \end{array} \tag{3.62}$$

and is often called the "commuting diagram property" (DOUGLAS–ROBERTS [A,B]). □

From Proposition 3.7, using Lemmas 1.6 and 1.7 and usual techniques, we easily have the following result.

Proposition 3.8: Let K be an affine element and ρ_K the interpolation operator: $W(K) \to M_k(K)$. There exists a constant c depending only on k and on the shape of K such that for $1 \le m \le \phi_M(k)$ we have

$$\|\text{div}(\underline{q} - \rho_K \underline{q})\|_{0,K} \le c h_K^m \, |\text{div } \underline{q}|_{m,K}, \tag{3.63}$$

where $\phi_M(k) = k$ for $BDM_k(K)$ or $BDM_{[k]}$ and $\phi_M(k) = k+1$ for the other choices. □

Remark 3.4: Proposition 3.8 shows that choosing RT_k, $RT_{[k]}$, $BDFM_{k+1}$ or $BDFM_{[k+1]}$ leads to the same accuracy in $H(\text{div}; K)$ as we have in $(L^2(K))^n$. This is not the case for BDM_k or $BDM_{[k]}$ where the accuracy in $(L^2(K))^n$ is of one order larger than the accuracy in $H(\text{div}; K)$. However, as we shall see in the next chapter, the commuting diagram property is so strong that this drawback can be circumvented. □

Remark 3.5: For nonaffine elements the situation is more complicated. In particular, we now have to define $D_k(K)$ and $\mathfrak{F}(D_k(\hat{K}))$, where $\hat{K}$ is the reference element and $\mathfrak{F}$ is defined in (1.35). On the other hand, $\text{div}(M_k(K))$ will be $\mathfrak{F}(J^{-1} \text{ div } M_k(\hat{K}))$. Hence, it is clear that Proposition 3.7 will not hold anymore. However, Lemma 3.5 will still have important consequences.

For instance, if $\underline{q} \in W(K)$, then

$$\pi_K \text{div } \underline{q} = 0 \Rightarrow \text{div } \underline{q} = 0. \tag{3.64}$$

Moreover, for any $\underline{q} \in W(K)$,

$$\text{div } \underline{q} \Rightarrow \text{div } \rho_K \underline{q} = 0. \tag{3.65}$$

On the other hand, Proposition 3.6 still holds (at least for RT elements) in the weaker form

$$\|\underline{q} - \rho_K \underline{q}\|_{s,K} \le c h_K^{m-s} \, (|\underline{q}|_{m,K} + h_K \, |\text{div } \underline{q}|_{m,K}) \tag{3.66}$$

(for $s = 0, 1$ and $1 \leq m \leq k+1$, see THOMAS [B]). This result is not optimal: a better one can be found in GIRAULT–RAVIART [A] for the case $m = 1$, in which the term $\operatorname{div}\underline{q}$ in the rigt-hand side does not appear. It is not known whether this better result can be obtained in the general case. Finally Proposition 3.8 does not hold (at least for RT-elements; see again THOMAS [B]). □

III.3.4 Approximation spaces for $H(\operatorname{div};\Omega)$

It is clear that the spaces defined in the previous sections can be used to define internal approximations of $H(\operatorname{div};\Omega)$. The choice of degrees of freedom has obviously been done in order to ensure continuity of $\underline{q} \cdot \underline{n}$ at interfaces of elements. We can then define, for each choice of $M_k(K)$, a space

$$M_k(\Omega, \mathcal{T}_h) = \{\underline{q} \in H(\operatorname{div};\Omega),\ \underline{q}|_K \in M_k(K)\}. \tag{3.67}$$

In a similar manner we have, in agreement with the notation (2.18),

$$\mathcal{L}^0(D_k, \mathcal{T}_h) = \{v \in L^2(\Omega),\ v|_K \in D_k(K)\}. \tag{3.68}$$

It is clear that for affine elements

$$\operatorname{div} M_k(\Omega, \mathcal{T}_h) \subset \mathcal{L}^0(D_k, \mathcal{T}_h). \tag{3.69}$$

Moreover, we can now define a *global* interpolation operator from

$$W = H(\operatorname{div};\Omega) \cap (L^s(\Omega))^n \tag{3.70}$$

(s fixed > 2) into $M_k(\Omega; \mathcal{T}_h)$ by simply setting

$$(\Pi_h \underline{q})_K = \rho_K(\underline{q}|_K). \tag{3.71}$$

By defining $P_h :=$ projection on $\mathcal{L}^0(D_k, \mathcal{T}_h)$ we have the following commuting diagram

$$\begin{array}{ccc} W & \xrightarrow{\operatorname{div}} & L^2(\Omega) \\ \Pi_h \big\downarrow & & P_h \big\downarrow \\ M_k(\Omega, \mathcal{T}_h) & \xrightarrow{\operatorname{div}} & \mathcal{L}^0(D_k, \mathcal{T}_h) \end{array} \tag{3.72}$$

This will imply in particular that

$$\operatorname{div} M_k(\Omega; \mathcal{T}_h) = \mathcal{L}^0(D_k, \mathcal{T}_h). \tag{3.73}$$

Finally, we have from Propositions 3.6 and 3.8 the following estimates for the interpolation operator Π_h.

Proposition 3.9: Let $\mathcal{T}_h$ be a regular family of decompositions of Ω, and let Π_h be defined as in (3.71). There exists a constant c independent of h such that

$$\|\underline{q} - \Pi_h \underline{q}\|_{0,\Omega} \leq ch^m \, |q|_{m,\Omega} \tag{3.74}$$

for $1 \leq m \leq k+1$. Moreover,

$$\|\mathrm{div}(\underline{q} - \Pi_h \underline{q})\|_{0,\Omega} \leq ch^s \, |\mathrm{div}\, \underline{q}|_{s,\Omega}, \tag{3.75}$$

where $s \leq k$ for BDM_k or $BDM_{[k]}$ and $s \leq k+1$ for the other choices of M_k. □

III.4 Concluding Remarks

This chapter is evidently not a complete presentation of finite element approximation methods. It cannot be, unless it becomes a book by itself. Our aim was therefore to present examples of the most classical cases and to consider a construction for the less standard case $H(\mathrm{div};\Omega)$. Other cases have been considered; for instance NEDELEC [A,B] developed approximations of the spaces $H(\mathrm{curl};\Omega)$. On the other hand, approximations of elasticity problems by the Hermann–Johnson technique will also require special spaces. They will be described in due time. We, however, believe that the present chapter will then provide a sound basis for these developments.

IV

Various Examples

This chapter will *rapidly* present various applications of the theory developed in Chapter II. It will give the reader a general idea of the possibilities offered by this theoretical framework. Many of our examples have already been considered in Chapter I. We shall consider here existence and uniqueness proofs, when they can be obtained, in a proper functional setting. Moreover, we shall give examples of discretizations and error estimates. Some of the problems considered here will be presented in a more detailed treatment in future chapters: this will be the place where special cases and exceptions will eventually be discussed; the present analysis is, in principle, restricted to simple and straightforward cases. We shall, therefore, successively consider non-standard methods for Dirichlet's problem, including hybrid methods. We shall then present approximations of the Stokes problem and of the linear elasticity problems. Fourth-order problems will also be considered either by mixed methods such as the ψ–ω method (CIARLET–RAVIART [C], MERCIER [A]) or à la MIYOSHI [A] or by dual hybrid methods. This list of examples is obviously not exhaustive and many applications have not been treated, in particular, equilibrium methods for which we refer to BREZZI–MARINI–QUARTERONI–RAVIART [A], HLAVACEK [A], HASLINGER–HLAVACEK [A]-[B] and BATOZ–BATHE–HO [A]. Other examples can be found in ROBERTS–THOMAS [A] and the references therein. Other applications and variants of the methods presented can also be found in BATOZ–BATHE–HO [A], KIKUCHI [A], and QUARTERONI [A,B], RANNACHER [A], and SCAPOLLA [A] for fourth-order problems. Time-dependent problems have been treated in QUARTERONI [C] and with a quite difffrent methodology in HUGHES–HULBERT [A]. Finally, let us point out the contribution (e.g., WHEELER–GONZALEZ [A]) of many people working on reservoir modeling to mixed methods.

IV.1 Nonstandard Methods for Dirichlet's Problem

IV.1.1 Description of the problem

This section presents a unified framework for the analysis of nonstandard methods for problems involving an elliptic, Laplacian-like equation in $\mathbb{R}^n$. Although we shall mainly consider the case $n = 2$, most results can be extended to the case $n = 3$ using the construction developed in Chapter III. We thus consider a problem of the following type:

$$(1.1)\qquad \begin{cases} -\operatorname{div} A(x)\,\underline{\operatorname{grad}}\, u = f \text{ in } \Omega, \\ u|_{\gamma_0} = g_1 \text{ on } D, \\ A(x)\,\underline{\operatorname{grad}}\, u \cdot \underline{n} = g_2 \text{ on } N, \end{cases}$$

where Ω is a bounded domain in $\mathbb{R}^n$ and $\Gamma = D \cup N = \partial\Omega$. We assume $A(x)$ to be an $n \times n$ positive definite matrix and that its smallest eigenvalue is bounded away from zero, uniformly with respect to x, that is,

$$(1.2)\qquad \langle A(x)\underline{q}, \underline{q}\rangle \geq \alpha |\underline{q}|^2_{\mathbb{R}^n}, \qquad \forall \underline{q} \in \mathbb{R}^n,$$

with α independent of x. We have already introduced this problem in Chapter I with $A(x) = I$. Restricting ourselves temporarily to the case $g_1 = 0$, the standard variational formulation is the following minimization problem (when A(x) is symmetric)

$$(1.3)\qquad \inf_{v \in H^1_{0,D}(\Omega)} \frac{1}{2}\int_\Omega A\,\underline{\operatorname{grad}}\, v . \underline{\operatorname{grad}}\, v\, dx - \int_\Omega f v\, dx - \int_N g_2 v\, ds,$$

where (cf. Chapter III)

$$(1.4)\qquad H^1_{0,D}(\Omega) = \{v \mid v \in H^1(\Omega),\ v|_D = 0\}.$$

It is classical that there exists a unique solution to this problem. We shall call problem (1.3) the *Primal Formulation.*

Using duality methods, we also transformed, in Chapter I, this problem to get a *Mixed Formulation*, namely for $f \in L^2(\Omega)$ and $g_2 = 0$,

$$(1.5)\qquad \inf_{\underline{q} \in H_{0,N}(\operatorname{div};\Omega)} \sup_{v \in L^2(\Omega)} \frac{1}{2}\int_\Omega A^{-1}\underline{q}.\underline{q}\, dx + \int_\Omega (\operatorname{div} \underline{q} + f)\, v\, dx + \int_D g_1\, \underline{q} \cdot \underline{n}\, ds,$$

where one has (cf. Section III.1.1)

$$(1.6)\qquad H_{0,N}(\operatorname{div};\Omega) = \{\underline{q} \mid \underline{q} \in H(\operatorname{div};\Omega),\ \underline{q} \cdot \underline{n}|_N = 0\},$$

the sense of $\underline{q} \cdot \underline{n}|_N = 0$ being defined as in Section III.1.1. Later we shall come back to the nonhomogeneous case $\underline{q} \cdot \underline{n} = g_2 \neq 0$. Problem (1.5) is equivalent to the *Dual Formulation*

$$\inf_{\substack{\underline{q} \in H_{0,N}(\mathrm{div};\Omega) \\ \mathrm{div}\, \underline{q} + f = 0}} \int_\Omega A^{-1} \underline{q}.\underline{q}\, dx + \int_D g_1\, \underline{q} \cdot \underline{n}\, ds. \tag{1.7}$$

Problem (1.7) is a constrained problem (in the sense of mathematical programming). The Mixed Formulation uses the Lagrange multiplier v to deal with the linear constraint $\mathrm{div}\, \underline{q} + f = 0$.

It must be remarked that problem (1.7) is not, strictly speaking, the dual of problem (1.3). That dual problem should be written with $\underline{q} \in L^2(\Omega))^n$ and $f \in H^{-1}(\Omega)$. Here we use a modified form using a stronger space for $\underline{q}$ while the regularity of v has been weakened. It must also be said that the approximation of this problem is not the main interest. The reason for such a detailed study is that it provides a simple framework that will later be generalized to other important problems.

This section will thus be entirely devoted to the study of problems (1.3), (1.5), and (1.7). We shall first consider approximations of problem (1.5), that is, *mixed finite element methods*. To work out such an approximation we shall have to use the finite element spaces approximating $H(\mathrm{div};\Omega)$ built in Chapter III.

To approximate the *dual problem*, we shall need to build vector functions $\underline{q}$ satisfying the condition

$$\mathrm{div}\, \underline{q} + f = 0. \tag{1.8}$$

This condition is the analogue of the *equilibrium condition* in elasticity theory, and approximations satisfying it will be called *equilibrium methods*. Finally *domain decomposition methods* will lead us to *hybrid finite element methods*. Hybrid methods will be called primal or dual, depending on the formulation being used. This distinction corresponds to assumed stress or assumed displacement hybrid methods in elasticity theory. Our analysis will rely directly on the properties of $H^1(\Omega)$ and $H(\mathrm{div};\Omega)$ and of their approximations considered in Chapter III.

IV.1.2 Mixed finite element methods for Dirichlet's problem

We are now able to consider in details the approximation of the mixed formulation

$$\inf_{\underline{q}} \sup_{v} \tfrac{1}{2} \int A^{-1}\, \underline{q}.\underline{q}\, dx + \int_\Omega (\mathrm{div}\, \underline{q} + f) v\, dx + \int_D g_1\, \underline{q} \cdot \underline{n}\, ds, \tag{1.9}$$

with $\underline{q} \in H_{0,N}(\text{div};\Omega)$ and $v \in L^2(\Omega)$. We can now see, by the results of Section IV.1.1 that the last term of (1.9) makes sense if $g_1 \in H^{1/2}(D)$ and that the boundary integral must be read as a formal way of writing the duality between $H^{1/2}$ and $H^{-1/2}$. Problem (1.9) is a saddle point problem. With the notation of Chapter II, we have

$$a(\underline{p},\underline{q}) = \int_\Omega A^{-1}\underline{p}.\underline{q}\,dx \tag{1.10}$$

and

$$b(v,\underline{q}) = \int_\Omega v\,\text{div}\,\underline{q}\,dx. \tag{1.11}$$

The optimality conditions for (1.9) can be written as

$$\begin{cases} a(\underline{p},\underline{q}) + b(\underline{q},u) = \langle g_1, \underline{q}\cdot\underline{n}\rangle, & \forall \underline{q} \in H_{0,N}(\text{div};\Omega), \\ b(\underline{p},v) = -\displaystyle\int_\Omega f v\,dx, & \forall v \in L^2(\Omega). \end{cases} \tag{1.12}$$

We work with the spaces $V = H_{0,N}(\text{div};\Omega)$, $Q \in L^2(\Omega)$. It is natural here to identify Q and its dual space Q'. The operator B is then the divergence operator from V into Q. This operator is surjective. Indeed if $f \in L^2(\Omega) = Q$ is given, we can solve the problem

$$\begin{aligned} -\Delta\phi &= f \text{ in } \Omega, \\ \phi|_D &= 0, \\ \frac{\partial\phi}{\partial n}|_N &= 0 \end{aligned}$$

to find $\phi \in H^1(\Omega)$. Taking $\underline{p} = \underline{\text{grad}}\,\phi$, we have found $\underline{p} \in H_{0,N}(\text{div};\Omega)$ with $\text{div}\,\underline{p} + f = 0$.

Remark 1.1: Note also that such a $\underline{p}$ will belong, for instance, to the space $(L^s(\Omega))^2$ for some $s > 2$. Setting

$$W = \{\underline{q} \mid \underline{q} \in (L^s(\Omega))^2,\ \text{div}\,\underline{q} \in L^2\} \cap H_{0,N}(\text{div};\Omega), \tag{1.13}$$

we have $\|\underline{p}\|_W \le c\,\|f\|_Q$. Hence, B has a continuous lifting from Q into W.

Moreover we have coercivenes of $a(.,.)$ on $\text{Ker}\,B$ although not on V. Using assumption (1.2) we have, in fact, whenever $\text{div}\,\underline{q}_0 = 0$,

$$a(\underline{q}_0,\underline{q}_0) \ge \alpha\,|\underline{q}_0|^2_{(L^2(\Omega))^n} = \alpha\,\|\underline{q}_0\|^2_{H(\text{div},\Omega)}. \tag{1.14}$$

The theory of Chapter II then applies in a straightforward way and we obtain *existence and uniqueness of a solution* $(\underline{p},u)$ *to this problem.* □

Remark 1.2: Uniqueness of the Lagrange multiplier u is consequence of the surjectivity of B which implies $\operatorname{Ker} B^t = \{0\}$. □

Remark 1.3: The reader should take care to the inversion of notations between the general theory of Chapter II and the present application. In the present case, $\underline{p}$ is the primal variable and u the Lagrange multiplier. □

The above results also enable us to consider a nonhomogeneous problem, that is, the case $g_2 \neq 0$ in (1.1). To do so we consider any $\underline{\tilde{q}}$ such that

$$A^{-1}\underline{\tilde{q}} \cdot \underline{n} = g_2 \text{ on } N. \tag{1.15}$$

This is possible and can be done explicitly by considering a classical solution to problem (1.1) with $f = 0$ and $g_1 = 0$ and then taking $\underline{\tilde{q}} = \underline{\operatorname{grad}}\, u$. We then look for $\underline{p} = \underline{\tilde{q}} + \underline{p}_0$ with $\underline{p}_0 \in H_{0,N}(\operatorname{div};\Omega)$. This leads us to the problem

$$\begin{cases} a(\underline{p}_0, \underline{q}_0) + b(\underline{q}_0, u) = \langle g_1, \underline{q}_0 \cdot \underline{n}\rangle - a(\underline{\tilde{q}}, \underline{q}_0), & \forall \underline{q} \in H_{0,N}(\operatorname{div};\Omega), \\ b(\underline{p}_0, v) = -\displaystyle\int_\Omega f v\, dx - b(\underline{\tilde{q}}, v), & \forall v \in L^2(\Omega). \end{cases} \tag{1.16}$$

This means that considering $g_2 \neq 0$ can be reduced to changing the right-hand side of (1.12).

We are therefore ready to consider the approximation of the Mixed Formulation.

In Chapter III, we built function spaces for the purpose of approximating $H(\operatorname{div};\Omega)$. We can now use to discretize problem (1.12) or (1.16) anyone of the spaces $M_k(\Omega, \mathcal{T}_h)$ introduced in Section III.3.4. The approximation of $Q = L^2(\Omega)$ is then implicitly done: Q_h must be $\mathcal{L}^0(\Omega, D_k)$. To fix ideas, we shall use, following RAVIART–THOMAS [A] $RT_k(\Omega, \mathcal{T}_h)$ and we define

$$V_h = \{\underline{q}_h \,|\, \underline{q}_h \in RT_k(\Omega, \mathcal{T}_h),\ \underline{q}_h \cdot \underline{n}|_N = 0\}. \tag{1.17}$$

Such a definition is possible if the partition into elements is made in such a way that there is no element across the interface between D and N on Γ. Having chosen V_h as in (1.17), we must take

$$Q_h = \mathcal{L}^0_h(\Omega) = \{v_h \,|\, v_h|_K \in P_k(K)\}. \tag{1.18}$$

We could replace this choice with any of the elements listed in Section III.3.3. In order to apply results of Chapter II without unnecessary technicalities, we shall restrict ourselves to the case of affine elements.

We may now introduce the discrete problem

$$\begin{cases} \displaystyle\int_\Omega A^{-1}\underline{p}_h \cdot \underline{q}_h\, dx + \int_\Omega u_h \operatorname{div} \underline{q}_h\, dx = \langle \underline{\tilde{g}}, \underline{q}_h\rangle, & \forall \underline{q}_h \in V_h, \\ \displaystyle\int_\Omega v_h \operatorname{div} \underline{p}_h\, dx + \langle \tilde{f}, v_h\rangle = 0, & \forall v_h \in Q_h, \\ (\underline{p}_h, u_h) \in V_h \times Q_h, \end{cases} \tag{1.19}$$

where $\tilde{f}$ and $\tilde{\underline{g}}$ may include nonhomogeneous boundary conditions as in problem (1.16), that is,

$$\langle \tilde{\underline{g}}, \underline{q} \rangle = \langle g_1, \underline{q} \cdot \underline{n} \rangle - a(\tilde{\underline{q}}, \underline{q}),$$
$$\langle \tilde{f}, v \rangle = \int_\Omega f v \, dx + b(\tilde{q}, v).$$

To apply the results of Chapter II, we must check that the bilinear form $a(\cdot,\cdot)$ is coercive on $\operatorname{Ker} B_h$ and the inf–sup condition. These properties will be an easy consequence of the commutative diagram (III.3.72). In particular, we already know from (III.3.73) that

$$\operatorname{div} V_h = Q_h. \tag{1.20}$$

This shows that B_h is nothing but the restriction to V_h of the divergence operator and that it is surjective so that $\operatorname{Ker} B_h^t = \{0\}$. Moreover ,we have

$$B_h = B|_{V_h} = \operatorname{div}|_{V_h} \tag{1.21}$$

and this implies obviously that we are in the special and interesting case where

$$\operatorname{Ker} B_h \subset \operatorname{Ker} B. \tag{1.22}$$

We can then rewrite (III.3.72) in the abstract form

$$\begin{array}{ccc} W & \xrightarrow{B} & Q \equiv Q' \\ \Pi_h \big\downarrow & & P_h \big\downarrow \\ V_h & \xrightarrow{B_h} & Q_h \equiv Q_h' \end{array} \tag{1.23}$$

with P_h the L^2-projection from Q onto Q_h. From Remark 1.1, we know that B has a continuous lifting from Q to W. Since the operators Π_h are uniformly bounded from W to V_h, we have

$$\begin{cases} \displaystyle\int_\Omega (\operatorname{div} \underline{q} - \operatorname{div} \Pi_h \underline{q}) v_h \, dx = 0, & \forall v_h \in Q_h, \\ \|\Pi_h \underline{q}\|_V \le c \, \|\underline{q}\|_W. \end{cases} \tag{1.24}$$

The first part of (1.24) is a direct consequence of the commuting property of diagram (1.23). Using (1.24) and Proposition II.2.8 we obtain that the discrete inf–sup condition is satisfied with a constant independent of h.

On the other hand, (1.22) implies that the coercivity of $a(\cdot,\cdot)$ on $\operatorname{Ker} B_h$ is trivial and follows directly from (1.14).

We can now apply our abstract results to get

Proposition 1.1: Problem (1.19) has a unique solution. Moreover, if $(\underline{p}, u)$ is the solution of problem (1.16), we have the estimates

(1.25)
$$\|\underline{p} - p_h\|_V \leq c \inf_{\underline{q} \in V_h} \|\underline{q} - \underline{q}_h\|_V,$$

(1.26)
$$\|u - u_h\|_Q \leq c \left(\inf_{v_h \in Q_h} \|u - v_h\|_Q + \inf_{\underline{q} \in V_h} \|\underline{p} - \underline{q}_h\|_V \right).$$

Proof: (1.25) is nothing but Proposition II.2.6 in the case where $\operatorname{Ker} B_h \subset \operatorname{Ker} B$. Then (1.26) follows from Proposition II.2.7. □

This direct use of Chapter II is optimal when the spaces RT or $BDFM$ are used but not with BDM. This comes from the fact that in $RT_k(\Omega, \mathcal{T})$ we have an estimate on $\inf_{\underline{q}_h \in V} \|\underline{q} - \underline{q}_h\|_0$ and $\inf_{\underline{q}_h \in V_h} \|\operatorname{div} \underline{q} - \operatorname{div} \underline{q}_h\|_0$ to the same order $O(h^{k+1})$, whereas the latter is only $O(h^k)$in $BDM_k(\Omega; \mathcal{T}_h)$ (Proposition III.3.9). We must, however, not despair. Denoting

(1.27)
$$H = (L^2(\Omega))^n.$$

we have as in Section II.2.5

(1.28)
$$\begin{cases} a(\underline{p}, \underline{q}) \leq \|\underline{p}\|_H \, \|\underline{q}\|_H, \\ a(\underline{q}, \underline{q}) \geq \alpha \|\underline{q}\|_H^2. \end{cases}$$

and indeed $\|\underline{q}\|_H = \|\underline{q}\|_V$ for any $\underline{q} \in \operatorname{Ker} B$. We can thus apply estimate (II.2.53) of Remark II.2.14 which yields

(1.29)
$$\|\underline{p} - \underline{p}_h\|_H \leq c \inf_{\underline{q}_h \in Z_h(g)} \|\underline{p} - \underline{q}_h\|_H \leq c \|\underline{p} - \Pi_h \underline{p}\|_H,$$

which is now optimal. It must be noted that the second term of (II.2.53) vanishes as we have $\operatorname{Ker} B_h \subset \operatorname{Ker} B$. Estimate (1.29) is now optimal for any of the spaces considered in Chapter III.

We can now join the above results with the approximation results (III.3.74) and (III.3.75).

Proposition 1.2: Let $M_k(\Omega, \mathcal{T}_h)$ be any of the spaces defined in (III.3.67) and (III.3.41) to (III.3.46) in the two-dimensional case or (III.3.51) to (III.3.56) in the three-dimensional case. Let $\mathfrak{L}^0(D_k, \mathcal{T}_h)$ be the corresponding space given by (III.3.68). Let $(\underline{p}, u)$ be the solution of problem (1.16). Let $(\underline{p}_h, u_h)$ be the solution in $V_h \times Q_h = M_k(\Omega, \mathcal{T}_h) \times \mathfrak{L}^0(D_k, \mathcal{T}_h)$ of problem (1.19). Then we have the estimates

(1.30)
$$\|\underline{p} - \underline{p}_h\|_{0,\Omega} \leq c h^s \|\underline{p}\|_{s,\Omega}$$

for $s \leq k + 1$. Moreover, we also have

(1.31)
$$\|u - u_h\|_{0,\Omega} \leq c h^s \left(\|\underline{p}\|_s + \|u\|_s \right)$$

for $s \leq k+1$ for the spaces RT and $BDFM$ and $s \leq k$ for the spaces BDM. □

Remark 1.4: The case of non-affine elements is somewhat more tricky. In that case the contravariant transformation $\mathfrak{G}$ of (III.1.45) no longer has a constant Jacobian and we no longer have $B_h = \operatorname{div}|_{V_h}$ because

$$\operatorname{div} \underline{q} = \operatorname{div}(\mathfrak{G}\hat{\underline{q}}) = \mathfrak{F}\Big(\frac{\operatorname{div} \hat{\underline{q}}}{J}\Big) \tag{1.32}$$

where $\mathfrak{F}$ is the standard change of variables (III.1.35). As the Jacobian is not constant in the general case, $\operatorname{div} \hat{q} \notin Q_h$. It can, however, be checked that $\operatorname{Ker} B_h \hookrightarrow \operatorname{Ker} B$. We refer to THOMAS [B] for a study of this case. □

Remark 1.5: In the affine case (where $\operatorname{div} V_h = Q_h$), a direct subtraction of the second equations in problems (1.16) and (1.19) yields

$$\int_\Omega (\operatorname{div} \underline{p} - \operatorname{div} \underline{p}_h) v_h \, dx = 0, \qquad \forall v_h \in Q_h. \tag{1.33}$$

This means that $\operatorname{div} \underline{p}_h$ is the $L^2(\Omega)$-projection of $\operatorname{div} \underline{p}$ onto Q_h. An estimate of $\|\operatorname{div} \underline{p} - \operatorname{div} \underline{p}_h\|$ then directly follows. □

We shall come back to this mixed method in Chapter V. We shall then consider sharper estimates and introduce Lagrange multipliers to deal with continuity of $\underline{p}_h \ \underline{n}$ at interfaces. This will allow us in particular to build an efficient solution method and to obtain from the results a better approximation of u. This method of Lagrange multipliers is in fact quite general and will lead to a more standard interpretation of otherwise non-standard methods. In particular, BDM spaces will recover in the scalar variable the same order of convergence as the other methods.

IV.1.3 Primal hybrid methods

We now consider for the first time a nonstandard method (cf. RAVIART–THOMAS [B]) based on domain decomposition. We place ourselves in the frame of Example I.3.4.

To avoid complicating unduly our presentation we shall restrict ourselves to problem (1.1) in which $D = \Gamma$, that is ,Dirichlet boundary conditions on the whole of Γ. This restriction is in no way essential and does not diminish the generality of our results. We thus want to find $u \in H_0^1(\Omega)$ solution of the minimization problem

$$\inf_{v \in H_0^1(\Omega)} \tfrac{1}{2} \int_\Omega A \, \underline{\operatorname{grad}} \, v \, \underline{\operatorname{grad}} \, v \, dx - \int_\Omega f v \, dx, \tag{1.34}$$

or equivalently of the variational problem

$$\int_\Omega A \, \underline{\operatorname{grad}} \, u \ \ \underline{\operatorname{grad}} \, v \, dx = \int_\Omega f v \, dx, \qquad \forall v \in H_0^1(\Omega). \tag{1.35}$$

Introducing now a partition of Ω into elements, it is natural (this is in fact one of the basic ideas of the Finite Element Method) to define u on each element and to impose continuity conditions at the interfaces. The standard assembly process is based on this idea. We now follow a slightly different route. We use $X(\Omega) = \prod_r H^1(K_r)$ as defined by (III.1.21) with the product norm (III.1.22). $H_0^1(\Omega)$ is then a closed subspace of $X(\Omega)$ and the fact of belonging to $H_0^1(\Omega)$ can be considered as a linear constraint on u. From this, we can transform (1.34) into a saddle point problem:

$$\text{(1.36)} \qquad \inf_{v \in X(\Omega)} \sup_{\underline{q} \in H(\mathrm{div},\Omega)} \sum_K \left\{ \frac{1}{2} \int_K A\, \underline{\mathrm{grad}}\, v \cdot \underline{\mathrm{grad}}\, v\, dx - \int_{\partial K} v \underline{q} \cdot \underline{n}\, ds - \int_K f v\, dx \right\},$$

where we formally write $\int_{\partial K} v \underline{q}\ \underline{n}\, ds$ for the duality between $H^{1/2}(\partial K)$ and $H^{-1/2}(\partial K)$. The optimality conditions of problem (1.36) are indeed

$$\text{(1.37)} \qquad \begin{cases} \displaystyle\sum_K \left\{ \int_K A\, \underline{\mathrm{grad}}\, u \cdot \underline{\mathrm{grad}}\, v\, dx - \int_{\partial K} v \underline{p} \cdot \underline{n}\, ds - \int_K f v\, dx \right\} = 0, & \forall v \in X(\Omega), \\ \displaystyle\sum_K \left\{ \int_{\partial K} u \underline{q} \cdot \underline{n}\, ds \right\} = 0, & \forall \underline{q} \in (\mathrm{div}, \Omega). \end{cases}$$

From Proposition III.1.1 we then have $u \in H_0^1(\Omega)$ so that u satisfies (1.35). Let us now set our problem in the framework of the general theory of Chapter II. Taking $V = X(\Omega)$ and $Q = H(\mathrm{div};\Omega)$; we then define,

$$\text{(1.38)} \qquad a(u,v) = \sum_K \left\{ \int_K A\, \underline{\mathrm{grad}}\, . \underline{\mathrm{grad}}\, v\, dx \right\}, \qquad \forall u, v \in V,$$

and

$$\text{(1.39)} \qquad b(v, \underline{q}) = \sum_k \left\{ - \int_{\partial K} v\, \underline{q} \cdot \underline{n}\, ds \right\}, \qquad \forall v \in V,\ \forall \underline{q} \in Q,$$

always using the formal integral notation for the duality between $H^{-1/2}(\partial K)$ and $H^{1/2}(\partial K)$. The bilinear form $b(v, \underline{q})$ defines an operator B from V into Q. We have from Propositions III.1.1 and III.1.2

$$\text{(1.40)} \qquad \mathrm{Ker}\, B = H_0^1(\Omega)$$

and

$$\text{(1.41)} \qquad \begin{aligned} \mathrm{Ker}\, B^t &= \{ \underline{q} \mid \underline{q} \in H(\mathrm{div};\Omega), \underline{q}\ \ \underline{n}|_{\partial K} = 0,\ \forall K \in \mathcal{T}_h \} \\ &= \prod_K H_{0,\partial K}(\mathrm{div}; K) \end{aligned}$$

Loosely speaking, the operator B associates to u its jumps on interelement interfaces. We could also have defined it from V *onto* the space

$$\mathfrak{M}^{1/2} = \prod_K H^{1/2}(\partial K). \tag{1.42}$$

We thus want to check the closedness of Im B by obtaining an inequality of the form

$$\sup_{v \in V} \frac{b(v, \underline{q})}{\|v\|_V} \geq k\, |\underline{q}|_{Q/\operatorname{Ker} B^t}. \tag{1.43}$$

In the present case it is obvious that one has

$$\sup_{v \in V} \frac{b(v, \underline{q})}{\|v\|_V} = \tfrac{1}{2}\Big\{\sum_K (\|\underline{q} \cdot \underline{n}\|_{-1/2, \partial K})^2\Big\}^{1/2} \tag{1.44}$$

and to obtain (1.43) it is sufficient to show that one has (on each element)

$$\inf_{\underline{q}_0 \in H_{0,\partial K(\mathrm{div},K)}} \|\underline{q} + \underline{q}_0\|_{H(\mathrm{div},K)} \leq \|\underline{q} \cdot \underline{n}\|_{-1/2,\partial K}, \tag{1.45}$$

But (1.45) is readily obtained by solving a Neumann problem

$$\int_K \underline{\mathrm{grad}}\, \phi \cdot \underline{\mathrm{grad}}\, v\, dx + \int_K \phi v\, dx = \int_{\partial K} \underline{q} \cdot \underline{n}\, v\, ds. \tag{1.46}$$

Setting $\hat{\underline{q}} = \underline{\mathrm{grad}}\, \phi$, we have $\hat{\underline{q}} \cdot \underline{n} = \underline{q} \cdot \underline{n}$ and $\mathrm{div}\, \hat{\underline{q}} = \phi \in L^2(K)$. Moreover, we have

$$\|\hat{\underline{q}}\|_{H(\mathrm{div},K)} = \|\phi\|_{H^1} \leq \|\underline{q} \cdot \underline{n}\|_{-1/2,\partial K}$$

and (1.45) follows.

Proposition 1.3: Let $f \in L^2(\Omega)$ be given. There exists a solution $(u, \underline{p})$ to problem (1.37). The first component is unique and the second one is defined up to an element of $\operatorname{Ker} B^t$ as defined by (1.41).

Proof: Assumption (1.2) made on A, implies that $a(\cdot,\cdot)$ is coercive on $\operatorname{Ker} B = H_0^1(\Omega)$. The result follows by the closedness of Im B and Theorem II. 1.1. □

Remark 1.6: The first component u is of course the unique solution of problem (1.1). The second component *can be chosen* so that $\operatorname{div} \underline{p} + f = 0$. Indeed taking $v = 1$ on K and 0 elsewhere, we have from (1.37a) for any solution $\underline{p}_0$,

$$\int_{\partial K} \underline{p}_0 \cdot \underline{n}\, ds = \int_K \operatorname{div} \underline{p}_0\, dx = -\int_K f\, dx. \tag{1.47}$$

It is then possible to solve on K the Neumann problem,

$$\begin{cases} -\Delta\phi = f + \operatorname{div} \underline{p}_0, \\ \left.\dfrac{\partial \phi}{\partial n}\right|_{\partial K} = 0 \end{cases}$$

The solution exists, and is defined up to an additive constant, as the right-hand side is compatible. Then $\underline{q}_0 = \underline{\operatorname{grad}}\, \phi \in \operatorname{Ker} B^t$ and $\underline{p}_1 = \underline{p}_0 + \underline{q}_0$ satisfies $\operatorname{div} \underline{p}_1 + f = 0$. □

Remark 1.7: It is moreover possible to choose $\underline{p} = A\, \underline{\operatorname{grad}}\, u$. Indeed there comes from the first equation of (1.37) that $\underline{p} \cdot \underline{n}|_{\partial K} = A\, \underline{\operatorname{grad}}\, u \cdot \underline{n}|_{\partial K}$ on any $K \in \mathcal{T}_h$. □

We are now able to consider a discretization of problem (1.37). We shall use, as an example,

$$V_h = \mathfrak{L}^0_{k+1}(\Omega) \subset V = X(\Omega) \tag{1.48}$$

$$Q_h = \{q_h \in H(\operatorname{div};\Omega),\ \underline{q}_h \cdot \underline{n} \in R_k(\partial K),\ \forall K \in \mathcal{T}_h\}. \tag{1.49}$$

Note that only the traces of vectors in Q_h are polynomials. Our space Q_h is in fact infinite dimensional. This is no problem in practice as only the (finite dimensional) traces are used in computing. We then solve the discrete problem

$$\begin{aligned} &\sum_K \Big\{ \int_K A\, \underline{\operatorname{grad}}\, u_h \cdot \underline{\operatorname{grad}}\, v_h\, dx \\ &\qquad - \int_{\partial K} v_h\, \underline{p}_h \cdot \underline{n}\, ds - \int_K f v_h\, dx \Big\} = 0, \quad \forall v \in V_h, \\ &\sum_K \Big\{ \int_{\partial K} u_h\, \underline{q}_h \cdot \underline{n}\, ds \Big\} = 0, \quad \forall q_h \in Q_h. \end{aligned} \tag{1.50}$$

The first step in the analysis of such a discretization is to examine the properties of the operator B_h associated with the bilinear form $b(v_h, q_h)$.

The first point that comes out is that we do not have (as in the previous example), $\operatorname{Ker} B_h \subset \operatorname{Ker} B$; that is, functions in $\operatorname{Ker} B_h$ do not belong to $H^1_0(\Omega)$.

It is, however, easy to see that their moments up to order k are continuous across interelements boundaries. This in turn implies, as the traces are polynomials of degree $k+1$ that we have continuity of the functions in $\operatorname{Ker} B_h$ at the $k+1$ Gauss–Legendre points, associated to a quadrature formula of degree $k+2$, on every interface. Eliminating the Lagrange multiplier $\underline{q}_h$ thus yields a nonconforming approximation of problem (1.37), namely, to find $u_h \in \operatorname{Ker} B_h$ solution of

$$(1.51) \qquad \sum_K \Big\{ \int_K A \,\underline{\operatorname{grad}}\, u_h \cdot \underline{\operatorname{grad}}\, v_h \, dx \Big\} = \int_\Omega f v_h \, dx, \quad \forall v_h \in \operatorname{Ker} B_h.$$

We already considered such approximations in Chapter III and their analysis is fairly well established (STRANG–FIX [A], CIARLET [B], CROUZEIX–RAVIART [A], CEA [B], STUMMEL [A], FRAEIJS DE VEUBEKE [B]).

One can therefore say that primal hybrid methods are another way of introducing nonconforming methods. The new point is to introduce an approximation of $\underline{p} = A \,\underline{\operatorname{grad}}\, u$ which is more regular than the approximation deduced directly from u_h. Moreover, this approximation can be built in order to satisfy the equilibrium conditions. Finally the convergence analysis through the saddle point approach is simpler than the standard one and permits one to introduce correctly the "patch test" arising in the analysis of nonconforming methods.

Before coming to this point, we first have to show existence and uniqueness of a solution. With respect to the existence and uniqueness of the solution u_h of (1.51) we fortunately have no problem It is obvious that

$$(1.52) \qquad |v_h|_{V_h} = \sqrt{a(v_h, v_h)}$$

defines on V_h a continuous seminorm. The kernel of this seminorm is

$$(1.53) \qquad M = \{ v_h \,|\, v_h \in L^2(\Omega),\ v_h|_K \in P_0(K) \} = \mathfrak{L}_0^0,$$

and we have $M \cap \operatorname{Ker} B_h = 0$ so that $a(u_h, v_h)$ is coercive on $\operatorname{Ker} B_h$:

$$(1.54) \qquad a(v_{0h}, v_{0h}) \geq \alpha_h \, |v_{0h}|^2_{V_h}, \qquad \forall v_{0h} \in \operatorname{Ker} B_h.$$

We do not know, however, how α_h depends on h. This would require a discrete Poincaré inequality and is quite technical to prove.

We shall rather obtain an error bound in the seminorm $|v_h|_{V_h}$ using Proposition II.2.14. In order to do so we must build an interpolate $\Pi_h \underline{p}$ of $\underline{p}$ such that

$$(1.55) \qquad b(v_h, \underline{p} - \Pi_h \underline{p}) = 0, \qquad \forall v_h \in M.$$

But this is immediate by taking $\Pi_h \underline{p}$ as defined by (III.3.71) provided $\underline{p}$ is at least in W defined by (1.13). We thus obtain the error bound

$$(1.56) \qquad |u - u_h|_{V_h} \leq c \Big(\inf_{v_h \in \operatorname{Ker} B_h} |u - v_h|_{V_h} + \|\underline{p} - \Pi_h \underline{p}\|_Q \Big).$$

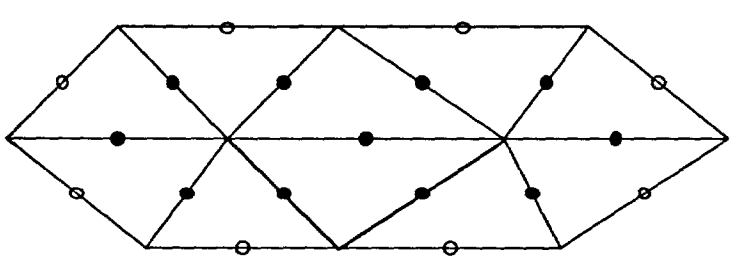
Such an estimate is typical of nonconforming methods. The first term is readily estimated by standard methods, that is, by the use of some interpolation operator. The second one has already been considered. It must be remarked that as $\underline{p}$ is defined only up to an element of $\operatorname{Ker} B^t$, this norm depends in fact only on the values of $\underline{p} \cdot \underline{n}$ and $\underline{p}_h \cdot \underline{n}$ on the boundary of the elements. If $\underline{p}$ is regular, we get the same order of accuracy as in the first term. We therefore recognize here a form of the classical patch test: moments up to order k must be continuous to get the optimal convergence rate (CEA [B]). This corresponds to the choice of multipliers belonging to $P_k(e_i)$ on interfaces and thus to the choice (1.49) for Q_h. *Consistency terms that appear in the analysis of nonconforming methods are nothing but the contribution of the dual variable to error estimates.* Choosing a poorer approximation would destroy convergence properties. The main difficulty in the present situation will be to study the convergence of $\underline{p}_h$. To do so we now have to check the inf–sup condition. We shall try to do it by the criterion of Proposition II.2.8, that is by building a proper interpolation operator for $u \in V$: given $u \in V = X(\Omega)$, one must find $\tilde{u}_h \in V_h$ such that

$$b(u - \tilde{u}_h, \underline{q}_h) = 0, \qquad \forall \underline{q}_h \in Q_h, \tag{1.57}$$

and depending continuously on u. This would prove $\operatorname{Ker} B_h^t \subset \operatorname{Ker} B^t$ and the inf–sup condition. We must then distinguish between two cases depending on whether k is even or odd. To make things simpler we shall restrict our presentation to $k = 1$ or 2 (which are, by far, the most important in practice).

Example 1.1: *Hybrid method, $k = 1$.*

This is the simplest case of primal hybrid method (or nonconforming method). Functions of V_h are piecewise linear and $\operatorname{Ker} B_h$ contains those of them that are continuous at mid-side points on interfaces (Figure IV.1).



• continuity point
∘ vanishing point

Figure IV.1: Ker B_h

The space $Q_h / \operatorname{Ker} B^t$ can be assimilated here to $RT_0(\Omega)$. Now taking $u \in X(\Omega)$, one readily builds $\tilde{u}_h$ by taking on each K

$$\int_{e_i} \tilde{u}_h \, ds = \int_{e_i} u \, ds, \quad i = 1, 2, 3. \tag{1.58}$$

It is then obvious that (1.57) holds; moreover, checking continuity is straightforward so that we have the error bound

$$(1.59)\quad \|\underline{p} - \underline{p}_h\|_{Q/\operatorname{Ker} B^t} \le c(\|\underline{p} - \underline{q}_h\|_{Q/\operatorname{Ker} B^t} + |u - u_h|_{V_h}), \quad \forall q_h \in Q_h.$$

In practice this means that one can extract from such a nonconforming formulation an approximation of $\underline{\operatorname{grad}}\, u$ that is better than the direct one, $\underline{\operatorname{grad}}\, u_h$. We shall see later (Chapter V) how this approximation can be easily deduced from the standard one by a simple post processing trick (MARINI [C]). □

Example 1.2: *Hybrid Method,* $k = 2$.

This hybrid formulation yields the next simpler case of a nonconforming method. Its use was long rejected because of a problem in the choice of the degrees of freedom. Although the functions of $\operatorname{Ker} B_h$ are continuous at two Gauss–Legendre points on each side, these points cannot be used as degrees of freedom because their values are linked by a linear relation. Indeed, let a_i $(1 \le i \le 6)$ be the six values of a second-degree polynomial on the six Gauss–Legendre points of the sides of a triangle (Figure IV.2), that is, $a_i = p_2(x_i)$.

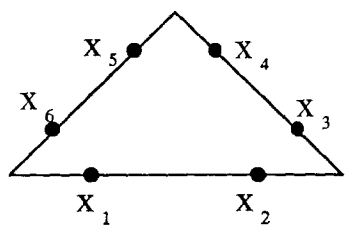


Figure IV.2

One then has

$$(1.60)\qquad (a_6 - a_5) + (a_4 - a_3) + (a_2 - a_1) = \int_{\partial K} \frac{\partial p_2}{\partial t} ds = 0$$

(FORTIN–SOULIE [A]). We shall call the *nonconforming bubble* the second-degree function vanishing at the six Gauss–Legendre points ($a_i = 0$) and taking value 1 at the barycenter of K. There also follows from (1.60) that one cannot define $\tilde{u}_h|_K$ by the six moments

$$(1.61)\qquad \int_{e_i} \tilde{u}_h\, \phi_i\, ds, \quad \phi_i \in P_1(e_i),$$

and this precludes checking (1.57) by the simple method of the previous example. Considering the problem a little more thoroughly, one then sees that $\operatorname{Ker} B_h^t \not\subset \operatorname{Ker} B^t$ and that (1.59) cannot hold.

Indeed $\operatorname{Ker} B_h^t$ contains one vector $\underline{q}_h^0$ that does not lie in $\operatorname{Ker} B^t$. It is sketched in Figure IV.3, where the symbols + and - represent equal absolute values of the normal component of $\underline{q}_h$.

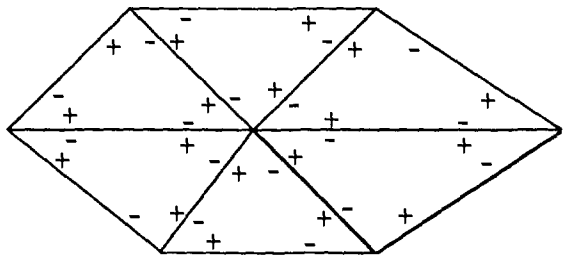


Figure IV.3: The vector $\underline{q}_h^0$

This is the first occurrence of a pathological situation where the inf–sup condition does not hold. In principle, this should imply some compatibility condition on the data. However, in the present case the second equation of (1.37) is always solved with a zero right-hand side and $Z_h(g) = Z_h(0) = \operatorname{Ker} B_h$ is always nonempty.

It must be noted that contrarily to other cases of spurious modes that we shall meet, for instance in Chapter VI, the existence of $\underline{q}_h^0$ does not depend on the mesh. Moreover, we know that its existence does not compromise the error estimates on u_h. One may therefore wonder if some convergence of $\underline{p}_h$ could not be obtained, modulo $\underline{q}_h^0$, that is, using an inf–sup condition of the form

$$\sup_{v_h \in V_h} \frac{b(v_h, \underline{q}_h)}{\|v_h\|_V} \geq k_0 \, \|\underline{q}_h\|_{Q_h/\operatorname{Ker} B_h^t}. \tag{1.62}$$

From Proposition II.2.8 this will hold if, given $u \in V$, $b(u, \underline{q}_h^0) = 0$, one can build $\tilde{u}_h \in V_h$ such that (1.57) holds.

This can, indeed, be done, through a construction that is not local and for which we do not know how to prove that the operation $\Pi_h : u \to u_h$ is uniformly continuous (with respect to h). We shall however be able to prove a partial result: $\underline{p}_h$ will converge in a quotient space Q_h/M_h with $\operatorname{Ker} B_h^t \subset M_h$. In order to see this, we first define on every element K

$$\underline{q}_K^0 = \underline{q}_h^0|_K \tag{1.63}$$

and we denote by Q_K^0 the one-dimensional space generated by $\underline{q}_K^0$. We then define $M_h = \sum_K Q_K^0$ and

$$Q_h^* = Q_h + M_h. \tag{1.64}$$

It must be noted that $Q_h^* \not\subset H(\mathrm{div};\Omega)$ so that we must now consider a nonconforming framework replacing Q by $Q^* = \prod_K H(\mathrm{div};K)$ as in Section II.2.6.

Let us first remark that the proof given for Im B to be closed is directly extended to the operator $B^* \cdot V \to Q^*$ now associated with the bilinear form $b(\cdot,\cdot)$ because this proof did not rely on any continuity property. It is also easy to check that one now has

$$\mathrm{Ker}\, B_h^{*t} = M_h + \mathrm{Ker}\, B^t \tag{1.65}$$

where B_h^* is evidently defined by the extension of $b(\cdot,\cdot)$ to $V_h \times Q_h^*$.

Considering now the problem

$$\begin{cases} a(u_h^*, v_h) + b(v_h, \underline{p}_h^*) = \langle f, v_h\rangle, & \forall v_h \in V_h, \\ b(u_h^*, \underline{q}_h^*) = 0, \qquad \forall \underline{q}_h^* \in Q_h^*, \end{cases} \tag{1.66}$$

it is easy to see that $u_h^* = u_h$. This comes from

$$b(u_h, \underline{m}_h) = 0, \qquad \forall \underline{m}_h \in M_h, \tag{1.67}$$

which is a direct consequence of (1.66). We thus have increased the indeterminacy of $\underline{q}_h$ without changing u_h. To prove convergence, we shall use Theorem VI.5.1, which is directly suitable. Let us thus define in the notations of this theorem

$$\begin{aligned} \tilde{Q}_h &= M_h, \\ \hat{Q}_h &= Q_h^* \backslash M_h, \\ \hat{V}_h &= V_h. \end{aligned} \tag{1.68}$$

From (1.67) we have $b(u_h, \underline{\tilde{q}}_h) = 0$, $\forall u_h \in V_h$, $\forall \underline{\tilde{q}}_h \in \tilde{Q}_h$, and there remains to prove that $b(\cdot,\cdot)$ satisfies an inf–sup condition on $V_h \times \hat{Q}_h$.

To do so, by Proposition II.2.8, one must build in a continuous way $\tilde{u}_h = \Pi_h u$ such that

$$b(\tilde{u}_h - u, \underline{\hat{q}}_h) = 0, \qquad \forall \underline{\hat{q}}_h \in \hat{Q}_h. \tag{1.69}$$

Working in $\hat{Q}_h$ (from which components $\underline{q}_K^0$ have been removed) now enables us to do it in a local way, that is, element by element. It is indeed sufficient, as in the previous example ($k = 1$) to interpolate u using its moments. This does not determine $\tilde{u}_h$ in a unique way and a minimum norm solution has to be chosen to get the desired uniform continuity property. We thus have the inf–sup condition (k being independent of h)

$$\sup_{v_h \in V_h} \frac{b(v_h, \underline{q}_h^*)}{\|v_h\|_{V_h}} \geq k_0 \|\underline{q}_h^*\|_{Q_h^*/M_h}, \quad \forall \underline{q}_h^* \in Q_h^*. \tag{1.70}$$

In Chapter VI, other instances are presented where a global mode such as $\underline{q}_K^0$ can be transformed into a local mode $\sum_K \alpha_K \underline{q}_K^0$. From the results of Section VI.5, we obtain that if the exact solution u satisfies

$$b(u, \underline{m}_h) = 0, \qquad \forall \underline{m}_h \in M_h = \tilde{Q}_h,$$

we have the estimate

$$(1.71)\quad \|\underline{p}-\underline{p}_h^*\|_{Q_h^*/M_h} \leq \Big(\inf_{q_h \in Q_h} \|\underline{p}-\underline{q}_h\|_Q + \|u-u_h\|_{V_h}\Big) + \inf_{\underline{q}_h^* \in Q_h^*} \|\underline{p}-\underline{q}_h^*\|_{Q^*}.$$

Remark 1.8: It must be noted that condition (1.71) is not as stringent as may appear. Indeed, given $u \in V$ and replacing Bu by $B\hat{u}_h$, with $\hat{u}_h$ the interpolate of u in V_h, introduces a perturbation of the problem which now has $\hat{u}_h$ as a solution. This means that, by a slight modification of the data, it is possible to switch from a noncompatible problem to a compatible one without really changing the solution. □

Remark 1.9: Knowing that $\underline{q}_h^* \in Q_h^*$ implies that $\underline{q}_h \cdot \underline{n}$ is continuous at mid-points of the interfaces. It can be checked, using the results of FORTIN–SOULIE [A], that the converging part of $\underline{p}_h$ is sometimes in fact equal to $\underline{\mathrm{grad}}\, u_h$ which satisfies the same continuity properties for some right-hand sides. However, the procedure sketched above can be extended to higher approximations, the case $k = 4$ for instance, where this equality will no longer hold. □

Remark 1.10: In the case $k = 2$, it is possible to build the solution $\underline{p}_h$ of (1.58) starting from $\underline{\mathrm{grad}}\, u_h \in \tilde{Q}_h$. The trick is to use a spanning tree of the elements: starting from the root, one can then adjust $\alpha_K\, q_K^0$ on each element so that $\underline{q}_h \cdot \underline{n}$. is continuous on the interfaces with previously visited elements. The properties of $\underline{\mathrm{grad}}\, u_h$ shown in FORTIN–SOULIE [A] enable us to do so in a unique way as $\alpha_K\, \underline{q}_K^0$ can be chosen arbitrarily on the root of the spanning tree. This is obviously not a local construction. Its continuity depends on the diameter of the spanning tree and thus of h and this leads us to believe that our result is probably optimal. (This is not the case for the construction for $k = 1$, described in Chapter V, which is local.) □

IV.1.4 Dual hybrid methods

We now turn to another use of domain decomposition, this time to solve the dual formulation (1.7) (RAVIART–THOMAS [C], THOMAS [A]), . In this formulation, the main difficulty is to work in the affine subspace of $H(\mathrm{div};\Omega)$,

$$(1.72)\qquad W_f = \{\underline{q}_f \,|\, \underline{q}_f \in H(\mathrm{div}\,\Omega),\ \mathrm{div}\,\underline{q}_f + f = 0\}.$$

When Neumann conditions are imposed on $N \subset \Gamma$, it is also necessary to ask for $\underline{q}_f$ to satisfy

$$\underline{q}_f \cdot \underline{n}|_N = g_2. \tag{1.73}$$

The idea of the dual hybrid formulation will again be to relax continuity, this time for the normal trace $\underline{q} \cdot \underline{n}$ at interfaces between elements. Condition (1.73) will also be treated weakly. We thus transform problem (1.7) into

$$\inf_{\underline{q}_f \in V_f} \sup_{v_{g_1} \in Q_{g_1}} \frac{1}{2} \int_\Omega A^{-1} \underline{q}_f \cdot \underline{q}_f \, dx + \sum_K \int_{\partial K} \underline{q}_f \cdot \underline{n} \, v_{g_1} \, ds - \int_N g_2 \, v_{g_1} \, ds, \tag{1.74}$$

where, denoting as in Chapter III, $Y(\Omega) = \prod_K H(\mathrm{div}; K)$, one sets

$$V_f = \{\underline{q} \,|\, \underline{q} \in Y(\Omega), \ \mathrm{div}\, \underline{q}|_K + f = 0, \ \forall K\}, \tag{1.75}$$

$$Q_{g_1} = \{v \,|\, v \in H^1(\Omega), \ v|_D = g_1\}. \tag{1.76}$$

Taking $\hat{\underline{q}}_f$ an arbitrary element of V_f and $\hat{v}_{g_1}$ an arbitrary element of Q_{g_1}, one may write (1.74) as

$$\begin{aligned} &\inf_{\underline{q}_0 \in V_0} \sup_{v_0 \in Q_0} \frac{1}{2} \int_\Omega A^{-1} (\underline{q}_0 + \hat{\underline{q}}_f) \cdot (\underline{q}_0 + \hat{\underline{q}}_f) \, dx \\ &\quad + \sum_K \int_{\partial K} (\underline{q}_0 + \hat{\underline{q}}_f) \cdot \underline{n} (v_0 + \hat{v}_{g_1}) ds - \int_N g_2 (v_0 + \hat{v}_{g_1}) ds, \end{aligned} \tag{1.77}$$

where V_0 and Q_0 are defined by (1.75) and (1.76) with $f = 0$ and $g_1 = 0$. Denoting as in the previous section

$$b(\underline{q}, v) = \sum_K \int_{\partial K} \underline{q} \cdot \underline{n} \, v \, ds, \tag{1.78}$$

problem (1.77) is equivalent to finding $(\underline{p}_0, u_0) \in V_0 \times Q_0$ the solution of

$$\int_\Omega A^{-1} \underline{p}_0 \cdot \underline{q}_0 \, dx + b(\underline{q}_0, u_0) = - \int_\Omega A^{-1} \hat{\underline{q}}_f \underline{q}_0 - b(\underline{q}_0, \hat{v}_{g_1}), \quad \forall \underline{q}_0 \in Q_0, \tag{1.79}$$

$$b(\underline{p}_0, v_0) = -b(\hat{\underline{q}}_f, v_0) + \int_N g_2 \, v_0 \, ds, \quad \forall v_0 \in V_0. \tag{1.80}$$

This is now in standard form and we shall try to apply the general theory. First note that we have

$$\mathrm{Ker}\, B^t = \prod_K H_0^1(K) \tag{1.81}$$

and

$$\text{(1.82)} \qquad \operatorname{Ker} B = \{\underline{q} | \underline{q} \in H_{0,N}(\operatorname{div};\Omega),\ \operatorname{div} \underline{q} = 0\}.$$

It is then clear that

$$\text{(1.83)} \qquad a(\underline{p}, \underline{q}) = \int_\Omega A^{-1} \underline{p} \cdot \underline{q}\, dx$$

is coercive on V_0, and to apply our general existence result one must show an inf–sup condition, that is, for all $v \in Q \equiv Q_0$,

$$\text{(1.84)} \qquad \sup_{\underline{q}_0 \in V_0} \frac{b(\underline{q}_0, v)}{\|\underline{q}_0\|_{H(\operatorname{div};\Omega)}} \geq k_0 \|v\|_{Q/\operatorname{Ker} B^t}.$$

To obtain this, we first select, v being given, $v_0 \subset Q$ such that

$$\text{(1.85)} \qquad \begin{cases} -\Delta v_0 = 0 \text{ on each element } K \in \mathcal{T}_h, \\ v_0|_{\partial K} = v|_{\partial K}. \end{cases}$$

Now we take $\underline{p}_0 = \underline{\operatorname{grad}}\, v_0$ and we have

$$\text{(1.86)} \qquad \int_\Omega |\underline{\operatorname{grad}}\, v_0|^2 dx = \sum_K \int_K |\underline{\operatorname{grad}}\, v_0|^2 dx = \sum_K \int_{\partial K} v_0 \underline{p}_0 \cdot \underline{n} ds = b(\underline{p}_0, v_0).$$

Moreover, $\operatorname{div} \underline{p}_0 = 0$ and, using Poincaré's inequality, we may write

$$\text{(1.87)} \qquad \|\underline{p}_0\|_{H(\operatorname{div},\Omega)} = \|\underline{p}_0\|_0 = \|\underline{\operatorname{grad}}\, v_0\|_0 \geq \frac{1}{C(\Omega)} \|v_0\|_{1,\Omega},$$

provided the domain is bounded and Dirichlet conditions are imposed on a part of $\partial\Omega$ (that is: $D \neq \emptyset$). From (1.85) and (1.87), we then have

$$\text{(1.88)} \qquad \|v\|_{Q/\operatorname{Ker} B^t} \leq \|v_0\|_{1,\Omega} \leq \frac{b(\underline{p}_0, v)}{\|\underline{p}_0\|_{H(\operatorname{div};\Omega)}} \leq \sup_{\underline{q}_0} \frac{b(\underline{q}, v)}{\|\underline{q}_0\|_{H(\operatorname{div};\Omega)}},$$

which is the desired result.

We now know that problem (1.79) and (1.80) has a unique solution (up to an element of $\operatorname{Ker} B^t$ for v_0). Our concern is now to introduce a discretization and, to do so, we shall again use the spaces defined in Chapter III. We define, in the notation of Propositions III.3.6 and III.3.7.

$$\text{(1.89)} \qquad V_h = \prod_K M_k(K)$$

where $M_k(K)$ is one of the approximations of $H(\text{div}; K)$ introduced in Section III.3. We suppose $f|_K \in D_k = \text{div}(M_k(K))$ so that it is possible to find $\hat{\underline{q}}_{hf}$ satisfying $\text{div}\,\hat{\underline{q}}_{hf} + f = 0$ in each K. In general, f can be approximated on D_k without loss of precision.

Using (III.2.6), we also set

$$(1.90)\qquad \begin{aligned} Q_h &= \{v_h | v_h \in H^1(\Omega),\ v_h|_{\partial K} \in T_{k+1}(\partial K),\ \forall K \in \mathcal{T}_h\},\\ Q_{0h} &= \{v_h \in Q_h |\ v_h|_D = 0\}. \end{aligned}$$

We now have again the unusual situation where the approximation Q_{0h} is infinite dimensional. However, only the traces on K are relevant to computation and we do not really have to worry about this. Moreover the choice

$$(1.91)\qquad V_{0h} = \{\underline{q}_h \in V_h, \text{div}\,\underline{q}_h|_K = 0,\ \forall K \in \mathcal{T}_h\}.$$

ensures that we have no problem with coerciveness of $a(\cdot,\cdot)$. This comes from the inclusion $V_{0h} \subset V_0$. The only crucial point with respect to convergence is to get a discrete inf–sup condition. We must now show that for any $v_{0h} \in Q_{0h}$, we have with k independent of h and v_{0h}

$$(1.92)\qquad \sup_{\underline{q}_0 \in V_{0h}} \frac{b(\underline{q}_{0h}, v_{0h})}{\|\underline{q}_{0,h}\|} \geq k\|v_{0h}\|_{Q/\,\text{Ker}\,B_h^t}.$$

The correct situation is of course obtained for $\text{Ker}\,B_h^t \subset \text{Ker}\,B^t$. To prove (1.92) we shall try, as usual, to use Proposition II.2.8. To do so, $\underline{q}_0 \in V_0$ being given, we should be able to build $\underline{q}_{0h} = \Pi_h \underline{q}_0$ such that

$$(1.93)\qquad \begin{cases} b(\underline{q}_0 - \underline{q}_{0h}, v_{0h}) = 0, & \forall v_{0h} \in Q_{0h},\\ \|\underline{q}_{0h}\|_0 \leq C\|\underline{q}_0\|_0 \end{cases}$$

with a constant C independent of h. To get this result, we shall in fact build for any $\underline{q} \in V$, $\underline{q}_h = \Pi_h \underline{q}$, in such a way that $\Pi_h \underline{q} \in V_{0h}$ if $\underline{q} \in V_0$, and satisfying

$$(1.94)\qquad b(\underline{q} - \underline{q}_h, v_h) = 0, \qquad \forall v_h \in Q_h.$$

From the definition of $b(\cdot,\cdot)$, this will, a fortiori, hold whenever one has

$$(1.95)\qquad \int_{\partial K} (\underline{q} - \Pi_h \underline{q}) \cdot \underline{n}\, v_h\, ds = 0, \qquad \forall v_h \in V_h,\ \forall K \in \mathcal{T}_h.$$

Condition (1.95) is, however, nothing but a small linear system,

$$(1.96)\qquad \int_{\partial K} \underline{q}_h \cdot \underline{n}\, v_h\, ds = \int_{\partial K} \underline{q} \cdot \underline{n}\, v_h\, ds, \qquad \forall v_h \in T_{k+1}(\partial K).$$

We have again the same problem as in the previous section: solving (1.96) for $\underline{q}_h$ depends on the degree of polynomials at hand. As we shall see, the cure is, however, much simpler here. To fix ideas we shall therefore consider two simple examples.

Example 1.3: ($k = 0$ *triangular elements*).

This is the simplest case, and it is easily seen that system (1.96) can always be solved. Indeed the degrees of freedom of $M_0(K)$ are the constant values $q_i = (\underline{q}_h \cdot \underline{n})_i$ on each side e_i of lenght ℓ_i of K when $D_0 = P_0(K)$. System (1.96) takes the form (cf. Figure IV.4)

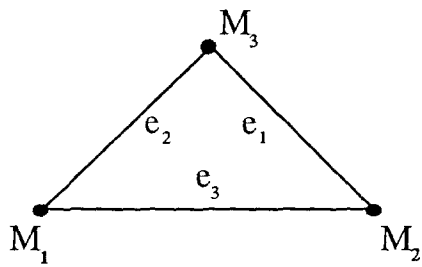


Figure IV.4

$$
\begin{aligned}
\tfrac{1}{2}[(q_2)\ell_2 + (q_3)\ell_3] &= \int_{\partial K} (\underline{q} \cdot \underline{n})\lambda_1 \, ds, \\
\tfrac{1}{2}[(q_3)\ell_3 + (q_1)\ell_1] &= \int_{\partial K} (\underline{q} \cdot \underline{n})\lambda_2 \, ds, \\
\tfrac{1}{2}[(q_2)\ell_2 + (q_1)\ell_1] &= \int_{\partial K} (\underline{q} \cdot \underline{n})\lambda_3 \, ds
\end{aligned}
\tag{1.97}
$$

which can always be solved. Moreover, (1.97) implies, by summing the three equations,

$$
q_1\ell_1 + q_2\ell_2 + q_3\ell_3 = \int_{\partial K} \underline{q}_h \cdot \underline{n} \, ds = \int_{\partial K} \underline{q} \cdot \underline{n} \, ds \tag{1.98}
$$

so that $\underline{q} \in V_0$ implies $\underline{q}_h \in V_{0h}$. □

Remark 1.11: Let us recall that any divergence-free function of $RT_k(K)$ is the curl of a stream function $\psi_h \in P_{k+1}$. If we want to check (1.96) only for divergence-free functions, which is sufficient to get the inf–sup condition, we can write, with $\partial\psi/\partial\tau$ denoting the tangential derivative of ψ on ∂K,

$$
\int_{\partial K} \frac{\partial \psi_h}{\partial \tau} v_h \, ds = \int_{\partial K} \frac{\partial \psi}{\partial \tau} v_h \, ds \quad \forall v_h \in T_{k+1}(\partial K), \tag{1.99}
$$

where ψ is the stream function associated to $\underline{q}$.

System (1.99) is then always singular, as for $v_h =$ constant on ∂K both sides vanish. For $k = 0$ (and all even k) the system can always be solved. For the case $k = 1$ of our next example, an extra linear dependence will appear among the equations. We refer to Lemma 5.2, where a similar situation will be encountered in the analysis of hybrid methods for fourth-order problems. □

Example 1.4: Case $k = 1$.

We now use the space $M_1(K) = RT_1(K)$ with $D_1 = P_1(K)$ (but this is not the only possible choice). The degrees of freedom of $\underline{q}_h$ are now given by two values (or moments) of the linear normal trace $\underline{q}_h \cdot \underline{n}$ on each side of K, plus two internal nodes which will be used to obtain the divergence-free condition on $\underline{q}_h$.

When trying to solve (1.96), we are again facing the same pathology that we had already met when studying primal hybrid methods: there exists a second-degree polynomial ϕ_K , which we already called the "non-conforming bubble" such that

$$\int_{\partial K} \underline{q}_h \cdot \underline{n}\, \phi_K \, ds = 0, \qquad \forall \underline{q}_h \in V_h \tag{1.100}$$

This implies that system (1.96) is not of maximal rank and cannot in general be solved for a general $\underline{q}$. Our *local construction* thus fails. It can, however be checked that $\operatorname{Ker} B_h^t \subseteq \operatorname{Ker} B^t$ because no function of Q_h can be built from nonconforming bubbles, satisfying Dirichlet conditions on a part of $\partial\Omega$. However, we know of no way to prove an inf–sup condition (if one holds).

The standard cure in such a situation is to use a richer space for V_0: we shall add to $M_1(K)$ one element of the next member of the family, that is, $RT_2(K)$. Let us then define $\psi_{3K} \in P_3(K)$ such that

$$\frac{\partial \psi_{3K}}{\partial \tau} = \phi_K . \tag{1.101}$$

We can now take $\psi_h \in P_2(K) \oplus$ span (ψ_{3K}) and the system (1.99) becomes of maximal rank and always has a solution. It is not unique and we may select the solution of minimal norm. □

Remark 1.12: Taking $\tilde{\underline{q}}_h = \underline{\operatorname{rot}}\, \psi_{3K}$ and working in the space $RT_1(K) \oplus$ span $\tilde{\underline{q}}_h$, we could have found a solution of (1.96) and made it divergence free using the internal nodes of $RT_1(K)$. □

The above examples are quite representative of situations generally encountered in hybrid methods: construction of approximations differ for odd or even degrees. Whenever a difficulty arises, enrichment of V_h can be used to cure the trouble. From a *computational point of view*, this enrichment is not troublesome, since degrees of freedom of $\underline{q}_h$ are *internal* to the element ($Y(\Omega)$ satisfies no continuity on interfaces). The standard practice is then to use "static condensation" and to reduce the problem to degrees of freedom in v_h. Dual hybrid methods can then be seen as a variant of standard conforming methods in which the shape of approximations inside K is not specified. As we shall see

in a latter part of this chaper, this will be very useful for higher-order problems where C^1 continuity is required in the construction of elements.

We have still a technical point to set. To apply Proposition II.2.8, we must show that Π_h is continuous, (uniformly in h) from V into V_{0h}. But this reduces to continuity in $L^2(K)$ as $\operatorname{div} \underline{q} = 0$ implies $\operatorname{div} \underline{q}_{0h} = 0$. This is easily obtained by a scaling argument or, equivalently, by transforming the problem to a reference element. We must, however, make this through the Piola transformation (III.1.45) to make $\hat{\underline{q}}_0$ divergence-free. Continuity of Π_h is then easily deduced from (III.1.51) and (III.1.52) with a standard condition on the shape of elements.

Remark 1.13: For an application of dual hybrid methods to the problem of the torsion of an elastic bar, the reader may refer to PIAN [A] and BREZZI [B] for the corresponding mathematical analysis. □

IV.2 Stokes Problem

We consider now another simple application of the abstract results of Chapter II. In particular we present a few examples of stable discretizations of the Stokes problem introduced in (I.3.11) and (II.1.31). Much more will be said on this problem in Chapter VI. Let us recall very briefly the notation and the results already obtained. We had

$$a(\underline{u}, \underline{v}) := 2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx,$$

$$b(\underline{v}, q) := -\int_\Omega q \operatorname{div} \underline{v}\, dx,$$

$$V = (H_0^1(\Omega))^2,\ Q = L^2(\Omega),\ \underline{f} \in (L^2(\Omega))^2 \text{ given}$$

and we were looking for $\underline{u} \in V$, $p \in Q$ such that

$$\text{(2.1)} \qquad \begin{cases} a(\underline{u}, \underline{v}) + b(\underline{v}, p) = (\underline{f}, \underline{v}), & \forall \underline{v} \in V, \\ b(\underline{u}, q) = 0, & \forall q \in Q. \end{cases}$$

We also saw that in this case Ker B^t is the one-dimensional subspace of constants and that (2.1) has a unique solution $(\underline{u}, p)$ in $V \times Q/\mathbb{R}$ (that is, the pressure is unique up to an additive constant). We have now to choose subspaces $V_h \subset V$ and $Q_h \subset Q$ so that some of the conditions of Chapter II can be applied. Our first attempt would be with the easiest choice

$$\text{(2.2)} \qquad V_h = (\mathcal{L}_1^1)^2 \cap (H_0^1(\Omega))^2,$$

$$\text{(2.3)} \qquad Q_h = \mathcal{L}_1^1.$$

However, it can be shown that this choice does not satisfy the inf–sup condition. Nevertheless, we can enrich (2.2) so that, in the end, the new choice will give a stable and convergent approximation to the Stokes problem (2.1). We set as in (III.2.22)

$$B_3 = \{b(x) \in H_0^1(\Omega),\ b(x)|_T \in P_3(T) \cap H_0^1(T),\ \forall T \in \mathcal{T}_h\}. \tag{2.4}$$

Hence each $b(x)$ of B_3 , on each triangle T, has the form $\alpha(T)\ \lambda_1(x)\ \lambda_2(x)\ \lambda_3(x)$ with $\alpha(T)$ constant in T. Following ARNOLD–BREZZI–FORTIN [A] we set

$$V_h := \{\mathcal{L}_1^1(\mathcal{T}_h) \oplus B_3\}^2 \tag{2.5}$$

and we want to show that (2.5) and (2.3) is a stable and convergent approximation of (2.1). For this we will apply Proposition II.2.8 with $W = V$ and $S = Q/\operatorname{Ker} B^t$ and we have therefore to construct an operator Π_h such that

$$\int_\Omega (\operatorname{div}(\underline{v} - \Pi_h \underline{v})\, q_h\, dx = 0,\ \forall q \in Q_h, \qquad \forall \underline{v} \in V, \tag{2.6}$$

$$||\Pi_h \underline{v}||_V \le c\, ||\underline{v}||_V, \qquad \forall \underline{v} \in V. \tag{2.7}$$

We shall use the technique of Proposition II.2.9 with $W = V$ so that we first take for Π_1 the operator r_h of Proposition III.2.1 and Corollary III.2.1. We, thus, set

$$\Pi_1 \underline{v}|_K = r_h \underline{v}|_K \tag{2.8}$$

which from (III.2.15) yields

$$|\underline{v} - \Pi_1 \underline{v}|_{m,K} \le c \Big(\sum_{\bar{K}' \cap \bar{K} \ne \emptyset} h_{K'}^{1-m} |\underline{v}|_{1,K'} \Big). \tag{2.9}$$

In particular, (2.9) implies the first condition of (II.2.30),

$$||\Pi_1 \underline{v}||_V \le c\, ||\underline{v}||_V. \tag{2.10}$$

We now define the operator $\Pi_2 : V \to (B_3)^2$ by means of

$$\int_\Omega \operatorname{div}(\Pi_2 \underline{v} - \underline{v}) q_h\, dx = \int_\Omega (\underline{v} - \Pi_2 \underline{v}) \cdot \underline{\operatorname{grad}}\, q_h\, dx = 0, \quad \forall q_h \in Q_h. \tag{2.11}$$

Since $\underline{\operatorname{grad}}\, q_h$ is piecewise constant, (2.11) is easily satisfied by choosing, in each K, bubbles with the same mean value as $\underline{v}$. It is easy to check that

$$||\Pi_2 \underline{v}||_{r,K} \le c h_K^{-r}\, ||\underline{v}||_{0,K}, \quad \forall \underline{v} \in V,\ r = 0, 1. \tag{2.12}$$

From (2.11) it is then immediate to check that the second condition of (II. 2.30) is fulfilled and from (2.12) and (2.9) we easily have the third condition.

We can thus apply Proposition II.2.9 and the inf–sup condition holds. Now we apply Theorem II.2.1 and we obtain

$$\|\underline{u} - \underline{u}_h\|_V + \|p - p_h\|_{Q/\mathbb{R}} \leq ch\,(\|\underline{u}\|_{2,\Omega} + |p|_1), \tag{2.13}$$

that is, an optimal error estimate.

We now analyze another possible simple discretization of the linear Stokes problem. In particular, we try to use a piecewise constant pressure field. It is easy to see that taking

$$V_h = (\mathcal{L}_2^1 \cap H_0^1(\Omega))^2, \tag{2.14}$$

$$Q_h = \mathcal{L}_0^0, \tag{2.15}$$

we can again apply (II.2.9) and define $\Pi_h : V \to V_h$ by

$$\Pi_h \underline{v} = \Pi_1 \underline{v} + \Pi_2(\underline{v} - \Pi_1 \underline{v}), \tag{2.16}$$

where Π_1 is still chosen as in the previous example and where $\Pi_2 : V \to V_h$ is defined by

$$\Pi_2 \underline{w} = 0 \text{ at the vertices; } \int_e (\Pi_2 \underline{w} - \underline{w})\, ds = 0, \quad \forall e \in \mathcal{E}_h. \tag{2.17}$$

It is now elementary to check that

$$\int_K \operatorname{div}(\underline{v} - \Pi_2 \underline{v}) q\, dx = \int_{\partial K} (\underline{v} - \Pi_2 \underline{v}) \cdot \underline{n}\, q\, ds = 0 \tag{2.18}$$

for all $q \in \mathcal{L}_0^0$ and that

$$\|\Pi_2 \underline{v}\|_{1,K} \leq ch_K^{-1} \|v\|_{0,K}, \quad \forall K \in \mathcal{T}_h. \tag{2.19}$$

Hence the inf–sup condition will be satisfied and again we can apply Theorem II.2.1 and get

$$\|\underline{u} - \underline{u}_h\|_V + \|p - p_h\|_{Q/\mathbb{R}} \leq ch\,(\|\underline{u}\|_{2,\Omega} + |p|_1). \tag{2.20}$$

A third possibility would be to use a *nonconforming* approximation of V. For instance we may choose again $Q_h = \mathcal{L}_0^0$ and

(2.21)

$$V_h := \{v_h \in (\mathcal{L}^{1,NC}(P_1, \mathcal{T}_h))^2, \text{ vanishing at the boundary midpoints}\}$$

$$a_h(\underline{u}_h, \underline{v}_h) = \sum_K \int_K \underline{\underline{\text{grad}}}\, \underline{u}_h : \underline{\underline{\text{grad}}}\, \underline{v}_h \, dx, \tag{2.22}$$

$$b_h(\underline{v}_h, q_h) = \sum_K \int_K \text{div}\, \underline{v}_h q_h \, dx, \tag{2.23}$$

and consider the problem

$$a_h(\underline{u}_h, \underline{v}_h) + b_h(\underline{v}_h, p_h) = (\underline{f}, \underline{v}_h), \quad \forall \underline{v}_h \in V_h, \tag{2.24}$$

$$b_h(\underline{u}_h, q_h) = 0, \qquad \forall q_h \in \mathfrak{L}_0^0. \tag{2.25}$$

We may now construct $\Pi_h : V \to V_h$ by

$$\int_{\partial K} (\Pi_h \underline{v} - \underline{v}) \cdot \underline{\phi} \, ds = 0, \qquad \forall \underline{\phi} \in R_0(\partial K), \tag{2.26}$$

and again it is easy to see that

$$|\Pi_h \underline{v}|_{1,h} \le c\, |\underline{v}|_{1,h}, \tag{2.27}$$

(where as usual $|v_h|_{1,h}^2 = \sum_k |v_h|_{1,k}^2$) and

$$b_h(\underline{v} - \Pi_h \underline{v}, q_h) = 0, \quad \forall q_h \in \mathfrak{L}_0^0, \tag{2.28}$$

which implies, by Proposition II.2.17,

$$\inf_{q \in Q_h/\mathbb{R}} \sup_{\underline{v} \in V_h} \frac{b_h(\underline{v}, q)}{|\underline{v}|_{1,h} \|q\|_{Q/\mathbb{R}}} \ge c > 0. \tag{2.29}$$

On the other hand, we also have

$$a_h(\underline{v}_h, \underline{v}_h) \ge \alpha \, \|\underline{v}_h\|_{1,h}^2, \qquad \forall v_h \in V_h. \tag{2.30}$$

We may now apply Proposition II.2.16 and get

$$|\underline{u} - \underline{u}_h|_{1,h} + \|p - p_h\|_{Q,\mathbb{R}} \le ch + E_h(\underline{u}, p), \tag{2.31}$$

where

$$\begin{aligned} E_h(\underline{u}, p) &= \sup_{\underline{v}_h \in V_h} |\underline{v}_h|_{1,h}^{-1} \{ a_h(\underline{u}, \underline{v}_h) + b_h(\underline{v}_h, p) - (\underline{f}, \underline{v}_h) \} \\ &= \sup_{\underline{v}_h \in V_h} |\underline{v}_h|_{1,h}^{-1} \sum_K \int_{\partial K} [(\underline{\text{grad}}\, \underline{u}) \cdot \underline{n}] \cdot \underline{v}_h \, ds \\ &\le ch \, \|\underline{u}\|_{2,\Omega} \end{aligned} \tag{2.32}$$

so that in the end we have the optimal estimate

$$|\underline{u} - \underline{u}_h|_{1,h} + \|p - p_h\|_{Q/\mathbb{R}} \le ch \, \|\underline{u}\|_{2,\Omega}.$$

We shall present many other approximations of Stokes problems in Chapter VI.

IV.3 Elasticity Problems

In our list of quick applications of the above abstract theory, we consider now some simple applications to linear elasticity problems. If we set $p = \lambda \operatorname{div} \underline{u}$,

$$a(\underline{u}, \underline{v}) = \int_\Omega 2\mu\, \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx,$$

$$b(\underline{u}, q) = \int_\Omega \operatorname{div} \underline{u}\, q\, dx,$$

$$V = (H_0^1(\Omega))^2, \quad Q = L^2(\Omega),$$

with the notation of Example I.2.2, we may write (I.2.21) as

$$\begin{aligned} a(\underline{u}, \underline{v}) + b(\underline{v}, p) &= (\underline{f}, \underline{v}), \qquad \forall \underline{v} \in V, \\ b(\underline{u}, q) &= \frac{1}{\lambda}(p, q), \qquad \forall q \in Q, \end{aligned} \tag{3.1}$$

(for the sake of simplicity we took $\Gamma_0 \equiv \partial\Omega$). When λ is "not to large," (3.1) has obviously a unique solution by the Lax–Milgram theorem and Korn's first inequality, that is,(cf.DUVAUT–LIONS [A])

$$\int_\Omega |\underline{\underline{\varepsilon}}(\underline{v})|^2\, dx \geq \alpha \|\underline{v}\|_{1,\Omega}^2, \qquad \forall \underline{v} \in (H_0^1(\Omega))^2. \tag{3.2}$$

Existence of a solution of (3.1), actually, is still true for any finite value of λ ; however, it is clear that, for λ larger and larger, we have to deal with a problem of the type (II.4.7), where now $1/\lambda$ plays the role of ε. Following the results of Proposition II.4.1 we see that *in order to deal with a linear elasticity problem where λ is large (nearly incompressible case) we must choose spaces V_h and Q_h which satisfy the inf–sup condition.* For instance with the choice (2.5) and (2.3) we have

$$\|\underline{u} - \underline{u}_h\|_V + \|p - p_h\|_Q \leq ch\, (\|\underline{u}\|_2 + |p|_1)$$

with c independent of h *and* λ. □

We want now to see what can be done in a *truly mixed* approach to (I.2.21), that is using

$$\Sigma = \underline{\underline{H}}(\underline{\operatorname{div}}; \Omega)_S, \quad U = (L^2(\Omega))^2, \tag{3.3}$$

$$a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) := \int_\Omega \left[\frac{1}{2\mu} \underline{\underline{\sigma}}^D : \underline{\underline{\tau}}^D + \frac{1}{(\lambda + \mu)} \operatorname{tr}(\underline{\underline{\sigma}})\ \operatorname{tr}(\underline{\underline{\tau}})\right] dx, \tag{3.4}$$

$$b(\underline{\underline{\tau}}, \underline{v}) := \int_\Omega \underline{\operatorname{div}}(\underline{\underline{\tau}}) \cdot \underline{v}\, dx. \tag{3.5}$$

Note that here we are using (Σ, U) instead of (V, Q) as in Chapter II. It is clear that, taking as in (2.5),

$$\Sigma_h = (\mathcal{L}_1^1 \oplus B_3)_s^4, \tag{3.6}$$

$$U_h = (\mathcal{L}_1^1)^2, \tag{3.7}$$

we have

$$a(\underline{\underline{\tau}}, \underline{\underline{\tau}}) \geq \frac{1}{2(\lambda+\mu)} \|\underline{\underline{\tau}}\|_0^2 \tag{3.8}$$

and that we may construct $\Pi_h : \Sigma \to \Sigma_h$ such that

$$b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, \underline{v}) = 0, \quad \forall \underline{q} \in U_h. \tag{3.9}$$

To see this last point, note that

$$b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, \underline{v}) = \int_\Omega \operatorname{div}(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}) \cdot \underline{v}\, dx = -\int_\Omega (\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx, \tag{3.10}$$

and that $\underline{\underline{\varepsilon}}(\underline{v})$ can be represented by three (independent) constants on each K. Hence we may proceed as in (2.12), defining

$$\Pi_h \underline{\underline{\tau}} = \Pi_1 \underline{\underline{\tau}} + \Pi_2(\underline{\underline{\tau}} - \Pi_1 \underline{\underline{\tau}}), \tag{3.11}$$

where Π_1 and Π_2 are defined as in (2.8) and (2.11). It is also clear that Π_h as given in (3.11) verifies

$$\|\Pi_h \underline{\underline{\tau}}\|_0 \leq c \|\underline{\underline{\tau}}\|_0, \tag{3.12}$$

so that using (3.9) and (3.12) we have as usual, using Proposition II.2.8, that

$$\inf_{\underline{\phi} \in U_h} \sup_{\underline{\underline{\tau}} \in \Sigma_h} \frac{b(\underline{\underline{\tau}}, \underline{\phi})}{\|\underline{\underline{\tau}}\|_0 |\underline{\phi}|_1} \geq c > 0. \tag{3.13}$$

Note that here we are using $\Sigma = (L^2(\Omega))_s^4$ and $U = (H_0^1(\Omega))^2$; this does not correspond to a *truly mixed* approach, but is allowed here due to the choice (3.7). Now from (3.8) and (3.13) we have the existence and uniqueness of the solution of the discretized problem and, from Theorem II.2.1, the usual error bounds

$$\begin{aligned} &\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}^h\|_0 + \|\underline{u} - \underline{u}^h\|_1 \\ &\qquad \leq c \left\{ \inf_{\underline{\underline{\tau}} \in \Sigma_h} \|\underline{\underline{\sigma}} - \underline{\underline{\tau}}\|_0 + \inf_{\underline{\phi} \in U_h} \|\underline{u} - \underline{\phi}\|_1 \right\} \leq c_1 h \|\underline{u}\|_2. \end{aligned} \tag{3.14}$$

A simple duality argument (as in Section II.2.7) would also show that

$$\|\underline{u} - \underline{u}^h\|_0 \leq c h^2 \|\underline{u}\|_2. \tag{3.15}$$

However, as we have noticed before, the troubles arise when we have to deal with a very large λ (nearly incompressible materials). In fact, it is clear from (3.8) that the constant c which appears in (3.14) goes like λ when $\lambda \to +\infty$ and the quality of the approximation is jeopardized. Actually, the situation is not as bad as it seems, because, as in Proposition II.2.4, we do not need (3.8) to hold for every $\underline{\underline{\tau}} \in \Sigma$ (or Σ_h) but only for $\underline{\underline{\tau}} \in \operatorname{Ker} B$ (respectively, $\operatorname{Ker} B_h$). In particular, the continuous formulation (I.3.48) does not break down when $\lambda \to \infty$ because of the following proposition.

Proposition 3.1: There exists a constant $c > 0$ such that, for every $\underline{\underline{\tau}} \in (L^2(\Omega))^4_s$ satisfying

$$\int_\Omega \operatorname{tr}(\underline{\underline{\tau}})\, dx = 0, \tag{3.16}$$

we have

$$\|\underline{\underline{\tau}}\|_0 \leq c\,(\|\underline{\underline{\tau}}^D\|_0 + \|\underline{\operatorname{div}}\,\underline{\underline{\tau}}\|_0). \tag{3.17}$$

Proof: It is clear that it is enough to show that

$$\|tr(\underline{\underline{\tau}})\|_0 \leq c\,(\|\underline{\underline{\tau}}^D\|_0 + \|\underline{\operatorname{div}}\,\underline{\underline{\tau}}\|_0). \tag{3.18}$$

For this, note that (3.16) implies the existence of a $\underline{v} \in (H_0^1)^2$ such that

$$\operatorname{div} \underline{v} = \operatorname{tr}(\underline{\underline{\tau}}), \tag{3.19}$$

$$\|\underline{v}\|_1 \leq c\,\|\operatorname{tr}(\underline{\underline{\tau}})\|_0. \tag{3.20}$$

Now from (3.19) and (3.16) we have

$$\left\{\begin{aligned} \|\operatorname{tr}(\underline{\underline{\tau}})\|_0^2 &= \int_\Omega tr(\underline{\underline{\tau}}) \operatorname{div} \underline{v}\, dx \\ &= \int_\Omega \underline{\underline{\tau}} : \underline{\underline{\delta}} \operatorname{div} \underline{v}\, dx \\ &= 2\int_\Omega \underline{\underline{\tau}} : (\underline{\underline{\operatorname{grad}}}\, \underline{v} - (\underline{\underline{\operatorname{grad}}}\, \underline{v})^D)\, dx \\ &= -2\int_\Omega \underline{\underline{\tau}}^D : \underline{\underline{\operatorname{grad}}}\, \underline{v}\, dx - 2\int_\Omega \underline{\operatorname{div}}\, \underline{\underline{\tau}} \cdot \underline{v}\, dx \\ &\leq 2\|\underline{\underline{\tau}}^D\|_0\, \|\underline{v}\|_1 + 2\|\underline{\operatorname{div}}\, \underline{\underline{\tau}}\|_0\, \|v\|_0 \end{aligned}\right. \tag{3.21}$$

and from (3.20) and (3.20) we get (3.18). □

It is clear that any stationary point of (I.3.48) must satisfy (3.16). If we work now in the subspace

$$\tilde{\Sigma} = \{\underline{\underline{\tau}} \,|\, \underline{\underline{\tau}} \in \Sigma,\ \int_\Omega \operatorname{tr}(\underline{\underline{\tau}})\, dx = 0\}, \tag{3.22}$$

we know that the set

$$\operatorname{Ker} B = \{\underline{\underline{\tau}} \,|\, \underline{\underline{\tau}} \in \tilde{\Sigma},\ b(\underline{\underline{\tau}}, \underline{v}) = 0\ \forall \underline{v} \in U\} \tag{3.23}$$

is precisely made of tensors satisfying (3.16) and

$$\underline{\operatorname{div}}\, \underline{\underline{\tau}} = 0. \tag{3.24}$$

Hence, from Proposition 3.1 we have

$$a(\underline{\underline{\tau}},\underline{\underline{\tau}}) \geq c(\mu)\,\|\underline{\underline{\tau}}\|_0^2 = c(\mu)\|\tau\|^2_{\underline{\underline{H}}(\mathrm{div},\Omega)_s}, \quad \forall \underline{\underline{\tau}} \in \operatorname{Ker} B. \tag{3.25}$$

If we had now $\operatorname{Ker} B_h \subseteq \operatorname{Ker} B$ we would get in (3.14) a constant c independent of λ. We might think to prove (3.18) for $\underline{\underline{\tau}}_h \in \operatorname{Ker} B_h$.

Unfortunately we can construct examples of $\underline{\underline{\tau}}_h \in \Sigma_h$ such that we have $\int \operatorname{tr}(\underline{\underline{\tau}}_h)\,dx = 0$ and

$$b(\underline{\underline{\tau}}_h, \underline{v}_h) = 0, \quad \forall \underline{v} \in U_h, \tag{3.26}$$

but $\underline{\underline{\tau}}_h^D \equiv 0$ and $\operatorname{tr}(\tau_h) \neq 0$. For instance, on a uniform mesh, take $\tau_{h11} - \tau_{h22} = \pm\phi$ (the unit bubble in each K) and $\tau_{h12} \equiv 0$ (the sign $\pm$ of ϕ changing on any pair of adjacent triangles). Globally, one has $\int \operatorname{tr}(\underline{\underline{\tau}}_h)\,dx = 0$ due to the alternating signs. It is also easy to see that (3.26) is satisfied. Hence, (3.17) is not true, in our case, for $\underline{\underline{\tau}}_h \in \operatorname{Ker} B_h$. We see again that the property $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ (that here is unfortunately false) is a very useful one.

The problem that we face is essentially due to the fact that the bilinear form $a(\cdot,\cdot)$ is not coercive on the whole space $\underline{\underline{H}}(\underline{\mathrm{div}},\Omega)_s$, but only on the subspace $\operatorname{Ker} B$ defined by (3.23). We shall now show that using the results of Section I.5, that is by modifying the variational formulation, we can obtain an approximate solution with error bounds independant of λ.

Let us indeed consider instead of (I.3.48) the saddle-point problem,

$$\left\{\begin{aligned} &\inf_{\underline{\underline{\sigma}}}\sup_{\underline{v}} \frac{1}{4\mu}\int_\Omega |\underline{\underline{\sigma}}^D|^2 dx + \frac{1}{2(\lambda+\mu)}\int_\Omega |\operatorname{tr}\underline{\underline{\sigma}}|^2 dx + \int_\Omega (\underline{\mathrm{div}}\,\underline{\underline{\sigma}}+\underline{f})\cdot v\,dx \\ &\qquad\qquad + \frac{\alpha}{2}\int_\Omega |\underline{\mathrm{div}}\,\underline{\underline{\sigma}}+\underline{f}|^2 dx, \quad \alpha \geq 0, \end{aligned}\right. \tag{3.27}$$

for which the optimality conditions are now

$$\left\{\begin{aligned} &\int_\Omega \frac{1}{2\mu}\underline{\underline{\sigma}}^D : \tau^D\,dx + \frac{1}{(\lambda+\mu)}\int_\Omega \operatorname{tr}\underline{\underline{\sigma}}\ \operatorname{tr}\underline{\underline{\tau}}\,dx + \alpha\int_\Omega (\underline{\mathrm{div}}\,\underline{\underline{\sigma}}+\underline{f})\cdot(\underline{\mathrm{div}}\,\underline{\underline{\tau}})\,dx \\ &\qquad\qquad + \int_\Omega (\underline{\mathrm{div}}\,\underline{\underline{\tau}})\cdot \underline{u}\,dx = 0, \quad \forall \underline{\underline{\tau}} \in (\underline{\underline{H}}(\underline{\mathrm{div}},\Omega)_s, \end{aligned}\right. \tag{3.28}$$

$$\int_\Omega (\underline{\mathrm{div}}\,\underline{\underline{\sigma}} + \underline{f})\cdot \underline{v}\,dx = 0 \quad \forall \underline{v} \in (L^2(\Omega))^2. \tag{3.29}$$

But (3.28) and (3.29) are clearly equivalent to the original formulation. However, we now have, instead of (3.4),

(3.30)

$$a(\underline{\underline{\sigma}},\underline{\underline{\tau}}) = \frac{1}{2\mu}\int_\Omega \underline{\underline{\sigma}}^D : \underline{\underline{\tau}}^D\,dx + \frac{1}{\lambda+\mu}\int_\Omega \operatorname{tr}\underline{\underline{\sigma}}\ \operatorname{tr}\underline{\underline{\tau}}\,dx + \alpha\int_\Omega \underline{\mathrm{div}}\,\underline{\underline{\sigma}}\cdot\underline{\mathrm{div}}\,\underline{\underline{\tau}}\,dx$$

and from Proposition 3.1, we get the coerciveness property

$$a(\underline{\underline{\sigma}},\underline{\underline{\sigma}}) \geq \alpha_0 \|\underline{\underline{\sigma}}\|^2_{\underline{\underline{H}}(\underline{\mathrm{div}},\Omega)}, \tag{3.31}$$

where α_0 depends on α, μ, and c but is independant of λ. On the other hand the bilinear form $b(\underline{\underline{\tau}},\underline{v})$ defined by (3.5) is unchanged but we now need the inf–sup condition (cf. ARNOLD–DOUGLAS–GUPTA [A])

$$\inf_{\underline{v}\in(L^2(\Omega))^2} \sup_{\underline{\underline{\tau}}\in(H^1(\Omega))^4_s} \frac{b(\underline{\underline{\tau}},\underline{v})}{\|\tau\|_1 \|v\|_0} \geq k > 0. \tag{3.32}$$

Now, we introduce the discretization already defined by (3.6) and (3.7). As we now have a coerciveness property on the whole space Σ, the only delicate point is to obtain a discrete inf–sup condition. We use Proposition II.2.8 and the operator Π_h defined by (3.11). Indeed as in section IV.2, we deduce,

$$\|\Pi_h \underline{\underline{\tau}}\|_\Sigma \leq c\,\|\underline{\underline{\tau}}\|_1. \tag{3.33}$$

But (3.32) and (3.33) imply, by Proposition II.2.8, that we have

$$\inf_{\underline{v}_h\in V_h} \sup_{\underline{\underline{\tau}}\in\Sigma_h} \frac{b(\underline{\underline{\tau}},\underline{v}_h)}{\|\underline{\underline{\tau}}_h\|_\Sigma |\underline{v}|_0} \geq k_0 > 0 \tag{3.34}$$

with k_0 independant of h. From the standard theory, we therefore obtain an error estimate

$$\begin{aligned} \|\underline{\underline{\sigma}}-\underline{\underline{\sigma}}_h\|_\Sigma + \|\underline{u}-\underline{u}_h\|_0 &\leq C\left\{\inf_{\tau_h\in\Sigma_h}\|\sigma-\tau_h\|_\Sigma + \inf_{v_h\in V_h}\|u-v_h\|_0\right\}. \\ &\leq Ch(\|\underline{u}\|_2 + \|\underline{\underline{\sigma}}\|_2). \end{aligned} \tag{3.35}$$

Augmented formulations, therefore, appear as a powerful tool to overcome difficulties associated with problems of coerciveness and enable us to bypass the inclusion of kernels property which is very difficult to obtain in practice. We shall present in Chapter V examples where one can also employ similar arguments to avoid the inf–sup condition. Examples of applications to elasticity problems can be found in FRANCA–STENBERG [A] and in BREZZI–FORTIN–MARINI [A].

We shall see other examples of elements (for linear elasticity) which can treat the nearly incompressible case in Chapter VII. Other variational formulations of the elasticity problem can be found in STENBERG [B].

IV.4 A Mixed Fourth-Order Problem

IV.4.1 The $\psi - \omega$ biharmonic problem

Let us now see, as a new example of application of the abstract results of Chapter II, some simple cases of fourth-order problems. We shall start with formulation (I.3.54) which we may now rewrite in the form (II.2.1) by setting

$$V = H^1(\Omega), \quad Q = H_0^1(\Omega), \tag{4.1}$$

$$a(\omega, \phi) = \int_\Omega \omega\phi \, dx, \quad \forall \omega, \phi \in H^1(\Omega), \tag{4.2}$$

$$b(\mu, \phi) = \int_\Omega \underline{\text{grad}}\, \mu \cdot \underline{\text{grad}}\, \phi \, dx =, \quad \forall \mu \in H_0^1(\Omega), \phi \in H^1(\Omega). \tag{4.3}$$

We shall denote by (ω, ψ) instead of (u, p) the solution of the problem in order to be consistent with the usual physical notations. It is easy to see that we are now in the situation of Section II.2.5: the bilinear form $a(\omega, \phi)$ is not coercive on V (nor is it on Ker B but only on $H = L^2(\Omega)$). A loss of accuracy is therefore to be expected. Another pitfall is that we cannot use the abstract existence results of Chapter II for the continuous problem and that we must deduce the existence of a solution through another channel. In the present case we know that the solution of our mixed problem: find $\psi \in H_0^1(\Omega)$ and $\omega \in H^1(\Omega)$ such that

$$\left\{ \begin{aligned} &\int_\Omega \omega\phi \, dx + \int_\Omega \underline{\text{grad}}\, \psi \cdot \underline{\text{grad}}\, \phi \, dx = 0, \quad \forall \phi \in H^1(\Omega), \\ &\int_\Omega \underline{\text{grad}}\, \omega \cdot \underline{\text{grad}}\, \mu \, dx = \int_\Omega f\mu \, dx, \quad \forall \mu \in H_0^1(\Omega) \end{aligned} \right. \tag{4.4}$$

should be a solution of a biharmonic problem

$$\Delta^2 \psi = f, \quad \psi \in H_0^2(\Omega). \tag{4.5}$$

From a regularity result on the biharmonic problem, we know, for instance if Ω is a convex polygon (LIONS–MAGENES [A], TEMAM [A], GRISVARD [A]), that for $f \in H^{-1}(\Omega)$, the solution of (4.5) belongs to $H^3(\Omega)$, so that $\omega = -\Delta\psi$ belongs to $H^1(\Omega)$. It is then direct to verify that we have thus obtained a solution of (4.4). This is an example of an "ill-posed" mixed problem. It should be remarked that the discussion of existence made above does not apply when the right-hand side of the first equation of (4.4) is not zero.

To get a discrete problem we take, in the notation of Chapter III, we set

$$V_h = \mathfrak{L}_k^1, \quad Q_h = \mathfrak{L}_k^1 \cap H_0^1(\Omega), \; k \geq 2. \tag{4.6}$$

The case $k = 1$ requires a more special analysis (FIX–GUNZBURGER–NICOLAIDES [A], SCHOLZ [C], GLOWINSKI [A]). We then have that the constant

$S(h)$, appearing in (II.2.51), can now be bounded by $S(h) \leq ch^{-1}$ so that a direct application of Proposition II.2.13 gives

$$\|\omega - \omega_h\|_0 + \|\psi - \psi_h\|_1 \leq ch^{k-1}. \tag{4.7}$$

Indeed, the inf–sup condition is quite straightforward. The operator B is nothing here but the Laplace operator from $H^1(\Omega)$ to $H^{-1}(\Omega)$, which is obviously surjective. To check the discrete condition we use the criterion of Proposition II.2.8; given $\omega \in H^1(\Omega)$ we want to built $\omega_h \in V_h$ such that

$$\int_\Omega \underline{\mathrm{grad}}\,\omega_h \cdot \underline{\mathrm{grad}}\,\mu_h\,dx = \int_\Omega \underline{\mathrm{grad}}\,\omega \cdot \underline{\mathrm{grad}}\,\mu_h\,dx, \quad \forall \mu_h \in Q_h. \tag{4.8}$$

We recall, however, that we have chosen $Q_h \subset V_h$ so that (4.8) will, a fortiori, hold if we take $\mu_h \in V_h$. But (4.8) is then nothing but a discrete Neumann problem for which a solution exists and can be chosen (it is defined up to an additive constant) so that

$$\|\omega_h\|_1 \leq c\,\|\omega\|_1. \tag{4.9}$$

It must be noted that the condition $Q_h \subset V_h$ is essential to the above result. In practice this is not a restriction as (4.6) is a natural and efficient choice. Result (4.7) is far from optimal and may suggest at a first sight that the method is not worth using. It can however be sharpened in two ways. First it is possible to raise the estimate on $|\omega - \omega_h|_0$ by half an order (SCHOLZ [D], FIX–GUNZBURGER–NICOLAIDES [A]) by a quite intricate analysis using L^∞-error estimates. The second way is a more direct variant of the duality method of Section II.2.7 and shows that the expected accuracy can be obtained for $\psi \in H^3(\Omega)$, that is,

$$\|\psi - \psi_h\|_1 \leq ch^k, \tag{4.10}$$

and under a supplementary regularity assumption

$$\|\psi - \psi_h\|_0 \leq ch^{k+1}. \tag{4.11}$$

We refer the reader to SCHOLZ [A,D], FALK [A], BRAMBLE–FALK [A] and FALK–OSBORN [A] for this analysis.

On the other hand, the particular structure of problem (4.4) allows the use of sophisticated but effective techniques for the numerical resolution, (cf. CIARLET–GLOWINSKI [A],GLOWINSKI [B], GLOWINSKI–PIRONNEAU [A]) so that this method and its variants have a considerable practical interest. In fact it provides a correct setting for the widely used $\psi - \omega$ approximations in numerical fluid dynamics. We refer to GIRAULT–RAVIART [A] for more informations on this subject.

Still in the case of fourth-order problems we could also consider instead formulation (I.3.59) which is more related to plate bending problems. We now set

$$V = (H^1(\Omega))^4_s, \quad Q = H^1_0(\Omega) \tag{4.12}$$

and we define, following (I.3.59) for $\underline{\underline{\sigma}}$ and $\underline{\underline{\tau}}$ in V,

$$a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) = \frac{12(1-\nu^2)}{Et^3} \int_\Omega [(1+\nu)]\, \underline{\underline{\sigma}} : \underline{\underline{\tau}} - \nu \operatorname{tr}(\underline{\underline{\sigma}}) \operatorname{tr}(\underline{\underline{\tau}})]\, dx. \tag{4.13}$$

In order to consider a weaker form of the saddle point problem (I.3.59), we introduce

$$b(v, \underline{\underline{\tau}}) = \int_\Omega (\underline{\operatorname{div}}\, \underline{\underline{\tau}}) \cdot \underline{\operatorname{grad}}\, v\, dx = \int_\Omega \sum_{i,j} \frac{\partial \tau_{ij}}{\partial x_j} \frac{\partial v}{\partial x_i} dx. \tag{4.14}$$

This enables us to look for $w \in H^1_0(\Omega)$ instead of $H^2_0(\Omega)$, the second boundary condition being implied by this variational formulation as a natural condition.

This is again an "ill-posed" mixed problem: we must obtain the existence of a solution through a regularity result on the standard problem.

Two approaches have been followed in the approximation of this mixed problem. One of them consists in taking (MIYOSHI [A]):

$$V_h = (\mathcal{L}^1_k)^4_s, \quad Q_h = \mathcal{L}^1_k \cap H^1_0(\Omega). \tag{4.15}$$

With respect to (4.14) it is, however, possible to use a second approach and to work not in $V = (H^1(\Omega))^4_s$ but in the weaker space (I.3.43)

$$\underline{\underline{H}}(\underline{\operatorname{div}} : \Omega)_s = \{\underline{\underline{\tau}} \mid \tau_{ij} = \tau_{ji},\ \tau_{ij} \in L^2(\Omega),\ \underline{\operatorname{div}}\, \underline{\underline{\tau}} \in (L^2(\Omega))^2\}. \tag{4.16}$$

Discretizations of this space can be built through composite elements as we shall see in Section VII.2 (JOHNSON–MERCIER [A], ARNOLD–DOUGLAS–GUPTA [A]) .

In the first case the results are the same as for the $\psi - \omega$ approximation discussed above. We get, by Proposition II.2.13 an error estimate which is (h^{k-1}). Duality methods (FALK–OSBORN [A]) enable one to lift the estimate on ψ at the right level. For the second case we can have optimal error estimates (see the above references).

IV.5 Dual Hybrid Methods for Plate Bending Problems

We consider now as a final example an application of our general theory to hybrid methods. We go back again to Example 3.8 of Chapter I and set, for the sake of simplicity, $\nu = 0$ and $Et^3/12 = 1$. The consideration of the true values would not change the mathematical structure of the problem, but would result in more lengthy formulas. The condition $D_2^*(\underline{\underline{\tau}}) = f$ in (I.3.63) is, in general, difficult to enforce directly. Hence, following PIAN–TONG [A], we may think of working with stresses satisfying $D_2^*(\underline{\underline{\tau}}) = f$ inside each element of a given decomposition. This will imply that we have to enforce some continuity of the stresses by means of a Lagrangian multiplier; moreover it will be convenient to assume $f \in L^2(\Omega)$. In order to make the exposition clearer, we need some Green's formula. We have indeed, on any triangle K of a triangulation $\mathcal{T}_h$ of Ω,

$$\int_K \underline{\underline{\tau}} : \underline{\underline{D}}_2(v)\, dx = \int_K D_2^*(\underline{\underline{\tau}}) v\, dx + \int_{\partial K} \Big(M_{nn}(\underline{\underline{\tau}}) \frac{\partial v}{\partial n} - \mathcal{K}_n(\underline{\underline{\tau}}) v \Big)\, ds \tag{5.1}$$

for all $\underline{\underline{\tau}} \in (H^2(K))^4_s$ and $v \in H^2(K)$, where,

$$M_{nn}(\underline{\underline{\tau}}) := (\underline{\underline{\tau}}\ \ \underline{n})\ \ \underline{n}, \tag{5.2}$$

$$\mathcal{K}_n(\underline{\underline{\tau}}) := \frac{\partial}{\partial n} \operatorname{tr}(\underline{\underline{\tau}}) - \frac{\partial}{\partial t}[(\underline{\underline{\tau}}\ \ \underline{n}) \cdot \underline{t}], \quad \underline{t} = \text{tangent unit vector.} \tag{5.3}$$

It is essential, in the definition of $\mathcal{K}_n$, to consider the derivative $\partial/\partial t$ *in the distributional sense*, that is, to take into account the *jumps* of $(\underline{\underline{\tau}} \cdot \underline{n}) \cdot \underline{t}$ at the corners of K (the so-called *corner forces*).

It is easy to check that the condition $D_2^*(\underline{\underline{\tau}}) = f$ in Ω is equivalent to

$$\begin{cases} D_2^*(\underline{\underline{\tau}}) = f \text{ in each } K, \\ \displaystyle\sum_K \int_{\partial K} [M_{nn}(\underline{\underline{\tau}}) \frac{\partial v}{\partial n} - \mathcal{K}_n(\underline{\underline{\tau}}) v]\, ds = 0, \ \forall v \in H_0^2(\Omega) \end{cases} \tag{5.4}$$

Setting

$$\begin{aligned} b(\underline{\underline{\tau}}, v) &:= \sum_K \int_{\partial K} \Big(M_{nn}(\underline{\underline{\tau}}) \frac{\partial v}{\partial n} - \mathcal{K}_n(\underline{\underline{\tau}}) v \Big)\, ds \\ &\equiv \int_\Omega \underline{\underline{\tau}} : \underline{\underline{D}}_2(v)\, dx - \sum_K \int_K D_2^*(\underline{\underline{\tau}}) v\, dx, \end{aligned} \tag{5.5}$$

$$V_f(\mathcal{T}_h) = \{\underline{\underline{\tau}} \in (L^2(\Omega))^4_s,\ D_2^*(\underline{\underline{\tau}}) = f \text{ in each } K\}, \tag{5.6}$$

the problem can now be written

$$\inf_{\underline{\underline{\tau}} \in V_f(\mathcal{T}_h)} \sup_{v \in H_0^2} \tfrac{1}{2}\|\underline{\underline{\tau}}\|_0^2 - b(\underline{\underline{\tau}}, v). \tag{5.7}$$

If now $\underline{\underline{\sigma}}^f$ is a given element of $V_f(\mathcal{T}_h)$, that is, a *particular solution* of $D_2^*(\underline{\underline{\sigma}}) = f$ in each K, we have

$$\begin{cases} (\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^f, \underline{\underline{\tau}}) - b(\underline{\underline{\tau}}, w) = 0, & \forall \underline{\underline{\tau}} \in V_0(\mathcal{T}_h), \\ b(\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^f, v) = 0, & \forall v \in H_0^2(\Omega), \end{cases} \tag{5.8}$$

where obviously $\underline{\underline{\sigma}}^0 + \underline{\underline{\sigma}}^f := \underline{\underline{\sigma}}$. Problem (5.8) has now the form (II.1.5), where $V = V_0(\mathcal{T}_h)$, $Q = H_0^2(\Omega)$, $a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) = (\underline{\underline{\sigma}}, \underline{\underline{\tau}})$, and $b(\underline{\underline{\tau}}, v)$ is given by (5.5). The right-hand side is obviously $-(\underline{\underline{\sigma}}^f, \underline{\underline{\tau}})$ for the first equation and $-b(\underline{\underline{\sigma}}^f, v)$ for the second equation. It is natural to use in V the L^2-norm, and in Q the norm $\|v\|_Q = \|\underline{\underline{D}}_2 v\|_V = \|\underline{\underline{D}}_2 v\|_0$. It is clear that condition (II.1.8), that is, the ellipticity of $a(\cdot,\cdot)$, is trivially satisfied in the whole V (and not only in Ker B) with $\alpha = 1$. A different value for E, t, and ν would obviously yield a different value for α but the V ellipticity will still be true. It is clear that Ker B^t cannot be empty; indeed, any v with support in a single K will satisfy $b(\underline{\underline{\tau}}, v) = 0$ for all $\underline{\underline{\tau}}$, and hence is a zero energy mode. However, it is not difficult to see that Im B is closed.

Proposition 5.1: The image of B is a closed subset of $Q' = H^{-2}(\Omega)$.

Proof: We have to show that if a sequence $\chi_n = B\underline{\underline{\tau}}_n$ converges to χ in H^{-2}, then $\chi = B\underline{\underline{\tau}}$ for some $\underline{\underline{\tau}} \in V_0(\mathcal{T}_h) = V$. We note first that

$$\text{if } \underline{\underline{\tau}} \in V_0(\mathcal{T}_h) \text{ and } \phi \in H_0^2(\Omega), \text{ then } b(\underline{\underline{\tau}}, \phi) \equiv (\underline{\underline{\tau}}, \underline{\underline{D}}_2\phi), \tag{5.9}$$

which is quite obvious from (5.5) and (5.6). Now let $\psi \in H_0^2(\Omega)$ be such that $\Delta^2\phi = \chi$ and let $\underline{\underline{\tau}} = \underline{\underline{D}}_2\phi$ (so that $D_2^*\underline{\underline{\tau}} = \chi$). For every $\phi \in H_0^2$ we have

$$\langle \chi, \phi\rangle_{H^{-2}\times H_0^2} = \langle D_2^*\underline{\underline{\tau}}, \phi)\rangle_{H^{-2}\times H_0^2} = (\underline{\underline{\tau}}, \underline{\underline{D}}_2\phi). \tag{5.10}$$

Now, since $\chi_n = B\underline{\underline{\tau}}_n \to \chi$ in H^{-2}, we have

$$(\underline{\underline{\tau}}_n, \underline{\underline{D}}_2\phi) = b(\underline{\underline{\tau}}_n, \phi) = \langle B\underline{\underline{\tau}}_n, \phi\rangle = \langle \chi_n, \phi\rangle \to \langle \chi, \phi\rangle = (\underline{\underline{\tau}}, \underline{\underline{D}}_2\phi), \tag{5.11}$$

that is, $(\underline{\underline{\tau}}_n - \underline{\underline{\tau}}, \underline{\underline{D}}_2\phi) \to 0$ for all $\phi \in H_0^2(\Omega)$. This easily implies $D_2^*\underline{\underline{\tau}} = 0$ in each T, so that $\underline{\underline{\tau}} \in V_0(\mathcal{T}_h)$. Hence $\langle \chi, \phi\rangle = (\underline{\underline{\tau}}, \underline{\underline{D}}_2\phi) = b(\underline{\underline{\tau}}, \phi) = \langle B\underline{\underline{\tau}}, \phi\rangle$, that is, $\chi \in$ Im B. □

Proposition 5.2: We have $\operatorname{Ker} B^t = \prod_K H_0^2(K)$.

Proof: It is obvious from (5.5) that if $\phi|_K \in H_0^2(K)$ for all K, then $b(\underline{\tau}, \phi) = 0\ \forall \underline{\tau}$, and hence $\phi \in \operatorname{Ker} B^t$. Therefore we need only to prove that $\operatorname{Ker} B^t \subset \prod_K H_0^2(K)$. For this, let $\phi \in \operatorname{Ker} B^t$, that is,

$$b(\underline{\tau}, \phi) \equiv (\underline{\tau}, \underline{\underline{D}}_2 \phi) = 0, \qquad \forall \underline{\underline{\tau}} \in V_0(\mathcal{T}_h). \tag{5.12}$$

We want to show that $\phi \in \prod_T (H^2 K))$, that is,

$$\phi|_K \in H_0^2(K), \quad \forall K. \tag{5.13}$$

Let ψ be defined in each K by

$$\psi \in H_0^2(K) \text{ and } \Delta^2 \psi = \Delta^2 \phi; \tag{5.14}$$

clearly, $(\underline{\tau}, \underline{\underline{D}}_2 \psi) = 0$ for all $\underline{\tau}$ in $V_0(\mathcal{T}_h)$ so that from (5.12)

$$b(\underline{\tau}, \psi - \phi) = (\underline{\tau}, \underline{\underline{D}}_2(\psi - \phi)) = 0, \quad \forall \underline{\underline{\tau}} \in V_0(\mathcal{T}_h). \tag{5.15}$$

But now $D_2^* \underline{\underline{D}}_2(\psi - \phi) = \Delta^2(\psi - \phi) = 0$ in each K, so that we can take $\underline{\underline{\tau}} = \underline{\underline{D}}_2(\psi - \phi)$ in (5.15) and obtain $\underline{\underline{D}}_2(\psi - \phi) \equiv 0$. Since both ψ and ϕ are in $H_0^2(\Omega)$, this implies $\psi = \phi$, so that from (5.14) we get (5.13). □

Proposition 5.3: We have

$$\|\phi\|_{Q/\operatorname{Ker} B^t} = \|\underline{\underline{D}}_2 \bar{\phi}\|_0, \tag{5.16}$$

where $\bar{\phi}$ is the function in $H_0^2(\Omega)$ such that

$$\phi - \bar{\phi} \in H_0^2(K) \quad \text{for each } K, \tag{5.17}$$

$$\Delta^2 \bar{\phi} = 0 \quad \text{in each } K. \tag{5.18}$$

Proof: By definition we have

$$\|\phi\|_{Q/\operatorname{Ker} B^t} = \inf_{\psi \in \operatorname{Ker} B^t} \|\phi - \psi\|_Q. \tag{5.19}$$

Now from Proposition 5.2 and the definition of $\|\chi\|_Q = \|\underline{\underline{D}}_2 \chi\|_0$ we have

$$\|\phi\|_{Q/\operatorname{Ker} B^t} = \inf_{\psi \in \prod_K H_0^2(K)} \|\underline{\underline{D}}^2(\phi - \psi)\|_{0,K}. \tag{5.20}$$

It is now an easy matter to check that, for each K,

$$\inf_{\psi \in H_0^2(K)} \|\underline{\underline{D}}_2(\phi - \psi)\|_{0\ K}^2 = \inf_{(\psi - \phi) \in H_0^2(K)} \|\underline{\underline{D}}_2 \psi\|_{0,K}^2 = \|\underline{\underline{D}}_2 \bar{\phi}\|_{0,K}^2 \tag{5.21}$$

for $\bar{\phi}$ defined in (5.17) an (5.18). Hence (5.21) and (5.20) prove (5.16). □

We are now able to prove the inf–sup condition

$$\sup_{\underline{\underline{\tau}}\in V_0(\mathcal{T}_h)} \frac{b(\underline{\underline{\tau}},\phi)}{\|\underline{\underline{\tau}}\|_0\|\phi\|_{Q/\operatorname{Ker} B^t}} = \sup_{\underline{\underline{\tau}}\in V_0(\mathcal{T}_h)} \frac{(\underline{\underline{\tau}},\underline{\underline{D}}_2\phi)}{\|\underline{\underline{\tau}}\|_0\|\underline{\underline{D}}_2\bar{\phi}\|_0} \tag{5.22}$$
$$\geq \frac{(\underline{\underline{D}}_2\bar{\phi},\underline{\underline{D}}_2\phi)}{\|\underline{\underline{D}}_2\bar{\phi}\|_0^2} = 1$$

because $\phi - \bar{\phi}$ is the projection (in Q) of ϕ onto $\operatorname{Ker} B^t$ so that $\bar{\phi}$ and $\phi - \bar{\phi}$ are orthogonal in Q.

Remark 5.1: A way of getting rid of $\operatorname{Ker} B^t$ (which is infinite dimensional) is to consider as a space of Lagrange multipliers the space

$$\tilde{Q} = \{\phi \mid \phi \in H_0^2(\Omega),\ \Delta^2\phi = 0 \text{ in each } T\}. \tag{5.23}$$

This is what has been done in BREZZI [C], BREZZI–MARINI [A]. The drawback in the choice (5.23) is that the actual transversal displacement w does not belong to $\tilde{Q}$ so that, as a solution, we have the unique function $\bar{w}$ in $\tilde{Q}$ that coincides with w (with its first derivatives) at the interelement boundaries (as in (5.17) and (5.18)). □

Let us continue our analysis of problem (5.8). We already noted that (II.1.8) is satisfied in our case. Hence, we have to check that the right-hand side of the second equation in (5.8) (that is, $-b(\underline{\underline{\sigma}}^f, v)$) is in $\operatorname{Im} B$; this means that we have to find a particular solution of the second equation of (5.8), which is obvious by taking $\underline{\underline{\sigma}}^0 := \underline{\underline{D}}_2 w - \underline{\underline{\sigma}}^f$.

We can now go to the discretization of (5.8); for this we have to choose subspaces $V_h \subset V_0(\mathcal{T}_h)$ and $Q_h \subset Q$. For instance, for any triple (m, r, s) of integers we may choose

$$V_h^m := (\mathfrak{L}_m^0(\mathcal{T}_h))_s^4 \cap V_0(\mathcal{T}_h), \tag{5.24}$$

$$Q_h^{r,s} := \{\phi \in H_0^2(\Omega),\ \phi|_{\partial T} \in T_r(\partial T),\ \left.\frac{\partial\phi}{\partial n}\right|_{\partial T} \in R_s(\partial T),\ \forall T \in \mathcal{T}_h\}. \tag{5.25}$$

Note that V_h is made of tensor-valued polynomials of degree $\leq m$ which are completely discontinuous from one element to another and verify $D_2^*\underline{\underline{\tau}} = 0$ in each T. On the other hand, Q_h is clearly infinite dimensional (which is quite unusual); however this does not show up in the computations, where only the values of ϕ and $\partial\phi/\partial n$ on $\mathcal{E}_h$ are considered. According to Proposition II.2.1 we now have to choose (m, r, s) in such a way that $\operatorname{Ker} B_h^t \in \operatorname{Ker} B^t$. This means, in our case, that we have to show

$$\begin{cases} \text{if } \phi \in Q_h^{r,s} \text{ and } b(\underline{\underline{\tau}},\phi) = 0,\ \forall \underline{\underline{\tau}} \in V_h^m \text{ (that is, if } \phi \in \operatorname{Ker} B_h^t), \\ \text{then } \phi = \operatorname{grad}\phi = 0 \text{ on } \mathcal{E}_h \text{(that is, } \phi \in \operatorname{Ker} B^t). \end{cases} \tag{5.26}$$

The proof of (5.26) (or, rather, the finding of sufficient conditions on m for having (5.26)) will be easier with the following characterization of V_h^m.

Lemma 5.1: We have

$$V_h^m \equiv \underline{\underline{S}}[(\mathcal{L}^0_{m+1}(\mathcal{T}_h))^2] \tag{5.27}$$

where $\underline{\underline{S}}$ is defined for $\underline{q} = (\alpha, \beta)$

$$\underline{\underline{S}} : \underline{q} \to \begin{pmatrix} \partial\alpha/\partial y & -\frac{1}{2}(\partial\alpha/\partial x + \partial\beta/\partial y) \\ -\frac{1}{2}(\partial\alpha/\partial x + \partial\beta/\partial y) & \partial\beta/\partial x \end{pmatrix}.$$

Proof: The inclusion $\underline{\underline{S}}[(\mathcal{L}^0_{m+1}(\mathcal{T}_h))^2] \subseteq V_h^m$ is trivial; the opposite inclusion is an exercise. (See BREZZI–MARINI [A] for more details.) □

We notice now that if $\underline{\underline{\tau}} = \underline{\underline{S}}(\underline{q})$,then

$$b(\underline{\underline{\tau}}, v) = \sum_K \int_{\partial K} \underline{\text{grad}}\, v \cdot \frac{\partial}{\partial t}\underline{q}\, ds, \tag{5.28}$$

where $\underline{t}$ is the tangent to ∂T. We also notice that

$$\begin{cases} \phi \in H_0^2(\Omega) \text{ and } \underline{\text{grad}}\,\phi = \text{ constant on } \mathcal{E}_h \\ \text{imply } \phi = 0 \text{ and } \underline{\text{grad}}\,\phi = 0 \text{ on } \mathcal{E}_h. \end{cases} \tag{5.29}$$

We may now use (5.27)–(5.29) in (5.26) which becomes

$$\begin{cases} \text{if } \phi \in Q_h^{r,s} \text{ and } \displaystyle\sum_K \int_{\partial K} \underline{\text{grad}}\,\phi \cdot \frac{\partial}{\partial t}\underline{q}\, ds = 0,\ \forall \underline{q} \in (\mathcal{L}^0_{m+1}(\mathcal{T}_h))^2, \\ \text{then } \underline{\text{grad}}\,\phi = \text{ constant on } \mathcal{E}_h. \end{cases} \tag{5.30}$$

Now (5.30) is implied by

$$\begin{cases} \text{if } \phi \in Q_h^{r,s} \text{ and } \displaystyle\int_{\partial K} \underline{\text{grad}}\,\phi \cdot \frac{\partial}{\partial t}\underline{q}\, ds = 0,\ \forall \underline{q} \in (P_{m+1}(K))^2, \\ \text{then } \underline{\text{grad}}\,\phi = \text{ constant on } \partial K. \end{cases} \tag{5.31}$$

(but not vice versa). Now let k be the degree of $\underline{\text{grad}}\,\phi$ on ∂T, that is,

$$k = \max(s, r-1). \tag{5.32}$$

The following technical lemma is proved for instance in BREZZI–MARINI [A].

Lemma 5.2: If $\phi \in H^1(K)$ and $\phi|_{e_i} \in P_k(e_i)\,(i = 1, 2, 3)$ and if

$$\int_{\partial T} \phi \frac{\partial q}{\partial t} ds = 0, \quad \forall q \in P_k(K), \tag{5.33}$$

then

$$\phi|_{e_i} = c\ell_k^i(s) + c_1 \quad (i = 1, 2, 3), \tag{5.34}$$

where, on each e_i, we define ℓ_k^i as the kth Legendre polynomial (normalized with value 1 in the second endpoint in the counterclockwise order). □

Formula (5.34), for k odd, implies directly that $\phi =$ constant on ∂K. We have therefore a first result.

Proposition 5.4: If $m + 1 = k = \max(r - 1, s)$ and k is odd, then (5.31) holds. □

If now $m+1$ is even, we can apply Lemma 5.2 to both $\partial\phi/\partial x$ and $\partial\phi/\partial y$ and get

$$\frac{\partial\phi}{\partial x} = c\ell_k^i + c_1, \qquad \frac{\partial\phi}{\partial y} = \gamma\ell_k^i + \gamma_1, \tag{5.35}$$

on each e_i. If now $r - 1 \neq s$, there must exist a combination of $\partial\phi/\partial x$ and $\partial\phi/\partial y$ on each e_i (to get $\partial\phi/\partial n$) which has degree lower than k). This easily implies that both $\partial\phi/\partial x$ and $\partial\phi/\partial y$ are constants on ∂K. We have, therefore, the following result:

Proposition 5.5: If $m + 1 = k = \max(r - 1, s)$ and $r - 1 \neq s$, then (5.31) holds. □

We are finally left with the last and worst case in which $r - 1 = s$ is *even*. We have several escapes. First, brutally, we may take $m + 1 = k + 1$. It is easy to see that then (5.31) always holds. As a second possibility, we may take $m + 1 = k$ and *enrich* $(\mathfrak{L}^0_{m+1}(\mathcal{T}_h))^2$ into $(\mathfrak{L}^0_{m+1}(\mathcal{T}_h))^2_{enr}$ by adding, in each K, a pair of functions $\underline{q}$ in $(P_{m+1})^2$ such that $\partial q_j/\partial t|_{e_i} = \ell_k^i$ ($j = 1, 2$ and $i = 1, 2, 3$). Again it is easy to check that (5.31) is satisfied if we take the enriched space $(\mathfrak{L}^0_{m+1}(\mathcal{T}_h))^2_{enr}$ instead of the original one. Then, of course, we must consider $V^m_{h,enr} = \underline{\underline{S}}[(\mathfrak{L}^0_{m+1}(\mathcal{T}_h))^2_{enr}]$ instead of V_h. Finally, we might give up (5.31) and go directly to (5.30). It is easy to check that in (5.35) the values of c, c_1, γ, and γ_1 must remain constants from one K to another, due to the continuity of $\underline{\text{grad}}\,\phi|_e$ across the edges. Hence, since $\phi \in H^2_0(\Omega)$, we must have $c = c_1 = \gamma = \gamma_1 = 0$ and *actually* (5.30) *holds for* $m + 1 = k = \max(r - 1, s)$ *in any case*, that is, also for $r - 1 = s =$ even. However, we shall see in a moment that (5.31) has other basic advantages over (5.30) that we are not very willing to give up. We summarize the results in the following theorem.

Theorem 5.1: The condition $\operatorname{Ker} B_h^t \subset \operatorname{Ker} B^t$ holds whenever

$$m + 1 \geq k = \max(r - 1, s). \tag{5.36}$$

Moreover, (5.31) holds when (5.36) is satisfied, unless $r-1=s=$ even. In that case (5.31) is satisfied by taking $m+1>k$ or by using an enriched $V_{h,enr}^{k-1}$ (between V_h^{k-1} and V_h^k) as described above. □

The condition $\operatorname{Ker} B_h^t = \operatorname{Ker} B^t$ implies, by Proposition II.2.1, the existence of an operator Π_h from $V_0(\mathcal{T}_h)$ to V_h^m such that

$$b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, v) = 0, \qquad \forall v \in Q_h^{r,s}. \tag{5.37}$$

However, in view of the use of Proposition II.2.8 we would also like to show that there exists a Π_h which satisfies (5.37) and

$$\|\Pi_h \underline{\underline{\tau}}\|_0 \le c\,\|\underline{\underline{\tau}}\|_0, \qquad \forall \underline{\underline{\tau}} \in V_0(\mathcal{T}_h), \tag{5.38}$$

with c *independent of* h. (Since V_h^m is finite dimensional, (5.38) will always hold, but the constant might depend on h.) Now, if (5.31) holds, we see that Π_h can be defined *element by element*. But, the dimension of $V_h^m|_K$ depends on m, but not on h. A continuous dependence argument on the shape of the element can now prove (5.38) without major difficulties (but, to be honest, not quickly); we refer to BREZZI–MARINI [A] for a detailed proof of (5.38). Once we have (5.37) and (5.38) we apply Proposition II.2.8 to prove the discrete inf–sup condition. Then Theorem II.2.1 gives immediately

$$\begin{aligned} &\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 + \|\underline{\underline{D}}_2(w - \tilde{w}_h\|_0 \\ &\qquad \le c\Big\{\inf_{\underline{\underline{\tau}} \in V_h^m} \|\underline{\underline{\sigma}}^0 - \underline{\underline{\tau}}\|_0 + \inf_{\phi \in Q_h^{r,s}} \|\underline{\underline{D}}_2(w-\phi)\|_0\Big\}, \end{aligned} \tag{5.39}$$

where $\tilde{w}_h$ is the (unique) element in $Q_h^{r,s}$ that satisfies $\Delta^2 \tilde{w}_h = f$ in each K and belongs to the set of discrete solutions.

Theorem 5.2: If $m+1 \ge \max(r-1, s)$ (and $m+1 > s$ for $r-1=s$ is even), we have

$$\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 + \|\underline{\underline{D}}_2(w - \tilde{w}_h)\|_0 \le c h^t \big(\|w\|_{t+2} + \sum_T \|\underline{\underline{\sigma}}^f\|_{t,K}^2\big)^{\frac{1}{2}} \tag{5.40}$$

with $t = \min(m+1, r-1, s)$.

The proof is obvious from inequality (5.39) and standard approximation results. □

We end this section with a few computational remarks. First we notice that our discretization of (5.8) has obviously the matrix structure

$$\begin{pmatrix} A & B \\ B^t & 0 \end{pmatrix}, \tag{5.41}$$

where A, corresponding to the approximation of the identity in V_h^m, is obviously block diagonal because V_h^m is made of discontinuous tensors. Hence one usually makes an a priori inversion of A, to end with the matrix $B^t A^{-1} B$, which operates on the w_h unknown and is symmetric and positive definite. However, the computation of the right-hand side is, in general, a weak point in the use of dual hybrid methods, unless f is very special (zero, Dirac mass, constant, etc.) and allows the use of a simple $\underline{\underline{\sigma}}^f$. A few computational tricks for dealing with more general cases can be found in BREZZI–MARINI [A], MARINI [A,B]. Here we recall, from BREZZI [D] a simple method that works for low-order approximations (more precisely, when t in Theorem 5.2 is ≤ 2). We define first the operator $R =$ orthogonal projection onto V_h. We remark then that the discretizations (5.24) and (5.25) of (5.8) may be written as

$$\text{(5.42)} \qquad \begin{cases} (\underline{\underline{\sigma}}_h^0 + \underline{\underline{\sigma}}^f, \underline{\underline{\tau}}) = (\underline{\underline{D}}_2 w_h, \underline{\underline{\tau}}), & \forall \underline{\underline{\tau}} \in V_h, \\ (\underline{\underline{\sigma}}_h + \underline{\underline{\sigma}}^f, \underline{\underline{D}}_2 \phi) = (f, \phi), & \forall \phi \in Q_h. \end{cases}$$

Solving a priori in $\underline{\underline{\sigma}}_h^0$ from the first equation and substituting into the second equation we obtain

$$\text{(5.43)} \qquad (R\underline{\underline{D}}_2 w_h, \underline{\underline{D}}_2 \phi) = (f, \phi) - (\underline{\underline{\sigma}}^f - R\underline{\underline{\sigma}}^f, \underline{\underline{D}}_2 \phi), \quad \forall \phi \in Q_h$$

Now the left-hand side of (5.43) corresponds to the matrix $B^t A^{-1} B$ acting on the unknown w_h. The right-hand side is actually *computable* because both $(f, \phi) - (\underline{\underline{\sigma}}^f, \underline{\underline{D}}_2 \phi)$ and $(R\underline{\underline{\sigma}}^f, \underline{\underline{D}}_2 \phi)$ depend (looking carefully) only on the values of ϕ and its gradient at the interelement boundaries. However, the computation, in general, is not easy. Therefore, in some cases, it can be convenient to use a rough approximation of it, for instance

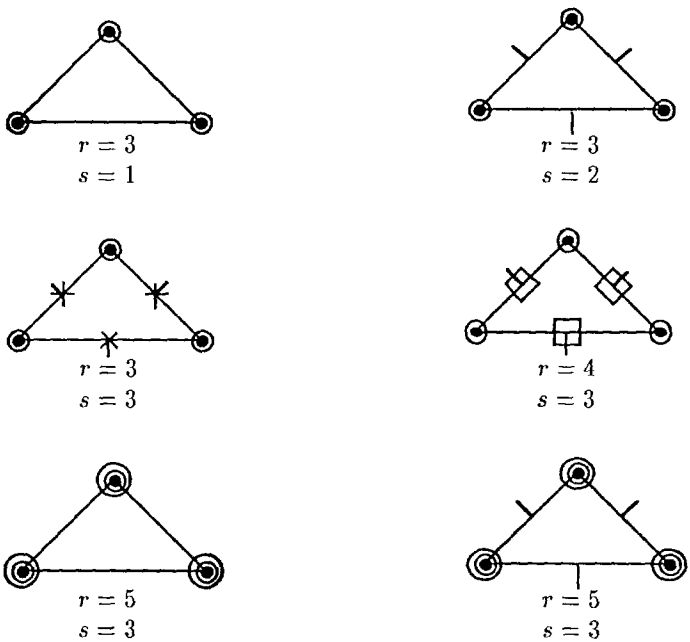
$$(f, \phi) - (\underline{\underline{\sigma}}^f - R\underline{\underline{\sigma}}^f, \underline{\underline{D}}_2 \phi) \simeq \sum_K \frac{meas(K)}{3} \sum_{j=1}^{3} f(V_j)\phi(V_j),$$

where the V_j are the vertices of K. It can be shown (BREZZI [D]) that this involves an additional error of order $O(h^2)$ (essentially because V_h contains *all* piecewise linear stress functions and therefore $\|\underline{\underline{\sigma}}^f - R\underline{\underline{\sigma}}^f\|_0 \leq ch^2$) and hence, this procedure is recommended whenever $t \leq 2$ in (5.40).

Finally, a few remarks on the choice of the degrees of freedom in V_h^m and $Q_h^{r,s}$. As we have seen, the $\underline{\underline{\sigma}}_h^0$ unknown is usually eliminated a priori at the element level due to the complete discontinuity of V_h^m. As a consequence, the choice of the degrees of freedom in V_h^m is of little relevance. In general, it is more convenient to start from $(\mathfrak{L}_{m+1}^0(T_h))^2$ and to derive V_h through (5.27).

When m is "large" (say $m \geq 4$, to fix the ideas), however, the resulting matrix A can be severely ill-conditioned unless the degrees of freedom in V_h^m are chosen in a suitable way. We refer to MARINI [A,B] for a discussion of

this point. On the other hand the degrees of freedom in $Q_h^{r,s}$ are the ones that count in the final stiffness matrix, and, besides, they have to take into account the C^1 -continuity requirements. We list in Figure IV.5 some commonly used choices for different values of r and s.



Symbol	Values of
●	ϕ
◉	$\underline{\text{grad}}\,\phi$
◎	$\partial^2\phi/\partial_n\partial_t$
—	$\underline{\underline{D}}_2\phi$
×	$\partial\phi/\partial n$
⊡	$\partial\phi/\partial t,\ \partial\phi/\partial n,\ \partial^2\phi/\partial_n\partial_t$

Figure IV.5

Remark 5.2: It is impossible to say what is, in general, the best choice for r and s. Numerical evidence shows obviously that the accuracy/number of degrees of freedom ratio is improved for large r and s, at least when the solution is smooth. However it is clear that the simplest (and most widely used) choice $r = 3$, $s = 1$ allows a much easier implementation. Similar considerations also hold with the choice of m, in particular in the case where $r - 1 = s$ is *even*, for instance for $r = 3, s = 2$. The use of the enriched $V^1_{h,enr}$ implies a smaller matrix to be inverted on each element than with the "brutal" choice V^2_h (11×11 instead of 17×17), but the latter may allow some simplification in writing the program. □

Remark 5.3: We have used, so far, homogeneous Dirichlet boundary conditions corresponding to a clamped plate. Nothing changes when considering nonhomogeneous Dirichlet conditions. If, instead, a part of the plate is simply supported ($w =$ given; $M_{nn} = 0$) or free ($M_{nn} = 0$; $\mathcal{K}_n = 0$), then we have two possibilities for dealing with them. Let us discuss a simple case: let $\partial\Omega = \Gamma_D \cup \Gamma_N$ and assume that $w = \partial w/\partial n = 0$ on Γ_D and $M_n = \mathcal{K}_n = 0$ on Γ_N. One possibility is to choose $Q^{r,s}_h$ so that its elements vanish only on Γ_D, and to let V^m_h unchanged. In this case the conditions $M_n = \mathcal{K}_n = 0$ on Γ_N will be satisfied only in a *weak* sense. A second possibility is to choose V^m_h in such a way that its elements satisfy, a priori, the boundary conditions $M_n = \mathcal{K}_n = 0$ on Γ_N. However, care must be taken in this case to enrich conveniently the stress field in the boundary elements, so that the inf–sup condition still holds. Otherwise, a loss in the order of convergence is likely to occur. □

Remark 5.4: One may think of using discretizations of the dual hybrid formulations other than the ones discussed here (see, for instance, the previous remarks). In any case, the inf–sup condition should be checked. Although this is not evident from our discussion (because we wanted to deal with many cases at the same time), nevertheless it is true that to check the inf–sup condition in hybrid methods is basically an *easy* task. What is really needed is the following: for any element K, the only displacement modes with zero energy on K, that is, the only modes ϕ such that

$$\int_{\partial K} \Big(M_{nn}(\underline{\underline{\tau}})\phi/n - \mathcal{K}_n(\underline{\underline{\tau}})\phi \Big)\, ds = 0, \quad \forall \underline{\underline{\tau}} \in V_h,$$

must be the *rigid* modes (that is, $\underline{\text{grad}}\phi =$ constant on K). If this condition is violated, one can expect troubles (minor or major, depending on the cases). □

V

Complements on Mixed Methods for Elliptic Problems

V.1 Numerical Solutions

V.1.1 Preliminaries

In this chapter we present some additional results on the application of the mixed finite element method to linear elliptic problems. In particular in Section V.1 we shall discuss some aspects of the numerical techniques that can be used for solving the linear system of equations that one obtains after discretization. The procedure suggested here is essentially due (to our knowledge) to Fraeijs de Veubeke and, as we shall see, involves the introduction of suitable interelement Lagrange multipliers λ. Such a trick has the remarkable effect of reducing the total number of unknowns and leads to solving a linear system for a matrix which is symmetric and positive definite instead of the original indefinite one. A rough analysis of the computational effort that this procedure requires for the various elements is presented in Section V.2. Moreover, as we shall see in Section V.3 , the new unknown λ's that are obtained by such a procedure allow the construction of a new approximation u_h^* of u, depending on λ and u_h, which is usually much closer to u. In a fourth section we sketch miscellaneous results on error estimates in different norms. Section V.5 is dedicated to an example of application to semiconductor devices simulation. Finally, Section V.6 presents, on a very simple problem, some examples of dicretization that do not work, and Section V.7 applications of augmented formulations introduced in Section I.5.

For the sake of simplicity we shall present the arguments on the model case (cf. Chapter IV).

$$(1.1)\qquad \begin{cases} \Delta u = f \text{ in } \Omega, \\ \quad u = g \text{ on } \Gamma_D, \\ \dfrac{\partial u}{\partial n} = 0 \text{ on } \Gamma_N, \end{cases}$$

although the range of generality is much wider. Similarly, we shall discuss in detail the simplest case of the approximation by means of the $RT_0 \equiv BDFM_1$ element and give statements and references for the proofs of the other cases.

V.1.2 Interelement multipliers

As we have seen in Section IV.1, the mixed formulation of (1.1) is

$$(1.2)\qquad \begin{cases} (\underline{p}, \underline{q}) + (u, \operatorname{div} \underline{q}) = \langle g, \underline{q} \cdot \underline{n} \rangle, & \forall \underline{q} \in H_{0,N}(\operatorname{div}; \Omega), \\ (v, \operatorname{div} \underline{p}) = (f, v), & \forall v \in L^2(\Omega), \end{cases}$$

where u, f, and g are the same as in (1.1) and $\underline{p} = \underline{\operatorname{grad}}\, u$. Assume, for the sake of simplicity, that Ω is a polygon and let $\mathcal{T}_h$ be a triangulation of Ω. We recall, following Section III.3.4, that the use of the RT_0 element for the approximation of (1.2) consists of the following steps. We had

$$(1.3)\qquad RT_0(K) = \{(a + bx_1, c + bx_2),\ a, b, c \in \mathbb{R}\} \subseteq (P_1(K))^2,$$

$$(1.4)\qquad \mathfrak{M}^0 = \{\underline{q} \mid \underline{q} \in (L^2(\Omega))^2,\ \underline{q}|_K \in RT_0(K),\ \forall K \in \mathcal{T}_h\},$$

$$(1.5)\qquad \mathfrak{M} = \mathfrak{M}^0 \cap H_{0,N}(\operatorname{div}; \Omega) = \{\underline{q} \in RT_0(\Omega, \mathcal{T}_h),\ \underline{q} \cdot \underline{n} = 0 \text{ on } \Gamma_N\}.$$

Thus, the elements of $\mathfrak{M}$ are the elements of $\mathfrak{M}^0$ such that $\underline{q} \cdot \underline{n}$ is continuous across the interelement boundaries and vanishes on Γ_N. The discretized version of (1.2) is now

$$(1.6)\qquad \begin{cases} (\underline{p}_h, \underline{q}_h) + (u_h, \operatorname{div} \underline{q}_h) = \langle g, \underline{q}_h \cdot \underline{n} \rangle, & \forall \underline{q}_h \in \mathfrak{M}, \\ (v_h, \operatorname{div} \underline{p}_h) = (f, v_h), & \forall v_h \in \mathfrak{L}_0^0, \end{cases}$$

where, clearly, $\underline{p}_h$ is sought in $\mathfrak{M}$ and u_h in $\mathfrak{L}_0^0$. We remind the reader that $\mathfrak{L}_0^0$ is the space of piecewise constant functions. The linear system of equations associated with (1.6) has the form (see (II.3.9))

$$(1.7)\qquad \begin{pmatrix} A & B^t \\ B & 0 \end{pmatrix} \begin{pmatrix} P \\ U \end{pmatrix} = \begin{pmatrix} G \\ F \end{pmatrix}$$

and its matrix is indefinite. This is definitely a considerable source of trouble. Therefore, following essentially the ideas of FRAEIJS de VEUBEKE [A] we introduce the space

$$\Lambda = \mathfrak{L}_0^0(\mathcal{E}_h) \tag{1.8}$$

of functions μ_h which are constant on each edge of the decomposition $\mathcal{T}_h$. For any function $\chi \in L^2(\Gamma_D)$ we consider

$$\Lambda_{\chi,D} = \{\mu_h \mid \mu_h \in \Lambda, \int_e (\mu_h - \chi)\, ds = 0, \ \ \forall e \in \mathcal{E}_h \cap \Gamma_D\}. \tag{1.9}$$

It will finally be convenient to set, for $\underline{q}_h \in \mathfrak{M}^0$ and $\mu_h \in \Lambda$,

$$c(\mu_h, \underline{q}_h) = \sum_K \int_{\partial K} \mu_h \ \underline{q}_h \cdot \underline{n}\, d\sigma. \tag{1.10}$$

The following lemma is a direct consequence of definition (1.5).

Lemma 1.1: Assume that $\underline{q}_h \in \mathfrak{M}^0$. Then

$$(c(\mu_h, \underline{q}_h) = 0, \ \ \forall \mu_h \in \Lambda_{0,D}) \Leftrightarrow (\underline{q}_h \in \mathfrak{M}). \ \square \tag{1.11}$$

Let now $(\underline{p}_h, u_h)$ be the solution of (1.6), and consider the linear mapping

$$\phi : \ \underline{q}_h \to (\underline{p}_h, \underline{q}_h) + (u_h, \operatorname{div} \underline{q}_h)_h - \langle g, \underline{q}_h \cdot \underline{n}\rangle, \tag{1.12}$$

where $(\chi, \psi)_h = \sum_K \int_K \chi\psi\, dx$. It is clear that $\phi(\underline{q}_h) = 0$ for all $\underline{q}_h \in \mathfrak{M}$. Therefore, (1.11) implies that there exists a $\lambda_{0h} \in \Lambda_{0,D}$ such that by Proposition II.1.2

$$\phi(\underline{q}_h) = c(\lambda_{0h}, \underline{q}_h), \qquad \forall \underline{q}_h \in \mathfrak{M}^0. \tag{1.13}$$

Let us show that such a λ_{0h} is unique. This will be an immediate consequence of the following lemma.

Lemma 1.2: If $\mu_h \in \Lambda_{0,D}$ and

$$c(\mu_h, \underline{q}_h) = 0, \qquad \forall \underline{q}_h \in \mathfrak{M}^0, \tag{1.14}$$

then $\mu_h \equiv 0$.

Proof: Let e^* be an edge in $\mathcal{E}_h$ and $K^* \in \mathcal{T}_h$ be a triangle such that $e^* \subset \partial K^*$. Let $\underline{q}_h^* \in \mathfrak{M}^0$ be such that

$$\underline{q}_h^*|_K = 0, \qquad \forall K \neq K^* \tag{1.15}$$

and defined on K^* by

$$\begin{cases} \underline{q}_h^* \cdot \underline{n} = 0 \text{ on the edges } e \neq e^*, \\ \underline{q}_h^* \cdot \underline{n} = 1 \text{ on } e^*. \end{cases} \tag{1.16}$$

Then $c(\mu_h, \underline{q}_h^*) = \int_{e^*} \mu_h \, ds$ and (1.14) implies that $\mu_h = 0$ on e^*. Since e^* was any edge in $\mathcal{E}_h$, this concludes the proof. □

Let us now define λ_h by means of

$$\lambda_h \in \Lambda_{g,D}, \quad \lambda_h \equiv \lambda_{0h} \text{ on } \mathcal{E}_h \backslash \Gamma_D. \tag{1.17}$$

Then (1.12) and (1.13) imply that

$$(\underline{p}_h, \underline{q}_h) + (u_h, \operatorname{div} \underline{q}_h)_h = c(\lambda_h, \underline{q}_h), \qquad \forall \underline{q}_h \in \mathfrak{M}^0. \ \square \tag{1.18}$$

We can summarize the results obtained so far in the next theorem.

Theorem 1.1: Let $(\underline{p}_h, u_h)$ be the solution of (1.6) and let λ_h be defined through (1.13) and (1.17). Then the triple $(\underline{p}_h, u_h, \lambda_h)$ is the unique solution of the following problem: find $(\underline{p}_h, u_h, \lambda_h)$ in $\mathfrak{M}^0 \times \mathfrak{L}_0^0 \times \Lambda_{g,D}$ such that

$$\begin{cases} (\underline{p}_h, \underline{q}_h) + (u_h, \operatorname{div} \underline{q}_h)_h = c(\lambda_h, \underline{q}_h), & \forall \underline{q}_h \in \mathfrak{M}^0, \\ (v_h, \operatorname{div} \underline{p}_h)_h = (f, v_h), & \forall v_h \in \mathfrak{L}_0^0, \\ c(\mu_h, \underline{p}_h) = 0, & \forall \mu_h \in \Lambda_{0,D}. \ \square \end{cases} \tag{1.19}$$

The matrix associated with (1.19) now has the form

$$\begin{pmatrix} \bar{A} & \bar{B}^t & \bar{C}^t \\ \bar{B} & O & O \\ \bar{C} & O & O \end{pmatrix} \begin{pmatrix} \bar{P} \\ \bar{U} \\ \bar{\Lambda} \end{pmatrix} = \begin{pmatrix} \bar{G} \\ \bar{F} \\ O \end{pmatrix}, \tag{1.20}$$

and we still do not see any improvement on (1.7). However, consider that the space $\mathfrak{M}^0$ is completely discontinuous from one element to another. This was not the case with $\mathfrak{M}$, which required the continuity of $\underline{q}_h \cdot \underline{n}$. As a consequence, we can choose in $\mathfrak{M}^0$ a basis, made of vectors $\underline{q}_h$ that are different from zero only on one triangle (as was the vector $\underline{q}_h^*$ in (1.15) and (1.16)). Then matrix $\bar{A}$ becomes *block diagonal*, each block being a 3×3 matrix corresponding to

a single element and we can eliminate the unknown $\bar{P}$ *at the element level* by solving

$$\bar{P} = \bar{A}^{-1}(\bar{G} - \bar{B}^t\bar{U} - \bar{C}^t\bar{\Lambda}). \tag{1.21}$$

We are left with the system

$$\begin{pmatrix} -\bar{B}\bar{A}^{-1}\bar{B}^t & -\bar{B}\bar{A}^{-1}\bar{C}^t \\ -\bar{C}\bar{A}^{-1}\bar{B}^t & -\bar{C}\bar{A}^{-1}\bar{C}^t \end{pmatrix} \begin{pmatrix} \bar{U} \\ \bar{\Lambda} \end{pmatrix} = \begin{pmatrix} -\bar{B}\bar{A}^{-1}\bar{G} + \bar{F} \\ 0 \end{pmatrix}. \tag{1.22}$$

Recall now that $\mathfrak{L}_0^0$ is made of piecewise constants. This means that the matrix $\bar{B}\bar{A}^{-1}\bar{B}^t$ is *diagonal* (in a more general case it will be block diagonal, each block corresponding again to a single element). This means that we can eliminate the unknown $\bar{U}$ at the element level by solving

$$\bar{U} = (\bar{B}\bar{A}^{-1}\bar{B}^t)^{-1}[-\bar{B}\bar{A}^{-1}\bar{C}^t\bar{\Lambda} + \bar{B}\bar{A}^{-1}\bar{G} - \bar{F}]. \tag{1.23}$$

We are finally left with a system of the form

$$H\bar{\Lambda} = R \tag{1.24}$$

with

$$H = \bar{C}\bar{A}^{-1}\bar{B}^t(\bar{B}\bar{A}^{-1}\bar{B}^t)^{-1}\bar{B}\bar{A}^{-1}\bar{C}^t - \bar{C}\bar{A}^{-1}\bar{C}^t$$

and

$$R = \bar{C}\bar{A}^{-1}\bar{B}^t(\bar{B}\bar{A}^{-1}\bar{B}^t)^{-1}[\bar{B}\bar{A}^{-1}\bar{G} - \bar{F}].$$

It is clear that H is symmetric and positive definite. It is easy to see that the procedure for getting from (1.20) to (1.24) is the most common procedure for eliminating internal degrees of freedom, better known as *static condensation.*

Clearly, all that we have described so far applies to the various RT_k, BDM_k, and $BDFM_k$ elements described in Chapter III for the mixed approximation of elliptic problems, as well as to their corresponding elements for quadrilaterals. More generally (and more philosophically) this procedure can be applied to systems of the form (1.7) whenever the matrix A corresponds to a bilinear continuous form on a space V which does not have continuity requirements at the vertices. See ARNOLD–BREZZI [A], BREZZI–DOUGLAS–MARINI [B], BREZZI–DOUGLAS–FORTIN–MARINI [A], BREZZI–DOUGLAS–DURAN–FORTIN [A] for the corresponding proofs in cases more general than the present one. Other examples in which the procedure applies will be presented in Chapter VII.

V.2 A Brief Analysis of the Computational Effort

As we have seen, the introduction of the interelement multiplier λ_h is in general the most effective way of solving a discrete version of (1.2). This implies that a comparison among different kinds of discretizations, as far as the computational effort is concerned, must be done by the light of the "λ–procedure."In this respect, two basic steps must be taken into account. The *first step* is the work which has to be done *at the element level*: basically the hard part of this work is the inversion of the matrix $\bar{A}$ (see (1.21)) and, if u_h has many degrees of freedom per element, also the inversion of $\bar{B}\bar{A}^{-1}\bar{B}^t$ (see (1.23)). This, in our example, was trivial, since $\bar{A}$, on each element, was a 3×3 matrix and $\bar{B}\bar{A}^{-1}\bar{B}^t$ a 1×1 matrix (that is, a scalar). In more general cases those numbers can be bigger. Therefore, it is always a good feature, for a space $\mathfrak{M}$ approximating $H(\mathrm{div};\Omega)$, to have a basis in which the two components are independently assumed. Let us make it clearer with a simple example. If K is a triangle, a reasonable choice for a basis in $RT_0(K)$ is

$$\underline{p}^1 = (1,0);\ \ \underline{p}^2 = (0,1);\ \ \underline{p}^3 = (x,y). \tag{2.1}$$

Now, since $\underline{p}^3$ has two components which are both different from zero, the corresponding local matrix

$$A^K_{ij} = \int_K \underline{p}^i \cdot \underline{p}^j \, dx\, dy \tag{2.2}$$

will have the structure

$$\begin{pmatrix} \otimes & 0 & \otimes \\ 0 & \otimes & \otimes \\ \otimes & \otimes & \otimes \end{pmatrix}, \tag{2.3}$$

where $\otimes$ means, a priori, a nonzero element. On the other hand, **if K is a** rectangle, one will choose as local basis for $RT_{[0]}(K)$

$$\underline{p}^1 = (1,0);\ \ \underline{p}^2 = (x,0);\ \ \underline{p}^3 = (0,1);\ \ \underline{p}^4 = (0,y). \tag{2.4}$$

Now each element of the basis (2.4) has one component identically zero, and the corresponding matrix A^K

$$A^K = \begin{pmatrix} \otimes & \otimes & 0 & 0 \\ \otimes & \otimes & 0 & 0 \\ 0 & 0 & \otimes & \otimes \\ 0 & 0 & \otimes & \otimes \end{pmatrix} \tag{2.5}$$

is block diagonal. An elementary inspection on the spaces RT_k, BDM_k and $BDFM_k$ gives the following outcome:

(2.6) *for* $K =$ *triangle*, then $BDM_k(\equiv (\mathcal{P}_k)^2)$ gives rise to a block diagonal elementary matrix A^k whereas $RT_{[k]}$ and $BDFM_{[k]}$ do not;

(2.7) *for* $K =$ *rectangle*, then $RT_{[k]}$ and $BDFM_{[k]}$ give rise to block diagonal elementary matrices whereas $BDM_{[k]}$ does not.

It must also be noted that the total dimension of A^K also comes into play. For instance, for $K =$ rectangle and $k = 2$ then $RT_{[k]}$ produces a matrix A^K (24×24), which is block diagonal and each of the two blocks is a 12×12 matrix, whereas $BDM_{[k]}$ produces a matrix A^K which is 14×14 (actually made from a (12×12) block-diagonal matrix with two 6×6 blocks, plus two full rows and columns). On the other hand, $BDFM_{[k+1]}$ gives two 9×9 blocks. Obviously, on a uniform mesh with constant coefficients, one just performs *one* inversion *once*, so that the total cost is nothing. But in a general case the inversion of A^K on each K might be expensive.

As far as the matrix $\bar{B}\bar{A}^{-1}\bar{B}^t$ is concerned, usually one gets, on each element, a full matrix so that the total dimension of it (that is, the number of degrees of freedom for u_h in each K) is the only way of comparison.

Let us consider now the second step which is the solution of the final system (1.24) in the unknown λ_h. It is easy to observe that the total number of degrees of freedom for λ_h equals the total number of degrees of freedom for $\underline{p}_h$ which lie on the edges in $\mathcal{E}_h$. In this respect BDM_k produces the same number of λ_h unknowns as RT_k while $BDFM_k$ produces the same number of λ_h unknowns as RT_{k-1}. The same is true for both triangles and rectangles.

Tables 2.1 to 2.4 summarize the computational effort for the various elements.

We have used, for the comparison, $BDFM_{k+1}$ rather than $BDFM_k$ because, as we shall see in the next section, the order of convergence of $BDFM_{k+1}$ is essentially the same as BDM_k or RT_k.

It must also be pointed out that the splitting of the vector space into two (or three) independent components is a crucial starting point for the use of ADI solvers. See for instance DOUGLAS–DURAN–PIETRA [A,B], DOUGLAS–PIETRA [A], BREZZI–DOUGLAS–FORTIN–MARINI [A], and BREZZI–DOUGLAS–DURAN–FORTIN [A].

Element	A^K	$(\bar{B}\bar{A}^{-1}\bar{B}^t)^K$	λ d.o.f. on ∂K
RT_k	$(k^2+4k+3)\times(k^2+4k+3)$	$\frac{(k+1)(k+2)}{2}\times\frac{(k+1)(k+2)}{2}$	$3k+3$
BDM_k	* $\frac{(k^2+3k+2)}{2}\times\frac{k^2+3k+2)}{2}$	$\frac{(k^2+k)}{2}\times\frac{(k^2+k)}{2}$	$3k+3$
$BDFM_{k+1}$	$(k^2+5k+3)\times(k^2+5k+3)$	$\frac{(k+1)(k+2)}{2}\times\frac{(k+1)(k+2)}{2}$	$3k+3$

Table V.1: K = triangle (* two times a system of the dimension indicated)

Element	Matrix A^K	Matrix $(\bar{B}\bar{A}^{-1}\bar{B}^t)^K$	λ d.o.f. on ∂K
$RT_{[k]}$	* $(k^2+3k+3)\times(k^2+3k+3)$	$(k^2+2k+1)\times(k^2+2k+1)$	$4k+4$
$BDM_{[k]}$	$(k^2+3k+4)\times(k^2+3k+4)$	$\frac{(k^2+k)}{2}\times\frac{(k^2+k)}{2}$	$4k+4$
$BDFM_{[k+1]}$	* $\frac{(k^2+5k+4)}{2}\times\frac{(k^2+5k+4)}{2}$	$\frac{(k^2+3k+2)}{2}\times\frac{(k^2+3k+2)}{2}$	$4k+4$

Table V.2 K = rectangle (* two times a system of the dimension indicated)

Element	A^K	$(\bar{B}\bar{A}^{-1}\bar{B}Tt)^K$	λ d.o.f. on ∂K
RT_k	$\frac{k^3+7k^2+14k+8}{2} \times \frac{k^3+7k+14k+8}{2}$	$\frac{k^3+6k^2+11k+6}{6} \times \frac{k^3+6k^2+11k+6}{6}$	$2k^2+6k+4$
BDM_k	* $\frac{k^3+6k^2+11k+6}{6} \times \frac{k^3+6k^2+11k+6}{6}$	$\frac{k^2+3k^2+2k}{6} \times \frac{k^3+3k^2+k}{6}$	$2k^2+4k+4$
$BDFM_{k+1}$	$\frac{k^3+9k^2+22k+16}{2} \times \frac{k^3+9k^2+22k+16}{2}$	$\frac{k^3+6k^2+11k+6}{6} \times \frac{k^3+6k^2+11k+6}{6}$	$2k^2+6k+4$

Table V.3 K = tetrahedron (* three times a system of the dimension indicated)

Element	A^K	$(\bar{B}\bar{A}^{-1}\bar{B}Tt)^K$	λ d.o.f. on ∂K
$RT_{[k]}$	* $(k^3+4k^2+5k+2)\times(k^3+4k^2+5k+2)$	$(k+1)^3 \times (k+1)^3$	$6(k+1)^2$
$BDM_{[k]}$	$\frac{k^3+6k^2+17k+12}{2} \times \frac{k^3+6k^2+17k+12}{2}$	$\frac{k^3+3k^2+2k}{6} \times \frac{k^3+3k^2+2k}{6}$	$\frac{6(k^2+3k+2)}{2}$
$BDFM_{[k+1]}$	* $\frac{k^3+9k^2+20k+12}{6} \times \frac{k^3+9k+20k+12}{6}$	$\frac{k^3+6k^2+11k+6}{6} \times \frac{k^3+6k^2+11k+6}{6}$	$\frac{6(k^2+3k+2)}{2}$

Table V.4 K = cube (* three times a system of the dimension indicated)

V.3 Error Analysis for the Multiplier

Let us consider again, for the sake of simplicity, the approximation of (1.2) by means of the discretization (1.6). We assume from now on that $\Gamma_N = \emptyset$ and $g = 0$ in the notation of Section IV.1. This, if Ω is for instance a convex polygon, will ensure at least H^2-regularity. We have seen in Chapter IV that

$$\|u - u_h\|_0 + \|\underline{p} - \underline{p}_h\|_{H(\mathrm{div};\Omega)} \leq ch\,(\|u\|_2 + \|f\|_1). \tag{3.1}$$

Now if we are going to solve (1.6) through the introduction of the interelement multiplier λ_h, we compute the λ_h unknown first, from (1.24), and then u_h and $\underline{p}_h$ out of it (this is done element by element). However, we still have computed λ_h (which physically must be an approximation of u) and we seek some further use of it. In order to do that, we first need an estimate which is somehow better than (3.1) and was proved first by DOUGLAS–ROBERTS [A]. If $\bar{u}_h$ is the L^2-projection of u onto $\mathfrak{L}_0^0$, then

$$\|\bar{u}_h - u_h\|_0 \leq ch^2\,(\|u\|_2 + \|f\|_1). \tag{3.2}$$

Estimates of this kind can be obtained from the abstract duality results of Chapter II. However, we found it more convenient to sketch a direct proof. To do so, let $\phi \in H^2(\Omega) \cap H_0^1(\Omega)$ be the solution of $\Delta\phi = \bar{u}_h - u_h$. Clearly we have

$$\|\phi\|_2 \leq c\,\|\bar{u}_h - u_h\|_0. \tag{3.3}$$

Set now $\underline{z} = \underline{\mathrm{grad}}\,\phi$ and let $\Pi_h\underline{z}$ be the interpolate of $\underline{z}$ in RT_0. Recall that $(\mathrm{div}\,(\Pi_h\underline{z} - \underline{z}), v_h) = 0$, $\forall v_h \in \mathfrak{L}_0^0$ (see Section III.3.4), so that, in particular, $\mathrm{div}\,\Pi_h\underline{z} = \bar{u}_h - u_h$. Then we have

$$\begin{aligned}
\|\bar{u}_h - u_h\|_0^2 &= (\mathrm{div}\,\Pi_h\underline{z}, \bar{u}_h - u_h) = (\mathrm{div}\,\Pi_h\underline{z}, u - u_h)\\
&= (\underline{p}_h - \underline{p}, \Pi_h\underline{z})\\
&= (\underline{p}_h - \underline{p}, \Pi_h\underline{z} - \underline{z}) + (\underline{p}_h - \underline{p}, \underline{z})\\
&= (\underline{p}_h - \underline{p}, \Pi_h\underline{z} - \underline{z}) + (\underline{p}_h - \underline{p}, \underline{\mathrm{grad}}\,\phi)\\
&= (\underline{p}_h - \underline{p}, \Pi_h\underline{z} - \underline{z}) + (\mathrm{div}\,(\underline{p}_h - \underline{p}), \phi).
\end{aligned} \tag{3.4}$$

Remember that $(\mathrm{div}\,(\underline{p}_h - \underline{p}), v_h) = 0$, $\forall v_h \in \mathfrak{L}_0^0$. Hence, if $\bar{\phi}_h = L^2$-projection of ϕ onto $\mathfrak{L}_0^0$, then (3.4) yields

$$\|\bar{u}_h - u_h\|_0^2 = (\underline{p}_h - \underline{p}, \Pi_h\underline{z} - \underline{z}) + (\mathrm{div}(\underline{p}_h - \underline{p}),\ \phi - \bar{\phi}_h). \tag{3.5}$$

Since

$$\|\underline{z} - \Pi_h\underline{z}\|_0 + \|\phi - \bar{\phi}_h\|_0 \leq ch\,\|\phi\|_2, \tag{3.6}$$

(3.2) follows from (3.1), (3.3), (3.5), and (3.6). □

We are now ready to get some extra information from λ_h. First, if $(\underline{p}, u)$ is the solution of (1.2) and $\underline{q}_h \in \mathfrak{M}^0$, then by Green's formula on each K we have

$$(3.7) \qquad (\underline{p}, \underline{q}_h) + (\operatorname{div} \underline{q}_h, u)_h = \sum_K \int_{\partial K} u\, \underline{q}_h \cdot \underline{n}\, ds = c(u, \underline{q}_h).$$

From (3.7) and the first equation of (1.19) one gets

$$(3.8) \qquad (\underline{p} - \underline{p}_h, \underline{q}_h) + (\operatorname{div} \underline{q}_h, \bar{u}_h - u_h)_h = c(u - \lambda_h, \underline{q}_h), \qquad \forall \underline{q}_h \in \mathfrak{M}^0,$$

where we were allowed to use $\bar{u}_h$ instead of u since div $\underline{q}_h$ is constant in each element. Let us define u_h^* and $\tilde{u}_h$ to be the interpolate in $\mathfrak{L}_k^{1,NC}$ (Section III.2.31) of λ_h and u, respectively, by means of

$$(3.9) \qquad \int_e (u_h^* - \lambda_h)\, ds = \int_e (\tilde{u}_h - u)\, ds = 0, \qquad \forall e \in \mathcal{E}_h.$$

Equation (3.8) implies

$$(3.10) \qquad \sum_K \int_{\partial K} (\tilde{u}_h - u_h^*)\, \underline{q}_h \cdot \underline{n}\, ds = (\underline{p} - \underline{p}_h, \underline{q}_h) + (\operatorname{div} \underline{q}_h, \bar{u}_h - u_h)_h, \qquad \forall \underline{q}_h \in \mathfrak{M}^0.$$

On the other hand, we have by Green's formula

$$(3.11) \qquad \int_{\partial K} (\bar{u}_h - u_h^*)\, \underline{q}_h \cdot \underline{n}\, ds = \int_K \underline{\operatorname{grad}}(\tilde{u}_h - u_h^*) \cdot \underline{q}_h\, dx + \int_K (\tilde{u}_h - u_h^*) \operatorname{div} \underline{q}_h\, dx.$$

A simple scaling argument shows that, for any $\tilde{v}_h \in \mathfrak{L}_1^{NC}$ and for any K in $\mathcal{T}_h$,

$$(3.12) \qquad \|\tilde{v}_h\|_{0,K} \le c \sup_{\underline{q}_h \in RT_0(K)} \frac{\int_K \underline{\operatorname{grad}}\, \tilde{v}_h \cdot \underline{q}_h\, dx + \int_K \tilde{v}_h \operatorname{div} \underline{q}_h\, dx}{h_K^{-1} \|\underline{q}_h\|_{0,K} + \|\operatorname{div} \underline{q}_h\|_{0,K}}$$

so that from (3.12), (3.11), and (3.10) we have

$$(3.13) \qquad \|\tilde{u}_h - u_h^*\|_{0,K} \le c\, \left[h_K\, \|\underline{p} - \underline{p}_h\|_{0,K} + \|\bar{u}_h - u_h\|_{0,K}\right], \quad \forall K \in \mathcal{T}_h,$$

which together with (3.1) and (3.2) gives

$$(3.14) \qquad \|\tilde{u}_h - u_h^*\|_0 \le ch^2\, (\|u\|_2 + \|f\|_1).$$

Since $\|u - \tilde{u}_h\|_0 \le ch^2\, \|u\|_2$ we get by the triangle inequality that

$$(3.15) \qquad \|u - u_h^*\|_0 \le ch^2\, (\|u\|_2 + \|f\|_1).$$

We can now summarize the above results in a theorem.

Theorem 3.1: Let $(\underline{p}_h, u_h, \lambda_h)$ be the solution of (1.19), let u be the solution of (1.1), and let u_h^* be the $\mathfrak{L}_1^{NC}$ interpolant of λ_h defined by (3.9). Then

$$\|u - u_h^*\|_0 \leq ch^2 \left(\|u\|_0 + \|f\|_1\right) \tag{3.16}$$

with c independent of h and u. □

Remark 3.1: The proof that we have given of Theorem 3.1 is somehow "unconventiona."The traditional proof (see for instance ARNOLD–BREZZI [A]) will, as an intermediate step, estimate first the distance of λ_h from the $L^2(\mathcal{E}_h)$ projection $\bar{\lambda}$ of u onto Λ, defined by

$$\int_e (u - \bar{\lambda})\, ds = 0, \qquad \forall e \in \mathcal{E}_h.$$

In particular, in our case one would get

$$\|\bar{\lambda} - \lambda_h\|_{h,-1/2} \leq ch^2 \left(\|u\|_2 + \|f\|_1\right), \tag{3.17}$$

where

$$\|\mu_h\|_{h,-1/2} := \Big(\sum_e |e|\, \|\mu_h\|_{o,e}^2\Big)^{1/2} \tag{3.18}$$

Then (3.15) would follow from (3.17) by extending λ_h in the interior of each K (in our case such an extension is u_h^*). □

Results of type (3.17) hold in much more general cases. For instance one has

$$\|\bar{\lambda} - \lambda_h\|_{h,-1/2} \leq ch^{k+2} \tag{3.19}$$

for RT_k or $RT_{[k]}$ or $BDFM_{k+1}$ or $BDFM_{[k+1]}$, whereas for BDM_k or $BDM_{[k]}$ one has

$$\|\bar{\lambda} - \lambda_h\|_{h,-1/2} \leq ch^{k+2} \quad (k \geq 2), \tag{3.20}$$

$$\|\bar{\lambda} - \lambda_h\|_{h,-1/2} \leq ch^2 \quad (k = 1). \tag{3.21}$$

In (3.19)–(3.21) λ_h is still the interelement multiplier, now in $\Lambda = \mathfrak{L}_k^0(\mathcal{E}_h)$, whereas $\bar{\lambda}$ is the $L^2(\mathcal{E}_h)$-projection of u onto Λ. For the proofs we refer to ARNOLD–BREZZI [A], BREZZI–DOUGLAS–MARINI [B], and BREZZI–DOUGLAS–FORTIN–MARINI [A]. One has now to extend λ_h in the interior of each element in order to derive from (3.19)–(3.21) estimates of type (3.15). This can be done in many ways. We shall indicate here one possible choice.

For K = triangle and k *even* we can define $u_h^* \in P_{k+1}(K)$ simply by setting

$$\int_{e_i} (u_h^* - \lambda_h)\, p_k = 0, \qquad \forall p_k \in P_k(e_i),\ i = 1,2,3, \tag{3.22}$$

$$\int_K (u_h^* - u_h) p_{k-2}\, dx = 0, \qquad \forall p_{k-2} \in P_{k-2}(K),\ \ (k \geq 2) \tag{3.23}$$

It is easy to check that (3.22)and (3.23) determine $u_h^* \in P_{k+1}(K)$ in a unique way. In order to show this, check first that the number of conditions in (3.22) and (3.23) matches correctly the dimension of P_{k+1}:

$$3(k+1) + \frac{(k-1)k}{2} = \frac{(k+2)(k+3)}{2}. \tag{3.24}$$

Then it is enough to show that if $\lambda_h = 0$ and $u_h = 0$, formulas (3.22) and (3.23) yield $u_h^* = 0$. First note that (3.22) (for $\lambda_h = 0$) implies that u_h^*, on each e_i, coincides with $\ell_{k+1}(e_i)$, the Legendre polynomial of degree $k+1$, up to a scaling factor. The continuity of u_h^* at the corners and the fact that for $k+1$ odd, ℓ_{k+1} is antisymmetric will then give $u_h^*|_{\partial K} = 0$. Hence, for $k \geq 2$, this means $u_h^* = b_3 p_{k-2}$ for some $p_{k-2} \in P_{k-2}(K)$ and where b_3 is the cubic bubble on K. Condition (3 23) will now give easily $u_h^* \equiv 0$.

Let us go now to the case K = triangle and k odd. Here the construction (3.22) and (3.23) does not work anymore. We shall indicate another choice that works. Other possible choices can be found in ARNOLD–BREZZI [A], BREZZI–DOUGLAS–MARINI [B]. Let us define, for k odd ≥ 1, ϕ_{k+2} as the polynomial $\in P_{k+2}$ such that $\phi_{k+2} = 0$ at the vertices of K and

$$\frac{\partial \phi_{k+2}}{\partial t}\Big|_{e_i} = \ell_{k+1}(e_i), \qquad i = 1,2,3, \tag{3.25}$$

$$\int_K \phi_{k+2}\, p_{k-1}\, dx = 0, \qquad \forall p_{k-1} \in P_{k-1}(K). \tag{3.26}$$

Note that in (3.25), $\partial/\partial t$ is the counterclockwise tangential derivative and $\ell_{k+1}(e_i)$ is the Legendre polynomial of degree $k+1$ taking the value 1 at the endpoints. We also define $\psi_{k+1} \in P_{k+1}(K)$ by

$$\psi_{k+1} = \frac{\partial \phi_{k+2}}{\partial t} \text{ on } \partial K, \tag{3.27}$$

$$\int_K \psi_{k+1}\, p_{k-2} dx = 0, \qquad \forall p_{k-2} \in P_{k-2}(K),\ \ k \geq 3. \tag{3.28}$$

Note that from (3.25)–(3.27), $\psi_{k+1}|_{e_i} = \ell_{k+1}(e_i)$ $(i = 1,2,3)$. Now we can set

$$S_{k+1} = P_{k+1} \oplus \{\phi_{k+2}\}. \tag{3.29}$$

Our extension u_h^* will be defined as the unique (we have to prove that!) element of S_{k+1} such that

$$\int_{e_i} (u_h^* - \lambda_h)\, p_k ds = 0, \qquad \forall p_k \in P_k(e_i)\ (i = 1, 2, 3), \tag{3.30}$$

$$\int_K (u_h^* - u_h)\, p_{k-2} dx = 0, \qquad \forall p_{k-2} \in P_{k-2}(K), \qquad k \geq 3, \tag{3.31}$$

$$\int_K (u_h^* - u_h)\, \Delta\psi_{k+1} dx = 0. \tag{3.32}$$

Note that the dimensional count (3.24), being independent of the parity of k, still holds since both sides are increased by one. Assume therefore that $\lambda_h = u_h = 0$ and let us show that (3.30)–(3.32) imply $u_h^* = 0$. For this, first note that for every $p_{k+1} \in P_{k+1}(K)$ we have

$$\int_{\partial K} p_{k+1} \frac{\partial \psi_{k+1}}{\partial t} ds = -\int_{\partial K} \frac{\partial p_{k+1}}{\partial t} \psi_{k+1} ds = 0. \tag{3.33}$$

On the contrary,

$$\int_{\partial K} \phi_{k+2} \frac{\partial \psi_{k+1}}{\partial t} ds = -\int_{\partial K} (\psi_{k+1})^2 ds \neq 0. \tag{3.34}$$

Hence (3.30), (with $\lambda_h = 0$) will first give $u_h^* \in P_{k+1}(K)$ (by taking $p_k|_{e_i} = \partial\psi_{k+1}/\partial t|_{e_i}$ and summing over i); then again (3.30) will imply that

$$u_h^* = \alpha\, \psi_{k+1} + b_{k+1} = \alpha\, \psi_{k+1} + b_3\, q_{k-2} \tag{3.35}$$

(where b_{k+1} is a bubble of degree $k+1$ and b_3 is the cubic bubble) for some $\alpha \in \mathbb{R}$ and some $q_{k-2} \in P_{k-2}(K)$. Now using (3.35), (3.28), and (3.31) with $u_h = 0$ we easily get $q_{k-2} = 0$. Finally, (3.32) gives $\alpha = 0$. □

A different approach for reconstructing an approximation $u_h^* \in P_{k+1}(K)$ of u which converges to u faster than u_h can be found for instance in STENBERG [C]. Basically one solves, in every K, a Neumann problem with $\underline{p}_h \cdot \underline{n}$ as boundary data by using u_h^* in order to fix the mean value in each element. Another approach (BRAMBLE–XU [A]) consists instead of taking u_h and p_h as the first and second terms of a suitable Taylor expansion. Note, however, that these methods will produce a completely discontinuous u_h^*, whereas the previous one was simply nonconforming.

Remark 3.2: Let us go back to the simplest case of the lowest-order element $RT_0 \equiv BDFM_1$. It has been proved by MARINI [C] that if we consider the space $\mathfrak{L}_1^{1,NC}$ and define $u_h^* \in \mathfrak{L}_1^{1,NC}$ to be the solution of

$$\sum_T \int_T \underline{\text{grad}}\, u_h^* \cdot \underline{\text{grad}}\, v_h \, dx = \int_\Omega f v_h \, dx, \qquad \forall v_h \in \mathfrak{L}_1^{NC}, \tag{3.36}$$

then, for f piecewice constant, one can compute, a posteriori, the solution $(\underline{p}_h, u_h)$ of (1.6) through the formulas

$$\underline{p}_h(\underline{x})|_K = \underline{\text{grad}}\, u_h^* + (\underline{x} - \underline{x}_K)\, \frac{f}{2}\bigg|_K, \qquad \forall K \in \mathcal{T}_h, \tag{3.37}$$

$$u_h|_K = \frac{1}{|\text{area}(K)|} \int_K u_h^* \, dx + O(h^2), \qquad \forall K \in \mathcal{T}_h, \tag{3.38}$$

where $\underline{x}_K$ is the barycenter of K. Formulas (3.37) and (3.38) (in particular (3.37)) are specially interesting because the *principle* of (3.36), and therefore its implementation and use, is much simpler that the principle of (1.6). On the other hand, experimental results show that in some applications the accuracy of (1.6), as far as $\underline{p}_h$ is concerned, is superior to the accuracy of the traditional methods (see, e.g., MARINI–SAVINI [A]) and that the correction (3.37) away from $\underline{x}_K$ (say, at ∂K) has a relevant improving effect on the accuracy. □

V.4 Error Estimates in Other Norms

We have seen in Chapter IV that all the families of mixed finite element methods for the Laplace operator (and hence for more general elliptic problems) satisfy the inf–sup condition and therefore provide optimal error estimates in the "natural norms,"which are the $H(\text{div};\Omega)$ norm for $\underline{p} - \underline{p}_h$ and the $L^2(\Omega)$-norm for $u - u_h$. We have also seen in this chapter that if one introduces Lagrange multipliers λ_h in order to solve (1.6) (and in general one does want to do so), then it is possible to obtain some additional information that allows one to construct a new approximation u_h^* of u that provides some extra accuracy for u in the $H^1(\Omega)$-norm. In this section we will present some other error estimates for $\underline{p} - \underline{p}_h$, $u - u_h$, and $u - u_h^*$ in other norms which might be interesting for applications. In particular, we shall deal with L^∞-norms and $H^{-s}(\Omega)$-norms. The interest of using L^∞-norms (especialy for $\underline{p} - \underline{p}_h$) is quite obvious in the applications: a large stress field in a very small region can have a small L^2-norm but will be very dangerous for safety reasons. The interest of having dual estimates, like the estimates in $H^{-s}(\Omega)$ $(s > 0)$, can only be understood as a prerequisite to the use of a "smoothing post-processor" (see, e.g., BRAMBLE–SCHATZ [A,B]). We shall not present here such smoothing post-processors; however ,we can describe their features: if you have a continuous solution (say u) and an approximate solution (say u_h) such that $u - u_h$ is small (say $O(h^{s+k})$)

in some dual norm $\|\cdot\|_{-s}$, then you can operate some "local" and relatively simple averages on u_h in order to produce a new approximation u_h^s such that $\|u - u_h^s\|_0 = O(h^{s+k})$. We refer to BRAMBLE–SCHATZ [A,B] for a more precise information.

Let us now list the error estimates which have been proved so far in the L^∞-norms. We shall only quote the more recent ones (and more accurate) obtained by GASTALDI–NOCHETTO [A,B]. Previous results where obtained by JOHNSON–THOMEE [A], DOUGLAS–ROBERTS [A,B], and SCHOLZ [B,D,E].

Let us see, for instance, the spaces RT_k or $RT_{[k]}$. Then one has

$$\|u-u_h\|_{L^\infty} + \|\underline{p}-\underline{p}_h\|_{L^\infty} \le ch^{k+1}. \tag{4.1}$$

Note that, for $k = 0$, (4.1) holds only if f is smooth enough inside each element. Moreover, for $k \ge 0$, the assumption $u \in W^{k+2,\infty}(\Omega)$ is obviously required. We also have a superconvergence result for $u_h - P_h u$ (here $P_h u = L^2(\Omega)$-projection of u onto $\mathfrak{L}_k^0$):

$$\|P_h u - u_h\|_{L^\infty} \le ch^{k+2}\,|\log h|, \tag{4.2}$$

where again u is assumed in $W^{k+2,\infty}(\Omega)$ and some extra regularity for f is needed for $k = 0$. Finally, for the case of rectangular elements one gets, for $k \ge 0$,

$$|u(S) - u_h(S)| \le ch^{k+2}\,|\log h|^2\,(\|f\|_{k,\infty,\Omega} + \delta_{k,0}\,\|f\|_{H^1}) \tag{4.3}$$

at the Gauss–Legendre points S of each element. It is also possible to study the error $u - u_h^*$. One has, for $k \ge 0$,

$$\|u - u_h^*\|_{L^\infty} \le ch^{k+2}\,|\log h|^2\,(\|f\|_{k,\infty,\Omega} + \delta_{k,0}\|f\|_{H^1}), \tag{4.4}$$

and, for $u \in W^{k+2,\infty}$ (and for smoother f if $k = 0$)

$$\|u - u_h^*\|_{L^\infty} \le ch^{k+2}\,|\log h|. \tag{4.5}$$

Similar results hold for BDM and $BDFM$ spaces and for their analogues in three dimensions. □

As far as the dual norms are concerned we have for RT_k and $RT_{[k]}$ or $BDFM_{k+1}$ and $BDFM_{[k+1]}$ elements, in two and three variables

$$\|u-u_h\|_{-s} + \|\underline{p}-\underline{p}_h\|_{-s} + \|\mathrm{div}(\underline{p}-\underline{p}_h)\|_{-s} \le ch^{k+s+1},\quad 0 \le s \le k+1, \tag{4.6}$$

whereas for BDM_k or $BDM_{[k]}$ elements, in two or three variables, one has

$$\|u - u_h\|_{-s} + \|\mathrm{div}(\underline{p} - \underline{p}_h)\|_{-s} \le ch^{k+s},\quad 0 \le s \le k, \tag{4.7}$$

and

$$\|\underline{p} - \underline{p}_h\|_{-s} \leq ch^{k+s+1}, \quad 0 \leq s \leq k-1. \tag{4.8}$$

For more precise estimates involving explicitly the regularity of the solution we refer to DOUGLAS–ROBERTS [B], BREZZI–DOUGLAS–MARINI [B], BREZZI–DOUGLAS–FORTIN–MARINI [A], BREZZI–DOUGLAS–DURAN–FORTIN [A]. Interior estimates can be found for instance in DOUGLAS–MILNER [A].

V.5 Application to an Equation Arising from Semiconductor Theory

We now consider a special case of application of mixed finite element methods that is interesting in the simulation of semiconductor devices. Let us assume that we have to solve an equation of the type

$$\operatorname{div}(\varepsilon \, \underline{\operatorname{grad}} \, u + \underline{u} \cdot \underline{\operatorname{grad}} \, \psi) = f \text{ in } \Omega, \tag{5.1}$$

and assume, for the sake of simplicity, that we have Dirichlet boundary conditions

$$u = g \text{ on } \partial\Omega. \tag{5.2}$$

Note, however, that, in practice, we will always have a Neumann boundary condition $\varepsilon \, \underline{\operatorname{grad}} \, u + \underline{u} \cdot \underline{\operatorname{grad}} \, \psi = 0$ on a part of $\partial\Omega$. In (5.1) we may assume ψ to be known, and in the computations we shall also assume that ψ is piecewise linear; this is realistic since in practice ψ will be the discretized solution of another equation (coupled with (5.1)). Assume moreover that ε is constant and small. In order to present the mixed exponential fitting approximation of (5.1) and (5.2), (BREZZI–MARINI–PIETRA [A,B,C]), we first introduce the Slotboom variable

$$\rho = e^{-\psi/\varepsilon} u \tag{5.3}$$

with its boundary value

$$\chi = e^{-\psi/\varepsilon} g. \tag{5.4}$$

In order to simplify the notation we shall often write

$$\phi = \psi/\varepsilon. \tag{5.5}$$

Using unknown ρ, problem (5.1) and (5.2) becomes

$$\begin{cases} \varepsilon \operatorname{div}(e^{-\phi} \underline{\operatorname{grad}} \, \rho) = f \text{ in } \Omega, \\ \rho = \chi \text{ on } \partial\Omega. \end{cases} \tag{5.6}$$

Note that the quantity

$$\underline{p} = \varepsilon\, e^{-\phi}\, \underline{\mathrm{grad}}\, \rho = \varepsilon\, \underline{\mathrm{grad}}\, u + \underline{u} \cdot \underline{\mathrm{grad}}\, \psi \tag{5.7}$$

(which has the physical meaning of the electric current $\underline{J}$ through the device), is the most relevant unknown of the problem.

We now apply a mixed method to the solution of (5.6). By choosing the lowest-order Raviart–Thomas method, formulation (1.19) becomes: find $(\underline{p}_h,\ \rho_h,\ \lambda_h) \in \mathfrak{M}^0 \times \mathfrak{L}_0^0 \times \Lambda$ such that

$$\begin{cases} (\varepsilon^{-1} e^{\phi} \underline{p}_h, \underline{q}_h) + (\rho_h, \operatorname{div} \underline{q}_h)_h = c(\lambda_h, \underline{q}_h), & \forall \underline{q}_h \in \mathfrak{M}^0, \\ (\sigma_h, \operatorname{div} \underline{p}_h)_h = (f, \sigma_h), & \forall \sigma_h \in \mathfrak{L}_0^0, \\ c(\mu_h, \underline{p}_h) = 0, & \forall \mu_h \in \Lambda_0, \end{cases} \tag{5.8}$$

where $(\cdot,\cdot)_h$ and $c(\mu, \underline{q}_h)$ are defined as in (1.19). By static condensation, (5.8) can be reduced as in (1.24) to the form

$$H\lambda = R, \tag{5.9}$$

with H symmetric and positive definite. We also point out that H will be an M-matrix (see for instance VARGA [A]) provided that the triangulation is of a weakly acute type. However, the scheme (5.8) (and the unknown ρ) are not suitable for actual computations. Indeed, one can see from (5.3) and (5.4) that ρ can become very large or very small in different parts of the domain Ω when ε is very small. Hence, we go back to the variable u. Since, as we have seen, λ_h in (5.8) will be an approximation of ρ at the interelement boundaries, we can use the inverse transformation of (5.3) in the form

$$u_h = e^{\bar{\phi}_h}\ \lambda_h \tag{5.10}$$

where $\bar{\phi}_h \in \mathfrak{L}_0^0(\mathcal{E}_h)$ is defined as

$$\int_{e_i} e^{\bar{\phi}_h}\, ds = \int_{e_i} e^{\phi_h}\, ds, \qquad \forall e_i \in \mathcal{E}_h. \tag{5.11}$$

Problem (5.8) now becomes: find $(\underline{p}_h,\ \rho_h,\ u_h) \in \mathfrak{M}^0 \times \mathfrak{L}_0^0 \times \Lambda_g$ such that

$$\begin{cases} (\varepsilon^{-1} e^{\bar{\phi}_h} \underline{p}_h, \underline{q}_h) + (\rho_h, \operatorname{div} \underline{q}_h)_h = c(e^{-\phi_h} u_h, \underline{q}_h), & \forall \underline{q}_h \in \mathfrak{M}^0, \\ (\sigma_h, \operatorname{div} \underline{p}_h)_h = (f, \sigma_h), & \forall \sigma_h \in \mathfrak{L}_0^0, \\ c(\mu_h, \underline{p}_h) = 0, & \forall \mu \in \Lambda_0 \end{cases} \tag{5.12}$$

The static condensation procedure applied to (5.12) now produces a system in the sole unknown u_h of the form

$$\tilde{H} u_h = \tilde{R}, \tag{5.13}$$

where the unknown, the coefficients, and the right-hand side have a reasonable size. Moreover, it is easy to check that the passage from H to $\tilde{H}$ involves only the multiplication of each row by a factor of the type $e^{-\phi_h}$, which does not alter the M-character of the matrix. Hence, if the decomposition is of weakly acute type, $\tilde{H}$ will be an M-matrix.

The most relevant feature of this approach is, however, that the approximation $\underline{p}_h$ of the current obtained by (5.12) will now have continuous normal components at the interelement boundaries. We have, therefore, a strong conservation of the current.

Remark 5.1: Problem (5.6) could also be discretized by dual hybrid methods. However, in this case the conservation of the current will hold only in a weak sense (BREZZI–MARINI–PIETRA [B]). □

Remark 5.2: It is easy to check that the one-dimensional version of this approach reproduces the celebrated Sharfetter–Gummel method, also known as exponential fitting method. The use and the analysis of nonstandard formulations (involving the harmonic average of the coefficients) in one dimension can be found in BABUŠKA–OSBORN [A]. □

Remark 5.3: It can be checked that, for very small ε, the scheme (5.13) produces an up-wind discretization of (5.1). See BREZZI–MARINI–PIETRA [C] for this kind of analysis. □

Remark 5.4. If (5.1) contains a zero-order term

$$\operatorname{div}(\varepsilon\, \underline{\operatorname{grad}}\, u + \underline{u} \cdot \underline{\operatorname{grad}}\, \psi) + cu = f,$$

then, in general, the matrix H in (5.9) will not be an M-matrix any longer, and the same will be true for the matrix $\tilde{H}$ in (5.13). To circumvent this difficulty, one can change the choice of the space $\mathfrak{M}^0$. We refer to MARINI–PIETRA [A] for a general theory of *nonconforming mixed methods* and to MARINI–PIETRA [B] for applications to semiconductor devices. □

V.6 How Things Can Go Wrong.

We considered so far in this chapter and in Chapter IV many examples of mixed finite element methods that can be used for linear elleptic problems. They are based on elements for the approximation of $H(\div\,.\,\Omega)$ presented in Chapter III. As it is normal, we only restricted our attention to the cases that work. This, together with the simplicity of the proofs, might lead the reader to think that we are dealing with a particularly easy, albeit trivial, case. The

following example shows that mixed methods for linear elliptic problems are, in a sense, very delicate, and that one has to handle them with care. Let us consider, to illustrate this assertion, the model problem

$$\begin{cases} u'' = 1 & \text{in } I =]-1,1[, \\ u(-1) = u(1) = 0. \end{cases} \tag{6.1}$$

It is clear that the exact solution of (6.1) is given by $u(x) = \frac{1}{2}(x^2 - 1)$. By setting, as usual,

$$\begin{cases} \sigma = u' & (= x \text{ in our case}), \\ \Sigma = H^1(I), & V = L^2(I), \end{cases} \tag{6.2}$$

we may write the mixed formulation of (6.1) as: find $\sigma \in \Sigma$ and $u \in V$ such that

$$\begin{cases} (\sigma, \tau) + (u, \tau') = 0, & \forall \tau \in \Sigma \\ (\sigma', v) = (1, v), & \forall v \in V. \end{cases} \tag{6.3}$$

Choosing two finite-dimensional subspaces $\Sigma_h \subset \Sigma$ and $V_h \subset V$, we have the discrete formulation: find $\sigma_h \in \Sigma_h$ and $u_h \in V_h$ such that

$$\begin{cases} (\sigma_h, \tau_h) + (u_h, \tau_h') = 0 & \forall \tau_h \in \Sigma_h \\ (\sigma_h', v_h) = (1, v_h) & \forall v_h \in V_h \end{cases} \tag{6.4}$$

It is very easy to check that, for a given uniform decomposition of I into subintervals I_h, if we choose $\Sigma_h = \mathfrak{L}_1^1$ and $V_h = \mathfrak{L}_0^0$, that is, piecewise constants, then (6.4) has a unique solution and $\sigma_h \equiv \sigma$. Moreover, u_h is first-order accurate (and second order accurate at the midpoints of the intervals).

Let us analyze this situation in the abstract framework of Chapter II. We have here

$$\begin{cases} a(\sigma, \tau) = \displaystyle\int_I \sigma\tau \, dx, \\ b(v, \tau) = \displaystyle\int_I v\tau' \, dx. \end{cases} \tag{6.5}$$

The operator $B\tau$ is here simply $d\tau/dx$. It is evidently surjective (hence the inf–sup condition) from $H^1(I)$ onto $L^2(I)$ and its kernel consists of constants. The bilinear form $a(\cdot,\cdot)$ is then coercive on the kernel of B, as needed, but not on $H^1(I)$.

Now with the discretization introduced above, we have exactly for the discrete operator B_h

$$\operatorname{Ker} B_h = \operatorname{Ker} B \cap \Sigma_h \subset \operatorname{Ker} B. \tag{6.6}$$

This highly desirable inclusion of kernels makes $a(\cdot,\cdot)$ coercive on $\operatorname{Ker} B$. The inf–sup condition can also be checked directly: for $v_h \in V_h$ (piecewise constant) one obviously can build $\hat{\tau}_h \in \Sigma_h$ such that $d\hat{\tau}_h/dx = v_h$. Hence

$$\sup_{\tau_h} \frac{\int_I v_h \tau_h' \, dx}{\|\tau_h\|_1} \geq \frac{\int_I v_h \hat{\tau}_h' \, dx}{\|\hat{\tau}_h\|_1} \geq \frac{\|v_h\|^2}{\|\hat{\tau}_h\|} \geq k\|v_h\|. \tag{6.7}$$

In fact, if N subintervals are used, the matrix corresponding to B_h is $N \times N+1$ and of maximal rank. $\operatorname{Ker} B_h$ is one-dimensional and consists of constants. Note that (6.6) implies, using (II.2.23), that we will have $\sigma_h \equiv \sigma$ for our model problem.

Now a naive idea, which is indeed correct in standard finite element formulations, is that using a richer approximation should yield better results. This would be true if one increased simultaneously the degrees of polynomials in both Σ_h and V_h e.g. moving to $\Sigma_h = \mathfrak{L}_2^1$, $V_h = \mathfrak{L}_1^0$. However, increasing Σ_h or V_h alone may lead to strange results.

It is not surprising that we cannot increase the space V_h too much without disatrous effects on the inf–sup condition. For example, taking $\Sigma_h = \mathfrak{L}_1^1$ and $V_h = \mathfrak{L}_1^0$ changes the matrix associated to B_h to a $2N \times N+1$ matrix which is still of rank N. (The additional components $\tilde{v}_h$ of V_h satisfy $\int_I \tilde{v}_h \tau_h' \, dx = 0$.)

Hence, we have only created here a nonzero kernel for B_h^T, introducing spurious zero energy modes in the solution. Doing so we make the matrix of the system singular. We shall come back in Chapter VI to this question of spurious modes.

Let us see what happens if instead we increase Σ_h by taking $\Sigma_h = \mathfrak{L}_2^1$ but keeping $V_h = \mathfrak{L}_0^0$.

This can be seen as adding to the previous approximation quadratic bubble function $b_k(x)$ in each subinterval I_k. An easy computation now shows that one has

$$\sigma_h(x) = x - \sum_k \frac{(x, b_k)}{\|b_k\|_0^2} b_k(x) \tag{6.8}$$

and that the error

$$\sigma - \sigma_h = \sum_k \frac{(x, b_k)}{\|b_k\|_0^2} b_k(x) \tag{6.9}$$

does not converge to zero in $L^2(I)$ (not even weakly) when the mesh size goes to zero. The reason for this is that, in this case, we lost the inclusion $\operatorname{Ker} B_h \subset \operatorname{Ker} B$ and the constant α_h^1 appearing in (II.2.8) is now given by

$$\alpha_h' = h^2/(h^2 + 10), \tag{6.10}$$

as one can see with a simple computation on the b_k's.

One, therefore, sees that we must keep a delicate balance between coerciveness on the kernel of B and the inf–sup condition which are in a sense conflicting conditions with respect to the choice of spaces.

Finally, let us give a last look at the solution (6.8). An equivalent form of problem (6.4) is the constrained minimization problem

$$\inf_{\tau_h \in \mathfrak{L}_2^1} \frac{1}{2} \int_I |\tau_h|^2 \, dx, \tag{6.11}$$

$$\int_I v_h \tau_h' \, dx = \int_I 1 v_h \, dx, \qquad \forall v_h \in \mathfrak{L}_0^0. \tag{6.12}$$

But the bubbles $b_k(x)$ are transparent with respect to (6.12) so that they are free to decrease the L^2-norm of τ_h in any way. Indeed, the norm of τ_h in (6.8) is smaller than the L^2-norm of the exact solution $\sigma = x$, and using a richer space spoils the solution instead of making it better.

V.7 Augmented Formulations (Galerkin Least Squares Methods)

We have seen in the previous section the importance of both conditions of the general theory of Chapter II, namely, the coerciveness on the kernel and the inf–sup condition. We shall now apply the ideas of Section I.5 to bypass one ore both of these conditions. The methods presented should be seen as modeling more complex situations and not as having a practical importance by themselves. We already considered a similar idea in Section IV.3 when studying elasticity problems.

Let us first consider the simplest modification, enabling us to obtain coerciveness on the whole space and not only on the kernel. We shall use the augmented formulation (I.5.2) for which we write the optimality conditions:

$$\begin{cases} \displaystyle\int_\Omega \underline{p}\cdot\underline{q} \, dx + \int_\Omega u \operatorname{div} \underline{q} \, dx + \beta \int_\Omega (\operatorname{div} \underline{p} + f) \operatorname{div} \underline{q} \, dx = 0, & \forall q \in H(\operatorname{div};\Omega), \\ \displaystyle\int_\Omega \operatorname{div} \underline{p} \, v \, dx + \int_\Omega f v \, dx = 0, & \forall v \in L^2(\Omega). \end{cases} \tag{7.1}$$

The bilinear form $a(\underline{p}, \underline{q})$ is now defined by

$$a(\underline{p}, \underline{q}) = \int_\Omega \underline{q}\cdot\underline{q} \, dx + \beta \int_\Omega \operatorname{div} \underline{p} \operatorname{div} \underline{q} \, dx \tag{7.2}$$

and we obviously have, for any $\beta > 0$ with $\gamma = \min(1, \beta)$

$$a(\underline{p}, \underline{p}) \geq \gamma \, \|\underline{p}\|_{H(\operatorname{div};\Omega)}. \tag{7.3}$$

Hence for the discretization of (7.1), we only have to worry about the inf–sup condition.

It is now obvious that, for instance, in Section V.6 we could now choose a quadratic approximation for p, and a constant approximation for u, without any problem.

Similarly, in more general two-dimensional cases, if we were interested in employing a continuous approximation for $\underline{p}$, we might choose, for instance, the MINI element introduced for the Stokes problem or the elasticity problems in Chapter IV or any other of the elements well suited for the Stokes problem which we shall study in Chapter VI.

If we now want to avoid as well the problem of the inf–sup condition, we can use formulation (I.5.3) for which the optimality condition can be written as

$$(7.4)\qquad \begin{cases} \displaystyle\int_\Omega \underline{p}\cdot\underline{q}\,dx + \int_\Omega u\,\mathrm{div}\,\underline{q}\,dx + \beta\int_\Omega (\mathrm{div}\,\underline{p}+f)\mathrm{div}\,\underline{q}\,dx \\ \qquad\qquad \displaystyle - \alpha\int_\Omega (\underline{p}-\underline{\underline{\mathrm{grad}}}\,u)\cdot\underline{q} = 0, \quad \forall \underline{q}\in H(\mathrm{div};\Omega), \end{cases}$$

$$(7.5)\qquad \begin{cases} \displaystyle\alpha\int_\Omega (\underline{\underline{\mathrm{grad}}}\,u-p)\cdot\underline{\underline{\mathrm{grad}}}\,v\,dx - \int_\Omega v\,\mathrm{div}\underline{p}\,dx \\ \qquad\qquad \displaystyle - \int_\Omega fv\,dx = 0 \quad \forall v\in H_0^1(\Omega). \end{cases}$$

It is easy to check that we now have a problem of the form (II.1.36)

$$(7.6)\qquad \begin{cases} a(\underline{p},\underline{q}) + b(u,\underline{q}) = \langle F,\underline{q}\rangle, & \forall \underline{q}\in H(\mathrm{div};\Omega) \\ b(v,\underline{p}) - c(u,v) = \langle G,v\rangle, & \forall v\in H_0^1(\Omega). \end{cases}$$

where,

$$(7.7)\qquad a(\underline{p},\underline{q}) = (1-\alpha)\int_\Omega \underline{p}\cdot\underline{q}\,dx + \beta\int_\Omega \mathrm{div}\underline{p}\,\mathrm{div}\,\underline{q}\,dx,$$

$$(7.8)\qquad b(u,\underline{q}) = (1-\alpha)\int_\Omega u\,\mathrm{div}\,\underline{q}\,dx,$$

$$(7.9)\qquad c(u,v) = \alpha\int_\Omega \underline{\underline{\mathrm{grad}}}\,u\cdot\underline{\underline{\mathrm{grad}}}\,v\,dx.$$

If we choose $0<\alpha<1$ and $\beta>0$, conditions (II.1.37) and (II.1.38) are satisfied and stability and optimal error estimates follow.

Remark 7.1: It is obviously also possible to use the variational formulation (5.11) instead of (5.6) and to obtain convergence for every $\alpha>0$ and $\beta>0$. □

We have rapidly presented here the basic idea of using augmented formulations. We refer for more details to BREZZI–FORTIN–MARINI [A].

VI

Incompressible Materials and Flow Problems

Although the approximation of incompressible flows by finite element methods has grown quite independently of the main stream of mixed and hybrid methods, it was soon recognized that a precise analysis requires the framework of mixed methods. In many cases, one may apply directly the techniques and results of Chapter II. In particular, the elements used are often standard elements or simple variants of standard elements. The specificity of Stokes problem has, however, led to the development of special techniques; we shall present some of them that seem particularly interesting. Throughout this study the main point will be to make a clever choice of elements leading to the satisfaction of the inf–sup condition. This chapter, after a summary of the problem, will present examples of elements and techniques of proof. It will not be possible to analyze fully all elements for which results are known; we shall try to group them by families which can be treated by similar methods. These families will be arbitrary and will overlap in many cases. Besides this presentation of elements, we shall also consider solution techniques by penalty methods and will develop the related problem of almost incompressible elastic materials. We shall consider the equivalence of penalty methods and mixed methods and some questions arising from it.

Finally, a section will be devoted to numerical considerations, in particular to the construction of a divergence-free basis for the discrete problems.

VI.1 Introduction

We have already considered in Chapter I (Example 3.1) the Stokes problem or creeping flow problem for an incompressible fluid. We had written it as a system of variational equations,

$$(1.1)\quad \begin{cases} 2\mu \displaystyle\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx - \int_\Omega \underline{f} \cdot \underline{v}\, dx - \int_\Omega p \operatorname{div} \underline{v}\, dx = 0, & \forall \underline{v} \in V, \\ \displaystyle\int_\Omega q \operatorname{div} \underline{u}\, dx = 0, & \forall q \in Q, \end{cases}$$

where $V = (H_0^1(\Omega))^2$ and $Q = L^2(\Omega)$. In this formulation, $\underline{u}$ is the velocity of the fluid and p its pressure. An analogue problem arises for the displacement of an incompressible elastic material.

For an elastic material we have, following Chapter I (Example 2.2), to solve the variational equation

$$(1.2)\quad 2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) \cdot \underline{\underline{\varepsilon}}(\underline{v})\, dx + \lambda \int_\Omega \operatorname{div} \underline{u} \operatorname{div} \underline{v}\, dx = \int_\Omega \underline{f} \cdot \underline{v}\, dx, \qquad \forall v \in V.$$

The case where λ is large, (or, equivalently, when $\nu = \lambda/2(\lambda + \mu)$ approaches $1/2$) can be considered as an approximation of (1.1) by a penalty method as in Section II.4. The limiting case is exactly (1.1) up to the fact that $\underline{u}$ is a displacement instead of a velocity. Problems where λ is large are quite common and correspond to almost incompressible materials. Equation (1.2) can be considered as a penalty approximation of (1.1). Results of Chapter II can be applied and give conditions under which error estimates can be found that do not depend on λ.

It is also worth recalling that, defining

$$(1.3)\qquad A\underline{u} = \begin{cases} \dfrac{\partial^2 u_1}{\partial x_1^2} + \dfrac{1}{2}\dfrac{\partial}{\partial x_2}\Big(\dfrac{\partial u_1}{\partial x_2} + \dfrac{\partial u_2}{\partial x_1}\Big), \\ \dfrac{\partial^2 u_2}{\partial x_2^2} + \dfrac{1}{2}\dfrac{\partial}{\partial x_1}\Big(\dfrac{\partial u_1}{\partial x_2} + \dfrac{\partial u_2}{\partial x_1}\Big), \end{cases}$$

that is, $A\underline{u} = \operatorname{div} \underline{\underline{\varepsilon}}(\underline{u})$, we have $2\mu\, A\underline{u} = \mu\, \Delta\underline{u} + \mu\, \underline{\operatorname{grad}} \operatorname{div} \underline{u}$. Problems (1.1) and (1.2) are then respectively equivalent to

$$(1.4)\qquad \begin{cases} -2\mu\, A\underline{u} + \underline{\operatorname{grad}}\, p = -\mu\, \Delta\underline{u} + \underline{\operatorname{grad}}\, p = \underline{f}, \\ \operatorname{div} \underline{u} = 0, \\ \underline{u}|_\Gamma = 0, \end{cases}$$

and

$$(1.5)\qquad -2\mu\, A\underline{u} - \lambda\, \underline{\operatorname{grad}} \operatorname{div} \underline{u} = -\mu\, \Delta\underline{u} - (\lambda + \mu)\, \underline{\operatorname{grad}} \operatorname{div} \underline{u} = \underline{f}.$$

Remark 1.1: The problems described above are, of course, physically unrealistic, as they involve body forces and homogeneous Dirichlet boundary conditions. The aim of doing so is to avoid purely technical difficulties and implies no loss of generality. The results obtained will be valid, unless otherwise stated, for all acceptable boundary conditions.

To approximate the Stokes problem, two approaches follow quite naturally from the preceding considerations. *The first* is to use system (1.1) and to discretize $\underline{u}$ and p by standard (or less standard) finite element spaces. *The second* one is to use formulation (1.2) with λ large as a penalty approximation to system (1.1).

It rapidly became clear that both these approaches could yield strange results. In particular, the first often led to nonconvergence of pressure and the second to a *locking mechanism*, the numerical solution being uniformly null.

For velocity–pressure approximations, empirical cures were found by HOOD and TAYLOR [A], HUGHES–ALLIK [A] and others. At about the same time some elements using discontinuous pressure fields were shown to work properly (FORTIN [B], CROUZEIX–RAVIART [A]) from the mathematical point of view.

For the penalty method, the cure was found in selective or reduced integration procedures. This consisted in evaluating terms like $\int_\Omega \operatorname{div} \underline{u} \operatorname{div} \underline{v}\, dx$ by quadrature formulas of low order. This *sometimes* led to good results.

It was finally stated (HUGHES–MALKUS [A]), even if the result was implicit in earlier works (BERCOVIER [A]), that the analysis underlying the two approaches must be the same. Penalty methods are equivalent to some mixed methods. A penalty method works if and only if the associated mixed method works (BERCOVIER [B]).

We now try to develop these results. We apologize from the beginning for not treating every aspect of the problem that is still the object of a rapidly growing literature.

VI.2 The Stokes Problem as a Mixed Problem

VI.2.1 Mixed formulation

We shall describe in this section how the Stokes problem (1.1) can be analyzed in the general framework of Chapter II. Defining as above $V = (H_0^1(\Omega))^2$, $Q = L^2(\Omega)$, and

$$a(\underline{u}, \underline{v}) = 2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx, \tag{2.1}$$

$$b(\underline{v}, q) = -\int_\Omega q \operatorname{div} \underline{v}\, dx, \tag{2.2}$$

problem (1.1) can clearly be written in the form

$$(2.3)\qquad \begin{cases} a(\underline{u},\underline{v}) + b(\underline{v},p) = (\underline{f},\underline{v}), & \forall \underline{v} \in V, \\ b(\underline{u},q) = 0, \qquad \forall q \in Q, \end{cases}$$

and is a saddle point problem in the sense of Chapter II. Indeed, we have already seen that p is the Lagrange multiplier associated with the incompressibility constraint.

In the notations of Chapter II, we can write

$$(2.4)\qquad B = -\mathrm{div} : \ (H_0^1(\Omega))^2 \to L^2(\Omega)$$

and

$$(2.5)\qquad B^t = \mathrm{grad} : \ L^2(\Omega) \to (H^{-1}(\Omega))^2.$$

It is clear that the kernel $\mathrm{Ker}\, B^t$ is one dimensional and consists of constant functions. On the other hand, we have (e.g., TEMAM [A])

$$(2.6)\qquad \mathrm{Im}\ B = \{q \mid \int_\Omega q\, dx = 0\}.$$

This is a closed subspace of $L^2(\Omega)$ and the operator $B = -\mathrm{div}$ possesses a continuous lifting. Now choosing an approximation $V_h \subset V$ and $Q_h \subset Q$ yields a discrete problem

$$(2.7)\qquad \begin{cases} 2\mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_h) : \underline{\underline{\varepsilon}}(\underline{v}_h) dx - \int_\Omega p_h \ \mathrm{div}\ \underline{v}_h\, dx = \int_\Omega \underline{f} \cdot \underline{v}_h\, dx, & \forall \underline{v}_h \in V, \\ \int_\Omega q_h \ \mathrm{div}\ \underline{u}_h\, dx = 0, \qquad \forall q_h \in Q_h. \end{cases}$$

As the bilinear form $\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_h) : \underline{\underline{\varepsilon}}(\underline{v}_h)\, dx$ is coercive on $(H_0^1(\Omega))^2$ there is no problem as to the existence of a solution $(\underline{u}_h, p_h)$ to problem (2.7). We thus try to obtain estimates of the errors $\|\underline{u} - \underline{u}_h\|$ and $\|p - p_h\|_Q$.

We can first see that the solution $\underline{u}_h$ *is not in general divergence-free.* Indeed the bilinear form $b(\underline{v}_h, q_h)$ defines a discrete divergence operator,

$$(2.8)\qquad B_h = -\mathrm{div}_h : \ V_h \to Q_h.$$

(It is convenient here to identify $Q = L^2(\Omega)$ and $Q_h \subset Q$ with their dual spaces). In fact, we have

$$(2.9)\qquad (\mathrm{div}_h\ \underline{u}_h, q_h)_Q = \int_\Omega \mathrm{div}\ \underline{u}_h\ q_h\, dx,$$

and, thus, $\mathrm{div}_h \underline{u}_h$ is the L^2-projection of div $\underline{u}_h$ on Q_h.

The discrete divergence operator coincides with the standard divergence operator if div $V_h \subset Q_h$. Refering to Chapter II, we see that obtaining error estimates requires a careful study of the properties of the operator $B_h = -\mathrm{div}_h$ and of its transpose that we denote by $\underline{\mathrm{grad}}_h$.

The first question is to characterize the kernel $\mathrm{Ker}\, B_h^t = \mathrm{Ker}(\underline{\mathrm{grad}}_h)$. It is clear from the definition of $b(\underline{v}_h, q_h)$ that $\mathrm{Ker}(\underline{\mathrm{grad}}_h)$ is *at least one-dimensional* and always contains constants. It may, however, be more than one-dimensional and we shall meet examples where this will occur. In these cases $\mathrm{Im}\, B_h = \mathrm{Im}\,(\mathrm{div}_h)$ will be *strictly smaller* than $P_{Q_h}(\mathrm{Im}\, B)$; this may lead to pathologies and may even imply trouble with the mere existence of the solution, as the following example shows.

Example 2.1: Let us consider problem (1.4) with *nonhomogeneous boundary conditions,* that is, with

$$\underline{u}|_\Gamma = \underline{r}, \qquad \int_\Gamma \underline{r} \cdot \underline{n}\, ds = 0. \tag{2.10}$$

It is classical to reduce this case to a problem with homogeneous boundary conditions by first introducing any function $\tilde{\underline{u}} \in (H^1(\Omega))^2$ such that $\tilde{\underline{u}}|_\Gamma = \underline{r}$. Setting $\underline{u} = \underline{u}_0 + \tilde{\underline{u}}$ with $\underline{u}_0 \in (H_0^1(\Omega))^2$ we then have to solve, with A defined by (1.3)

$$\begin{cases} -2\mu A \underline{u}_0 + \underline{\mathrm{grad}}\, p = \underline{f} + 2\mu A \tilde{\underline{u}} = \tilde{\underline{f}}, \\ \mathrm{div}\, \underline{u}_0 = -\mathrm{div}\, \tilde{\underline{u}} = g, \qquad \underline{u}_0|_\Gamma = 0. \end{cases} \tag{2.11}$$

We thus find a problem with a constraint $B\underline{u}_0 = g$ where $g \neq 0$. We have seen in Chapter II that the associated discrete problem may fail to have a solution, because $g_h = P_{Q_h} g$ does not necessarily belongs to $\mathrm{Im}\, B_h$, whenever $\mathrm{Ker}\, B_h^t \not\subset \mathrm{Ker}\, B^t$. Discretizations where $\mathrm{Ker}(\underline{\mathrm{grad}}_h)$ is more than one-dimensional *can therefore lead to ill-posed problems* in particular for some non-homogeneous boundary conditions. Examples of such conditions can be found in GRESHO–GRIFFITHS–LEE–SANI [A]. In general, any method that relies on extra compatibility conditions will sooner or later be a source of trouble when applied to more complex (nonlinear, time-dependent, etc.) problems. □

We have given in Chapter II, Proposition 2.2, a criterion ensuring that $\mathrm{Ker}\, B_h^t \subset \mathrm{Ker}\, B^t$. We have seen that this is equivalent to the existence of an operator Π_h from V to V_h satisfying

$$b(\underline{u} - \Pi_h \underline{u}, q_h) = \int_\Omega (\mathrm{div}\, \underline{u} - \mathrm{div}\, \Pi_h \underline{u})\, q_h\, dx = 0, \qquad \forall q_h \in Q_h. \tag{2.12}$$

We can also write this as $\operatorname{div}_h(\Pi_h \underline{u}) = P_{Q_h}(\operatorname{div} \underline{u})$. Building such an operator will also be the key to verify the inf–sup condition, which is in the present case,

$$\sup_{\underline{v}_h \in V_h} \frac{\int_\Omega q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_V} \geq k_h \|q_h\|_{L^2(\Omega)/\mathbb{R}}, \quad k_h \geq k_0 > 0, \tag{2.13}$$

where the constant k_0 is independent of h. Indeed, Proposition II.2.8 tells us that (2.13) will hold if the operator defined in (2.12) is continuous from V into V_h, that is, if one has

$$\|\Pi_h \underline{u}\|_V \leq c \, \|\underline{u}\|_V. \tag{2.14}$$

In most cases where Π_h can be explicitly built, (2.14) will also hold, thus proving the inf–sup condition and the error estimate

$$\begin{aligned} &\|\underline{u} - \underline{u}_h\|_V + \|p - p_h\|_{Q/\mathbb{R}} \\ &\qquad \leq c\Big\{ \inf_{\underline{v}_h \in V_h} \|\underline{u} - \underline{v}_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_{Q/\mathbb{R}} \Big\}. \end{aligned} \tag{2.15}$$

Usually, we shall use quite standard elements to approximate $\underline{u}$ and p and it will be quite classical to evaluate the right-hand side.

Remark 2.1: We shall also meet cases in which the constant k_h is not bounded from below by k_0. We shall then try to know precisely how it depends on h and to see whether convergence to a lower order can still be expected. When $\operatorname{Ker}(\underline{\operatorname{grad}}_h)$ is more than one dimensional, we are interested in a weaker form of (2.13),

$$\sup_{\underline{v}_h \in V_h} \frac{\int_\Omega q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_V} \geq k_h \inf_{q \in \operatorname{Ker}(\underline{\operatorname{grad}}_h)} \|q_h - q\|_{L^2(\Omega)}, \tag{2.16}$$

and in the dependence of k_h with respect to h. □

Several methods have also been proposed to get a more direct and intuitive evaluation of the quality of finite element approximations to divergence-free functions. One of them is the *constraint ratio*, which we shall denote C_r and define by

$$C_r = (\dim Q_h - 1)/\dim V_h. \tag{2.17}$$

It is, therefore, the ratio of the number of linearly independent constraints arising from the discrete divergence-free condition to the total number of degrees of freedom of the discrete velocity.

The value of C_r has no direct interpretation, *unless it is larger than* 1, which obviously means that, as the number of constraints exceeds that of variables,

Example 3.1: $P_1 - P_0$ *approximation*

This is probably the simplest element one can imagine for the approximation of an incompressible flow: one uses a standard P_1 approximation for velocity and a piecewise constant approximation for pressure. As the divergence of a P_1 velocity field is piecewise constant, this would lead to a divergence-free approximation. However, it is easy to see that such an element will not work for a general mesh. Indeed consider a triangulation of a (simply connected) domain Ω and let us denote

— t, the number of triangles,

— v_I, the number of internal vertices,

— v_B, the number of boundary vertices.

We shall thus have $2v_I$ degrees of freedom (d.o.f.) for space V_h (as the velocity must vanish on the boundary) and t d.o.f. for pressure leading to $(t-1)$ independent divergence-free constraints. By Euler's relations, we have

$$t = 2v_I + v_B - 2 \tag{3.1}$$

and thus

$$(t-1) \geq 2(v_I - 1). \tag{3.2}$$

A function $\underline{u}_h \in V_h$ is thus overconstrained and a *locking phenomenon* will occur: in general the only divergence-free discrete function is $\underline{u}_h \equiv 0$. When the mesh is built under certain restrictions, it is, however, possible that some linear constraints become dependent: this will be the case for the cross-grid macroelement (Figure VI.3) that will be analysed in section VI.5.2. □

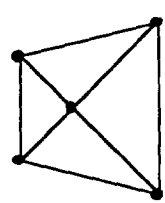


FIgure VI.3: The cross-grid element

Example 3.2: *Rectangular approximations.*

Let us consider a rectangular mesh with a Q_1 approximation of velocity. Using a P_1 pressure field would yield a divergence-free velocity field. It is, however, easy to see that this approximation is strongly overconstrained. In the same way it is not possible to build a divergence-free approximation from a Q_2 velocity field. □

Example 3.3: $P_2 - P_1$ *approximation.*

Coming back to triangles and using a P_2 velocity field, a divergence-free approximation is obtained from a P_1 (discontinuous) pressure field. This can be shown to work with a special cross-grid mesh (Section VI.3.3). On a general mesh there exist nontrivial divergence-free discrete velocity fields. They will, however, generate a poor approximation as *the smallest representable vortices will require a large patch of elements.* For instance, on the mesh of Figure VI.4, three divergence-free independent functions will remain from the original 74 degrees of freedom.

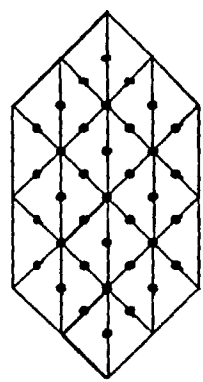


Figure VI.4

An analysis of such a method is, however, far from being straightforward, although the macroelement technique of section VI.5.3 could probably be applied. It results from this, the fact that genuinely divergence-free approximations cannot be built on a general mesh using low-degree elements. Indeed, it was shown in FORTIN [B] that building a general divergence-free triangular element requires fourth-degree polynomials. This was afterwards analyzed in detail by VOGELIUS [A] and SCOTT and VOGELIUS [A]. We shall present this precise result later.

An obvious way to circumvent this problem is to weaken the discrete divergence-free condition by using a smaller space Q_h. Thus, the discrete divergence operator $\mathrm{div}_h = P_{Q_h}(\mathrm{div})$ will generate weaker constraints provided $\mathrm{div}\, V_h \not\subset Q_h$. For instance, using a piecewise constant pressure field with a P_2 velocity field will be shown in Example 3.6 to yield a convergent and stable approximation.

Using continuous pressure fields is another way of weakening the discrete divergence-free condition even if this fact was not the initial reason for their introduction. Before coming back to the analysis of some stable discontinuous pressure methods, we shall consider some methods of this kind.

the only discrete divergence-free function is zero. We then have a *locking phenomenon.*

Conversely a small value of C_r implies a poor approximation of the divergence-free condition. It must, however, be emphasized that such a use of the constraint ratio has only a limited empirical value.

Another heuristic evaluation can be found by looking at the smallest representable vortex for a given mesh. This will be closely related to building a divergence-free basis (cf. Section VI.6). The idea behind (FORTIN [D]) is that a discrete divergence-free function can be expressed as a sum of small vortices that are, indeed, basis functions for $\operatorname{Ker} B_h$. The size of the smallest vortices can be thought of as the equivalent of the smallest representable wavelength in spectral methods.

In this context, we shall refer to a regular mesh of n^2 rectangles, n^3 cubes or $2n^2$ triangles (Figure VI.1).

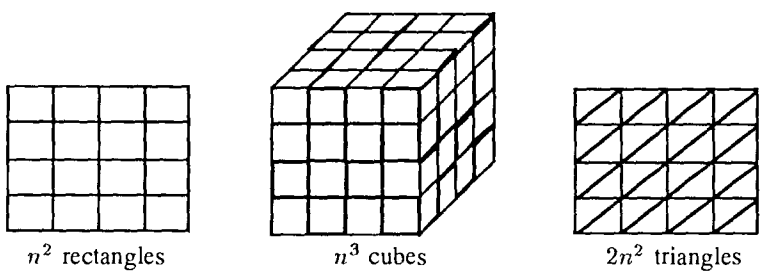


Figure VI.1

We must also quote the results of ZIENKIEWICZ–QU–TAYLOR–NAKAZAWA [A] who introduced a "patch test" to analyze similar problems. This patch test is only heuristic and does not yield a proof of stability. Although it can provide an intuitive indication, such a test may be misleading in several cases.

VI.3 Examples of Elements for Incompressible Materials

This section will provide a survey of elements for incompressible materials. This implies some classification of the known elements and such a classification necessarily contains a large part of arbitrariness. We shall base our classification on the techniques required for their analysis rather than on the elements themselves. A major distinction also appears between continuous pressure and discontinuous pressure elements. Variational formulation (1.1) contains no pressure derivative so that no continuity is required for the approximation of this variable. Users of finite element methods, however, feel better at ease with continuous approximations; this "inertial" reason, along with the desire to compute pressure at boundaries, has led people to approximate pressure by standard continuous elements.

Our main task will be to check the inf–sup condition. We shall first present the simplest examples both for continuous and discontinuous pressure elements. From this point we shall introduce some stabilization techniques enabling us to use some initially pathological elements. Finally, we consider special techniques, among them the use of macroelements, permitting one to prove the inf–sup condition.

We shall use the notations of Chapter III. In particular, $(P_k\text{-}P_\ell)$, $(Q_k\text{-}P_\ell)$ and $(Q_k\text{-}Q_\ell)$ will respectively mean elements in which velocities are approximated by polynomials of degree k and pressure by polynomials of degree ℓ or polynomials of degree k and ℓ in each variable. Special notations will be introduced for enriched versions of these elements and will be defined when needed.

VI.3.1 Simple examples

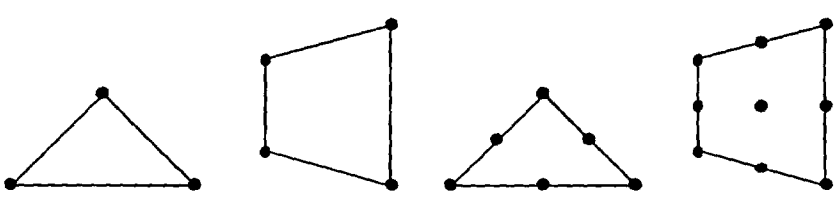
When attempting to define a finite element approximation for problem (1.1), it is quite natural to try using standard elements as developed for problems of plane elasticity or for thermal problems. In this respect the most natural approaches are probably

— use a standard element for velocity and try to impose the divergence-free condition everywhere,

— use the same standard element for both velocity and pressure ("Equal interpolation methods").

The first approach leads to discontinuous pressure approximations in which $\operatorname{div}(V_h) \subset Q_h$ and the second one to continuous pressure formulations. We shall first see that both these approaches need to be altered if correct results are to be obtained. We shall then present a few classical elements and show how they can be analyzed.

The most standard elements, represented schematically on Figure VI.2, are the P_1, P_2, Q_1, and Q_2 approximations.



a) P_1 b) P_2 c) Q_1 d) Q_2

Figure VI.2

We first deal with the first approach: building true divergence-free elements. We first give a few examples showing that it is not possible with low order elements.

Example 3.4: *Equal interpolation methods.*

To fix ideas let us consider a very simple case, that is, a P_1 continuous interpolation for both velocity and pressure. A simple count shows that if the number of triangles is large enough, there exist nontrivial functions satisfying the *discrete* divergence-free condition. Thus, no locking will occur and a solution can be computed. Users of such methods (for instance $(P_2 - P_2)$, $(Q_1 - Q_1)$, etc.) soon became aware that their results were strongly mesh dependent. In particular, pressure exhibited a very strange instability. This comes from the fact that for some meshes the kernel of the discrete gradient operator is more than one dimensional and contains nonconstant functions. This means that the solution obtained is determined only up to a given number of *spurious pressure modes,* (GRESHO–GRIFFITHS–LEE–SANI [A]) and that, at best, some filtering will have to be done before accurate results are available. We shall come back later to this phenomenon also named checkerboarding after the behavior of the $(Q_1 - P_0)$ approximation of Example 3.8. To fully understand the nature of spurious pressure modes, the reader may check the results of Figure VI.5 in which different symbols denote points where functions in $\mathrm{Ker}(\underline{\mathrm{grad}}_h)$ must have equal values for a $(P_1 - P_1)$ approximation.

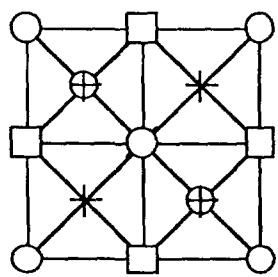


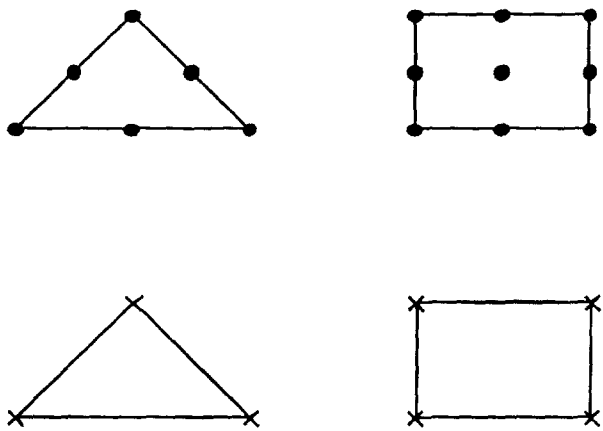
Figure VI.5

Apart from the constant pressure mode, we thus have in this case, *three* spurious pressure modes. *This also shows that there exists on this mesh one nontrivial discrete divergence-free function, whereas a direct count would predict locking.* □

This unpredictable behavior of equal interpolation methods led to the introduction of more sophisticated methods. In particular, HOOD–TAYLOR [A,B] experimentally discovered that using an approximation for pressure one degree lower than the approximation of velocity led to acceptable results. From this emerged the methods of the following example.

Example 3.5: *Taylor and Hood elements.*

From purely experimental considerations it was soon recognized that a $(P_2 - P_1)$ or a $(Q_2 - Q_1)$ approximation (Figure VI.6) yielded stable and convergent results.



$P_2 - P_1$ $\qquad$ $Q_2 - Q_1$

Figure VI.6

The analysis of these elements was first done by BERCOVIER–PIRONNEAU [A] and requires special techniques; it will be presented in Section VI.6. We shall see later how the above elements can be slightly modified to make their analysis simpler (and incidentally making them more accurate). □

Example 3.6: $(P_2 - P_0)$ *and Crouzeix–Raviart elements.*

Having realized that low-degree divergence-free elements could not be built (except on special meshes and with a few difficulties), an obvious idea is to weaken the constraint. Let us consider the case of a P_2 velocity field on triangles and a (piecewise constant) P_0 pressure field. The discrete divergence-free condition can then be written as

$$\int_K \operatorname{div} \underline{u}_h \, dx = \int_{\partial K} \underline{u}_h \cdot \underline{n} \, ds = 0, \qquad \forall K \in \mathcal{T}_h, \tag{3.3}$$

that is, as a conservation of mass on every element. This is intuitively an approximation of div $\underline{u}_h = 0$, directly related to the physical meaning of this condition. It is, however, clear from error estimate (2.15) and standard approximation results (cf. Chapter III) that such an approximation will lead to the loss

introducing the cure for this problem, that is, the Crouzeix–Raviart element, we give a first glance at the analysis of the $P_2 - P_0$ element as this will be the basis for most discontinuous pressure approximations. If one tries to check the inf–sup condition by building an operator Π_h satisfying (2.12) one is led, $\underline{u}$ being given, to build $\underline{u}_h = \Pi_h \underline{u}$ such that

$$\int_K \operatorname{div}(\underline{u} - \underline{u}_h)\, q_h \, dx = 0, \qquad \forall q_h \in Q_h \tag{3.4}$$

or, equivalently, as q_h is constant on every element $K \in \mathcal{T}_h$,

$$\int_K \operatorname{div}(\underline{u} - \underline{u}_h)\, dx = \int_{\partial K} (\underline{u} - \underline{u}_h) \cdot \underline{n}\, ds = 0. \tag{3.5}$$

This last condition could be satisfied if $\underline{u}_h$ were built in the following way. Let us denote by M_i and e_i, $i = 1, 2, 3$, the vertices and the sides of the triangular element K (Figure VI.7); mid-side nodes are denoted M_{ij}. We then define

$$\underline{u}_h(M_i) = \underline{u}(M_i), \quad i = 1, 2, 3 \tag{3.6}$$

$$\int_{e_i} \underline{u}_h \, ds = \int_{e_i} \underline{u} \, ds, \quad i = 1, 2, 3. \tag{3.7}$$

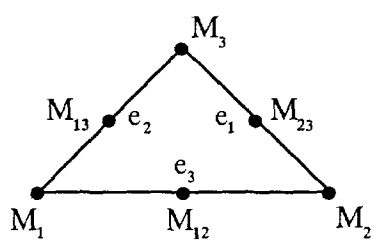


Figure VI.7

Condition (3.7) can be fulfilled by a correct choice of $u_h(M_{ij})$. Moreover, this construction can be done at element level as the choice of $u_h(M_{ij})$ is compatible on adjacent elements.

Although this is the basic idea, some technicalities must be introduced before a real construction is obtained. Indeed, for $\underline{u} \in (H_0^1(\Omega))^2$, condition (3.6) has no sense. We shall use in Section VI.3.2 the same technique as in Chapter IV to overcome this. We now turn our attention to the Crouzeix–Raviart element, which is an enrichment of the previous one. If a piecewise linear pressure is to be used, condition (3.5) must be changed to

The previous construction (provided it is correctly justified) enables us to choose $\underline{u}_h$ satisfying this for $p_1|_K =$ constant. Moreover,this choice of $\underline{u}_h$ depends only on the values of $\underline{u}$ on ∂K. The idea of Crouzeix and Raviart was to increase the number of d.o.f. of the element by adding *bubble functions* (that is, shape functions vanishing on ∂K). Thus we use as a velocity field, a finite element space such that

$$\underline{u}_h|_K = \underline{p}_2 + \underline{\alpha}_K \lambda_1 \lambda_2 \lambda_3, \ \underline{p}_2 \in (P_2(K))^2, \quad \underline{\alpha}_K \in \mathbb{R}^2, \tag{3.9}$$

where the λ_i are the barycentric coordinates of K. Writing now (3.8) in the form

$$\int_K (\underline{u} - \underline{u}_h) \cdot \underline{\mathrm{grad}}\, p_1 \, dx = \int_{\partial K} (\underline{u} - \underline{u}_h) \cdot \underline{n} \, p_1 \, ds, \tag{3.10}$$

it is possible to choose $\underline{\alpha}_K$ to satisfy this equation. It must be noted that $\underline{u}_h \cdot \underline{n}$, in the right-hand side of (3.10) is already specified by (3.6) and (3.7). It is now easy to see that error estimate (2.15) will be optimal, that is, $O(h^2)$, with respect to the elements used. As we shall see in the next sections, adding internal nodes to stabilize elements and permit the use of higher-degree pressure fields is a fundamental idea that will also prove useful for continuous pressure elements shown in the next example. □

Example 3.7: *Stable continuous pressure elements (MINI and related elements).*

Let us come back to the $(P_1 - P_1)$ element of Example 3.4. Following ARNOLD–BREZZI–FORTIN [A], we add, as we have already seen in Chapter IV, to the velocity field bubble functions

$$\underline{u}_h|_K = \underline{p}_1 + \underline{\alpha}_K \lambda_1 \lambda_2 \lambda_3, \qquad \forall K \in \mathcal{T}_h, \ \underline{p}_1 \in (P_1(K))^2, \ \underline{\alpha}_K \in \mathbb{R}^2. \tag{3.11}$$

As the pressure field is continuous, integrating (3.4) by parts reduces to

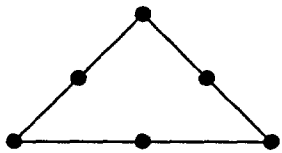
$$\int_\Omega \underline{u}_h \cdot \underline{\mathrm{grad}}\, p_1 \, dx = \int_\Omega \underline{u} \cdot \underline{\mathrm{grad}}\, p_1 \, dx, \qquad \forall p_1 \in \mathfrak{L}_1^1, \tag{3.12}$$

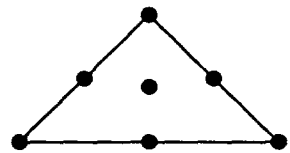
as boundary terms cancel. But (3.12) will, a fortiori, be satisfied if one has

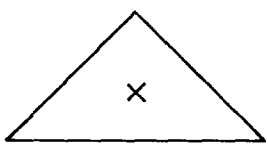
$$\int_K \underline{u}_h \, dx = \int_K \underline{u} \, dx, \qquad \forall K. \tag{3.13}$$

Indeed, $\underline{\mathrm{grad}}\, p_1$ is piecewise constant and (3.13) evidently implies (3.12). It is easily seen that a proper choice of $\underline{\alpha}_K$ makes (3.13) hold and thus adding a bubble function to a standard element again makes it stable. The same trick will yield an enriched version of the Taylor–Hood element: adding bubbles enables us to use the technique sketched in (3.12) and (3.13). □

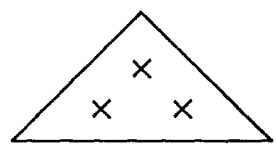
Remark 3.1: Let us denote by P_2^+ the space of P_2 polynomials enriched by bubbles. We present in Figure VI.8 a diagram of the elements introduced above.



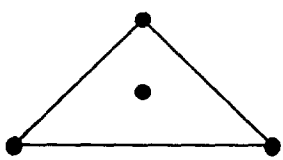


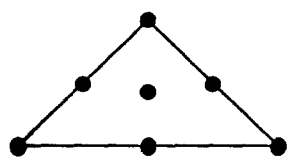


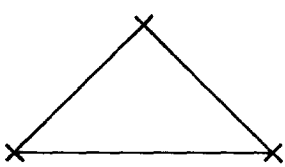
a) $P_2 - P_0$



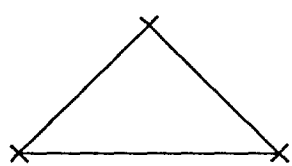
b) $P_2^+ - P_1$ **(Crouzeix–Raviart)**







c) **Mini**



d) $P_2^+ - P_1$ **continuous**

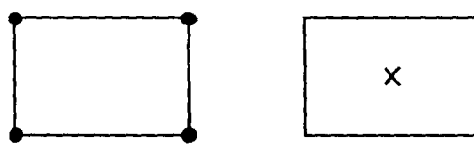
Figure VI.8

The continuous pressure $(P_2^+ - P_1)$ element can be considered as a sub-element of the Crouzeix–Raviart element. The inf–sup condition for this element is a direct consequence of the result for the Crouzeix–Raviart element. Both these

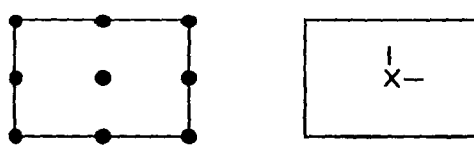
practical interest of such a fact is somewhat limited. □

Example 3.8: *The $Q_1 - P_0$ element.*

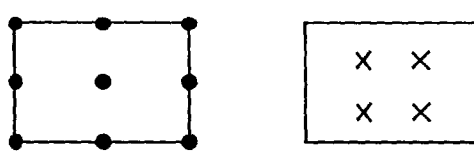
Our previous examples of discontinuous pressure elements were built on triangles. Quadrilaterals are also widely used and among quadrilateral elements, the $Q_1 - P_0$ element (Figure VI.9 a)) is the first that comes to mind. It uses a standard Q_1 velocity field and a piecewise constant pressure field. This element is strongly related, for rectangular meshes, to some finite-difference methods (FORTIN–PEYRET–TEMAM [A]). Its first appearance in a finite element context seems to be in HUGHES–ALLIK [A].



a) $Q_1 - P_0$ element



b) $Q_2 - P_1$ element



c) $Q_2 - Q_1$ element

● value of velocity

× value of pressure

×— values of pressure and its first derivatives

Figure VI.9

However simple it may look, the $Q_1 - P_0$ element is one of the hardest elements to analyze and many questions are still open about its properties. This element does not satisfy the inf–sup condition: it strongly depends on the mesh. For a regular mesh the kernel of the discrete gradient is two dimensional. More precisely, $\text{grad}_h\ p_h = 0$ implies that p_h is constant on the

c_1	c_2	c_1
c_2	c_1	c_2
c_1	c_2	c_1

Figure VI.10

This means that two singular values (cf. Chapter II) of the operator $B_h = \mathrm{div}_h$ are null. Moreover, it was verified by computation (MALKUS [A]) that a large number of nonzero singular values converge to zero when h becomes small. JOHNSON–PITKARANTA[A] indeed proved that the constant k_h is $O(h)$ and cannot be bounded from below (see also ODEN–JACQUOTTE [A]). The $Q_1 - P_0$ element has been the subject of a vast literature. We shall summarize some of the facts known about it in Section VI.5.4. □

Example 3.9: $(Q_2 - P_1)$ *and* $(Q_2 - Q_1)$ *elements*

The $Q_2 - P_1$ element is probably the most popular quadrilateral two-dimensional element at the present time. It was apparently discovered around a blackboard at the Banff Conference on Finite Elements in Flow Problems (1979). This element is sketched in Figure VI.9; it satisfies the inf–sup condition, the proof of this being essentially the same as for the Crouzeix–Raviart element, besides some technical details specific to quadrilaterals. This element is a relatively late comer in the field; the reason for this is that using P_1 pressure on a quadrilateral is not a standard procedure. It appeared as a cure for the instability of the $(Q_2 - Q_1)$ element (also sketched in Figure VI 9) which appears quite naturally in the use of reduced integration penalty methods (BERCOVIER [B]). This last element is essentially related to the $Q_1 - P_0$ element and suffers the same problem although to a lesser extent. Another cure can be obtained by adding internal nodes (FORTIN–FORTIN [A]). □

Example 3.10: *Non-conforming elements*

Another classical way (CROUZEIX–RAVIART [A]) of obtaining elements satisfying the inf–sup condition is to use a nonconforming velocity field. The simplest element of this kind is for sure the $P_1^{NC} - P_0$ element of Figure VI.11.

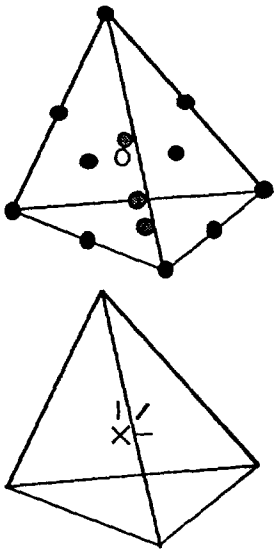
$P_1^{NC} - P_0$ element

Figure VI.11

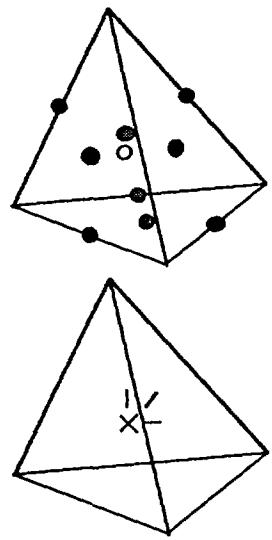
Mid-side nodes enable to control the flux through the side and to build an interpolation operator satisfying (2.12) or rather a nonconforming version of it. Other nonconforming methods of odd degrees can be found in CROUZEIX–RAVIART [A]. The second degree case has been treated in FORTIN–SOULIE [B]. It must also be said that coerciveness may be a problem for the $P_1^{NC} - P_0$ element as it does not satisfy the discrete version of Korn's inequality. □

Example 3.11: *Three-dimensional elements.*

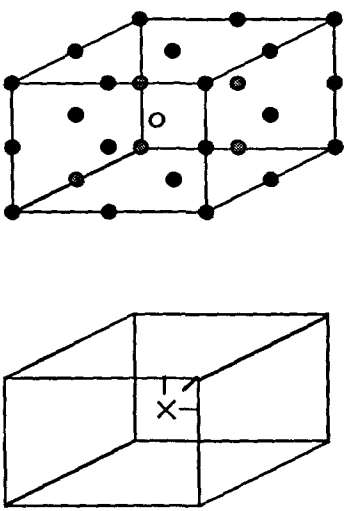
Many of the elements presented above have a three-dimensional counterpart. For instance, the Crouzeix–Raviart element of Example 3.6 can easily be generalized to the three-dimensional case: one needs bubble functions at the faces of a tetrahedron to enable control of the normal flow and internal bubbles to stabilize the nonconstant part of pressure (Figure VI.12a). This yields an element containing fourth-degree shape functions. A nonconforming counterpart has been built in FORTIN [E]. In this case shape functions remain in $P_2(K)$ and the degrees of freedom associated with vertices can be deleted without loosing accuracy. (Figure VI.12b). The $Q_2 - P_1$ (FIgure VI.12c) element has a direct extension and has been used numerically (BERCOVIER–ENGELMAN–SANI–GRESHO [A]). The Taylor–Hood $P_2 - P_1$ element has also been used (Figure VI.12d). Adding to it a bubble function as in Example 3.7 makes its stability obvious. Numerical evidence (BERTRAND–DHATT–FORTIN–OUELLET–SOULAIMANI [A]) indicates that this also improves accuracy and could, therefore, be worthwhile to use. The MINI element of Example 3.7 is also readily extended to tetrahedra (Figure VI.12e). It is probably the simplest stable three-dimensional element along with the *nonconforming* $P_1 - P_0$ element (Figure VI.12f) which has been studied in HECHT [A] for instance. The most popular of three-dimensional elements up to now is probably the $Q_1 - P_0$ element (Figure VI.12g). It suffers from the same problems as its two-dimensional counterpart. We shall analyze it in detail in Section VI.5. Development of three-dimensional elements is still an active area (FORTIN [E], RUAS [A]); implementations have been mostly restricted to first-order elements because of computer limitations but one can expect a shift to second-order elements in the next years.



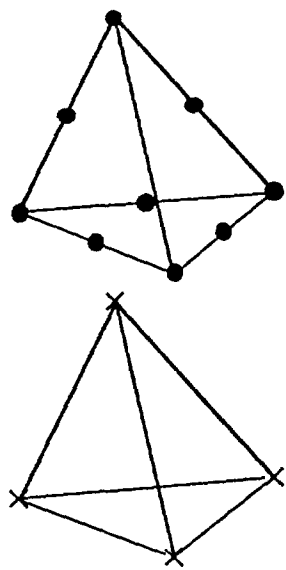
a) Crouzeix–Raviart element



b) Non-conforming P_2 –

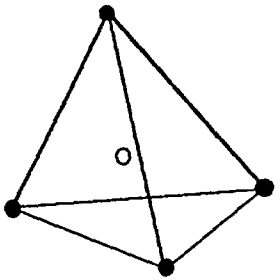


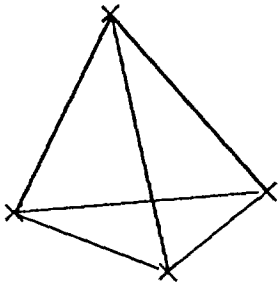
c) Q_2–P_1 element



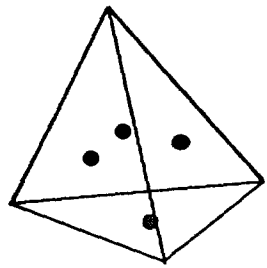
d) Taylor–Hood

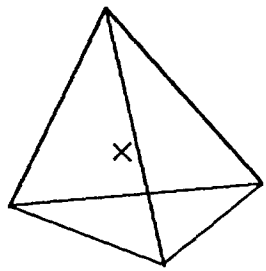
Figure VI.12



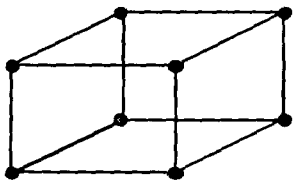


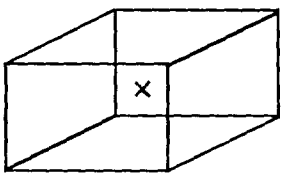
e) MINI element





f) Non-conforming P_1 –





g) $Q_1 - P_0$ element

● value of velocity

○ value of velocity at barycen

× value of pressure

values of pressure and its first derivatives

Figure VI.12: Three-dimensional elements

VI.4 Standard Techniques of Proof for the inf–sup Condition

We now consider in this section the proof of the inf–sup stability condition a large class of elements. In the present problem, the most general way to this is to build some interpolation operator Π_h satisfying

(4.1) $\int \operatorname{div}($

and

$$\|\Pi_h \underline{u}\|_V \le c\,\|\underline{u}\|_V, \tag{4.2}$$

where the V norm is the $(H_0^1(\Omega))^2$ norm. As it is shown in Chapter II, condition (4.1) is equivalent to $\mathrm{Ker}(\underline{\mathrm{grad}}_h) \subset \mathrm{Ker}(\underline{\mathrm{grad}})$. An element with this property will present no spurious pressure mode. We shall develop a technique to build Π_h for a fairly general class of elements. We shall follow, the approach of Proposition II.2.9. We have, therefore, to build two operators $\Pi_1 \in \mathcal{L}(V, V_h)$ and $\Pi_2 \in \mathcal{L}(V, V_h)$ satisfying

$$\|\Pi_1 \underline{v}\|_V \le c_1\,\|\underline{v}\|_V, \qquad \forall \underline{v} \in V, \tag{4.3}$$

$$\|\Pi_2(I - \Pi_1)\underline{v}\| \le c_2\,\|\underline{v}\|_V, \qquad \forall \underline{v} \in V, \tag{4.4}$$

$$\int_\Omega \mathrm{div}(\underline{v} - \Pi_2 \underline{v})\, q_h\, dx = 0, \qquad \forall \underline{v} \in V,\ \forall q_h \in Q_h, \tag{4.5}$$

where the constants c_1 and c_2 must be independent of h. Then the operator Π_h satisfying (4.1) and (4.2) will be found as

$$\Pi_h \underline{u} = \Pi_1 \underline{u} + \Pi_2(\underline{u} - \Pi_1 \underline{u}). \tag{4.6}$$

In many cases, Π_1 will be the interpolation operator of CLEMENT [A] (cf. Proposition III.2.1) in $H^1(\Omega)$; we then have

$$\sum_K h_K^{2r-2} |\underline{v} - \Pi_1 \underline{v}|_{r,K}^2 \le c\,\|\underline{v}\|_{1,\Omega}^2, \quad r = 0, 1. \tag{4.7}$$

Setting $r=1$ in (4.7), and the triangle inequality $\|\Pi_1 \underline{v}\|_V \le \|\underline{v} - \Pi_1 \underline{v}\|_V + \|\underline{v}\|_V$ yield (4.3).

On the contrary, the choice of Π_2 will vary from one case to the other, according to the choice of V_h and Q_h. However, as we have already seen in Chapter IV, the common feature of the various choices for Π_2 will be the following one: the operator Π_2 is constructed on each element K in order to satisfy (4.5). In many cases it will be such that

$$\|\Pi_2 \underline{v}\|_{1,K} \le c(h_K^{-1} \|\underline{v}\|_{0,K} + |\underline{v}|_{1,K}). \tag{4.8}$$

It is clear that (4.8) and (4.7) imply (4.4) since

$$\begin{aligned} \|\Pi_2(I - \Pi_1)\underline{v}\|_{1,\Omega}^2 &= \sum_K \|\Pi_2(I - \Pi_1)\underline{v}\|_{1,K}^2 \\ &\le c \sum_K \{h_K^{-2} \|(I - \Pi_1)\underline{v}\|_{0,K}^2 + |(I - \Pi_1)\underline{v}|_{1,K}^2\} \le c\|\underline{v}\|_{1,\Omega}^2. \end{aligned} \tag{4.9}$$

We can summarize this in the following proposition.

Proposition 4.1: Let V_h be such that a "Clement's operator": $\Pi_1 : V \to V_h$ exists and satisfies (4.7). If we can construct an operator $\Pi_2 : V \to V_h$ such that (4.5) and (4.8) hold, then the operator Π_h defined by (4.6) satisfies (4.1) and (4.2) and, therefore, the discrete inf–sup condition holds. □

We have already seen in Chapter IV some examples of the above procedure. Let us briefly recall the following example.

Example 4.1: *The $P_2 - P_0$ element.*

As we did in Chapter IV, we define, for every $K \in \mathcal{T}_h$ and every $\underline{v} \in (H_0^1(\Omega))^2$, $\Pi_2\underline{v}|_K$ by the conditions

$$(4.10)\qquad \begin{cases} \Pi_2\underline{v}|_K \in P_2(K), & \\ \Pi_2\underline{v}|_K(M) = 0, & \forall M = \text{vertex of } K, \\ \displaystyle\int_e \Pi_2\underline{v}\, ds = \int_e \underline{v}\, ds, & \forall e = \text{ edge of } K. \end{cases}$$

It is now elementary to check that Π_2, as defined in (4.10), satisfies (4.5) and (4.8). Actually (4.5) follows from

$$(4.11)\qquad \int_K \operatorname{div}(\Pi_2\underline{v} - \underline{v})\, dx = \int_{\partial K} (\Pi_2\underline{v} - \underline{v}) \cdot \underline{n}\, ds = 0$$

and (4.8) follows by a scaling argument (see Section III.2.4)

$$(4.12)\qquad |\Pi_2\underline{v}|_{1,K} = |\widehat{\Pi_2\underline{v}}|_{1,\hat{K}} < c(k,\theta_0)\|\hat{\underline{v}}\|_{1,\hat{K}} \le c(k,\theta_0)(h_K^{-1}|\underline{v}|_{0,K} + |\underline{v}|_{1,K}).$$

The above proof can easily be extended to more general cases. It can be applied to the $Q_2 - P_0$ quadrilateral element provided the usual regularity assumptions on quadrilateral meshes. A simple modification will hold for elements in which only the normal component of velocity is used as a degree of freedom at mid-side nodes (FORTIN M.[D], FORTIN, A. [A], BERNARDI–RAUGEL [A]). Indeed if only the normal component of $\underline{u}_h$ were used as a degree of freedom, the $P_2 - P_0$ element would become the element of Figure VI.13 in which, on each side, the normal component of $\underline{u}_h$ is quadratic, whereas the tangential is only linear. We now define $\Pi_2\underline{v}$ in $P_2(K)$ by asking $(\Pi_2\underline{v})(M_i) = 0$ $(i = 1, 2, 3)$ and

(4.13)

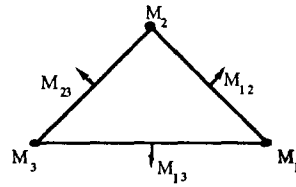


Figure VI.13

The above proof applies directly. The same argument can be done for the $Q_2 - P_0$ quadrilateral element.

For three-dimensional elements, one needs to correct some mid-face node instead of a mid-side node in order to control the normal flow on the face. From this point, the idea of the proof is the same. Finally, we shall present examples in the next section in which a patch of elements (a macroelement) will have to be used. A correction of $\underline{u}_h$ at some cleverly chosen node on the boundary of the patch will ensure that the inf–sup condition will hold for piecewise constant pressure. □

In a certain number of situations, the operator Π_h that we have constructed here for the $P_2 - P_0$ element will be used, when applying Proposition II.2.9, as an operator Π_1. We, therefore, change its name: *we denote by* $\tilde{\Pi}_1$ *the operator* Π_h *constructed for the* $P_2 - P_0$ *element* and we recall its properties

$$
\begin{aligned}
&\|\tilde{\Pi}_1 \underline{v}\|_1 \le c\,\|\underline{v}\|_1, \qquad \forall \underline{v} \in V, \\
&\int_K \operatorname{div}(\underline{v} - \tilde{\Pi}_1 \underline{v})\, dx = 0, \qquad \forall \underline{v} \in V,\ \forall K \in \mathcal{T}_h.
\end{aligned}
\tag{4.14}
$$

As an example in which this choice of Π_1 is convenient, let us consider the following example.

Example 4.2: *The Crouzeix–Raviart element.*

We already said that we shall choose $\Pi_1 = \tilde{\Pi}_1$ in applying Proposition II.2.9 . We now choose Π_2. Since V_h is locally made by P_2 polynomials plus bubble functions, we choose $\Pi_2 \quad V \to (B_3)^2$. That means $\Pi_2 \underline{v}$ in each K will be a pair of P_3-bubble functions (with coefficients to be chosen).

Actually we shall define $\Pi_2 \underline{v}$ only in the case when $\operatorname{div} \underline{v}$ has zero mean value in each K. This will be sufficient since we shall use in practice $\Pi_2(I - \tilde{\Pi}_1)$ and $\tilde{\Pi}_1$ satisfies the second equation of (4.14). For all K and for all $\underline{v}$ with

$$
\int_K \operatorname{div} \underline{v}\, dx = 0, \tag{4.15}
$$

we then set $\Pi_2 \underline{v}$ as the unique solution of

$$\int_K \operatorname{div}(\Pi_2 \underline{v} - \underline{v})\, q_h \, dx = 0, \qquad \forall q_h \in P_1(K). \tag{4.17}$$

Note that (4.16) and (4.17) is a linear system of three equations ($\dim P_1(K) = 3$) in two unknowns ($\dim(B_3(K))^2 = 2$) that is compatible since $\underline{v}$ is assumed to satisfy (4.15) and, on the other hand, for every $\underline{b} \in (B_3(K))^2$ we clearly have

$$\int_K \operatorname{div} \underline{b} \, dx = 0$$

We have only to prove that

$$\|\Pi_2 \underline{v}\|_{1,K} \leq c \, \|\underline{v}\|_{1,K} \tag{4.18}$$

for all $\underline{v} \in V$ satisfying (4.15). Indeed (4.17) can be rewritten as

$$\int_K (\Pi_2 \underline{v}).\underline{\operatorname{grad}}\, q_h \, dx = \int_K \operatorname{div} \underline{v} \, (q_h - \bar{q}_h) \, dx, \tag{4.19}$$

where $\bar{q}_h$ is any piecewise constant approximation of q_h. A scaling argument as in Section III.2.4 yields

$$|\widehat{\Pi_2 \underline{v}}|_{0,\hat{K}} \leq |\hat{\underline{v}}|_{1,\hat{K}} \, c(\theta_0) \tag{4.20}$$

which easily implies (4.18). □

The above proof applies quite directly, with minor changes due to the technicalities special to quadrilaterals, to the $Q_2 - P_1$ element. It can also be used to create nonstandard elements. For instance, in FORTIN–FORTIN [A], bubble functions were added to a $Q_1 - P_0$ element in order to use a P_1 pressure field. This element is not more, but not less, stable than the standard $Q_1 - P_0$ but gives better results in some cases.

Before moving to general results, we recall the proof of the stability of the MINI element (Example 3.7) of ARNOLD–BREZZI–FORTIN [A] already presented in Section IV.2.

Example 4.3: *Stability proof for the MINI element.*

Here we use, again, the operator Π_1 of Clement as in Proposition 4.1, and we have now to construct Π_2. Since in each element we have $\underline{v}_h \in (P_1(K))^2 + (B_3(K))^2$, we can again build $\Pi_2 \,.\, V \to B_3$ and define it, on each K, as the solution of

$$\begin{aligned} &\Pi_2 \underline{v}|_K \in (B_3(K))^2, \\ &\int_K (\Pi_2 \underline{v} - \underline{v}) \cdot \underline{\operatorname{grad}} \, q_h \, dx = 0 \qquad \forall q_h \in P_1(K) \end{aligned} \tag{4.21}$$

It is clear that (4.21) has a unique solution and that (4.8) will hold by a simple scaling argument On the other hand, taking advantage of the fact that now $Q_h \subset H^1(\Omega)$, we have

$$(4.22) \quad \int_\Omega \operatorname{div}(\Pi_2 \underline{v} - \underline{v})\, q_h \, dx = -\int_\Omega (\Pi_2 \underline{v} - \underline{v}).\underline{\operatorname{grad}}\, q_h \, dx$$
$$= \sum_K \int_K (\Pi_2 \underline{v} - \underline{v}) \;\; \underline{\operatorname{grad}}\, q_h \, dx = 0,$$

so that (4.5) is also satisfied and we can apply Proposition 4.1 to obtain the inf–sup condition.

VI.4.1 General results

This subsection will present a general framework containing the previous examples and providing a general tool for the analysis of finite element approximations to incompressible materials problems. This technique will be further extended in Section VI.5 to the case of composite elements, but for the sake of comprehension it is worth considering first the simpler case.

The basic idea has been used several times on particular cases, starting from CROUZEIX–RAVIART [A] for discontinuous pressures and from ARNOLD–BREZZI–DOUGLAS [A], and ARNOLD–BREZZI–FORTIN [A] for continuous pressures. We are going to present it in its final general form given by BREZZI–PITKARANTA [A]. It consists essentially in stabilizing an element by adding suitable bubble functions to the velocity field.

In order to do that, we first associate to every finite element discretization $Q_h \subset L^2(\Omega)$ the space

$$(4.23) \quad B(\underline{\operatorname{grad}}\, Q_h) = \{\underline{\beta} \in (H_0^1(\Omega))^2, \underline{\beta}|_K$$
$$= b_{3,K}\, \underline{\operatorname{grad}}\, q_h|_K \quad \text{for some } q_h \in Q_h\}.$$

In other words, the restriction of a $\underline{\beta} \in B(\underline{\operatorname{grad}} Q_h)$ to an element K is the product of the P_3 bubble functions $b_{3,K}$ times the gradient of a function of $Q_h|_K$.

Remark 4.1: Notice that the space $B(\underline{\operatorname{grad}}\, Q_h)$ is not defined through a basic space $\hat{B}$ on the reference element. This can be easily done, if one wants to, in the case of *affine* elements for all the reasonable choices of Q_h. However, this is clearly *unnecessary*: if we know how to compute q_h on K we also know how to compute $\underline{\operatorname{grad}}\, q_h$ and there is no need for a reference element. □

We can now prove our basic results concerning the two cases of continuous or discontinuous pressures.

Proposition 4.2: (*Stability of continuous pressure elements*). Let the following assumptions hold:

$$\text{there exists } \Pi_1 \in \mathcal{L}(V, V_h) \text{ satisfying (4.7)} \tag{4.24}$$

$$Q_h \subset H^1(\Omega), \tag{4.25}$$

$$V_h \supset B(\underline{\text{grad}}\, Q_h) \quad \text{(defined as in (4.23))}. \tag{4.26}$$

Then the pair (V_h, Q_h) is a stable element, in the sense that it satisfies the inf–sup condition.

Proof: We shall use Proposition 4.1. We already have our operator Π_1 by assumption (4.24). We only need to construct Π_2. We define $\Pi_2 \cdot V \to B(\underline{\text{grad}}\, Q_h)$ on each element by requiring

$$\begin{cases} \Pi_2 \underline{v}|_K \in B(\underline{\text{grad}}\, Q_h)|_K = b_{3,K}\, \underline{\text{grad}}\, Q_h|_K, \\ \displaystyle\int_K (\Pi_2 \underline{v} - \underline{v}) \cdot \underline{\text{grad}}\, q_h \, dx = 0, \qquad \forall q_h \in Q_h|_K. \end{cases} \tag{4.27}$$

Problem (4.27) has obviously a unique solution. As in (4.22), it is clear that Π_2 satisfies (4.5). Finally (4.8) follows by a scaling argument. Hence, Proposition 4.1 gives us the desired result. □

Corollary 4.1: Assume that $Q_h \subset Q$ is any space of continuous piecewise smooth functions. If $V_h \supset (\mathcal{L}_1^1)^2 \oplus B(\underline{\text{grad}}\, Q_h)$, then the pair (V_h, Q_h) satisfies the inf–sup condition.

Proof: Continuity and piecewise smoothness imply (4.25). The condition $(\mathcal{L}_1^1)^2 \subset V_h$ implies (4.24), and condition $B(\underline{\text{grad}}\, Q_h) \subset V_h$ is (4.25). Hence we can apply Proposition 4.2. □

The above results apply, for instance, to the enriched Taylor–Hood element and, if one wants, to the families (ARNOLD–BREZZI–FORTIN [A])

$$\begin{aligned} V_h &= (\mathcal{L}_k^1 \oplus B_{k+1})^2, \quad Q_h = \mathcal{L}_{k-1}^1, \\ V_h &= (\mathcal{L}_k^1 \oplus B_{k+2})^2, \quad Q_h = \mathcal{L}_k^1 \end{aligned} \tag{4.28}$$

We turn now to the case of discontinuous pressure elements

Proposition 4.3: (*Stability of discontinuous pressure elements*). Let the following assumptions hold:

$$\text{there exists } \tilde{\Pi}_1 \in \mathfrak{L}(V, V_h) \text{ satisfying (4.14)}, \tag{4.29}$$

$$V_h \supset B(\underline{\text{grad}}\, Q_h), \text{ (defined in (4.23))}. \tag{4.30}$$

Then the pair (V_h, Q_h) is a stable element in the sense that it satisfies the inf–sup condition.

Proof: We are going to proceed as in Example 4.2, that is, by essentially applying Proposition II.2.9. We take $\tilde{\Pi}_1$ (given by (4.29)) as operator Π_1. As in Example 4.2, we are not going to define Π_2 on all V, but only in the subspace

$$V^0 = \{\underline{v} \mid \underline{v} \in V, \int_K \text{div}\, \underline{v}\, dx = 0, \ \forall K \in \mathcal{T}_h\}. \tag{4.31}$$

For every $\underline{v} \in V^0$ we construct $\Pi_2 \underline{v} \in B(\underline{\text{grad}}\, Q_h)$ by requiring that, in each element K,

$$\begin{aligned} &\Pi_2 \underline{v}|_K \in B(\underline{\text{grad}}\, Q_h)|_K = b_{3,K}\, \underline{\text{grad}}\, Q_h|_K, \\ &\int_K \text{div}(\Pi_2 \underline{v} - \underline{v})\, q_h\, dx = 0, \qquad \forall q_h \in Q_h|_K. \end{aligned} \tag{4.32}$$

Note that (4.32) is uniquely solvable if $\underline{v} \in V^0$, since the divergence of a bubble function has always zero mean value (hence, the number of nontrivial equations is equal to $\dim(Q_h|_K) - 1$, which is equal to the number of unknowns; the nonsingularity then follows easily). It is obvious that Π_2, as given by (4.32), will satisfy (4.5) for all $\underline{v} \in V^0$. We have to check that

$$\|\Pi_2 \underline{v}\|_1 \le c\, \|\underline{v}\|_1, \tag{4.33}$$

which actually follows by the same scaling argument as in (4.19) and(4.20). It is then easy to see that the operator

$$\Pi_h = \tilde{\Pi}_1 - \Pi_2(I - \tilde{\Pi}_1) \tag{4.34}$$

satisfies (4.1) and (4.2). Hence, the inf–sup condition follows from Proposition II.2.8. □

Corollary 4.2: (Bidimensional case). Assume that $Q_h \subset Q$ is any space of piecewise smooth functions. If $V_h \supset (\mathfrak{L}_2^1)^2 \oplus B(\underline{\text{grad}}\, Q_h)$, then the pair (V_h, Q_h) satisfies the inf–sup condition.

Proof: Condition $(\mathcal{L}_2^1)^2 \subset V_h$ implies (4.29) as we have seen in Example 4.1. On the other hand, the condition $B(\underline{\mathrm{grad}}\, Q_h) \subset V_h$ is (4.30), so that we can apply Proposition 4.3. □

Proposition 4.1, 4.2, and 4.3 are worth a few comments. They show that almost any element can be stabilized by using bubble functions. For a continuous pressure element this procedure is mainly useful in the case of triangular elements. For discontinuous pressure elements it is possible to fully stabilize elements which are already partially stable,that is, stable for a piecewise constant pressure field. Examples of such a procedure can be found in FORTIN–FORTIN [A]. Stability with respect to piecewise constant pressure implies that at least one degree of freedom on each side or face of the element is linked to the normal component of velocity (FORTIN [D]).

VI.4.2 Higher-order methods

In this subsection we shall recall the statement of a basic result by SCOTT–VOGELIUS [A] which, roughly speaking, says: under minor assumptions on the decomposition $\mathcal{T}_h$ (in triangles) the pair $V_h = (\mathcal{L}_k^1)^2$, $Q_h = \mathcal{L}_{k-1}^0$ satisfies the inf–sup condition for $k \geq 4$. This, in a sense, settles the matter as far as higher-order methods are concerned, and leaves only the problem of finding stable lower-order approximations.

In order to state in a precise way the restrictions that have to be made on the triangulation, we assume first that Ω is a polygon, and that its boundary $\partial\Omega$ has no double points. In other words there exists two continuous piecewise linear maps $x(t)$, $y(t)$ from $[0,1[$ into $\mathbb{R}$ such that

$$(x(t_1) = x(t_2) \text{ and } y(t_1) = y(t_2)) \text{ implies } t_1 = t_2,$$
$$\partial\Omega = \{(x,y) \mid x = x(t),\ y = y(t) \text{ for some } t \in [0,1[\}.$$

Clearly we will have $\lim_{t\to 1} x(t) = x(0)$ and $\lim_{t\to 1} y(t) = y(0)$. Note that we already restricted ourselves to a less general case than the one treated by SCOTT–VOGELIUS [A]. We shall make further restrictions in what follows, so that we are actually going to present a particular case of their results.

Let now V be a vertex of a triangulation $\mathcal{T}_h$ of Ω and let $\theta_1, \ldots, \theta_n$, be the angles, at V, of all the triangles meeting at V, ordered, for instance, in the counterclockwise sense. If V is an internal vertex, we also set $\theta_{n+1} := \theta_1$. Now we define $S(V)$ by

$$\text{if } n = 1, \text{ then } S(V) = 0, \tag{4.35}$$

$$\text{if } n > 1 \text{ and } V \in \partial\Omega, \text{ then } S(V) = \max_{i=1,n-1} (\pi - \theta_i - \theta_{i+1}), \tag{4.36}$$

(4.37) if $V \notin \partial\Omega$, then $S(V) =$ [illegible]

It is easy to check that $S(V) = 0$ if and only if all the edges of $\mathcal{T}_h$ meeting at V fall on two straight lines. In this case V is said to be singular (SCOTT–VOGELIUS [A]). If $S(V)$ is positive but very small, then V will be "almost singular". Thus, $S(V)$ measures how close V is to be singular.

We are now able to state the following result.

Proposition 4.4: (SCOTT–VOGELIUS [A]). Assume that there exists two positive constants c and δ such that

$$ch \leq h_K, \qquad \forall K \in \mathcal{T}_h,$$

and

$$S(V) \geq \delta, \qquad \forall V \text{ vertex of } \mathcal{T}_h. \tag{4.38}$$

Then the choice $V_h = (\mathfrak{L}_k^1)^2$, $Q_h = \mathfrak{L}_{k-1}^0$, $k \leq 4$, satisfies the inf–sup condition with a constant depending on c and δ but not on h. □

Condition (4.38) is worth a few comments. The trouble is that $S(V) = 0$ makes the linear constraints on $\underline{u}_h$, arising from the divergence-free condition, linearly dependent. We shall meet other instances of this case in Examples 5.1 and 5.2. When this linear dependence appears, some part of the pressure becomes unstable. In the present case, this unstable part could be filtered out so that condition (4.38) could in reality be avoided. But this would require the methods developed in the next section.

VI.5 Macroelement Techniques and Spurious Pressure Modes

This section will introduce two general techniques for the analysis of finite element approximations to the Stokes problem. We shall first, after some remarks about spurious pressure modes, consider an abstract convergence result which generalizes some results of Chapter II. On the other hand, introducing the concept of macroelement enables us to extend the techniques of Section VI.4 to a new class of approximations. Finally, the joint use of these methods will enable us to make a partial analysis of the $Q_1 - P_0$ element and some other elements suffering from global spurious pressure modes.

VI.5.1 Some remarks about spurious pressure modes

We already introduced, in Section VI.3, the concept of spurious pressure mode, the classical example being the checkerboard mode of the $Q_1 - P_0$ element. The underlying problem is essentially algebraic in nature; in the framework of Chapter II we shall say that spurious pressure modes occur whenever,

that is, when the discrete gradient operator vanishes on nonconstant functions. From the results of Chapter II, It is clear that, in such a situation, the standard inf–sup condition (where quotient norms with respect to Ker B^t are used), cannot hold. There arises the question of whether a weaker form could be obtained; such a weaker condition would explain the success of some numerical computations using such pathological elements.

We already discussed in Example 3.8 the checkerboard pressure mode associated with the $Q_1 - P_0$ element and in Example 3.4 some modes associated with equal interpolation elements. We shall now present a few more examples and distinguish between local and global spurious pressure modes. This will lead us to the concept of macroelement or composite element that will be useful in Section VI.5.3.

Example 5.1: *Cross-grid $P_1 - P_0$ element.*

Let us consider a mesh of quadrilaterals divided into four triangles by their diagonals (Figure VI.14). It is well known that using a piecewise linear approximation for velocity and piecewise constant pressure leads to locking, that is, to a null velocity field. On the mesh introduced above, it is easy to see, however, that nonzero divergence-free functions can be obtained. The divergence is constant on each triangle. This means four linear relations between the values of the partial derivatives. It is easily seen that one of them can be expressed as a combination of the others, this fact being caused by equality of tangential derivatives along the straight-sided diagonal. To make things simple we consider the case where the diagonals are orthogonal (Figure VI.15) and we label by A, B, C, D the subtriangles. We then have by taking locally the coordinates axes along the diagonals and denoting u^K the approximation on element K

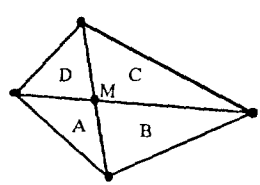


Figure VI.14

(5.2) $\partial u_1^K \quad \partial u_2^K$

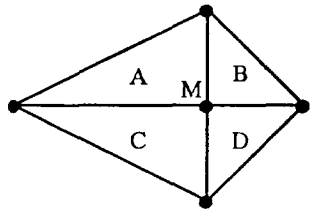


Figure VI.15

On the other hand, one has at point M,

$$\frac{\partial u_2^A}{\partial x_2} = \frac{\partial u_2^B}{\partial x_2}, \quad \frac{\partial u_1^A}{\partial x_1} = \frac{\partial u_1^C}{\partial x_1}, \quad \frac{\partial u_2^C}{\partial x_2} = \frac{\partial u_2^D}{\partial x_2}, \quad \frac{\partial u_1^B}{\partial x_1} = \frac{\partial u_1^D}{\partial x_1}, \tag{5.3}$$

It is easy to check that this makes one of the four conditions (5.2) redundant. The reader may check the general case by writing the divergence operator in a nonorthogonal coordinate system.

The consequence of the above discussion is that on each composite quadrilateral one of the four constant pressure values will be undetermined. The dimension of $\operatorname{Ker} B_h^t$ will be *at least as large as the number of quadrilaterals.*

Thus, three constraints remain on each composite quadrilateral element. If we admit that two of them can be accounted for, using the methods of Section VI.4, by the "internal" node M, we obtain an element that is very similar to the $Q_1 - P_0$ element with respect to degrees of freedom. Indeed, it can be checked that, on a regular mesh, an additional checkerboard mode occurs and that the behavior of this approximation is essentially the same as that of the $Q_1 - P_0$ element that will be discussed in details in Section VI.5.3. □

The above example clearly shows the existence of two kinds of *spurious pressure modes.* Let us consider an element where $\operatorname{Ker} B_h^t \supset \operatorname{Ker} B^t$ and thus where $\dim \operatorname{Ker} B_h^t > \dim \operatorname{Ker} B^t$.

In the first kind of spurious pressure mode $\dim \operatorname{Ker} B_h^t$ grows when $h \to 0$ and there exists a basis of $\operatorname{Ker} B_h^t$ with local support (that is, the support of each basis function can be restricted to a macroelement). Such pressure modes can be eliminated by considering a composite mesh (in the example a mesh of quadrilaterals instead of triangles) and using a smaller space for pressure by deleting some degrees of freedom from the composite elements. We shall then speak of *local pressure modes.*

In the second kind, the dimension of $\operatorname{Ker} B_h^t$ does not grow when $h \to 0$ and no basis can be found with a local support. We then have a *global pressure mode* which cannot be eliminated as easily as the local ones. Global modes usually appear on special (regular) meshes and are symptoms that the behavior

Some elements may generate both local and global modes as we have seen in the above example.

It must be emphasized that local spurious modes are source of trouble only when one prefers to work directly on the original mesh and not on the composite mesh on which they could easily be filtered out by a simple projection on each macroelement. We shall prove this in Section VI.5.2, where a more precise framework will be given.

Example 5.2: *Cross-grid $P_2 - P_1$ element.*

Another simple example where a local mode occurs is the straightforward extension of the previous example to the case of a $P_2 - P_1$ approximation (Figure VI.16). This yields, on each quadrilateral, 12 discrete divergence-free constraints, and it is easily seen by the argument of Example 5.1, written at point M, that one of them is redundant. Thus, one spurious mode will appear for each composite quadrilateral. However, in this case, no global mode will appear and we shall be able to analyze this element by the macroelement technique of section 5.2. The analysis of this element is also related to the work of CIAVALDINI–NEDELEC [A] by considering the stream function associated with a divergence-free function.

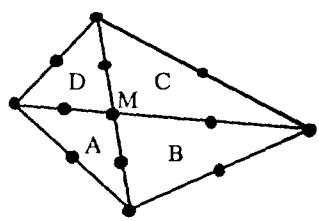


Figure VI.16

The presence of spurious modes can be interpreted as a signal that the pressure field used is in some sense too rich. We, therefore, can hope to find a cure by using a strict subspace $\hat{Q}_h$ of Q_h as the space of discrete pressure to obtain a stable approximation. The question arises whether or not this stability can be used to prove at least a partial result on the original approximation. One can effectively get some result in this direction. We now introduce an abstract setting that will be the key to many results of the next sections.

VI.5.2 An abstract convergence result

In this section, we shall work in the abstract framework of Chapter II. Even if known applications of the results come from Stokes problem, their proof

problem

$$a(u_h, v_h) + b(v_h, p_h) = (f, v_h), \qquad \forall v_h \in V_h,\ u_h \in V_h, \tag{5.4}$$

$$b(u_h, q_h) = (g, q_h), \qquad \forall q_h \in Q_h,\ p_h \in Q_h, \tag{5.5}$$

for which we suppose a solution to exist. This problem is, of course, an approximation of the corresponding infinite-dimensional problem posed in $V \times Q$ and we suppose $V_h \subset V$ and $Q_h \subset Q$. We now consider cases where $\operatorname{Ker} B_h^t \not\subset \operatorname{Ker} B^t$ and where the inf–sup constant k_h may depend on h. We also suppose that $a(\cdot,\cdot)$ is symmetric and V-elliptic (cf.(II.1.38)).

We now suppose that there exists subspaces $\hat{V}_h \subset V_h$ and $\hat{Q}_h \subset Q_h$ defining a stable approximation of the problem. We, thus, have

$$\sup_{\hat{v}_h \in \hat{V}} \frac{b(\hat{v}_h, \hat{q}_h)}{\|\hat{v}_h\|_V} \geq k_0 \, \|\hat{q}_h\|_{Q/\operatorname{Ker} B^t}, \quad \forall \hat{q}_h \in \hat{Q}_h. \tag{5.6}$$

We recall some notation from Chapter II: then let

$$Z(g) = \{v \mid b(v, q) = (g, q),\ \forall q \in Q\} \tag{5.7}$$

and

$$Z_h(g) = \{v_h \mid b(v_h, q_h) = (g, q_h),\ \forall q_h \in Q_h\}, \tag{5.8}$$

and let us introduce

$$\hat{Z}_h(g) = \{\hat{w}_h \in \hat{V}_h \mid b(\hat{w}_h, \hat{q}_h) = (g, \hat{q}_h),\ \forall \hat{q}_h \in \hat{Q}_h\}. \tag{5.9}$$

Using (5.6) and Proposition II.2.5, we have

$$\inf_{\hat{w}_h \in \hat{Z}_h(g)} \|u - \hat{w}_h\|_V \leq c \inf_{\hat{v}_h \in \hat{V}_h} \|u - \hat{v}_h\|_V. \tag{5.10}$$

In the context of finite element approximation, we shall usually require that the quantities $\inf_{\hat{w}_h \in \hat{V}_h} \|u - \hat{w}_h\|_V$ and $\inf_{w_h \in V_h} \|u - w_h\|_V$ can be estimated to the same order of accuracy.

Let us denote $\tilde{Q}_h$ the orthogonal complement of $\hat{Q}_h$ in Q_h. We shall make the following strong hypothesis (which is nevertheless satisfied in many cases):

$$b(\hat{v}_h, \tilde{q}_h) = 0, \qquad \forall \tilde{q}_h \in \tilde{Q}_h, \qquad \forall \hat{v}_h \in \hat{V}_h. \tag{5.11}$$

Condition (5.11) implies, in particular, that for $\hat{v}_h \in \operatorname{Ker} \hat{B}_h = \hat{Z}_h(0)$, one has

(5.12) $\qquad b(\hat{v}_h, \tilde{q}_h) = 0, \quad \forall \tilde{q}_h \in \tilde{Q}_h$

that is

$$\text{Ker}\,\hat{B}_h \subset \text{Ker}\,B_h. \tag{5.13}$$

We would also like to have

$$\hat{Z}_h(g) \subset Z_h(g). \tag{5.14}$$

This will hold by (5.11) provided g satisfies the condition

$$(g, \tilde{q}_h) = 0, \qquad \forall \tilde{q}_h \in \tilde{Q}_h. \tag{5.15}$$

This last condition will not be difficult to check in practical cases. It will, of course, hold in the important case where $g = 0$.

We can now consider our convergence result. Let (u, p) be the solution of the continuous problem; we can easily get from Proposition II.2.4

$$\|u - u_h\|_V \le c\Big(\inf_{w_h \in Z_h(g)} \|u - w_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q\Big). \tag{5.16}$$

From (5.14) we have

$$\inf_{w_h \in Z_h(g)} \|u - w_h\|_V \le \inf_{\hat{w}_h \in \hat{Z}_h(g)} \|u - \hat{w}_h\|_V. \tag{5.17}$$

Using (5.10) we have

$$\inf_{\hat{w}_h \in \hat{Z}_h(g)} \|u - \hat{w}_h\|_V \le c \inf_{\hat{v}_h \in \hat{V}_h} \|u - \hat{v}_h\|_V, \tag{5.18}$$

and we finally obtain

$$\|u - u_h\|_V \le c\Big(\inf_{\hat{v}_h \in \hat{V}_h} \|u - \hat{v}_h\|_V + \inf_{q_h \in Q_h} \|p - q_h\|_Q\Big) \tag{5.19}$$

and the problem is now a standard approximation problem in $\hat{V}_h$.

Remark 5.1: As we shall see when considering examples, it will be possible when dealing with *local pressure modes* to take $\hat{V}_h = V_h$. Estimate (5.19) then shows that local modes cause no loss in accuracy. *This can, in fact, be used as a precise definition of local modes.* □

We shall now try to get an estimate for pressure. As can be expected, we shall have to consider convergence for the component (say, $\hat{p}_h$) of p_h in $\hat{Q}_h$ and not for p_h itself. This corresponds to the practice of filtering out spurious pressure modes. We substract the continuous and discrete equation and write

$$(5.20)\quad a(u-u_h,v_h)+b(v_h,p-\hat{q}_h) + b(\hat{q}_h-p_h,v_h)=0, \quad \forall v_h\in V_h,\ \forall \hat{q}_h\in \hat{Q}_h,$$

and we separate p_h into its components

$$(5.21)\qquad p_h=\hat{p}_h+\tilde{p}_h, \qquad \hat{p}_h\in\hat{Q}_h, \tilde{p}_h\in\tilde{Q}_h.$$

By hypothesis (5.6) we can find $\hat{v}_h\in\hat{V}_h$ such that

$$(5.22)\qquad b(\hat{v}_h,\hat{q}_h-\hat{p}_h)=\|\hat{q}_h-\hat{p}_h\|_Q^2, \quad \|\hat{v}_h\|_V<\frac{1}{k_0}\|\hat{q}_h-\hat{p}_h\|_Q.$$

This yields, using (5.11) to get rid of the term $b(\hat{v}_h,\tilde{p}_h)$,

$$(5.23)\qquad \|\hat{q}_h-\hat{p}_h\|_Q\le c\,(\|u-u_h\|_V+\|p-q_h\|_Q).$$

By the triangle inequality, we thus have

$$(5.24)\quad \|p-\hat{p}_h\|_Q\le c\,(\|u-u_h\|_V+\inf_{q_h\in Q_h}\|p-q_h\|_Q)+\inf_{\hat{q}_h\in\hat{Q}_h}\|p-\hat{q}_h\|_Q$$

and no loss of accuracy will occur provided $\hat{Q}_h$ approximates Q with the same order as Q_h.

We can now summarize these results in the following:

Theorem 5.1: Let $\hat{V}_h\subset V_h\subset V$ and $\hat{Q}_h\subset Q_h\subset Q$ satisfy hypotheses (5.6) and (5.11). Let g satisfy (5.15) and let (u_h,p_h) be the solution of (5.4) and (5.5), $\hat{p}_h$ denoting the projection of p_h on $\hat{Q}_h$. One then has constants c_1 and c_2, independent of h, such that

$$(5.25)\qquad \|u-u_h\|_V\le c_1\Big(\inf_{\hat{v}_h\in\hat{V}_h}\|u-\hat{v}_h\|_V+\inf_{q_h\in Q_h}\|p-q_h\|_Q\Big)$$

$$(5.26)\quad \|p-\hat{p}_h\|_Q\le c_2\Big(\|u-u_h\|_V+\inf_{q_h\in Q_h}\|p-q_h\|_Q\Big)+\inf_{\hat{q}_h\in\hat{Q}_h}\|p-\hat{q}_h\|_Q. \ \square$$

Remark 5.2: Let $\hat{u}_h$ and $\hat{p}_h$ be the solution of the problem

$$a(\hat{u}_h, \hat{v}_h) + b(\hat{v}_h, \hat{p}_h) = (f, \hat{v}_h), \tag{5.27}$$

$$b(\hat{u}_h, \hat{q}_h) = (g, \hat{q}_h). \tag{5.28}$$

Making $v_h = \hat{v}_h$ and $q_h = \hat{q}_h$ in (5.4) and (5.5) and substracting, we get

$$a(u_h - \hat{u}_h, \hat{v}_h) + b(\hat{v}_h, p_h - \hat{p}_h) = 0, \qquad \forall \hat{v}_h \in \hat{V}_h, \tag{5.29}$$

$$b(u_h - \hat{u}_h, \hat{q}_h) = 0, \qquad \forall \hat{q}_h \in \hat{Q}_h. \tag{5.30}$$

Making $\hat{v}_h \in \operatorname{Ker} \hat{B}_h$ in (5.29) we have

$$\hat{u}_h = \hat{P} u_h, \tag{5.31}$$

where $\hat{P}$ is the projection operator on $\hat{Z}_h(g)$ with respect to the scalar product $a(\cdot, \cdot)$. □

Remark 5.3: The key of the above result is of course hypothesis (5.11). It is possible that extensions could be obtained if one could get *an estimate* on $b(\hat{v}_h, \tilde{q}_h)$ instead of requiring it to be zero. □

In order to apply Theorem 5.1 to the examples of Section VI.5.1 and to other cases, we shall need to introduce another special technique, namely, the use of macroelements. This will ultimately lead us to the analysis of the $Q_1 - P_0$ element in Section VI.5.4.

VI.5.3 Macroelement techniques

This section may be considered as a generalization of Section VI.4 in which we proved general stability results for a large class of continuous or discontinuous pressure elements through the use of internal nodes. We also refer to BOLAND–NICOLAIDES [A] for related results in a somewhat different setting. It is, however, clear that the notion of internal node can be extended by introducing composite elements or, in the language of STENBERG [A–E], macroelements. For instance, the quadrilateral element of Figure VI.17 is built from two quadratic triangles and node P may be considered as internal with respect to the quadrilateral.

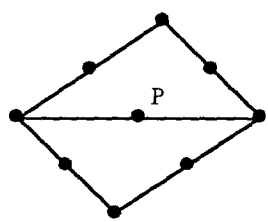


Figure VI.17

Let us now define more precisely what a mesh of macroelements may be. We first suppose given a standard partition of Ω into triangles or quadrilaterals. This partition bears a more or less standard finite element approximation of the problem at hand.

We shall denote V_h and Q_h these finite element spaces. *By a macroelement we now mean the union of a fixed number of adjacent elements along a well-defined pattern.* Indeed, a macroelement M should be equivalent, through a proper change of variables, to a reference macroelement $\hat{M}$. (STENBERG [C]). We now define, supposing that one has

$$M = \bigcup_{j=1}^{m} K_j, \tag{5.32}$$

the spaces

$$V_{0,M} = \{\underline{v} \mid \underline{v}_h \in V_h,\ \underline{v}_h = 0 \text{ in } \Omega \backslash M\}, \tag{5.33}$$

$$N_M = \{q_M \mid\ q_M = q_h|_M,\ q_h \in Q_h,\ \int_M q_h \operatorname{div} \underline{v}_h\, dx = 0,\ \forall \underline{v}_h \in V_{0,M}\}, \tag{5.34}$$

Supposing now that a mesh of macroelements is used, it is possible to try building on this mesh an interpolation operator Π_h such that one would have

$$\int_\Omega (\underline{u} - \Pi_h \underline{u}) \cdot \underline{\chi}_h\, dx = 0, \qquad \forall \underline{\chi}_h \in \Phi_h \supset \underline{\operatorname{grad}}\, Q_h \tag{5.35}$$

in the case of continuous pressure fields, or

$$\int_\Omega \operatorname{div}(\underline{u} - \Pi_h \underline{u})\, q_h\, dx = 0, \qquad \forall q_h \in Q_h \tag{5.36}$$

for discontinuous pressure fields. This would prove stability provided Π_h is a continuous operator.

It should now be clear to the reader that functions of $V_{0,M}$ associated with *internal nodes with respect to the macro-element*, will play a special role in building the operator Π_h. Two ways are indeed open. For continous pressure fields one may try directly to build Π_h satisfying (5.35). For discontinuous pressure fields one can split condition (5.36) into local elementwise conditions. These are again split into

$$\int_K \operatorname{div}(\underline{u} - \Pi_h \underline{u})\, q_M\, dx = 0, \qquad \forall q_M \in N_M^\perp \tag{5.37}$$

and

$$\int_K \operatorname{div}(\underline{u} - \Pi_h \underline{u})\, q_M\, dx = 0, \qquad \forall q_M \in N_M. \tag{5.38}$$

Condition (5.37) can be handled using the internal nodes associated to $V_{0,M}$. As to (5.38), it reduces in most cases to ensuring conservation of mass by properly choosing boundary nodes.

Remark 5.4: It should also be recalled that whenever stability has been proven for a class of discontinuous pressure field elements, any element built by using a subspace of the pressure space Q_h will also be stable. In particular, a subspace of continuous pressures can thus be shown to be stable. This is implicitly used in the work of STENBERG [C]. □

Remark 5.5: In Section VI.4 the choice of internal nodes had been made in to order to make the equivalent of (5.35) immediate. In the present case, one has to deal with internal nodes as they come from the building of the macroelement. In order to prove the equivalent of Proposition 4.1, we shall have to introduce an additional assumption. This assumption is essentially algebraic and has to do with the rank of a small linear system. □

Proposition 5.1: Let us suppose that V_h is defined on a mesh of macroelements and can be written as

$$V_h = \tilde{V}_h \oplus \Big(\bigoplus_M V_{0,M}\Big). \tag{5.39}$$

Let us moreover suppose $Q_h \subset H^1(\Omega)$ (that is, we use a continuous approximation for pressure). Suppose that on every M there is a space $\Phi_M \supset \underline{\text{grad}}\, Q_h|_M$ such that the matrix associated with

$$\int_M \underline{v}_h \cdot \underline{\phi}_h \, dx, \qquad \forall \underline{v}_h \in V_{0,M}, \quad \forall \underline{\phi}_h \in \Phi_M, \tag{5.40}$$

has rank $= \dim \Phi_M$. Let us suppose that on V_h we have an interpolation operator Π_1 such that

$$\sum_M h_M^{2r-2} |\underline{v} - \Pi_1 \underline{v}|^2_{r,M} \le c\, \|\underline{v}\|^2_{1,\Omega}, \quad r = 0, 1, \quad \forall \underline{v} \in V \tag{5.41}$$

Then the inf–sup condition holds.

Proof: We want to build Π_h satisfying (5.35). As in Proposition 4.1, one sets

$$\Pi_h \underline{u} = \Pi_1 \underline{u} + \Pi_2 \underline{w} \tag{5.42}$$

where $\underline{w} = \underline{u} - \Pi_1 \underline{u}$ and $(\Pi_2 \underline{w})|_M \in V_{0,M}$. We can then by hypothesis find $\Pi_2 \underline{w}$ by solving (and choosing a minimal solution if uniqueness fails)

$$\int_M \Pi_2 \underline{w} \cdot \underline{\phi}_h dx = \int_M (\underline{u} - \Pi_1 \underline{u}) \ \underline{\phi}_h \, dx \quad \forall \underline{\phi} \in \Phi_M \tag{5.43}$$

There remains to estimate $\Pi_2 \underline{w}$, which uses the same scaling argument as in Proposition 4.1. □

The key is thus to check that the matrix associated with (5.40) has full rank. This may be in some cases a tedious task but is nevertheless, a priori, a simple problem. In fact a closer look to the result shows that what we actually need is

$$\inf_{q_h \in Q_h|_M} \sup_{\underline{v}_h \in V_{0,M}} \frac{\int_M \operatorname{div} \underline{v}_h \, q_h \, dx}{\|\underline{v}_h\|_{1,M} \|q_h\|_{0,M/\mathbb{R}}} \geq \beta > 0, \tag{5.44}$$

that is, in each subdomain M the choice of $V_{0,M}$ for velocity fields and $Q_h|_M$ for pressure leads to a well-posed problem. This condition is strongly related to the *patch test* used by engineers (cf., e.g., ZIENKIEWICZ–QU–TAYLOR–MAKAZAWA [A]) although their counting of degrees of freedom is clearly insufficient. It is clear that the rank condition for (5.40) is a sufficient condition for (5.44) to hold.

We now turn to the case of discontinuous pressure elements. As in Proposition 4.2 we shall have to control separately the constant part of pressure by using nodes on M and use internal nodes for the remaining part. We refer the reader to STENBERG [C] from which we now quote the following result.

Proposition 5.2: Let us suppose on Ω a partition into macroelements such that

$$N_M \text{ is one-dimensional.} \tag{5.45}$$

Suppose, moreover, that there exists an interpolation operator $\tilde{\Pi}_h : V \to V_h$ such that one has

$$\begin{aligned} &\int_M \operatorname{div}(\tilde{\Pi}_h \underline{u} - \underline{u}) \, dx = 0, \qquad \forall M, \\ &\|\tilde{\Pi}_h \underline{u}\|_V \leq c \, \|\underline{u}\|_V . \end{aligned} \tag{5.46}$$

Then the inf–sup stability condition is satisfied. □

This contains the case where M is built from one element and thus Proposition 4.2. The proof is a generalization of the ideas of Section VI.4 and we refer the reader to the work of Stenberg for details. Note that in practice, (5.45) will follow from (5.44) and that (5.46) means that the pair (V_h, M_h) is stable, where M_h is the space of piecewise constants on the macroelements.

We shall now consider on a few examples, applications of the above result.

Example 5.3: *Taylor–Hood element* $(Q_2 - Q_1)$.

We consider, following STENBERG [C] a patch of quadrilaterals as in Figure VI.18. The velocity field will be approximated by a standard nine-node

patch we thus have six degrees of freedom for pressure that could be matched by the six internal velocity degrees of freedom associated with nodes M_1, M_2, and M_3 to use Proposition 5.1. STENBERG [C] instead, checks that N_M is one dimensional so that boundary nodes can be, in the classical way, used to get (5.46). Proposition 5.2 then proves stability for a macrowise discontinuous pressure field (in this case a most "unnatural" approximation). Stability for the continuous field is then obvious by Remark 5.1. The original proof of convergence for this element was given by BERCOVIER–PIRONNEAU [A].

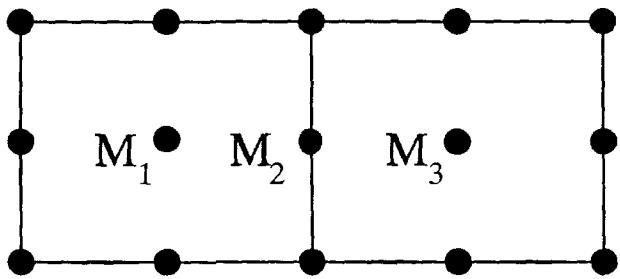


Figure VI.18

□

Example 5.4: *Cross-grid divergence-free elements.*

We consider elements presented in Examples 5.1 and 5.2, namely, the cross-grid $P_1 - P_1$ and $P_2 - P_1$ elements. Taking as the macroelement M the quadrilateral, we already checked that N_M is not one but two dimensional. We, however know, that in the case of Example 5.2, condition (5.46) holds. We thus introduce the space $\hat{Q}_h \subset Q_h$ by deleting from the pressure the spurious mode on each macroelement M (that is, a checkerboard patterned discontinuous function on M). Theorem 5.1 can then be applied using $\hat{V}_h = V_h$ for (5.11) is a direct consequence of the construction (as indeed in all cases of local modes). The same method enables us to eliminate the local mode in Example 5.1. We have however no way to check condition (5.46) and we cannot conclude from this the stability of the element. Indeed on a regular rectangular mesh a global mode arises and the analysis of this element will have to follow the analysis of the $Q_1 - P_0$ element in the next section. □

Example 5.5: *The Union Jack element.*

We consider a composite element made from piecewise linear triangular element following the pattern of Figure VI.19. This "Union Jack" element is used as an approximation for velocity while pressure is taken to be linear on M (and discontinuous). This composite element has the same degrees of freedom as the

direct from Proposition 5.2. In practice, such an element could be advantageous because its elementary matrix can be computed much more economically than the matrix associated with the $Q_2 - P_1$ element and the structure of the global matrix is sparser. There remains to see if this is sufficient to suffer the loss of one order in accuracy. □

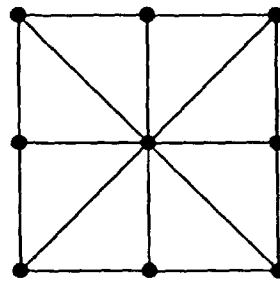


Figure VI.19

VI.5.4 The bilinear velocity-constant pressure ($Q_1 - P_0$) element

We now consider the analysis of what is probably the most popular of all elements for incompressible flow problems (Figure VI.20). This is perhaps also the hardest to analyze and as we shall see only partial results are known (at least at the time of this writing). Origins of this element can be traced back to finite-difference methods (FORTIN–PEYRET–TEMAM [A]) and its peculiar properties were soon recognized. In particular, the checkerboard pressure mode was already a familiar feature long before the scheme used was written in terms of finite elements.

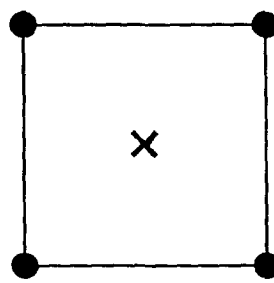


Figure VI.20

Let us summarize the basic facts. On a regular mesh, for a problem with Dirichlet boundary conditions, two singular values (cf. Section II.3.2) of the matrix associated with the discrete divergence vanish instead of one. We thus have a pure spurious pressure mode in the terminology of GRESHO–GRIFFITHS–LEE–SANI [A]. When the mesh is slightly distorted, or for other types of bound-

thus implying an ill-conditioning of the problem. This ill-conditioning is fortunately almost restricted to pressure: we have an impure pressure mode which can be eventually filtered but does not seem to affect (at least substantially) the computation of velocity. This is still not, however, the whole story. One could indeed hope, from all this, that an inf–sup stability condition could hold for the third singular value instead of the second and that we could have stability in a simple quotient space. Experimental evidence showed this hope to be false: on a regular mesh, a large number of eigenvalues converge to zero at order h (MALKUS [A]). JOHNSON–PITKÄRANTA [A] indeed proved the constant k_h to be $O(h)$ (see also ODEN–JACQUOTTE [A], BOLAND–NICOLAIDES [B,C], MANSFIELD [A]). The standard estimates then led to the conclusion that no convergence occurred, in complete contradiction with experience. The paper of Johnson and Pitkäranta provided a first result by showing, on a regular mesh, that under stricter regularity assumptions than usual on the solution convergence still takes place.

STENBERG–PITKÄRANTA [A] proved a convergence result, without special regularity assumptions, for a special type of mesh. We shall now consider a new proof of these results using the technique of Section VI.5.2. To make things simpler we shall first consider the case of a regular rectangular mesh. We shall thus try to find subspaces $\hat{V}_h$ and $\hat{Q}_h$ satisfying the stability condition and condition (5.11). For this purpose, we consider a macroelement (Figure VI.22) M formed of four quadrilaterals. On this macroelement a piecewise constant pressure has four degrees of freedom. We introduce a local basis on M, ϕ_1, ϕ_2, ϕ_3, and ϕ_4 described symbolically in the figure.

+1	+1
+1	+1

a) $\phi_{1,M}$

-1	+1
-1	+1

b) $\phi_{2,M}$

-1	-1
+1	+1

c) $\phi_{3,M}$

-1	+1
+1	-1

d) $\phi_{4,M}$

Figure VI.21: Pressure basis functions on M

A checkerboard mode will obviously take its roots in ϕ_4. We, therefore, introduce quite naturally the space

$$\hat{Q}_h = \sum \Big(\sum^{3} \alpha_{i,M}\, \phi_{i,M} \Big) \tag{5.47}$$

and, therefore,

$$\tilde{Q}_h = \sum_M \alpha_{4M}\, \phi_{4,M}. \tag{5.48}$$

The choice of $\hat{V}_h$ can then be inferred from other well-known elements. We use as degrees of freedom the two values of velocity at the vertices of M and at its barycenter and the normal value (rather a correction to this value) at mid-side nodes (Figure VI.22). This normal node is readily used to control the condition (5.46) and, in $\hat{Q}_h$, N_M is one dimensional ($= \phi_{1M}$). Stability of $\hat{V}_h$, $\hat{Q}_h$ is, thus, immediate from Proposition 5.2. In order to apply Theorem 5.1 we need to check (5.11) that is now

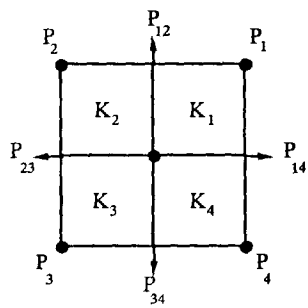


Figure VI.22

$$\int_M \phi_{4M} \operatorname{div} \underline{\hat{v}}_h \, dx = 0, \qquad \forall M,\ \forall \underline{\hat{v}}_h. \tag{5.49}$$

In order to check this, let us consider the shape function $\underline{w}_1$ associated to vertex P_1, for instance, which is the function of $Q_1(M)$ having the whole of M as its support. A straightforward computation then shows that one has

$$\begin{aligned} \int_M \phi_{4M} \operatorname{div} \underline{w}_1 \, dx &= \int_{\partial K_1} \underline{w}_1 \cdot \underline{n}\, ds - \int_{\partial K_2} \underline{w}_1 \cdot \underline{n}\, ds \\ &\quad + \int_{\partial K_3} \underline{w}_1 \ \underline{n}\, ds - \int_{\partial K_4} \underline{w}_1 \ \underline{n}\, ds = 0. \end{aligned} \tag{5.50}$$

In the same way the shape function $\underline{w}_{12}$ associated with node P_{12} satisfies

$$\int_M \phi_{4M} \operatorname{div} \underline{w}_{12} \, dx = \int_{\partial K_1} \underline{w}_{12} \ \underline{n}\, ds - \int_{\partial K_2} \underline{w}_{12} \cdot \underline{n}\, ds = 0 \tag{5.51}$$

and this is also true in the adjacent element because the mesh is aligned. The shape function associated with the barycenter trivially satisfies the condition. Condition (5.11), therefore, holds and we have by Theorem 5.1

$$(5.52) \qquad \|\underline{u}-\underline{u}_h\|_V \leq \Big(\inf_{\hat{\underline{v}}_h \in \hat{V}_h} \|\underline{u}-\hat{\underline{v}}_h\|_V + \inf_{q_h \in Q_h} \|p-q_h\|_Q \Big).$$

In the present case it is clear that an error estimate in $\hat{V}_h$ has the same order as an estimate in V_h and the result is therefore almost optimal. We also have convergence of (filtered) pressure in $\hat{Q}_h$ by estimate (5.26). Following PITKÄRANTA–STENBERG [A] we can now extend this result to the case where the mesh is made from super macroelements as in Figure VI.23. A general quadrilateral is divided in a regular way into sixteen quadrilaterals. (It is well known (FORTIN [D]) that on a nonrectangular mesh at least a 4×4 patch of elements is needed to generate a nontrivial discrete divergence-free function.) We, thus, have four "submacros" similar to the previous case. The space of filtered pressures $\hat{Q}_h$ is taken exactly as on the regular mesh and is still defined by (5.47). The space $\hat{V}_h$ is defined by the following degrees of freedom: the values of velocity at the vertices of the M_i, the values at the barycenters of the M_i, and a correction of the component of velocity parallel to the mesh at the mid-side nodes of the M_i internal to SM. It is readily checked that on SM Proposition 5.2 applies. One can also directly build an interpolation operator enabling us to check the inf–sup condition. Mid-side nodes of SM control the part of pressure which is constant on the whole of SM. Internal mid-side nodes ensure mass balance on each M_i and the nodes at the barycenters of the M_i end the job. It must be remarked that the alignment of mid-side velocities along the mesh is an essential feature of the construction.

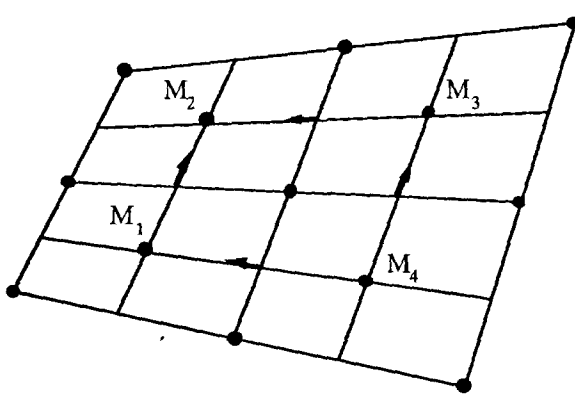


Figure VI.23: A supermacro SM and its submacros

In order to prove condition (5.11) the only difficult point is to check that (5.49) still holds on every M_i. We refer the reader to PITKARANTA–STENBERG

[A] for this proof. It is then possible to use Theorem 5.1 and to get optimal error estimates.

This is still not the whole story about this peculiar element. It is also possible to prove stability on meshes built from macroelements like in Figure VI.24 (STENBERG [C], LE TALLEC–RUAS [A]) without filtering or other subterfuge. This is coherent with the known experimental fact that on a general distorted mesh pressure modes disappear and the inf–sup constant is independent of h. This last fact is still resisting analysis. It is our hope that the above technique could be generalized to yield the complete result.

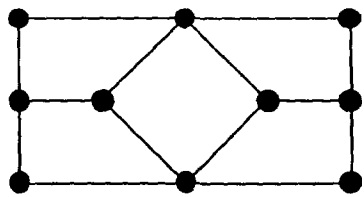
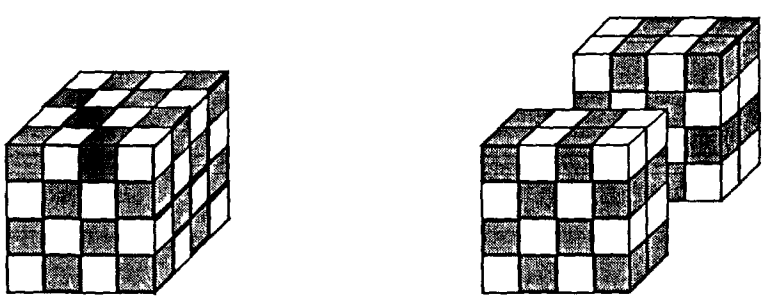


Figure VI.24

The above discussion can be extended to the three-dimensional case. Things are made still more complicated by the fact that on a regular mesh (let say a $n \times n \times n$ assembly of elements to fix ideas) we do not have one spurious pressure mode but $3n - 2$ of them. (This will also mean the same number of compatibility conditions on data so that trouble should be expected from time to time when using apparently reasonable boundary conditions). These spurious modes are depicted in the Figure VI.25. One of them is the genuine three-dimensional checkerboard mode (Figure VI.25a). The other ones are built from an assembly of two-dimensional modes. In Figure VI.25b we have sliced the mesh in order to make apparent the internal structure of this mode. There are $3(n - 1)$ possible slices so that we find the number of modes stated above.



a) b)

Figure VI.25: Pressure modes

We now sketch the extension of the above proof to the three-dimensional case. We shall only present the rectangular case to avoid lengthening unduly this exposition. We, thus, suppose the mesh is built from $2 \times 2 \times 2$ macroelements (Figure VI.26). Our pressure space $\hat{Q}_h$ will be built from Q_h deleting on each macro-element four ($= 3 \times 2 - 2$)spurious modes sketched in Figure VI.26.

a) b)

Figure VI.26

The mode depicted in Figure VI.26b has obviously two other symmetrical counterparts. On each macro-element we thus keep the three-dimensional analogues of the basis functions $\phi_{1,M}$, $\phi_{2,M}$,and $\phi_{3,M}$ of Figure VI.21 (we have obviously four of them now). We must now introduce $\hat{V}_h$. This is done again by taking off some degrees of freedom from V_h. The remaining ones are sketched in Figure VI.27.

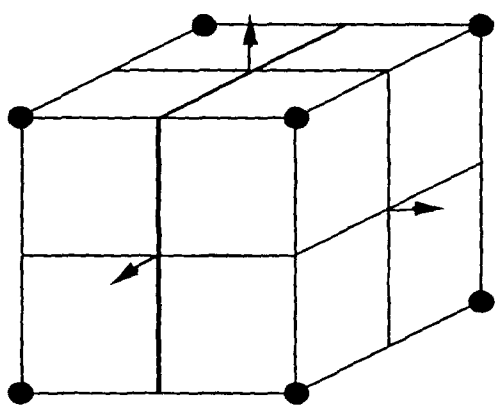


Figure VI.27: Degrees of freedom for $\hat{V}_h$

The internal node at the barycenter of the element is also used. It is now clear that $\hat{V}_h$, $\hat{Q}_h$ is a stable pair that provides $O(h)$ convergence. There remains to check condition (5.11), that is, that $\hat{V}_h$ is transparent with respect to $\tilde{Q}_h$. This is done exactly as in the two-dimensional case by a simple check of flow

balance at the surface of elements. Theorem 5.1 then applies and we get $0(h)$ convergence for velocities and filtered pressures.

It could be hoped that the same kind of analysis could be done for equal interpolation continuous pressure methods such as the $Q_1 - Q_1$ approximation. Unfortunately, we could find no way in which condition (5.11) could be made to hold and an analysis of the convergence properties of these approximations remains an open question. We can, however, introduce an alternate way of stabilizing such approximations and this is done is the following section.

VI.5.5 Other stabilization procedures (Augmented formulations)

It is clear or should be clear from the results presented in this chapter that the key of success in stabilizing incompressible elements is in weakening the discrete divergence-free condition. This was done up to now by reducing the space Q_h of pressures or by enriching the space V_h of velocity field. Another still unexplored possibility is to explicitly weaken the condition $\operatorname{div}_h \underline{u}_h = 0$ by changing it to

$$\operatorname{div}_h \underline{u}_h = g_h \tag{5.53}$$

when g_h is a (well-chosen) "small" function.

One step in this direction has been done in the work of BREZZI–PITKÄRANTA [A] who considered the relaxed condition

$$-\int_\Omega \operatorname{div} \underline{u}_h \, q_h \, dx + \alpha \sum_K h_K^2 \int_K \underline{\operatorname{grad}}\, p_h . \underline{\operatorname{grad}}\, q_h \, dx = 0 \tag{5.54}$$

in the case of a continuous pressure approximation (that is, $Q_h \subset H^1(\Omega)$). In (5.54), α is any positive real number and can be chosen equal to 1. On a regular mesh this is a discrete form of

$$-\operatorname{div} \underline{u} = -\alpha h^2 \, \Delta p. \tag{5.55}$$

It is easy to understand that appearance of oscillations due to spurious pressure modes will make $\Delta_h \, p_h$ large. This will relax the divergence-free condition thus preventing the growth of such oscillations. It is easy to understand that appearance of oscillations due to spurious pressure modes will make $\Delta_h \, p_h$ large. This will relax the divergence-free condition thus preventing the growth of such oscillations. It is also clear that (5.55) introduces a perturbation in the problem and that a consistency error (of order h) is unavoidable. A way of circumventing this problem can be found, essentially, in the use of the augmented formulations presented in Section I.1.5. According to this section, we can now consider the following augmented formulation:

$$\begin{aligned} \inf_{\underline{v}_h \in V_h} \sup_{q_h \in Q_h} \{ \mu \int_\Omega |\underline{\underline{\varepsilon}}(\underline{v}_h)|^2 \, dx - \int_\Omega q_h \operatorname{div} \underline{v}_h \, dx - \int_\Omega \underline{f} \cdot \underline{v}_h \, dx \\ + \sum_K \alpha(K) \int_K |A\underline{v}_h - \underline{\operatorname{grad}}\, q_h + \underline{f}|^2 \, dx \}. \end{aligned} \tag{5.56}$$

This gives rise to the variational formulation

$$(5.57)\quad \mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_h) : \underline{\underline{\varepsilon}}(\underline{v}_h)\, dx - \int_\Omega p_h \operatorname{div} \underline{v}_h\, dx - \int_\Omega \underline{f} \cdot \underline{v}_h\, dx$$
$$+ \sum_K \alpha(K) \int_K (A\underline{u}_h - \underline{\operatorname{grad}}\, p_h + \underline{f}) \cdot A\underline{v}_h\, dx = 0, \qquad \forall \underline{v}_h \in V_h,$$

$$(5.58)\quad \int_\Omega q_h \operatorname{div} \underline{u}_h\, dx$$
$$- \sum_K \alpha(K) \int_K (A\underline{u}_h - \underline{\operatorname{grad}}\, p_h + \underline{f}) \cdot \underline{\operatorname{grad}}\, q_h\, dx = 0, \quad \forall q_h \in Q_h.$$

It can be seen easily that, by taking

$$\alpha(K) = \bar{\alpha} h_K^2$$

with $\bar{\alpha} > 0$ small enough (in comparison to the inverse inequality constant c which appears in (III.2.5)), we can get stability and optimal error bounds in the norm

$$|||(\underline{u}, p)|||^2 = \|\underline{u}\|_{1,\Omega}^2 + \sum_K h_K^2 \|\underline{\operatorname{grad}} p\|_{0,K}^2$$

for *any choice* of the finite element spaces V_h and Q_h. We refer to FRANCA–HUGHES [A] for the details of this idea which was first introduced in a non-symmetric form in HUGHES–FRANCA–BALESTRA [A], guided by the basic ideas introduced in the theory of the approximation of hyperbolic equations by HUGHES–BROOKS [A] and JOHNSON [B]. A variant of the original non-symmetric approach can be found in BREZZI–DOUGLAS [A]. A very interesting variant of the symmetric approach (5.57) and (5.58) was introduced by DOUGLAS–WANG [A]. As we have seen in (I.5.11), it consists in considering, instead of (5.58),

$$(5.61)\quad \int_\Omega q_h \operatorname{div} \underline{u}_h\, dx$$
$$+ \sum_K \alpha(K) \int_K (A\underline{u}_h - \underline{\operatorname{grad}}\, p_h + \underline{f}) \cdot A\underline{\operatorname{grad}}\, q_h\, dx = 0, \quad \forall q_h \in Q_h.$$

It can be seen (cf. DOUGLAS–WONG [A]) that the choice (5.59) for $\alpha(K)$ now makes the problem defined by (5.57) and (5.61) stable and optimally convergent for any choice of the finite element spaces V_h, Q_h, and for any choice of $\bar{\alpha} > 0$. The following result is proved in FRANCA–STENBERG [A].

Proposition 5.3 Assume that one of the following conditions is satisfied

(i) $(\mathcal{L}_1^1 \cap H_0^1(\Omega))^2 \subset V_h$ and $Q_h \subset C^0(\bar{\Omega})$

(ii) the pair $(V_H, \mathcal{L}_0^0)$ satisfies the inf–sup condition

Then the method (5.57) and (5.58) (for $\bar{\alpha} > 0$ small enough), and the method (5.57) and (5.61) (for any $\bar{\alpha} > 0$) are stable and yield optimal error bounds. □

An interesting aspect which is still to be investigated in an exhaustive way is the relationship between the use of augmented formulations and the use of suitable bubble functions to augment the velocity space. To give a simple hint, let us consider what happens to the MINI element of Example 3.7, when the bubbles are eliminated by static condensation, that is, by Gaussian elimination at element level. For this, let $(\underline{u}_h, p_h)$ be the discrete solution of the Stokes problem

$$(5.62)\qquad \begin{cases} a(\underline{u}_h, \underline{v}_h) + \displaystyle\int_\Omega \underline{v}_h \cdot \underline{\mathrm{grad}}\, p_h \, dx = \int_\Omega \underline{f} \cdot \underline{v}_h \, dx, & \forall \underline{v}_h \in V_h, \\ \displaystyle\int_\Omega \underline{u}_h \cdot \underline{\mathrm{grad}}\, q_h \, dx = 0, & \forall q_h \in Q_h, \end{cases}$$

with the MINI element, that is, $V_h = (\mathcal{L}_1^1 \cap H_0^1(\Omega) \oplus B_3)^2$, $Q_h = \mathcal{L}_1^1$, and let us split the solution $\underline{u}_h$ as

$$(5.63)\qquad \underline{u}_h = \underline{u}_H^1 + \underline{u}_h^b; \quad \underline{u}_h^1 \in (\mathcal{L}_1^1 \cap H_0^1(\Omega))^2, \underline{u}_h^b = \sum_K \underline{\beta}_K b_K \in (B_3)^2.$$

We first note that, for $\underline{v}_h \in (B_3)^2$ and $\underline{w}_h \in (\mathcal{L}_1^1)^2$, we have

$$(5.64)\qquad a(\underline{v}_h, \underline{w}_h) = a(\underline{w}_h, \underline{v}_h) = 0.$$

Therefore, if we take $\underline{v}_h \in (B_3)^2$ in the first equation of (5.62) we easily obtain

$$(5.65)\qquad \delta_K \underline{\beta}_K = \int_K (\underline{f} - \underline{\mathrm{grad}}\, p_h) \underline{b}_k \, dx$$

where,

$$(5.66)\qquad \delta_K = \mu \int_K |\underline{\varepsilon}(\underline{b}_K)|^2 \, dx.$$

If we assume now that $\underline{f}$ is piecewise constant, we can rewrite (5.65) in the form

$$(5.67)\qquad \delta \underline{\beta}_K = \gamma_K (\underline{f} - \underline{\mathrm{grad}}\, p_h),$$

$$\gamma_k = \int_\Omega b_K \, dx. \tag{5.68}$$

From (5.67), we the get

$$\underline{u}_h^b = \sum_K (\gamma_K/\delta_K)(\underline{f} - \underline{\text{grad}}\, p_h).|_K \tag{5.69}$$

Substituting the value of $\underline{u}_h^b$ given by (5.69) into (5.63), we can now rewrite the second equation of (5.62) as

$$\begin{aligned} &\int_\Omega \underline{u}_h^1 \cdot \underline{\text{grad}}\, q_h \, dx \\ &+ \sum_K (\gamma_K/\delta_K)(\underline{f} - \underline{\text{grad}}\, p_h)|_K \cdot \int_K b_K \, \underline{\text{grad}}; q_h \, dx = 0, \quad \forall q_h \in Q_h, \end{aligned} \tag{5.70}$$

or equivalently

$$\begin{aligned} &\int_\Omega \underline{u}_h^1 \cdot \underline{\text{grad}}\, q_h \, dx \\ &\qquad + \sum_K (\gamma_K^2/\delta_K)(\underline{f} - \underline{\text{grad}}\, p_h)|_K \cdot \underline{\text{grad}}\, q_h = 0, \quad \forall q_h \in Q_h \end{aligned} \tag{5.71}$$

and finally

$$\begin{aligned} &\int_\Omega \underline{u}_h^1 \cdot \underline{\text{grad}}\, q_h \, dx \\ &+ \sum_K (\gamma_K^2/\delta_K \text{Meas}(K)) \int_K (\underline{f} - \underline{\text{grad}}\, p_h)|_K \cdot \underline{\text{grad}}\, q_h \, dx = 0, \quad \forall q_h \in Q_h \end{aligned} \tag{5.72}$$

which is

$$\begin{aligned} &\int_\Omega \underline{u}_h^1 \cdot \underline{\text{grad}}\, q_h \, dx \\ &\qquad + \sum_K \alpha(K) \int_K (\underline{f} - \underline{\text{grad}}\, p_h)|_K \cdot \underline{\text{grad}}\, q_h \, dx = 0, \quad \forall q_h \in Q_h, \end{aligned} \tag{5.73}$$

with

$$\alpha(K) = \gamma_K^2/\delta_K \text{Meas}(K) \sim h_K^2. \tag{5.74}$$

Note that we now have, using (5.64) and the first equation of (5.62)

$$a(\underline{u}_h^1, \underline{v}_h) + \int_\Omega \underline{v}_h \cdot \underline{\text{grad}}\, p_h \, dx = \int_\Omega \underline{f} \cdot \underline{v}_h \, dx, \qquad \forall \underline{v}_h \in (\mathfrak{L}_1^1 \cap H_0^1(\Omega))^2 \tag{5.75}$$

and it is easy to check that (5.74) and (5.75) coincide with (5.57) and (5.58) since $A\underline{v}_h = 0$. Hence, in a sense, we can say that the introduction of a bubble function and its elimination by static condensation leads to the augmented formulation (5.56) for a suitable choice of $\alpha(K)$, given here by (5.74), depending on the shape of the bubble. For more details about this subject, for the MINI element, we refer to PIERRE [A,B]. However, the relationship between bubbles and augmented formulations is probably larger as suggested by a simple comparison between Proposition 5.3 and Propositions 4.2 and 4.3. Indeed, if one of the two conditions (i) or (ii) is satisfied, it is always possible, from Corollary 4.1 or Corollary 4.2, to stabilize the proble by adding bubble functions to the velocity space.

As presented above, the stabilization procedure cannot be used for discontinuous pressure elements where expressions such as (5.54) make no sense. This is however not a definite obstacle. We may indeed consider two cases.

Suppose that we have an approximation $V_h \times Q_h$ that does not satisfy the inf–sup condition but that a subspace $V_h \times \hat{Q}_h$ does. If $\hat{Q}_h$ contains piecewise constants we could use an elementwise analogue of (5.54) in the same way as adding bubbles to stabilize. We could even manage so that the $\hat{Q}_h$ part of pressure is not affected. This could be used for instance to stabilize the $Q_2 - Q_1$ approximation that is known to suffer from spurious pressure modes.

Another possibility is offered by employing a mixed method to implement (5.54). To illustrate this point, we shall now show how $Q_1 - P_0$ approximation (Section VI.5.4) can be stabilized.

To simplify the exposition we shall consider the case of a rectangular mesh but the techniques of Chapters III through V could provide an obvious extension to a general mesh. Let us then consider the $Q_1 - P_0$ of section VI.5.4 and let us introduce $\underline{\chi}_h \in RT_{[0]}(\Omega, \mathcal{T}_h)$. On our rectangular mesh, we have

$$\underline{\chi}_h|_K = \{(k_1, k_2) | k_1 = a_0 + a_1 x,\ k_2 = b_0 + b_2 y\}, \tag{5.76}$$

and we have div $\underline{\chi}_h|_K \in P_0(K)$. We may, thus, write instead of (5.54)

$$\begin{cases} (\operatorname{div} \underline{u}_h, q_h) = -\varepsilon(\operatorname{div} \underline{\chi}_h, q_h), & \forall q_h \in Q_h, \\ (\underline{\chi}_h, \underline{m}_h) = (p_h, \operatorname{div} \underline{m}_h), & \forall \underline{m}_h \in RT_{[0]}(\Omega). \end{cases} \tag{5.77}$$

This, of course, is a mixed approximation of div $\underline{u} = \varepsilon \Delta p$.

Using the λ-trick of Chapter V, one can eliminate $\underline{\chi}_h$ and obtain a discrete problem having exactly the same structure as in continuous pressure elements, its unknown being velocities at vertices and pressure at mid-side points. This has a major drawback: one can no longer eliminate pressure by a penalty method (cf. Section VI.7) as this is usually done for a $Q_1 - P_0$ approximation. One possible

cure would be to proceed iteratively, that is, to use $\varepsilon\Delta p_n$ to obtain u^{n+1} and p^{n+1}. For instance the following procedure can be shown to converge (although slowly in practice):

$$(5.78)\qquad \begin{cases} a(\underline{u}^{n+1},\underline{v}) - (p^{n+1},\operatorname{div}\underline{v}) = (\underline{f},\underline{v}), \\ (\operatorname{div}\underline{u}^{n+1},q) - \varepsilon_1(p^{n+1},q) = -\varepsilon_2(\underline{\operatorname{grad}}\,p^n,\underline{\operatorname{grad}}\,q) - \varepsilon_1(p^n,q), \end{cases}$$

provided $\varepsilon_1/\varepsilon_2$ is large enough.

As we shall see in Section VI.7 solving (5.78) reduces to a penalty method. It would from this be quite interesting to find an efficient way to compute the solution by a more clever iterative procedure.

Finally, let us point out another way in which the $Q_1 - P_0$ approximation can be stabilized. The idea is to consider directly (5.53) and to make g_h a control variable. This supposes that we can define an objective function to be minimized. We shall show how this can be done using the space $\tilde{Q}_h$ defined in Section VI.5.4 by (5.48). We can thus consider, provided the mesh is formed of 2×2 macroelements

$$(5.79)\qquad F(p_h) = P_{\tilde{Q}_h}(p_h).$$

This projection is readily computed by a local process. We can then consider the control problem

$$(5.80)\qquad \inf_{g_h\in Q_h}\int_\Omega |F(p_h)|^2\,dx$$

under the constraint

$$(5.81)\qquad \begin{cases} a(\underline{u}_h,\underline{v}_h) - (p_h,\operatorname{div}\underline{v}_h) = (\underline{f},\underline{v}_h), & \forall \underline{v}_h\in V_h, \\ (\operatorname{div}\underline{u}_h,q_h) = (g_h,q_h), & \forall q_h\in Q_h. \end{cases}$$

The solution of this is equivalent to the solution with p_h and q_h chosen in $\hat{Q}_h$ as defined by (5.47), which is known to be stable. The advantage here is to avoid manipulating the problem in $\hat{Q}_h$ that requires the explicit construction of macroelements.

A gradient method to solve (5.80) and (5.81) can easily be built and leads to solving a sequence of Stokes problems (with the same matrix).

Let g_h^0 be chosen arbitrarily (e.g., $g_h^0 = 0$). Supposing g_h^n known we solve,

$$(5.82)\qquad \begin{cases} a(\underline{u}_h^n,\underline{v}_h) - (p_h^n,\operatorname{div}\underline{v}_h) = (\underline{f},\underline{v}_h), & \forall \underline{v}_h\in V_h, \\ (\operatorname{div}\underline{u}_h^n,q_h) = (g_h^n,q_h), & \forall q_h\in Q_h. \end{cases}$$

and then the adjoint problem

$$(5.83)\qquad \begin{cases} a(\underline{\lambda}_h^n, \underline{v}_h) - (\pi_h^n, \operatorname{div} \underline{v}_h) = 0, & \forall \underline{v}_h \in V_h, \\ (\operatorname{div} \underline{\lambda}_h^n, q_h) = (F(p_h^n), F(q_h)), & \forall q_h \in Q_h. \end{cases}$$

One can update g_h by

$$(5.73)\qquad g_h^{n+1} = g_h^n + \rho_n\, \pi_h^n.$$

The coefficient ρ_n can be determined in order to get a steepest descent method. Such a procedure could be extended to other approximations where a subspace of Q_h is known to yield a stable computation but is hard to build explicity. It must also be emphasized that (5.82) and (5.83) use the same matrix and that solving a sequence may be a small increase in computing costs if this matrix is factored once and for all.

VI.6 An Alternative Technique of Proof and Generalized Taylor–Hood Elements

We now consider a very popular choice of element for incompressible problems. Velocity is approximated by a standard P_k element and pressure by a standard (continuous) P_{k-1}; that is, using the notation of Chapter III, $\underline{v}_h \in (\mathfrak{L}_k^1)^2$, $p_h \in \mathfrak{L}_{k-1}^1$. This choice has an analogue on rectangles using $(\mathfrak{L}_{[k]}^1)^2$ for velocities and $\mathfrak{L}_{[k-1]}^1$ for pressure. Such an approximation has arisen from experimental considerations. First attempts which used equal interpolation ($P_2 - P_2$, for instance) yielded either locking or spurious pressures. It was realized that lowering the approximation of pressure by one degree was a way to get good results. These elements do not fit in the results of the previous section: it is possible to get a proof of stability for the $P_2 - P_1$ element without adding bubble functions. The first proof of convergence for this element was given in BERCOVIER–PIRONNEAU [A] using a weaker form of the inf–sup condition The analysis was subsequently improved by VERFÜRTH [A] who showed that the classical inf–sup condition is indeed satisfied. We give an alternate proof below.

Proposition 6.1: Assume that every triangle K in $\mathcal{T}_h$ has at most one edge on ∂K. Then the choice $V_h = (H_0^1 \cap \mathfrak{L}_2^1)$ and $Q_h = \mathfrak{L}_1^1$ satisfies the inf–sup condition.

Proof: Let $q_h \in Q_h$ and let $\bar{q}_h$ be its L^2-projection onto $\mathfrak{L}_0^0$ (piecewise constants). Moreover, let $\underline{v} \in H_0^1(\Omega)$ be such that

$$(6.1)\qquad \int_\Omega \operatorname{div} \underline{v}\, \bar{q}_h\, dx = \|\bar{q}_h\|_{0/\mathbb{R}}, \qquad \|\underline{v}\|_1 \le c_1\, \|\bar{q}_h\|_{0/\mathbb{R}}.$$

We first construct $\underline{w}_h \in V_h$ as $\underline{w}_h = \tilde{\Pi}_1 \underline{v}$, with $\tilde{\Pi}_1$ given by **(4.14)**. Then we have

$$
\begin{aligned}
\int_\Omega \operatorname{div} \underline{w}_h \, q_h \, dx &= \int_\Omega \operatorname{div} \underline{w}_h \, \bar{q}_h \, dx + \int_\Omega \operatorname{div} \underline{w}_h (q_h - \bar{q}_h) \, dx \\
&= |\bar{q}_h|^2_{0/\mathbb{R}} + \int_\Omega \operatorname{div} \underline{w}_h \, (q_h - \bar{q}_h) \, dx \\
&\geq \|\bar{q}_h\|^2_{0/\mathbb{R}} - \|\underline{w}_h\|_1 \, \|q_h - \bar{q}_h\|_0 .
\end{aligned} \tag{6.2}
$$

Recalling **(4.14)**, we have

$$
\|\underline{w}_h\|_1 = \|\tilde{\Pi}_1 \underline{v}\|_1 \leq c_2 \, \|\underline{v}\|_1 \leq \gamma \, \|\bar{q}_h\|_{0/\mathbb{R}} \tag{6.3}
$$

so that **(6.2)** becomes

$$
\int_\Omega \operatorname{div} \underline{w}_h \, q_h \geq \|\bar{q}_h\|^2_{0/\mathbb{R}} - \gamma \, \|\bar{q}_h\|_{0/\mathbb{R}} \|q_h - \bar{q}_h\|_0 . \tag{6.4}
$$

We are now going to construct a $\underline{z}_h \in V_h$ which takes care of the "nonconstant" part, $q_h - \bar{q}_h$. We first define $\underline{z}_h$ to be zero at all vertices of $\mathcal{T}_h$, and $\underline{z}_h \cdot \underline{n}$ to be zero at all mid-points of the edges of $\mathcal{T}_h$. Since $\underline{z}_h$ must be in $(H^1_0)^2$, we also need $\underline{z}_h \cdot \underline{t}$ (tangential component of $\underline{z}_h$) to vanish on the mid-points on $\partial\Omega$. Then for every edge e internal we define $\underline{z}_h \cdot \underline{t}$ at the mid-point M of e by

$$
\underline{z}_h \;\; \underline{t}(M) := -|e|^2 \, (\underline{\operatorname{grad}} \, q_h \cdot \underline{t})(M). \tag{6.5}
$$

An easy scaling argument shows that

$$
\|z_h\|_1 \leq c_2 \, \|q_h - \bar{q}_h\|_0 . \tag{6.6}
$$

We now have , for every K in $\mathcal{T}_h$,

$$
\begin{aligned}
\int_K \operatorname{div} \underline{z}_h \, q_h \, dx &= - \int_K \underline{z}_h . \underline{\operatorname{grad}} \, q_h \, dx \\
&= - \sum_M \frac{|K|}{3} (\underline{z}_h \cdot \underline{\operatorname{grad}} \, q_h)(M) \\
&= \sum_{M \notin \partial\Omega} |e|^2 \frac{|K|}{3} |\underline{\operatorname{grad}} \, q_h . \underline{t}(M)|^2 \\
&:= |||\underline{\operatorname{grad}} \, q_h|||^2_{0,K} ,
\end{aligned} \tag{6.7}
$$

using the fact that $\underline{z}_h . \underline{\operatorname{grad}} \, q_h$ is in $P_2(K)$. Now, since $\underline{\operatorname{grad}} \, q_h$ is constant in K and since K has at least two different mid-points $M \notin \partial\Omega$, we can check that $|||\underline{\operatorname{grad}} q_h|||_{0,K}$ is zero if and only if q_h is constant. Another scaling argument then shows

$$
|||\underline{\operatorname{grad}} q_h|||^2_{0,K} \geq c_3 \, \|q_h - \bar{q}_h\|^2_{0,K} . \tag{6.8}
$$

From (6.7) and (6.8), summing in K, we have

$$\int_\Omega \operatorname{div} \underline{z}_h \, q_h \, dx \geq c_3 \, \|q_h - \bar{q}_h\|_{0,\Omega}^2. \tag{6.9}$$

We look now for $\underline{v}_h \in V_h$ of the type

$$\underline{v}_h = \underline{w}_h + \beta \underline{z}_h, \tag{6.10}$$

with β to be chosen. We have from (6.4) and (6.9)

$$\int_\Omega \operatorname{div} \underline{v}_h \, q_h \, dx \geq \|\bar{q}_h\|_{0/\mathbb{R}}^2 - \gamma \, \|\bar{q}_h\|_{0/\mathbb{R}} \|q_h - \bar{q}_h\|_0 + \beta c_3 \, \|q_h - \bar{q}_h\|_0^2. \tag{6.11}$$

Choosing $\beta = 1/2 + \gamma^2/(2c_3)$, we have

$$\int_\Omega \operatorname{div} \underline{v}_h \, q_h \geq \frac{1}{2}\|\bar{q}_h\|_{0/\mathbb{R}}^2 + \frac{c_3}{2}\|q_h - \bar{q}_h\|_0^2 \geq \tfrac{1}{2} \min(1, c_3) \, \|q_h\|_{0/\mathbb{R}}^2. \tag{6.12}$$

On the other hand, from (6.3), (6.6), and (6.10) we have

$$\|\underline{v}_h\|_1 \leq \|\underline{w}_h\|_1 + \beta \|\underline{z}_h\|_1 \leq (\gamma + \beta c_2) \, \|q_h\|_{0/\mathbb{R}}, \tag{6.13}$$

and from (6.12) and (6.13)

$$\frac{\int_\Omega \operatorname{div} \underline{v}_h \, q_h \, dx}{\|\underline{v}\|_1} \geq \frac{1}{2} \frac{\min(1, c_3)}{(\gamma + \beta c_2)} \|q_h\|_{0/\mathbb{R}}. \; \square \tag{6.14}$$

A similar proof can be given for quadrilateral elements but this result can be obtained in a simple way using the macroelement technique of the previous sections. It is also worth noting that adding a bubble function to the previous Taylor–Hood element makes stability quite straightforward. Our numerical experience has been that the bubble function does improve the accuracy of the element.

Remark 6.1: Another element that has been used because of the simplicity of its shape functions is the so-called "P_1 iso P_2" element that is sketched in Figure VI.28. It is a composite element assembled from four piecewise linear elements for velocity while pressure remains linear on the macroelement. The above analysis can extended to this case showing that the inf–sup condition holds. One could use this subspace of a $P_1 - P_1$ equal approximation method in the context of the stabilization procedure described in Section VI.5.5.

Remark 6.2: The technique employed in Proposition 6.1 is quite general. It has been extended by BREZZI–FALK [A] to show a convergence result for the $P_3 - P_2$ element which is an obvious generalization of the Taylor–Hood element studied above. They also show that the essential ingredients, namely, that the space of velocities can stand a piecewise constant pressure and a clever use of a quadrature rule, can be used to prove stability of the $Q_k - Q_{k-1}$ rectangular elements with continuous pressures for every $k \geq 2$. They also show that $Q_k - Q_{k-1}$ elements with discontinuous pressure exhibit spurious pressure modes on a regular mesh for every $k \geq 1$ which shows that a general stability proof cannot be obtained. □

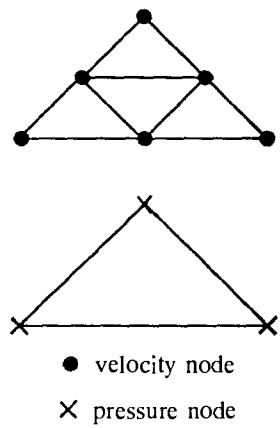


Figure VI.28: The "P_1 iso P_2" element

□

In some cases, the technique used in the proof of Proposition 6.1 does not apply directly, but a variant of it, based on Verfürth's trick (VERFÜRTH [A]), can be employed. We shall illustrate this by proving the stability of the three-dimensional Hood–Taylor element. The original proof by STENBERG [D] was done through the macroelement technique.

Verfürth's trick is essentially based on two steps. The first step is quite general and can be summarized in the following lemma.

Lemma 6.1: Let Ω be a bounded domain in $\mathbb{R}^N$ with Lipschitz continuous boundary. Let $V_h \subset (H_0^1(\Omega))^2 = V$ and $Q_h \subset H^1(\Omega)$ be closed subspaces. Assume that there exists a linear operator Π_h^0 from V into V_h and a constant c (independent of h) such that

$$(6.15) \quad \|\underline{v}_h - \Pi_h^0\, \underline{v}\|_{r,\Omega} \leq c \sum_K \left(h_K^{2-2r}\, \|\underline{v}\|_{1,K}^2 \right)^{1/2}, \qquad \forall \underline{v} \in V,\ r = 0, 1,$$

where $\Omega = \cup K$ and h_K is the diameter of K. Then there exist two positive constants c_1 and c_2 such that, for every $q_h \in Q_h$

$$\sup_{\underline{v}_h \in V_h} \frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} \geq c_1 \|q_h\|_{0/\mathbb{R}} - c_2 \sum_K (h_K^2 \, |\underline{\operatorname{grad}}\, q_h|_{0,K}^2)^{1/2}. \tag{6.16}$$

Proof. We remark first that the inf–sup condition for the continuous problem,

$$\inf_{q \in L^2(\Omega)/\mathbb{R}} \sup_{\underline{v} \in V} \frac{\int q \operatorname{div} \underline{v} \, dx}{\|\underline{v}\|_1 \, \|q\|_{0/\mathbb{R}}} \geq \beta \tag{6.17}$$

holds in the N-dimensional case (TEMAM [A]). This implies that for every $q_h \in Q_h$ there exists a $\bar{\underline{v}} \in V$ such that

$$\begin{aligned} \sup_{\underline{v}_h \in V_h} \frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} &\geq \frac{\int q_h \operatorname{div} \Pi_h^0 \bar{\underline{v}} \, dx}{\|\Pi_h^0 \bar{\underline{v}}\|_1} \geq \frac{1}{2c} \frac{\int q_h \operatorname{div} \Pi_h^0 \bar{\underline{v}} \, dx}{\|\bar{\underline{v}}\|_1} \\ &= \frac{1}{2c} \frac{\int q_h \operatorname{div} \bar{\underline{v}} \, dx}{\|\bar{\underline{v}}\|_1} + \frac{1}{2c} \frac{\int q_h \operatorname{div} (\Pi_h^0 \bar{\underline{v}} - \bar{\underline{v}}) dx}{\|\bar{\underline{v}}\|_1} \end{aligned} \tag{6.18}$$

from which we get,

$$\begin{aligned} \sup_{\underline{v}_h \in V_h} \frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} &\geq \frac{\beta}{4c} \|q_h\|_{0/\mathbb{R}} + \frac{1}{2c} \frac{\int \underline{\operatorname{grad}}\, q_h . (\Pi_h^0 \bar{\underline{v}} - \bar{\underline{v}}) dx}{\|\bar{\underline{v}}\|_1} \\ &\geq \frac{\beta}{4c} \|q_h\|_{0/\mathbb{R}} - \Big(\tfrac{1}{2} \sum_K h_K^2 |\underline{\operatorname{grad}}\, q_h|_{0,K}^2 \Big)^{1/2}. \quad \square \end{aligned} \tag{6.19}$$

The second step in Verfürth's trick is to prove a kind of inf–sup condition where the zero norm of q_h is substituted by $h|q_h|_1$. This will be done, for the three-dimensional Hood–Taylor element, in the next lemma. To be precise, however, we first have to introduce the element. For this, let Ω be a polyhedron in $\mathbb{R}^3$ and $\mathcal{T}_h$ a decomposition of Ω into tetrahedra K with the usual "nondegeneracy" condition. We assume, moreover, that

(6.20) every $K \in \mathcal{T}_h$ has at least three internal edges.

We now set

$$V_h = (\mathfrak{L}_2^1(\mathcal{T}_h))^3 \cap (H_0^1(\Omega))^3, \tag{6.21}$$

$$Q_h = \mathfrak{L}_1^1(\mathcal{T}_h). \tag{6.22}$$

Lemma 6.2: Under the assumptions (6.20)–(6.22) there exists a positive constant c_3 such that, for every $q_h \in Q_h$,

$$\sup_{\underline{v}_h \in V_h} \frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} \geq c_3 \Big(\sum_K h_K^2 \, |q_h|_{1,K}^2 \Big)^{1/2}. \tag{6.23}$$

Proof: We shall prove (6.23) by a suitable construction of $\underline{v}_h$. Let q_h be given in Q_h and let K be an element of $\mathcal{T}_h$. We define $\underline{v}|_K$ by the following conditions

$$\underline{v}_h = 0 \text{ at the vertices of } K, \tag{6.24}$$

$$\underline{v}_h = -\underline{t}^e (\underline{\text{grad}}\, q_h . \underline{t}^e) |e|^2 \tag{6.25}$$

at the midpoint of every edge e of K, where $|e|$ is the length of e and $\underline{t}^e$ is the unit unit tangent vector to e (the orientation is immaterial, but has obviously to be chosen once and for all in $\mathcal{T}_h$). It is easy to check that

$$\|\underline{v}_h\|_{1,K} \leq c \, h_K \, |q_h|_{1,K}. \tag{6.26}$$

We shall use the following well-known integration formula

$$\int_K p_2(x) \, dx = \Big(\sum_M \frac{p_2(M)}{5} - \sum_V \frac{p_2(V)}{20} \Big) \text{Meas}(K) \tag{6.27}$$

(where M and V vary over midpoints of edges and vertices respectively). for all p_2 polynomial of degree ≤ 2 on K. We have, with the choice (6.24) and (6.25) for $\underline{v}_h$,

$$\begin{aligned} \int_\Omega q_h \operatorname{div} \underline{v}_h \, dx &= - \int_\Omega \underline{\text{grad}}\, q_h \cdot \underline{v}_h \, dx \\ &= - \sum_K \int_K \underline{\text{grad}}\, q_h \cdot \underline{v}_h \, dx \\ &= - \sum_K \sum_M (\underline{\text{grad}}\, q_h \cdot \underline{v}_h)(M) \frac{\text{Meas}(K)}{5} \\ &= \sum_K \sum_M |\frac{\partial q_h}{\partial t}|^2 |e|^2 \frac{\text{Meas}(K)}{5} \\ &\geq C \sum_K h_K^2 \, \|\underline{\text{grad}}\, q_h\|_{0,K}^2, \end{aligned} \tag{6.28}$$

where in the last inequality we used the nondegeneracy condition $|e| \geq \sigma h_K$ and assumption (6.20). From (6.26) and (6.28) we obtain

$$\frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} \geq c_3 \frac{\sum_K h_K^2 \, \|\underline{\text{grad}}\, q_h\|_{0,K}^2}{\Big(\sum_K h_K^2 \, \|\underline{\text{grad}}\, q_h\|_{0,K}^2 \Big)^{1/2}}, \tag{6.29}$$

which is (6.23). □

The last step of Verfürth's trick is then to multiply (6.16) by c_3 and (6.23) by c_2 and sum. We have

$$(c_3 + c_2) \sup_{\underline{v}_h \in V_h} \frac{\int q_h \operatorname{div} \underline{v}_h \, dx}{\|\underline{v}_h\|_1} \geq c_1 c_3 \, \|q_h\|_{0/\mathbb{R}}$$

that is, the inf–sup condition. □

Remark 6.3: The above proof could also be applied to the two-dimensional case of Proposition 6.1. We presented both to make explicit the use of two different techniques. □

VI.7 Nearly Incompressible Elasticity, Reduced Integration Methods, and Relation with Penalty Methods

VI.7.1 Variational formulations and admissible discretizations

We have already seen in Chapter I that there are problems associated with approximations of nearly incompressible materials when using the standard variational principle. Consider, to make things simple, a problem with homogeneous Dirichlet conditions

$$\inf_{\underline{v} \in (H_0^1(\Omega))^2} \mu \int_\Omega |\underline{\varepsilon}(\underline{v})|^2 \, dx + \frac{\lambda}{2} \int_\Omega |\operatorname{div} \underline{v}|^2 \, dx - \int_\Omega \underline{f} \cdot \underline{v} \, dx. \tag{7.1}$$

We already noted in Section VI.1 that this problem is closely related to a penalty method to solve the Stokes problem.

It was soon recognized in practice that a brute force use of (7.1) could lead, for large values of λ, to bad results, the limiting case being the locking phenomenon that is an identically zero solution. A cure was found in using a reduced (that is inexact) numerical quadrature when evaluating the term $\lambda \int_\Omega |\operatorname{div} \underline{v}|^2 \, dx$ associated with compressibility effects. We refer the reader to the papers of HUGHES–MALKUS [A] and BERCOVIER [B] for a discussion of the long history of this idea. We shall rather develop in detail on this example the relations between reduced integrations and mixed methods and try to make clear to what extent they may be claimed to be equivalent. For this we first recall from Chapter I, that problem (7.1) can be transformed by a straightforward application of duality techniques into a saddle point problem

$$\inf_{\underline{v}} \sup_{q} \mu \int_\Omega |\underline{\varepsilon}(\underline{v})|^2 \, dx - \frac{2}{2\lambda} \int_\Omega |q|^2 \, dx + \int_\Omega q \operatorname{div} \underline{v} \, dx - \int_\Omega \underline{f} \cdot \underline{v} \, dx \tag{7.2}$$

for which optimality conditions are, denoting by $(\underline{u}, p)$ the saddle point,

$$(7.3)\quad \mu \int_\Omega \underline{\underline{\varepsilon}}(\underline{u}) : \underline{\underline{\varepsilon}}(\underline{v})\, dx + \int_\Omega p \operatorname{div} \underline{v}\, dx = \int_\Omega \underline{f} \cdot \underline{v}\, dx, \quad \forall \underline{v} \in (H_0^1(\Omega))^2,$$

$$(7.4)\qquad \int_\Omega \operatorname{div} \underline{u}\, q\, dx = \frac{1}{\lambda} \int_\Omega pq\, dx, \quad \forall q \in L^2(\Omega).$$

This is obviously very close to a Stokes problem and is also an **example of the** problem studied in Chapter II, that is, find $u \in V$, $p \in Q$ such that

$$(7.5)\qquad a(u,v) + b(v,p) = (f,v), \qquad \forall v \in V,$$

$$(7.6)\qquad b(u.q) - c(p,q) = (g,q), \qquad \forall q \in Q.$$

We then know from Chapter II that an approximation of (7.3) and (7.4) (that is, a choice of an approximation for both $\underline{u}$ and p), leading to error estimates independent of λ, must be a good approximation for the Stokes problem. The preceding sections of this chapter, therefore, give us a good idea of what should (or should not) be used as an approximation. What we shall now see is that reduced integration methods correspond to *an implicit choice* of a mixed approximation. The success of the reduced integration method will thus rely on the qualities of this underlying mixed method.

VI.7.2 Reduced integration methods

Let us consider a (more or less) standard approximation of the original problem (7.1). An exact evaluation of the "penalty term" $\lambda \int_\Omega |\operatorname{div} \underline{v}|^2 dx$ means that for λ large one tries to get an approximation of $\underline{u}$ which is *exactly* divergence-free. But as we have already seen few finite elements can stand such a condition that will in most cases lead to locking phenomenon due to overconstraining. In a mixed formulation one relaxes the incompressibility condition by the choice of the approximation for p. Let us now see how this will be translated as a reduced integration method at least in some cases. Let us then consider $V_h \subset V = (H_0^1(\Omega))^2$, $Q_h \subset Q = L^2(\Omega)$, these approximation spaces being built from finite elements defined on a partition of Ω. On each element K, let there be given a set of k points x_i and weights ω_i defining a numerical quadrature formula (Figure VI.29),

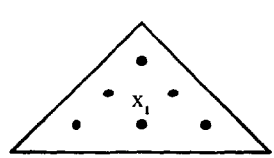


Figure VI.29

$$\int_K f(x)\,dx = \sum_{i=1}^{k} \omega_i\, f(x_i). \tag{7.7}$$

Remark 7.1: It will be convenient to define the numerical quadrature on a reference element $\hat{K}$ and to evaluate integrals by a change of variables.

$$\int_K f(x)\,dx = \int_{\hat{K}} f(\hat{x})\,J(\hat{x})\,d\hat{x} = \sum_{i=1}^{k} \omega_i\, f(\hat{x}_i)\,J(\hat{x}_i) \tag{7.8}$$

The presence of the Jacobian $J(x)$ should be taken into account when discussing the precision of the quadrature rule on K. □

Let us now make the hypothesis that for $\underline{v}_h \in V_h$ and $p_h,\ q_h \in Q_h$, one has exactly

$$\int_K q_h \operatorname{div} \underline{v}_h\,dx = \sum_{i=1}^{k} \omega_i\, \hat{q}_h(\hat{x}_i)\widehat{\operatorname{div}\,\underline{v}_h}(\hat{x}_i)\,J(\hat{x}_i) \tag{7.9}$$

and

$$\int_K p_h\, q_h\,dx = \sum_{i=1}^{k} \omega_i\, \hat{p}_h(\hat{x}_i)\,\hat{q}_h(\hat{x}_i)\,J(\hat{x}_i). \tag{7.10}$$

Let us now consider the discrete form of (7.4)

$$\int_\Omega \operatorname{div} \underline{u}_h\, q_h\,dx = \frac{1}{\lambda}\int_\Omega p_h\, q_h\,dx, \qquad \forall q_h \in Q_h. \tag{7.11}$$

When the space Q_h is built from discontinuous functions, this can be read element by element,

$$\int_K q_h \operatorname{div} \underline{u}_h\,dx = \frac{1}{\lambda}\int_K p_h\, q_h\,dx, \qquad \forall q_h \in Q_h, \tag{7.12}$$

so that by using (7.9) and (7.10) one gets

$$\hat{p}_h(\hat{x}_i) = \lambda\widehat{\operatorname{div}\,\underline{u}_h}(\hat{x}_i) \text{ or } p_h(x_i) = \lambda \operatorname{div} u_h(x_i), \tag{7.13}$$

provided the values of q_h at the quadrature points can be used as degrees of

are unisolvent. Formula (7.8) can, in turn, be used in the discrete form of (7.3) which now gives

$$(7.14)\qquad \begin{cases} 2\mu \displaystyle\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_h)\cdot\underline{\underline{\varepsilon}}(\underline{v}_h)\,dx + \lambda \sum_K \Big(\sum_{i=1}^{k} \omega_i\, J(\hat{x}_i)(\widehat{\operatorname{div}\underline{u}_h}(\hat{x}_i)(\widehat{\operatorname{div}\underline{v}_h}(\hat{x}_i))\Big) \\ \qquad\qquad = \displaystyle\int_\Omega \underline{f}\cdot\underline{v}_h\,dx \end{cases}$$

In general the term $\sum_K \left(\sum_{i=1}^{k} \omega_i\ J(\hat{x}_i)(\widehat{\operatorname{div}\underline{u}_h}(\hat{x}_i)(\widehat{\operatorname{div}\underline{v}_h}(\hat{x}_i))\right)$ *is not* an exact evaluation of $\int_\Omega \operatorname{div}\underline{u}_h \operatorname{div}\underline{v}_h\ dx$ and reduced integration is effectively introduced. In the case where (7.9) and (7.10) hold there is a perfect equivalence between the mixed method and the use of reduced integration. Whatever will come from one can be reduced to the other one. It will, however, not be general in possible to get equalities (7.9) and (7.10) so that a further analysis will be needed. But we shall first consider some examples of this complete equivalence case.

Example 7.1: Let us consider the $Q_1 - P_0$ approximation on a *rectangle* and a one-point quadrature rule. It is clear that $\operatorname{div}\underline{u}_h \in P_1(K)$ and is integrated exactly. In the same way a one-point rule is exact for $\int_\Omega p_h\ q_h\ dx$ whenever $p_h,\ q_h \in P_0(K)$. There is thus a perfect equivalence between reduced integration and the exact penalty method defined by (7.11). □

Example 7.2: We now consider again the same $Q_1 - P_0$ element on a general quadrilateral (Figure VI.30). To show that we still have equivalence requires a somewhat more delicate analysis. Indeed, at first sight the quadrature rule is not exact for $\int_{\hat{K}} \widehat{\operatorname{div}\underline{u}_h}\ J_K(\hat{x})\ d\hat{x}$.

Let us however consider in detail the term $\widehat{\operatorname{div}\underline{u}_h} = \widehat{\partial u_1/\partial x_1} + \widehat{\partial u_2/\partial x_2}$. Let $B = DF$ be the Jacobian matrix of the transformation F from $\hat{K}$ into K. Writing explicitly

$$(7.15)\qquad F = \begin{cases} a_0 + a_1\hat{x} + a_2\hat{y} + a_3\hat{x}\hat{y} \\ b_0 + b_1\hat{x} + a_2\hat{y} + b_2\hat{x}\hat{y} \end{cases}$$

one has

$$(7.16)\qquad B = \begin{pmatrix} a_1 + a_3\hat{y} & b_1 + b_3\hat{y} \end{pmatrix}$$

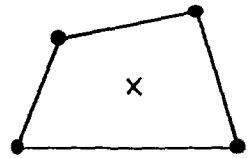


Figure VI.30

so that we get

$$B^{-1} = \frac{1}{J(\hat{x})}\begin{pmatrix} b_2 + b_3\hat{x} & -b_1 - b_3\hat{y} \\ -a_2 - a_3\hat{x} & a_1 + a_3\hat{y} \end{pmatrix}. \tag{7.17}$$

But

$$\widehat{\frac{\partial u_1}{\partial x_1}} = \Big(\frac{\partial \hat{u}_1}{\partial \hat{x}_1}(b_2 + b_3\hat{x}) - \frac{\partial \hat{u}_1}{\partial \hat{x}_2}(b_1 - b_3\hat{y})\Big)\frac{1}{J(\hat{x})},$$

$$\widehat{\frac{\partial u_2}{\partial x_2}} = \Big(\frac{\partial \hat{u}_2}{\partial \hat{x}_1}(-a_2 - a_3\hat{x}) + \frac{\partial \hat{u}_2}{\partial \hat{x}_2}(a_1 + a_3\hat{y})\Big)\frac{1}{J(\hat{x})}.$$

When computing $\int_{\hat{K}} \widehat{\operatorname{div} \underline{u}_h}\, J(\hat{x})\, d\hat{x}$, Jacobians cancel and one is left with the integral of a function that is linear in each variable and that can be computed exactly by a one-point formula. □

Example 7.3: Using a four-point integration formula on a straight-sided quadrilateral can be seen as in the previous example to be exactly equivalent to a $Q_2 - Q_1$ approximation (BERCOVIER [A,B]).

The above equivalence is however not the general rule. Consider the following examples.

Example 7.4: We want to use a reduced integration procedure to emulate the Crouzeix–Raviart $P_2 - P_1$ element (cf. Section VI.3). To define a P_1 pressure, we need three integration points (Figure VI.31) which can generate a formula that will be exact for second degree polynomials (but not more). The bubble function included in velocity however makes $\operatorname{div} \underline{u}_h \in P_2(K)$ and $\int_K \operatorname{div} \underline{u}_h\, q_h dx$ will not be evaluated exactly. □

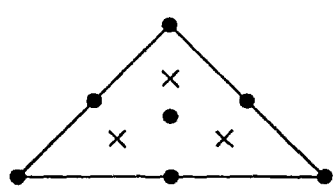


Figure VI.31

Example 7.5: A full iso-parametric $Q_2 - Q_1$ element is not equivalent to its four-point reduced integration analogue. □

Example 7.6: A $Q_2 - P_0$ approximation is not, even on rectangles, equivalent to a one-point reduced integration method, since div $\underline{u}_h$ contains second-order terms which are not taken into account by a one-point quadrature. □

VI.7.3 Effects of inexact integration

If we now consider into more detail the cases where a perfect equivalence does not hold between the mixed method and some reduced integration procedure we find ourselves in the setting of Section II.2.6. In particular, $b(\underline{v}_h, q_h)$ is replaced by an approximate bilinear form $b_h(\underline{v}_h, q_h)$. We shall suppose, for the sake of simplicity, that the scalar product on Q_h is exactly evaluated. Two questions must then be answered.

— Does $b_h(.,.)$ satisfy the inf–sup condition ?

— Do error estimates still hold without loss of accuracy ?

We have already introduced in Section II.2.6 a general setting in which this situation can be analyzed. We shall first apply Proposition II.2.19 to the verification of the inf–sup condition for two examples and give an example where inexact integral changes the nature of the problem. We shall then consider consistency error on those three examples.

Example 7.7: We in fact come back to Example 7.6 and study, on a rectangular mesh, the $Q_2 - P_0$ approximation with a one-point quadrature rule. This is not, as we have said, equivalent to the standard $Q_2 - P_0$ approximation. We now want to check using Proposition II.2.19, that it satisfies the inf–sup condition. We, thus, have to build a continuous operator (in $H^1(\Omega)$-norm) such that

$$\int_\Omega \operatorname{div} \underline{u}_h \, q_h \, dx = \sum_K [(\operatorname{div} \Pi_h \underline{u}_h)(M_{0,K}) q_K] \operatorname{area}(K), \tag{7.18}$$

where $M_{0,K}$ is the barycenter of K and q_K the restriction of q_h to K. We can restrict our analysis to one element as q_h is discontinuous and we study both sides of equality (7.18). We have, taking $q_K = 1$,

(7.19) $\int$ [illegible] dx = $\int$ [illegible] dx

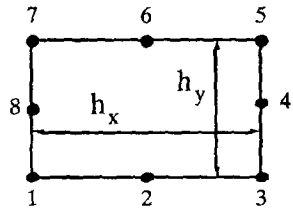


Figure VI.32

Using the numbering of Figure VI.32 and denoting by u_i and v_i the horizontal and vertical components of velocity at node i, we can write (7.19), by Simpson's quadrature rule, in the form

$$\int_K \operatorname{div} \underline{u}_h \, dx = \frac{h_y}{6}[u_5 + 4u_4 + u_3] - \frac{h_y}{6}[u_1 + 4u_8 + u_7] + \frac{h_x}{6}[v_7 + 4v_6 + v_5] - \frac{h_x}{6}[v_1 + 4v_2 + v_3]. \tag{7.20}$$

If we write

$$u_4 = \frac{u_5 + u_3}{2} + \hat{u}_4, \quad u_8 = \frac{u_1 + u_7}{2} + \hat{u}_8,$$
$$v_6 = \frac{v_5 + v_7}{2} + \hat{v}_6, \quad v_2 = \frac{v_1 + v_3}{2} + \hat{v}_2,$$

where $\hat{u}_4, \hat{u}_6, \hat{v}_6$, and $\hat{v}_2$ are corrections with respect to a bilinear interpolation, we may rewrite (7.20) as

$$\int_K \operatorname{div} \underline{u}_h \, dx = \frac{h_y}{2}[u_5 + u_3 + \frac{4}{3}\hat{u}_4] - \frac{h_y}{2}[u_1 + u_7 + \frac{4}{3}\hat{u}_8] + \frac{h_y}{2}[v_7 + v_5 + \frac{4}{3}\hat{v}_6] - \frac{h_x}{2}[v_1 + v_3 + \frac{4}{3}\hat{v}_2]. \tag{7.21}$$

On the other hand, area (K) div $\underline{u}_h(M_{0,K})$ can be seen to be equal to

$$\frac{h_y}{2}[u_5 + u_3 + 2\hat{u}_4] - \frac{h_y}{2}[u_1 + u_7 + 2\hat{u}_8] - \frac{h_x}{2}[u_7 + v_5 + 2\hat{v}_6] - \frac{h_x}{2}[v_1 + v_3 + 2\hat{u}_2]. \tag{7.22}$$

If we split $\underline{u}_h$ into a bilinear part $\underline{u}_h^0$ and a mid-point correction part $\hat{\underline{u}}_h$, one can define $\Pi_h \underline{u}_h$ by setting

$$\begin{cases} (\Pi_h \underline{u}_h)^0 = \underline{u}_h^0, \\ (\widehat{\Pi_h \underline{u}_h}) = \frac{2}{3}\hat{\underline{u}}_h. \end{cases} \tag{7.23}$$

Equality (7.19) will then hold and (7.23) is clearly continuous with a continuity constant independent of h. □

Example 7.8: We come back to Example 7.4, that is, a three-point quadrature rule used in conjunction with the Crouzeix–Raviart element. We shall not give the analysis in detail but only sketch the ideas. The problem again is to check that the inf–sup condition holds through Proposition II.2.19. As the quadrature rule *is exact* when q_h is *piecewise constant*, the obvious idea is to build $\Pi_h \underline{u}_h$ by leaving invariant the trace of $\underline{u}_h$ on ∂K and *only modifying the coefficients of the bubble functions*. This can clearly be done. Continuity is now to be checked and the proof is essentially the same as the standard proof of the inf–sup condition (Section VI.3). □

Example 7.9: *A modified $Q_1 - P_0$ element.*

We now present a puzzling example (BREZZI–MARINI [A]) of an element which is stable but for which convergence is tricky due to a consistency error term. We have here a case where using a one-point quadrature rule will change the situation with respect to the inf–sup condition. In fact, it will make a stable element from an unstable one but will also introduce an essential change in the problem. The departure point is, thus, the standard $Q_1 - P_0$ element that was studied in Section VI.5 and that, as we know, does not satisfy the inf–sup condition. We now make it richer by adding to velocity $\underline{u}_h|_K = \{u_1, u_2\}$ what we shall call wave functions. On the reference element $\hat{K} =]-1,1[\times]-1,1[$, those functions are defined by

$$\begin{cases} w_1 = \hat{x}\, b_2(\hat{x},\hat{y}), \\ w_2 = \hat{y}\, b_2(\hat{x},\hat{y}), \end{cases} \tag{7.24}$$

where $b_2(\hat{x},\hat{y}) = (1-\hat{x}^2)(1-\hat{y}^2)$ is the Q_2 bubble function. If we now consider

$$\hat{\underline{u}}_h|_K = \{u_1 + \alpha_K w_1, u_2 + \alpha_K w_2\} = \underline{u}_h|_K + \alpha_K \underline{w}_K, \tag{7.25}$$

we obtain a new element with an internal degree of freedom. The wave functions that we added vanish on the boundary and nothing is changed for the stability of the mixed method with exact integration. If we use a one-point quadrature rule, things become different. We shall, indeed, check that the modified bilinear form $b_h(\hat{\underline{v}}_h, q_h)$ satisfies the inf–sup condition. Thus, we have to show that

$$\sup_{\hat{\underline{v}}_h} \frac{\sum_K \operatorname{div} \hat{\underline{u}}_h(M_{0,K}) p_K\, h_K^2}{\|\hat{\underline{u}}_h\|_1} \geq k_0\, \|p_h\|_0. \tag{7.26}$$

This is easily checked by posing on K (we suppose a rectangular mesh to simplify)

$$\hat{\underline{u}}_h|_K = h_K\, p_K\, \underline{w}_K. \tag{7.27}$$

We then have $\operatorname{div} \hat{\underline{u}}_h = 4p_h$ at the integration points, and

$$\|\hat{\underline{u}}_h\|_{1,K} = h_K\, p_K\, \|\underline{w}_K\|_{1,K}, \tag{7.28}$$

which implies

$$\|u_h\|_1 \leq c\,\|p_h\|_0, \tag{7.29}$$

and (7.26) follows. A remarkable point here is that even the hydrostatic mode has disappeared. This is an indication that something incorrect has been introduced in the approximation. An analysis of *consistency error* indeed shows that usual error estimates fail and that we are actually approximating a continous problem in which the incompressibility condition has been replaced by $\operatorname{div} \underline{u} + kp = 0$, where $k = 1575/416$ (UGLIETTI [A]). We then see that if in general for the Stokes problem, making the space of velocities richer improves (at least does not reduce) the quality of the method, this fact can become false when numerical integration is used. □

Let us now turn our attention to the problem of error estimation. From Propositions II.2.16 and (II.2.74) and(II.2.75), all we have to do is to estimate the consistency terms

$$\sup_{\underline{v}_h} \frac{|b(\underline{v}_h, p) - b_h(\underline{v}_h, p)|}{\|\underline{v}_h\|_V} \tag{7.30}$$

and

$$\sup_{q_h} \frac{|b(\underline{u}, q_h) - b_h(\underline{u}, q_h)|}{\|q_h\|_{0/\operatorname{Ker} B'}}. \tag{7.31}$$

We thus have to estimate quadrature errors. It is not our intent to enter here into detail and we refer the reader to CIARLET [B] where examples of such analysis are presented exhaustively. The first step is to transform (7.30) into a form which is sometimes more tractable. We may indeed write

$$\begin{aligned} b(\underline{v}_h, p) &- b_h(\underline{v}_h, p) \\ &= (b(\underline{v}_h, p - q_h) - b_h(\underline{v}_h, p - q_h)) + (b(\underline{v}_h, q_h) - b_h(\underline{v}_h, q_h)) \end{aligned} \tag{7.32}$$

and

$$\begin{aligned} b(\underline{u}, q_h) &- b_h(\underline{u}, q_h) \\ &= (b(\underline{u} - \underline{v}_h, q_h) - b_h(\underline{u} - \underline{v}_h, q_h)) + (b(v_h, q_h) - b_h(v_h, q_h)). \end{aligned} \tag{7.33}$$

The first parenthesis in the right-hand sides of (7.32) and (7.33) can be reduced to an approximation error. The second parentheses imply only polynomials.

Let us therefore consider (7.33) for the three approximations introduced above. For the Crouzeix–Raviart triangle taking $\underline{v}_h$ the standard interpolate of $\underline{u}$ makes the second parentheses vanish whereas the first yields an $O(h)$ estimate. For the two other approximations taking $\underline{v}_h$ to be a standard bilinear approximation of $\underline{u}$ makes the second parenthesis vanish, whereas the first yields on $O(h)$ estimate, which is the best that we can hope, anyway. The real trouble is, therefore, with (7.30) with or without (7.32). In the case of the Crouzeix–Raviart triangle, we can use directly (7.30) and the following result of CIARLET [B]

Proposition 7.1: Let $f \in W_{k,q}(\Omega)$, $p_k \in P_k(K)$ and denote by $E_k(fp_k)$ the quadrature error on element K when numerical integration is applied to fp_k. Let us suppose that $E_K(\hat{\phi}) = 0$, $\forall \hat{\phi} \in P_{2k-2}(K)$, then one has for $k - q/n > 0$

$$|E_K(fp_k)| \leq ch_K^k \, (\text{meas}(K))^{1/2-1/q} |f|_{k,q,K} |p_k|_{1,K} \quad \square \tag{7.34}$$

Taking $k = 2$, $q = \infty$ and using the inverse inequality to go from $|p_k|_1$ to $|p_k|_0$ one gets an $O(h^2)$ estimate for (7.30).

The two other appproximations cannot be reduced to Proposition 7.1 and must be studied through (7.32). We must study a term like

$$\sup_{\underline{v}_h} \frac{|b(\underline{v}_h, q_h) - b_h(\underline{v}_h, q_h)|}{\|\underline{v}_h\|_1} \tag{7.35}$$

This can at best be *bounded.* For instance, in the case of the $Q_2 - P_0$ approximation we can check by hand that the quadrature error on K reduces to $h_K^3 \, |\text{div } \underline{v}_h|_{2,K} \, p_K$.

VI.8 Divergence-Free Basis, Discrete Stream Functions

We have dealt in this chapter with the mixed formulation of the Stokes problem and we have built finite element approximations in which discrete divergence-free functions approximate the continuous ones. It is sometimes useful to consider directly the constrained minimization problem

$$\inf_{\underline{v}_0 \in V_0} \frac{1}{2} \int_\Omega |\underline{\underline{\varepsilon}}(\underline{v}_0)|^2 \, dx - \int_\Omega \underline{f} \cdot \underline{v}_0 \, dx, \tag{8.1}$$

where V_0 is the subspace of divergence-free functions. In this subspace we have a standard minimization problem and the discrete form would lead to a positive definite linear system. Indeed, the solution $\underline{u}_0 \in V_0$ of problem (8.1) satisfies the variational equation

$$\int_\Omega \underline{\underline{\varepsilon}}(\underline{u}_0) : \underline{\underline{\varepsilon}}(\underline{v}_0) \, dx = \int_\Omega \underline{f} \cdot \underline{v}_0 \, dx, \qquad \forall \underline{v}_0 \in V_0. \tag{8.2}$$

In the discrete problem, if one knows a basis $\{\underline{w}_0, \ldots, \underline{w}_m\}$ of V_{0h}, the solution is reduced to the solution of the linear system

$$A_0 \, U_0 = F_0, \tag{8.3}$$

where

$$a_{ij}^0 = \int_\Omega \underline{\underline{\varepsilon}}(\underline{w}_i^0) : \underline{\underline{\varepsilon}}(\underline{w}_j^0) \, dx, \; f_i^0 = \int_\Omega \underline{f} \cdot \underline{w}_i^0 \, dx$$

and

$$A_0 = \{a_{ij}^0\}, \ F_0 = \{f_i^0\}.$$

Building a basis for the divergence-free subspace could therefore lead to a neat reduction of computational costs: pressure is eliminated, along with a certain amount of velocity degrees of freedom. System (8.3) is smaller than the original one. It must however be noted that with respect to the condition number, (8.3) is behaving like a fourth order problem, (VERFÜRTH [C]) which makes its practical usefulness often dubious. As to pressure, it can be recovered a posteriori (CAUSSIGNAC [A,B]).

The construction of such a basis is not, however, a very popular method and is considered as a hard task although it has been numerically implemented (GRIFFITHS [B], THOMASSET [A], HECHT [A]).

As we shall see the two-dimensional case is quite readily handled in many cases. The degrees of freedom can be associated with those of a discrete stream function. The three-dimensional problem is harder to handle: a generating system can often easily be found but the construction of a basis requires the elimination of some degrees of freedom in a not so obvious way.

It is also possible to define a numerical procedure, related to static condensation (FORTIN–FORTIN [A]) for the construction of a partly divergence-free basis.

Finally, we want to emphasize that the construction that we describe will make sense only if the finite element approximation is good so that the previous analysis is still necessary even if it might seem to be bypassed.

We first consider a simple example of a divergence-free basis.

Example 8.1: *The nonconforming $P_1 - P_0$ element.*

We consider the classical non-conforming element introduced in CROUZEIX–RAVIART [A] (cf. Section VI.3) in which mid-side nodes are used as degrees of freedom for velocity. This generates a piecewise linear nonconforming approximation; pressure is taken constant on each element (Figure VI.33). The restriction to an element K of $\underline{u}_h \in V_h$ is then exactly divergence-free and is, therefore, locally the curl of a quadratic polynomial. This discrete stream function cannot be continuous on interfaces but must have continuous derivatives at mid-side points: it can be built from Morley's triangle (cf. Example III.2.5). The degrees of freedom of the divergence-free basis can be associated to the degrees of freedom of this nonconforming stream function (Figure VI.34). This assigns a basis function to each vertex and to each mid-side node. They are depicted schematically in Figure VI.35. One observes a general pattern: divergence-free functions are made from small vortices. □

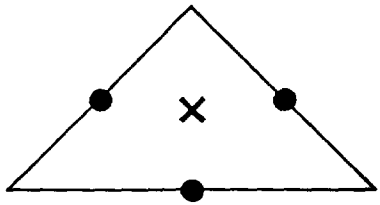


Figure VI.33

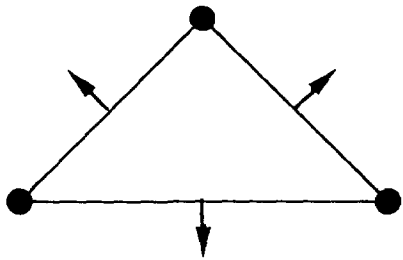


Figure VI.34

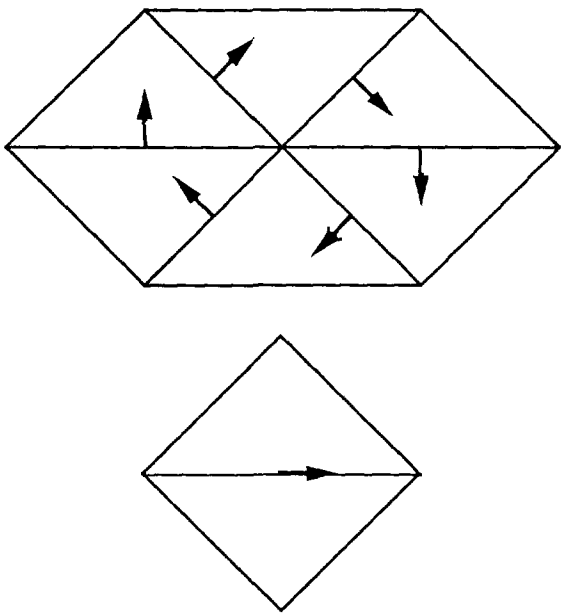


Figure VI.35: Basis functions for a divergence free $P_1 - P_0$ non-conforming element

Remark 8.2: The kind of basis obtained in the previous example is typical of a domain without holes with homogeneous Dirichlet boundary conditions. Whenever a hole is present, an extra basis function must be added in order to ensure circulation around the hole (Figure VI.36). This function is not local.

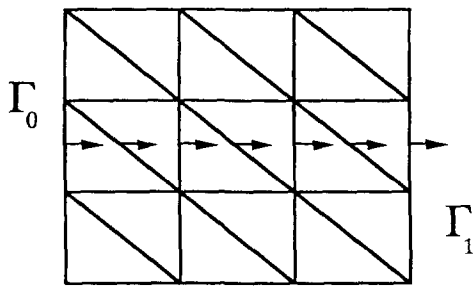


Figure VI.36

In the same way when the flow is entering on a part Γ_0 of $\partial\Omega$ and outgoing on a part Γ_1, a basis function must be provided to link those parts and to take into account the potential part of the flow (Figure VI.37). □

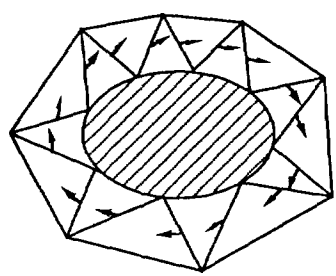


Figure VI.37

□

We now consider a conforming approximation, namely, the popular $Q_2 - P_1$ element.

Example 8.3: *The conforming $Q_2 - P_1$ element.*

We shall sketch in this example the construction of a divergence-free basis for the $Q_2 - P_1$ element. To make things simple we shall assume that the mesh is formed of 2×2 macroelements. The general case can easily be deduced. Let us first look for divergence-free (in the discrete sense, of course) functions with their support on a macroelement. We have 18 degrees of freedom for velocity (Figure VI.38) linked by $(12 - 1) = 11$ linear constraints. This leaves

7 linearly independent functions which can be described by the diagrams of Figure VI.39.Three of them are associated with the center and one to each mid-side node. It must be noted that internal nodes are no longer degrees of freedom. □

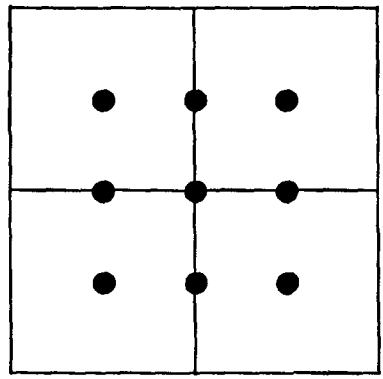


Figure VI.38

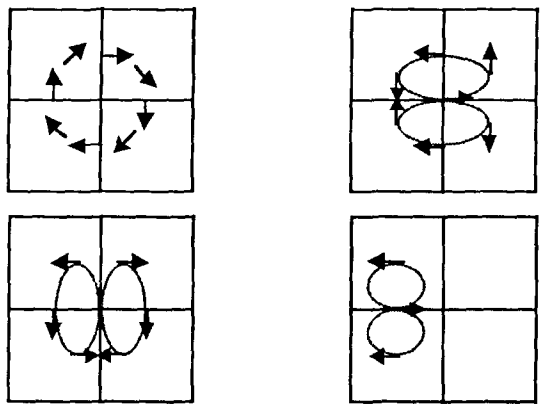


Figure VI.39

Remark 8.4: The "divergence-free" functions described above cannot be taken as the curl of a stream function as they are not exactly divergence-free. However, a discrete stream function ψ_h can nevertheless be built. Its trace on ∂K can be totally determined by integrating $\underline{u}_h \cdot \underline{n}$ along the boundary As the flow is conserved at element level this defines $\psi_h|_{\partial K}$ that is a piecewise third degree polynomial such that $\partial\psi_h/\partial\tau = \underline{u}_h \cdot \underline{n}$. This stream-function could be built from the element of Figure VI.40 (close to Adini's element) but $\underline{u}_h$ must be deduced by taking a *discrete curl operation.*

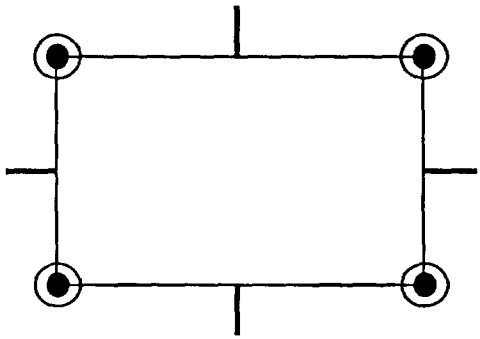


Figure VI.40

Other methods which have been studied for elasticity problems can be extended to the Stokes problem. For instance, the Hellan–Hermann–Johnson mixed method for plate bending that will be described in Chapter VII has been extended to the $\psi-\omega$ formulation for Stokes by BREZZI–LE TELLIER–OLIER [A]. □

VI.9 Other Mixed and Hybrid Methods for Incompressible Flows

We have considered in this chapter only the most standard applications to the Stokes problem using primitive variables. This is not, by far, the only possibility; we already considered in Chapter IV the $\psi - \omega$ decomposition of the biharmonic problem. This can clearly be applied to the Stokes problem. Indeed any divergence-free functions $\underline{u} \in (H_0^1(H))^2$ can be written in the form

$$\underline{u} = \underline{\text{curl}}\ \psi, \quad \psi \in H_0^2(\Omega). \tag{9.1}$$

From (9.1) we get

$$\text{curl}\ \underline{u} = \omega = -\Delta\psi. \tag{9.2}$$

On the other hand, taking the curl of (1.1) gives

$$-\Delta\omega = \text{curl}\ \underline{f} = f_1. \tag{9.3}$$

This procedure can be extended to the Navier–Stokes equation (indeed in many ways) including, if wanted, some upwinding procedure for the non linear tems (FORTIN–THOMASSET [A], JOHNSON [A]). The reader will find a fairly complete study of such procedures in GLOWINSKI–PIRONNEAU [A], and PIRONNEAU [B]. It must be noted that the simplest case of such a procedure, using for ψ_h a bilinear approximation, yields as an approximation of $\underline{u}$ the famous MAC scheme (Figure VI.41).

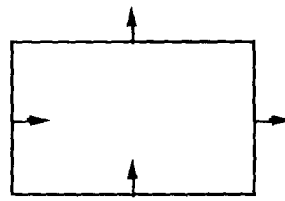


Figure VI.41

Indeed, this is nothing but the space $RT_{[0]}$ for which the subspace of divergence-free functions can be obtained from a bilinear stream function. The Hellan–Hermann–Johnson mixed method for elasticity described in Chapter I can also be applied to the Stokes problem with $\underline{u}_h$ chosen in some approximation of $H(\mathrm{div};\Omega)$. A direct approach precludes to use a symmetric tensor and forces to use $\underline{\underline{\mathrm{grad}}}\ \underline{u}$ instead of $\underline{\underline{\varepsilon}}(\underline{u})$ as a dual variable (ARNOLD–FALK [A]). This difficulty has been circumvented by MGHAZLI [A] by enriching the spaces by the trick of AMARA–THOMAS [A] or ARNOLD–BREZZI–DOUGLAS [A] or BREZZI–DOUGLAS–MARINI [B].

Finally it must be said that dual hybrid methods have been applied by ATLURI–YANG [A] to the Stokes problem. As in Section IV.1.5 (but in a simpler setting), this generates elements that are defined only by the traces at the boundaries and for which internal values can be chosen arbitrarily. This can be seen as the ultimate case of enrichment by bubble functions: enriching by a (potentially infinite dimensional) space enables to use exactly divergence-free function, provided the inf–sup condition is satisfied for *piecewise constant pressure.*

VII

Other Applications

In this chapter we shall present a few among the many other applications of mixed methods. In the first section we shall describe a mixed method for linear thin plates theory, in the second section we shall discuss some applications of mixed methods to linear elasticity with a particular stress on the nearly incompressible case, and in the third section we shall report some recent results on the discretization of the Mindlin–Reissner formulation for moderately thick plates.

VII.1 Mixed Methods for Linear Thin Plates

Let us go back to the variational formulation of the problem discussed in Chapter I and let us recall it here for the convenience of the reader. We had

$$L(\underline{\underline{\sigma}}, w) = \inf_{\underline{\underline{\tau}} \in L^2(\Omega))^4_s} \sup_{\phi \in H_0^2(\Omega)} L(\underline{\underline{\tau}}, \phi), \tag{1.1}$$

where

$$\begin{aligned} L(\underline{\underline{\tau}}, \phi) := \frac{1}{2}\left(\frac{12}{Et^3}\right) \int_\Omega [(1+\nu)\underline{\underline{\tau}} \cdot \underline{\underline{\tau}} - \nu(\mathrm{tr}(\underline{\underline{\tau}}))^2]\, dx \\ - \int_\Omega \underline{\underline{\tau}} \cdot \underline{\underline{D}}_2 \phi\, dx + \int_\Omega f\phi\, dx; \end{aligned} \tag{1.2}$$

$$E = \text{ Young's modulus,} \tag{1.3}$$

$$t = \text{ thickness of the plate,} \tag{1.4}$$

$$\nu = \text{ Poisson's ratio,} \tag{1.5}$$

$$f = \text{ transversal load / unit surface,} \tag{1.6}$$

$$w = \text{ transversal displacement,} \tag{1.7}$$

$$\underline{\underline{\sigma}} = \text{ stresses (in the Kirchoff assumption).} \tag{1.8}$$

In order to use a more compact notation we set

$$C\underline{\underline{\tau}} := \frac{1}{2}Et^3\big((1+\nu)\underline{\underline{\tau}} - \nu\,\mathrm{tr}(\underline{\underline{\tau}})\underline{\underline{\delta}}\big) \tag{1.9}$$

and write $L(\underline{\underline{\tau}}, \phi)$ as

$$L(\underline{\underline{\tau}}, \phi) = \tfrac{1}{2}(C\underline{\underline{\tau}}, \underline{\underline{\tau}}) - (\underline{\underline{\tau}}, \underline{\underline{D}}_2\phi) + (f, \phi) \tag{1.10}$$

Assume now that we are given a triangulation $\mathcal{T}_h$ of Ω, and that we are willing to discretize the stress field $\underline{\underline{\sigma}}$ by means of piecewise polynomials for which the normal bending moment

$$M_{nn}(\underline{\underline{\sigma}}) = (\underline{\underline{\sigma}}\,\underline{n})\,\underline{n} \tag{1.11}$$

is continuous from one element to another. We recall the following Green's formulas:

$$\int_K \underline{\underline{\tau}}\;\underline{\underline{D}}_2\phi\,dx = -\int_K \underline{\mathrm{div}}\,\underline{\underline{\tau}}\cdot\underline{\mathrm{grad}}\,\phi\,dx + \int_{\partial K} M_{nn}(\underline{\underline{\tau}})\frac{\partial\phi}{\partial n}ds + \int_{\partial K} M_{nt}(\underline{\underline{\tau}})\frac{\partial\phi}{\partial t}\,ds, \tag{1.12}$$

$$-\int_K \underline{\mathrm{div}}\,\underline{\underline{\tau}}\;\underline{\mathrm{grad}}\,\phi\,dx = \int_K D_2^*(\underline{\underline{\tau}})\,\phi\,dx - \int_{\partial K} Q_n(\underline{\underline{\tau}})\,\phi\,ds \tag{1.13}$$

valid for all $\underline{\underline{\tau}}$ and ϕ smooth in K; we recall again that here $\underline{t}$ is the unit tangent (counterclockwise) vector and

$$M_{nt}(\underline{\underline{\tau}}) = (\underline{\underline{\tau}}\;\;\underline{n})\cdot\underline{t}, \quad Q_n(\underline{\underline{\tau}}) = \underline{\mathrm{div}}(\underline{\underline{\tau}})\;\;\underline{n} \tag{1.14}$$

If now $M_{nn}(\underline{\underline{\tau}})$ is continuous and ϕ is smooth we can write

$$L(\underline{\underline{\tau}}, \phi) = \frac{1}{2}(C\underline{\underline{\tau}}, \underline{\underline{\tau}}) + \sum_K\Big\{\int_K \underline{\mathrm{div}}(\underline{\underline{\tau}})\cdot\underline{\mathrm{grad}}\,\phi\,dx - \int_{\partial K} M_{nt}(\underline{\underline{\tau}})\frac{\partial\phi}{\partial t}ds\Big\} + (f, \phi) \tag{1.15}$$

A little functional analysis shows that every integral in (1.15) makes sense (at least as a suitable duality pairing) provided $\underline{\underline{\tau}}$ and ϕ are, respectively, in the following spaces:

$$V = \{\underline{\underline{\tau}} \mid \underline{\underline{\tau}}|_K \in (H^1(K))^4_s,\ M_{nn}(\underline{\underline{\tau}}) \text{ continuous }\} \tag{1.16}$$

$$Q = W^{1,p}(\Omega), \quad p > 2 \tag{1.17}$$

Remark 1.1: (for mathematicians). We have to choose $p > 2$ in (1.17) because for $\phi \in H^1(K)$ we have $\partial\phi/\partial t \in H^{-1/2}(\partial K)$, whereas $M_{nt}(\underline{\tau})$ is in $\prod_{e_i} H^{1/2}(e_i)$ but *not* in $H^{1/2}(\partial K)$. On the other hand, for $\phi \in W^{1,p}$ we have that $\partial\phi/\partial t \in W^{-1/p,p}(\partial K)$. Since $M_{nt}(\underline{\tau})$ is in $H^s(\partial K)$ for all $s < 1/2$ and since $W^{-1/p,p}(\partial K) \subset H^{-s}(\partial K)$ for $s > 1/p$, the boundary integral which appears in (1.15) can now be interpreted as a duality pairing between $H^{-s}(\partial K)$ and $H^s(\partial K)$ for $1/p < s < 1/2$ (which is possible since $p > 2$). □

The Euler equations of (1.15) can now be written as

$$(1.18)\quad (C\underline{\underline{\sigma}}, \underline{\underline{\tau}}) + \sum_K \Big\{ \int_K \underline{\mathrm{div}}(\underline{\underline{\tau}}) \cdot \underline{\mathrm{grad}}\, w\, dx - \int_{\partial K} M_{nt}(\underline{\underline{\tau}}) \frac{\partial \phi}{\partial t}\, ds \Big\} = 0,\ \forall \underline{\underline{\tau}} \in V,$$

$$(1.19)\quad \sum_K \Big\{ \int_K \underline{\mathrm{div}}(\underline{\underline{\tau}}) \cdot \underline{\mathrm{grad}}\, \phi\, dx - \int_{\partial K} M_{nt}(\underline{\underline{\tau}}) \frac{\partial \phi}{\partial t} ds \Big\} = (-f, \phi),\ \forall \phi \in Q,$$

which has the form (II.2.1) if we set

$$(1.20)\qquad a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) = (C\underline{\underline{\sigma}}, \underline{\underline{\tau}}),$$

$$(1.21)\qquad b(\underline{\underline{\sigma}}, \phi) = \sum_K \Big\{ \int_K \underline{\mathrm{div}}(\underline{\underline{\tau}}) \cdot \underline{\mathrm{grad}}\, \phi\, dx - \int_{\partial K} M_{nt}(\underline{\underline{\tau}}) \frac{\partial \phi}{\partial t} ds \Big\}.$$

Unfortunately, problem (1.18), (1.19), as it stands, does not satisfy any of the conditions given in Chapter II in order to have a well-posed problem. However, we know that the original problem (I.2.30) has a solution w. If $\underline{\underline{\sigma}} = C^{-1}(\underline{\underline{D}}_2 w)$ is in $H^1(\Omega)$, that is, if the solution w of (I.2.30) is smooth enough, it is easy to check that the pair $(\underline{\underline{\sigma}}, w)$ solves (1.18) and (1.19). Hence we only have to prove the *uniqueness* of the solution of (1.18), (1.19).

Proposition 1.1: Problem (1.18) and (1.19) has a unique solution.

Proof: It is obvious that

$$(1.22)\qquad a(\underline{\underline{\tau}}, \underline{\underline{\tau}}) \geq \alpha\, \|\underline{\underline{\tau}}\|_0^2, \qquad \forall \underline{\underline{\tau}} \in V$$

Now let us check a weaker inf–sup condition. For every ϕ in Q, let us define $\underline{\underline{\tau}}(\phi)$ by

$$(1.23)\qquad \tau_{11} = \tau_{22} = \phi, \quad \tau_{12} = \tau_{21} = 0.$$

It is immediate to check that $M_{nt}(\underline{\underline{\tau}}))$ is continuous across the interelement boundaries, so that

$$\sum_K \int_{\partial K} M_{nt}(\underline{\underline{\tau}}(\phi)) \frac{\partial \phi}{\partial t} ds = 0 \tag{1.24}$$

and, therefore,

$$b(\underline{\underline{\tau}}(\phi), \phi) = |\phi|^2_{1,\Omega}. \tag{1.25}$$

It is also easy to check, using (1.23) and the Poincaré's inequality (I.2.7), that

$$||\underline{\underline{\tau}}(\phi)||_V \leq c\, |\phi|_{1,\Omega}; \tag{1.26}$$

hence we have from (1.25) and (1.26),

$$\left\{ \begin{aligned} \inf_{\phi \in H^1_0(\Omega)} \sup_{\underline{\underline{\tau}} \in V} \frac{b(\underline{\underline{\tau}}, \phi)}{||\underline{\underline{\tau}}||_V\, |\phi|_{1,\Omega}} &\geq \inf_{\phi \in H^1_0(\Omega)} \frac{b(\underline{\underline{\tau}}(\phi), \phi)}{||\underline{\underline{\tau}}(\phi)||_V\, |\phi|_{1,\Omega}} \\ &\geq \frac{|\phi|_{1,\Omega}}{||\underline{\underline{\tau}}(\phi)||_V} \geq \frac{1}{c} > 0. \end{aligned} \right. \tag{1.27}$$

Now using (1.22) and (1.27) we have the desired uniqueness by standard arguments.

We are now ready to discretize our problem. Following BREZZI–RAVIART [A] and JOHNSON [A], for any integer $k \geq 0$ we set

$$V_h = (\mathfrak{L}^0_k)^4_s \cap V, \tag{1.28}$$

$$Q_h = \mathfrak{L}^1_{k+1} \tag{1.29}$$

with the notation of Chapter III. Note that the space V_h in (1.28) is made of tensors whose normal bending moment is continuous across the interelement boundaries. The degrees of freedom for Q_h will be the usual ones (see Section III.2). As degrees of freedom for V_h we may choose, for instance the following ones:

$$\int_e M_{nn}(\underline{\underline{\tau}}) p(s)\, ds, \qquad \forall p \in P_k(e), \forall e \in \mathcal{E}_h, \tag{1.30}$$

$$\int_K \underline{\underline{\tau}} : \underline{\underline{p}}\, dx, \qquad \forall \underline{\underline{p}} \in (P_{k-1}(K))^4_s, \forall K \in \mathcal{T}_h, \ (k \geq 1) \tag{1.31}$$

The possibility of choosing (1.30) and (1.31) as degrees of freedom in V_h is shown by the following lemma and by a standard dimensional count.

Lemma 1.1: Let $\underline{\underline{\tau}} \in (P_k(K))^4_s$ be such that

$$(1.32) \qquad \int_{e_i} M_{nn}(\underline{\underline{\tau}})p(s)\, ds = 0, \qquad \forall p \in P_k(e_i) \;\; (i = 1,2,3),$$

$$(1.33) \qquad \int_K \underline{\underline{\tau}} \cdot \underline{\underline{p}}\, dx = 0, \qquad \forall \underline{\underline{p}} \in (P_{k-1}(K))^4_s \;\; (k \geq 1),$$

then $\underline{\underline{\tau}} \equiv 0$.

Proof: (Hint). From (1.32) we get $M_{nn}(\underline{\underline{\tau}}) = 0$. We first show that $D_2^*(\underline{\underline{\tau}}) = 0$. This is trivial for $k \leq 1$; for $k > 1$ take $\underline{\underline{p}} = \underline{\underline{D}}_2 b$ with $b = b_3\, D_2^* \underline{\underline{\tau}}$ in (1.33) to get $\int_K b_3 (D_2^*(\underline{\underline{\tau}}))^2 dx = 0$ and hence, $D_2^*(\underline{\underline{\tau}}) = 0$. Now use the formula (see Section IV.5.1))

$$(1.34) \qquad \int_K \underline{\underline{\tau}} : \underline{\underline{D}}_2 \phi\, dx = \int_K D_2^*(\underline{\underline{\tau}})\, \phi\, dx + \int_{\partial K} [M_{nn}(\underline{\underline{\tau}}) \frac{\partial \phi}{\partial n} - \mathcal{K}_n(\underline{\underline{\tau}})\phi]\, ds$$

for $\phi \in P_{k+1}(K)$; thus we get

$$\int_{\partial K} \mathcal{K}_n(\underline{\underline{\tau}})\, \phi\, ds = 0, \quad \forall \phi \in P_{k+1}(K),$$

and easily, that $\mathcal{K}_n(\underline{\underline{\tau}}) = 0$. It is now simple to show that $\underline{\underline{\tau}} = \underline{\underline{S}}(\underline{q})$ (see (IV.5.27) for the definition of $\underline{\underline{S}}$) for some $\underline{q} \in (P_{k+1}(K))^2$ with $\underline{q} = 0$ on $\partial = K$. Therefore, q_1 (for instance) has the form $b_3 z$ with $z \in P_{k-2}(K)$. Now let us choose in (1.33), p_{11} such that $\partial p_{11}/\partial y = z$ and $p_{12} = p_{22} = 0$; we get

$$0 = \int_K \tau_{11} p_{11}\, dx = \int_K \frac{\partial q_1}{\partial y} p_{11} dx = -\int_K q_1\; z\; dx = -\int_K b_3 z^2 dx$$

so that $z = 0$ and $q_1 = 0$. Similarly, one proves that $q_2 = 0$. □

We are now able to define the operator Π_h. We set, for $\underline{\underline{\tau}} \in V$,

$$(1.35) \qquad \int_e M_{nn}(\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}})\, p(s)\, ds = 0, \quad \forall p \in P_k(e),\ \forall e \in \mathcal{E}_h,$$

$$(1.36) \qquad \int_K (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) \cdot \underline{\underline{p}}\, ds = 0, \quad \forall \underline{\underline{p}} \in (P_{k-1}(K))^4_s,\ \forall K \in \mathcal{T}_h.$$

Lemma 1.2: Let Π_h be defined by (1.35) and (1.36). Then we have

$$(1.37) \qquad \|\Pi_h \underline{\underline{\tau}}\|_V \leq c\, \|\underline{\underline{\tau}}\|_V\,, \qquad \forall \underline{\underline{\tau}} \in V,$$

and

$$(1.38) \qquad b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, \phi_h) = 0\,, \qquad \forall \underline{\underline{\tau}} \in V,\ \forall \phi_h \in Q_h.$$

Proof: Formula (1.37) is easy to check. Let us prove (1.38). From (1.12) and (1.21) we have

$$(1.39)\quad b(\underline{\underline{\tau}}-\Pi_h\underline{\underline{\tau}},\phi) = -\sum_K\Big\{\int_K(\underline{\underline{\tau}}-\Pi_h\underline{\underline{\tau}}) : \underline{\underline{D}}_2\phi\,dx - \int_{\partial K} M_{nn}(\underline{\underline{\tau}}-\Pi_h\underline{\underline{\tau}})\frac{\partial\phi}{\partial n}ds\Big\}$$

and from (1.39), (1.35), and (1.36) we get (1.38). □

Lemma 1.3: If $\underline{\underline{\tau}}_h \in V_h$ is such that

$$(1.40)\qquad b(\underline{\underline{\tau}}_h,\phi_h) = 0, \qquad \forall\phi_h \in Q_h,$$

then

$$(1.41)\qquad b(\underline{\underline{\tau}}_h,\phi) = 0, \qquad \forall\phi \in Q.$$

Proof: We have from (1.13) and (1.21)

$$(1.42)\quad b(\underline{\underline{\tau}}_h,\phi) = -\sum_K\Big\{\int_K D_2^*(\underline{\underline{\tau}}_h)\,\phi\,dx+\int_{\partial K}[M_{nt}(\underline{\underline{\tau}}_h)\frac{\partial\phi}{\partial t}-Q_n(\underline{\underline{\tau}}_h)\phi]\,ds\Big\}.$$

Integrating $\int_{\partial K} M_{nt}\partial\phi/\partial t ds$ by parts and recalling the definition of $\mathcal{K}_n$ in (IV.5.3) we then have

$$(1.43)\qquad b(\underline{\underline{\tau}}_h,\phi) = -\sum_K\Big\{\int_K D_2^*(\underline{\underline{\tau}}_h)\,\phi\,dx - \int_{\partial K}\mathcal{K}_n(\underline{\underline{\tau}}_h)\,\phi\,ds\Big\}.$$

Note that (1.43) holds for any $\underline{\underline{\tau}}_h$ and ϕ piecewise smooth. If now (1.40) holds, we first have $D_2^*(\underline{\underline{\tau}}_h) = 0$ by choosing $\phi|_K = b_3 D_2^*(\underline{\underline{\tau}}_h)$, (for $k \geq 2$, otherwise the property is trivial). Hence, we are left with

$$(1.44)\qquad \sum_K\int_{\partial K}\mathcal{K}_n(\underline{\underline{\tau}}_h)\phi_h\,ds = 0, \qquad \forall\phi \in Q_h$$

Since $\mathcal{K}_n$ is made of Dirac measures at the vertices plus polynomials of degree $\leq k-1$ on each edge, it is easy to see that (1.44) implies $\mathcal{K}_n(\underline{\underline{\tau}}_h) = 0$. Therefore, we have proved that if $\underline{\underline{\tau}}_h \in V_h$ satisfies (1.40), then $D_2^*(\underline{\underline{\tau}}_h) = 0$ and $\mathcal{K}_n(\underline{\underline{\tau}}_h) = 0$. Now we insert those two equations into (1.43) and we get (1.41). □

This last property was denoted, in Chapter II, as $Z_h(0) \subset Z(0)$. We have seen that, together with the existence of the operator Π_h, this property is so important that it can provide optimal error estimates even in desperate situations (no ellipticity, no inf–sup condition) like ours.

Actually we remark first that (1.27) and Lemma 1.2 provide, through Proposition II.2.8, the following inf–sup type condition:

$$\inf_{\phi_h \in Q_h} \sup_{\underline{\underline{\tau}}_h \in V_h} \frac{b(\underline{\underline{\tau}}_h, \phi_h)}{\|\underline{\underline{\tau}}_h\|_V |\phi_h|_1} \geq c > 0 \quad (c \text{ independent of } h). \tag{1.45}$$

On the other hand, since Q_h and V_h are finite dimensional, (1.22) and (1.45) ensure that the discrete problem has a unique solution. We are now ready for error estimates.

Proposition 1.2: If $(\underline{\underline{\sigma}}, w)$ is the solution of (1.18) and (1.19) and $(\underline{\underline{\sigma}}_h, w_h)$ is the discrete solution of (1.18) and (1.19) through (1.28) and (1.29), we have

$$\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 \leq c\, \|\underline{\underline{\sigma}} - \Pi_h \underline{\underline{\sigma}}\|_0. \tag{1.46}$$

The proof is immediate from Proposition II.2.4. □

From (1.46) and standard approximation results we then have

$$\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 \leq c h^{k+1} \|\underline{\underline{\sigma}}\|_{k+1}. \tag{1.47}$$

Proposition 1.3: With the notation of Proposition 1.2, we have

$$\|w - w_h\|_1 \leq c\, \{h^{k+1} \|\underline{\underline{\sigma}}\|_{k+1} + h^{k+1} \|w\|_{k+2}\}. \tag{1.48}$$

Proof: Let $\phi_h \in Q_h$ to be chosen. From (1.45) we have for some $\underline{\underline{\tau}}_h \in V_h$

$$\begin{aligned} c\|\phi_h - w_h\|_1 \|\underline{\underline{\tau}}_h\|_V &\leq b(\underline{\underline{\tau}}_h, \phi_h - w_h) \\ &= b(\underline{\underline{\tau}}_h, \phi_h - w) + b(\underline{\underline{\tau}}_h, w - w_h) \\ &= b(\underline{\underline{\tau}}_h, \phi_h - w) + a(\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h, \underline{\underline{\tau}}_h). \end{aligned} \tag{1.49}$$

It is now elementary to see that ϕ_h can be chosen in such a way that

$$\int_e p \frac{\partial}{\partial t}(w - \phi_h)\, ds = 0, \quad \forall p \in P_k(e),\ \forall e \in \mathcal{E}_h, \tag{1.50}$$

$$\|w - \phi_h\|_1 \leq c h^{k+1} \|w\|_{k+2}. \tag{1.51}$$

With such a choice we have

$$\begin{aligned} b(\underline{\underline{\tau}}_h, w - \phi_h) &= \sum_K \int_K \underline{\mathrm{div}}(\underline{\underline{\tau}}_h) \cdot \underline{\mathrm{grad}}(w - \phi_h) dx \\ &\leq \|\underline{\underline{\tau}}_h\|_V \|w - \phi_h\|_1 \\ &\leq c h^{k+1} \|\underline{\underline{\tau}}_h\|_V \|w\|_{k+2} \end{aligned} \tag{1.52}$$

so that from (1.49), (1.52), and (1.47) we get (1.48). □

Remark 1.2: Result (1.48) is not optimal as far as the regularity of w is involved. Actually it states

$$\|w - w_h\|_1 \le ch^s \, \|w\|_{s+2} \quad (s \le k+1),$$

while an $(s+1)$-norm on w should be enough for optimality. A more sophisticated analysis (FALK–OSBORN [A], BABUŠKA–OSBORN–PITKÄRANTA [A]) shows that

$$\|w - w_h\|_r \le ch^{s-r} \|w\|_s \quad (s \le k+2,\ 0 \le r \le 1) \tag{1.53}$$

for $k \ge 1$ and

$$\|w - w_h\|_0 \le ch^2 \, \|w\|_4 \quad \text{for } k = 0. \tag{1.54}$$

In particular, the approach of BABUŠKA–OSBORN–PITKÄRANTA [A] is of special interest because, by a suitable use of mesh-dependent norms in V_h and Q_h, they can show that the discretized problem (in the new norms) satisfy the abstract assumptions (II.2.34) and (II.2.35) so that optimal error estimates (in the new norms) can be directly obtained by Theorem II.2.1. Their approach also works for other fourth-order mixed methods, like those analyed in Sections IV.4 and IV.5. □

Remark 1.3: In the actual solution of the discretized problem, the most convenient way is to disconnect the continuity of $\underline{\underline{\sigma}}_h \cdot \underline{n}$ and to enforce it back via Lagrange multipliers λ_h. Then one eliminates $\underline{\underline{\sigma}}_h$ at the element level and one solves a symmetric and positive definite system in the unknowns λ_h and w_h. The procedure is identical to the one described in Section V.1 and we refer to it for a detailed description. As far as the error estimates for the Lagrange multipliers λ_h are concerned, recent results have been obtained by COMODI [A]. □

Remark 1.4: It is interesting to analyze the relationship between the mixed methods described here and some nonconforming methods for fourth-order problems. For instance, the following result is proved in ARNOLD–BREZZI [A]. Let us consider the space built by means of the Morley element $\mathfrak{L}_2^{2,NC}$ described in Example III.2.5 and let us define

$$a_h(\psi_h, \phi_h) := \frac{Et^3}{12(1-\nu^2)} \sum_K \int_K [(1-\nu)\underline{\underline{D}}_2\psi_h \cdot \underline{\underline{D}}_2\phi_h + \nu\Delta\psi_h\Delta\phi_h]\, dx. \tag{1.55}$$

For every $\phi_h \in \mathfrak{L}_2^{2,NC}$, let ϕ_h^I be the piecewise linear interpolant of ϕ_h (that is, $\phi_h^I \in \mathfrak{L}_1^1$ and $\phi_h^I = \phi_h$ at the vertices). Consider now the modified Morley problem: find ψ_h in $\mathfrak{L}_2^{2,NC}$ such that

$$a_h(\psi_h, \phi_h) = (f, \phi_h^I), \qquad \forall \phi_h \in \mathfrak{L}_2^{2,NC}. \tag{1.56}$$

Then we have

$$\underline{\underline{D}}_2\psi_h = \underline{\underline{\sigma}}_h, \quad \psi_h^I = w^h, \tag{1.57}$$

where $(\underline{\underline{\sigma}}_h, w_h)$ is the discrete solution of the mixed problem (1.18), (1.19) through (1.28), (1.29) for $k = 0$. We note explicitly that, in the case of variable coefficients, the equivalence is more complicated. Note also that $\partial\psi_h/\partial n|_e = \lambda_h|_e$ for all $e \in \mathcal{E}_h$, where λ_h is the Lagrange multiplier introduced in the previous remark. Note also that we have from ARNOLD–BREZZI [A]

$$\|\psi_h - w\|_{1,h} \leq ch^2 \, \|w\|_3 \tag{1.58}$$

which improves on (1.48) and (1.54) since it requires only H^3-regularity on w. This is particularly striking since the cost for computing ψ_h is cheaper (or equal, using λ_h) than the cost for computing $(\underline{\underline{\sigma}}_h, w_h)$. □

VII.2 Mixed Methods for Linear Elasticity Problems

We shall now present some among the many mixed approximations of the two-dimensional linear elasticity problem. For convenience of the reader we recall here the mixed formulation which we already introduced in Chapter IV. We set

$$\Sigma = \{\underline{\underline{\tau}} \in \underline{\underline{H}}(\underline{\mathrm{div}}, \Omega), \int_\Omega \mathrm{tr}(\underline{\underline{\tau}})dx = 0\}, \quad U = (L^2(\Omega))^2, \tag{2.1}$$

$$a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) = \int_\Omega \left(\tfrac{1}{2\mu}\underline{\underline{\sigma}}^D : \underline{\underline{\tau}}^D + \frac{1}{(\lambda+\mu)} \, \mathrm{tr}(\underline{\underline{\sigma}}) \, \mathrm{tr}(\underline{\underline{\tau}})\right) dx, \tag{2.2}$$

$$b(\underline{\underline{\tau}}, \underline{v}) = \int_\Omega \underline{\mathrm{div}}(\underline{\underline{\tau}}) \cdot \underline{v} \, dx \tag{2.3}$$

We recall that $\mathrm{tr}(\underline{\underline{\tau}}) = \tau_{11} + \tau_{22}$ and that $\underline{\underline{\tau}}^D = \underline{\underline{\tau}} - \underline{\underline{\delta}}tr(\underline{\underline{\tau}})/2$. Note that we are using (Σ, U) instead of (V, Q) as in the abstract theory. We recall as well that λ and μ are the Lamé coefficients. Considering again, for the sake of simplicity, the case of homogeneous Dirichlet boundary conditions, we may write the problem as follows: find $\underline{\underline{\sigma}} \in \Sigma$ and $\underline{u} \in U$ such that

$$\begin{cases} a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) + b(\underline{\underline{\tau}}, \underline{u}) = 0, & \forall \underline{\underline{\tau}} \in \Sigma, \\ b(\underline{\underline{\sigma}}, \underline{v}) = (\underline{f}, \underline{v}), & \forall \underline{v} \in U \end{cases} \tag{2.4}$$

It is very easy to check that

$$\inf_{\underline{v} \in U} \sup_{\underline{\underline{\tau}} \in \Sigma} \frac{b(\underline{\underline{\tau}}, \underline{v})}{\|\underline{\underline{\tau}}\|_1 \|\underline{v}\|_0} \geq c > 0 \tag{2.5}$$

and that

$$a(\underline{\underline{\tau}},\underline{\underline{\tau}}) \geq \frac{1}{2(\lambda+\mu)}\|\underline{\underline{\tau}}\|_0^2, \qquad \forall \underline{\underline{\tau}} \in \Sigma \tag{2.6}$$

Moreover, setting (as in Chapter II)

$$\operatorname{Ker} B = \{\underline{\underline{\tau}} \in \Sigma,\ b(\underline{\underline{\tau}},\underline{v}) = 0,\ \forall \underline{v} \in U\} = \{\underline{\underline{\tau}} \in \Sigma,\ \underline{\operatorname{div}}\, \underline{\underline{\tau}} = 0\}$$

and defining

$$\|\underline{\underline{\tau}}\|_\Sigma^2 = \|\underline{\underline{\tau}}\|_0^2 + \|\underline{\operatorname{div}}\, \underline{\underline{\tau}}\|_0^2\ (\leq \|\underline{\underline{\tau}}\|_1^2), \tag{2.7}$$

we have from Proposition IV.3.1 that

$$a(\underline{\underline{\tau}},\underline{\underline{\tau}}) \geq c(\mu)\, \|\underline{\underline{\tau}}\|_\Sigma^2, \quad \forall \underline{\underline{\tau}} \in \operatorname{Ker}\ B. \tag{2.8}$$

Hence from (2.5), (2.7) and (2.8) we know that the problem (2.4) is well posed in the sense of Chapter II.

If we now choose some finite-dimensional subspaces Σ_h and U_h, we must be careful to have the discrete analogues of (2.5) and (2.8) verified. However we see here a delicate point. In order to prove an inequality of type (2.8) we *needed*, in Proposition IV.3.1, $\underline{\operatorname{div}}\underline{\underline{\tau}} = 0$. Hence, our life would be a lot easier if we had the "inclusion of the kernels" property: $\operatorname{Ker} B_h \subset \operatorname{Ker} B$. In other words, we must require from our spaces Σ_h and U_h the following property:

$$\begin{aligned} \operatorname{Ker} B_h &= \{\underline{\underline{\tau}}_h \in \Sigma_h\ ,\ b(\underline{\underline{\tau}}_h, \underline{v}_h) = 0,\ \forall \underline{v}_h \in U_h\} \\ &\subset \operatorname{Ker} B = \{\underline{\underline{\tau}} \in \Sigma\ ,\ \underline{\operatorname{div}}\, \underline{\underline{\tau}} = 0\} \end{aligned} \tag{2.9}$$

At the same time, we still need the existence of an operator $\Pi_h : \boldsymbol{\Sigma} \to \boldsymbol{\Sigma}_h$ such that

$$b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, \underline{v}_h) = 0, \quad \forall \underline{v}_h \in U_h, \tag{2.10}$$

$$\|\Pi_h \underline{\underline{\tau}}\|_\Sigma \leq c\, \|\underline{\underline{\tau}}\|_\Sigma, \quad \forall \underline{\underline{\tau}} \in \Sigma \tag{2.11}$$

We saw many examples of discrete spaces satisfying (2.9) , (2.10) and (2.11) in Section III.3, for the approximation of spaces of type $H(\operatorname{div},\Omega)$ and $L^2(\Omega)$. It seems, at the first sight, that we could just use a *pair* of vectors in $H(\operatorname{div};\Omega)$ to approximate Σ, but we should not forget the symmetry of the tensors in $\underline{\underline{H}}(\underline{\operatorname{div}};\Omega)$. The problem of finding subspaces of Σ and U satisfying (2.9), (2.10), and (2.11) is actually very difficult. Let us make a few rough computations in order to understand why: assume that we take U_h of type $\mathcal{L}_k^0$ (it is reasonable to take discontinuous functions if we are willing to get (2.9)).

Then, in each triangle, we try using polynomials of degree $k+1$ for Σ_h. We have, in each triangle,

$$3 \times \dim(P_{k+1}) = \frac{3(k+2)(k+3)}{2} \tag{2.12}$$

unknowns. In order to obtain (2.10) we try building an operator Π_h such that

$$\int_{e_i} [(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}) \cdot \underline{n}] \cdot \underline{p}_k(s)\, ds = 0, \quad \forall \underline{p}_k \in (P_k(e_i))^2, \ (i = 1,2,3), \tag{2.13}$$

and

$$\int_K (\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}) : \underline{\underline{\varepsilon}}(\underline{p}_k)\, dx = 0, \qquad \forall \underline{p}_k \in (P_k(K))^2. \tag{2.14}$$

Now (2.13) amounts to $3 \times 2 \times (k+1)$ conditions and (2.14) to $(2\dim(P_k(K)) - 3)$ conditions, that is, $(k+1)(k+2) - 3$ conditions. Now, we still need $\Pi_h \underline{\underline{\tau}}$ to be in $\underline{\underline{H}}(\underline{\mathrm{div}}; \Omega)$ and this requires the continuity of $(\Pi_h \underline{\underline{\tau}}) \cdot \underline{n}$ at the edges and introduces six additional conditions. Consider finally that, in $(P_{k+1})^4_s$, there are tensors which surely satisfy $\underline{\underline{\tau}} \cdot \underline{n} = 0$ and $\underline{\mathrm{div}}\, \underline{\underline{\tau}} = 0$. They have the form: $\tau_{11} = \partial^2 b/\partial y^2$, $\tau_{12} = \tau_{21} = -\partial^2 b/\partial x \partial y$, $\tau_{22} = \partial^2 b/\partial x^2$ with $b \in P_{k+3} \cap H_0^2(K)$. Thus, we have $(k-1)(k-2)/2 (= \dim(P_{k+3} \cap H_0^2)$ tensors to throw away because they are insensitive to (2.13) and (2.14). We are left with

$$\frac{3(k+2)(k+3)}{2} - 6(k+1) - (k+1)(k+2) + 3 - 6 - \frac{(k-1)(k-2)}{2} = -3$$

and we do anticipate trouble. At our present knowledge there are basically three possibilities at hand:

(1) to give up the symmetry of $\underline{\underline{\tau}}$ and enforce it back in a weaker form by some Lagrange multiplier;

(2) To give up the use of polynomial functions, for instance going for composite elements (hence, using piecewise polynomials inside each K);

(3) to employ an augmented formulation in the sense of Section I.5

We shall present here some examples of each one of the first two possibilities and give an hint about the third one.

Therefore, we start by giving a short idea on the approximation of (2.4) by means of discrete tensor spaces which are not symmetric. This idea, to our knowledge, was first used by FRAEIJS DE VEUBEKE [C] and his school; it was then used by AMARA–THOMAS [A] and more recently by ARNOLD–BREZZI–DOUGLAS [A]. Other recent results can be found in BREZZI–DOUGLAS–MARINI [C], MORLEY [A] and STENBERG [F,G]. The example

that we are going to present here is very close to all those previous works but has the merit to allow a shorter presentation.

We consider first the space obtained basically from the $BDFM_k$ element (cf. Section III.3) with $k = 2$,

$$\tilde{\Sigma}_h = \{\underline{\underline{\tau}}_h \mid \underline{\underline{\tau}}_h \in (\mathcal{L}_2^0)^4,\ \underline{\underline{\tau}}_h \cdot \underline{n} \text{ continuous and of degree} \leq 1 \text{ on each } e \in \mathcal{E}_h\} \tag{2.15}$$

and its subspace

$$\Sigma_h = \{\underline{\underline{\tau}}_h \in \tilde{\Sigma}_h,\ \int_\Omega \operatorname{tr}(\underline{\underline{\tau}}_h)\, dx = 0,\ \int_\Omega \operatorname{as}(\underline{\underline{\tau}}_h)\, p\, dx = 0,\ \forall p \in \mathcal{L}_1^0\} \tag{2.16}$$

where we used the notation

$$\operatorname{as}(\underline{\underline{\tau}}) = \tau_{21} - \tau_{12} \quad (= \text{ asymmetry of } \underline{\underline{\tau}}) \tag{2.17}$$

We note that, $\underline{\underline{\tau}}_h$ being locally of degree 2, the orthogonality of as $(\underline{\underline{\tau}})$ to P_1 enforces only a *weak* symmetry. Hence, $\Sigma_h \not\subset \Sigma$ (defined in (2.11)) and it must therefore be regarded as a nonconforming approximation of Σ. Next we take

$$U_h = (\mathcal{L}_1^0)^2 \subset U \tag{2.18}$$

and we consider the following discretized problem: find $(\underline{\underline{\sigma}}_h, \underline{u}_h) \in \Sigma_h \times U_h$ such that

$$\begin{cases} a(\underline{\underline{\sigma}}_h, \underline{\underline{\tau}}_h) + b(\underline{\underline{\tau}}_h, \underline{u}_h) = 0, & \forall \underline{\underline{\tau}}_h \in \Sigma_h, \\ b(\underline{\underline{\sigma}}_h, \underline{v}_h) = (\underline{f}, \underline{v}_h), & \forall \underline{v}_h \in U_h. \end{cases} \tag{2.19}$$

We shall first show that Σ_h has a local basis; then we will show that (2.19) has a unique solution and finally we will give optimal error bounds.

Proposition 2.1: Any $\underline{\underline{\tau}} \in (P_2(K))^4$ with $\underline{\underline{\tau}} \cdot \underline{n} \in P_1(e_i))^2$ $(i = 1, 2, 3)$, is uniquely determined by the values of

$$\int_{e_i} (\underline{\underline{\tau}} \cdot \underline{n}) \cdot \underline{p}\, ds, \qquad \forall \underline{p} \in (P_1(e_i))^2\ (i = 1, 2, 3), \tag{2.20}$$

$$\int_K \underline{\underline{\tau}} : \underline{\underline{\varepsilon}}(\underline{p})\, dx, \qquad \forall \underline{p} \in (P_1(K))^2, \tag{2.21}$$

$$\int_K \operatorname{as}(\underline{\underline{\tau}})\, p\, dx, \qquad \forall p \in P_1(K). \tag{2.22}$$

Proof: Check first that (2.20)–(2.22) define $18 = 12 + 3 + 3$ conditions which equals the dimension of $(P_2)^4$ with $\underline{\underline{\tau}} \cdot \underline{n} \in (P_1(e_i))^2$ on each e_i. Hence, we only have to check that the homogeneous system has only the trivial solution. Equating (2.20) to zero gives $\underline{\underline{\tau}} \cdot \underline{n} \equiv 0$ on ∂K while equating (2.21) and (2.22) to zero gives $\underline{\mathrm{div}}\, \underline{\underline{\tau}} = 0$. This implies $\underline{\underline{\tau}} = \underline{\underline{\mathrm{curl}}}(\underline{\phi})$ with $\underline{\phi} \in (B_3(K))^2$. Note now that as $(\underline{\underline{\mathrm{curl}}}(\underline{\phi})) = -\mathrm{div}\, \underline{\phi}$ so that equating again (2.22) to zero we have

$$\int_K \underline{\phi} \cdot \underline{\mathrm{grad}}\, p \, dx = 0, \ \forall p \in P_1(K),$$

which implies $\underline{\phi} = 0$. □

Remark 2.1: The above results do not really imply a local basis for Σ_h, if we keep $\int_\Omega \mathrm{tr}(\underline{\underline{\tau}}) dx = 0$. We saw, however, that this condition is used only to simplify some proofs (in this case it can be imposed a posteriori) and is not really used on the computer unless $\lambda = +\infty$. □

Now we will look for sufficient conditions in order to prove existence and uniqueness of the solution of (2.19). We have the following immediate result.

Proposition 2.2: Let $\underline{\underline{\tau}}_h \in \Sigma_h$ satisfy

$$b(\underline{\underline{\tau}}_h, \underline{v}_h) = 0, \qquad \forall \underline{v}_h \in U_h. \tag{2.23}$$

Then $\underline{\underline{\tau}}_h$ also satisfies $\underline{\mathrm{div}}\, \underline{\underline{\tau}}_h = 0$.

The proof is obvious. □

Next we have

Proposition 2.3: For any $\underline{\underline{\tau}}_h \in \mathrm{Ker}\, B_h$ we have

$$a(\underline{\underline{\tau}}_h, \underline{\underline{\tau}}_h) \geq c(\mu) \, \|\underline{\underline{\tau}}_h\|_\Sigma^2, \quad \underline{\underline{\tau}} \in \mathrm{Ker}\ B_h, \tag{2.24}$$

with $c(\mu)$ independent of $\lambda \in [0, +\infty)$.

Again the proof is an easy adaptation of Proposition IV.3.1. □

Proposition 2.4: There exists a linear operator $\Pi_h : (H^1(\Omega))^4_s \to \Sigma_h$ such that

$$b(\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}, \underline{v}_h) = 0, \qquad \forall \underline{v}_h \in U_h, \tag{2.25}$$

$$\|\Pi_h \underline{\underline{\tau}}\|_\Sigma \leq c \, \|\underline{\underline{\tau}}\|_1 \tag{2.26}$$

and, moreover,

$$\|\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}\|_1 \leq ch^2 \, \|\underline{\underline{\tau}}\|_2. \tag{2.27}$$

Proof: Using Proposition 2.1 we define $\Pi_h \underline{\underline{\tau}}$ by

$$(2.28)\qquad \int_{\mathcal{E}_h} (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) \cdot \underline{n} \cdot \underline{p}\, ds = 0, \qquad \forall \underline{p} \in (\mathfrak{L}_1^0(\mathcal{E}_h))^2,$$

$$(2.29)\qquad \int_K (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) : \underline{\underline{\varepsilon}}(\underline{p})\, dx = 0, \qquad \forall \underline{p} \in (P_1(K))^2,\ \forall K \in \mathcal{T}_h,$$

$$(2.30)\qquad \int_\Omega \text{as}\,(\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) p\, dx = 0, \qquad \forall p \in \mathfrak{L}_1^0(\mathcal{T}_h).$$

Note that (2.30) actually implies $\int_\Omega \text{as}\,(\Pi_h \underline{\underline{\tau}})\ p\ dx = 0$ since $\underline{\underline{\tau}} \in \Sigma$ implies $\text{as}\,(\underline{\underline{\tau}}) \equiv 0$. Note also that if $\int_\Omega \text{tr}(\underline{\underline{\tau}}) dx = 0$, then (2.29) implies $\int_\Omega \text{tr}(\Pi_h \underline{\underline{\tau}})\ dx = 0$, taking $\underline{p} = (x, y)$ so that $\underline{\underline{\varepsilon}}(\underline{p}) = \underline{\underline{\delta}}$. Now it is clear that integrating (2.29) by parts and using (2.30) and (2.28) we have, for all $\underline{v}_h \in U_h \equiv (\mathfrak{L}_1^0)^2$,

$$(2.31)\qquad \int_K \underline{\text{div}}(\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) \cdot \underline{v}_h\, dx$$
$$= -\int_K (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) : \underline{\underline{\varepsilon}}(\underline{v}_h)\, dx + \int_{\partial K} (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) \cdot \underline{n} \cdot \underline{v}_h\, ds = 0$$

which proves (2.25). Then (2.26) and (2.27) follow from (2.28)–(2.30) with a simple scaling argument (see Section III.2.4). □

Now, proceeding as in Proposition II.2.8 we have

Proposition 2.5: The following inf–sup condition holds:

$$(2.32)\qquad \inf_{\underline{v}_h \in U_h} \sup_{\underline{\underline{\tau}}_h \in \Sigma_h} \frac{b(\underline{\underline{\tau}}_h, \underline{v}_h)}{\|\underline{\underline{\tau}}_h\|_\Sigma \|\underline{v}_h\|_0} \geq c > 0$$

with c independent of h. □

From Propositions 2.3 and 2.5 and Theorem II.1.1 we then have

Theorem 2.1: Problem (2.19) has a unique solution. □

We are now ready for the error estimates.

Theorem 2.2: If $(\underline{\underline{\sigma}}, \underline{u})$ is the solution of (2.4) and $(\underline{\underline{\sigma}}_h, \underline{u}_h)$ is the solution of (2.19), we have

$$(2.33)\qquad \|\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}\|_0 + \|\underline{u}_h - \underline{u}\|_0 \leq c(\mu) h^2\, (\|\underline{u}\|_3 + \|\underline{\underline{\sigma}}\|_2).$$

Proof: We just have to pay some care to the nonconformity $\Sigma_h \not\subseteq \Sigma$. We have, using (2.24)

$$
\begin{aligned}
c(\mu)\,\|\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}\|_0^2 &\le a(\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}, \underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}) \\
&= a(\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}, \underline{\underline{\sigma}}_h - \Pi_h\underline{\underline{\sigma}}) + a(\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}, \Pi_h\underline{\underline{\sigma}} - \underline{\underline{\sigma}}).
\end{aligned}
\tag{2.34}
$$

Now set $\underline{\underline{\tau}}_h := \underline{\underline{\sigma}}_h - \Pi_h\underline{\underline{\sigma}}$ and note that from (2.4), (2.19), (2.25), and Proposition 2.2 we have

$$
\underline{\mathrm{div}}\ \underline{\underline{\tau}}_h = 0, \tag{2.35}
$$

which implies, using (2.19)

$$
a(\underline{\underline{\sigma}}_h, \underline{\underline{\tau}}_h) = 0\,. \tag{2.36}
$$

On the other hand, recalling the definition of $a(\cdot,\cdot)$, (2.2), and the relation $\underline{\underline{\sigma}}^D/\mu + 1/2(\lambda+\mu)\ \mathrm{tr}(\underline{\underline{\sigma}})\underline{\underline{\delta}} = \underline{\underline{\varepsilon}}(\underline{u})$, we have

$$
a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}_h) = \int_\Omega \underline{\underline{\tau}}_h : \underline{\underline{\varepsilon}}(\underline{u})\ dx. \tag{2.37}
$$

Now from (2.16) and (2.35) we have, for every $\underline{v}_h \in (\mathfrak{L}_2^1 \cap H_0^1)^2$,

$$
\int_\Omega \underline{\underline{\tau}}_h : \underline{\underline{\varepsilon}}(\underline{v}_h)\ dx = \int_\Omega \underline{\underline{\tau}}_h : \underline{\underline{\mathrm{grad}}}\ \underline{v}_h\ dx = \int_\Omega -\underline{\mathrm{div}}(\underline{\underline{\tau}}_h)\cdot \underline{v}_h\ dx = 0 \tag{2.38}
$$

so that from (2.37), (2.38), and standard approximation theory we get

$$
|a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}_h)| \le ch^2\ \|\underline{u}\|_3 \|\tau_h\|_0\,. \tag{2.39}
$$

From (2.34), (2.27), (2.36), and (2.39) we have

$$
\|\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}\|_0 \le c(\mu)h^2\ (\|\underline{\underline{\sigma}}\|_2 + \|\underline{u}\|_3)\,. \tag{2.40}
$$

Now let $\bar{\underline{u}}_h$ be the orthogonal projection of $\underline{u}$ onto $(\mathfrak{L}_1^0)^2$; from (2.32) we know that there exists a $\underline{\underline{\tau}}_h \in \Sigma_h$ such that

$$
c\|\bar{\underline{u}}_h - \underline{u}_h\|_0\ \|\underline{\underline{\tau}}_h\|_\Sigma \le b(\underline{\underline{\tau}}_h, \bar{\underline{u}}_h - \underline{u}_h) = b(\underline{\underline{\tau}}_h, \underline{u} - \underline{u}_h), \tag{2.41}
$$

the last equality following from the definition of $\bar{\underline{u}}_h$. Using again the fact that $\int_\Omega \underline{\underline{\tau}}_h \cdot \underline{\underline{\varepsilon}}(\underline{v}_h)\ dx = \int_\Omega \underline{\underline{\tau}}_h \cdot \underline{\underline{\mathrm{grad}}}(\underline{v}_h)\ dx$ for all $\underline{v}_h \in (\mathfrak{L}_2^1)^2$, we now have

$$
\begin{aligned}
|b(\underline{\underline{\tau}}_h, \underline{u}) + \int_\Omega \underline{\underline{\tau}}_h : \underline{\underline{\varepsilon}}(\underline{u})\ dx| &= |\int_\Omega [-\underline{\underline{\tau}}_h : \underline{\underline{\mathrm{grad}}}\ \underline{u} + \underline{\underline{\tau}}_h : \underline{\underline{\varepsilon}}(\underline{u})]\ dx| \\
&\le ch^2\ \|\underline{\underline{\tau}}_h\|_0 \|\underline{u}\|_3
\end{aligned}
\tag{2.42}
$$

which combined with (2.41) yields (using (2.4) and (2.19))

$$
c\|\bar{\underline{u}}_h - \underline{u}_h\|_0\ \|\underline{\underline{\tau}}_h\|_\Sigma \le ch^2\ \|\underline{\underline{\tau}}_h\|_0 \|\underline{u}\|_3 + a(\underline{\underline{\sigma}}_h - \underline{\underline{\sigma}}, \underline{\underline{\tau}}_h), \tag{2.43}
$$

so that, from (2.43), (2.40), and standard estimates on $\|\underline{u} - \bar{\underline{u}}_h\|_0$ we obtain (2.33). □

Remark 2.2: If we define

$$\omega := \frac{1}{2}\left\{\frac{\partial u_2}{\partial x_1} - \frac{\partial u_1}{\partial x_2}\right\} \tag{2.44}$$

and

$$c(\underline{\underline{\tau}}, \gamma) := \int_\Omega \operatorname{as}(\underline{\underline{\tau}})\, \gamma \, dx, \tag{2.45}$$

we may see that, in general, an approximation of (2.4) with relaxed symmetry requirements corresponds to a conforming approximation of the following continuous problem: find $(\underline{\underline{\sigma}}, \underline{u}, \omega) \in (H(\operatorname{div};\Omega))^2 \times (L^2(\Omega))^2 \times L^2(\Omega)$ such that

$$\begin{cases} a(\underline{\underline{\sigma}}, \underline{\underline{\tau}}) + b(\underline{\underline{\tau}}, \underline{u}) + c(\underline{\underline{\tau}}, \omega) = 0, & \forall \underline{\underline{\tau}} \in (H(\operatorname{div};\Omega))^2, \\ b(\underline{\underline{\sigma}}, \underline{v}) = (\underline{f}, \underline{v}), & \forall \underline{v} \in (L^2(\Omega))^2, \\ c(\underline{\underline{\sigma}}, \gamma) = 0, & \forall \gamma \in L^2(\Omega). \end{cases} \tag{2.46}$$

In particular, the method described above corresponds to an approximation of (2.46) where $(H(\operatorname{div};\Omega))^2$ is approximated by pairs of $BDFM_2$, $(L^2(\Omega))^2$ is approximated by $(\mathcal{L}_1^0)^2$, and $L^2(\Omega)$ by $\mathcal{L}_1^0$. Moreover, in the formulation (2.19) we did work *directly in the kernel* of C_h:

$$\operatorname{Ker} C_h = \{\underline{\underline{\tau}}_h \mid c(\underline{\underline{\tau}}_h, \gamma_h) = 0,\ \forall \gamma_h \in \mathcal{L}_1^0\}. \tag{2.47}$$

This was possible because we were able to construct a local basis of weakly symmetric tensors. If we had approximated the problem in the full form (2.46), we would also obtain

$$\|\omega - \omega_h\|_0 \le ch^2. \tag{2.48}$$

Note, however, that in our analysis we made "a wild use" of the degrees of freedom (2.20)–(2.22). At our knowledge this approach (that is, to work directly in the kernel of C_h) cannot be used with "smaller choices" for Σ_h. However, if one wants to use an element with lower degree (and less degrees of freedom), one can go back to form (2.46) which allows the use of less ambitious elements. For instance, one can think of approximating (2.46) by means of:

(i) pairs of Raviart–Thomas elements of lowest degree for $\underline{\underline{\sigma}}^h$;

(ii) piecewise constants ($\equiv (\mathcal{L}_0^0)^2$) for $\underline{u}_h$;

(iii) piecewise linear continuous elements ($\equiv \mathcal{L}_1^1$) for ω_h.

This does not work. However, we may enrich the tensor space, with $\underline{\underline{\operatorname{curl}}}(\underline{b})$ for $\underline{b} \in (B_3)^2$ (thus adding two more degree of freedom per triangle). In that way one gets the PEERS element of ARNOLD–BREZZI–DOUGLAS [A] which

converges with an $O(h)$ rate (uniformly in $\gamma \in [0, +\infty)$). Another possibility is to use pairs of BDM elements of degree 1 for $\underline{\underline{\sigma}}_h$ and again $\underline{u}_h$ and $\omega_h \in \mathcal{L}_1^1$. It is proved in BREZZI–DOUGLAS–MARINI [B] that, enriching again the tensor space with curl $(B_3)^2$, we have

$$\|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 + \|\bar{\underline{u}}_h - \underline{u}_h\|_0 + \|\omega - \omega_h\|_0 \leq ch^2, \tag{2.49}$$

where $\bar{\underline{u}}_h$ is the projection of $\underline{u}$ onto $(\mathcal{L}_0^0)^2$.

Obviously many other choices are possible. For equilibrium methods (instead of mixed; but the difference, at a second sight, is negligible) AMARA–THOMAS [A] considered a family of elements with reduced symmetry, whose element of lowest degree coincides (at third sight) to the one described here. In the same framework see also STENBERG [B].

If one discretizes (2.46) with continuous $\underline{u}$ and ω (and discontinuous $\underline{\underline{\sigma}}$), one can reach (after elimination of $\underline{\underline{\sigma}}$ by static condensation) a final matrix in the unknowns $\underline{u}$ and ω. This is particularly interesting in applications to shell problems since it makes an explicit use of the "drilling degrees of freedom" (see HUGHES–BREZZI [A]). □

Remark 2.3: From the computational point of view, the best approach for solving, say, (2.19) (but the same holds for all the other elements of this type) is to relax the continuity requirements on $\underline{\underline{\tau}} \cdot \underline{n}$ by means of Lagrange multipliers $\underline{\lambda}_h$ at the interelement boundaries. Then we can proceed as in Section V.1 and eliminate $\underline{\underline{\sigma}}^h$ at the element level. In the case of (2.19) we can also eliminate $\underline{u}_h$ afterwards and be left with a matrix which is symmetric and positive definite in the unknowns $\underline{\lambda}_h$ (generalized displacements). This would also produce an $O(h^3)$ approximation of $\underline{u}$. □

We turn to the second possibility considered at the begining of this chapter, that is considering test and trial functions which are *not polynomials* in each K. The simplest example would be to use, in each K, functions which are piecewise polynomial on a given subdivision of K into subelements. We obtain in that way the so-called composite elements which have been widely used in particular for constructing conforming approximations of spaces like $H^2(\Omega)$ requiring (essentially) C^1-continuity. To our knowledge the use of composite elements for the mixed formulation of elasticity problems was introduced in JOHNSON–MERCIER [A]. More recent results in this direction have been obtained by ARNOLD–DOUGLAS–GUPTA [A]. Here we shall present first the Johnson–Mercier element (for triangles) and then give a short idea on the Arnold–Douglas–Gupta family of elements.

Assume then that each triangle $K \in \mathcal{T}_h$ is subdivided into three subtriangles K_j $(j = 1, 2, 3)$ as indicated in Figure VII.1, where, for instance, B is the barycenter of K.

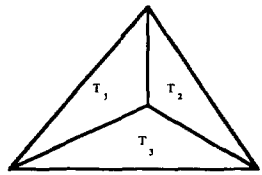


Figure VII.1

Now consider the space

$$(2.50)\quad JM_1(K) = \{\underline{\underline{\tau}} \mid \underline{\underline{\tau}} \in \underline{\underline{H}}(\underline{\mathrm{div}}; K),\quad \underline{\underline{\tau}}|_{K_j} \in (P_1(K_j))^4_s \ \ (j = 1,2,3)\}.$$

Note that in (2.50), the condition $\underline{\underline{\tau}} \in \underline{\underline{H}}(\underline{\mathrm{div}}; K)$ requires that $\underline{\underline{\tau}} \cdot \underline{n}$ be continuous from one subtriangle to another. We need the following lemma.

Lemma 2.1: If ξ is a function on K such that $\xi|_{K_j}$ be constant $(j = 1,2,3)$ and

$$(2.51)\qquad \int_K \xi p \, dx = 0, \qquad \forall p \in P_1(K),$$

then $\xi = 0$.

The proof is left as an exercise. □

We may now choose the degrees of freedom in $JM_1(K)$.

Proposition 2.6: An element $\underline{\underline{\tau}}$ of $JM_1(K)$ is uniquely determined by the following 15 degrees of freedom

$$(2.52)\qquad \int_{e_i} (\underline{\underline{\tau}} \cdot \underline{n}) \cdot \underline{p} \, ds, \quad \forall \underline{p} \in (P_1(e_i))^2,\ i = 1,2,3,$$

$$(2.53)\qquad \int_K \underline{\underline{\tau}} : \underline{\underline{p}} \, dx, \qquad \forall \underline{\underline{p}} \in (P_0(K))^4_s.$$

Proof: It is clear that $\dim(JM_1(K)) \geq 15$ (that is, $3 \times \dim(P_1)^4_s$ minus the 12 conditions needed to enforce the continuity of $\underline{\underline{\tau}} \cdot \underline{n}$). Hence we must show that if $\underline{\underline{\tau}}$ is such that

$$(2.54)\qquad \begin{aligned} &\int_{e_i} (\underline{\underline{\tau}} \ \ \underline{n}) \cdot \underline{p} \, ds = 0, \qquad \forall \underline{p} \in (P_1(e_i))^2,\ i = 1,2,3, \\ &\int_K \underline{\underline{\tau}} : \underline{\underline{p}} \, dx, \qquad \forall \underline{\underline{p}} \in (P_0(K))^4_s, \end{aligned}$$

there follows $\underline{\underline{\tau}} = 0$. Clearly (2.53) implies $\underline{\underline{\tau}} \cdot \underline{n} = 0$ on ∂K. Now for every $\underline{v} \in (P_1(K))^2$ we have $\underline{\underline{\varepsilon}}(\underline{v}) \in (P_0(K))^4_s$ so that

$$\int_K \underline{\mathrm{div}}\, \underline{\underline{\tau}} \cdot \underline{v}\, dx = -\int_K \underline{\underline{\tau}} : \underline{\underline{\varepsilon}}(\underline{v})\, dx = 0 .$$

Applying Lemma 2.1 this implies $\underline{\mathrm{div}}\, \underline{\underline{\tau}} = 0$. It is easy to see that then $\underline{\underline{\tau}}$ must have the form

$$\underline{\underline{\tau}} = A\underline{\underline{i}}ry(\psi) = \begin{pmatrix} \partial^2\psi/y^2 & -\partial^2\psi/\partial x\partial y \\ -\partial^2\psi/\partial x\partial y & \partial^2\psi/\partial x^2 \end{pmatrix} \tag{2.55}$$

and that $\underline{\underline{\tau}} \cdot \underline{n} = 0$ implies that ψ in (2.55) can be chosen in such a way that $\psi = \partial\psi/\partial n = 0$ on ∂K.

Since (2.55) also implies $\psi \in P_3(K_j)$ $(j = 1, 2, 3)$ and $\psi \in C^1(K)$, then we conclude that $\psi = 0$ (see, for instance, CIARLET [B] on the unisolvence of the Hsieh–Clough–Tocher element). □

We are now ready to construct approximations for Σ and U as in (2.1). We set

$$\begin{aligned} \Sigma_h = \{\underline{\underline{\tau}}_h \mid \underline{\underline{\tau}}_h \in \underline{\underline{H}}(\underline{\mathrm{div}};\Omega),\ \underline{\underline{\tau}}_h|_K \in JM_1(K), \\ \forall K \in \mathcal{T}_h,\ \int_\Omega \mathrm{tr}(\underline{\underline{\tau}}_h)dx = 0\}, \end{aligned} \tag{2.56}$$

$$U_h = \mathfrak{L}_1^0. \tag{2.57}$$

It is very easy to show that, from Lemma 2.1, we have $\mathrm{Ker}\, B_h \subset \mathrm{Ker}\, B$, that is,

$$\{\underline{\underline{\tau}}_h \in \Sigma_h,\ b(\underline{\underline{\tau}}_h, \underline{v}_h) = 0,\ \forall \underline{v}_h \in U_h\} \subseteq \{\underline{\underline{\tau}} \mid \underline{\mathrm{div}}\, \underline{\underline{\tau}} = 0\}. \tag{2.58}$$

Now we have, from Proposition IV.3.1 that

$$a(\underline{\underline{\tau}}_h, \underline{\underline{\tau}}_h) \geq c(\mu)\, \|\underline{\underline{\tau}}_h\|_0^2, \qquad \forall \underline{\underline{\tau}}_h \in \mathrm{Ker}\, B_h. \tag{2.59}$$

Finally, it is easy to check that the operator Π_h defined by

$$\int_{\mathcal{E}_h} (\Pi_h \underline{\underline{\tau}}_h - \underline{\underline{\tau}}) \cdot \underline{p}\, ds = 0, \qquad \forall \underline{p} \in (\mathfrak{L}_1^0(\mathcal{E}_h))^2, \tag{2.60}$$

$$\int_\Omega (\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}) \cdot \underline{\underline{p}}\, dx = 0, \qquad \forall \underline{\underline{p}} \in (\mathfrak{L}_0^0)^4_s, \tag{2.61}$$

satisfies

$$b(\Pi_h \underline{\underline{\tau}} - \underline{\underline{\tau}}, \underline{v}_h) = 0, \qquad \forall \underline{v}_h \in U_h, \tag{2.62}$$

$$\|\Pi_h \underline{\underline{\tau}}\|_\Sigma \leq c\, \|\underline{\underline{\tau}}\|_1 \tag{2.63}$$

and

$$\|\underline{\underline{\tau}} - \Pi_h \underline{\underline{\tau}}\|_0 \leq ch^2\, \|\underline{\underline{\tau}}\|_2. \tag{2.64}$$

We can, therefore, use Theorem II.2.1 and get the following result.

Theorem 2.3: If $(\underline{\underline{\sigma}}, \underline{u})$ is the solution of (2.4) and $(\underline{\underline{\sigma}}_h, \underline{u}_h)$ is the solution of the discretized problem corresponding to the choice (2.56) and (2.57), then

$$\begin{aligned} \|\underline{\underline{\sigma}} - \underline{\underline{\sigma}}_h\|_0 + \|\underline{u} - \underline{u}_h\|_0 &\le c(\mu)\, [\|\underline{\underline{\sigma}} - \Pi_h \underline{\underline{\sigma}}\|_0 + \inf_{\underline{v}_h \in U_h} \|\underline{u} - \underline{v}_h\|_0] \\ &\le c(\mu)\, h^2\, [\|\underline{\underline{\sigma}}\|_2 + \|\underline{u}\|_2]. \ \square \end{aligned} \tag{2.65}$$

Remark 2.4: The space (2.56) contains 15 degrees of freedom per triangle, whereas the space (2.16) had 18 degrees of freedom per triangle. However the elements of (2.16) are polynomials in each K and the ones in (2.56) are only piecewise polynomials in each K, and then more difficult to deal with. The choice between them should therefore be made on a "case by case" basis. $\square$

The composite element presented in (2.56) corresponds (roughly) to the choice $k = 1$. The Arnold–Douglas–Gupta families starts with $k = 2$ and considers, on each $K = \bigcup_j K_j$, subspaces of

$$\begin{aligned} W_k = \{\underline{\underline{\tau}}_h \mid \underline{\underline{\tau}}_h \in \underline{\underline{H}}(\underline{\mathrm{div}}; K),\ \underline{\underline{\tau}}_h|K_j \in (P_k(K_j))^4_s,\ (j = 1,2,3), \\ \underline{\mathrm{div}}\, \underline{\underline{\tau}}_h \in (P_{k-1}(K))^2\}. \end{aligned} \tag{2.66}$$

Now we have to choose (arbitrarly, but some reasonable choice is indicated in ARNOLD–DOUGLAS–GUPTA [A]) three tensors in $W_k \cdot \underline{\underline{\tau}}_h^{(1)}, \underline{\underline{\tau}}_h^{(2)}, \underline{\underline{\tau}}_h^{(3)}$, the only constraint being

$$c_1 \underline{\underline{\tau}}_h^{(1)} + c_2 \underline{\underline{\tau}}_h^{(2)} + c_3 \underline{\underline{\tau}}_h^{(3)} \notin (P_k(K))^4_s, \quad \forall (c_1, c_2, c_3) \neq (0,0,0). \tag{2.67}$$

The tensors $\underline{\underline{\tau}}_h^{(r)}$ $(r = 1,2,3)$ clearly depend on k, and will be the only "composite" part in Σ_h. We set now

$$ADG_k(K) = \mathrm{span}\{(P_k(K))^4_s,\ \underline{\underline{\tau}}_h^{(1)},\ \underline{\underline{\tau}}_h^{(2)},\ \underline{\underline{\tau}}_h^{(3)}\}. \tag{2.68}$$

The corresponding degrees of freedom are

$$\int_{e_i} (\underline{\underline{\tau}} \cdot \underline{n}) \cdot \underline{p}\, ds, \qquad \forall \underline{p} \in (P_k(e_i))^2\ (i = 1,2,3), \tag{2.69}$$

$$\int_K \underline{\underline{\tau}} : \underline{\underline{\varepsilon}}(\underline{p})\, dx, \qquad \forall \underline{p} \in (P_{k-1}(K))^2, \tag{2.70}$$

$$\int_K \underline{\underline{\tau}} : A\underline{\underline{i}}ry(\underline{p})\, dx, \qquad \forall \underline{p} \in (B_{k+2}(K))^2 \cap H_0^2(K). \tag{2.71}$$

Then, we set for $k \ge 2$,

$$\begin{aligned} \Sigma_h = \{\underline{\underline{\tau}}_h \mid \underline{\underline{\tau}}_h \in \underline{\underline{H}}(\mathrm{div}; \Omega), \\ \underline{\underline{\tau}}_h|_T \in ADG_k(K),\ \forall K \in \mathcal{T}_h,\ \int_\Omega tr(\underline{\underline{\tau}}_h) dx = 0\}, \end{aligned} \tag{2.72}$$

$$U_h = \mathcal{L}^0_{k-1}. \tag{2.73}$$

In particular Arnold–Douglas–Gupta proved that if Π_h is the operator from $(H^1(\Omega))^4_s$ into Σ_h naturally associated with the degrees of freedom (2.69)–(2.71) and if P_h is the orthogonal projection on $\mathcal{L}^0_{k-1}$, then we have the commuting diagram

$$\begin{array}{ccc} (H^1)^4_s & \xrightarrow{\underline{\mathrm{div}}} & U \\ \Pi_h \big\downarrow & & P_h \big\downarrow \\ \Sigma_h & \xrightarrow{\underline{\mathrm{div}}} & U_h \end{array} \tag{2.74}$$

which, as we have seen, implies all sorts of optimal error bounds. We refer to ARNOLD–DOUGLAS–GUPTA [A] for the proofs and additional results.

Remark 2.5: The previous composite element (2.56),(2.57) of Johnson–Mercier did not satisfy (2.74) because, in this case, $\underline{\mathrm{div}}(\Sigma_h) \not\subseteq U_h$. On the other hand, the element (2.16), (2.18) satisfies (2.74) somehow, but $\Sigma_h \not\subseteq \Sigma$. □

Remark 2.6: We might also consider a reduced ADG element. For instance we can define

$$\begin{aligned} \Sigma_h = \{\underline{\underline{\tau}}_h \mid \underline{\underline{\tau}}_h \in \underline{\underline{H}}(\underline{\mathrm{div}};\Omega),\ \underline{\underline{\tau}}_h|_K \in ADG_k,\ \underline{\underline{\tau}}_h \cdot \underline{n}|_e \in (P_{k-1}(e))^2, \\ \forall K \in \mathcal{T}_h,\ \forall e \in \mathcal{E}_h\} \end{aligned} \tag{2.75}$$

(plus, always formally, the condition $\int_\Omega \mathrm{tr}(\underline{\underline{\tau}}_h)\, dx = 0$). For $k = 2$ we obtain an element which has, on each triangle, only 12 degrees of freedom, satisfies (2.74), and has an $O(h^2)$ accuracy (in $L^2(\Omega)$) for both $\underline{\underline{\sigma}}$ and $\underline{u}$.

Remark 2.7: In recent times, relevant progress has been made on this subject by the use of stabilizing techniques such as those described in Section I.5 and Section VI.5.5. For the present context they correspond in adding to formulation (2.4) a term like

$$\delta h^2 \sum_K \left\{ \int_K \underline{\mathrm{div}}\underline{\underline{\sigma}} \cdot \underline{\mathrm{div}}\underline{\underline{\tau}}\, dx - \int_K \underline{f} \cdot \underline{\mathrm{div}}\, \underline{\underline{\tau}}\, dx \right\}$$

(clearly vanishing if $\underline{\underline{\sigma}}$ is the continuous solution). The parameter δ has to be conveniently chosen. Note however that this will work for discretizations using continuous displacements and discontinuous stresses, that is, in a functional context which is different from (2.1) (and less "mixed"). We refer to FRANCA–HUGHES [A] (and references therein) for additional information. □

VII.3 Moderately Thick Plates

VII.3.1 Generalities

We end this chapter with a hint on the theory for the so-called "Mindlin–Reissner plates." The corresponding model stands somehow in between the standard three-dimensional linear elasticity and the two-dimensional Kirchoff theory for thin plates. Let us recall it briefly. Assume that we are given a three-dimensional elastic body that, in the absence of forces, occupies the region $\Omega\times]-t,t[$, where $\Omega\subset I\!R^2$ is a bounded smooth domain and $t>0$ is " small" (but not "too small") with respect to diam(Ω). This is what we call a "moderately thick" plate. We shall assume, for the sake of simplicity, that the plate is clamped along the entire boundary $\partial\Omega\times]-t,t[$ and acted by a vertical load $\underline{f}=(0,0,f_3)$. The Mindlin model assumes that the "in plane" displacements u_1 and u_2 have the form

$$u_1(x,y,z)=-z\beta_1(x,y),\quad u_2(x,y,z)=-z\beta_2(x,y) \tag{3.1}$$

and that the "transversal" displacement u_3 has the form

$$u_3(x,y,z)=w(x,y). \tag{3.2}$$

The corresponding strain field therefore takes the form:

$$\begin{cases}\varepsilon_{11}=-z\,\partial\beta_1/\partial x;\ \varepsilon_{22}=-z\,\partial\beta_2/\partial y;\ \varepsilon_{33}=0;\\ 2\varepsilon_{12}=-z(\partial\beta_1/\partial y+\partial\beta_2/\partial x);\ 2\varepsilon_{13}=\partial w/\partial x-\beta_1;\ \boldsymbol{2\varepsilon_{23}=\partial w/\partial y-\beta_2};\end{cases} \tag{3.3}$$

and assuming a linear elastic material the stress field is

$$\begin{cases}\sigma_{11}=(\varepsilon_{11}+\nu\varepsilon_{22})\,E/(1-\nu^2);\ \sigma_{22}=(\varepsilon_{22}+\nu\varepsilon_{11})\boldsymbol{E/(1-\nu^2)};\\ \sigma_{ij}=\varepsilon_{ij}E/(1+\nu);\quad i,j=1,2,3,\ i\neq j.\end{cases} \tag{3.4}$$

If we now write the total potential energy

$$\Pi=\tfrac{1}{2}\int_{\Omega\times]-t,t[}(\underline{\underline{\sigma}}:\underline{\underline{\varepsilon}}-2\,\underline{f}\cdot\underline{u})\,dx\,dy\,dz \tag{3.5}$$

in terms of $\underline{\beta}$ and w through (3.1)–(3.4) we obtain (after some calculations)

$$\Pi=\frac{t^3}{2}a(\underline{\beta},\underline{\beta})+\frac{\lambda t}{2}\int_\Omega|\underline{\mathrm{grad}}\,w-\underline{\beta}|^2\,dx\,dy-\int_{\Omega\times]-t,t[}f_3\,w\,dx\,dy\,dz, \tag{3.6}$$

where

$$\begin{aligned}a(\underline{\beta},\underline{\beta}):=\frac{E}{12(1-\nu^2)}\int_\Omega\Big[&\Big(\frac{\partial\beta_1}{\partial x}+\frac{\nu\partial\beta_2}{\partial y}\Big)\frac{\partial\beta_1}{\partial x}\\&+\Big(\frac{\nu\partial\beta_1}{\partial x}+\frac{\partial\beta_2}{\partial y}\Big)\frac{\partial\beta_2}{\partial y}+\frac{(1-\nu)}{2}\Big(\frac{\partial\beta_1}{\partial y}+\frac{\partial\beta_2}{\partial x}\Big)^2\Big]dx\,dy,\end{aligned} \tag{3.7}$$

$$\lambda = \frac{E\,k}{2(1+\nu)} \tag{3.8}$$

and k is a correction factor which is often used to account for the "nonconformity" of (3.4). Indeed from (3.1)–(3.4) we deduce that σ_{13} and σ_{23} are *constants* in z, whereas the physical problem has $\sigma_{13} = \sigma_{23} = 0$ on the upper and lower face of the plate: $\Omega \times \{t\}$ and $\Omega \times \{-t\}$; hence (3.4) is often corrected by assuming that σ_{13} and σ_{23} behave parabolically in z, vanishing for $z = \pm t$, and assuming the value (3.4) for $z = 0$. To tell the truth, this explanation is not 100% satisfactory for our mathematical minds. However, after all, it is not our business here to discuss the validity of a model as far as its application gives answers that are considered good enough by engineers. We however refer to DESTUYNDER [A] and CIARLET [B] for a precise discussion. The assumed boundary conditions lead to the kinematic constraints

$$\beta_1 = \beta_2 = w = 0 \quad \text{on } \partial\Omega. \tag{3.9}$$

Hence, we define the spaces

$$H = (H_0^1(\Omega))^2; \quad W = H_0^1(\Omega); \quad V = H \times W; \tag{3.10}$$

the generic element of V will be denoted $\underline{v} = (\underline{\eta}, \zeta)$ with $\underline{\eta} = (\eta_1, \eta_2) \in H$ and $\zeta \in W$. We finally recall the Korn's inequality,

$$\exists \alpha > 0 \text{ such that } a(\underline{\eta}, \underline{\eta}) \geq \alpha \, \|\underline{\eta}\|_1^2, \qquad \forall \underline{\eta} \in H. \tag{3.11}$$

It is easy to check that, for any fixed $t > 0$, functional (3.5) has a unique minimizer $(\underline{\beta}, w)$ on V which satisfies

$$t^3 \, a(\underline{\beta}, \underline{\eta}) + \lambda t \int_\Omega (\underline{\text{grad}}\, w - \underline{\beta}) \cdot \underline{\eta} \, dx\, dy = 0, \quad \forall \underline{\eta} \in H, \tag{3.12}$$

$$\lambda t \int_\Omega (\underline{\text{grad}}\, w - \underline{\beta}) \cdot \underline{\text{grad}}\, \zeta \, dx\, dy = \int_{\Omega \times]-t,t[} f_3 \, \zeta \, dx\, dy\, dz, \quad \forall \zeta \in W. \tag{3.13}$$

In particular, we have

$$\frac{t^3}{2} \, a(\underline{\eta}, \underline{\eta}) + \frac{\lambda t}{2} \int_\Omega |\underline{\text{grad}}\, \zeta - \underline{\eta}|^2 dx\, dy \geq c(t) \, (\|\underline{\eta}\|_1^2 + \|\zeta\|_1^2) \tag{3.14}$$

for any $\underline{v} = (\underline{\eta}, \zeta) \in V$. Note that for fixed t, (3.14) always guarantees that (3.12), (3.13) is a nice linear elliptic problem, so that, for instance, any reasonable conforming approximation of V will have optimal order of convergence.

The troubles start when we take a *small* t; then the constant in (3.14) deteriorates and so does the constant in front of the optimal error bound. In practice,

it is well known that if we use "any reasonable conforming approximation of V", we will get pretty bad answers for small t. Here we shall make an analysis of the nature of the trouble. We shall also give some sufficient conditions on the discretization so that it stays good for t smaller and smaller. The one dimensional case was treated by ARNOLD [A], but the two dimensional case, as we shall see, is more complicated.

The first thing that we have to do is to construct a *sequence* of physical problems $\mathcal{P}_t$ $(t > 0)$ that fulfill the following requirements

(1) each $\mathcal{P}_t$ is of type (3.12), (3.13) and so has a unique solution $\underline{\beta}(t)$, $w(t)$;

(2) there exists two constants c_1, c_2 with $0 < c_1 < c_2$ such that

$$c_1 \leq \|\underline{\beta}(t)\|_1 + \|w(t)\|_1 \leq c_2. \tag{3.15}$$

A possible answer is to fix Ω, E, and ν, and to choose, for each $t > 0$, the load $f_3(x, y, z)$ of the form

$$f_3(x, y, z) := \frac{t^2}{2} g(x, y) \tag{3.16}$$

with $g(x, y)$ fixed (once and for all) independent of t. It is clear that **(3.16)** implies

$$\int_{\Omega\times]-t,t[} f_3\, w\, dx\, dy\, dz = t^3 \int_\Omega g w\, dx\, dy = t^3 (g, w) \tag{3.17}$$

so that, dividing (3.6) by t^3 , each problem $\mathcal{P}_t$ will amount to minimize, **in V,**

$$\Pi_t = \frac{1}{2} a(\underline{\beta}, \underline{\beta}) + \frac{t^{-2}\lambda}{2} \int |\underline{\text{grad}}\, w - \underline{\beta}|^2\, dx\, dy - (g, w). \tag{3.18}$$

Proposition 3.1: Let $\underline{\beta}(t)$, $w(t)$ be the minimizer of (3.18) in V. Then (3.15) holds with c_1 and c_2 independent of t.

Proof: We obviously have

$$a(\underline{\beta}, \underline{\beta}) + t^{-2}\lambda \|\underline{\text{grad}}\, w - \underline{\beta}\|_0^2 = (g, w). \tag{3.19}$$

Using (3.11) and a little algebra we deduce from (3.19) that

$$\|\underline{\beta}\|_1^2 + \|w\|_1^2 \leq c(\alpha, \lambda) \|g\|_0\, \|w\|_1, \tag{3.20}$$

which implies the boundedness of $\|\underline{\beta}\|_1 + \|w\|_1$ from above. **Then one checks** that

$$\frac{1}{2} a(\underline{\beta}, \underline{\beta}) + \frac{t^{-2}\lambda}{2} \|\underline{\text{grad}}\, w - \underline{\beta}\|_0^2 - (g, w) \leq -c < 0 \tag{3.21}$$

with c independent of t (this is checked easilyy: we minimize Π_t on $V_0 = \{(\underline{\eta}, \zeta) | \underline{\eta} = \underline{\text{grad}}\, \zeta\}$ and we get a negative minimum independent of t). Now from (3.19) and (3.21) we deduce

$$\frac{1}{2} a(\underline{\beta}, \underline{\beta}) + \frac{t^{-2}\lambda}{2} \|\underline{\text{grad}}\, w - \underline{\beta}\|_0^2 \geq c > 0, \tag{3.22}$$

which implies that $\|\underline{\beta}\|_1 + \|w\|_1$ is bounded from below by a positive constant. This completes the proof. □

It will be convenient, in order to carry on the analysis, to introduce the auxiliary variable

$$\underline{\gamma}(t) := \lambda t^{-2}(\underline{\text{grad}}\, w(t) - \underline{\beta}(t)), \tag{3.23}$$

which is related to the shear stresses but does not go to zero with t. We can now write the Euler equations for Π_t in the form

$$a(\underline{\beta}, \underline{\eta}) + (\underline{\gamma}, \underline{\text{grad}}\, \zeta - \underline{\eta}) = (g, \zeta), \quad \forall (\underline{\eta}, \zeta) \in V, \tag{3.24}$$

$$\underline{\gamma} = \lambda t^{-2}(\underline{\text{grad}}\, w - \underline{\beta}). \tag{3.25}$$

This is now taking the form of the abstract problem studied in Chapter II and it is clear that if we are going to find a uniform bound for $\underline{\gamma}(t)$, this will be in the dual space of the space that is the image of V through the mapping

$$B : (\underline{\eta}, \zeta) \longrightarrow (\underline{\text{grad}}\, \zeta - \underline{\eta}). \tag{3.26}$$

In what follows, we are going to use the notation:

$$\underline{\text{rot}} : \phi \longrightarrow \underline{\text{rot}}\, \phi = \left\{ \frac{\partial \phi}{\partial y}, -\frac{\partial \phi}{\partial x} \right\},$$

$$\text{rot} : \underline{\chi} \longrightarrow \text{rot}\, \underline{\chi} = -\frac{\partial \chi_1}{\partial y} + \frac{\partial \chi_2}{\partial x}$$

This is different from the notation we used in other parts of this book, but is more consistent with the current literature for Reissner–Mindlin plates. Note as well that (for the same reason) we are using here (x, y, z) instead of (x_1, x_2, x_3).

Proposition 3.2: The mapping B is surjective from V onto the space $\Gamma = H_0(\text{rot}, \Omega)$ defined by

$$H_0(\text{rot}; \Omega) = \{\underline{\chi} | \underline{\chi} \in (L^2(\Omega))^2,\ \text{rot}\underline{\chi} \in L^2(\Omega),\ \underline{\chi} \cdot \underline{t} = 0 \text{ on } \partial\Omega\} \tag{3.27}$$

$$\|\underline{\chi}\|^2_{H_0(\text{rot},\Omega)} := \|\underline{\chi}\|_0^2 + \|\text{rot}\underline{\chi}\|_0^2 \tag{3.28}$$

(where $\underline{t}$ is the unit tangent to $\partial\Omega$) and admits a continuous lifting.

Proof: We shall show that for every $\underline{\chi} \in H_0(\text{rot};\Omega)$ there exists $(\underline{\eta},\zeta) \in V$ such that

$$\underline{\chi} = \underline{\text{grad}}\,\zeta - \underline{\eta}, \tag{3.29}$$

$$\|\zeta\|_1 + \|\underline{\eta}\|_1 \le c\,\|\underline{\chi}\|_{H_0(\text{rot},\Omega)}. \tag{3.30}$$

For this we first choose $\underline{v} \in (H_0^1)^2$ such that

$$\text{div}\,\underline{v} = -\text{rot}\ \underline{\chi}, \tag{3.31}$$

$$\|\underline{v}\|_1 \le c\|\text{rot}\ \underline{\chi}\|_0, \tag{3.32}$$

this is obviously possible because

$$\int_\Omega \text{rot}\ \underline{\chi}\ dx\,dy = \int_{\partial\Omega} \underline{\chi}\cdot\underline{t}\ ds = 0. \tag{3.33}$$

Then we set

$$\underline{\eta} = (\eta_1,\eta_2) := (-v_2, v_1), \tag{3.34}$$

so that from (3.31) and (3.32) we have

$$\text{rot}\ \underline{\eta} = -\text{rot}\ \underline{\chi}, \tag{3.35}$$

$$\|\underline{\eta}\|_1 \le \|\text{rot}\ \underline{\chi}\|_0 \tag{3.36}$$

Now choose ζ as the unique solution in $H_0^1(\Omega)$ of

$$\Delta\zeta = \text{div}\,\underline{\chi} + \text{div}\,\underline{\eta} \in H^{-1}(\Omega); \tag{3.37}$$

we have, using (3.36) and (3.37)

$$\|\zeta\|_1 \le c\,(\|\text{div}\,\underline{\chi}\|_{-1} + \|\text{div}\,\underline{\eta}\|_{-1}) \le c\,(\|\underline{\chi}\|_0 + \|\text{rot}\ \underline{\chi}\|_0). \tag{3.38}$$

We now have

$$\begin{cases} \text{div}(\underline{\text{grad}}\,\zeta - \underline{\eta}) = \text{div}\ \underline{\chi} \text{ in } \Omega, \\ \text{rot}(\underline{\text{grad}}\,\zeta - \underline{\eta}) = \text{rot}\ \underline{\chi} \text{ in } \Omega, \\ (\underline{\text{grad}}\,\zeta - \underline{\eta})\cdot\underline{t} = \underline{\chi}\cdot\underline{t} = 0 \text{ on } \partial\Omega, \end{cases} \tag{3.39}$$

which easily implies (3.29). On the other hand, (3.30) follows from (3.36) and (3.38). □

From Propositions 3.2 and II.1.2, we have the following result.

Proposition 3.3: Let $(\underline{\beta}, w, \underline{\gamma})$ be the solution of (3.24). Then we have

$$\|\underline{\gamma}(t)\|_{\Gamma'} \leq c, \tag{3.40}$$

where

$$\begin{aligned} \Gamma' &:= (H_0(\mathrm{rot};\Omega))' \\ &= H^{-1}(\mathrm{div};\Omega) \\ &= \{\underline{\gamma} \mid \underline{\gamma} \in (H^{-1}(\Omega))^2,\ \mathrm{div}\,\underline{\gamma} \in H^{-1}(\Omega)\} \end{aligned} \tag{3.41}$$

with the norm

$$\|\underline{\gamma}\|^2_{\Gamma'} := \|\underline{\gamma}\|^2_{-1} + \|\mathrm{div}\,\underline{\gamma}\|^2_{-1}\,. \tag{3.42}$$

Proof: From Propositions 3.2 and II.1.2, we have

$$\inf_{\underline{\chi}\in\Gamma'} \sup_{(\underline{\eta},\zeta)\in V} \frac{\int_\Omega (\underline{\mathrm{grad}}\,\zeta - \underline{\eta}) \cdot \underline{\chi}\, dx\, dy}{\|(\underline{\eta},\zeta)\|_V \|\underline{\chi}\|_{\Gamma'}} \geq c > 0 \tag{3.43}$$

and from (3.43), (3.24), and (3.15) we have (3.40). It remains to check that the norm (3.42) is equivalent to the natural dual norm induced by (3.28)–(3.41), which is an exercise of functional analysis. □

Remark 3.1: In the case of beam problems, the space Γ' is replaced by L^2, which makes things much easier. □

We are now able to make the result of Proposition 3.1 more precise.

Theorem 3.1: Let $(\underline{\beta}(t), w(t), \underline{\gamma}(t))$ be the solution of (3.24), (3.25). Then we have for $t \to 0$

$$\begin{aligned} &\underline{\beta}(t) \rightharpoonup \underline{\beta}_0 \text{ in } (H^1_0(\Omega))^2, \\ &w(t) \rightharpoonup w_0 \text{ in } H^1_0(\Omega), \\ &\underline{\gamma}(t) \rightharpoonup \underline{\gamma}_0 \text{ in } \Gamma', \end{aligned} \tag{3.44}$$

where $\underline{\beta}_0, w_0, \underline{\gamma}_0$ satisfy

$$a(\underline{\beta}_0, \underline{\eta}) + \langle \underline{\gamma}_0, \underline{\mathrm{grad}}\,\zeta - \underline{\eta}\rangle = (g, \zeta), \quad \forall (\underline{\eta}, \zeta) \in V, \tag{3.45}$$

$$\underline{\beta}_0 = \underline{\mathrm{grad}}\, w_0, \tag{3.46}$$

$$E\Delta^2 w_0 = 12(1-\nu^2)g. \tag{3.47}$$

Proof: The weak convergences (a priori, up to a subsequence) in (3.44) just follow from (3.15) and (3.40). A passage to the limit in (3.24) gives (3.45), whereas (3.46) follows from (3.25). Now putting (3.46) into (3.45) and using (3.7) yields (3.47). □

Remark 3.2: Additional results in this direction can be found in DESTUYNDER [A]. □

We can now apply the result of Proposition II.4.3 to estimate the convergence rate as a function of t^2 which plays here the role of ε. This leads us to a convergence rate in $\sqrt{\varepsilon} = t$. In order to improve this bound and also to enable us later to get sharper error estimates, we now introduce a decomposition principle for (3.24) and (3.25). We shall first prove

Proposition 3.4: Every element $\underline{\gamma} \in \Gamma'$ can be written in a unique way as

$$\underline{\gamma} = \underline{\mathrm{grad}}\, \psi + \underline{\mathrm{rot}}\, p \tag{3.48}$$

with $\psi \in H_0^1(\Omega)$, $p \in L^2(\Omega)/\mathbb{R}$, and $\underline{\mathrm{rot}}\, p = (\partial p/\partial y, -\partial p/\partial x)$. Moreover, we may use

$$\|\underline{\gamma}\|_{\Gamma'}^2 = \|\psi\|_{H_0^1(\Omega)}^2 + \|p\|_{L^2(\Omega)/\mathbb{R}}^2 \tag{3.49}$$

as a norm on Γ'.

Proof: Set $\xi = \mathrm{div}\, \underline{\gamma} \in H^{-1}(\Omega)$. We define ψ to be the unique solution of $-\Delta\psi = \xi$, $\psi \in H_0^1(\Omega)$, and we set $\underline{\alpha} = \underline{\gamma} - \underline{\mathrm{grad}}\, \psi$. One has $\mathrm{div}\, \underline{\alpha} = 0$ so that $\underline{\alpha} = \underline{\mathrm{rot}}\, p$ and p is determined up to a constant in $L^2(\Omega)$. Condition (3.49) is then immediate. □

Remark 3.3: The decomposition introduced in Proposition 3.4 also holds for $(L^2(\Omega))^2$, $H(\underline{\mathrm{rot}};\Omega)$, and $H_0(\underline{\mathrm{rot}};\Omega)$, the difference between these spaces being in the regularity of the p component. Indeed, taking $\underline{\gamma} = \underline{\mathrm{grad}}\, \psi + \underline{\mathrm{rot}}\, p$ with $\psi \in H_0^1(\Omega)$, we have

$$\underline{\gamma} \in (L^2(\Omega))^2 \Leftrightarrow p \in H^1(\Omega)/\mathbb{R}, \tag{3.50}$$

$$\underline{\gamma} \in H(\mathrm{rot};\Omega) \Leftrightarrow p \in H^2(\Omega)/\mathbb{R}, \tag{3.51}$$

$$\underline{\gamma} \in H_0(\mathrm{rot};\Omega) \Leftrightarrow p \in H^2(\Omega)/\mathbb{R} \text{ and } \frac{\partial p}{\partial n} = 0 \text{ on } \partial\Omega. \tag{3.52}$$ □

It is now a simple exercise to transform problem (3.24), (3.25). We write (3.25) in $(L^2(\Omega))^2$ using $\underline{\gamma} = \underline{\mathrm{grad}}\, \psi + \underline{\mathrm{rot}}\, p$ and we multiply by suitable test functions. We then get the following proposition:

Proposition 3.5: Any solution of (3.24) and (3.25) is a solution of the following problem (and conversely):

$$(\underline{\text{grad}}\,\psi(t), \underline{\text{grad}}\,\xi) = (g, \xi), \quad \forall \xi \in H_0^1(\Omega), \tag{3.53}$$

$$\begin{cases} a(\underline{\beta}(t), \underline{\eta}) - (\underline{\text{rot}}\ p(t), \underline{\eta}) = (\underline{\text{grad}}\,\psi(t), \underline{\eta}), & \forall \underline{\eta} \in (H_0^1(\Omega))^2, \\ -(\underline{\beta}(t), \underline{\text{rot}}\ q) = \dfrac{t^2}{\lambda}(\underline{\text{rot}}\ p(t), \underline{\text{rot}}\ q), & \forall q \in H^1(\Omega)/\mathbb{R}, \end{cases} \tag{3.54}$$

$$\begin{aligned} (\underline{\text{grad}}\, w(t), \underline{\text{grad}}\,\chi)) = (\underline{\beta}(t), \underline{\text{grad}}\,\chi)) + t^2(\underline{\text{grad}}\,\psi(t), \underline{\text{grad}}\,\chi), \\ \forall \chi \in H_0^1(\Omega). \ \square \end{aligned} \tag{3.55}$$

It must be noted that (3.54) implies $\partial p/\partial n|_{\partial\Omega} = 0$ and $p \in H^2(\Omega)$ so that $\underline{\gamma} = \underline{\text{grad}}\ \psi + \underline{\text{rot}}\ p$ is indeed an element of $\Gamma = H_0(\underline{\text{rot}}; \Omega)$. Note also that $\psi(t)$ is actually *independant* of t.

We have thus reduced, through Proposition 3.5 our original problem to the following sequence

— a Dirichlet problem (3.53) that is independent of t,

— a "Stokes-like" problem (3.54),

— a Dirichlet problem (3.55).

This decomposition shows us that it is the p component of $\underline{\gamma}$ which depends on t. Before coming back to the quantification of this dependency, we rapidly develop the analogy between (3.54) and a Stokes problem. Let us set $\underline{\eta}^\perp = \{-\eta_2, \eta_1\}$. We can write (3.54) in the form

$$\begin{cases} a(\underline{\beta}^\perp, \underline{\eta}^\perp) + (p, \text{div}\ \underline{\eta}^\perp) = \lambda\ (\underline{\text{grad}}\ \psi, \underline{\eta}^\perp), & \forall \underline{\eta}^\perp \in (H_0^1(\Omega))^2, \\ (\text{div}\ \underline{\beta}^\perp, q) = t^2\ (\underline{\text{grad}}\ p, \underline{\text{grad}}\ q), & \forall q \in H^1(\Omega)/\mathbb{R}. \end{cases} \tag{3.56}$$

The limit problem ($t = 0$) is, thus, a standard Stokes problem and we shall be able to rely on results of Chapter VI to build approximations. We shall not analyze here the case $t \neq 0$ in too much detail . However, it is important to see the behavior of p as $t \to 0$.

Proposition 3.6: Let $\underline{\beta}(t)$, $w(t)$, $p(t)$ and ψ be the solution of (3.53)–(3.55). We then have

$$\|\underline{\beta}(t)\|_2 + \|w(t)\|_2 + \|\psi(t)\|_2 + \|p(t)\|_1 + t\ \|p(t)\|_2 \leq c\ \|g\|_0, \tag{3.57}$$

where the constant c is independent of t. $\square$

We refer to BREZZI–FORTIN [A] for the proof of this result which is based essentially on the regularity properties of the Dirichlet problem and the Stokes problem. □

An important point is that (3.57) does not improve for a more regular g (even in a smooth domain). It is not possible to bound $\|p(t)\|_2$ uniformly in t. The reason is that the normal derivative of $p(t)$ vanishes although this is not the case for the solution $p(0)$ of the limit problem. We, thus, have a boundary layer effect which has been studied by ARNOLD–FALK [C]. Their analysis shows that an analogue of (3.57) exists for $\|\underline{\beta}\|_{5/2}$ and $\|p\|_{3/2/R}$ but not for more regular spaces.

Remark 3.4: We can now try to apply Remarks II.4.4 and II.4.5 to our problem. Denoting $W_+ = \{p \mid p \in H^2(\Omega)/\mathbb{R},\ \partial p/\partial n|_{\partial\Omega} = 0\}$, it is clear that we have

$$|(\underline{\mathrm{rot}}\ p, \underline{\mathrm{rot}}\ q)| \le c\ \|p\|_{W_+}\ |q\|_{L^2(\Omega)/\mathbb{R}}. \tag{3.58}$$

Whenever the solution $p(0)$ of the limit problem is regular enough (this is the case for smooth data and a smooth domain) we shall have,

$$p(0) \in [L^2(\Omega), W_+]_\theta, \quad \forall\theta < \tfrac{3}{4}. \tag{3.59}$$

No improvement is possible because of the fact that $\partial p(0)/\partial n \neq 0$. We can thus apply Remark II.4.5 to get for $\theta < \frac{3}{4}$

$$\|\underline{\beta}(t) - \underline{\beta}(0)\|_1 + \|p(t) - p(0)\|_0 + \|w(t) - w(0)\|_1 \le ct^{2\theta}\ \|p(0)\|_\theta, \tag{3.60}$$

where $\|p(0)\|_\theta$ is the norm of $p(0)$ in $[L^2(\Omega), W_+]_\theta$. We can summarize (3.60) by saying that we have an $O(t^{3/2-\varepsilon})$ convergence. This requires, however, a smooth domain. In the case where $\partial\Omega$ is only Lipschitzian, the best we can get is $O(t)$. □

VII.3.2 Discretization of the problem

We now turn our attention to the discretization of our problem (3.24), (3.25). Let us assume that we are given finite-dimensional subspaces H_h and W_h of H and W and use $V_h = H_h \times W_h$ as a subspace of V. We also discretize the space $\Gamma = H_0(\underline{\mathrm{rot}};\Omega)$ by Γ_h and we consider the discretized problem:find $(\underline{\beta}^h, w_h, \underline{\gamma}_h)$ such that

$$\begin{cases} a(\underline{\beta}_h, \underline{\eta}_h) + (\underline{\gamma}_h, \underline{\mathrm{grad}}\ \zeta_h - \underline{\eta}_h) = (g, \zeta_h), & \forall(\underline{\eta}_h, \zeta_h) \in V_h, \\ (\underline{\mathrm{grad}}\ w_h - \underline{\beta}_h, \underline{\chi}_h) - (t^2/\lambda)\,(\underline{\gamma}_h, \underline{\chi}_h) = 0, & \forall\underline{\chi}_h \in \Gamma_h. \end{cases} \tag{3.61}$$

Note that we *do not have*, in general, $\underline{\gamma}_h = \lambda t^{-2}(\underline{\mathrm{grad}}\ w_h - \underline{\beta}_h)$ unless we take precisely $\Gamma_h = \underline{\mathrm{grad}}\ W_h - H_h$.

We can also introduce the limit problem: find $(\underline{\beta}_{0h}, w_{0h}, \underline{\gamma}_{0h}) \in H_h \times W_h \times \Gamma_h$ such that

$$(3.62)\quad \begin{cases} a(\underline{\beta}_{0h}, \underline{\eta}_h) + (\underline{\gamma}_{0h}, \underline{\mathrm{grad}}\, \zeta_h - \underline{\eta}_h) = (g, \zeta_h), & \forall (\underline{\eta}_h, \zeta_h) \in V_h, \\ (\underline{\chi}_h, \underline{\mathrm{grad}}\, w_{0h} - \underline{\beta}_{0h}) = 0, & \forall \underline{\chi}_h \in \Gamma_h . \end{cases}$$

Thus we have a problem of the form (II.1.5). It also comes from the results of Section II.4 that to get a good approximation of (3.24) and (3.25) by (3.61) (that is, with convergence properties independent of t), it is necessary for (3.62) to be a good approximation of (3.45) and (3.46). Therefore, we should choose H_h, W_h, and Γ_h so that the following properties hold.

First, we must have a coercivenes property:

$$(3.63)\qquad a(\underline{\eta}_h, \underline{\eta}_h) \geq \alpha_0 \left(\|\underline{\eta}_h\|_1^2 + \|\zeta_h\|_1^2\right), \qquad \forall (\underline{\eta}_h, \zeta_h) \in \mathrm{Ker}\, B_h,$$

where, of course

$$(3.64)\qquad \mathrm{Ker}\, B_h = \{(\underline{\eta}_h, \zeta_h) \in V_h \mid (\underline{\chi}_h, \underline{\mathrm{grad}}\, \zeta_h - \underline{\eta}_h) = 0,\ \forall \underline{\chi}_h \in \Gamma_h\}.$$

Condition (3.63) is the usual condition that $a(\cdot,\cdot)$ be coercive on the kernel of B_h. This is, here, a nontrivial condition because $a(\cdot,\cdot)$ is not coercive on V_h as this was the case in the Stokes problem of Chapter VI.

We also know from Chapter II that we must satisfy in our choices of discrete spaces an inf–sup condition

$$(3.65)\qquad \inf_{\underline{\chi}_h \in \Gamma_h} \sup_{(\underline{\eta}_h, \zeta_h) \in V_h} \frac{(\underline{\chi}_h, \underline{\mathrm{grad}}\, \zeta_h - \underline{\eta}_h)}{\|\underline{\chi}_h\|_{\Gamma'}\, \|(\underline{\eta}_h, \zeta_h)\|_V} \geq k_0 > 0.$$

The constants α_0 and k_0 must obviously be independent of h if we want to get the proper result. Now we can start to see why the problem is so nasty. If we use the strategy of making V_h richer in order to satisfy the inf–sup condition (3.65), we will get a large $\mathrm{Ker}\, B_h$ and (3.63) may fail. Making V_h smaller could cure (3.63) but then (3.65) will give us trouble. It is somehow a "short cover" problem: either your arms or your feet will freeze! Moreover, we do not like to use one space with poorer approximation properties than the other, as errors will sum up. Choosing spaces satisfying (3.63)–(3.65) is not an easy task as we shall see below. Let us first apply the results of Chapter II to our case. We first have

Proposition 3.7: If (3.63) holds, then (3.62) has at least one solution and $(\underline{\beta}_{0h}, w_{0h})$ is unique. Moreover if $(\underline{\beta}_0, w_0)$ is the solution of (3.45), (3.46) we have the estimate

$$(3.66)\quad \|\underline{\beta}_0 - \underline{\beta}_{0h}\|_1 + \|w_0 - w_{0h}\|_1 \leq c\Big\{ \inf_{(\underline{\eta}_h, \zeta_h) \in \mathrm{Ker}\, B_h} \left(\|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1\right) + \inf_{\underline{\chi}_h \in \Gamma_h} \|\underline{\gamma}_0 - \underline{\chi}_h\|_{\Gamma'} \Big\}.$$

This is a direct application of Proposition II.2.4. □

We are,therefore, led to the study of $\operatorname{Ker} B_h$ as this will be the key to ellipticiy of the bilinear from $a(\cdot,\cdot)$ and to error estimation. We shall first consider the most "naive" case.

Example 3.1: *The direct approach.*

Let us suppose given $H_h \subset H$ and $W_h \subset W$, and let us choose

$$\Gamma_h = \underline{\text{grad}}(W_h) - H_h \tag{3.67}$$

This choice implies that

$$\operatorname{Ker} B_h = \{(\underline{\eta}_h, \zeta_h) \mid \underline{\eta}_h = \underline{\text{grad}}\, \zeta_h\} \subset \operatorname{Ker} B, \tag{3.68}$$

so that (3.63) evidently holds.

It is important to note that the choice (3.67) is very easy to use on the computer. Actually, in all applications, one deals with a positive thickness t and, therefore, with problem (3.24), (3.25). Now condition (3.67) means that you are actually minimizing the functional Π_t given by (3.18) on $V_h = H_h \times W_h$ and that you *do not even see* $\underline{\gamma}_h$ (nor Γ_h). Condition (3.67) is then one of the most widely used choices for Γ_h, although, in general, one does not realize it.

Now a quick glance to $\operatorname{Ker} B_h$ will make us understand that we have a long way to go. Consider $\underline{\eta}_h^{\perp} = \{-\eta_{2h}, \eta_{1h}\}$, that is, a rotation of $\pi/2$ of $\underline{\eta}_h$. It is clear that if $(\underline{\eta}_h, \zeta_h)$ belongs to $\operatorname{Ker} B_h$, we then have by (3.68)

$$\operatorname{div} \underline{\eta}_h^{\perp} = \operatorname{rot} \underline{\eta}_h = 0. \tag{3.69}$$

Therefore, with choice (3.67), the infimum which appears in (3.66) is actually taken on a subset of functions $\underline{\eta}_h$ satisfying (3.69). But we have already seen in Chapter VI, for the linear Stokes problem, that it is not recommended to work with velocity fields which are exactly incompressible (because there are too few of them in general). A direct application of (3.67) is likely to lead to bad results (e.g., locking) unless a very special choice of H_h and W_h has been done. □

We can also be guided by the decomposition principle of Propositions 3.4 and 3.5 in which a Stokes-like problem explicitly appears. This yields the kind of approximation described in the following.

Example 3.2: *Solution through the decomposition principle.*

We shall, instead of directly approximating $\underline{\gamma}$, explicitly approximate each component of its decomposition into $\underline{\text{grad}}\, \psi_h + \underline{\text{rot}}\, p_h$. Moreover as (3.54) shows us that $\underline{\beta}_h$ and p_h are analogous to a velocity field and a pressure field in a Stokes

problem, we shall try to use some results of Chapter VI to build a suitable approximation.

Let then H_h be built by employing the MINI element of Chapter VI (figure VII.2), that is, in the notation of Chapter III.

$$\begin{cases} H_h &= (\mathfrak{L}_1^1 \cap H_0^1(\Omega))^2 \oplus B_3, \\ W_h &= \mathfrak{L}_1^1 \cap H_0^1(\Omega). \end{cases} \tag{3.70}$$

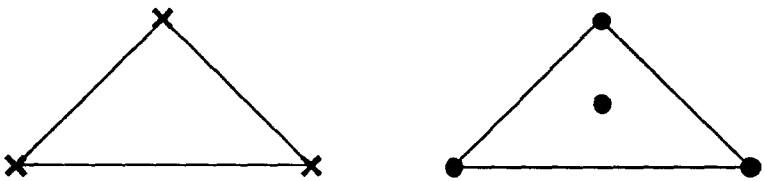


Figure VII.2

These are spaces of piecewise linear polynomials enriched by a bubble function in the case of H_h. We also introduce

$$\Gamma_h = \underline{\mathrm{grad}}(\mathfrak{L}_1^1 \cap H_0^1(\Omega)) \oplus \underline{\mathrm{rot}}\ \mathfrak{L}_1^1. \tag{3.71}$$

This space is then a strict subspace of piecewise constant vector functions for which we imposed, a priori, the analogue of the decomposition principle of Proposition 3.4 and Remark 3.3.

It is then straightforward to check that $\mathrm{Ker}\, B_h$ is made of the pairs $(\underline{\eta}_h, \zeta_h)$ in $H_h \times W_h$ such that,

$$(\underline{\eta}_h, \underline{\mathrm{rot}}\ q_h) = 0, \qquad \forall q_h \in \mathfrak{L}_1^1, \tag{3.72}$$

$$(\underline{\mathrm{grad}}\ \zeta_h, \underline{\mathrm{grad}}\ \phi_h) = (\underline{\eta}_h, \underline{\mathrm{grad}}\ \phi_h), \qquad \forall \phi_h \in \mathfrak{L}_1^1 \cap H_0^1(\Omega). \tag{3.73}$$

Now condition (3.73) is especially nice as it implies

$$\|\zeta_h\|_1 \le c\, \|\underline{\eta}_h\|_1, \qquad \forall (\underline{\eta}_h, \zeta_h) \in \mathrm{Ker}\, B_h, \tag{3.74}$$

and hence, (3.63) holds from (3.11) and we can apply Proposition 3.7. To get a really usable result we, however, need to replace in (3.66) the infimum on $\mathrm{Ker}\, B_h$ by an infimum over the whole space. As we have learned in Chapter II, this amounts to checking the inf–sup condition (3.65) and we can do it using Proposition II.2.8.

Given $(\underline{\beta}, \zeta)$, we must be able to build $\underline{\beta}_h$, ζ_h such that

$$(\underline{\gamma}_h, \underline{\beta}_h - \underline{\mathrm{grad}}\ \zeta_h) - (\underline{\gamma}_h, \underline{\beta} - \underline{\mathrm{grad}}\ \zeta) = 0, \qquad \forall \underline{\gamma}_h \in \Gamma_h, \tag{3.75}$$

$$\|\underline{\beta}_h\|_1 + \|\zeta_h\|_1 \leq c\,(\|\underline{\beta}\|_1 + \|\zeta\|_1), \tag{3.76}$$

where, in (3.76), c is independent of h. But from the construction of Γ_h condition (3.75) can be decomposed into

$$\begin{cases} (\underline{\mathrm{grad}}\ \phi_h, \underline{\beta}_h - \underline{\mathrm{grad}}\ \zeta_h) - (\underline{\mathrm{grad}}\ \phi_h, \underline{\beta} - \underline{\mathrm{grad}}\ \zeta) = 0, \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \forall \phi_h \in \mathfrak{L}_1^1 \cap H_0^1(\Omega), \\ (\underline{\mathrm{rot}}\ p_h, \underline{\beta}_h - \underline{\mathrm{grad}}\ \zeta_h) - (\underline{\mathrm{rot}}\ p_h, \underline{\beta} - \underline{\mathrm{grad}}\ \zeta) = 0, \quad \forall p_h \in \mathfrak{L}_1^1. \end{cases} \tag{3.77}$$

To do such a construction we use the trick already introduced in Chapter VI to deal with the inf–sup condition for the MINI element. We first build $\underline{\beta}_h$ by taking a standard interpolate of $\underline{\beta}$ and then, adjusting the bubble function so that one has for any piecewise constant $\underline{m}_h \in (\mathfrak{L}_0^0)^2$,

$$(\underline{\beta}_h - \underline{\beta}, \underline{m}_h) = 0. \tag{3.78}$$

We have seen in Chapter VI that this can be done in a continuous way, that is,

$$\|\underline{\beta}_h\|_1 < c\,\|\underline{\beta}\|_1. \tag{3.79}$$

By (3.78), the fact that $\underline{\mathrm{grad}}\ \phi_h$ is piecewise constant and that $(\underline{\mathrm{rot}}\ p_h, \underline{\mathrm{grad}}\ \zeta_h) = (\underline{\mathrm{rot}}\ p_h, \underline{\mathrm{grad}}\ \zeta) = 0$, (3.76) reduces to

$$(\underline{\mathrm{grad}}\ \phi_h, \underline{\mathrm{grad}}\ \zeta_h) = (\underline{\mathrm{grad}}\ \phi_h, \underline{\mathrm{grad}}\ \zeta), \qquad \forall \phi_h \in W_h, \tag{3.80}$$

and this is a discrete Dirichlet problem for which we have $\|\zeta_h\|_1 \leq c\,\|\zeta\|_1$, yielding the second part of (3.75). This proves the inf–sup condition and we can, therefore, apply to problem (3.62) the basic results of Chapter II. We can summarize this in the following proposition:

Proposition 3.8: Problem (3.62) with the choice (3.70) and (3.71) has a unique solution. Moreover if $(\underline{\beta}_0, w_0, \underline{\gamma}_0)$ is the solution of (3.45) and (3.46), we have

$$\begin{aligned} \|\underline{\beta}_0 - \underline{\beta}_{0h}\|_1 + \|w_0 - w_{0h}\|_1 + \|\underline{\gamma}_0 - \underline{\gamma}_{0h}\|_{\Gamma'} \\ \leq ch\,\{\|w_0\|_3 + \|\underline{\gamma}_0\|_{H(\mathrm{div};\Omega)}\}. \ \square \end{aligned} \tag{3.81}$$

Remark 3.5: The result of Proposition 3.5 can be applied to the discrete problem in the present case. Indeed, we built, a priori, Γ_h in order to obtain a decomposition principle. Problem (3.61) can be written in the form

$$(\underline{\mathrm{grad}}\ \psi_h, \underline{\mathrm{grad}}\ \zeta_h) = (g, \zeta_h), \qquad \forall \zeta_h \in H_0^1(\Omega) \cap \mathfrak{L}_1^1, \tag{3.82}$$

$$(3.83)\qquad \begin{cases} a(\underline{\beta}_h, \underline{\eta}_h) - (\underline{\mathrm{rot}}\ p_h, \underline{\eta}_h) = (\underline{\mathrm{grad}}\ \psi_h, \underline{\eta}_h), & \forall \underline{\eta}_h \in H_h, \\ -(\underline{\beta}_h, \underline{\mathrm{rot}}\ q_h) = \dfrac{t^2}{\lambda}(\underline{\mathrm{rot}}\ p_h, \underline{\mathrm{rot}}\ q_h), & \forall q_h \in \mathfrak{L}_1^1, \end{cases}$$

$$(3.84)\quad (\underline{\mathrm{grad}}\, w_h, \underline{\mathrm{grad}}\, \chi_h) = (\underline{\beta}_h, \underline{\mathrm{grad}}\, \chi_h) + t^2(\underline{\mathrm{grad}}\, \psi_h, \underline{\mathrm{grad}}\, \chi_h),\ \forall \chi_h \in W_h.$$

These problems can be solved sequentially and (3.83) is a Stokes-like problem using the MINI element of Chapter VI. This approximation has been introduced and studied for $t \neq 0$ in BREZZI–FORTIN [A]. Using this decomposition and Proposition 3.8, recalling that

$$\|\underline{\gamma}\|_{\Gamma'} = \|\psi\|_1 + |p|_{0/\mathbb{R}},$$

and bringing in the regularity result of Proposition 3.6, we have, for $t = 0$, the following estimate:

$$(3.85)\quad \|\psi_{0h} - \psi_0\|_1 + |p_0 - p_{0h}|_{0/\mathbb{R}} \leq ch\ \{\|w_0\|_2 + \|\psi_0\|_2 + \|p_0\|_1\} \leq ch\ \|g\|_0.$$

From a numerical point of view (3.82)–(3.84) can lead to an efficient method, provided one can dispose of a Stokes solver. □

Remark 3.6: An easy duality argument would also show that we have the estimate

$$(3.86)\qquad \|\underline{\beta}_0 - \underline{\beta}_{0h}\|_0 + \|w_0 - w_{0h}\|_0 \leq ch^2\ \{\|w_0\|_3 + \|\underline{\gamma}_0\|_{H(\mathrm{div},\Omega)}\}.\ \square$$

To end this example, we rapidly show how the results of Section II.2.4 can be applied to the case $t \neq 0$. We consider the error estimate (II.2.48) from Remark II.2.13, where we denote $V = (H_0^1(\Omega))^2 \times H_0^1(\Omega)$, $Q = \Gamma'$, and $W = (L^2(\Omega))^2$. The parameter λ is, of course, t^2 in the present case. It is easily verified that all conditions are satisfied and that we have, taking into account regularity properties of Remark 3.3,

$$(3.87)\qquad \begin{aligned} &\|\underline{\beta}(t) - \underline{\beta}_h(t)\|_1^2 + \|w(t) - w_h(t)\|_1^2 + \|\underline{\gamma}(t) - \underline{\gamma}_h(t)\|_{\Gamma'}^2 \\ &\quad + t^2\,\|\underline{\gamma}(t) - \underline{\gamma}_h(t)\|_0^2 \leq \inf_{\underline{\eta}_h} \|\underline{\beta}(t) - \underline{\eta}_h\|_1^2 + \inf_{q_h} \|w(t) - q_h\|_1^2 \\ &\qquad + \inf_{\underline{\delta}_h} \|\underline{\gamma}(t) - \underline{\delta}_h\|_{\Gamma'}^2 + \inf_{\underline{\delta}_h} t^2 \|\underline{\gamma}(t) - \underline{\delta}_h\|_0^2. \end{aligned}$$

Using the decomposition principle and the estimate (3.57) we can find the result of BREZZI–FORTIN [A]:

$$(3.88)\qquad \begin{aligned} &\|\underline{\beta}(t) - \underline{\beta}_h(t)\|_1^2 + \|w(t) - w_h(t)\|_1^2 + \|\psi(t) - \psi_h(t)\|_1^2 \\ &\quad + |p(t) - p_h(t)|_0^2 + t^2\ \|p(t) - p_h(t)\|_1^2 \\ &\leq ch^2\{\|\underline{\beta}(t)\|_2^2 + \|w(t)\|_2^2 + \|\psi(t)\|_2^2 \\ &\quad + |p(t)|_1^2 + t^2\|p(t)\|_2^2\}, \end{aligned}$$

that is, an $O(h)$ convergence. This result cannot be (much) improved because of the boundary layer effect already described. □

The above example is, although interesting, rather remote from the actual engineering practice in which one tries to stick as closely as possible to the original formulation. What we have to avoid is a spurious locking of the solution whenever, for t small, one nearly enforces the condition $(\underline{\text{grad}}\, w - \underline{\beta}) = 0$. As we have seen in Chapter VI, it is not, in general, possible to introduce directly such a condition in a finite element method.

The most common escape from the troubles that we are facing is to use some kind of numerical integration for the term $\lambda t^{-2}\|\underline{\text{grad}}\, w - \underline{\beta}\|^2$ which appears in (3.18), thus weakening condition (3.69). A way of formalizing it is the following. We assume that we are given a linear operator r which maps $H_h \times W_h$ into (for instance) $L^2(\Omega)$. To fix the ideas, let us consider the possible, but not recommended, choice:

$$r(\underline{\eta}, \zeta) \in \mathfrak{L}_0^0 \text{ and } r(\underline{\eta}, \zeta)|_K = \text{ value of } (\underline{\text{grad}}\, \zeta - \underline{\eta}) \text{ at the barycenter of } K. \tag{3.89}$$

Then one minimizes, instead of Π_t (as in (3.18)), the functional

$$M_t^r := \frac{1}{2} a(\underline{\beta}, \underline{\beta}) + \frac{\lambda t^{-2}}{2} \|r(\underline{\beta}, w)\|_0^2 - (g, w) \tag{3.90}$$

on $H_h \times \boldsymbol{W_h}$. This corresponds to the choice

$$\Gamma_h := r(H_h, W_h) \tag{3.91}$$

and leads to the limit problem (for $t = 0$): find $(\underline{\beta}_h, w_h, \underline{\gamma}_h) \in H_h \times W_h \times \Gamma_h$ such that

$$\begin{cases} a(\underline{\beta}_h, \underline{\eta}_h) + (\underline{\gamma}_h, r(\underline{\eta}_h, \zeta_h)) - (g, \zeta_h) = 0, & \forall (\underline{\eta}_h, \zeta_h) \in V_h, \\ (\underline{\chi}_h, r(\underline{\beta}_h, w_h)) = 0, & \forall \underline{\chi}_h \in \Gamma_h \end{cases} \tag{3.92}$$

We then have implicitly defined, as in Section II.2.6, a bilinear form

$$b_h(\{\underline{\eta}_h, \zeta_h\}, \underline{\gamma}_h) = (\underline{\gamma}_h, r(\underline{\eta}_h, \zeta_h)), \tag{3.93}$$

where $\underline{\gamma}_h$ is to be looked for in the range of $r(\underline{\eta}_h, \zeta_h)$. We shall have to characterize this space and to analyze

$$\text{Ker}\, B_h = \{(\underline{\eta}_h, \zeta_h) | r(\underline{\eta}_h, \zeta_h) = 0\}, \tag{3.94}$$

which will enforce $\underline{\eta}_h = \underline{\text{grad}}\, \zeta_h$ only in a weak sense.

Theorem 3.2: Let Ker B_h be defined by (3.94). If (3.63) holds, then (3.92) has for a unique solution $(\underline{\beta}_{0h}, w_{0h})$. Moreover if $(\underline{\beta}_0, w_0)$ is the solution of (3.45) and (3.46) we have:

(3.95)
$$\|\underline{\beta}_0-\underline{\beta}_{0h}\|_1 + \|w_0-w_{0h}\|_1 \le \frac{c}{\alpha}\Big\{\inf_{(\underline{\eta}_h,\zeta_h)\in \operatorname{Ker} B_h} \|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1\Big\} + \sup_{(\underline{\eta}_h,\zeta_h)\in \operatorname{Ker} B_h} \frac{(\underline{\gamma}_0, r(\underline{\eta}_h,\zeta_h)-(\underline{\operatorname{grad}}\,\zeta_h-\underline{\eta}_h))}{\|\zeta_h\|_1 + \|\underline{\eta}\|_1},$$

where c is independent of h and α is given by (3.63) (and, a priori, might depend on h).

Proof: We have for all $(\underline{\eta}_h, \zeta_h) \in \operatorname{Ker} B_h$,

$$\begin{aligned}\alpha(\|\underline{\eta}_h - \underline{\beta}_{0h}\|_1^2 + \|\zeta_h - w_{0h}\|_1^2) &\le a(\underline{\eta}_h - \underline{\beta}_{0h}, \underline{\eta}_h - \underline{\beta}_{0h})\\ &= a(\underline{\eta}_h-\underline{\beta}_0, \underline{\eta}_h-\underline{\beta}_{0h}) + a(\underline{\beta}_0-\underline{\beta}_{0h}, \underline{\eta}_h-\underline{\beta}_{0h})\\ &\le \|\underline{\eta}_h - \underline{\beta}_0\|_1\, \|\underline{\eta}_h - \underline{\beta}_{0h}\|_1\\ &\qquad + (\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h}) - (g, \zeta_h - w_{0h}).\end{aligned}$$

Moreover,

$$(\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h}) = (\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h} - \underline{\operatorname{grad}}\,(\zeta_h - w_{0h})) + (g, \zeta_h - w_{0h}),$$

so that

$$\begin{aligned}(\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h}) - (g, \zeta_h - w_{0h}) &= (\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h} - \underline{\operatorname{grad}}\,(\zeta_h - w_{0h}))\\ &= (\underline{\gamma}_0, \underline{\eta}_h - \underline{\beta}_{0h} - \underline{\operatorname{grad}}(\zeta_h - w_{0h}) + r(\underline{\eta}_h - \underline{\beta}_{0h}, \zeta_h - w_{0h}))\end{aligned}$$

and (3.95) follows easily. □

The above result does not require that Γ_h, the range of $r(\underline{\eta}_h, \zeta_h)$, be known explicitly. On the other hand, it does not yield an estimate on the multiplier $\underline{\gamma}_{0h}$ and requires one to build an error estimate in Ker B_h. As we know from the results of Chapter II, the cure to this is to suppose an inf–sup condition. From Section II.2.6, in particular Proposition II.2.16 and the remarks that follow, we have the following result.

Theorem 3.3: If (3.63) holds and if, moreover, $b_h(\cdot,\cdot)$ satisfies an inf–sup condition, that is

$$\inf_{\underline{\chi}_h\in\Gamma_h}\ \sup_{(\underline{\eta}_h,\zeta_h)\in V_h} \frac{(\underline{\chi}_h, r(\underline{\eta}_h,\zeta_h))}{\|\underline{\chi}_h\|_{\Gamma'}\, \|(\underline{\eta}_h,\zeta_h)\|_V} \ge k_0 > 0, \tag{3.96}$$

then (3.92) has a unique solution $(\underline{\beta}_{0h}, w_{0h}, \underline{\gamma}_{0h})$. Moreover, if $(\underline{\beta}_0, w_0, \underline{\gamma}_0)$ is the solution of (3.45) and (3.46), we have

$$
\begin{aligned}
\|\underline{\beta}_0 - \underline{\beta}_{0h}\|_1 + \|w_0 - w_{0h}\|_1 + \|\underline{\gamma}_0 - \underline{\gamma}_{0h}\|_{\Gamma'} \\
\le c\Big\{ & \inf_{(\underline{\eta}_h, \zeta_h) \in V_h} (\|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1) \\
& + \inf_{\underline{\delta}_h \in \Gamma_h} \|\underline{\gamma}_0 - \underline{\delta}_h\|_{\Gamma'} \\
& + \sup_{(\underline{\eta}_h, \zeta_h) \in V_h} \frac{(\underline{\gamma}_0, r(\underline{\eta}_h, \zeta_h) - (\underline{\mathrm{grad}}\, \zeta_h - \underline{\eta}_h))}{\|\zeta_h\|_1 + \|\underline{\eta}_h\|_1} \\
& + \sup_{\underline{\chi}_h \in \Gamma_h} \frac{(\underline{\chi}_h, r(\underline{\beta}_0, w_0) - (\underline{\mathrm{grad}}\, w_0 - \underline{\beta}_0))}{\|\underline{\chi}_h\|_{\Gamma'}} \Big\}. \quad \square
\end{aligned}
\tag{3.97}
$$

This is nothing but Proposition II.2.16 using (II.2.75) and (II.2.76). To be able to use this estimate we should be able to do some extra work, namely, ,

— make explicit the space Γ_h;

— check the inf–sup condition (3.96);

— properly estimate the two consistency terms of (3.97).

Before considering an example where part of this can be done, some remarks have to be made.

At least formally, formulation (3.92), for this limit problem, includes all the previous ones. In particular, any choice of Γ_h in (3.62) can be interpreted in the form (3.92) by defining

$$
r(\underline{\eta}, \zeta) = P_h(\underline{\mathrm{grad}}\, \zeta - \underline{\eta}); \quad P_h := \text{ projection onto } \Gamma_h. \tag{3.98}
$$

However, if one starts with a Γ_h independently assumed, it will usually be simpler to use directly the approach (3.62) instead of using (3.92) and (3.98).

Let us try to summarize the few results that we have obtained so far. For any given choice of V_h (that is, H_h and W_h) we have to make a decision on the treatment of the term $\lambda t^{-2}\|\underline{\mathrm{grad}}\, w - \underline{\beta}\|_0^2$ in (3.18). We can choose to introduce the space Γ_h and to look for the additional variable $\underline{\gamma}_h$ in Γ_h (as in (3.61)) or we may choose to make use of a reduction operator r as in (3.90). The study of the discrete kernel is a crucial step in the analysis. Then we have to

— prove that (3.63) holds in $\operatorname{Ker} B_h$ possibly with α independent of h;

— estimate $\inf_{(\underline{\eta}_h, \zeta_h) \in \operatorname{Ker} B_h} (\|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1)$ (and possibly the other term in (3.95)).

Once this is done, Theorems 3.2 and 3.3 will provide the error estimate. We recall once more that if $(\underline{\eta}_h, \zeta_h) \in \operatorname{Ker} B_h$, then $\operatorname{div} R\underline{\eta}_h \simeq 0$ and $\underline{\eta}_h \simeq \underline{\mathrm{grad}}\, \zeta_h$,

where $R\underline{\eta} = \underline{\eta}^{\perp}$ is a rotation of $\pi/2$ of $\underline{\eta}$ as in (3.69). Hence, *the analysis of Ker B_h will be, in general, related to some discrete solution of a problem of Stokes type.*

Example 3.3: *A "reduced integration" approximation.*

We consider now an example of application of formulation (3.92). We assume that Ω is a convex polygon and that we are given a sequence $\{\mathcal{T}_h\}$ of partitions of Ω into triangles. We set, with the notation of Chapter III,

$$tH_h = (\mathfrak{L}_2^1 \oplus B_3)^2; \qquad W_h = \mathfrak{L}_1^1 \tag{3.99}$$

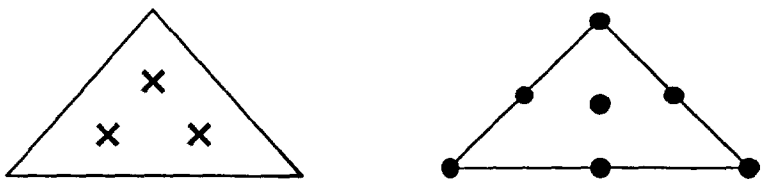


Figure VII.3

One then sees that H_h is similar to the Crouzeix–Raviart element of Chapter VI (Figure VII.3). We now have to define the operator r. We set

$$\Gamma_h = \{\underline{\gamma}_h \mid \underline{\gamma}_h \in R(RT_1)\}, \tag{3.100}$$

where R is, as previously, the $\pi/2$ rotation operator. Note that Γ_h is now an approximation of $H_0(\underline{\mathrm{rot}};\Omega)$, whereas RT_1 was an approximation of $H(\mathrm{div};\Omega)$, and that the *tangential components* are continuous across the interelement boundaries instead of the normal components. Now given $\underline{\eta}_h \in H_h$, we may consider its interpolant $\Pi_h\underline{\eta}_h = \underline{\tilde{\eta}}_h$ by means of

$$\int_e (\underline{\eta}_h - \underline{\tilde{\eta}}_h)\cdot \underline{t}\,\mu_1\,d\sigma = 0, \qquad \forall \mu_1 \in L_1^0(\mathcal{E}_h), \tag{3.101}$$

$$\int_K (\underline{\eta}_h - \underline{\tilde{\eta}}_h)\,dx = 0, \qquad \forall K \in \mathcal{T}_h. \tag{3.102}$$

We are now able to define the operator r by

$$r(\underline{\eta}_h, \zeta_h) = \underline{\mathrm{grad}}\,\zeta_h - \Pi_h\underline{\eta}_h \in \Gamma_h. \tag{3.103}$$

We remark that, using essentially Proposition III.3.8, one can easily check that rot $\underline{\tilde{\eta}}_h$ is the L^2-projection of rot $\underline{\eta}_h$ onto $\mathfrak{L}_1^0(\mathcal{T}_h)$. Hence, the kernel of B_h as defined in (3.94) is now easily characterized as the set of $(\underline{\eta}_h, \zeta_h)$ such that

$$\underline{\tilde{\eta}}_h = \underline{\mathrm{grad}}\,\zeta_h, \tag{3.104}$$

$$\int_\Omega (\text{rot}\, \underline{\eta}_h)\, q_h\, dx = 0, \qquad \forall q_h \in \mathcal{L}_1^0(\mathcal{T}_h). \tag{3.105}$$

Since $\|\tilde{\underline{\eta}}_h\|_0 \le \|\underline{\eta}_h\|_1$, we clearly have

$$a(\underline{\eta}_h, \underline{\eta}_h) \ge c\,(\|\underline{\eta}_h\|_1 + \|\zeta_h\|_1), \qquad \forall (\underline{\eta}_h, \zeta_h) \in \text{Ker}\, B_h, \tag{3.106}$$

that is, condition (3.63). We can now apply Theorem 3.2 but for this we need a result for Ker B_h. This is, however, a direct consequence of estimates obtained in Chapter VI for the Crouzeix–Raviart element and we have

Proposition 3.9: Let $\underline{\beta}_0, w_0$ be the solution of (3.45), (3.46) and let Ker B_h be given by (3.104), (3.105). Then we have

$$\begin{aligned} &\inf_{(\underline{\eta}_h, \zeta_h) \in \text{Ker}\, B_h} (\|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1) \\ &\qquad \le C \inf_{\substack{\underline{\theta}_h \in H_h \\ \phi_h \in W_h}} (\|\underline{\beta}_0 - \underline{\theta}_h\|_1 + \|w_0 - \phi_h\|_1). \end{aligned} \tag{3.107}$$

Proof: $\underline{\beta}_0 = \underline{\text{grad}}\, w_0$ and, thus, rot $\underline{\beta}_0 = 0$. From Chapter VI we know that we can approximate it to the right order with the Crouzeix–Raviart element by $\underline{\eta}_h$ satisfying (3.105). But then $\tilde{\underline{\eta}}_h$ is equal to $\underline{\text{grad}}\, \zeta_h$ and the result follows from the continuity of the operator Π_h. □

To apply Theorem 3.2, we must, however, also bound the consistency term of (3.95) that is given by

$$\sup_{(\underline{\eta}_h, \zeta_h) \in V_h} \frac{(\underline{\gamma}_0, r(\underline{\eta}_h, \zeta_h) - (\underline{\text{grad}}\, \zeta_h - \underline{\eta}_h))}{\|\zeta_h\|_1 + \|\underline{\eta}_h\|_1}. \tag{3.108}$$

In the present case we have

$$(\underline{\gamma}_0, r(\underline{\eta}_h, \zeta_h) - (\underline{\text{grad}}\, \zeta_h - \underline{\eta}_h)) = (\underline{\gamma}_0, \underline{\eta}_h - \Pi_h \underline{\eta}_h). \tag{3.109}$$

Using (3.102), we also see that we can write for any $\underline{\eta}_h$

$$(\underline{\gamma}_0, \underline{\eta}_h - \Pi_h \underline{\eta}_h) = (\underline{\gamma}_0 - \bar{\underline{\gamma}}_0, \underline{\eta}_h - \Pi_h \underline{\eta}_h), \tag{3.110}$$

where $\bar{\underline{\gamma}}_0$ is the projection of $\underline{\gamma}_0$ onto $(\mathcal{L}_0^0)^2$. From (3.110) we obtain

$$\begin{aligned} (\underline{\gamma}_0, \underline{\eta}_h - \Pi_h \underline{\eta}_h) &\le \inf_{\bar{\underline{\gamma}}_0 \in (\mathcal{L}_0^0)^2} \|\underline{\gamma}_0 - \bar{\underline{\gamma}}_0\|_0 \|\underline{\eta}_h - \Pi_h \underline{\eta}_h\|_0 \\ &\le c_1 h\, \|\underline{\gamma}_0\|_1 c_2 h\, \|\underline{\eta}_h\|_1, \end{aligned} \tag{3.111}$$

so that finally we have

$$\sup_{(\underline{\eta}_h,\zeta_h)\in V_h} \frac{(\underline{\gamma}_0, r(\underline{\eta}_h,\zeta_h) - (\underline{\mathrm{grad}}\,\zeta_h - \underline{\eta}_h))}{\|\zeta_h\|_1 + \|\underline{\eta}_h\|_1} \le ch^2\,\|\underline{\gamma}_0\|_1. \tag{3.112}$$

Now to finish our error estimate we have to consider the approximation errors involved in (3.107). There are two of them and they are straightforward from the results of Chapter III. Indeed we can immediately write

$$\begin{aligned}\inf_{(\underline{\eta}_h,\zeta_h)\in V_h} (\|\underline{\beta}_0 - \underline{\eta}_h\|_1 + \|w_0 - \zeta_h\|_1) &\le ch^2\,(\|w_0\|_3 + \|\underline{\beta}_0\|_3)\\ &\le ch^2\,\|\underline{\beta}_0\|_3.\end{aligned} \tag{3.113}$$

We can therefore summarize the previous results in

Theorem 3.4: Let $(\underline{\beta}_0, w_0)$ be the solution of (3.45), (3.46) and $(\underline{\beta}_{0h}, w_{0h})$ be the solution of (3.92) obtained by (3.99)–(3.103). We then have

$$\begin{aligned}&\|\underline{\beta}_0 - \underline{\beta}_{0h}\|_1 + \|w_0 - w_{0h}\|_1\\ &\qquad \le C_1 h^2\,(\|w_0\|_3 + \|\underline{\beta}_0\|_3) + C_2 h^2\,\|\underline{\gamma}_0\|_1 \quad \square\end{aligned} \tag{3.114}$$

As we already noted, this estimate is overoptimistic because it ignores the boundary layer effects. From the results of ARNOLD–FALK [C] an $O(h^{3/2})$ convergence rate should be expected.

Remark 3.7: Similar estimates have been obtained in BATHE–BREZZI–FORTIN [A] for the presently discussed element and related ones, including elements defined on quadrilaterals. More refined estimates can be found in BREZZI–FORTIN–STENBERG [A]. □

Remark 3.8: The choice of second-order accuracy has been done only for the sake of simplicity. Higher-order element are possible and we shall indicate at the end of this chapter a general framework within which they could be built. On the contrary, lower-order elements are more difficult to get; see, for instance, BATHE–BREZZI [A] for the convergence analysis of a similar method, which is only $O(h)$ accurate (HUGHES–TEZDUYAR [A], BATHE–DVORKIN [A]). We also refer to BATHE–BREZZI [A], BREZZI–FORTIN [A], ARNOLD–FALK [B] and PITKÄRANTA [A] for other examples. □

Remark 3.9: It is possible to use a duality argument to get an $O(h^3)$ estimate for $\|\underline{\beta}_0 - \underline{\beta}_{0h}\|_0$ and $\|w_0 - w_{0h}\|_0$. See BREZZI–FORTIN–STENBERG [A]. □

The above result, although interesting, still leaves part of the question unanswered as we have no estimate on the shear stress $\underline{\gamma}_h$. Moreover, we would like to be able to check that we indeed have an uniform bound in $t > 0$. Theorem 3.3 would be a first step in that direction and we, therefore, should try to obtain some inf–sup condition.

In order to do so, we shall need to use the following result on the structure of Γ_h, which is in fact nothing but the discrete analogue of Proposition 3.4.

Proposition 3.10: For any $\underline{\gamma}_h \in \Gamma_h$, there exists $\phi_h \in \mathcal{L}_2^1 \cap H_0^1(\Omega) = W_h$ and $p_h \in \mathcal{L}_1^0/\mathbb{R} = Q_h$ such that

$$(\underline{\gamma}_h, \underline{\delta}_h)_0 = (\underline{\mathrm{grad}}\, \phi_h, \underline{\delta}_h) + (p_h, \mathrm{rot}\, \underline{\delta}_h), \qquad \forall \underline{\delta}_h \in \Gamma_h. \tag{3.115}$$

Proof: Given $\underline{\gamma}_h$ we solve in Γ_h a mixed problem similar to those studied in Chapter V, namely,

$$\begin{cases} (\underline{\alpha}_h, \underline{\delta}_h) - (p_h, \mathrm{rot}\, \underline{\delta}_h) = 0, & \forall \underline{\delta}_h \in \Gamma_h, \\ (\mathrm{rot}\, \underline{\alpha}_h, q_h) = (\mathrm{rot}\, \underline{\gamma}_h, q_h), & \forall q_h \in Q_h. \end{cases} \tag{3.116}$$

In fact, this is a "$\pi/2$ rotation" of the problem of Chapter V and we can find $\underline{\alpha}_h$ and p_h (up to an additive constant) solution of this problem. Now, $\mathrm{rot}(\underline{\gamma}_h - \underline{\alpha}_h) = 0$ and Corollary III.3.2 enables us to write $\underline{\gamma}_h - \underline{\alpha}_h = \underline{\mathrm{grad}}\, \phi_h$, which completes the proof. □

This result will later lead us to a decomposition principle analogous to (3.82)–(3.84). But we shall first proceed to prove

Proposition 3.11: The operator r defined by (3.103) is surjective on Γ_h. Given $\underline{\gamma}_h \in \Gamma_h$, one can find $\underline{\beta}_h \in H_h$ and $w_h \in W_h$ with

$$\begin{cases} r(\underline{\beta}_h, w_h) = \underline{\gamma}_h, \\ \|\underline{\beta}_h\|_1 + \|w_h\|_1 \le c\,(\|\underline{\gamma}_h\|_0 + \|\mathrm{rot}\, \underline{\gamma}_h\|_0) \le c\, \|\underline{\gamma}_h\|_{H_0(\mathrm{rot},\Omega)} \end{cases} \tag{3.117}$$

with a constant c independent of h.

Proof: Given $\underline{\gamma}_h$, we first solve a "Stokes-like" problem: find $(\underline{\beta}_h, k_h) \in H_h \times Q_h$ such that

$$\begin{cases} a(\underline{\beta}_h, \underline{\eta}_h) - (k_h, \mathrm{rot}\, \underline{\eta}_h) = 0, & \forall \underline{\eta}_h \in H_h, \\ (\mathrm{rot}\, \underline{\beta}_h, q_h) = (\mathrm{rot}\, \underline{\gamma}_h, q_h), & \forall q_h \in Q_h. \end{cases} \tag{3.118}$$

The couple $H_h \times Q_h$ is the Crouzeix–Raviart element for Stokes problem. This makes (3.118) well posed and there exists a unique $\underline{\beta}_h$. It satisfies

$$\|\underline{\beta}_h\|_1 \le c\,\|\mathrm{rot}\underline{\gamma}_h\|_0 \tag{3.119}$$

and

$$\mathrm{rot}(\Pi_h\underline{\beta}_h) = \mathrm{rot}\ \underline{\gamma}_h. \tag{3.120}$$

Thus, $\mathrm{rot}(\underline{\gamma}_h - \Pi_h\underline{\beta}_h) = 0$ and we can find (again from Corollary III.3.2) $w_h \in W_h$ such that $\underline{\gamma}_h - \Pi_h\underline{\beta}_h = \underline{\mathrm{grad}}\ w_h$. In fact, w_h can be obtained by solving the discrete Dirichlet problem

$$(\underline{\mathrm{grad}}\ w_h, \underline{\mathrm{grad}}\ \phi_h) = (\underline{\gamma}_h - \Pi_h\underline{\beta}_h, \underline{\mathrm{grad}}\ \phi_h), \qquad \forall \phi_h \in W_h.$$

From this, (3.119), and continuity of Π_h, we get (3.117). □

But now we have, in fact, proved a weak form of the inf–sup condition for it comes from Proposition 3.11 that B_h coincides with the operator r. What we have obtained is, in fact,

$$\sup_{(\underline{\eta}_h, w_h)\in V_h} \frac{(\underline{\gamma}_h, \Pi_h\underline{\eta}_h - \underline{\mathrm{grad}}\ w_h)}{\|\Pi_h\underline{\eta}_h - \underline{\mathrm{grad}}\ w_h\|_{H_0(\mathrm{rot},\Omega)}} \ge k_0\|\underline{\gamma}_h\|_{\Gamma_h'}, \tag{3.121}$$

where the norm $\|\underline{\gamma}_h\|_{\Gamma_h'}$ is the *discrete dual norm*

$$\|\underline{\gamma}_h\|_{\Gamma_h'} = \sup_{\underline{\eta}_h\in\Gamma_h} \frac{(\underline{\gamma}_h, \underline{\eta}_h)}{\|\underline{\eta}_h\|_{H_0(\mathrm{rot},\Omega)}}. \tag{3.122}$$

To apply Theorem 3.3, we would need (3.121) to hold with

$$\|\underline{\gamma}_h\|_{\Gamma'} = \sup_{\eta\in\Gamma} \frac{(\underline{\gamma}_h, \underline{\eta})}{|\underline{\eta}|_{H_0(\underline{\mathrm{rot}},\Omega)}} \ge \|\underline{\gamma}_h\|_{\Gamma_h'}.$$

We are thus slightly short of an optimal result. The intuitive reason is that $\underline{\gamma}_h$ cannot be written $\underline{\gamma}_h = \underline{\mathrm{grad}}\ \psi_h + \underline{\mathrm{rot}}\ p_h$ but only as $\underline{\mathrm{grad}}\ \psi_h + \underline{\mathrm{rot}}_h\ p_h$ where $\underline{\mathrm{rot}}_h$ is a discrete rotational operator as in (3.115).

On the other hand, $\underline{\gamma}_h$ (as any other element of $H_0(\mathrm{rot},\Omega)$) can be written as

$$\underline{\gamma}_h = \underline{\mathrm{grad}}\ \tilde{\psi} + \underline{\mathrm{rot}}\ \tilde{p} \tag{3.123}$$

and (see (3.49))

$$\|\underline{\gamma}_h\|_{\Gamma'} \simeq \|\tilde{\psi}\|_1 + \|\tilde{p}\|_{0/\mathbb{R}}, \tag{3.124}$$

whereas from (3.115) and (3.122) we easily get

$$\|\underline{\gamma}_h\|_{\Gamma_h'} \simeq \|\phi_h\|_1 + \|p_h\|_{0/\mathbb{R}}. \tag{3.125}$$

We shall now try to get an estimate on ϕ_h and p_h and for this, we are led to introduce a decomposition principle analogous to (3.82)–(3.84). This will, of course, be based on Proposition 3.10. We consider now the full case $t \neq 0$ and we start from (3.90) which we rewrite as

$$\inf_{(\underline{\beta}_h, w_h) \in V_h} \frac{1}{2} a(\underline{\beta}_h, \underline{\beta}_h) + \frac{t^{-2}}{2} |\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h|^2 - (g, w_h), \tag{3.126}$$

where $\underline{\beta}_h$ and w_h are chosen as in Example 3.3 and Π_h is again defined by (3.101) and (3.102). Note that, for the sake of simplicity we set $\lambda = 1$. The optimality conditions for the problem are

$$a(\underline{\beta}_h, \underline{\eta}_h) - t^{-2}(\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h, \underline{\eta}_h) = 0, \qquad \forall \underline{\eta}_h \in H_h, \tag{3.127}$$

$$t^{-2}(\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h, \underline{\mathrm{grad}}\, \phi_h) = (g, \phi_h), \qquad \forall \phi_h \in W_h. \tag{3.128}$$

Denoting $\underline{\gamma}_h = t^{-2}(\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h)$ and recalling that $\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h$ is surjective on Γ_h, this becomes equivalent to

$$a(\underline{\beta}_h, \underline{\eta}_h) - (\underline{\gamma}_h, \Pi_h \underline{\eta}_h) = 0, \qquad \forall \underline{\eta}_h \in H_h, \tag{3.129}$$

$$(\underline{\gamma}_h, \underline{\mathrm{grad}}\, \phi_h) = (g, \phi_h), \qquad \forall \phi_h \in W_h, \tag{3.130}$$

$$(\underline{\gamma}_h, \underline{\delta}_h) = t^{-2}(\underline{\mathrm{grad}}\, w_h - \Pi_h \underline{\beta}_h, \underline{\delta}_h), \qquad \forall \underline{\delta}_h \in \Gamma_h. \tag{3.131}$$

Now from Proposition 3.10 we can decompose $\underline{\gamma}_h$ into two components,

$$(\underline{\gamma}_h, \underline{\delta}_h) = (\underline{\mathrm{grad}}\, \psi_h, \underline{\delta}_h) + (p_h, \mathrm{rot}\, \underline{\delta}_h), \quad \forall \underline{\delta}_h \in \Gamma_h, \tag{3.132}$$

with $\psi_h \in W_h$, $p_h \in Q_h$. From (3.130), we now have

$$(\underline{\mathrm{grad}}\, \psi_h, \underline{\mathrm{grad}}\, \phi_h) = (g, \phi_h), \qquad \forall \phi_h \in W_h, \tag{3.133}$$

so that ψ_h is immediately known. Now we insert (3.132) into (3.129), recalling that $(\mathrm{rot}\, \Pi_h \underline{\eta}_h, q_h) = (\mathrm{rot}\, \underline{\eta}_h, q_h)$, $\forall q_h \in Q_h$. This gives

$$a(\underline{\beta}_h, \underline{\eta}_h) - (p_h, \mathrm{rot} \underline{\eta}_h) = (\underline{\mathrm{grad}}\, \psi_h, \Pi_h \underline{\eta}_h). \tag{3.134}$$

We still have to manage with (3.131). First let us take $\underline{\delta}_h = \underline{\mathrm{grad}}\, \phi_h$. We then get

$$\begin{aligned} (\underline{\mathrm{grad}}\, w_h, \underline{\mathrm{grad}}\, \phi_h) &= (\Pi_h \underline{\beta}_h, \underline{\mathrm{grad}}\, \phi_h) + t^2(\underline{\mathrm{grad}}\, \psi_h, \underline{\mathrm{grad}}\, \phi_h) \\ &= (\Pi_h \underline{\beta}_h, \underline{\mathrm{grad}}\, \phi_h) + t^2(g, \phi_h). \end{aligned} \tag{3.135}$$

This makes it possible to compute w_h whenever $\underline{\beta}_h$ is known. Finally we take $\underline{\delta}_h$ in the orthogonal of $\underline{\text{grad}}\, W_h$. We have in (3.131) from (3.132)

$$(3.136) \qquad (p_h, \text{rot}\, \underline{\delta}_h) = t^{-2}(\Pi_h \underline{\beta}_h, \underline{\delta}_h), \qquad \forall \underline{\delta}_h \in (\underline{\text{grad}}\, W_h)^{\perp}.$$

Let now $\underline{\alpha}_h$ be the element of Γ_h defined by

$$(3.137) \qquad (\underline{\alpha}_h, \underline{\delta}_h) = (p_h, \text{rot}\, \underline{\delta}_h), \qquad \forall \underline{\delta}_h \in \Gamma_h,$$

then (3.136) can be written as

$$(3.138) \qquad (\underline{\alpha}_h, \underline{\delta}_h) = -t^{-2}(\Pi_h \underline{\beta}_h, \underline{\delta}_h), \qquad \forall \underline{\delta}_h \in (\underline{\text{grad}}\, W_h)^{\perp}.$$

But for $\underline{\delta}_h \in (\underline{\text{grad}}\, W_h)^{\perp}$, there exists $q_h \in Q_h$ such that

$$(3.139) \qquad (\underline{\delta}_h, \underline{\chi}_h) = (q_h, \text{rot}\, \underline{\chi}_h), \qquad \forall \underline{\chi}_h \in (\underline{\text{grad}}\, W_h)^{\perp},$$

and (3.138) then becomes (using again $(\text{rot}\, \Pi_h \underline{\eta}_h, q_h) = (\text{rot}\, \underline{\eta}_h, q_h)$, $\forall q_h$)

$$(3.140) \qquad (\text{rot}\, \underline{\alpha}_h, q_h) = t^{-2}(\text{rot}\, \underline{\beta}_h, q_h), \quad \forall q_h \in Q_h.$$

We can summarize this in

Proposition 3.11: Using the hypotheses of Theorem 3.4, the solution $(\underline{\beta}_h, w_h, \underline{\gamma}_h)$ of problem (3.129)–(3.131) can be found by solving the following problem:find $(\underline{\beta}_h, w_h, \psi_h, p_h, \underline{\alpha}_h)$ in $H_h \times W_h \times W_h \times Q_h \times \Gamma_h$ such that,

$$(3.141) \quad (\underline{\text{grad}}\, \psi_h, \underline{\text{grad}}\, \phi_h) = (g, \phi_h), \qquad \forall \phi_h \in W_h,$$

$$(3.142) \quad a(\underline{\beta}_h, \underline{\eta}_h) - (p_h, \text{rot}\, \underline{\eta}_h) = (\underline{\text{grad}}\, \psi_h, \Pi_h \underline{\eta}_h), \qquad \forall \underline{\eta}_h \in H_h,$$

$$(3.143) \quad (\text{rot}\, \underline{\beta}_h, q_h) + t^2(\text{rot}\, \underline{\alpha}_h, q_h) = 0, \qquad \forall q_h \in Q_h,$$

$$(3.144) \quad (\underline{\alpha}_h, \underline{\delta}_h) - (p_h, \text{rot}\, \underline{\delta}_h) = 0, \qquad \forall \underline{\delta}_h \in \Gamma_h,$$

$$(3.145) \quad (\underline{\text{grad}}\, w_h, \underline{\text{grad}}\, \phi_h) = (\Pi_h \underline{\beta}_h, \underline{\text{grad}}\, \phi_h) + t^2(g, \phi_h), \quad \forall \phi_h \in W_h.$$

and setting

$$(3.146) \qquad \underline{\gamma}_h = \underline{\text{grad}}\, \psi_h + \underline{\alpha}_h. \ \square$$

Now (3.141) is a standard Dirichlet problem and for $t = 0$, (3.142), (3.143) is a "Stokes problem" discretized by the Crouzeix–Raviart element (Chapter VI). Again one must emphasize the relations between Stokes problem and Mindlin's problem. For $t \neq 0$ we have on the right-hand side of (3.143) and in (3.144) a mixed formulation of the Neumann problem using the Raviart–Thomas element. This could be brought to a more computable form using the " λ-trick" of Chapter V. However, in opposition with the situation of Example 3.2, where the decomposition principle (3.82)–(3.84) was the only way to solve the problem, it is now more suitable from a computational point of view to use (3.127) and (3.128). Formulation (3.141), (3.142) however makes it possible to easily refine the error estimate of (3.114). Following BREZZI–FORTIN–STENBERG [A] (see also PEISKER–BRAESS [A]) we now prove the following theorem.

Theorem 3.5: Let $(\underline{\beta}(t), w(t)$ and $\underline{\gamma}(t)$ be the solution of (3.24), (3.25), let $\psi(t)$ and $p(t)$ be defined by Proposition 3.4 (see (3.48)), and set

$$\underline{\alpha}(t) = \underline{\text{rot}}\ p(t). \tag{3.147}$$

Let $\underline{\beta}_h(t), w_h(t), \psi_h(t), p_h(t), \underline{\alpha}_h(t)$ be the solution of (3.141)–(3.145). Then there exist two constants $c, c_1 > 0$ independent of h and t such that

$$\begin{cases} \|\underline{\beta}_h(t)-\underline{\beta}(t)\|_1 + \|w_h(t)-w(t)\|_1 + \|\psi_h(t)-\psi(t)\|_1 \\ \qquad + \|p_h(t)-p(t)\|_{0/\mathbb{R}} + t\,\|\underline{\alpha}_h(t)-\underline{\alpha}(t)\|_0 \\ \le c\Big\{ \inf\limits_{\underline{\eta}_h \in H_h} \|\underline{\beta}(t)-\underline{\eta}_h\|_1 + \inf\limits_{\zeta_h \in W_h} \|w(t)-\zeta_h\|_1 + \inf\limits_{\phi_h \in W_h} \|\psi(t)-\phi_h\|_1 \\ \qquad + \inf\limits_{q_h \in Q_h} \|p(t)-q_h\|_{0/\mathbb{R}} + \inf\limits_{\underline{\delta}_h \in \Gamma_h} t\,\|\underline{\alpha}(t)-\underline{\delta}_h\|_0 + \|\underline{\beta}-\Pi_h\underline{\beta}\|_0 \\ \qquad + \sup\limits_{\underline{\eta}_h \in H_h} \dfrac{(\underline{\text{grad}}\ \psi(t), \underline{\eta}_h - \Pi_h\underline{\eta}_h)}{\|\underline{\eta}_h\|_1} \Big\} \\ \le C_1 h^s \Big\{ \|\underline{\beta}(t)\|_{s+1} + \|w(t)\|_{s+1} + \|\psi(t)\|_{s+1} + \|p(t)\|_{s/\mathbb{R}} \\ \qquad + \|\underline{\alpha}(t)\|_s \Big\} \end{cases} \tag{3.148}$$

for every $s \in [0, 2]$.

Proof: We remark first that $\underline{\beta}(t), w(t), \psi(t), p(t), \underline{\alpha}(t)$ is a solution of

$$(\underline{\text{grad}}\ \psi, \underline{\text{grad}}\ \phi) = (g, \phi), \qquad \forall \phi \in H_0^1(\Omega), \tag{3.149}$$

$$a(\underline{\beta}, \underline{\eta}) - (p, \text{rot}\ \underline{\eta}) = (\underline{\text{grad}}\ \psi, \underline{\eta}), \qquad \forall \underline{\eta} \in (H_0^1(\Omega))^2, \tag{3.150}$$

$$-(\text{rot}\ \underline{\beta}, q) - t^2(\text{rot}\ \underline{\alpha}, q) = 0, \qquad \forall q \in L^2(\Omega), \tag{3.151}$$

$$(\underline{\alpha}, \underline{\delta}) - (p, \text{rot}\ \underline{\delta}) = 0, \qquad \forall \underline{\delta} \in H_0(\text{rot}; \Omega), \tag{3.152}$$

$$(\underline{\text{grad}}\ w, \underline{\text{grad}}\ \phi) = (\underline{\beta}, \underline{\text{grad}}\ \phi) + t^2(g, \phi), \quad \forall \phi \in H_0^1(\Omega). \tag{3.153}$$

We now look for functions $(\underline{\beta}_h^I, w_h^I, \psi_h^I, p_h^I, \underline{\alpha}_h^I)$ in $H_h \times W_h \times W_h \times Q_h \times \Gamma_h$ close to $(\underline{\beta}, w, \psi, p, \underline{\alpha})$. For this we set first

$$(\underline{\text{grad}}\ \psi_h^I, \underline{\text{grad}}\ \phi_h) = (g, \phi_h), \qquad \forall \phi_h \in W_h, \tag{3.154}$$

that is, actually, $\psi_h^I = \psi_h$. Then we consider the solution $(\underline{\beta}_h^I, \tilde{p}_h^I)$ of the discrete Stokes-like problem

$$(3.155)\qquad a(\underline{\beta}_h^I, \underline{\eta}_h) + (\tilde{p}_h^I,\ \mathrm{rot}\ \underline{\eta}_h) = a(\underline{\beta}, \underline{\eta}_h), \qquad \forall \underline{\eta}_h \in H_h,$$

$$(3.156)\qquad (\mathrm{rot}\ \underline{\beta}_h^I, q_h) = (\mathrm{rot}\ \underline{\beta}, q_h), \qquad \forall q_h \in Q_h$$

Since the pair (H_h, Q_h) is a stable pair for Stokes problem (see Chapter VI) we shall have that $\underline{\beta}_h^I$ is an optimal approximation of $\underline{\beta}$. Now we consider the mixed problem: find $(\underline{\alpha}_h^I, p_h^I) \in \Gamma_h \times Q_h$ such that

$$(3.157)\qquad (\mathrm{rot}\ \underline{\beta}_h^I, q_h) + t^2(\mathrm{rot}\ \underline{\alpha}_h^I, q_h) = 0, \qquad \forall q_h \in Q_h,$$

$$(3.158)\qquad (\underline{\alpha}_h^I, \underline{\delta}_h) - (p_h^I, \mathrm{rot}\ \underline{\delta}_h) = 0, \qquad \forall \underline{\delta}_h \in \Gamma_h.$$

Note that, from (3.156), $(\underline{\alpha}_h^I, p_h^I)$ will be the mixed finite element solution of $\underline{\alpha} = \underline{\mathrm{rot}}\ p$ and $\mathrm{rot}\ \underline{\alpha} = t^{-2}\mathrm{rot}\ \underline{\beta}$ and hence an optimal approximation of $(\underline{\alpha}, p)$. Finally we can solve

$$(3.159)\qquad (\underline{\mathrm{grad}}\ w_h^I, \underline{\mathrm{grad}}\ \phi_h) = (\Pi_h \underline{\beta}_h, \underline{\mathrm{grad}}\ \phi_h) + t^2(g, \phi_h), \quad \forall \phi_h \in W_h$$

Let us look now at the difference $(\underline{\beta}_h - \underline{\beta}_h^I, \ldots, \underline{\alpha}_h - \underline{\alpha}_h^I)$. It is a solution of

$$(3.160)\qquad (\underline{\mathrm{grad}}(\psi_h - \psi_h^I),\ \underline{\mathrm{grad}}\ \phi_h) = 0, \qquad \forall \phi_h \in W_h$$

$$(3.161)\qquad \begin{aligned} a(\underline{\beta}_h - \underline{\beta}_h^I, \underline{\eta}_h) - (p_h - p_h^I, \mathrm{rot}\,\underline{\eta}_h) &= a(\underline{\beta} - \underline{\beta}_h^I, \underline{\eta}_h) - (p - p_h^I, \mathrm{rot}\,\underline{\eta}_h) \\ &\quad - (\underline{\mathrm{grad}}\ \psi, \underline{\eta}_h - \Pi_h \underline{\eta}_h) + (\underline{\mathrm{grad}}\,(\psi_h - \psi), \Pi_h \underline{\eta}_h), \quad \forall \underline{\eta}_h \in H_h, \end{aligned}$$

$$(3.162)\qquad (\mathrm{rot}(\underline{\beta}_h - \underline{\beta}_h^I), q_h) + t^2(\mathrm{rot}(\underline{\alpha}_h - \underline{\alpha}_h^I), q_h) = 0, \qquad \forall q_h \in Q_h,$$

$$(3.163)\qquad (\underline{\alpha}_h - \underline{\alpha}_h^I, \underline{\delta}_h) - (p_h - p_h^I,\ \mathrm{rot}\ \underline{\delta}_h) = 0, \qquad \forall \underline{\delta}_h \in \Gamma_h,$$

$$(3.164)\qquad (\underline{\mathrm{grad}}(w_h - w_h^I), \underline{\mathrm{grad}}\ \phi_h) = (\Pi_h \underline{\beta}_h - \Pi_h \underline{\beta}, \underline{\mathrm{grad}}\ \phi_h), \quad \forall \phi_h \in W_h.$$

From (3.160) we have again $\psi_h = \psi_h^I$. This easily implies

$$(3.165)\qquad \|\psi_h - \psi\|_1 \le \inf_{\phi_h \in W_h} \|\psi - \phi_h\|_1.$$

Now we take $\underline{\eta}_h = \underline{\beta}_h - \underline{\beta}_h^I$ in (3.161), $q_h = p_h - p_h^I$ in (3.162), $\underline{\delta}_h = t^2(\underline{\alpha}_h - \underline{\alpha}_h^I)$ in (3.163), and we sum the three equations. We obtain

$$(3.166)\qquad \begin{aligned} \|\underline{\beta}_h - \underline{\beta}_h^I\|_1^2 + t^2\|\underline{\alpha}_h - \underline{\alpha}_h^I\|_0^2 &\le c_2\Big\{\|\underline{\beta} - \underline{\beta}_h^I\|_1 + \|p - p_h^I\|_{0/\mathbb{R}} \\ &\quad + \sup_{\underline{\eta}_h} \frac{(\underline{\mathrm{grad}}\ \psi, \underline{\eta}_h - \Pi_h \underline{\eta}_h)}{\|\underline{\eta}_h\|_1} + \|\psi - \psi_h\|_1\Big\}\|\underline{\beta}_h - \underline{\beta}_h^I\|_1 \end{aligned}$$

Using now the triangle inequality, the optimality of the solutions of (3.155), (3.156) and (3.157), (3.158) and (3.165) we have

$$
\begin{aligned}
\|\underline{\beta} - \underline{\beta}_h\|_1 \le c_3 \Big\{ & \inf_{\underline{\eta}_h \in H_h} \|\underline{\beta} - \underline{\eta}_h\|_1 + \inf_{q_h \in Q_h} \|p - q_h\|_{0/\mathbb{R}} \\
& + \inf_{\phi_h \in W_h} \|\psi - \phi_h\|_1 + \sup_{\underline{\eta}_h \in H_h} \frac{(\underline{\mathrm{grad}}\, \psi, \underline{\eta}_h - \Pi_h \underline{\eta}_h)}{\|\underline{\eta}_h\|_1} \Big\},
\end{aligned} \tag{3.167}
$$

$$
\begin{aligned}
t\,\|\underline{\alpha}_h - \underline{\alpha}\|_0 \le C_4 \Big\{ & \inf_{\underline{\eta}_h \in H_h} \|\underline{\beta} - \underline{\eta}_h\|_1 + \inf_{q_h \in Q_h} \|p - q_h\|_{0/\mathbb{R}} \\
& + \inf_{\phi_h \in W_h} \|\psi - \phi_h\|_1 + \inf_{\underline{\delta}_h \in \Gamma_h} t\,\|\underline{\alpha} - \underline{\delta}_h\|_0 \\
& + \sup_{\underline{\eta}_h \in H_h} \frac{(\underline{\mathrm{grad}}\, \psi, \underline{\eta}_h - \Pi_h \underline{\eta}_h)}{\|\underline{\eta}_h\|_1} \Big\}.
\end{aligned} \tag{3.168}
$$

Using now the inf–sup condition for Stokes problems in (3.161) (and again the triangle inequality) we have from (3.161) and the previous estimates

$$
\begin{aligned}
\|p - p_h\|_{0/\mathbb{R}} \le C_5 \Big\{ & \inf_{\underline{\eta}_h \in H_h} \|\underline{\beta} - \underline{\eta}_h\|_1 + \inf_{q_h \in Q_h} \|p - q_h\|_{0/\mathbb{R}} \\
& + \inf_{\phi_h \in W_h} \|\psi - \phi_h\|_1 + \sup_{\underline{\eta}_h \in H_h} \frac{(\underline{\mathrm{grad}}\, \psi, \underline{\eta}_h - \Pi_h \underline{\eta}_h)}{\|\underline{\eta}\|_1} \Big\}.
\end{aligned} \tag{3.169}
$$

Finally from the approximation properties of Dirichlet problem for the Laplace operator we have, from (3.164),

$$
\|w - w_h\|_1 \le C_6 \Big\{ \inf_{\phi_h \in W_h} \|w - \phi_h\|_1 + \|\underline{\beta} - \Pi_h \underline{\beta}\|_0 + \|\underline{\beta} - \underline{\beta}_h\|_1 \Big\}. \tag{3.170}
$$

Collecting (3.166)–(3.170) we have that (3.148) follows from known approximation properties, bounding the consistency term (with the sup) as in (3.110)–(3.112). □

Remark 3.10: If the triangulation is quasi-uniform (see, e.g., CIARLET [A]) we can use an inverse inequality in (3.163) and obtain

$$
\|\underline{\alpha}_h - \underline{\alpha}_h^I\|_0 \le c h^{-1} \|p_h - p_h^I\|_0, \tag{3.171}
$$

which, in turn, produces a bound for $h\|\underline{\alpha} - \underline{\alpha}_h\|_0$ similar to (3.148). □

Remark 3.11: From Theorem 3.5 we have optimal error estimates for the variable $\underline{\beta}$ and w. As far as $\underline{\gamma}$ is concerned, we have, on the one hand, optimal error bounds in the norm (3.125). On the other hand, the estimates on ψ and $\underline{\alpha}$ give an estimate for $t\,\|\underline{\gamma} - \underline{\gamma}_h\|_0$ and for $(t+h)\,\|\underline{\gamma} - \underline{\gamma}_h\|_0$ for quasi uniform meshes. It is not clear whether the (natural) norm in Γ' (see (3.42) or (3.49)) is equivalent to (3.125) or not. □

Now to end this lengthy section, we are in a position to present general guidelines for the discretization of Mindlin–Reissner problems.

We must emphasize again that the decomposition principle makes apparent a direct link with the Stokes problem. Indeed, all examples for which a satisfactory analysis could be achieved contained an already proven Stokes element. If we distinguish the case of continuous pressure approximation and the case of discontinuous pressure element, we get two types of strategies.

VII.3.3 Continuous pressure approximations

— Suppose one knows $H_h \times Q_h$ to be a good approximation of the Stokes problem with $Q_h \subset H^1(\Omega)$.

— Choose W_h an approximation of $H_0^1(\Omega)$ of the same order of accuracy.

— Write $\Gamma_h = \underline{\text{grad}}\, W_h + \underline{\text{rot}}\, Q_h$.

In this context, the definition of Γ_h does not lead, in general, to a standard space and the decomposition principle (3.82)–(3.84) is the only way to handle things from a computational point of view. It may, however happen, for a clever choice of W_h and Q_h, that Γ_h turns out to be a standard polynomial space. Such a situation has been encountered in ARNOLD–FALK [B] where, using for $H_h \times Q_h$ the MINI element as in Example 3.2, but taking W_h to be $\mathfrak{L}_1^{1,NC}$, that is, a nonconforming P_1 approximation of $H_0^1(\Omega)$, Γ_h comes to be the whole space $(\mathfrak{L}_0^0)^2$ and not a proper subspace as in Example 3.2. No such construction is known (at the time we write this) to extend this result to $(\mathfrak{L}_1^0)^2$ or to larger spaces.

VII.3.4 Discontinuous pressure elements

This second class of approximations to the Stokes problem has been the basis for the "reduced integration" method of Example 3.4. We shall try to outline here the principal features of this example in order to provide a guide for possible extensions, some of which can be found in BATHE–BREZZI–FORTIN [A].

(1) Here again our starting point is an approximation of the Stokes problem $H_h \times Q_h$, Q_h being a space of discontinous polynomial functions. This approximation should, of course, satisfy the inf–sup condition.

(2) We need to match this with an approximation Γ_h of $H_0(\text{rot};\Omega)$. More precisely we need a couple of spaces (Γ_h, Q_h) (where Q_h is the same as

before) and a uniformly bounded linear operator $\Pi_h \to \Gamma_h$ such that we have the commuting diagram:

$$\begin{array}{ccc} H & \xrightarrow{\text{rot}} & L^2(\Omega) \\ {\scriptstyle \Pi_h}\downarrow & & {\scriptstyle P_h}\downarrow \\ \Gamma_h & \xrightarrow{\text{rot}} & Q_h, \end{array}$$

where $H = (H^1(\Omega))^2 \cap H_0(\text{rot};\Omega)$ and P_h is the L^2-projection operator.

(3) We finally need a space $W_h \subset H_0^1(\Omega)$ such that

$$\underline{\text{grad}}\, W_h = \{\underline{\delta}_h \in \Gamma_h,\ \text{rot}\ \underline{\delta}_h = 0\}.$$

Ingredients (1), (2), (3) will produce a plate element for which one can essentially repeat the proof of Theorem 3.5 and obtain optimal error estimates for $\underline{\beta}$, and w, and for $\underline{\gamma}$ (in the norm (3.125)).

BABUŠKA I., OSBORN J., PITKARANTA J.
[A] Analysis of mixed methods using mesh-dependent norms, *Math. Comp.*, **35**, 1039–1079 (1980).

BARBU V., PRECUPANU T.
[A] *Convexity and optimization in Banach spaces*, Editura Academiei, Bucarest (1978).

BATHE K.J.
[A] *Finite Element Procedures in Engineering Analysis*, Prentice–Hall, Englewood Cliffs, N.J. (1982).

BATHE K.J., BREZZI F.
[A] The convergence of a four-node plate bending element based on Mindlin–Reissner plate theory and a mixed interpolation, *Proceedings of the Conference on Mathematics of Finite Elements and Applications V* (J.R. Whiteman, ed.), Academic Press, New York (1985), pp. 491–503.
[B] A simplified analysis of two plate bending elements. The MITC4 and MITC9 elements, *Numerical Techniques for Engineering Analysis and Design. Numeta 87, Vol 1* (G.N. Pande and J. Middleton, eds.), Martinus Nijhoff, Amsterdam (1987).

BATHE K.J., DVORKIN E.
[A] A continuum mechanics based four-node shell element for general nonlinear analysis, *J. Eng. Comp.*, **1**, 77–78 (1984).

BATHE K.J., SUSSMAN T.
[A] A Finite Element Formulation for Nonlinear Incompressible Elastic and Inelastic Analysis, *J. Comp. and Structure,* **26**, 357–409 (1987).

BATOZ J.L., BATHE K.J., HO L.W.
[A] A study of three-node triangular plate bending elements, *Int. J. Numer. Methods Eng.*, **15**, 1771–1812 (1980).

BERCOVIER M.
[A] Régularisation duale des problèmes variationnels mixtes. Thèse de doctorat d'Etat, Université de Rouen (1976).
[B] Perturbation of a mixed variational problem, applications to mixed finite element methods, *R.A.I.R.O. Anal. Numer.*, **12**, 211–236 (1978).

BERCOVIER M., ENGELMAN M., GRESHO P.
[A] Consistent and reduced integration penalty methods for incompressible media using several old and new elements. *Int. J. Numer. Methods Fluids*, **2**, 25–42 (1982).

BERCOVIER H., HASBANI Y., GITON Y., BATHE K.J.
[A] On a finite element procedure for non-linear incompressible elasticity in Hybrid and Mixed Finite Element Methods, S.N. Attluri, R.H. Rallagher O.C. Zienkiewicz eds, John Wiley and Sons, New York (1983).

BERCOVIER M., PIRONNEAU O.A.
[A] Error estimates for finite element method solution of the Stokes problem in the primitive variables, *Numer. Math.*, **33**, 211–224 (1977).

BERGH J., LOFSTROM J.
[A] *Interpolation Spaces: An Introduction*, Springer-Verlag, Berlin (1976).

BERNARDI C., CANUTO C., MADAY Y.
[A] Generalized Inf-Sup condition for Chebyshev approximation of the Navier–Stokes equations, *SIAM J. Numer. Anal.*, **25**, 1237–1265 (1988).

BERNARDI C., RAUGEL G.
[A] Méthodes d'éléments finis mixtes pour les équations de Stokes et de Navier-Stokes dans un polygone non convexe, *Calcolo*, **18**, 255–291 (1981).

BERTRAND F., DHATT G., FORTIN M., OUELLET Y., SOULAIMANI A.
[A] Simple continuous pressure elements for two and three-dimensional incompressible flows, *Comp. Methods Appl. Mech. Eng.*, **62**, 47–69 (1987).

BOLAND J.M., NICOLAIDES
[A] Stability of finite elements under divergence constraints, *SIAM J. Numer. Anal.*, **20**, 722–731 (1983).
[B] On the stability of bilinear-constant velocity-pressure finite elements, *Numer. Math.*, **44**, 219–222 (1984).
[C] Stable and semistable low order finite elements for viscous flows, *SIAM J. Numer. Anal.*, **22**, 474–492 (1985).

BRAESS D., PEISKER P.
[A] Uniform Convergence of Mixed Interpolated elements for Reissner-Mindlin Plates, *Preprint* **142** *(1990) Fakultät und Institut für Mathmatik der Ruhr-Universität Bachum Universitatsstaße* 150, D-4630 Bochum.

BRAMBLE J.H.
[A] The Lagrange multiplier method for Dirichlet's problem, *Math. Comp.*, **37**, 1–11 (1981).

BRAMBLE J.H., FALK R.S.
[A] Two mixed finite element methods for the simply supported plate problem, *R.A.I.R.O. Anal. Numer.*, **17**, 337–384 (1983).

BRAMBLE J.H., SCHATZ A.H.
[A] Estimates for spline projections, *R.A.I.R.O. Anal. Numer.*, **10**, 5–37 (1976).
[B] Higher order local accuracy by averaging in the finite element method, *Math. Comp.*, **31**, 94–111 (1977).

BRAMBLE J.H., XU J.M.
[A] local post-processing technique for improving the accuracy in mixed finite element approximations, *SIAM J. Numer. Anal.* **26**, 1267–1275 (1989).

BREZZI F.

[A] On the existence, uniqueness and approximation of saddle point problems arising from lagrangian multipliers, *R.A.I.R.O. Anal. Numer.*, **8**, 129–151 (1974).

[B] Sur une méthode hybride pour l'approximation du problème de la torsion d'une barre élastique, *Ist. Lombardo (Rend. Sc.)*, **A108**, 274–300 (1974).

[C] Sur la méthode des éléments finis hybrides pourle problème biharmonique, *Numer. Math.*, **24**, 103–131 (1975).

[D] Hybrid approximations of non-linear plate bending problems, *Hybrid and Mixed Finite Element Methods* (S.N.Atluri, R.H.Gallagher, O.C. Zienkiewicz, eds.), John Wiley and Sons (1983).

BREZZI F., BATHE K.J., FORTIN M.

[A] Mixed-Interpolated elements for Reissner–Mindlin plates, *Int. J. Numer. Methods Eng.*, **28**, 1787–1801 (1989).

BREZZI F., DOUGLAS J.

[A] Stabilized mixed methods for the Stokes problem, *Numer. Math.*, **53**, 225–235 (1988).

BREZZI F., DOUGLAS J., DURAN R., FORTIN M.

[A] Mixed finite elements for second order elliptic problems in three variables, *Numer. Math.*, **51**, 237–250 (1987).

BREZZI F., DOUGLAS J., FORTIN M., MARINI L.D.

[A] Efficient rectangular mixed finite elements in two and three space variables, *Math. Model. Numer. Anal.*, **21**, 581–604 (1987).

BREZZI F., DOUGLAS J., MARINI L.D.

[A] Two families of mixed finite elements for second order elliptic problems, *Numer. Math.*, **47**, 217–235 (1985).

[B] Recent results on mixed finite element methods for second order elliptic problems, in *Vistas in Applied Math., Numerical Analysis, Atmospheric Sciences, Immunology* (Balakrishanan, Dorodnitsyn, and Lions, eds.), Optimization Software Publications, New York (1986).

BREZZI F., FALK R.S.

[A] Stability of a higher-order Hood–Taylor method *SIAM J. Num. AnaL.* **28** (1991).

BREZZI F., FORTIN M.

[A] Numerical approximation of Mindlin–Reissner plates, *Math. Comp.*, **47**, 151–158 (1986).

BREZZI F., FORTIN M., STENBERG R.

[A] Quasi-optimal error bounds for approximation of shear-stresses in Mindlin–Reissner plate models, (to appear in *Math. Models and Methods in Appl. Sci.*, 1) (1991).

BREZZI F., LE TELLIER J., OLIER T.
[A] Mixed finite element approximation for the stationary Navier–Stokes equations (in Russian), Meeting INRIA Novosibirsk, Paris 1988, *Viceslitelnia Metodii V. Prikladnoi Mathematiceskie*, NAUKA, Novosibirsk (1982), pp. 96–108.

BREZZI F., MARINI L.D.
[A] On the numerical solution of plate bending problems by hybrid methods, *R.A.I.R.O. Anal. Numer.*, 5–50 (1975).

BREZZI F., MARINI L.D., PIETRA P.
[A] Méthodes d'éléments finis mixtes et schéma de Scharfetter–Gummel, *C.R.A.S. Paris*, **305**, I, 599–604 (1987).
[B] Two dimensional exponential fitting and application to drift-diffusion models, *SIAM. J. Numer. Anal.*, **26**, 1347–1355 (1989)
[C] Numerical simulation of semi conductor devices, *Comp. Math. Appl. Mech. Eng.*, **75**, 493–514 (1989).

BREZZI F., MARINI L.D., QUARTERONI A., RAVIART P.A.
[A] On an equilibrium finite element method for plate bending problems, *Calcolo*, **17**, 271–291 (1980).

BREZZI F., PITKÄRANTA J.
[A] On the stabilization of finite element approximations of the Stokes equations, in *Efficient Solutions of Elliptic Systems*, Notes on Numerical Fluid Mechanics, Vol 10 (W. Hackbush, ed.), Braunschweig, Wiesbaden (1984).

BREZZI F., RAVIART P.A.
[A] Mixed finite element methods for 4th order elliptic equations, in *Topics in Numerical Analysis III* (J. Miller, ed.), Academic Press, New York (1978).

BROOKS A., HUGHES T.J.R.
[A] Streamline upwind/Petrov–Galerkin formulation for convection dominated flows with particular emphasis on the incompressible Navier–Stokes equations, *Comp. Methods Appl. Mech. Eng.*, **32**, 199–259 (1982).

CANUTO C.
[A] Eigenvalue approximations by mixed methods, *R.A.I.R.O. Anal, . Numer.*, **12**, 27–50 (1978).
[B] A hybrid finite element method to compute the free vibration frequencies of a clamped plate, *R.A.I.R.O. Anal. Numer.*, **15**, 101–118 (1981).

CAUSSIGNAC P.
[A] Explicit basis functions of quadratic and improved quadratic finite element spaces for the Stokes problem, *Comm. Appl. Numer. Methods*, **2**, 205–211 (1986).

[B] Computation of pressure from the finite element vorticity stream-function approximation of the Stokes problem, *Comm. Appl. Numer. Methods*, **3**, 287–295 (1987).

CEA J.

[A] Approximation variationnelle des problèmes aux limites, *Ann. Inst. Fourier*, **14**, 2 (1964).

[B] Approximation variationnelle et convergence des éléments finis; un test, *Journées Eléments Finis*, Université de Rennes, Rennes (1976).

CHAVENT G., JAFFRE J.

[A] Mathematical models and finite elements for reservoir simulation , *Studies in Mathematics and its Applications 17*, North-Holland, Amsterdam (1986).

CIARLET P.G.

[A] *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam (1978).

[B] *Mathematical Elasticity, Volume 1. Three-Dimensional Elasticity*, North-Holland, Amsterdam (1988); *Volume 2. Lower-Dimensional Theories of Plates and Rods*, North-Holland, Amsterdam, (to appear).

CIARLET P.G., DESTUYNDER P.

[A] A justification of the two-dimensional linear plate model, *J. Mécanique*, **18**, 315–344 (1979).

CIARLET P.G., GLOWINSKI R.

[A] Dual iterative techniques for solving a finite element approximation of the biharmonic equation, *Comp. Methods Appl. Mech. Eng.*, **5**, 227–295 (1975).

CIARLET P.G., RAVIART P.A.

[A] Interpolation theory over curved elements with applications to finite element methods, *Comp. Methods Appl. Mech. Eng.*, **1**, 217–249 (1972).

[B] General Lagrange and Hermite interpolation in $\mathbb{R}^n$ with applications to finite element methods., *Arch. Rat. Mech. Anal.*, **46**, 177–199 (1972).

[C] A mixed finite element method for the biharmonic equation, in *Mathematical Aspects of Finite Element in Partial Differential Equations* (C. de Boor, ed.), Academic Press, New York (1974).

CIAVALDINI J.F., NEDELEC J.C.

[A] Sur l'élément de Fraeijs de Veubeke et Sander, *R.A.I.R.O. Anal. Numer.*, **8**, 29–45 (1974).

CLEMENT P.

[A] Approximation by finite element functions using local regularization, *R.A.I.R.O. Anal. Numer.*, **9**, 77–84 (1975).

COMODI L.
[A] The Hellan–Hermann–Johnson method: error estimates for the Lagrange multipliers and post processing, *Math. Comp.*, **52**, 17–30 (1989).

CROUZEIX M., RAVIART P.A.
[A] Conforming and non-conforming finite element methods for solving the stationary Stokes equations, *R.A.I.R.O. Anal. Numer.*, **7**, 33–76 (1973).

DAUTRAY R., LIONS J.L.
[A] *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Collection Commissariat à l'Energie Atomique, Masson, Paris (1984).

DESTUYNDER P.
[A] *Une théorie asymptotique des plaques minces en élasticité linéaire*, Masson, Paris (1986).

DOUGLAS J., DURAN R., PIETRA P.
[A] Alternating-direction iteration for mixed finite element methods, *Computing Methods in Applied Sciences and Engineering* (R. Glowinski and J.L. Lions, eds.), North-Holland, Amsterdam (1986), p.7.
[B] Formulation of alternating-direction iterative methods for mixed methods in three space, *Numerical Approximation of Partial Differential Equations* (E.L.Ortiz, ed.), North-Holland, Amsterdam (1987).

DOUGLAS J., MILNER F.
[A] Interior and superconvergence estimates for mixed methods for second order elliptic problems, *Math. Modelling Numer. Anal.*, **19**, 397–428 (1985).

DOUGLAS J., PIETRA P.
[A] A description of some alternating-direction iterative techniques for mixed finite element methods, *Mathematical and Computational Methods in Seismic Exploration and Reservoir Modeling* (W.E. Fitzgibbon, ed.), SIAM, Philadelphia (1986).

DOUGLAS J., ROBERTS J.E.
[A] Mixed finite element methods for second order elliptic problems, *Math. Appl. Comp.*, **1**, 91–103 (1982).
[B] Global estimates for mixed methods for second order elliptic equations, *Math. Comp.*, **44**, 39–52 (1985).

DOUGLAS J., WANG, J.
[A] An absolutely stabilized finite element method for the stokes problem, *Math. Comput.* **52**, 495–508 (1989).

DUPONT T., SCOTT L.R.
[A] Polynomial approximation of functions in Sobolev spaces, *Math. Comp.*, **34**, 441–463 (1980).

DUVAUT G., LIONS J.L.
[A] *Les inéquations en mécanique et en physique*, Dunod, Paris (1972).

EKELAND I., TEMAM R.
[A] *Analyse convexe et problèmes variationnels*, Dunod Gauthier Villars, Paris (1974).

FALK R.S.
[A] A Ritz method based on a complementary variational principle, *R.A.I.R.O. Anal. Numer.*, **10**, 39–48 (1976).
[B] Approximation of the biharmonic equation by a mixed finite element method, *SIAM J. Numer. Anal.*, **15**, 556–567 (1978).

FALK J., OSBORN J.
[A] Error estimates for mixed methods, *R.A.I.R.O. Anal. Numer.*, **4**, 249–277 (1980).

FIX G.J., GUNZBURGER M.D., NICOLAIDES R.A.
[A] Theory and applications of mixed finite element methods, in *Constructive Approaches to Mathematical Models* (C.V. Coffman and G.J. Fix, eds.), Academic Press, New York (1979), pp. 375–393.

FORTIN A.
[A] *Méthodes d'éléments finis pour les équations de Navier–Stokes*, **Thèse**, Université Laval (1984).

FORTIN A., FORTIN M.
[A] Newer and newer elements for incompressible flow, *Finite Elements in Fluids 6* (R.H. Gallagher, G.F. Carey, J.T. Oden and O.C. Zienkiewicz, eds.), Wiley, New York (1985).

FORTIN M.
[A] *Calcul numérique des écoulements des fluides de Bingham et des fluides newtoniens incompressibles par la méthode des éléments finis*, Thèse, Université Paris VI (1972).
[B] Utilisation de la méthode des éléments finis en mécanique des fluides, *Calcolo*, **12**, 405–441 (1975).
[C] An analysis of the convergence of mixed finite element methods, *R.A.I.R.O. Anal. Numer.*, **11**, 341–354 (1977).
[D] Old and new finite elements for incompressible flows, *Int. J. Numer. Methods Fluids*, **1**, 347–364 (1981).

FORTIN M., GLOWINSKI R.
[A] *Méthodes de Lagrangien augmenté*, Dunod, Paris (1982).
[B] *Augmented Lagrangian Methods*, North-Holland, Amsterdam, (1983).

FORTIN M., PEYRET R., TEMAM R.
[A] Résolution numérique des équations de Navier-Stokes pour un fluide visqueux incompressible, *Mécanique*, **10**, 3, 357–390 (1971).

FORTIN M., PIERRE R.
[A] Stability analysis of discrete generalized Stokes problems (to appear in Mathematical Methods for Partial Differential Equations).

FORTIN M., SOULIE M.
[A] A non-conforming piecewise quadratic finite element on triangles, *Int. J. Numer. Methods Eng.*, **19**, 505–520 (1983).

FORTIN M., THOMASSET F.
[A] Mixed finite element methods for incompressible flow problems, *J. Comp. Phy.*, **37**, 173–215 (1979).

FRAEIJS de VEUBEKE B.
[A] Displacement and equilibrium models in the finite element method, in *Stress Analysis* (O.C. Zienkiewicz and G. Holister, eds.), John Wiley and Sons, New York (1965).
[B] Variational principles and the patch test, *Int. J. Numer. Methods Eng.*, **8**, 783–801 (1974).
[C] Stress function approach, in *World Congress on the Finite Element Method in Structural Mechanics*, Bournemouth, Dorset, England (1975).

FRANCA L.P.
[A] New Mixed Finite Flement Methods, *Ph. D. Thesis, Appl. Mech. Divi.*, Stanford University (1987).

FRAEIJS de VEUBEKE B., SANDER G.
[A] An equilibrium model for plate bending, *Int. J. Solids Structures*, **4**, 447–468 (1968).

FRANCA L.P., HUGHES T.J.R.
[A] Two classes of finite element methods, *Comp. Meth. Appl. Mech. Eng.*, **69**, 89–129 (1988).

FRANCA L.P., STENBERG R.
[A] Error Analysis of Some Galerkin Least-Squares Methods of the Elasticity Equations, Rapport INRIA 1054, INRIA , Domaine de Voluceau, Rocquencourt, B.P. 105, 78153, Le Chesnay, CEDEX, France (1989).

GASTALDI L., NOCHETTO R.
[A] Optimal L^∞-error estimates for non-conforming and mixed finite element methods of lowest order, *Numer. Math.*, **50**, 587–611 (1987).
[B] Sharp maximum norm error estimates for general mixed finite element approximations to second-order elliptic equations.,*Model. Math. An.Numer.*, **23**,103–128 (1989).

GIRAULT V., RAVIART P.A.
[A] *Finite Element Approximation of Navier–Stokes Equations*, Lecture Notes in Math. 749, Springer–Verlag, Berlin (1979).
[B] *Finite Element Methods for Navier-Stokes Equations, Theory and Algorithms*, Springer-Verlag, Berlin (1986).

GLOWINSKI R.

[A] *Numerical Methods for Nonlinear Variational Problems*, Springer–Verlag, Berlin (1984).

[B] Approximations externes par éléments finis d'ordre 1 et 2 du problème de Dirichlet pour l'opérateur biharmonique, Méthode itérative de résolution des problèmes approchés; *Topics in Numerical Analysis*, J. Miller ed., Academic Press, New York (1973).

GLOWINSKI R., PIRONNEAU O.

[A] Numerical methods for the first biharmonic equation and for the two-dimensional Stokes problem, *SIAM Review*, **17**, 167–212 (1979).

GRESHO P., GRIFFITHS D., LEE R., SANI R.

[A] The cause and cure of the spurious pressures generated by certain GFEM solutions of the incompressible Navier–Stokes equations, *Int. J. Numer. Methods Fluids*, **1**, (part 1), 17–44; (part 2) 171–204 (1981).

GRIFFITHS D.

[A] The construction of approximately divergence-free finite elements, in *Mathematics of Finite Elements and Applications* (J.R. Whiteman, ed.) Academic Press (1979).

[B] Finite elements for incompressible flow, *Math. Methods Appl. Sci.*, **1**, 16–31 (1979).

GRISVARD P.

[A] *Boundary Value Problem in Non-Smooth Domains*, Lectures Notes, 19, University of Maryland (1980).

[B] *Elliptic Problems in Non-Smooth Domains*, Pitman, Marshfields, Mass. (1985).

HARLOW F.H., WELCH J.E.

[A] Numerical calculation of time dependant viscous incompressible flow, *Phys. Fluids*, **8**, 2182 (1965).

HASLINGER J., HLAVACEK I.

[A] Convergence of an equilibrium finite element method based on the dual variational principle, *Appl. Math.*, **21**, 43–65 (1976).

[B] A mixed finite element method close to the equilibrium model, *Numer. Math.*, **26**, 85–97 (1976).

HECHT F.

[A] Construction d'une base de fontions P1 non-conformes à divergence nulle dans $\mathbb{R}^3$, *R.A.I.R.O. Anal. Numer.*, **15**, 119–150 (1981).

HELLAN K.

[A] Analysis of elastic plates in flexure by a simplified finite element method, *Acta Polytechnica Scandinavia, Civil Engineering Series*, 46, Trondheim (1967).

HENNART J.P., JAFFRE J., ROBERTS J.E.
[A] A constructive method for deriving finite elements of nodal type, *Numer. Math.*, **55**, 701–738 (1988).

HERRMANN L.R.
[A] Finite element bending analysis for plates, *J. Eng. Mech. Div. ASCE*, **93**, EM5, 13–26 (1967).

HESTENES M.
[A] Multiplier and gradient methods, *J. Opt. Theory Appl.*, **4**, 303–320 (1969).

HLAVACEK I.
[A] Convergence of an equilibrium finite element model for plane elastostatics, *Apl. Mat.*, **24**, 427–457 (1979).

HOOD P., TAYLOR C.
[A] Numerical solution of the Navier–Stokes equations using the finite element technique, *Comput. Fluids*, **1**, 1–28 (1973).
[B] Navier-Stokes equations using mixed interpolation, *Finite Element Methods in Flow Problems* (J.T. Oden, ed.), UAH Press, Huntsville, Alabama (1974).

HUGHES T.J.R.
[A] *The Finite Element Method: Linear Static and Dynamic Finite Element Analysis*, Prentice-Hall, Englewood Cliffs N.J. (1987).

HUGHES T.J.R., ALLIK H.
[A] Finite elements for compressible and incompressible continua, in *Proceedings of the Symposium on Civil Engineering*, Vanderbilt University, Nashville Tenn. (1969), pp. 27–62.

HUGHES T.J.R., BREZZI F.
[A] On drilling degrees of freedom, *Comp. Methods Appl. Mech. Eng.*, **72**, 105–121 (1989).

HUGHES T.J.R., BROOKS A. (See BROOKS A., HUGHES T.J.R.)

HUGHES T.J.R., FRANCA L.P.
[A] A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions, symmetric formulations that converge for all velocity-pressure spaces, *Comp. Methods. Appl. Mech. Eng.*, **65**, 85–96 (1987).
[B] A mixed finite element formulation for Reissner–Mindlin plate theory: uniform convergence of all higher-order spaces, *Comp. Methods. Appl. Mech. Eng.*, **67**, 223–240 (1988).

HUGHES T.J.R., FRANCA L.P., BALESTRA, M.
[A] A new finite element formulation of computational fluid dynamics: a stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations, *Comp. Methods. Appl. Mech. Eng.*, **59**, 85–99 (1986).

HUGHES T.J.R., HULBERT G.M.
[A] Space-time finite element methods for elastodynamics: Formulations and error estimates, *Comp. Methods Appl. Mech. Eng.*, **66**, 339–363 (1988).

HUGHES T.J.R., TEZDUYAR T.E.
[A] Finite elements based upon Mindlin plate theory with particular reference to the four-node bilinear isoparametric element, *J. Appl. Mech.*, **48**, 587–596 (1981).

IRONS B.M., RAZZAQUE A.
[A] Experience with the patch-test for convergence of finite elements, in *Mathematics of Finite Element Method with Applications to Partial Differential Equations* (A.K. Aziz, ed.), Univ. of Maryland, Baltimore, (1972).

JAMET P.
[A] Estimation d'erreur pour des éléments finis droits presque dégénérés, *R.A.I.R.O. Anal. Numer.*, **10**, 3, 43–62 (1976).

JOHNSON C.
[A] On the convergence of a mixed finite element method for plate bending problems, *Numer. Math.*, **21**, 43–62 (1973).
[B] *Numerical Solution of Partial Differential Equations by the Finite Element Method*, Cambridge University Press, Cambridge (1987).

JOHNSON C., MERCIER B
[A] Some equilibrium finite element methods for two-dimensional elasticity problems, *Numer. Math.*, **30**, 103–116 (1978).

JOHNSON C., PITKÄRANTA J.
[A] Analysis of some mixed finite element methods related to reduced integration, *Math. Comp.*, **38**, 375–400 (1982).

JOHNSON C., THOMEE V.
[A] Error estimates for some mixed finite element methods for parabolic type problems, *R.A.I.R.O. Anal. Numer.*, **15**, 41–78 (1981).

KIKUCHI F.
[A] On a finite element scheme based on the discrete Kirchhoff assumption, *Numer. Math.*, **24**, 211–231 (1975).

KIKUCHI F., ANDO Y.
[A] On the convergence of a mixed finite element scheme for plate bending, *Nucl. Eng. Design*, **24**, 357–373 (1973).

KIKUCHI N., ODEN J.T.
[A] *Contact problems in Elasticity: A Study of Variational Inequalities and Finite Element Methods*, SIAM Studies in Applied Mathematics, SIAM, Philadelphia (1988).

LADYZHENSKAYA O.A.
[A] *The Mathematical Theory of Viscous Incompressible Flow*, Gordon and Breach, New York (1969).

LASCAUX P., LESAINT P.
[A] Some non-conforming finite elements for the plate bending problem, *R.A.I.R.O. Anal. Numer.*, **9**, 9–53 (1975).

LEROUX M.-N.
[A] A mixed finite element method for a weighted elliptic problem, *R.A.I.R.O. Anal. Numer.*, **16**, 243–273 (1982).

LESAINT P.
[A] Nodal methods for the transport equation, *The Mathematics of Finite Elements and Applications V*, MAFELAP 984, Proc 5th Conf., Oxbridge, England (1985), pp. 563–569.

LE TALLEC P.
[A] Existence and approximation results for nonlinear mixed problems. Application to incompressible finit elasticity, *Numer. Math.*, **38**, 365–382 (1982).

LE TALLEC P., RUAS V.
[A] On the convergence of the bilinear velocity-constant pressure finite method in viscous flow, *Comp. Methods Appl. Mech. Eng.*, **54**, 235–243 (1986).

LIONS J.L.
[A] *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod Gauthier Villars, Paris (1968).

LIONS J.L., MAGENES E.
[A] *Problèmes aux limites non-homogènes et applications*, Dunod, Paris (1968).

MALKUS D.S.
[A] Eigenproblems associated with the discrete LBB-condition for incompressible finite elements, *Int. J. Eng. Sci.*, **19**, 1299–1310 (1981).

MALKUS D.S, HUGHES T.J.R.
[A] Mixed finite element methods. Reduced and selective integration techniques : a unification of concepts, *Comp. Methods Appl. Mech. Eng.*, **15**, 63–81 (1978).

MANSFIELD L.

[A] On finite element subspaces on quadrilateral and hexahedral meshes for incompressible viscous flow problems, *Numer. Math.*, **45**, 165–172 (1984).

MARINI L.D.

[A] Implementation of hybrid finite element methods and associated numerical problems, part 1, *Publ. 36*, IAN-CNR, Pavia (1976).

[B] Implementation of hybrid finite element methods and associated numerical problems, part 2, *Publ. 182*, IAN-CNR, Pavia (1978).

[C] An inexpensive method for the evaluation of the solution of the lower order Raviart–Thomas mixed method, *SIAM J. Numer. Anal.*, **22**, 493–496 (1985).

MARINI L.D., PIETRA P.

[A] An abstract theory for mixed approximations of second order elliptic problems, *Mat. Aplic. Comp.*, **8** (3), 219–239 (1989).

[B] New mixed finite element schemes for current continuity equations, *Compel.* **9** (4), 257-268 (1990).

MARINI L.D., SAVINI A.

[A] Accurate computation of electric field in reverse biased semi conductor devices: a mixed finite element approach, *Compel*, **3**, 123–135 (1984).

MARSDEN J.E., HUGHES T.J.R.

[A] *Mathematical Foundations of Elasticity*, Prentice-Hall, Englewood Cliffs, N.J. (1983).

MERCIER B.

[A] Numerical solution of the biharmonic problem by mixed finite elements of class C^0, *Boll. U.M.I.*, **10**, 133–149 (1974).

MERCIER B., OSBORN J., RAPPAZ J., RAVIART P.A.

[A] Eigenvalue approximation by mixed and hybrid methods, *Math. Comp.*, **36**, 427–453 (1981).

MGHAZLI Z.

[A] *Une méthode mixte pour les équations de l'hydrodynamique*, Thèse, Université de Montréal (1987).

MIYOSHI T.

[A] A finite element method for the solution of fourth order partial differential equations, *Kumamoto J. Sci. (Math.)*, **9**, 87–116 (1973).

MORLEY M.

[A] A family of mixed finite elements for linear elasticity, *Numer. Math.* **55**, 633–666 (1989).

NEČAS J.

[A] *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris (1967).

NEDELEC J.C.

[A] Mixed finite elements in $\mathbb{R}^3$, *Numer. Math.*, **35**, 315–341 (1980).

[B] A new family of mixed finite elements in $\mathbb{R}^3$, *Numer. Math.*, **50**, 57–81 (1986).

NICOLAIDES R.A.

[A] Existence, uniqueness and approximation for generalized saddle point problems, *SIAM J. Numer. Anal.*, **19**, 349–357 (1982).

NITSCHE J.

[A] Ein kriterium fur die quasi-optimalitat des Ritzchen Verfahrens, *Numer. Math.*, 346–348 (1968).

ODEN J.T., JACQUOTTE O.

[A] Stability of some mixed finite element methods for Stokesian flows, *Comp. Methods Appl. Mech Eng.*, **43**, 231–247 (1984).

[B] Stable and unstable RIP/perturbed lagrangian methods for two-dimensional viscous flow problems, *Finite Elements in Fluids V*, John Wiley and Sons, New York (1984).

ODEN J.T., KIKUCHI N., SONG Y.J

[A] Penalty finite element methods for stokesian flows, *Comp. Meth. Appl. Mech. Eng.*, **31**, 297–329 (1982).

ODEN J.T., REDDY J.N.

[A] On mixed finite element approximations, *SIAM J. Numer. Anal.* **13**, 393–404 (1976).

PIAN T.H.H.

[A] Formulations of finite element methods for solid continua, in *Recent Advances in Matrix Methods Structural Analysis and Design* (R.H.Gallagher, Y. Yamada and J.T. Oden, eds.), The University of Alabama Press (1971).

[B] Finite element formulation by variational principles with relaxed continuity requirements, in *The Mathematical Foundations of the Finite Element Method with Applications to Partial Differential Equations* (A.K. Aziz, ed.), Academic Press, New York (1972).

PIAN T.H.H., TONG P.

[A] Basis of finite element methods for solid continua, *Int. J. Numer. Methods Eng.*, **1**, 3–28 (1969).

PIERRE R,

[A] Simple C^0-approximations for the computation of incompressible flows, *Comp. Methods. Appl Mech. Eng.* **68** 205–227 (1988).

[B] Regularization procedures of mixed finite element approximation of the Stokes problem, *Numerical Methods for Partial Differential Equations*, **5**, 241-258 (1989).

[C] Convergence properties and numerical approximation of the solution of the Mindlin plate bending problem, *Math. Comp.*, **51**, 15–25 (1988).

PIRONNEAU O.A.

[A] *Méthodes d'éléments finis pour les fluides*, Masson, Paris (1988).

[B]*Finite Element Methods for Fluids*, John Wiley and Sons (1989). (English Version of [A].)

PIRONNEAU O.A., RAPPAZ J.

[A] Numerical analysis for compressible viscous isentropic stationary flows, *Impact Comput. Sci. Eng.*, **1**, 109–137 (1989).

PITKÄRANTA J.

[A] Analysis of some low-order finite element schemes for Mindlin–Reissner and Kirchoff plates, *Numer. Math.*, **53**, 237–254 (1988).

PITKÄRANTA J., STENBERG R.

[A] Analysis of some mixed finite element methods for plane elasticity equations, *Math. Comp.*, **41**, 399–423 (1983).

POWELL M.J.D.

[A] A method for non-linear constraints in minimization problems, in *Optimization* (R. Fletcher, ed.), Academic Press, London (1969).

QUARTERONI A.

[A] Error estimates for the assumed stresses hybrid methods in the approximation of fourth order elliptic equations, *R.A.I.R.O. Anal. Numer.*, **13**, 355–367 (1979).

[B] On mixed methods for fourth order problems, *Comp. Methods Appl. Mech. Eng.*, **24**, 13–24 (1980).

[C] Mixed approximations of evolution problems, *Comp. Methods Appl. Mech. Eng.*, **24**, 137–163 (1980).

RANNACHER R.

[A] On nonconforming and mixed finite element methods for plate bending problems. The linear case, *R.A.I.R.O. Anal. Numer.*, **13**, 369–387 (1979).

RAVIART P.A.

[A] *Méthode des éléments finis*, Université Paris VI, Paris (1972).

RAVIART P.A., THOMAS J.M.

[A] A mixed finite element method for second order elliptic problems, *Mathematical Aspects of the Finite Element Method* (I. Galligani, E. Magenes, eds.), Lectures Notes in Math. 606, Springer-Verlag, New York (1977).

[B] Primal hybrid finite element methods for second order elliptic equations, *Math. Comp.*, **31**, 391–413 (1977).

[C] Dual finite element models for second order elliptic problems, in *Energy Methods in Finite Element Analysis* (R. Glowinski, E.Y. Rodin and O.C. Zienkiewicz, eds.), John Wiley and Sons, Chichester (1979).

[D] Introduction à l'analyse numérique des équations aux dérivées partielles, Masson, Paris (1983).

REISSNER E.,

[A] On a variational theorem in elasticity, *J. Math. Physics,* **29**, 90–95 (1958).

[B] On a variational theorem for finite elastic deformations *J. Math. Physics.* **32**, 129-135 (1953).

ROBERTS J.E., THOMAS J.M.

[A] Mixed and hybrid methods, in *Handbook of Numerical Analysis*, (P.G. Ciarlet and J.L. Lions, eds.), *Vol.II, Finite Element Methods (Part 1)*, North-Holland, Amsterdam (1989).

ROCKAFELLAR R.T.

[A] *Convex Analysis*, Princeton University Press, Princeton, N.J. (1970).

RUAS V.

[A] Finite element solution of three-dimensional viscous flow problems using non standard degrees of freedom, *Japan J. Appl. Math.*, **2**, 415–431 (1985).

SCAPOLLA T.

[A] A new abstract framework for special mixed methods with an application to plate bending problems, *Math. Appl. Comp.*, **4**, 219–244 (1985).

[B] A mixed finite element method for the biharmonic problem, *R.A.I.R.O. Anal. Numer.*, **14**, 55–79 (1980).

SCHOLZ R.

[A] Approximation von sattelpunkten mit finiten elementen, *Bonner Mathematischen Schriften*, **89**, 54–66 (1976).

[B] L^∞-convergence of saddle point approximations for second order problems, *R.A.I.R.O. Anal. Numer.*, **11**, 209–216 (1977).

[C] A mixed method for fourth order problems using linear finite elements, *R.A.I.R.O. Anal. Numer.*, **12**, 85–90 (1978).

[D] A remark on the rate of convergence for a mixed finite element method for second order problems, *Numer. Funct. Anal. Optim.*, **4**, 3, 269–277 (1982).

[E] Optimal L^∞-estimates for a mixed finite element method for elliptic and parabolic problems, *Calcolo*, **20**, 355–377 (1983).

SCOTT L.R., VOGELIUS M.

[A] Norm estimates for a maximal right inverse of the divergence operator in spaces of piecewise polynomials, *Math. Modelling Numer. Anal.*, **9**, 11–43 (1985).

STENBERG R.

[A] Analysis of mixed minite mlement methods for the Stokes problem: a unified approach, *Math. Comp.*, **42**, 9–23 (1984).

[B] On the construction of optimal mixed finite element methods for the linear elasticity problem, *Numer. Math.*, **48**, 447–462 (1986).

[C] On the postprocessing of mixed equilibrium finite element methods, in *Numerical Tehchniques in Continuum Mechanics* (W. Hackbusch, K. Witsch, eds.), Proceedings of the Second GAMM-Seminar, Kiel 1986, Veiweg, Braunschweig (1987).

[D] On Some three-dimensional finite elements for incompressible media, *Comp. Methods. Appl. Mech. Eng.*, **63**, 261–269 (1987).

[E] Error Analysis of Some Finite Element Methods for the Stokes Problem, Rapport de recherche 948, INRIA, Domaine de Voluceau, B.P.105, 78153, Le Chesnay, France (1988).

[F] A family of mixed finite elements for the elasticity problem, *Numer. Math.*, **53**, 513–538 (1988).

[G] Two low-order mixed methods for the elasticity problem, *The Mathematics of Finite Elements and Applications VI*, J.R. Whiteman ed, Academic Press, London, pp 271–280 (1988).

STRANG G., FIX G.J.

[A] An Analysis of the Finite Element Method, Prentice–Hall, New York (1973). (Now published by Wellesley-Cambridge Press.)

STUMMEL F.

[A] The generalized patch test *SIAM, J. Numer. Anal.*, **16**, 449–471 (1979).

TEMAM R.

[A] Navier–Stokes Equations, North-Holland, Amsterdam (1977).

THOMAS J.M

[A] Méthode des éléments finis hybrides duaux pour les problèmes du second ordre, *R.A.I.R.O. Anal. Numer.*, **10**, 51–79 (1976).

[B] Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes, Thèse d'Etat, Université Pierre et Marie Curie, Paris (1977).

THOMASSET F.

[A] *Implementation of Finite Element Methods for Navier–Stokes Equations*, Springer Series in Comp. Physics, Springer–Verlag, Berlin (1981).

TONG P., PIAN T.H.H.

[A] A variational principle and the convergence of a finite element method based on assumed stress distribution, *Int. J. Solids Struct.*, **5**, 463–472 (1969).

VARGA R.

[A] *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J. (1962).

VERFÜRTH R.

[A] Error estimates for a mixed finite element approximation of the Stokes equation, *R.A.I.R.O. Anal. Numer.*, **18**, 175–182 (1984).

[B] A combined conjugate gradient multi-grid algorithm for the numerical solution of the Stokes problem, *IMA J. Numer. Anal.*, **4**, 441–455 (1984).

[C] On the preconditioning of non-conforming solenoidal finite element approximations of the Stokes equation (to appear).

VOGELIUS M.

[A] An analysis of the p-version of the finite element method for nearly incompresssible material - Uniformly valid, optimal error estimates, *Numer. Math.*, **41**, 39–53 (1983).

WHEELER M., GONZALEZ R.

[A] Mixed finite element methods for petroleum reservoir engineering problems, in *Computing Methods in Applied sciences and Engineering VI* (R. Glowinski and J.L. Lions, eds.), North-Holland, Amsterdam (1984).

YOSIDA K.

[A] *Functional Analysis*, Springer-Verlag, New York (1966).

ZIENKIEWICZ O.C.

[A] *The Finite Element Method*, McGraw-Hill, London (1977).

ZIENKIEWICZ O.C., QU S., TAYLOR R.L., NAKAZAWA S.

[A] The patch text for mixed formulations, *Int. J. Numer. Methods Eng.*, **23**, 1873–1883 (1986).

Index

A
Adini's element, 273
Admissible discretizations, 259-260
Affine elements, 107-108, 226
Affine finite elements, 104-105
Approximations
    external, 110-111
    internal, 102
    non polynomial, 114
Argyris' triangle, 107
Arnold-Douglas-Gupta families, 295
Aubin-Nitsche duality technique, 72

B
Babŭska-Brezzi condition, 58
Barycentric coordinate, 117
Basis function, 271-272
Beam problems, 302
Bending moments, 22
Bibliography, 326-344
Biharmonic problem, 65, 166
    decomposition of, 20-21
Bilinear form, 46
    duality methods for, 23
Bilinear velocity-constant pressure elements, 242-248
Block diagonal matrix, 182-183
Boundary conditions
    Dirichlet, 144, 196
    homogeneous, 206-207
    non-homogeneous, 206
Bramble-Hilbert lemma, 108
Brezzi-Douglas-Fortin-Marini space, 121
Bubble functions, 110
    adding, 215

C
Checkerboard pressure modes, 230, 231
Clement's operator, 223
Coercive bilinear form, 23
Coercive form, 3
Coerciveness, 64
    inf-sup condition and, 201
Commuting diagram property, 133
Complementary energy, 20
Complementary energy principle, 18
Composite quadrilateral elements, 232
Condition number, 89
Conforming elements, 272-273
    standard, 104
Conforming methods, 102-110
Conjugate function, 13
Consistency error, 267
Consistency terms, 69
Constitutive law, 9-10
Constrained minimization problem, 268-269
Constraint ratio, 207-208
Continuous bilinear form, 23
Continuous boundary, Lipschitz, 4, 5, 257
Continuous functions, 102
Continuous interpolate, 17
Continuous lifting, 40
    uniformly, 59
Continuous operator, 264
Continuous pressure approximations, 324
Continuous pressure elements
    stability of, 226-227
    stable, 215-216
Convergence rate, 302
Convex function, 13
Corner forces, 169
Creeping flow problem, *see* Stokes problem
Cross-grid divergence-free elements, 241
Cross-grid elements, 210, 231-233
Crouzeix-Raviart elements, 213-215
    stability proof for, 224-225
    for Stokes problem, 317

Stokes problem discretized by, 320
Cubes, mesh of, 208
Curl, 120
Curl operation, discrete, 273

D
Decomposition principle, 307-309
Deviatoric, 8
Dirac measures, 281
Dirichlet boundary conditions, 144,196
Dirichlet conditions, non-homogeneous, 17
Dirichlet problem, 6
discrete, 309
domain decomposition for, 45
domain decomposition method for, 25-26
dual, weak form of, 18-20
dualization of, 17-18
mixed finite element methods for,139-144
mixed formulation of, 44
non-standard methods for, 137-158
Discontinuous pressure elements, 324-325
stability of, 227-228
Discrete curl operation, 273
Discrete Dirichlet problem, 309
Discrete divergence operator, 205-206
Discrete stream functions, 268-274
Discretization methods, 33
Discretizations
admissible, 259-260
of Mindlin-Reissner problems, 323-325
stable, of Stokes problem, 158-162
Divergence-free approximations, 211-212
Divergence-free condition, 14
Divergence-free elements, cross-grid, 241
Divergence-free subspaces, 269
Divergence-free vectors, 120
Divergence operator, 212
discrete, 205-206
standard, 206
Domain, partition of, 3
Domain decomposition method, 138
for Dirichlet problem, 25-26, 45
dual problem of, 27
Dual Formulation problem, 38
Dual hybrid methods, 28, 153-158
for plate bending problems, 169-178
Dual norm, 93
Dual problem, 13
discrete form of, 76
of domain decomposition method, 27
for Stokes problem, 15-16
Duality methods, 12-23
for nearly imcompressible elasticity, 16-17
for non-symmetric bilinear forms, 23
Dualization
of Dirichlet problem, 17-18
for linear elasticity problem, 20

E
Eigenvalue problem, inf-sup condition and, 77-80
Elastic body, three-dimensional, 296-297
Elasticity
Hellan-Hermann-Johnson meth in, 28-30
linear, 7-10
nearly incompressible, 259-268
Elasticity problems, 12, 162-165
linear, *see* Linear elasticity problems
Elements, 96; *see also* Pressure elements
Adini's, 273

affine, 107-108, 226
bilinear velocity-constant, 242-248
choice of, 202
conforming, *see* Conforming elements
cross-grid, 210, 231-233
cross-grid divergence-free, 241
Crouzeix-Raviart, *see* Crouzeix-Raviart elements
finite, *see* Finite elements
Hermite type, 104, 106-107
for incompressible materials, 208-221
inequality for, 114-115
isoparametric triangular, 105
Lagrange type, 104
macro-elements, *see* Macro-element *entries*
MINI, *see* MINI elements
Morley, 283
non-conforming, 218-219, 270-272
quadrilateral, *see* Quadrilateral elements
Raviart-Thomas, 320
reference, 98
second-order, 219
shape of, 108
Taylor-Hood, *see* Taylor-Hood elements
three-dimensional, 219-221
Union-Jack, 241-242
Elliptic problems, linear, mixed finite element method for, 179-201
Ellipticity, 53
Energy functional, 2
Equal interpolation methods, 209, 212-213
Equilibrium condition, 3
Equilibrium methods, 138
Error analysis for interelement multipliers, 186-194
Error estimates, 194-195
Exponential fitting method, 198
External approximations, 110-111

F
Family of triangulations, 109
Finite element method, 1
Finite elements
affine, 104-105
defining, 3
mixed methods for, *see* Mixed finite element methods
serendipity, 106
Flow problems, incompressible materials and, 202-275
Fourth-order problem, mixed, 166-168
Function space, *see* Functional spaces
Functional spaces, 3, 92-135
finite element approximations of, 102-115
partitioning, 96-98
properties of, 4-6

G
Galerkin's method, 3
Generalized Taylor-Hood elements, 253-259
Global pressure modes, 232
Gradient method, 253
Green operator, 16
Green's formula, 94-95

H
Hellan-Hermann-Johnson method, 28-30
Hermite type elements, 104, 106-107
Higher order methods, 229-230
Homogeneous boundary conditions, 206-207
Hood elements, *see* Taylor-Hood elements
Hybrid methods, 24-30, 138
dual, *see* Dual hybrid methods

primal, *see* Primal hybrid methods

I
Incompressibility condition, 10
Incompressible elasticity, nearly, 259-268
duality method for, 16-17
Incompressible flow, viscous, Stokes problem for, 10
Incompressible materials
almost, 203
elements for, 208-221; *see also* Elements
flow problems and, 202-275
Inequality for elements, 114-115
Inexact integration effects, 264-268
inf-sup condition, 58-62
checking, 209
coerciveness and, 201
continuous, 60
discrete, 60, 155
eigenvalue problem and, 77-80
importance of, 80-82
standard techniques of proof for, 221-230
for Stokes problem, 323
Injectivity, 53
Integration effects, inexact, 264-268
Integration methods, reduced, 260-264
Interelement multipliers, 180-183
error analysis for, 186-194
Internal approximations, 102
Internal nodes, 237
choice of, 238
Interpolate, 107
Interpolation operator, 127, 221
Interpolation spaces, 5
Invertibility, 43
Isoparametric quadrilateral elements,105-106
Isoparametric triangular elements, 105

J
Jacobians, 261-263
Jump, 29, 169

K
Kernel, 38
characterizing, 206
Kernels property, 285
Korn's inequality, 162, 298

L
Lagrange multiplier, 26
Lagrange type elements, 104
Lagrangian algorithm, augmented, 90
Lamé coefficients, 9
Lax-Milgram theorem, 38, 162
Linear constraints, quadratic problems under, 38-45
Linear continuous operator, 38
Linear elasticity, 7-10
Linear elasticity problems, 162-165
dualization for, 20
mixed methods for, 284-296
Linear thin plates, mixed methods for, 276-284
Lipschitz continuous boundary, 4, 5, 257
Local pressure modes, 232, 235-236
Locking mechanism, 204
Locking phenomenon, 81, 208, 210

M
MAC cells, 126, 274
Macro-element techniques, 230-253, 237-241
Macro-elements, 210
Matrix form of discrete problem, 75-76
Mesh
of cubes, 208
quasi-uniform, 250
rectangular, 208, 251-252
of triangles, 208
Mindlin model, 297
Mindlin-Reissner plates, 296-325

Mindlin-Reissner problems,
discretization of, 323-325
MINI elements, 215-216, 219
stability proof for, 225-226
Minimization problem, 6
constrained, 268-269
Mixed approach, truly, 163-164
Mixed finite element methods, 138
for Dirichlet's problem, 139-144
for linear elliptic problems, 179-21
for semi-conductor devices, 196-198
Mixed formulation, 20
of Dirichlet problem, 44
Mixed Formulation problem, 137-138
Mixed fourth-order problem, 166-168
Mixed methods
for finite elements, *see* Mixed finite element methods
for linear elasticity problems, 284-296
for linear thin plates, 276-284
penalty methods and, 202, 204
Mixed type problem, 6
Stokes problem as, 204-208
Moderately thick plates, 296-325
Morley elements, 283
Morley's triangle, 113, 270
Multiplier(s)
interelement, *see* Interelement multipliers
Lagrange, 26

N
Natural boundary conditions, 11
Natural norms, 194
Navier-Stokes equation, 274
Nearly incompressible elasticity, *see* Incompressible elasticity, nearly
Neumann boundary conditions, 196
Neumann conditions, 19
Neumann problem, 7
using Raviart-Thomas elements, 320
variational, 93
Non-conforming elements, 218-219, 270-272
Non-conforming methods, 67, 110-113
Non-homogeneous boundary conditions, 206
Non-homogeneous Dirichlet conditions, 178
Non polynomial approximations, 114
Normal trace, 18
Numerical integration concept, 67
Numerical quadrature formula, 260-261

O
Optimality, 111
Oscillations, appearance of, 249

P
Particular solutions, 170
Partition of domain, 3
Patch-test, 111, 208, 240
Penalty methods
mixed methods and, 202, **204**
solution by, 83-89
stabilization by, 88-89
standard, 90
Penalty term, 84
exact evaluation of, 260
Piola's transformation, 100
Poincaré inequality, 6
Point values, 104
Polynomial spaces, 103
Pressure, in Stokes problem, 13-15
Pressure approximations, continuous, 324
Pressure elements, *see also* Elements
bilinear, velocity-constant, 242-248
continuous, *see* Continuous pressure elements

discontinuous, *see* Discontinuous pressure elements
Pressure modes, 247
checkerboard, 230, 231
global, 232
local, 232, 235-236
spurious, 200, 212, 230-233
Primal Formulation problem, 137
Primal hybrid methods, 144-152
simplest case of, 148-149

Q

Quadratic problems under linear constraints, 38-45
Quadrature errors, estimating, 267-268
Quadrature formula, numerical, 260-261
Quadrilateral elements, 217-218
composite, 232
isoparametric, 105-106
Quasi-uniform mesh, 250
Quasi-uniform triangulation, 323

R

Raviart-Thomas elements, Neumann problem using, 320
Raviart-Thomas space, 121
Rectangular approximations, 211
Rectangular mesh, 208, 251-252
Reduced integration approximation, 313-314
Reduced integration methods, 260-264
Reference elements, 98
Regularity results, 7
Rigid modes, 178
Ritz's method, 2

S

Saddle point, 21
Saddle point condition, 3
Saddle point problems, 37-91
approximation of, 52-74
discrete, numerical properties of, 75-82
dual error estimates for, 72-74
error estimates for, 55-59
existence and uniqueness of solutions for, 37-52
extensions of error estimates for, 62-64
generalizations of error estimates for, 64-66
iterative solution methods, 89-90
perturbations of, 67-71
solution by penalty methods, 83-89
Scaling arguments, 114-115
Second-order elements, 219
Semi-conductor devices, mixed finite element methods for, 196-198
Semi-norm, 66
Serendipity finite elements, 106
Shape of elements, 108
Sharfetter-Gummel method, 198
Simple functions, 3
Singular value, generalized, 77
Singular value problem, generalized, 78
Slotboom variable, 196
Smoothing post-processors, 194
Sobolev spaces, 4-5, 92-94
of fractional order, 5
Spaces, 114
functional, *see* Functional spaces
interpolation, 5
polynomial, 103
Spurious pressure modes, 200, 212, 230-233
Stabilization by penalty methods, 88-89
Stabilization procedures, 248-253
Stable continuous pressure elements, 215-216
Standard conforming elements, 104
Standard divergence operator, 206
Static condensation, 183
Stokes problem, 43-44, 203

approximating, 204
Crouzeix-Raviart elements for, 317
discretized by Crouzeix-Raviart elements, 320
dual problem for, 15-16
inf-sup condition for, 323
as mixed problem, 204-208
pressure in, 13-15
specificity of, 202
stable discretizations of, 158-162
for viscous incompressible flow, 10
Strang's lemma, 111
Stream-function, 120
discrete, 268-274
Subspaces, 97
divergence-free, 269
Superconvergence property, 66
Surjective trace operator, 95-96
Surjectivity, 53

T
Tangential components, 314
Taylor-Hood elements, 213, 240
generalized, 253-259
Thin clamped plate problem, 10-11
Thin plate bending problem
decomposition of, 21-22
dual hybrid methods for, 169-178
Thin plates, linear, mixed methods for, 276-284
Three-dimensional elastic body, 296-297
Three-dimensional elements, 219-221
Trace operator, 94
surjective, 95-96
Trace(s)
of functions, 5
normal, 18
Transmission problem, 24-25
Transposition methods, 33-35
Triangles, mesh of, 208
Triangular elements, isoparametric, 105
Triangulation(s)
family of, 109
quasi-uniform, 323
Truly mixed approach, 163-164

U
Union-Jack elements, 241-242
Uzawa's algorithm, 89-90

V
Variational equations, 2, 203
augmented, 30-33
Variational Neumann problem, 93
Variational principle, 3
Velocity-constant pressure elements, bilinear, 42-248
Velocity-pressure approximations, 20
Verfurth's trick, 256-259
Viscous incompressible flow, Stokes problem for, 10

W
Wave functions, 266

Z
Zero eigenvalues, 78

SPRINGER SERIES IN COMPUTATIONAL MATHEMATICS

Editorial Board R.L. Graham, J. Stoer, R. Varga

Computational Mathematics is a series of outstanding books and monographs which study the applications of computing in numerical analysis, optimization, control theory, combinatorics, applied function theory, and applied functional analysis The connecting link among these various disciplines will be the use of high-speed computers as a powerful tool The following list of topics best describes the aims of **Computational Mathematics**: finite element methods, multigrade methods, partial differential equations, numerical solutions of ordinary differential equations, numerical methods of optimal control, nonlinear programming, simulation techniques, software packages for quadrature, and p d e solvers **Computational Mathematics** is directed towards mathematicians and appliers of mathematical techniques in disicplines such as engineering, computer science, economics, operations research and physics

Volume 1

R. Piessens, E. de Doncker-Kapenga, C.W. Uberhuber, D.K. Kahaner

QUADPACK

A Subroutine Package for Automatic Integration

1983 VII, 301 pp. 26 figs.
Hardcover ISBN 3-540-12553-1

Volume 2

J.R. Rice, R.F. Boisvert

Solving Elliptic Problems Using ELLPACK

1985 X, 497 pp. 53 figs.
Hardcover ISBN 3-540-90910-9

Volume 3

N.Z. Shor

Minimization Methods for Non-Differentiable Functions

Translated from the Russian by K.C. Kiwiel, A. Ruszczynski
1985 VIII, 162 pp
Hardcover ISBN 3-540-12763-1

Volume 4

W. Hackbusch

Multi-Grid Methods and Applications

1985 XIV, 377 pp 43 figs. 48 tabs.
Hardcover ISBN 3-540-12761-5

Volume 5

V. Girault, P-A. Raviart

Finite Element Methods for Navier-Stokes Equations

Theory and Algorithms

1986. X, 374 pp. 21 figs
Hardcover ISBN 3-540-15796-4

Volume 6

F. Robert

Discrete Iterations

A Metric Study

Translated from the French by J. Rokne
1986 XVI, 195 pp 126 figs
Hardcover ISBN 3-540-13623-1

Volume 7

D. Braess

Nonlinear Approximation Theory

1986. XIV, 290 pp
38 figs
Hardcover
ISBN
3-540-13625-8

Springer-Verlag
New York
Berlin
Heidelberg
London
Paris
Tokyo
Hong Kong
Barcelona