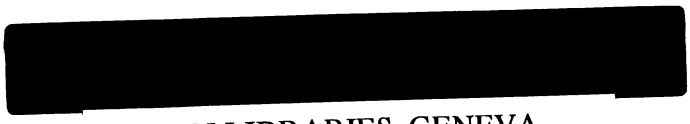
EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

CERN-Data Handling Division
DD/77/17
O. Axelsson
October 1977



PRECONDITIONING OF INDEFINITE PROBLEMS BY REGULARIZATION

(Submitted to SIAM Journal of Numerical Analysis)

EUROPEAN ORGANIZATION FOR NUCLEAR RESEARCH

PRECONDITIONING OF INDEFINITE PROBLEMS BY REGULARIZATION

O. Axelsson*)

DD Division, CERN, Geneva, Switzerland

ABSTRACT

It is shown how mixed formulations of boundary-value problems with constraints, such as the Stokes problem and the equilibrium formulation in elasticity, may be solved iteratively by preconditioning with a regularized operator. Advantages of using this approach in comparison with the frequently-used augmented Lagrangian method are pointed out.

Geneva - 28 October 1977

*) On leave from Chalmers University of Technology, Göteborg, Sweden.

1. INTRODUCTION

Many problems in physics and engineering, like the Stokes problem in fluid flow or the equilibrium formulation of the elasticity equations, lead to boundary-value problems with constraints.

From a practical point of view, it is often difficult to satisfy the constraints directly; instead a mixed formulation is used. Other problems, like the biharmonic plate problem, may be formulated, with great advantage, from a computational point of view, as a coupled problem in different variables, which again leads to a mixed formulation (see, for example [8]).

The mixed formulation approach will be successful as long as there exists a saddle point to the associated Lagrangian function. The existence of a saddle point has thus to be proved. Another difficulty with the mixed formulation is that the resulting equations are indefinite. The solution of such equations are most easily done by some appropriate iterative algorithm (which would be needed anyway if the problem is non-linear). In this paper we discuss a method to deal with such problems which is based on regularization, equivalent to penalization of the energy functional, if this exists.

This approach is taken in order to prove the existence of a saddle point and to provide a preconditioning matrix for the numerical solution. The approach to prove existence by regularization (and then go to the limit) is well known, see, for example, [11], [15].

Bercovier [4] (see also [15]) has used this method to approximate the solution of the original problem by the solution of the penalized one. His main contribution is to prove that by use of a proper hypothesis, the bounds of the error due to regularization and to discretization are independent of each other. The present paper is based on that idea, but a simplified and slightly more general proof of the error estimate is given here. More important, however, we present a method which actually solves the original problem (naturally only a discretized version of it). This is accomplished by the solution of a finite sequence of positive definite linear problems, the solution of which is much simpler than that of the original problem. This latter may be non-linear, geometrically and/or materially, but for ease of presentation, we have limited the study to linear problems.

A comparison is made with Uzawa's algorithm, a frequently used method for constrained optimization (see, in particular, [7]). Since our approach is not based on an energy functional, it is more generally applicable and, furthermore, for the outer iteration sequence, we may easily adopt a fast convergent and parameter-free method, such as the conjugate gradient method, instead of the often quite slowly convergent gradient method with a fixed parameter, commonly used in Uzawa's algorithm.

Another preconditioning method for indefinite problems, based on a particular incomplete factorization of the given matrix, is presented in [2]. The main advantage with preconditioning as opposed to a direct solver is to save storage, since there is no, or minor, fill-in. This is even more important in mixed formulations, where owing to the additional variables introduced, the bandwidth may be quite large. The additional complication of having to deal with an indefinite matrix may also necessitate pivoting in a direct solver.

In three-dimensional and/or in time-dependent problems an iterative solver is usually also faster, even for moderate-sized problems with many right-hand sides present (see, for example, [1], [9]).

2. MIXED FORMULATIONS

Let V be a Hilbert space with dual V^* and let $\langle \cdot,\cdot \rangle_V$ denote the duality pairing on $V \times V^*$. We let H be another Hilbert space with inner product $(\cdot,\cdot)$ and let its dual H^* be identified with H. The norm in H is denoted by $\|\cdot\|$. Further, let $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ be two continuous bilinear forms on $V \times V$ and $V \times H^*$, respectively. Denote by $A \in L(V,V^*)$, $B \in L(V,H)$, and the adjoint $B^* \in L(H^*,V^*)$, the linear operators associated with the bilinear forms, i.e.

$$\langle Au,v \rangle_V = a(u,v) \qquad \forall\ u,\ v \in V$$

$$(Bu,p) = \langle Bu,p \rangle_H = \langle u,B^*p \rangle_V = b(u,p) \qquad \forall\ u \in V,\ p \in H^* .$$

We suppose that

(i) $a(v,v) + r_0 \|Bv\|^2 \geq \alpha \|v\|_V^2, \ \alpha > 0 \qquad v \in V$

for some $r_0 \geq 0$. In particular, this means that A is positive definite over $\mathcal{N}(B)$, the null space of B. Further assume that the Brezzi-Babuška condition (see [5])

(ii) $\sup_{u \in V - \{0\}} \frac{b(u,p)}{\|u\|_V} > \gamma \|p\|, \quad \gamma > 0 \quad \forall p \in \overset{\circ}{H}{}^* = H^*/\mathcal{N}(B^*)$

holds. Given $g_1 \in V^*$ and $g_2 \in \overset{\circ}{H}$, where

$$\overset{\circ}{H} = \{g \in H;\ (g,p) = 0 \quad \forall p \in \mathcal{N}(B^*)\} ,$$

we will prove that there exists a saddle point $(\hat{u},\hat{p})$, unique in $V \times \overset{\circ}{H}{}^*$, to the Lagrangian

$$\ell(u,p) = \frac{1}{2}\, a(u,u) - \langle u,g_1\rangle_V + b(u,p) - (g_2,p) . \tag{2.1}$$

The corresponding stationary relations are

$$\begin{aligned} a(\hat{u},v) + b(v,\hat{p}) &= \langle g_1,v\rangle \quad \forall v \in V \\ b(\hat{u},p) &= (g_2,p) \quad \forall p \in H^* . \end{aligned} \tag{2.2}$$

It is now an easy matter to prove that there is at most one solution in $V \times \overset{\circ}{H}{}^*$ to (2.2). To prove the existence of such a saddle point, we will first study a corresponding regularized problem, and then go to the limit. Thus, in order to get a Lagrangian which is coercive in both u and p, we add a (small) term to $\ell(u,p)$, making the perturbed functional concave in p. Thus let

$$\ell_\varepsilon(u,p) = \ell(u,p) - \frac{1}{2}\,\varepsilon\, \|p\|^2 , \qquad \varepsilon > 0 .$$

We observe, that by choosing $p = \sqrt{r_0}\, Bu$, we find using (i) that

$$\ell(u,p) = \frac{1}{2}\, a(u,u) + r_0 \|Bu\|^2 - \langle u,g_1\rangle - (g_2,Bu) \to +\infty$$

as $\|u\|_V \to \infty$. By the addition of the penalty term we also obtain

$$\ell_\varepsilon(u,p) \to -\infty$$

when $\|p\| \to \infty$. Thus by a classical theorem [12], there exists a saddle point to $\ell_\varepsilon(u,p)$, unique in $V \times \overset{\circ}{H}{}^*$ by strong convexity (concavity).

Remark 2.1. We observe that if a(u,v) is symmetric, then

$$f_\varepsilon(u) = \sup_{p \in H^*} \ell_\varepsilon(u,p) = \frac{1}{2}\, a(u,u) - \langle u,g_1\rangle + \frac{1}{2\varepsilon}\, \|Bu - g_2\|^2 ,$$

since to $\sup_p \ell_\varepsilon(u,p)$ we have the extremality relation

$$\varepsilon p = Bu - g_2 .$$

The corresponding primal problem is thus

$$(\mathcal{P}'_\varepsilon) \qquad \inf_{u \in V} f_\varepsilon(u) = \inf_{u \in V} \sup_{p \in H^*} \ell_\varepsilon(u,p) \ .$$

Owing to (i), f_ε is continuous, increasing, proper, and strongly convex, which means that there exists a unique minimizer of f_ε.

We will now prove an *a priori* bound of the saddle point of $\ell_\varepsilon(u,p)$.

Theorem 2.1. Let $\hat{u}_\varepsilon$, $\hat{p}_\varepsilon$ be the saddle point of $\ell_\varepsilon(u,p)$, $u,p \in V \times H^*$. Then for $0 \le \varepsilon \le \varepsilon_0$, $\varepsilon_0 r_0 \le \min\left[1, (\alpha\gamma/3M)\right]$, we have

$$\|\hat{u}_\varepsilon\|_V + \|\hat{p}_\varepsilon\| \le C(\|g_1\|_V + \|g_2\|) \ .$$

Here M is the boundedness constant of a, i.e.

$$|a(u,v)| \le M\|u\|_V \, \|v\|_V \quad \forall\, u,v \in V \ .$$

Proof. The extremality relations corresponding to $\ell_\varepsilon(u,p)$ are (neglecting the circumflexes),

$$(\mathcal{P}_\varepsilon) \qquad \begin{aligned} a(u_\varepsilon,u) + b(v,p_\varepsilon) &= \langle v,g_1\rangle \quad \forall\, v \in V \\ b(u_\varepsilon,p) - \varepsilon(p_\varepsilon,p) &= (g_2,p) \quad \forall\, p \in H^* \ . \end{aligned}$$

These relations may formally be written

$$\begin{aligned} Au_\varepsilon + B^* p_\varepsilon &= g_1 \\ Bu_\varepsilon - \varepsilon p_\varepsilon &= g_2 \ . \end{aligned} \tag{2.3}$$

We let $u_\varepsilon = u_1 + u_2$, $p_\varepsilon = p_1 + p_2$, where

$$a(u_1,v) + b(v,p_1) = \langle v,g_1\rangle \quad \forall\, v \in V \tag{2.4a}$$

$$b(u_1,p) - \varepsilon(p_1,p) = 0 \quad \forall\, p \in H^* \tag{2.4b}$$

and

$$a(u_2,v) + b(v,p_2) = 0 \quad \forall\, v \in V \tag{2.5a}$$

$$b(u_2,p) - \varepsilon(p_2,p) = (g_2,p) \quad \forall\, p \in H^* \tag{2.5b}$$

Consider the first equation (2.4). We get from (2.4b)

$$\varepsilon p_1 = Bu_1 \ .$$

Hence with $u = u_1$, $p = p_1$, and $\varepsilon \le 1/r_0$, we get from (2.4a) and from (i):

$$\alpha\|u_1\|_V^2 \le a(u_1,u_1) + \frac{1}{\varepsilon}\|Bu_1\|^2 = \langle u_1,g_1\rangle \le \frac{\alpha}{2}\|u_1\|_V^2 + \frac{1}{2\alpha}\|g_1\|_{V^*}^2 \ ,$$

i.e.

$$\|u_1\|_V \le \frac{1}{\alpha} \|g_1\|_{V^*} ,$$

that is $u_1 \in V$, all ε, $0 < \varepsilon \le 1/r_0$. Hence $Au_1 \in V^*$ and from the extremality relation

$$B^* p_1 = g_1 - Au_1$$

we get

$$\|B^* p_1\|_{V^*} \le \|g_1\|_{V^*} + \|Au_1\|_{V^*} \le C_0 \|g_1\|_{V^*} , \qquad C_0 = 1 + M/\alpha .$$

Hence, from (ii),

$$\|p_1\| \le \frac{1}{\gamma} C_0 \|g_1\|_{V^*} .$$

Consider now the second equation (2.5). From (2.5a) we get

$$a(u_2,v) = -b(v,p_2) ,$$

i.e. since $a(\cdot,\cdot)$ is continuous,

$$|b(v,p_2)| \le M\|u_2\|_V \|v\|_V .$$

Hence from (ii),

$$\|p_2\| \le \frac{1}{\gamma} \sup_v \frac{b(v,p_2)}{\|v\|_V} < \frac{M}{\gamma} \|u_2\|_V . \tag{2.6}$$

Now let $v = u_2, p = p_2$. Then

$$a(u_2,u_2) + \varepsilon\|p_2\|^2 = -(g_2,p_2)$$

and using (2.6) and the Cauchy-Schwarz inequality:

$$a(u_2,u_2) \le C\|g_2\|^2 + \frac{\alpha}{3} \|u_2\|_V^2 .$$

Moreover, from the extremality relation $Bu_2 - \varepsilon p_2 = g_2$ we get

$$\|Bu_2\| \le \|g_2\| + \varepsilon\|p_2\| \le \|g_2\| + \varepsilon \frac{M}{\gamma} \|u_2\|_V .$$

Summing up, we have

$$\alpha\|u_2\|_V^2 \le a(u_2,u_2) + r_0\|Bu_2\|^2 \le \left(\frac{\alpha}{3} + \varepsilon r_0 \frac{M}{\gamma}\right) \|u_2\|_V^2 + C\|g_2\|^2 .$$

Thus, with ε_0 small enough so that, for example,

$$\varepsilon_0 r_0 \frac{M}{\gamma} \le \frac{\alpha}{3} ,$$

we get

$$\|u_2\|_V \le C\|g_2\| \ .$$

Finally, from (2.6) and the triangle inequality

$$\|u_\varepsilon\| \le \|u_1\| + \|u_2\| \ , \quad \|p_\varepsilon\| \le \|p_1\| + \|p_2\| \ ,$$

the theorem is proved.

Now this theorem implies that there exists a subsequence $\{u_\varepsilon\}$ which converges to a (weak) limit $\hat{u}$,

$$\hat{u}_\varepsilon \rightharpoonup \hat{u} \in V \ , \quad \varepsilon \to 0$$

and hence from the extremality relations (2.3) that

$$B^*\hat{p}_\varepsilon = g_1 - Au_\varepsilon \rightharpoonup g_1 - A\hat{u} \in V^* \ .$$

Likewise

$$\|\hat{p}_\varepsilon\| \le C \ , \qquad 0 < \varepsilon \le \varepsilon_0 \ ,$$

implies that $\exists\, \hat{p}$ ∋

$$\hat{p}_\varepsilon \rightharpoonup \hat{p} \in H^* \ , \quad \varepsilon \to 0 \ ,$$

which in turn satisfies the extremality relation

$$B^*\hat{p} = g_1 - A\hat{u} \ .$$

This means that $\ell(u,p)$ has a saddle point $(\hat{u},\hat{p}) \in V \times \mathring{H}^*$.

Remark 2.2. We observe that the solution p_ε of (2.3) does not belong to $\mathcal{N}(B^*)$, since $g_2 \in \mathring{H} = \mathcal{N}(B^*)$.

Remark 2.3. If $a(u,u)$ is V-elliptic, i.e if (i) is satisfied with $r_0 = 0$, then Theorem 2.1 is valid for all $\varepsilon > 0$.

Remark 2.4. In some applications (like almost incompressible elastic materials) the regularized problem $(\mathcal{P}_\varepsilon)$ also has a slightly changed bilinear form $a_\varepsilon = \rho(\varepsilon)a$, where $0 < \rho(\varepsilon) = 1 + O(\varepsilon)$. Thus, now let

$$\begin{aligned} a_\varepsilon(u_\varepsilon,v) + b(v,p_\varepsilon) &= \langle v,g_1\rangle \quad \forall\, v \in V \\ b(u_\varepsilon,p) - \varepsilon(p_\varepsilon,p) &= (g_2,p) \quad \forall\, p \in H^* \ . \end{aligned}$$

It is easily seen that Theorem 2.1 is valid also in this case, if

$$\varepsilon r_0/\rho(\varepsilon) \le \min\left(1, \frac{\alpha}{3}\frac{\gamma}{M}\right) \ , \qquad 0 < \varepsilon \le \varepsilon_0 \ .$$

It is now a straightforward matter to derive an error estimate for

$$e_\varepsilon = \hat{u} - u_\varepsilon \ , \qquad \delta_\varepsilon = \hat{p} - \hat{p}_\varepsilon \ .$$

We have

$$\begin{aligned} a(\hat{u},v) + b(v,\hat{p}) &= \langle v,g_1\rangle \quad \forall\, v \in V \\ b(\hat{u},p) &= (g_2,p) \quad \forall\, p \in H^* \end{aligned}$$

and

$$\begin{aligned} a_\varepsilon(\hat{u}_\varepsilon,v) + b(v,\hat{p}_\varepsilon) &= \langle v,g_1\rangle \quad \forall\, v \in V \\ b(\hat{u}_\varepsilon,p) - \varepsilon(\hat{p}_\varepsilon,p) &= (g_2,p) \quad \forall\, p \in H^* \ , \end{aligned}$$

where $a_\varepsilon = \rho(\varepsilon)a$ (see Remark 2.4). By subtraction we get

$$\begin{aligned} a_\varepsilon(e_\varepsilon,v) + b(v,\delta_\varepsilon) &= a_\varepsilon(\hat{u},v) - a(\hat{u},v) \\ b(e_\varepsilon,p) - \varepsilon(\delta_\varepsilon,p) &= -\varepsilon(\hat{p},p) \ , \end{aligned}$$

where $a_\varepsilon(u,v) - a(u,v) = \left[\rho(\varepsilon) - 1\right] a(\hat{u},v)$ and $\rho(\varepsilon) - 1 = O(\varepsilon)$. Now from the previous theorem and Remark 2.4, we get at once

$$\|e_\varepsilon\|_V + \|\delta_\varepsilon\| \le O(\varepsilon)(\|A\hat{u}\|_{V^*} + \|\hat{p}\|) = O(\varepsilon)(\|\hat{u}\|_V + \|\hat{p}\|) \ .$$

This proves that the error in both $\hat{u}_\varepsilon$ and $\hat{p}_\varepsilon$, goes to zero not slower than linearly with ε.

Remark 2.5. (Extrapolation). By some additional hypothesis on regularity of the given data, one may prove that there exists an expansion

$$\hat{u} = \hat{u}_\varepsilon + \varepsilon u^{(1)} + \varepsilon^2 u^{(2)} + \cdots$$

where (formally), in the case $\rho(\varepsilon) = 1$,

$$\begin{aligned} Au^{(m)} + B^* p^{(m)} &= 0 \\ Bu^{(m)} &= p^{(m-1)} \ , \qquad p^{(m)} \in \overset{\circ}{H}{}^* \ , \qquad m = 1, 2, \cdots, \ p^{(0)} = \hat{p} \ . \end{aligned}$$

This means, of course, that we may calculate u_ε for a few different values of ε and then extrapolate to the limit $\varepsilon \to 0$ in order to get a more accurate approximation. In Section 4 we will, however, present a technique where the same goal is reached without these additional assumptions on regularity.

3. THE AUGMENTED LAGRANGIAN METHOD

Let V_h and S_h be finite-dimensional subspaces of V and H, respectively, such that the conditions (i) and (ii) are still satisfied. For different choices of such subspaces, see, for example, [6], [14], and [10].

Let $\{\phi_i\}$ and $\{\psi_i\}$ be (finite element) basis functions in V_h and S_h, respectively, and define the matrices

$$\underset{\sim}{A} = [a(\phi_j,\phi_i)] \ , \ \underset{\sim}{B} = [b(\phi_j,\psi_i)] \ , \ \underset{\sim}{G} = (\psi_i,\psi_j) \ .$$

The indefinite finite dimensional problem, corresponding to (2.2) has then the form

$$\begin{aligned} \underset{\sim}{A}\underset{\sim}{u} + \underset{\sim}{B}^T\underset{\sim}{p} &= \underset{\sim}{g}_1 \\ \underset{\sim}{B}\underset{\sim}{u} &= \underset{\sim}{g}_2 \ , \qquad (u,p) \in V_h \times S_h \ , \end{aligned} \tag{3.1}$$

where $\underset{\sim}{g}_1 = [\langle g_1,\phi_i\rangle]$, $\underset{\sim}{g}_2 = [(g_2,\psi_i)]$ and $\underset{\sim}{u}$, $\underset{\sim}{p}$, etc., indicate the nodal point vectors corresponding to the functions u, p, etc. To solve (3.1) we will at first recall some properties of the augmented Lagrangian method. (For a general discussion of this method in connection with boundary value problems, see [7].) We suppose that the bilinear form a(u,v) is symmetric, so that $\underset{\sim}{A}$ is symmetric. The augmented Lagrangian now takes the form

$$f_r(u,p) = \frac{1}{2}\, a(u,u) - \langle g_1,u\rangle + b(u,p) - (g_2,p) + \frac{r}{2}\, \|Bu - g_2\|^2 \ ,$$

where $r \geq 0$ is a (penalty) parameter. The solution $\hat{u}$, $\hat{p}$ of (3.1) is obviously a saddle point of f_r. This implies that the functional

$$\phi(p) = f_r(\hat{u}(p),p) = \min_{u \in V_h} f_r(u,p)$$

has a maximum at $\hat{p}$. Since $\nabla_u f_r(u,p)\big|_{u=\hat{u}(p)} = 0$, we have

$$\nabla\phi(p) = \left[\nabla_u f_r(u,p)\ \nabla_p \hat{u}(p) + \nabla_p f_r(u,p)\right]\Big|_{u=\hat{u}(p)} = \nabla_p f_r(u,p)\Big|_{u=\hat{u}(p)} = \underset{\sim}{B}\underset{\sim}{u} - \underset{\sim}{g}_2 \ .$$

Further, its Jacobian is

$$\kappa_\phi = \nabla^2\psi(p) = \nabla_u(\nabla\phi(p))\ \nabla_p\hat{u}(p) = -\underset{\sim}{B}(\underset{\sim}{A} + r\underset{\sim}{B}^T\underset{\sim}{B})^{-1}\ \underset{\sim}{B}^T \ . \tag{3.2}$$

For r large, κ_ϕ is close to a projection operator onto the column space $\mathcal{R}(B)$. Thus if $\underset{\sim}{B}$ has maximal rank, we have

$$\kappa_\phi \simeq -\frac{1}{r}\, I \ , \qquad r \gg 1 \ , \tag{3.3}$$

where I is the identity matrix.

For the calculation of the maximizer of ϕ, a gradient method may be used. At each step of this we have to calculate an approximation of $\hat{u}(p)$. Thus the algorithm, known as Uzawa's algorithm, takes the form

$$\begin{aligned} &\underset{\sim}{p} := \underset{\sim}{p}^0 \ ; \\ R : &\underset{\sim}{u} := \text{minimizer of } f_r(u,p) \ ; \\ &\underset{\sim}{c} := \underset{\sim}{B}\underset{\sim}{u} - \underset{\sim}{g}_2 \ ; \\ &\underset{\sim}{p} := \underset{\sim}{p} + \tau \underset{\sim}{c} \ ; \\ &\text{IF } \|\underset{\sim}{c}\|^2 > \varepsilon \text{ GOTO R} \ ; \end{aligned}$$

For the calculation of the minimizer of f_r, a system of equations with matrix $\underset{\sim}{A} + r\underset{\sim}{B}^T\underset{\sim}{B}$ has to be solved.

The parameter τ has to be small enough for convergence. If $r >> 1$, the algorithm will converge fast owing to (3.3), and then $\tau = r$ is a good choice. We observe that the use of so-called second-order information [κ_ϕ in (3.2)] is likely to be too time-consuming. Also function evaluations of $\phi(p)$ are costly in a non-linear problem, since for every p, $\hat{u}(p)$ has to be calculated by an optimization algorithm.

4. PRECONDITIONING BY REGULARIZATION

We now present an algorithm which is based on preconditioning of the given matrix

$$\underset{\sim}{\mathcal{A}} = \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B}^T \\ \underset{\sim}{B} & 0 \end{bmatrix}$$

with the help of the perturbed problem (2.3). We will notice that the resulting algorithm, besides being applicable to non-symmetric forms a(u,v), also more easily lends itself to the use of fast-convergent methods, like the (parameter-free) conjugate-gradient algorithm, than was the case in the augmented Lagrangian method.

At each step (ℓ) in the new algorithm, we solve for $\rho^{(\ell)}$, $\gamma^{(\ell)}$, where

$$\begin{aligned} a(\rho^{(\ell)},v) + b(v,\gamma^{(\ell)}) &= a(u^{(\ell)},v) + b(v,p^{(\ell)}) - \langle g_1,v\rangle \quad \forall\, v \in V_h \\ b(\rho^{(\ell)},q) - \varepsilon(q,\gamma^{(\ell)}) &= b(u^{(\ell)},q) \qquad\qquad - (q,g_2) \quad \forall\, q \in S_h \end{aligned}$$

with u^0, p^0 being arbitrary. With

$$\mathcal{A}_\varepsilon = \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B}^T \\ \underset{\sim}{B} & -\varepsilon \underset{\sim}{C} \end{bmatrix}$$

we get

$$\mathcal{A}_\varepsilon \begin{bmatrix} \underset{\sim}{\rho}^{(\ell)} \\ \underset{\sim}{\gamma}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{r}^{(\ell)} \\ \underset{\sim}{s}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \underset{\sim}{u}^{(\ell)} \\ \underset{\sim}{p}^{(\ell)} \end{bmatrix} - \begin{bmatrix} \underset{\sim}{g}_1 \\ \underset{\sim}{g}_2 \end{bmatrix} . \tag{4.1}$$

From Theorem 2.1, it follows that for $0 < \varepsilon \leq \varepsilon_0$, $\mathcal{A}_\varepsilon$ is non-singular.

A new approximation is then calculated from $\rho^{(\ell)}$, $\gamma^{(\ell)}$ by

$$u^{(\ell+1)} = u^{(\ell)} + \tau_\ell \rho^{(\ell)}$$
$$p^{(\ell+1)} = p^{(\ell)} + \tau_\ell \gamma^{(\ell)} ,$$

as in the steepest descent or similarly in the conjugate gradient algorithms. $\{\rho^{(\ell)}, \gamma^{(\ell)}\}$ are search directions. The convergence of the algorithm is determined by the condition number or by the distribution of eigenvalues of $\mathcal{A}_\varepsilon^{-1}\mathcal{A}$ (see, for example, [13] and [1]). A similar process is applicable to non-linear operators (cf. [3]).

In the modified minimum residual conjugate gradient algorithm (see, for example, [2]), we know that the rate of convergence is determined by how fast a linear combination of the Krylov sequence

$$\mathcal{A}_\varepsilon^{-1}\mathcal{A} \begin{bmatrix} \underset{\sim}{r}^0 \\ \underset{\sim}{s}^0 \end{bmatrix} , \ (\mathcal{A}_\varepsilon^{-1}\mathcal{A})^2 \begin{bmatrix} \underset{\sim}{r}^0 \\ \underset{\sim}{s}^0 \end{bmatrix} , \ \cdots , \ (\mathcal{A}_\varepsilon^{-1}\mathcal{A})^k \begin{bmatrix} \underset{\sim}{r}^0 \\ \underset{\sim}{s}^0 \end{bmatrix} ,$$

may approximate $\begin{bmatrix} \underset{\sim}{r}^0 \\ \underset{\sim}{s}^0 \end{bmatrix}$.

If the eigenvalues of $\mathcal{A}_\varepsilon^{-1}\mathcal{A}$, which may be complex, are located close to 1, we thus have fast convergence. This follows, since we will never get slower convergence than by the geometrically convergent sequence

$$\{[I - \tau\mathcal{A}_\varepsilon^{-1}\mathcal{A}]^k - I\} \begin{bmatrix} \underset{\sim}{r}^0 \\ \underset{\sim}{s}^0 \end{bmatrix} , \qquad \tau \in R .$$

From Theorem 2.1 we know that if ε is small enough, the absolute value of the smallest eigenvalue is $\geq \delta > 0$, where δ is independent of ε. Furthermore,

$$\|\mathcal{A}_\varepsilon^{-1}\mathcal{A}\| \leq C\|\mathcal{A}\|$$

is also independent of ε, as well as on the number of unknowns.

The number of iterations in the conjugate gradient method is thus independent of ε and the number of unknowns, if ε is small enough.

Furthermore, if we choose u^0 and p^0 as solutions of the perturbed system

$$\underset{\sim}{A}\underset{\sim}{u}^0 + \underset{\sim}{B}^T\underset{\sim}{p}^0 = \underset{\sim}{g}_1$$
$$\underset{\sim}{B}\underset{\sim}{u}^0 - \varepsilon\underset{\sim}{p}^0 = \underset{\sim}{g}_2 \ ,$$

we get $\underset{\sim}{r}^0 = 0$ and $\underset{\sim}{s}^0 = \varepsilon\underset{\sim}{p}^0$, so already the initial approximation may be quite accurate.

We will now consider the solution of (4.1). There are several possible approaches for this. One would be to factorize the matrix $\mathcal{A}_\varepsilon$ approximately, which is possible in many problems without pivoting (see [2]).

However, a more general approach would be to eliminate the variable $\underset{\sim}{\gamma}^{(\ell)}$. To simplify this, we use a modified matrix

$$\mathcal{A}'_\varepsilon = \begin{bmatrix} \underset{\sim}{A} & \underset{\sim}{B}^T \\ \underset{\sim}{B} & -\varepsilon I \end{bmatrix} .$$

We observe that the condition number of the Gramian matrix is bounded above by a number, independent of the number of unknowns, so the identity matrix I should be an appropriate replacement of G (cf. the process of "lumping" in dynamical problems). It is readily seen that this modification only slightly influences the condition number of $\mathcal{A}_\varepsilon^{-1}\mathcal{A}$. It is still bounded and the eigenvalues are positive for all sufficiently small ε, with a bound independent of the number of unknowns.

Then we have

$$\underset{\sim}{A}\underset{\sim}{\rho}^{(\ell)} + \underset{\sim}{B}^T\underset{\sim}{\gamma}^{(\ell)} = \underset{\sim}{r}^{(\ell)}$$
$$\underset{\sim}{B}\underset{\sim}{\rho}^{(\ell)} - \varepsilon\underset{\sim}{\gamma}^{(\ell)} = \underset{\sim}{s}^{(\ell)}$$

so that

$$\underset{\sim}{\gamma}^{(\ell)} = \frac{1}{\varepsilon}\left(\underset{\sim}{B}\underset{\sim}{\rho}^{(\ell)} - \underset{\sim}{s}^{(\ell)}\right)$$

and

$$\left(\underset{\sim}{A} + \frac{1}{\varepsilon}\underset{\sim}{B}^T\underset{\sim}{B}\right)\underset{\sim}{\rho}^{(\ell)} = \underset{\sim}{r}^{(\ell)} + \frac{1}{\varepsilon}\underset{\sim}{B}^T\underset{\sim}{s}^{(\ell)} \ .$$

This latter system is now solved for $\underset{\sim}{\rho}^{(\ell)}$ and the solution is substituted into the equation for $\underset{\sim}{\gamma}^{(\ell)}$. We observe that from (ii) and since $1/\varepsilon \geq r_0$,

$$\underset{\sim}{A} + \frac{1}{\varepsilon}\underset{\sim}{B}^T\underset{\sim}{B}$$

is *positive definite*.

(4.3) in its turn may be solved by many algorithms. Since the same matrix $\underset{\sim}{A} + (1/\varepsilon)\underset{\sim}{B}^T\underset{\sim}{B}$ in a linear problem appears at every iterative step, often the fastest way is to LU-factorize it once and for all. If ε is not too small, this is easily done without any cancellation.

Since the number of (outer) iterations will be few, a more economical approach may be to solve the system (4.3) by a preconditioned conjugate gradient method. One then introduces two sparse, non-singular, triangular matrices $\underset{\sim}{E}$ and $\underset{\sim}{F}$. By transformation of the variable $\underset{\sim}{u}$ by $\underset{\sim}{E}$ and of $\underset{\sim}{p}$ by $\underset{\sim}{F}^T$, we then have to solve a system with matrix $\tilde{\underset{\sim}{A}} + (1/\varepsilon)\,\tilde{\underset{\sim}{B}}^T\tilde{\underset{\sim}{B}}$, where

$$\tilde{\underset{\sim}{A}} = \underset{\sim}{E}^{-T}\underset{\sim}{A}\underset{\sim}{E}^{-1},\quad \tilde{\underset{\sim}{B}} = \underset{\sim}{F}^{-1}\underset{\sim}{B}\underset{\sim}{E}^{-1}\ .$$

$\underset{\sim}{E}$ and $\underset{\sim}{F}$ are chosen so that the eigenvalues cluster in two groups, independently of ε.

Then the effective condition number is essentially emanating from the group with the largest condition number. That means that apart from rounding errors, i.e. for not too small ε the number of iterations is independent of ε (cf. [13] and [1]). Since the conjugate gradient method only demands matrix-vector multiplications, there will only be triangular systems of equations to solve at each step. This method may thus be faster than a direct (factorization) method and, furthermore, will save computer storage. However, if ε is too small, rounding errors may make the method converge less fast. On the other hand, we have noticed that the condition number $\kappa(\mathcal{A}_\varepsilon^{-1}\mathcal{A})$ may be close to 1 even for ε not very small. Anyhow, as we have seen, the minimal residual conjugate gradient method will converge even if there are negative eigenvalues of $\mathcal{A}_\varepsilon^{-1}\mathcal{A}$.

5. CONCLUSIONS

In comparing Uzawa's algorithm for the augmented Lagrangian method and the method preconditioned by regularization, we observe that the matrix $\underset{\sim}{A} + (1/\varepsilon)\underset{\sim}{B}^T\underset{\sim}{B}$ appears at each step of both methods. Actually, in a linear problem, an elementary calculation will reveal that the two algorithms will result in identical sequences for $p^{(\ell)}$ (apart from rounding errors), if we choose a fixed sequence $\tau_\ell = -\tau/\varepsilon$ and $1/\varepsilon = r$. However, the preconditioning approach (4.1) opens up the possibility of

choosing $\mathcal{A}_\varepsilon$ in different ways, for instance as an incomplete factorization. Then we will not have a coupled outer-inner sequence of iterations to perform [assuming that we solve (4.3) by iteration]. It is also easier to accelerate the outer process (4.1) by a conjugate gradient method. This latter is parameter-free (the corresponding parameters τ_ℓ are calculated *a posteriori* during the iterations) and will converge for any choice of $\varepsilon \leq 1/r_0$. Thus no accurate estimate of ε is needed. In particular, if a(u,v) is V-elliptic, any $\varepsilon \geq 0$ will do.

Furthermore, non-linear problems (with non-linear operators A) may be solved in this way, with $\mathcal{A}_\varepsilon$ still being a linear operator (matrix), which, however, then should be updated during the (outer) iterative steps. The augmented Lagrangian, on the other hand, usually needs the solution of an inner non-linear optimization problem for each outer iteration.

Finally, the augmented Lagrangian is based on an "energy" functional, i.e. is applicable only when the form a(u,v) is symmetric.

When this is the case, we have seen in Remark 2.1 that the process of regularization followed by elimination of one of the variables is identical to penalization of the corresponding energy functional to get rid of the constraint. Thus, our process means that we have to solve a sequence of such problems. As was already observed in our method (as in the augmented Lagrangian method), it is not necessary to choose a particularly small value of ε as opposed to the penalty method. Bercovier [4] reports that $\varepsilon = \delta h^2$, where $\delta = 10^{-1}$ to 10^{-2}, say, in order to get a good accuracy. In our method, ε may be chosen independent of h, typically $\varepsilon = \delta$ or even larger values, are to be recommended. Thus the well-known difficulties in the penalty method of ill-conditioning associated with a large penalty parameter $1/\varepsilon$ should not appear.

Numerical experiments with different iterative methods for non-definite problems will be reported elsewhere.

6. APPLICATIONS

Boundary value problems with equality constraints appear in many applications. A well-known example is the Stokes problem,

$$
\begin{aligned}
-\nu\Delta\vec{u} + \operatorname{grad} p &= \vec{g} \quad \text{in } \Omega \\
-\operatorname{div} \vec{u} &= 0 \quad \text{in } \Omega \\
\vec{u} &= 0 \quad \text{in } \partial\Omega \ .
\end{aligned}
$$

Here $A = -\nu\Delta$, $B = -\operatorname{div}$ and $B^* = \operatorname{grad}$. Further

$$
V = \overset{\circ}{H}{}^1(\Omega)^n \ , \quad H = L^2(\Omega) \ ,
$$

$$
H^*/\mathcal{N}(B^*) = L^2(\Omega)/R = \{p \in L^2(\Omega) \ ; \ \int_\Omega p\,d\Omega = 0\} \ .
$$

Apparently (i) is satisfied with $r_0 = 0$ since

$$
a(\vec{u},\vec{v}) = \frac{\nu}{2} \sum_{k,\ell} \int_\Omega \frac{\partial u_k}{\partial x_\ell} \frac{\partial v_k}{\partial x_\ell} \, d\Omega \ , \qquad \vec{u},\vec{v} \in V
$$

is V-elliptic. If A is non-symmetric it may be necessary to have $r_0 > 0$, however, for instance in problems with large Reynolds numbers. It is also readily seen that (ii) is satisfied. For the choice of finite dimensional subspaces, see, for example [6] and [4].

Consider now the elasticity equations for incompressible and almost incompressible materials. Thus let an incompressible, isotropic body fill a set $\Omega \subset R^3$. The body is fixed to a rigid support on $\Gamma_0 \subset \partial\Omega$, with meas $(\Gamma_0) > 0$. On $\Gamma_1 = \Gamma - \Gamma_0$ we have given surface forces $\vec{g}$. For notational convenience we let the body forces $= 0$. Let $\vec{u} \in V$ be the displacement and

$$
V = \{v \in H^1(\Omega) \ ; \ \gamma_0 u = 0 \text{ on } \Gamma_0\}^3 \ ,
$$

where γ_0 is the trace operator. Let $H = L^2(\Omega)$ and denote its inner product by $(\cdot,\cdot)$. On $V \times V$ we define the bilinear form

$$
a(\vec{u},\vec{v}) = \sum_{i,j} \varepsilon_{ij}(\vec{u}) \, \varepsilon_{ij}(\vec{v}) \, d\Omega \ ,
$$

$$
\varepsilon_{ij}(\vec{u}) = \frac{1}{2}\left(\frac{\partial u_i}{\partial x_j} + \frac{\partial u_j}{\partial x_i}\right)
$$

and the norm

$$
\|\vec{u}\|_V^2 = \sum_i \int_\Omega \nabla u_i{}^T \nabla u_i \, d\Omega \ .
$$

On $V \times H$ we let

$$
b(\vec{u},p) = \int_\Omega \operatorname{div}(\vec{u}) \, p \, d\Omega \ .
$$

With Poisson's ratio ν and Young's modulus E, we have the Lamé constants

$$\mu = E/(2(1 + \nu)) , \quad \lambda = 2\mu\nu/(1 - 2\nu) .$$

From the constitutive equations,

$$\sigma_{ij} = \delta_{ij}\lambda \sum_{s=1}^{3} \varepsilon_{ss} + 2\mu\varepsilon_{ij} ,$$

we get the pressure

$$p = \frac{1}{E} \operatorname{tr}(\sigma) = \frac{1}{1 - 2\nu} \operatorname{div}(\vec{u}) ,$$

so that

$$b(\vec{u},q) - (1 - 2\nu)(p,q) = 0 \quad \forall q \in H .$$

From the equilibrium equation we get

$$2\mu a(\vec{u},\vec{v}) + 2\nu\mu b(\vec{u},p) = \int_{\Gamma} \vec{\vec{g}}\vec{v} \, d\Omega , \quad \forall \vec{v} \in V .$$

Thus we have the following mixed variational formulation, due to Herrmann

$$\begin{aligned} \frac{1}{\nu} a(\vec{u},\vec{v}) + b(\vec{v},p) &= \frac{1}{2\nu\mu} \int_{\Gamma} \vec{\vec{g}}\vec{v} \, d\Omega \quad \forall \vec{v} \in V \\ b(\vec{u},q) - (1 - 2\nu)(p,q) &= 0 \qquad\qquad \forall q \in H . \end{aligned}$$

For the incompressible case ($\nu = 0.5$) we get a system of type (2.2)

$$\begin{aligned} 2a(\vec{u},\vec{v}) + b(\vec{u},p) &= \frac{1}{\mu_0} \int_{\Gamma} \vec{\vec{g}}\vec{v} \, d\Omega \quad \forall \vec{v} \in V \\ b(\vec{u},q) &= 0 \qquad\qquad \forall q \in H \end{aligned} \tag{6.1}$$

where $\mu_0 = E/2$.

Note that $(\mathcal{J}_\varepsilon)$ is the regularization corresponding to (6.1) of the form as in Remark 2.4 and has the variational formulation of an almost incompressible material, when $\varepsilon = 1 - 2\nu > 0$ is small. With $p = \operatorname{div}(\vec{u})/(1 - 2\nu)$, we get the variational formulation corresponding to the penalized energy functional

$$(\mathcal{J}'_\varepsilon) \qquad 2\mu a(\vec{u},\vec{v}) + \lambda \int_{\Omega} \operatorname{div} \vec{u} \operatorname{div} \vec{v} \, d\Omega = \int_{\Gamma} \vec{\vec{g}}\vec{v} \, d\Omega , \quad \forall \vec{v} \in V .$$

where

$$\lambda = \frac{1}{\varepsilon} 2\nu\mu .$$

From Korn's inequality it follows that $a(\vec{u},\vec{u})$ is V-elliptic. We may thus choose $\varepsilon > 0$ (or rather $\nu < \frac{1}{2}$) freely. One also realizes that (ii) is satisfied, and with Remark 2.4 that Theorem 2.1 is applicable.

For different choices of finite element subspaces, see, for example, [4] and [10].

REFERENCES

[1] O. Axelsson, A class of iterative methods for finite element equations, Computer Methods in Applied Mechanics and Engineering, 9 (1976), pp. 123-137.

[2] O. Axelsson and N. Munksgaard, A class of preconditioned conjugate gradient methods for the solution of a mixed finite-element discretization of the biharmonic operator, Institute for Numerical Analysis report NI-77-14, Technical University of Denmark, Lyngby, Denmark, 1977.

[3] O. Axelsson and U. Nävert, On a graphical package for non-linear partial differential equation problems, Information processing 77, B. Gilchrist (ed.), IFIP, North-Holland, Amsterdam, 1977.

[4] M. Bercovier, Perturbation of mixed variational problems, application to mixed finite element methods, Report, Department of Mathematics, Hebrew University, Jerusalem, 1977.

[5] F. Brezzi, On the existence, uniqueness and approximation of saddle point problems arising from Lagrangian multipliers, R.A.I.R.O., R-2 (1974), pp. 129-151.

[6] M. Crouzeix and P.A. Raviart, Conforming and non-conforming finite element methods for solving the stationary Stokes equations, R.A.I.R.O., R-3 (1973).

[7] R. Glowinski, J.L. Lions and R. Tremolieres, Analyse numérique des inéquations variationelles (Tome 2), Dunod-Bordas, Paris, 1976.

[8] R. Glowinski and O. Pironneau, Numerical methods for the first biharmonic equations and for the two-dimensional Stokes problem, Stanford University report STAN-CS-77-615, 1977.

[9] I. Gustafsson, A class of first order factorization methods, Computer Sciences Report 77.04R, Chalmers University of Technology, Göteborg, Sweden, 1977.

[10] C. Johnson and B. Mercier, Some equilibrium finite element methods for two-dimensional elasticity problems, Computer Sciences Report 77.08R, Chalmers University of Technology, Göteborg, Sweden, 1977.

[11] J.L. Lions and G. Stampacchia, Variational inequalities, Comm. Pure Appl. Math., 20 (1967), pp. 493-519.

[12] D.G. Luenberger, Optimization by vector space methods, Wiley, New York, 1969.

[13] D.G. Luenberger, Introduction to linear and non-linear programming, Addison-Wesley, Reading, Mass., 1973.

[14] P.A. Raviart and J.M. Thomas, A mixed finite element method for second order elliptic problems *in* Proc. Symposium on the Mathematical Aspects of the Finite Element Method, Rome, 1975.

[15] M. Sibony, Methodes itératites pour les équations et inéquations aux dérivées partielles non-linéares de type monotone, Calcolo, 12 (1970), pp. 65-184.