FINITE ELEMENT APPOXIMATION AND ITERATIVE SOLUTION OF A CLASS OF MILDLY NON-LINEAR ELLIPTIC EQUATIONS

by

Tony Chan and Roland Glowinski

STAN-CS-78-674
NOVEMBER 1978

COMPUTER SCIENCE DEPARTMENT
School of Humanities and Sciences
STANFORD UNIVERSITY

FINITE ELEMENT APPROXIMATION AND ITERATIVE SOLUTION OF A CLASS OF MILDLY NON-LINEAR ELLIPTIC EQUATIONS

Tony CHAN*
Computer Science Department
Stanford University
Serra House, Serra Street
Stanford, California
94305

and

Roland GLOWINSKI**
Université Pierre et Marie Curie
Analyse Numérique L.A. 189
4, Place Jussieu
75230 Paris Cedex 05
France

and

IRIA-LABORIA

*Now at Applied Mathematics 101-50, Caltech, Pasadena, CA 91125.

**This work was supported in part by Department of Energy contract No. EY-76-S-03-0326 PA#30.

Abstract

We describe in this report the numerical analysis of a particular class of nonlinear Dirichlet problems. We consider an equivalent variational inequality formulation on which the problems of existence, uniqueness and approximation are easier to discuss. We prove in particular the convergence of an approximation by piecewise linear finite elements. Finally, we describe and compare several iterative methods for solving the approximate problems and particularly some new algorithms of augmented lagrangian type, which contain as special case some well-known alternating direction methods. Numerical results are presented.

TABLE OF CONTENTS

Acknowledgments

The results in this report have been obtained while the second author was visiting the Computer Science Department of Stanford University with the support of DOE contract EY-76-S-03-0326 PA#30.

We would like to thank Professor Gene H. Golub for his interest in this work.

1. - INTRODUCTION

In this report we would like to discuss the numerical analysis of mildly non linear elliptic partial differential equations of the following type

$$(1.1)\qquad \begin{cases} Au + \phi(u) = f \\ u_{|\Gamma} = 0, \end{cases}$$

where in (1.1) :

- A is a second order elliptic operator, possiblynotself adjoint,
- $\phi : \mathbb{R} \to \mathbb{R}$, $\phi \in C^0(\mathbb{R})$ and is non decreasing,
- f is a function defined on Ω.

In fact some of the results and methods to be described here may be extended to more complicated problems or to problems with other boundary conditions. In Section 2 we give a variational formulation of (1.1) (problem (P)) and then introduce an equivalent variational inequality (problem (π)) for which the existence and uniqueness properties, as well as the numerical analysis, are easier to study. We also prove an existence and uniqueness theorem and various lemmas useful in the numerical analysis sections of this report. In Section 3 we study a finite element approximation of (1.1) and prove that the approximate solutions converge to the solution of the continuous problem for some Sobolev norms.

In Section 4 we describe various standard methods which can be used to solve the approximate problem obtained in Sec. 3. Some of these methods are.: Gradient and Conjugate gradient methods, Newton's method, SOR, ADI.

In Sec. 5 we introduce some new methods based on the simultaneous use of penalty and lagrange multipliers which contain some ADI algorithms as particular cases.

In Sec. 6 we use the above methods to solve a test problem. Comparisons between the standard methods of Sec. 4 and the methods of Sec. 5 suggest the superiority and much more robustness for the new algorithms.

We may find in BARTELS-DANIEL [1], DOUGLAS-DUPONT [1](resp. EISENSTAT-SCHULTZ-SHERMAN [1]) conjugate gradient algorithms (resp. Newton's algorithms) for solving equations like (1.1), once a suitable approximation has been made.
For the material concerning the Sobolev spaces, (definitions, properties, etc...) we refer to the classical treatises of ADAMS [1], NECAS [1].

We refer also to D.J. FIGUEIREDO [1] where the reader interested by the theoretical aspects of non linear elliptic equations will find a survey of the various techniques which can be used to study these problems, including the most recent results of the theory of monotone operators.

2. - A CLASS OF MILDLY NON LINEAR ELLIPTIC EQUATIONS

2.1. Formulation of the problems

Let Ω be a bounded domain of $\mathbb{R}^N$ $(N \geq 2)$ with a smooth boundary Γ. We consider

- $V = H_o^1(\Omega) = \{v \mid v \in L^2(\Omega), \frac{\partial v}{\partial x_i} \in L^2(\Omega)\ i=1,\ldots N,\ v|_\Gamma = 0\}$.
- $L : V \to \mathbb{R}$, i.e. $L(v) = \langle f, v\rangle$ where $f \in V' = H^{-1}(\Omega)$ (V' is the dual space of V and $\langle \cdot,\cdot \rangle$ the duality pairing between V' and V).
- $a : V \times V \to \mathbb{R}$ bilinear, continuous and V-elliptic, i.e. $\exists \alpha > 0$ such that

$$(2.1) \qquad a(v,v) \geq \alpha \|v\|_V^2 \ \forall v \in V$$

where

$$(2.2) \qquad \|v\|_V = \left(\int_\Omega |\nabla v|^2 dx\right)^{1/2} ;$$

we don't assume that $a(*,*)$ is symmetric.

- $\phi : R' R$, $\phi \in C^o(\mathbb{R})$, non decreasing with $\phi(0) = 0$.

We then consider the following non linear variational equation

$$(P) \quad \begin{cases} \text{Find } u \in V \text{ such that } \phi(u) \in L^1(\Omega) \cap V' \text{ and} \\ a(u,v) + \langle \phi(u), v \rangle = \langle f, v \rangle \ \forall\ v \in V. \end{cases}$$

It follows from the Riesz Representation Theorem (see, e.g. YOSIDA [1]) that there exists $A \in \mathcal{L}(V,V')$ such that

$$a(u,v) = \langle Au, v \rangle \ \forall\, u,v \in V \ ;$$

therefore (P) is equivalent to

$$(2.3) \quad \begin{cases} Au + \phi(u) = f, \\ u \in V, \\ \phi(u) \in L^1(R) \cap V'. \end{cases}$$

Example 2.1 : Let us consider $a_o \in L^\infty(\Omega)$ such that

$$(2.4) \qquad a_o(x) \geq \alpha > 0 \text{ a.e. on } \Omega.$$

Define $a(\cdot,\cdot)$ by

$$(2.5) \qquad a(u,v) = \int_\Omega a_o(x) \nabla u \cdot \nabla v \, dx + \int_\Omega \beta \cdot \nabla u \, v \, dx$$

where β is a constant vector in $\mathbf{R}^N$.

From the properties of a_o and using the fact that

$$\int_\Omega \beta \cdot \nabla v \, v \, dx = 0 \ \forall\, v \in H_o^1(\Omega)$$

we clearly have

$$a(v,v) \geq \alpha \|v\|_V^2 \text{ so that } a(\cdot,\cdot) \text{ is V-elliptic.}$$

From (2.5) we obtain

$$Au = - \nabla\cdot(a_o \nabla u) + \beta\cdot\nabla u \ ,$$

hence in this particular case (2.3) becomes

$$\begin{cases} - \nabla\cdot(a_o \nabla u) + \beta\cdot\nabla u + \phi(u) = f, \\ \text{UEV}, \quad \phi(u) \in L^1(\Omega). \end{cases}$$

Remark 2.1 : If N=1 we have $H_o^1(\Omega) \subset C^o(\overline{\Omega})$. From this inclusion there is no difficulty in the study of one dimensional problems of type (P). If $N \geq 2$ the main difficulty is precisely related to the fact that $H_o^1(\Omega)$ is not contained in $C^o(\overline{\Omega})$.

Remark 2.2 : The analysis given below may be extended to problems in which either $V = H^1(\Omega)$ or V is a convenient closed subspace of $H^1(\Omega)$.

2.2. A variational inequality related to (P).

2.2.1. Definitions

Let

$$(2.6) \qquad \Phi(t) = \int_0^t \phi(\tau)d\tau$$

$$(2.7) \qquad D(\Phi) = \{v \in V, \Phi(v) \in L^1(\Omega)\} \ .$$

The functional $j : L^2(\Omega) \to \overline{\mathbb{R}}$ is defined by(*)

$$(2.8) \qquad j(v) = \begin{cases} \int_\Omega \Phi(v)dx \text{ if } \Phi(v) \bullet L^1(\Omega), \\ +\infty \quad \text{if } \Phi(v) \notin L^1(\Omega). \end{cases}$$

Instead of studying the problem (P) directly, it is natural to associate with (P) the following elliptic variational inequality(**) .

(*) $\overline{\mathbb{R}} = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$.

(**) For variational inequalities and their approximation see GLOWINSKI-LIONS-TREMOLIERES [1], [2], GLOWINSKI [1], [2].

$$(\pi)\qquad \begin{cases} a(u,v-u) + j(v) - j(u) \geq L(v-u) \ \forall\, v \in V, \\ u \in V. \end{cases}$$

If $a(\cdot,\cdot)$ is symmetric, a standard method to study (P) would have been to consider it as the formal Euler equation of the following minimization problem encountered in the Calculus of Variations :

$$(Q)\qquad \begin{cases} J(u) \leq J(v) \ \forall\, v \in V, \\ u \in V \end{cases}$$

where

$$J(v) = \frac{1}{2}\, a(v,v) + \int_\Omega \Phi(v)dx - L(v).$$

Therefore associating (π) to (P) is a natural generalization of this approach.■
We clearly have

Proposition 2.1 : $D(\Phi)$ is a convex, non empty subset of V.

2.2.2. Properties of $j(\cdot)$:
Since $\phi : \mathbb{R} \to \mathbb{R}$ is C^0, non decreasing with $\phi(0) = 0$ we have

$$(2.9)\qquad \Phi \in C^1(\mathbb{R}),\ \Phi \text{ convex},\ \Phi(0) = 0,\ (s(t) \geq 0\ \forall t \in \mathbb{R}.$$

The properties of $j(\cdot)$ are given by the following lemma :

Lemma 2.1 : The functional $j(\cdot)$ is convex, proper(*) and l.s.c.(**) over $L^2(\Omega)$.

Proof : Since $j(v) \geq 0\ \forall\, v \in L^2(\Omega)$ it follows that $j(\cdot)$ is proper. The convexity of $j(\cdot)$ is obvious from the fact that Φ is convex.

(*) $j(\cdot)$ proper means that $j(v) > -\infty\ \forall v \in V$, $j \not\equiv +\infty$.

(**) l.s.c. : abreviation for lower semi continuous.

Let us prove that $j(\cdot)$ is l.s.c. :
Let $(v_n)_n$, $v_n \in L^2(\Omega)\ \forall\, n$, be such that

$$\lim_{n \to +\infty} vn = v \ \underline{\text{strongly}} \text{ in } L^2(\Omega).$$

Then we have to prove that

$$\text{(2.10)} \qquad \lim_{n \to +\infty} \inf\, j(v_n) \geq j(v).$$

If $\lim_{n \to +\infty} \inf\, j(v_n) = +\infty$ the property is proved. Therefore assume that $\lim_{n \to +\infty} \inf\, j(v_n) = \ell < +\infty$; hence we can extract a subsequence $(v_{n_k})_{n_k}$ such that

$$\text{(2.11)} \qquad \lim_{k \to +\infty} j(v_{n_k}) = \ell ,$$

$$\text{(2.12)} \qquad v_{n_k} \to v \text{ a.e. in } \Omega.$$

Since $\Phi \bullet C^1(\mathbb{R})$, (2.12) implies

$$\text{(2.13)} \qquad \lim_{k \to +\infty} \Phi(v_{n_k}) = \Phi(v) \text{ a.e. .}$$

Moreover $\Phi(v) \geq 0$ a.e. and (2.11) implies that

$$\text{(2.14)} \qquad \{\Phi(v_{n_k})\}_k \text{ is } \underline{\text{bounded}} \text{ in } L^1(\Omega).$$

Hence by Fatou's lemma, it follows from (2.13) (2.14) that we have

$$\text{(2.15)} \qquad \begin{cases} \Phi(v) \in L^1(\Omega), \\ \lim\inf_{k \to +\infty} \int_\Omega \Phi(v_{n_k})dx \geq \int_\Omega \Phi(v)dx. \end{cases}$$

From (2.11) and (2.15) we obtain (2.10) ; this proves the lemma. a

Corollary 2.1 : The functional $j(\cdot)$ restricted to V is convex, proper and l.s.c. .

2.2.3. Existence and uniqueness results for (π).

Theorem 2.1 : Under the above hypothesis on V, a, L and ϕ, problem (π) has a unique solution in $V \cap D(\Phi)$.

Proof : From the above properties of V, a, L and j, we can apply some standard results concerning elliptic variational inequalities (see, e.g., LIONS-STAMPACCHIA [1], LIONS [1], EKELAND-TEMAM [1], GLOWINSKI [1], [2]) which imply that (π) has a unique solution u in V.
Let us now show that $u \in D(\Phi)$:
Taking $v=0$ in (Π) we obtain

$$a(u,u) + j(u) \leq L(u) \leq \|f\|_* \, \|u\|_V \tag{2.16}$$

where

$$\|f\|_* = \sup_{v \in V-\{0\}} \frac{|\langle f,v \rangle|}{\|v\|} .$$

Since $j(u) \geq 0$ and using the ellipticity of $a(\cdot,\cdot)$ we obtain

$$\|u\|_V \leq \frac{\|f\|_*}{\alpha} \tag{2.17}$$

which implies, combined with (2.16) that

$$j(u) \leq \frac{\|f\|_*^2}{\alpha} \tag{2.18}$$

This implies $u \in D(\Phi)$. ∎

Remark 2.3 : If $a(\cdot,\cdot)$ is symmetric, (π) is equivalent to the minimization problem (Q) of Sec. 2.2.1.

2.3. Equivalence between (P) and (Π).
In this Section we shall prove that (P) and (π) are equivalent. We prove first that the unique solution of (π) is also a solution of (P). In order

to prove this last result we need to prove that $\phi(u)$ and $u\phi(u) \in L^1(\Omega)$.

Proposition 2.2 : Let u be the unique solution of (π). Then $u\phi(u)$ and $\phi(u)$ belong to $L^1(\Omega)$.

Proof : Here we use a truncation technique. Let n be a positive integer. Define

$$K_n = \{v \in V, |v(x)| \leq n \text{ a.e.l } .$$

Since K_n is a closed, convex, non empty subset of V, the following variational inequality

$$(\pi_n) \quad \left\{ \begin{array}{l} a(u_n, v-u_n)+j(v)-j(u_n) \geq L(v-u_n) \forall v \in K_n, \\ \\ \blacklozenge_n \bullet K_n \end{array} \right.$$

has a unique solution. Now we prove that

$$\lim_{n \to +\infty} u_n = u \text{ weakly in } V,$$

where u is the solution of (π).

Since $0 \in K_n$, taking v=0 in (π_n) we obtain, as in Theorem 2.1 of this Section, that

$$\|u_n\|_V \leq \frac{\|f\|}{\alpha} , \tag{2.19}$$

$$j(u_n) \leq \frac{\|f\|^2}{\alpha} . \tag{2.20}$$

It follows from (2.19) that there exist a subsequence - still denoted by $\{u_n\}_n$ - and $u^* \in V$ such that

$$\lim_{n \to +\infty} u = u^* \text{ weakly in } V. \tag{2.21}$$

Moreover from the compactness of the canonical injection from $H_o^1(\Omega)$ into $L^2(\Omega)$ (see, e.g., NEČAS [1]), and from (2.21), it follows that

$$\lim_{n\to+\infty} u_n = u \text{ strongly in } L^2(\Omega). \tag{2.22}$$

Relation (2.22) implies that we can extract a subsequence - still denoted by $\{u_n\}_n$ - such that

$$\lim_{n\to+\infty} u_n = u \text{ a.e. in } \Omega. \tag{2.23}$$

Now let $v \in V \cap L^\infty(\Omega)$, then for n large enough we have $v \in K_n$ and

$$a(u_n,u_n)+j(u_n) \le a(u_n,v)+j(v)-L(v-u_n). \tag{2.24}$$

Since

$$\liminf_{n\to+\infty} a(u_n,u_n) \ge a(u^*,u^*)$$

and

$$\liminf_{n\to+\infty} j(u_n) \ge j(u^*)$$

it follows from (2.21) and (2.24) that

$$\begin{cases} a(u^*,u^*)+j(u^*) \le a(u^*,v)+j(v)-L(v-u^*) & \forall\, v \in L^\infty(\Omega) \cap V, \\ u^* \in V \end{cases}$$

which can also be written as

$$\begin{cases} a(u^*,v-u^*)+j(v)-j(u^*) \ge L(v-u^*) \ \forall\, v \in V \cap L^\infty(\Omega), \\ u^* \in V. \end{cases} \tag{2.25}$$

For $n > 0$ define $\tau_n : V \to K_n$ by (see Figure 2.1)

$$\tau_n v = \inf\,(n, \sup(-n, v)). \tag{2.26}$$

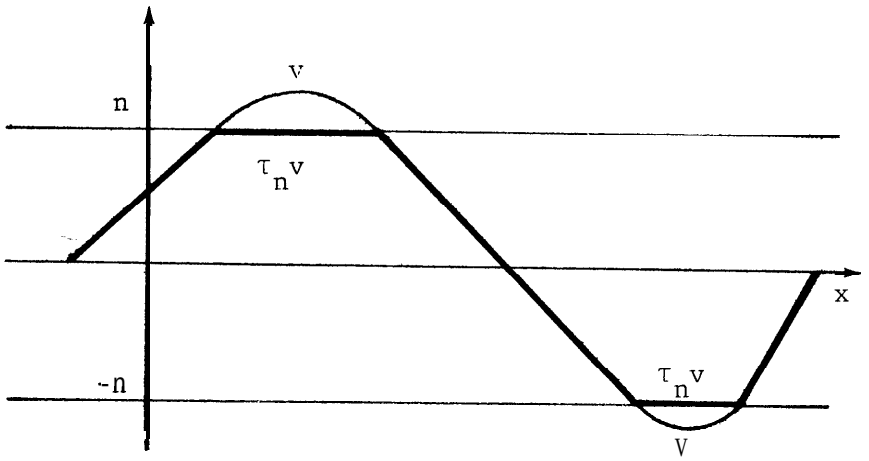


Figure 2.1

Then from STAMPACCHIA [1] we have

$$\tau_n v \in V = H_o^1(\Omega) \quad \forall n, \tag{2.27}$$

$$\begin{cases} \lim_{n \to +\infty} \tau_n v = v \text{ strongly in } V, \\ \lim_{n \to +\infty} \tau_n v = v \text{ a.e. in } \Omega. \end{cases} \tag{2.28}$$

Moreover we obviously have

$$|\tau_n v(x)| \leq |v(x)| \quad \text{a.e.}, \tag{2.29}$$

$$v(x)\tau_n v(x) \geq 0 \text{ a.e.} . \tag{2.30}$$

It follows then from (2.28)-(2.30) that

(2.31) $$0 \le \Phi(\tau_n v) \le \Phi(v) \text{ a.e.},$$

(2.32) $$\lim_{n\to+\infty} \Phi(\tau_n v) = \Phi(v) \text{ a.e. .}$$

Since $\tau_n v \in L^\infty(\Omega) \cap V$ it follows from (2.25) that

(2.33) $$\begin{cases} a(u^*, \tau_n v - u^*) + j(\tau_n v) - j(u^*) \ge L(\tau_n v - u^*) \quad \forall v \in V, \\ u^* \in V. \end{cases}$$

If $v \notin D(\Phi)$ then by Fatou's Lemma

$$\lim_{n\to+\infty} j(\tau_n v) = +\infty .$$

If $v \in D(\Phi)$ it follows from (2.31) and (2.32), by applying Lebesgue's Dominated Convergence Theorem that

$$\lim_{n\to+\infty} j(\tau_n v) = j(v).$$

From these convergence properties, and from (2.28), it follows by taking the limit in (2.33) that

$$\begin{cases} a(u^*, v - u^*) + j(v) - j(u^*) \ge L(v - u^*) \ \forall v \in V, \\ u^* \in V. \end{cases}$$

Then u^* is solution of (π) and from the uniqueness property we have $u^* = u$. This proves that $\lim_{n\to+\infty} u_n = u$ weakly in V.

Let us show that $\phi(u), u\phi(u) \in L^1(\Omega)$.

Let $v \in K_n$; then $u_n + t(v - u_n) \in K_n \ \forall t \in]0,1]$. Replacing v by $u_n + t(v - u_n)$ in (π_n) and dividing both sides of the inequality by t we obtain

(2.34) $$\begin{cases} a(u_n, v - u_n) + \int_\Omega \frac{\Phi(u_n + t(v - u_n)) - \Phi(u_n)}{t} dx \ge L(v - u_n) \\ \forall v \in K_n. \end{cases}$$

Since $\Phi \in C^1(\mathbb{R})$ with $\Phi' = \phi$ we have

$$\lim_{\substack{t \to 0 \\ t>0}} \frac{\Phi(u_n + t(v-u_n)) - \Phi(u_n)}{t} = \phi(u_n)(v-u_n) \text{ a.e. .} \tag{2.35}$$

Moreover, since Φ is convex, we also have $\forall\, t \in]0,1]$

$$\begin{cases} \phi(u_n)\,(v-u_n) \le \dfrac{\Phi(u_n + t(v-u_n)) - \Phi(u_n)}{t} \le \\ \\ \le \Phi(v) - \Phi(u_n) \text{ a.e. in } \Omega. \end{cases} \tag{2.36}$$

From (2.35), (2.36) and using Lebesgue's Dominated Convergence Theorem in (2.34), we obtain

$$a(u_n, v-u_n) + \int_\Omega \phi(u_n)(v-u_n)\,dx \ge L(v-u_n) \;\forall\, v \in K_n. \tag{2.37}$$

Then taking $v=0$ in (2.37) we have

$$a(u_n, u_n) + \int_\Omega \phi(u_n) u_n\,dx \le L(u_n) \;\forall\, n$$

which implies, using (2.19), that

$$\int_\Omega \phi(u_n) u_n\,dx \le \frac{\|f\|_*^2}{a} \quad \forall\, n\ . \tag{2.38}$$

Since $\phi(v)v \ge 0$ a.e. $\forall v \in V$, it follows from (2.38) that

$$\phi(u_n)u_n \text{ is bounded in } L^1(\Omega).$$

Moreover for some subsequence - still denoted $\{u_n\}_n$ - we have

$$\lim_{n \to +\infty} \phi(u_n)u_n = \phi(u)u \text{ a.e. in } \Omega.$$

Then by Fatou's Lemma it follows that

$$u\phi(u) \in L^1(\Omega),$$

and this implies, obviously, that $\phi(u) \in L^1(\Omega)$ and completes the proof of the Proposition.

Incidentally, when proving the convergence of $\{u_n\}_n$ to u, we have proved the following useful result :

Lemma 2.2 : The solution u of (IT) is characterized by

$$(2.39) \qquad \begin{cases} a(u,v-u)+j(v)-j(u) \geq L(v-u) \ \forall\, v \in V \cap L^\infty(\Omega), \\ u \in V, \ \Phi(u) \in L^1(\Omega). \ \blacksquare \end{cases}$$

In view of proving that (IT) implies (P) we also need the two following lemmas :

Lemma 2.3 : The solution u of (π) is characterized by

$$(2.40) \qquad \begin{cases} a(u,v-u)+ \int_\Omega \phi(u)\,(v-u)dx \geq L(v-u) \ \forall\, v \in V \cap L^\infty(\Omega), \\ u \in V, \ u\phi(u) \in L^1(\Omega). \end{cases}$$

Proof : We first prove that

(i) (π) implies (2.40).

Let $v \in L^\infty(\Omega) \cap V$; then $v \in D(\Phi)$ and since $D(\Phi)$ is convex we have $u+t(v-u) \in D(\Phi) \ \forall\, t \in]0,1]$. Replacing v by u+t(v-u) in (IT) and dividing by t we obtain $\forall\, t \in]0,1]$

$$(2.41) \qquad \begin{cases} a(u,v-u)+ \int_\Omega \frac{\Phi(u+t(v-u))-\Phi(u)}{t}\, dx \geq L(v-u) \\ \forall v \in V \cap L^\infty(\Omega). \end{cases}$$

Since Φ is C^1 and is convex, we have

$$(2.42) \qquad \lim_{\substack{t\to 0 \\ t>0}} \frac{\Phi(u+t(v-u))-\Phi(u)}{t} = \phi(u)\,(v-u) \ \text{a.e.},$$

$$(2.43)\qquad \phi(u)(v-u) \le \frac{\Phi(u+t(v-u))-\Phi(u)}{t} \le \Phi(v)-\Phi(u)\ .$$

By Proposition 2.2 we have $\phi(u)$, $\phi(u)u \in L^1(\Omega)$. Hence $\phi(u)(v-u) \in L^1(\Omega)$, and $\Phi(u)$, $\Phi(v) \in L^1(\Omega)\ \forall\ v \in V \cap L^\infty(\Omega)$. Then using the Lebesgue's Dominated Convergence Theorem it follows from (2.42),(2.43) that

$$\lim_{t\to 0} \int_\Omega \frac{\Phi(u+t(v-u))-\Phi(u)}{t}\, dx = \int_\Omega \phi(u)(v-u)dx.$$

Using the above relation and (2.41) we obtain (2.40).

(ii) We next prove that (2.40) implies (I_T).

Let u be a solution of (2.40). Since Φ is convex it follows that

$$-\Phi(u) = \Phi(0)-\Phi(u) \ge \phi(u)(0-u) = -\phi(u)u.$$

This implies $0 \le \Phi(u) \le u\phi(u)$ and $\Phi(u) \in L^1(\Omega)$.

Let $v \in V \cap L^\infty(\Omega)$; then from the inequality

$$\phi(u)(v-u) \le @(v)-@(u) \text{ a.e. in } \Omega$$

we obtain by integration

$$\int_\Omega \phi(u)(v-u)dx \le j(v)-j(u)\ \forall\ v \in V \cap L^\infty(\Omega),$$

which when combined with (2.40) and $\Phi(u) \in L^1(\Omega)$ implies (2.39). Hence from Lemma 2.2 we obtain that (2.40) implies (π). ■

Lemma 2.4 : Let u be the solution of (I_T), then u is characterized by

$$(2.44)\qquad \begin{cases} a(u,v) + \int_\Omega \phi(u)v\, dx = L(v)\ \forall\ v \in L^\infty(\Omega) \cap V, \\ \\ UEV,\ \phi(u) \in L^1(51). \end{cases}$$

Proof : (i) We first prove that (π) implies (2.44).
Let $v \in V \cap L^{\infty}(\Omega)$. If u is the solution of (π), then u is also the unique solution of (2.40). Let τ_n be defined by (2.26), then $\tau_n u \in V \cap L^{\infty}(\Omega)$. Replacing v by $\tau_n u+v$ in (2.40) we obtain

$$(2.45) \qquad \begin{cases} a(u,v)+\int_{\Omega}\phi(u)v\,dx + a(u,\tau_n u-u)+\int_{\Omega}\phi(u)(\tau_n u-u) \\ \geq L(v) + L(\tau_n u-u)\ \forall\, v \in V \cap L^{\infty}(\Omega). \end{cases}$$

It follows from (2.26), (2.28)-(2.30) that

$$(2.46) \qquad \begin{cases} \lim_{n\to+\infty} a(u,\tau_n u-u) = 0, \\ \lim_{n\to+\infty} L(\tau_n u-u) = 0, \end{cases}$$

$$(2.47) \qquad \lim_{n\to+\infty} \phi(u)(\tau_n u-u) = 0 \text{ a.e.},$$

$$(2.48) \qquad 0 \leq \phi(u)(u-\tau_n u) \leq 2u\phi(u) \text{ a.e.}.$$

Then by Lebesgue's Dominated Convergence Theorem and (2.47), (2.48) we obtain

$$(2.49) \qquad \lim_{n\to+\infty} \phi(u)(\tau_n u-u) = 0 \text{ strongly in } L^1(\Omega).$$

Then (2.45), (2.46), (2.49) imply

$$a(u,v) + \int_{\Omega}\phi(u)v\,dx \geq L(v) \quad \forall\, v \in V \cap L^{\infty}(\Omega).$$

Since the above relation also holds for -v we have

$$(2.50) \qquad a(u,v) + \int_{\Omega}\phi(u)v\,dx = L(v) \quad \forall\, v \in V \cap L^{\infty}(\Omega).$$

By Proposition 2.2 we have $\phi(u) \in L^1(\Omega)$, combining this property with (2.50) we obtain (2.44).

This proves that (IT) implies (2.44).

(ii) Next we prove that (2.44) implies (π).

We have

$$a(u,v) + \int_\Omega \phi(u)v\,dx = L(v) \qquad \forall\, v \in V \cap L^\infty(\Omega),$$

then

$$(2.51) \qquad a(u,\tau_n u) + \int_\Omega \phi(u)\tau_n u\,dx = L(\tau_n u) \quad \forall n.$$

Since $\tau_n u \to \tilde{u}$ strongly in V, $\{\int_\Omega \phi(u)\tau_n u\,dx\}_n$ is bounded. But $\phi(u)\tau_n u \geq 0$ a.e., hence we obtain that

$$\phi(u)\tau_n u \text{ is bounded in } L^1(\Omega).$$

We also have

$$\lim_{n\to+\infty} \tau_n u\ \phi(u) = u\phi(u) \text{ a.e.},$$

hence by Fatou's Lemma we have

$$(2.52) \qquad u\phi(u) \in L^1(\Omega).$$

But now we observe that

$$0 \leq \phi(u)\tau_n u \leq \phi(u)u \text{ a.e. .}$$

Hence by Lebesgue's Dominated Convergence Theorem

$$\lim_{n\to+\infty} \int_\Omega \phi(u)\tau_n u\,dx = \int_\Omega \phi(u)u\,dx$$

which along with (2.51) gives

$$a(u,u) + \int_\Omega \phi(u)u \, dx = L(u). \tag{2.53}$$

Then by substracting (2.53) from (2.44) we obtain

$$\begin{cases} a(u,v-u) + \int_\Omega \phi(u)(v-u)dx = L(v-u) \quad \forall v \in V \cap L^\infty(\Omega), \\ u \in V, \ u\phi(u) \in L^1(\Omega) \end{cases} \tag{2.54}$$

and obviously (2.54) implies (2.40).
This completes the proof of the Lemma.

Corollary 2.2 : If u is the solution of (π) then u is also a solution of (P).

Proof : We recall that $V' = H^{-1}(\Omega) \subset \mathcal{D}'(\Omega)$ (*) and that

$$a(u,v) = \langle Au, v \rangle \ \forall u,v \in V,$$
$$L(u) = \langle f, v \rangle \ \forall v \in V.$$

Let u be a solution of (π). Then u is characterized by (2.44) and since $\mathcal{D}(\Omega) \subset V$ (where $\mathcal{D}(\Omega) = \{v \in C^\infty(\overline{\Omega})$, v has a compact support in $\Omega\}$) we obtain

$$\langle Au, v \rangle + \int_\Omega \phi(u)v \, dx = \langle f, v \rangle \ \forall v \in \mathcal{D}(\Omega). \tag{2.55}$$

It follows then from (2.55) that

$$Au + \phi(u) = f \text{ in } \mathcal{D}'(\Omega). \tag{2.56}$$

Since Au and $f \in V'$ we have $\phi(u) \in V'$; hence

$$\phi(u) \in L^1(\Omega) \cap H^{-1}(\Omega)$$

and from (2.56) we obtain that u is a solution of (P).■

(*) $\mathcal{D}'(\Omega)$: space of distributions on Ω

We have proved up to this point that the unique solution of (π) is also a solution of (P). Now we prove the reciprocal property, that is, every solution of (P) is a solution of (π) and hence (P) has a unique solution. In order to prove this we shall use the following density lemma :

Lemma 2.5 : $\mathcal{D}(\Omega)$ is dense in $V \cap L^{\infty}(\Omega)$, $V \cap L^{\infty}(\Omega)$ being equipped with the strong topology of V and the weak * topology of $L^{\infty}(\Omega)$.

Proof : Let $v \in V \cap L^{\infty}(\Omega)$. Since $\overline{\mathcal{D}(\Omega)}^{H^1(\Omega)} = V$, there exists a sequence $\{v_n\}_n$, $v_n \in \mathcal{D}(\Omega)$ such that

$$\lim_{n \to +\infty} v_n = v \text{ strongly in } V. \tag{2.57}$$

Let us define w_n by (see Figure 2.2)

$$w_n = \min\ (v^+, v+n) - \min(v^-, v_n^-). \tag{2.58}$$

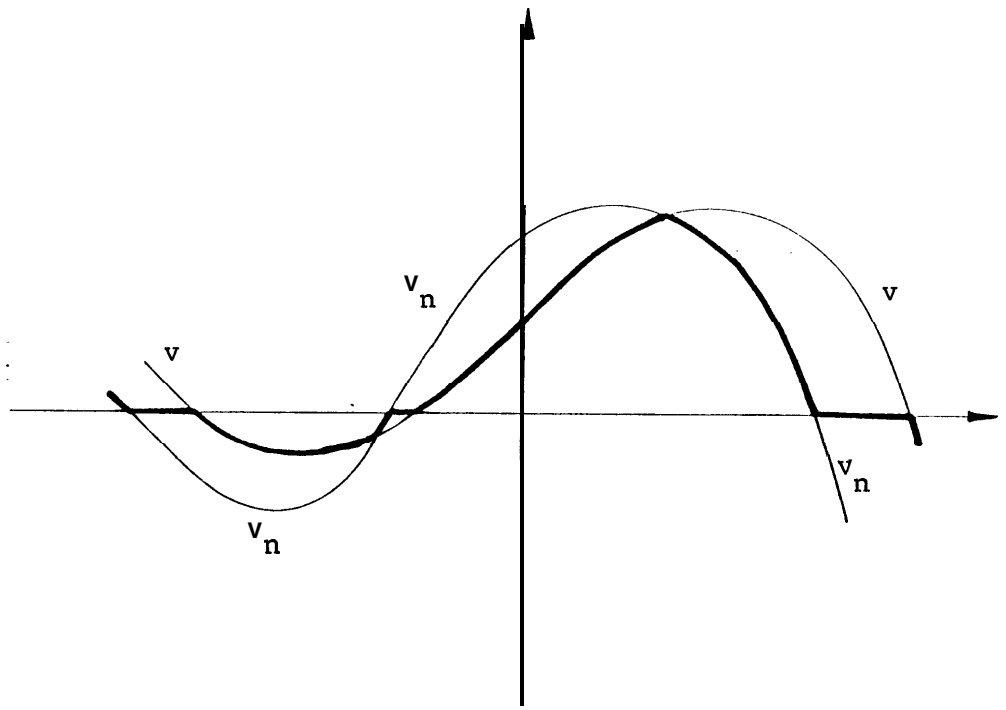


Figure 2.2

(The reinforced curve is the graph of w_n)

Then

(2.59) w_n has a compact support in Ω,

(2.60) $\|w_n\|_{L^\infty(\Omega)} \leq \|v\|_{L^\infty(\Omega)} \quad \forall n.$

Moreover, since (cf. STAMPACCHIA [1 **1**) the mapping

$$v \to \{v^+, v^-\}$$

is continuous from $H^1(\Omega)$ to $H^1(\Omega) \times H^1(\Omega)$ (resp. v to $v \times V$), we have from (2.57) that

(2.61) $\lim_{n \to +\infty} w_n = v$ strongly in V.

From (2.60),(2.61) we obtain that

$$\lim_{n \to +\infty} w_n = v \text{ in the weak * topology of } L^\infty(\Omega).$$

Thus we have proved that

$$\mathcal{V} = \{v \in V \cap L^\infty(\Omega),\ v \text{ has a compact support in } \Omega\}$$

is dense in $V \cap L^\infty(\Omega)$ for the topology given in the statement of the Lemma. Let $v \in \mathcal{V}$, and $(\rho_n)_n$ be a mollifying sequence, i.e.

$$\begin{cases} \rho_n \in \mathcal{D}(\mathbb{R}^N),\ \rho_n \geq 0, \\ \int_{\mathbb{R}^N} \rho_n(y)\,dy = 1, \\ \lim_{n \to +\infty} \text{support } (\rho_n) = \{0\}\ (*). \end{cases}$$

Then define $\tilde{v}$ and $\tilde{v}_n$ by

(*) i.e., for n large enough, support (ρ_n) is contained in any given neighbourhood of 0.

$$\tilde{v}(x) = \begin{cases} v(x) \text{ if } x \in \Omega, \\ 0 \text{ if } x \notin \Omega , \end{cases} \tag{2.62}$$

$$\tilde{v}_n = \rho_n * \tilde{v} \tag{2.63}$$

i.e.

$$\tilde{v}_n(x) = \int_{\mathbb{R}^N} \rho_n(x-y)\tilde{v}(y)dy \quad \forall x \in \mathbb{R}^N. \tag{2.64}$$

It is well known then (see, e.g., LIONS [2],NECAS [1]) that

$$\tilde{v}_n \in \mathcal{D}(\mathbb{R}^N), \lim_{n \to +\infty} \tilde{v}_n = \tilde{v} \text{ strongly in } H^1(\mathbb{R}^N), \tag{2.65}$$

$$\tilde{v}_n \text{ has a compact support in } \Omega \text{ for } n \text{ large enough.} \tag{2.66}$$

Let $v_n = \tilde{v}_n|_\Omega$, then for n large enough

$$\begin{cases} v_n \in \mathcal{D}(\Omega), \\ \lim_{n \to +\infty} v_n = v \text{ strongly in } V. \end{cases}$$

Since $\|\tilde{v}\|_{L^\infty(\mathbb{R}^N)} = \|v\|_{L^\infty(\Omega)}$ it follows from (2.64) that

$$|v_n(x)| \leq \int_{\mathbb{R}^N} \rho_n(x-y)|\tilde{v}(y)|dy \leq \|v\|_{L^\infty(\Omega)} . \tag{2.67}$$

It follows then from (2.67) that for n large enough we have

$$\|v_n\|_{L^\infty(\Omega)} \leq \|v\|_{L^\infty(\Omega)} . \tag{2.68}$$

Summarizing the above results we have proved that $\forall v \in L^\infty(\Omega) \cap V$, there exists a sequence $(v_n)_n$, $v_n \in \mathcal{D}(\Omega)$ such that

$$\lim_{n \to +\infty} v_n = v \text{ strongly in } V, \tag{2.69}$$

$$\|v_n\|_{L^\infty(\Omega)} \leq \|v\|_{L^\infty(\Omega)} \quad \forall n. \tag{2.70}$$

Hence from (2.69),(2.70) we obtain that

$$v_n \to v \text{ in } L^\infty(\Omega) \text{ weak } *.$$

This completes the proof of the Lemma.

Theorem 2.2 : Under the above hypothesis on V, $a(\cdot,\cdot)$, L and ϕ, problems (π) and (P) are equivalent.

Proof : We have already proved that (π) implies (P). We need only to prove that (P) implies (π).
From the definition of (P) we have

$$(2.71) \qquad \begin{cases} a(u,v) + \langle \phi(u), v \rangle = L(v) \quad \forall\, v \in V, \\ u \in V,\ \phi(u) \in H^{-1}(\Omega) \cap L^1(\Omega). \end{cases}$$

It follows from (2.71) that

$$(2.72) \qquad a(u,v) + \int_\Omega \phi(u) v \, dx = L(v) \quad \forall v \in \mathcal{D}(\Omega).$$

If $v \in V \cap L^\infty(\Omega)$ we know from Lemma 2.5 that there exists a sequence $(v_n)_n$, $v_n \in \mathcal{D}(\Omega)$ $\forall n$, such that

$$(2.73) \qquad \lim_{n \to +\infty} v_n = v \text{ strongly in } V,$$

$$(2.74) \qquad \lim_{n \to +\infty} v_n = v \text{ in } L^\infty(\Omega) \text{ weak } *.$$

Since $v_n \in \mathcal{D}(\Omega)$ we have, from (2.72)

$$(2.75) \qquad a(u,v_n) + \int_\Omega \phi(u) v_n \, dx = L(v_n) \quad \forall\, n.$$

It follows from (2.73) that

$$\lim_{n \to +\infty} a(u,v_n) = a(u,v), \quad \lim_{n \to +\infty} L(v_n) = L(v),$$

and since $\phi(u) \in L^1(\Omega)$, (2.74) implies that

$$\lim_{n \to +\infty} \int_\Omega \phi(u) v_n dx = \int_\Omega \phi(u) v \, dx.$$

Thus taking the limit in (2.75) we obtain

$$a(u,v) + \int_\Omega \phi(u) v \, dx = L(v) \quad \forall v \in V \cap L^\infty(\Omega).$$

Therefore (P) implies (2.43) which in turn implies (π) (see Lemma 2.4)). This completes the proof of the Theorem.

2.4. Some comments on the continuous problem.

We have studied (P) and (π) with rather weak hypothesis, namely $\phi \in C^0(\mathbb{R})$ and is non decreasing, and $f \in H^{-1}(\Omega)$. The proof we have given for the equivalence between (P) and (π) can be made shorter using more sophisticated tools of Convex Analysis and the theory of monotone operators (see LIONS [1] and the bibliography therein)(*). However our proof is very elementary and some of the lemmas we have obtained will be useful for the numerical analysis of the problem (P).

Regularity results for problems a little more complicated than (P) and (π) are given in BREZIS-CRANDALL-PAZY [1]; in particular for $f \in L^2(\Omega)$ and with convenient smoothness hypothesis for A, the $H^2(\Omega)$-regularity of u is proved there.

3. - A FINITE ELEMENT APPROXIMATION OF (π) and (P).

3.1. Definition of the approximate problem

Let Ω be a bounded polygonal(**) domain of $\mathbb{R}^2$ and $\mathcal{T}_h$ be a triangulation of Ω satisfying

(i) $T \subset \bar{\Omega}, \forall T \in \mathcal{T}_h, \bigcup_{T \in \mathcal{T}_h} T = \bar{\Omega}$,

(ii) $\mathring{T} \cap \mathring{T}' = \emptyset, \forall T, T' \in \mathcal{T}_h, T \neq T'$; $\bigcup_{T \in \mathcal{T}_h} T = \bar{\Omega}$,

(iii) If $T, T' \in \mathcal{T}_h$, $T \neq T'$ then $T \cap T' = \emptyset$ or T and T' have either only one common vertex or a whole common edge ; as usual h will be the length of the largest edge of $\mathcal{T}_h$.

(*) See also OSBORN-SATHER C1].
(**) This assumption is not essential.

We approximate V by

$$V_h = \{v_h \in C^0(\bar{\Omega}) \ , \ v_h|_\Gamma = 0, v_h|_T \in P_1 \quad \forall T \in \mathcal{T}_h\}$$

where P_1 = space of polynomials in two variables of degree ≤ 1. It is then natural to approximate (P) and (π) respectively by

$$(P_h^*) \quad \begin{cases} a(u_h,v_h) + \int_\Omega \phi(u_h)v_h \, dx = L(v_h) \ \forall \ v_h \in V_h, \\ u_h \in V_h \end{cases}$$

$$(\pi_h^*) \quad \begin{cases} a(u_h,v_h-u_h) + j(v_h)-j(u_h) \geq L(v_h-u_h) \quad \forall v_h \in V_h, \\ u_h \in V_h \end{cases}$$

with $j(v_h) = \int_\Omega \Phi(v_h)dx$.

Obviously (P_h^*) and (π_h^*) are equivalent.

From a computational point of view we cannot use in general (P_h^*) and (π_h^*) directly since they involve the computation of integrals which cannot be done exactly. For this reason we shall have to modify (π_h^*) and (P_h^*) by using some numerical integration procedures.

In fact we have to approximate $a(\cdot,\cdot)$, L and $j(\cdot)$, but since the approximation of $a(\cdot,\cdot)$ and L is studied in, e.g., STRANG-FIX [1], CIARLET-RAVIART [1], ODEN-REDDY [1], CIARLET [1], [2] we shall assume that we still work with $a(\cdot,\cdot)$ and L, but we shall approximate $j(\cdot)$.

Hence using the notation of Figure 3.1 below, we approximate $j(\cdot)$ by

$$(3.1) \qquad j_h(v_h) = \sum_{T \in \mathcal{T}_h} \frac{\text{meas.}(T)}{3} \sum_{i=1}^{3} \Phi\left(v_h(M_{iT})\right) \ \forall v_h \in V_h.$$

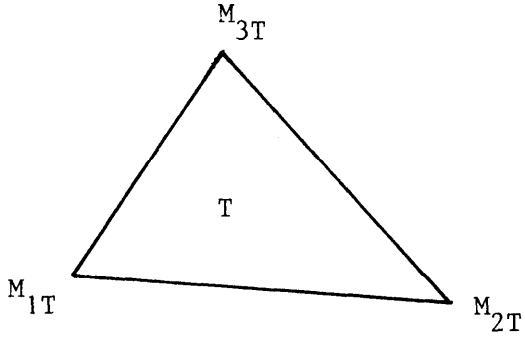


Figure 3.1

Actually $j_h(v_h)$ may be viewed as the exact integral of some piecewise constant function. More precisely let us denote by Σ_h the set of nodes of $\mathcal{T}_h$; assume that Σ_h has been ordered by $i=1,2,\ldots N_h$ where N_h = Card (Σ_h).

Let $M_i \in \Sigma_h$; we define a domain Ω_1 by joining, as in Figure 3.2, the centroids of the triangles having M_i as a common vertex, to the midpoints of the edges having M_i as a common extremity (if M_i is a boundary point the modification of Figure 3.2 is trivial to do).

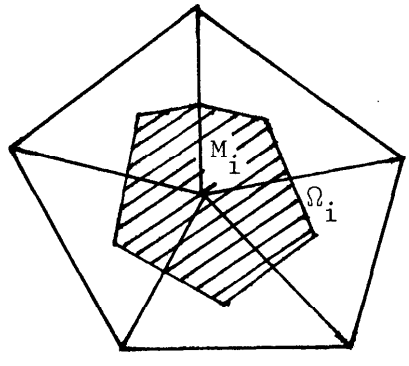


Figure 3.2

Let us define L_h, a space of piecewise constant functions by

$$(3.2) \qquad L_h = \{\mu_h \mid \mu_h = \sum_{i=1}^{N_h} \mu_i \chi_i,\ \mu_i \in \mathbb{R}\ \forall\, i=1,\ldots N_h\} \ ,$$

where χ_i is the characteristic function of Ω_i , i.e.,

$$\begin{cases} \chi_i(x) = 1 & \text{if } x \in \Omega_i, \\ \chi_i(x) = 0 & \text{if } x \notin \Omega_i. \end{cases}$$

Then we define $q_h : C^o(\bar{\Omega}) \cap H_o^1(\Omega) \to L_h$ by

$$(3.3) \qquad q_h v = \sum_{i=1}^{N_h} v(M_i)\chi_i .$$

It follows then from (3.1), (3.2), (3.3) that

$$(3.4) \qquad j_h(v_h) = \int_\Omega \Phi(q_h v_h)\, dx \quad \forall v_h \in V_h .$$

We also have

$$(3.5) \qquad j_h(v_h) = j(q_h v_h) \quad \forall v_h \in V_h .$$

Then we approximate (P) and (π) by

$$(P_h) \quad \begin{cases} a(u_h, v_h) + \int_\Omega \phi(q_h u_h)\, q_h v_h\, dx = L(v_h) \quad \forall v_h \in V_h, \\ u_h \in V_h \end{cases}$$

and

$$(\pi_h) \quad \begin{cases} a(u_h, v_h - u_h) + j_h(v_h) - j_h(u_h) \geq L(v_h - u_h) \quad \forall v_h \in V_h, \\ u_h \in V_h . \end{cases}$$

We have then the obvious

Theorem 3.1 : Problems (P_h) and (π_h) are equivalent and have a unique solution.

3.2. Convergence of the approximate solution

Theorem 3.2 : If as $h \to 0$ the angles of $\mathcal{C}_h$ are uniformly bounded below by $\theta_o > 0$, then

$$\lim_{h\to 0} \|u_h - u\|_V = 0,$$

where u and u_h are respectively the solutions of (P) and (P_h).

Proof : (i) A priori estimates for u_h.

Taking $v_h = 0$ in (π_h) we obtain $\forall h$

$$\|u_h\|_V \leq \frac{\|f\|_*}{\alpha}, \tag{3.6}$$

$$0 \leq \int_\Omega \Phi(q_h u_h)\,dx \leq \frac{\|f\|_*^2}{\alpha}. \tag{3.7}$$

(ii) Weak convergence of u_h

It follows from (3.6) and from the compactness of the injection of V in $L^2(\Omega)$ that we can extract from $(u_h)_h$ a subsequence, still denoted by $(u_h)_h$, such that

$$u_h \to u^* \text{ weakly in } V, \tag{3.8}$$

$$u_h \to u^* \text{ strongly in } L^2(\Omega), \tag{3.9}$$

$$u_h \to u^* \text{ a.e. in } \Omega. \tag{3.10}$$

Admitting for the moment the following inequality (which we shall prove later)

$$\begin{cases} \|q_h v_h - v_h\|_{L^p(\Omega)} \leq \frac{2h}{3} \|\nabla v_h\|_{L^p(\Omega)\times L^p(\Omega)} \\ \forall v_h \in V_h, \forall p \text{ with } 1 \leq p \leq +\infty, \end{cases} \tag{3.11}$$

it follows from (3.6) and (3.9) that

$$q_h u_h \to u^* \text{ strongly in } L^2(\Omega). \tag{3.12}$$

Then, modulo another extraction of a subsequence, we have

$$(3.13)\qquad \begin{cases} q_h u_h \to u^* \ \underline{a.e.}\ \text{in } \Omega, \\ \Phi(q_h u_h) \to \Phi(u^*) \ \underline{a.e.}\ \text{in } \Omega. \end{cases}$$

Taking $v \in \mathcal{D}(\Omega)$ it follows from STRANG-FIX [1], CIARLET [1],[2] that under the assumptions on $\mathcal{T}_h$ in the statement of the Theorem we have

$$(3.14)\qquad \| r_h v - v \|_{W^{1,\infty}(\Omega)} \leq Ch \| v \|_{W^{2,\infty}(\Omega)} \quad \forall\, v \in \mathcal{D}(\Omega),$$

$$(3.15)\qquad \| r_h v - v \|_{L^{\infty}(\Omega)} \leq Ch^2 \| v \|_{W^{2,\infty}(\Omega)} \quad \forall\, v \in \mathcal{D}(\Omega),$$

where :

- C is--a constant independent of v and h,
- r_h is the usual linear interpolation operator over $\mathcal{T}_h$ i.e.

$$\begin{cases} r_h v \in V_h \quad \forall\, v \in H_o^1(\Omega) \cap C^o(\overline{\Omega}), \\ r_h v(P) = v(P) \quad \forall\, P \in \Sigma_h. \end{cases}$$

Moreover (3.11) with $p = +\infty$, (3.14),(3.15) imply that

$$(3.16)\qquad \| q_h r_h v - v \|_{L^{\infty}(\Omega)} \to 0 \quad \forall\, v \in \mathcal{D}(\Omega).$$

Taking $r_h v$ in (Π_h) we obtain

$$(3.17)\qquad \begin{cases} a(u_h, u_h) + \int_\Omega \Phi(q_h u_h)\,dx \leq a(u_h, r_h v) + \\ \int_\Omega \Phi(q_h r_h v)\,dx - L(r_h v - u_h) \quad \forall\, v \in \mathcal{D}(\Omega). \end{cases}$$

From (3.8), (3.12) and from Lemma 2.1 we have

$$a(u^*, u^*) + \int_\Omega \Phi(u^*)\,dx \leq \liminf_{h \to 0} \left(a(u_h, u_h) + \int_\Omega \Phi(q_h u_h)\,dx \right).$$

Moreover

$$\lim_{h\to 0} \int_\Omega \Phi(q_h r_h v)dx = \int_\Omega \Phi(v)dx = j(v) \quad \forall v \in \mathcal{D}(\Omega).$$

Then in the limit in (3.17) we obtain

$$(3.18) \quad \begin{cases} a(u^*,u^*) + j(u^*) \le a(u^*,v) + j(v) - L(v-u^*) \\ \forall v \in \mathcal{D}(\Omega). \end{cases}$$

From Fatou's **Lemma** applied to (3.7),(3.13) we obtain

$$(3.19) \quad \Phi(u^*) \in L^1(\Omega).$$

It follows then from (3.18), (3.19) that u^* satisfies

$$(3.20) \quad \begin{cases} a(u^*,v-u^*)+j\ (v>-j\ (u^*) \ge L(v-u^*)\ \forall v \in \mathcal{D}(\Omega), \\ u^* \in v,\ \Phi(u^*)\ EL'(n). \end{cases}$$

We now take $v \in V \cap L^\infty(\Omega)$; it follows from Lemma 2.5 that there exists a sequence $\{v_n\}_n$ such that $v_n \in \mathcal{D}(\Omega)$ and

$$(3.21) \quad \lim_{n\to+\infty} v_n = v \text{ strongly in } V,$$

$$(3.22) \quad \lim_{n\to+\infty} v_n = v \text{ in } L^\infty(\Omega) \text{ weak } *.$$

We have from (3.20) that

$$(3.23)\qquad \begin{cases} a(u^*, v_n - u^*) + j(v_n) - j(u^*) \geq L(v_n - u^*) \quad \forall n, \\ u^* \in V, \Phi(u^*) \in L^1(\Omega). \end{cases}$$

We obviously have from (3.21) that

$$\lim_{n \to +\infty} a(u^*, v_n - u^*) = a(u^*, v - u^*),$$

$$\lim_{n \to +\infty} L(v_n - u^*) = L(v - u^*).$$

Since $v_n \to v$ in the weak * topology of $L^\infty(\Omega)$, we have

$$(3.24)\qquad \|v_n\|_{L^\infty(\Omega)} \leq \text{Const.} \quad \forall n.$$

Moreover, for some subsequence, (3.21) implies that

$$(3.25)\qquad \lim_{n \to +\infty} v_n = v \ \underline{\text{a.e.}} \text{ in } \Omega.$$

From (3.25) we obtain that

$$(3.26)\qquad \Phi(v_n) \to \Phi(v) \text{ a.e. in } \Omega.$$

From (3.25),(3.26) one can easily see that the Lebesgue's Dominated Convergence Theorem can be applied to $\{\Phi(v_n)\}_n$. Hence we obtain

$$\lim_{n \to +\infty} j(v_n) = \lim_{n \to +\infty} \int_\Omega \Phi(v_n)\,dx = \int_\Omega \Phi(v)\,dx = j(v).$$

Therefore in the limit in (3.23) we have

$$(3.27)\qquad \begin{cases} a(u^*, v - u^*) + j(v) - j(u^*) \geq L(v - u^*) \ \forall\, v \in V \cap L^\infty(\Omega), \\ u^* \text{ EV}, \ \Phi(u^*) \in L^1(\Omega). \end{cases}$$

Since from Lemma 2.2 we know that (3.27) is equivalent to (π) we have thus proved that $u^* = u$ where u is the solution of (π) (and (P)).
From the uniqueness of the solution of (π) it follows that the whole sequence $(u_h)_h$ converges to u.

(iii) Strong convergence of $(u_h)_h$.

It follows from (π_h) and from the V-ellipticity of $a(\cdot,\cdot)$ that

$$(3.28)\qquad \left\{ \begin{array}{l} \alpha \|u_h-u\|_V^2 + j_h(u_h) \le a(u_h-u,u_h-u)+j_h(u_h) = \\ a(u,u)-a(u_h,u)-a(u,u_h)+a(u_h,u_h)+j_h(u_h) \le \\ \le a(u,u)-a(u_h,u)-a(u,u_h)+a(u_h,r_h v)+j_h(r_h v)-L(r_h v-u_h) \quad \forall\, v \in \mathfrak{D}(\Omega). \end{array} \right.$$

Using the various convergence results of Part (ii) we obtain from (3.28) that

$$(3.29)\qquad \left\{ \begin{array}{l} j(u) \le \liminf j_h(u_h) \le \liminf \left(\alpha \|u_h-u\|_V^2 + j_h(u_h)\right) \le \\ \le \limsup \left(\alpha \|u_h-u\|_V^2 + j_h(u_h)\right) \le \\ \le a(u,v-u)+j(v)-L(v-u) \quad \forall\, v \in \mathfrak{D}(\Omega). \end{array} \right.$$

Using as in Part (ii) the density of $\mathfrak{D}(\Omega)$ in $V \cap L^\infty(\Omega)$ (for the strong topology of V and the weak * topology of $L^\infty(\Omega)$) we obtain that (3.29) also holds for all $v \in V \cap L^\infty(\Omega)$.
Taking τ_n like in Sec. 2.3, relation (2.26)' we have then

$$(3.30)\qquad \left\{ \begin{array}{l} j(u) \le \liminf j_h(u_h) \le \liminf \left(\alpha \|u_h-u\|_V^2 + j_h(u_h)\right) \le \\ \le \limsup \left(\alpha \|u_h-u\|_V^2 + j_h(u_h)\right) \le \\ \le a(u,\tau_n v-u)+j(\tau_n v)-L(\tau_n v-u) \quad \forall\, v \in V, \forall\, n. \end{array} \right.$$

From the properties of τ_n (see Sec. 2.3), we have at the limit in (3.30)

$$
(3.31)\qquad \left\{ \begin{array}{l} j(u) \le \lim\ \inf\left(\alpha \| u_h - u \|_V^2 + j_h(u_h) \right) \le \\ \\ \le \lim\ \sup\left(\alpha \| u_h - u \|_V^2 + j_h(u_h) \right) \le \\ \\ \le a(u, v-u) + j(v) - L(v-u) \quad \forall\, v \in V. \end{array} \right.
$$

Taking $v=u$ in (3.31) we obtain that

$$
\lim_{h \to 0} j_h(u_h) = j(u),
$$

$$
\lim_{h \to 0} \| u_h - u \|_V = 0.
$$

This proves the theorem modulo the proof of (3.11).

Lemma 3.1 : We have $\forall p,\ 1 \le p \le +\infty$,

$$
\| q_h v_h - v_h \|_{L^p(\Omega)} \le \frac{2}{3}\, h\, \| \nabla v_h \|_{L^p(\Omega) \times \ L^p(\Omega)} \quad \forall\, v_h \in V_h .
$$

Proof : We use the notation of Sec. 3.1.

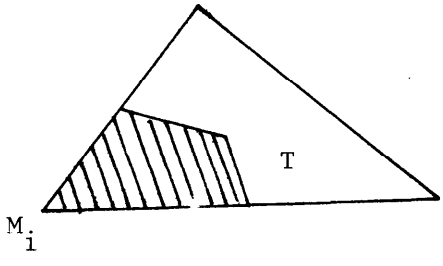


Figure 3.3

We have (see Figure 3.3)

$$
(3.32)\qquad \left| q_h v_h(M) - v_h(M) \right| = \left| v_h(M_i) - v_h(M) \right| \quad \forall M \in \Omega_i \cap T.
$$

But since $v_{h|T} \in P_1$ we have

$$v_h(M) = v_h(M_i) + \overrightarrow{M_i M} \cdot \nabla v_h \quad \forall M \in \Omega_i \cap T$$

from which it follows, combined with (3.32)' that

$$|q_h v_h(M) - v_h(M)| \leq |\overrightarrow{M_i M}| \; |\nabla v_h| \quad \forall M \in \Omega_i \cap T.$$

It follows from the definition of h that we have

$$|M_i M| \leq \frac{2}{3} h \quad \forall M \in \Omega_i \cap T, \quad \forall T$$

from which it follows that

$$|q_h v_h(x) - v_h(x)| \leq \frac{2}{3} h \, |\nabla v_h(x)| \text{ a.e. in } \Omega, \; \forall v_h \in V_h.$$

This implies

$$\| q_h v_h - v_h \|_{L^p(\Omega)} \leq \frac{2}{3} h \, \| \nabla v_h \|_{L^p(\Omega) \times L^p(\Omega)} .$$

This proves the lemma.

Remark 3.1 : The numerical analysis of problems like (P) but with much stronger hypothesis on $a(\cdot,\cdot)$, ϕ, f is considered in CIARLET-SCHULTZ-VARGA [1] where error estimates are given.

4. - A SURVEY OF ITERATIVE METHODS FOR SOLVING (P_h).

4.1. Orientation

-In this section we briefly describe some iterative methods which may be useful for computing the solution of (P_h) (and (π_h)). Actually most of these methods may be extended to other non linear problems. Many of the methods to be described here can be found in ORTEGA-RHEINBOLDT [1].

A method based on penalty and duality techniques will be described in Sec. 5.

4.2. Formulation of the discrete problem

Here we are using the notation of the continuous problem. Taking as unknowns the values of u_h at the interior nodes of $\mathcal{T}_h$, the problem (P_h) reduces to the finite dimensional non linear problem

(4.1) $Au + D\phi(u) = f$

where A is a $N \times N$ positive definite matrix, D is a diagonal matrix with positive diagonal elements d_i's and where

$$u = \{u_1, \ldots u_N\} \in \mathbb{R}^N \text{ , } f \in \mathbb{R}^N,$$

$$\phi(u) \in \mathbb{R}^N \text{ with } (\underset{\sim}{\phi(u)})_i = \phi(u_i).$$

Clearly from the properties of A,D,ϕ,f we can see that (4.1) has a unique solution.

4.3. Gradient mehtods

The basic algorithm with constant step (see CEA [1]) is given by

(4.2) $u^o \in \mathbb{R}^N$ given,

(4.3) $\underset{\sim}{U}^{n+1} = \underset{\sim}{u}^n - \rho S^{-1}(A\underset{\sim}{u}^n + D\phi(\underset{\sim}{u}^n) - \underset{\sim}{f}), \ \rho > 0.$

In (4.3), S is a symmetric and positive definite matrix. A canonical choice is obviously $S = I_N$, but in most problems it will give a slow speed of convergence. If A is symmetric the natural choice is S=A and if A# A^t we can take $S = \frac{A+A^t}{2}$. If ϕ is locally Lipschitz continuous (i.e. Lipschitz continuous on the bounded sets of $\mathbb{R}$) then algorithm (4.2),(4.3) converges to the unique solution u of (4.1) if ρ is taken sufficiently small. Obviously the closer u^o is to u, the faster is the convergence.

Remark 4.1 : If $A=A^t$ then $Av+D\phi(\underset{\sim}{v})-\underset{\sim}{f}$ is the gradient at v of the convex functional

$$J(v) = \frac{1}{2}(A\underset{\sim}{v}, \underset{\sim}{v}) + \sum_{i=1}^{N} d_i \Phi(v_i) - (\underset{\sim}{f}, \underset{\sim}{v})$$

where $(\bullet, \bullet)$ denotes the usual inner-product of $\mathbb{R}^N$ and $Q(t) = \int_0^t \phi(\tau)d\tau$.

Remark 4.2 : In each specific case ρ has to be determined ; this can be done theoretically, experimentally, or by using an automatic procedure, which will not be described here.

Remark 4.3 : Let us define g^n by

$$\underset{\sim}{g}^n = A\underset{\sim}{u}+D\phi(\underset{\sim}{u}^n)-\underset{\sim}{f}.$$

Instead of using a constant parameter ρ we can use a family $(\rho_n)_n$ of positive parameters in (4.3). Therefore (4.3) can be written as

$$\underset{\sim}{u}^{n+1} = \underset{\sim}{u}^n-\rho_n S^{-1}\underset{\sim}{g}^n. \tag{4.4}$$

Suppose $A=A^t$, then if we use (4.2),(4.4) with ρ_n defined by

$$\begin{cases} J(\underset{\sim}{u}^n-\rho_n S^{-1}\underset{\sim}{g}^n) \leq J(\underset{\sim}{u}^n-\rho S^{-1}\underset{\sim}{g}^n) \quad \forall \rho \in \mathbb{R}, \\ \rho_n \in \mathbb{R} , \end{cases} \tag{4.5}$$

the resulting algorithm is a steepest descent method. This algorithm is convergent for $\phi \in C^0(\mathbb{R})$ (we recall that ϕ is non decreasing in this report). We observe that at each iteration the determination of ρ_n requires the solution of a one-dimensional problem (a "line search") ; for the solution of such one-dimensional problems see POLAK [1], BRENT [1].

Remark 4.4 : At each iteration of (4.2),(4.3) or (4.2),(4.4),(4.5) we have to solve a linear system related to S. Since S is symmetric and positive definite this system can be solved using the Cholesky method, provided the factorization

$$S = LL^t$$

has been done.

From a practical point of view it is obvious that the factorization of S should be made in the beginning once for all. Then at each iteration we just have to solve two triangular systems, which is a trivial operation.

4.4. Newton's method

The Newton's algorithm is given by

(4.6) $u^o \in \mathbb{R}^N$ given,

(4.7) $\underset{\sim}{U}^{n+1} = (A+D\phi'(\underset{\sim}{u}^n))^{-1}(D\phi'(\underset{\sim}{u}^n)\underset{\sim}{u}^n - D\phi(\underset{\sim}{u}^n)+\underset{\sim}{f})$,

where $\phi'(v)$ denotes the diagonal matrix

$$\begin{pmatrix} \phi'(v_1) & & 0 \\ & \ddots & \\ 0 & & \phi'(v_n) \end{pmatrix},$$

with $\phi' = \frac{d\phi}{dx}$.

Since ϕ is non decreasing we have $\phi' \geq 0$, this implies that $A+\phi'(v)$ is positive definite $\forall\, v \in \mathbb{R}^N$.

Remark 4.5 : At each iteration we have to solve a linear system. Since the matrix $A+\phi'(u^n)$ depends on n, the Newton's method above may be not efficient for large N. However a variant of Newton's method avoiding partly that difficulty may be found in EISENSTAT-SCHULTZ-SHERMAN [1]. The idea is to replace the complete solving of (4.7) by a few cycles of an iterative method for solving linear systems ; for more details see the above paper.

Remark 4.6 : The choice of u^o is very important when using Newton's method moreover the convergence requires more regularity for ϕ (let say $\phi \in C^2$) than in most of the methods to be described in the following sections.

4.5. Relaxation and Overrelaxation methods

We use the following notation

$$A = (a_{ij})_{1 \leq i,j \leq N}\ ,\ \underset{\sim}{f} = \{f_1, \ldots f_N\}.$$

Since A is positive definite we have $a_{ii} > 0 \ \forall i=1,2,...N$. Here we shall describe three algorithms :

Algorithm 1 :

(4.8) $\quad \underset{\sim}{u}^0 \in \mathbb{R}^N$ given,

then for u^n known we compute u^{n+1}, component by component, using

$$(4.9) \qquad a_{ii}\,\bar{u}_i^{n+1} + d_i\phi(\bar{u}_i^{n+1}) = f_i - \sum_{j<i} a_{ij}\,u_j^{n+1} - \sum_{j>i} a_{ij}\,u_j^n \ ,$$

$$(4.10) \qquad u_i^{n+1} = u_i^n + \omega(\bar{u}_i^{n+1} - u_i^n)$$

for $i=1,2,...N$.

Since $a_{ii} > 0$, $d_i > 0$, $\phi \in C^0(\mathbb{R})$ and ϕ is a non decreasing function, (4.9) has a unique solution.
If $\omega=1$ we recover an ordinary relaxation method ; in this case it follows from CEA-GLOWINSKI [1] that if $A=A^t$ and since ϕ is C^0 and non decreasing, then the sequence $\{\underset{\sim}{u}^n\}_n$ associated with (4.8)-(4.10) converges to the solution u of (4.1).

If in (4.1), A is not symmetric or $\omega \neq 1$, some sufficient conditions of convergence may be found in ORTEGA-RHEINBOLDT [1] and S. SCHECHTER [1], [2],[3].

Algorithm 2 : This algorithm is the variant of (4.8)-(4.10) obtained by replacing (4.9), (4.10) by

$$(4.11) \qquad \begin{cases} a_{ii}u_i^{n+1} + d_i\phi(u_i^{n+1}) = (1-\omega)(a_{ii}\,u_i^n + d_i\,\phi(u_i^n)) + \\ + \omega(f_i - \sum_{j<i} a_{ij}\,u_j^{n+1} - \sum_{j>i} a_{ij}\,u_j^n) \end{cases}$$

for $i=1,2,...N$.

Remark 4.7 : If $\omega = 1$ or ϕ is linear the two algorithms coincide. In the general case the convergence of (4.8)-(4.11) seems to be an open question. However, from our numerical experiments it seems that the algorithm 2 is more "robust" than algorithm 1 ; may be because it is more implicit. Furthermore it can be used even if ϕ is only defined on a bounded or semi-bounded interval $]\alpha,\beta[$ of R such that $\phi(\alpha) = -\infty$, $\phi(\beta) = +\infty$; in such a case if $\phi \in C^0(]\alpha,\beta[)$ and ϕ is increasing, then (4.1) still has a unique solution but the use of (4.8)-(4.10) with $\omega>1$ may be dangerous.

Remark 4.8 : If $\phi \in C^1(\mathbb{R})$, an efficient method to compute $\bar{u}_i^{n+1}$ in (4.9) and u_1^{n+1} in (4.11) is the one dimensional Newton's method :

Let $g \in C^1(\mathbb{R})$. In this case the Newton's algorithm for solving the equation $g(x) = 0$ is

(4.12) $\quad x^0 \in \mathbb{R}$ given,

(4.13) $$x^{n+1} = x^n - \frac{g(x^n)}{g'(x^n)}$$

If in the computation of $\bar{u}_i^{n+1}$ and u_1^{n+1} we use only one iteration of Newton's method, starting from u_i^n, then the resulting algorithms are identical and we obtain

Algorithm 3 :

(4.14) $\quad u^0 \in \mathbb{R}^N$ given,

then for $n \geq 0$

(4.15) $$u_i^{n+1} = \frac{\left(\sum_{j<i} a_{ij}u_j^{n+1} + \sum_{j>i} a_{ij}u_j^n + d_i\phi(u_i^n) - f_i\right)}{a_{ii} + d_i\phi'(u_i^n)}, \quad i=1,2,\ldots N.$$

Sufficient conditions for the convergence of (4.14),(4.15) are given in S. SCHECHTER [1], [2], [3].

Remark 4.9 : We may find in GLOWINSKI-MARROCCO [1], [2], applications of relaxation methods for solving the non linear elliptic equations modelling the magnetic state of electrical machines.

4.6. Alternating direction methods

In this section we take $\rho>0$. Here we will give two numerical methods for solving (4.1).

First method :

(4.16) $u^0 \in \mathbb{R}^N$ given,

once $\underset{\sim}{u}^n$ is known, we compute $\underset{\sim}{u}^{n+1/2}$ by

(4.17) $\rho\underset{\sim}{u}^{n+1/2}+A\underset{\sim}{u}^{n+1/2} = \rho\underset{\sim}{u}^n-D\phi(\underset{\sim}{u}^n)+\underset{\sim}{f}$,

then u^{n+1} by

(4.18) $\rho u^{n+1} + D\phi(u^{n+1}) = pu^{n+1/2}-Au^{n+1/2}+f$

For the convergence of (4.16)-(4.18) see, e.g., R.B. KELLOG [1].

Second method :

(4.19) $\underset{\sim}{u}^0 \in \mathbb{R}^N$ given,

knowing u^n we compute $u^{n+1/2}$ by

(4.20) $\rho u^{n+1/2} + Au^{n+1/2} = \rho u^n-D\phi(u^n)+f$,

then u^{n+1} by

(4.21) $\rho u^{n+1}+D\phi(u^{n+1}) = \rho u^{n}-Au^{n+1/2}+f$

($\rho \underset{\sim}{u}^{n+1/2}$ in (4.18) has been replaced by ρu^{n}).

Using the results of LIEUTAUD [1] it can be proved that for all $\rho > 0$, $u^{n+1/2}$ and u^{n} converge to u if A, D and ϕ satisfy the hypothesis given in Sec. 4.2.

Remark 4.10 : At each iteration we have to solve a linear system whose matrix is independent of n if we use a constant step ρ. This is an advantage from a computational point of view (see Remark 4.4).
We also have to solve a non linear system of N equations, but in fact these equations -are independent from each other and reduce to N non linear equations in one variable, which can be easily solved.

Remark 4.11 : Variant of (4.16)-(4.18) and (4.19)-(4.21) are obtained by inversion of the order in which we solve the linear and non linear problems. Doing so we obtain from (4.16)-(4.18)

(4.22) $u^{o} \in \mathbb{R}^{N}$ given,

and for $n \geq 0$

(4.23) $\rho \underset{\sim}{u}^{n+1/2}+D\phi(\underset{\sim}{u}^{n+1/2}) = \rho u^{n}-A\underset{\sim}{u}^{n}+\underset{\sim}{f}$,

(4.24) $\rho \underset{\sim}{u}^{n+1}+A\underset{\sim}{u}^{n+1} = \rho \underset{\sim}{u}^{n+1/2}-D\phi(\underset{\sim}{u}^{n+1/2})+\underset{\sim}{f}$.

From (4.19)-(4.21) we obtain

(4.25) $u^{o} \in \mathbb{R}^{N}$ given,

and for $n \geq 0$

(4.26) $$\rho \underset{\sim}{u}^{n+1/2} + D\phi(u^{n+1/2}) = \rho u^n - A\underset{\sim}{u}^n + \underset{\sim}{f},$$

(4.27) $$\rho u^{n+1} + Au^{n+1} = \rho u^n - D\phi(u^{n+1/2}) + f.$$

If (4.16)-(4.18) and (4.22)-(4.24) may be viewed as the same algorithm (with different starting procedures) this is not the case for (4.19)-(4.21) and (4.25)-(4.27) ; indeed for the class of problems under consideration it appears that for the same ρ the convergence of (4.25)-(4.27) is faster than the convergence of (4.19)-(4.21).

4.7. Conjugate gradient methods.

In this section wa assume that $A=A^t$. For a detailled study of conjugate gradient methods we refer, e.g., to POLAK [1], DANIEL [1], CONCUS-GOLUB ill. If the functional J defined in Remark 4.1 (see also (4.28) below) is not quadratic (i.e. if ϕ is non linear), several conjugate gradient methods can be used. Let us describe two of them, the convergence of which is studied in POLAK [1].

Let J given by

(4.28) $$J(v) = \frac{1}{2}(Av,v) + \sum_{i=1}^{N} d_i \Phi(v_i) - (f,v),$$

where $\Phi(t) = \int_0^t \phi(\tau)d\tau$, ϕ being, as above, a non decreasing continuous function on $\mathbb{R}$, with $\phi(0) = 0$. Let S be a $N\times N$ symmetric, positive definite matrix.

First method : (Fletcher-Reeves)

(4.29) $$u^o \in \mathbb{R}^N \text{ given,}$$

(4.30) $$\underset{\sim}{g}^o = S^{-1}(A\underset{\sim}{u}^o + \phi(\underset{\sim}{u}^o) - \underset{\sim}{f}),$$

(4.31) $$w^0 = \underset{\sim}{g}^o.$$

Then, assuming that u^n and w^n are known, we compute u^{n+1} by

$$\underset{\sim}{u}^{n+1} = \underset{\sim}{u}^n - \rho_n \underset{\sim}{w}^n, \tag{4.32}$$

where ρ_n is the solution of the one dimensional minimization problem

$$\begin{cases} J(\underset{\sim}{u}^n - \rho_n \underset{\sim}{w}^n) \leq J(u^n - \rho w^n) \quad \forall \rho \in \mathbb{R}, \\ \rho_n \in \mathbb{R}. \end{cases} \tag{4.33}$$

Then we compute $\underset{\sim}{g}^{n+1}$ and w^{n+1} by,

$$\underset{\sim}{g}^{n+1} = S^{-1}(A\underset{\sim}{u}^{n+1} + \phi(\underset{\sim}{u}^{n+1}) - \underset{\sim}{f}), \tag{4.34}$$

$$\underset{\sim}{w}^{n+1} = \underset{\sim}{g}^{n+1} + \lambda_n \underset{\sim}{w}^n \tag{4.35}$$

where

$$\lambda_n = \frac{(S\underset{\sim}{g}^{n+1}, \underset{\sim}{g}^{n+1})}{(S\underset{\sim}{g}^n, \underset{\sim}{g}^n)}. \tag{4.36}$$

Second method : (Polak-Ribière)

This method is like the previous method except that (4.36) is replaced by

$$\lambda_n = \frac{(S\underset{\sim}{g}^{n+1}, \underset{\sim}{g}^{n+1} - \underset{\sim}{g}^n)}{(S\underset{\sim}{g}^n, \underset{\sim}{g}^n)}. \tag{4.37}$$

Remark 4.12 : For the computation of ρ_n in (4.33)' see Remark 4.3.

Remark 4.13 : It follows from POLAK [1], that if ϕ is sufficiently smooth, then the convergence of the above algorithms is super linear, i.e. faster than the convergence of any geometric sequence.

Remark 4.14 : The above algorithms are fairly sensitive to round off errors ; hence double precision may be required for some problems. Moreover it may be convenient to take periodically $w^n = g^n$ (in this direction see POWELL [1] where more sophisticated restarting procedures are discussed).

Remark 4.15 : We have to solve at each iteration a linear system related to S, Remark 4.4 applied to these algorithms also.

Remark 4.16 : Since the matrix S is symmetric and positive definite, an obvious choice is $S = I_N$, but in some problems it may give a slow convergence. Since A is symmetric and positive definite another obvious choice is S = A.

In BARTELS-DANIEL [1] and DOUGLAS-DUPONT [1] cne may find applications of conjugate gradient methods (very similar to those of this section) to the numerical solution of mildly non linear second order elliptic equations like

$$\begin{cases} -\nabla\cdot(a_o(x)\nabla u) + \phi(u) = f \text{ on } \Omega, \\ u|_\Gamma = g. \end{cases} \tag{4.38}$$

Assuming that (4.38) has been discretized (by finite differences or finite elements) the above authors take for S a discrete analogue of -A ; in the case of finite difference approximations, this choice allows them to use Fast Poisson Solvers. We refer to BARTELS-DANIEL and DOUGLAS-DUPONT, loc. cit., for more details (see also the very recent paper of CONCUS-GOLLJB-O'LEARY L-11).

4.8. Comments

The methods of this Section 4 are fairly classical and may be applied to more general non linear systems than (4.1). They can be applied of course to the solution of the finite dimensional systems obtained by discretization of elliptic problems like

$$\begin{cases} -\nabla\cdot(a_o(x)\nabla u)+\beta\cdot\nabla u+\phi(x,u) = f \text{ in } \Omega, \\ + \text{ suitable boundary conditions,} \end{cases}$$

where, for fixed x, the function $t \to \phi(x,t)$ is continuous and non decreasing on $\mathbb{R}$.

5. - NUMERICAL SOLUTION OF (Ph) BY PENALTY-DUALITY ALGORITHMS

5.1. Formulation of the discrete problem. Orientation.

We use the notation of Section 4. We have seen in Sec. 4.2 that (P_h) reduces to a non linear system like

$$Au + D\phi(u) = f \tag{5.1}$$

where A is a $N \times N$ positive definite matrix, D is a diagonal matrix with positive diagonal elements d_i's and where

$$\underset{\sim}{u} = \{u_1, \ldots u_N\} \in \mathbb{R}^N, \quad \underset{\sim}{f} \in \mathbb{R}^N,$$

$$\phi(u) \in \mathbb{R}^N \text{ with } (\phi(\underset{\sim}{u}))_i = \phi(u_i) .$$

If the bilinear form a(*,*) of Sec. 2.1. is symmetric then A is also symmetric.

Following FORTIN-GLOWINSKI [1] and GLOWINSKI [2, Ch. 5] we shall describe in the following sections two algorithms for solving (5.1). These two algorithms are based on a decomposition-coordination principle, via penalty-duality (they are strongly related to augmented Lagrangian methods ; see Remark 5.2 for motivation). The proof of the convergence of these algorithms are not given here, since they follow from general results which may be found in the two references above.

Numerical applications of these methods to problems like (4.38) and comparisons with other methods are given in Sec. 6.

5.2. Description of the algorithms. Remarks.

5.2.1. A first algorithm.

Let r be a positive parameter.

Let us denote by ALG1 the following algorithm :

(5.2) $\lambda^0 \in \mathbb{R}^N$, arbitrary given,

then for $n \geq 0$ we define $\underset{\sim}{u}^n, \underset{\sim}{p}^n, \underset{\sim}{\lambda}^{n+1}$ by

$$(5.3)\quad \begin{cases} ru^n + D\phi(\underset{\sim}{u}^n) \ni \underset{\sim}{f} + r\underset{\sim}{p}^n - \underset{\sim}{\lambda}^n , \\ (rI+A)\underset{\sim}{p}^n = ru^n + \lambda^n , \end{cases}$$

$$(5.4)\quad \lambda^{n+1} = \lambda^n + \rho(\underset{\sim}{u}^n - \underset{\sim}{p}^n).$$

Remark 5.1 : Looking at (5.2)-(5.4) it appears that the main difficulty when using this algorithm is the solution of the nonlinear system (5.3). Fortunately (5.3) has a very special structure making it very suitable for a solution by block-relaxation (or under or over relaxation) methods. More precisely (5.3) is a particular case of the following nonlinear system in R^{2N}

$$(5.5)\quad \begin{cases} rx + D\phi(x) \ni ry + \underset{\sim}{f}_1 , \\ (rI+A)y = rx + \underset{\sim}{f}_2 . \end{cases}$$

A block relaxation algorithm for solving (5.5) is the following

(5.6) $y^0 \in \mathbb{R}^N$ given,

then for $m \geq 0$ we compute x^{m+1} and y^{m+1} by

(5.7) $rx^{m+1} + D\phi(x^{m+1}) \ni ry^m + \underset{\sim}{f}_1$,

(5.8) $(rI+A)y^{m+1} = rx^{m+1} + \underset{\sim}{f}_2$.

We observe that if y^m is known in (5.7) then the computation of x^{m+1} is easy since it is reduced to the solution of N independent, single variable nonlinear equations of the following type

$$rt + d\phi(t) = b \tag{5.9}$$

with $d > 0$. Since $r > 0$ and ϕ is C^0 and non decreasing, then (5.9) has a unique solution which can be computed by various standard methods (see, e.g. HOUSEHOLDER [1], BRENT [1]).

Similarly if x^{m+1} is known in (5.8), we obtain y^{m+1} by solving a linear system whose matrix is $rI+A$. Since r is fixed it is very convenient in some cases to prefactorize $rI+A$ (by Cholesky or Gauss methods).

If $A=A^t$, then (5.5) is equivalent to

$$\left\{ \begin{array}{l} j(\underset{\sim}{x},\underset{\sim}{y}) \leq j(\underset{\sim}{\xi},\underset{\sim}{\eta}) \quad \forall \{\underset{\sim}{\xi},\underset{\sim}{\eta}\} \in \mathbb{R}^{2N} , \\ \\ \{\underset{\sim}{x},\underset{\sim}{y}\} \in \mathbb{R}^{2N} , \end{array} \right. \tag{5.10}$$

where

$$j(\underset{\sim}{\xi},\underset{\sim}{\eta}) = \frac{1}{2}(A\underset{\sim}{\eta},\underset{\sim}{\eta}) + \sum_{i=1}^{N} d_i \Phi(\xi_i) - (\underset{\sim}{f}_1,\underset{\sim}{\xi}) - (\underset{\sim}{f}_2,\underset{\sim}{\eta}) + \frac{r}{2}\|\underset{\sim}{\xi}-\underset{\sim}{\eta}\|^2 . \tag{5.14}$$

Since j is a C^1 strictly convex function of $\{\underset{\sim}{\xi},\underset{\sim}{\eta}\}$, such that

$$\lim_{(\|\xi\|+\|\eta\|)\to+\infty} j(\underset{\sim}{\xi},\underset{\sim}{\eta}) = +\infty$$

it follows from CEA-GLOWINSKI [1] that the sequence $\{\underset{\sim}{x}^m,\underset{\sim}{y}^m\}$ given by (5.6)-(5.8) converges to the unique solution $\{\underset{\sim}{x},\underset{\sim}{y}\}$ of (5.5) (and (5.10)).

When using (5.6)-(5.8) to solve (5.3) an obvious choice for y^o is p^{n-1}.

Remark 5.2 : We suppose that $A=A^t$. Let us define

$$\mathcal{L}_r : \mathbb{R}^{3N} \to \mathbb{R}$$

by

$$(5.12) \qquad \mathcal{L}_r(\underset{\sim}{v},\underset{\sim}{q},\underset{\sim}{\mu}) = \frac{1}{2}(A\underset{\sim}{q},\underset{\sim}{q}) + \sum_{i=1}^{N} d_i\Phi(v_i) - (\underset{\sim}{f},\underset{\sim}{v}) + \frac{r}{2}\|\underset{\sim}{v}-\underset{\sim}{q}\|^2 + (\underset{\sim}{\mu},\underset{\sim}{v}-\underset{\sim}{q}) .$$

Then $\mathcal{L}_r$ is an augmented lagrangian (see, e.g., HESTENES [1], GABAY-MERCIER [1], FORTIN-GLOWINSKI [1] for more details) related to the minimization problem

$$(5.12) \qquad \underset{\{\underset{\sim}{v},\underset{\sim}{q}\} \in W}{\text{Min}} \{\frac{1}{2}(A\underset{\sim}{q},\underset{\sim}{q}) + \sum_{i=1}^{N} d_i\,\Phi(v_i) - (\underset{\sim}{f},\underset{\sim}{v})$$

where

$$W = \{\{\underset{\sim}{v},\underset{\sim}{q}\} \in \mathbb{R}^{2N} , \ \underset{\sim}{v}-\underset{\sim}{q} = 0\} .$$

The minimization problem (5.12) is obviously equivalent to

$$\underset{v\in\mathbb{R}^N}{\text{Min}} \{\frac{1}{2}(A\underset{\sim}{v},\underset{\sim}{v}) + \sum_{i=1}^{N} d_i\,\Phi(v_i) - (\underset{\sim}{f},\underset{\sim}{v})\}$$

i.e. to

$$Au + D\phi(u) = f,$$

which is the nonlinear system (5.1) under consideration.

One may easily prove that if $\underset{\sim}{u}$ is the solution of (5.1), then $\{\underset{\sim}{u},\underset{\sim}{u},A\underset{\sim}{u}\}$ is $\forall\ r \geq 0$, the unique saddle-point of $\mathcal{L}_r$ over R^{3N}.

From these properties it appears that the algorithm (5.2)-(5.4) (ALG1) may be interpreted as an Uzawa algorithm (see GLOWINSKI-LIONS-TREMOLIERES [1, Ch. 2], EKELAND-TEMAM [1]) for computing the above saddle-point of $\mathcal{L}_r$. Moreover $\lambda = Au$ appears as a Lagrange multiplier related to the linear constraints $\underset{\sim}{v}-\underset{\sim}{q} = 0$.

5.2.2. A second algorithm.

With r as in Sec. 5.2.1., let us denote by ALG2 the following algorithm

(5.13) $\{\underset{\sim}{p}^0, \underset{\sim}{\lambda}^1\} \in \mathbf{R}^{2N}$ given,

then for $n \geq 1$ we define $\underset{\sim}{u}^n, \underset{\sim}{p}^n, \underset{\sim}{\lambda}^{n+1}$ by

(5.14) $ru^n + D\phi(u^n) = f + rp^{n-1} - \lambda^n$,

(5.15) $(rI+A)p^n = ru^n + \lambda^n$,

(5.16) $\lambda^{n+1} = \lambda^n + \rho(\underset{\sim}{u}^n - \underset{\sim}{p}^n)$.

Remark 5.3 : Assume that in ALG1 we use the block-relaxation algorithm (5.6)-(5.8) to solve (5.3). Then if we use $\underset{\sim}{y}^0 = p^{n-1}$ as a starting vector, and if we only do one iteration of (5.6)-(5.8), then ALG1 reduces to ALG2.

Remark 5.4 : Suppose that $\rho = r$ in ALG2 ; we have then

$$(5.17)\quad \left|\begin{array}{l} ru^n + D\phi(u^n) = f + rp^{n-1} - \lambda^n, \\ rp^n + Ap^n = ru^n + \lambda^n, \\ \lambda^{n+1} = \lambda^n + r(\underset{\sim}{u}^n - \underset{\sim}{p}^n). \end{array}\right.$$

It follows from (5.17) that

(5.18)' $\lambda^{n+1} = Ap^n$.

Then from (5.17), (5.18) we obtain

(5.19) $ru^n + D\phi(u^n) + Ap^{n-1} = f + rp^{n-1}$,

(5.20) $rp^n + Ap^n + D\phi(u^n) = f + rp^{n-1}$.

Therefore, if $\rho = r$, ALG2 reduces (with different notation) to the alternating direction method described in Sec. 4.6 by (4.25)-(4.27).

5.3. Convergence of ALG1, ALG2. Further Remarks

5.3.1. Convergence results.

It follows from FORTIN-GLOWINSKI [1], GLOWINSKI [2, Ch. 5] that the properties of A, D, ϕ imply that we have convergence of $\{\underset{\sim}{u}^n, \underset{\sim}{p}^n, \underset{\sim}{\lambda}^n\}$ to $\{\underset{\sim}{u}, \underset{\sim}{u}, \underset{\sim}{Au}\}$ if in ALG1 (resp. ALG2) we take

(5.21) $\quad 0 < \rho < 2r$

(resp.

(5.22) $\quad 0 < \rho < \frac{1+\sqrt{5}}{2} r$).

5.3.2. On the choice of ρ and r.

If r is given our computational experiments with ALG1 and ALG2 seem to indicate that the best choice for ρ is $\rho = r$. The choice of r is not clear and ALG2 appears to be more sensitive to the choice of r than ALG1. In fact ALG1 seems to be more robust on very stiff problems than ALG2. We mean that the choice of the parameter r is less critical and that the computational time with ALG1 may become much shorter than with ALG2 for a given problem.

Remark 5.5 : (On the choice of r in ALGI).

About the choice of r in ALG1 it can be proved that theoretically the largest is r, the fastest is the convergence ; practically the situation is not so simple for the following reasons : the largest is r, the worse is the conditioning of the problem (5.3). Then since (5.3) is numerically (and not exactly) solved at each iteration, an error is done in the calculation of $\{\underset{\sim}{u}^n, \underset{\sim}{p}^n\}$. The analysis of this error and the effect ot it on the gloval behaviour of ALG1 is a very complicated problem since we have to take into account the conditioning of (5.3), the stop-- ping test of the algorithms solving (5.3), round-off errors, etc...
Fortunately it seems that the combining effect of all these factors is to give an algorithm which is not very sensitive to the choice of r.

6. - NUMERICAL EXPERIMENTS AND COMPARISONS WITH OTHER METHODS.

6.1. The test problem.

We consider the following test problem

$$(6.1)\qquad \begin{cases} -\Delta u + \phi(u) = f \text{ on } \Omega, \\ u|_{\partial\Omega} = 0 , \end{cases}$$

where $\Omega =]0,1[\times]0,1[$,

$$\phi(t) = \operatorname{sgn}(t)\,|t|^{\ell} = t|t|^{\ell-1} , \quad \ell > 0 .$$

If (with $x = \{x_1, x_2\}$) we define u by

$$u(x) = \sin 2\pi x_1 \sin 2\pi x_2 ,$$

then for f given by

$$f = 8\pi^2 u + |u|^{\ell-1} u$$

the exact solution of (6.1) is u.

The behaviour of ϕ is shown on fig. 6.1

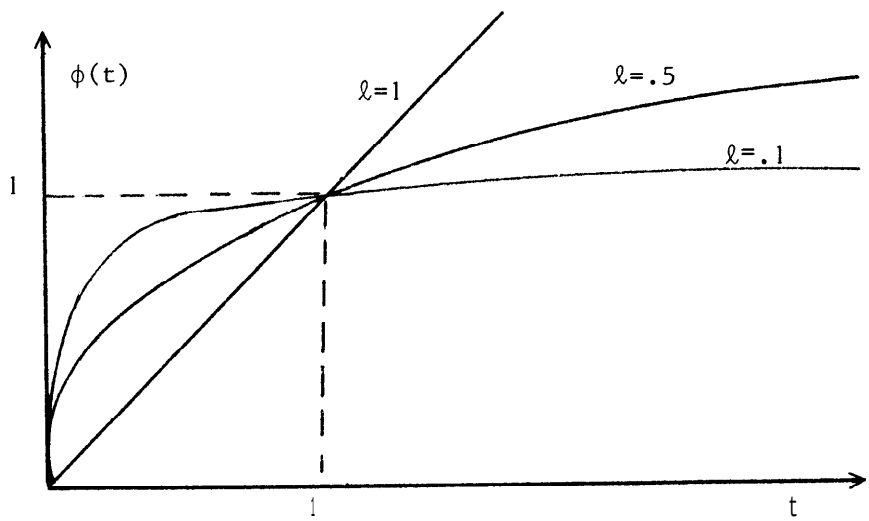


Figure 6.1.

We observe that ϕ is not smooth near t=0 if $0 < \ell < 1$; hence Newton-type methods are not very suitable for this kind of ϕ's.

If we discretize Ω using an uniform square grid with equal grid spacing in both the x_1 and x_2 directions, our matrix A, in ALG1 and ALG2, will be the usual discrete laplacian matrix (since Ω is a square, a finite difference approximation of (6.1) is very convenient). So in both our algorithms ALG1 and ALG2 we have to solve the discrete Helmhotz's equation, namely the discrete formulation of

(6.2) $\quad -Au + ru = f.$

There existfastdirectsolvers of Helmhotz's equation on a uniform mesh in a rectangular domain. We used such a solver called TBPSDN, written by B.L. Buzbee (cf. BUZBEE-GOLUB-NIELSON [1]) at Los Alamos Scientific Laboratory, and tested and modified for Lawrence Berkeley Laboratory by Gary A. Sod. This solver incorporates the truncated Buneman's algorithm, using the standard five point difference approximation for the laplacian.

We have seen in Sec. 5 that each iteration of ALG1 and ALG2 requires the solution of one-dimensional nonlinear equations of the form

(6.3) $\quad r\xi + d\phi(\xi) = RHS$

with $d > 0$ (since we are using finite differences we have in fact d=1).

We do not want to use Newton's method to solve this equation because :

(i) If $\phi \notin C^1$ we may have troubles with Newton's method,

(ii) We think that an efficient method not using ϕ' may be more interesting in view of more general problems.

There exists one-dimensional nonlinear equation solvers which do not require derivatives. We used such a routine, called ZEROIN and due to Richard Brent. This method is described in BRENT [1]. ZEROIN will always locate a root within a given interval where it is known to lie, to within a given accuracy TOL.

From the facts that $\phi(0) = 0$ and that ϕ is non-decreasing, we can easily deduce that the solution ξ of (6.3) is in the interval $[0, \frac{RHS}{r}]$ if RHS > 0 and in the interval $[\frac{RHS}{r}, 0]$ if RHS < 0.

For the inner loop (5.7),(5.8) convergence test, in ALG1, we have used the ℓ_1 norm (the actual norm used is not important ; we have also used the ℓ_2 norm and obtained similar results).

In our experiments, for the purpose of comparisons, we stop our iterations when $\| \underset{\sim}{p}^n - \underset{\sim}{u}_h \|_2 \leq ACC$ where $\underset{\sim}{u}_h$ is the exact solution of the discretized system

$$A\underset{\sim}{u}_h + \phi(\underset{\sim}{u}_h) = f \text{ (here } D=I),$$

with a uniform spacing h (we recall that if ρ is well chosen, cf. Sec. 5.3.1., then $\{\underset{\sim}{u}^n, \underset{\sim}{p}^n, \underset{\sim}{\lambda}^n\}$ converges to $\{\underset{\sim}{u}_h, \underset{\sim}{u}_h, A\underset{\sim}{u}_h\}$).

In practice, $\underset{\sim}{u}_h$ is not known and so some other kind of stopping criteria has to be used, e.g.

$$\frac{\| \underset{\sim}{p}^{n+1} - \underset{\sim}{p}^n \|}{\| \underset{\sim}{p}^{n+1} \|} \leq ACC$$

in some suitable norm.

Remark 6.1 : We can determine u_h with a very good precision by running ALG1 or ALG2 on the test problem until $(\underset{\sim}{u}^n - \underset{\sim}{p}^n)$ is very, very small. Notice that if $\underset{\sim}{u}^n = \underset{\sim}{p}^n$ then $A\underset{\sim}{p}^n + \phi(\underset{\sim}{u}^n) = \underset{\sim}{f}$ and hence $\underset{\sim}{u}_h = \underset{\sim}{u}^n = \underset{\sim}{p}^n$. Incidentally the closeness of $\underset{\sim}{u}^n$ to $\underset{\sim}{p}^n$ can be used as another stopping criteria or as a check on the final iterate $\underset{\sim}{p}^{\tilde{n}}$.

6.2. Study of parameters in ALG1 and ALG2.

We would like to study the effect on the general performance of the algorithms ALG1 and ALG2, of the following parameters :

ALG1 : $\{\hat{u}^o, \hat{p}^o, \lambda^o, r, \rho, \varepsilon, \text{TOL}\}$;

ε is the tolerance parameter for the stopping test of the inner loops of ALG1. The parameters $\lambda^o, r, \rho, \text{TOL}$ have been defined before. The vectors $\hat{u}^o, \hat{p}^o$ deserve some more explanations. It follows from Sec. 5.2.1, Remark 5.1, that the non-linear system (5.3) may be solved by the block-relaxation algorithm (5.6)-(5.8) ; we have used precisely that last algorithm for solving the examples of Sec. 6, taking u^{n-1}, p^{n-1} as starting vectors to compute u^n, p^n. Therefore to compute u^o, p^o from λ^o, we need some starting vectors which are precisely what we have denoted by $\hat{u}^o, \hat{p}^o$ above. Since u^n and p^n converge to the same limit we have systematically taken $\hat{u}^o = \hat{p}^o$.

ALG2 : $\{\hat{u}^o, p^o, \lambda^1, r, \rho, \text{TOL}\}$;

We recall that in ALG2, p^o and λ^1 are given (see Sec. 5.2.2.). Since u^o is computed, from p^o and λ^1, by an-iterative method we need an initial guess say $\hat{u}^o$. In fact we have systematically taken $\hat{u}^o = p^o$.

In addition we want also to study, to some extent, the effects of the smoothness of ϕ on the algorithms. This smoothness can be controlled by ℓ, since

$$\phi(t) = t|t|^{\ell-1} \quad , \ \ell > 0 \ .$$

6.2.1. Effects of $\hat{u}^o, p^o$.

Basically ALG1 (resp. ALG2) will converge for any starting vector $\hat{u}^o (= \hat{p}^o)$ (resp. $p^o (= \hat{u}^o)$). Obviously when an approximation of the solution u_h of

$$(6.4) \qquad A u_h + \phi(u_h) = f$$

is known we should use it. But most often (i.e. for general f) we do not know what the solution is like. So we are often forced to start with some constant value like $\hat{u}^o = 0$ (resp. $p^o = 0$) or $\hat{u}^o$ (resp. p^o) with constant components.

Intuitively if ϕ has a sharp jump at t^* (in our case $t^* = 0$ if $\ell \in]0,1[$ since $\phi'(0) = +\infty$) we would expect that the points of the grid where u_h is near t^*

will produce the slowest convergence for the corresponding component of $\underset{\sim}{u}_h$. This was observed in our experiments. For our test problem, ϕ has a sharp jump at $t^* = 0$ (if $\ell \in]0,1[$) ; if we start with $\hat{\theta}_1^0 = 10$, $\forall i=1,\ldots N$, (resp. $p_1^0 = 10$, $\forall i=1,\ldots N$), the convergence is generally fast except for points where $\underset{\sim}{u}_h$ is close to zero. In all cases the maximum final errors occured at such points. However if we start with $\underset{\sim}{\hat{u}}^0 = \underset{\sim}{0}$ (resp. $p^0 = 0$) no difficulties were observed with these points. Our guess is that is we start with 0 we are starting with a good guess of the components of $\underset{\sim}{u}_h$ at which ϕ has a sharp jump and which we expect slow convergence. Very often we know where ϕ has a sharp jump and we can take advantage of this knowledge (at least if ϕ is not too complicated).

Therefore we can in general recommend the following :

(i) If ϕ is known to have a sharp jump at t^*, and if we don't have a good approximation of $\underset{\sim}{u}_h$ to start with, use $\underset{\sim}{\hat{u}}^0 = \underset{\sim}{t}^*$ (resp. $p^0 = t^*$) (i.e. set $\hat{\theta}_1^0 = t^* \ \forall i=1,\ldots N$, idem for p^0).

(ii) Otherwise, start with the best approximation available.

For example, Figure 6.2 shows that ALG2 (i.e. with no inner loop c-test) works just fine with $p^0 = 0$ but has problem if started with $\underset{\sim}{p}^0 = \underset{\sim}{10}$. Luckily, as we shall see later, ALG1 (with inner-loop E-test) will overcome this trouble.

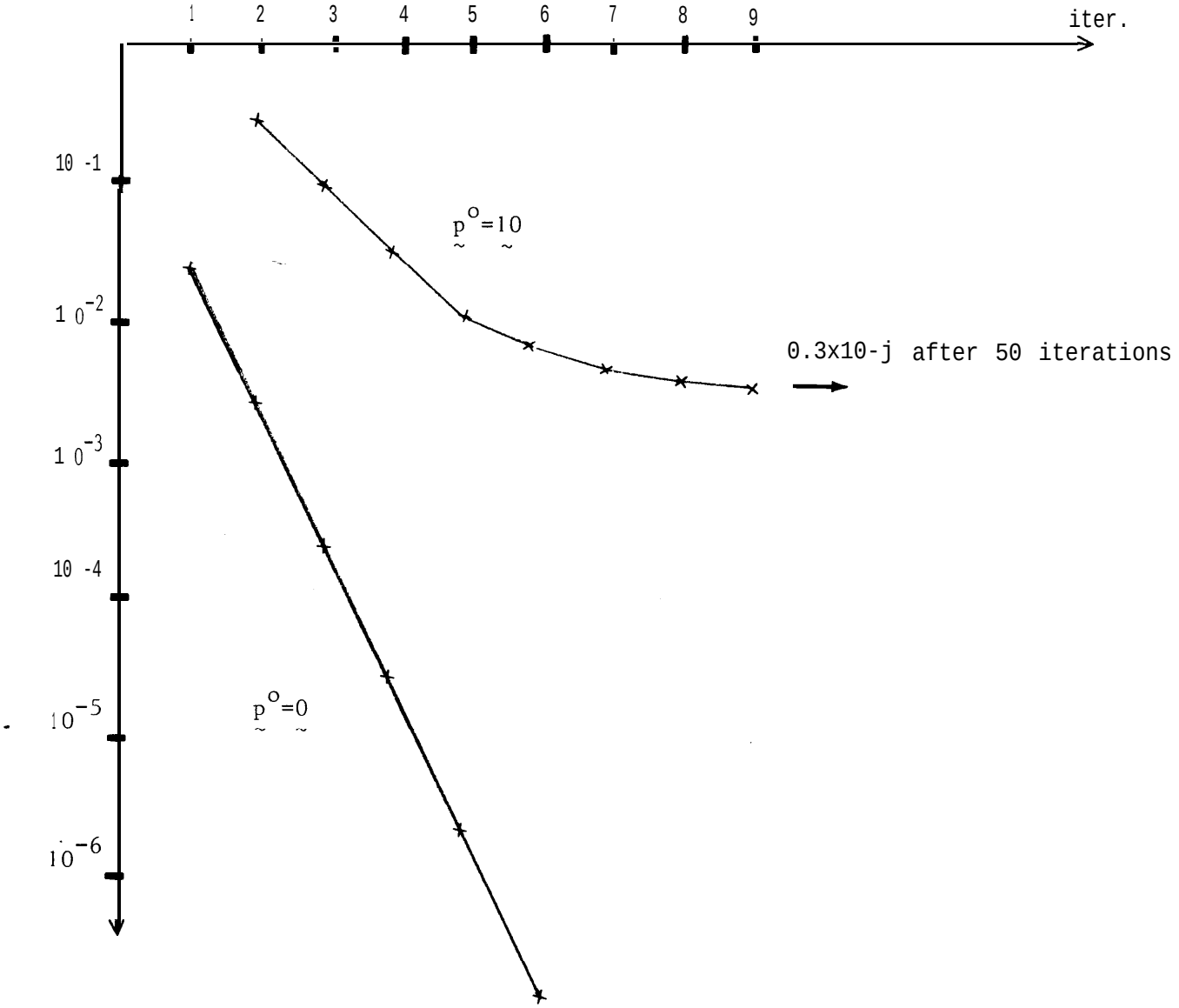
Form Figure 6.2 it appears that the convergence is linear if $p^0 = 0$ and sub-linear if $p^0 = 10$.

Effect of starting vector p^o

ALG2 : 17 ×17 grid , ℓ = .l , ACC = 10^{-6}

$r = \rho = 5$, $\lambda^o = 0$.

L_2 error of p^n



Difficulty : $\phi(t)$ has a sharp jump at t=0.

Figure 6.2

6.2.2. Effects of λ^o

Note that with the exact solution $\underset{\sim}{u}_h$ we have

$$\lambda = A\underset{\sim}{u}_h = f - \phi(\underset{\sim}{u}_h).$$

So two natural choices for λ^o are

(i) $A\hat{u}^o$ (resp. $A\, p^o$) ,

(ii) $f - \phi(\hat{u}^o)$ (resp. $f - \phi(p^o)$) .

If $\hat{u}^o$ (resp. p^o) is a good approximation to $\underset{\sim}{u}_h$, then $\underset{\sim}{\lambda}^o$ will be a good approximation to λ.

For our test problem, if $\hat{u}^o = 0$ (resp. $p^o = 0$), then the two choices for λ^o are 0 and f. We have tried both and conclude that the convergence of ALG1 and ALG2 was quite insensitive to these choices of λ^o.

6.2.3. Choice of ρ.

We actually found that the best choice for ρ is ρ=r for our test problem. Similar observations have been done by GLOWINSKI-MARROCCO [3], GABAY-MERCIER [1], for algorithms like ALG1 , ALG2, applied to the solution of other classes of nonlinear problems.

6.2.4. Choice of TOL

TOL measures how accurately we want to solve, with ZEROIN, the one variable nonlinear equations, obtained from ALG2 and the inner loops of ALG1. From our experiences we recommend a value of TOL = ACC. Intuitively this makes sense because if TOL > ACC we won't be able to obtain the required accuracy in the final solution because our intermediate steps are not solved accurately enough. If TOL << ACC, we spend more work in each inner loop than it is necessary and from our experience, this doesn't improve the convergence of the algorithms.

6.2.5. Choice of ε.

This parameter is used in the inner loop of ALG1 to decide when to stop the inner iterations and update λ^n. Note that if $\varepsilon \to +\infty$, ALG1 $\to$ ALG2 because we will be updating λ^n after only one iteration every time. In general, ALG1 (with a reasonably small ε) is more robust than ALG2. ALG2 will work a little better (takes less iterations) than ALG1 if we start with a "good" guess at the solution ($p^o = 0$ in our test problem) but if we start with a "bad" guess ($p^o = 0$, e.g.) ALG2 will have problem at points where $\tilde{\ }$ has a sharp jump, as explained earlier. In fact ALG1 solved the inner equations more accurately and thus its updating of λ^n will be more accurate and this is often enough to bring us very close to the solution we want. (See Figure 6.3. For the test problem, $\varepsilon = 10^{-4}$ seems to be a good choice.) In other words, ALG1's 'cautiousness" in updating λ^n pays off. ALG1 may lose a little bit in the early itarations by spending" too much time in the inner loop but it gives a better chance of obtaining a solution to within the required accuracy. Therefore, in general, ALG1 is to be recommended. We think that some reasonably small value for ε, like $\varepsilon = \sqrt{ACC}$, will work fine. Another approach is to use variable ε i.e. a sequence $\{\varepsilon_n\}_n$; this requires further investigation.

6.2.6. Choice of r .

We complete the Remark 5.5 of Sec. 5.3.2.. The parameter r controls the relative weight of the penalty term in the augmented lagrangian (see (5.10)). This penalty term has the effect of providing some global convergence steering. We varied the value of r in ALG1 with the test problem and found that the convergence is surprisingly insensitive to r (see Fig. 6.4). This is an advantage over SOR and AD1 type methods which are very sensitive to their parameters (as will be shown later).

Effect of ε

ACC = 10^{-6} , ℓ = .1 , 17 × 17 grid,

r = ρ = 5 , λ^o = 0

N.B. : For ALG1 we count the inner iterations.

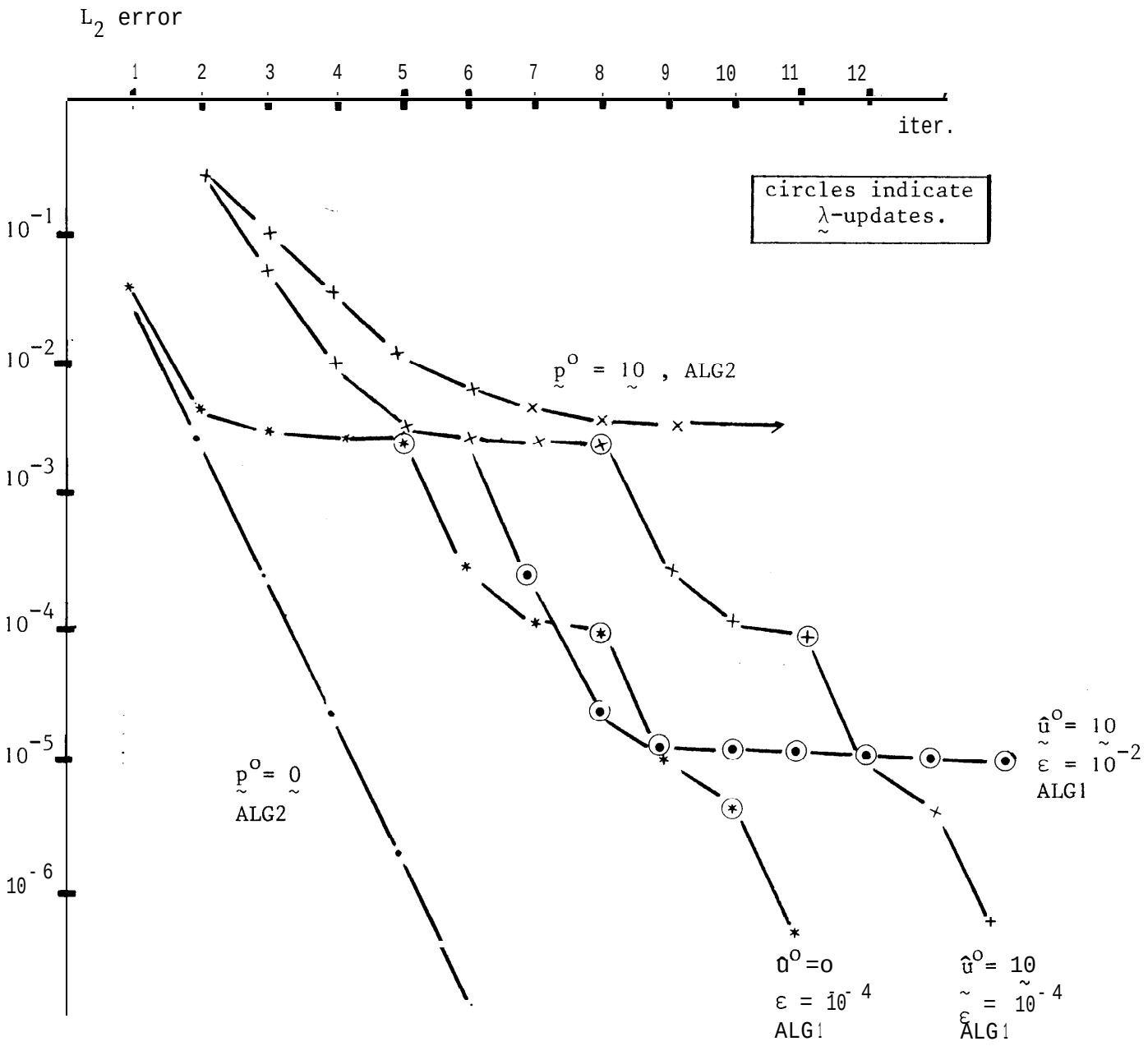


Fig. 6.3

6.2.7. Effect of the smoothness of ϕ.

The smoothness of ϕ can be controlled by ℓ in our test problem. We ran ALG1 with ℓ=.1 and ℓ=.5 and found that ALG1 actually performs a little bit better for ℓ=.1 than for the "smoother" ℓ=.5. We also determined the optimal r for ℓ=.5 and found it to be about the same as that for ℓ=.1.

6.3. Comparisons with other methods.

6.3.1. Description of the other methods.

For a given accuracy ACC on $\|\underset{\sim}{p}^n - \underset{\sim}{u}_h\|_2$ we want to compare the efficiencies of ALG1, ALG2 and other methods for solving the discretized problem

$$A\underset{\sim}{u}_h + \phi(\underset{\sim}{u}_h) = \underset{\sim}{f} \ .$$

Among these methods compared are the Successive over-relaxation method (SOR) and the alternating direction implicit (ADI) methods discussed in Sec. 4.6.

These methods are reproduced below :

1) SOR : We can look at the discretized equation

$$A\underset{\sim}{u}_h + \phi(\underset{\sim}{u}_h) = \underset{\sim}{f}$$

as a system of non linear equations

$$\begin{aligned} &f_1(u_1,u_2 \ldots u_N) = 0 \ , \\ &\vdots \\ &f_i(u_1,u_2,\ldots u_N) = 0, \\ &\vdots \\ &f_N(u_1,u_2,\ldots u_N) = 0 \ . \end{aligned} \qquad (6.5)$$

Then we can use (cf. Sec. 4.5) the two following variants of SOR :

SOR 1 :

(6.6) $\underset{\sim}{u}^{o}$ given

at step n, with u^n known, we compute $\underset{\sim}{u}^{n+1}$ by :

For i from 1 to N, solve

(6.7) $$f_i(u_1^{n+1},\ldots u_{i-1}^{n+1},U_i^{n+1/2},u_{i+1}^{n}\ldots) = 0 ,$$

then

(6.8) $$U_i^{n+1} = u_i^n + \omega(u_i^{n+1/2}-u_i^n) .$$

SOR 2 :

(6.9) $\underset{\sim}{u}^{o}$ given,

at step n, with u^n known, compute u^{n+1} by :

For i from 1 to N, solve

(6.10) $$f_i(u_1^{n+1},\ldots u_i^{n+1},u_{i+1}^{n}\ldots) = (1-\omega)f_i(u_1^{n+1}\ldots u_{i-1}^{n+1},u_i^n,\ldots).$$

2) ADI : We consider the following variants of ADI

ADI 1 : We iterate on the following

(6.11) u^o given,

then for $n \geq 0$

(6.12) $$(\rho I+A)u^{n+1/2} = f + \rho u^n-\phi(u^n) ,$$

(6.13) $$\rho u^{n+1} + \phi(u^{n+1}) = f + \rho u^n - Au^{n+1/2}$$

ADIlM : We replace (6.13) by

$$\rho u^{n+1} + \phi(u^{n+1}) = f + \rho u^{n+1/2} - Au^{n+1/2} \tag{6.14}$$

ADI2 : It is defined by

$$u^{o} \text{ given} \tag{6.15}$$

then for $n \geq 0$

$$U^{n+1/2} + \phi(u^{n+1/2}) = f + \rho u^{n} - Au^{n} , \tag{6.16}$$

$$\rho u^{n+1} + Au^{n+1} = f + \rho u^{n} - \phi(u^{n+1/2}) . \tag{6.17}$$

ADI2M : We replace (6.17) by

$$\rho u^{n+1} + Au^{n+1} = f + \rho u^{n+1/2} - \phi(u^{n+1/2}) . \tag{6.18}$$

6.3.2. Comments. Further Remarks.

One of the main problem with SOR and ADI is the sensitivity to the parameters ω and ρ respectively. Hence we first study the convergence of SOR and AD1 as a function of their respective parameter ω and ρ. See Fig. 6.5, 6.6, 6.7, 6.8.

Remark 6.2 : ADI1 (and ADIlM) both didn't work well and their plots are left out. The difficulty may be due to solving the linear part first instead of the nonlinear part first.

Remark 6.3 : From these plots we can see that both ADI2M and SOR are quite sensitive to their parameters whereas ALG1 and ALG2 are not (specially ALG1). For linear problems one can usually find some good estimates for the optimal parameters. However, for nonlinear problems this is often difficult.

Remark 6.4 : It follows from Remark 5.3 that ALG2 and AD12 are in fact the same algorithm. This appears clearly in Table 6.1 which summarize some of our computational experiments. From this table we can see that ALG1 performs the best if U^0 is not close to the solution u_h.

6.4. CONCLUSION.

From our experiments on the test problem, we can make the following empirical statements :

(i) The convergences of ALG1 and ALG2 are not very sensitive to their parameters, in particular the penalty parameter r.

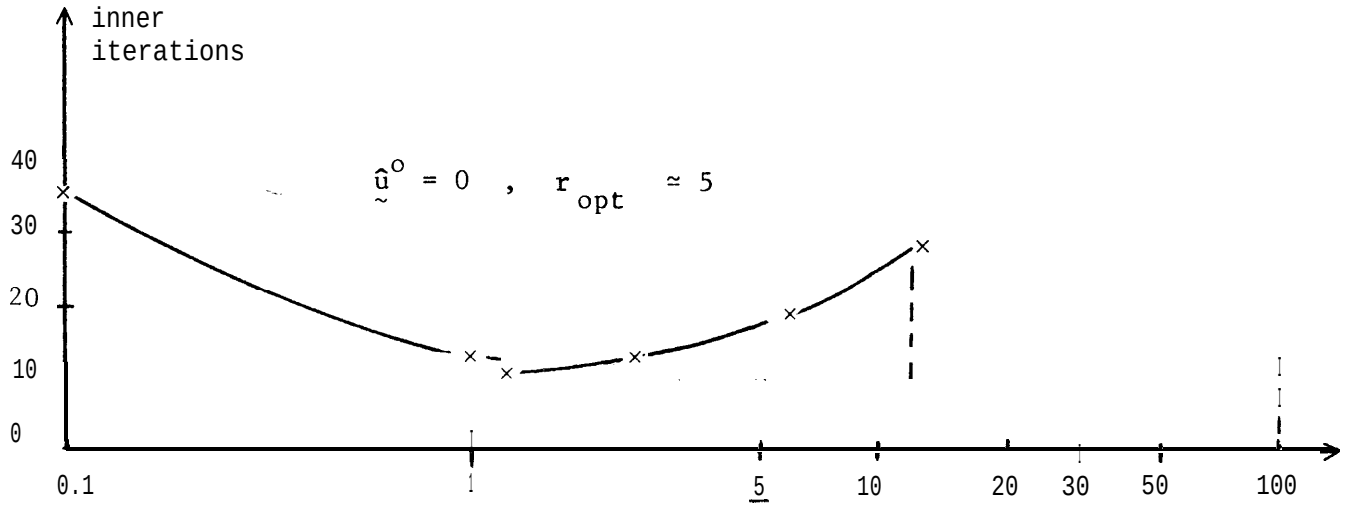
(ii) ALG1 is more robust and as efficient as ALG2 in general.

(iii) ALG1 is more efficient than SOR and ADI for functions ϕ that are not smooth.

Effect of r (=ρ)

$\ell = .1$, 17×17 grid, ACC = 10^{-6} , $\underset{\sim}{\lambda}^{o} = \underset{\sim}{0}$

with $\varepsilon = 10^{-4}$



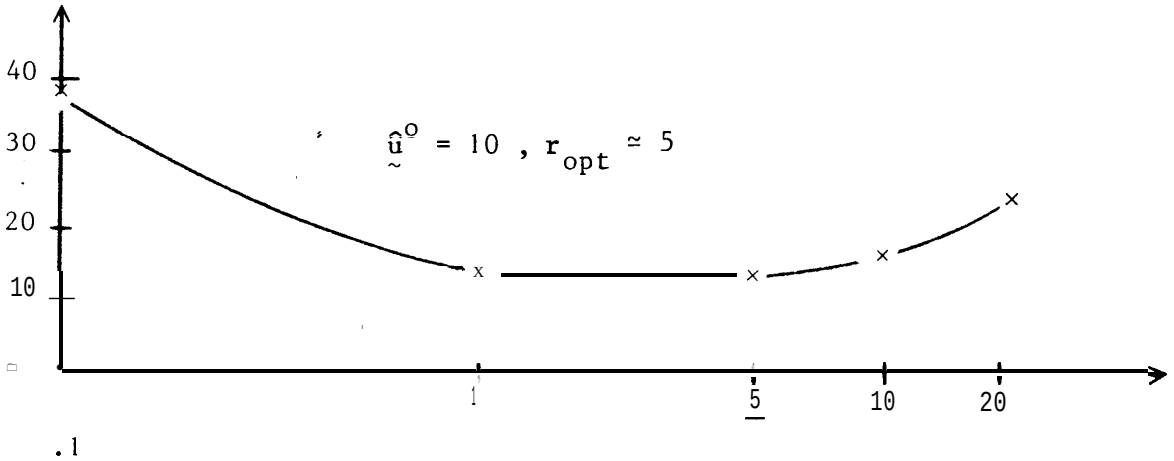
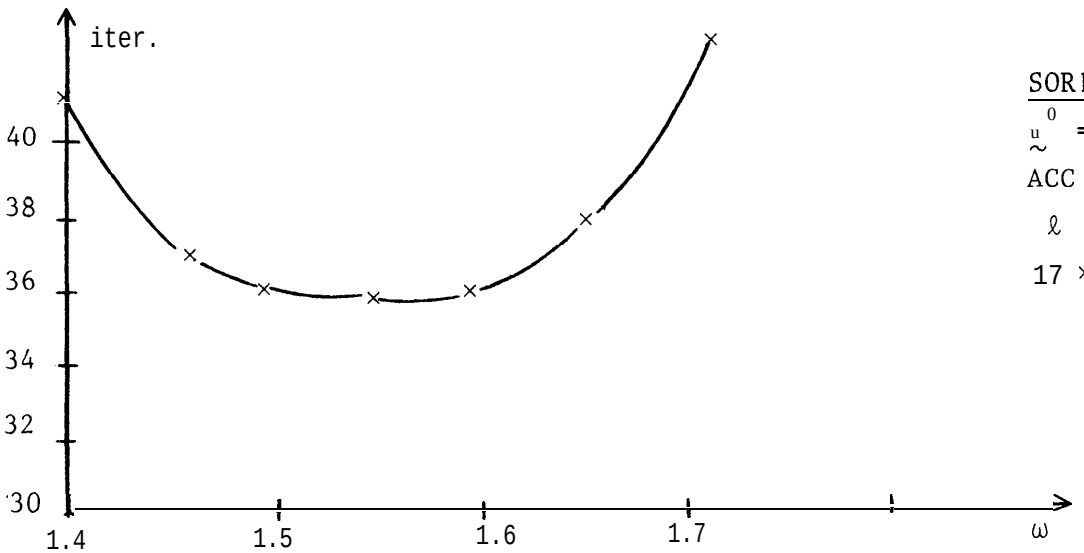


Fig. 6.4

SOR



SOR1

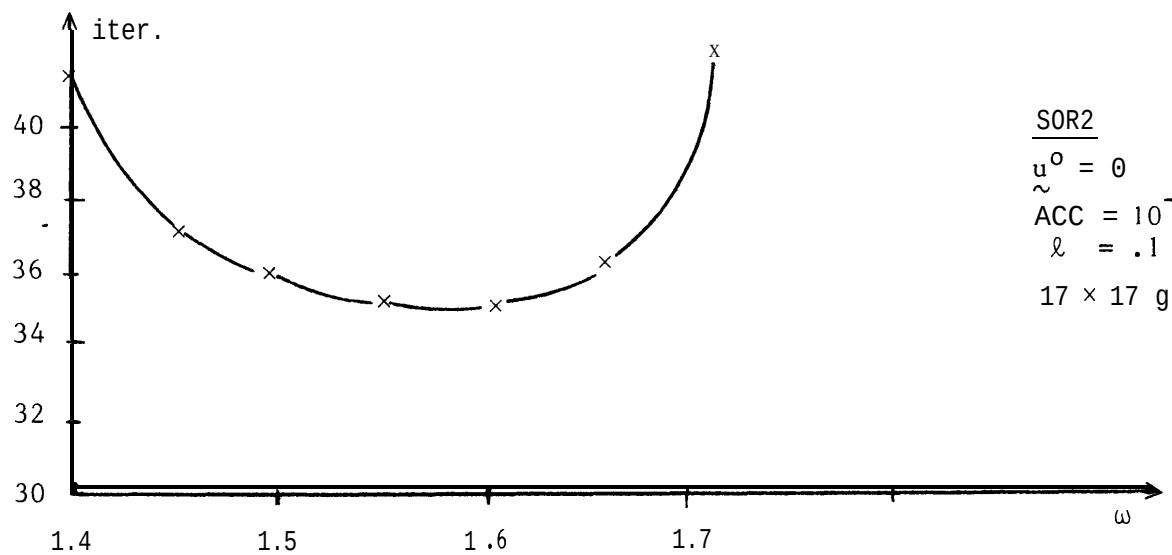
$\underset{\sim}{u}^0 = \underset{\sim}{0}$

ACC = 10^{-6}

$\ell = .1$

17 × 17 grid points

$\omega_{opt} \simeq 1.57$



SOR2

$u^o = 0$

ACC = 10^{-6}

$\ell = .1$

17 × 17 grid points

$\omega_{opt} \simeq 1.59$

Fig. 6.5

AD1

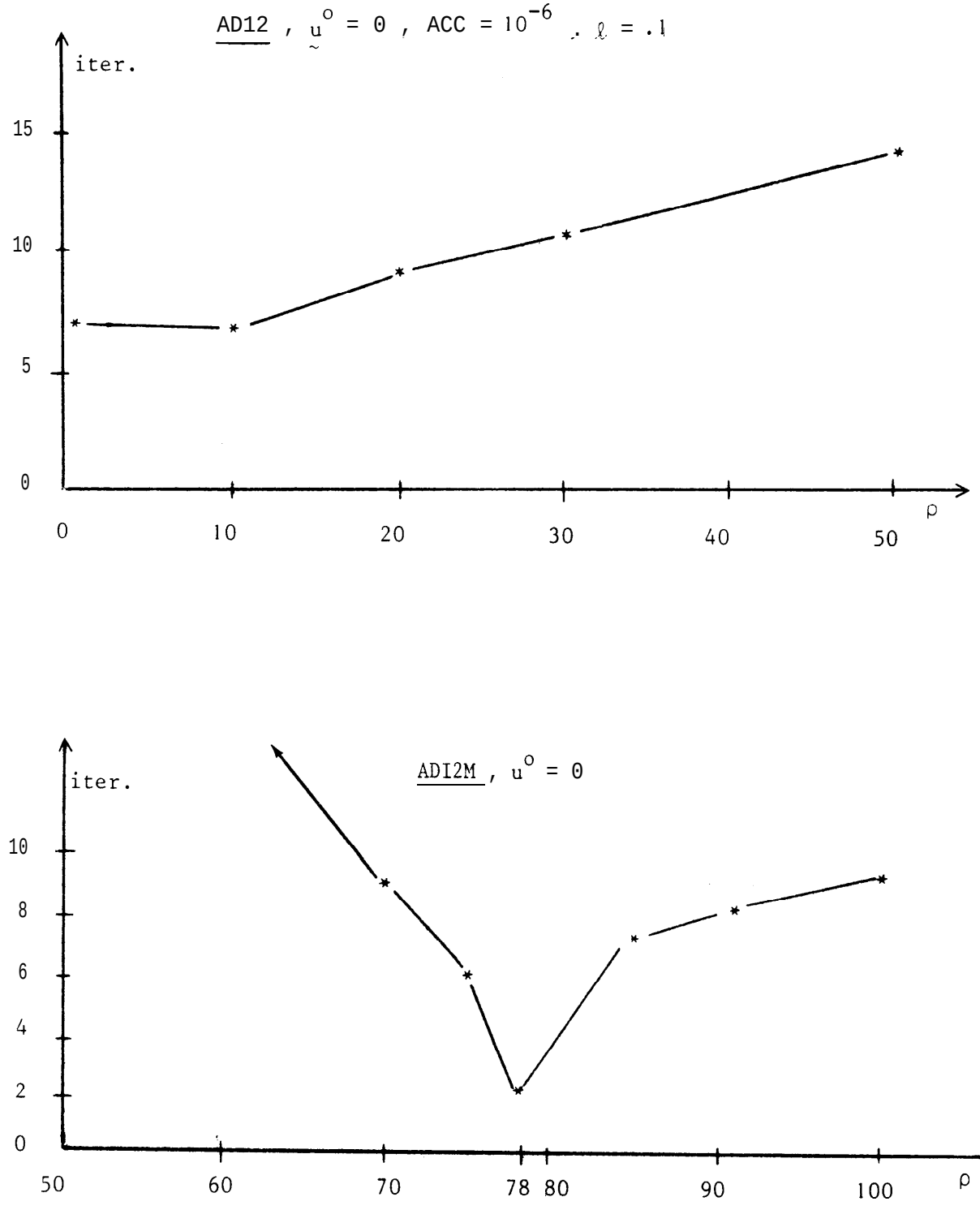
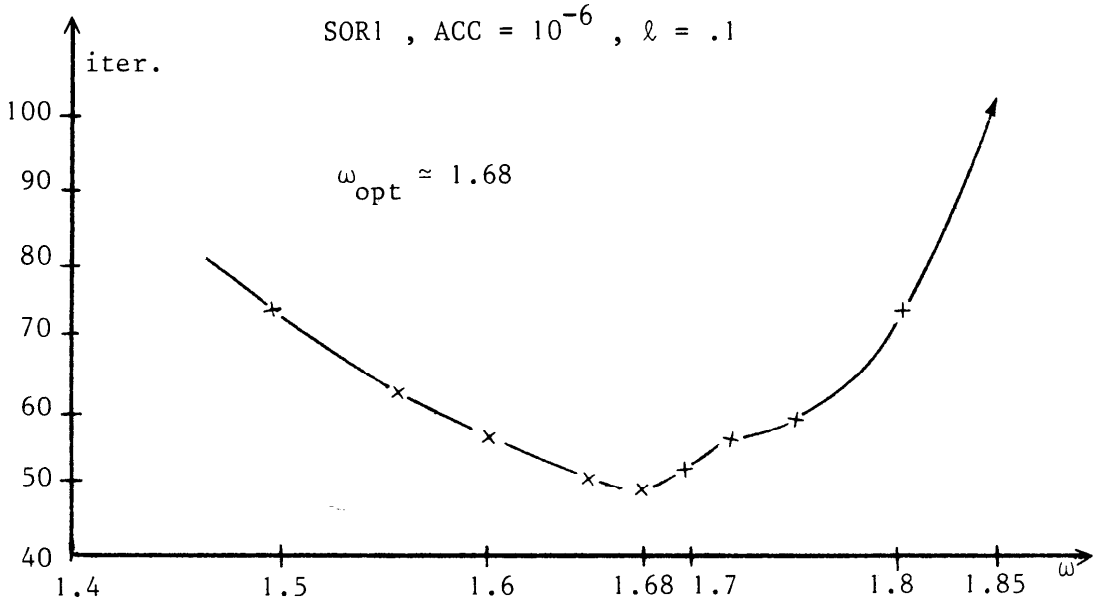

AD12 , $u^o = 0$, ACC = 10^{-6} , $\ell = .1$
iter.
15
10
5
0
0
10
20
30
40
50
ρ
ADI2M , $u^o = 0$
iter.
10
8
6
4
2
0
50
60
70
78
80
90
100
ρ

Fig. 6.6

Optimal Parameters (SOR), $\underset{\sim}{u}^{o} = \underset{\sim}{10}$



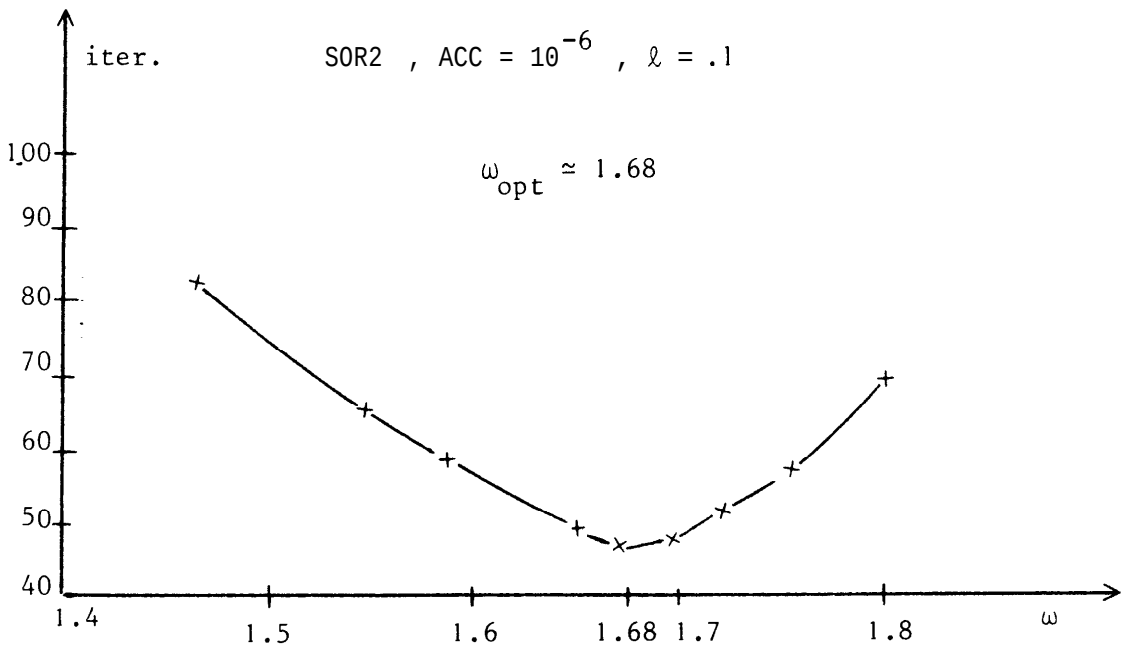


Fig. 6.7

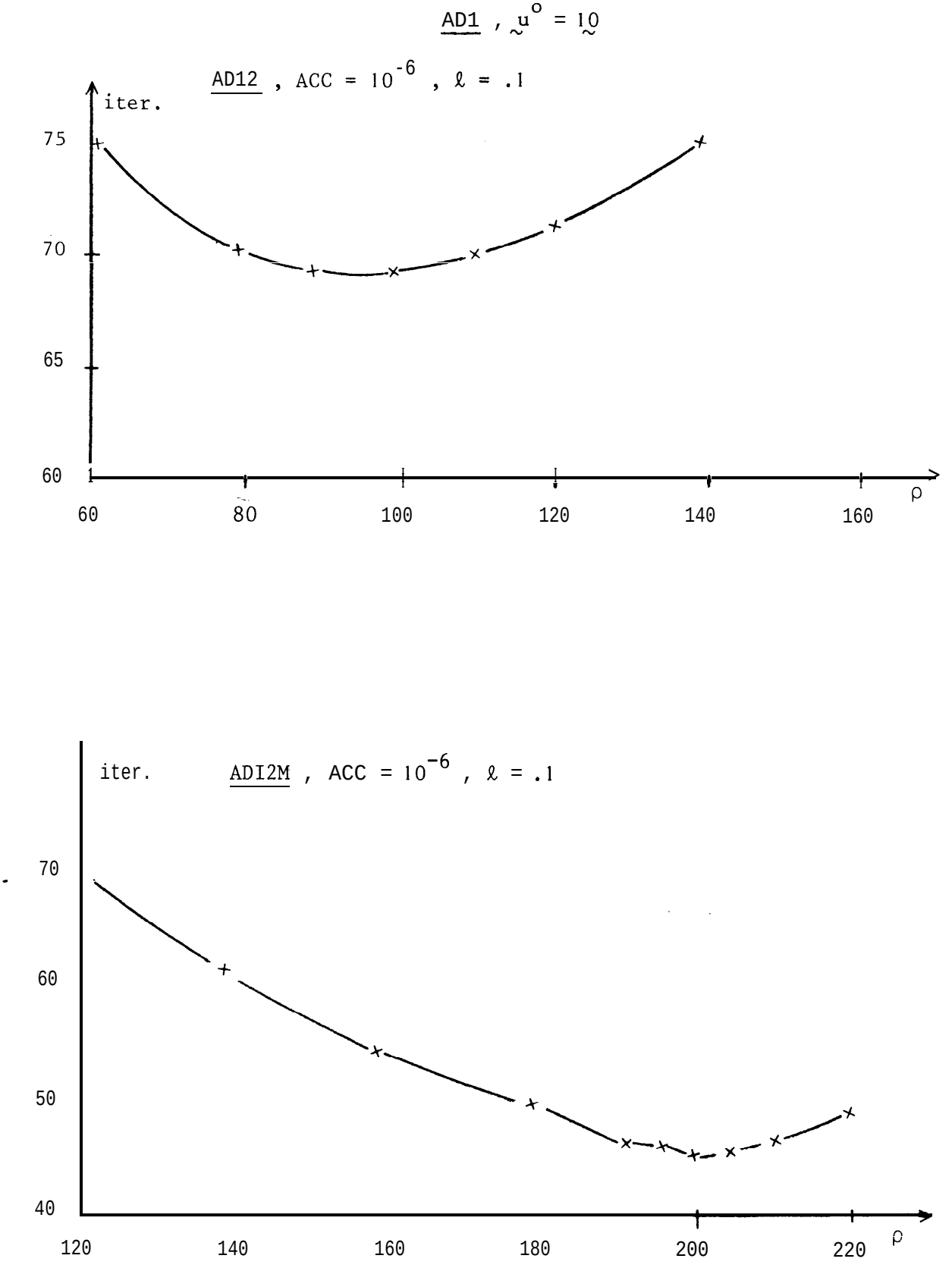

AD1 , $\underset{\sim}{u}^o = 1\underset{\sim}{0}$
AD12 , ACC = 10^{-6} , ℓ = .1
iter.
75
70
65
60
60
80
100
120
140
160
ρ
iter.
ADI2M , ACC = 10^{-6} , ℓ = .1
70
60
50
40
120
140
160
180
200
220
ρ

Fig. 6.8

Comnarison of the different algorithms

Required accuracy : L_2 error< 10^{-6}

ℓ = ● ☒ , $\phi(u)$ ▪ $\mathrm{sgn}(u)\,|u|^{\ell}$ ▱ 17 × 17 grid points

Optimal parameters are used.

u^0 =o

Algorithm	Time (using optimal parameter)	Iterations
ALG1($\varepsilon=10^{-4}$)	2.54 sec.	11
ALG2	1.56 sec.	6
SOR1	5.47 sec.	36
SOR2	5.9 sec.	36
AD12	1.67 sec.	7
ADI2M	0.65 sec.	2

u^0 = 10

Algorithm	Time (using optimal parameter)	Iterations
ALG1($\varepsilon=10^{-4}$)	3.1 sec.	14
ALG2	11.7 sec.	50
SOR1	7.0 sec.	48
SOR2	7.48 sec.	47
ADI2	11.5 sec.	69
ADI2M	7.5 sec.	45

IBM 370/168 , FORTRAN H , OPT = 2. - Double precision

Table 6.1

R E F E R E N C E S

ADAMS R.A.

[I] Sobolev spaces, Academic Press, New-York, 1975.

BARTELS R., DANIEL J.W.

[I] A conjugate gradient approach to nonlinear elliptic boundary value problems in irregular regions, in Conference on the Numerical Solution of Differential Equations, Dundee 1973, G.A. Watson Ed., Lecture Notes in Math., Vol. 363, Springer-Verlag, Berlin, 1974.

BREZIS H., CRANDALL M., PAZY A.,

[1] Perturbations of nonlinear monotone sets in Banach spaces, Comm. Pure Applied Math., Vol. 23, (1970), pp. 123-144.

BRENT R.

[1] Algorithms for minimization without derivatives, Prentice Hall, Englewood Cliff N.J., 1973.

BUZBEE B.L., GOLUB G.H., NIELSON C.W.

[1] On direct methods for solving Poisson's equations, SIAM J. Num. Anal., 7, (1970), pp. 627-656.

CEA J.

[1] Optimisation : Théorie et Algorithmes, Dunod, 1970.

CEA J., GLOWINSKI R.

[1] Sur des méthodes d'optimisation par relaxation, Revue Française d'Automatique, Informatique et Rech. Operationnelle, Dec. 1973, R-3, pp. 5-32.

CIARLET P.G.

[1] The finite element method for elliptic problems, North Holland, Amsterdam, 1978.

[2] Numerical Analysis of the finite element method, Séminaire de Math. Supérieures de l'Université de Montreal , Presse de l'Université de Montreal, 1977.

CIARLET P.G., RAVIART P.A.

[1] The combined effect of curved boundaries and numerical integration in isoparametric finite element methods, in The Mathematical foundations of the finite element method with applications to partial differential equations, A.K. Aziz Ed., Acad. Press, New-York, 1972, pp. 409-474.

CIARLET P.G., SCHULTZ M.H., VARGA R.S.

Cl] Numerical methods of high-order accuracy for nonlinear boundary value problems, V. Monotone Operator Theory, Numerish Math., 13, (1969), pp. 51-77.

CONCUS P., GOLUB G.H.

[1] A generalized conjugate gradient method for non symmetric systems of linear equations, in Computing Methods in Applied Sciences and Engineering, R. Glowinski, J.L. Lions Ed., Lecture Notes in Economics and Math. Systems, Vol. 134, Springer Verlag, Berlin, 1976, pp. 56-65.

CONCUS P., GOLUB G.H., O'LEARY D.,

[1] Numerical solution of nonlinear partial differential equations by a generalized conjugate gradient method, Computing, 19, (1977), 4, pp. 321-340.

DANIEL J.W.

Cl] The approximate minimization of functionals, Prentice Hall, Englewood Cliff N.J., 1970.

DOUGLAS J., DUPONT T.

[1] Preconditioned Conjugate Gradient Iterations applied to Galerkin Methods for . a Mildly Nonlinear Dirichlet Problem, in Sparse Matrix Computations, J.R. Bunch, D.J. Rose Ed., Academic Press, New-York, 1976, pp. 333-348.

EISENSTAT S., SCHULTZ M.H., SHERMAN A.,

[1] The application of sparse matrix methods to the numerical solution of nonlinear elliptic partial differential equations, in Constructive and Computational Methods for Differential and Integral Equations, D.L. Colton, R.P. Gilbert Ed., Lecture Notes in Math., Vol. 430, Springer-Verlag, Berlin, 1974.

EKELAND I., TEMAM R.,

[1] Analyse Convexe et Problèmes Variationnels, Dunod -Gauthier-Villars, 1974.

FIGUEIREDO D.G.

[1] Equações Elipticas não lineares, Published by the Instituto de Matematica Pura e Aplicada do C.N.Pq. (IMPA), Rio do Janeiro, 1977.

FORTIN M., R. GLOWINSKI R.

[1] Chapter 3 of Resolution numérique de problèmes aux limites par des méthodes de lagrangien augment@, M. Fortin, R. Glowinski Ed., (to appear).

GABAY D., MERCIER B.,

[1] A dual algorithm for the solution of nonlinear variational problems via finite element approximation, Comp. and Math. with Applic., 2, (1976), pp. 17-40.

GLOWINSKI R.

[1] Introduction to the approximation of elliptic variational inequalities, Rep. 76006, Laboratoire d'Analyse Numerique, Université Pierre et Marie Curie, 1976.

[2] Numerical Analysis of Nonlinear Variational Problems, Lecture Notes, Tata Institute, Bombay and Bangalore (to appear).

GLOWINSKI R., LIONS J.L., TREMOLIERES R.,

[1] Analyse Numerique des Inequations Variationnelles, Vol. 1, Théorie générale, premieres applications, Dunod-Bordas, Paris, 1976.

[2] Analyse Numerique des Inequations Variationnelles, Vol. 2, Applications aux phénomènes stationnaires et d'évolution, Dunod-Bordas, Paris, 1976.

GLOWINSKI R., MARROCCO A.,

[1] Analyse Numerique du champ magnetique d'un alternateur par elements finis et surrelaxation ponctuelle non lineaire, Computer Meth. Applied Mech. Engineering, 3, (1974), pp. 55-85.

[2] Etude numérique du champ magnetique dans un alternateur tétrapolaire par la méthode des elements finis, in Computing Methods in Applied Sciences and Engineering, Part 1, R. Glowinski, J.L. Lions Ed., Lecture Notes in Comp. Science, Vol. 10, Springer-Verlag, Berlin, 1974, pp. 392-409.

[3] Sur l'approximation par elements finis d'ordre un, et la resolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non lineaire, Revue Française d'Automatique, Informatique, Recherche Operationnelle, R-2, Août 1975, pp. 41-76.

HESTENES M.,

[1] Multiplier and gradient methods, J. of Optimization Theory and Applications, 4, (1969), 5, pp. 303-320.

HOUSEHOLDER A.S.,

[I] The numerical treatment of a single nonlinear equation, Mc Graw-Hill, New-York, 1970.

KELLOG R.B.,

[1] A nonlinear alternating direction method, Math. of Comp , 23, (1969), 105, pp. 23-27.

LIEUTAUD J.

[1] Approximation par des methodes de decomposition, Doctoral Dissertation, University of Paris VI, 1968.

LIONS J.L.

[1] Quelques méthodes de resolution des problèmes aux limites non linéaires, Dunod, Gauthier-Villars, Paris, 1968.

[2] Problèmes aux limites dans les equations aux dérivées partielles, Séminaire de Mathematiques Supérieures de l'Université de Montreal, Presses de l'Université de Montreal, 1962.

LIONS J.L., STAMPACCHIA G.,

[1] Variational Inequalities, Comm. Pure Applied Math., 20, (1967), pp. 493-519.

NECAS J.,

[1] Les méthodes directes en théorie des equations elliptiques, Masson, Paris, 1967.

ODEN J.T., REDDY J.N.,

[1] An introduction to the mathematical theory of finite element, John Wiley and Sons, New-York, 1976.

ORTEGA J., RHEINBOLDT W.C.,

[1] Iterative solution of nonlinear equations in several variables, Academic Press, New-York, 1970.

OSBORN J.E., SATHER D.,

[1] Alternative problems for nonlinear operators, J. of Differential Equations, 17, (1975), pp. 12-31.

POLAK E.

[1] Computational methods in optimization, Acad. Press, New-York, 1971.

POWEL M.J.D.,

[1] Restart procedure for the conjugate gradient method, Mathematical Programming, 12, (1977), pp. 241-254.

SCHECHTER S.

[1] Iteration methods for nonlinear problems, Trans. Amer. Math. Society, 104, (1962), pp. 601-612.

[2] Minimization of convex functions by relaxation, Chapter 7 of Integer and nonlinear Programming, J. Abadie Ed., North-Holland, Amsterdam, 1970, pp. 177-189.

[3] Relaxation methods for convex problems, Siam J. Num. Anal., 5, (1968), pp. 601-612.

STAMPACCHIA G.,

[1] Equations elliptiques du second ordre à coefficients discontinus, Séminaire de Mathematiques Supérieures de l'Université de Montreal, Presses de L'université de Montreal, 1965.

STRANG G., FIX G.,

[1] An analysis of the finite element methods, Prentice Hall, Englewood Cliff N.J., 1973.

YOSIDA K.,

[1] Functional Analysis, Springer Verlag, 1966.