Computer Methods in Applied Mechanics and Engineering 110 (1993) 237–262
North-Holland

An iterative penalty method for the finite element solution of the stationary Navier–Stokes equations

Ramon Codina

Escola Tècnica Superior d'Enginyers de Camins, Canals i Ports, Universitat Politècnica de Catalunya, Gran Capità s/n, Mòdul C1, 08034 Barcelona, Spain

Received 23 October 1991

The objective of this paper is to analyse an iterative procedure for the finite element solution of the Stokes and Navier–Stokes stationary problems. For the latter case, the usual condition on the viscosity and the data that ensures uniqueness is assumed. The method is based on the iterative imposition of the incompressibility condition via penalization. Theoretical and numerical results show that this constraint can be approximated iteratively within the same iterative loop used to deal with the nonlinear term of the equations. Two particular iterative schemes are analysed, namely those based on the Picard and Newton–Raphson algorithms.

1. Introduction

The mixed velocity–pressure finite element solution of the incompressible Navier–Stokes equations has several disadvantages due to the zero divergence condition for the velocity field. If the standard Galerkin formulation is used, the first problem to be faced is the use of compatible spaces for the velocity and the pressure, in the sense that they have to satisfy the inf–sup or LBB condition [1, 2] (see also [3] for a simple derivation of this condition for the discrete problem). A remedy that is gaining popularity is the use of the Galerkin Least Squares approach introduced in [4, 5], especially effective when a continuous interpolation for the pressure is used (otherwise, high order velocity interpolation or non-standard assembly algorithms have to be employed [6]). Recently, quasi-optimal convergence of this method has been proved in [7] for the time-dependent Navier–Stokes equations using space–time linear elements. In any case, the formulation depends on an algorithmic parameter whose physical meaning and optimal values are not yet known. This, and the fact that pressures appear as nodal variables (see below) may decide the user in favor of the Galerkin formulation, perhaps with an upwind technique for high Reynolds number flows.

If the Galerkin formulation is used, the matrix of the discrete algebraic system resulting from the finite element discretization has zero diagonal terms. The use of iterative solvers or inefficient renumbering algorithms seems to be the only available remedy for solving this system of equations. However, the penalty approach circumvents this problem and has other interesting features. If the pressure interpolation is discontinuous, one can eliminate the element unknowns of this field in terms of the velocity nodal unknowns. Substitution of the obtained expression in the momentum equation leads to a system whose only degrees of freedom are velocities. The reduction of the number of nodal unknowns and the fact that the method is known to work well, have made the penalty method very popular, especially in the engineering literature [8–11]. Perhaps the only drawback of this approach is the ill-conditioning of the stiffness matrix when the penalty parameter is very small. A lower bound is determined basically by the computer and the arithmetic precision used in the calculation.

Our objective in this paper is to present and analyse an iterative penalty finite element method for

Correspondence to: Prof. Ramon Codina, E.T.S. Ingenieros de Caminos, Universitat Politècnica de Catalunya, Gran Capitán s/n, Mòdul C1, 08034 Barcelona, Spain.

the stationary Stokes and Navier–Stokes equations. The goal is the convergence of the iterates to the true incompressible solution. The main advantage of this approach is that larger penalty parameters can be used, thus alleviating the ill-conditioning mentioned above. The basic idea is solving the penalized equations in each iteration but adding a right-hand side term that is basically the residual of the incompressibility equation of the previous iterate. For the Stokes equations, this approach is the only reason for an iterative scheme and the conditions under which convergence is achieved are only determined by the iterative penalization. However, the Navier–Stokes equations must be solved iteratively. The question that naturally arises is whether the iterative scheme employed can be coupled with the iterative penalization or not. We prove that, under not very restrictive conditions, the answer is yes.

The exposition is organized as follows. In the next section, notation and the statement of the problem are presented. The Stokes problem is considered in Section 3, where the idea of the iterative penalization is described in detail. Section 4 deals with the Navier–Stokes equations when the Picard (or successive substitution) algorithm is used for the nonlinear term and Section 5 when the Newton–Raphson scheme is employed. The decoupling of the nonlinear and penalization iterative loops is studied in Section 6. Section 7 contains the results of numerical experiments performed for two well known benchmark tests (the driven cavity flow and the flow over a backward facing step) and finally some comments are made and conclusions are drawn.

2. Notation and statement of the problem

Let Ω be an open bounded domain of $\mathbb{R}^N$ ($N=2$ or 3) and $\Gamma=\partial\Omega$ its boundary, assumed to be locally Lipschitz. The Navier–Stokes problem for an incompressible fluid moving in Ω with, for simplicity, homogeneous boundary conditions, consists in finding a velocity field $\boldsymbol{u}$ and a pressure p such that

$$\begin{aligned} &(\boldsymbol{u}\cdot\nabla)\boldsymbol{u}-\nu\,\Delta\boldsymbol{u}+\nabla p=\boldsymbol{f} && \text{in } \Omega\,,\\ &\nabla\cdot\boldsymbol{u} && \text{in } \Omega\,,\\ &\boldsymbol{u}=0 && \text{on } \Gamma\,, \end{aligned} \tag{2.1}$$

where $\boldsymbol{f}$ is a given body force and ν is the kinematic viscosity. In order to write the weak form of problem (2.1), we introduce the spaces

$$V=H_0^1(\Omega)^N\,,\qquad Q=L^2(\Omega) \tag{2.2}$$

and the multilinear forms

$$\begin{aligned} &a(\boldsymbol{u},\boldsymbol{v})=\nu\int_\Omega \nabla\boldsymbol{u}:\nabla\boldsymbol{v}\,,\\ &b(q,\boldsymbol{v})=\int_\Omega q\nabla\cdot\boldsymbol{v}\,,\\ &c(\boldsymbol{u},\boldsymbol{v},\boldsymbol{w})=\int_\Omega((\boldsymbol{u}\cdot\nabla)\boldsymbol{v})\cdot\boldsymbol{w}\,,\\ &l(\boldsymbol{v})=\langle\boldsymbol{f},\boldsymbol{v}\rangle\,, \end{aligned} \tag{2.3}$$

defined on $V\times V$, $Q\times V$, $V\times V\times V$ and V, respectively. $\langle\cdot,\cdot\rangle$ denotes the duality paring between V and its topological dual V' ($=H^{-1}(\Omega)^N$). If the viscous term in (2.1) is written as $-\nabla\cdot(2\nu\boldsymbol{\varepsilon}(\boldsymbol{u}))$, where $\boldsymbol{\varepsilon}(\boldsymbol{u})$ is the symmetric part of $\nabla\boldsymbol{u}$, the bilinear form a to be considered is

$$a(\boldsymbol{u},\boldsymbol{v})=2\nu\int_\Omega \boldsymbol{\varepsilon}(\boldsymbol{u}):\boldsymbol{\varepsilon}(\boldsymbol{v})\,, \tag{2.4}$$

instead of that appearing in (2.3). Continuity of a, b and l is obvious. Continuity of c follows from Sobolev's imbedding Theorem (if $\boldsymbol{u}$ and $\boldsymbol{v} \in V$, then $\boldsymbol{u}$ and $\boldsymbol{v} \in L^4(\Omega)^N$ for $N = 2, 3$) and from Hölder's inequality (if u_j, $v_k \in L^4(\Omega)$, then $u_j v_k \in L^2(\Omega)$, u_j and v_k being the components of $\boldsymbol{u}$ and $\boldsymbol{v}$). See [12, 13] for details. Since a, c and l are continuous, we can define their 'norms' by

$$N_a = \sup \frac{a(\boldsymbol{v}_1, \boldsymbol{v}_2)}{\|\boldsymbol{v}_1\| \|\boldsymbol{v}_2\|}, \qquad N_c = \sup \frac{c(\boldsymbol{v}_1, \boldsymbol{v}_2, \boldsymbol{v}_3)}{\|\boldsymbol{v}_1\| \|\boldsymbol{v}_2\| \|\boldsymbol{v}_3\|}, \qquad N_l = \sup \frac{l(\boldsymbol{v}_1)}{\|\boldsymbol{v}_1\|}, \tag{2.5}$$

where the supremum is taken over all the $\boldsymbol{v}_1$, $\boldsymbol{v}_2$, $\boldsymbol{v}_3 \in V - \{0\}$ and $\|\cdot\|$ denotes the usual norm in V. We will use the same symbol for the norm in Q and $(\cdot,\cdot)$ for the scalar product both in the spaces V and Q.

Define the space

$$Z = \{q \in Q \,|\, b(q, \boldsymbol{v}) = 0 \; \forall \boldsymbol{v} \in V\} . \tag{2.6}$$

For b given by (2.3), $Z = \mathbb{R}$. In the quotient space Q/Z, the following norm is defined:

$$\|q\|_{Q/Z} = \inf_{z \in Z} \|q + z\| . \tag{2.7}$$

Having introduced all this notation, the weak form of problem (2.1) can be written as follows: find $\boldsymbol{u} \in V$ and $p \in Q/Z$ such that

$$\begin{aligned} &c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) + a(\boldsymbol{u}, \boldsymbol{v}) - b(p, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &b(q, \boldsymbol{u}) = 0 \qquad\qquad\qquad\qquad \forall q \in Q . \end{aligned} \tag{2.8}$$

Besides the continuity of all the forms involved in (2.8), we will assume that the bilinear form a is coercitive and that b satisfies the LBB condition, i.e., there exist positive constants K_a and K_b such that

$$a(\boldsymbol{v}, \boldsymbol{v}) \geqslant K_a \|\boldsymbol{v}\|^2 \qquad \forall \boldsymbol{v} \in V , \tag{2.9}$$

$$\sup \frac{b(q, \boldsymbol{v})}{\|\boldsymbol{v}\|} \geqslant K_b \|q\|_{Q/Z} , \quad \boldsymbol{v} \in V - \{0\}, \forall q \in Q . \tag{2.10}$$

Condition (2.9) follows from Poincaré–Friedrics inequality if a is given by (2.3) and from Korn's inequality if it is given by (2.4). Condition (2.10) holds for V and Q given by (2.2).

For the trilinear form c it will be assumed that

$$c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in V . \tag{2.11}$$

If $\boldsymbol{u}$ is the solution of problem (2.8), it is easy to see that condition (2.11) is satisfied. However, we will be interested in velocity fields that do not exactly satisfy the incompressibility condition. In this case, instead of the form c given in (2.3), its skew-symmetrized form is used:

$$c_\sigma(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) + \frac{1}{2} \int_\Omega (\nabla \cdot \boldsymbol{u}) \boldsymbol{v} \cdot \boldsymbol{w} .$$

It can be easily checked that $c_\sigma(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{v}) = 0$ and that $c_\sigma(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w}) = c(\boldsymbol{u}, \boldsymbol{v}, \boldsymbol{w})$ if $\boldsymbol{u}$ is the solution of problem (2.8). Continuity of c_σ can be proved as for c. Thus, c_σ can be used instead of c in (2.8) and condition (2.11) will hold. In any case subscript σ is omitted.

Finally, we will assume that

$$\rho = \frac{N_c N_l}{K_a^2} < 1 . \tag{2.12}$$

Under all these conditions, existence and uniqueness of solution of (2.8) can be proved [12].

In what follows, the spaces V and Q will be those given by (2.2) or finite dimensional subspaces V_h and Q_h arising from the finite element discretization of Ω (internal approximation). Conditions (2.9), (2.11) and (2.12) will be automatically satisfied if V and Q are replaced by V_h and Q_h. However, (2.10) has to be explicitly required for each pair of finite element spaces V_h, Q_h.

3. Iterative penalization for the Stokes problem

The problem we consider in this section is (2.8) with $c = 0$, i.e., to find $\boldsymbol{u} \in V$ and $p \in Q/Z$ such that

$$\begin{aligned} &a(\boldsymbol{u}, \boldsymbol{v}) - b(p, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &b(q, \boldsymbol{u}) = 0 \qquad\qquad\quad \forall q \in Q . \end{aligned} \tag{3.1}$$

The iterative penalty method that will be analysed is particularly simple to introduce for this linear problem.

3.1. Motivation and statement of the algorithm

If the penalty method is applied to solve (3.1), one has to find $\boldsymbol{u}^{\varepsilon(1)} \in V$ and $p^{\varepsilon(1)} \in Q$ such that

$$\begin{aligned} &a(\boldsymbol{u}^{\varepsilon(1)}, \boldsymbol{v}) - b(p^{\varepsilon(1)}, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &\varepsilon(p^{\varepsilon(1)}, q) + b(q, \boldsymbol{u}^{\varepsilon(1)}) = 0 \qquad \forall q \in Q , \end{aligned} \tag{3.2}$$

where ε is a small positive number. Convergence (in norm) of $\boldsymbol{u}^{\varepsilon(1)}$ to $\boldsymbol{u}$ and of $p^{\varepsilon(1)}$ to p when $\varepsilon \to 0$ is a well known result [11, 12, 14]. Once $\boldsymbol{u}^{\varepsilon(1)}$ and $p^{\varepsilon(1)}$ are found, define $\delta\boldsymbol{u}$ and δp such that $\boldsymbol{u} = \boldsymbol{u}^{\varepsilon(1)} + \delta\boldsymbol{u}$ and $p = p^{\varepsilon(1)} + \delta p$. Then, $\delta\boldsymbol{u}$ and δp will be the solution of

$$\begin{aligned} &a(\delta\boldsymbol{u}, \boldsymbol{v}) - b(\delta p, \boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in V , \\ &b(q, \delta\boldsymbol{u}) = \varepsilon(p^{\varepsilon(1)}, q) \quad \forall q \in Q . \end{aligned}$$

Now, this problem can also be solved using the penalty method. If $\delta\boldsymbol{u}^{\varepsilon}$, δp^{ε} is the penalized solution and we define $\boldsymbol{u}^{\varepsilon(2)} = \boldsymbol{u}^{\varepsilon(1)} + \delta\boldsymbol{u}^{\varepsilon}$, $p^{\varepsilon(2)} = p^{\varepsilon(1)} + \delta p^{\varepsilon}$, we will have that

$$\begin{aligned} &a(\boldsymbol{u}^{\varepsilon(2)}, \boldsymbol{v}) - b(p^{\varepsilon(2)}, \boldsymbol{v}) = l(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in V , \\ &\varepsilon(p^{\varepsilon(2)}, q) + b(q, \boldsymbol{u}^{\varepsilon(2)}) = \varepsilon(p^{\varepsilon(1)}, q) \quad \forall q \in Q . \end{aligned} \tag{3.3}$$

It should be remarked that the pressures that are solution of (3.2) and (3.3) belong to the space

$$Q_0 = \left\{ q \in Q \,\middle|\, \int_\Omega q = 0 \right\} \tag{3.4}$$

that is isomorphic to Q/Z for b defined in (2.3) [15]. Pressures that are solution of (2.8) and (3.1) are determined up to an additive constant that can be fixed seeking p either in Q_0 or in Q/Z. From now onwards, the former choice will be employed since penalized solutions automatically belong to Q_0.

The argument used above to arrive at problem (3.3) may be applied iteratively. This leads to the following algorithm: given $p^{\varepsilon(0)} \in Q_0$, for $i = 1, 2, \ldots$, find $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ such that

$$\begin{aligned} &a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) - b(p^{\varepsilon(i)}, \boldsymbol{v}) = l(\boldsymbol{v}) \qquad \forall \boldsymbol{v} \in V , \\ &\varepsilon(p^{\varepsilon(i)}, q) + b(q, \boldsymbol{u}^{\varepsilon(i)}) = \varepsilon(p^{\varepsilon(i-1)}, q) \quad \forall q \in Q . \end{aligned} \tag{3.5}$$

Existence and uniqueness of solution follows considering the problem in the space $V \times Q$ and applying Lax–Milgram's lemma. This algorithm will be analyzed below and extended to the Navier–Stokes equations. Before that, some other existing methods are discussed.

3.2. Some remarks on related methods

From (3.5) we see that the penalized equation implies

$$p^{\varepsilon(i)} = p^{\varepsilon(i-1)} - \frac{1}{\varepsilon} \nabla \cdot \boldsymbol{u}^{\varepsilon(i)} \tag{3.6}$$

in the space Q for b given by (2.3). For the continuous problem, (3.6) holds in the classical sense, since for $\boldsymbol{u}^{\varepsilon(i)}$ belonging to $H_0^1(\Omega)^N$ the pressure space defined by (3.6) is precisely $L^2(\Omega)$. However, for the discrete finite element problem only the weak version in (3.5) makes sense. Otherwise, once the velocity space V is defined, the pressure space Q is given by (3.6) and the pair V, Q does not necessarily satisfy the LBB condition. If one starts imposing (3.6) for the continuous problem and then eliminating the pressure in the momentum equation, reduced integration techniques (RIP methods) have to be applied when the equations are discretized in order to obtain a stable (or semi-stable) pressure space [11, 16]. The connection between this approach and the weak penalty method employed here is now well understood [9, 15, 17]. RIP methods are not considered in this work.

Algorithm (3.5) for the Stokes problem is not new (cf. [14]) and may be obtained from different approaches. One of them is the Augmented Lagrangian method, provided that Uzawa's algorithm is used to uncouple the pressure computation (see, e.g. [13, 18]). In this case, the resulting problem will be different if the continuous or the discrete Lagrangian is augmented with the constraint. In the former option, the same remarks as for RIP methods may be applied. In any case, pressure updating is slightly different from the approach used here. The term $\varepsilon(p^{\varepsilon(i)}, q)$ in (3.5) leads to a Gramm matrix once pressure is interpolated whereas the identity matrix appears in Uzawa's updating. The main point, however, is that we do not assume that the bilinear form $a(\cdot, \cdot)$ is symmetric (although it certainly is for the problem we consider) and thus our analysis is not based on the existence of an associated minimization problem for (3.1).

Another way to obtain algorithm (3.5) is to introduce a false transient only for the pressure, assuming the fluid to be slightly compressible (artificial compressibility method) and then to discretise the temporal derivative using the backward Euler scheme (see e.g. [13]). In this case, the penalty parameter ε would be the inverse of $c^2 \Delta t$, where c is the speed of sound in the fluid and Δt is the time step. This approach makes sense for algorithm (3.5) and for algorithm (6.1) considered in Section 6. However, it ceases to be valid when the second equation in (3.5) is coupled with a linearized form of the Navier–Stokes equations, whereas the residual argument used above can be easily extended in these cases.

3.3. Convergence of the algorithm

Before studying the convergence of the iterates of (3.5), let us state two simple results.

LEMMA 3.1. If $q \in Q_0$, then $\|q\|_{Q/Z} = \|q\|$.

PROOF. It follows directly from the definition of $\|\cdot\|_{Q/Z}$ and the fact that, in our case, $Z = \mathbb{R}$:

$$\begin{aligned} \|q\|_{Q/Z} &= \inf_{c \in \mathbb{R}} \|q + c\| \\ &= \inf_{c \in \mathbb{R}} \{\|q\|^2 + \|c\|^2 + 2(q, c)\}^{1/2} \\ &= \inf_{c \in \mathbb{R}} \{\|q\|^2 + \|c\|^2\}^{1/2} \\ &= \|q\| . \quad \square \end{aligned}$$

This lemma allows us to omit the subscript Q/Z when using condition (2.10). We also need the following a priori estimates.

LEMMA 3.2. Let $\boldsymbol{u}$ *and* p *be the solution of the Stokes problem (3.1). Then*

$$\|\boldsymbol{u}\| \leq \frac{N_l}{K_a}, \tag{3.7}$$

$$\|p\| \leq \frac{N_l}{K_b}\left(1+\frac{N_a}{K_a}\right). \tag{3.8}$$

PROOF. Taking $\boldsymbol{v}=\boldsymbol{u}$ in (3.1), we obtain

$$K_a\|\boldsymbol{u}\|^2 \leq a(\boldsymbol{u},\boldsymbol{u}) = l(\boldsymbol{u}) \leq N_l\|\boldsymbol{u}\|$$

and (3.7) follows. On the other hand, condition (2.10) implies that there exists $\boldsymbol{v}\in V-\{0\}$ such that

$$\begin{aligned} K_b\|p\|\,\|\boldsymbol{v}\| &\leq b(p,\boldsymbol{v}) \\ &= a(\boldsymbol{u},\boldsymbol{v}) - l(\boldsymbol{v}) \\ &\leq (N_l + N_a\|\boldsymbol{u}\|)\|\boldsymbol{v}\| \end{aligned}$$

and using (3.7) we obtain (3.8). □

We now proceed to the main result of this section.

THEOREM 3.3. Let $(\boldsymbol{u},p)\in V\times Q_0$ *be the solution of the Stokes problem (3.1) and* $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)})\in V\times Q_0$ *the solution of (3.5). Define*

$$\bar{\varepsilon} = \varepsilon\frac{N_a^2}{K_aK_b^2}.$$

If $\bar{\varepsilon}<1$, *then*

$$\lim_{i\to\infty}\|p-p^{\varepsilon(i)}\| = 0, \qquad \lim_{i\to\infty}\|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)}\| = 0.$$

Moreover, convergence is linear with $\bar{\varepsilon}$:

$$\|p-p^{\varepsilon(i)}\| \leq \bar{\varepsilon}\|p-p^{\varepsilon(i-1)}\|, \tag{3.9}$$

$$\|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)}\| \leq \bar{\varepsilon}\frac{K_b}{N_a}\|p-p^{\varepsilon(i-1)}\|. \tag{3.10}$$

PROOF. Subtracting (3.1) and (3.5), one finds

$$\begin{aligned} &a(\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)},\boldsymbol{v}) - b(p-p^{\varepsilon(i)},\boldsymbol{v}) = 0 \qquad \forall\boldsymbol{v}\in V, \\ &\varepsilon(p^{\varepsilon(i-1)}-p^{\varepsilon(i)},q) + b(q,\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)}) = 0 \quad \forall q\in Q. \end{aligned} \tag{3.11}$$

On the other hand, we have that

$$\begin{aligned} 0 &\leq (p-p^{\varepsilon(i)}, p-p^{\varepsilon(i)}) \\ &= (p^{\varepsilon(i-1)}-p^{\varepsilon(i)}, p-p^{\varepsilon(i)}) + (p-p^{\varepsilon(i-1)}, p-p^{\varepsilon(i)}) \end{aligned}$$

and then

$$(p^{\varepsilon(i)} - p^{\varepsilon(i-1)}, p - p^{\varepsilon(i)}) \leq (p - p^{\varepsilon(i-1)}, p - p^{\varepsilon(i)}) . \tag{3.12}$$

This inequality is used several times. Taking $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}$ and $q = p - p^{\varepsilon(i)}$ in (3.11) and using (3.12), we obtain

$$\begin{aligned} K_a \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|^2 &\leq a(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}) \\ &= b(p - p^{\varepsilon(i)}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}) \\ &= \varepsilon(p^{\varepsilon(i)} - p^{\varepsilon(i-1)}, p - p^{\varepsilon(i)}) \\ &\leq \varepsilon(p - p^{\varepsilon(i-1)}, p - p^{\varepsilon(i)}) \\ &\leq \varepsilon \|p - p^{\varepsilon(i-1)}\| \, \|p - p^{\varepsilon(i)}\| . \end{aligned} \tag{3.13}$$

Using the LBB condition, there exists $\boldsymbol{v} \in V - \{0\}$ such that

$$\begin{aligned} K_b \|p - p^{\varepsilon(i)}\| \, \|\boldsymbol{v}\| &\leq b(p - p^{\varepsilon(i)}, \boldsymbol{v}) \\ &= a(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) \\ &\leq N_a \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \, \|\boldsymbol{v}\| \end{aligned}$$

and hence

$$\|p - p^{\varepsilon(i)}\| \leq \frac{N_a}{K_b} \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| . \tag{3.14}$$

Combining (3.13) and (3.14), relations (3.9) and (3.10) are found. Applying these inequalities inductively, we obtain

$$\|p - p^{\varepsilon(i)}\| \leq \bar{\varepsilon}^i \|p - p^{\varepsilon(0)}\| , \qquad \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \leq \bar{\varepsilon}^i \frac{K_b}{N_a} \|p - p^{\varepsilon(0)}\| .$$

The Theorem follows from the fact that $\bar{\varepsilon} < 1$. □

If we take $p^{\varepsilon(0)} = 0$ and apply Lemma 3.2, we see that

$$\|p - p^{\varepsilon(i)}\| \leq C_1 \bar{\varepsilon}^i , \qquad \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \leq C_2 \bar{\varepsilon}^i , \tag{3.15}$$

where the constants C_1 and C_2 are

$$C_1 = \frac{N_l}{K_b}\left(1 + \frac{N_a}{K_a}\right) , \qquad C_2 = \frac{N_l}{N_a}\left(1 + \frac{N_a}{K_a}\right) .$$

The rates of convergence (3.15) are checked numerically in Section 7.

4. A Picard-based iterative algorithm for the Navier–Stokes equations

The Stokes problem is linear and iterating it is the price to be paid if one wants to satisfy (weakly) the constraint $\nabla \cdot \boldsymbol{u} = 0$ up to a certain tolerance with a given penalty parameter ε. However, an iterative algorithm is needed for the Navier–Stokes equations and if the iteration loop could be coupled with the iterative penalization, we would have satisfied the incompressibility constraint at a low computational cost. The purpose of this section is to investigate whether this is possible when the Picard scheme is

used to deal with the nonlinear term. Theorem 4.2 gives sufficient conditions under which the final algorithm is convergent.

The problem to be considered now is (2.8) with c given by (2.3) (or its skew-symmetric form): find $(\boldsymbol{u}, p) \in V \times Q_0$ such that

$$\begin{aligned} &c(\boldsymbol{u}, \boldsymbol{u}, \boldsymbol{v}) + a(\boldsymbol{u}, \boldsymbol{v}) - b(p, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &b(q, \boldsymbol{u}) = 0 \qquad\qquad\qquad\qquad\quad \forall q \in Q . \end{aligned} \tag{4.1}$$

The Picard or successive substitution algorithm for this problem is: given $\boldsymbol{u}^{(0)} \in V$, for $i = 1, 2, \ldots$, find $(\boldsymbol{u}^{(i)}, p^{(i)}) \in V \times Q_0$ such that

$$\begin{aligned} &c(\boldsymbol{u}^{(i-1)}, \boldsymbol{u}^{(i)}, \boldsymbol{v}) + a(\boldsymbol{u}^{(i)}, \boldsymbol{v}) - b(p^{(i)}, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &b(q, \boldsymbol{u}^{(i)}) = 0 \qquad\qquad\qquad\qquad\quad \forall q \in Q . \end{aligned} \tag{4.2}$$

We note that sometimes the name *Picard algorithm* is reserved for the case when all arguments of the nonlinear term are evaluated in the previous iteration. Convergence of algorithm (4.2) for any intial guess $\boldsymbol{u}^{(0)}$ assuming that condition (2.12) holds is a well known result (see, e.g. [19] for the case of RIP methods). The scheme we propose and analyse is the following: given $(\boldsymbol{u}^{\varepsilon(0)}, p^{\varepsilon(0)}) \in V \times Q_0$, for $i = 1, 2, \ldots$, find $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ such that

$$\begin{aligned} &c(\boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) + a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) - b(p^{\varepsilon(i)}, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &\varepsilon(p^{\varepsilon(i)}, q) + b(q, \boldsymbol{u}^{\varepsilon(i)}) = \varepsilon(p^{\varepsilon(i-1)}, q) \qquad \forall q \in Q . \end{aligned} \tag{4.3}$$

Once again, the existence and uniqueness of a solution for (4.3) follows considering the problem in the space $V \times Q$ and applying Lax–Milgram's lemma, since coercitivity of the associated bilinear form is a consequence of (2.9) and (2.11). Before proving convergence in norm of the iterates of (4.3) to the solution of (4.1) we state the following a priori estimates to be used later.

LEMMA 4.1. Let $(\boldsymbol{u}, p)$ *and* $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)})$ *be the solutions of (4.1) and (4.3), respectively. Then*

$$\|\boldsymbol{u}\| \leq \frac{N_l}{K_a} , \tag{4.4}$$

$$\|\boldsymbol{u}^{\varepsilon(i)}\| \leq \frac{N_l}{K_a} + \sqrt{\frac{\varepsilon}{K_a}} \left(\|p - p^{\varepsilon(i-1)}\| + \|p\| \right) . \tag{4.5}$$

PROOF. Estimate (4.4) is obtained exactly as (3.7) noting (2.11). To prove (4.5), take $\boldsymbol{v} = \boldsymbol{u}^{\varepsilon(i)}$ and $q = p^{\varepsilon(i)}$ in (4.3). We obtain

$$\begin{aligned} l(\boldsymbol{u}^{\varepsilon(i)}) &= a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u}^{\varepsilon(i)}) + \varepsilon(p^{\varepsilon(i)} - p^{\varepsilon(i-1)}, p^{\varepsilon(i)}) \\ &= a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u}^{\varepsilon(i)}) + \frac{\varepsilon}{2} \|p^{\varepsilon(i)} - p^{\varepsilon(i-1)}\|^2 + \frac{\varepsilon}{2} \left(\|p^{\varepsilon(i)}\|^2 - \|p^{\varepsilon(i-1)}\|^2 \right) \\ &\geqslant a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u}^{\varepsilon(i)}) - \frac{\varepsilon}{2} \|p^{\varepsilon(i-1)}\|^2 \end{aligned}$$

and hence

$$\begin{aligned} 2K_a \|\boldsymbol{u}^{\varepsilon(i)}\|^2 &\leq 2N_l \|\boldsymbol{u}^{\varepsilon(i)}\| + \varepsilon \|p^{\varepsilon(i-1)}\|^2 \\ &\leq \frac{N_l^2}{K_a} + K_a \|\boldsymbol{u}^{\varepsilon(i)}\|^2 + \varepsilon \|p^{\varepsilon(i-1)}\|^2 \end{aligned}$$

and (4.5) follows easily applying the triangle inequality to $\|p^{\varepsilon(i-1)} - p + p\|$. □

We next establish convergence of the algorithm (4.3).

THEOREM 4.2. Let $(\boldsymbol{u}, p) \in V \times Q_0$ *and* $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ *be the solutions of (4.1) and (4.3), respectively. Assume that*

$$\varepsilon < \frac{K_a K_b^2}{N_c^2}$$

and for any $\alpha \geq 2$, *define the following constants*:

$$M = \left[\frac{N_l}{K_a} + \sqrt{\frac{\varepsilon}{K_a}} \left(\alpha \frac{N_a}{K_b} \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| + \|p - p^{\varepsilon(0)}\| + \|p\| \right) \right] \left[1 - \sqrt{\frac{\varepsilon}{K_a}} \frac{N_c}{K_b} \right]^{-1},$$

$$C_\alpha = \frac{1}{K_b}(N_c M + \alpha N_a), \qquad \beta = \frac{1}{2} + \frac{1}{2}\left(1 + \frac{2}{\alpha}\right)^{1/2},$$

$$\bar{\varepsilon} = \varepsilon \frac{1}{K_a} C_\alpha^2, \qquad \bar{\rho} = \beta(\rho + \bar{\varepsilon}),$$

with ρ *defined in (2.12). Suppose that* $\bar{\rho} < 1$ *and that the initial guess for the velocity satisfies* $\|\boldsymbol{u}^{\varepsilon(0)}\| \leq M$. *Then, the following holds*:

$$\|\boldsymbol{u}^{\varepsilon(i)}\| \leq M, \quad i = 1, 2, \ldots, \tag{4.6}$$

$$\lim_{i \to \infty} \|p - p^{\varepsilon(i)}\| = 0, \qquad \lim_{i \to \infty} \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| = 0. \tag{4.7}$$

Moreover, convergence is linear with $\bar{\rho}$, *that is, there exist constants C and C' such that, for* $i = 1, 2, \ldots,$

$$\|p - p^{\varepsilon(i)}\| \leq C\bar{\rho}^i, \qquad \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \leq C'\bar{\rho}^i. \tag{4.8}$$

PROOF. Only (4.8) has to be proved, since (4.7) follows from the fact that $\bar{\rho} < 1$. We proceed by induction. By hypothesis, (4.6) holds for $i = 0$. Assume that it is true up to $i - 1$, with $i \geq 1$ fixed. Subtracting the equations of (4.3) from (4.1) and using the fact that for a bilinear form g, we have $g(u_1, v_1) - g(u_2, v_2) = g(u_1 - u_2, v_1) + g(u_2, v_1 - v_2)$, we obtain

$$\begin{aligned} &c(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}, \boldsymbol{v}) + c(\boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) + a(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) - b(p - p^{\varepsilon(i)}, \boldsymbol{v}) = 0 \quad \forall \boldsymbol{v} \in V, \\ &\varepsilon(p^{\varepsilon(i-1)} - p^{\varepsilon(i)}, q) + b(q, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}) = 0 \qquad \forall q \in Q. \end{aligned}$$

Taking $\boldsymbol{v} = \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}$ and $q = p - p^{\varepsilon(i)}$, we obtain

$$a(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}) = \varepsilon(p^{\varepsilon(i)} - p^{\varepsilon(i-1)}, p - p^{\varepsilon(i)}) - c(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)})$$

and using the coercitivity of a, (3.12) and Lemma 4.1,

$$\begin{aligned} K_a \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|^2 &\leq N_c \|\boldsymbol{u}\| \, \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}\| \, \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| + \varepsilon \|p - p^{\varepsilon(i-1)}\| \, \|p - p^{\varepsilon(i)}\| \\ &\leq K_a \rho \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}\| \, \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| + \varepsilon \|p - p^{\varepsilon(i-1)}\| \, \|p - p^{\varepsilon(i)}\|. \end{aligned} \tag{4.9}$$

The LBB condition implies that there exists $\boldsymbol{v} \in V - \{0\}$ such that

$$\begin{aligned} K_b \|p - p^{\varepsilon(i)}\| \, \|\boldsymbol{v}\| &\leq b(p - p^{\varepsilon(i)}, \boldsymbol{v}) \\ &= a(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) + c(\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}, \boldsymbol{v}) + c(\boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) \\ &\leq (N_c \|\boldsymbol{u}\| \, \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i-1)}\| + N_c \|\boldsymbol{u}^{\varepsilon(i-1)}\| \, \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| + N_a \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|) \|\boldsymbol{v}\|. \end{aligned}$$

Using again Lemma 4.1 and (4.6) for $i-1$,

$$\|p-p^{\varepsilon(i)}\| \leq \frac{K_a}{K_b}\rho\|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i-1)}\| + \frac{N_c M}{K_b}\|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)}\| + \frac{N_a}{K_b}\|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(i)}\| . \tag{4.10}$$

Inequalities (4.9) and (4.10) can be written as

$$U^{(i)^2} \leq \rho U^{(i-1)}U^{(i)} + \varepsilon\frac{1}{K_a}P^{(i-1)}P^{(i)} , \tag{4.11}$$

$$P^{(i)} \leq \frac{K_a}{K_b}\rho U^{(i-1)} + \left(\frac{N_c M}{K_b}+\frac{N_a}{K_b}\right)U^{(i)} , \tag{4.12}$$

where we have defined

$$U^{(j)} = \|\boldsymbol{u}-\boldsymbol{u}^{\varepsilon(j)}\| , \qquad P^{(j)} = \|p-p^{\varepsilon(j)}\| ,$$

for $j=i$ or $i-1$. Using (4.12) in (4.11), we obtain

$$U^{(i)^2} \leq A_1 U^{(i)} + A_2 , \tag{4.13}$$

where

$$A_1 = \rho U^{(i-1)} + \varepsilon\frac{1}{K_a}P^{(i-1)}\left(\frac{N_c M}{K_b}+\frac{N_a}{K_b}\right) ,$$

$$A_2 = \varepsilon\frac{1}{K_a}P^{(i-1)}\rho\frac{K_a}{K_b}U^{(i-1)} .$$

Since

$$A_1 \leq A_0 = \rho U^{(i-1)} + \varepsilon\frac{1}{K_a}P^{(i-1)}C_\alpha ,$$

we see from (4.13) that

$$U^{(i)^2} \leq A_0 U^{(i)} + A_2 . \tag{4.14}$$

A_0^2 contains the term

$$2\rho U^{(i-1)}\varepsilon\frac{1}{K_a}P^{(i-1)}C_\alpha ,$$

and, since $K_a \leq N_a$ we obtain

$$A_2 \leq \frac{\varepsilon}{K_a}P^{(i-1)}\rho\frac{N_a}{K_b}U^{(i-1)} \leq \frac{\varepsilon}{K_a}P^{(i-1)}\rho\frac{1}{\alpha}C_\alpha U^{(i-1)} .$$

Hence $A_0^2 \geq 2\alpha A_2$ and from (4.14) we obtain

$$\begin{aligned} U^{(i)} &\leq \frac{1}{2}A_0 + \frac{1}{2}(A_0^2+4A_2)^{1/2} \\ &\leq \left[\frac{1}{2}+\frac{1}{2}\left(1+\frac{2}{\alpha}\right)^{1/2}\right]A_0 \\ &= \beta A_0 . \end{aligned}$$

Substitution of this inequality in (4.12) and writing the expression of A_0 leads to

$$
\begin{aligned}
U^{(i)} &\leq \beta\rho U^{(i-1)} + \beta\varepsilon \frac{1}{K_a} C_\alpha P^{(i-1)} \\
&= \beta\rho U^{(i-1)} + \beta\bar{\varepsilon} C_\alpha^{-1} P^{(i-1)} , \\
P^{(i)} &\leq \frac{K_a}{K_b} \rho U^{(i-1)} + \beta\rho C_1 U^{(i-1)} + \beta\varepsilon \frac{1}{K_a} C_1 C_\alpha P^{(i-1)} \\
&\leq \beta\rho C_2 U^{(i-1)} + \beta\varepsilon \frac{1}{K_a} C_\alpha C_\alpha P^{(i-1)} \\
&\leq \beta\rho C_\alpha U^{(i-1)} + \beta\bar{\varepsilon} P^{(i-1)} .
\end{aligned}
$$

From this and assuming that (4.8) holds up to $i-1$, one easily finds that (4.8) is also true for the given i with the constants appearing in (4.8), $C = C_\alpha \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| + \|p - p^{\varepsilon(0)}\|$, $C' = \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| + C_\alpha^{-1} \|p - p^{\varepsilon(0)}\|$. It only remains to show that (4.6) holds for this iteration. Applying Lemma 4.1 and the fact that $P^{(i-1)} \leq C_\alpha U^{(0)} + P^{(0)}$, we obtain

$$
\begin{aligned}
\|\boldsymbol{u}^{\varepsilon(i)}\| &\leq \frac{N_l}{K_a} + \sqrt{\frac{\varepsilon}{K_a}} \left[C_\alpha \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| + \|p - p^{\varepsilon(0)}\| + \|p\| \right] \\
&= \frac{N_l}{K_a} + \sqrt{\frac{\varepsilon}{K_a}} \left[\left(\frac{N_c}{K_b} M + \alpha \frac{N_a}{K_b} \right) \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| + \|p - p^{\varepsilon(0)}\| + \|p\| \right] \\
&= M ,
\end{aligned}
$$

and the induction is complete. □

It is interesting to compare the result stated by this theorem with that obtained for the Picard algorithm (4.2). First, in this case convergence is achieved regardless of the initial guess $\boldsymbol{u}^{(0)}$, whereas for (4.3) we have seen that $\boldsymbol{u}^{\varepsilon(0)}$ has to have a norm bounded by the constant M. For practical purposes, this does not present any problem, since one usually starts taking $\boldsymbol{u}^{\varepsilon(0)} = 0$.

For (4.2) it is known that (4.8) holds with ρ instead of $\bar{\rho}$. However, from the definition of C_α, β and $\bar{\varepsilon}$, we see that if α is taken of order ε^{-q}, with $q < \frac{1}{2}$, then $\bar{\varepsilon} \to 0$ and $\beta \to 1$ when $\varepsilon \to 0$, that is, $\bar{\rho} \to \rho$. So, for ε small, the convergence of algorithm (4.3) is the same as that of (4.2). One can obtain α such that $\bar{\rho}$ is minimized (under the restriction $\alpha \geq 2$). For example, taking the norms and the coercitivity constants of the forms involved in the problem equal to unity, one obtains that the optimal condition for achieving convergence is $\rho < 0.7511$ for $\varepsilon = 10^{-1}$ and $\rho < 0.9744$ for $\varepsilon = 10^{-4}$. The fact that, in any case, $\bar{\rho} \geq \rho$ is the cost of converging to the true incompressible solution using a penalized scheme.

5. A Newton–Raphson-based iterative algorithm for the Navier–Stokes equations

The objective of this section is to analyse the algorithm obtained when the Newton–Raphson scheme is coupled with the iterative penalization in the sense used in the previous section for the Picard scheme. Once again, we will see that the usual convergence requirements of the Newton–Raphson method have to be slightly restricted. In this case, there is another important issue to be considered. It is well known that convergence is quadratic for Newton–Raphson's iterates. The question is whether this rate of convergence will be inherited by the scheme we propose. The answer is that this is true up to a certain iteration. From here onwards, the convergence rate is only linear. However, the numerical experiments we have performed, some of which are presented in Section 7, indicate that the situation is not so bad as it might seem. For small penalty parameters, usually much larger than those used in classical penalty methods, convergence is achieved before its rate turns from quadratic to linear.

The Newton–Raphson algorithm applied to problem (4.1) reads as follows: given $\boldsymbol{u}^{(0)} \in V$, for $i = 1,$

$2, \ldots,$ find $(\boldsymbol{u}^{(i)}, p^{(i)}) \in V \times Q_0$ such that

$$\begin{aligned} &c(\boldsymbol{u}^{(i-1)}, \boldsymbol{u}^{(i)}, \boldsymbol{v}) + c(\boldsymbol{u}^{(i)}, \boldsymbol{u}^{(i-1)}, \boldsymbol{v}) + a(\boldsymbol{u}^{(i)}, \boldsymbol{v}) - b(p^{(i)}, \boldsymbol{v}) \\ &\quad = c(\boldsymbol{u}^{(i-1)}, \boldsymbol{u}^{(i-1)}, \boldsymbol{v}) + l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &b(q, \boldsymbol{u}^{(i)}) = 0 \quad \forall q \in Q . \end{aligned} \tag{5.1}$$

That this algorithm is convergent if the initial guess is sufficiently close to the exact solution and that convergence is quadratic is a well known result. In [12], this is proved when the solution of (4.2) belongs to a nonsingular branch. The modified algorithm we consider is the following: given $(\boldsymbol{u}^{\varepsilon(0)}, p^{\varepsilon(0)}) \in V \times Q_0$, for $i = 1, 2, \ldots,$ find $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ such that

$$\begin{aligned} &c(\boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) + c(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{v}) + a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) - b(p^{\varepsilon(i)}, \boldsymbol{v}) \\ &\quad = c(\boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{u}^{\varepsilon(i-1)}, \boldsymbol{v}) + l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V , \\ &\varepsilon(p^{\varepsilon(i)}, q) + b(q, \boldsymbol{u}^{\varepsilon(i)}) = \varepsilon(p^{\varepsilon(i-1)}, q) \quad \forall q \in Q . \end{aligned} \tag{5.2}$$

Our analysis is based on the assumption that (2.12) holds.

THEOREM 5.1. Let $(\boldsymbol{u}, p) \in V \times Q_0$ *and* $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ *be the solutions of (4.1) and (5.2), respectively. Let* $\alpha \geq 2$ *be given and define the following constants*:

$$C = \frac{2N_c}{K_a(1-\rho)}, \qquad C_\alpha = \frac{K_a}{2K_b}(1+3\rho) + \alpha \frac{N_a}{K_b},$$

$$\beta = \frac{1}{2} + \frac{1}{2}\left(1 + \frac{2}{\alpha}\right)^{1/2}, \qquad \bar{\varepsilon} = \varepsilon \frac{CC_a^2}{N_c}.$$

Assume that the following conditions are satisfied:

$$\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(0)}\| < \frac{1}{C\beta}\frac{\sigma}{\gamma}, \quad \text{with } \sigma < 1, \gamma > 1 , \tag{H1}$$

$$\bar{\varepsilon} < \frac{\gamma - 1}{\beta\gamma}\sigma , \tag{H2}$$

$$\bar{\varepsilon}\|p - p^{\varepsilon(0)}\| < \frac{\gamma - 1}{\beta^2\gamma^2}\frac{C_\alpha}{C}\sigma^2 \quad \text{or} \quad \|p - p^{\varepsilon(0)}\| < \frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma}. \tag{H3}$$

Under these conditions, we have that, for $i = 1, 2, \ldots,$

$$\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| < \frac{1}{C\beta}\frac{\sigma}{\gamma}, \qquad \|p - p^{\varepsilon(i)}\| < \frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma}, \tag{T1}$$

$$K_a - N_c\|\boldsymbol{u}^{\varepsilon(i)}\| > \frac{K_a}{2}(1-\rho) , \tag{T2}$$

$$\lim_{i\to\infty}\|p - p^{\varepsilon(i)}\| = 0 , \qquad \lim_{i\to\infty}\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| = 0 . \tag{T3}$$

Moreover, if for a certain I,

$$\bar{\varepsilon} < \frac{\gamma - 1}{\beta\gamma}\sigma^{2I-1} , \tag{H4}$$

then convergence is quadratic up to this I:

$$\|u-u^{\varepsilon(i)}\|<\frac{1}{C\beta}\frac{\sigma^{2^i}}{\gamma}, \qquad \|p-p^{\varepsilon(i)}\|<\frac{C_\alpha}{C\beta}\frac{\sigma^{2^i}}{\gamma}, \quad 1\leq i\leq I. \tag{T4}$$

PROOF. We proceed by induction. For $i=0$, (T1) is precisely (H1) and (H3) if the second option in this hypothesis is taken. In fact, for $i=1, 2, \ldots$, we will see that any of the two possibilities in (H3) are sufficient for proving (T1). On the other hand, (T2) for $i=0$ follows from (H1) and (4.4):

$$\begin{aligned} K_a-N_c\|u^{\varepsilon(0)}\| &\geq K_a-N_c\|u-u^{\varepsilon(0)}\|-N_c\|u\| \\ &> K_a-\frac{N_c}{\beta}\frac{K_a(1-\rho)}{2N_c}-K_a\rho \\ &>\frac{1}{2}K_a(1-\rho), \end{aligned} \tag{5.3}$$

since $\beta>1$. Now, let i be fixed and assume (T1) and (T2) hold up to $i-1$. In order to prove the existence and uniqueness of the solution for (5.2), let us write this problem as follows: find $(u^{\varepsilon(i)}, p^{\varepsilon(i)})\in V\times Q_0$ such that

$$\mathcal{A}_{i-1}(u^{\varepsilon(i)}, p^{\varepsilon(i)}; v, q)=\mathcal{L}_{i-1}(v, q) \quad \forall (v, q)\in V\times Q,$$

where

$$\mathcal{A}_{i-1}(u, p; v, q)=c(u^{\varepsilon(i-1)}, u, v)+c(u, u^{\varepsilon(i-1)}, v)+a(u, v)-b(p, v)+\varepsilon(p, q)+b(q, u),$$

$$\mathcal{L}_{i-1}(v, q)=l(v)+c(u^{\varepsilon(i-1)}, u^{\varepsilon(i-1)}, v)+\varepsilon(p^{\varepsilon(i-1)}, q).$$

If we prove that $\mathcal{A}_{i-1}$ is coercitive in $V\times Q$, existence and uniqueness will follow from Lax–Milgram's lemma. We see that

$$\begin{aligned} \mathcal{A}_{i-1}(v, q; v, q) &= c(v, u^{\varepsilon(i-1)}, v)+a(v, v)+\varepsilon(q, q) \\ &\geq (K_a-N_c\|u^{\varepsilon(i-1)}\|)\|v\|^2+\varepsilon\|q\|^2 \\ &\geq \min\{K_a-N_c\|u^{\varepsilon(i-1)}\|, \varepsilon\}(\|v\|^2+\|q\|^2) \end{aligned}$$

and the fact that $K_a-N_c\|u^{\varepsilon(i-1)}\|>0$ is a consequence of (T2) used inductively.

Applying the same arguments as in Theorem 4.2 to arrive at (4.9) and (4.10), we now obtain

$$\begin{aligned} K_a\|u-u^{\varepsilon(i)}\|^2 &\leq N_c\|u-u^{\varepsilon(i-1)}\|^2\|u-u^{\varepsilon(i)}\|, \\ &\quad +N_c\|u-u^{\varepsilon(i)}\|^2\|u^{\varepsilon(i-1)}\|+\varepsilon\|p-p^{\varepsilon(i-1)}\|\,\|p-p^{\varepsilon(i)}\|, \end{aligned} \tag{5.4}$$

$$\begin{aligned} K_b\|p-p^{\varepsilon(i)}\| &\leq N_c\|u\|\,\|u-u^{\varepsilon(i)}\|+N_c\|u-u^{\varepsilon(i-1)}\|^2 \\ &\quad +N_c\|u^{\varepsilon(i-1)}\|\,\|u-u^{\varepsilon(i)}\|+N_a\|u-u^{\varepsilon(i)}\|. \end{aligned} \tag{5.5}$$

If we call

$$U^{(j)}=\|u-u^{\varepsilon(j)}\|, \qquad P^{(j)}=\|p-p^{\varepsilon(j)}\|, \quad j=i \text{ or } i-1,$$

$$A_1=K_a-N_c\|u^{\varepsilon(i-1)}\|,$$

$$A_2=N_cU^{(i-1)^2}+\varepsilon P^{(i-1)}\left(\frac{N_c}{K_b}\|u\|+\frac{N_c}{K_b}\|u^{\varepsilon(i-1)}\|+\frac{N_a}{K_b}\right),$$

$$A_3=\varepsilon P^{(i-1)}\frac{N_c}{K_b}U^{(i-1)^2},$$

we find from the previous inequalities that

$$A_1 U^{(i)^2} \leq A_2 U^{(i)} + A_3 . \tag{5.6}$$

From (T2), $A_1 > 0$. Note now that, from (T1) and (4.4),

$$\begin{aligned} \frac{N_c}{K_b}\|\boldsymbol{u}\| + \frac{N_c}{K_b}\|\boldsymbol{u}^{\varepsilon(i-1)}\| + \frac{N_a}{K_b} &< \frac{N_c}{K_b}\|\boldsymbol{u}\| + \frac{N_c}{K_b}\frac{1}{C} + \frac{N_c}{K_b}\|\boldsymbol{u}\| + \frac{N_a}{K_b} \\ &= \frac{2K_a}{K_b}\rho + \frac{N_a}{K_b} + \frac{K_a(1-\rho)}{2K_b} \\ &< C_\alpha , \end{aligned}$$

and so we have

$$A_2 < A_0 = N_c U^{(i-1)^2} + \varepsilon P^{(i-1)} C_\alpha .$$

A_0^2 contains the term

$$2N_c U^{(i-1)^2} \varepsilon P^{(i-1)} \alpha \frac{N_a}{K_b} .$$

Since $K_a \leq N_a$, we have that

$$2A_1 A_3 < 2K_a \varepsilon P^{(i-1)} \frac{N_c}{K_b} U^{(i-1)^2} < \frac{1}{\alpha} A_0^2 ,$$

and hence, from (5.6) and using that $A_2 < A_0$ and (T2):

$$\begin{aligned} U^{(i)} &< \frac{1}{2}\frac{A_0}{A_1} + \frac{1}{2A_1}(A_0^2 + 4A_1 A_3)^{1/2} \\ &< \frac{1}{2}\frac{A_0}{A_1} + \frac{1}{2A_1}\left(1 + \frac{2}{\alpha}\right)^{1/2} A_0 \\ &= \beta \frac{A_0}{A_1} \\ &< \frac{2\beta}{K_a(1-\rho)}(N_c U^{(i-1)^2} + \varepsilon P^{(i-1)} C_\alpha) \\ &= C\beta U^{(i-1)^2} + \varepsilon\beta \frac{C}{N_c} C_\alpha P^{(i-1)} . \end{aligned}$$

On the other hand, for the pressures we obtain from (5.5),

$$\begin{aligned} P^{(i)} &\leq \frac{N_c}{K_b} U^{(i-1)^2} + \left(\frac{N_c}{K_b}\|\boldsymbol{u}^{\varepsilon(i-1)}\| + \frac{N_c}{K_b}\|\boldsymbol{u}\| + \frac{N_a}{K_b}\right) U^{(i)} \\ &< \frac{N_c}{K_b} U^{(i-1)^2} + C_1 U^{(i)} \\ &< \frac{N_c}{K_b} U^{(i-1)^2} + C_1 C\beta U^{(i-1)^2} + \varepsilon\beta C_\alpha \frac{C_1}{N_c} P^{(i-1)} . \end{aligned}$$

Noting that

$$\frac{N_a}{K_b}\beta C = \beta\frac{N_a}{K_b}\frac{2N_c}{K_a(1-\rho)} > \frac{N_c}{K_b},$$

we finally obtain that (5.4) and (5.5) imply

$$U^{(i)} < C\beta U^{(i-1)^2} + \bar{\varepsilon}\beta C_\alpha^{-1} P^{(i-1)}, \tag{5.7}$$

$$P^{(i)} < CC_\alpha\beta U^{(i-1)^2} + \bar{\varepsilon}\beta P^{(i-1)}. \tag{5.8}$$

Now, using (H2) and (T1) for $i-1$, we obtain

$$\begin{aligned} U^{(i)} &< C\beta\frac{1}{C^2\beta^2}\frac{\sigma^2}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma\frac{C_\alpha^{-1}C_\alpha}{C\beta}\frac{\sigma}{\gamma} \\ &= \frac{1}{C\beta}\frac{\sigma^2}{\gamma}, \\ P^{(i)} &< CC_\alpha\beta\frac{1}{C^2\beta^2}\frac{\sigma^2}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma\frac{C_\alpha}{C\beta}\frac{\sigma}{\gamma} \\ &= \frac{C_\alpha}{C\beta}\frac{\sigma^2}{\gamma}. \end{aligned}$$

Since $\sigma < 1$, the induction for (T1) is closed. Observe that either of the two possibilities in (H3) suffices for proving this part of the thesis. (T2) is obtained from (T1) using the same steps as in (5.3). In order to prove (T3), from (5.7) and (5.8), we see that the required condition is

$$\tau^{(i)} = C\beta U^{(i)} + \bar{\varepsilon}\beta < 1.$$

From (T1) and (H2), it follows that $\tau^{(i)} < \sigma < 1$, so convergence of the algorithm is ensured. Inequalities (5.7) and (5.8) show that the rate of convergence will be at least linear.

Now, suppose that (H4) holds. For $1 \leq i \leq I$, we obtain from (5.7) and (5.8) and assuming (T4) to be true up to $i-1$, that

$$\begin{aligned} U^{(i)} &< C\beta\frac{1}{C^2\beta^2}\frac{\sigma^{2^i}}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma^{2^{i-1}}\frac{C_\alpha^{-1}C_\alpha}{C\beta}\frac{\sigma^{2^{i-1}}}{\gamma} \\ &= \frac{1}{C\beta}\frac{\sigma^{2^i}}{\gamma}, \\ P^{(i)} &< CC_\alpha\beta\frac{1}{C^2\beta^2}\frac{\sigma^{2^i}}{\gamma^2} + \frac{\gamma-1}{\gamma}\sigma^{2^{i-1}}\frac{C_\alpha}{C\beta}\frac{\sigma^{2^{i-1}}}{\gamma} \\ &= \frac{C_\alpha}{C\beta}\frac{\sigma^{2^i}}{\gamma}. \end{aligned}$$

This proves (T4) and completes the proof of the theorem. □

This theorem states that if the initial guess is close enough to the final solution and ε is sufficiently small, algorithm (5.2) will converge. The situation is similar for the standard Newton–Raphson scheme (5.1). The only difference in the requirement for the initial velocity guess is that (H1) has to hold but with $\beta = 1$. Nevertheless, the same remarks as in Theorem 4.2 apply and in our case α can be taken such that $\beta \to 1$ when $\varepsilon \to 0$. Another observation is that in (H3) we can choose either having a 'good' initial pressure guess or limiting the value of ε.

6. Decoupling of the iterative penalization

In Sections 4 and 5, we have seen that if the iterative penalization is coupled with the iterative scheme used to deal with the convective term of the Navier–Stokes equations, the conditions under which this scheme converges have to be restricted. There is also the possibility of decoupling the iteration due to the nonlinearity of the equations and the imposition on the incompressibility constraint. The purpose of this section is the analysis of the following algorithm: given $p^{\varepsilon(0)} \in Q_0$, for $i = 1, 2, \ldots$, find $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ such that

$$
\begin{aligned}
&c(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) + a(\boldsymbol{u}^{\varepsilon(i)}, \boldsymbol{v}) - b(p^{\varepsilon(i)}, \boldsymbol{v}) = l(\boldsymbol{v}) \quad \forall \boldsymbol{v} \in V\,, \\
&\varepsilon(p^{\varepsilon(i)}, q) + b(q, \boldsymbol{u}^{\varepsilon(i)}) = \varepsilon(p^{\varepsilon(i-1)}, q) \qquad \forall q \in Q\,.
\end{aligned} \tag{6.1}
$$

For a given i, the existence and uniqueness of solution is a consequence of assumption (2.12) [12]. For each iteration of this algorithm a nonlinear problem has to be solved. We assume that the solution found in this process is exact. In this case, we obtain a result similar to the one encountered for the Stokes problem in Section 3.

THEOREM 6.1. Let $(\boldsymbol{u}, p) \in V \times Q_0$ *and* $(\boldsymbol{u}^{\varepsilon(i)}, p^{\varepsilon(i)}) \in V \times Q_0$ *be the solutions of (4.1) and (6.1), respectively. Define the following constants*:

$$
M = \frac{N_l}{K_a} + \sqrt{\frac{\varepsilon}{K_a}}\left(\|p - p^{\varepsilon(0)}\| + \|p\|\right), \qquad C_1 = K_a(1-\rho)\,,
$$

$$
C_2 = \frac{K_a}{K_b}\rho + \frac{N_c}{K_b}M + \frac{N_a}{K_b}, \qquad \bar{\varepsilon} = \frac{\varepsilon}{C_1}C_2^2\,.
$$

If $\bar{\varepsilon} < 1$, *then*

$$
\|\boldsymbol{u}^{\varepsilon(i)}\| \leq M\,, \quad i = 1, 2, \ldots, \tag{6.2}
$$

$$
\lim_{i\to\infty} \|p - p^{\varepsilon(i)}\| = 0\,, \qquad \lim_{i\to\infty} \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| = 0\,. \tag{6.3}
$$

Moreover, convergence is linear with $\bar{\varepsilon}$:

$$
\|p - p^{\varepsilon(i)}\| \leq \bar{\varepsilon}^i \|p - p^{\varepsilon(0)}\|\,, \qquad \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \leq C_2^{-1}\bar{\varepsilon}^i \|p - p^{\varepsilon(0)}\|\,. \tag{6.4}
$$

PROOF. First observe that (4.5) holds for the solution $\|\boldsymbol{u}^{\varepsilon(i)}\|$ of (6.1). This can be proved exactly as in Lemma 4.1. Thus, (6.2) is verified for $i = 1$. Let $i > 1$ be given and assume this is true up to this iteration. If the ideas used to arrive at (4.9) and (4.10) in Theorem 4.2 are now applied, one finds

$$
K_a\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|^2 \leq N_c\|\boldsymbol{u}\|\,\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|^2 + \varepsilon\|p - p^{\varepsilon(i-1)}\|\,\|p - p^{\varepsilon(i)}\|\,, \tag{6.5}
$$

$$
K_b\|p - p^{\varepsilon(i)}\| \leq N_c\|\boldsymbol{u}\|\,\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| + N_c\|\boldsymbol{u}^{\varepsilon(i)}\|\,\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| + N_a\|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\|\,. \tag{6.6}
$$

Combining inequalities (6.5) and (6.6), using the estimates (4.4) for $\boldsymbol{u}$ and (6.2) for $\boldsymbol{u}^{\varepsilon(i)}$ and considering the definition of the constants in the statement of the Theorem, we arrive at

$$
\|p - p^{\varepsilon(i)}\| \leq \bar{\varepsilon}\|p - p^{\varepsilon(i-1)}\|\,, \qquad \|\boldsymbol{u} - \boldsymbol{u}^{\varepsilon(i)}\| \leq \bar{\varepsilon}C_2^{-1}\|p - p^{\varepsilon(i-1)}\|\,,
$$

from which (6.4) follows. Since $\bar{\varepsilon} < 1$, we have that $\|p - p^{\varepsilon(i)}\| < \|p - p^{\varepsilon(0)}\|$ and we obtain from (4.5) that (6.2) holds for $i + 1$. This closes the induction. Finally, (6.4) implies (6.3) for $\bar{\varepsilon} < 1$. □

7. Numerical results

The finite element implementation of the algorithms studied in this work is not considered here (see e.g. [14]). We remark that the only additional cost of the methods presented compared with standard penalty methods is that the pressure has to be computed and stored in each iteration.

The first issue to be considered is the finite element spaces for the velocity and the pressure. We have programmed several stable elements with discontinuous pressures for both 2-D and 3-D problems, including the bilinear velocity–constant pressure pair. Although this element does not satisfy the LBB condition, its use is widespread since it is known to work well if pressure spurious modes are removed. Moreover, it can be stabilized either by using iterative procedures [20] or by discretizing the domain in macroelements composed of this element [21]. The two-dimensional results presented in this section have been obtained using the biquadratic-linear element (continuous biquadratic interpolation for the velocities and piecewise linear discontinuous pressures), known to yield very good results for incompressible flow problems [22, 23]. We have also tested the quadratic simplicial element enriched with a bubble function for velocities and piecewise linear pressure with similar results.

The pressure computed once convergence has been achieved is discontinuous. In order to obtain a continuous pressure field, a smoothing technique has been used. Let p_s be the smoothed pressure, interpolated as the velocity components. If p_c is the computed pressure, the nodal values of p_s are obtained by minimizing $\|p_c - p_s\|_{L^2}$. A numerical quadrature rule with the integration points placed on the element nodes is then used to evaluate the components of the Gramm matrix of the resulting algebraic system. This results in a diagonal matrix the inversion of which is trivial. See [24] for details.

In the examples below, convergence has been checked only in velocities, using the norm of the residual over the norm of the last iterate as the parameter to decide whether this convergence has been achieved or not. Since we are mainly interested in the satisfaction of the incompressibility constraint, the norm of the discrete divergence of the velocity has also been computed.

All the calculations have been carried out on a CONVEX-C120 computer using double arithmetic precision.

7.1. The driven cavity flow

In this example, the Stokes problem with $\nu = 1$ in the unit square $[0, 1] \times [0, 1]$ has been solved. The boundary conditions have been taken as $\boldsymbol{u} = (1, 0)$ for $y = 1$, $0 \leq x \leq 1$ (x, y being the Cartesian coordinates) and $\boldsymbol{u} = (0, 0)$ on the rest of the boundary. External body forces have been taken as zero. The domain has been discretized using a uniform mesh of 21×21 nodal points (10×10 elements). Figure 1 shows the convergence history for the values of the penalty parameter $\varepsilon = 10^{-1}$, 10^{-2}, 10^{-3} and 10^{-4}. Observe that the difference in the slope of the curves agrees with the theoretical prediction (3.15).

Once the finite element discretization has been performed, the term $\nabla \cdot \boldsymbol{u}$ leads to $\boldsymbol{Hu}$, where $\boldsymbol{H}$ is the discrete divergence matrix. In order to study the convergence of the iterates to the incompressible solution, the norm of $\boldsymbol{Hu}^{\varepsilon(i)}$ has been computed. Figure 2 shows the results obtained for different values of the penalty parameter. The curves correspond to 1, 2 and 3 iterations in the algorithm (3.5). Once again, their relative slope agrees with what (3.15) predicts. Observe that $\|\boldsymbol{Hu}^{\varepsilon(i)}\|$ will be bounded by $\varepsilon G \|p^{\varepsilon(i)} - p^{\varepsilon(i-1)}\|$, where G is the norm of the Gramm matrix whose components are the scalar products of the basis functions for the pressure (see (3.5)).

The pressure contours and the velocity vectors solution of this problem are plotted in Figs. 3 and 4, respectively.

7.2. Flow over a forward step

The purpose of this example is to present some numerical results concerning the algorithms studied in Sections 4 and 5 for the incompressible Navier–Stokes equations. We have chosen this well known benchmark problem because a large number of numerical results are available. The computational domain we have taken is the rectangle $[0, 22] \times [0, 1.5]$ with a step of length 3 and height 0.5 placed in

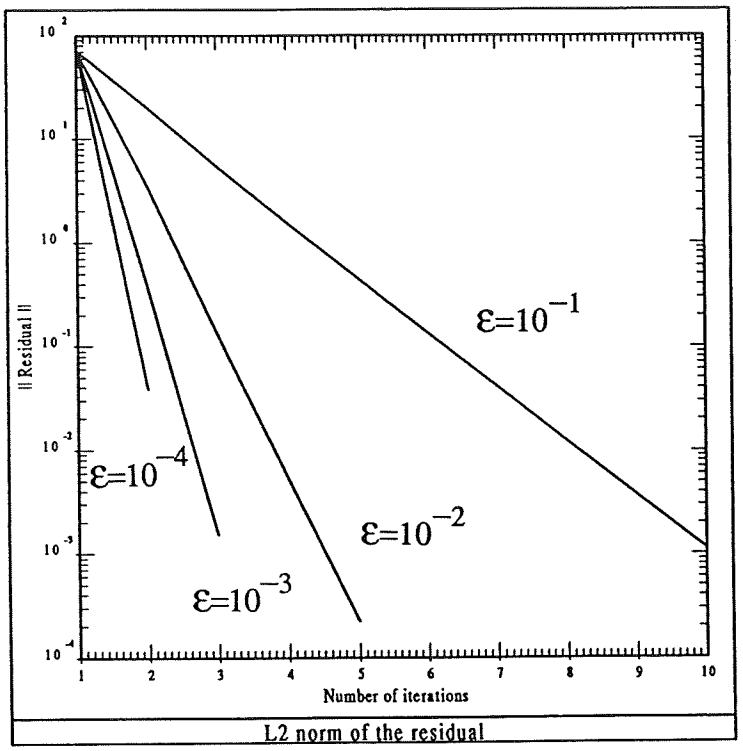

Fig. 1. Convergence history of the cavity flow example using different penalty parameters.

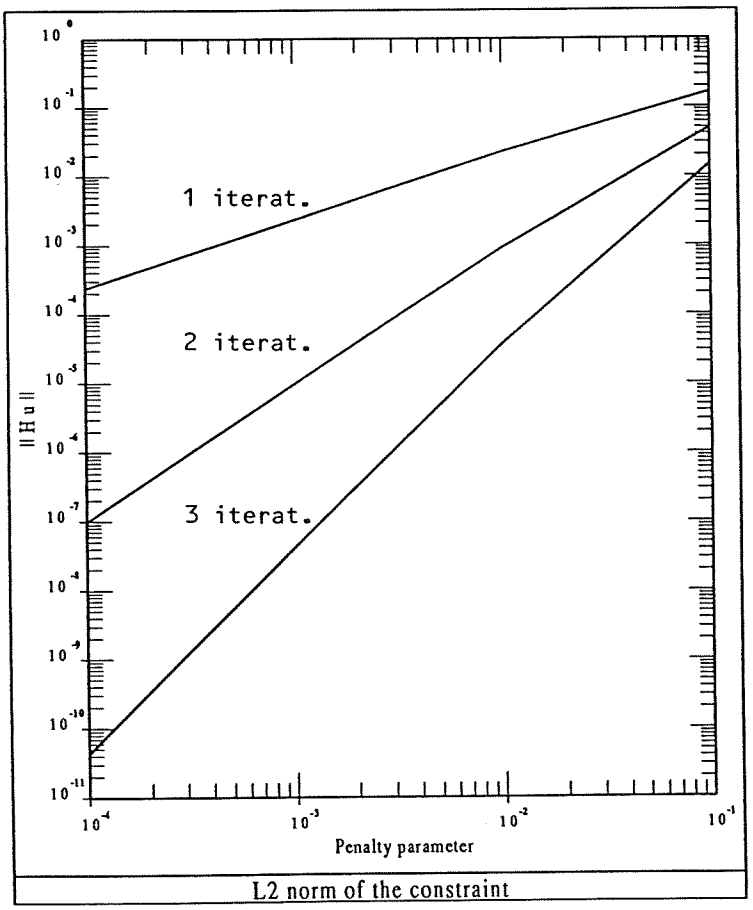

Fig. 2. Norm of the discrete divergence for the cavity flow example for different number of iterations

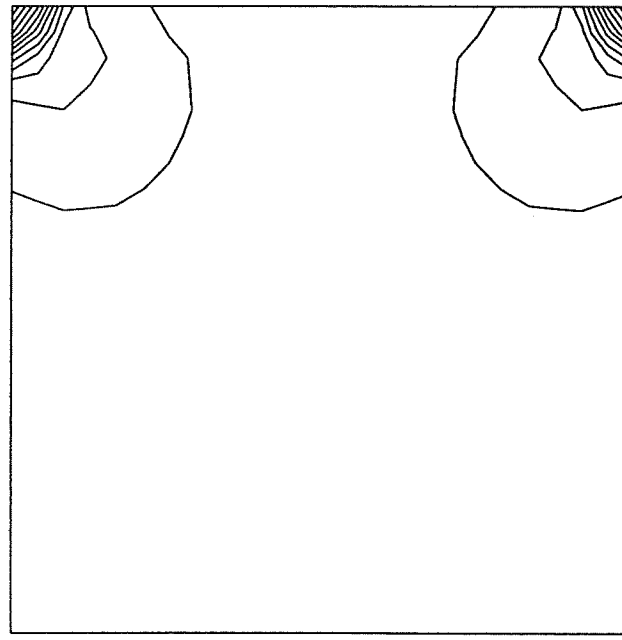

Fig. 3. Pressure contours for the cavity flow problem.

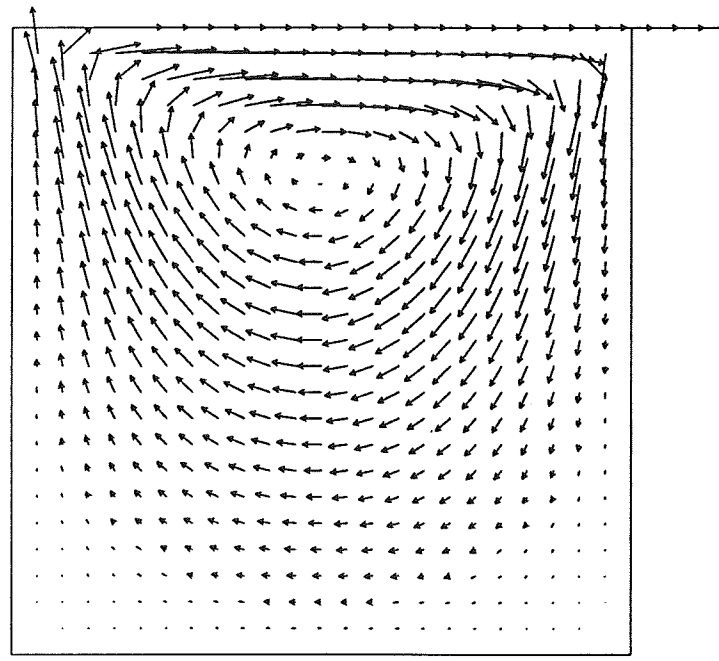

Fig. 4. Velocity vectors for the cavity flow problem.

the lower left corner. A detail of the mesh used in the calculation is shown in Fig. 5. This mesh is composed of 408 biquadratic elements (for the velocity interpolation) and 1721 nodal points.

On the left boundary $x = 0$, a parabolic velocity profile with maximum value $(1, 0)$ has been prescribed. The viscosity has been taken as $\nu = 0.005$. Thus, the Reynolds number based on the inflow profile and the step height is $\mathrm{Re} = 100$. The outflow boundary $x = 22$, $0 < y < 1.5$ has been left free. We have employed the expression (2.4) for the viscous term. In this case, the associated natural boundary condition is zero traction. On the rest of the boundary, the no-slip condition $\boldsymbol{u} = 0$ has been imposed. External body forces are zero.

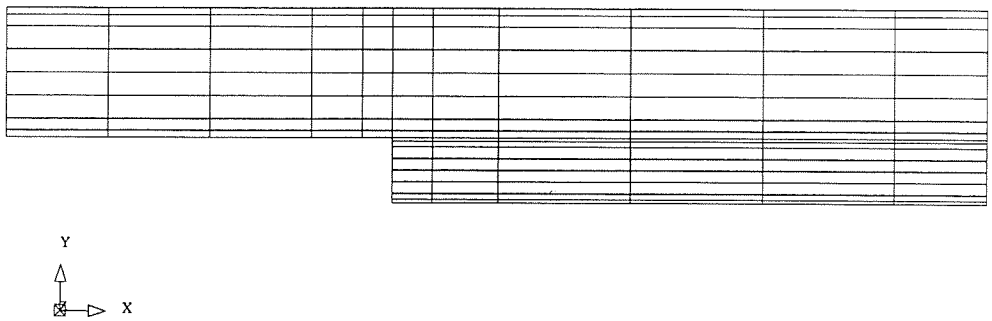

Fig. 5. Detail of the mesh for the forward step problem.

The computed pressure contours and a detail of the streamlines are plotted in Figs. 6 and 7. These results have been obtained using a penalty parameter $\varepsilon = 10^{-4}$ and with a tolerance of $10^{-4}\%$. The iterative scheme employed has been (5.2). Now we discuss the performance of the algorithms (4.3) and (5.2) for this problem when the classical penalty method and the iterative penalization proposed in this paper are used. In the former case, the right-hand side in the second equation of both (4.3) and (5.2) is zero.

Consider first algorithm (4.3). Figure 8 shows the convergence history in the discrete L^2 norm when both the classical and the iterative penalty methods are used. No difference can be observed in the plot even though the parameter that defines convergence in the former case is ρ whereas in the latter it is $\bar{\rho} > \rho$ (see (4.8)). The values of the relative norm of the residual in iteration number 18 are 0.87215×10^{-2} for the classical penalty method and 0.87108×10^{-2} for the iterative penalization. However, the important issue is the evolution of $\|\boldsymbol{H}\boldsymbol{u}^{\varepsilon(i)}\|$ shown in Fig. 9. For the penalty method, this norm remains constant (and, as expected, of order ε). On the other hand, the velocity solution of algorithm (4.3) converges linearly to a (weakly) solenoidal field. Similar results are obtained when the penalty parameter is $\varepsilon = 10^{-1}$. Figures 10 and 11 show the convergence history and the evolution of $\|\boldsymbol{H}\boldsymbol{u}^{\varepsilon(i)}\|$. It is interesting to observe that for this large penalty number, the residual norm using (4.3) is only slightly larger than using the classical penalty method.

The same experiments discussed above have been performed using the Newton–Raphson-based

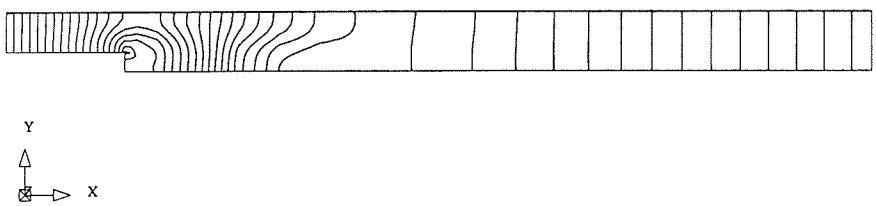

Fig. 6. Pressure contours for the forward step problem.

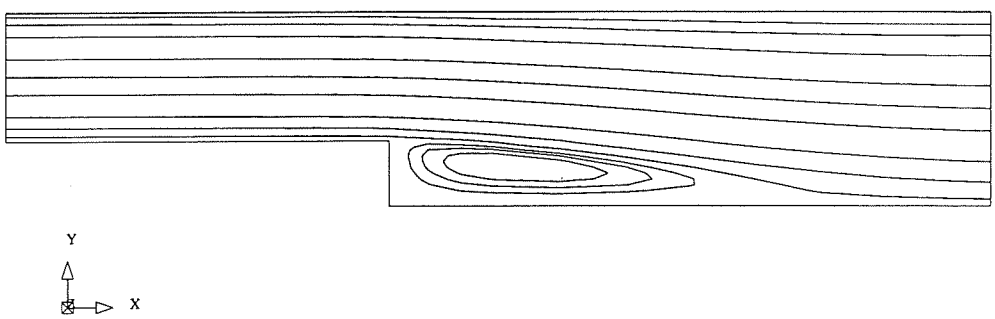

Fig. 7. Detail of the streamlines for the forward step problem.

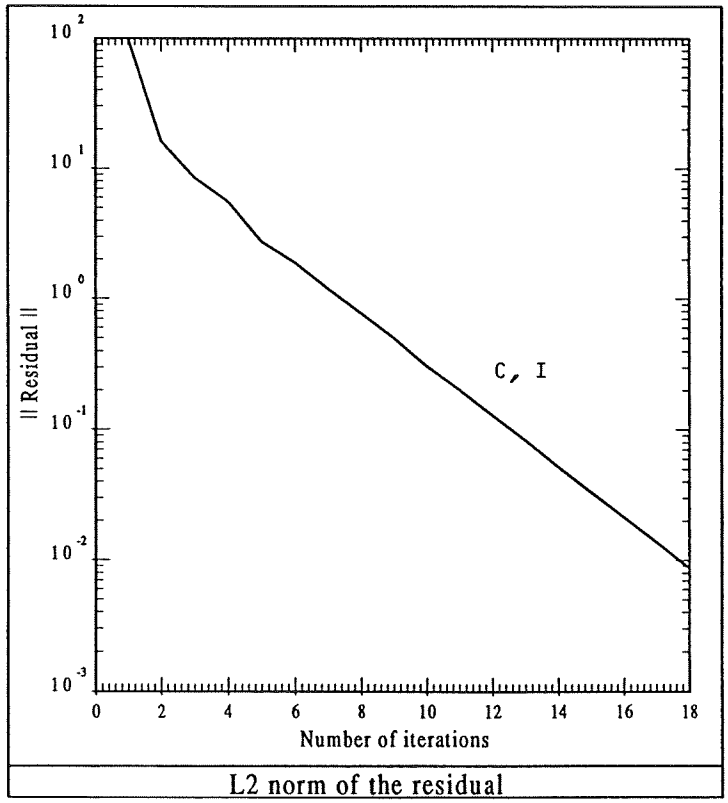

Fig. 8. Convergence history for the Picard-based algorithm with $\varepsilon = 10^{-4}$. Classical (C) and iterative (I) penalization.

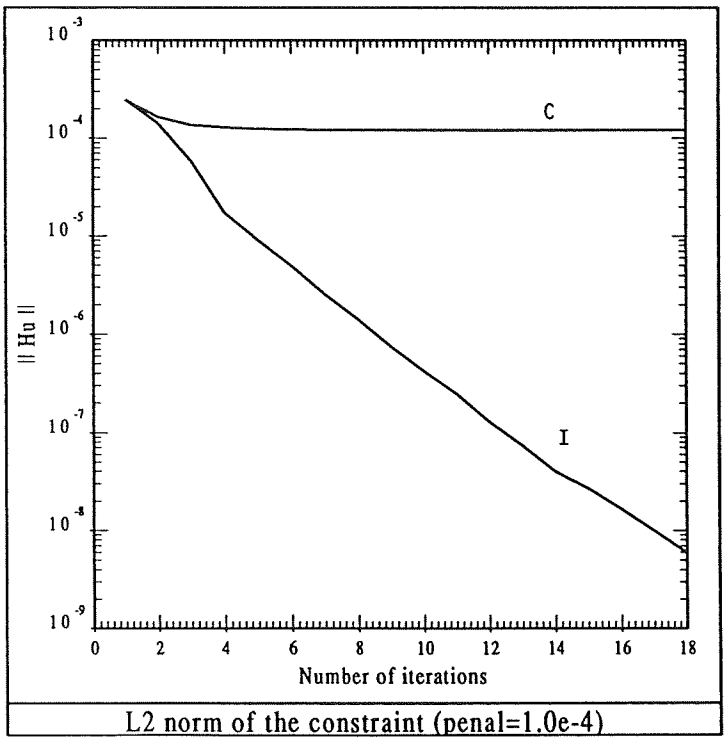

Fig. 9. Evolution of the norm of the discrete divergence for the Picard-based algorithm with $\varepsilon = 10^{-4}$. Classical (C) and iterative (I) penalization.

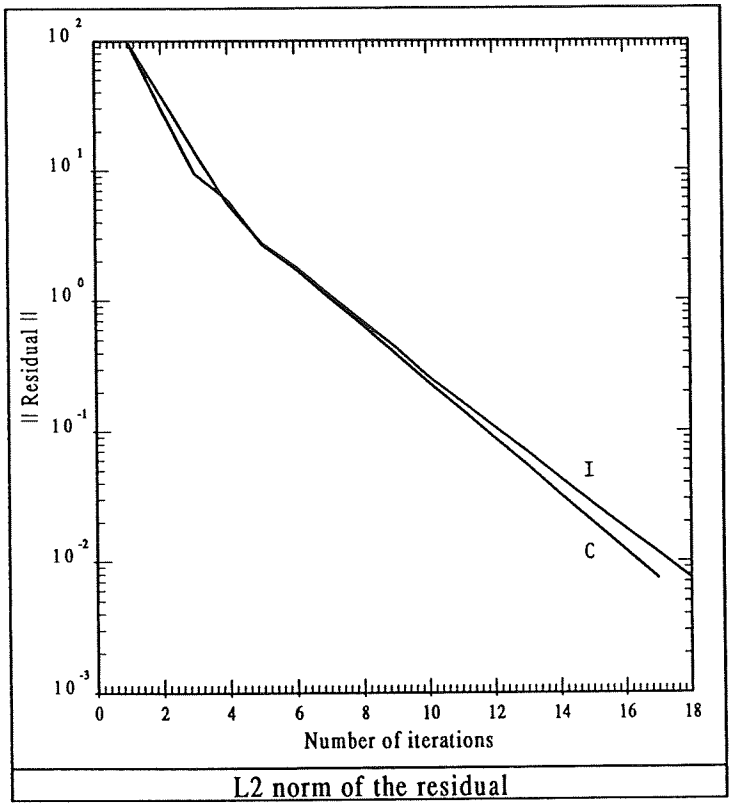

Fig. 10. Convergence history for the Picard-based algorithm with $\varepsilon = 10^{-1}$. Classical (C) and iterative (I) penalization.

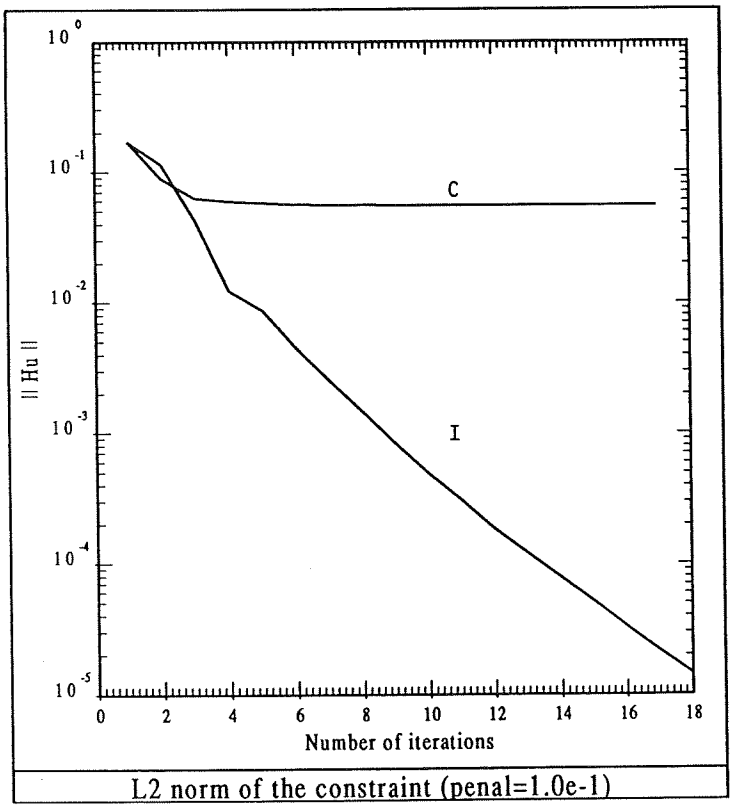

Fig. 11. Evolution of the norm of the discrete divergence for the Picard-based algorithm with $\varepsilon = 10^{-1}$. Classical (C) and iterative (I) penalization.

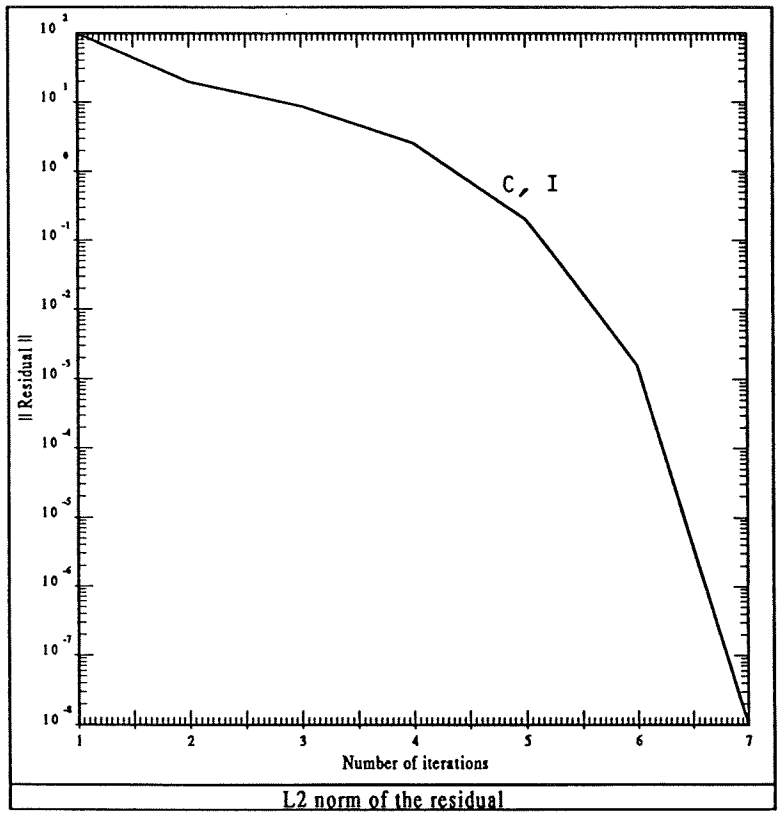

Fig. 12. Convergence history for the Newton–Raphson-based algorithm with $\varepsilon = 10^{-4}$. Classical (C) and iterative (I) penalization.

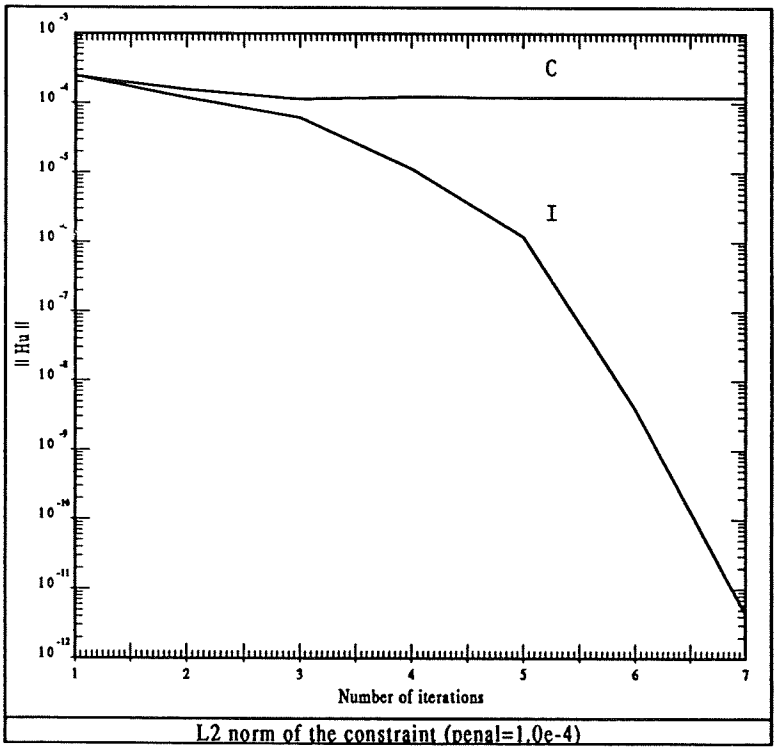

Fig. 13. Evolution of the norm of the discrete divergence for the Newton–Raphson-based algorithm with $\varepsilon = 10^{-4}$. Classical (C) and iterative (I) penalization.

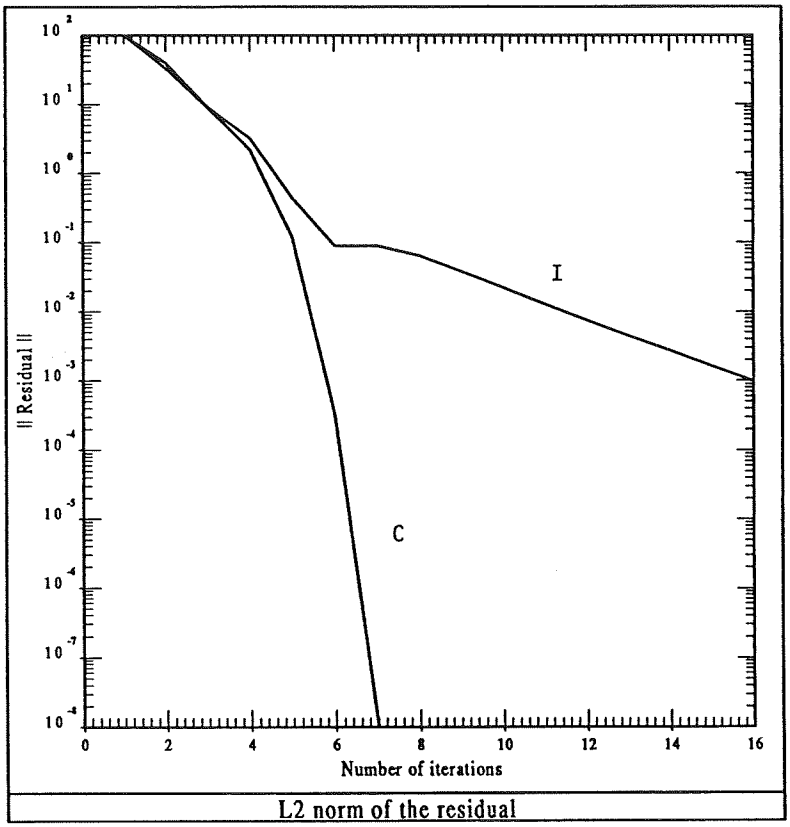

Fig. 14. Convergence history for the Newton–Raphson-based algorithm with $\varepsilon = 10^{-1}$. Classical (C) and iterative (I) penalization.

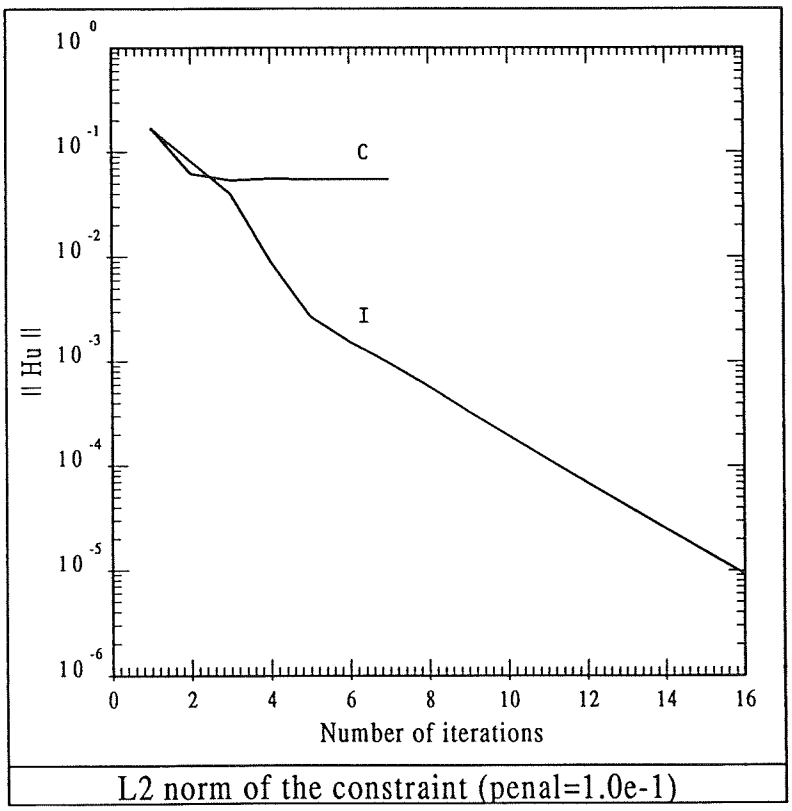

Fig. 15. Evolution of the norm of the discrete divergence for the Newton–Raphson-based algorithm with $\varepsilon = 10^{-1}$. Classical (C) and iterative (I) penalization.

algorithm (5.2). If the initial guess is taken as $\boldsymbol{u}^{\varepsilon(0)} = 0$, $p^{\varepsilon(0)} = 0$ the scheme does not converge. In order to obtain a good initial guess, at least two iterations of algorithm (4.3) have to be performed, both for the classical and the iterative penalty methods. For $\varepsilon = 10^{-4}$, the convergence history of the two methods shown in Fig. 12 is the same. The relative norm of the residual reaches the value 0.9986×10^{-8} at iteration number 7 for the classical penalty method and 0.9781×10^{-8} if (5.2) is used. The evolution of the norm of the discrete divergence (Fig. 13) is certainly very different for the two methods. Whereas the penalty method yields a constant value, the iterative penalization converges quadratically to a zero divergence velocity field. Of special interest is the case $\varepsilon = 10^{-1}$. The convergence rate for the iterates of (5.2) (see Fig. 14) is quadratic up to iteration number 6 (except for the two first iterations, in which scheme (4.3) has been used). From there on, this rate turns out to be linear. This possibility was already predicted in Theorem 5.1. The classical penalty method has a global quadratic convergence rate, but $\|\boldsymbol{H}\boldsymbol{u}^{\varepsilon(i)}\|$ remains constant in the iterative process (Fig. 15) at an unacceptable value.

8. Discussion and conclusions

We believe that the method proposed and analysed in this paper has very interesting features. The penalty method for the incompressible Navier–Stokes equations in its classical form is attractive. It reduces the number of nodal unknowns and yields good results. This is a very important attribute if three-dimensional problems have to be solved on medium-size computers. However, small penalty parameters lead to ill-conditioned stiffness matrices. Usually, this ill-conditioning is not a trouble if direct solvers are used. But, still thinking in the numerical simulation of 3-D flows, iterative solvers are almost imperative when a real problem has to be faced. These solvers are very sensitive to the condition number of the stiffness matrix and this seriously limits the feasibility of the classical penalty method. The iterative penalization presented here tries to circumvent, at least in part, this inconvenience. It allows the use of much larger penalty parameters, thus yielding matrices whose condition numbers are much smaller. Whether this will be enough for using iterative solvers or not is something that experience has to provide. We have performed some tests for the Stokes problem using the conjugate gradient method with encouraging results.

The iterative penalization presented here may be obtained from different approaches conceptually different. For the Stokes problem, it reduces to the Augmented Lagrangian method combined with the Uzawa algorithm to uncouple the pressure. It can also be interpreted as the introduction of an artificial compressibility and a false transient only for the pressure whenever the temporal derivative is discretized using the backward Euler scheme. However, we prefer the residual argument described in Section 3 since it is still valid when the iterative equation for the pressure is coupled with a linearized form of the momentum equations in the Navier–Stokes problem. When the Picard method is used to obtain this linearization, the analysis of the algorithm (Theorem 4.2) reveals that the rate of convergence is smaller than for the classical penalty method. For the Newton–Raphson scheme, the attraction ball of the exact solution happens to be smaller and quadratic convergence can only be ensured up to a certain iteration (Theorem 5.1). However, numerical experiments indicate that these effects are only apparent when the penalty parameter is 'very large', compared with the standards of the classical approach.

In practice, it is common to use penalties of order $10^{-6}\nu^{-1}$ to $10^{-9}\nu^{-1}$. We have already said that those values can be easily handled using direct solvers. However, there are some practical cases in which the viscosity varies several orders of magnitude in the fluid domain, as in quasi-Newtonian fluids with thermal dependent physical properties. In these cases, the above rule has to be applied using the smallest value of the viscosity, thus relaxing in excess the incompressibility constraint in the high viscosity zones. We have also applied the iterative penalty method in these cases where the nonlinearity comes from the constitutive law with very good results.

References

[1] I. Babuška, Errors bounds for finite element method, Numer. Math. 16 (1971) 322–333.
[2] F. Brezzi, On the existence, uniqueness and approximation of saddle point problems arising from Lagrange multipliers, RAIRO, Ser. Rouge Anal. Numer. 8 (1976) R-2.

[3] F. Brezzi and K.J. Bathe, A discourse on the stability conditions for mixed finite element formulations, Comput. Methods Appl. Mech. Engrg. 82 (1990) 27–57.
[4] T.J.R. Hughes and L.P. Franca, A new finite element formulation for computational fluid dynamics: VII. The Stokes problem with various well-posed boundary conditions: Symmetric formulations that converge for all velocity/pressure spaces, Comput. Methods Appl. Mech. Engrg. 65 (1987) 85–96.
[5] T.J.R. Hughes, L.P. Franca and M. Balestra, A new finite element formulation for computational fluid dynamics: V. Circumventing the Babuška–Brezzi condition: A stable Petrov–Galerkin formulation of the Stokes problem accommodating equal-order interpolations, Comput. Methods Appl. Mech. Engrg. 59 (1986) 85–99.
[6] L. Franca and R. Stenberg, Error analysis of some Galerkin least-squares methods for the elasticity equations, INRIA, Rapports de Recherche, 1989.
[7] P. Hansbo and A. Szepessy, A velocity–pressure streamline diffusion finite element method for the incompressible Navier–Stokes equations, Comput. Methods Appl. Mech. Engrg. 84 (1990) 175–192.
[8] G.F. Carey and R. Krishnan, Penalty finite element methods for the Navier–Stokes equations, Comput. Methods Appl. Mech. Engrg. 42 (1984) 183–223.
[9] M.S. Engelman, R.L. Sani, P.M. Gresho and M. Bercovier, Consistent vs reduced integration penalty methods for incompressible media using several old and new elements, Internat. J. Numer. Methods Fluids 2 (1983) 25–42.
[10] T.J.R. Hughes, W.K. Liu and A. Brooks, Finite element analysis of incompressible viscous flows by the penalty function formulation, J. Comput. Phys. 30 (1979) 1–60.
[11] J.T. Oden, N. Kikuchi and Y.J. Song, Penalty-finite element methods for the analysis of Stokesian flows, Comput. Methods Appl. Mech. Engrg. 31 (1982) 297–329.
[12] V. Girault and P.A. Raviart, Finite Element Methods for Navier–Stokes Equations (Springer, Berlin, 1986).
[13] R. Temam, Navier–Stokes Equations (North-Holland, Amsterdam, 1984).
[14] C. Cuvelier, A Segal and A. van Steenhoven, Finite Element Methods and Navier–Stokes Equations (Reidel, Dordrecht, 1986).
[15] C. Johnson and J. Pitkaranta, Analysis of some mixed finite element methods related to reduced integration, Math. Comp. 38 (1982) 375–400.
[16] O.C. Zienkiewicz, R.L. Taylor and J.M. Too, Reduced integration technique in general analysis of plates and shells, Internat. J. Numer. Methods Engrg. 3 (1971) 275–290.
[17] D.S. Malkus and T.J.R. Hughes, Mixed finite element methods – reduced and selective integration techniques: a unification of concepts, Comput. Methods Appl. Mech. Engrg. 15 (1978) 63–81.
[18] R. Glowinski, Numerical Methods for Nonlinear Variational Problems (Springer, Berlin, 1984).
[19] G.F. Carey and R. Krishnan, Convergence of iterative methods in penalty finite element approximations of the Navier–Stokes equations, Comput. Methods Appl. Mech. Engrg. 60 (1987) 1–29.
[20] M. Fortin and S. Boivin, Iterative stabilization of the bilinear velocity-constant pressure element, Internat. J. Numer. Methods Fluids 10 (1990) 125–140.
[21] P. Le Tallec and V. Ruas, On the convergence of the bilinear-velocity constant-pressure finite element method in viscous flow, Comput. Methods Appl. Mech. Engrg. 54 (1986) 235–243.
[22] M. Fortin, Old and new finite elements for incompressible flows, Internat. J. Numer. Methods Fluids 3 (1981) 347–364.
[23] J.C. Nagtegaal, D.M. Parks and J.R. Rice, On numerically accurate finite element solutions in the fully plastic range, Comput. Methods Appl. Mech. Engrg. 4 (1974) 153–177.
[24] T.J.R. Hughes, The Finite Element Method. Linear Static and Dynamic Analysis (Prentice Hall, Englewood, Cliffs, NJ, 1987).