

NASA Contractor Report 195354

11-39
48303

p. 152

Analysis and Development of Finite Element Methods for the Study of Nonlinear Thermomechanical Behavior of Structural Components

J. Tinsley Oden
The University of Texas at Austin
Austin, Texas

May 1995

Prepared for
Lewis Research Center
Under Grant NAG3-329



National Aeronautics and
Space Administration

N95-26387

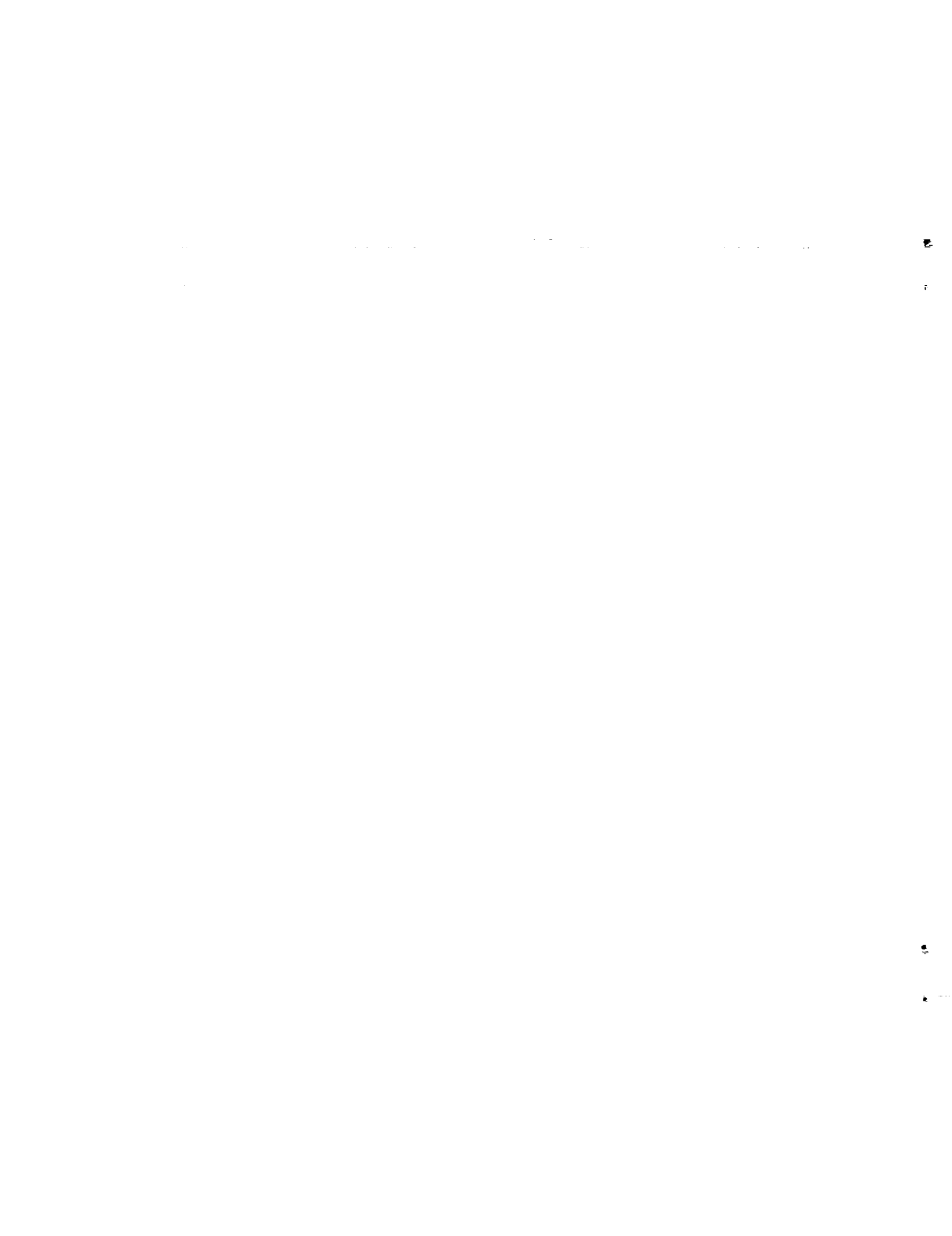
Unclass

G3/39 0048303

(NASA-CR-195354) ANALYSIS AND
DEVELOPMENT OF FINITE ELEMENT
METHODS FOR THE STUDY OF NONLINEAR
THERMOMECHANICAL BEHAVIOR OF
STRUCTURAL COMPONENTS Final
Contractor Report (Texas Univ.)
152 p

ABSTRACT

Underintegrated methods are investigated with respect to their stability and convergence properties. The focus was on identifying regions where they work and regions where techniques such as hourglass viscosity and hourglass control can be used. Results obtained show that underintegrated methods typically lead to finite element stiffness with spurious modes in the solution. However, problems exist (scalar elliptic boundary value problems) where underintegrated with hourglass control yield convergent solutions. Also, stress averaging in underintegrated stiffness calculations does not necessarily lead to stable or convergent stress states.



0. INTRODUCTION

One of the most important and widespread numerical procedures used in contemporary finite element analysis of nonlinear problems in structural mechanics is the use of so-called reduced or underintegration. Promoted strongly in the late seventies and early eighties as a means for dramatically reducing computational times in large-scale calculations, the use of underintegrated finite element methods has become common practice in a large majority of all nonlinear calculations.

A question of overriding importance that has perplexed many users of underintegrated finite element techniques for some years is whether or not these underintegrated methods are really satisfactory. It is known that underintegrated methods are frequently unstable, but these instabilities can be dampened out by the use of various types of "hourglass viscosity" or "hourglass control." It is also known that in many cases the underintegrated solutions can seem to converge at a rate equal to that of the fully integrated solutions. What are the true properties of underintegrated methods? When do they work? What criteria must hold in order that they can be used with confidence?

These are the questions that were addressed in the research project reported in this document. This final report summarizes the results of a two-year research project, supported by the NASA Lewis Research Center and carried out by Professor J. Tinsley Oden and his students at The University of Texas. Some of the principal conclusions of the work are listed as follows:

- 1) Underintegration of finite element stiffnesses generally leads to the introduction of spurious modes in the finite element solution.

These spurious modes arise from two distinctly different mechanisms: underintegration of constraint terms, which gives rise to checkerboard instabilities and the underintegration of primary stiffness terms, which gives rise to hourglass instabilities.

2) The spurious modes actually arise from expanded kernels of constraint operator and the governing differential operator. For example, an improperly underintegrated stiffness matrix will be ranked deficient and these ranked deficiencies correspond to additional modes supplied to the rigid body modes that appropriately belong to the kernel of this operator. In a similar fashion, underintegration of constraint terms leads to checkerboard modes, which belong to an expanded kernel of the constraint operator.

3) There is a significant class of problems in which, with appropriate filtering, can be shown that an underintegrated solution with hourglass control can yield very satisfactory answers, and produce a finite element method which has the same rate of convergence as the fully integrated method. The fact that this does indeed hold has been rigorously proved in the enclosed document for a class of scalar elliptic boundary value problems.

4) Unfortunately, underintegrated with hourglass control does not work uniformly on all linear or nonlinear problems, and it can lead to solutions which, while looking reasonable to the unsuspecting eye, may be grossly in error. The success of underintegrated methods seems to depend strongly on the regularity of the solution. Underintegration seems to work well in the presence of smooth solutions.

5) Most of the better known and often used underintegrated methods for constrained problems are actually unstable, but the instabilities are subtle and may be manifested only in cases in which irregular meshes

are used or in which there are irregularities in the data. In general, these unstable methods should be avoided in code development.

6) For the underintegration of constraints, such as those occurring involved with the incompressibility condition and Stokes problems are incompressible elasticity or incompressible plasticity, a necessary condition for the numerical stability of underintegrated methods is the satisfaction of a specific LBB condition. Some excellent underintegrated elements which satisfy this condition are discussed in the report. Stress averaging in underintegrated stiffness calculations does not necessarily lead to a stable or convergent stress.

0.1. Major Publications and Presentations

A number of significant papers and reports were published during the contract period. These are listed as follows:

Oden, J.T., "Penalty Method and Reduced Integration for the Analysis of Fluids," Proceedings, Symposium on Penalty Finite Element Methods in Mechanics, ASME Winter Annual Meeting, November 14-19, 1982, Phoenix, AZ.

Oden, J.T. and O.-P. Jacquotte, "Stability of Some Mixed Finite Element Methods for Stokesian Flows," Computer Methods in Applied Mechanics and Engineering, 1984, Vol. 43, No. 2, pp. 231-248.

Kikuchi, N., Oden, J.T., and Song, Y.J., "Convergence of Modified Penalty Methods and Smoothing Schemes of Pressure for Stokes Flow Problems," Finite Elements in Fluid Dynamics, Vol. V, John Wiley & Sons, Ltd., London, 1984.

Oden, J.T. and Jacquotte, O.-P., "Stable and Unstable RIP/Perturbed Lagrangian Methods for Two-Dimensional Viscous Flow Problems," Finite Elements in Fluid Dynamics, Vol. V, John Wiley & Sons, Ltd., London, 1984, pp. 127-146.

Endo, T., Oden, J.T., Becker, E. and T. Miller, "A Numerical Analysis of Contact and Limit-Point Behavior in a Class of Problems of Finite Elastic Deformation," Computers and Structures, 1984, Vol. 18, No. 5, pp. 899-910.

Jacquotte, O.-P. and Oden, J.T. Analysis of Hourglass Instabilities and Control in Underintegrated Finite Element Methods," Computer Methods in Applied Mechanics and Engineering, 1984, Vol. 44, pp. 339-363.

Jacquotte, O.-P. and Oden, J.T. "Analysis and Treatment of Hourglass Instabilities in Underintegrated Finite Element Methods," Proceedings, Symposium on Innovative Methods for Nonlinear Mechanics, ASME Winter Annual Meeting, December 12-15, 1984, New Orleans, LA.

Jacquotte, O.-P., "Stability, Accuracy, and Efficiency of Some Underintegrated Methods in Finite Element Computations," Computer Methods in Applied Mechanics and Engineering, (to appear).

Oden, J.T., Jacquotte, O.-P. and Becker, E.B., "Numerical Control of Hourglass Instability," Computers and Structures, (to appear).

Oden, J.T. and Jacquotte, O.-P., "Convergence and Stability of Underintegrated Finite Element Methods," To appear in Proceedings, ASCE/ASME Mechanics Meeting, June 24-26, 1985 at Albuquerque, NM.

0.2 Dissertations:

Jacquotte, Olivier-P., "Underintegration in Finite Element Methods," Ph.D. thesis, University of Texas, Austin, Texas, 1985.

0.3 Oral Presentations

There were three oral presentations during the research period. They are as follows:

Oden, J.T., "Stability and Convergence of Underintegrated Finite Element Approximations," Presented at the NASA-LeRC/INDUSTRY/UNIVERSITY Workshop on Nonlinear Analyses for Engine Structures, April 19-20, 1983, in Cleveland, OH.

Jacquotte, O.-P., "Analysis and Treatment of Hourglass Instabilities in Underintegrated Finite Element Methods," Presented at the ASME Winter Annual Meeting December 12-15, 1984 in New Orleans, LA.

Oden, J.T., "Convergence and Stability of Underintegrated Finite Element Methods," Presented at the ASCE/ASME Mechanics Meeting June 24-26, 1985 in Albuquerque, NM.

0.4 Personnel

The following individuals worked on technical aspects of the project during the report period:

Prof. J.T. Oden, Principal Investigator

Mr. O.-P. Jacquotte, Graduate Research Assistant

Mssrs. Lin, Martins, Strouboulis, Wu, and Manifold worked on it for a small percentage of their time.

0.5 Outline of the Technical Report

This technical report is divided into two major parts: in Part I a numerical analysis of underintegrated constraints is presented. Particular

attention is focused on the Stokes problem with a constraint divergence $u = 0$ and on construction of an appropriate LBB condition for stability.

Part II deals with underintegration and hourglass control. There projection methods, error estimates, and a large collection of numerical results are described.

PART I: STABILITY OF SOME MIXED FINITE
ELEMENT METHODS FOR STOKESIAN FLOWS

1.1. Introduction

In so-called primitive variable formulations of problems of flow of viscous, incompressible, Stokesian fluids, two fields appear as unknowns: the velocity field \underline{u} and the pressure field p , the latter representing a Lagrange multiplier associated with the incompressibility constraint, $\text{div } \underline{u} = 0$. Finite element methods based on such formulations were first introduced over a decade ago [42]. Since the mid-1970s, interest in these methods was rekindled by the appearance of several new techniques which provided for very efficient calculation of the element pressures. These included mixed methods which employ pressure approximations which are discontinuous at interelement boundaries as well as the closely related mixed-type methods which employ an exterior penalty approximation of the incompressibility condition and reduced integration of the penalty terms. All of these methods have the attractive feature that the discontinuous element pressures can be eliminated element by element, reducing the problem to one only involving velocities. Upon determining velocities, element pressures can then be evaluated through a simple post-processing operation.

Methods of this type were developed and discussed by several authors, and we mention in particular the works of Malkus [30,31], Hughes [23], Malkus and Hughes [33], Reddy [43], Bercovier [6], Engleman and Sani [15], and the references therein. In 1980, however, mathematical analyses indicated that some of the more popular discontinuous-pressure/mixed-methods might be numerically unstable [34,35-40,41]. It was discovered that while certain of these

methods perform well in problems with smooth solutions for which regular uniform meshes are employed, serious oscillations in the pressure approximation can occur when the data or the mesh pattern are mildly irregular, and these oscillations increase in amplitude as the mesh is refined.

Oden, Kikuchi, and Song [41] attributed the deficiency of these unstable methods to their failure to satisfy a key stability criterion which they referred to as the "LBB-condition," making reference to the work of Ladyzhenskaya [29] on existence theorems of viscous flow problems and of Babuška [1] and Brezzi [8] on the approximation of elliptic problems with constraints. The discrete LBB-condition of Oden, Kikuchi, and Song is basically the requirement that the discrete approximation B_h^* of the transpose B^* of the constraint operator $B \approx \text{div}$ be bounded below as a linear operator mapping the space of approximate pressures onto the dual of the space of approximate velocities. For example, one form of this condition is that there exist an $\alpha_h > 0$ such that for all $q_h \in Q_h^*$,

$$\alpha_h \|q_h\|_{L^2(\Omega)/\ker B_h^*} \leq \sup_{v_h} \frac{(q_h, \text{div } v_h)}{\|v_h\|_1}$$

Related conditions for mixed finite elements were discussed by Fortin [18] and Girault and Raviart [22]. The possibility of unstable pressure approximations is signalled by the existence of a parameter α_h which depends upon the mesh size h . Indeed, the fact that a mesh-dependent α_h corresponds to methods with "spurious pressure modes" is supported by the theoretical and numerical results of Oden et al [41] and by extensive numerical experiments of Malkus [32]. Equally important, the behavior of

* Definitions of terms displayed here are given in Section 1.3.

α_h as a function of h governs the asymptotic rate of convergence of such mixed methods.

An important question that has arisen from these considerations is whether or not stable mixed methods exist which converge at optimal rates in the energy- and L^2 -norms. The present paper is directed at resolving this question for a restricted class of problems by estimating the stability parameter α_h in the corresponding discrete LBB-condition.

The methods of proof of the LBB-condition basically fall into two categories depending on whether or not $\ker B_h^* = \ker B^*$ or $\ker B_h^* \subsetneq \ker B^*$, where $B^* = -\text{gradient} + \text{boundary conditions}$ and B_h^* is its finite element approximation. In the former case, a general method of proof can be constructed which is inspired by the work of Girault and Raviart [22], and which will be discussed in the first section. We shall concentrate next on the latter case, and present another general constructive technique for estimating α_h for uniform meshes which makes use of a discrete Poincaré-type inequality.

These two methods of proof will be presented and used to establish the LBB-condition for two elements. In the first category of stable mixed method with discontinuous pressure for which $\ker B_h^* = \ker B^*$, we shall analyze in detail the Q_2/P_1 -element (biquadratic velocity/linear discontinuous pressure), and prove that it does in fact satisfy the LBB-condition with α_h independent of h .

As far as the second category of method (for which $\ker B_h^* \subsetneq \ker B^*$) is concerned, we shall prove that the $I8/P_1$ -element (eight node isoparametric velocity/linear discontinuous pressure) satisfies the LBB-condition with α_h of order h , and therefore appears to be unstable. However, certain ways to stabilize this element are suggested and their implementation in codes have led to stable and accurate solutions.

Also falling into this second category is the Q_1/P_0 -element (bilinear velocity/piecewise constant pressure). This element will be briefly discussed.

Finally the various values of the LBB-constant and the rate of convergence expected from the most used rectangular elements and using discontinuous pressure will be summarized.

1.2 Statement of the Problem

Let Ω denote an open bounded region of \mathbb{R}^2 with boundary $\partial\Omega$. We consider the two-dimensional Stokes problem on Ω , which involves finding a velocity field $\underline{u} = (u_1, u_2)$ and a pressure field p such that

$$\left. \begin{aligned} -\nu \Delta \underline{u} + \nabla p &= \underline{f} & \text{in } \Omega \\ \operatorname{div} \underline{u} &= 0 & \text{in } \Omega \\ \underline{u} &= \underline{0} & \text{on } \partial\Omega \end{aligned} \right\} \quad (2.1)$$

where ν is the viscosity of the fluid, ($\nu = \text{const.} > 0$), and \underline{f} is the body force, assumed to be a prescribed vector field with components $f_i \in L^2(\Omega)$.

We recast (2.1) in a weaker variational framework by introducing the spaces

$$V = (H_0^1(\Omega))^2, \quad Q = L^2(\Omega) \quad (2.2)$$

and the forms

$$\left. \begin{aligned} a: V \times V &\rightarrow \mathbb{R}, \quad f: V \rightarrow \mathbb{R} \\ a(\underline{u}, \underline{v}) &= \nu (\underline{u}, \underline{v})_1, \quad f(\underline{v}) = \sum_{i=1}^2 (f_i, v_i) \end{aligned} \right\} \quad (2.3)$$

for all $\underline{u}, \underline{v} \in V$, where $(\cdot, \cdot)_1$ and (\cdot, \cdot) are inner products on V and Q , respectively, and are given by

$$\left. \begin{aligned} (v, w) &= \int_{\Omega} v w dx; \quad v, w \in Q \\ (\underline{u}, \underline{v})_1 &= \sum_{i,j=1}^2 \left(\frac{\partial u_i}{\partial x_j}, \frac{\partial v_i}{\partial x_j} \right); \quad \underline{u}, \underline{v} \in V \end{aligned} \right\} \quad (2.4)$$

The partial derivatives in $(2.4)_2$ are interpreted in a distributional sense.

We proceed by considering the problem of finding $(\underline{u}, p) \in V \times Q$ such that

$$\left. \begin{aligned} a(\underline{u}, \underline{v}) - (p, \operatorname{div} \underline{v}) &= f(\underline{v}) \quad \forall \underline{v} \in V \\ (q, \operatorname{div} \underline{u}) &= 0 \quad \forall q \in Q \end{aligned} \right\} \quad (2.5)$$

It is easily verified that any solution of (2.1) satisfies (2.5); any solution of (2.5) satisfies equations of the form (2.1) in a distributional sense. Under the conditions stated, it is also known that (2.5) possesses a solution (\underline{u}, p) , with \underline{u} uniquely determined by each choice of f and p unique up to an arbitrary constant.

Problem (2.1) can also be interpreted as the characterization of a saddle point of the functional

$$L: V \times Q \rightarrow \mathbb{R}$$

$$L(\underline{v}, q) = \frac{1}{2} a(\underline{v}, \underline{v}) - f(\underline{v}) - (q, \operatorname{div} \underline{v}) \quad (2.6)$$

with q clearly a Lagrange multiplier associated with the constraint, $\operatorname{div} \underline{v} = 0$ in Q .

We remark the saddle point problem for the functional $L(\cdot, \cdot)$ of (2.6), can be pre-conditioned by introducing the perturbed Lagrangian

$$L_{\varepsilon}: V \times Q \rightarrow \mathbb{R}$$

$$L_{\varepsilon}(\underline{v}, q) = L(\underline{v}, q) - \frac{1}{2\varepsilon} (q, q) \quad (2.7)$$

for all $q \in Q$, which represents a regularization of $L(\cdot, \cdot)$ with respect to the multipliers q . For each $\varepsilon > 0$, saddle points $(\underline{u}_\varepsilon, p_\varepsilon)$ of $L(\cdot, \cdot)$ are characterized by

$$a(\underline{u}_\varepsilon, \underline{v}) - (p_\varepsilon, \operatorname{div} \underline{v}) = f(\underline{v}) \quad \forall \underline{v} \in V \quad (2.8)$$

$$(\varepsilon p_\varepsilon + \operatorname{div} \underline{u}_\varepsilon, q) = 0 \quad \forall q \in Q$$

Upon solving the last equation in (2.8) for p_ε , we obtain

$$p_\varepsilon = -\frac{1}{\varepsilon} \operatorname{div} \underline{u}_\varepsilon \quad \text{in } Q \quad (2.9)$$

The forms $a(\cdot, \cdot)$ and $f(\cdot)$ are continuous and $a(\cdot, \cdot)$ is V -elliptic. In addition, Ladyzhenskaya [29] has shown that a constant $\alpha > 0$ exists such that

$$\alpha \|q\|_{L^2(\Omega)/\mathbb{R}} \leq \sup_{\underline{v} \in V} \frac{(q, \operatorname{div} \underline{v})}{\|\underline{v}\|_1} \quad \forall q \in Q \quad (2.10)$$

where $\|\underline{v}\|_1 = \sqrt{(v, v)_1}$. Under these conditions, the sequence $\{(\underline{u}_\varepsilon, p_\varepsilon)\}_{\varepsilon > 0}$ of solutions converge strongly in $V \times Q/\mathbb{R}$ to the saddle point (u, p) of the functional $L(\cdot, \cdot)$ in (2.6).

Finally, it is interesting to note that when (2.9) is introduced into the first equation in (2.8), one obtains

$$a(\underline{u}_\varepsilon, \underline{v}) + \frac{1}{\varepsilon} (\operatorname{div} \underline{u}_\varepsilon, \operatorname{div} \underline{v}) = f(\underline{v}) \quad \forall \underline{v} \in V \quad (2.11)$$

which is equivalent to an exterior penalty formulation of the constraint, $\operatorname{div} \underline{u} = 0$.

1.3 Finite Element Approximations

We shall outline briefly features of certain finite element approximations of (2.5) or (2.8). We confine our attention to cases in which Ω is rectangular or is the union of rectangles and, for simplicity, to uniform meshes of rectangular elements of maximum length h . For a family of such meshes with $E = E(h)$ elements, we introduce the discrete (finite-dimensional) spaces,

$$V^h = \{v_h = (v_{h1}, v_{h2}) \mid v_{hi} \in C^0(\bar{\Omega}),$$

$$v_{hi} \mid_{\Omega_e} \in Q_k(\bar{\Omega}_e); v_{hi} = 0 \text{ on } \partial\Omega, \quad ,$$

$$1 \leq e \leq E, i = 1, 2\} \quad (3.1)$$

$$Q^h = \{q_h \in L^2(\Omega) \mid q_h \mid_{\Omega_e} \in P_r(\Omega_e);$$

$$1 \leq e \leq E, r \geq 0\} \quad (3.2)$$

Here $Q_k(\bar{\Omega}_e)$ is the space of tensor products of complete polynomials in x_1 and x_2 of degree $\leq k$ defined on finite element $\bar{\Omega}_e$ and $P_r(\Omega_e)$ is the space of complete polynomials of degree $\leq r$ defined on Ω_e . The elements Q_2/P_1 and Q_1/P_0 clearly correspond to the values (2.1) and (1.0) of the parameters (k, r) .

In addition to the spaces V^h , we shall also consider cases in which V^h is constructed using I8-elements:

I8 = eight-node isoparametric elements

This element is also referred to as a serendipity element [46]. We also consider composite elements which employ both Q_2 and I8-subelements.

Clearly, for every h ,

$$V^h \subset V \quad \text{and} \quad Q^h \subset Q \quad (3.3)$$

The finite element approximation of the formulation (2.5) consists of seeking $u_h \in V^h$ and $p_h \in Q^h$ such that

$$\left. \begin{aligned} a(u_h, v_h) - (p_h, \operatorname{div} v_h) &= f(v_h) & \forall v_h \in V^h \\ (q_h, \operatorname{div} u_h) &= 0 & \forall q_h \in Q^h \end{aligned} \right\} \quad (3.4)$$

while the approximation of (2.8) is of the form

$$\left. \begin{aligned} a(u_h^\varepsilon, v_h) - (p_h^\varepsilon, \operatorname{div} v_h) &= f(v_h) & \forall v_h \in V^h \\ (\varepsilon p_h^\varepsilon + \operatorname{div} u_h^\varepsilon, q_h) &= 0 & \forall q_h \in Q^h \end{aligned} \right\} \quad (3.5)$$

The solvability of (3.4) depends upon a compatibility condition between the spaces V^h and Q^h which resembles (2.10) and which we record below.

Likewise, while (3.5) is uniquely solvable for u_h^ε and p_h^ε for any $\varepsilon > 0$ (under the stated conditions on $a(\cdot, \cdot)$), the behavior of u_h^ε and p_h^ε as ε or h tend to zero also depends upon more delicate features of the approximation.

Let B_h and B_h^* denote the discrete operators,

$$B_h : V^h \rightarrow Q^h ; \quad B_h^* : Q^h \rightarrow V^h$$

$$[q_h, B_h v_h] = \langle v_h, B_h^* q_h \rangle \equiv (q_h, \operatorname{div} v_h)$$

$$\forall q_h \in Q^h, \forall v_h \in V^h \quad (3.6)$$

where $[\cdot, \cdot]$ and $\langle \cdot, \cdot \rangle$ denote duality pairings on $Q' \times Q$ ($Q = Q' = L^2(\Omega)$) and $V' \times V$ respectively (i.e., B_h and B_h^* are the discrete approximations of div and $-\text{grad}$ plus boundary conditions defined by (\cdot, \cdot)). Then, the discrete LBB-condition for problems (3.4) and (3.5) is as follows:

There exists a number $\alpha_h > 0$ such that

$$\alpha_h \|q_h\|_{L^2(\Omega)/\ker B_h^*} \leq \sup_{\underline{v}_h \in V^h} \frac{(q_h, \text{div } \underline{v}_h)}{\|\underline{v}_h\|_1} \quad (3.7)$$

for all $q_h \in Q^h$

The behavior of α_h as h tends to zero and the structure of $\ker B_h^*$ governs the stability of these types of mixed methods. In particular, let $E_h(\underline{u}, p)$ denote the distance function

$$E_h(\underline{u}, p) \equiv \inf_{\underline{v}_h \in V^h} \|\underline{u} - \underline{v}_h\|_1 + \inf_{q_h \in Q^h} \|p - q_h\|_{L^2(\Omega)} \quad (3.8)$$

defined on $V \times \tilde{Q}$, $\tilde{Q} = \{q \in Q \mid \int_{\Omega} q dx = 0\}$. Then one can show (see Oden and Kikuchi [40]) that if (\underline{u}, p) is the solution of (2.5) and $(\underline{u}_h^{\varepsilon}, p_h^{\varepsilon})$ is the solution of (3.5) in $V^h \times Q^h$,

$$\|\underline{u} - \underline{u}_h^{\varepsilon}\|_1 \leq C(1 + \alpha_h^{-1}) (E_h(\underline{u}, p) + \varepsilon) \quad (3.9)$$

$$\|p - p_h^{\varepsilon}\|_{L^2(\Omega)} \leq C(1 + \alpha_h^{-1} + \alpha_h^{-2}) (E_h(\underline{u}, p) + \varepsilon)$$

where C is a generic constant independent of \underline{u} , p , ε , and h .

The remainder of this part is devoted to the study of (3.7) and estimations of the stability parameter α_h for different approximation spaces

v^h and Q^h ((3.1), (3.2)). As noted in the Introduction, we will focus our study on the following approximations:

- 1) Q_2/P_1 elements [biquadratic velocities, piecewise linear pressures]
- 2) $I8/P_1$ elements [eight-node isoparametric elements for velocities, piecewise linear pressures]
- 3) Composite elements [elements consisting of two or more of the above]
- 4) Q_1/P_0 elements [bilinear velocities, piecewise constant pressures]

Again, we note that in all of the cases we study, we shall assume that

$$\left. \begin{aligned} \Omega = \Omega_h \text{ is a rectangle (or a union of} \\ \text{rectangles) discretized by a uniform} \\ \text{mesh of rectangular finite elements;} \\ v^h \subset V, \text{ with } V = (H_0^1(\Omega))^2 \text{ and} \\ Q^h \subset Q = L^2(\Omega). \end{aligned} \right\} \quad (3.10)$$

The principal results concerning $\ker B_h^*$ and the LBB-constant α_h are stated in the following theorems.

Theorem I. *Let conditions (3.10) hold and let the discrete spaces v^h and Q^h be constructed using Q_2/P_1 -elements. Then $\ker B_h^* = \ker B^*$ and the stability parameter α_h in the discrete LBB-condition (3.7) is a positive constant independent of h ($\alpha_h = O(1)$). □*

Theorem II. *Let conditions (3.10) hold and suppose that v^h and Q^h are defined by $I8/P_1$ -elements. Then $\dim \ker B_h^* = 3$ and the stability parameter α_h in the discrete LBB-condition (3.7) depends linearly on h :*

$$\alpha_h = O(h) \quad \square$$

Theorem III. Let conditions (3.10) hold and suppose that V^h and Q^h are defined using composite $I_8/P_1 - Q^2/P_1$ -elements of the type shown in Fig. 1. Then $\dim \ker B_h^* = 1$ and the stability parameter α_h appearing in the discrete LBB-condition is a positive constant independent of h ; ($\alpha_h = O(1)$). \square

Theorem IV. Under the assumptions of Theorem II, if V^h and Q^h are defined using Q_1/P_0 -elements, then $\dim \ker B_h^* = 2$ and $\alpha_h = O(h)$. \square

1.4 The LBB-Condition For Q_2/P_1 Elements

In this section, we describe a general method for establishing the LBB-condition when $\ker B_h^*$ and $\ker B^*$ coincide. This procedure will then be used to prove Theorem I for Q_2/P_1 -elements, and can also be used for Theorem III. The method is embodied in the following four steps.

I. Let q_h be an arbitrary element in Q^h . Construct a vector $u_h \in V^h$ such that

$$\begin{aligned} (q_h, \operatorname{div} u_h) &= \|q_h\|_0^2 \\ \|u_h\|_1 &\leq C \|q_h\|_0 \end{aligned} \tag{4.1}$$

where $\|\cdot\|_0 = \|\cdot\|_{L^2(\Omega)}$ and C is a constant. Then

$$\sup_{v_h \in V^h} \frac{(q_h, \operatorname{div} v_h)}{\|v_h\|_1} \geq \frac{(q_h, \operatorname{div} u_h)}{\|u_h\|_1} \geq \frac{1}{C} \|q_h\|_0$$

so that $\alpha_h = 1/C$.

To construct such a u_h , we continue as follows.

* It suffices to define q_h only to within an arbitrary constant or to demand that all q_h be such that $(1, q_h) = 0$.

II. For each $q_h \in Q^h \subset Q$, $q_h \neq \text{constant}$, it can be shown (Ladyszhenkaya [29]) that a $v_h \in V$ can be found such that

$$\operatorname{div} v_h = q_h \text{ in } \Omega \text{ and } \|v_h\|_1 \leq C_1 \|q_h\|_0 \quad (4.2)$$

Let w_h denote the V -orthogonal projection of v_h onto V^h :

$$(w_h - v_h, v_h)_1 = 0 \quad \forall v_h \in V^h \quad (4.3)$$

Then

$$\|w_h\|_1 \leq \|v_h\|_1 \leq C_1 \|q_h\|_0$$

III. Set

$$e = v_h - w_h \quad (4.4)$$

We attempt to construct a u_h with the desired property (4.1) by demanding that

$$(q_h, \operatorname{div}(e - e_h)) \equiv \sum_{e=1}^E \int_{\Omega_e} q_h \operatorname{div}(e - e_h) dx = 0 \quad (4.5)$$

where

$$e_h = u_h - w_h \quad (4.6)$$

Then it is clear that

$$(q_h, \operatorname{div} v_h) = \|q_h\|_0^2 = (q_h, \operatorname{div} u_h)$$

which is (4.1)₁, and it remains only to verify that (4.1)₂ holds. Assuming that this is possible, we see that the original problem reduces to one of constructing a u_h such that (4.5) holds.

IV. To satisfy (4.5), it is sufficient to require that

$$\begin{aligned} \int_{\Omega_e} q_h \operatorname{div}(\underline{e} - \underline{e}_h) dx &= - \int_{\Omega_e} \nabla q_h \cdot (\underline{e} - \underline{e}_h) dx \\ &+ \int_{\partial\Omega_e} q_h \underline{n} \cdot (\underline{e} - \underline{e}_h) dx \\ &= 0 \end{aligned}$$

for each finite element Ω_e , \underline{n} being a unit outward normal to $\partial\Omega_e$. In many finite element meshes, each Ω_e is the image of a fixed master element $\hat{\Omega}$ under an invertible affine map F_e ,

$$F_e : \hat{\Omega} \rightarrow \Omega_e, \quad F_e \hat{x} = \underline{x} = T_e \hat{x} + \underline{b}_e \quad (4.7)$$

T_e being a 2×2 matrix and \underline{b}_e a translation vector. Then it is sufficient to construct \underline{u}_h such that

$$\int_{\hat{\Omega}} \hat{q} \cdot (\hat{e} - \hat{e}_h) d\hat{x} - \int_{\partial\hat{\Omega}} q \hat{n} \cdot (\hat{e} - \hat{e}_h) d\hat{x} = 0 \quad (4.8)$$

where $\hat{q} = q_h \circ F_e^{-1}$, $\hat{e} = \underline{e} \circ F_e^{-1}$, etc.

Remark: This procedure is next used for the Q_2/P_1 -element. But in the case where Ω is partitioned into $18/P_1$ elements, except one Q_2/P_1 element, we can show that

$$\ker B_h^* = \ker B^*$$

For this mesh, the construction II, III, IV can theoretically be made, but the essential estimate of (4.1) cannot be obtained. This remark suggests the introduction of the composite element described in Theorem III.

For the Q_2/P_1 element, the two discrete finite-dimensional spaces V^h and Q^h are defined as

$$V^h = \{ \underline{v}_h = (v_{h1}, v_{h2}) \mid v_{hi} \in C^0(\bar{\Omega});$$

$$v_{hi}|_{\Omega_e} \in Q_2(\bar{\Omega}_e), \quad v_{hi} = 0 \text{ on}$$

$$\partial\Omega, \quad \{ \leq e \leq E, \quad i = 1, 2 \}$$

$$Q^h = \{ q_h \in L^2(\Omega) \mid q_h|_{\Omega_e} \in P_1(\bar{\Omega}_e) \mid \leq e \leq E \}$$

and then using the definition

$$\begin{aligned} \ker B_h^* &= \{ q_h \in Q^h \text{ such that } \int_{\Omega} q_h \operatorname{div} \underline{v}_h \, dx = 0 \\ &\text{for all } \underline{v}_h \in V^h \} \end{aligned} \quad (4.9)$$

a simple calculation reveals that

$$\ker B_h^* = \ker B^* = \mathbb{R} \quad (4.10)$$

Then it suffices to construct a \underline{u}_h such that (4.8) holds for the master element $\hat{\Omega}$ shown in Fig.2 and to then show that \underline{u}_h satisfies (4.1)₂. We use the notation indicated in the figure; the integral appearing on the left side of (4.8) is denoted \hat{I} , and we seek \hat{u} with $\hat{u}_1 \in Q_2(\hat{\Omega})$. Observe that the shape functions associated with the indicated nodes are of the form

$$\hat{\psi}_5 = (1 - \hat{x}^2)(1 - \hat{y}^2), \quad \hat{\psi}_{12} = \frac{1}{2}(\hat{x}^2 - 1)\hat{y}(1 - \hat{y})$$

$$\hat{\psi}_{23} = \frac{1}{2}\hat{x}(1 + \hat{x})(1 - \hat{y}^2), \quad \hat{\psi}_{34} = \frac{1}{2}(1 - \hat{x}^2)\hat{y}(1 + \hat{y})$$

$$\hat{\psi}_{41} = \frac{1}{2}\hat{x}(\hat{x} - 1)(1 - \hat{y}^2), \text{ etc.}$$

and that each $\hat{q} \in P_1(\hat{\Omega})$ is of the form

$$\hat{q} = q_0 + q_1 \hat{x} + q_2 \hat{y} \quad \text{with} \quad \nabla \hat{q} = (q_1, q_2)$$

where $q_\alpha, \alpha = 0, 1, 2$ are real numbers. A simple calculation reveals that

$$\begin{aligned} \hat{I} &= q_0 \int_{\partial \hat{\Omega}} (\hat{e}_{\sim h} - \hat{e}) \cdot \hat{n} \, d\hat{s} \\ &+ q_1 \left[- \int_{\hat{\Omega}} (\hat{e}_{h1} - \hat{e}_1) d\hat{x}d\hat{y} + \int_{\partial \hat{\Omega}} \hat{x} (\hat{e}_{\sim h} - \hat{e}) \cdot \hat{n} \, d\hat{s} \right] \\ &+ q_2 \left[- \int_{\hat{\Omega}} (\hat{e}_{h2} - \hat{e}_2) d\hat{x}d\hat{y} + \int_{\partial \hat{\Omega}} \hat{y} (\hat{e}_{\sim h} - \hat{e}) \cdot \hat{n} \, d\hat{s} \right] \end{aligned} \quad (4.11)$$

It is clear that we can make $\hat{I} = 0$ by choosing $\hat{e}_{\sim h}$ (equivalently, choosing a $\hat{u}_{\sim h}$) such that the following five conditions hold:

$$(i) \quad \hat{e}_{\sim h}(\hat{a}^i) = 0, \quad 1 \leq i \leq 4$$

$$(ii) \quad \hat{e}_{\sim h}(a^{ij}) \cdot \hat{t}^{ij} = 0, \quad 1 \leq i \leq j \leq 4$$

where \hat{t} is the unit vector tangent to $\partial \hat{\Omega}$

$$(iii) \quad \int \hat{e}_{\sim h} \cdot \hat{n} \, d\hat{s} = \int \hat{e} \cdot \hat{n} \, d\hat{s}, \quad 1 \leq i \leq j \leq 4$$

$$(iv) \quad - \int_{\hat{\Omega}} \hat{e}_{h1} d\hat{x}d\hat{y} + \oint_{\partial \hat{\Omega}} \hat{x} \hat{e}_{\sim h} \cdot \hat{n} \, d\hat{s} = - \int_{\hat{\Omega}} \hat{e}_1 d\hat{x}d\hat{y} + \oint_{\partial \hat{\Omega}} \hat{x} \hat{e} \cdot \hat{n} \, d\hat{s}$$

$$(v) \quad - \int_{\hat{\Omega}} \hat{e}_{h2} d\hat{x}d\hat{y} + \oint_{\partial \hat{\Omega}} \hat{y} \hat{e}_{\sim h} \cdot \hat{n} \, d\hat{s} = - \int_{\hat{\Omega}} \hat{e}_2 d\hat{x}d\hat{y} + \oint_{\partial \hat{\Omega}} \hat{y} \hat{e} \cdot \hat{n} \, d\hat{s}$$

This set of conditions must determine the 18 independent components of

$$\hat{e}_{\sim h} \quad (\hat{e}_{hi} \in Q_2(\hat{\Omega})).$$

Conditions (i) and (ii) make 12 of the 18 degrees of freedom of $\hat{e}_{\sim h}$ zero. We are left with six coefficients:

$$\hat{e}_{h1} = \hat{e}_{h1}(a^5) \hat{\psi}_5 + \hat{e}_{h1}(a^{23}) \hat{\psi}_{23} + \hat{e}_{h1}(a^{14}) \hat{\psi}_{14}$$

$$\hat{e}_{h2} = \hat{e}_{h2}(a^5) \hat{\psi}_5 + \hat{e}_{h2}(a^{12}) \hat{\psi}_{12} + \hat{e}_{h2}(a^{34}) \hat{\psi}_{34}$$

But four of these coefficients are immediately determined from (iii) by a direct integration:

$$\begin{aligned}\hat{e}_{h2}(a^{12}) &= \frac{3}{4} \int_{a^1}^{a^2} \hat{e}_2 d\hat{s} ; & \hat{e}_{h1}(a^{23}) &= \frac{3}{4} \int_{a^2}^{a^3} \hat{e}_1 d\hat{s} \\ \hat{e}_{h2}(a^{34}) &= -\frac{3}{4} \int_{a^3}^{a^4} \hat{e}_2 d\hat{s}; & \hat{e}_{h1}(a^{14}) &= -\frac{3}{4} \int_{a^4}^{a^1} \hat{e}_1 d\hat{s}\end{aligned}$$

Thus, it remains only to determine $e_{hi}(a^5)$, $i = 1, 2$, using the last two conditions, (iv) and (v). But a direct calculation leads to the pair of equalities

$$\begin{aligned}\frac{16}{9} \hat{e}_{h1}(a^5) &= \int_{\Omega} \hat{e}_1 d\hat{x}d\hat{y} + \int_{a^1}^{a^2} \hat{x} \hat{e}_2 d\hat{s} - \frac{1}{3} \int_{a^2}^{a^3} \hat{e}_1 d\hat{s} \\ &\quad - \int_{a^3}^{a^4} \hat{x} \hat{e}_2 d\hat{s} - \frac{1}{3} \int_{a^4}^{a^1} \hat{e}_1 d\hat{s} \\ \frac{16}{9} \hat{e}_{h2}(a^5) &= \int_{\Omega} \hat{e}_2 d\hat{x}d\hat{y} - \frac{1}{3} \int_{a^1}^{a^2} \hat{e}_2 d\hat{s} - \int_{a^2}^{a^3} \hat{y} \hat{e}_1 d\hat{s} \\ &\quad + \frac{1}{3} \int_{a^3}^{a^4} \hat{e}_2 d\hat{s} - \int_{a^4}^{a^1} \hat{y} \hat{e}_1 d\hat{s}\end{aligned}$$

Hence, conditions (i)-(v) determine a vector $\hat{e}_{\sim h}$ for which $\hat{I} = 0$, as required. We easily verify that $e_{\sim h} \in v^h, e_{\sim h} |_{\Omega_e} = \hat{e}_{\sim h} \circ F_e$.

It remains to be proven that the vector $u_h = e_h + w_h$ satisfies (4.1)₂. We note that it is sufficient to prove that

$$\|e_h\|_1 \leq c_1 \|\hat{e}\|_1 \quad (4.12)$$

because

$$\begin{aligned} \|u_h\|_1 - \|w_h\|_1 &\leq \|u_h - w_h\|_1 \leq c_1 \|v_q - w_h\|_1 \\ &\leq c_1 (\|v_q\|_1 + \|w_h\|_1) \end{aligned}$$

so, since $\|w_h\|_1 \leq \|v_q\|_1$

$$\|u_h\|_1 \leq (1 + 2c_1) \|v_q\|_1 \leq C \|q_h\|_0$$

To establish (4.12), we note that for the master element*,

$$\hat{e}_h = \sum_{i=1}^9 \hat{e}_h(b_i) \hat{\psi}_i, \quad \|\hat{e}_h\|_{1,\hat{\Omega}}^2 \leq C \sum_{i=1}^9 \|\hat{e}_h(b_i)\|^2 \quad (4.13)$$

where $\{b_i\}$ are the 9 nodes of the element and $\|\cdot\|$ denotes the euclidean norm in \mathbb{R}^2 . Using the fact that

$$\begin{aligned} \left| \int_{a_i}^{a_j} \hat{e} \cdot \hat{n} \, d\hat{s} \right| &\leq C \|\hat{e}\|_{0,\partial\hat{\Omega}} \leq C \{ \|\hat{e}\|_{0,\hat{\Omega}}^2 + \|\hat{e}\|_{1,\hat{\Omega}}^2 \}^{1/2} \\ \left| \int_{a_i}^{a_j} \hat{x} \hat{e} \cdot \hat{n} \, d\hat{s} \right| &\leq C \|\hat{e}\|_{0,\partial\hat{\Omega}}, \text{ etc.} \end{aligned}$$

and the previously computed nodal values of \hat{e}_h obtained via steps (i)-(v) above, we can verify that $\|\hat{e}_h(b_i)\|^2 \leq C \|\hat{e}\|_{0,\hat{\Omega}}^2 + \|\hat{e}\|_{1,\hat{\Omega}}^2$. Thus, a constant C exists such that

$$\|\hat{e}_h\|_{1,\hat{\Omega}} \leq C \{ \|\hat{e}\|_{0,\hat{\Omega}}^2 + \|\hat{e}\|_{1,\hat{\Omega}}^2 \}^{1/2} \quad (4.14)$$

* Here and elsewhere in this paper, C denotes a generic constant independent of h and does not necessarily have the same value throughout.

We next transform this result so that it applies to a typical element Ω_e of the mesh and sum over all elements to obtain

$$\| \underline{e}_h \|_{1, \Omega} \leq C \{ h^{-2} \| \underline{e} \|_0^2 + \| \underline{e} \|_1^2 \}^{1/2} \quad (4.15)$$

In this last calculation, we used the affine map F_e of (4.7), the fact that $\| | T_e \| | \leq C_1 h$, $\| | T_e^{-1} \| | \leq C_2 h^{-1}$, and standard relations between $\| \hat{e} \|_{1, \hat{\Omega}}$ and $\| \underline{e} \|_{1, \Omega}$.

We shall next verify that

$$\| \underline{e} \|_0 \leq Ch \| \underline{e} \|_1 \quad (4.16)$$

We will then arrive at (4.12) via (4.15) and thereby complete the proof of the theorem.

To prove (4.16), we employ a duality argument of Girault and Raviart [22]. Note that

$$\| e_i \|_{0, \Omega} = \sup_{v \in L^2(\Omega)} \frac{(e_i, v)}{\| v \|_{0, \Omega}}, \quad i = 1, 2 \quad (4.17)$$

Let g be in $L^2(\Omega)$ and ϕ_g be the solution to the Dirichlet problem

$$-\Delta \phi_g = g \quad (4.18)$$

$$\phi_g |_{\partial \Omega} = 0$$

Then

$$\phi_g \in H^2(\Omega) \cap H_0^1(\Omega) \quad \text{and} \quad \| \phi_g \|_{2, \Omega} \leq C \| g \|_{0, \Omega} \quad (4.19)$$

The variational formulation for the problem (4.18) is:

$$(\phi_g, v)_{1, \Omega} = (g, v)_{0, \Omega} \quad \forall v \in H_0^1(\Omega)$$

It is permissible to take $v = e_i$ so that $(\phi_g, e_i)_{1, \Omega} = (g, e_i)_{0, \Omega}$. But e_i is orthogonal to V^h ; hence, $(v_h, e_i)_{1, \Omega} = 0$, $\forall v_h \in V^h$.

It follows that $(e_i, g)_{0, \Omega} = (e_i, \phi_g - v_h)_{1, \Omega} \quad \forall v_h \in V^h$ and

$$|(g, e_i)_{0, \Omega}| \leq \|e_i\|_{1, \Omega} \|\phi_g - v_h\|_{1, \Omega} \quad \forall v_h \in V^h.$$

Choosing $v_h = \phi_g^h$ to be the interpolant in V^h of ϕ_g , we have

$$\|\phi_g - \phi_g^h\|_{1, \Omega} \leq C_h \|\phi\|_{2, \Omega} \leq Ch \|g\|_{0, \Omega}$$

Hence,

$$-\frac{|(g, e_i)|}{\|g\|_{0, \Omega}} \leq Ch \|e\|_{1, \Omega} \quad \text{and by (5.7),}$$

$$\|e\|_{0, \Omega} \leq Ch \|e\|_{1, \Omega}$$

This completes the proof of the theorem. \square

1.5 The LBB Condition For I8/P₁ Elements

This section is devoted to the proof of Theorem II and in particular to obtaining the kernel of the operator B_h^* and an estimate of the LBB constant α_h

$$V^h = \{v_h = (v_{h1}, v_{h2}) \mid v_{hi} \in C^0(\Omega) ;$$

$$v_{hi}|_{\Omega_e} \in Q'_2(\Omega_e), v_{hi} = 0 \text{ on}$$

$$\partial\Omega, 1 \leq e \leq E, i = 1, 2\}$$

$$Q'_2(\Omega_e) = Q_2(\Omega_e) - \{x^2, y^2, \lambda \in \mathbb{R}, (x, y) \in \bar{\Omega}_e\}$$

$$Q^h = \{q_h \mid q_h|_{\Omega_e} \in P_1(\bar{\Omega}_e), 1 \leq e \leq E\}$$

where Q'_2 is the subspace of Q_2 used to define the serendipity element.

We shall work with a master element $\hat{\Omega}$. Each element $q_h \in Q^h$ has

three degrees of freedom q_α , $\alpha = 0,1,2$, and each element and can be chosen such that

$$q|_{\Omega^e} = q_0 + q_1x + q_2y ; \nabla q = (q_1, q_2)$$

When referred to the element $\hat{\Omega}$, these degrees of freedom will be noted by \hat{q}_α , $\alpha = 0,1,2$:

$$\hat{q}|_{\hat{\Omega}} = \hat{q}_0 + \hat{q}_1\hat{x} + \hat{q}_2\hat{y} ; \nabla\hat{q} = (\hat{q}_1, \hat{q}_2)$$

Then

$$q_0 = \hat{q}_0, \hat{q}_1 = h\hat{q}_1, \text{ and } q_2 = h\hat{q}_2.$$

Lemma 5.1. Under the conditions of Theorem II, $\dim \ker B_h^* = 3$.

Moreover, $\ker B_h^* = \text{span} \{1, \chi_1, \chi_2\}$, where χ_1, χ_2 are discontinuous functions of the type shown in Fig. 3.

Proof. It suffices to confine our attention to the collection of four reference elements $\hat{\Omega}_i$, $i = 1,2,3,4$, shown in Fig. 4a.

We wish to characterize all $q_h \in \ker B_h^*$; i.e., $(q_h, \text{div } \underline{v}_h) = 0$ $\forall \underline{v}_h \in V^h$. We begin by choosing \underline{v}_h in V^h such that $\underline{v}_h = \underline{0}$ at all nodes except a^{14} where $v_{h1}(a^{14}) = 1$. Then,

$$\text{in } \hat{\Omega}_1 : v_{h1} = -4x(1+x)(1-y), \quad q_h = \hat{q}_0^1 + \hat{q}_1^1(x + \frac{1}{2}) + \hat{q}_2^1(y - \frac{1}{2})$$

$$\text{in } \hat{\Omega}_4 : v_{h4} = -4x(1+x)(1+y), \quad q_h = \hat{q}_0^4 + \hat{q}_1^4(x + \frac{1}{2}) + \hat{q}_2^4(y + \frac{1}{2})$$

where \hat{q}_α^e , $\alpha = 0,1,2$, are the degrees of freedom (coefficients) of q_h in subelement $\hat{\Omega}_e$, $e = 1,2,3,4$. Then we find that

$$\int_{\hat{\Omega}_1 \cup \hat{\Omega}_4} q_h \text{div } \underline{v}_h \, dx = 0$$

and this implies that

$$\hat{q}_1^4 = -\hat{q}_1^1$$

Similarly, choosing $v_{h2}(a^{14}) = 1$ with $v_h = 0$ at other nodes, leads to the conclusion that $\hat{q}_0^1 = \hat{q}_0^4$. We continue this procedure at nodes a^{12} , a^{23} , and a^{34} to eliminate 7 of the 12 degrees-of-freedom of q_h . Collecting these results, we are left with the 5 coefficients indicated in Fig. 4 b.

We next choose $v_h = 0$ at all nodes except that $v_{h1}(0) = 1$. We find that

$$\int_{\hat{\Omega}_1 \cup \hat{\Omega}_2 \cup \hat{\Omega}_3 \cup \hat{\Omega}_4} q_h \operatorname{div} v_h \, dx dy = 0 \Rightarrow \hat{q}_2 = \hat{q}_2^1$$

A similar calculation with $v_{h2}(0) = 1$ yields $\hat{q}_1 = \hat{q}_1^1$.

Collecting all of the results, we are left with three independent coefficients, q_0, q_1, q_2 and these define the q_h -pattern indicated in Fig. 1. Reciprocally, a linear combination q_h of $1, \chi_1$ and χ_2 satisfies $(q_h, \operatorname{div} v_h) = 0$ for all v_h in V^h . A typical member χ_1 of $\ker B_h^*$ is indicated in Fig. 3; χ_2 is obtained by rotating the x -, y -axes 90-degrees. \square

Proof of Theorem III. We now return to the completion of proof of Theorem II. On each element Ω_e , we evaluate the product $q_h \operatorname{div} v_h$ using 16 degrees of freedom of v_h (eight for each component v_1 and v_2) and the 3 degrees of freedom of q_h (the coefficients \hat{q}_0, \hat{q}_1 , and \hat{q}_2 defined earlier). We get

$$\int_{\Omega_e} q_h \operatorname{div} \underline{v}_h \, dx = \frac{h}{2} \left\{ \begin{aligned} & \frac{1}{6} v_1(a^1) [-\hat{q}_0 + 2\hat{q}_1 + \hat{q}_2] + \frac{1}{6} v_2(a^1) [-\hat{q}_0 + \hat{q}_1 + 2\hat{q}_2] \\ & + \frac{1}{6} v_1(a^2) [+ \hat{q}_0 + 2\hat{q}_1 - \hat{q}_2] + \frac{1}{6} v_2(a^2) [-\hat{q}_0 - \hat{q}_1 + 2\hat{q}_2] \\ & + \frac{1}{6} v_1(a^3) [+ \hat{q}_0 + 2\hat{q}_1 + \hat{q}_2] + \frac{1}{6} v_1(a^3) [+ \hat{q}_0 + \hat{q}_1 + 2\hat{q}_2] \\ & + \frac{1}{6} v_1(a^4) [-\hat{q}_0 + 2\hat{q}_1 - \hat{q}_2] + \frac{1}{6} v_2(a^4) [+ \hat{q}_0 - \hat{q}_1 + 2\hat{q}_2] \\ & + \frac{2}{3} [-v_1(a^{12})\hat{q}_1 - v_2(a^{12})\hat{q}_0] \\ & + \frac{2}{3} [+v_1(a^{23})\hat{q}_0 - v_2(a^{23})\hat{q}_2] \\ & + \frac{2}{3} [-v_1(a^{34})\hat{q}_1 + v_2(a^{34})\hat{q}_0] \\ & + \frac{2}{3} [-v_1(a^{14})\hat{q}_0 - v_2(a^{14})\hat{q}_2] \end{aligned} \right\}$$

$$(q_h, \operatorname{div} \underline{v}_h) = \sum_{e=1}^E \int_{\Omega_e} q_h \operatorname{div} \underline{v}_h \, dx$$

This summation over the elements can be replaced by a summation over the nodes as shown in Fig. 5. Three types of nodes can be distinguished as shown in the Fig. 6. The indices for the pressure with respect to each node is also shown. Then, using the numbering scheme shown, we have

$$\begin{aligned} 2h^{-1}(q_h, \operatorname{div} \underline{v}_h) &= \frac{1}{6} \sum_{e_I} \left(v_1(e_I) [-\hat{q}_0^1 + 2\hat{q}_1^1 + \hat{q}_2^1 + \hat{q}_0^2 + \hat{q}_1^2 - \hat{q}_2^2 + \hat{q}_0^3 + 2\hat{q}_1^3 + \hat{q}_2^3 - \hat{q}_0^4 + 2\hat{q}_1^4 - \hat{q}_2^4] \right. \\ & \quad \left. + v_2(e_I) [-\hat{q}_0^1 + \hat{q}_1^1 + 2\hat{q}_2^1 - \hat{q}_0^2 - \hat{q}_1^2 + 2\hat{q}_2^2 + \hat{q}_0^3 + \hat{q}_1^3 + 2\hat{q}_2^3 + \hat{q}_0^4 - \hat{q}_1^4 + 2\hat{q}_2^4] \right) \\ & \quad + \frac{2}{3} \sum_{e_{II}} \left(v_1(e_{II}) [-\hat{q}_1^1 - \hat{q}_1^4] + v_2(e_{II}) [-\hat{q}_0^1 + \hat{q}_0^4] \right) \\ & \quad + \frac{2}{3} \sum_{e_{III}} \left(v_1(e_{III}) [-\hat{q}_0^1 + \hat{q}_0^2] + v_2(e_{III}) [-\hat{q}_2^1 - \hat{q}_2^2] \right) \end{aligned}$$

If we choose:

$$\begin{aligned}
 v_1(e_I) &= 6(-\hat{q}_0^1 + 2\hat{q}_1^1 + \hat{q}_2^1 + \hat{q}_0^2 + 2\hat{q}_1^2 - \hat{q}_2^2 + \hat{q}_0^3 + 2\hat{q}_1^3 + \hat{q}_2^3 - \hat{q}_0^4 + 2\hat{q}_1^4 - \hat{q}_2^4) \\
 v_2(e_I) &= 6(-\hat{q}_0^1 + \hat{q}_1^1 + 2\hat{q}_2^1 - \hat{q}_0^2 - \hat{q}_1^2 + 2\hat{q}_2^2 + \hat{q}_0^3 + \hat{q}_1^3 + 2\hat{q}_2^3 + \hat{q}_0^4 - \hat{q}_1^4 + 2\hat{q}_2^4) \\
 v_2(e_{II}) &= \frac{3}{2} (-\hat{q}_1^1 - \hat{q}_1^4) \\
 v_2(e_{II}) &= \frac{3}{2} (-\hat{q}_0^1 + \hat{q}_0^4) \\
 v_1(e_{III}) &= \frac{3}{2} (-\hat{q}_0^2 + \hat{q}_0^2) \\
 v_2(e_{III}) &= \frac{3}{2} (-\hat{q}_2^1 - \hat{q}_2^2)
 \end{aligned} \tag{5.1}$$

and $\underline{u} = 0$ at the nodes on the boundary,

then

$$2h^{-1}(q_h, \operatorname{div} \underline{v}_h) = \sum_{n=1}^N \left(v_1(n)^2 + v_2(n)^2 \right) \geq c \|\underline{v}_h\|_1^2 \tag{5.2}$$

where the summation is over all N nodes.

Now it will be shown that the choice (5.1) implies

$$\|\underline{v}_h\|_1 \geq c \|q_h\|_{0/\ker B_h^*} \tag{5.3}$$

Then (5.2) and (5.3) complete the proof. \square

With the expression chosen in (5.1) we can reorder $\|\underline{v}_h\|_1^2$ and, using the numbering scheme of Fig. 5, obtain

$$\begin{aligned}
\|v_h\|_1^2 \geq c \sum_i \left\{ & (-\hat{q}_0^1 + 2\hat{q}_1^1 + \hat{q}_2^1 + \hat{q}_0^2 + 2\hat{q}_1^2 - \hat{q}_2^2 + \hat{q}_0^3 + 2\hat{q}_1^3 + \hat{q}_2^3 - \hat{q}_0^4 + 2\hat{q}_1^4 - \hat{q}_2^4)^2 \right. \\
& + (-\hat{q}_0^1 + \hat{q}_1^1 + 2\hat{q}_2^1 - \hat{q}_0^2 - 2\hat{q}_1^2 + 2\hat{q}_2^2 + \hat{q}_0^3 + \hat{q}_1^3 + 2\hat{q}_2^3 + \hat{q}_0^4 - \hat{q}_1^4 + 2\hat{q}_2^4)^2 \\
& + (\hat{q}_1^1 + \hat{q}_1^4)^2 + (\hat{q}_1^2 + \hat{q}_1^3)^2 + (\hat{q}_2^1 + \hat{q}_2^2)^2 + (\hat{q}_2^3 + \hat{q}_2^4)^2 \\
& \left. + (\hat{q}_0^1 - \hat{q}_0^2)^2 + (\hat{q}_0^2 - \hat{q}_0^3)^2 + (\hat{q}_0^3 - \hat{q}_0^4)^2 + (\hat{q}_0^4 - \hat{q}_0^1)^2 \right\} \quad (5.4)
\end{aligned}$$

Here we find a quadratic form, whose kernel is precisely the kernel of B_h^* defined in the Lemma 5.1.

Then, it can be written:

$$\begin{aligned}
\|v_h\|_1^2 \geq c \sum_i \left\{ & (\hat{q}_0^1 - \hat{q}_0^2)^2 + (\hat{q}_0^2 - \hat{q}_0^3)^2 + (\hat{q}_0^3 - \hat{q}_0^4)^2 + (\hat{q}_0^4 - \hat{q}_0^1)^2 \right. \\
& + (\hat{q}_1^1 - \hat{q}_1^2)^2 + (\hat{q}_1^2 + \hat{q}_1^3)^2 + (\hat{q}_1^3 - \hat{q}_1^4)^2 + (\hat{q}_1^4 + \hat{q}_1^1)^2 \\
& \left. + (\hat{q}_2^1 + \hat{q}_2^2)^2 + (\hat{q}_2^2 - \hat{q}_2^3)^2 + (\hat{q}_2^3 + \hat{q}_2^4)^2 + (\hat{q}_2^4 - \hat{q}_2^1)^2 \right\} \quad (5.5)
\end{aligned}$$

The passage from (5.4) to (5.6) comes from the fact that both quadratic forms in bracket in these expressions provide the same kernel and therefore define two equivalent semi-norms on

$$v_e = \{q_j^i ; i = 1, 4 ; j = 0, 2\}$$

Now if we pay our attention to the quadratic form on the first line of (5.5), it can be interpreted as the L^2 -norm of the gradient of a piecewise bilinear function ϕ_0 , defined by q_0^i at the corresponding centroid of each element. This function ϕ_0 belongs to H^1 and, as proved in TEMAM [45], there exists a constant C such that

$$\|\nabla\phi_0\|^2 \geq c \|\phi_0\|_{L^2/\mathbb{R}}^2 \quad (5.6)$$

This procedure can also be applied in the construction of functions ϕ_1 and ϕ_2 from the q_1^i 's and q_2^i 's. Finally we obtain the inequality (5.3) from (5.5) and (5.6) noticing that the summation of the three squared L^2/\mathbb{R} norms of ϕ_j ($j = 0, 1, 2$) precisely corresponds to $\|q_h\|_{0/\ker B_h^*}^2$.

1.6 The LBB Conditions For Q_1/P_0 Elements

The general proof of Theorem III can also be used in the analysis of the Q_1/P_0 element (bilinear velocities, constant pressures). For this element we maintain that the kernel of B_h^* consists of checkerboard nodes which are characterized by alternating values a and b in each neighbor element. In this case

$$\dim \ker B_h^* = 2$$

Using the same notation as in the previous section, we can define for each element and each node their integer component J and K ; the element $e = 1$ in the corner (resp. $e = 2$) corresponds to $J = K = 1$ (resp. $J = 2, K = 1$). (See Fig. 7.) Using the elements e satisfying $J + K = \text{even}$ and constructing a piecewise bilinear continuous function defined by q_e at the centroid of these elements, we can apply the inequality (5.6) and obtain

$$\sum_{\substack{i=1 \\ J+K \text{ even}}}^N (q_i^{\circ 3} - \bar{q}_i^{\circ 3})^2 + \sum_{\substack{i=1 \\ J+K \text{ odd}}}^N (q_i^{\circ 2} - \bar{q}_i^{\circ 4})^2 \geq Ch^2 \sum_{\substack{e=1 \\ J+K \text{ even}}}^E q_e^{\circ 2}$$

We obtain a similar inequality considering the element e such that $J + K$ is odd, and by addition, we arrive at

$$\sum_{i=1}^N (q_i^{\circ 1} - q_i^{\circ 3})^2 + (q_i^{\circ 2} - q_i^{\circ 4})^2 \geq Ch^2 \sum_{e=1}^E q_e^{\circ 2} \quad (6.1)$$

The improvement h^2 instead of h^4 in the estimate of Oden, Kikuchi, and Song [41] allows us to obtain an LBB constant $\alpha_h = O(h)$ for this element.

1.7 Summary of Some Stability Results

A mathematical analysis of the discrete Babuska-Brezzi condition (3.7) has been made by Oden and Kikuchi [35], Oden, Kikuchi and Song [40], and Oden and Jacquotte [37,38] for several finite elements for a model two-dimensional Stokes' problem on a uniform mesh. We shall summarize these results here which pertain to the behavior of the "LBB-constant" α_h and the stability of the pressure calculations. We use the notations

P_k = space of complete piecewise polynomials of degree k
over an element

Q_k = space of tensor products of complete polynomials of
degree k

I8 = the eight-node isoparametric element

Results are summarized in Table 1. In this table, examples 1, 2, and 7 "lock" at small values of the penalty parameter ϵ . This means that for a given mesh size h , ϵ cannot be taken arbitrarily small, as noted earlier. Of course, for an acceptable ϵ for reasonable mesh sizes, ϵ is so large that the constraint of incompressibility is not adequately satisfied. Hence these elements should generally be avoided. Elements 2, 4, 5, 8, 11, and 14 are unstable since $\alpha_h = O(h)$. Remarkably, these instabilities frequently are not observed on uniform meshes when the solution is very smooth. Mild

irregularities in the solution or small perturbations in the mesh may, however, produce violent oscillations in computed pressures the magnitudes of which increase without bound as h tends to zero. In many cases, however, these oscillations disappear upon "filtering" the pressure solutions (i.e. upon averaging the pressures over one or more elements). In the case of elements 2 and 14 it has been proven mathematically [20] that certain filtering schemes will produce a stable and convergent method. However, it is not known if filtering can be used to stabilize and salvage the remaining unstable elements.

Elements 6 and 10 lead to stable and convergent schemes and are quite robust in the sense that they are insensitive to singularities in the solution. However, they are not too accurate and converge at a suboptimal rate.

Element 9 is clearly the superior of any listed: it is unconditionally stable, it provides both velocity and pressure approximations which converge at the optimal rate, and

$$\ker \nabla_h = \ker \nabla$$

Element 13 is somewhat of a novelty. While element 5 yields unstable pressure approximations, Oden and Jacquotte [37] have shown that a composite of three Q_2/P_1 elements (no. 9) and one $I8/P_1$ element (no. 5) is stable.

The behavior of elements 11 and 12, marked with an asterisk, is only conjectured here and has not been rigorously proven.

Extensions of these results to three-dimensional elements are straightforward.

1.8 Numerical Examples

The results of several numerical experiments are described which are designed to verify the theoretical results with regard to the Q_2/P_1 , $I8/P_1$,

and the composite elements described earlier. We also investigate numerically the effects of a pressure filtering operation.

As a first example, we consider an L-shaped domain Ω partitioned into 64 square subdomains, as shown in Fig. 8. The fluid is subjected to a constant body force $\underline{f} = (0, -100)$. We take $\nu = 333$, and the penalty parameter $\epsilon = 10^{-5}$. We will be interested in the computed hydrostatic pressure across the section AA' defined by: $y = 0.80$. Each subdomain corresponds to a finite element; the velocity on each element is interpolated at 8 or 9 nodes and the pressure by its value at 3 points. Thus, various choices of how to handle the ninth node lead to meshes with $I8/P_1$, Q_2/P_1 , or Composite/ P_1 elements. We will be interested in three cases involving these elements:

- Mesh 1: All the elements are Q_2/P_1 elements
- Mesh 2: All the elements are $I8/P_1$ elements
- Mesh 3: Adding 16 centroid nodes, we obtain 16 composite elements as shown in Fig. 9.

The results reported here were obtained using the FIDAP code for problems of incompressible viscous flow [14].

Figures 10 and 11 show the comparison between the results obtained with the Q_2/P_1 element (Mesh 1) and those obtained with meshes 2 and 3. Fig. 10 illustrates the major difference between the Q_2/P_1 and the $I8/P_1$ element; the former involves a pressure which seems to be smoothly distributed along the section AA' while the latter yields a pressure with severe oscillations. We note, however, that the values of the pressure obtained at the centroid of each element are close to the values obtained with the Q_2/P_1 element, which suggest that this unstable solution can be stabilized by a filtering operation which effectively uses these averaged values of pressure.

It is also remarked that the oscillations seem to come from the spurious modes in $\ker B_h^*$. The smoothing device may be equivalent to an a posteriori elimination of these spurious modes. However, it turns out that these spurious modes do not solely come from $\ker B_h^*$: Figure 11 contains results obtained by adding one node in the elements in the corner (point C_1 in Fig. 8). For this mesh, $\ker B_h^* = \mathbb{R}$ but the results still exhibit pressure oscillations. However, for this mesh, the solution seems to be much smoother than in the 18-case.

Finally, the composite elements lead to a quite smooth solution as indicated in Fig. 12, which is close to the solution obtained with 9-node elements, except that for this element $h^2 = 0.25$, while for the Q_2/P_1 element h^2 was equal to 0.0625.

We also note that when the body force \underline{f} derives from a potential: $\underline{f} = -\nabla v$, then the unique solution for the Stokes Problem is

$$\begin{cases} \underline{u} = \underline{0} \\ p = -v \end{cases}$$

In this example $\underline{f} = (0, -100)$ and $v = 100y = -p$. The numerical results obtained by these different methods are summarized in Table II. We observe that the error in the filtered pressures is around 33 percent greater than that of the Q_2/P_1 elements for this particular example.

As a second example, we consider a Dirichlet Stokes Problem, which is designed for numerical verification of the convergence theory for three schemes considered:

- a) A uniform mesh of Q_2/P_1 elements.

- b) A uniform mesh of $18/P_1$ elements.
 c) A uniform mesh of $18/P_1$ elements with pressure averaging.

We consider the unit square domain partitioned into square subdomain, and the following body forces $\underline{f} = (f_1, f_2)$ are applied:

$$\begin{aligned} f_1 = & -4y + 12x^2 + 24xy + 12y^2 - 24x^3 - 48x^2y - 72xy^2 \\ & -8y^3 + 12x^4 + 48x^3y + 72x^2y^2 + 48xy^3 - 24x^4y - 48x^2y^3 \\ & -2(x - x_0) + \alpha(x) \end{aligned}$$

$$\begin{aligned} f_2 = & 4x - 12x^2 - 24xy - 12y^2 + 8x^3 + 72x^2y + 48xy^2 + 24y^3 \\ & -12y^4 - 48xy^3 - 72x^2y^2 - 48x^3y + 24xy^4 + 48x^3y^2 \end{aligned}$$

where $\alpha(x) = -1$ if $0 \leq x \leq x_0$, $\alpha(x) = 1$ if $x_0 < x \leq 1$. Then

(\underline{u}, p) is defined by

$$\underline{u} = (u_1, u_2) ; \begin{cases} u_1 = x^2(1-x)^2(2y - 6y^2 + 4y^3) \\ u_2 = (-2x + 6x^2 - 4x^3)y^2(1-y)^2 \end{cases}$$

and

$$\begin{cases} p = x_0 - x - (x - x_0)^2 & \text{if } 0 \leq x \leq x_0 \\ p = x - x_0 - (x - x_0)^2 & \text{if } x_0 < x \leq 1 \end{cases}$$

(\underline{u}, p) satisfies:

$$\begin{cases} \underline{u}|_{\Gamma} = 0 \\ \operatorname{div} \underline{u} = 0 & \text{in } \Omega \\ -\Delta \underline{u} + p = \underline{f} & \text{in } \Omega \end{cases}$$

As before, we construct a plot of the computed pressures across a section of the domain. Figure 13 shows the results obtained by partitioning the domain Ω in 64 square subdomains. For this mesh, h is equal to $1/8$. The computations are made with Q_2 -on I8-elements. Whereas the Q_2 -solution seems to be stable, clearly the I8-solution shows oscillations around the exact solution. However, it is noted that both solutions coincide at the centroid of the elements and this again suggests that the "smoothed I8-solution," obtained using only the pressure at the centroid, is stable, and may converge at a rate of $O(h^2)$.

Finally, Fig. 14 confirms this suspicion showing the computed rate of convergence is precisely $O(h^2)$ for the pressure for the Q_2 -element, and for the smoothed I8-element. However, it is also observed once again that the Q_2/P_1 -pressures are considerably more accurate than the filtered I8/ P_1 -pressures for all mesh sizes considered.

With the results from these examples we can conclude that

- The Q_2/P_1 element is stable and the optimal L^2 -rate of convergence of the pressures of $O(h^2)$ is attained.
- The I8/ P_1 element yields unstable pressure approximations, but these can apparently be stabilized considering only the values at the centroids.
- Spurious oscillations (checkerboarding) can also appear when $\ker B_h^* = \mathbb{R}$
- Filtering the pressures in the I8/ P_1 -element by using only the centroidal value leads to a pressure approximation which may converge in L^2 at a rate of $O(h^2)$; however, the accuracy of the filtered scheme is quite inferior to that of the Q_2/P_1 -elements.

These computed results underline once again the critical role played by the LBB-condition in studying the stability of finite element schemes by reduced integration. These and other results we have computed also indicate that the estimates obtained in Section 4 for the discrete LBB-constant α_h are sharp. Indeed, the theoretical result that the use of a composite element of the type employed here leads to a stable pressure field, while not of great practical value, is fully confirmed by the numerical results. This suggests again that these calculated estimates of α_h are a good indication of the actual numerical performances of these methods.

PART II: ANALYSIS OF INSTABILITIES IN
UNDERINTEGRATED FINITE ELEMENT METHODS

2.1 Introduction

For many years, a special type of numerical instability has been observed in finite difference approximations of flow fields, which has been referred to as "hourglassing", "keystoning", or "chickenwiring". These graphic terms refer to geometrical patterns which appear in computed flow fields (e.g. velocities) and which emerge as spurious oscillations superimposed on an otherwise smooth field, the spurious oscillations often taking a zig-zag form which resembles an hourglass or a chickenwire mesh. These spurious modes can be amplified upon refining the mesh, and to control such numerical instabilities, various schemes for incorporating "hourglass viscosity" or "hourglass damping" have been proposed by some authors.

It is now known that hourglass modes can arise from an incomplete (or poor) approximation of the kernel of the operators in the momentum equations in flow or solid mechanics problems (or, more generally, of the principal part of the operator in the governing differential equation of a given boundary-value problem). For example, in addition to the rigid body motions residing in the kernel of the standard operators appearing in the equilibrium (momentum) equations of solid and fluid mechanics, one finds hourglass modes in various crude discrete models of these operators.

In recent years, the occurrence of hourglass instabilities in

underintegrated finite element approximations has been observed. In the implementation of most finite element methods, integrals defining stiffness matrices are evaluated using numerical quadrature schemes. To improve computational efficiency, the practice of *underintegration* is often employed, by which is meant the use of a quadrature rule of an order lower than that required to integrate polynomial integrands exactly. This can produce rank-deficient stiffness matrices or, equivalently, an expanded kernel of the equilibrium operation which contains spurious hourglass modes, and the result is again a numerically unstable scheme.

In order to overcome this difficulty, artificial stiffness or viscosity methods, or other stabilization methods have been proposed by several authors (e.g. [2-5, 17, 28]). These methods involve computing an underintegrated matrix, and then adding a stabilization matrix which effectively eliminates the hourglass modes. They turn out to be fairly general and have been used for a long time in numerous codes. Whereas all of these methods based on intuitive feeling give good numerical results, their mathematical study remains often non-existent.

The most interesting challenge is to solve the problem using only the crude rank-deficient underintegrated stiffness matrix, the solution is obtained up to within an arbitrary spurious mode, and then to eliminate these modes from the solution in a post-processing operation.

Unfortunately, even when the stiffness matrix is rank-sufficient, similar oscillations are observed when underintegration is used. In that case, the process of the excitations of modes similar to the hourglass modes is not completely understood and these modes have never been

mathematically studied.

In this report, we would like to give precise mathematical justifications and answers to the questions previously mentioned. The next section (Section 2.2) is devoted to the proof that the Stabilization Method is mathematically justified. Then, in Section 2.3 we present a method which involves solving an underintegrated and not well-posed problem, then in a-posteriori eliminating the unknown degree of freedom. The proof of the accuracy of the method is given in Section 2.4, and its numerical aspects and results are described in Section 2.5. In Section 2.6, we examine the case where the spurious oscillations cannot be predicted from the rank-deficiency of the stiffness matrix and we analyze why these modes may be excited. Finally, we apply the previous considerations to the Linear Elasticity Problem.

It should be noted that the method and its results cannot be embedded in a classical elliptic theory: Strang's ellipticity condition [44] is here violated and this non-elliptic method cannot be studied by the classical theory of finite element methods and numerical integration [10, 11]. We also refer to Girault [20, 21] for his approach to the same kind of problem, non-elliptic because of the use of partially underintegrated stiffness matrix, but where hourglass modes did not appear.

2.2 A-Priori Hourglass Control

2.2.1 Introduction. This section is devoted to giving a mathematical support to several methods consisting of adding a stabilization matrix to the underintegrated matrix. For clarity, we shall confine our

attention to a simple model problem. Let Ω be a regular domain in \mathbb{R}^2 with boundary $\partial\Omega$ and consider the model Neumann problem,

(P₀) Find $u = u(x,y)$ such that

$$\left. \begin{aligned} -\Delta u &= f && \text{on } \Omega \\ \frac{\partial u}{\partial n} &= 0 && \text{on } \partial\Omega \end{aligned} \right\} \quad (2.1)$$

where f is an $L^2(\Omega)$ -function satisfying

$$\int_{\Omega} f \, dx dy = 0 \quad (2.2)$$

The questions of the existence and uniqueness of solutions to (2.1) (which are well-known) are taken up in Part II.

We shall first consider a finite element approximation of (2.1) constructed using Q_1 -elements, i.e., four-node quadrilateral elements over which bilinear shape functions are used. Most of our notations and results are reproductions of those of Flanagan [17] and Belytschko [2, 4, 5].— Then we will attempt to extend our results to the Q_2 -elements (nine-node, biquadratic elements) and will indicate in which ways they differ from Belytschko's [3].

The construction of finite element approximations of (2.1) involves the calculation of the stiffness matrix K_e for a typical finite element Ω_e , which is given by the formula,

$$K_e = \int_{\Omega} \tilde{N}^t \cdot \tilde{N} \, dx dy \quad (2.3)$$

where \tilde{N} is a vector representing the bilinear or biquadratic shape functions in each element Ω_e , $1 \leq e \leq E$.

When Q_1 - (respectively Q_2 -) elements are used to discretize the domain Ω , K_e is a 4×4 matrix (resp. 9×9) and the N 's contain four bilinear (resp. nine biquadratic) shape functions. We will distinguish exact-, full-, and under-integrations. The full integration is obtained using the number of Gauss integration points necessary to obtain the exact integration on regular square elements: 4 (resp. 9) points in our study. The underintegration will involve the Gauss rule of lower order: 1 (resp. 4) points. The stiffness matrix associated with a rule involving k points will be denoted $K_e^{(k)}$, $k = 1, 4, 9$.

Several authors [2, 4, 5, 17, 28] proposed to add to the underintegrated stiffness matrix a stabilization matrix which exhibits several special properties. In this section, we will prove that these properties are indeed satisfied and that the exact stiffness matrix K_e can be computed by this method. This will be accomplished by first carrying out the integration (2.3) exactly.

We first introduce some notations. Suppose that element Ω_e is defined by the coordinates of its nodes (x^I, y^I) , $1 \leq I \leq p$, $p = 4$ or 9 . We introduce the isoparametric mapping from a master element

$$\hat{\Omega} = \left[-\frac{1}{2}, +\frac{1}{2} \right] \times \left[-\frac{1}{2}, +\frac{1}{2} \right]$$

to Ω_e such that

$$\left. \begin{aligned} x &= \sum_{I=1}^P x^I N_I(\xi, \eta) \\ y &= \sum_{I=1}^P y^I N_I(\xi, \eta) \end{aligned} \right\} \quad (2.4)$$

where N_I , $1 \leq I \leq p$, are the shape functions for the quadrilateral element on the master element. The node numbering convention is shown in Figure 15.

The stiffness matrix \tilde{K}_e is evaluated in (2.3) using the mapping (2.4) from $\hat{\Omega}$ to Ω_e :

$$\left. \begin{aligned} \tilde{K}_e &= \int_{\Omega} \tilde{\nabla N}^T \tilde{\nabla N} \, dx dy \\ &= \int_{\hat{\Omega}} |J| \hat{\nabla N}^T \left[\frac{d\xi}{dx} \right]^T \left[\frac{d\xi}{dx} \right] \hat{\nabla N} \, d\xi d\eta \end{aligned} \right\} \quad (2.5)$$

where $\left[\frac{d\xi}{dx} \right]$ is the Jacobian matrix of the mapping from $\hat{\Omega}$ to Ω_e , J is the inverse of its determinant, and where the gradients of the shape functions are derived with respect to the master element coordinates (ξ, η) . These matrices can be computed using (2.4):

$$\left. \begin{aligned} x &= \tilde{x}^T \cdot \tilde{N} \\ y &= \tilde{y}^T \cdot \tilde{N} \end{aligned} \right\} \quad (2.6)$$

$$\left[\frac{dx}{d\xi} \right] = \begin{bmatrix} \tilde{x}^T \frac{dN}{d\xi} & \tilde{y}^T \frac{dN}{d\xi} \\ \tilde{x}^T \frac{dN}{d\eta} & \tilde{y}^T \frac{dN}{d\eta} \end{bmatrix} \quad (2.7)$$

$$\left[\frac{d\xi}{dx} \right] = \left[\frac{dx}{d\xi} \right]^{-1} = \frac{1}{J} \begin{bmatrix} \tilde{y}^T \frac{dN}{d\eta} & -\tilde{y}^T \frac{dN}{d\xi} \\ -\tilde{x}^T \frac{dN}{d\eta} & \tilde{x}^T \frac{dN}{d\xi} \end{bmatrix} \quad (2.8)$$

$$\hat{V}N = \begin{bmatrix} \frac{dN^T}{d\xi} \\ \frac{dN^T}{d\eta} \end{bmatrix} \quad (2.9)$$

where J is the Jacobian of the mapping

$$J = \det \frac{dx}{d\xi} \quad (2.10)$$

Finally, we obtain the expression,

$$\underline{K}_e = \int_{\hat{\Omega}} \left(\frac{\underline{A}_{xx}^T \underline{A}^T}{\underline{y}^T \underline{A}x} + \frac{\underline{A}_{yy}^T \underline{A}^T}{\underline{y}^T \underline{A}x} \right) d\xi d\eta \quad (2.11)$$

where \underline{A} is the antisymmetric matrix

$$\underline{A} = \frac{dN}{d\eta} \frac{dN^T}{d\xi} - \frac{dN}{d\xi} \frac{dN^T}{d\eta} \quad (2.12)$$

the Jacobian J can be expressed as $\underline{y}^T \underline{A}x$.

A study of \underline{K}_e expressed as in (2.11) and the properties of the matrix \underline{A} will then enable us to study the effect of the underintegration of the stiffness matrix. We will first concentrate on the 4 node element and derive the exact expression of the stabilization matrix. Then we will discuss what form this matrix may take for the 9-node element.

2.2.2. The stabilization matrix for the bilinear element. For the bilinear element, the shape function vector can be written as

$$\underline{N} = \frac{1}{4} \underline{t} - \frac{\xi}{2} \underline{s} + \frac{\eta}{2} \underline{s}' + \xi\eta \underline{h} \quad (2.13)$$

where

$$\begin{aligned}
 \underline{\underline{s}}^T &= (1, -1, -1, 1) \\
 \underline{\underline{s}}'^T &= (1, 1, -1, -1) \\
 \underline{\underline{t}}^T &= (1, 1, 1, 1) \\
 \underline{\underline{h}}^T &= (1, -1, 1, -1)
 \end{aligned} \tag{2.14}$$

then the explicit form of the (4x4) matrix A is

$$\underline{\underline{A}} = \frac{1}{4}(\underline{\underline{s}}'\underline{\underline{s}}^T - \underline{\underline{s}}\underline{\underline{s}}'^T) + \frac{\xi}{2}(\underline{\underline{s}}\underline{\underline{h}}^T - \underline{\underline{h}}\underline{\underline{s}}^T) + \frac{\eta}{2}(\underline{\underline{h}}\underline{\underline{s}}'^T - \underline{\underline{s}}'\underline{\underline{h}}^T) \tag{2.15}$$

for $(\xi, \eta) = (0, 0)$, we obtain $\underline{\underline{A}}_0$ which satisfies

$$\underline{\underline{y}}^T \underline{\underline{A}}_0 \underline{\underline{x}} = \underline{\underline{y}}^T \underline{\underline{A}} \Big|_{\xi=\eta=0} \underline{\underline{x}} = \frac{1}{2}(y_{24}x_{13} + y_{31}x_{24}) \tag{2.16}$$

which is merely the area of the element Ω_e , noted $|\Omega_e|$.

At this point, we can precisely see the matrix resulting from a 1-point rule; this underintegrated matrix denoted by $\underline{\underline{K}}_e^{(1)}$ is given by

$$\underline{\underline{K}}_e^{(1)} = \frac{\underline{\underline{A}}_0 \underline{\underline{x}} \underline{\underline{x}}^T \underline{\underline{A}}_0^T}{|\Omega_e|} + \frac{\underline{\underline{A}}_0 \underline{\underline{y}} \underline{\underline{y}}^T \underline{\underline{A}}_0^T}{|\Omega_e|} \tag{2.17}$$

Also, if we note $\underline{\underline{B}} = (\underline{\underline{b}}_1, \underline{\underline{b}}_2)^T$ the discrete approximations of the gradient ∇N evaluated at the integration point we can remark that

$$\begin{cases} \underline{\underline{b}}_1 = -\underline{\underline{A}}_0 \underline{\underline{y}} \\ \underline{\underline{b}}_2 = \underline{\underline{A}}_0 \underline{\underline{x}} \end{cases} \tag{2.18}$$

and therefore (2.17) takes the usual form

$$\underline{\underline{K}}_e^{(1)} = \frac{1}{|\Omega_e|} (\underline{\underline{b}}_1 \underline{\underline{b}}_1^T + \underline{\underline{b}}_2 \underline{\underline{b}}_2^T) \tag{2.19}$$

the rank deficiency of $\underline{K}_e^{(1)}$ can now be verified.

Indeed, from (2.14) and (2.15) we can simply note that

$$\left. \begin{aligned} \underline{A}_{0\sim} h &= 0 \\ \underline{A}_{0\sim} t &= 0 \end{aligned} \right\} \quad (2.20)$$

and then

$$\left. \begin{aligned} \underline{K}_e^1 h &= 0 \\ \underline{K}_e^1 t &= 0 \end{aligned} \right\} \quad (2.21)$$

Therefore, if we consider H and T the global hourglass and translation, and $\underline{K}^{(1)}$ the assembled underintegrated stiffness matrix, we have

$$\left. \begin{aligned} \underline{K}^{(1)} \cdot \underline{H} &= \sum_e \underline{K}_e^{(1)} \cdot \underline{H} = \sum_e \underline{K}_e^{(1)} \cdot \underline{h} = 0 \\ \text{also } \underline{K}^{(1)} \cdot \underline{T} &= 0 \end{aligned} \right\} \quad (2.22)$$

and this proves the rank deficiency of $\underline{K}^{(1)}$.

Note that this "+1" pattern is independent of the regularity of the mesh and that H will take alternating values +1 and -1 at neighbor nodes as shown in Fig. 16.

Our goal will now be to calculate a matrix $\underline{K}_e^{\text{stab}}$ such that, if added to $\underline{K}_e^{(1)}$, we obtain the exact stiffness matrix \underline{K}_e given by (2.11). This expression does not seem easy to integrate, but the image of certain vectors mapped by this matrix can be easily computed using orthogonality relations previously obtained (2.20) and the fact that

$$\underline{b}_1^T \underline{x} = \underline{b}_2^T \underline{y} = |\Omega_e|$$

$$\underline{b}_1^T \underline{y} = \underline{b}_2^T \underline{x} = 0 \quad (2.23)$$

we can obtain:

$$\left. \begin{aligned} \underline{K}_e \underline{t} &= \underline{0} \\ \underline{K}_e \underline{x} &= \underline{b}_1 \\ \underline{K}_e \underline{y} &= \underline{b}_2 \end{aligned} \right\} \quad (2.24)$$

Equation (2.24) is not sufficient to compute \underline{K}_e , because it gives only 9 out of the 10 coefficients of \underline{K}_e (4 x 4, symmetric). It is enough to know $\underline{X}^T \underline{K}_e \underline{X}$, where \underline{t} , \underline{x} , \underline{y} , and \underline{X} form a set of independent vectors. That is the case for $\underline{X} = \underline{h}$ because

$$\det(\underline{x}, \underline{y}, \underline{t}, \underline{h}) = 4A \neq 0$$

provided the element is not singular. Then the knowledge of $\underline{h}^T \underline{K}_e \underline{h}$ and the relations define uniquely \underline{K}_e . If we set

$$\underline{h}^T \underline{K}_e \underline{h} = 16 \bar{\epsilon} \quad (2.25)$$

then \underline{K}_e is given by

$$\underline{K}_e = \underline{K}_e^{(1)} + \bar{\epsilon} \underline{\gamma} \underline{\gamma}^T \quad (2.26)$$

whereas, again, given by (2.25), $\bar{\epsilon}$ is a scalar, and

$$\underline{\gamma} = \underline{h} - \frac{\underline{h}^T \underline{x}}{|\underline{\Omega}_e|} \underline{b}_1 - \frac{\underline{h}^T \underline{y}}{|\underline{\Omega}_e|} \underline{b}_2 \quad (2.27)$$

While it is difficult to express $\bar{\epsilon}$ nicely in function of \underline{x} and \underline{y} , its exact value can be written

$$\bar{\epsilon} = \frac{i}{4} \int_{\hat{\Omega}} \frac{\left[\xi(\tilde{s}'_x) - \eta(\tilde{s}'_x) \right]^2 + \left[\xi(\tilde{s}'_y) - \eta(\tilde{s}'_y) \right]^2}{|\Omega_e| + \xi(y_{43}x_{12} + y_{12}x_{34}) + \eta(y_{32}x_{14} + y_{14}x_{23})} d\xi d\eta \quad (2.28)$$

We observe that for parallelogram elements, the denominator is constant and its value is the area of the domain $|\Omega_e|$. In this case

$$\bar{\epsilon} = \frac{1}{24|\Omega_e|} (x_{13}^2 + x_{24}^2 + y_{13}^2 + y_{24}^2) \quad (2.29)$$

or for rectangular elements

$$\bar{\epsilon} = \frac{\frac{\ell_x^2}{x} + \frac{\ell_y^2}{y}}{12 \frac{\ell_x \ell_y}{x y}} \quad (2.30)$$

where ℓ_x and ℓ_y are the lengths of the sides of the rectangular element. Also note that for such parallelogram elements \tilde{y} reduces to \tilde{h} .

The expression (2.26) is often used to a-priori eliminate spurious modes for the kernel of K but the determination of $\bar{\epsilon}$ remains a problem. The choice $\bar{\epsilon} = 0$ leads to the underintegrated matrix and to the method to be studied in the next section. On the other hand, a cheaper way than the full integration of the whole matrix would be to fully integrate $\bar{\epsilon}$ given by (2.28). This method would lead again to the full integration and is cheaper because it needs only one 4×4 integration by element instead of 10. A more common use is to take for $\bar{\epsilon}$ a simple value independent of the geometry of the element, which is often the value obtained for a square $1/6$ or sometimes any arbitrary constant, as used in [4, 5].

2.2.3. The stabilization matrix for the biquadratic element. In this section, we will study the effect of the underintegration of the 9×9 stiffness matrix obtained in (2.11) with nine-node elements when a 4- Gauss integration point rule is used. Whereas Belytschko and al. [3] intuitively obtain another " $\underline{\gamma} \cdot \underline{\gamma}^T$ " stabilization, we prove that this decomposition is not even valid for regular meshes. We then propose a decomposition derived on regular mesh.

But first we exhibit the spurious modes out of $K_e^{(4)}$. For the biquadratic element, the shape function vector can be written as

$$\underline{N} = \underline{S} \underline{\xi} \quad (2.31)$$

where \underline{S} is the 9×9 matrix

$$\underline{S} = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 & -2 & -2 & 4 \\ 0 & 0 & 0 & -1 & 0 & 0 & 2 & -2 & 4 \\ 0 & 0 & 0 & 1 & 0 & 0 & 2 & 2 & 4 \\ 0 & 0 & 0 & -1 & 0 & 0 & -2 & 2 & 4 \\ 0 & 0 & -1 & 0 & 0 & 2 & 0 & 4 & -8 \\ 0 & 1 & 0 & 0 & 2 & 0 & -4 & 0 & -8 \\ 0 & 0 & 1 & 0 & 0 & 2 & 0 & -4 & -8 \\ 0 & -1 & 0 & 0 & 2 & 0 & 4 & 0 & -8 \\ 1 & 0 & 0 & 0 & -4 & -4 & 0 & 0 & 16 \end{bmatrix} \quad (2.32)$$

and

$$\underline{\xi} = [1, \xi, \eta, \xi\eta, \xi^2, \eta^2, \xi\eta^2, \xi^2\eta, \xi^2\eta^2]^T \quad (2.33)$$

The integration rule we are interested in involves four integration points $(\xi^\alpha, \eta^\alpha)$, $\alpha = 1, 4$. Associated to each of them, we note \underline{A}_α and \underline{J}_α the corresponding matrix \underline{A} and Jacobian. The underintegrated matrix is then

$$\underline{K}_e^{(4)} = \sum_{\alpha=1}^4 \frac{A_{\alpha xx}^T A_{\alpha} + A_{\alpha yy}^T A_{\alpha}}{J_{\alpha}} \quad (2.34)$$

and can also be written

$$\underline{K}_e^{(4)} = \sum_{\alpha=1}^4 \frac{1}{J_{\alpha}} (b_{-1}^{\alpha} b_{-1}^{\alpha T} + b_{-2}^{\alpha} b_{-2}^{\alpha T}) \quad (2.35)$$

where

$$\underline{B}^{\alpha} = \begin{bmatrix} b_{-1}^{\alpha T} \\ b_{-2}^{\alpha T} \end{bmatrix} = \begin{bmatrix} (-A_{\alpha} y)^T \\ (A_{\alpha} x)^T \end{bmatrix} \quad (2.36)$$

generalizes (2.18) to a 4- point rule.

The rank deficiency of $\underline{K}_e^{(4)}$ can now be verified. Indeed if we call \underline{t} and \underline{h} the vectors defined by

$$\begin{aligned} \underline{t}^T &= [1, 1, 1, 1, 1, 1, 1, 1, 1] \\ \underline{h}^T &= [1, 1, 1, 1, -1, -1, -1, -1, 0] \end{aligned} \quad (2.37)$$

we easily obtain:

$$\underline{t}^T \cdot \underline{N} = \underline{t}^T \cdot \underline{S} \cdot \underline{\xi} = 1$$

and

$$\underline{h}^T \cdot \underline{N} = \underline{h}^T \cdot \underline{S} \cdot \underline{\xi} = -4(\xi^2 + \eta^2 - 12\xi^2\eta^2)$$

and then differentiating these expressions and using (2.12) we get:

$$\underline{A} \cdot \underline{t} = 0$$

$$\underline{A} \cdot \underline{h} = -8 \left[\eta(1-12\xi^2) \frac{dN}{d\xi} + \xi(1-12\eta^2) \frac{dN}{d\eta} \right]$$

the second expression vanishes when the point (ξ, η) is one of the four Gauss integration points of $\hat{\Omega}$

$$(\xi^\alpha, \eta^\alpha) = \left(\pm \frac{1}{2\sqrt{3}}, \pm \frac{1}{2\sqrt{3}} \right); \alpha = 1, 4 \quad (2.38)$$

therefore we have

$$\underline{A}_\alpha \cdot \underline{t} = \underline{A}_\alpha \cdot \underline{h} = 0 \quad \alpha = 1, 4 \quad (2.39)$$

and consequently

$$\underline{K}_e^{(4)} \cdot \underline{t} = \underline{K}_e^{(4)} \cdot \underline{h} = 0 \quad (2.40)$$

which proves the rank deficiency of $\underline{K}_e^{(4)}$. Once again we note that the pattern of \underline{h} defined in (2.37) is independent of the geometry of the element and is therefore valid for a rectangular element as well as for an irregular element.

The search for decomposition for this element cannot be completed in a manner as complete as it has been for the 4-node element. However, Belytschko and co-workers [3] have intuitively come up with a decomposition similar to (2.26) where $\underline{\gamma}$ and $\underline{\epsilon}$ are

$$\underline{\gamma} = \underline{h} - \frac{1}{4} \underline{h}^T \cdot \underline{x} \sum_{\alpha=1}^4 \frac{b_1^\alpha}{J_\alpha} - \frac{1}{4} \underline{h}^T \cdot \underline{y} \sum_{\alpha=1}^4 \frac{b_2^\alpha}{J_\alpha} \quad (2.41)$$

$$\underline{\epsilon} = \frac{1}{100} \sum_{\alpha=1}^4 \frac{1}{J_\alpha} \left[\underline{b}_1^{\alpha T} \cdot \underline{b}_1^\alpha + \underline{b}_2^{\alpha T} \cdot \underline{b}_2^\alpha \right] \quad (2.42)$$

This decomposition does in fact satisfy several properties also satisfied by the exact matrix, as

$$\underline{K}_e \cdot \underline{t} = 0$$

$$\underline{K}_e \cdot \underline{x} = \sum_{\alpha=1}^4 \underline{b}_1^\alpha$$

$$\underline{K}_e \cdot \underline{y} = \sum_{\alpha=1}^4 \underline{b}_2^\alpha.$$

but for a simple square element*, \underline{K}_e and its decomposition (2.26) does not coincide. Indeed, for this simple geometry, the calculation of (2.11) can be carried out explicitly and the polynomial in (ξ, η) obtained can be split into one part exactly integrated with 4 Gauss points, and another part of higher order that requires 9 points. This calculation leads to the decomposition:

$$\left. \begin{aligned} K_{xx}^{(9)} &= K_{xx}^{(4)} + \Omega_e \left(\frac{1}{45} \underline{s}_7 \cdot \underline{s}_7^T + \frac{4}{135} \underline{s}_9 \cdot \underline{s}_9^T \right) \\ K_{yy}^{(9)} &= K_{yy}^{(4)} + \Omega_e \left(\frac{1}{45} \underline{s}_8 \cdot \underline{s}_8^T + \frac{4}{135} \underline{s}_9 \cdot \underline{s}_9^T \right) \end{aligned} \right\} \quad (2.43)$$

where

$$\left. \begin{aligned} K_{xx} &= \int_{\hat{\Omega}} \frac{\underline{A} \cdot \underline{y} \cdot \underline{y}^T \cdot \underline{A}^T}{\underline{y}^T \underline{A} \underline{x}} d\xi d\eta \\ K_{yy} &= \int_{\hat{\Omega}} \frac{\underline{A} \cdot \underline{x} \cdot \underline{x}^T \cdot \underline{A}^T}{\underline{y}^T \underline{A} \underline{x}} d\xi d\eta \end{aligned} \right\} \quad (2.44)$$

and \underline{s}_7 , \underline{s}_8 , \underline{s}_9 are the 7th, 8th and 9th column vector of \underline{S} (2.32). These vectors correspond to the higher order of $\underline{\xi}$ (2.33) that cannot be exactly integrated by a 4-point rule. The form taken by the stabilization matrix involves now three matrices $(\underline{s}_i \cdot \underline{s}_i^T, i = 7, 8, 9)$, is exact for a square element and cannot coincide with the decomposition found in [3]. Finally, we note that both decompositions were used in our a-posteriori control described in Section 2.7 on a regular mesh,

* or also for a geometry for which the Jacobian is constant.

and optimal rates of convergence were only obtained with the decomposition (2.43).

2.2.4. The stabilization matrix for a general heat transfer equation. In this paragraph we would like to give the stabilization matrix for a slightly more complicated operator. The case of the linear elasticity operator will be discussed later.

Let us consider the case where the operator is defined by

$$A = \underset{\sim}{\beta}^T \underset{\sim}{C} \underset{\sim}{\beta} \quad (2.45)$$

where

$$\underset{\sim}{\beta} = \left(\frac{\partial}{\partial x} \quad \frac{\partial}{\partial y} \right)^T$$

and

$$\underset{\sim}{C} = \begin{pmatrix} C_{11} & C_{21} \\ C_{12} & C_{22} \end{pmatrix}$$

then the stiffness matrix associated with this operator is given by

$$\underset{\sim}{K}_e = \int_{\Omega} \underset{\sim}{VN}^T \cdot \underset{\sim}{C} \cdot \underset{\sim}{VN} \, dx dy \quad (2.46)$$

The generalization of the stabilization decomposition when Q_1 elements are used can then be written

$$\underset{\sim}{K}_e = \underset{\sim}{K}_e^{(1)} + \bar{\epsilon} \underset{\sim}{Y} \cdot \underset{\sim}{Y}^T \quad (2.47)$$

where

$$\underset{\sim}{K}_e^{(1)} = \frac{1}{|\Omega_e|} \underset{\sim}{B}^T \underset{\sim}{C} \underset{\sim}{B} \quad (2.48)$$

$$\bar{\epsilon} = C_{11} \bar{\epsilon}_{xx} + (C_{12} + C_{21}) \bar{\epsilon}_{xy} + C_{22} \bar{\epsilon}_{yy} \quad (2.49)$$

and

$$\left. \begin{aligned} \bar{\epsilon}_{xx} &= \frac{1}{4} \int_{\hat{\Omega}} \frac{1}{J} (\xi \underline{s}^T \cdot \underline{y} - \eta \underline{s}'^T \cdot \underline{y})^2 d\xi d\eta \\ \bar{\epsilon}_{yy} &= \frac{1}{4} \int_{\hat{\Omega}} \frac{1}{J} (\xi \underline{s}^T \cdot \underline{x} - \eta \underline{s}'^T \cdot \underline{y})^2 d\xi d\eta \\ \bar{\epsilon}_{xy} &= \frac{1}{4} \int_{\hat{\Omega}} \frac{1}{J} (\xi \underline{s}^T \cdot \underline{x} - \eta \underline{s}'^T \cdot \underline{x}) (\xi \underline{s}^T \cdot \underline{y} - \eta \underline{s}'^T \cdot \underline{y}) d\xi d\eta \end{aligned} \right\} (2.50)$$

The quantities $\underline{\gamma}$, \underline{B} and J are the ones previously defined. Expressions similar to those given in (2.29) and (2.30) can be used to simplify $\bar{\epsilon}$.

For a regular geometry, and corresponding to (2.29) and (2.30) we have

$$\left. \begin{aligned} \bar{\epsilon}_{xx} &= \frac{1}{24(\Omega_e)} (y_{13}^2 + y_{24}^2) \\ \bar{\epsilon}_{yy} &= \frac{1}{24(\Omega_e)} (x_{13}^2 + x_{24}^2) \\ \bar{\epsilon}_{xy} &= \frac{1}{24(\Omega_e)} (x_{13} y_{13} + x_{24} y_{24}) \end{aligned} \right\} (2.51)$$

As far as the 9-node element is concerned, the decomposition can be obtained only for regular elements. First we note that

$$\underline{K}_{xy}^{(9)} = \underline{K}_{xy}^{(4)} \quad (2.52)$$

where the notations are similar to (2.43) and (2.44). Therefore, the decomposition can be written:

$$\begin{aligned} \underline{K}_e^{(9)} &= \underline{K}_e^{(4)} + \Omega_e C_{11} \left(\frac{1}{45} \underline{s}_7 \underline{s}_7^T + \frac{4}{135} \underline{s}_9 \underline{s}_9^T \right) \\ &\quad + \Omega_e C_{22} \left(\frac{1}{45} \underline{s}_8 \underline{s}_8^T + \frac{4}{135} \underline{s}_9 \underline{s}_9^T \right) \end{aligned} \quad (2.53)$$

2.3 A-Posteriori Hourglass Control

2.3.1. Introduction and preliminaries. The basic ideas are more easily understood when demonstrated for the same simple model problem. We still focus on the model Neumann problem P_0 or its variational equivalent P .

Let Ω be a regular (e.g. Lipschitz) domain in \mathbb{R}^2 with boundary $\partial\Omega$ and let f be a given L^2 -function. Problem P_0 is then,

(P_0) Find u such that

$$\left. \begin{aligned} -\Delta u &= f \text{ in } \Omega \\ \frac{\partial u}{\partial n} &= 0 \text{ on } \partial\Omega \end{aligned} \right\} \quad (3.1)$$

where the data f satisfies the compatibility condition ,

$$\int_{\Omega} f dx = 0 \quad (3.2)$$

Later we shall put further restrictions on Ω and on f (e.g. we will need $f \in H(\Omega)$). The kernel of the governing operator $A = (-\Delta, \frac{\partial}{\partial n})$ in (3.1) is, of course, the space of constants. Thus, whenever (3.2) holds, there exists a solution to (3.1) which is unique up to an arbitrary constant.

To formulate a variational statement of problem P_0 , we introduce the spaces and inner products^{*},

* The elements of V (and $L^2(\Omega)/\mathbb{R}$) are cosets $[v]$ such that $u \in [v]$ implies that $u, v \in H^1(\Omega)$ (or $L^2(\Omega)$) and $v - u \in \mathbb{R}$. Throughout this paper we frequently refer to functions v in V , meaning, of course, that v is a representative function in the coset $[v]$.

$$V = H^1(\Omega) / \mathbb{R}$$

$(u, v)_1$ = an inner product on V

$$= \int_{\Omega} \nabla u \cdot \nabla v dx ; u, v \in V$$

$(f, g)_0$ = an inner product on $L^2(\Omega) / \mathbb{R}$

$$= \int_{\Omega} fg dx - \frac{1}{\text{meas } \Omega} \int_{\Omega} f dx \int_{\Omega} g dx \quad (3.3)$$

Three remarks are in order:

i) The norm $\|\cdot\|_0$ associated with the inner product $(\cdot, \cdot)_0$ is the canonical norm on the quotient space $L^2(\Omega) / \mathbb{R}$,

$$\|f\|_0 = \inf_{\lambda \in \mathbb{R}} \|f + \lambda\|_{L^2(\Omega)} \quad (3.4)$$

ii) According to Temam [45], there exists a constant C_0 , depending only on Ω , such that

$$\|v\|_1 \geq C_0 \|v\|_0 \quad \forall v \in V \quad (3.5)$$

iii) For all f satisfying the compatibility condition (3.2) and any $v \in V$, we have

$$\begin{aligned} (f, v)_0 &= \int_{\Omega} f v dx \leq \|f\|_0 \|v\|_0 \\ &\leq \frac{1}{C_0} \|f\|_0 \|v\|_1 \end{aligned} \quad (3.6)$$

With these relations now established, we consider the variational statement of P_0 as problem P :

(P) Find $u \in V$ such that

$$(u, v)_1 = (f, v)_0 \quad \forall v \in V \quad (3.7)$$

We can easily verify that any solution of P_0 is a solution of P and, conversely, the solution of P satisfies the condition of P_0 in at least, a distributional sense. Moreover, since the bilinear form $(\cdot, \cdot)_1$ is continuous and coercive on V and since the linear form $(f, \cdot)_0$ is continuous on V if (and only if) f satisfies (3.2), the following result is an immediate consequence of the Lax-Milgram Theorem:

THEOREM V. *Let f satisfy (3.2).*

Then there exists one and only one solution $u \in V$ to problem P and this solution depends continuously on the data f . \square

We now consider a finite element approximation of the problem P . Let us now construct a finite element approximation of problem P . We begin by introducing a partition \mathcal{Q} of Ω into E finite elements so that

$$\Omega = \bigcup_{e=1}^E \Omega_e$$

We shall assume that Ω is such that it can be partitioned in this fashion into four-node quadrilateral elements over which bilinear shape functions are defined. Thus, if

$$Q_1(\Omega_e) = \text{space of bilinear functions defined on } \Omega_e$$

we can introduce the finite-dimensional space

$$V^h = \left\{ v^h \in C^0(\Omega) \text{ such that } v^h|_{\Omega_e} \in Q_1(\Omega_e), 1 \leq e \leq E / \mathbb{R} \subset V \right\} \quad (3.8)$$

wherein, as usual, the label h is the mesh parameter (e.g. ,
 $h = \max_{1 \leq e \leq E} \text{dia}(\Omega_e)$). The functions in V^h are continuous and are still
 defined up to an arbitrary constant.

Our finite-element approximation of problem P is embodied in the
 discrete problem,

$$(P_h) \quad \text{Find } u^h \in V^h \text{ such that} \tag{3.9}$$

$$(u^h, v^h)_1 = (f, v^h)_0 \quad \forall v^h \in V^h$$

where, again, f satisfies condition (3.2).

In analogy with Theorem V, we have:

THEOREM VI. *Let f satisfy (3.2). Then there is one and only
 one solution u^h to problem P_h in V^h and this solution depends
 continuously on the data f . \square*

In examining the convergence of such finite element approximations,
 we shall confine our attention throughout this study to regular mesh
 refinements. In such cases, we have the a priori asymptotic error
 estimates,

$$\|u - u^h\|_1 = O(h), \quad \|u - u^h\|_0 = O(h^2) \tag{3.10}$$

2.3.2 The underintegrated problem. We now focus our attention
 on finite element approximations of problem P in which incomplete
 quadratures are used to evaluate the bilinear form $(\cdot, \cdot)_1$. To simpli-
 fy this study, we shall now introduce some additional assumptions:

- i) Ω is the unit square,

$$\Omega = (0,1) \times (0,1)$$

ii) The finite elements are the squares,

$$\Omega_{ij} = \left(\frac{i-1}{N}, \frac{i}{N}\right) \times \left(\frac{j-1}{N}, \frac{j}{N}\right)$$

$$1 \leq i, j \leq N$$

$$\bar{\Omega} = \bigcup_{1 \leq i, j \leq N} \bar{\Omega}_{ij}$$

iii) The data f is L^2 -integrable; e.g.

$$f \in L^2(\Omega) \quad (3.11)$$

In this case, we take

$$h = \frac{1}{N}, \dim V^h = (N+1)^2 - 1 = O(h^{-2}) \quad (3.12)$$

In P_h we can replace f by f^h , its L^2 -projection on V^h is defined by

$$(f^h, v^h)_0 = (f, v^h)_0 \quad \forall v^h \in V^h \quad (3.13)$$

For further use, we note that the projection satisfies

$$\|f^h\|_0 \leq \|f\|_0 \quad (3.14)$$

and can be chosen such that

$$\int_{\Omega} f^h dx = 0 \quad (3.15)$$

Now we turn to the issue of numerical integration of the stiffnesses. Let $I(\cdot, \cdot)$ denote a discrete inner product on $C^0(\Omega)$ defined by a numerical quadrature rule as follows:

$$I_G(f,g) = \sum_{e=1}^E I_e^G(f,g) \quad (3.16)$$

$$I_e^G(f,g) = \sum_{j=1}^G W_j^e f(\xi_j^e) g(\xi_j^e)$$

Here W_j^e are the quadrature weights and ξ_j^e are the quadrature points for element e and G is the number of quadrature points used.

Assuming that Gaussian quadrature is used, the choice $G=4$ (2x2 - Gauss rule) leads to an exact integration of the stiffnesses for each element:

$$(u^h, v^h)_1 = I_4(u^h, v^h) = \underline{u}^T \underline{K} \underline{v} \quad (3.17)$$

for any $u^h, v^h \in V^h$. Here \underline{K} is the fully-integrated stiffness matrix and \underline{u} and \underline{v} are vectors of nodal degrees of freedom of u^h and v^h , respectively.

Instead of the correct bilinear form in (3.18), we wish to consider an underintegrated approximation to $(\cdot, \cdot)_1$ in which only one integration point per element is used:*

$$(u^h, v^h)_{1,h} = I_1(u^h, v^h) = \underline{u}^T \underline{K}^{(1)} \underline{v} \quad (3.18)$$

$$\forall u^h, v^h \in V^h$$

Here $\underline{K}^{(1)}$ is the underintegrated stiffness matrix. The difference between $(\cdot, \cdot)_1$ and $(\cdot, \cdot)_{1,h}$ (on V^h) is denoted $a'(\cdot, \cdot)$ and the corresponding stiffness matrix is $\underline{K}^{\text{stab}}$:**

* Recall Section (2.2).

** Recall that $\underline{K}^{\text{stab}} = \bar{\epsilon} \underline{\gamma} \underline{\gamma}^T$ where $\bar{\epsilon} = 1/6$ for a rectangular mesh and $\underline{\gamma}$ is given by (2.27).

$$\begin{aligned}
 a'(u^h, v^h) &= (u^h, v^h)_1 - (u^h, v^h)_{1,h} \\
 &= \underline{\underline{u}}^T \underline{\underline{K}}^{\text{stab}} \underline{\underline{v}} \quad \forall u^h, v^h \in V^h
 \end{aligned} \tag{3.19}$$

The "underintegrated problem",

(P_h^{*}) Find $u^h \in V^h$ such that

$$(u^h, v^h)_{1,h} = (f^h, v^h)_0 \quad \forall v^h \in V^h \tag{3.20}$$

is, in general, meaningless. This problem, in general, has no solution except for the special case in which f^h is orthogonal to the one-dimensional space of hourglass modes,

$$H = \{H \in V^h \mid (H, v^h)_{1,h} = 0 \quad \forall v^h \in V^h\} \tag{3.21}$$

A way to overcome this difficulty is to note that the underintegration of the righthand side also leads to a rank-deficient linear form

$(\cdot, \cdot)_{0,h}$:

$$\begin{aligned}
 (f^h, H)_{0,h} = 0 \quad , \quad & \forall f^h \in V^h \\
 & \forall H \in H
 \end{aligned}$$

Note that if f^h satisfies (3.15) we also have

$$(f^h, 1)_{0,h} = 0 \tag{3.23}$$

Therefore we now consider the underintegrated problem \bar{P}_h :

(\bar{P}_h) Find $\bar{u}^h \in \bar{V}^h$ such that

$$(\bar{u}^h, v^h)_{1,h} = (f^h, v^h)_{0,h} \quad \forall v^h \in \bar{V}^h \tag{3.24}$$

where

$$\bar{V}^h : V^h/H$$

We can now state and prove

THEOREM VII: There exists one and only one solution \bar{u}^h to \bar{P}_h .

Proof: This is an immediate consequence of the Lax-Milgram theorem. Since

$$(u^h, u^h)_{1,h} = 0 \iff u^h = \gamma_1 H + \gamma_2$$

we can consider $(\cdot, \cdot)_{1,h}^{1/2}$ as a norm on \bar{V}^h . It is therefore coercive and continuous on \bar{V}^h . As far as the continuity of the righthand side is concerned, a simple calculation shows that for any v^h in V^h we have

$$|(f^h, v^h)_{0,h}| \leq \|f^h\|_0 \|v^h\|_0$$

Also for any constants γ_1 and γ_2

$$|(f^h, v^h + \gamma_1 + \gamma_2 H)_{0,h}| = |(f^h, v^h)_{0,h}|$$

therefore

$$\begin{aligned} |(f^h, v^h)_{0,h}| &\leq \|f^h\|_0 \|v^h + \gamma_1 + \gamma_2 H\|_0 \quad \forall \gamma_1, \gamma_2 \\ &\leq \|f^h\|_0 \|v^h + \gamma_2 H\|_1 \quad \forall \gamma_2 \\ &\leq \alpha_h \|f^h\|_0 \|v^h\|_{1,h} \end{aligned}$$

Here we successively used (2.23), (2.22), (3,6) and the equivalence between the canonical norm of \bar{V}^h and the norm $\|\cdot\|_{1,h}$ \square

We have obtained a solution to the underintegrated problem \bar{P}_h .

This solution is unique in \bar{V}^h , from a computational point of view it is defined up to within an arbitrary hourglass mode. We now need a projection to obtain a reasonable solution from any representative \bar{u}^h chosen.

2.3.3. Projection of the underintegrated solution. In order to construct this projection, we remark that, since u^h is a solution of P_h and since $H \in V^h$, u^h satisfies

$$(u^h, H)_1 = (f^h, H)_0 \quad (3.26)$$

We wish to extend \bar{u}^h to all of V^h so that a new function $\tilde{u}^h \in V^h$ is obtained which contains an hourglass mode and which also satisfies (4.8). Thus, if π is an operator from \bar{V}^h into V^h , we define

$$\tilde{u}^h = \pi \bar{u}^h = \bar{u}^h + \lambda_0 H, \quad \lambda_0 \in \mathbb{R} \quad (3.27)$$

$$(\tilde{u}^h, H)_1 = (f^h, H)_0$$

This latter requirement determines λ_0 uniquely as

$$\lambda_0 = \frac{1}{\|H\|_1^2} \left[(f^h, H)_0 - (\bar{u}^h, H)_1 \right] \quad (3.29)$$

so that \tilde{u}^h is uniquely determined as the function

$$\tilde{u}^h = \bar{u}^h + \frac{(f^h, H)_0}{\|H\|_1^2} H - \frac{(u^h, H)_1}{\|H\|_1^2} H$$

It is instructive to consider a geometrical interpretation of our projection defined in (4.9). Note that the "component" of the fully integrated solution u^h orthogonal (in V^h) to H^\perp is $(u^h, H)_1 = (f^h, H)_0$.

as indicated in Fig. 17. The solutions \bar{u}_h of \bar{P}_h constitute the vectors generating a line "parallel to" the space H in the figure. The projection \tilde{u}^h is then the vector defined by the orthogonal projection of u^h onto this line. Indeed, by construction,

$$(\tilde{u}^h - u^h, H)_1 = 0$$

At this point, we have established the following procedure for processing an underintegrated finite element approximation of problem P.

- i) Compute the underintegrated bilinear and linear forms $(\cdot, \cdot)_{1,h}$ and $(f^h, \cdot)_{0,h}$
- ii) Solve problem \bar{P}_h for \bar{u}^h
- iii) Compute $(\bar{u}^h, H)_1$
- iv) Construct the enhanced solution \tilde{u}^h using (3.30).

Thus, this procedure involves the computation of an underintegrated solution \bar{u}_h to a reduced problem \bar{P}_h and its enrichment via a post-processing operation to obtain a new approximation \tilde{u}^h . We shall now show that these post-processed solutions \tilde{u}^h converge to the exact solution u of problem P as the mesh is refined, and, remarkably, these approximations converge at precisely the same rate as the fully-integrated solution!

Indeed we have:

THEOREM VIII: Let u , u^h and \bar{u}^h be the solutions of P, P_h and \bar{P}_h , let f be in $L^2(\Omega)$ and satisfy (3.2). Let \tilde{u}^h be ob-

tained by the projection of \bar{u}^h defined in (3.30). Then we have the following error estimates for $s = 0$ and 1

$$\|u^h - \tilde{u}^h\|_s \leq C_1 h^{2-s} \|f\|_0 \quad (3.31)$$

and

$$\|u - \tilde{u}^h\|_s \leq C_1 h^{2-s} \|f\|_0 \quad (3.32)$$

The next section will prove this theorem.

2.3. Convergence of the A-Posteriori Control

2.4.1 Introduction. This section is devoted to the proof of Theorem VIII. The method of proof relies on the tensor properties of the bilinear element and of the Gauss integration rules. The problems P_h and \bar{P}_h will be explicitly solved using an orthonormal basis of eigenvectors of $(\cdot, \cdot)_1$, $(\cdot, \cdot)_{1,h}$ and $(\cdot, \cdot)_{0,h}$. Then we note that for a regular domain and mesh, $f \in L^2(\Omega)$ implies $u \in H^2(\Omega)$ and that

$$\|u - u^h\|_1 < Ch \|f\|_0 \quad (4.1)$$

Likewise, the Aubin-Nitsche method provides also

$$\|u - u^h\|_0 \leq C'h^2 \|f\|_0 \quad (4.2)$$

By the triangle inequality,

$$\|u - \tilde{u}^h\|_1 \leq Ch \|f\|_0 + \|u^h - \tilde{u}^h\|_1 \quad (4.3)$$

with a similar estimate in the $\|\cdot\|_0$ -norm.

Thus, it suffices to estimate the relative error

$$e^h = \tilde{u}^h - u^h \quad (4.4)$$

The L^2 - and H^1 -norms of this relative error will be explicitly calculated and estimated.

2.4.2. Some one-dimensional results. For reasons to be made clear in the next subsection, it is convenient to review briefly some results on one-dimensional piecewise-linear approximations on a uniform mesh for $\Omega = (0,1)$. Our aim here is to establish concrete relationships between various bilinear forms $(\cdot, \cdot)_{0,h}$, $(\cdot, \cdot)_1$, and $(\cdot, \cdot)_{1,h}$ on spaces of piecewise-linear functions.

Let $\tilde{D}(k, \alpha)$ and \tilde{I} denote the $N+1$ -order matrices

$$\tilde{D}(k, \alpha) = \begin{bmatrix} k & \alpha & \cdot & \cdot & \cdot & 0 & 0 \\ \alpha & 2k & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 2k & \alpha \\ 0 & 0 & \cdot & \cdot & \cdot & \alpha & k \end{bmatrix}, \tilde{I}' = \begin{bmatrix} 1 & 0 & \cdot & \cdot & \cdot & 0 & 0 \\ 0 & 2 & \cdot & \cdot & \cdot & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & \cdot & \cdot & \cdot & 2 & 0 \\ 0 & 0 & \cdot & \cdot & \cdot & 0 & 1 \end{bmatrix} \quad (4.5)$$

(i.e. $\tilde{I}' = \tilde{D}(1,0)$). Then, for $\alpha \neq 0$, one can show that

$$\begin{aligned} \det \tilde{D}(k, \alpha) &= (-\alpha)^{N+1} \det \tilde{D}\left(-\frac{k}{\alpha}, -1\right) \\ &= (-\alpha)^{N+1} \det \tilde{D}\left(-\frac{k}{\alpha}\right) \end{aligned} \quad (4.6)$$

where

$$\tilde{D}(k) \stackrel{\text{def}}{=} \tilde{D}(k, -1) \quad (4.7)$$

The values of k for which $\det \tilde{D}(k)$ vanishes are

$$k_i = \cos \frac{i\pi}{N}, \quad 0 \leq i \leq N \quad (4.8)$$

and the corresponding vectors $(D(k_i)v_i = 0)$ are

$$v_i = \left\{ \cos \frac{ij\pi}{N} \right\}, \quad 0 \leq j \leq N \quad (4.9)$$

The significance of the above matrices is that in one dimension, the discrete $H^1(0,1)$ - , $L^2(0,1)$ - and underintegrated $L^2(0,1)$ -norms, on the space V_1^h of piecewise linear C_N^0 -functions on a uniform mesh of N elements on $(0,1)$,

$$V_1^h = \{v^h \in C^0(0,1) \mid v^h \text{ is linear on } [eh, (e+1)h], e=0, \dots, N-1\} \quad (4.10)$$

are associated with the matrices

$$A_0 = \frac{h}{3} D(1, \frac{1}{2}), \quad A_1 = \frac{1}{h} D(1, -1) \quad \text{and} \quad A_{0,h} = \frac{h}{4} D(1,1) \quad (4.11)$$

respectively. In other words,

$$\|v^h\|_s^2 = v^T A_s v \quad s = 0, 1, (0,h) \quad (4.12)$$

where v is the vector of nodal values of v^h .

By using (4.6) through (4.8), one can verify that the numbers α_i and β_i which render $A_{0,h} - \alpha_i A_{i=0}$ and $A_1 - \beta_i A_{i=0}$ singular are

$$\alpha_i = \frac{3(1 + \cos \frac{i\pi}{N})}{2(2 + \cos \frac{i\pi}{N})} \quad (4.13)$$

$$\beta_i = \frac{6}{h^2} \frac{1 - \cos \frac{i\pi}{N}}{2 + \cos \frac{i\pi}{N}} \quad (4.14)$$

In particular, let $\phi^i = \phi^i(x)$, $x \in [0,1]$ denote the piecewise

linear functions associated with the vectors v_i :

$$\left. \begin{aligned} \phi^i(jh) &= \cos \frac{ij\pi}{N}, \quad 0 \leq i, j \leq N \\ \text{span } \{\phi^i\}_{0 \leq i \leq N} &= V_1^h \end{aligned} \right\} \quad (4.15)$$

Then,

$$(v^h, \phi^i)_{0,h} = \alpha_i (v^h, \phi^i)_0 \quad (4.16)$$

$$(v^h, \phi^i)_1 = \beta_i (v^h, \phi^i)_0 \quad (4.17)$$

$v^h \in V_1^h$

Notice that the base functions ϕ^i are orthogonal for each of the scalar products under consideration.

The following remarks are in order:

- i) The denominators in (4.13) and (4.14) are non-zero.
- ii) For $i = N$, $\alpha_i = 0$ and the corresponding eigenfunction is the one-dimensional hourglass mode:

$$(1, -1, 1, -1, \dots)$$
- iii) For $i = 0$, $\beta_i = 0$ and the corresponding eigenfunction is constant. Then we have the condition $(v^h, 1)_1 = 0$ as expected.

2.4.3. Discrete norms for two dimensional meshes. The extension of the above results to two-dimensional rectangular meshes is straightforward. Since the bilinear basis functions for V^h are tensor products of piecewise linear functions of one variable, we can define

$$\phi^{ij}(x,y) = \phi^i(x)\phi^j(y)$$

$$0 \leq i, j \leq N \quad (4.18)$$

Further, let us normalize these basis functions so that

$$\|\phi^{ij}\|_0 = 1$$

We can then establish the following:

Lemma 4.1. For $v^h \in V^h$, we have

$$(v^h, \phi^{ij})_{0,h} = \alpha_i \alpha_j (v^h, \phi^{ij})_0 \quad (4.19)$$

$$(v^h, \phi^{ij})_1 = (\beta_i + \beta_j) (v^h, \phi^{ij})_0 \quad (4.20)$$

$$(v^h, \phi^{ij})_{1,h} = (\alpha_j \beta_i + \alpha_i \beta_j) (v^h, \phi^{ij})_0 \quad (4.21)$$

Moreover, if arbitrary $v^h \in V^h$ is expressed in the form,

$$\left. \begin{aligned} v^h &= \sum_{0 \leq i, j \leq N} v_{ij} \phi^{ij} \\ v_{ij} &= (v^h, \phi^{ij})_0 \end{aligned} \right\} \quad (4.22)$$

then

$$\|v^h\|_0^2 = \sum_{0 \leq i, j \leq N} v_{ij}^2 \quad (4.23)$$

$$\|v^h\|_1^2 = \sum_{0 \leq i, j \leq N} (\beta_i + \beta_j) v_{ij}^2 \quad (4.24)$$

Proof: First note that

$$\begin{aligned} (v^h, \phi^{ij})_{0,h} &= I(v^h \phi^i(x) \phi^j(y)) \\ &= \alpha_i \alpha_j \int_{\Omega} v^h \phi^i(x) \phi^j(y) \, dx dy \\ &= \alpha_i \alpha_j (v^h, \phi^{ij})_0 \end{aligned}$$

We also have

$$\begin{aligned}
 (v^h, \phi^{ij})_1 &= \int_{\Omega} \frac{\partial v^h}{\partial x} \phi^{i'}(x) \phi^j(y) \\
 &\quad + \frac{\partial v^h}{\partial y} \phi^{j'}(y) \phi^i(x) \, dx dy \\
 &= \beta_i \int_{\Omega} v^h \phi^i(x) \phi^j(y) \, dx dy \\
 &\quad + \beta_j \int_{\Omega} v^h \phi^i(x) \phi^j(y) \, dx dy \\
 &= (\beta_i + \beta_j) (v^h, \phi^{ij})_0
 \end{aligned}$$

Finally

$$\begin{aligned}
 (v^h, \phi^{ij})_{1,h} &= I_1 \left(\frac{\partial v^h}{\partial x} \phi^{i'} \phi^j + \frac{\partial v^h}{\partial y} \phi^i \phi^{j'} \right) \\
 &= \beta_i \alpha_j (v^h, \phi^{ij})_0 + \beta_j \alpha_i (v^h, \phi^{ij})_0
 \end{aligned}$$

The norms (4.23) and (4.24) are then directly obtained \square

In analogy with our remarks on the one-dimensional case, we observe that for $i = j = N$, $\phi^{ij} = H$, the two dimensional hourglass mode. Then

$$(v^h, H)_1 = \frac{24}{h^2} (v^h, H)_0 \quad (4.25)$$

and

$$(v^h, H)_{0,h} = 0 \quad (4.26)$$

Also, for $i = j = 0$, $\phi^{ij} = 1$ and the equilibrium condition (3.2) can be written

$$f_{00} = 0 \quad (4.27)$$

with

$$f_{ij} = (f^h, \phi^{ij})_0 \quad (4.28)$$

2.4.4. Explicit resolution of P_h and $(\bar{P}_h + \pi)$. With the above results in hand, let us now return to the fully-integrated finite-element approximate problem P_h given in (3.9). The solution u^h to that problem can be written

$$\left. \begin{aligned} u^h &= \sum_{0 \leq i, j \leq N} u_{ij} \phi^{ij} \\ u_{ij} &= (u^h, \phi^{ij})_0 \end{aligned} \right\} \quad (4.29)$$

and since for $[v^h = \phi^{ij} \text{ in (3.9)}]$,

$$\begin{aligned} (u^h, \phi^{ij})_1 &= (\beta_i + \beta_j) (u^h, \phi^{ij})_0 \\ &= (f^h, \phi^{ij})_0 = f_{ij} \end{aligned}$$

we have

$$u_{ij} = \frac{1}{\beta_i + \beta_j} f_{ij} ; (i,j) \neq (0,0) \quad (4.30)$$

Using constructions similar to those in (4.29) for the fully-integrated problem, we easily verify that the solution \bar{u}^h to the underintegrated problem \bar{P}_h is representable in the form,

$$\bar{u}^h = \sum_{\substack{(i,j) \neq (N,N) \\ (i,j) \neq (0,0)}} \bar{u}_{ij} \phi^{ij} \quad (4.31)$$

with

$$\bar{u}_{ij} = \frac{\alpha_i \alpha_j}{\alpha_i \beta_j + \alpha_j \beta_i} f_{ij} ; (i,j) \neq (0,0) \text{ and } (N,N) \quad (4.32)$$

The cases $(i,j) = (N,N)$ and $(i,j) = (0,0)$ correspond to the arbitrary hourglass mode and arbitrary constant, respectively.

The projected approximation \tilde{u}^h defined by $\tilde{u}^h = \pi \bar{u}^h$ is constructed so that projections of \tilde{u}^h and u^h coincide; i.e.

$$\left. \begin{aligned} \tilde{u}_{ij} &= \bar{u}_{ij} & (i,j) \neq (0,0) \text{ and } (N,N) \\ \tilde{u}_{N,N} &= \bar{u}_{N,N} \end{aligned} \right\} \quad (4.33)$$

2.4.5. Proof of theorem VIII. Since the error function $e^h = u^h - \tilde{u}^h$ is in V^h , we use (4.29) and (4.31) to obtain

$$e^h = \sum_{\substack{(i,j) \neq (N,N) \\ (i,j) \neq (0,0)}} e_{ij} \phi^{ij} \quad (4.34)$$

where

$$\begin{aligned} e_{ij} &= (e^h, \phi^{ij})_0 \\ &= (\tilde{u}^h, \phi^{ij})_0 - (u^h, \phi^{ij})_0 \\ &= \tilde{u}_{ij} - u_{ij} \end{aligned}$$

Thus, from (4.30) and (4.33),

$$e_{ij} = \left(\frac{\alpha_i \alpha_j}{\alpha_i \beta_j + \alpha_j \beta_i} - \frac{1}{\beta_i + \beta_j} \right) f_{ij} \quad (4.35)$$

Then, using (4.13) and (4.14), e_{ij} can be written as

$$e_{ij} = h^2 K_{ij} f_{ij} \quad (4.36)$$

where

$$K_{ij} = K\left(\cos \frac{i\pi}{N}, \cos \frac{j\pi}{N}\right) \quad (4.37)$$

and

$$K(x,y) = \frac{1}{4} \frac{(1+x)(1+y)}{(1+x)(1-y)+(1+y)(1-x)} - \frac{1}{6} \frac{(2+x)(2+y)}{(2+x)(1-y)+(2+y)(1-x)} \quad (4.38)$$

On the square $S = [-1,+1] \times [-1,+1] \setminus \{(-1,-1), (1,1)\}$, $K(\cdot, \cdot)$ is bounded and there exists a positive constant K such that

$$|K(x,y)| \leq K \quad \forall (x,y) \in S \quad (4.39)$$

Therefore we have

$$|K_{ij}| \leq K \quad \forall (i,j) \neq (0,0) \text{ and } (N,N) \quad (4.40)$$

and we can obtain using (3.13) and (4.23)

$$\begin{aligned} \|e^h\|_0^2 &= h^4 \sum_{\substack{(i,j) \neq (0,0) \\ (i,j) \neq (N,N)}} K_{ij}^2 f_{ij}^2 \\ &\leq h^4 K^2 \|f^h\|_0^2 \leq h^4 K^2 \|f\|_0^2 \end{aligned}$$

Also, after calculation and use of (4.24), (4.14) and (3.14), we have

$$\begin{aligned} \|e^h\|_1^2 &= h^4 \sum_{\substack{(i,j) \neq (0,0) \\ (i,j) \neq (N,N)}} (\beta_i + \beta_j) K_{ij}^2 f_{ij}^2 \\ &\leq 12 h^2 K^2 \|f\|_0^2 \quad \blacksquare \end{aligned}$$

2.5. Implementation and Numerical Results of the A-Posteriori Control

For the Laplace Equation.

In this section we first would like to indicate how the a-posteriori

control method is implemented, and how its time efficiency compares to the a-priori method. Then several numerical results will be given, illustrating the accuracy of the method and confirming the results obtained in the previous sections.

2.5.1. Implementation of the a-posteriori method. First let us indicate that from a mathematical point of view the problem \bar{P}^h is well-posed but computationally, the matrix obtained from this formulation is singular and the dimension of its kernel is 2. Consequently, we must pick two nodes, fix them a value, and solve. The first value fixes the constant mode, and the second one fixes the hourglass mode to be eliminated later. Let us fix \bar{u}^h equal to zero at the origin and at the next point on the boundary (coordinates : $h,0$) (Figure 18.a). According to the error estimates (3.30) and (3.31), we may write

$$u^h = \bar{u}^h + \lambda H + O(h^{2-s}) \quad (5.1)$$

and therefore, if we normalize H such that its nodal values are 0 or 1, λ measures precisely the value of u^h at $(h,0)$ (Figure 18.b), and approaches $u(h,0)$

$$\lambda = u(h,0) + O(h^\sigma) \quad (5.2)$$

But $u(h,0)$ is $O(h^2)$ for a smooth enough solution ($u(0,0) = 0$, $\partial u / \partial \eta(0,0) = 0$) and using L^∞ -estimates 12, σ can be evaluated to $2-\epsilon$, ϵ arbitrary. Finally, we have the estimate

$$\lambda = O(h^{2-\epsilon}), \quad \epsilon \text{ arbitrary} \quad (5.3)$$

Also, the choice of H leads to

$$\begin{aligned} \|H\|_0 &= 1/3 \\ \|H\|_1 &= \frac{2\sqrt{2}}{\sqrt{3}h} \end{aligned} \quad (5.4)$$

and therefore we obtain

$$\|\lambda H\|_s = O(h^{2-s-\epsilon}) \quad s = 0,1 \quad (5.5)$$

and that proves that the post processor contribution λH can be neglected if the fixed nodes are chosen as indicated for this type of boundary condition. The error estimates of Theorem I still hold up to within $h^{-\epsilon}$.

Unfortunately this remark has two major drawbacks: it supposes that u is smooth ($u \in H^2(\Omega)$) and it is not valid to 9-node elements that will later be discussed.

Before discussing the implementation of (3.30), we indicate that this projection can be simplified. Indeed, taking $v^h = H$ in (4.25) we obtain

$$\left\| \frac{(f_1^h H)_0}{\|H\|_1^2} H \right\| \leq \|f^h\|_0 \frac{\|H\|_0^2}{\|H\|_1^2} = \frac{h^2}{24} \|f^h\|_0 \quad (5.6)$$

$$\left\| \frac{(f_1^h H)_0}{\|H\|_1^2} H \right\|_1 \leq \|f^h\|_0 \frac{\|H\|_0}{\|H\|_1} = \frac{h}{2\sqrt{6}} \|f^h\|_0$$

Therefore we have

$$\left\| \frac{(f_1^h H)_0}{\|H\|_1^2} H \right\|_s \leq C \|f^h\|_0 h^{2-s}, \quad s = 0,1 \quad (5.8)$$

and this term can be neglected without affecting the estimate of Theorem

VIII. The formula used in the post processor is then

$$\bar{u}^h = \bar{u}^h - \lambda H \quad (5.9)$$

$$\lambda = (\bar{u}^h, H)_1 \|H\|_1^{-2} \quad (5.10)$$

In order to preserve the efficiency of the method provided by underintegration, one must find an efficient way to compute the parameter λ in (5.9). One way that suggests itself is to calculate the H^1 inner products of (5.10) using numerical integration. The use of a one point rule would be absurd and would lead to a ratio 0/0. The use of a 4 Gauss point rule has been numerically implemented and gives good results (similar to those to be presented next) but cost of this integration is expensive, as shown in Table 3. This method is therefore rejected.

We shall now describe a more efficient method with related numerical results shown in the next subsection. This method relies on the fact that, for the bilinear element, the stiffness matrix can be decomposed into two parts, one of which contains H in its kernel. The other part is such that the image of H is cheap to calculate.

This decomposition proved in Section 2.2.2 can be written as

$$K_{\approx e}^{\text{exact}} = K_{\approx e}^{\text{under}} + \bar{\epsilon}_e \gamma_e \gamma_e^T \quad (5.11)$$

where $K_{\approx e}^{\text{exact}}$ and $K_{\approx e}^{\text{under}}$ respectively are the exact element stiffness matrix and its under-integrated form, $\bar{\epsilon}_e$ and γ_e are obtained from (2.27) and (2.28). In particular

$$\gamma_e = h - \frac{h \cdot x}{|\Omega_e|} b_1 - \frac{h \cdot y}{|\Omega_e|} b_2 \quad (5.12)$$

Even though ϵ_e given by (2.27) is still difficult to calculate, note that, during the element calculations, the vectors $b_{\sim 1}$ and $b_{\sim 2}$ and the Jacobian $|\Omega_e|$ are necessarily computed. Therefore γ_e is very easy to calculate.

The inner product $(u^{\sim h}, H)_1$ may, therefore, be calculated by using the decomposition (2.1) of K_{\sim}^{exact} . Introducing the nodal vectors \bar{U} and \bar{H} associated with the function $u^{\sim h}$ and H , we have

$$\begin{aligned} (u^{\sim h}, H)_1 &= U_{\sim}^T K_{\sim}^{\text{exact}} H_{\sim} = U_{\sim}^T \cdot \sum_e \bar{\epsilon}_e \gamma_e \cdot \gamma_e^T \cdot H_{\sim} \\ &= \sum_e \bar{\epsilon}_e (\gamma_e^T \cdot \bar{U}_e) (\gamma_e^T \cdot H_e) \end{aligned}$$

where \bar{U}_e and H_e are the values of \bar{U} and \bar{H} at the nodes of the element e . We note that if the values of H_e are $+1$ or -1 , the scalar vector product $\gamma_e^T \cdot H_e$ is always ± 4 . Therefore

$$(u^{\sim h}, H)_1 = 4 \sum_e \pm \bar{\epsilon}_e \sum_{i=1}^4 \gamma_e^i u_e^i \quad (5.13)$$

and

$$(H, H)_1 = 16 \sum_e \bar{\epsilon}_e \quad (5.14)$$

These expressions are still exact since no approximation has been made on $\bar{\epsilon}_e$. If we suppose that the Jacobian of the element is approximately constant (true for parallelogram element), $\bar{\epsilon}_e$ is simply expressed as

$$\bar{\epsilon}_e = \frac{1}{12 |\Omega_e|} (b_1^T b_1 + b_2^T b_2) \quad (5.15)$$

The calculation of the approximate projection can be summarized in

the following algorithm:

- Loop on Elements
 - ↑ Calculate $\underline{\gamma}_e, \varepsilon_e$ using (2.2) and (2.6)
 - Calculate $\pm \varepsilon_e (\underline{\gamma}_e^T \cdot \bar{U}_e)$
 - Add
 - $\left\{ \begin{array}{l} \lambda_2 = \lambda_1 \pm \varepsilon_e (\underline{\gamma}_e^T \cdot \bar{U}_e) \\ \lambda_2 = \lambda_2 + \varepsilon_e \end{array} \right.$
- $\lambda = \lambda_1 / 4\lambda_2$
- Loop on Nodes
 - ↑ $\tilde{u}^h|_{\text{Node}} = \bar{u}^h|_{\text{Node}} \pm \lambda$

Remark: The notations previously used are essentially those found in the work of Belytschko and co-workers [4,5] on stabilization methods. These methods rely on the decomposition (5.11) but the stabilization term $\varepsilon \underline{\gamma} \cdot \underline{\gamma}^T$ is a-priori added to the under-integrated matrix to prevent the spurious modes from the kernel of the stiffness matrix; whereas our control method uses the very same term a-posteriori, after solving with the underintegrated matrix. Therefore, our method seems to be cheaper than the stabilization methods as summarized in Table 4.

2.5.2. Numerical results.

2.5.2.a. Regular mesh of 4-node elements. In order to illustrate what has been stated, we have considered the Laplacian problem solved on a square domain partitioned into $N^2 (= h^{-2})$ subdomains, for various values of N and we have studied the norms of the difference between the solution obtained with a full integration u^h (4 point rule

and with underintegration (1 point rule) \bar{u}^h and \tilde{u}^h (before and after post processing). The results are shown as plot of $\text{Log} \|u^h - \bar{u}^h\|$ or $\text{Log} \|u - \tilde{u}^h\|_s$ in function of $|\text{Log } h|$, for $s = 0, 1$. Data of various regularities have been used:

i) f_1 is a C^0 -function, but not C^1 :

$$\begin{cases} f_1(x,y) = \frac{3}{2}(1-x) - \frac{14}{9}y & \text{if } Y_1(x,y) \geq 0 \\ f_1(x,y) = y(\frac{3}{2}(1-x) - y - \frac{5}{9}) & \text{if } Y_1(x,y) \leq 0 \end{cases}$$

where the C^1 -discontinuity line is

$$Y_1(x,y) = \frac{3}{2}(1-x) - y$$

ii) f_2 is a non-continuous function

$$\begin{cases} f_2(x,y) = 1 & \text{if } Y_1(x,y) > 0 \\ f_2(x,y) = -2 & \text{if } Y_1(x,y) < 0 \end{cases}$$

where Y_1 is the same as in i).

Remark: Both of these functions satisfy the compatibility condition (3.2).

Results obtained with the continuous function f_1 are shown in Figure 19. When the solution has been treated by the post processor (Fig. 19a.), for both L^2 and H^1 norms, the representing points lie on lines of slope 2. This proves that whereas the estimate (1-13) is optimal for the L^2 norm ($s = 0$), it is not in the H^1 norm and seems to be in fact better than what was expected in our study. This does not affect in any case the comparison with the exact solution (1.14).

Figure 19.b shows the comparison with the crude solution \bar{u}^h , obtained with two fixed nodes, and not treated by the post-processor.

Slopes 2 and 1 are observed and the loss (1. instead of 2. for the H^1 norm) corroborates the final remark of Section 3.32.

When the function f_2 is used (Figure 20), the points show oscillations around two lines of slope 2. (for the L^2 -norm) and 1.65 (for the H^1 -norm) proving that the estimate (3.31) still holds (Fig. 20.a). When the solution has not been treated by the post-processor (Fig. 20.b), the slope 1.65 becomes 1. as expected.

The next series of examples was intended to study the influence of a singularity (at the origin) for a unit square domain regularly partitioned. The data functions are of the form

$$f_\alpha(x,y) = r^\alpha - C, \quad \alpha > -2 \quad (5.16)$$

where C is a real number chosen such that the equilibrium condition (2.2) is satisfied. The family $\{f_\alpha\}$ represents various regularities of data:

$$f_\alpha \in H^s(\Omega) \quad \alpha > s - 1 \quad (5.17)$$

The result shown in Fig. 21.a is a plot of α (regularity) versus σ (rate of convergence of $\| \tilde{u}^h - u^h \|_{s=0,1}$). The pattern of the (α, σ) plot seems to show a linear increase of slope 1 towards the maximum value 2 reached for $f \in L^2$ ($\alpha = -1$) for the L^2 -norm ($s=0$). As far as the H^1 -norm ($s=1$) of the error is concerned, the linear increase of slope 1 reached 1 for $f \in L^2$ but keeps increasing towards 2. This shows that the expected error estimate

$$\| u^h - \tilde{u}^h \|_s \leq C h^k \| f \|_m, \quad s = 0,1 \quad (5.18)$$

where

$$k = 1 + \min(1, m) - s \quad (5.19)$$

is not optimal for $s = 1$ and $m > 0$. The estimate (3.32) remains optimal however.

In conclusion, these numerical results prove that the method is accurate for regular mesh and that no accuracy is lost.

2.5.2.b. Regular mesh of 8- and 9-node elements. Since the beginning of Part II we have not discussed the underintegration of the stiffness matrix of the 8-node elements. It is well known that this matrix is not rank-deficient, and the practice of the underintegration has been widely used with good results when the mesh is regular. Since there is not any spurious mode, the a-posteriori control previously described is not needed.

Unfortunately, the method of proof presented in Section 2.4 cannot be used because this element does not possess the nice tensor product properties on which the method relies. The only hope for a proof of convergence would be to obtain the result as a by-product of a result for the 9-node elements.

As far as this element is concerned (9-node element), we have proved (Section 2.2.3) that the underintegration of this element leads to a rank-deficient matrix; in fact, the procedure described in Section 2.3, for the resolution of the underintegrated problem and the projection of its solution is completely applicable to a mesh of 9-node elements. Thus, Theorem VII is valid and the projection defined in (3.3) can be used to eliminate the spurious mode. As far as the existence of

When a convergence theorem is concerned, one can establish generalizations of (4.16) and (4.17) to 3-node, one-dimensional elements: there exist α_i , β_i , ϕ^i and ψ^i such that

$$\begin{aligned} (v^h, \phi^i)_{0,h} &= \alpha_i (v^h, \phi^i)_0 \\ (v^h, \psi^i)_1 &= \beta_i (v^h, \psi^i)_0 \end{aligned} \quad \forall v^h$$

Unfortunately, the basis functions ϕ^i and ψ^i are different for the L^2 -underintegrated and H^1 -norms and a lemma as Lemma 4.1 cannot be obtained.

However, in this subsection we will show numerical results obtained by use of the projection (3.30) for regular meshes of 9-node elements. Note that two types of control have been tested with similar results: the control only involving the term in $\gamma \cdot \gamma^T$ predicted by Belytschko [3] and the complete control calculated with $s_i s_i^T$, $i = 7, 8, 9$. (See Section 2.2.3). The results obtained with either of them are similar for this operator $(-\Delta)$.

For 8 and 9-node elements, the optimal rates of convergence are given by

$$\|u - u^h\|_s \leq C h^k \|f\|_m \quad s = 0, 1 \quad (5.20)$$

where

$$k = 2 + \min(1, m) - s \quad (5.21)$$

and the best rates of convergence $O(h^{3-s})$ are obtained when $f \in H^1(\Omega)$. The results obtained with functions presenting a singularity line (such as the functions f_1 and f_2 previously defined and others) are pre-

sented in Table 5 (first and second lines). We obtained 1.99 and 1.74 for a discontinuous function ($f \in L^2$), then 2.43 and 1.97 for a continuous, not C^1 , function ($f \in H^1$), 2.95 and 1.95 for a C^1 , not C^2 function ($f \in H^2$) and finally 4 and 3 for a C^∞ data. Therefore, the rates 3 and 2 are reached when f is at least H^2 or equivalent when the solution u is in H^4 . In this case, the convergence rate (5.1) does not seem to be reached.

The second series of data involving the singularity at the origin (5.16) has been tested and results are shown in Fig. 20.b. The pattern of the (α, σ) plot shows linear increases of slope 1, the predicted values 3 and 2 are reached for $f \in H^1(r)$ according to (5.21), but the maximum values 4 and 3 are reached for $f \in H^2(\Omega)$.

2.5.2.c. Irregular mesh of 4- and 9- node elements. Finally, the method has been tested on the quarter unit disk shown in Fig. 22 with

$$f = r^\alpha - \frac{2}{\alpha+2} \quad \alpha > -2$$

The plot (α, σ) is shown in Fig. 23 and we can point out:

- The general pattern is respected (linear increase towards a maximum value)
- The maximum values 4 and 3 (9-node elements) are reduced to values slightly lower than 3 and 2.

2.6 Excitation of Spurious Modes

The previous sections were devoted to the study of the Laplace equations with Neumann boundary conditions. The choice of these boundary conditions is convenient for the analysis of the hourglass instabil-

ities because these modes appear explicitly in the kernel of the underintegrated discrete operator. When Dirichlet conditions are applied on a part of the boundary, even though the kernel of the underintegrated stiffness matrix is not rank-deficient, instabilities may appear.

In this section we would like to study the influence the boundary conditions have on the solution of the underintegrated problem, and obtain results analogous to Theorem VIII. Also we would like to explain how the oscillations may be excited in certain problems. The method of proof is similar to that presented in Section 2.4. For various boundary conditions, we are able to exhibit the exact eigenvalues and eigenfunctions of the various linear and bilinear form involved. The explanation of the excitation of oscillations will result from the comparison of these eigenvalues. The procedure also allows us to study the underintegration of the operator $-\Delta+1$ and the control of resulting spurious modes. Numerical results will illustrate the theory.

2.6.1. The underintegrated problem with Dirichlet or mixed boundary conditions. This section is devoted to a generalization of the results obtained in Section 2.4 to the Laplacian equation with Dirichlet or mixed Dirichlet-Neumann boundary conditions. Only proofs for the Dirichlet case will be given in this section, but their equivalent for the mixed case can be found in Appendix A.

The Dirichlet case is simpler than the Neumann case because the hourglass mode does not belong to the new approximation space defined to handle the boundary condition. Therefore the stiffness matrix is no longer singular and can be normally inverted. In the variational formulation, similar to (3.7), the projection of the data function is not

The values for which $\det D(k)$ vanishes are

$$k_i = \cos \frac{i\pi}{N} \quad 1 \leq i \leq N-1 \quad (6.2)$$

and the corresponding vectors ($D(k_i)v_i = 0$) are

$$v_i = \left\{ \sin \frac{ij\pi}{N} \right\}_{j=1, N-1} \quad (6.3)$$

Let $\phi^i = \phi^i(x)$, $x \in [0, 1]$, denote the piecewise linear function associated with the vector v_i :

$$\phi^i(jh) = \sin \frac{ij\pi}{N} \quad 1 \leq i, j \leq N-1 \quad (6.4)$$

$$\text{span}\{\phi^i\}_{1 \leq i \leq N-1} = V_{1,0}^h \quad (6.5)$$

where

$$V_{1,0}^h = \{v^h \in C^0(0,1), v^h(0) = v^h(1) = 0$$

$$v^h \text{ is linear on } [eh, (e+1)h], 0 \leq e \leq N-1\}$$

From this point, the remainder of the proof goes as in Section 2.4 and the variational problem and its underintegrated formulation can be explicitly solved and the decomposition (4.29), (4.30) and (4.31), (4.32) are obtained for $1 \leq i, j \leq N-1$, and we finally obtain the result for Dirichlet boundary conditions:

THEOREM IX: Let f be a function in $L^2(\Omega)$. Let u be the solution of P :

$$P : \text{Find } u \in H_0^1 / (u, v)_1 = (f, v)_0 \quad \forall v \in H_0^1 \quad (6.6)$$

Let f^h be the L^2 -projection of f onto V^h and let \bar{u}^h be the solution of \bar{P}^h :

$$\bar{P}^h: \text{Find } \bar{u}^h \in V_0^h / (\bar{u}^h, v^h)_{1,h} = (f^h, v^h)_{0,h} \quad \forall v^h \in V_0^h \quad (6.7)$$

Then we have the following error estimate:

$$\|u - \bar{u}^h\|_s \leq C h^{2-s} \|f\|_0 \quad s = 0,1 \quad \square \quad (6.8)$$

This theorem proves that the use of the underintegrated matrix does not affect the rate of convergence of the solution. The method is therefore accurate and efficient.

Various regularities of data have been tested for meshes of 4-, 8-, and 9-node elements, with various boundary conditions. Results are summarized in Tables 5 and 6. They indicate that the optimal rates of convergence for $f \in L^2(\Omega)$ for the 4-node case and $f \in H^2(\Omega)$ for the 8- and 9-node case.

2.6.2. The underintegration of the operator $-\Delta+1$. In this subsection we consider the underintegration of the operator associated with the problem

P_0 : Find $u \in H^1(\Omega)$ such that

$$\begin{cases} -\Delta u + u = f & \text{in } \Omega \\ \frac{\partial u}{\partial n} = 0 & \text{on } \partial\Omega \end{cases} \quad (6.9)$$

The usual variational formulation of P_0 is

P : Find $u \in H^1(\Omega)$ such that

$$(u, v)_1 + (u, v)_0 = (f, v)_0, \quad v \in H^1(\Omega) \quad (6.10)$$

The results of existence, uniqueness of P are well-known and so are

-the ones for its discrete formulation:

P_h : Find $u^h \in V^h$ such that

$$(u^h, v^h)_1 + (u^h, v^h)_0 = (f, v^h)_0, \quad \forall v^h \in V^h \quad (6.11)$$

where V^h is an approximation of $H^1(\Omega)$ using bilinear elements.

The underintegration of $(\cdot, \cdot)_1 + (\cdot, \cdot)_0$ leads to the following under-integrated problem:

\bar{P}_h : Find $\bar{u}^h \in \bar{V}^h$ such that

$$(\bar{u}^h, v^h)_{1,h} + (\bar{u}^h, v^h)_{0,h} = (f, v^h)_{0,h}, \quad \forall v^h \in \bar{V}^h \quad (6.12)$$

where the choice of approximation space

$$\bar{V}^h = V^h/H \quad (6.13)$$

is justified by

$$(v^h, H)_{1,h} + (v^h, H)_{0,h} = 0, \quad \forall v^h \in V^h \quad (6.14)$$

Then, the method of proof used in Section 2.3 allows us to obtain the existence and uniqueness of \bar{u}^h . A projection similar to (3.30) can be obtained by analogy: we have

$$(u^h, H)_1 + (u^h, H)_0 = (f, H)_0 \quad (6.15)$$

we therefore construct the projection as:

$$\hat{u}^h = \pi \bar{u}^h = \bar{u} + \lambda_0 H \quad (6.16)$$

$$(\hat{u}^h, H)_1 + (\hat{u}^h, H)_0 = (f, H)_0 \quad (6.17)$$

This defines uniquely λ_0 as

$$\lambda_0 = \frac{(f, H)_0 - (\bar{u}, H)_1 - (\bar{u}, H)_0}{\|H\|_1^2 + \|H\|_0^2} \quad (6.18)$$

Similarly to what was done in Section 2.5.1, we can use (4.25), (5.6) through (5.8), simplify λ_0 without any loss of accuracy and still use (5.9), (5.10) for the projection

$$\tilde{u}^h = \bar{u}^h - \lambda H \quad (5.9) \text{ repeat}$$

$$\lambda = (\bar{u}^h, H)_1 \|H\|_1^{-2} \quad (5.10) \text{ repeat}$$

The proof of the convergence of \tilde{u}^h towards u is again done by direct calculation of u^h and \tilde{u}^h : the explicit resolution of P_h and $(\bar{P}_h + \pi)$ leads to :

$$u_{i,j} = \frac{1}{1 + \beta_i + \beta_j} f_{ij} \quad 0 \leq i, j \leq N \quad (6.19)$$

and

$$\left\{ \begin{array}{l} \tilde{u}_{i,j} = \frac{\alpha_i \alpha_j}{\alpha_i \alpha_j + \alpha_i \beta_j + \alpha_j \beta_i} f_{ij} \quad \begin{array}{l} 0 \leq i, j \leq N \\ (i, j) \neq (N, N) \end{array} \\ \tilde{u}_{NN} = u_{NN} \end{array} \right. \quad (6.20)$$

These decompositions allow us to obtain $u^h - \tilde{u}^h$ as done in Section (2.4). Provided that $f \in L^2(\Omega)$ we can obtain

$$\|u^h - \tilde{u}^h\|_s \leq C h^{2-s} \|f\|_0 \quad s = 0, 1 \quad (6.21)$$

Once again, the underintegration does not seem to affect the rate of convergence. The result can also be obtained with various boundary

conditions. Numerical results are joined in Tables 5 and 6 and assert the theory.

2.6.3. Excitation of oscillations. The existence of spurious oscillations when underintegration is used is not only encountered when Neumann boundary conditions are applied on the whole boundary. In this subsection, we would like to analyze precisely how modes that oscillate with wavelength of order h are excited when underintegration is used, whereas they are damped when the integration is exact.

For this discussion we consider the unit square

$$\Omega =]0,1[\times]0,1[\quad (6.22)$$

discretized into $N \times N$ elements. We consider the Laplace equation on Ω

$$\left. \begin{aligned} -\Delta u &= f && \text{in } \Omega \\ u &= 0 && \text{on } \partial\Omega \cap \{x = 0\} \\ \frac{\partial u}{\partial n} &= g && \text{on } \partial\Omega / \{x = 0\} \end{aligned} \right\} \quad (6.23)$$

For the first time we include two kinds of load: body forces and surface loads, and we will observe separately the effects of each of them.

The eigenfunctions associated with these particular mixed boundary conditions are constructed as in Section 2.4.

$$\chi^{ij}(x,y) = \psi^i(x)\phi^j(y), \quad \begin{array}{l} 1 < i < N \\ 0 < \underline{j} < \underline{N} \end{array} \quad (6.24)$$

where ϕ^j is defined in (4.15) (associated with Neumann boundary conditions at both ends) and ψ^i is similarly defined (see Appendix A). These functions are defined through sine and cosine functions and therefore oscillate. Among them we will distinguish "smooth" modes with

longer wavelengths (0(1)) from "irregular" modes with shorter wavelengths (0(h)). Smooth (respectively irregular) modes correspond to smaller (respectively longer) values of i or j . Examples of each extreme are shown in Fig. 25 for $N = 10$.

The resolution of the fully integrated problem leads to the search for coefficients u_{ij} such that

$$u^h = \sum_{\substack{1 < i < N \\ 0 < j < N}} u_{ij} \chi^{ij} \quad (6.25)$$

The basis $\{\chi^{ij}\}$ is an eigenbasis for $(\cdot, \cdot)_1$ and therefore we have

$$u_{ij} = A_{ij} \left[(f, \chi^{ij})_0 + (g, \chi^{ij})_{0, \partial\Omega} \right] \quad (6.26)$$

where

$$A_{ij} = \frac{1}{\beta'_i + \beta_j} \quad (6.27)$$

with

$$\left. \begin{aligned} \beta'_i &= \frac{6}{h^2} \frac{1 - \cos\left(\frac{i\pi}{N} - \frac{\pi}{2N}\right)}{2 + \cos\left(\frac{i\pi}{N} - \frac{\pi}{2N}\right)} \\ \beta_j &= \frac{6}{h^2} \frac{1 - \cos\left(\frac{j\pi}{N}\right)}{2 + \cos\left(\frac{j\pi}{N}\right)} \end{aligned} \right\} \quad (6.28)$$

The values A_{ij} have been calculated exactly with these formulae and their values are reported in Table 7.a for $N = 10$. The 20 highest values are in the shaded zone. We clearly can observe that

- i) these values range from the highest value to 1% of this value,
- ii) these values are associated with smooth modes (tensor products of smooth modes).

Contrarily, the eigenvalues of irregular modes are smaller and because of this, these modes will be damped; only smooth modes will contribute in (6.25).

When the underintegration is used, and when g is zero, the solution \tilde{u}^h is

$$\tilde{u}^h = \sum_{\substack{1 < i < N \\ 0 < j < N}} \tilde{u}_{ij} \chi^{ij} \quad (6.29)$$

with

$$\tilde{u}_{ij} = \tilde{A}_{ij}(f, \chi^{ij})_0 \quad (6.30)$$

where

$$\tilde{A}_{ij} = \frac{\alpha'_i \alpha_j}{\alpha_j \beta'_i + \alpha'_i \beta_j} \quad (6.31)$$

and

$$\left. \begin{aligned} \alpha'_i &= \frac{3(1 + \cos(\frac{i\pi}{N} - \frac{\pi}{2N}))}{2(2 + \cos(\frac{i\pi}{N} - \frac{\pi}{2N}))} \\ \alpha_j &= \frac{3(1 + \cos(\frac{j\pi}{N}))}{2(2 + \cos(\frac{j\pi}{N}))} \end{aligned} \right\} \quad (6.32)$$

Again, the values of \tilde{A}_{ij} have been calculated exactly and they are reported in Table 7b. The 20 highest values are in the shaded zone. The comparison between Tables 7a and b shows that these 20 values are approximately the same and they are associated with the same smooth modes. In this case, irregular modes will still be damped, and one can predict that no oscillation will occur.

When a load is only applied on the boundary ($f = 0, g \neq 0$),

\tilde{u}_{ij} is now

$$\tilde{u}_{ij} = \bar{A}_{ij} (g, \chi^{ij})_{0, \partial\Omega} \quad (6.33)$$

where

$$\bar{A}_{ij} = \frac{1}{\alpha_j \beta'_i + \alpha'_i \beta_j} \quad (6.34)$$

Again the values of \bar{A}_{ij} are reported in Table 7.c and the 20 highest values are in the shaded zone. Among these 20 values, three correspond to very irregular modes. In particular, the third value is associated with $\chi^{10,10}$. Therefore, we can predict a strong contribution of irregular modes within the solution \tilde{u}^h , which will show oscillations.

Finally, one could wonder if the calculation of the boundary integral can be calculated such that $(g, \chi^{ij})_{0, \partial\Omega}$ is damped for large i and j . Unfortunately, no precise method has been obtained. In particular, if the load g is a concentrated load at (x_0, y_0) , then

$$(g, \chi^{ij})_{0, \partial\Omega} = \chi^{ij}(x_0, y_0)$$

and this value is not necessarily zero. The procedure, consisting of splitting g between neighboring nodes, seems to give satisfactory results, but is more ad hoc than general.

2.7. The Practice of Underintegration in Linear Elasticity

We devote this section to the discussion and the effects of the underintegration of the linear elasticity operator. Our goal is: 1) to exhibit the kernel of the underintegrated operator; 2) to obtain a post-processor formula similar to (3.30) to control, a-posteriori, the

spurious modes; 3) to indicate how to implement this control, and 4) finally, to show numerical results. This study is entirely qualitative - the basis function obtained in Section 2.4 cannot be used at this point to obtain basis functions for the elasticity operator. However, both 4- and 9-node elements will be discussed.

2.7.1. The kernel of the discrete underintegrated linear elasticity operator. We consider the linear elasticity operator defined by

$$\underset{\approx}{A} = \underset{\approx}{\beta}^T \underset{\approx}{C} \underset{\approx}{\beta} \quad (7.1)$$

where

$$\underset{\approx}{\beta} = \begin{pmatrix} \frac{\partial}{\partial x} & 0 & \frac{\partial}{\partial y} \\ 0 & \frac{\partial}{\partial y} & \frac{\partial}{\partial x} \end{pmatrix}^T \quad (7.2)$$

and $\underset{\approx}{C}$ is a 3x3 symmetric matrix. In the plane strain case we may particularize $\underset{\approx}{C}$:

$$\underset{\approx}{C} = \begin{pmatrix} \lambda+2\mu & \lambda & 0 \\ \lambda & \lambda+2\mu & 0 \\ 0 & 0 & \lambda \end{pmatrix} \quad (7.3)$$

In order to exhibit the spurious modes we consider the operator A associated with Neumann boundary conditions. In that case, the kernel of A consists of the usual 2-dimensional rigid body modes denoted by \underline{t}_x , \underline{t}_y and \underline{r}

$$\text{RBM} = \text{span} \left\{ \underline{t}_x = \begin{pmatrix} 1 \\ 0 \end{pmatrix}; \underline{t}_y = \begin{pmatrix} 0 \\ 1 \end{pmatrix}; \underline{r} = \begin{pmatrix} -y \\ x \end{pmatrix} \right\} \quad (7.4)$$

We consider the problem

P : Find $\underline{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in [H^1(\Omega)]^2/\text{RBM}$ such that

$$\int_{\Omega} \underline{u}^T \underline{\beta}^T C \underline{\beta} \underline{v} \, dx dy = \int_{\Omega} \underline{f}^T \cdot \underline{v} \, dx dy \quad (7.5)$$

where $\underline{f} = (f_1, f_2)^T$ is a force satisfying the equilibrium conditions

$$\underline{f} \in \text{RBM}^T \quad (7.6)$$

or equivalently

$$\left. \begin{aligned} \int_{\Omega} f_1 \, dx dy &= \int_{\Omega} f_2 \, dx dy = 0 \\ \int_{\Omega} (f_1 y - f_2 x) \, dx dy &= 0 \end{aligned} \right\} \quad (7.7)$$

The existence and uniqueness of a solution for P are well known. The construction of finite element approximations of (7.5) involves the calculation of the $(2N \times 2N)$ stiffness matrix K_e for a typical element Ω_e , which is given by the formula

$$K_e = \int_{\Omega} \underline{N}^T \underline{\beta}^T C \underline{\beta} \underline{N} \, dx dy \quad (7.8)$$

where \underline{N} is a vector representing the bilinear ($N=4$) or biquadratic ($N=9$) shape functions in each element Ω_e , $1 \leq e \leq E$. In computational applications K_e is evaluated using an integration rule:

$$K_e = \sum_{\alpha=1}^L w_{\alpha} \underline{B}^{\alpha T} C \underline{B}^{\alpha} \quad (7.9)$$

where, similar to (2.36),

$$\underline{B}^{\alpha} = \begin{pmatrix} b_1^{\alpha} & 0 & b_2^{\alpha} \\ 0 & b_2^{\alpha} & b_1^{\alpha} \end{pmatrix}^T \quad (7.10)$$

0-2

and w_α is the weight at the integration point α . Simple rank considerations allow us to predict the rank of $K_{\approx e}$. Indeed, since

$$\left. \begin{aligned} \text{rk}(A \cdot B) &\leq \max(\text{rk } A, \text{rk } B) \\ \text{and } \text{rk}(A + B) &\leq \text{rk } A + \text{rk } B \end{aligned} \right\} \quad (7.11)$$

we have

$$\text{rk } K_{\approx e} \leq 3L \quad (7.12)$$

When the full integration is used, (7.12) does not tell us anything, but we know that $K_{\approx e}^{\text{full}}$ has the correct kernel containing only rigid body modes. However, when underintegration ($L=1$) is used on 4-node elements, we have

$$\text{rk } K_{\approx e} \leq 3 \quad (7.13)$$

Therefore, the 8x8 matrix K_e possesses at least two spurious modes. In fact, two is the exact number. Similarly, when underintegration ($L=4$) is used and 8- or 9-node elements, we have

$$\text{rk } K_{\approx e} \leq 12 \quad (7.14)$$

This inequality predicts one spurious mode for the 16x16 matrix associated with 8-node elements, but when the procedure is repeated for two neighboring 8-node elements, the spurious modes can no longer exist in the global matrix [46]. We can also interpret this elimination of the spurious mode by noticing that neighboring element cannot share the mode [13].

As far as 9-node elements are concerned, the inequality (7.14) tells us that the 18x18 stiffness matrix has at least three spurious modes. In fact, there are exactly three such modes and they can be

shared by adjacent elements. Next the modes will be explicitly described.

2.7.1.a. The spurious modes for 4-node elements. Let $\underline{H}_{\underline{x}}$ and $\underline{H}_{\underline{y}}$ be the two hourglass vectors defined as

$$\underline{H}_{\underline{x}} = \begin{pmatrix} h \\ 0 \end{pmatrix} \quad \underline{H}_{\underline{y}} = \begin{pmatrix} 0 \\ h \end{pmatrix} \quad (7.15)$$

where h^* is the hourglass nodal displacement defined in (2.14). Then when $L=1$, we obtain from (2.18) and (2.20)

$$\underline{B}_{\underline{x}}^1 \cdot \underline{H}_{\underline{x}} = \begin{pmatrix} \underline{b}_1^T \cdot h \\ 0 \\ \underline{b}_2^T \cdot h \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

and similarly

$$\underline{B}_{\underline{y}}^1 \cdot \underline{H}_{\underline{y}} = 0$$

Therefore,

$$\underline{K}_{\underline{e}}^{(1)} \cdot \underline{H}_{\underline{x}} = \underline{K}_{\underline{e}}^{(1)} \cdot \underline{H}_{\underline{y}} = 0 \quad (7.16)$$

These element displacements can be put together to obtain two global spurious modes, also denoted by $\underline{H}_{\underline{x}}$ and $\underline{H}_{\underline{y}}$ and we have :

$$\text{Ker } \underline{K}^{\text{under}} = \{ \text{span } \underline{t}_{\underline{x}}, \underline{t}_{\underline{y}}, \underline{r}, \underline{H}_{\underline{x}}, \underline{H}_{\underline{y}} \} \quad (7.17)$$

This defines entirely the kernel of the underintegrated matrix and the spurious modes for 4-node elements.

Remark: In problems where symmetry is used for simplifications, the kernel of $\underline{K}^{\text{under}}$ must respect the symmetry. If one axis of sym-

* In this section, nodal values and associated functions will be denoted by the same letter, the underlining "" differentiating them. The nodal values are expressed component by component.

metry (say, the x-axis) exists, then

$$\text{Ker } \underset{\sim}{K}^{\text{under}} = \text{span} \{ \underset{\sim}{t}_x, \underset{\sim}{H}_x \}$$

If the problem has two axes of symmetry (x- and y-axis)

$$\text{Ker } \underset{\sim}{K}^{\text{under}} = \{0\}$$

The spurious modes are eliminated by the symmetry conditions.

2.7.1.b. The spurious modes for 9-node elements. Let $\underset{\sim}{H}_x$ and $\underset{\sim}{H}_y$ be the two vectors defined as

$$\underset{\sim}{H}_x = \begin{pmatrix} h \\ 0 \end{pmatrix} \quad \underset{\sim}{H}_y = \begin{pmatrix} 0 \\ h \end{pmatrix} \quad (7.18)$$

where h is the spurious mode of 9-node elements defined in (2.37).

Using (2.36), (2.39) and (7.9), we easily get

$$\underset{\sim}{K}_e^{(4)} \cdot \underset{\sim}{H}_x = \underset{\sim}{K}_e^{(4)} \cdot \underset{\sim}{H}_y = 0 \quad (7.19)$$

Therefore, $\underset{\sim}{H}_x$ and $\underset{\sim}{H}_y$ are two out of the three spurious modes of $\underset{\sim}{K}_e^{(4)}$. We remark that the pattern (2.37) defining them does not depend on the geometry of the mesh. As far as the third spurious mode denoted by

$$\underset{\sim}{W} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} \quad (7.20)$$

is concerned, one can show that the equations defining it are

$$\underset{\sim}{b}_1^{\alpha T} \cdot \underset{\sim}{w}_1 = \underset{\sim}{b}_2^{\alpha T} \cdot \underset{\sim}{w}_2 = \underset{\sim}{b}_1^{\alpha T} \cdot \underset{\sim}{w}_2 + \underset{\sim}{b}_2^{\alpha T} \cdot \underset{\sim}{w}_1 = 0 ; \alpha=1,4 \quad (7.21)$$

or equivalently :

$$\underset{\sim}{y}^T \underset{\sim}{A}^{\alpha} \underset{\sim}{w}_1 = 0 \quad \Bigg\}$$

$$\left. \begin{aligned} \underline{x}^T \underline{A}^\alpha \underline{w}_2 &= 0 \\ \underline{x}^T \underline{A}^\alpha \underline{w}_1 &= \underline{y}^T \underline{A}^\alpha \underline{w}_2 \end{aligned} \right\} \alpha = 1, 4 \quad (7.22)$$

Note that for this system of 12 equations, we have 18 unknowns. If we add 5 orthogonality equations between \underline{W} and \underline{t}_x , \underline{t}_y , \underline{r} , \underline{H}_x and \underline{H}_y , the system will define only one \underline{W} (up to within a multiplicative factor). For a general geometry of Ω_e , one cannot exhibit an explicit form for \underline{W} ; however, when Ω_e is a quadrilateral, and when \underline{x} and \underline{y} are of the form

$$\begin{aligned} \underline{x} = (x_1, x_2, x_3, x_4, \frac{1}{2}(x_1 + x_2), \frac{1}{2}(x_2 + x_3), \\ \frac{1}{2}(x_3 + x_4), \frac{1}{2}(x_4 + x_1), \frac{1}{2}(x_1 + x_2 + x_3 + x_4)) \end{aligned} \quad (7.23)$$

we can prove that one candidate for \underline{W} can be written as

$$\left. \begin{aligned} \underline{w}_1 &= \underline{T} \underline{y}' \\ \underline{w}_2 &= -\underline{T} \underline{x}' \end{aligned} \right\} \quad (7.24)$$

where

$$\underline{T} = \begin{pmatrix} 4 & 2 & 0 & 2 & -1 & 0 & 0 & -1 & 0 \\ -2 & -4 & -2 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 2 & 4 & 2 & 0 & -1 & -1 & 0 & 0 \\ -2 & 0 & -2 & -4 & 0 & 0 & 1 & 1 & 0 \end{pmatrix}^T \quad (7.25)$$

and \underline{x}' , \underline{y}' are the vectors constructed with the first four components of \underline{x} and \underline{y} . An example on \underline{W} for a geometry satisfying (7.23) is shown in Figure 25 and can be constructed as follows:

- i) the displacement of a mid-side node is normal to the side, alternatively inwardly and outwardly oriented, with magnitude

proportional to the length of the side.

ii) the displacement of a corner is obtained by multiplication of -2 of the sum of the two displacements of the closest mid-side nodes.

iii) the displacement of the centroid is zero.

On a square, the pattern of \tilde{W} is well-known:

$$\tilde{W} = \begin{pmatrix} -2, 2, 2, -2, 0, -1, 0, 1, 0 \\ 2, 2, -2, -2, -1, 0, 1, 0, 0 \end{pmatrix} \quad (7.26)$$

Contrary to 8-node elements, and because of the presence of H_x and H_y , this mode can "propagate" from one element to another. For example, on a square mesh, if the nodal displacement vector is \tilde{W} given by (7.26) on an element Ω_0 , then the displacement vectors $\tilde{W} + 3H_x + t_x$ and $\tilde{W} - 3H_y - t_y$ on the elements to the right of Ω_0 and above Ω_0 allow us to construct a continuous global displacement also denoted \tilde{W} , on the mesh as shown in Figure 26.

We finally have

$$\text{Ker } K^{\text{under}} = \text{span} \{t_x, t_y, r, H_x, H_y, W\} \quad (7.27)$$

Remark: Similar to what we have with 4-node elements, the existence of one axis of symmetry (say, the x-axis) reduces the kernel of the underintegrated stiffness matrix:

$$\text{Ker } K^{\text{under}} = \text{span} \{t_x, H_1, H_2 + H_3\} \quad (7.28)$$

where

$$\left. \begin{aligned} H_1 &= 3/2(H_x - t_x) \\ H_2 &= -3/2(H_y - t_y) \end{aligned} \right\}$$

$$\tilde{H}_3 = \tilde{W} + 2(t_x - t_y) \quad \Bigg\} \quad (7.29)$$

have been chosen such that the displacements of these modes are zero at the intersection of both axes for a square mesh. Contrary to the 4-node case, we still have a spurious mode when two axes of symmetry exist:

$$\text{Ker } K^{\text{under}} = \text{span} \{ \tilde{H}_1 + \tilde{H}_2 + \tilde{H}_3 \} \quad (7.30)$$

This mode is shown in Figure 27.

It is also important to point out that whereas the pattern of the spurious modes \tilde{H}_x and \tilde{H}_y are independent of both the geometry and the element, the mode \tilde{W} depends upon both of them. Moreover, we can see by construction on a square mesh that the amplitude varies strongly when we consider successive elements. In fact, the pattern we may observe is a succession of pattern \tilde{H}_x and \tilde{H}_y with increasing amplitude.

2.7.2. The a-posteriori control in linear elasticity. In this subsection we wish to generalize (3.30) with regard to the discrete operators, using various kernels discussed in the previous subsection. We consider the general case where

$$\text{Ker } \tilde{K}^{\text{under}} = \text{RBM} \oplus \text{span} \{ H_i, i = 1, I \} \quad (7.31)$$

where I may have the values 1, 2 or 3. We recall that for $I = 1$, we obtained a control formula similar to

$$\tilde{u}^h = \tilde{u}^h - \frac{a(\tilde{u}^h, H_1)}{a(H_1, H_1)} H_1 \quad (7.32)$$

where the bilinear form $a(\cdot, \cdot)$ was obtained in the variational formulation of the initial problem. This projection satisfies:

$$a(\tilde{u}^h, H_1) = 0 \quad (7.33)$$

or, in other words, \tilde{u}^h is orthogonal to the spurious mode. We generalize this property to the elasticity problem by supposing the projection to be orthogonal to all the spurious modes. Therefore the control will consist of looking for I constants λ_i ($i = 1, I$) such that

$$\begin{cases} \tilde{u}^h = \bar{u}^h - \sum_{i=1, I} \lambda_i H_i \\ a(\tilde{u}^h, H_i) = 0 \quad \text{for } i=1, 3 \end{cases} \quad (7.34)$$

This leads to the system of I equations with I unknowns :

Find λ_i , $i = 1, I$ such that

$$\sum_{j=1, I} \lambda_j a(H_j, H_i) = a(\bar{u}^h, H_i), \quad i=1, I \quad (7.35)$$

The computations involved in the control are computations of products of \bar{u}^h and the spurious modes by themselves. The implementation of these computations are to be discussed in the next section.

2.7.3. Implementation of the spurious modes control. For the computation of the coefficients in (7.35) we again use the decomposition

$$\tilde{K}^{\text{full}} = \tilde{K}^{\text{under}} + \tilde{K}^{\text{sp}} \quad (7.36)$$

where \tilde{K}^{under} satisfies

$$\tilde{K}^{\text{under}} \cdot H_i = 0 \quad (7.37)$$

then

$$a(\tilde{u}^h, H_i) = \sum_{e=1, E} \bar{U}^T \cdot \tilde{K}^{\text{sp}} \cdot H_i \quad (7.39)$$

The expressions used for \underline{K} are next given for 4- or 9-node elements.

2.7.3.a. Control for 4-node elements. For the operator defined in (7.1) and (7.2) with

$$\underline{C} = \begin{pmatrix} C_{11} & C_{12} & C_{13} \\ C_{21} & C_{22} & C_{23} \\ C_{31} & C_{32} & C_{33} \end{pmatrix} \quad (7.40)$$

we have the *exact* decomposition for *any* geometry of Ω_e :

$$\underline{K}_{\approx e}^{\text{exact}} = \underline{K}_{\approx}^{\text{under}} + \begin{pmatrix} \alpha_{11} \underline{\gamma} \cdot \underline{\gamma}^T & \alpha_{12} \underline{\gamma} \cdot \underline{\gamma}^T \\ \alpha_{21} \underline{\gamma} \cdot \underline{\gamma}^T & \alpha_{22} \underline{\gamma} \cdot \underline{\gamma}^T \end{pmatrix}$$

where

$$\begin{bmatrix} \alpha \\ \approx \end{bmatrix} = \begin{pmatrix} C_{11} \bar{\epsilon}_{xx} + (C_{13} + C_{31}) \bar{\epsilon}_{xy} + C_{33} \bar{\epsilon}_{yy}; C_{13} \bar{\epsilon}_{xx} + (C_{12} + C_{33}) \bar{\epsilon}_{xy} + C_{32} \bar{\epsilon}_{yy} \\ C_{31} \bar{\epsilon}_{xx} + (C_{21} + C_{33}) \bar{\epsilon}_{xy} + C_{23} \bar{\epsilon}_{yy}; C_{33} \bar{\epsilon}_{xx} + (C_{32} + C_{23}) \bar{\epsilon}_{xy} + C_{22} \bar{\epsilon}_{yy} \end{pmatrix} \quad (7.42)$$

The vector $\underline{\gamma}$ and the $\bar{\epsilon}$'s are defined in Section 2 ((2.41) and (2.50)). For practical use, the expressions (2.51) are used for $\bar{\epsilon}$.

For linear isotropic linear material, \underline{C} is given by (7.3) and

$$\begin{bmatrix} \alpha \\ \approx \end{bmatrix} = \begin{pmatrix} (\lambda + 2\mu) \bar{\epsilon}_{xx} + \mu \bar{\epsilon}_{yy} & \mu \bar{\epsilon}_{xy} \\ \mu \bar{\epsilon}_{xy} & \mu \bar{\epsilon}_{xx} + (\lambda + 2\mu) \bar{\epsilon}_{yy} \end{pmatrix} \quad (7.43)$$

This expression of $\begin{bmatrix} \alpha \\ \approx \end{bmatrix}$ can be compared to the general strain-stress relationship:

$$\begin{bmatrix} \sigma \\ \approx \end{bmatrix} = \begin{pmatrix} (\lambda+2\mu)\epsilon_x + \mu\epsilon_y & \mu\epsilon_{xy} \\ \mu\epsilon_{xy} & \mu\epsilon_x + (\lambda+2\mu)\epsilon_y \end{pmatrix} \quad (7.44)$$

An algorithm similar to the one presented in Section 2.5.1 can be constructed. It involves the computation of $\underline{\gamma}$, ϵ and $\underline{\alpha}$, then the computation of $a(H_i, H_j)$ and $a(\underline{u}^h, H_i)$, and finally the coefficients λ_i are obtained by resolution of a NxN system, N measuring the rank deficiency of K^{under} (N=1 or 2).

2.7.4.b. Control for 9-node elements. In this subsection, devoted to 9-node elements, we would like first to show why the results obtained by Belytschko are not sufficient to obtain a generalization of the Linear Elasticity Problem, and then to propose an implementation of the control that leads to a stable solution converging to the exact solution with the optimal rate of convergence. However, for 9-node elements, we have not yet been able to obtain a computationally easy way to exhibit the third spurious mode, and the proposed results are only applicable to regular discretizations of a domain.

As far as the stabilization method proposed in [3] is concerned, algebra similar to that in Subsection 2.7.3.a leads to (7.41) where $\underline{\gamma}$ was defined in 2.41. But, whereas the stabilization matrix constructed with the submatrix $\underline{\gamma} \cdot \underline{\gamma}^T$ eliminates \underline{H}_x and \underline{H}_y from the kernel of the stiffness matrix, it does not take \underline{W} into account. Indeed, we have

$$\begin{pmatrix} \alpha_{11} \underline{\gamma} \cdot \underline{\gamma}^T & \alpha_{21} \underline{\gamma} \cdot \underline{\gamma}^T \\ \alpha_{12} \underline{\gamma} \cdot \underline{\gamma}^T & \alpha_{11} \underline{\gamma} \cdot \underline{\gamma}^T \end{pmatrix} \cdot \underline{W} = \underline{0}$$

Therefore this procedure cannot be used to control \tilde{W} .

In order to obtain an accurate control, we have to consider a generalization of (2.43). Now we have

$$\begin{pmatrix} \alpha_{11} s_i \cdot s_i & \alpha_{21} s_i s_i \\ \alpha_{12} s_i \cdot s_i & \alpha_{22} s_i s_i \end{pmatrix} \cdot H_j \neq 0$$

for $7 < i < 9$
 $1 < j < 3$

where the vectors s_1 are defined in Section 2.2.3. Finally, using (2.43) and (2.52), we have

$$\begin{aligned} K_{\approx e}^{(9)} = K_{\approx e}^{(4)} + \frac{4\Omega}{135} e & \begin{bmatrix} (C_{11}+C_{33})s_9s_9^T & (C_{13}+C_{32})s_9s_9^T \\ (C_{31}+C_{23})s_9s_9^T & (C_{22}+C_{33})s_9s_9^T \end{bmatrix} \\ + \frac{\Omega}{45} e & \left(\begin{bmatrix} C_{11}s_7s_7^T & C_{13}s_7s_7^T \\ C_{31}s_7s_7^T & C_{33}s_7s_7^T \end{bmatrix} + \begin{bmatrix} C_{33}s_8s_8^T & C_{32}s_8s_8^T \\ C_{23}s_8s_8^T & C_{22}s_8s_8^T \end{bmatrix} \right) \end{aligned} \quad (7.45)$$

Similarly, for the 4-node case, the algorithm for the computations of the coefficients in (7.35) has been obtained and implemented. Numerical results agree with our presumptions concerning a " $\tilde{\gamma} \cdot \tilde{\gamma}^T$ "-type of control and incline in favor of the decomposition (7.45). On a square domain discretized with $N \times N$ elements, we have calculated and compared the solutions obtained with full (exact) and underintegration for various boundary and symmetry conditions. The rates of convergence were calculated by comparing the error norms ($s=0$: L^2 /RBM norm; $s=1$: energy norm) obtained with $N=5,6$ and 7 . We consistently got the rate $O(h^{2-s})$ using a $\tilde{\gamma} \cdot \tilde{\gamma}^T$ decomposition and $O(h^{3-s})$ with (7.45) for homogeneous

materials under the action of gravity ($f \in C^\infty$); the order 3-s being optimal, we may conclude that the method presented below is accurate. It is also efficient: for one second taken for the fully integrated stiffness matrix, only .61 are taken when the underintegration is used and only .05 seconds are taken for the control.

Remark: The analysis of the excitation of the spurious modes carried for the simple Laplace equation cannot be done for the elasticity because we are not able to exhibit eigenbasis of the discrete operator for neither 4- nor 9-node elements. Numerical computations [7] seem to indicate that the same phenomenon occurs: several "irregular" modes appear within the smooth, high wavelength modes, and are therefore excited. The shape of these modes and their mathematical knowledge would allow their elimination or damping.

2.8 Conclusions and Further Research

The underintegration seems to be a very attractive way to obtain more efficient computations in solid or fluid mechanics. The spurious modes this practice introduces and the precise way they are excited have long remained unstudied. Several authors previously mentioned proposed several interpretations based on the intuition. We have here tried to study this phenomenon from a rigorous mathematical point of view, and we have precisely answered all the questions concerning one simple problem. Unfortunately, the algebra involved in more sophisticated problems (9-node elements, linear elasticity) does not allow us such a complete study. We would like to indicate that several generalizations of our results will help in the control of the spurious modes: i) a discrete

eigenbasis of the linear elasticity would allow an interpretation of the excitation of spurious modes, and hence a way of damping them;

ii) an accurate control formula for 9-node elements would help in a-priori as well as a-posteriori control of the widely known spurious mode \underline{W} . This could also help in preventing bad behavior of 8-node elements in certain geometries.

APPENDIX A

As far as mixed boundary conditions are concerned, we suppose that a Dirichlet boundary condition is applied at 0 and a Neumann boundary condition at 1. For the interval $[0,1]$, we consider the $N \times N$ matrix

$$D(k) = \left\{ \begin{array}{ccc} 2k & -1 & 0 \\ -1 & & \\ & & 2k & -1 \\ 0 & & -1 & k \end{array} \right\} \quad (A.1)$$

The values for which $\det K(k)$ vanishes are:

$$k_i = \cos \left(\frac{-\pi}{2N} + \frac{i\pi}{N} \right), \quad 1 \leq i \leq N \quad (A.2)$$

and the corresponding vectors $(D(k_i)v_i = 0)$ are :

$$v_i = \left\{ \sin \frac{ij\pi}{2N} \right\} \quad 1 \leq j \leq N \quad (A.3)$$

The corresponding approximation space $V_{1, \frac{1}{2}}^h$ with basis $\{\phi^j\}$ is constructed as in §25] or in Section 2.4. Then, depending upon the sides where the various boundary conditions (D or N) are applied, tensor product of V_1^h , $V_{1,0}^h$ or $V_{1, \frac{1}{2}}^h$ are to be considered. The results of Theorem II hold for the Mixed Problem.

REFERENCES

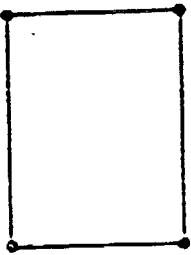
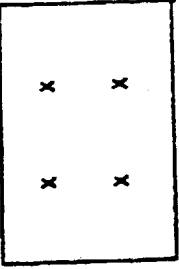
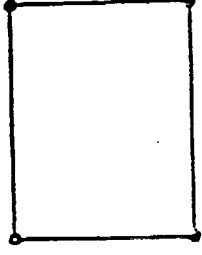
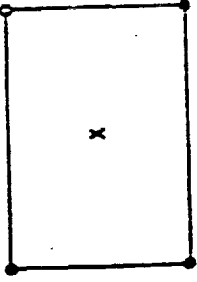
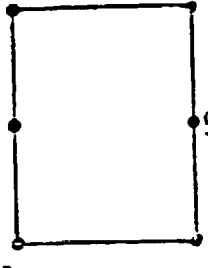
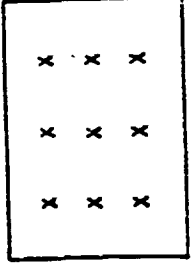
1. BABUSKA, I., "The Finite Element Method with Lagrange Multipliers", Num. Math., Vol.20, 1973.
2. BELYTSCHKO, T. and KENNEDY, J.M., "Computer Models for Sub-assembly Simulation", Nuclear Engineering & Design, Vol.49, pp.17-38, July 1978.
3. BELYTSCHKO, ONG, J., S.-J. and LIU, W.K., "A Consistent Control of Spurious Singular Modes in the 9-Node Lagrange Element for the Laplace and Mindlin Plate Equation", Comp. Meth. in Appl. Mech. & Engrg., Vol.44, pp.269-295, 1984.
4. BELYTSCHKO, T. and TSAY, C.S., "A Stabilization Procedure for the Quadrilateral Plate Element with One-Point Quadrature", Comp. Meth. in Appl. Mech. & Engrg., Vol.19, pp.409-419, 1983.
5. BELYTSCHKO, T., TSAY, C.S. and LIU, W.K., "A Stabilization Matrix for the Bilinear Mindlin Plate Element", Comp. Meth. in Appl. Mech. & Engrg., Vol.29, pp.313-327, 1981.
6. BERCOVIER, M., "Perturbation of Mixed Variational Problems - Applications to Mixed Finite Element Methods", R.A.I.R.O., Vol.12., No.3, 1978.
7. BICANIC, N. and HINTON, E., "Spurious Modes in Two-Dimensional Isoparametric Elements", Int. J. Num. Methods in Engrg., Vol.14, pp.1545-1557, 1979.
8. BREZZI, R., "On the Existence, Uniqueness and Approximation of Saddle-Point Problems Arising from Lagrangian Multipliers", R.A.I.R.O., Numerical Analysis 8, 1974.
9. CAREY, G.F. and KRISHNAN, R., "Penalty Approximation of Stokes Flow, Part I: Stability Analysis, Part II: Error Estimates and Numerical Results", TICOM Report, 82-5, June 1982.
10. CIARLET, P., Numerical Analysis of the Finite Element Method for Elliptic Boundary Problems, North Holland, Amsterdam, 1978.
11. CIARLET, P. and RAVIART, P.A., "The Combined Effect of Curved Boundaries and Numerical Integration in Isoparametric Finite Element Methods", The Mathematical Foundations of the Finite Element Method with Application to Partial Differential Equations, A.K.Aziz(editor), Academic Press, New-York, pp.409-474, 1972.

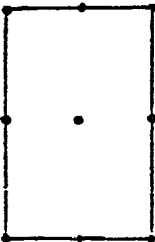
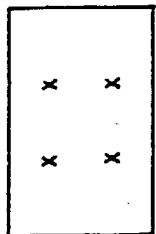
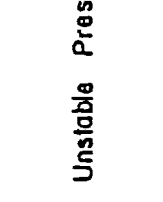
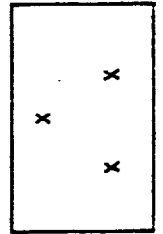

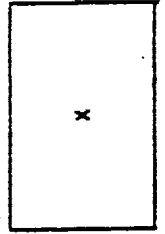
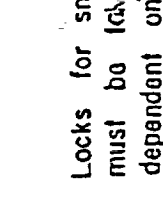
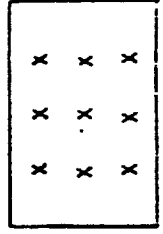
12. COOK, R.D., Concepts and Applications of Finite Element Analysis, New-York, Willey, 1981.
13. COOK, R.D. and ZHAO-HUA, F., "Control of Spurious Modes in the Nine-Node Quadrilateral Element", Int. J. Num. Methods in Engrg, Vol. 18, pp.1576-1580, 1982.
14. ENGLEMAN, M., FIDAP User's Manual, Boulder, Colorado, August 1981.
15. ENGLEMAN, M. AND SANI, R., The Mathematics of Finite Elements, Edited by J.R. Whiteman, Academic Press, Ltd., London, 1982.
16. FALK, R.S., "An Analysis of the Penalty Method and Extrapolation for the Stationary Stokes Problem", Advances in Computer Methods for Partial Differential Equations, Edited by R. Vichnevetsky, AICA Publication, pp.66-69, 1975.
17. FLANAGAN, D.P. and BELYTSCHKO, T., "A Uniform Strait Hexahedron and Quadrilateral with Orthogonal Hourglass Control", Int. J. Num. Methods in Engrg., Vol.17, pp.676-706, 1981.
18. FORTIN, M., "A Analysis of Convergence of Mixed Finite Element Methods", R.A.I.R.O., Vol.II, No.4, 1977.
19. FORTIN, M., "Old and New Finite Elements for Incompressible Flows", Int. J. Num. Methods in Fluids, Vol.1, pp.347-364, 1979.
20. GIRAULT, V., "Theory of a Finite Difference Method on Irregular Networks", SIAM J. Numer. Anal., Vol.II, pp.409-474, 1974.
21. GIRAULT, V., "Nonelliptic ppxrimation of a Classof Partial Differential Equations with Neumann Boundary Conditions", Mathematics of Computation, Vol.30, No.133, pp.68-91, January 1976.
22. GIRAULT, V. and RAVIART, P.A., Lecture Notes in Mathematics 749, Finite Element Approximation of the Navier-Stokes Equations, Springer-Verlag, Berlin, 1979.
23. HUGHES, T.J.R., "Equivalence of Finite Elements for Nearly Incompressible Elasticity", J. Appl.Mech., March 1977.
24. JACQUOTTE, O.-P., "Stability, Accuracy and Efficiency of Some Underintergrated Methods in Finite Element Computations", Comp. Meth. in Appl. Mech. & Engrg., (to appear).
25. JACQUOTTE, O.-P. and ODEN, J.T., "Analysis of Hourglass Instabilities and Control in Underintegrated Finite Element Methods", Comp. Meth. in Appl. Mech. & Engrg., Vol.44, pp.339-363, 1984.


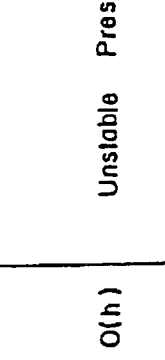
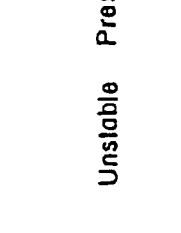
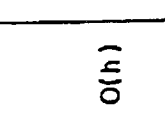

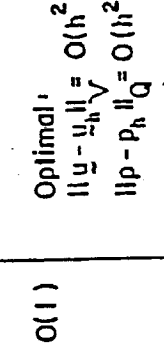
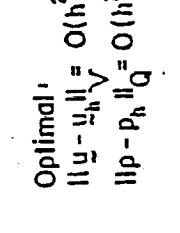
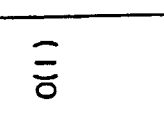
26. JACQUOTTE, O.-P. and ODEN, J.T., "Analysis and Treatment of Hourglass Instabilities in Underintegrated Finite Element Methods", Innovative Methods in Nonlinear Computational Mechanics, Edited by T. Belytschko, K.C. Park and W.K. Liu, Pineridge Press, Swansea, 1984 [presented at the ASEM Winter Annual Meeting, New Orleans, Louisiana, December 9-14, 1984)
27. JACQUOTTE, O.-P., ODEN, J.T. and BECKER, E.B., "Numerical Control of the Hourglass Instabilities", Computers and Structures, (to appear).
28. KOSLOFF, D. and FRAZIER, G.A., "Treatment of Hourglass Patterns in Low Order Finite Element Codes", Int. J. Num. Anal. Meth. in Geomech., Vol.2, pp.57-72, 1978.
29. LADYSZHENSKAYA, O.A., The Mathematical Theory of Viscous Incompressible Flows, Gordon Breach, New York, 1969.
30. MALKUS, D.S., "Finite Element Analysis of Incompressible Solids", Dissertation, Department of Mathematics, Boston University, 1972.
31. MALKUS, D.S., "A Finite Element Displacement Model Valid for Any Value of the Compressibility", International Journal of Solids and Structures, Vol.12, pp.731-738, 1976.
32. MALKUS, D.S., "Finite Elements with Penalties for Incompressible Elasticity: A Progress Report", Illinois Inst. of Technology, Chicago, 1981.
33. MALKUS, D.S. AND HUGHES, T.J.R., "Mixed Finite Element Methods - Reduced and Selective Integration Technique - A Unification of Concepts", Comp. Meth. in Appl. Mech. & Engrg, Vol.15, pp.63-81, 1978.
34. ODEN, J.T., "R.I.P. Methods for Stokesian Flows", Finite Elements in Fluids, Vol.IV, Edited by Gallagher et al., John Wiley and Sons, Ltd, London, 1982 [presented at the Third Symposium of Finite Elements in Flow Problems, Bnaff, Canada, June 1980].
35. ODEN, J.T., "Penalty Methods for Constrained Problems in Nonlinear Elasticity", Proceedings, IUTAM Symposium on Finite Elasticity, Edited by D.E. Carlson and R.T. Shield, (Lehigh 1980), Martenus Nijhoff, The Hague, pp.281-300, 1982.
36. ODEN, J.T., "Penalty Method and Reduced Integration for the Analysis of Fluids", Proceedings, Symposium on Penalty Finite Element Methods, ASME Winter Annual Meeting, Phoenix, Arizona, November 14-19, 1982.
37. ODEN, J.T. and JACQUOTTE, O.-P., "Stable and Unstable RIP/Perturbed Lagrangian Methods for Two-Dimensional Viscous Flow Problems", Finite Elements in Fluids, Vol.V, ed. R.H. Gallagher et al, John Wiley and Sons, Ltd., London, 1982.

38. ODEN, J.T. and JACQUOTTE, O.-P., "Analysis of Hourglass Instabilities and Control in Underintegrated Finite Element Methods," Computer Methods in Applied Mechanics and Engineering, 1984, Vol. 44, pp. 339-363.
39. ODEN, J.T., "Penalty Method and Reduced Integration for the Analysis of Fluids," Proceedings, Symposium on Penalty Finite Element Methods in Mechanics, ASME Winter Annual Meeting, November 14-19, 1982, Phoenix, AZ.
40. ODEN, J.T., and JACQUOTTE, O.-P., "Stability of Some Mixed Finite Element Methods for Stokesian Flows," Computer Methods in Applied Mechanics and Engineering, 1984, Vol. 43, No. 2, pp. 231-248.
41. ODEN, J.T., KIKUCHI, N. and SONG, Y.J., "Convergence of Modified Penalty Methods and Smoothing Schemes of Pressure for Stokes Flow Problems," Finite Elements in Fluid Dynamics, Vol. V, John Wiley & Sons, Ltd., London, 1984.
42. ODEN, J.T., and JACQUOTTE, O.-P., "Stable and Unstable RIP/Perturbed Lagrangian Methods for Two-Dimensional Viscous Flow Problems," Finite Elements in Fluid Dynamics, Vol. V, John Wiley & Sons, Ltd., London, 1984, pp. 127-146.
43. ODEN, J.T., "Stability and Convergence of Underintegrated Finite Element Approximations," Presented at the NASA-LeRC/Industry/University Workshop on Nonlinear Analysis for Engine Structures, April 19-20, 1983, in Cleveland, OH.
44. ODEN, J.T., and JACQUOTTE, O.-P., "Convergence and Stability of Underintegrated Finite Element Methods," to appear in Proceedings, ASCE/ASME Mechanics Meeting, June 24-26, 1985, Albuquerque, NM.
45. ODEN, J.T., ENDO, T., BECKER, E. and MILLER, T, "A Numerical Analysis of Contact and Limit-Point Behavior in a Class of Problems of Finite Elastic Deformation," Computers and Structures, 1984, Vol. 18, No. 5, pp. 899-910.
46. ODEN, J.T., and JACQUOTTE, O.-P., "Analysis and Treatment of Hourglass Instabilities in Underintegrated Finite Element Methods," Proceedings, Symposium on Innovative Methods for Nonlinear Mechanics, ASME Winter Annual Meeting, December 12-15, 1984, New Orleans, LA.

TABLE 1

| Velocity Approx. V^h | Quadrature Rule (Pressure Approx. Q^h) | α_h | Rate of Convergence |
|---|--|--------------------------|--|
| <p>1</p>  <p style="text-align: center;">q_1</p> |  <p style="text-align: center;">Q_2</p> | <p>$O(1)$</p> | <p>Locks for small ϵ, ϵ must be taken as dependent on h.</p> |
| <p>2</p>  <p style="text-align: center;">q_1</p> |  <p style="text-align: center;">Q_0</p> | <p>$O(h)$</p> | <p>Unstable Pressure</p> |
| <p>3</p>  <p style="text-align: center;">IB</p> |  <p style="text-align: center;">Q_2</p> | <p>$O(1)$</p> | <p>Locks for small ϵ, ϵ must be taken as dependent on h.</p> |

| \mathcal{V}_h | Q^h | α_h | Rate of Convergence |
|--|--|------------|---|
| 4  18 |  Q_1 | $O(h)$ | Unstable Pressure |
| 5  18 |  P_1 | $O(h)$ | Unstable Pressure |
| 6  18 |  Q_0 | $O(1)$ | Suboptimal ($O(h)$) in velocity error in energy norm |
| 7  18 |  Q_2 | $O(1)$ | Locks for small ϵ ; ϵ must be taken as dependent on h |

| v^h | q^h | α_h | Rate of Convergence |
|---|--|---|--|
| <p>8 </p> <p>9 </p> | <p></p> <p></p> | <p>$O(h)$</p> <p>$O(1)$</p> | <p>Unstable Pressure</p> <p>Optimal: $\ \bar{u} - u_h \ = O(h^2)$ $\ p - p_h \ _{Q^2} = O(h^2)$</p> |
| <p>10 </p> <p>11 </p> | <p></p> <p></p> | <p>$O(1)$</p> <p>$O(h)$</p> | <p>Suboptimal ($O(h)$) in velocity error in energy norm</p> <p>Unstable Pressure</p> |

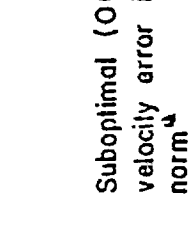
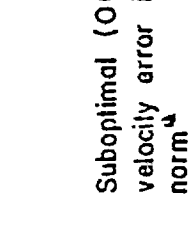


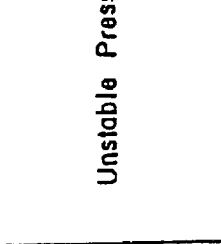
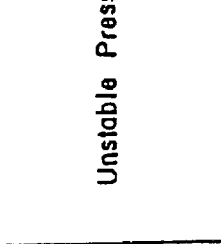
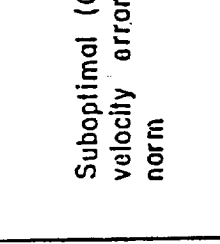
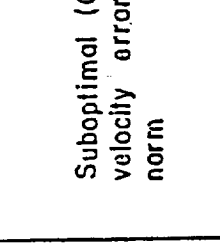
| v^h | q^h | α_h | Rate of Convergence |
|--|---|--------------------------|--|
| <p>12</p>  <p>Q_3</p> |  <p>Q_0</p> | <p>$O(1)$</p> | <p>Suboptimal ($O(h)$) in velocity error in energy norm^a</p> |
| <p>13</p>  <p>Composite $Q_2/18$</p> |  <p>Composite $4P_1$</p> | <p>$O(1)$</p> | <p>Optimal</p> |
| <p>14</p>  <p>Composite $4P_1$</p> |  <p>Composite $4P_0$</p> | <p>$O(h)$</p> | <p>Unstable Pressures</p> |
|  <p>P_2</p> |  <p>P_0</p> | <p>$O(1)$</p> | <p>Suboptimal ($O(h)$) in velocity error in energy norm</p> |

Table 2: Norm-evaluation obtained by:

- Method 1: Q_2/P_1 elements
 Method 2: composite elements
 Method 3: $I8/P_1$ elements and filtering of the pressures
 by using only the centroidal value

Exact Solution: $\|p\|_{L^2(\Omega)/\mathbb{R}} = 100 \sqrt{\frac{37}{12}} \approx 175.5942$; $h^2 = 0.0625$

| | $\ p_\epsilon^h\ _{L^2(\Omega)/\mathbb{R}}$ | $\ p-p_\epsilon^h\ _{L^2(\Omega)/\mathbb{R}}$ | $\frac{\ p-p_\epsilon^h\ _{L^2(\Omega)/\mathbb{R}}}{\ p\ _{L^2(\Omega)/\mathbb{R}}}$ |
|----------|---|---|--|
| Method 1 | 167.1254 | 20.0310 | 0.1141 |
| Method 2 | 171.5448 | 36.3181 | 0.2068 |
| Method 3 | 171.5845 | 26.6219 | 0.1516 |

Table 3: Cost of computation of with a full integration.

| | 4 Node Element Full Integration: 4 points Under Integration: 1 point | 9 Node Element Full Integration: 9 points Under Integration: 4 points |
|---------------------------|--|---|
| Stiffness with Full Int. | 1. | 1. |
| Stiffness with Under-Int. | .41 | .52 |
| Control with Full Int. | .51 | .34 |

Table 4: Operations Cost per Element for Both
Stabilization and A-posteriori Methods

| Stabilization Method | Operations Cost | | A-posteriori Method |
|---|-----------------|-----------|---|
| Computations of $\bar{\epsilon}_e, \underline{\gamma}_e$ | 20x ; 21+ | 20x ; 21+ | Computations of $\bar{\epsilon}_e, \underline{\gamma}_e$ |
| Multiply $\underline{\gamma}_e \cdot \underline{\gamma}_e^T$ | 16x | 4x | Multiply $U_e^T \cdot \underline{\gamma}_e$ |
| Multiply $\epsilon_e \cdot \underline{\gamma}_e \underline{\gamma}_e^T$ | 16x | 1x | Multiply $\epsilon_e \cdot U_e^T \underline{\gamma}_e$ |
| Add $K_e + \epsilon_e \underline{\gamma}_e \underline{\gamma}_e^T$ | 16+ | 2+ | Add $\underline{\gamma}_1 \pm \epsilon_e U_e^T \underline{\gamma}_e$ $\underline{\gamma}_2 + \epsilon_e$ |
| | | 4+ | <u>Then: (4 nodes/element)</u> Add $\underline{u}^h \pm \lambda$ |
| TOTAL | 52x ; 37+ | 29x ; 27+ | TOTAL |

Table 5: Rate of Convergence $|\text{Log} \|e\|_s|$ v.s $|\text{Log } h|$ ($0=0,1$)
for 8- and 9-node Elements

| REGULARITY | $f \in L^2$ | $f \in H^1$ | $f \in H^2$ | $f \in C^\infty$ |
|---------------------------------|--------------|--------------|-------------|------------------|
| BOUNDARY CONDITIONS AND ELEMENT | $\notin H^1$ | $\notin H^2$ | $\in H^3$ | |
| NEUMANN, 9-NODE | 1.99 | 2.43 | 2.95 | 4.00 |
| +SPECIAL PROJECTION | 1.74 | 1.97 | 1.94 | 3.00 |
| NEUMANN, 8-NODE EL. | 1.99 | 2.00 | 2.97 | 4.00 |
| | 1.79 | 1.93 | 1.95 | 3.00 |
| DIRICHLET, 9-NODE EL. | 2.35 | 2.85 | 2.99 | 3.00 |
| | 1.47 | 1.99 | 1.99 | 2.00 |
| DIRICHLET, 8-NODE EL. | 2.30 | 2.71 | 2.99 | 3.00 |
| | 1.46 | 2.12 | 1.99 | 2.00 |
| MIXED, 9-NODE EL. | 2.00 | 2.00 | 3.83 | 4.00 |
| | 1.67 | 2.28 | 2.84 | 3.00 |
| MIXED, 8-NODE EL. | 2.00 | 2.00 | 3.84 | 4.00 |
| | 1.74 | 2.27 | 2.84 | 3.00 |

Table 6: Rate of Convergence $|\text{Log } \|e\|_s|$ v.s. $|\text{Log } h|$ for
4-node Elements ($s=0,1$)

| REGULARITY | | $f \in L^2$ | $f \in H^1$ | $f \in C^\infty$ |
|-------------------------------------|-------------|--------------|--------------|------------------|
| OPERATOR AND BOUNDARY CONDITIONS | | $\notin H^1$ | $\notin H^2$ | |
| - Δ | NEUMANN | 1.99 | 2.00 | 2.00 |
| | +PROJECTION | 1.61 | 2.00 | 2.00 |
| | DIRICHLET | 1.99 | 2.00 | 2.00 |
| | | 1.50 | 1.85 | 2.00 |
| - $\Delta + 1$ | MIXED | 2.00 | 2.00 | 2.00 |
| | | 1.50 | 1.99 | 1.99 |
| | NEUMANN | 2.00 | 2.00 | 2.00 |
| | +PROJECTION | 1.50 | 2.00 | 2.00 |
| - $\Delta + 1$ | DIRICHLET | 2.00 | 2.00 | 2.00 |
| | | 1.50 | 1.85 | 2.00 |
| | MIXED | 2.00 | 2.00 | 2.00 |
| | | 1.50 | 1.85 | 2.00 |

Table 7: Arrays of Eigen Values A_{ij} , \bar{A}_{ij} and $\bar{\bar{A}}_{ij}$ Table 7a. Array A_{ij}

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|-------|------|------|------|------|------|------|------|------|------|
| 0 | 40.45 | 4.42 | 1.54 | 0.75 | 0.43 | 0.27 | 0.18 | 0.13 | 0.10 | 0.08 |
| 1 | 8.05 | 3.07 | 1.34 | 0.70 | 0.41 | 0.26 | 0.17 | 0.12 | 0.10 | 0.08 |
| 2 | 2.31 | 1.58 | 0.95 | 0.57 | 0.36 | 0.24 | 0.17 | 0.12 | 0.09 | 0.08 |
| 3 | 1.02 | 0.85 | 0.62 | 0.44 | 0.30 | 0.21 | 0.15 | 0.11 | 0.09 | 0.08 |
| 4 | 0.55 | 0.49 | 0.41 | 0.32 | 0.24 | 0.18 | 0.13 | 0.10 | 0.08 | 0.07 |
| 5 | 0.33 | 0.31 | 0.27 | 0.23 | 0.19 | 0.15 | 0.12 | 0.09 | 0.08 | 0.07 |
| 6 | 0.21 | 0.21 | 0.19 | 0.17 | 0.14 | 0.12 | 0.10 | 0.08 | 0.07 | 0.06 |
| 7 | 0.15 | 0.14 | 0.14 | 0.12 | 0.11 | 0.10 | 0.08 | 0.07 | 0.06 | 0.05 |
| 8 | 0.11 | 0.11 | 0.10 | 0.10 | 0.09 | 0.08 | 0.07 | 0.06 | 0.05 | 0.05 |
| 9 | 0.09 | 0.09 | 0.08 | 0.08 | 0.07 | 0.07 | 0.06 | 0.05 | 0.05 | 0.04 |
| 10 | 0.08 | 0.08 | 0.08 | 0.07 | 0.07 | 0.06 | 0.06 | 0.05 | 0.04 | 0.04 |

Table 7b. Array \bar{A}_{ij}

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|-------|------|------|------|------|------|------|------|------|------|
| 0 | 40.36 | 4.34 | 1.46 | 0.67 | 0.34 | 0.18 | 0.09 | 0.04 | 0.01 | 0.00 |
| 1 | 7.99 | 3.02 | 1.27 | 0.62 | 0.33 | 0.18 | 0.09 | 0.04 | 0.01 | 0.00 |
| 2 | 2.24 | 1.53 | 0.90 | 0.52 | 0.30 | 0.17 | 0.09 | 0.04 | 0.01 | 0.00 |
| 3 | 0.94 | 0.79 | 0.58 | 0.39 | 0.25 | 0.15 | 0.09 | 0.04 | 0.01 | 0.00 |
| 4 | 0.47 | 0.43 | 0.36 | 0.28 | 0.20 | 0.13 | 0.08 | 0.04 | 0.01 | 0.00 |
| 5 | 0.25 | 0.24 | 0.21 | 0.18 | 0.14 | 0.11 | 0.07 | 0.04 | 0.01 | 0.00 |
| 6 | 0.13 | 0.13 | 0.12 | 0.11 | 0.10 | 0.08 | 0.05 | 0.03 | 0.01 | 0.00 |
| 7 | 0.06 | 0.06 | 0.06 | 0.06 | 0.05 | 0.05 | 0.04 | 0.03 | 0.01 | 0.00 |
| 8 | 0.03 | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 | 0.02 | 0.02 | 0.01 | 0.00 |
| 9 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
| 10 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 7c. Array $\bar{\bar{A}}_{ij}$

| i | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
|----|-------|------|------|------|------|------|------|------|------|------|
| 0 | 40.45 | 4.42 | 1.54 | 0.75 | 0.43 | 0.27 | 0.18 | 0.13 | 0.10 | 0.08 |
| 1 | 8.08 | 3.11 | 1.36 | 0.71 | 0.42 | 0.26 | 0.18 | 0.13 | 0.10 | 0.09 |
| 2 | 2.32 | 1.62 | 0.99 | 0.61 | 0.39 | 0.26 | 0.18 | 0.13 | 0.10 | 0.09 |
| 3 | 1.02 | 0.87 | 0.67 | 0.48 | 0.34 | 0.24 | 0.18 | 0.13 | 0.10 | 0.09 |
| 4 | 0.55 | 0.51 | 0.44 | 0.37 | 0.29 | 0.23 | 0.17 | 0.14 | 0.11 | 0.10 |
| 5 | 0.33 | 0.32 | 0.30 | 0.27 | 0.24 | 0.20 | 0.17 | 0.14 | 0.12 | 0.11 |
| 6 | 0.22 | 0.21 | 0.21 | 0.20 | 0.19 | 0.18 | 0.17 | 0.16 | 0.14 | 0.14 |
| 7 | 0.15 | 0.15 | 0.15 | 0.15 | 0.15 | 0.16 | 0.17 | 0.17 | 0.18 | 0.19 |
| 8 | 0.11 | 0.11 | 0.11 | 0.12 | 0.13 | 0.14 | 0.16 | 0.20 | 0.26 | 0.33 |
| 9 | 0.09 | 0.09 | 0.09 | 0.10 | 0.11 | 0.13 | 0.16 | 0.23 | 0.42 | 0.97 |
| 10 | 0.08 | 0.08 | 0.09 | 0.09 | 0.10 | 0.12 | 0.16 | 0.25 | 0.57 | 4.57 |

ORIGINAL PAGE IS
OF POOR QUALITY

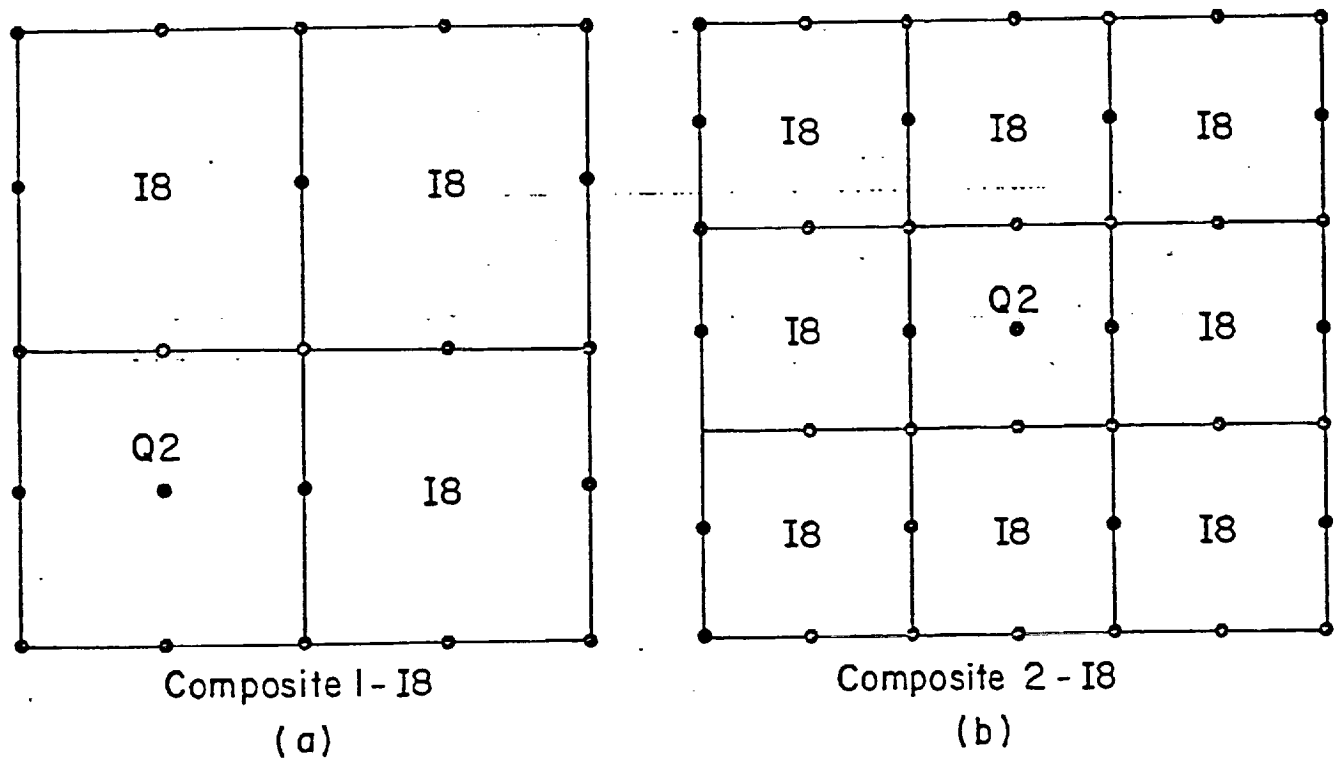


Figure 1. Examples of stable composite elements.

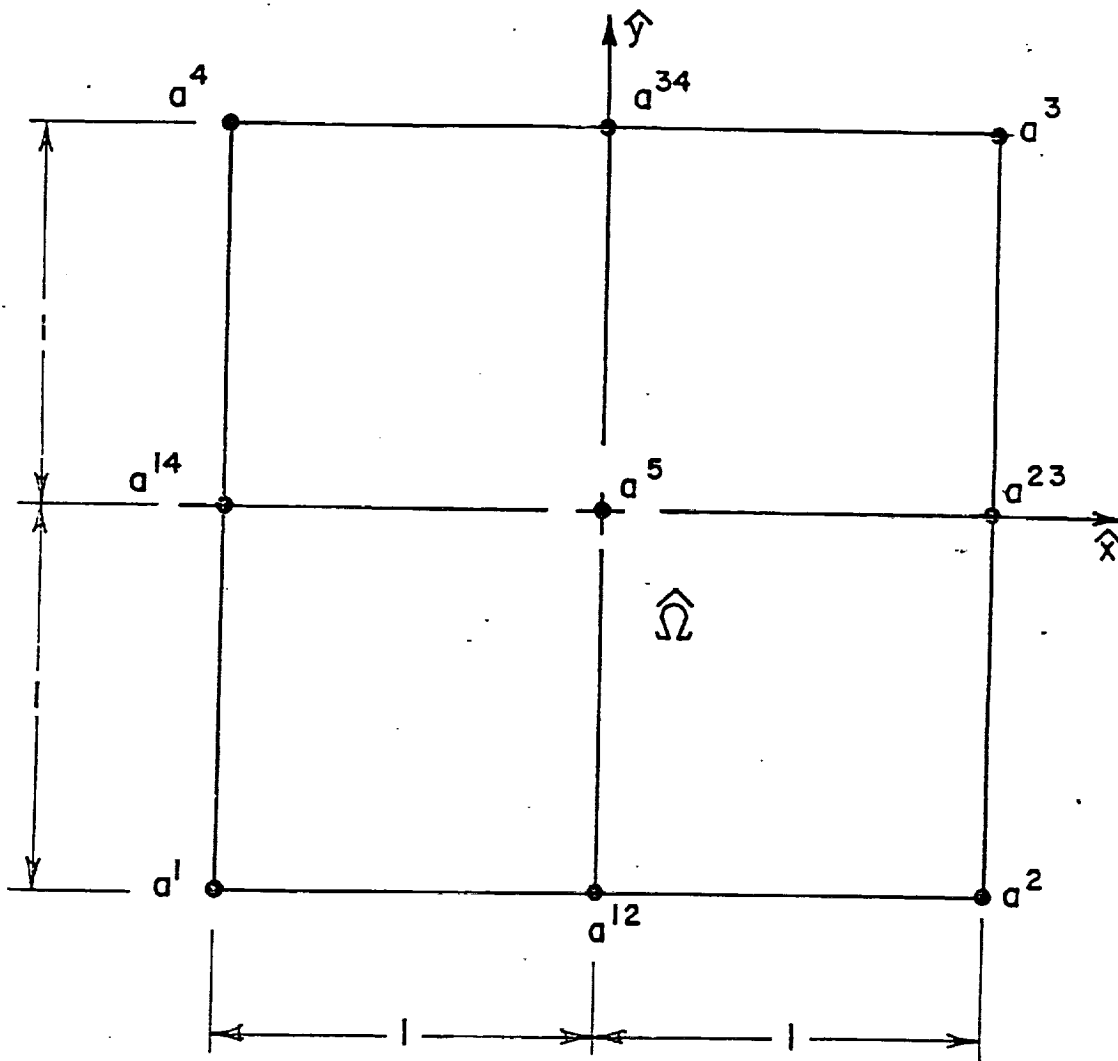


Figure 2. Geometry and node-numbering for a master Q_2/P_1 element $\hat{\Omega}$.

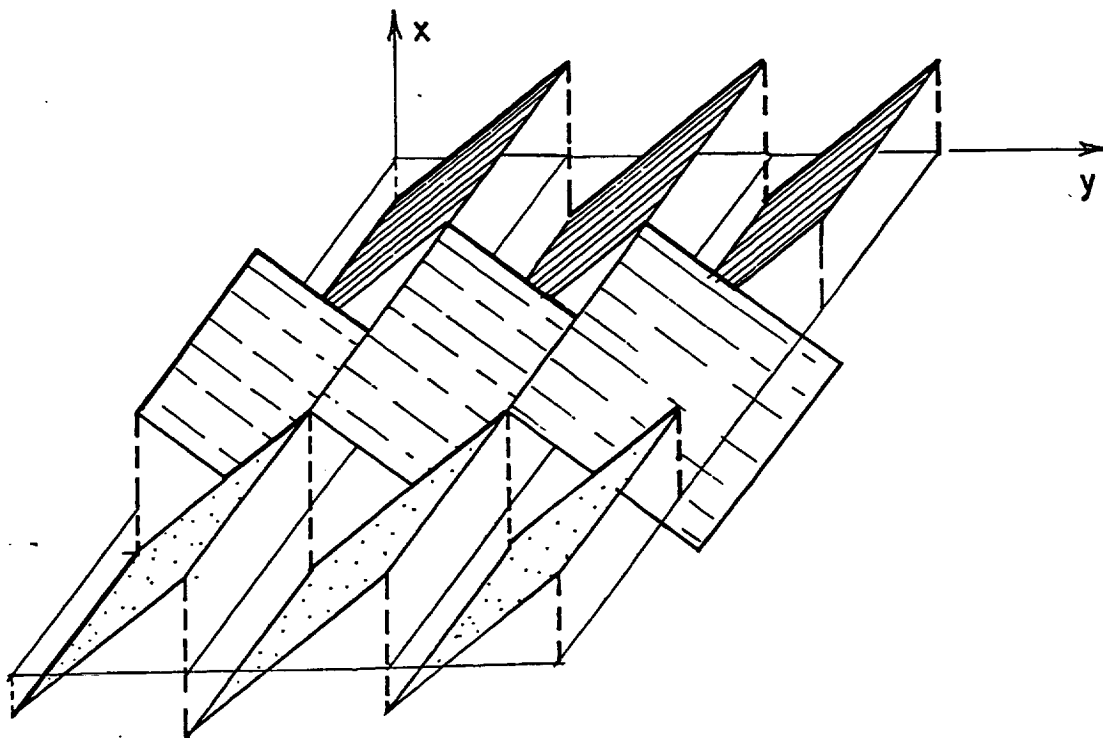
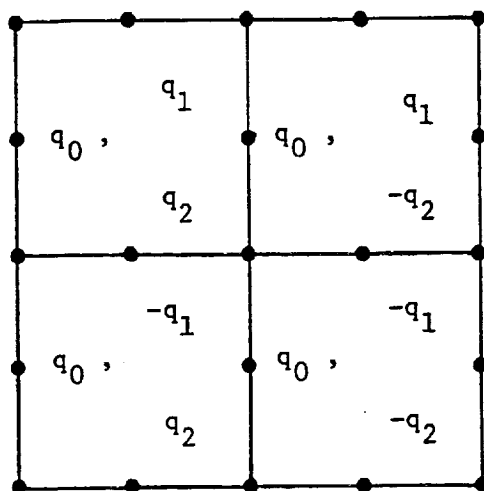
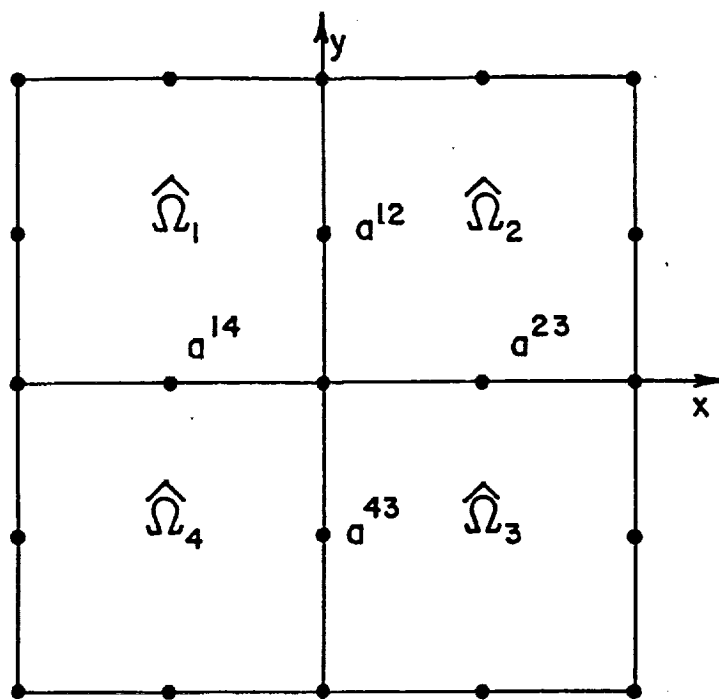
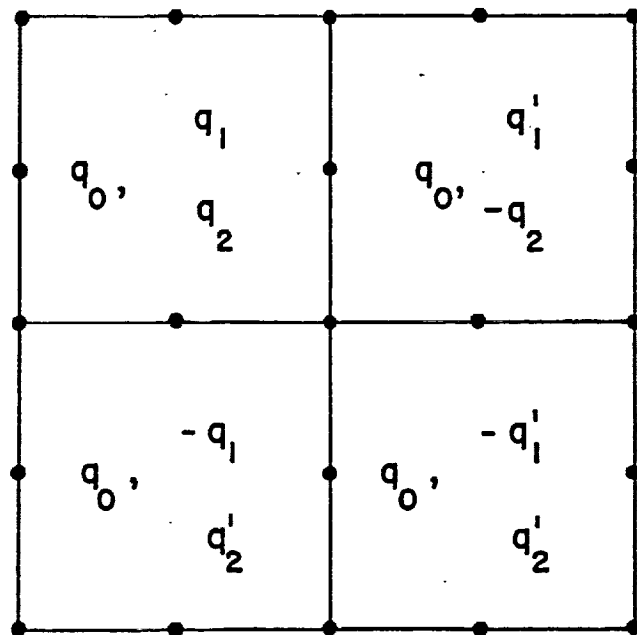


Figure 3. Kernel functions in $\ker B_h^*$ for $I8/P_1$ - elements.

25A



(a)



(b)

Figure 4. a) a composite of four master elements and b) the degrees of freedom in $\ker B_h^*$ at an intermediate step in the proof of Lemma 5.1.

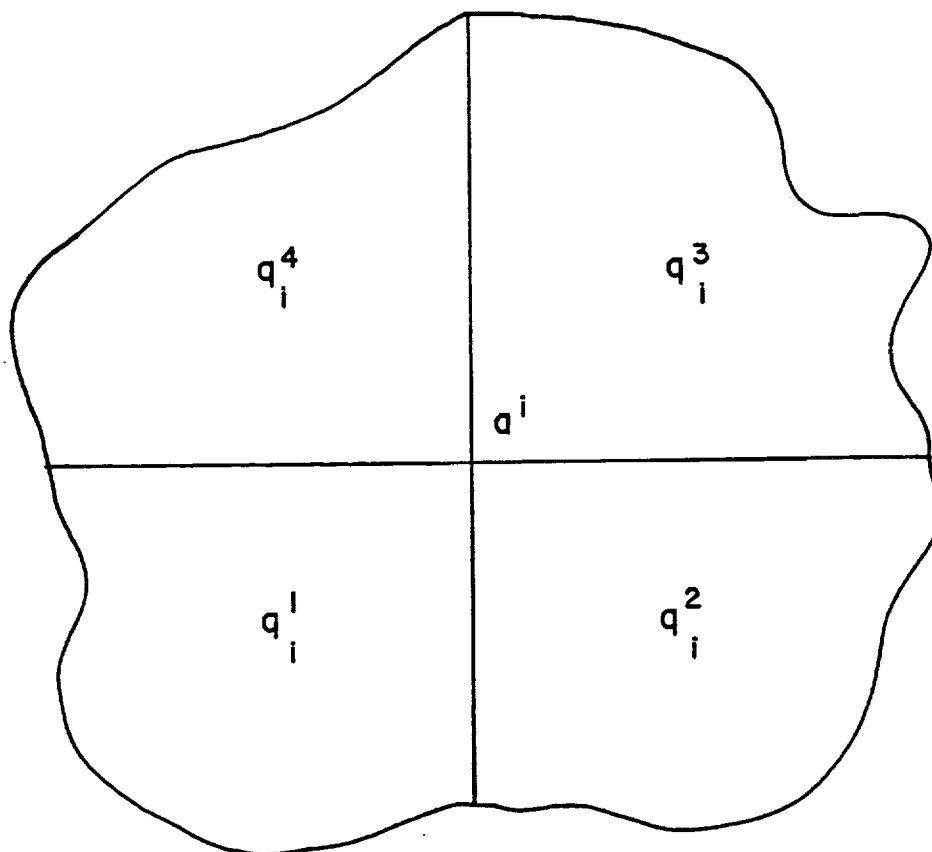


Figure 5. Numbering scheme for values of q_i^j surrounding node a^i .

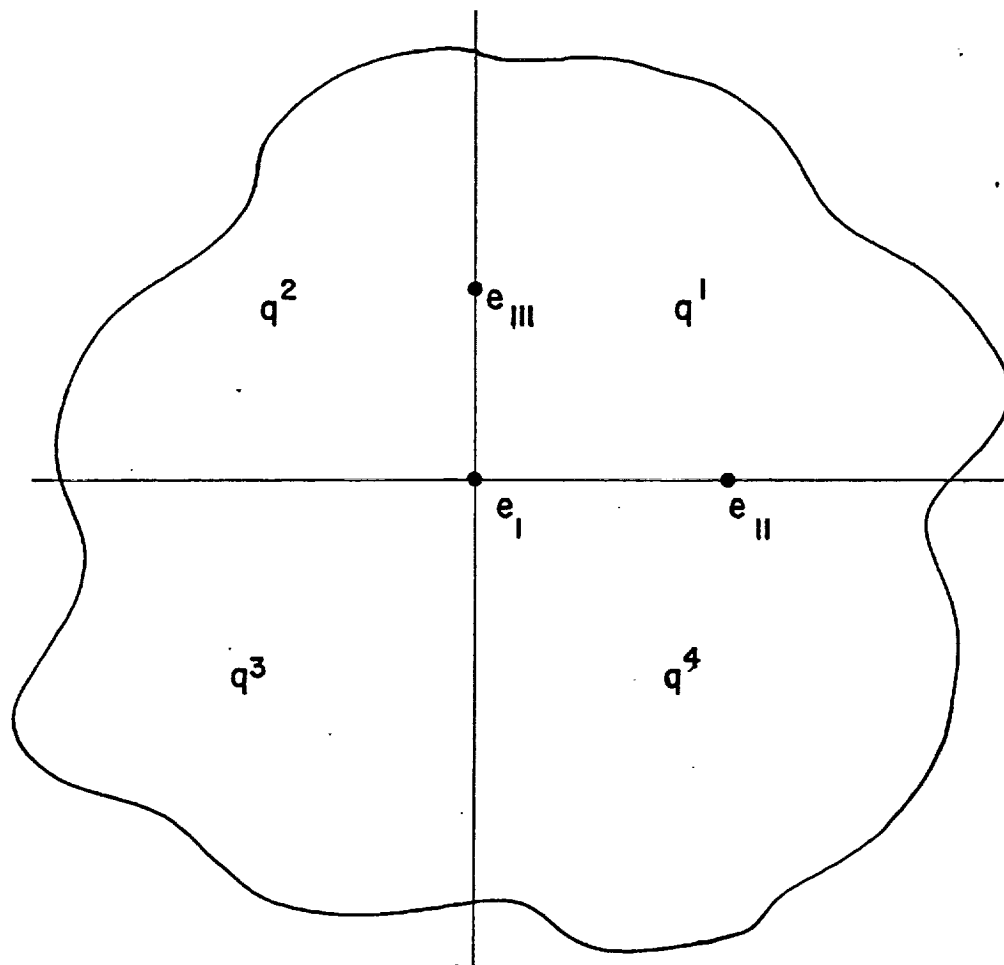
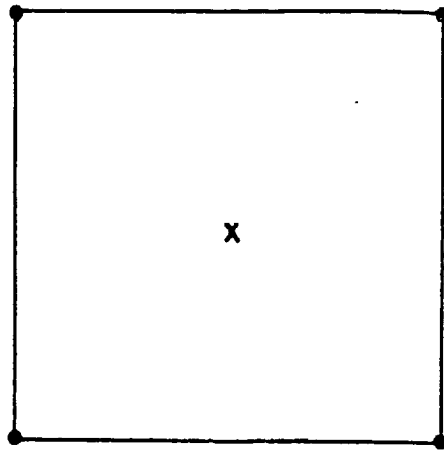
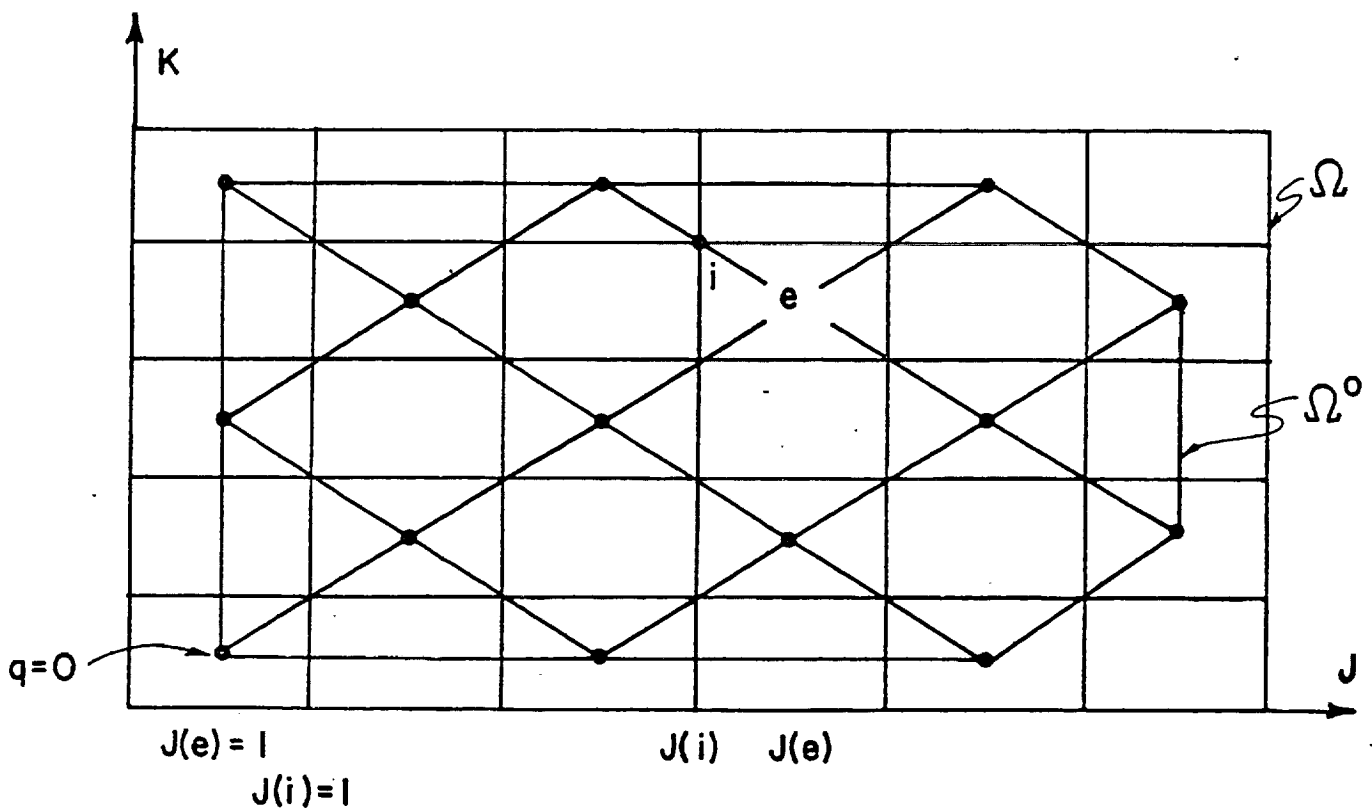


Figure 6. Nodes e_I , e_{II} , and e_{III} .



(a)



(b)

Figure 7. a) The Q_1/P_0 - element and b) a mesh path needed to construct the auxiliary function ϕ with values q_e for e such that $I + J$ is even, so as to obtain (7.1) and (7.2).

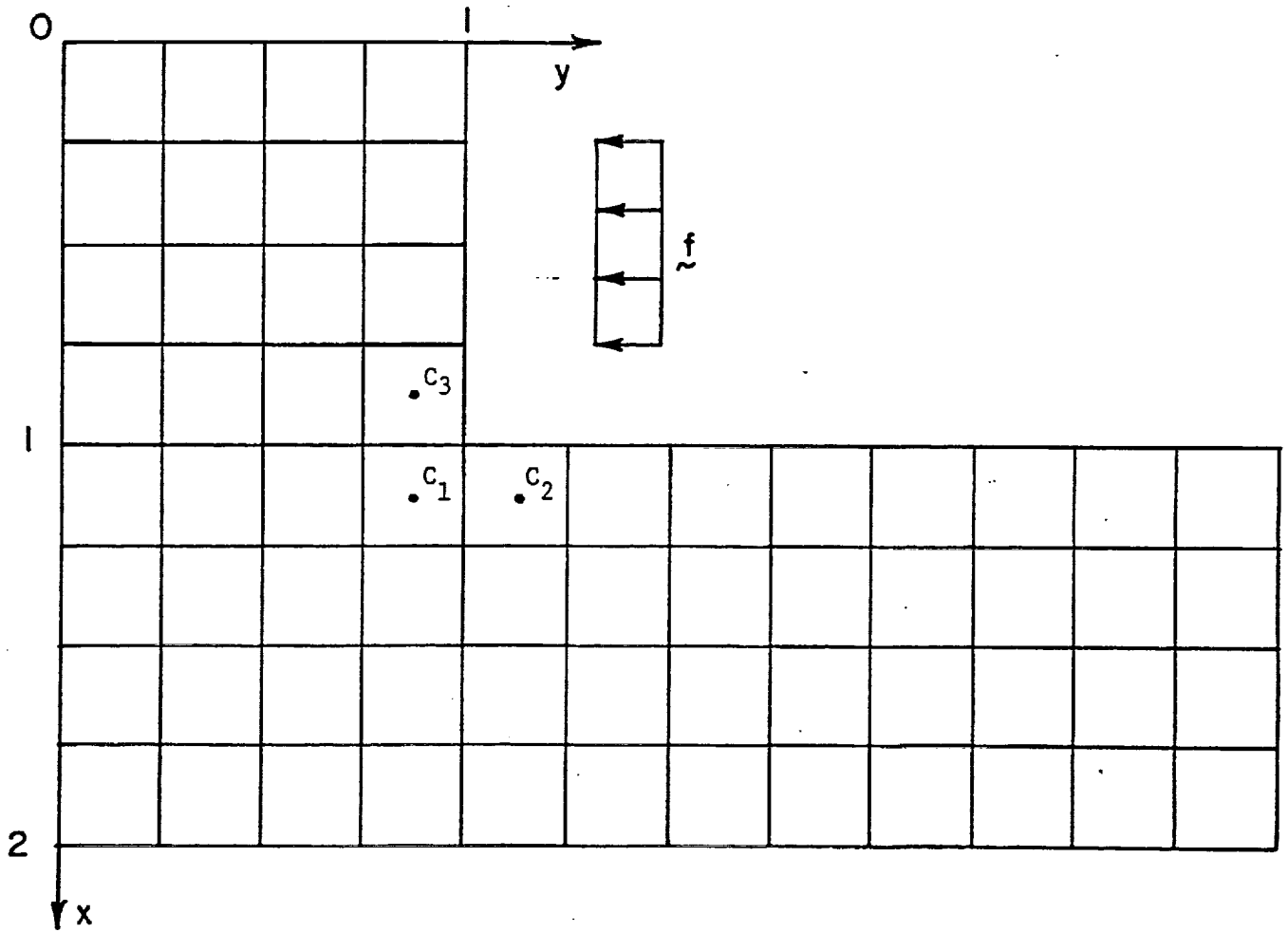


Figure 8. A mesh of 64 elements on an L-shaped domain.

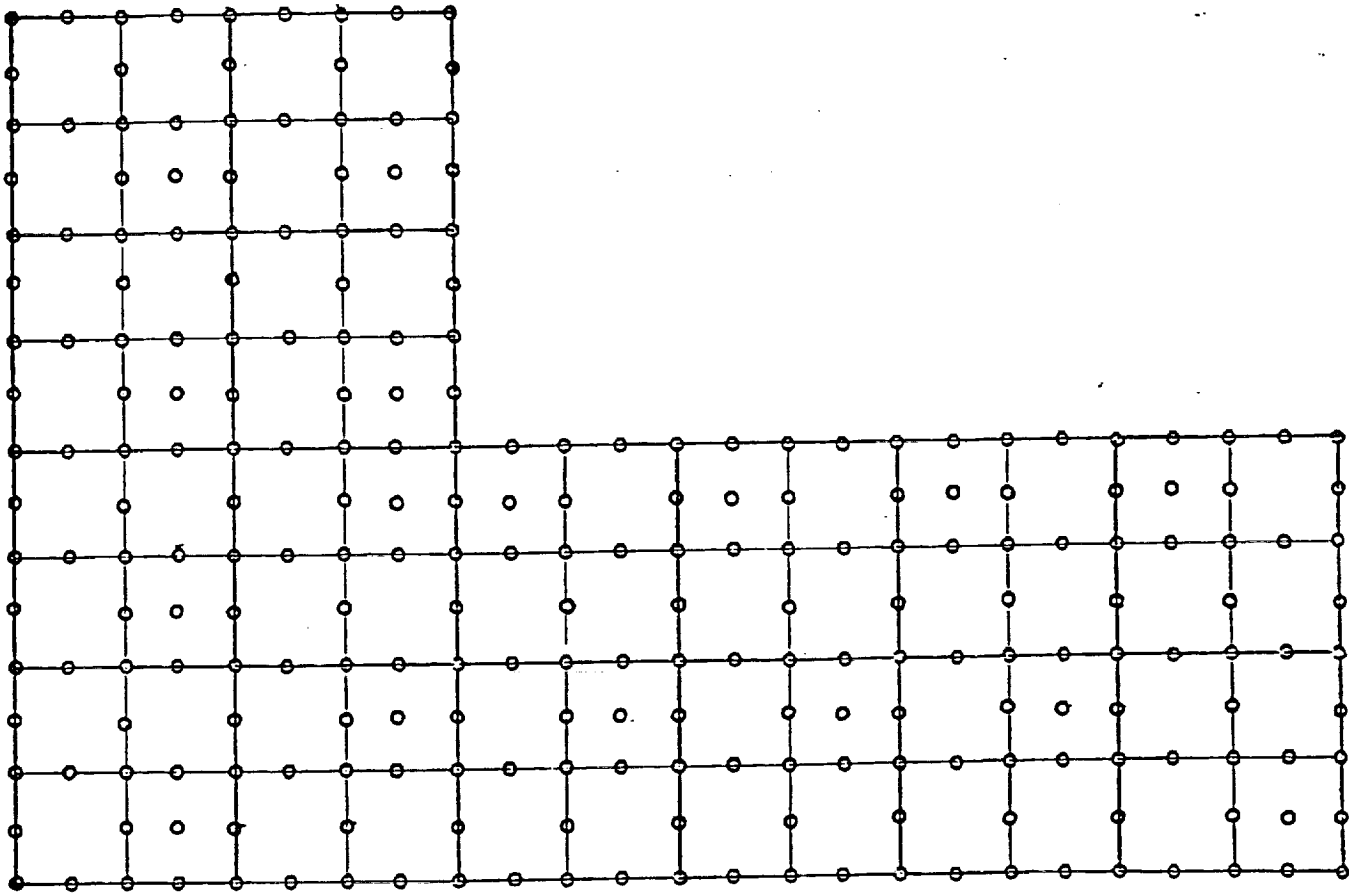


Figure 9. Mesh with composite-elements.

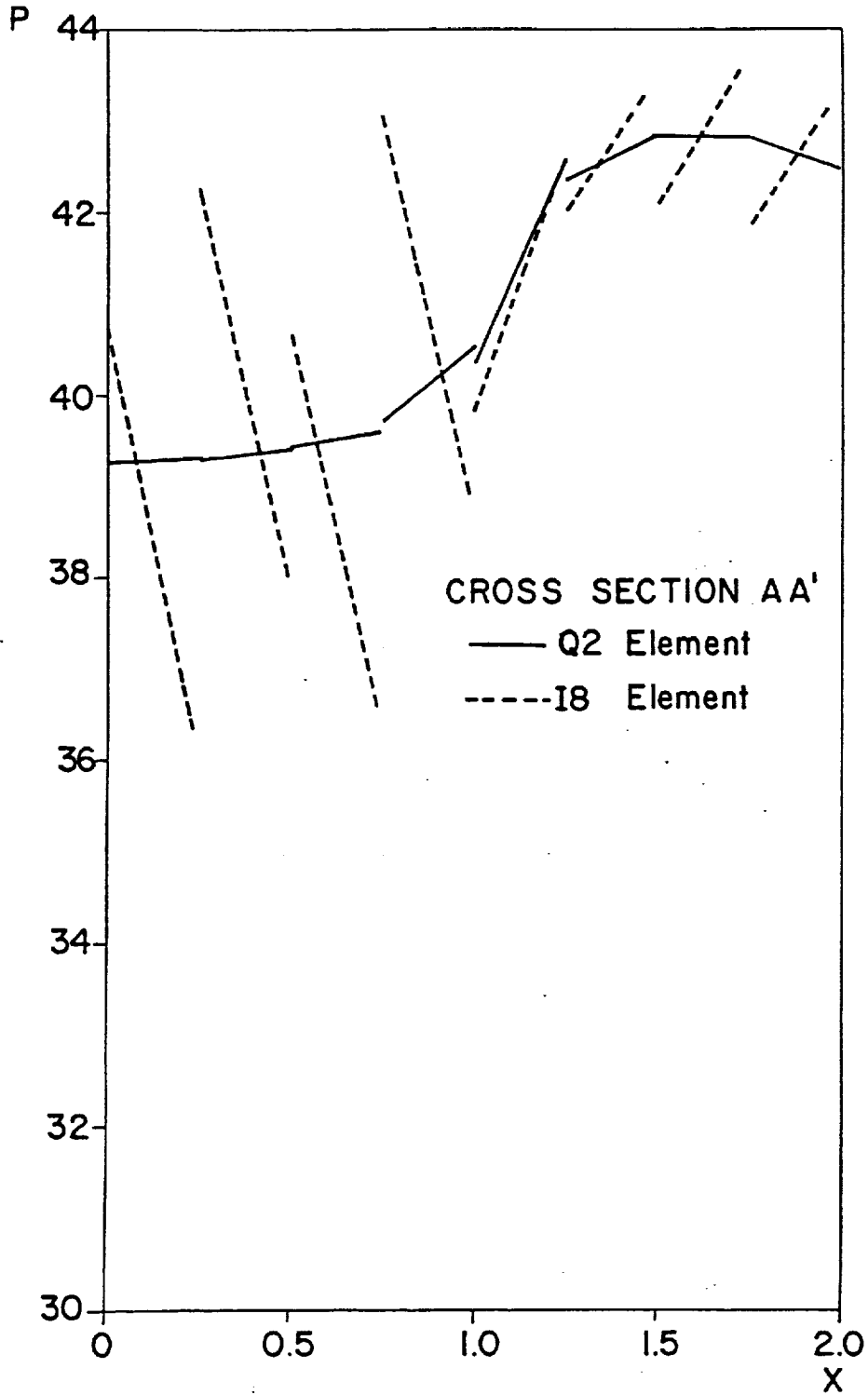


Figure 10. Computed pressure profiles along Section AA'.

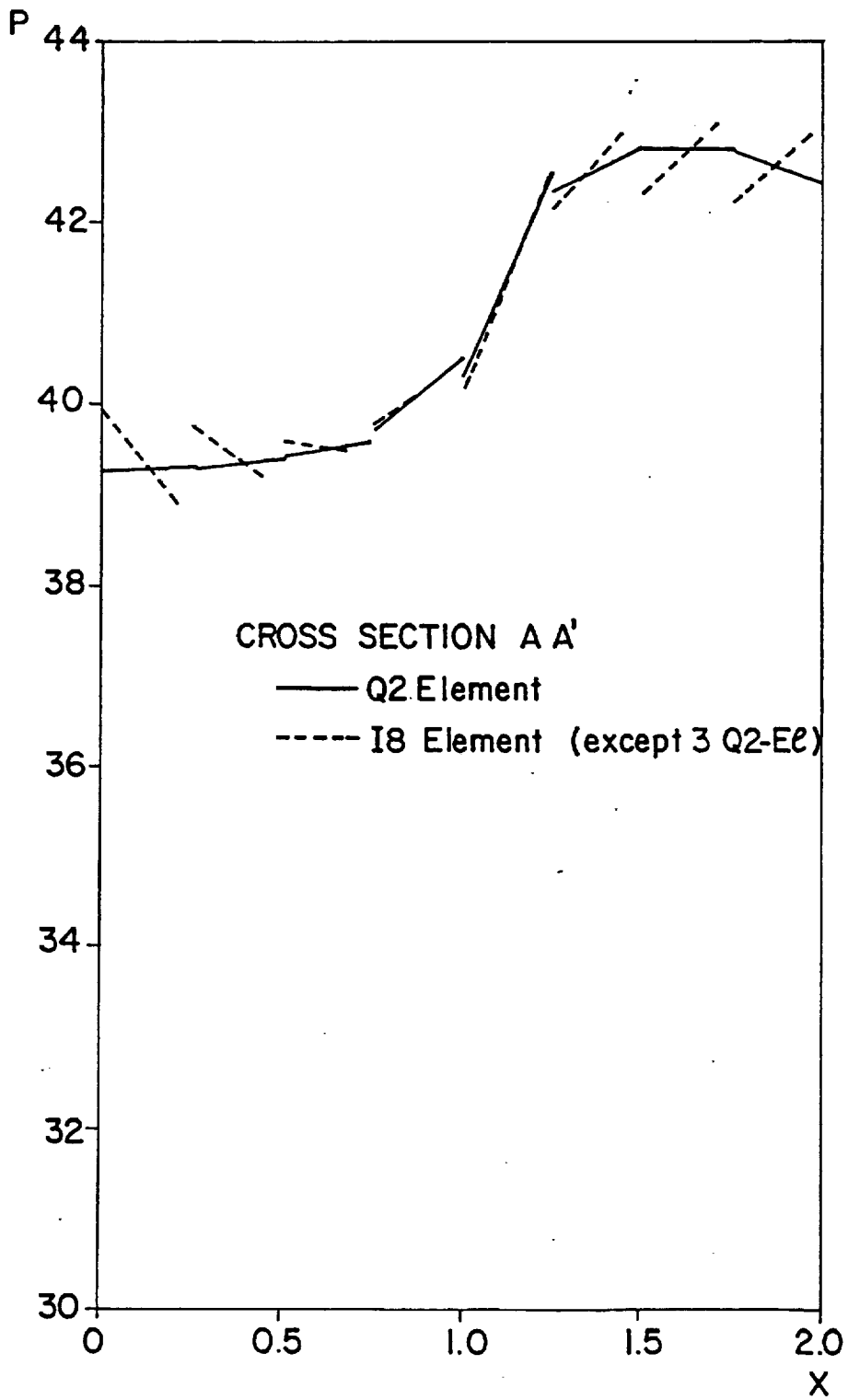


Figure 11. Computed pressure profiles along Section AA'.

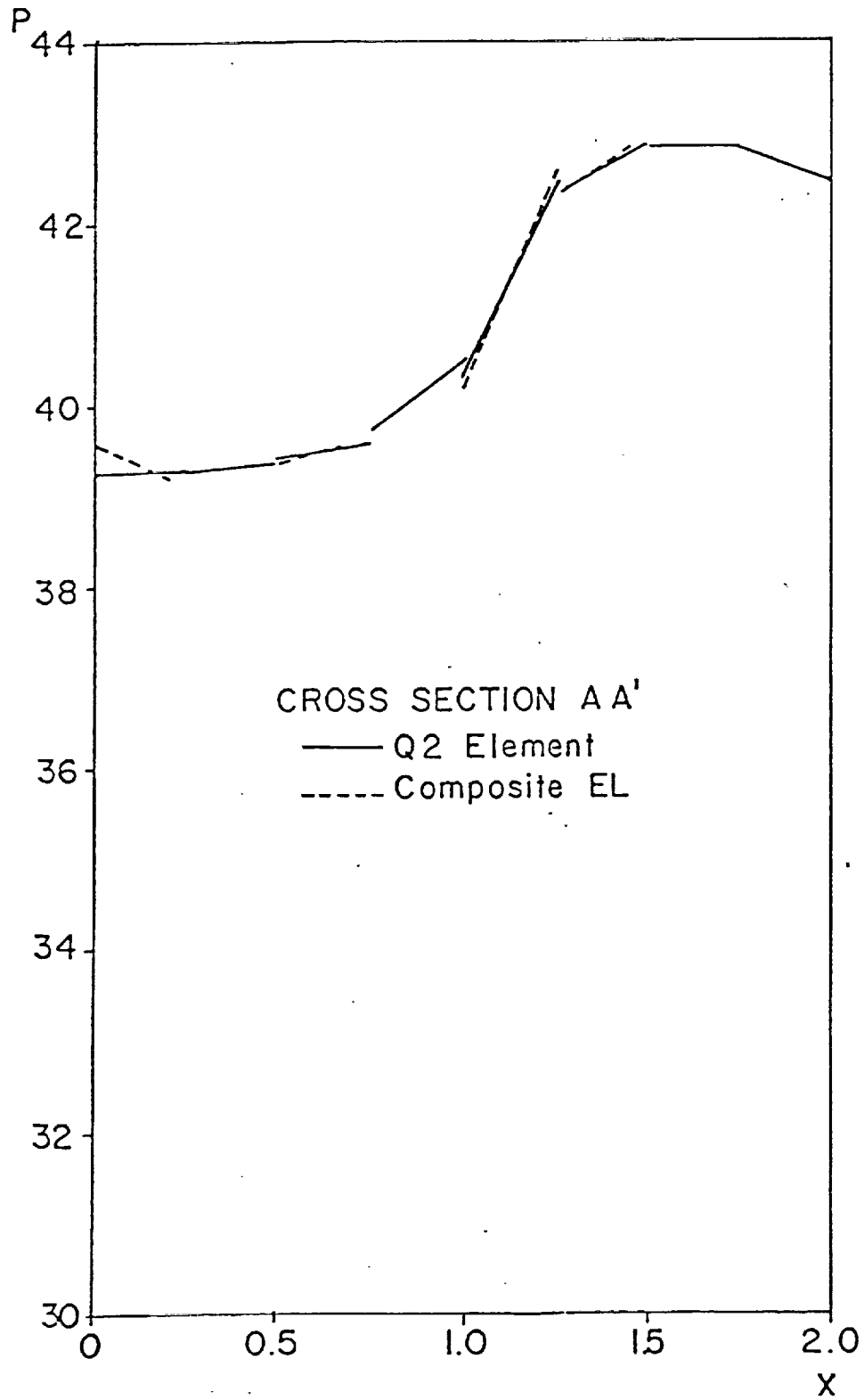


Figure 12. Computed pressure profiles along Section AA'.

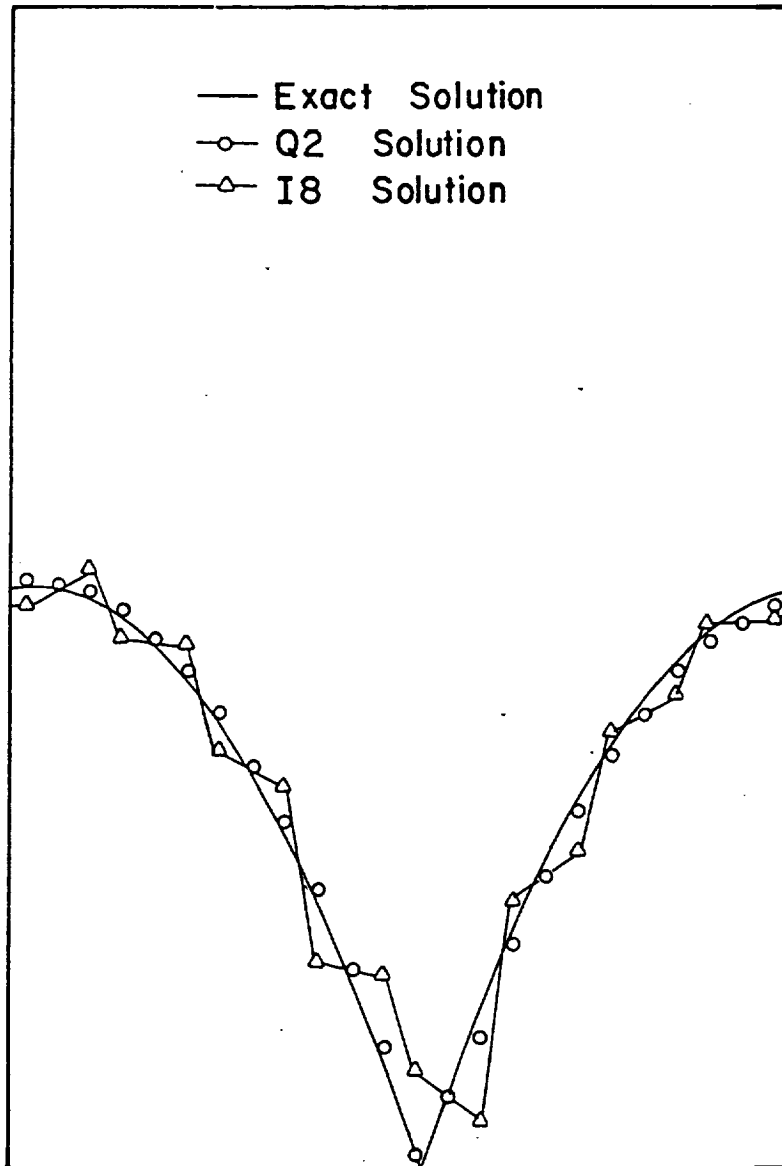


Figure 13. Pressure profiles for second example.

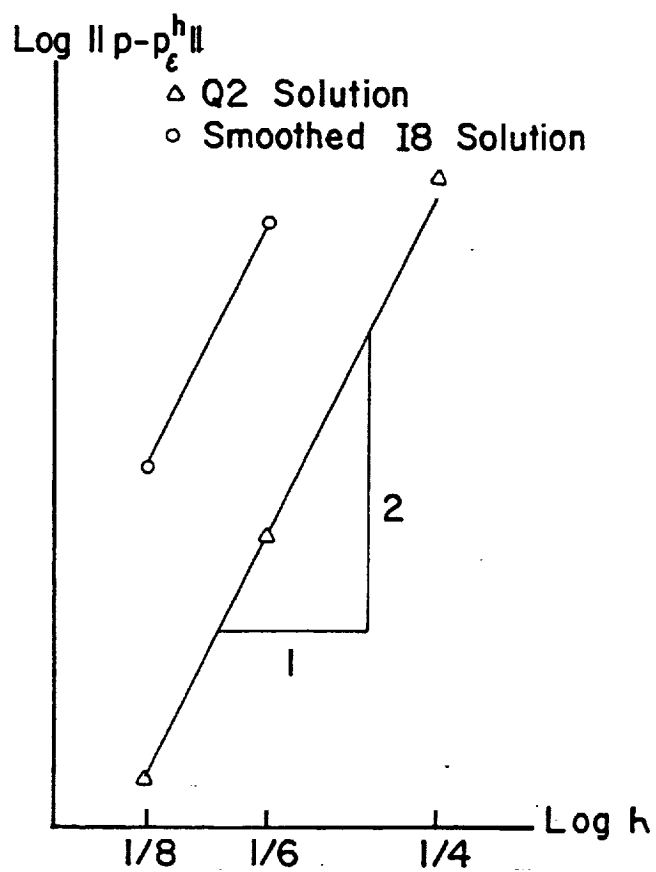


Figure 14. Computed rate of convergence of the I8/P₁ - pressures in L² - with and without filtering.

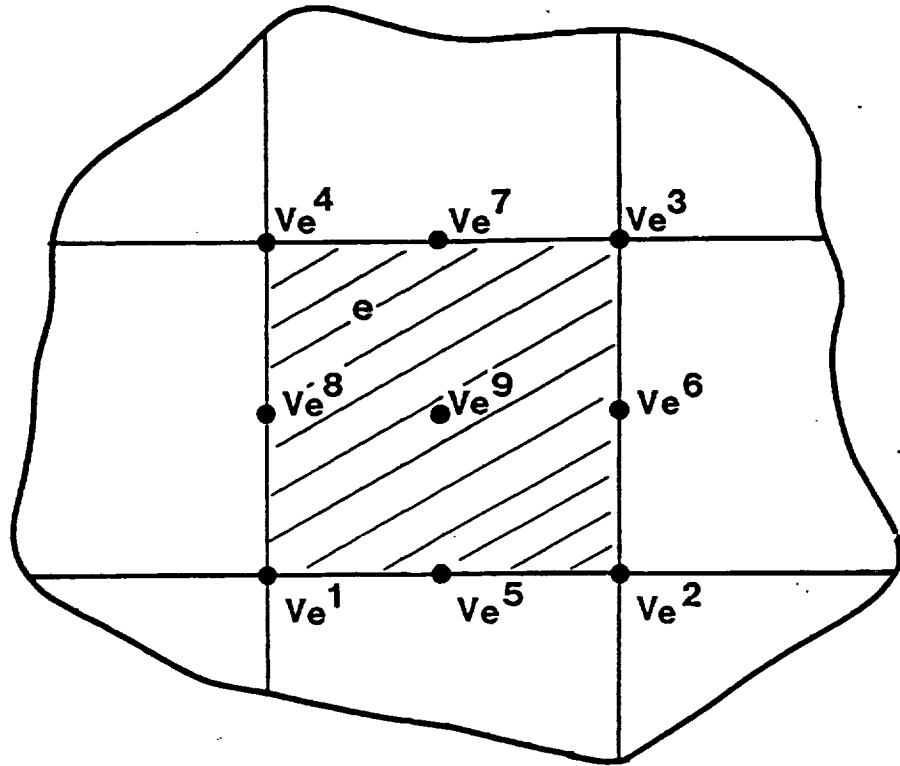
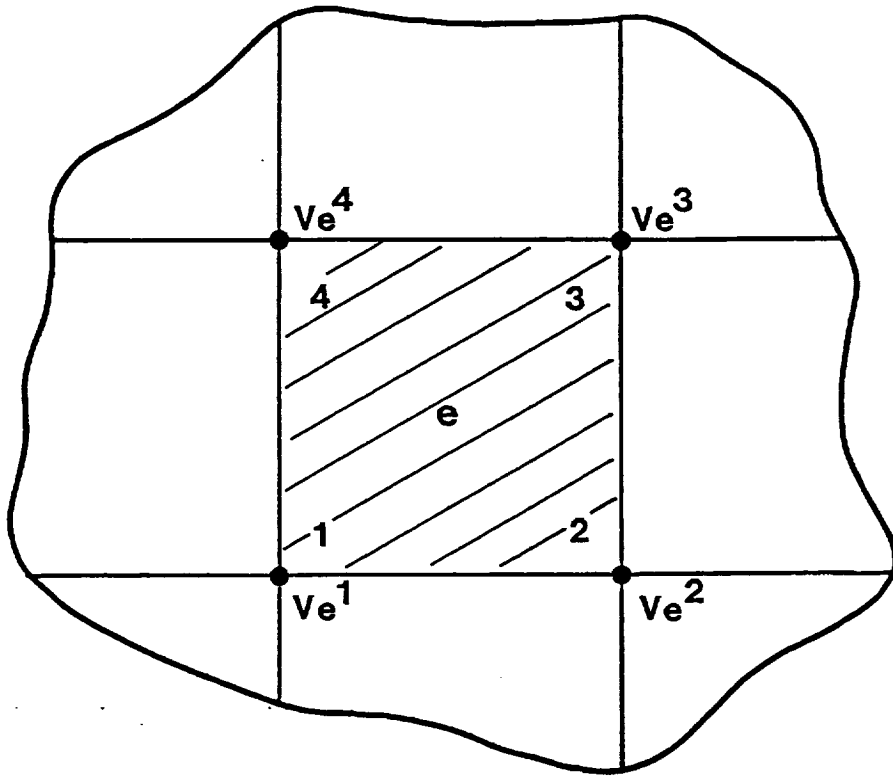


Figure 15: Node numbering convention for 4, and 9, node elements.

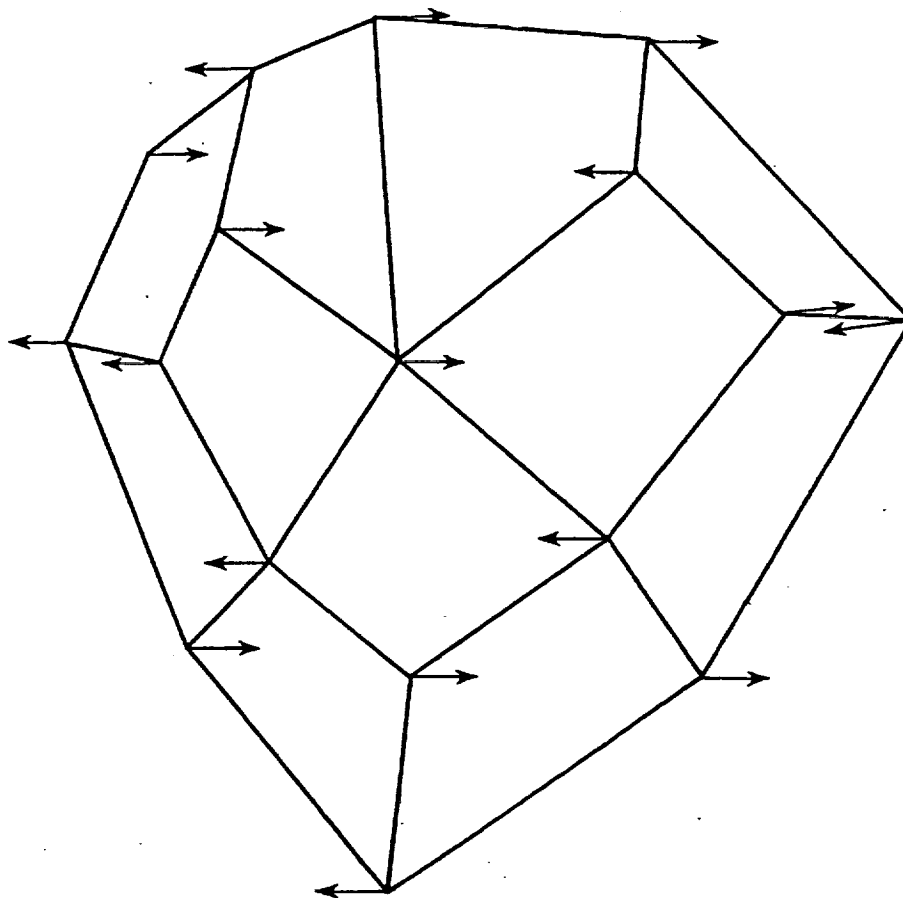


Figure 16. "1" pattern of the hourglass mode in an arbitrary mesh.

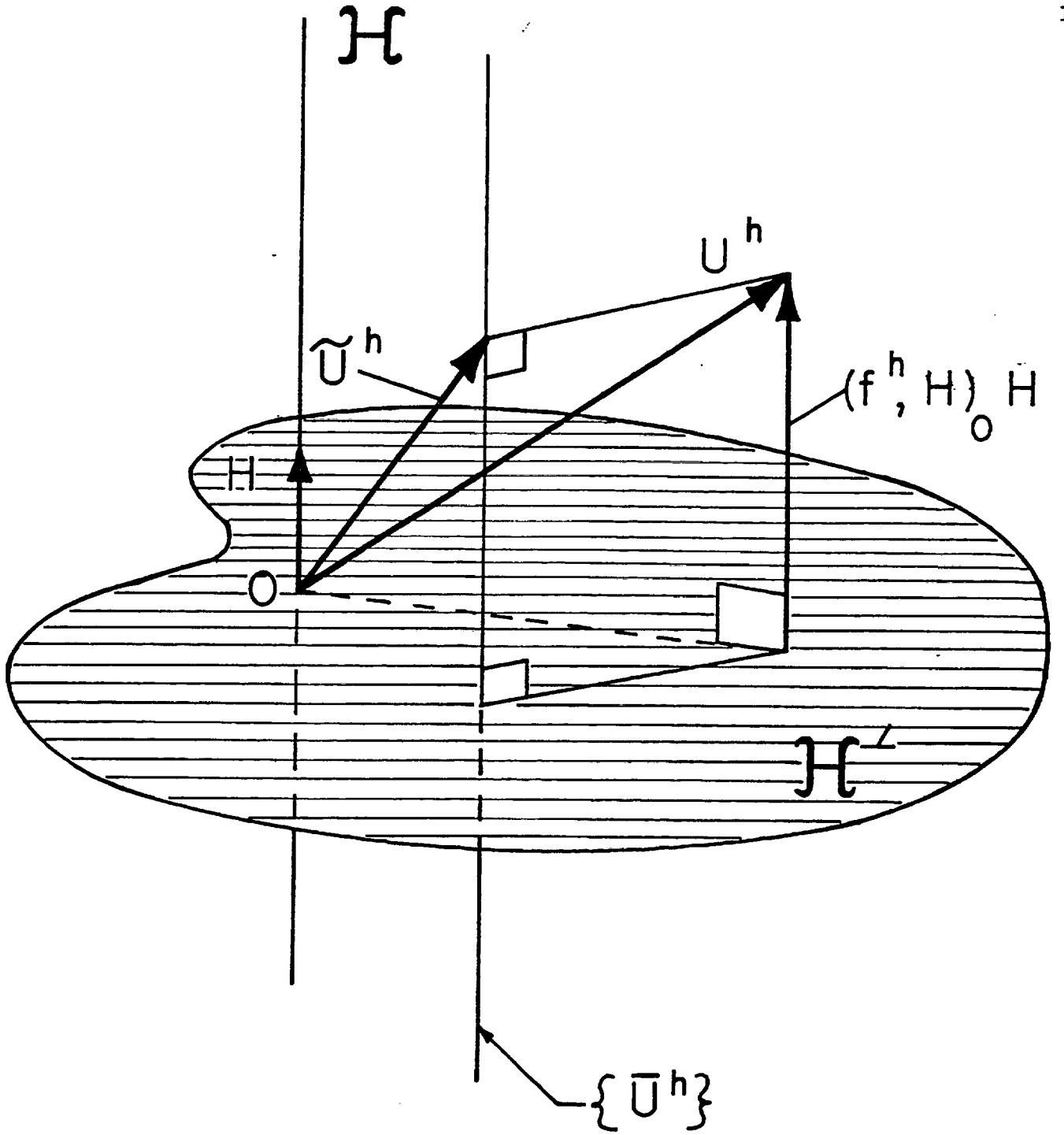


Figure 17. Geometrical interpretation of the projection $\tilde{u}^h = \pi_{\mathcal{H}} u^h$.

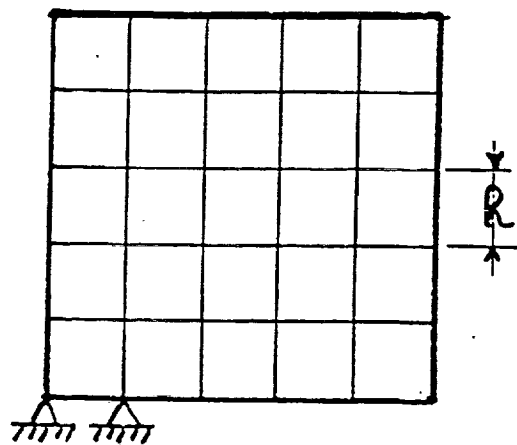


Figure 18a

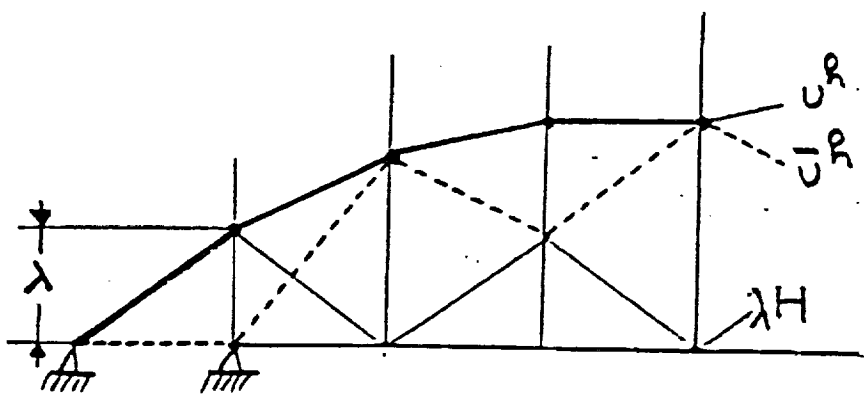


Figure 18b. $U^h = \bar{U}^h + \lambda H + O(h^{2-s})$

Figure 18. Justification of the omission of the projection

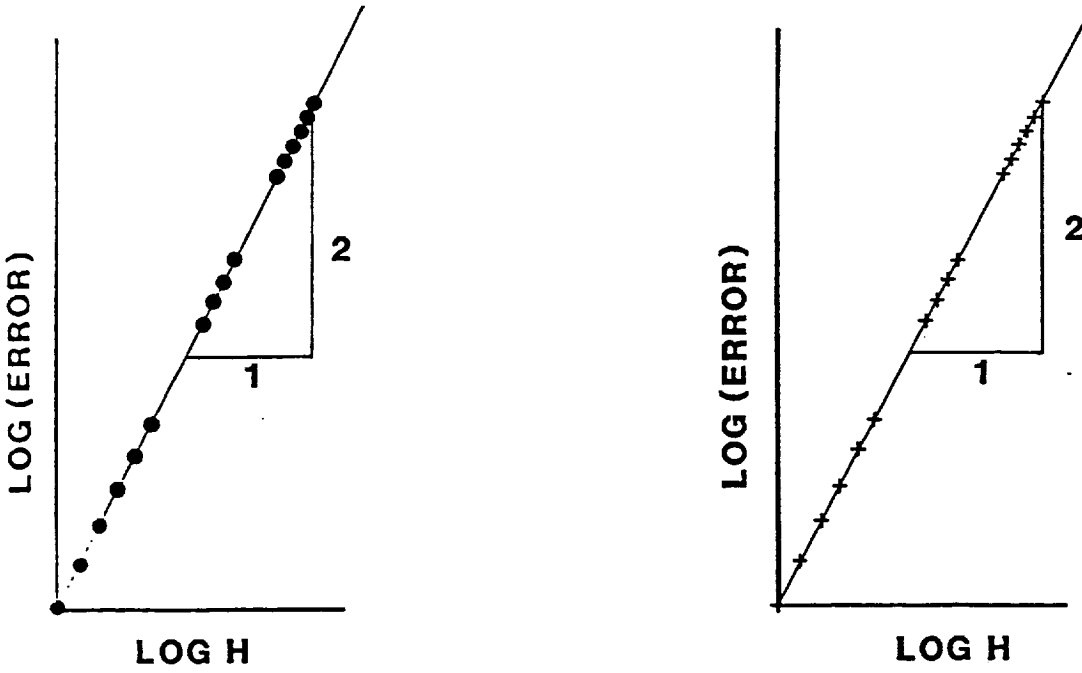


Figure 19a. Comparison between u^h and \bar{u}^h .

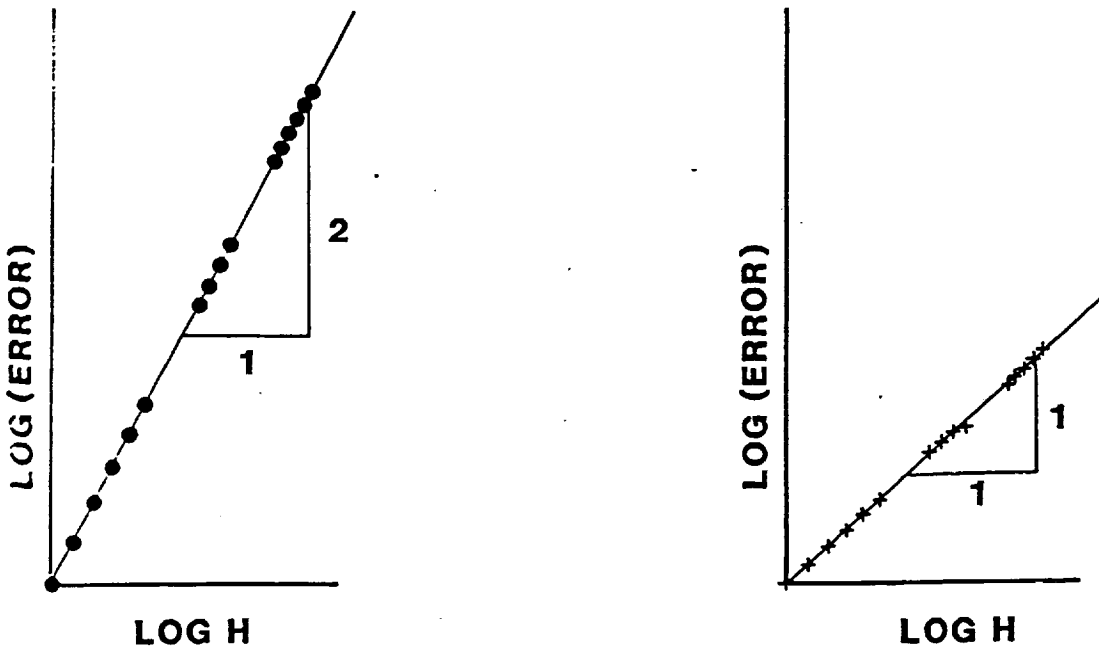


Figure 19b. Comparison between u^h and \bar{u}^h .

Figure 19. Results obtained with a continuous data function.

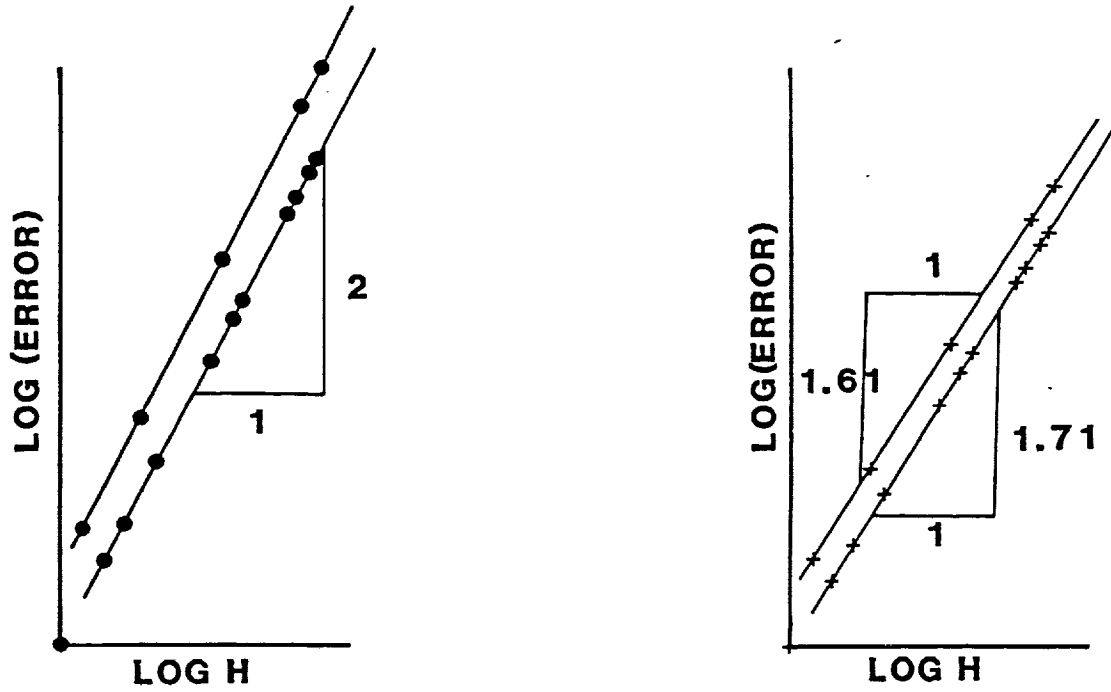


Figure 20a. Comparison between u^h and \bar{u}^h .

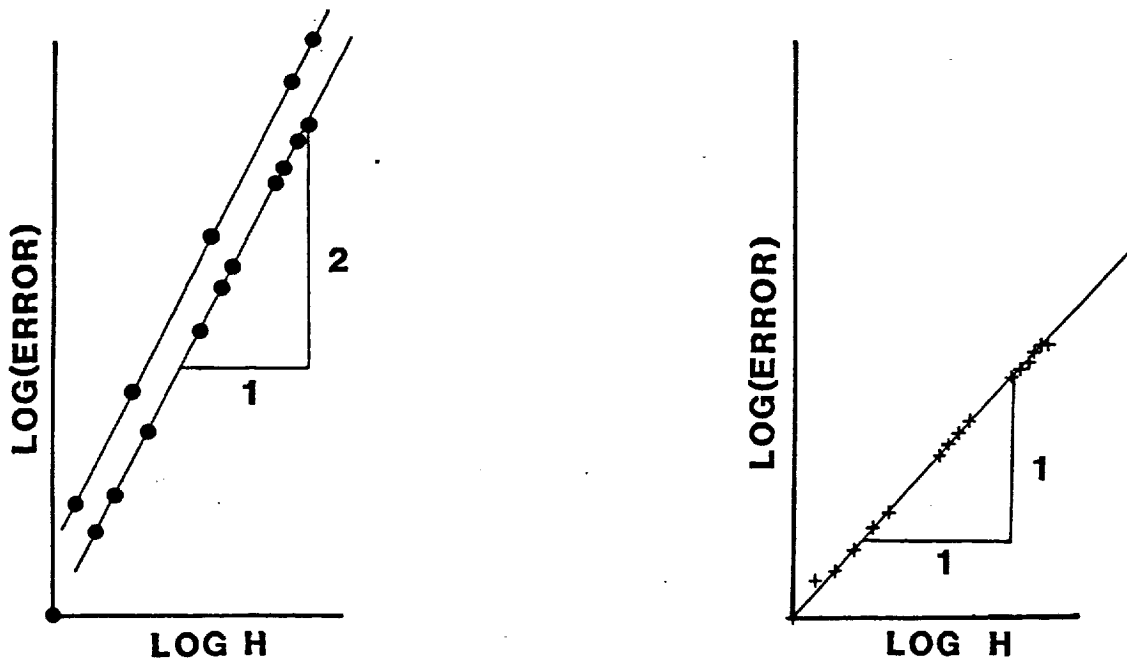
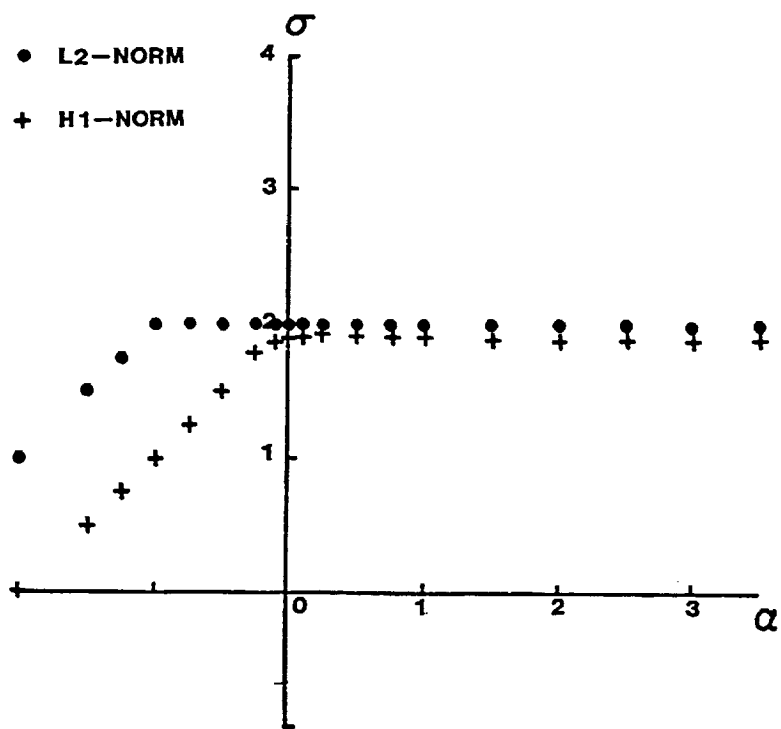
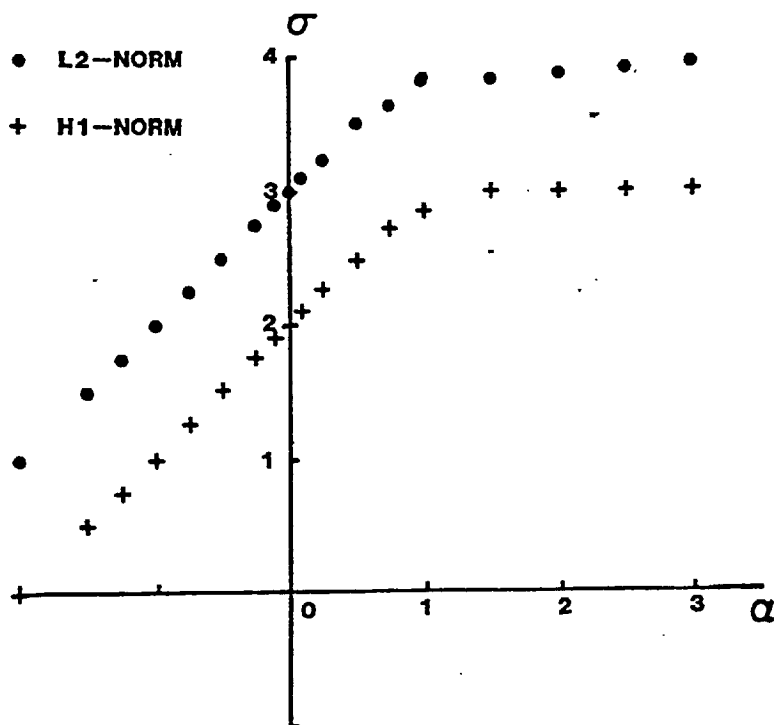


Figure 20b. Comparison between u^h and \bar{u}^h .

Figure 20. Results obtained with the discontinuous data function f_2 , for Neumann boundary conditions.



a: Bilinear elements



b: Biquadratic elements

Figure 21. (α, σ) Plot for a square domain.

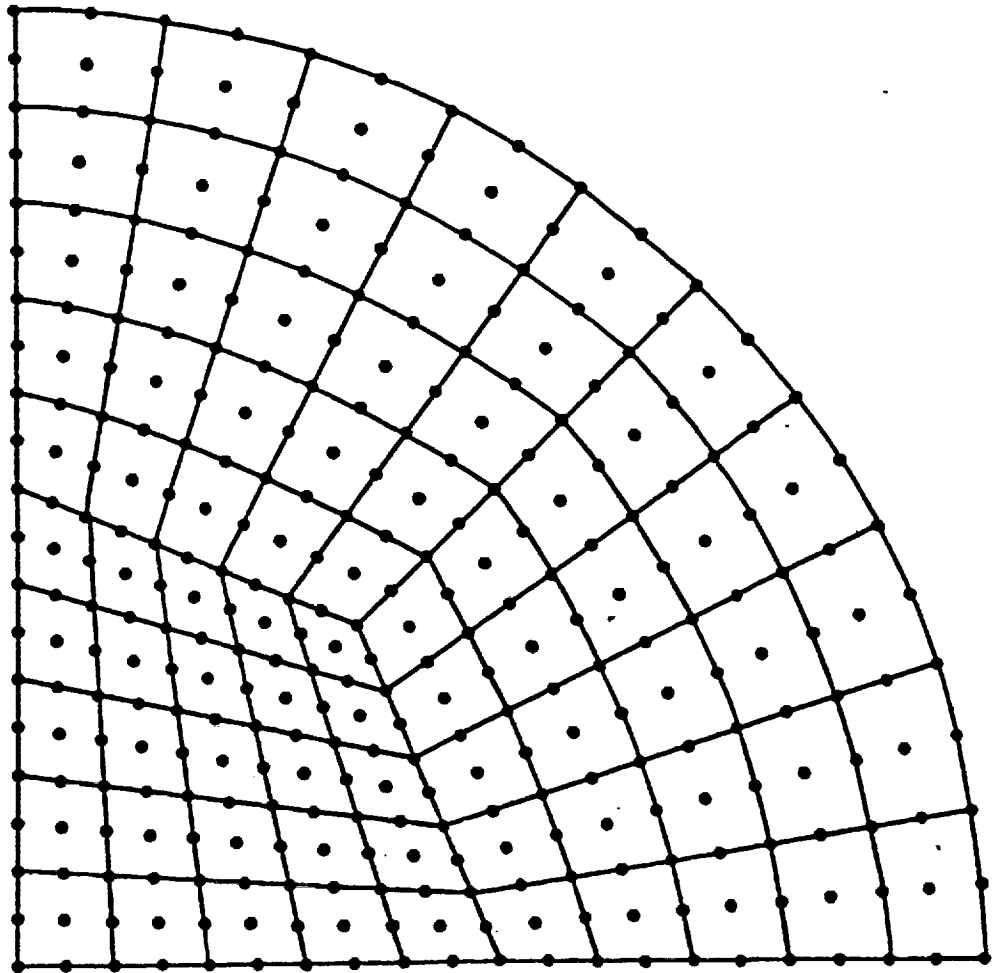
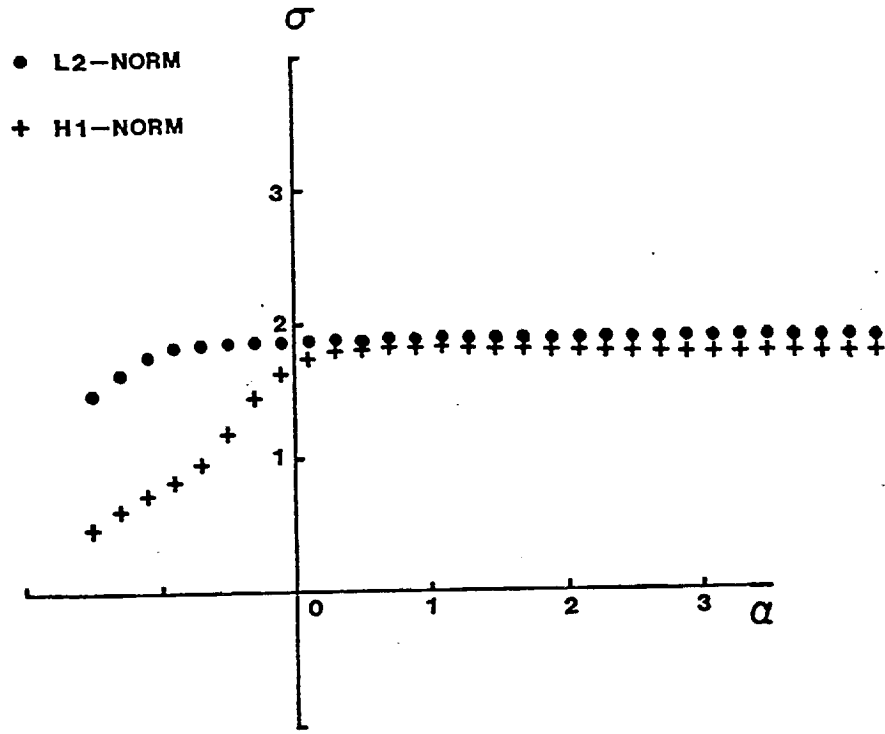
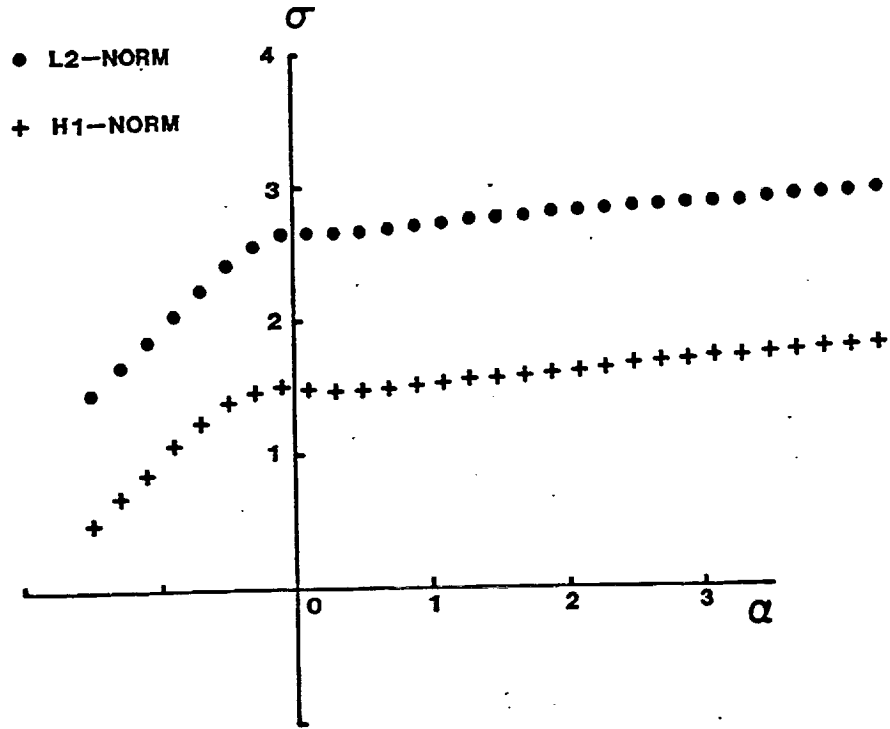


Figure 22. Typical mesh on a quarter circle domain.



a: Bilinear elements



b: Biquadratic elements

Figure 23. (α, σ) Plot for a quarter circle domain

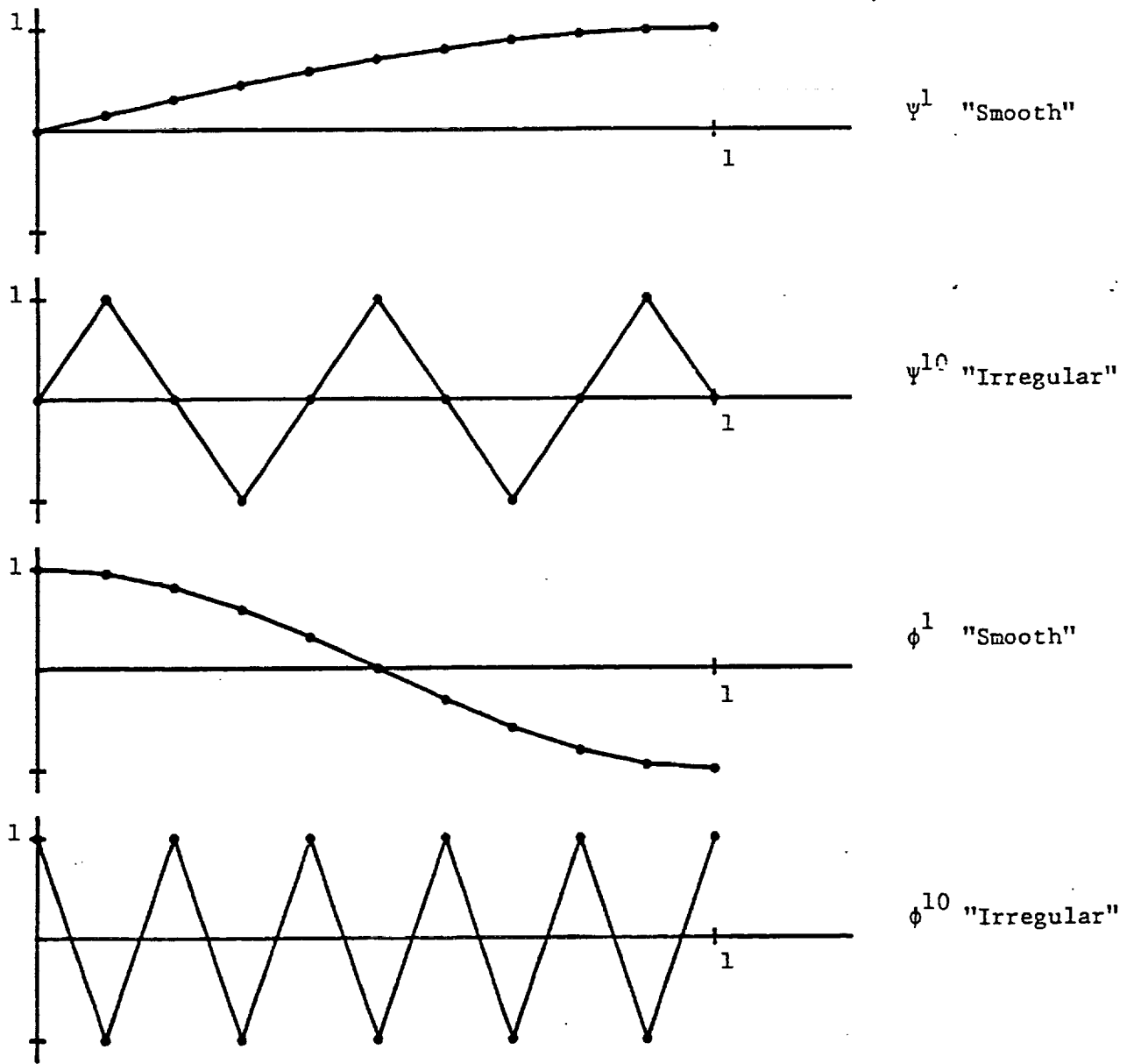


Figure 24. Examples of "Smooth" or "Irregular" Eigenfunctions.

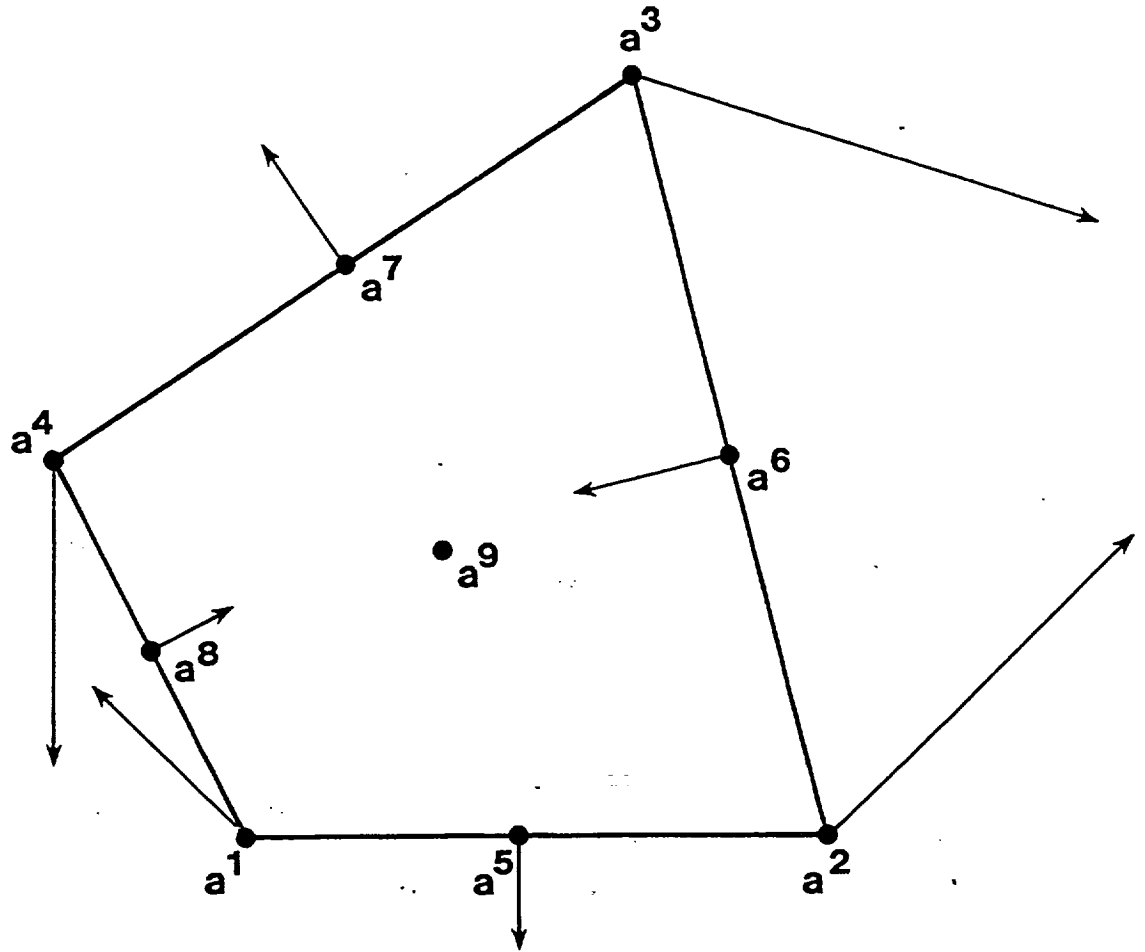
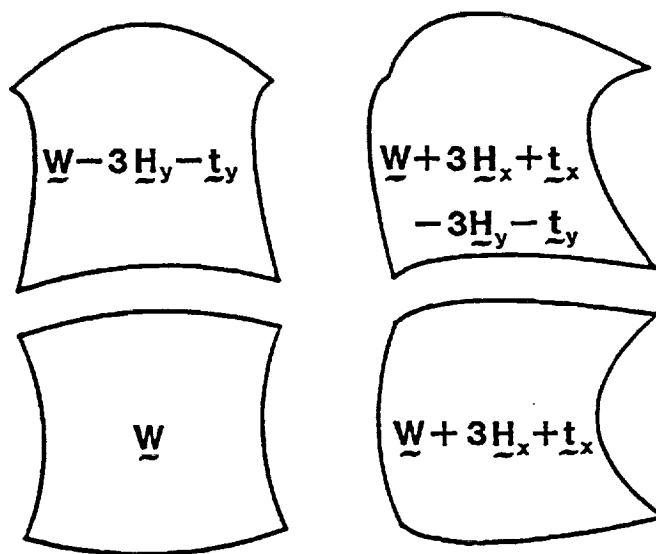
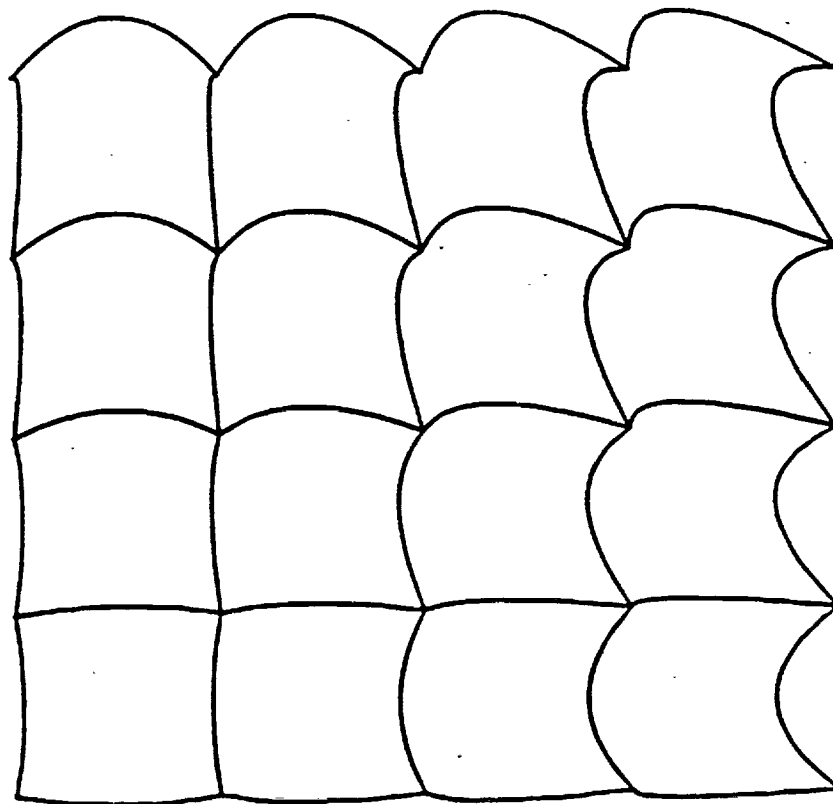


Figure 25. Spurious Mode W on a Quadrilateral Element.

26a. Construction of the Mode \underline{W} 26b. The spurious mode \underline{W} on a 16 element meshFigure 26: The Spurious Mode \underline{W}

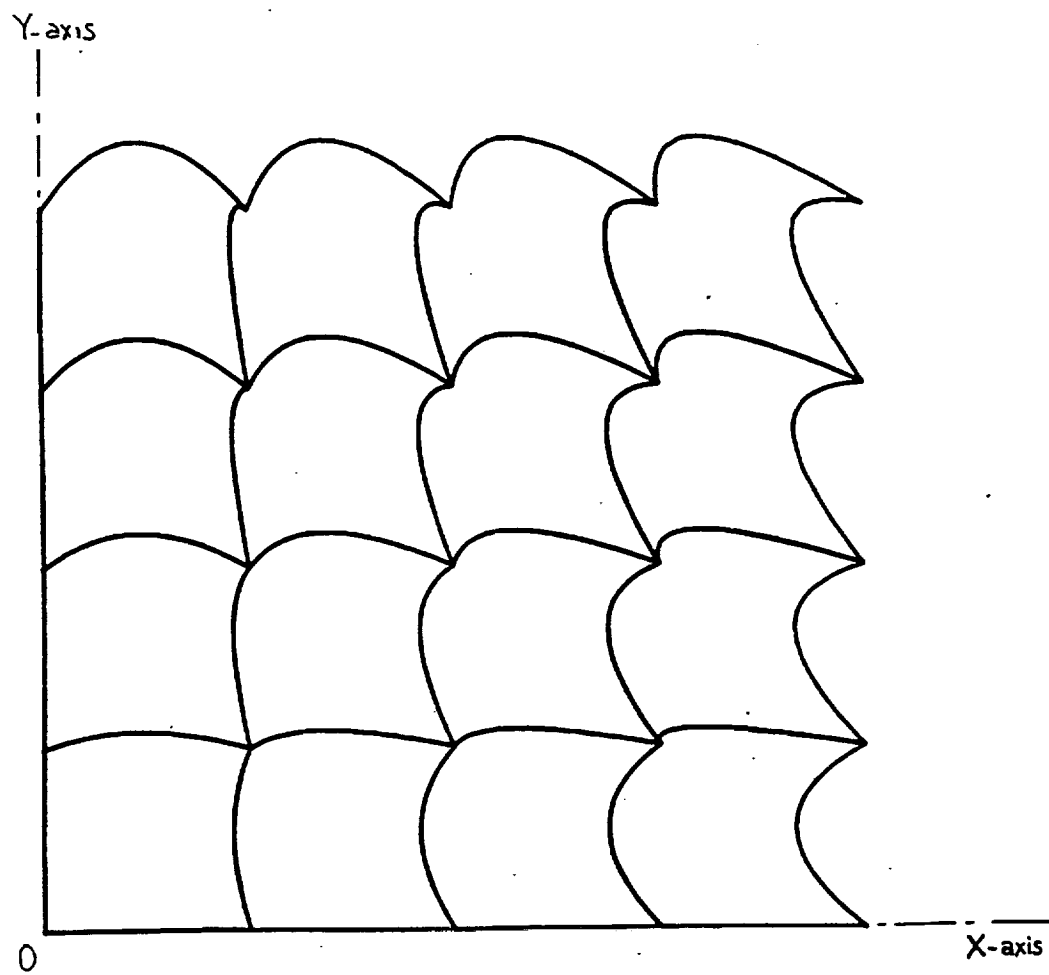


Figure 27. The Spurious mode $H_1 + H_2 + H_3$

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

| | | | |
|---|--|--|----------------------------|
| 1. AGENCY USE ONLY (Leave blank) | 2. REPORT DATE May 1995 | 3. REPORT TYPE AND DATES COVERED Final Contractor Report | |
| 4. TITLE AND SUBTITLE Analysis and Development of Finite Element Methods for the Study of Nonlinear Thermomechanical Behavior of Structural Components | | 5. FUNDING NUMBERS WU-505-63-5B G-NAG3-329 | |
| 6. AUTHOR(S) J. Tinsley Oden | | 8. PERFORMING ORGANIZATION REPORT NUMBER E-9005 | |
| 7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) University of Texas at Austin Aerospace Engineering and Engineering Mechanics Department Austin, Texas 78705 | | 10. SPONSORING/MONITORING AGENCY REPORT NUMBER NASA CR-195354 | |
| 9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Lewis Research Center Cleveland, Ohio 44135-3191 | | 11. SUPPLEMENTARY NOTES Project Manager, Christos C. Chamis, Structures Division, NASA Lewis Research Center, organization code 5200, (216) 433-3252. | |
| 12a. DISTRIBUTION/AVAILABILITY STATEMENT Unclassified - Unlimited Subject Category 39 This publication is available from the NASA Center for Aerospace Information, (301) 621-0390. | | 12b. DISTRIBUTION CODE | |
| 13. ABSTRACT (Maximum 200 words) Underintegrated methods are investigated with respect to their stability and convergence properties. The focus was on identifying regions where they work and regions where techniques such as hourglass viscosity and hourglass control can be used. Results obtained show that underintegrated methods typically lead to finite element stiffness with spurious modes in the solution. However, problems exist (scalar elliptic boundary value problems) where underintegrated with hourglass control yield convergent solutions. Also, stress averaging in underintegrated stiffness calculations does not necessarily lead to stable or convergent stress states. | | | |
| 14. SUBJECT TERMS Underintegration; Stability; Convergence; Hourglass control; Spurious modes; Constrained problems; Incompressibility | | 15. NUMBER OF PAGES 152 | |
| | | 16. PRICE CODE A08 | |
| 17. SECURITY CLASSIFICATION OF REPORT Unclassified | 18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified | 19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified | 20. LIMITATION OF ABSTRACT |

**National Aeronautics and
Space Administration**

Lewis Research Center
21000 Brookpark Rd.
Cleveland, OH 44135-3191

Official Business
Penalty for Private Use \$300

POSTMASTER: If Undeliverable — Do Not Return

