NUMERICAL ANALYSIS PROJECT
MANUSCRIPT NA-91-04

OCTOBER 1991

# Fast Iterative Solution of Stabilised Stokes Systems

# Part I: Using Simple Diagonal Preconditioners

by

Andrew Wathen
and
David Silvester

NUMERICAL ANALYSIS PROJECT
COMPUTER SCIENCE DEPARTMENT
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305

# Fast Iterative Solution of Stabilised Stokes Systems
## Part I: Using Simple Diagonal Preconditioners

*by*

**Andrew Wat hen**

**University of Bristol, UK**

**and**

**David Silvester**

**University of Manchester Institute of Science and Technology, UK**

*3rd October, 1991*

## Abstract

Mixed finite element approximation of the classical Stokes problem describing slow viscous incompressible flow gives rise to symmetric indefinite systems for the discrete velocity and pressure variables. Iterative solution of such indefinite systems is feasible and is an attractive approach for large problems. The use of stabilisation methods for convenient (but unstable) mixed elements introduces stabilisation parameters. We show how these can be chosen to obtain rapid iterative convergence.

We propose a conjugate gradient-like method (the method of preconditioned conjugate residuals) which is applicable to symmetric indefinite problems, describe the effects of stabilisation on the algebraic structure of the discrete Stokes operator and derive estimates of the eigenvalue spectrum of this operator on which the convergence rate of the iteration depends. Here we discuss the simple case of diagonal preconditioning. Our results apply to both locally and globally stabilised mixed elements as well as to elements which are inherently stable. We demonstrate that convergence rates comparable to that achieved using the diagonally scaled conjugate gradient method applied to the discrete Laplacian are approachable for the Stokes problem.

## 1. Introduction.

Mixed finite element solution of the Stokes equations describing slow incompressible viscous flow leads to a symmetric indefinite discrete system for the pressure and velocity components. Since such systems are usually large and their solution is frequently part of an (outer) iterative scheme to solve the Navier-Stokes equations( [GL]), rapid solution methods for such indefinite systems are desirable.

Given a flow domain $\Omega$ of $\Re^d (d = 2 \text{or} 3)$ with boundary $\partial\Omega$, a function f and appropriate boundary conditions, the classical form of the Stokes problem is to find the velocity u and pressure p satisfying

$$-\nu\nabla^2 \mathbf{u} + \text{gradp} = \text{f} \qquad \text{inn}, \tag{1.1}$$

$$\text{divu} = 0 \qquad \text{in}\,\Omega. \tag{1.2}$$

For simplicity, let

$$\mathbf{u} = 0 \qquad \text{on}\,\partial\Omega \tag{1.3}$$

representing 'no-flow' on the boundary. A weak form of the problem is obtained by multiplying (1.1) by an arbitrary test velocity $\text{v} \in \text{V}$ and (1.2) by an arbitrary test pressure $q \in P$ and integrating over $\Omega$. Here

$$P = L_0^2(\Omega) = \{q \mid q \in L^2(\Omega), \int_\Omega q d\Omega = 0\}, \qquad v = [H_0^1(\Omega)]^d \tag{1.4}$$

with

$$L^2(\Omega) = \{\phi \mid (\phi, \phi) < \infty\}, \qquad (\phi, \psi) = \int_\Omega \phi\psi d\Omega,$$

$$H_0^1(\Omega) = \{\phi | \phi \in L^2(\Omega), \frac{\partial\phi}{\partial x_i} \in L^2(\Omega), i = 1, \ldots, d, \phi = 0 \,\text{on}\,\partial\Omega\}.$$

The resulting weak form is: find $\text{u} \in \text{V}$ and $p \in P$ satisfying

$$\begin{aligned} (\text{gradu}, \text{gradv}) - (p, \text{divv}) &= (\text{f}, \text{v}) \qquad \forall \text{v} \in \text{V} \\ -(q, \text{divu}) &= 0 \qquad \forall q \in P \end{aligned} \tag{1.5}$$

Discretisation of (1.5) is achieved by introducing finite dimensional subspaces $\mathbf{V}_h \subset \mathbf{V}$ and $P_h \subset P$ and the discrete Stokes problem is then: find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in P_h$ such that

$$\begin{aligned} (\text{gradu}_h, \text{gradv}) - (p_h, \text{divv}) &= (\text{f}, \text{v}) \qquad \forall \text{v} \in \text{V}_h \\ -(q, \text{divu}_h) &= 0 \qquad \forall q \in P_h. \end{aligned} \tag{1.6}$$

Independently of the choice of $\mathbf{V}_h$ and $P_h$, the discrete system (1.6) can be written in block matrix form as

$$\begin{pmatrix} A & B^t \\ B & O \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ O \end{pmatrix} \tag{1.7}$$

where $u$ is a vector of the discrete velocity variables and $p$ a vector of the discrete pressure variables with respect to appropriate bases for $\mathbf{V}_h$ and $P_h$ respectively, $A$ represents the terms $(\mathrm{grad}\mathbf{u}_h, \mathrm{grad}\mathbf{v})$ and $B$ the coupling terms $-(q, \mathrm{div}\mathbf{u}_h)$ expressed in terms of these bases. The vector f represents the body force terms. The coefficient matrix is symmetric, but necessarily indefinite because of the zero diagonal block. The Dirichlet boundary conditions (1.3) for the velocity ensure that $A$ is positive definite, and by applying the congruence transform

$$\begin{pmatrix} A & B^t \\ B & O \end{pmatrix} = \begin{pmatrix} A & O \\ B & I \end{pmatrix} \begin{pmatrix} A^{-1} & O \\ O & -BA^{-1}B^t \end{pmatrix} \begin{pmatrix} A & B^t \\ O & I \end{pmatrix} \tag{1.8}$$

it is apparent that the coefficient matrix in *(1.7)* is non-singular if and only if $B$ has full row rank.

For a mixed finite element which is stable in the LBB sense ([BF], page 75), there exists a constant $\gamma > 0$ independent of the mesh spacing, $h$, such that

$$\sup_{\mathbf{u}_h \in \mathbf{V}_h - \{0\}} \frac{(p^h, \mathrm{div}\mathbf{u}_h)}{\|\mathbf{u}_h\|_{\mathbf{V}}} \geq \gamma \|p_h\|_P, \qquad \forall p_h \in P_h \tag{1.9}$$

where $\|.\|_{\mathbf{V}}$ and $\|.\|_P$ are norms in the underlying spaces V and $P$. Making the specific (and common) choice

$$\|\mathbf{u}_h\|_{\mathbf{V}} = (\mathrm{grad}\mathbf{u}_h, \mathrm{grad}\mathbf{u}_h)^{\frac{1}{2}}, \qquad \|p_h\|_P = (p_h, p_h)^{\frac{1}{2}}$$

leads to the matrix form

$$\max_{u \in \Re^n - \{0\}} \frac{p^t B u}{[u^t A u]^{\frac{1}{2}}} \geq \gamma [p^t M_p p]^{\frac{1}{2}}, \qquad \boldsymbol{vp} \in \Re^m. \tag{1.10}$$

Here $m$ is the total number of discrete pressure variables (the dimension of $P_h$), $n$ the total number of discrete velocity variables. Throughout the paper, $\Re^m$ is to be interpreted as excluding vectors corresponding to the hydrostatic pressure mode, i.e. vectors representing constant functions $p_h$ and $\Re^n$ is to be interpreted as discrete velocities satisfying the boundary condition (1.3). In (1.10), $M_p$ is the pressure mass matrix, i.e.

the Grammian matrix of basis functions for $P_h$. It is symmetric and positive definite and has condition number independent of $h$ for any usual finite element basis for $P_h$ ([F],[W]). Also, as $A$ is a discrete representation of a second order operator, the Dirichlet boundary conditions *(1.3)* ensure that there exists a positive constant C such that $u^tAu/u^tu \geq Ch^2$ for all u $\in \Re^n - \{0\}$ ([AB]). Thus (1.10) implies the existence of a positive constant C such that

$$\min_{p\in\Re^m-\{0\}} \max_{u\in\Re^n-\{0\}} \frac{p^tBu}{[p^tp]^{\frac{1}{2}}\,[u^tu]^{\frac{1}{2}}} \geq Ch \min_{p\in\Re^m-\{0\}} \frac{[p^tM_pp]^{\frac{1}{2}}}{[p^tp]^{\frac{1}{2}}}. \tag{1.11}$$

The left hand side of (1.11) is a definition of the smallest singular value of $B$ ([GVL]), hence $B$ is of full rank and we see that the Stokes system (1.7) for an LBB stable element is uniquely solvable on any finite computational grid.

Conversely, if

$$\min_{p\in\Re^m-\{0\}} \max_{u\in\Re^n-\{0\}} \frac{p^tBu}{[p^tp]^{\frac{1}{2}}\,[u^tu]^{\frac{1}{2}}} = 0 \tag{1.12}$$

then $B$ is rank deficient since its smallest singular value (at least) is zero. Returning to the system (1.7), we see that the coefficient matrix for the Stokes problem is singular with null vectors which have only non-zero entries in the discrete pressure variables, $p$. These null vectors are precisely the spurious pressure modes of which the 'chequerboard' mode for the $Q_1$–$P_0$ element is a well known example.

We can learn more about stable approximations from the LBB condition than just the above: using the 'discrete' LBB condition (1.10) we have that for any $p \in \Re^m$

$$\gamma\,[p^tM_pp]^{\frac{1}{2}} \leq \max_{u\in\Re^n-\{0\}} \frac{p^tBu}{[u^tAu]^{\frac{1}{2}}} = \max_{z=A^{\frac{1}{2}}u\neq 0} \frac{p^tBA^{\frac{1}{2}}z}{[z^tz]^{\frac{1}{2}}}. \tag{1.13}$$

The maximum is attained when $z = (p^tBA^{\frac{1}{2}})^t$ and gives the value

$$(p^tBA^{-1}B^tp)^{\frac{1}{2}}.$$

Thus

$$\gamma^2\,\frac{p^tM_pp}{p^tp} \leq \frac{p^t(BA^{-1}B^t)p}{p^tp} \qquad \forall p \in \Re^m - \{0\}. \tag{1.14}$$

Similarly, whether we use a stable element or not, if we assume boundedness of $B$, i.e. that there exists $\Gamma$ with

$$p^tBu \leq \Gamma\,[p^tM_pp]^{\frac{1}{2}}\,[u^tAu]^{\frac{1}{2}} \qquad \forall u \in \Re^n \text{ and } \forall p \in \Re^m \tag{1.15}$$

then

$$\Gamma\,[p^t M_p p]^{\frac{1}{2}} \geq \max_{u \in \Re^n - \{0\}} \frac{p^t B u}{[u^t A u]^{\frac{1}{2}}} = \max_{z = A^{\frac{1}{2}} u \neq 0} \frac{p^t B A^{\frac{1}{2}} z}{[z^t z]^{\frac{1}{2}}} \qquad (1.16)$$
$$= (p^t B A^{-1} B^t p)^{\frac{1}{2}}.$$

So for an LBB stable element

$$\gamma^2 \leq \frac{p^t (B A^{-1} B^t) p}{p^t M_p p} \leq \Gamma^2 \qquad \forall p \in \Re^m - \{0\}. \qquad (1.17)$$

For an unstable element, the upper bound (only) in (1.17) holds. Since the spectral condition number of $M_p$ is independent of the grid size $h$, (1.17) simply states that the 'Schur complement' matrix $BA^{-1}B^t$ has spectral condition number independent of $h$ for any stable element. Hence any system with $BA^{-1}B^t$ as coefficient matrix may be rapidly solved by a conjugate gradient or other iterative method (see e.g. [GVL]). This has lead a number of authors to propose nested iterative solution strategies based on the block factorisation (1.8). See [V],[BP],[BWY].

In this paper, however, we address the possibility of a single non-nested iterative solution of Stokes systems. In particular, we consider the important effect of stabilisation on the convergence of such an iteration for both unstable as well as stable elements. In section 2 we review a Krylov space (conjugate-gradient-like) method which is applicable to symmetric indefinite systems (the Preconditioned Conjugate Residual method). The relevent convergence analysis reveals that a certain minimax polynomial approximation problem on the eigenvalue spectrum of the coefficient matrix describes the rate of convergence of the iteration in an analogous way to the positive definite case. Section 3 covers the analytic description of global and local stabilisation strategies for unstable mixed elements. In Section 4 we establish estimates for the eigenvalue spectrum of the stable and stabilised Stokes operator with simple diagonal preconditioning. In section 5 we present computational results obtained with the Preconditioned Conjugate Residual method in the three cases of a stable element, a globally stabilised element and a locally stabilised element, and relate these to our analytic estimates. With regard to these results and the characterisation of iterative convergence given by the polynomial approximation problem described in section 2, we consider in each case 'good' choices of stabilisation parameters which ensure rapid convergence of the Preconditioned Conjugate Residual method.

## 2. Iterative Methods for Indefinite Systems

The applicability and efficiency of Conjugate Gradient methods ([GVL]) for solving symmetric positive definite systems is widely appreciated. For symmetric indefinite problems such as the discrete Stokes problem, block elimination yields a definite Schur complement (for a stable element) and back substitution gives also a definite system, hence using nested iteration ([BWY],[BP]) is an attractive approach. However, Conjugate Gradient methods exist for indefinite systems also, and a non-nested iterative solution of the Stokes problem is perfectly feasible using, for example, the method of Preconditioned Conjugate Residuals (PCR) which is alternatively called the MINRES algorithm ([AMS]).

The PCR algorithm for solving $\mathcal{A}x = b$ with symmetric indefinite d and symmetric positive definite preconditioning matrix $M$ is expressable in either the Orthodir or Orthomin forms ([JY],[AMS]). The robust Orthodir form, for example, is

$$r_0 = b - \mathcal{A}x_0, p_0 = r_0$$

$$\alpha_i = p_i^t \mathcal{A}\mathcal{M}^{-1} r_i / p_i^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A} p_i$$

$$x_{i+1} = \text{Xi} + \alpha_i p_i$$

$$r_{i+1} = r_i - \alpha_i \mathcal{A} p_i$$

$$\gamma_i = p_i^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A}\mathcal{A} p_i / p_i^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A} p_i$$

$$\sigma_i = p_{i-1}^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A}\mathcal{A} p_i / p_{i-1}^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A} p_{i-1}$$

$$p_{i+1} = \mathcal{A} p_i - \gamma_i p_i - \sigma_i p_{i-1}.$$

This method can be implemented with only two matrixxvector products at each iteration. Only a single matrixxvector product is needed at each iteration in general if a hybrid Orthomin/Orthodir form is used ([AMS]).

If we define the norm

$$|||y||| = y^t \mathcal{A}\mathcal{M}^{-1} \mathcal{A} y,$$

then the PCR iterates have the property that

$$|||x - x_k||| \leq |||x - y|||$$

for all y in the (affine) Krylov space

$$x_0 + \text{span}\{\mathcal{M}^{-1}r_0, (\mathcal{M}^{-1}\mathcal{A})\mathcal{M}^{-1}r_0, \ldots, (\mathcal{M}^{-1}\mathcal{A})^{k-1}\mathcal{M}^{-1}r_0\}.$$

Thus if $\Pi_{k-1}$ is the set of real polynomials of degree $k-1$, we have

$$\|x - x_k\| = \min_{p \in \Pi_{k-1}} \|\mathcal{A}^{-1}r_0 - p(\mathcal{M}^{-1}\mathcal{A})\mathcal{M}^{-1}r_0\| \tag{2.2}$$

¿from which we may further deduce that

$$\|x - x_k\| \leq \min_{p \in \Pi_k^1} \max_{i \in \{1,\ldots,N\}} |p(\lambda_i)| \, \|r_0\|_{\mathcal{M}^{-1}} \tag{2.3}$$

where $\{\lambda_i : i = 1,\ldots, N\}$ are the eigenvalues of $\mathcal{M}^{-1}\mathcal{A}$, $\Pi_k^1$ is the set of $k$th degree polynomials with constant term one and $\|y\|_{\mathcal{M}^{-1}} = y^t\mathcal{M}^{-1}y$. The $\lambda_i$ are real since positive definiteness of $M$ implies that $\mathcal{M}^{1/2}$ exists and so $\mathcal{M}^{-1}\mathcal{A}$ is similar to the symmetric matrix $\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}$. The analysis is entirely analogous to that used in the case of positive definite $\mathcal{A}$ (see for example Axelsson & Barker [AB]).

The convergence estimate (2.3) indicates that the convergence of the PCR iteration depends crucially on the spectrum of $\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}$. The PCR method can be applied when $\mathcal{A}$ is positive definite, in which case the spectrum is contained in a real positive interval $[\lambda_{\min}, \lambda_{\max}]$ (using an obvious notation), and the estimate (2.3) can be expressed in terms of shifted Chebyshev polynomials, which are the minimax polynomials on this interval. Using this approach, one can derive the estimate

$$\|x - x_k\| \leq 2\left(\frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}}\right)^{2k} \|x - x_0\| \tag{2.4}$$

for the PCR algorithm where $\kappa = \frac{\lambda_{\max}(\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}})}{\lambda_{\min}(\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}})}$ is the spectral condition number. The corresponding estimate for the common Preconditioned version of the Hestenes-Stiefel CG algorithm, which is applicable when $\mathcal{A}$ and $M$ are both symmetric and positive definite, is

$$\|x - x_k\|_{\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}} \leq 2\left(\frac{1 - \sqrt{\kappa}}{1 + \sqrt{\kappa}}\right)^{2k} \|x - x_0\|_{\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}} \tag{2.5}$$

where $\|y\|_{\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}} = y^t\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}}y$. (Positive definiteness of $M$ is not in fact necessary, see [AMS]). The Hestenes-Stiefel algorithm requires a single matrix×vector

product at each iteration and thus is the method of choice in the positive definite case. If there are significant gaps in the spectrum of A, then the Chebyshev estimate (2.5) is pessimistic and more accurate estimates can be derived by considering the polynomial minimax approximation problem in (2.3) on disjoint subintervals which contain the spectrum ([AB],[G],[R]).

When $\mathcal{A}$ is indefinite (but nonsingular), PCR is still applicable, but the Hestenes-Stiefel method can fail. In this case the spectrum is contained in $[-b,-a] \cup [a, b]$ for some $a, b > 0$ and again analytic expressions for the minimax *errors* on this set in terms of Chebyshev polynomials are available ([L]). However, such an estimate is most likely to be pessimistic unless the eigenvalues are essentially symmetric about the origin. In section 5 we show minimax polynomials on the discrete eigenvalue sets which come from some of the specific problems considered in section 4.

The role of preconditioning for conjugate gradient solution of symmetric matrix equations is to make the spectrum more clustered: this is often replaced by the simpler goal of reducing the condition number $\kappa$. Unless there is some symmetry in the eigenvalues about the origin which is preserved under preconditioning, then the simple reduction of $\kappa$ seems a less appropriate goal in the indefinite case.

In section 4 we will establish inclusion intervals for the eigenvalues of various indefinite discrete representations of the Stokes problem when only an elementary diagonal scaling is used. More sophisticated preconditioning strategies are developed in [SW]. The simple diagonal preconditioning case serves to illustrate the difference between various elements, and in particular to highlight the fact that stabilisation has an effect on iterative convergence.

## 3. Stabilisation.

In this section, the idea of 'regularising' the discrete Stokes problem (1.6) as a means of ensuring the compatibility of arbitrary mixed approximations is reviewed. Our objective here is to summarise the theoretical framework covering the case of low-order elements like the $P_1$–$P_0$ and $P_1$–$P_1$ triangle or tetrahedron which are not stable in a conventional (Babuška-Brezzi) sense. For a more complete discussion including the generalisation to higher-order mixed approximations, see the recent review [FHS].

With the notation (1.4), finite element subspaces of V and $P$ are characterised by $\tau_h$, a partitioning of $\overline{\Omega}$ into triangles/quadrilaterals or tetrahedra/hexahedra, assumed to be regular in the usual sense. The mesh parameter $h$ is given by $h = \max(h_K)$ where $h_K$ is the diameter of element K. The set of all interelement boundaries (edges in $\Re^2$,faces in $\Re^3$) will be denoted by $\Gamma_h$, and the length of edge $e \in \Gamma_h$ in $\Re^2$ or the diameter of face $e \in \Gamma_h$ in $\Re^3$ will be denoted by $h_e$.

We assume below for ease of exposition, that the discrete velocity space $\mathbf{V}_h$ is either piecewise linear (in the triangle or tetrahedral case) or else is the usual bilinear/trilinear isoparametric approximation (in the quadrilateral/hexahedral case). The generalisation to higher-order mixed approximations is straightforward. Once the approximations $\mathbf{V}_h$ and $P_h$ have been defined, a *stabilised* discrete formulation of the Stokes problem is: find $\mathbf{u}_h \in \mathbf{V}_h$ and $p_h \in P_h$ such that

$$\begin{aligned} (\mathrm{grad}\mathbf{u}_h, \mathrm{grad}\mathbf{v}) - (p_h, \mathrm{div}\mathbf{v}) &= (\mathbf{f}, \mathbf{v}) \qquad \forall \mathbf{v} \in \mathbf{V}_h \\ -(q, \mathrm{div}\mathbf{u}_h) - \beta C_h(p_h, q) &= 0 \qquad \forall q \in P_h \end{aligned} \tag{3.1}$$

Where $\beta > 0$ is the so-called stabilisation parameter, and $C_h(\cdot,\cdot)$ is a symmetric continuous bilinear form which is positive semi-definite on $P_h \times P_h$, and which satisfies a weak stabilisation *condition:*

$$(p, \mathrm{div}\mathbf{v}) = 0 \quad \forall \mathbf{v} \in \mathbf{V}_h \qquad \Rightarrow \qquad C_h(p,p) \neq 0. \tag{3.2}$$

Ideally the stabilisation should be set up so that it satisfies a consistency *condition:* any classical solution $(\mathbf{u}, p)$ satisfying (1.1),(1.2),(1.3) must also satisfy (3.1) for *all* values of $h$ and for any $\beta > 0$. Note that such a consistency condition excludes standard perturbations of the Stokes system, so-called penalty methods, which are not true stabilisation methods (even though the perturbed system satisfies (3.2)). In section 4 we derive some theoretical results which demonstrate why the use of iterative solvers applied to 'penalised systems' is doomed to failure.

Expressed in matrix form the system (3.1) is

$$\begin{pmatrix} A & B^t \\ B & -\beta C \end{pmatrix} \begin{pmatrix} u \\ p \end{pmatrix} = \begin{pmatrix} f \\ g \end{pmatrix} \tag{3.3}$$

where C is a symmetric positive semi-definite 'stabilisation matrix'.

We can justify the use of the term 'regularisation' by considering the eigenvalue distribution of the systems (1.7) and (3.3). In particular, constructing the congruence transformation

$$\begin{pmatrix} A & B^t \\ B & -\beta C \end{pmatrix} = \begin{pmatrix} A & O \\ B & I \end{pmatrix} \begin{pmatrix} A^{-1} & O \\ O & -BA^{-1}B^t - \beta C \end{pmatrix} \begin{pmatrix} A & B^t \\ O & I \end{pmatrix} \quad (3.4)$$

and applying Sylvester's law of inertia, it is clear that the coefficient matrix in (3.3) is always non-singular, whereas the coefficient matrix in (1.7) will have zero eigenvalues whenever the matrix $B$ is rank deficient (giving rise to instability associated with spurious pressure modes). These zero eigenvalues are transformed into strictly negative eigenvalues by the stabilisation condition (3.2). In simple terms the condition (3.2) ensures that the matrix C automatically filters any 'spurious' pressures. Two types of stabilised formulation of the form (3.1) will be distinguished.

Starting from a mixed approximation based on a *continuous* pressure, the obvious route to stabilisation is by adding a stabilisation term $C_h(\cdot,\cdot)$ which controls gradients in pressure:

$$C_h(p_h, q_h) = \beta \sum_{K \in \tau_h} h_K^2 \int_K \int \mathrm{grad} p_h \cdot \mathrm{grad} q_h \, \mathrm{d}K \quad (3.5)$$

in $\Re^2$ for example. This type of regularisation was first suggested by Brezzi and Pitkäranta [BPi] in the context of the $P_1$–$P_1$ triangular element. The generalisation of (3.5) to cover higher order elements is constrained by the fact that the consistency condition is not satisfied unless the stabilised formulation is generalised to a full 'least squares formulation', as is done implicitly in [HF]. Continuous pressure elements which depend on internal velocity bubble functions for their stability, for example the popular 'mini element' introduced by Arnold et al in [ABF], can also be expressed as a stabilised method of the form (3.1) with $C_h(\cdot,\cdot)$ given by (3.5) (after static condensation of the bubble terms).' Note that in this case the magnitude of $\beta$ cannot be arbitrarily chosen, it is fixed by the underlying mixed approximation. See [P] for details in the mini element case. An important feature of (3.5) is the 'global' nature of the stabilisation. The point is that (3.5) represents an approximation to the Laplacian of the pressure, defined over the entire domain $\Omega$.

Mixed approximations based on discontinuous pressure can be stabilised in a similar way to that above by means of a stabilisation term $C_h(\cdot,\cdot)$ which controls inter-element

jumps in pressure. Stabilisation methods of this type were first introduced by Hughes and Franca [HF], and correspond to the following definition of the stabilisation term

$$C_h(p_h, q_h) = \beta \sum_{e \in \Gamma_h} h_e \int_e [\![p_h]\!]_e [\![q_h]\!]_e \, ds \qquad (3.6)$$

which, in the limit of $\beta \to \infty$, leads to a *globally continuous* pressure solution. Here $[\![\cdot]\!]_e$ represents the jump across edge or face e. To see that this type of stabilisation is also 'global', consider using *(3.6)* to stabilise a piecewise constant pressure approximation defined on a uniform grid of square elements. As the contributions to a particular element pressure comprise differences with the four neighbouring pressures, the resulting stabilisation matrix C is the standard five-point finite difference approximation to the Laplacian. Henceforth, both *(3.5)* and *(3.6)* till be referred to as *global stabilisation* methods.

A second general class of stabilisation methods, which are particularly applicable in the case of discontinuous pressure approximations, are those based on macroelements. They will be referred to here as *local stabilisation* methods. Given any subdivision $\tau_h$, a macroelement partitioning $M_h$ is defined to be a union of macroelements $M$, each macroelement being a union of neighbouring elements from $\tau_h$, such that the interior is simply connected. In addition, every element $K$ must he in exactly one macroeleme,nt, which implies that macroelements do not overlap. For each $M$, the set of interelement boundaries which are strictly in the interior of $M$ is denoted by $\Gamma_M$. A typical example of a locally stabilised method is that developed by Kechkar and Silvester [KS] in the case of piecewise constant pressure:

$$C_h(p_h, q_h) = \beta \sum_{M \in \mathcal{M}_h} \sum_{e \in \Gamma_M} h_e \int_e [\![p_h]\!]_e [\![q_h]\!]_e \, ds \qquad (3.7)$$

which, in the limit of $\beta \to \infty$, leads to a *locally* (ie. within macroelements) *continuous* pressure solution.

Note that if the space $N_h$ of spurious pressure modes:

$$N_h = \{q_h \in P_h; \quad (q_h, \mathrm{div}\mathbf{v}) = 0 \ \forall \mathbf{v} \in \mathbf{V}_h\}$$

can be explicitly characterised (for all values of $h$), then a very natural way of generating a consistent stabilisation of the form (3.1) is to define a stabilisation term such that:

$$C_h(p_h, q_h) = \beta(\Pi_h p_h, \Pi_h q_h) \qquad (3.8)$$

Here $\Pi_h$ denotes the orthogonal projection of the space $P_h$ onto the space $N_h$, so that the stabilisation condition *(3.2)* is automatically satisfied. When the space $N_h$ is one dimensional, e.g. in the case of the $Q_1$–$P_0$ rectangular element, constructing a $C_h$ satisfying *(3.8)* is straightforward. Furthermore if this projection can be done locally (e.g. using a $2 \times 2$ macro-element construction), then (3.8) is also a local stabilisation. For further details in the $Q_1$–$P_0$ *case, see* [PS]. In more general cases however, for example stabilising the $P_1$–$P_0$ triangle, this idea is perhaps not quite so straightforward to implement.

The stabilisation condition *(3.2)* is a necessary condition for a well-posed discrete problem since it ensures that all the eigenvalues of the matrix in *(3.3)* are non-zero, so a solution of the discrete system always exists and is unique. In simple terms the condition *(3.2)* establishes the existence of a strictly positive Babuška-Brezzi constant for any fixed value of *h;* in practice we need to ensure that this constant is independent of *h,* which implies that a slightly stronger stabilisation condition must be satisfied. In this work we will need to make a precise definition of such a condition, in particular for any specific choice of $C_h$ we assume that there exists constants $\gamma,\Gamma$ independent of *h* such that the *(uniform-)* stabilisation *condition:*

$$\gamma^2 \leq \frac{p^t(BA^{-1}B^t + \beta C)p}{p^t M_p p} \leq \Gamma^2, \qquad \forall p \in \Re^m - \{0\} \tag{3.9}$$

is satisfied for all $\beta > 0$. Note that this is the obvious extension of *(1.17)* to the stabilised *case.*

As stated earlier our objective in this paper is to analyse the effect of stabilisation on the convergence of iterative solvers applied to the discrete system *(3.3).* The point of our analysis is to show how to choose a stabilisation parameter $\beta$ which is optimal in a fast solution sense. Numerical experience (see for example [SK]) shows that if fast solution is the objective, then there is a clear advantage in using a local stabilisation approach ((3.7) or (3.8) above), rather than a global stabilisation approach such as *(3.5).* The basic problem with a global stabilisation is that the numerical solutions tend to be quite sensitive to the particular choice of $\beta$, and in particular, the accuracy deteriorates in the limit of arbitrarily large $\beta$. Some examples of this in the case of $P_1$–$P_1$ stabilised via (3.5), and in the $Q_1$–$P_0$ case stabilised via (3.6), are given in [P] and [SK]

respectively. This limitation does not apply when using a locally stabilised method, in which case there is far greater scope for tuning the magnitude of $\beta$ to improve the rate of convergence of the iterative solver, without adversely affecting accuracy. This issue is discussed in more detail in section 5.

## 4. Eigenvalue Estimates.

We are interested in solving systems with the symmetric coefficient matrix

$$\mathcal{A} = \begin{pmatrix} A & B^t \\ B & -\beta C \end{pmatrix} \tag{4.1}$$

where $A$ represents a block diagonal matrix of discrete Laplacians, $\beta$ is the stabilisation parameter, -C the stabilisation matrix, and $B$ the coupling terms between velocities and pressure. $A$ is positive definite and C is positive semi-definite. We will denote by $n$ the order of the square matrix $A$ and by $m$ the order of C. (Thus $B$ is $m \times n$). In all practical cases $n > m$. For a stable element, we may simply take C = 0 throughout this section.

We choose the positive definite preconditioner

$$\mathcal{M} = \begin{pmatrix} D_A & O \\ O & \beta D_C \end{pmatrix} \tag{4.2}$$

where $D_A$ = diag($A$), $D_C$ = diag( C) if C $\neq$ 0, else $D_C = h^d I$ where $d$ is the spatial dimension as before. Certainly $D_A$ is positive definite with (diagonal) entries of O(1) in $\Re^2$ and $O(h)$ in $\Re^3$. For all types of stabilisation described in the previous section, $D_C$ is also positive definite with the diagonal entries being $O(h^2)$ in 2-dimensions, and $O(h^3)$ in 3-dimensions. The definition of $D_C$ in the stable case is designed to satisfy a corresponding scaling with the mesh size. The important point is simply that $D_C$ be spectrally equivalent to the pressure mass matrix, i.e. that there exist constants $\theta, \Theta$ independent of $h$ such that

$$\theta^2 \leq \frac{p^t M_p p}{p^t D_C p} \leq \Theta^2, \qquad vp \in \Re^m - \{0\}. \tag{4.3}$$

Using the results of [W], $D_C$ = diag($M_p$) satisfies (4.3) and is thus another reasonable choice for both stable and unstable elements. As the parameter $\beta$ does not arise in (4.1) in the stable case and the scaling of $D_C$ with $h$ is prescribed, it is not appropriate to take $\beta$ to be other than unity in the case of an unstabilised stable element.

Recalling the convergence estimate (2.3) we are interested in the optimal minimax polynomials of increasing degree on the eigenvalue spectrum of the diagonally scaled matrix

$$\mathcal{M}^{-\frac{1}{2}}\mathcal{A}\mathcal{M}^{-\frac{1}{2}} = \begin{pmatrix} D_A^{-\frac{1}{2}}AD_A^{-\frac{1}{2}} & \frac{1}{\sqrt{\beta}}D_A^{-\frac{1}{2}}B^tD_C^{-\frac{1}{2}} \\ \frac{1}{\sqrt{\beta}}D_C^{-\frac{1}{2}}BD_A^{-\frac{1}{2}} & -D_C^{-\frac{1}{2}}CD_C^{-\frac{1}{2}} \end{pmatrix} = \begin{pmatrix} \tilde{A} & \frac{1}{\sqrt{\beta}}\tilde{B}^t \\ \frac{1}{\sqrt{\beta}}\tilde{B} & -\tilde{C} \end{pmatrix} = \tilde{\mathcal{A}}. \tag{4.4}$$

(Note $\tilde{C} = 0$ in the unstabilised case).

By applying Sylvester's Law of Inertia to the congruence transform of $\tilde{\mathcal{A}} - \xi I$

$$\begin{pmatrix} I & O \\ \frac{1}{\sqrt{\beta}}\tilde{B}(\tilde{A}-\xi I)^{-1} & I \end{pmatrix}\begin{pmatrix} \tilde{A}-\xi I & O \\ O & -\frac{1}{\beta}\tilde{B}(\tilde{A}-\xi I)^{-1}\tilde{B}^t - \tilde{C} - \xi I \end{pmatrix}\begin{pmatrix} I & \frac{1}{\sqrt{\beta}}(\tilde{A}-\xi I)^{-1}\tilde{B}^t \\ O & I \end{pmatrix} \tag{4.5}$$

for $\xi = 0$ we see that the inertia of $\tilde{\mathcal{A}}$ is invariant with $h > 0$ and $\beta > 0$ provided C (equivalently $C_h$) satisfies the *weak stabilisation* condition (3.2) or the element is stable. This follows since in algebraic form (3.2) is

$$p^tBv = 0 \qquad \forall v \in \Re^n \qquad \Rightarrow \qquad p^tCp \neq 0 \tag{4.6}$$

or in scaled form

$$p^t\tilde{B}v = p^tD_C^{-\frac{1}{2}}BD_A^{-\frac{1}{2}}v = 0 \qquad \forall w \in \Re^n \qquad \Rightarrow \qquad p^tD_C^{-\frac{1}{2}}CD_C^{-\frac{1}{2}}p = p^t\tilde{C}p \neq 0. \tag{4.7}$$

We thus denote the eigenvalues of $\tilde{\mathcal{A}}$ by

$$\mu_{-m} \leq \mu_{-m+1} \leq \ldots \mu_{-1} \leq 0 < \mu_1 \leq \mu_2 \leq \ldots \leq \mu_n. \tag{4.8}$$

Correspondingly, we denote the eigenvalues of $\tilde{A}$ by $(0 <) \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ and the eigenvalues of $-\tilde{C}$ by $\lambda_{-m} \leq \lambda_{-m+1} \leq \ldots \leq \lambda_{-1} (\leq 0)$. Note $\lambda_{-i} = 0$, $i = 1, \ldots, m$ for an unstabilised method. We also (break the usual convention to consistently) write the singular values of $\tilde{B}$ as $(0 \leq) \sigma_1 \leq \sigma_2 \leq \ldots \leq \sigma_m$.

Since $\tilde{A}$ is a set of discrete Laplacians, it follows that $\lambda_1 = O(h^2)$, $\lambda_n = O(1)$ for small $h$ (see e.g. [AB]). Our first task is to prove that $\sigma_m$ is bounded independent of $h$.

¿From (1.17), whether we are using a stable or unstable element, we have

$$\frac{p^t(BA^{-1}B^t)p}{p^tM_pp} \leq \Gamma^2 \qquad \forall p \in \Re^m - \{0\}. \tag{4.9}$$

Also, by construction, $D_C$ satisfies (4.3), so

$$\frac{p^t(BA^{-1}B^t)p}{p^tD_Cp} = \frac{q^tD_C^{-\frac{1}{2}}BA^{-1}B^tD_C^{-\frac{1}{2}}q}{q^tq} \leq \Gamma^2\Theta^2 \tag{4.10}$$

for all $q = D_C^{\frac{1}{2}}p(\neq 0)$. Further, since $\lambda_{\max}(D_A^{-1}A) = O(1)$, there exists $\Upsilon$ with

$$\Upsilon^2 \leq \frac{u^tA^{-1}u}{u^tD_A^{-1}u} \qquad vu \in \Re^n - \{0\}. \tag{4.11}$$

Thus in (4.10), for *all* $q$ with $B^tD_C^{-\frac{1}{2}}q \neq 0$, we have

$$\frac{q^tD_C^{-3}BA^{-1}B^tD_C^{-\frac{1}{2}}q}{q^tD_C^{-\frac{1}{2}}BD_A^{-1}B^tD_C^{-\frac{1}{2}}q} \quad \frac{q^tD_C^{-\frac{1}{2}}BD_A^{-1}B^tD_C^{-\frac{1}{2}}q}{q^tq} \leq \Gamma^2\Theta^2$$

which using (4.11) and the definition (4.4) of $\widetilde{B}$ gives

$$\frac{q^t\widetilde{B}\widetilde{B}^tq}{q^tq} \leq \frac{\Gamma^2\Theta^2}{\Upsilon^2}. \tag{4.12}$$

For $q \neq 0$ with $B^tD_C^{-\frac{1}{2}}q = 0$, (4.12) is trivially satisfied, so (4.12) holds for all $q \in \Re^m - \{0\}$. Thus, since the singular values of $\widetilde{B}$ are the square roots of the eigenvalues of $\widetilde{B}\widetilde{B}^t$, it follows that $\sigma_m \leq \Gamma\Theta/\Upsilon$. That is, $\sigma_m$ is bounded independent of $h$.

We now write

$$\begin{pmatrix} \widetilde{A} & \frac{1}{\sqrt{\beta}}\widetilde{B}^t \\ \frac{1}{\sqrt{\beta}}\widetilde{B} & -\widetilde{C} \end{pmatrix} = \begin{pmatrix} \widetilde{A} & 0 \\ 0 & -\widetilde{C} \end{pmatrix} + \frac{1}{\sqrt{\beta}}\begin{pmatrix} 0 & \widetilde{B}^t \\ \widetilde{B} & 0 \end{pmatrix} \tag{4.13}$$

and apply the minimax result on the eigenvalues of the sum of two symmetric matrices ([Wi], page 101-102). The eigenvalues of

$$\begin{pmatrix} 0 & \widetilde{B}^t \\ \widetilde{B} & 0 \end{pmatrix}$$

are $-\sigma_m, \ldots, -\sigma_1, 0\,(n-m \text{ times}), \sigma_1, \ldots, \sigma_m$ ([GVL], page 286), thus the eigenvalues of $\widetilde{\mathcal{A}}$ satisfy the bounds

$$\lambda_i - \frac{\sigma_m}{\sqrt{\beta}} \leq \mu_i \leq \lambda_i + \frac{\sigma_m}{\sqrt{\beta}}, \qquad i = -m, \ldots, -1, 1, \ldots, n. \tag{4.14}$$

Since $\sigma_m = O(1)$, this gives simple P-dependent inclusion intervals for the eigenvalues of $\widetilde{\mathcal{A}}$. The intervals are centred on the eigenvalues of $\widetilde{A}$ and $-\widetilde{C}$. Note that the inclusion

intervals are small for large $\beta$. For small $\beta$, we may interchange the roles of the matrices in (4.13) to obtain

$$
\begin{aligned}
-\frac{\sigma_{-i}}{\sqrt{\beta}} - \lambda_{\max} \le \mu_i \le -\frac{\sigma_{-i}}{\sqrt{\beta}} + \lambda_{\max}, \qquad & i = -m, \ldots, -1 \\
-\lambda_{\max} \le \mu_i \le \lambda_{\max}, \qquad & i = 1, \ldots, n-m \\
\frac{\sigma_{i-n+m}}{\sqrt{\beta}} - \lambda_{\max} \le \mu_i \le \frac{\sigma_{i-n+m}}{\sqrt{\beta}} + \lambda_{\max}, \qquad & i = n-m+1, \ldots, n.
\end{aligned}
\tag{4.15}
$$

Here, $\lambda_{\max} = \max\{-\lambda_{-m}, \lambda_n\}$ is independent of $h$ in general. Certain refinement of (4.14) is possible: for any $0 < \xi < \text{Xi}$, application of the Sylvester Law of Inertia to the congruence transform (4.5)shows that $\lambda_1 \le \mu_1$ for any $\beta$. Thus the positive spectrum of $\tilde{A}$ lies in the interval $[\lambda_1, \lambda_n + \frac{\sigma_m}{\sqrt{\beta}}]$ for any stabilisation. We see in practice that the smallest positive eigenvalue of $\tilde{A}$ moves away from the origin as $\beta$ is decreased for a fixed value of $h$.

The negative part of the spectrum of $\tilde{A}$ will depend on the form of stabilisation $C_h$ which gives rise to the stabilisation matrix C. For a stable element (without stabilisation), $\tilde{C} = 0$, so (4.14) says how far the negative eigenvalues can move away from the origin. For a typical penalty method, $C = D_C = h^d I$, so $\lambda_{-i} = -1$ for $i = 1, \ldots, m$. In this case, further use of the Sylvester Law of Inertia on (4.5) with $-1 < \xi < 0$ reveals

$$
-1 - \frac{\sigma_m}{\sqrt{\beta}} \le \mu_{-i} \le -1, \qquad i = 1, \ldots, m,
$$

but as $\beta$ must necessarily be small for reasons of consistency, the lower bound must necessarily be large. Furthermore, (4.15) shows that some of the positive eigenvalues must also be large. Computations do indeed indicate severe degradation of iterative convergence due to the presence of large eigenvalues with such a method.

For stabilisation methods such as (3.5),(3.6),(3.7) and (3.8) a key issue is how the zero eigenvalues of $\tilde{C}$ move away from the origin for finite values of $\beta$. Certainly $\lambda_{-j} < 0, j = 1, \ldots, m$ for any finite $\beta$ since the inertia is invariant with any bounded value of $\beta$ and any $h$. Some further analysis, which also applies in the case of stable elements, helps answer this question:

Once again we make use of the Sylvester Law of Inertia. From (4.5), if we can show the existence of $\xi < 0$ with

$$
\frac{1}{\beta} \frac{p^t \tilde{B}(\tilde{A} - \xi I)^{-1} \tilde{B}^t p}{p^t p} + \frac{p^t \tilde{C} p}{p^t p} > -\xi \qquad \forall p \in \Re^m - \{0\} \tag{4.16}
$$

then it follows that $\mu_{-1} < \xi$. (Note $(\widetilde{A}-\xi I)^{-1}$ is positive definite for all $\xi \leq 0$). Firstly using the definition (4.4) of $\widetilde{A}, \widetilde{B}$ and $\widetilde{C}$ we see that the left hand side of (4.16) is

$$\frac{1}{\beta}\frac{q^tB(A-\xi D_A)^{-1}B^tq}{q^tD_Cq} + \frac{q^tCq}{q^tD_Cq} \tag{4.17}$$

$$\geq \frac{\theta^2}{\beta}\frac{q^tB(A-\xi D_A)^{-1}B^tq}{q^tM_pq} + \theta^2\frac{q^tCq}{q^tM_pq} \tag{4.18}$$

¿from (4.3) with $q = D_C^{-\frac{1}{2}}p$. For an unstable element, if $B^tq = 0$ then from (3.9)

$$\frac{q^tCq}{q^tM_pq} \geq \frac{\gamma^2}{\beta}$$

thus (4.16) holds for any such $q = D_C^{-\frac{1}{2}}p$ with $-\xi = O(1)$. For any other $q$ satisfying $B^tq \neq 0$, then (4.18) is

$$\theta^2\frac{z^t(I-\xi A^{-\frac{1}{2}}D_AA^{-\frac{1}{2}})^{-1}z}{z^tz}\frac{1}{\beta}\frac{q^tBA^{-1}B^tq}{q^tM_pq} + \theta^2\frac{q^tCq}{q^tM_pq} \tag{4.19}$$

with $z = A^{-\frac{1}{2}}B^tq\,(\neq 0)$. Now writing $y = (I - \xi A^{-\frac{1}{2}}D_AA^{-\frac{1}{2}})^{-\frac{1}{2}}z\,(\neq 0)$, we have

$$\begin{aligned}\frac{z^t(I-\xi A^{-\frac{1}{2}}D_AA^{-\frac{1}{2}})^{-1}z}{z^tz} &= \left[\frac{y^t(I-\xi A^{-\frac{1}{2}}D_AA^{-\frac{1}{2}})y}{y^ty}\right]^{-1} \\ &= \left[1 - \xi\frac{w^tD_Aw}{w^tAw}\right]^{-1}, w = A^{-\frac{1}{2}}y \neq 0.\end{aligned} \tag{4.20}$$

Now (4.20) is as small as possible for w satisfying $w^tAw/w^tD_Aw = v^2h^2$ for any $\xi < 0$ and for some constant v. In (4.19) we thus have

$$\begin{aligned}&\theta^2\frac{z^t(I-\xi A^{-\frac{1}{2}}D_AA^{-\frac{1}{2}})^{-1}z}{z^tz}\frac{1}{\beta}\frac{q^tBA^{-1}B^tq}{q^tM_pq} + \theta^2\frac{q^tCq}{q^tM_pq} \\ &\geq \theta^2\left[1 - \xi v^{-2}h^{-2}\right]^{-1}\frac{1}{\beta}\frac{q^tBA^{-1}B^tq}{q^tM_pq} + \theta^2\frac{q^tCq}{q^tM_pq} \\ &> \theta^2\left[1 - \xi v^{-2}h^{-2}\right]^{-1}\left[\frac{1}{\beta}\frac{q^tBA^{-1}B^tq}{q^tM_pq} + \frac{q^tCq}{q^tM_pq}\right]\end{aligned} \tag{4.21}$$

as $\xi < 0 \quad \Rightarrow \quad \left[1 - \xi v^{-2}h^{-2}\right] > 1$. But now we may use (3.9) (which is (1.17) in the stable case, C = 0) giving

$$\frac{1}{\beta}\frac{p^t\widetilde{B}(\widetilde{A}-\xi I)^{-1}\widetilde{B}^tp}{p^tp} + \frac{p^t\widetilde{C}p}{p^tp} \geq \frac{\theta^2\gamma^2}{\beta}\left[1 - \xi v^{-2}h^{-2}\right]^{-1} \tag{4.22}$$

for any $\xi < 0$. Finally,

$$\frac{\theta^2\gamma^2}{\beta}\left[1 - \xi v^{-2}h^{-2}\right]^{-1} \geq -\xi \tag{4.23}$$

is certainly satisfied for

$$\xi = -h\,\theta\gamma v/\sqrt{\beta} + h^2\,v/2. \tag{4.24}$$

Thus $\mu_{-1}$ is not closer to the origin than $O(h/\sqrt{\beta})$.

In the case of a large stabilisation parameter, a more refined estimate of $\mu_{-1}$ can be derived by noting that for equality in (4.23), we have that

$$\xi = \frac{v^2h^2}{2} - \frac{v^2h^2}{2}\sqrt{1 + \frac{4\theta^2\gamma^2}{\beta v^2h^2}}, \tag{4.25}$$

so that if $\beta > 4\theta^2\gamma^2/v^2h^2 = O(h^{-2})$, we can expand in terms of the binomial series to give

$$\xi = -\frac{v^2h^2}{2}\left(\frac{2\theta^2\gamma^2}{\beta v^2h^2} - \frac{2\theta^4\gamma^4}{\beta^2v^4h^4} + \ldots\right) \approx -\frac{\theta^2\gamma^2}{\beta} \tag{4.26}$$

for $\beta$ sufficiently large. In this case $\mu_{-1}$ is not closer to the origin than $O(1/\beta)$, ie. it is independent of $h$.

One comment on these estimates in the stable case (for which they hold by simply setting C = 0 throughout) is appropriate: the choice of $\beta$ (> 0) in the preconditioner (4.2) is apparently unconstrained, however choosing other than $\beta = 1$ affects the bounds (4.14), (4.15) on $\mu_{-m}$ and (4.24) on $\mu_{-1}$ in a compensating manner, i.e. making one better makes the other worse. Our computations in the next section in any case indicate that (4.24) is pessimistic at least for the particular stable element considered there. The method behaves more like (4.26) for that element.

## 5. Discussion.

In this section, the question of how to choose the stabilisation parameter so as to enhance the rate of convergence of the PCR algorithm of Section 2 is addressed. We compare the performance of the different types of stabilisation introduced in Section 3 with that of an a-priori stable method. For convenience we restrict attention to two-dimensional elements and consider only the lowest order continuous velocity approximation, i.e. based on linear triangle or bilinear square elements.

As a test example we solved the 'leaky' two-dimensional lid-driven cavity problem in a unit square domain with a flow solution, calculated using the stable method below, as illustrated in Fig. 1. This problem was also discussed by Pierre [P], wherein he showed the sensitivity of the solution accuracy to the choice of the stabilisation parameter using (globally-) stabilised $P_1$–$P_1$ and $Q_1$–$Q_1$ methods. In this work, only half the domain was modelled exploiting the natural symmetry of the solution about the line x = 1/2. Rectangular and triangular element grids were both used; starting from a uniform subdivision of N x $2N$ square elements, the triangular grids were constructing by dividing each square into two. In all cases an initial solution guess of zero was used, and the tolerance for convergence was a reduction of $10^{-6}$ in the $L_2$–norm of the residual. All computations were done using Pro-MATLAB on an SGI $4D/35$ Iris-workstation. The stable case is discussed first.

### *5.1 Using an LBB stable method.*

Restricting ourselves to linear triangular or bilinear rectangular elements necessarily implies that the pressure approximation must be defined on a coarser grid if the element is to satisfy the LBB condition (1.9). Using a standard four-triangle macro-element (with internal edges connecting the mid-points of the macro-element edges), we can construct a stable $P_1$–$P_1$ method by using a continuous pressure approximation defined by the macro-element vertices. The method is commonly referred to as the $P_1 iso P_2$ method, c.f. [BF], p.255. For the uniform grids we used, the asymptotic ratio of velocity to pressure degrees of freedom is 8:1 as $h \to 0$ which is somewhat high. Thus from an approximation point of view this method is probably too under-constrained to be the 'best' fist order method (ie. with an $O(h)$ error for velocity in the $H^1$–norm). Note that the tetrahedral analogue of this method is also LBB stable.

Solving the test problem using the PCR algorithm gave the iteration counts in Table 1. Results for $\beta = 1$ with two choices of the matrix $D_C$ (in (4.2)) are presented. Here $D_{M_p}$ is the diagonal of $M_p$. These results vividly illustrate the importance of having the right scaling for the 'pressure part' of the Stokes operator.

A nice feature of these results is the fact that in the case where the 'correct' scaling is used, the iteration count behaves like $O(h^{-1})$, as would be expected using diagonally

| Grid | $D_C = I$ | $D_C = D_{M_p}$ |
|---|---|---|
| $2 \times 4$ | 17 | 16 |
| $4 \times 8$ | 88 | 56 |
| $8 \times 16$ | 261 | 130 |
| $16 \times 32$ | * | 276 |

**Table 1.**

**Number of PCR iterations in the stable case.**

scaled CG to solve Laplace's equation on a uniform sequence of grids. This behaviour can be explained by considering the actual eigenvalue distribution of the preconditioned Stokes operator (4.4). To get a flavour of this, the eigenvalue distribution for the $4 \times 8$ grid is illustrated in figure 2. Also plotted in figure 2 is the optimal polynomial approximation (of degree 11) on the discrete set of values, ie. the polynomial is constructed such that the PCR error contraction estimate (2.3) is minimised. Each vertical bar represents an eigenvalue of (4.4) and is of height equal to twice the minimax error.

| Grid | $\mu_{-m}$ | $\mu_{-1}$ | $\mu_1$ | $\mu_n$ |
|---|---|---|---|---|
| $2 \times 4$ | -0.2842 | -0.0286 | 0.4822 | 1.7307 |
| $4 \times 8$ | -0.3538 | -0.0405 | 0.1925 | 1.9255 |
| $8 \times 16$ | -0.3688 | -0.0407 | 0.0505 | 1.9809 |
| $16 \times 32$ | -0.3721 | -0.0404 | 0.0127 | 1.9952 |

**Table 2.**

**Eigenvalues of $\tilde{\mathcal{A}}$ in the stable case**

To see the way the eigenvalue distribution changes as $h \to 0$, the extremal eigenvalues $\mu_{-m}, \mu_{-1}, \mu_1, \mu_n$ are listed in Table 2. The key point is that the extremal positive eigenvalues indeed behave like those of a scaled Laplacian, whilst the negative eigenvalues remain in a fairly tight cluster which is bounded away from the origin independently of $h$. The behaviour of the positive eigenvalues is consistent with our perturbation analysis, but the fact that the eigenvalue $\mu_{-1}$ appears to be independent of $h$ is slightly surprising. Our estimate (4.24) is clearly pessimistic in this case.

To conclude our discussion of Table 1, when we solved the test problem on the finest

grid using the naive preconditioner $D_C = I$ (and $\beta = 1$), the PCR algorithm broke down after 576 iterations; after a residual reduction of about $10^{-5}$, the IEEE number NaN was calculated. This erratic behaviour seems to emphasise the importance of preconditioning the Stokes operator so as to ensure the right scaling of the pressure and velocity with respect to $h$.

*5.2 Using* a *globally stabilised* method.

We will use the continuous pressure $P_1$–$P_1$ triangle stabilised using (3.5), as a representative globally stabilised method here. A very attractive feature of this method is its inherent simplicity, both the velocity and the pressure being defined by the same piecewise polynomial basis set. Another important feature, is that the approximating power of the method is much better than that of the stable method above. If the effect of the stabilisation term is ignored, then the asymptotic velocity to pressure constraint count is 2:1, which is 'optimal' in two-dimensions. Perhaps the only negative feature of this method is the fact that solution accuracy is known to deteriorate if the stabilisation parameter is not chosen correctly, see [P] for details. On the one hand, if the parameter is too small then the method might not be sufficiently stable to give good results. For example, solving our test problem with $\beta < 10^{-2}$ gave rise to oscillatory pressure solutions. On the other hand, it is easily seen that in the limiting case of $\beta \to \infty$ the pressure solution tends to a constant, which implies that the corresponding velocity field is nowhere near divergence-free. Solving the test problem with $\beta > 10$ gave poor solutions on all of the grids we considered.

Returning now to our main concern; that of finding the 'optimal' choice of stabilisation parameter in the sense that the contraction factor in (2.3) is minimised. PCR iteration count's using preconditioner (4.2) for a range of values of $\beta$ are listed in Table 3. These results illustrate that the efficiency of the PCR solution method is also crucially dependent on the choice of stabilisation parameter.

The characteristic feature of globally stabilised methods is the fact that the stabilisation matrix C represents some discrete approximation to the Laplacian (with Neumann boundary conditions). This means that the eigenvalues $\lambda_{-m}, \ldots, \lambda_{-1}$ of $-\tilde{C}$ are fairly evenly spread within the interval $[-2, 0]$. For comparison with the a-priori stable

| Grid | $\beta = 0.01$ | $\beta = 0.025$ | $\beta = 0.1$ | $\beta = 1.0$ | $\beta = 10.0$ |
|---|---|---|---|---|---|
| $2 \times 4$ | 22 | 21 | 22 | 23 | 23 |
| $4 \times 8$ | 71 | 61 | 65 | 91 | 93 |
| $8 \times 16$ | 169 | 135 | 150 | 285 | 359 |
| $16 \times 32$ | 400 | 307 | 369 | 741 | 1279 |

**Table 3.**

**Number of PCR iterations in the globally stabilised case.**

case above, we plot the eigenvalue distribution corresponding to the case of $\beta = 0.025$ in figure 3, together with the optimal minimax polynomial (of degree 11).

The results in Table 3 closely agree with our perturbation analysis. For a small value of $\beta$ the estimate (4.15) applies, and the extremal eigenvalues $\mu_{-m}$ and $\mu_n$ are forced to move out towards $\pm\infty$ (by an amount proportional to $1/\sqrt{\beta}$) independently of $h$. Applying (4.15) in the case of the smallest positive eigenvalue $\mu_1$, we see that its movement away from the origin is bounded independently of $\beta$ by $\lambda_{\max}(= 2)$. Hence the condition of the preconditioned system necessarily deteriorates as $\beta \rightarrow 0$, as reflected by the iteration counts in the case of $\beta = 0.01$. For a large value of $\beta$ the estimate (4.14) applies, showing that the spectrum of the preconditioned system is 'close' to that of $\tilde{A}$ on the positive side, and to that of $-\tilde{C}$ on the negative side. In particular as $\beta \rightarrow \infty$ the spectrum becomes essentially symmetric about the origin. In this situation, it is well known that the application of Conjugate Residuals gives essentially the same rate of convergence as would be obtained using CG to solve the 'normal equations', ie. the condition number of the system is effectively squared ([Fr]). Thus for large values of $\beta$, we could expect the number of iterations to increase by a factor of four for each succesive refinement of the grid, as can be seen in the case of $\beta = 10$. Interestingly, the 'optimal' choice of $\beta = 0.025$ ¿from the table turns out to be a very natural choice. It is precisely the value (see [P]) which generates the system that would result using the analogous subdivision of $P_1 - P_1$ Mini-elements, after elimination of the internal velocity bubble terms.

In Table 4 we illustrate the variation of the extremal eigenvalues with grid refinement in the Mini-element case (ie. with $\beta = 0.025$). Comparing these values with those

| Grid | $\mu_{-m}$ | $\mu_{-1}$ | $\mu_1$ | $\mu_n$ |
|---|---|---|---|---|
| $2 \times 4$ | -2.2251 | -0.6888 | 0.6986 | 2.5778 |
| $4 \times 8$ | -2.3667 | -0.5122 | 0.1974 | 2.5916 |
| $8 \times 16$ | -2.4553 | -0.2846 | 0.0507 | 2.5931 |
| $16 \times 32$ | -2.4865 | -0.1476 | 0.0128 | 2.5931 |

**Table 4.**

**Eigenvalues of $\tilde{\mathcal{A}}$ in the globally stabilised case with parameter $\beta = 0.025$.**

in Table 2, note that there is a fundamental difference in the behaviour of the largest negative eigenvalue $\mu_{-1}$, our estimate (4.24) of $O(h)$ movement of the eigenvalue $\mu_{-1}$ appears to be tight in this case.

*5.3* Using a *locally* stabilised method.

Finally we discuss the performance of a representative locally stabilised method, namely the $Q_1$–$P_0$ quadrilateral, locally stabilised over 2 x 2 macroelements via (3.7). The attractive feature of this method (apart from its simplicity) is the fact that it also has the 'optimal' approximation property of having an asymptotic velocity to pressure constraint count of 2:1. The use of a discontinuous pressure is especially alluring since it gives the method a 'local bias', for example, it leads to conservation of mass at an element level. Despite it's inherent instability the raw $Q_1$–$P_0$ method is often used in practical computations without any stabilisation. Indeed, in terms of accuracy the velocity solutions are usually reasonable, and pressures often appear to be realistic after post-processing. The lack of inherent stability has to be overcome in our case; the crucial point is that if the method is not stabilised then the eigenvalues of the Schur complement $BA^{-1}B^t$ in (1.17) are not independent of $h$, (the LBB constant $\gamma$ is $O(h)$), hence the performance of the PCR solver rapidly deteriorates as $h \to 0$. Such deterioration in convergence must also be expected with a nested iterative strategy based on the Schur complement.

As in the globally stabilised case, solution accuracy tends to deteriorate if the stabilisation parameter is not sufficiently large. For example, solving our test problem,

oscillatory pressure solutions are evident if $\beta < 10^{-2}$; see [SK] for some related results. In contrast to the globally stabilised case however, solution accuracy is retained in the limit of an arbitrarily large stabilisation parameter. This implies that there is more freedom when seeking to optimise the choice of stabilisation parameter to speed up the rate of convergence of the PCR solver. Of course if the effect of the stabilisation is localised, then we might expect that varying $\beta$ might have less effect on the iteration counts. This expectation is boume out by the results in Table 5.

| Grid | $\beta = 0.01$ | $\beta = 0.1$ | $\beta = 1$ | $\beta = 10$ | $\beta = 100$ |
|---|---|---|---|---|---|
| $2 \times 4$ | 17 | 17 | 17 | 17 | 16 |
| $4 \times 8$ | 73 | 55 | 55 | 59 | 65 |
| $8 \times 16$ | 202 | 142 | 156 | 152 | 147 |
| $16 \times 32$ | * | 347 | 375 | 329 | 263 |

**Table 5.**

**Number of PCR iterations in the locally stabilised case.**

The characteristic feature of locally stabilised methods is the fact that the stabilisation matrix C must always be block diagonal, since it represents a discrete approximation of some local operator. The block size corresponds to the number of discrete pressure variables on a macroelement. In the case above the blocks are $4 \times 4$ matrices, all having eigenvalues 0,1,1,2 after diagonal scaling. The repeated eigenvalue structure of the operator $C_h$ (after scaling) means that for large $\beta$ the preconditioned Stokes matrix will have well clustered negative eigenvalues as illustrated by the spectrum plotted in figure 4.

Comparing the results in Table 5 with those of Table 3, we see that in both cases, the performance is poor if $\beta$ is too small. In the locally stabilised case with $\beta = 0.01$ the PCR algorithm breaks down on the finest grid (in the same way as when using the stable method with the 'wrong' scaling). As discussed above this poor behaviour is explained by our eigenvalue theory of the last section. On the other hand, for a large stabilisation parameter the eigen-spectrum of the locally stabilised method is fundamentally different to that in the globally stabilised case, and this is reflected in the iteration counts. In the locally stabilised case, the clustering on the negative side of the spectrum can be

exploited by the PCR algorithm and the nice behaviour of the iteration count growing like $O(h^{-1})$ is retained. This is in stark contrast to the globally stabilised case with large values of $\beta$. Note also that our analysis clearly shows that $\beta$ must not be too large, otherwise the smallest negative eigenvalue $\mu_{-1}$ (ie. the right-most cluster) will be close to the origin, which will certainly slow down convergence.

In Table 6 we show the variation of the extremal eigenvalues for the optimal choice of parameter from the table above, (ie. with $\beta = 100$). Comparing with the values in Table 4, we see that the fundamental difference is the fact that with this value of $\beta$ the largest negative eigenvalue seems to be insensitive to $h$, much as in the stable case (cf. Table 2). This is perhaps to be expected since any locally stabilised method must tend to an a-priori stable method (in this case the $P_0 iso Q_2$ method) in the limit of $\beta \to \infty$. For more moderate values of $\beta$ (for example, $\beta = 1$) the eigenvalue $\mu_{-1}$ moves towards the origin like $O(h)$, so that the convergence behaviour is more like the Mini-element case above.

| Grid | $\mu_{-m}$ | $\mu_{-1}$ | $\mu_1$ | $\mu_n$ |
|---|---|---|---|---|
| $2 \times 4$ | -2.0004 | -0.0019 | 0.3983 | 1.2526 |
| $4 \times 8$ | -2.0011 | -0.0016 | 0.1135 | 1.4297 |
| $8 \times 16$ | -2.0012 | -0.0013 | 0.0309 | 1.4839 |
| $16 \times 32$ | -2.0012 | -0.0012 | 0.0091 | 1.4982 |

Table 6.
Eigenvalues of $\tilde{A}$ in the locally stabilised case
with parameter $\beta = 100$.

Our conclusions from this are as follows: using an a-priori stable method, convergence rates analogous to those which would be expected solving the diagonally scaled Laplacian can be obtained in the indefinite case, as long as the preconditioner (4.2) enforces the 'correct' scaling. Using a globally stabilised method, good convergence rates can only be achieved by making the correct choice of stabilisation parameter. However, even when chosen optimally the iteration counts are likely to be asypmtotically inferior to those which would be obtained using an a-priori stable method as above. Using a

locally stabilised method leads to reasonable rates of convergence and solution accuracy as long as the stabilisation parameter is not too small. Making the 'optimal choice' in the locally stabilised case as discussed above, again gives the convergence behaviour obtained in the a-priori stable case.

In part II of this work ([SW]), we extend our analysis to cover the case of more sophisticated preconditioners, for example, based on the Laplacian part of the Stokes operator.


## Acknowledgements

This research was completed whilst the authors were visiting the Computer Science Department at Stanford University. We would like to thank Gene Golub for his continued encouragement and support. Financial assistance from the Fulbright Commission, the NSF under grant CCR-8821078-A2 and the SERC under grant GR/H20299 is also gratefully acknowledged.

## References.


[ABF] D. Arnold, F. Brezzi & M. Fortin, "A stable finite element for the Stokes equations," *Calcolo*, v. 21, 1984, pp. 337-344.

[AMS] S.F. Ashby, T.A. Manteuffel & P.E. Saylor, "A taxonomy for conjugate gradient methods," *SIAM J. Numer. Anal.*, v. 27, 1990, pp. 1542–1568.

[AB] 0. Axelsson & V.A. Barker, *Finite Element Solution of Boundary Value Problems: Theory and Computation,* Academic Press, New York, 1984.

[BWY] R.E. Bank, B.D. Welfert & H. Yserentant, "A class of iterative methods for solving saddle point problems," *Numer. Math.*, v. 56, 1990, pp. 645–666.

[BP] J. Bramble & J. Pasciak, "A preconditioning technique for indefinite systems resulting ¿from mixed approximations of elliptic problems," *Math.* Comput., v. 50, 1988, pp. 1–17.

[BF] F. Brezzi & M. Fortin, *Mixed and Hybrid finite Element Methods,* Springer, New York, 1991.

[BPi] F. Brezzi & J. Pitkäranta, "On the stabilisation of finite element approximations of the Stokes problem," *Efficient Solutions of Elliptic Systems,* Notes on Numerical Fluid Mechanics Vol.10 (W. Hackbusch, ed.), Vieweg, Braunschweig, 1984, pp. 11–19.

[FHS] L.P. Franca, T.J.R. Hughes & R. Stenberg, "Stabilised finite element methods for the Stokes problem," to appear in *Incompressible Computational Fluid Dynamics - Trends and Advances* (R.A. Nicolaides and M.D. Gunzburger eds.

[Fr] R. Freund, "On polynomial preconditioning for indefinite hermitian matrices," RIACS Technical Report 89.32, RIACS, NASA Ames Research Center, 1989.

[F] I. Fried, "The $l_2$ and $l_\infty$ condition numbers of the finite element stiffness and mass matrices and the pointwise convergence of the method," *The Mathematics of Finite Elements and Applications*, (J.R. Whiteman, ed.), Academic Press, 1983, pp.

[GL] R.L. Glowinski, *Numerical Methods for Nonlinear Variational Principles*, Springer, New York, 1984.

[GVL] G.H. Golub & C.F. Van Loan, *Matrix Computations*, John Hopkins University Press, Baltimore, 1983.

[G] A. Greenbaum, "Comparison of splittings used with the conjugate gradient algorithm," *Numer.* Math., v. 33, 1979, pp. 181–193.

[HF] T.J.R. Hughes & L.P. Franca, "A new finite element formulation for CFD: VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces," *Comput. Methods Appl. Mech. Engrg., v.* 65, 1987, pp. 85-96.

[JY] W.D. Joubert & D.M. Young, "Necessary and sufficient conditions for the simplification of generalised conjugate gradient algorithms," *Linear Algebra Appl., v.* 88/89, 1987, pp *449-485.*

[KS] N. Kechkar & D.J. Silvester, "Analysis of locally stabilised mixed finite element methods for the Stokes problem", to appear in Math. *Comp.,* 1992.

[L] V.I. Lebedev, "Iterative methods for solving operator equations with a spectrum contained in several intervals," *USSR Comput.* Math. *and* Math. *Phys. v.* 9, 1969, pp. 17–24.

[P] R. Pierre, "Simple $C^0$ approximations for the computation of incompressible flows", Comput. Methods *Appl. Mech. Engrg., v. 68,* 1988, pp. 205–227.

[PS] J. Pitkäranta & T. Saarinen, "A multigrid version of a simple finite element method for the Stokes problem," Math. *Comp., v. 45,* 1985, pp. 1–14.

[R] A. Ramage, Ph.D. Thesis, University of Bristol, UK, 1991.

[SK] D.J. Silvester & N. Kechkar, "Stabilised bilinear-constant velocity-pressure finite elements for the conjugate gradient solution of the Stokes problem", *Comput.* Meth*ods Appl. Mech. Engrg., v. 79, 1990,* pp. 71-86.

[SW] D.J. Silvester & A.J. Wathen, "Fast iterative solution of stabilised Stokes systems Part II: Using block preconditioners," in preparation.

[V] R. Verfurth, "A combined conjugate gradient-multigrid algorithm for the numerical solution of the Stokes problem",*IMA J.* Numer. *And., v. 4,* 1984, pp. 441–455.

[W] A.J. Wathen, "Realistic eigenvalue bounds for the Galerkin Mass Matrix", *IMA J. Numer. And., v. 7,* 1987, pp. 449457.

[Wi] J.H. Wilkinson, *The Algebraic Eigenvalue Problem,* Oxford University Press, 1965.

Fig 1a: Stokes Problem Velocity Plot

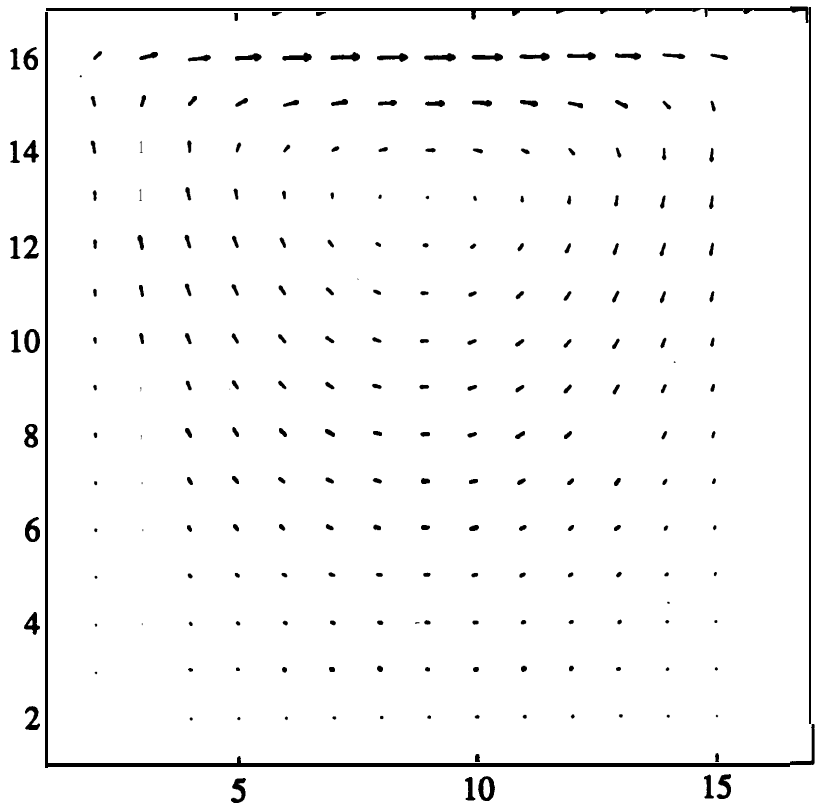
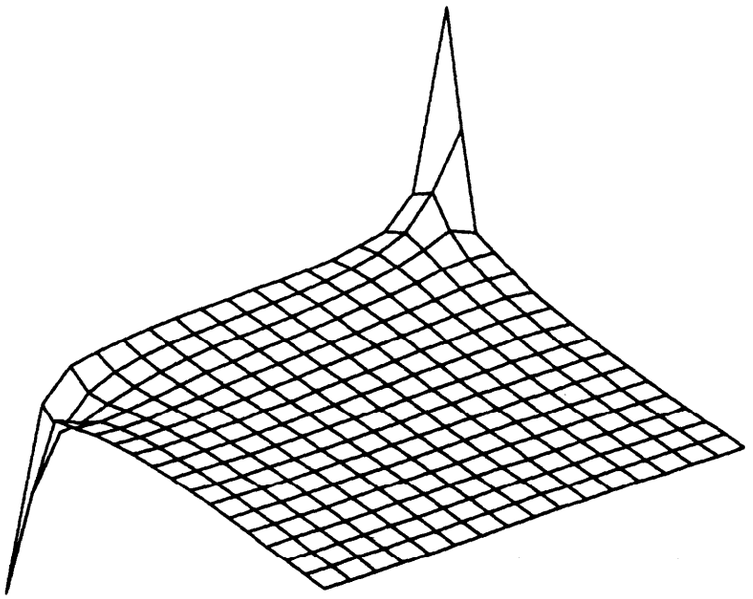


Fig 1b: Stokes Problem Pressure Distribution



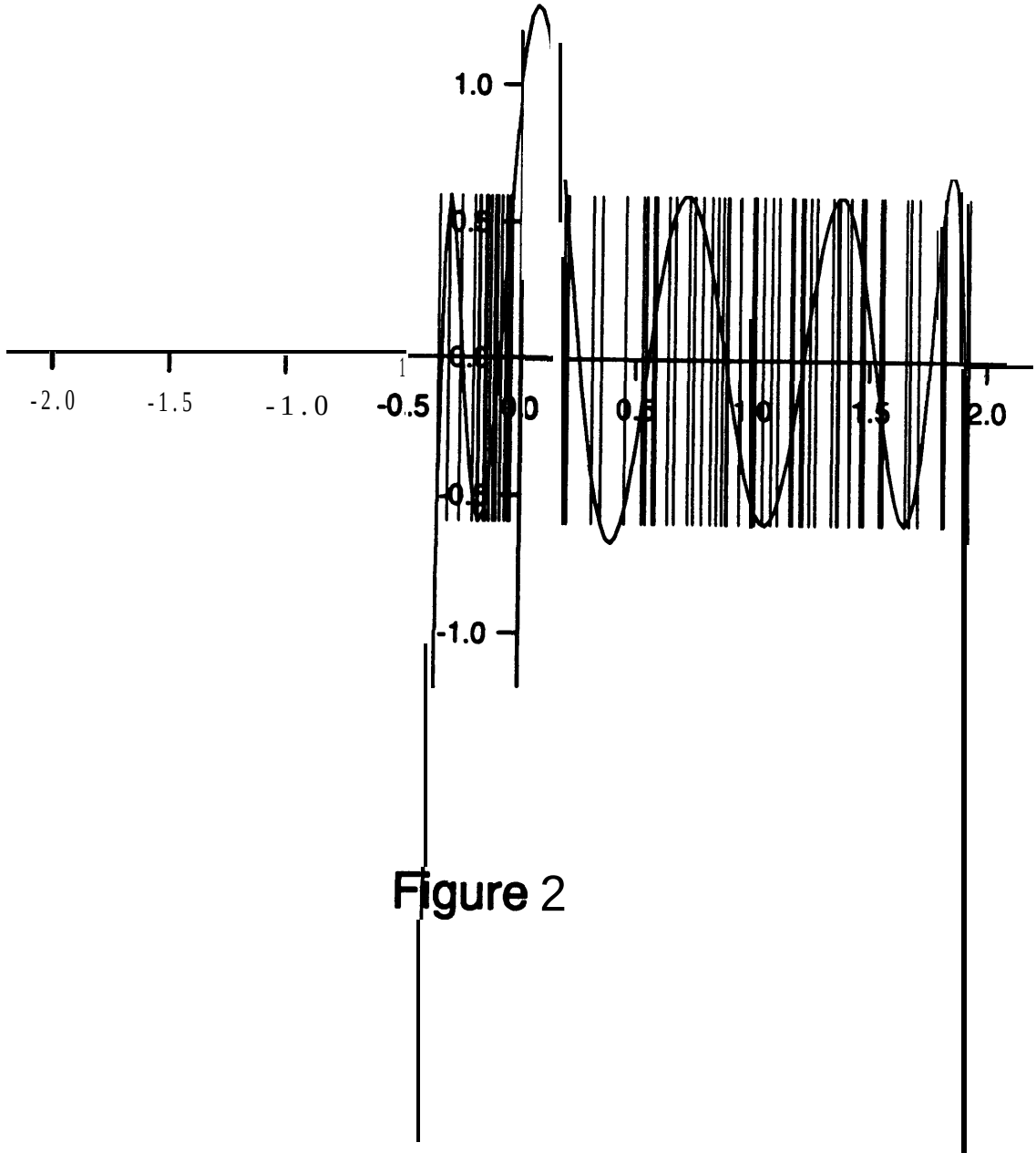



Figure 2

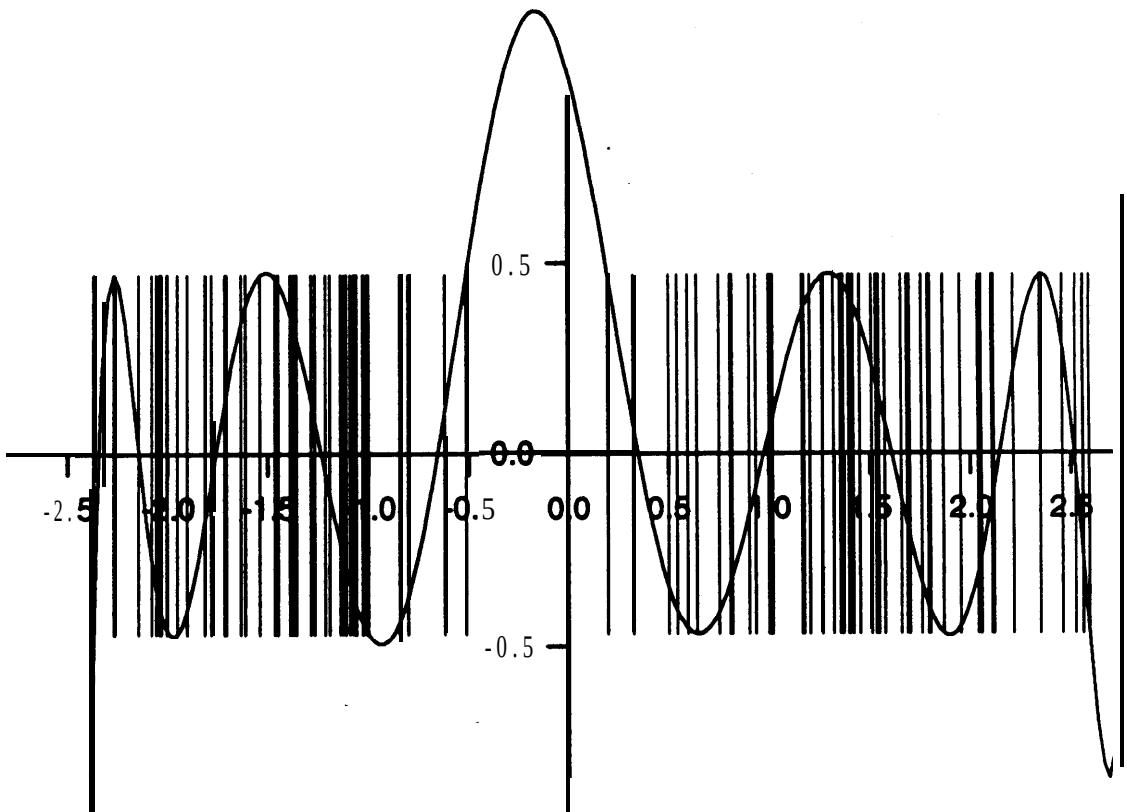



Figure 3

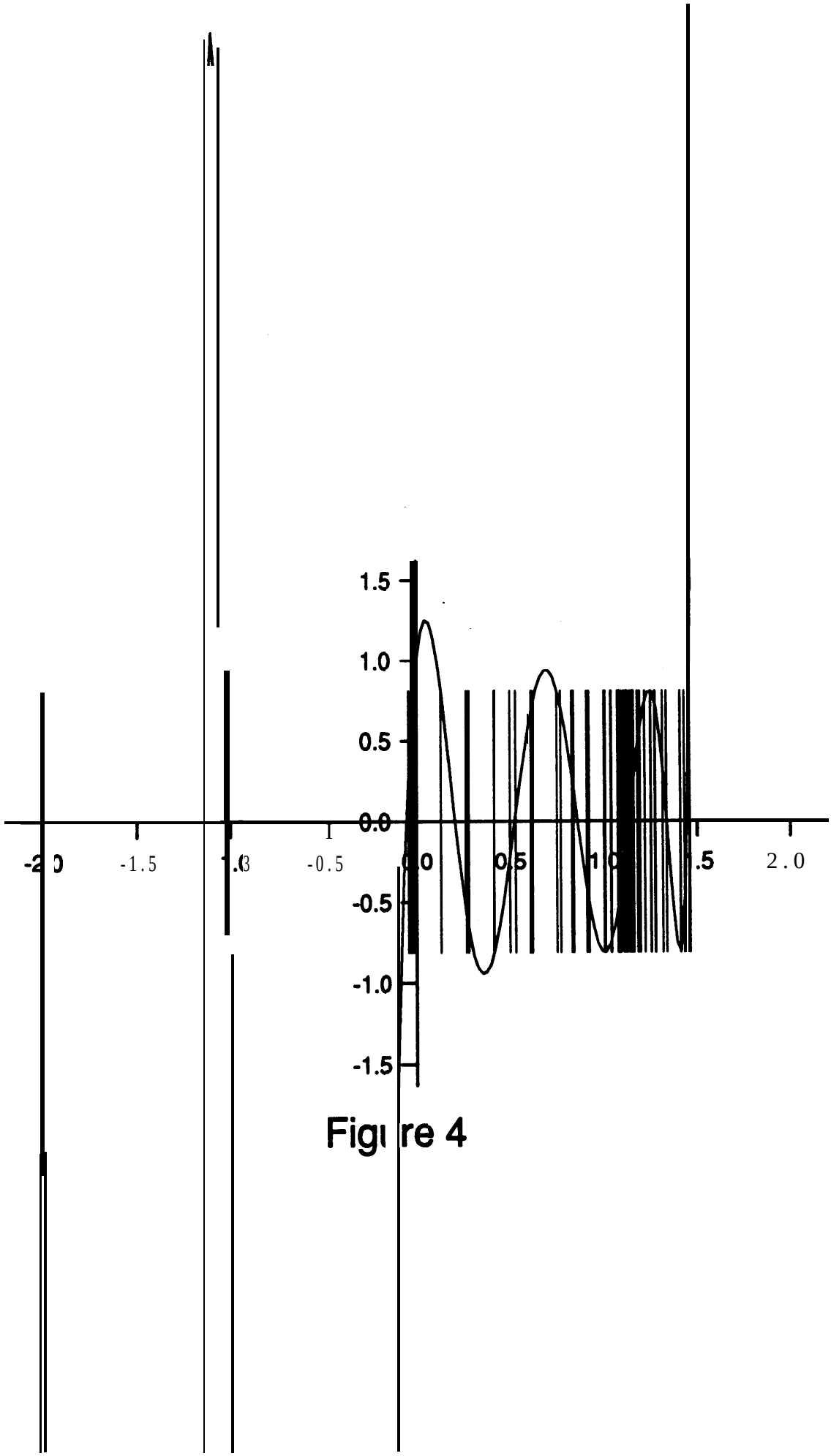



Figure 4