INTERNATIONAL JOURNAL FOR NUMERICAL METHODS IN FLUIDS, VOL. 20, 1237–1252 (1995)

THE INFLUENCE OF THE STABILIZATION PARAMETER ON THE CONVERGENCE FACTOR OF ITERATIVE METHODS FOR THE SOLUTION OF THE DISCRETIZED STOKES PROBLEM

C. VINCENT

Université de la Réunion, IREMIA, 15 Avenue René Cassin, F-97489 Saint-Denis Cedex, Ile de la Réunion, France

SUMMARY

This paper discusses the influence of the stabilization parameter on the convergence factor of various iterative methods for the solution of the Stokes problem discretized by the so-called *locally stabilized* Q1–P0 finite element. Our objective is to point out optimal parameters which ensure rapid convergence.

The first part of the paper is concerned with the dual formulation of the problem. It gives the theoretical precision and practical developments of our *stabilized context Uzawa-type algorithm*. We assert that the convergence factor of such a method is majored independently of the mesh size by a function of the stabilization parameter. Moreover, we point out that there exists an optimal value of this parameter that minimizes this upper bound. This gives a theoretical justification of pre-existing numerical results. We show that the optimal parameter can be determined *a priori*. This is a key point when the method has to be implemented. Finally, we base an interpretation of the *iterated penalty method* numerical behaviour on some theoretical results about the minimum eigenvalue of the stabilized dual operator. This algorithm involves a penalty parameter and a stabilization parameter and we discuss a strategy for choosing optimal parameters.

The mixed formulation of the problem is dealt with in the second part of the paper, which proposes several preconditioned conjugate-gradient-type methods. The indefinite character of the problem makes it intrinsically hard. However, if one chooses a suitable preconditioner, this difficulty is overcome, since the preconditioned operator becomes positive definite. We study the eigenvalue spectrum of the preconditioned operator and thereby the convergence factor of the algorithm. In contrast with the two previous formulations, we show that this convergence factor is majored independently of the stabilization parameter. More precisely, we point out convergence factors comparable with those obtained for Poisson-type problems. Finally, we present a variant of the latter method which uses our so-called *macroblock-type preconditioner*. A comparison with the simple case of diagonal preconditioning is addressed and the improved performance of the macroblock-type preconditioner is evidenced.

Various 2D numerical experiments are given to corroborate the theories presented herein.

KEY WORDS Stokes equations; mixed finite elements; stabilization; conjugate gradient methods; preconditioning

1. INTRODUCTION

The finite element discretization of the Stokes equations leads to the solution of mixed systems such as

$$\mathbf{A}X = \begin{pmatrix} A & B^{\mathrm{T}} \\ B & 0 \end{pmatrix} \begin{Bmatrix} u \\ p \end{Bmatrix} = \begin{Bmatrix} f \\ 0 \end{Bmatrix} = F. \tag{1}$$

It is a well-known result that problem (1) has a unique solution for discretizations (X_h, M_h) satisfying

CCC 0271–2091/95/111237–15
© 1995 by John Wiley & Sons, Ltd.
Received 7 January 1993
Revised 15 July 1994

the classical compatibility condition of Babuska[1] and Brezzi:[2] there is a constant $k > 0$ independent of h such that

$$\sup_{u_h \in X_h} \frac{|\langle \operatorname{div} u_h,\ q_h \rangle|}{\|u_h\|_1} \geq k \|q_h\|_0 \quad \forall q_h \in M_h. \tag{2}$$

Unfortunately, this condition is not satisfied by the (so-called) Q1–P0 finite element, which is one of the most commonly used finite elements, particularly in the three-dimensional case. The discretization is unstable. Among other difficulties clearly pointed out by Sani *et al.*,[3] iterative methods for solving (1) can be totally inefficient because of the 'bad' condition number of the operators.[4–6]

In order to get round the difficulty caused by instability, one can replace the discrete incompressibility constraint $Bu = 0$ by $Bu - \beta Cp = 0$. Thus problem (1) becomes

$$\mathbf{A}_\beta X = \begin{pmatrix} A & B^{\mathrm{T}} \\ B & -\beta C \end{pmatrix} \begin{Bmatrix} u \\ p \end{Bmatrix} = \begin{Bmatrix} f \\ 0 \end{Bmatrix} = F, \tag{3}$$

where C is a stabilization operator and β is a positive stabilization parameter.

Taking advantage of a suitable stabilization procedure, one can obtain efficient iterative solvers from the normal choice $\beta = 1$.[6] This paper aims essentially to show that the stabilization parameter β influences the convergence factor of various iterative techniques. Our objective is to point out some optimal values of β that minimize the convergence factor and lead to a significant improvement in the method.

Therefore the choice of the (so-called) *locally stabilized method* of Kechkar and Silvester[7] is very important. Currently, this method, unlike many others, allows one to tune the magnitude of the stabilization parameter in order to get the best convergence factors without adversely affecting the accuracy of the solution. The improved robustness of this method is justified theoretically in Reference 7 and various numerical experiments can be found in Reference 8.

As far as we know, the benefit of stabilization for iterative Stokes solvers was first brought forward for multigrid technique settings.[4] However, other approaches were quickly investigated.

An important new research category lies in the development of what we have called *stabilized context Uzawa-type algorithms*. Such algorithms have been presented after a very short delay for two of the most popular finite elements, namely the Q1–P0 element[9] and the '*mini-élément*'.[10]

A stabilized context Uzawa-type algorithm leads to some conjugate gradient solvers for the stabilized dual problem[6,11–13]

$$L_\beta p = \left(BA^{-1}B^{\mathrm{T}} + \beta C\right)\{p\} = BA^{-1}f = g. \tag{4}$$

The first part of the paper is concerned with this approach. Section 2.1 reviews some fairly recent developments of the original Uzawa algorithm. Section 2.2 explains the efficiency of stabilized context algorithms by the properties of the theoretical upper bound for the spectral condition number K of the operator L_β. This efficiency has already been supported by extensive numerical results in References 6, 11 and 12 (2D and 3D) and Reference 13 (2D).

The main results,[9] obtained for uniform 2D meshes of regular macroelements,[14] are:

(i) L_β is symmetric positive definite
(ii) $K(L_\beta)$ is *majored independently of the mesh size* by a function $\gamma(\beta)$ of the stabilization parameter β
(iii) $\gamma(\beta)$ can be minimized for an optimal value of this parameter, say β^*.

Moreover, we show that the optimal parameter can be determined *a priori*, which is essential when the method has to be implemented.

Section 2.3 establishes estimates for the convergence factor of the method. We give various numerical experiments to illustrate our theoretical results. These results agree closely with those of Atanga and Silvester,[13] especially with the optimal parameter $\beta^* = O(10^{-1})$ they obtained from various simulations.

Finally, in Section 2.4 we base an interpretation of the *iterated penalty method*[13] numerical behaviour on some theoretical results about the minimum eigenvalue of the stabilized dual operator. This algorithm has a classical embedded outer–inner iteration structure. A penalty parameter and a stabilization parameter influence both the outer iteration behaviour and the inner iteration behaviour. We will discuss a strategy for choosing optimal parameters.[15]

The second part of the paper applies iterative methods directly to the mixed stabilized problem (3). Section 3.1 presents recent research about these techniques. The indefinite character of the problem makes it intrinsically hard. However, if one chooses an adequate preconditioner, this difficulty is overcome, since the preconditioned operator becomes positive definite. Ewing *et al.*[16] and Bank *et al.*[17] have presented such a procedure in the case of inherently stable mixed approximations. We can extend this technique to a stabilized formulation. A suitable preconditioner[9] is then

$$\bar{\mathbf{A}}_\beta X = \begin{pmatrix} \bar{A} & B^{\mathrm{T}} \\ B & -\beta C \end{pmatrix}, \quad \text{where } \bar{A} = \operatorname{diag}(A). \tag{5}$$

Section 3.2 computes some algebraic properties[9] of the preconditioned operator $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$, especially its positivity which allows the use of a conjugate gradient method:

(i) $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$ is symmetric positive definite

(ii) $K(\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta) \leq K(\bar{A}^{-1}A)$.

In contrast with the two previous formulations, the condition number of the preconditioned mixed operator $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$, and thereby the convergence factor of the algorithm, is majored *independently of the stabilization parameter*. More precisely, we point out convergence factors comparable with those obtained for the diagonally scaled Laplacian.

Section 3.3 discusses the implementation of the algorithm, which has a classical embedded outer–inner iteration structure since every preconditioning step requires the solution of a dual-type system. The outer iteration behaviour is almost independent of β but the inner iteration behaviour depends on β. We show that the optimal value is $\beta = O(10^2)$. Finally, we give a variant of the previous method based on our *macroblock-type preconditioner*.[6,9] A comparison with the very simple diagonal preconditioner[13] is addressed and Section 3.4 discusses numerical results obtained for the three approaches. The improved performance of our macroblock-type preconditioner is evidenced.

The numerical results are given to corroborate the theories presented in this paper. All the results come from the uniform lid-driven cavity problem. It is a common test problem which has been solved in most of the papers written about Stokes flows.[4–6,9–13,15–21] We employ a mixed approximation by Q1–P0 elements using the local stabilization procedure of Reference 7. Structured meshes of regular macroelements are considered.

2. A STABILIZED CONTEXT UZAWA-TYPE ALGORITHM

2.1. Introduction

Among the various iterative techniques which can be used to solve mixed problems, the Uzawa algorithm[22] is certainly one of the most common choices. This algorithm has been improved in very different ways; see e.g. References 5, 18 and 19. In those papers the algorithms are presented in the

context of stable discretizations satisfying the classical compatibility condition. In this case one can retain the main property that the corresponding dual operator is symmetric, positive definite and provides a condition number bounded independently of the mesh size. Thus good results have been presented for the Q1$^+$–P1 element[5] and the P2–P1 element.[18,19]

On the other hand, Uzawa-type algorithms have been clearly demonstrated to be inefficient for the convenient (but unstable) Q1–P0 element.[5,6] Reference 6 shows that these difficulties can be overcome by using an appropriate stabilization procedure. Such an improvement, updating the original (non-stabilized) Uzawa algorithm, is quite recent.[6,9–13] The next subsection is concerned with both the theoretical aspects and the practical development of such new algorithms. More precisely, we will study the effect of the stabilization parameter on the condition number of the stabilized dual operator, for which we show the existence of an optimal value.

2.2. *Properties of the stabilized dual operator*

The initial indefinite problem (3) for the velocity and pressure can be transformed into an equation for the pressure:

$$L_\beta p = (BA^{-1}B^{\mathrm{T}} + \beta C)p = BA^{-1}f = g. \tag{6}$$

The properties of the stabilized dual operator L_β are given in Theorem 1.

Theorem 1

For the locally stabilized method of Kechkar and Silvester[7] and for uniform 2D meshes of regular macroelements:[14]

(i) $L_\beta = BA^{-1}B^{\mathrm{T}} + \beta C$ is symmetric positive definite
(ii) $K(L_\beta) \leq \gamma(\beta)$, $\gamma(\beta) = (a + b\beta)/\min(c, 4\beta)$, where a, b and c are positive constants *independent of the mesh size h*
(iii) the optimal choice for β in the sense of minimizing $\gamma(\beta)$ is $\beta^* = c/4$.

Proof. A proof of this theorem is given in References 9 and 12. Let us remember that (i) follows easily from the 'stabilization condition'.[13]

The estimate (ii) is based on the following properties.

1. A standard and useful property of the original (non-stabilized) dual operator:[23]

$$\langle BA^{-1}B^* q,\ q\rangle = \sup_u \frac{\langle Bu,\ q\rangle^2}{\langle Au,\ u\rangle} \quad \forall q \in M_h. \tag{7}$$

2. A natural decomposition of the pressure space:

$$M_h = N_h \oplus N_h^{\perp}, \quad \text{where } N_h = \ker(B^*). \tag{8}$$

3. A weak Babuska–Brezzi-type stability condition for the Q1–P0 element:[24]

$$C_1|q|_h \geq \sup_u \frac{\langle \operatorname{div} u,\ q\rangle}{\|u\|_1} \geq C_2|q|_h \quad \forall q \in M_h, \tag{9a}$$

$$C_3 h\|q\|_0 \leq |q|_h \leq C_4\|q\|_0 \quad \forall q \in N_h^{\perp}, \tag{9b}$$

in which $|\ |_h$ is the mesh-dependent seminorm introduced by Johnson and Pitkaranta.[24]

4. Some properties of the stabilization operator:

$$\lambda_i = 0,\ 2,\ 2,\ 4 \text{ are the eigenvalues of } C. \tag{10}$$

From (7)–(10) we have the estimates

$$\left\langle \left(BA^{-1}B^* + \beta C\right)q,\ q\right\rangle \leq (a + b\beta)\|q\|_0^2, \tag{11}$$

$$\left\langle \left(BA^{-1}B^* + \beta C\right)q,\ q\right\rangle \geq \min(c,\ 4\beta)\|q\|_0^2, \tag{12}$$

where a, b and c are positive constants independent of the mesh size h,[9,12] which demonstrates (ii) if one comes to the matricial form.

Finally, it is obvious that the minimum of $\gamma(\beta) = (a + b\beta)/\min(c, 4\beta)$ occurs at $\beta^* = c/4$, which demonstrates (iii) and completes the proof. □

Unfortunately, information about β^* is unavailable. The question of interest here is the *a priori* determination of the optimal parameter β^*. Note the important result (iii) which implies that β^* is independent of the mesh size. Thus we can determine the optimal parameter in the following way.

1. Compute an approximation β_H of the optimal parameter from a sequence of runs made on a coarse grid such as $H = 1/4$.
2. Extrapolate the solution (u_H, p_H) in order to obtain a 'good' initial guess on the desired grid. Then start the stabilized context Uzawa-type algorithm with $\beta = \beta_H \approx \beta^*$.

2.3. *Estimates for the convergence factor of the algorithm*

A measure of the convergence property of the algorithm is given by

$$\delta = \left(R^n/R^0\right)^{1/n}, \tag{13}$$

where n denotes the number of iteration steps to achieve convergence at the desired relative accuracy, i.e. $R^n \leq R^0 \cdot \varepsilon$, and R^i denotes the norm of the residual, $R^i = \left\|L_\beta p^i - g\right\|$.

Remark. ε is equal to 10^{-6} in our computations.

δ is strongly related to the condition number K of the stabilized dual operator as shown in the classical upper-bound estimation[18]

$$\delta \leq \frac{\sqrt{K} - 1}{\sqrt{K} + 1} \tag{14a}$$

derived for the conjugate gradient approach. The previous result (ii) implies

$$\delta \leq \left(\frac{\sqrt{\gamma(\beta)} - 1}{\sqrt{\gamma(\beta)} + 1}\right). \tag{14b}$$

In Figures 1(a) and 1(b) we have plotted the convergence factor δ as a function of β. These results might be expected from the study of the function

$$\beta \to \delta(\beta) = \left(\frac{\sqrt{\gamma(\beta)} - 1}{\sqrt{\gamma(\beta)} + 1}\right).$$

In fact, a simple calculation yields

$$\delta' = \frac{\gamma'}{\sqrt{\gamma}(\sqrt{\gamma}+1)^2} \qquad \delta'' = \frac{2\gamma''\gamma(\sqrt{\gamma}+1) - \gamma'^2(1+3\sqrt{\gamma})}{2\gamma\sqrt{\gamma}(\sqrt{\gamma}+1)^3}.$$

Thus we have the following.

1. For $\beta \geq \beta^*$, $\gamma(\beta) = (a+b\beta)/c$, $\gamma'(\beta) = b/c$, $\gamma''(\beta) = 0$. Thus $\delta' > 0$ and $\delta'' < 0$.
2. For $\beta < \beta^*$, $\gamma(\beta) = (a+b\beta)/4\beta$, $\gamma'(\beta) = -a/\beta^2$, $\gamma''(\beta) = a/2\beta^3$. Thus $\delta' < 0$ and $\delta'' > 0$.

Remark. Similar behaviour is observed in the three-dimensional case. See Reference 9 or 12.

Figures 1(a) and 1(b) show how the convergence factor of the algorithm depends on the stabilization parameter. The value of the optimal parameter β^* is found to be $\beta^* = 0{\cdot}1$ in 2D and $\beta^* = 0{\cdot}075$ in 3D.[9,12] These results agree closely with those of Atanga and Silvester,[13] especially with the optimal parameter $\beta^* = O(10^{-1})$ they obtained from various 2D simulations.

The improved performance of the algorithm is evidenced since the optimal choice $\beta = \beta^*$ required about two times fewer iterations than $\beta = 1$, itself an important improvement compared with the original (non-stabilized) case $\beta = 0$. A measure of improvement is proposed in Figure 2, where the reader can appreciate the benefits coming from the contribution of a classical choice $\beta = 1$ and from the contribution of an optimal choice $\beta = \beta^*$.

Remark. References 9 and 12 compare our stabilized algorithm with other Uzawa-type algorithms for two inherently stable discretizations, namely P2–P1[18] and $Q1^+$–P1.[5] This comparison shows that the operator L_{β^*} is well-conditioned, as is the standard operator $L = BA^{-1}B^T$ corresponding to the $Q1^+$–P1 discretization. See Reference 10 for the 'mini-élément' $P1^+$–P1 or the stabilized P1–P1 element after static condensation of the 'bubble'.

With appropriate *stabilized context Uzawa-type algorithms* including an automatic determination of the stabilization optimal parameter, the popular Q1–P0 element can be used very efficiently, particularly in the three-dimensional case. Moreover, the algorithm is open to considerable improvement since it takes advantage of available preconditioners for A or Poisson solvers. We can think about the *incomplete Uzawa algorithm* of Robichaud *et al.*[5] and the *combined conjugate gradient–multigrid algorithm* of Verfurth.[19] These algorithms perform well on inherently stable discretizations but break down on unstable ones.[5] In Reference 11 we have presented an improvement of the incomplete algorithm by stabilization and a stabilized version of the combined algorithm of

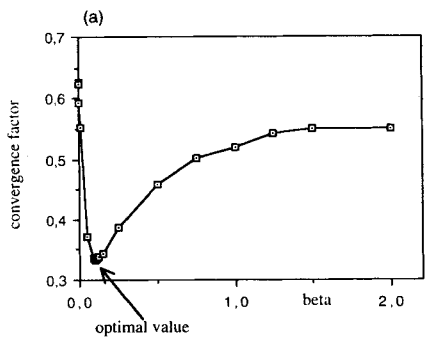

Figure 1(a). Convergence factor of Uzawa-type algorithm, $\beta \in [0; 2]$ ($h = 1/16$)

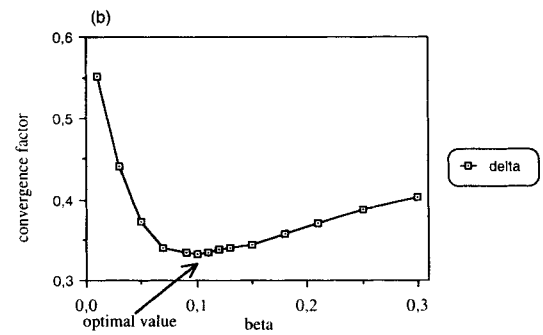

Figure 1(b). Convergence factor of Uzawa-type algorithm, $\beta \in [0{\cdot}01; 0{\cdot}3]$ ($h = 1/16$)

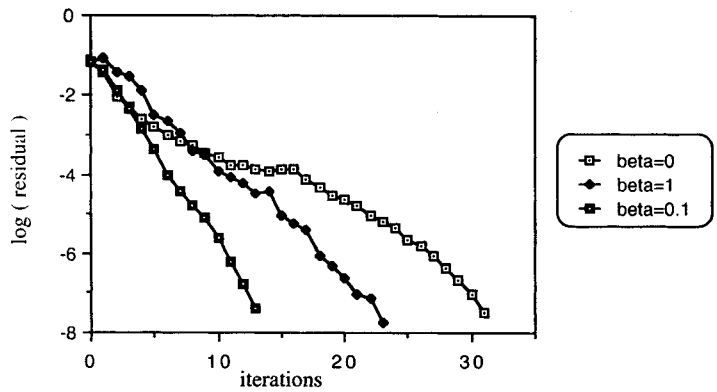

Figure 2. Measure of improvement ($h = 1/16$)

Verfurth is under study. Instead of the multigrid approach we prefer the FAC algorithm of MacCormick[25] when the mesh is irregular but can be broken up into regular submeshes. The case of (so-called) composite grids[25] is favourable for the FAC method. Consequently, we expect that an approach combining our stabilized Uzawa-type algorithm and an FAC preconditioner for A will perform well on composite grids. This will be studied in future work.

2.4. *Further information about the iterated penalty algorithm of Atanga and Silvester*

2.4.1. Introduction. This subsection gives some theoretical justifications[15] about the numerical behaviour of the (so-called) *iterated penalty algorithm* of Atanga and Silvester.[13] The iterative process runs as

$$\begin{pmatrix} A & B^{\mathrm{T}} \\ B & -(\beta C + \varepsilon M) \end{pmatrix} \begin{Bmatrix} u^{i+1} \\ p^{i+1} \end{Bmatrix} = \begin{Bmatrix} f \\ -\varepsilon M p^{i} \end{Bmatrix}, \tag{15}$$

where M is the pressure mass matrix and ε is a positive penalty parameter. Equation (15) can be decomposed as

$$A_{\varepsilon,\beta} u^{i+1} = \left[A + B^{\mathrm{T}}(\beta C + \varepsilon M)^{-1} B\right] u^{i+1} = f - B^{\mathrm{T}}(\beta C + \varepsilon M)^{-1} \varepsilon M p^{i}, \tag{16a}$$

$$p^{i+1} = (\beta C + \varepsilon M)^{-1}\left(B u^{i+1} + \varepsilon M p^{i}\right). \tag{16b}$$

For the solution of (16a) a conjugate gradient method may be used. Then the resulting algorithm has a classical embedded inner–outer iteration structure. At this stage it is first interesting to study the convergence of the inner and outer iterations separately.

2.4.2. Study of the outer iteration. Every step (15) of the algorithm leads to an iteration for the pressure error $\bar{p}^{i} = p^{i} - p$:[13]

$$K^{-1}\bar{p}^{i+1} = \left[(1/\varepsilon) M^{-1/2}\left(B A^{-1} B^{\mathrm{T}} + \beta C\right) M^{-1/2} + I\right] \bar{p}^{i+1} = \bar{p}^{i}. \tag{17}$$

Its convergence behaviour is determined by the spectral radius of K,

$$\rho(K) = \varepsilon/(\varepsilon + \lambda^{*}), \tag{18}$$

where λ^{*} is the minimum eigenvalue of the matrix $M^{-1/2}(BA^{-1}B^{\mathrm{T}} + \beta C)M^{-1/2}$.

Remark. The penalty parameter behaves like an acceleration parameter.

If one denotes by λ_0^* the maximum eigenvalue of the operator $L_\beta = BA^{-1}B^* + \beta C$, it is obvious that

$$\lambda^* = \inf_p \frac{\langle M^{-1/2}(BA^{-1}B^{\mathrm{T}} + \beta C)M^{-1/2}p,\ p\rangle_{\mathrm{e}}}{\langle p,\ p\rangle_{\mathrm{e}}}$$
$$= \inf_p \frac{\langle (BA^{-1}B^{\mathrm{T}} + \beta C)p,\ p\rangle_{\mathrm{e}}}{\langle Mp,\ p\rangle_{\mathrm{e}}} = \lambda_0^*,$$

where $\langle\ ,\ \rangle_{\mathrm{e}}$ denotes the Euclidean scalar product. For uniform 2D meshes made of regular macroelements, one can conclude from (12) that $\lambda_0^* \geq \min(c,\ 4\beta))$. Thus

$$\rho(K) \leq \varepsilon/[\varepsilon + \min(c,\ 4\beta)]. \tag{19}$$

This theoretical result proves the observations in Table I of Reference 13. For a given penalty parameter ε, $\rho(K) \leq \varepsilon/(\varepsilon + c)$ for $\beta \geq \beta^*$. In this range β has no influence on the convergence behaviour of the outer iteration. This appears clearly in Figure 3.

2.4.3. Study of the inner iteration. The weak point of the algorithm is the fact that an 'inner system' (16) has to be solved at each outer iteration (15). This requires most of the computational time and the overall performance of the algorithm depends on the efficiency of the solver for the inner system.

First note the particular structure of $A_{\varepsilon,\beta}$ which occurs in the *iterated penalty algorithm*. In each macroelement the components of the velocity are coupled by the term $B^{\mathrm{T}}(\beta C + \varepsilon M)^{-1}B$, which is not the case with the classical penalization matrix $A_{\varepsilon,0}$:

$$A_{\varepsilon,0} = A + (1/\varepsilon)B^{\mathrm{T}}M^{-1}B. \tag{20}$$

More precisely, the matrix $B^{\mathrm{T}}(\beta C + \varepsilon M)^{-1}B$ corresponding to the Q1–P0 discretization has the same structure as the matrix $B^{\mathrm{T}}M^{-1}B$ corresponding to the Q2–discontinuous Q1 discretization and the size of the system is consequently increased. This remains a problem when a direct matrix method is used.

If one uses a conjugate gradient solution of (16a), the matrix $A_{\varepsilon,\beta}$ is never built in practice. Unfortunately, the matrix $A_{\varepsilon,\beta}$ is ill-conditioned for a 'small' value of the penalty parameter, which makes iterative solution of (16a) difficult. The main result we want to point out is the fact that the use

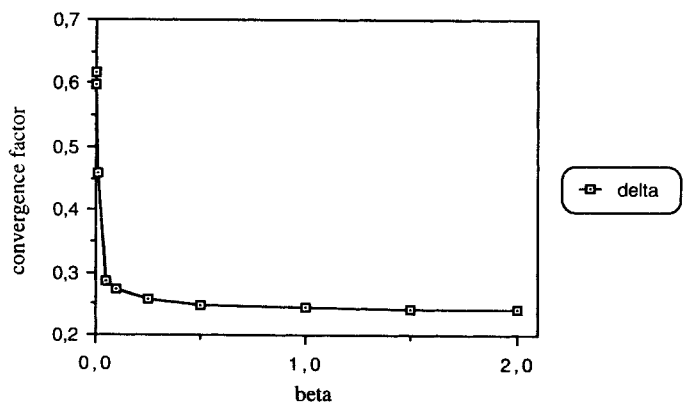

Figure 3. Convergence factor of iterated penalty algorithm ($h = 1/8$, $\varepsilon = 10^{-1}$)

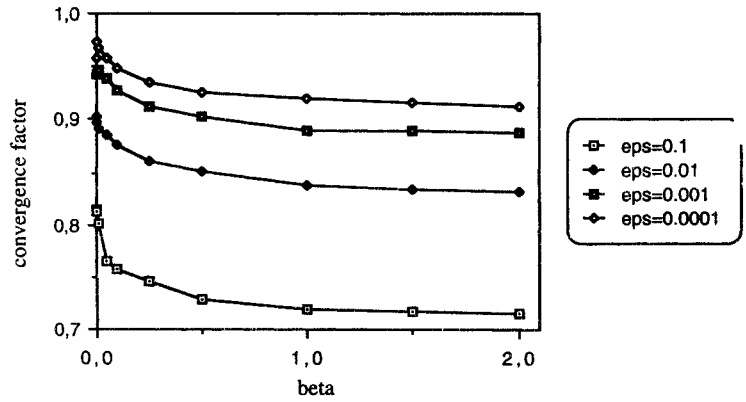

Figure 4. Convergence factor of inner iteration ($h = 1/16$)

of a large value of the stabilization parameter has a preconditioning effect[11] on the 'inner problem' and increases the convergence of (16a) significantly (Figure 4).

2.4.4. A strategy for choosing optimal parameters ε and β. From these different points of view we propose the following strategy for choosing the optimal parameters.[11]

1. Choose a moderate value of the penalty parameter in order to have a reasonable condition number for $A_{\varepsilon,\beta}$. The greater $1/\varepsilon$ is, the faster is the convergence of the outer iteration but the bigger is the condition number of the inner system. Numerical experiments show that a value of $\varepsilon = 10^{-4}$ is a good compromise between the previous considerations.
2. Now choose the stabilization parameter large enough to (a) raise the optimal convergence behaviour of the outer iteration (this is obtained for $\beta \geq \beta^*$) and (b) ensure that the matrix $A_{\varepsilon,\beta}$ will be as well-conditioned as possible (this is obtained for $\beta \geq 1$).

By carefully selecting the parameters ε and β, one can significantly enhance the overall performance of the algorithm.

3. A PRECONDITIONED CONJUGATE GRADIENT METHOD FOR THE MIXED FORMULATION

3.1. Introduction

For mixed formulations such as (1) or (3) the difficulty comes partly from the indefinite character of the operator, which has either positive or negative eigenvalues. The stabilization operator allows us to filter the spurious modes so that the zero eigenvalues of the operator $\mathbf{A}$ are changed into strictly negative eigenvalues for $\mathbf{A}_\beta$.[13] Unfortunately, the indefinite character of the stabilized operator remains as a problem. An alternative is to choose a suitable preconditioner $\bar{\mathbf{A}}_\beta$ in order to obtain a positive definite preconditioned operator $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$ and then be able to use the conjugate gradient algorithm to solve the generalized problem

$$\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta X = \bar{\mathbf{A}}_\beta^{-1}F. \tag{21}$$

This idea has been presented by Ewing *et al.*[16] and Bank *et al.*[17] for stable discretization ($\beta = 0$). We

can extend this technique to our stabilized problem (3). The suitable preconditioner[9] is given by

$$\bar{\mathbf{A}}_\beta = \begin{pmatrix} \bar{A} & B^{\mathrm{T}} \\ B & -\beta C \end{pmatrix}, \quad \text{where } \bar{A} = \mathrm{diag}(A). \tag{22}$$

It leads to a positive definite operator $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$ whose properties are given in Theorem 2.

3.2 Properties of the mixed preconditioned operator

The reader is also referred to the recent papers of Silvester–Whaten,[20,21] where the mixed operator $\mathbf{A}_\beta$ is preconditioned by a symmetric positive definite preconditioner $\bar{\mathbf{A}}$:

$$\bar{\mathbf{A}} = \begin{pmatrix} \bar{A} & 0 \\ 0 & \bar{C} \end{pmatrix}, \tag{23}$$

where $\bar{A}$ and $\bar{C}$ are both symmetric positive definite matrices. This leads to a symmetric indefinite preconditioned operator. Although different, the two approaches (22), (23) lead to the same result: the preconditioning of the Laplacian determine the spectrum of the preconditioned mixed operator (see Theorem 2 (ii) and References 20 and 21).

Theorem 2

For an operator C satisfying the stabilization condition[13] and for any given $\beta > 0$:

(i) $\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta$ is symmetric positive definite.

Under the hypothesis that the Laplacian preconditioner A satisfies (H1) $\theta_{\mathrm{m}} \leq 1$ and $\theta_{\mathrm{M}} \geq 1$:

(ii) $K\left(\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta\right) \leq K\left(\bar{A}^{-1}A\right)$

where θ_{m} represents the minimum eigenvalue of $\bar{A}^{-1}A$ and θ_{M} represents the maximum eigenvalue of $\bar{A}^{-1}A$.

Proof. Let us consider the generalized eigenvalue problem

$$Au + B^{\mathrm{T}}p = \lambda\bar{A}u + \lambda B^{\mathrm{T}}p, \tag{24a}$$

$$Bu - \beta Cp = \lambda Bu - \lambda\beta Cp. \tag{24b}$$

(a) *$\lambda = 1$ is always an eigenvalue of this problem*, corresponding to the eigenvector $(0, p)$.

(b) *Let us now consider the eigenvalues $\lambda \neq 1$.* Equation (24b) implies $(\lambda - 1)(Bu - \beta Cp) = 0$ and thus $Bu = \beta Cp$, since $\lambda \neq 1$. The corresponding eigenvector (u, p) is such that $u \neq 0$. In fact, $u = 0$ would lead to $Cp = 0$ and $B^{\mathrm{T}}p = 0$, since $\lambda \neq 1$. According to the stabilization condition,[13] we would have $p = 0$. Combining (24a) and $u^{\mathrm{T}}B^{\mathrm{T}} = (Bu)^{\mathrm{T}} = \beta p^{\mathrm{T}}C$ gives $u^{\mathrm{T}}Au = \lambda u^{\mathrm{T}}\bar{A}u + \beta(\lambda - 1)p^{\mathrm{T}}Cp$. Letting $\gamma = p^{\mathrm{T}}Cp/u^{\mathrm{T}}\bar{A}u$, we obtain

$$\theta_{\mathrm{m}} \leq (1 + \beta\gamma)\lambda - \beta\gamma \leq \theta_{\mathrm{M}}, \tag{25a}$$

$$(\theta_{\mathrm{m}} + \beta\gamma)/(1 + \beta\gamma) \leq \lambda \leq (\theta_{\mathrm{M}} + \beta\gamma)/(1 + \beta\gamma). \tag{25b}$$

and then

$$\theta_{\mathrm{m}} \leq \lambda \leq \theta_{\mathrm{M}} \tag{25c}$$

since θ_{m} and θ_{M} satisfies (H1) and since we have

$$\frac{a + x}{b + x} \leq \frac{a}{b} \text{ for all } 0 \leq x,\ 0 < b \leq a$$

and

$$\frac{a}{b} \leq \frac{a+x}{b+x} \text{ for all } 0 \leq x,\ 0 < a \leq b.$$

The desired results (i) and (ii) follows from (a), (b), (25c) and (H1). This completes the proof. □

Remarks.

(1) In applications, the preconditioner $\bar{A}$ may not satisfy (H1) *a priori*. Nevertheless, it is easy to compute an appropriate factor ρ for the preconditioner so that $\rho\bar{A}$ satisfies (H1):

$$(\rho\bar{A})^{-1}A = (1/\rho)\bar{A}^{-1}A \text{ with } \rho \text{ such that } (1/\rho)\theta_m < 1 < (1/\rho)\theta_M, \text{ i.e. } \theta_M < \rho < \theta_M.$$

(2) If there exist $u_m \neq 0$ and $u_M \neq 0$ such that:

$$\bar{A}^{-1}Au_m = \theta_m u_m \text{ and } Bu_m = 0,$$
$$\bar{A}^{-1}Au_M = \theta_M u_M \text{ and } Bu_M = 0,$$

one can see that

$$\theta_m \text{ is eigenvalue of } \bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta \text{ corresponding to } (u_m,\ 0),$$
$$\theta_M \text{ is eigenvalue of } \bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta \text{ corresponding to } (u_M,\ 0),$$

Thus $K(\bar{\mathbf{A}}_\beta^{-1}\mathbf{A}_\beta) = K(\bar{A}^{-1}A)$.

The latter result Theorem 2 (ii) obtained with the mixed formulation of the problem is inherently different from the corresponding result $K(L_\beta) \leq \gamma(\beta)$ obtained with the dual formulation. The condition number of the preconditioned operator, and thereby the convergence factor of the method, is now majored *independently of the stabilization parameter*. Thus there is certainly no great advantage in varying β. This appears clearly in the numerical results of Section 3.4. However, the algorithm has a classical embedded outer–inner iteration structure and the inner iteration behaviour depends on β. Thus an optimal parameter is available for the overall algorithm.

Like the Uzawa-type algorithm, the mixed algorithm takes full advantage of available preconditioners for A. For example, an incomplete Cholesky factorization as preconditioning has been used successfully. On the other hand, $\bar{A}$ can be defined implicitly from a small number of iterations of some iterative methods for solving Poisson-type problems.[17] Iterative techniques giving convergence factors bounded independently of the mesh size are certainly among the best choices. We can use multigrid[18] or FAC[25] methods in the case of (so-called) composite grids.

3.3. *Implementation and variants of the method*

For each iteration step the method requires the solution of the preconditioned system

$$\begin{pmatrix} \bar{A} & B^T \\ B & -\beta C \end{pmatrix} \begin{pmatrix} z_u \\ z_p \end{pmatrix} = \begin{pmatrix} r_u \\ r_p \end{pmatrix}, \tag{26}$$

which can be reformulated as

$$\text{solve } (B\bar{A}^{-1}B^T + \beta C)z_p = B\bar{A}^{-1}r_u - r_p, \tag{27a}$$
$$\text{compute } z_u = \bar{A}^{-1}(r_u - B^T z_p). \tag{27b}$$

A dual-type problem (27a) has to be solved at each step (26). This can be done iteratively, possibly by the conjugate gradient method. As for the previous penalty-iterated algorithm, the overall algorithm has a classical embedded inner–outer iteration structure.

Since one wants the cost of (27a) to be small, Bank *et al.*[17] have suggested solving (27a) only approximately by performing fewer iterations in (27a) than are required for exact termination. See Reference 17 for details in the stable mixed approximation case.

Here we propose variant of the algorithm in which (27a) is replaced by

$$\bar{L}_\beta z_p = B\bar{A}^{-1} r_u - r_p, \tag{28}$$

where $\bar{L}_\beta$ is an approximation of $B\bar{A}^{-1}B^{\mathrm{T}} + \beta C$. We find it convenient to use our (so-called) *macroblock-type preconditioner*[6,9]

$$\bar{L}_\beta = \operatorname{diag}\left(B\bar{A}^{-1}B^{\mathrm{T}}\right) + \beta C.$$

It has been proven to be a good preconditioner for the stabilized dual operator $BA^{-1}B^{\mathrm{T}} + \beta C$ and significant savings in work are obtained compared with the first choice (27a). The resulting algorithm has a much simpler single-level iteration structure, since (28) is solved directly.

3.4. Estimate for the convergence factor of the algorithm

In Figure 5 we plot the convergence factor $\delta = (R^n/R^0)^{1/n}$ of the preconditioned conjugate gradient method (PCGM) as a function of the stabilization parameter. The previous preconditioners are investigated. For ease of notation in the figures, the method is called PCGM-S (resp. PCGM-MB) when the preconditioner (27) (resp. (28)) is employed.

Remarks.

(1) In this case the norm of the residual is

$$R^i = \left(\left\|f - Au^i - B^{\mathrm{T}}p^i\right\|^2 + \left\|Bu^i - \beta Cp^i\right\|^2\right)^{1/2}.$$

(2) The algorithm with the preconditioner (27) is denoted PCGM-S because of its resemblance to the SIMPLE pressure correction algorithm.[26]

All the numerical experiments indicate that the convergence factor of PCGM-S is almost insensitive to the value of the stabilization parameter β. In agreement with (ii), this factor is comparable with that obtained for the diagonally scaled Laplacian. However, the convergence of the inner iteration depends on β. For the solution of the dual-type problem (27a) we use the conjugate gradient method with the macroblock-type preconditioner (28). The numerical tests show that an increase in β improves the

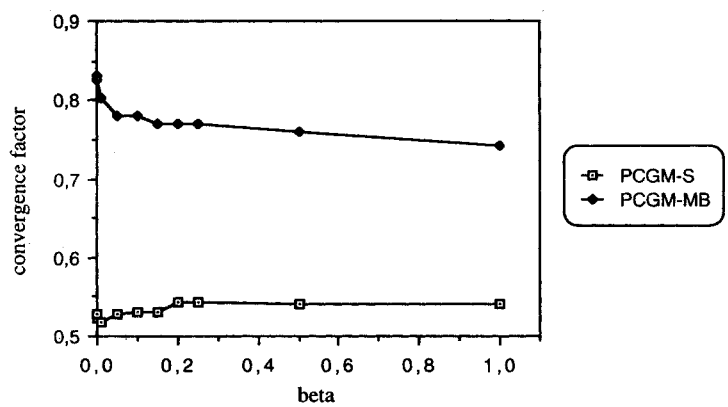

Figure 5. Convergence factor of PCGM for mixed formulation ($h = 1/16$)

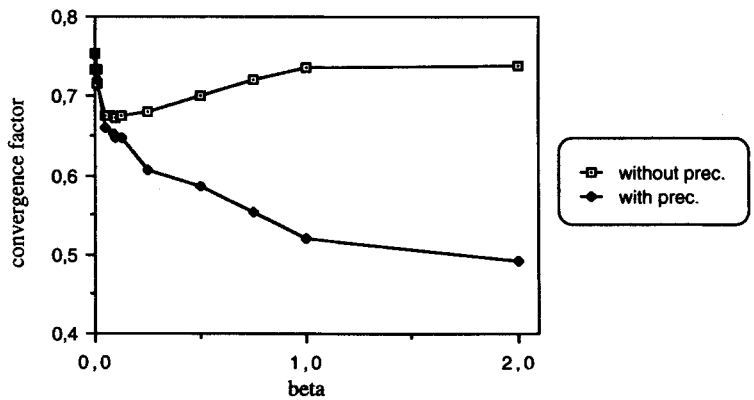

Figure 6. Solution of dual-type system ($h = 1/16$)

convergence of (27a). This appears clearly in Figure 6. The 'best' convergence factors are obtained when β is large, say 10^2. Thus $\beta^* = 10^2$ is the optimal choice for the overall algorithm.

On the other hand, the convergence factor of PCGM-MB varies with β. It seems to be a decreasing function of the stabilization parameter β and the 'best' convergence factors are obtained when $\beta \geq 1$.

The next numerical results are intended to compare the numerical performance of the algorithms with the very simple case of diagonal preconditioning[13] (PCGM-D):

$$\begin{pmatrix} \text{diag}(A) & 0 \\ 0 & \beta\ \text{diag}(C) \end{pmatrix}. \tag{29}$$

In any case the optimal choice $\beta = \beta^*$ is chosen. The optimal parameter for PCGM-D is $\beta^* = O(10^2)$ according to Reference 13.

The convergence of each algorithm is analysed by plotting the norm of the residual versus the number of iterations. A typical iteration history is shown in Figure 7.

PCGM-S shows a monotonic decrease in residual. With the other two methods the residual is an erratic function of the iteration number. The first preconditioner is clearly more efficient in terms of convergence factor. The disadvantage of this method is the amount of work due to the solution of a dual-type problem at each step of the algorithm. PCGM-S is obviously the most expensive per iteration step in terms of CPU time and storage.

PCGM-MB involves nearly the same amount of work as PCGM-D, since the preconditioning step (28) is intended to solve independent subsystems of very small size.[6,9] However, PCGM-MB performs well compared with PCGM-D. It requires about three times fewer iterations than PCGM-D. For these reasons we recommend the macroblock-type preconditioner PCGM-MB. Its attractive features are its inherent simplicity and its reasonable CPU costs and memory requirements.

Also of significance is the fact that the preconditioning step (28) is easily parallelizable. Attempts towards parallelization concern not only (28) but also all the other steps of the algorithm, i.e. the computations of Ax, $B^T y$, Bx and Cy for given x and y. See Reference 9 for a fuller discussion of this issue.

4. CONCLUSIONS

We have studied the influence of the stabilization parameter on various iterative methods for the solution of the Stokes problem discretized by the (so-called) locally stabilized Q1–P0 element. Both theoretical and numerical results have been presented for new approaches updating classical algorithms

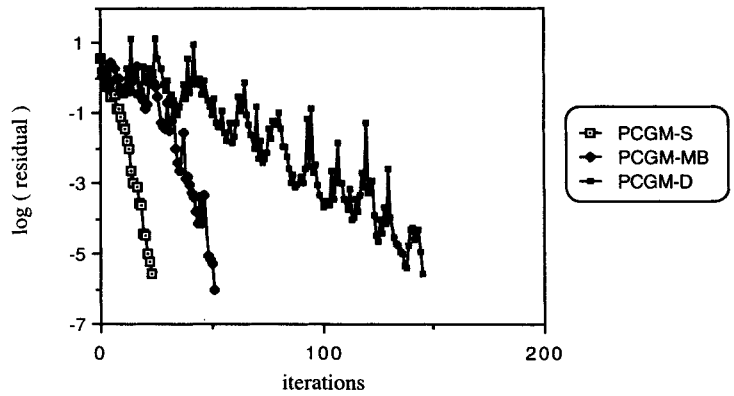

Figure 7. Efficiency of preconditioners ($h = 1/16$)

for either a dual formulation, a penalty formulation or a mixed one. Our main conclusion is that rapid convergence can be obtained with a suitable choice of the stabilization parameter.

The ideas of the analysis can be used to analyse some others iterative methods for the stabilized Stokes problem. See References 4, 9 and 10 for a multigrid approach.

Moreover, we think that the results presented in this paper are open to considerable improvement since they take full advantage of available preconditioners or solvers for the Laplacian. Further developments are still under way for taking into account the case of composite grids. We think that an algorithm combining one of our stabilized context Stokes solvers and an FAC approach for the Laplacian will work very well. This is the main direction of our research.

REFERENCES

1. I. Babuska, 'The finite element method with Lagrangian multipliers', *Numer. Math.*, **20**, 179–192 (1973).
2. F. Brezzi, 'On the existence, uniqueness and approximation of saddle-point problems arising from Lagrange multipliers', *RAIRO*, **8**, 129–151 (1974).
3. R. L. Sani, R. M. Gresho, R. L. Lee and D. F. Griffiths, 'The cause and cure (?) of the spurious pressures generated by certain FEM solutions of the incompressible Navier–Stokes equations: Part 1', *Int. J. Numer. Methods Fluids*, **1**, (1981).
4. J. Pitkaranta and T. Saarinen, 'A multigrid version of a simple finite element method for the Stokes problem', *Math. Comput.*, **45**, 1–14 (1985).
5. M. P. Robichaud, P. A. Tanguy and M. Fortin, 'An iterative implementation of the Uzawa algorithm for 3-D fluid flow problems', *Int. J. Numer. Methods Fluids*, **10**, 429–442 (1990).
6. C. Vincent and R. Boyer, 'A preconditioned conjugate gradient Uzawa- type method for the solution of the Stokes problem by mixed Q1–P0 stabilized finite elements', *Int. J. Numer. Methods Fluids*, **14**, 289–298 (1992).
7. N. Kechkar and D. J. Silvester, 'Analysis of locally stabilized mixed finite element methods for the Stokes problem', *Math. Comput.*, **58**, 1–10 (1992).
8. D. J. Silvester and N. Kechkar, 'Stabilized bilinear–constant velocity–pressure finite elements for the conjugate gradient solution of the Stokes problem', *Comput. Methods Appl. Mech. Eng.*, **79**, 71–86 (1992).
9. C. Vincent, 'Méthodes de gradient conjugué préconditionné et techniques multigrilles pour la résolution du problème de Stokes par éléments finis mixtes Q1–P0 stabilisés–applications', *Thèse*, Université de Provence, Marseille, 1991.
10. E. Wabbo, 'Résolution du problème de Stokes: méthodes multigrilles adaptées a certains schémas stabilisés', *Thèse*, Ecole Centrale de Lyon, 1990.
11. C. Vincent, 'Influence of the stabilization parameter on the convergence factor of Uzawa-type algorithms for the solution of the Stokes problem', *Math. Modell. Sci. Comput. Sci. Technol.*, in press.
12. C. Vincent, 'On the benefit of stabilizing the Q1–P0 finite element in Uzawa-type algorithms for the Stokes problem', *Modell. Sci. Comput.*, in press.
13. J. Atanga and D. Silvester, 'Iterative methods for stabilized mixed velocity–pressure finite elements', *Int. J. Numer. Methods Fluids*, **14**, 71–81 (1992).
14. R. Stenberg, 'Analysis of finite elements methods for the Stokes problem: a unified approach', *Math. Comput.*, **42**, 9–33 (1984).

15. C. Vincent, 'Stabilized-context iterative methods for the solution of the discretized Stokes problem', *Math. Modell. Sci. Comput. Sci. Technol.*, in press.
16. R. E. Ewing, R. D. Lazarov, Z. Z. Penglu and P. S. V. Assilevski, 'Preconditioning indefinite systems arising from mixed finite element discretization of second-order elliptic problems', *Numerical Analysis Report*, Department of Mathematics, University of Wyoming, 1995.
17. R. E. Bank, B. D. Welfert and H. Yserentant, 'A class of iterative methods for solving saddle point problems', *Numer. Math.*, **56**, 645–666 (1990).
18. J. Cahouet and J. P. Chabard, 'Some fast 3D finite element solvers for the generalized Stokes problem', *Int. J. Numer. Methods Fluids*, **8**, 869–895 (1988).
19. R. Verfurth, 'A combined conjugate gradient multigrid algorithm for the numerical solution of the Stokes problem', *IMA J. Numer. Anal.*, **4**, 441–455 (1984).
20. A. J. Whaten and D. J. Silvester, 'Fast iterative solution of stabilized Stokes systems, Part I: using simple diagonal preconditioners', *SIAM J. Num. Anal.*, **30**, 630–649 (1993).
21. D. J. Silvester and A. J. Whaten, 'Fast iterative solution of stabilized Stokes systems, Part II: using general block preconditioners', *SIAM J. Num. Anal.*, **31**, 1352–1367 (1994).
22. M. Fortin and R. Glowinski, *Augmented Lagrangian Methods*, North-Holland, Amsterdam, 1983.
23. J. H. Bramble and J. E. Pasciak, 'A preconditioning technique for indefinite systems resulting from mixed approximations of elliptic problems', *Math. Comput.*, 1–17 (1988).
24. C. Johnson and J. Pitkaranta, 'Analysis of some mixed finite element methods related to reduced integration', *Math. Comput.*, **38**, 375–400 (1982).
25. MacCormick, 'The F.A.C. (fast adaptative composite grid) method for elliptic boundary value problems', *Math. Comput.*, **46**, 439–536 (1986).
26. S. Sivalogavathan and G. S. Shaw, 'On the smoothing properties of the SIMPLE pressure-correction algorithm', *Int. J. Numer. Methods Fluids*, **8**, 441–461 (1988).