STUDIES IN MATHEMATICS AND ITS APPLICATIONS

J.L. Lions
G. Papanicolaou
R.T. Rockafellar
H. Fujita
Editors

15

AUGMENTED LAGRANGIAN METHODS: APPLICATIONS TO THE NUMERICAL SOLUTION OF BOUNDARY-VALUE PROBLEMS

M. Fortin
R. Glowinski

NORTH-HOLLAND

Augmented Lagrangian Methods: Applications to the numerical solution of boundary-value problems

STUDIES IN MATHEMATICS AND ITS APPLICATIONS

VOLUME 15

NORTH-HOLLAND – AMSTERDAM • NEW YORK • OXFORD

AUGMENTED LAGRANGIAN METHODS: APPLICATIONS TO THE NUMERICAL SOLUTION OF BOUNDARY-VALUE PROBLEMS

MICHEL FORTIN
Professor at the Université Laval, Quebec

ROLAND GLOWINSKI
Professor at the Université Pierre et Marie Curie, Paris
Scientific Director at INRIA

1983

NORTH-HOLLAND – AMSTERDAM • NEW YORK • OXFORD

ISBN 0 444 86680 9

Translation of:
Méthodes de Lagrangien Augmenté
Applications à la résulotion numérique
de problèmes aux limites.
© Bordas (Dunod), Paris, 1982

Publishers:
ELSEVIER SCIENCE PUBLISHERS B.V.
P.O. BOX 1991
1000 BZ AMSTERDAM
THE NETHERLANDS

Sole distributors for the U.S.A. and Canada:
ELSEVIER SCIENCE PUBLISHING COMPANY, INC.
52 VANDERBILT AVENUE, NEW YORK, N.Y. 10017

English version edited, prepared and produced by
TRANS-INTER-SCIENTIA
P.O. Box 16, Tonbridge, TN11 8DY, England

Library of Congress Cataloging in Publication Data
Fortin, Michel.
Augmented Lagrangian methods.

(Studies in mathematics and its applications ; v. 15)
Translation of: Méthodes de Lagrangien augmenté.
Bibliography: p.
1. Boundary value problems--Numerical solutions.
2. Differential equations, Partial--Numerical solutions.
3. Lagrangian functions. I. Glowinski, R. II. Trans-inter-scientia (Firm) III. Title. IV. Series.
QA379.F6713 1983 515.3'5 83-6802
ISBN 0-444-86680-9

PRINTED IN THE NETHERLANDS

INTRODUCTION

The essential purpose of this volume is to present the principles of the *Augmented Lagrangian Method*, together with numerous applications of this method to the *numerical solution* of *boundary-value problems for partial differential equations or inequalities* arising in *Mathematical Physics*, in the *Mechanics of Continuous Media* and in the *Engineering Sciences*.

Simultaneous developments in *computers* and in *Numerical Analysis* - in particular the *Finite Element Method* - have gradually led research workers, engineers, etc... to use *mathematical models* of greater and greater complexity for the representation of the phenomena which arise in their respective disciplines. Certain calculations which hitherto had been considered impracticable have become routine matters, and it has become possible to abandon certain simplifying assumptions - in particular, linearity - and thereby gain a more realistic simulation of the phenomena considered.

This development has led, for *Numerical Analysts* in particular, to a search for efficient methods for solving the "*large systems*" which arise from the discretisation of *nonlinear boundary-value problems, variational inequalities, optimal control problems*, etc...

Many of these problems can be expressed in the form of a *search for a minimum of a functional* - or, at any rate can be easily reduced to this. It is therefore tempting to apply to the solution of these problems methods taken from *Mathematical Programming* (i.e. the field of Optimisation Algorithms).

It should be remarked that, whilst accepting that the existing techniques of Mathematical Programming provide a safe and reliable

starting point, their application to the solution of problems involving *partial differential equations or inequalities* requires certain precautions to be taken; there are in reality few methods capable of efficiently minimising functionals which depend on several thousand variables, over sets defined by similar numbers of linear or nonlinear constraints. As one might expect, it is necessary to adapt the general methods and take account of the particular structure of the problems to be solved. This is what we shall be attempting to do in this book, by showing how a small number of very simple ideas can be applied to the solution of problems which a priori appear completely different.

The approach which we shall be following may at first seem surprising; in fact, we shall very often be modifying the original problem by introducing supplementary constraints and variables, which at first sight, may appear to have the effect of increasing its complexity. However, this complication will then be seen to be largely compensated by the *structural simplification* which it introduces.

The basic principles which will serve to guide us throughout this work will be the following:

(i) There are at the present time efficient methods available for solving *linear systems*, even of very large order; this is particularly true for systems in which the matrix is *symmetric* and *positive definite*.

(ii) The average cost of the above solutions increases only marginally if we solve, using a direct method, linear systems relating to a common matrix, internally within an iterative process.

(iii) *Nonlinear* problems which depend on a *small number of variables* (say, $<<$ 10 to be more specific) are easy to solve (or at any rate much easier to solve than those which depend on a large number of variables).

Methods for solving large nonlinear systems are universally based (see ORTEGA-RHEINBOLDT [1]) on a linearisation embedded within an iterative process. The updating of the linearised problem and its solution generally constitute the most expensive phase of the overall solution process. In this perspective, a solution method which uses *the same matrix all the time* may be extremely attractive.

The methods which we propose utilise a *decomposition-coordination* principle (in the sense of BENSOUSSAN-LIONS-TEMAM [1]) which means

that the nonlinearity is treated at a *local* level, and in which the coordination is effected through the *simultaneous* use of *Lagrange multipliers* and a *penalisation* method; in this we follow a methodology first introduced around 1970 by HESTENES [1] and POWELL [1]. This principle of *localisation of the nonlinearities* is in fact one of the governing principles of this volume. Additionally, the methods of decomposition which will be found in this book are well suited to *Parallel Computation*, which certainly appears to be one of the directions of the future in Numerical Analysis, having regard to the architectures now being adopted for large modern scientific computers. We therefore have reason to hope that the methods developed in this book will find wide application in future years.

We now come on to the contents of this volume:

Chapter I introduces the Augmented Lagrangian method in the classical context of *Quadratic Programming with linear constraints in finite dimensions*. We here study in detail some standard algorithms - and others which are rather less so - and give a number of results, some of them new, relating to their convergence.

In Chapter II we apply the results of the previous chapter to the solution of the *Stokes and Navier-Stokes equations for incompressible viscous fluids*. Certain of the results of Chapter I are verified experimentally on the basis of numerical tests.

In Chapter III we introduce, within a rather general Hilbertian framework, the principle of decomposition-coordination on which the rest of the book will be based. Here we study in particular the convergence, under quite general assumptions, of the two basic algorithms, ALG1 and ALG2, to which the remainder of the book is primarily devoted. This chapter is illustrated by numerous examples taken from Mechanics and from Physics.

In Chapter IV we apply the results of the preceding chapter to the solution of mildly nonlinear problems of the form

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

where A is a linear elliptic operator of order two and ϕ is a numerical function. We also consider in this context the use of hybrid finite elements.

In Chapter V, we apply the results of Chapter III to the solution of second-order nonlinear partial differential equations and inequalities, in which the nonlinearity relates to the gradient of the solution. The methods which are described here apply in particular

to the solution of boundary-value problems for partial differential equations such as

$$-\nabla\cdot(\nu(x,\nabla u)\nabla u) = f.$$

It will be established that such problems can be solved with remarkable efficiency.

In Chapters VI and VII we consider applications of the methods of Chapter III to the solution of, respectively, problems in *Elasto-Plasticity* and in the *steady and time-dependent flow of visco-plastic fluids of Bingham type in two-dimensional cavities*. The problem considered in Chapter VI is formulated initially as an *elliptic variational inequality* relating to the *Linear Elasticity* operator, while that considered in Chapter VII is reformulated in terms of a *variational inequality of order* 4, which may be *elliptic* or *parabolic* depending on the circumstances, via the introduction of a stream function.

Whilst the problems dealt with in Chapters III to VII all fall within the scope of *Convex Analysis* and *Monotone Operators*, those considered in Chapter VIII quite definitely depart from this framework; these are nonlinear problems arising from Finite Nonlinear Elasticity, and reduce to the minimisation of functionals (which may be convex) over non-convex sets. Nonetheless, the decomposition-coordination principles of Chapter III still lead to extremely powerful algorithms for solving these problems, even though strictly speaking we are no longer within the range of application of these methods.

In Chapters II to VIII results of numerical experiments are given which enable the efficiency of the proposed methods of solution to be assessed.

Chapter IX, which concludes this volume, is much more abstract in nature; it takes up, in a general setting, certain of the ideas originally considered in Chapters III and IV and shows, in particular, the links which exist between the algorithms ALG1 and ALG2 of Chapter III and certain classical *alternating direction* methods. It also provides a theoretical framework which is well suited to the study of the convergence of a number of algorithms based on the use of Lagrange multipliers.

Augmented Lagrangian methods have been the subject of numerous publications, and it is very difficult to select from these a bibliography which is anything like complete. We have therefore indicated in this book only references with which we are personally

familiar and which have a direct relevance to the questions addressed herein; we thus advise the reader interested in obtaining further information to refer to the bibliographies of these volumes, together with the following journals:

Journal of Optimization Theory and Applications,
Mathematical Programming,
Siam Journal of Control and Optimization.

We would like to thank Messrs Bégis, Bourgat, Chan, Gabay, Le Tallec, Marrocco, Mercier and Thomasset who participated in the preparation of this book. We would also like to acknowledge our particular indebtedness to Mrs. Françoise Weber of INRIA who painstakingly typed the entire French original of this work (the typed equations of which have been retained in the present English language edition), to the Translators, Messrs B. Hunt and D. Spicer, and their wives who helped produce this work for "*Trans-Inter-Scientia*", and to the *North-Holland Publishing Company* for agreeing to publish this translation in the series

Studies in Mathematics and its Applications.

The first editor (M. Fortin) wishes to thank CRSNG (Canada) and the Ministry of Education of Quebec for the financial aid which they have given to this work, and, since a large part of the work was drafted during a period spent by the second editor (R. Glowinski) at the Mathematical Research Center (M.R.C.) of the University of Wisconsin at Madison (financed under contract DAA 629-80-C-0041), we would like to thank Professor John Nohel, the director of the M.R.C., for the facilities which were made available for carrying out the final drafting of the work.

Last but not least, our thanks go to the Management of INRIA for allowing various people at that institution to participate in the preparation of this book; in particular, it was at INRIA that the majority of the numerical experiments presented herein were carried out.

Madison, U.S.A.
19th August 1981

Michel Fortin,
Roland Glowinski.

The following individuals contributed to the production of the original French-language edition of this book published by Editions Dunod:

D. BEGIS	INRIA, B.P. 105, 78153 Le Chesnay Cedex, France.
J.F. BOURGAT	INRIA, B.P. 105, 78153 Le Chesnay Cedex, France.
T.F. CHAN	Computer Science Department, Yale University, Box 2158, New Haven, Connecticut 06520, U.S.A.
M. FORTIN	Département de Mathématiques, Université Laval, Faculté des Sciences, Québec G1K 7P4, Canada.
D. GABAY	C.N.R.S., Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique LA 189, 4, Place Jussieu, 75230 Paris Cedex 05, and INRIA.
R. GLOWINSKI	Université Pierre et Marie Curie, Laboratoire d'Analyse Numérique LA189, 4 Place Jussieu, 75230 Paris Cedex 05, and INRIA.
P. LE TALLEC	Laboratoire Central des Ponts et Chaussées, Service Mathématiques, 58 Boulevard Lefebvre, 75732 Paris Cedex 15.
A. MARROCCO	INRIA, B.P. 105, 78153 Le Chesnay Cedex, France.
B. MERCIER	C.E.A., Service MA, Centre d'Etudes de Limeil, B.P. 27, 94190 Villeneuve Saint Georges, France.
F. THOMASSET	INRIA, B.P. 105, 78153 Le Chesnay Cedex, France.

The present English-language edition was translated by Messrs. B. Hunt and D.C. Spicer, and was produced by *Trans-Inter-Scientia*.

o o
o

TABLE OF CONTENTS

This Page Intentionally Left Blank

CHAPTER I

AUGMENTED LAGRANGIAN METHODS IN QUADRATIC PROGRAMMING

M. Fortin, R. Glowinski

1. PRINCIPLES OF THE METHOD

In this chapter and with a view to simplifying the presentation, we shall limit ourselves to a particularly simple finite-dimensional problem:

Let A be a symmetric, positive definite $N \times N$ matrix and suppose that $b \in \mathbb{R}^N$; with A and b we associate the quadratic functional $J: \mathbb{R}^N \to \mathbb{R}$ defined by

$$J(v) = \frac{1}{2}(Av,v) - (b,v), \tag{1.1}$$

where in (1.1), $(.,.)$ denotes the canonical Euclidian inner product in $\mathbb{R}^N$. Let B be a linear mapping from $\mathbb{R}^N$ into $\mathbb{R}^M$, this thus being identifiable with an $M \times N$ matrix. We consider the *minimisation* problem

$$\begin{cases} J(u) \leq J(v) & \forall v \in \text{Ker } B = \{v \in \mathbb{R}^N,\ Bv=0\}, \\ u \in \text{Ker } B. \end{cases} \tag{1.2}$$

It is a classical result that (1.2) admits *a unique solution*.

Following a well known technique, we introduce a *Lagrange multiplier* $p \in \mathbb{R}^M$ which transforms (1.2) into an *unconstrained* problem[1], namely

[1] We also use $(.,.)$ for the inner product in $\mathbb{R}^M$, there being no danger of ambiguity.

$$\text{(1.3)} \qquad \min_{v \in \mathbb{R}^N} \{J(v) + (p, Bv)\}.$$

The Lagrange multiplier p appears as an extra unknown which may, for example, be obtained through the solution of a *saddle-point problem*. More precisely, we define $\mathcal{L} : \mathbb{R}^N \times \mathbb{R}^M \to \mathbb{R}$ by

$$\text{(1.4)} \qquad \mathcal{L}(v,q) = J(v) + (q, Bv)$$

and we recall that $\{u,p\}$ will be a *saddle-point* of $\mathcal{L}$ on $\mathbb{R}^N \times \mathbb{R}^M$, if

$$\text{(1.5)} \qquad \mathcal{L}(u,q) \leq \mathcal{L}(u,p) \leq \mathcal{L}(v,p) \qquad \forall v \in \mathbb{R}^N,\ q \in \mathbb{R}^M,$$

and also that (1.5) implies

$$\text{(1.6)} \qquad \min_{v \in \mathbb{R}^N} \max_{q \in \mathbb{R}^M} \mathcal{L}(v,q) = \max_{q \in \mathbb{R}^M} \min_{v \in \mathbb{R}^N} \mathcal{L}(v,q) = \mathcal{L}(u,p).$$

It can be shown that $\mathcal{L}$ admits at least one saddle-point $\{u,p\}$ on $\mathbb{R}^N \times \mathbb{R}^M$, where u is the solution of (1.2) and is common to all the saddle-points of $\mathcal{L}$ on $\mathbb{R}^N \times \mathbb{R}^M$. A *necessary* and *sufficient* condition of *uniqueness* for $\{u,p\}$, in fact for p, is that B be *surjective*, i.e. Rank B = M.

The following (classical) result is essential to the subsequent discussion:

THEOREM 1.1: *The solution* u *of* (1.2) *is characterised by the existence of* $p \in \mathbb{R}^M$ *such that*

$$\text{(1.7)} \qquad \begin{cases} Au + B^t p = b, \\ Bu = 0. \end{cases}$$

The relations (1.7) *also characterise all the saddle-points of* $\mathcal{L}$ *on* $\mathbb{R}^N \times \mathbb{R}^M$.

■

Following HESTENES [1], and POWELL [1] we introduce the *augmented Lagrangian* $\mathcal{L}_r$ defined, for $r > 0$, by

$$\mathcal{L}_r(v,q) = J(v) + (q,Bv) + \frac{r}{2}|Bv|^2 = \mathcal{L}(v,q) + \frac{r}{2}|Bv|^2, \tag{1.8}$$

where, in (1.8), $|.|$ denotes the canonical Euclidian norm on $\mathbb{R}^M$. It can easily be proved that any saddle-point of $\mathcal{L}_r$ is a saddle-point of $\mathcal{L}$ and vice-versa (this is due to the fact that $r|Bv|^2$ vanishes when the constraint $Bv = 0$ is satisfied).

Remark 1.1: It should be noted that for $q = 0$ we have

$$\mathcal{L}_r(v,0) = J(v) + \frac{r}{2}|Bv|^2, \tag{1.9}$$

this being the classical *penalised functional* relative to the constraint $Bv = 0$.

The advantage of the augmented Lagrangian is that, because of the presence of the term (q,Bv), the *exact* solution of problem (1.2) can be determined without making r tend to infinity, unlike ordinary penalisation methods where this has the effect of causing a deterioration in the *conditioning* of the systems to be solved. Furthermore the addition of the quadratic term $\frac{r}{2}|Bv|^2$ to the Lagrangian $\mathcal{L}$ will improve the convergence properties of the duality algorithms described later. ∎

Remark 1.2: The case where B is injective is of no interest since in this case $u = 0$ is the unique solution of (1.2). In the following text we shall thus assume $\mathrm{Ker}\, B \neq \{0\}$. ∎

2. A FIRST ALGORITHM FOR SADDLE-POINT CALCULATION

2.1 Description of the algorithm

It follows from Section 1 that there is equivalence between solving (1.2) and finding a saddle-point of $\mathcal{L}_r$ on $\mathbb{R}^N \times \mathbb{R}^M$; from

ARROW-HURWICZ-UZAWA [1], GLOWINSKI-LIONS-TREMOLIERES[2] [1, Chapter II, Section 4], etc, such a saddle-point can be calculated using the following algorithm, the variants of which we shall denote in the following text under the general name of *Uzawa's algorithm:*

(2.1) $p^0 \in \mathbb{R}^M$, *specified arbitrarily;*
with p^n *known, calculate* u^n *then* p^{n+1} *by*

(2.2) $$\begin{cases} \mathcal{L}_r(u^n,p^n) \leq \mathcal{L}_r(v,p^n) \quad \forall v \in \mathbb{R}^N, \\ u^n \in \mathbb{R}^N, \end{cases}$$

(2.3) $$p^{n+1} = p^n + \rho_n Bu^n \ , \ \rho_n > 0. \ \blacksquare$$

We note that (2.2) is equivalent to

(2.4) $$(A + rB^tB)u^n + B^tp^n = b.$$

2.2 Convergence results

Regarding the convergence of the algorithm, we now prove the following:

THEOREM 2.1: *For* $0 < \alpha_o \leq \rho_n \leq 2r$ *and for all* $p^o \in \mathbb{R}^M$ *the sequence* u^n *defined by* (2.1) - (2.3) *converges to the solution* u *of* (1.2).

Proof: This follows G.L.T. [1, Chapter 2, Section 4][3]. Suppose $\{u,p\}$ is a saddle-point of $\mathcal{L}_r$. From Theorem 1.1, this saddle-point is characterised by

(2.5) $$(A + rB^tB)u + B^tp = b \ ,$$

(2.6) $$Bu = 0 \iff p = p + \rho_n Bu, \ \forall n.$$

---

[2] hereinafter abbreviated to G.L.T. (see also Appendix 2 of G.L.T. [2] (English translation of G.L.T. [1])).

[3] See also G.L.T. [2].

We introduce $\bar{u}^n = u^n - u$, $\bar{p}^n = p^n - p$. Then by subtracting (2.5) from (2.4) and (2.6) from (2.3) we obtain

$$(A + r B^t B)\, \bar{u}^n + B^t \bar{p}^n = 0 \tag{2.7}$$

$$\bar{p}^{n+1} = \bar{p}^n + \rho_n B\bar{u}^n . \tag{2.8}$$

We deduce from (2.8) that

$$|\bar{p}^{n+1}|^2 = |\bar{p}^n|^2 + 2\rho_n (\bar{p}^n, B\bar{u}^n) + \rho_n^2 |B\bar{u}^n|^2,$$

and hence that

$$|\bar{p}^n|^2 - |\bar{p}^{n+1}|^2 = -2\rho_n(\bar{p}^n, B\bar{u}^n) - \rho_n^2 |B\bar{u}^n|^2 . \tag{2.9}$$

It follows from (2.7) that

$$(A\bar{u}^n, \bar{u}^n) + r|B\bar{u}^n|^2 = -(\bar{p}^n, B\bar{u}^n),$$

and hence by substitution in (2.9) we have

$$|\bar{p}^n|^2 - |\bar{p}^{n+1}|^2 = 2\rho_n(A\bar{u}^n, \bar{u}^n) + \rho_n(2r-\rho_n) |B\bar{u}^n|^2. \tag{2.10}$$

If we then take

$$0 < \alpha_o \le \rho_n \le 2r \tag{2.11}$$

the sequence $|\bar{p}^n|$ is decreasing. Being *bounded below* by 0, it is convergent and hence $|\bar{p}^n|^2 - |\bar{p}^{n+1}|^2 \to 0$; (2.10) then implies

$$\lim_{n\to+\infty} (A\bar{u}^n, \bar{u}^n) = 0 , \tag{2.12}$$

and since A is *positive definite* it follows from (2.12) that $\bar{u}^n \to 0$, and also since $\bar{u}^n = u^n - u$, we have $\lim_{n\to+\infty} u^n = u$. ∎

Remark 2.1: It follows from (2.10) that we actually have convergence of (2.1) - (2.3) under the following condition, which is less restrictive than (2.11):

$$0 < \alpha_o \le \rho_n \le \alpha_1 < 2\left(r + \frac{1}{\beta^2}\right) \tag{2.13}$$

where β^2 is defined by

$$\beta^2 = \operatorname*{Max}_{v \neq 0} \frac{|Bv|^2}{(Av,v)} \tag{2.14}$$

which implies that β^2 is *the largest eigenvalue* of $A^{-1}B^tB$. ■

With a view to studying the behaviour of the sequence p^n it is worth noting that

$$(\text{Im } B)^{\perp} = \text{Ker } B^t, \tag{2.15}$$

hence

$$\mathbb{R}^M = \text{Im } B \oplus \text{Ker } B^t. \tag{2.16}$$

We thus have, for $\forall q \in \mathbb{R}^M$, the *unique* decomposition

$$q = q_1 + q_2 \ , \ q_1 \in \text{Im } B \ , \ q_2 \in \text{Ker } B^t. \tag{2.17}$$

Denoting by P_1 (resp. P_2) the *projector* of $\mathbb{R}^M$ onto Im B (resp. Ker B^t), we thus have

$$\left| \begin{array}{l} P_i \in \mathcal{L}(\mathbb{R}^M,\mathbb{R}^M) \quad \forall i = 1,2, \\ \\ P_i(q) = q_i \quad \forall q \in \mathbb{R}^M \ , \ \forall i=1,2. \end{array} \right. \tag{2.18}$$

If, additionally, p is a Lagrange multiplier for (1.2), (1.4), then we can deduce from Theorem 1.1 that the same is also true of $p+q$, $\forall q \in \text{Ker } B^t$. It follows from this that there exists a *unique* $\hat{p} \in \text{Im } B$, such that the Lagrange multipliers of (1.2), (1.4) are all of the form

$$p = \hat{p}+q \ , \ q \in \text{Ker } B^t. \tag{2.19}$$

The vector $\hat{p}$ thus appears as the Lagrange multiplier of (1.2), (1.4) with *minimal norm* in $\mathbb{R}^M$.

From Theorem 2.1 and the above properties of the Lagrange multipliers we shall now deduce:

THEOREM 2.2: *If* ρ_n *satisfies* (2.13) *then the sequence* p^n *defined by algorithm* (2.1) - (2.3) *converges to* $\hat{p} + p_2^o$ *where* $p_2^o = P_2(p^o)$ *is the component of* p^o *in* Ker B^t. *In particular if* $p^o = 0$ *then* $\lim_{n\to+\infty} p^n = \hat{p}$.

Proof: The relation (2.3) immediately implies

$$P_2(p^{n+1}) = P_2(p^n) = P_2(p^o) \quad \forall n. \tag{2.20}$$

Under the condition (2.13), we have $\lim_{n\to+\infty} \bar{u}^n = 0$. It then follows from (2.7) that

$$\lim_{n\to+\infty} B^t \bar{p}^n = 0. \tag{2.21}$$

Since the quantity $|B^t q|$ defines a norm on ImB, (2.21) implies by projection onto ImB that

$$\lim_{n\to+\infty} P_1(p^n) = \hat{p}$$

hence, with (2.20), $\lim_{n\to+\infty} p^n = \hat{p} + P_2(p^o)$. ∎

2.3 Interpretation of algorithm (2.1) - (2.3). Rate of convergence if $\rho_n = \rho$, and choice of r.

Algorithm (2.1)-(2.3) is in fact a *gradient* type algorithm applied to the minimisation of the dual functional $J_r^* : \mathbb{R}^M \to \mathbb{R}$ defined by

$$\left\{ \begin{array}{l} J_r^*(q) = -\min_{v\in\mathbb{R}^N} \mathcal{L}_r(v,q) = \frac{1}{2}(BA_r^{-1}B^t q, q) - (BA_r^{-1}b, q) + \frac{1}{2}(A_r^{-1}b, b) \\ \quad \textit{where} \qquad A_r = A + r\,B^tB. \end{array} \right. \tag{2.22}$$

More precisely, by eliminating u^n, algorithm (2.1)-(2.3) can be re-expressed as

$$p^o \in \mathbb{R}^M, \quad \textit{specified arbitrarily;} \tag{2.23}$$

$$p^{n+1} = p^n - \rho_n(BA_r^{-1}B^t p^n - BA_r^{-1}b). \tag{2.24}$$

Remark 2.2: The advantage of formulation (2.1)-(2.3) compared

with (2.23), (2.24) is that we do not have to construct A_r^{-1} explicitly; in certain applications in partial differential equations this would in practice be unrealizable since A_r^{-1} will be a full matrix of very large order. The formulation (2.23), (2.24) will on the other hand be very useful as a theoretical basis for studying the influence of r and ρ_n on the convergence of (2.1)-(2.3). ∎

In similar fashion, by eliminating p^n, (2.1)-(2.3) can be re-expressed as

$$u^o = A_r^{-1}(b - B^t p^o) , \tag{2.25}$$

$$u^{n+1} = u^n - \rho_n A_r^{-1} B^t B u^n , \tag{2.26}$$

and Remark 2.2 holds equally for the algorithm (2.25), (2.26). ∎

To study the influence of r and ρ_n on the convergence of (2.1)-(2.3), we observe that if $\bar{p}^n = p^n - (\hat{p} + P_2(p^o))$ then from (2.20), (2.24) we have

$$\begin{cases} \bar{p}^n \in \operatorname{Im} B \quad \forall n \geq 0 \\ \\ \bar{p}^{n+1} = \bar{p}^n - \rho_n B A_r^{-1} B^t \bar{p}^n . \end{cases} \tag{2.27}$$

From the second relation in (2.27) we deduce that

$$A^{-1}B^t\bar{p}^{n+1} = A^{-1}B^t\bar{p}^n - \rho_n A^{-1}B^t B A_r^{-1} B^t \bar{p}^n . \tag{2.28}$$

We have

$$A_r^{-1} = (I + rA^{-1}B^tB)^{-1}A^{-1} \tag{2.29}$$

which together with (2.28) implies

$$A^{-1}B^t\bar{p}^{n+1} = A^{-1}B^t\bar{p}^n - \rho_n A^{-1}B^tB(I + rA^{-1}B^tB)^{-1}A^{-1}B^t\bar{p}^n . \tag{2.30}$$

We put $y^n = A^{-1}B^t\bar{p}^n$; given the *linear* relations existing between y^n and $\bar{u}^n$, $\bar{p}^n$ the convergence of y^n towards zero tells us that $\{\bar{u}^n, \bar{p}^n\}$ converges towards $\{0,0\}$. We deduce from (2.30) that

$$y^{n+1} = (I - \rho_n A^{-1}B^tB(I + rA^{-1}B^tB)^{-1})\, y^n. \tag{2.31}$$

We shall express (2.31) in a *basis of eigenvectors* of $A^{-1}B^tB$, but before doing so we shall first indicate, without proof, several properties of the eigenvectors and eigenvalues of $A^{-1}B^tB$:

PROPOSITION 2.1: *The eigenvalues of* $A^{-1}B^tB$ *are* ≥ 0 *and the eigenvectors corresponding to two distinct eigenvalues are* A*-orthogonal, i.e. if*

$$\begin{cases} A^{-1}B^t B\, w_i = \lambda_i\, w_i \\ A^{-1}B^t B\, w_j = \lambda_j\, w_j \end{cases}$$

with $\lambda_i \neq \lambda_j$ *then*

$$(A\, w_i, w_j) = 0. \qquad \blacksquare$$

PROPOSITION 2.2: *If* 0 *is an eigenvalue of* $A^{-1}B^tB$ *then the corresponding eigen-subspace is* $\operatorname{Ker} B$ *and* $\operatorname{Im} A^{-1}B^t$ *is the subspace of* $\mathbb{R}^N$ *which is* A*-orthogonal to* $\operatorname{Ker} B$. $\operatorname{Im} A^{-1}B^t$ *is thus spanned by the eigenvectors of* $A^{-1}B^tB$ *associated with the eigenvalues which are distinct from* 0. $\blacksquare$

We have, of course, $\dim \operatorname{Im} A^{-1}B^t = \operatorname{Rank} B^t = \operatorname{Rank} B$. In the following text we shall denote by N_1 the rank of B and by λ_m, λ_M, respectively, the smallest non-zero eigenvalue and the largest eigenvalue of $A^{-1}B^tB$. Suppose $\mathcal{B} = \{w_i\}_{i=1}^{N_1}$ is a basis of $\operatorname{Im} A^{-1}B^t$, with w_i the eigenvector of $A^{-1}B^tB$ associated with the eigenvalue λ_i. If $y \in \operatorname{Im} A^{-1}B^t$ we thus have

$$y = \sum_{i=1}^{N_1} y_i\, w_i\ , \quad y_i \in \mathbb{R} \quad \forall i.$$

We then deduce from $y^n \in \operatorname{Im} A^{-1}B^t\ \forall n \geq 0$, and from (2.31) that

$$\begin{cases} y_i^{n+1} = \left(1 - \dfrac{\rho_n \lambda_i}{1+r\lambda_i}\right) y_i^n = \left(\dfrac{1 + (r-\rho_n)\lambda_i}{1 + r\lambda_i}\right) y_i^n \quad \forall n \geq 0, \\ \forall i = 1, \ldots N_1. \end{cases} \tag{2.32}$$

In the remainder of Section 2.3 we shall assume that $\rho_n = \rho \ \forall n$, postponing until Section 3 the study of algorithms of the type (2.1)-(2.3) with variable step size ρ_n. We thus have

$$y_i^{n+1} = \frac{1 + (r-\rho)\lambda_i}{1 + r\lambda_i} y_i^n. \tag{2.33}$$

We shall first examine a number of possible choices for ρ, with r *given*, as this will allow us to draw a number of conclusions concerning the choice of r.

(i) *The case* $r = 0$.

We here have

$$y_i^{n+1} = (1 - \rho\lambda_i)y_i^n \quad \forall n \geq 0, \ \forall i = 1, \ldots N_1. \tag{2.34}$$

From an inspection of the (classically familiar) Figure 2.1

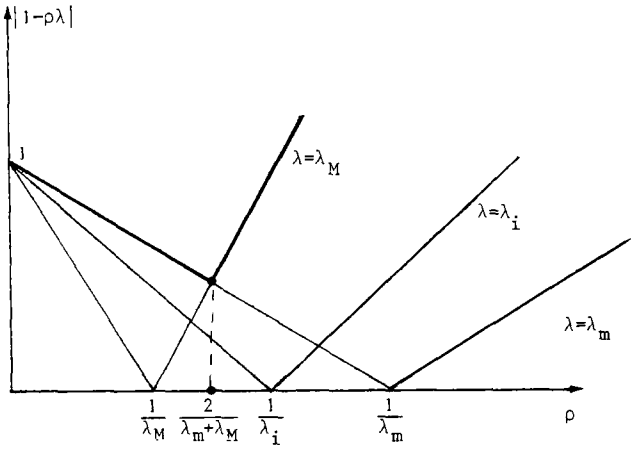


Figure 2.1

the optimal choice for ρ can be seen to be

$$\rho_{opt} = \frac{2}{\lambda_m + \lambda_M} \ ; \tag{2.35}$$

for $\rho = \rho_{opt}$ we then have

$$|y_i^{n+1}| \leq \left(\frac{1 - \frac{\lambda_m}{\lambda_M}}{1 + \frac{\lambda_m}{\lambda_M}} \right) |y_i^n| \qquad \forall i=1,\ldots N_1 \ , \ \forall n \geq 0 \ ; \tag{2.36}$$

we deduce from this that the *convergence rate* R of the method satisfies, for $\rho = \rho_{opt}$,

$$R \leq \frac{1 - \frac{\lambda_m}{\lambda_M}}{1 + \frac{\lambda_m}{\lambda_M}} \ . \tag{2.37}$$

■

Remark 2.3: If $r = 0$, algorithm (2.1)-(2.3) reduces to Uzawa's algorithm applied to the Lagrangian $\mathcal{L}$ defined by (1.4), namely

$$p^o \in \mathbb{R}^M, \quad \textit{specified arbitrarily;} \tag{2.38}$$

$$Au^n + B^t p^n = b \tag{2.39}$$

$$p^{n+1} = p^n + \rho_n Bu^n. \tag{2.40}$$

The result (2.37) is a standard one in the study of the convergence of fixed step gradient methods (see for example, CEA [1], MARCHOUK-KUZNETSOV [1] and the bibliography of these volumes).

(ii) *The case* $\rho = 2r$.

In this case we deduce from (2.33) that

$$y_i^{n+1} = \frac{1 - r\lambda_i}{1 + r\lambda_i} y_i^n \qquad \forall i=1,\ldots N_1, \ \forall n \geq 0. \tag{2.41}$$

As indicated by Figure 2.2, the optimal choice for r is

$$r_{opt} = \frac{1}{\sqrt{\lambda_m \lambda_M}} \ . \tag{2.42}$$

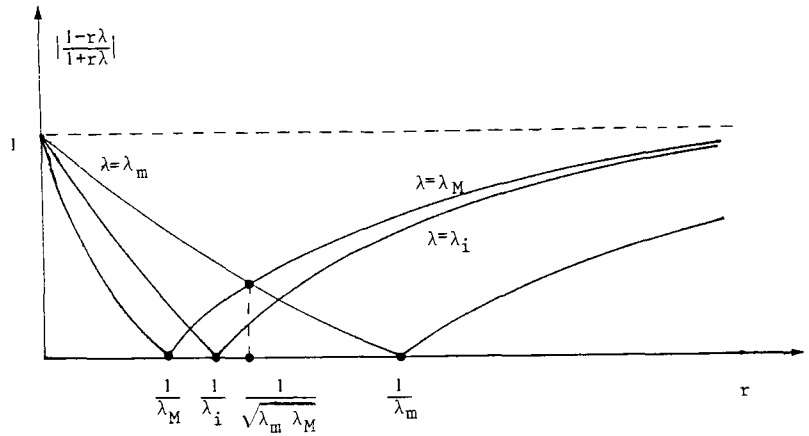


Figure 2.2

and for $r = r_{opt}$, $\rho = 2r_{opt}$, we have

$$|y_i^{n+1}| \leq \left(\frac{1 - \sqrt{\frac{\lambda_m}{\lambda_M}}}{1 + \sqrt{\frac{\lambda_m}{\lambda_M}}} \right) |y_i^n| \quad \forall i=1,\dots N_1 \ , \ \forall n \geq 0. \tag{2.43}$$

We thereby deduce that for this choice of r and ρ the convergence rate R of the method satisfies

$$R \leq \frac{1 - \sqrt{\frac{\lambda_m}{\lambda_M}}}{1 + \sqrt{\frac{\lambda_m}{\lambda_M}}} \ . \tag{2.44}$$

We deduce from (2.37), (2.44) (and from the behaviour of the function $\xi \to \frac{1-\xi}{1+\xi}$) that algorithm (2.1)-(2.3) with $r = \frac{1}{\sqrt{\lambda_m \lambda_M}}$, $\rho = \frac{2}{\sqrt{\lambda_m \lambda_M}}$ is *iteratively faster* than algorithm (2.38)-(2.40) with $\rho_n = \rho = \frac{2}{\lambda_m + \lambda_M}$, which we recall corresponds to algorithm (2.1)-(2.3) with $r = 0$ and $\rho_n = \rho = \frac{2}{\lambda_m + \lambda_M}$. ■

(iii) *The case* $\rho = r$.

This choice is the standard one; in this case we can deduce from (2.33) that

$$y_i^{n+1} = \frac{1}{1 + r\,\lambda_i}\, y_i^n \qquad \forall i=1,\ldots N_1, \quad \forall n \geq 0,$$

and hence that

$$|y_i^{n+1}| \leq \frac{1}{1 + r\,\lambda_m}\, |y_i^n| \qquad \forall i=1,\ldots N_1, \quad \forall n \geq 0. \tag{2.45}$$

We deduce from (2.45) that the *convergence rate* R satisfies

$$R \leq \frac{1}{1 + r\,\lambda_m} \,. \tag{2.46}$$

On inspection of (2.46) it appears that for $\rho_n = \rho = r$ algorithm (2.1)-(2.3) becomes faster, *iteratively*, as the value of r gets larger. If, in particular, $r \ll \frac{1}{\lambda_m}$, it follows from (2.46) that algorithm (2.1)-(2.3) will in general be iteratively slow. We note that if $r = \frac{1}{\lambda_m}$ then $R \leq \frac{1}{2}$. ∎

Remark 2.4: Although relation (2.46) appears to indicate that it is advantageous to work with $\rho_n = \rho = r$ as large as possible, one must realise that *all other things being equal* the determination of u^n in (2.2), i.e. the solution of the linear system

$$(A + r\,B^tB)u^n = b - B^t\,p^n, \tag{2.47}$$

is more costly (in computation time and/or in memory requirements) the larger the value of r. In fact, as we shall see in the following discussion, the matrix $A_r = A + r\,B^tB$ becomes progressively more ill-conditioned the larger r becomes.

Using, once again, a basis of eigenvectors of $A^{-1}B^tB$, it can be shown that the *condition number* of the matrix $A^{-1}B^tB(I + rA^{-1}B^tB)^{-1}$ *restricted* to $\mathrm{Im}A^{-1}B^t$ (i.e. to the subspace A-orthogonal to Ker B) *tends to* 1 when $r \to +\infty$. This property clearly corresponds to the fact that the theoretical convergence of algorithm (2.1)-(2.3) is faster the larger the value of r, in the case where $\rho = r$. ∎

(iv) *Optimal choice of* ρ *with given* r.

From inspection of Figure 2.3, it follows from (2.33)

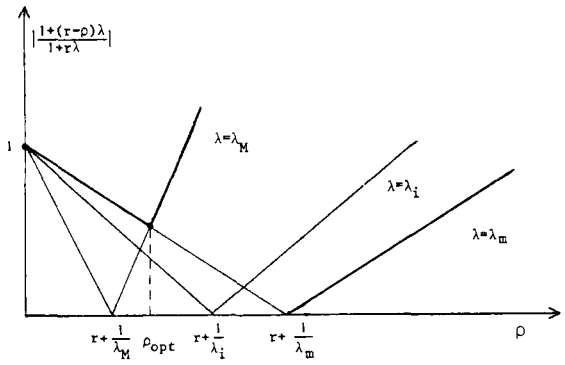


Figure 2.3

that the optimal value of ρ is the solution of the linear equation

$$\frac{1 + (r-\rho)\,\lambda_m}{1 + r\,\lambda_m} = -\frac{1 + (r-\rho)\,\lambda_M}{1 + r\,\lambda_M}$$

hence

$$\rho_{opt} = 2\,\frac{1 + r(\lambda_m + \lambda_M) + r^2\,\lambda_m\,\lambda_M}{(\lambda_m + \lambda_M) + 2r\,\lambda_m\,\lambda_M} = r + \frac{2 + r(\lambda_m + \lambda_M)}{2r\,\lambda_m\,\lambda_M + (\lambda_m + \lambda_M)}. \tag{2.48}$$

We deduce from (2.48) that for $\rho = \rho_{opt}$ we have

$$|y_i^{n+1}| \leq \frac{1 - \frac{\lambda_m}{\lambda_M}}{1 + \frac{\lambda_m}{\lambda_M} + 2r\lambda_m}\,|y_i^n| \qquad \forall i=1,\ldots N_1, \quad \forall n \geq 0, \tag{2.49}$$

and hence for the convergence rate R, we have

$$R \leq \frac{1 - \frac{\lambda_m}{\lambda_M}}{1 + \frac{\lambda_m}{\lambda_M} + 2r\lambda_m} . \tag{2.50}$$

We note that $\rho_{opt} > r$ and that for *the same value of* r algorithm (2.1)-(2.3) is *iteratively faster* with ρ given by (2.48) than with $\rho = r$; of course, (2.48) involves λ_m and λ_M, quantities which in general are not known a priori. Remark 2.4 is again valid for this choice of ρ, with (2.46) replaced by (2.50)

Condition number of A_r*; choice of* r.

We now go to finish off Remark 2.4, but first it is appropriate to define some notation. We shall denote by $|v|$ the standard Euclidian norm on $\mathbb{R}^N$ and for a linear operator L defined on $\mathbb{R}^N$ we shall denote by $\|L\|$ the norm associated with $|.|$, namely

$$\|L\| = \sup_{v \in \mathbb{R}^N - \{0\}} \frac{|Lv|}{|v|} = \sup_{v \in S} |Lv|$$

where

$$S = \{v | v \in \mathbb{R}^N, \ |v| = 1\}.$$

For the condition number of A_r when $r \to +\infty$, we then have the following:

PROPOSITION 2.3: *The condition number* $\mathcal{K}(A_r)$ *of* A_r *satisfies*

$$\mathcal{K}(A_r) \simeq r \frac{\|B\|^2}{\sigma} \quad \textit{when} \quad r \to +\infty, \tag{2.51}$$

where

$$\sigma = \inf_{v \in \text{Ker } B - \{0\}} \frac{(Av, v)}{|v|^2} .$$

Proof: We have

$$\mathcal{K}(A_r) = \|A_r\| \ \|A_r^{-1}\| .$$

It can easily be shown that

$$r\,\|B\|^2 + \frac{1}{\|A^{-1}\|} \leq \|A_r\| \leq r\,\|B\|^2 + \|A\| . \tag{2.52}$$

We have moreover

$$\frac{1}{\|A_r^{-1}\|} = \operatorname*{Min}_{v \in S} (A_r v, v) = \operatorname*{Min}_{v \in S} \left[(Av,v) + r\,|Bv|^2\right] ;$$

there then exists $u_r \in S$ such that

$$\frac{1}{\|A_r^{-1}\|} = (Au_r, u_r) + r|Bu_r|^2. \tag{2.53}$$

We now study the behaviour of the family $(u_r)_r$ when $r \to +\infty$.

We have

$$(Au_r, u_r) + r|Bu_r|^2 \leq (Av,v) \ , \ \forall v \in \operatorname{Ker} B \cap S, \ \forall r \tag{2.54}$$

which implies

$$(Au_r,u_r) + r\,|Bu_r|^2 \leq \|A\| \ , \ \forall r \ . \tag{2.55}$$

Since the sequence u_r is bounded, we can extract from it a subsequence, also denoted by u_r, converging to an element u^* of $\mathbb{R}^N$. We then deduce from (2.55) that

$$|Bu^*|^2 = \lim_{r \to +\infty} |Bu_r|^2 = \lim_{r \to +\infty} \frac{\|A\|}{r} = 0, \tag{2.56}$$

hence it follows that

$$u^* \in \operatorname{Ker} B \cap S. \tag{2.57}$$

In view of (2.57) we can take $v = u^*$ in (2.54). We then deduce that

$$(Au_r,u_r) \leq (Au_r,u_r) + r\,|Bu_r|^2 \leq (Au^*, u^*) \quad \forall r$$

and hence that

$$\lim_{r \to +\infty} r|Bu_r|^2 = 0,$$

which together with (2.53) implies

$$\lim_{r \to +\infty} \frac{1}{\|A_r^{-1}\|} = (Au^*, u^*). \tag{2.58}$$

It therefore follows from the above that

$$\begin{cases} (Au^*,\ u^*) \leq (Av,v) \quad \forall v \in \mathrm{Ker}\ B \cap S, \\ u^* \in \mathrm{Ker}\ B \cap S \end{cases}$$

thus

$$\lim_{r \to +\infty} \|A_r^{-1}\| = \frac{1}{(Au^*,u^*)} = \frac{1}{\sigma} \tag{2.59}$$

which with (2.52) implies (2.51). ∎

We have thus proved that the condition number of A_r is, *asymptotically, proportional to* r, which thus has the effect as r increases, of making it more difficult, (other things being equal), to solve the system

$$A_r\, u^n = b - B^t\, p^n. \tag{2.60}$$

If we solve (2.60) by an *iterative method*, the convergence, being linked to the condition number, will become slower as the value of r increases, and this may lead to a large number of iterations to solve (2.60) to an appropriate accuracy even if, in the obvious manner, we initialise the calculation of u^n with u^{n-1}. Furthermore, if we solve (2.60) by a direct method, the sensitivity to *rounding-error accumulation* will be greater when r is large. In a large number of problems, therefore, a "good strategy" would seem to be the following:

* Work in *"double precision"*

* With the parameter r having a fixed value, as large as possible (!), carry out *once and for all* the Cholesky factorisation of the matrix A_r, which we recall is symmetric and positive definite.

* Take $\rho_n = \rho = r$. ∎

Remark 2.5: In the case where an iterative method is used to solve (2.60) we can, in the early stages of algorithm (2.1)-(2.3), make do with a low accuracy in the determination of u^n. This effect can be obtained, for example, by choosing to use a *fixed* (and "*small*") number of iterations (in the solution of (2.60)).

We shall return to this subject in Section 4, in connection with the method of *Arrow-Hurwicz*. ■

Remark 2.6: With regard to the solution of (2.60) by an *iterative method*, it may be advantageous to use a parameter r which varies with n, giving in fact a sequence $(r_n)_n$. Certain authors recommend the use of a sequence $(r_n)_n$ such that

$$\begin{cases} r_o \geq 0 \\ r_{n+1} \geq r_n \quad \forall n \geq 0, \\ \lim_{n \to +\infty} r_n = +\infty. \end{cases}$$

The optimal choice for $(r_n)_n$ seems to be an open question. The use of such a method combined with a *direct* solution of (2.60) is of little interest, since the factorisation of A_{r_n} would need to be carried out every time that $r_n > r_{n-1}$, this being costly in general. ■

3. VARIABLE STEP-LENGTH ALGORITHMS. CONJUGATE GRADIENT METHOD.

3.1 General notes

We have shown in Section 2.3 that algorithm (2.1)-(2.3) can be interpreted as a gradient algorithm applied to the minimisation of the dual functional J^* defined by (2.22). With this interpretation in mind it is natural to seek to apply, to the minimisation of J^* on $\mathbb{R}^M$, the standard *iterative methods* for *minimisation of quadratic functionals* (see for example, CEA [1], the review article of MARCHOUK-KUZNETSOV [1] and the corresponding bibliography for a thorough study of these methods). In order to clarify the presentation we shall now give some review material on these methods.

Suppose then that $\mathcal{A}$ is an $M \times M$ symmetric, positive definite matrix and suppose $\beta \in \mathbb{R}^M$; we associate with $\mathcal{A}$ and β the functional $\mathcal{J}$ defined by

$$\mathcal{J}(q) = \frac{1}{2} (\mathcal{A}q, q) - (\beta, q). \tag{3.1}$$

The *minimisation* problem

$$\text{(3.2)} \qquad \begin{cases} \mathcal{J}(p) \le \mathcal{J}(q) \quad \forall q \in \mathbb{R}^M , \\ \\ p \in \mathbb{R}^M , \end{cases}$$

admits a *unique solution* which is also a solution of the linear system

$$\text{(3.3)} \qquad \mathcal{A}\, p = \beta .$$

To solve (3.2), (3.3) we now consider descent methods of the general type:

$$\text{(3.4)} \qquad \begin{cases} p^0 \in \mathbb{R}^M, \quad \textit{specified arbitrarily;} \\ \\ p^{n+1} = p^n - \rho_n w_n . \end{cases}$$

The *descent direction* w_n will in general be deduced from the *direction of the gradient* of $\mathcal{J}$ at the point p^n. For a given descent direction, we shall choose ρ_n in such a way as to optimise a criterion related to the problem. In practice, we shall confine ourselves to the following methods:

① *STEEPEST DESCENT METHOD:* The descent is made in the direction opposite to the gradient, hence

$$\text{(3.5)} \qquad w_n = g_n = \operatorname{grad} \mathcal{J}(p^n) = \mathcal{A}\, p^n - \beta .$$

The choice of ρ_n is made by minimising, with respect to ρ, the function

$$\text{(3.6)} \qquad \rho \to \mathcal{J}(p^n - \rho\, g_n) .$$

We have

$$\text{(3.7)} \qquad \mathcal{J}(p^n - \rho\, g_n) = \mathcal{J}(p^n) - \rho |g_n|^2 + \frac{\rho^2}{2} (\mathcal{A} g_n, g_n) ,$$

ρ_n is thus given by

$$\text{(3.8)} \qquad \rho_n = \frac{|g_n|^2}{(\mathcal{A} g_n, g_n)} .$$

② *MINIMUM RESIDUAL METHOD:* The descent is still carried out in the gradient direction, hence $w_n = g_n$; we choose ρ_n so as to minimise, with respect to ρ, the residual $|\mathcal{A}(p^n - \rho g^n) - \beta|$.

Expanding, we have

$$|\mathcal{A}(p^n-\rho g_n)-\beta|^2 = |g_n-\rho\mathcal{A}g_n|^2 = |g_n|^2 - 2\rho(\mathcal{A}g_n,g_n) + \rho^2|\mathcal{A}g_n|^2; \tag{3.9}$$

the optimal ρ is therefore given by:

$$\rho_n = \frac{(\mathcal{A}g_n,g_n)}{|\mathcal{A}g_n|^2} \quad . \tag{3.10}$$

③ *CONJUGATE GRADIENT METHOD:* The *conjugate gradient* method is especially attractive for solving quadratic problems because *theoretically* (i.e. ignoring rounding errors) it *converges in a finite number of iterations* ($\leq M$) and because moreover in the general case it leads to *quadratic convergence*[4]. It would be too lengthy and inappropriate to study here the convergence of this method; we instead refer the reader to E. POLAK [1], J. DANIEL [1], CEA [1], MARCHOUK-KUZNETSOV [1], CONCUS-GOLUB [1] etc..., for a detailed analysis of this algorithm.

The algorithm proceeds as follows:

1) We carry out a *first steepest descent step* by taking $w_o = g_o$, with ρ_o given by (3.8).

2) In the subsequent stages we construct descent directions w_n which are $\mathcal{A}$*-conjugates*, i.e. $(\mathcal{A}w_i, w_j) = 0\ \forall i,j,\ i\neq j$. More precisely, assuming that we know the descent direction w_{n-1} and that we are able to calculate p^n in terms of p^{n-1}, we seek w_n in the form

$$w_n = g_n + \lambda_n w_{n-1} ,$$

such that

$$(\mathcal{A}w_n, w_{n-1}) = 0. \tag{3.11}$$

We hereby deduce that λ_n must take the value

$$\lambda_n = -\frac{(\mathcal{A}g_n, w_{n-1})}{(\mathcal{A}w_{n-1},w_{n-1})} . \tag{3.12}$$

---

[4] For M "large" quadratic convergence becomes a greater attraction than convergence in a finite number of iterations.

We can further prove the orthogonality relations

$$(3.13)\qquad \begin{cases} (\mathcal{A} w_i, w_j) = 0 & \text{if } i \neq j, \\ (g_i, g_j) = 0 & \text{if } i \neq j, \\ (g_i, w_j) = 0 & \text{if } i > j. \end{cases}$$

By virtue of these relations, (3.12) can be reduced by elementary manipulations to

$$(3.14)\qquad \lambda_n = \frac{|g_n|^2}{|g_{n-1}|^2} .$$

3) λ_n and thus w_n being known, we carry out an optimal descent in the direction w_n, i.e. we choose ρ_n to minimise the function

$$\rho \to \mathcal{J}(p^n - \rho w_n) .$$

By a calculation analogous to that carried out in ① we obtain

$$(3.15)\qquad \rho_n = \frac{|g_n|^2}{(g_n, \mathcal{A} w_n)} \left(= \frac{|g_n|^2}{(\mathcal{A} g_n, g_n)} \right). \qquad \blacksquare$$

3.2 Application to the minimisation of J_r^*.

We shall see here how the methods described in Section 3.1 can be applied to the minimisation of J_r^*. We recall that we have

$$(3.16)\qquad J_r^*(q) = \tfrac{1}{2}(BA_r^{-1}B^t q, q) - (BA_r^{-1}b, q) + \tfrac{1}{2}(A_r^{-1}b, b).$$

The constant term $(A_r^{-1}b, b)$ obviously plays no part in the minimisation. We thus have, with the notation of Section 3.1

$$(3.17)\qquad \mathcal{A} = BA_r^{-1}B^t,$$

$$(3.18)\qquad \beta = BA_r^{-1}b.$$

The matrix $\mathcal{A}$ given by (3.17) is only *positive semi-definite*, but it can easily be shown that this does not affect the algorithms considered in Section 3.1, which always converge in the quotient space $\mathbb{R}^M/\mathrm{Ker}\, B^t$. We can prove as in Theorem 2.2 that the component of p^n in $\mathrm{Ker}\, B^t$ is in fact constant, and therefore equal to that of p^o.

Under these conditions the three algorithms considered in Section 3.1 become as follows:

① *Steepest descent algorithm:*

We have $w_n = g_n = \mathcal{A} p^n - \beta$; thus

$$w_n = BA_r^{-1}B^t p^n - BA_r^{-1} b = BA_r^{-1}(B^t p^n - b). \tag{3.19}$$

It follows from (2.4) that we have

$$A_r u^n = b - B^t p^n, \tag{3.20}$$

hence

$$w_n = - Bu^n. \tag{3.21}$$

The optimal value of ρ_n is then, from (3.8),

$$\rho_n = - \frac{|Bu^n|^2}{(Bu^n, BA_r^{-1}B^t g_n)} . \tag{3.22}$$

The calculation of ρ_n thus necessitates the solution, with respect to z_n, of the linear system

$$A_r z_n = B^t g^n = - B^t Bu^n , \tag{3.23}$$

hence

$$\rho_n = - \frac{|Bu^n|^2}{(Bu^n, Bz_n)} .$$

In addition, with p^{n+1} known, we note that u^{n+1} satisfies

$$A_r u^{n+1} = b - B^t p^{n+1} = b - B^t p^n - \rho_n B^t Bu^n = A_r u^n + \rho_n A_r z_n , \tag{3.24}$$

hence

$$u^{n+1} = u^n + \rho_n z_n. \tag{3.25}$$

It follows from the above formulae that at *each iteration* we have to solve only *a single* linear system, with matrix A_r. We can, then, re-express algorithm (2.1)-(2.3) in the following equivalent form:

(3.26) p^o *arbitrarily specified in* $\mathbb{R}^M$, *and* $u^o = A_r^{-1}(b - B^t p^o)$;

with p^n, u^n *known, determine* z_n, ρ_n, p^{n+1}, u^{n+1} *by*

$$z_n = - A_r^{-1} B^t B u^n , \tag{3.27}$$

$$\rho_n = - \frac{|Bu^n|^2}{(Bu^n, Bz_n)} , \tag{3.28}$$

$$p^{n+1} = p^n + \rho_n Bu^n , \tag{3.29}$$

$$u^{n+1} = u^n + \rho_n z_n . \tag{3.30}$$

② *Minimum-residual algorithm.*

We have shown in Section 3.1 that, in the case of the minimum residual algorithm, ρ_n is given by

$$\rho_n = \frac{(\mathcal{A} g_n, g_n)}{|\mathcal{A} g_n|^2}$$

hence the following variant of algorithm (3.26)-(3.30):

(3.31) $p^o \in \mathbb{R}^M$ *arbitrarily specified, and* $u^o = A_r^{-1}(b - B^t p^o)$, *then with* p^n, u^n *known, determine* z_n, ρ_n, p^{n+1}, u^{n+1} *by*

$$z_n = - A_r^{-1} B^t B u^n \tag{3.32}$$

$$\rho_n = - \frac{(Bu^n, Bz_n)}{|Bz_n|^2} , \tag{3.33}$$

$$p^{n+1} = p^n + \rho_n Bu^n , \tag{3.34}$$

$$u^{n+1} = u^n + \rho_n z_n . \tag{3.35}$$

③ *Conjugate-gradient algorithm.*

From Section 3.1, at each iteration a descent direction w_n, conjugate to w_{n-1}, has to be calculated; thus

$$w_n = g_n + \lambda_n w_{n-1} = - Bu^n + \lambda_n w_{n-1} , \tag{3.36}$$

the value of λ_n being given by

$$\lambda_n = \frac{|g_n|^2}{|g_{n-1}|^2} = \frac{|Bu^n|^2}{|Bu^{n-1}|^2} . \tag{3.37}$$

The descent direction w_n being known, we then need to calculate

$$\rho_n = \frac{|g_n|^2}{(g_n, \mathcal{A} w_n)} \quad \left(= \frac{|g_n|^2}{(\mathcal{A} g_n, g_n)} \right) ; \tag{3.38}$$

to do this we introduce, as in the two preceding algorithms, z_n such that

$$z_n = A_r^{-1} B^t w_n, \tag{3.39}$$

and we again have

$$\rho_n = - \frac{|Bu^n|^2}{(Bu^n, Bz_n)} . \tag{3.40}$$

In summary, the conjugate-gradient algorithm can thus be written:

$$p^o \text{ specified arbitrarily in } \mathbb{R}^M; \tag{3.41}$$

$$u^o = A_r^{-1}(b - B^t p^o); \tag{3.42}$$

at iteration n *calculate the descent direction* w_n *by*

$$w_o = g_o = -Bu^o \text{ if } n=0, \tag{3.43}$$

$$w_n = - Bu^n + \lambda_n w_{n-1} \text{ if } n \geq 1, \tag{3.44}$$

$$\lambda_n = \frac{|Bu^n|^2}{|Bu^{n-1}|^2} \quad \text{if } n \geq 1, \tag{3.45}$$

then

$$z_n = A_r^{-1} B^t w_n, \tag{3.46}$$

$$\rho_n = - \frac{|Bu^n|^2}{(Bu^n, Bz_n)} , \tag{3.47}$$

$$p^{n+1} = p^n - \rho_n w_n, \tag{3.48}$$

$$u^{n+1} = u^n + \rho_n z_n. \tag{3.49}$$

■

Remark 3.1: In the three algorithms described above, we have to solve at each iteration only a single linear system, with matrix A_r. Compared with algorithm (2.1)-(2.3) used with a fixed ρ, these algorithms require the additional presence in memory of z^n, a vector which is of the same order as u^n, having N components. In

the case of the conjugate-gradient algorithm, we also need to retain w_{n-1} in memory. This increased memory requirement will be justified if the automatic determination of the step length ρ_n leads to a very clear improvement in the speed of convergence compared with algorithm (2.1)-(2.3) used with $\rho_n = \rho = r$; *this, however, does not always appear to be the case.* ∎

Remark 3.2: In the case of the problem

$$\left\{ \begin{array}{l} J_r^*(p) \leqslant J_r^*(q) \quad \forall q \in \mathbb{R}^M, \\ p \in \mathbb{R}^M \end{array} \right. \tag{3.50}$$

the conjugate-gradient algorithm, i.e. (3.41)-(3.49), converges *theoretically* in N_1 (= rank B) iterations at most. Given that *rounding errors* are present, this result no longer holds in practice. Furthermore, bearing in mind the large size of problems arising from the discretisation of partial differential equations, it is desirable that, with an adequate termination test, convergence should be obtained in a number of iterations *considerably less* than N_1. This property will depend essentially on the *condition number of* $BA_r^{-1}B^t$ *restricted to* Im B, this quantity being henceforth denoted by $\mathcal{K}(BA_r^{-1}B^t)_{ImB}$. It can be shown that

$$\lim_{r \to +\infty} \mathcal{K}(BA_r^{-1}B^t)_{ImB} = 1. \tag{3.51}$$

We note that

$$BA_r^{-1}B^t = B(I+rA^{-1}B^tB)A^{-1}B^t \tag{3.52}$$

and that the matrix corresponding to the case $r = 0$ is $BA^{-1}B^t$. It follows from these properties that the replacement of the Lagrangian $\mathcal{L}$ defined by (1.4) by the *augmented* Lagrangian $\mathcal{L}_r$, defined by (1.8), may be considered as a method of *preconditioning*, in a sense close to that of AXELSSON [1], the preconditioning matrix being $(I+rA^{-1}B^tB)$; this remark is true not only for the conjugate-gradient algorithm but also for all the other algorithms studied in the preceding sections. In particular, in view of this preconditioning, it would seem that it should not be necessary to carry out a reinitialisation (of the type $w_n = g_n$) in the conjugate-gradient algorithm, in order to counteract the accumulation of rounding errors. ∎

4. ON CERTAIN VARIANTS OF THE METHODS OF SECTION 2: INTRODUCTION OF A RELAXATION PARAMETER; METHOD OF ARROW-HURWICZ.

4.1 Synopsis

We shall show in Section 4.2 that we can *improve the speed of convergence* of algorithm (2.1)-(2.3) by utilising a *relaxation parameter* ω; this leads us to the algorithm

(4.1) $u^0 \in \mathbb{R}^N$, $p^0 \in \mathbb{R}^M$ *arbitrarily specified; with* u^n, p^n *known, calculate* $u^{n+\frac{1}{2}}$, *then* u^{n+1} *and* p^{n+1} *by*

(4.2) $$\begin{cases} \mathcal{L}_r(u^{n+\frac{1}{2}}, p^n) \leq \mathcal{L}_r(v, p^n) \quad \forall v \in \mathbb{R}^N, \\ u^{n+\frac{1}{2}} \in \mathbb{R}^N, \end{cases}$$

(4.3) $$u^{n+1} = u^n + \omega(u^{n+\frac{1}{2}} - u^n),$$

(4.4) $$p^{n+1} = p^n + \rho B u^{n+1};$$

(4.2) is equivalent to

(4.5) $$(A + rB^tB)u^{n+\frac{1}{2}} + B^t p^n = b.$$

For $\omega = 1$, we once again have (with a slightly different notation) algorithm (2.1)-(2.3) with $\rho_n = \rho$, $\forall n$[5].

In the sections following Section 4 we shall be studying algorithms of the type

(4.6) $u^0 \in \mathbb{R}^N$, $p^0 \in \mathbb{R}^M$ *arbitrarily specified;*

(4.7) $$u^{n+1} = u^n - \omega_n S_r^{-1} \left((A + rB^tB)u^n + B^t p^n - b \right),$$

(4.8) $$p^{n+1} = p^n + \rho_n B u^{n+1},$$

where, in (4.7), the auxiliary operator S_r is *symmetric* and *positive definite*. We observe that (4.6)-(4.8) is an algorithm of the *Arrow-Hurwicz* type (see ARROW-HURWICZ-UZAWA [1], G.L.T. [1,

[5] We can also consider parameters ω_n and ρ_n which vary with n.

Chapter 2], EKELAND-TEMAM [1], etc...). It should also be noted that (4.1)-(4.4) is a particular case of (4.6)-(4.8) obtained by taking $S_r = A+rB^tB$ $(= A_r)$ and constant values for ω_n and ρ_n.

4.2 Study of algorithm (4.1)-(4.4)

For *given* r, we now study the convergence of algorithm (4.1)-(4.4) and in particular the optimal choice of the parameters ω and ρ. Proceeding as in Section 2.3, we put

$$(4.9)\qquad \begin{cases} \bar{u}^n = u^n - u\ ,\ \bar{u}^{n+\frac{1}{2}} = u^{n+\frac{1}{2}} - u, \\ \bar{p}^n = p^n - (\hat{p}+P_2(p^o))\ ,\ y^n = A^{-1}B^t\bar{p}^n. \end{cases}$$

We then have

$$(4.10)\qquad (A+rB^tB)\bar{u}^{n+\frac{1}{2}} + B^t\bar{p}^n = 0,$$

$$(4.11)\qquad \bar{u}^{n+1} = \bar{u}^n + \omega(\bar{u}^{n+\frac{1}{2}} - \bar{u}^n),$$

$$(4.12)\qquad \bar{p}^{n+1} = \bar{p}^n + \rho B\bar{u}^{n+1}.$$

By elimination of $\bar{u}^n$ and $\bar{u}^{n+\frac{1}{2}}$ we deduce from (4.9)-(4.12)

$$(4.13)\qquad y^{n+1} + (\rho\omega\, A^{-1}B^tB(I+rA^{-1}B^tB)^{-1} + (\omega-2)I)y^n + (1-\omega)y^{n-1} = 0.$$

By means of (4.13) we have reduced the study of the convergence of algorithm (4.1)-(4.4) to the study of the convergence of a *doubly recurrent* sequence. In fact, taking into account the relations (4.9)-(4.12), the convergence of y^n to zero implies that of $\{\bar{u}^n, \bar{p}^n\}$ to $\{0, 0\}$ (see Sections 2.2, 2.3).

Proceeding as in Section 2.3, we now express (4.13) on a *basis* of $\mathrm{Im}A^{-1}B^t$ formed from the *eigenvectors* of $A^{-1}B^tB$ *associated with the strictly positive eigenvalues;* thus, with the notation of Section 2.3, we have

$$(4.14)\qquad \begin{cases} y_i^{n+1} - 2\left(1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right)\right)y_i^n + (1-\omega)\, y_i^{n-1} = 0 \\ \forall i = 1, \ldots, N_1. \end{cases}$$

A necessary and sufficient condition of convergence for (4.1)-(4.4)

$\forall \{u^o, p^o\} \in \mathbb{R}^N \times \mathbb{R}^M$ *will thus be that the roots of the characteristic equation*

$$\xi^2 - 2\left(1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right)\right)\xi + 1 - \omega = 0 \tag{4.15}$$

associated with (4.14) *be of modulus strictly less than* 1, $\forall i = 1, \ldots N_1$. The *convergence rate* R will satisfy

$$R \leqslant \operatorname*{Max}_{1 \leqslant i \leqslant N_1} \operatorname{Max}\left(|\xi_i^+|, |\xi_i^-|\right), \tag{4.16}$$

where ξ_i^+, ξ_i^- are the roots of (4.15), namely

$$\begin{cases} \xi_i^+ = 1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right) + \sqrt{\left(1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right)\right)^2 + \omega - 1}\ , \\ \xi_i^- = 1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right) - \sqrt{\left(1 - \frac{\omega}{2}\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right)\right)^2 + \omega - 1}\ . \end{cases} \tag{4.17}$$

We are now going to study, with *given* ρ and with r still *fixed*, the behaviour of $|\xi_i^+|$ and $|\xi_i^-|$ as a function of ω.

The behaviour of these two roots is particularly straightforward in the case where they are imaginary (conjugates); in fact we then have

$$|\xi_i^+| = |\xi_i^-| = \sqrt{1-\omega}\ . \tag{4.18}$$

We shall have this situation if, in (4.17), the quantity under the square-root sign is *negative*, which is the case if

$$\omega < \omega_i^*(\rho) = 1 - \left(\frac{1+(r-\rho)\lambda_i}{1+(r+\rho)\lambda_i}\right)^2 . \tag{4.19}$$

For $\omega > \omega_i^*(\rho)$, both roots are real and we deduce from (4.17) that the graphs of ξ_i^+ and ξ_i^- are arcs of hyperbolas, respectively *asymptotic* in the (ξ, ω) plane to the straight lines whose equations are

$$\xi = \frac{1}{1 + \dfrac{\rho\lambda_i}{1+r\lambda_i}} \qquad \text{and} \qquad \xi = -\left(1 + \frac{\rho\lambda_i}{1+r\lambda_i}\right)\omega + 2 - \frac{1}{1 + \dfrac{\rho\lambda_i}{1+r\lambda_i}}\ .$$

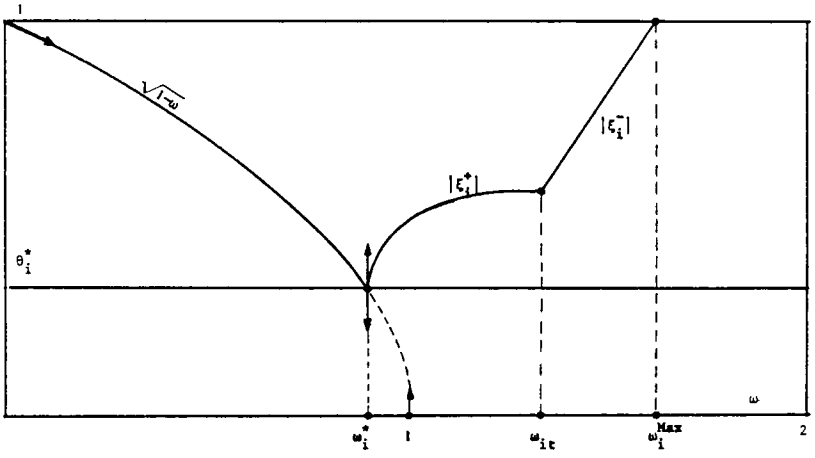


Figure 4.1 ($\rho < r + \frac{1}{\lambda_i}$)

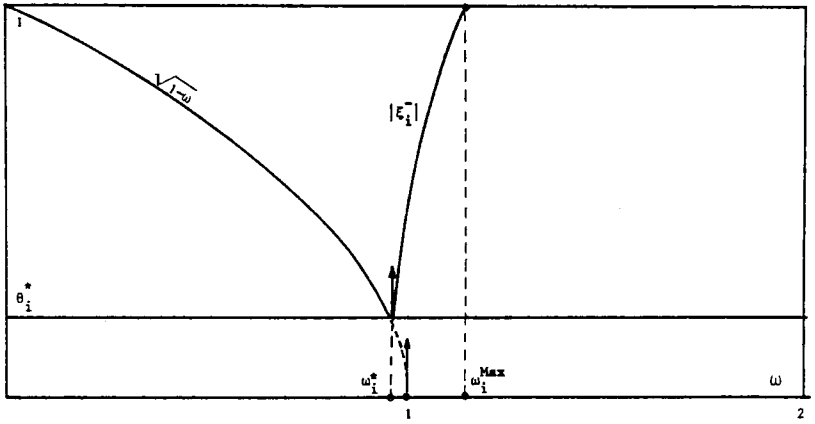


Figure 4.2 ($\rho > r + \frac{1}{\lambda_i}$)

We have shown in Figures 4.1 and 4.2 the behaviour of $\max(|\xi_i^+|, |\xi_i^+|)$ as a function of ω, when, respectively, $\rho < r + \frac{1}{\lambda_i}$ and $\rho > r + \frac{1}{\lambda_i}$ (the scalar ω_{it} in Figure 4.1 has the value

$$2\left(1 - \frac{\rho\lambda_i}{1+(\rho+r)\lambda_i}\right) .$$

It can easily be shown that $|\xi_i^-| > 1$ if $\omega > \omega_i^{Max}(\rho)$ where

$$\omega_i^{Max}(\rho) = 2\,\frac{1 + r\lambda_i}{1+(r+\frac{\rho}{2})\lambda_i} \quad . \tag{4.20}$$

It can then be shown (see also the above figures) that, for given ρ and λ_i, the convergence rate is optimal if $\omega = \omega_i^*$, and that it is then equal to

$$\theta_i^*(\rho) = \sqrt{1-\omega_i^*(\rho)} = \left|\frac{1+(r-\rho)\lambda_i}{1+(r+\rho)\lambda_i}\right| . \tag{4.21}$$

If we plot, as a function of ρ, the convergence rates θ_i^* corresponding to the different eigenvalues λ_i, we obtain, from (4.21), the graphs in Figure 4.3.

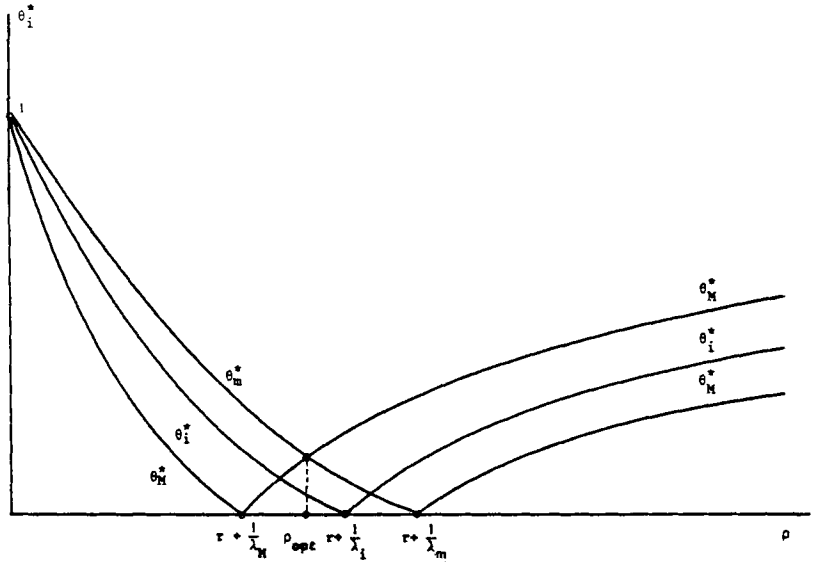


Figure 4.3

It then follows that for a given r, the convergence rate will be optimal if $\theta_M^*(\rho) = \theta_m^*(\rho)$; this will be the case if $\rho = \rho_{opt}$ where

$$\rho_{opt} = \sqrt{(r + \frac{1}{\lambda_M})(r + \frac{1}{\lambda_m})} . \tag{4.22}$$

Substituting this optimal value of ρ into (4.19), with $\lambda_i = \lambda_M$ (or $\lambda_i = \lambda_m$), we obtain for ω the optimal value

$$\omega_{opt} = 1 - \left(\frac{1 + (r-\rho_{opt})\lambda_M}{1 + (r+\rho_{opt})\lambda_M}\right)^2 = 1 - \left(\frac{1 - \sqrt{\frac{r\lambda_m + \frac{\lambda_m}{\lambda_M}}{r\lambda_m + 1}}}{1 + \sqrt{\frac{r\lambda_m + \frac{\lambda_m}{\lambda_M}}{r\lambda_m + 1}}}\right)^2 \tag{4.23}$$

which gives us the optimal convergence rate

$$\theta_{opt} = \frac{1 - \sqrt{\frac{r\lambda_m + \frac{\lambda_m}{\lambda_M}}{r\lambda_m + 1}}}{1 + \sqrt{\frac{r\lambda_m + \frac{\lambda_m}{\lambda_M}}{r\lambda_m + 1}}} . \tag{4.24}$$

Remark 4.1: For $\rho = \rho_{opt}$ and $\omega = \omega_{opt}$, the asymptotic rate of convergence is better than the optimal convergence rate of algorithm (2.1)-(2.3), which corresponds to the case $\omega = 1$. In particular, for $r = 0$, the rates are respectively, (writing $\alpha = \lambda_m/\lambda_M$), $\frac{1-\alpha}{1+\alpha}$ in the case $\omega = 1$ and $\frac{1-\sqrt{\alpha}}{1+\sqrt{\alpha}}$ in the case $\omega = \omega_{opt}$. ∎

Remark 4.2: For given ρ, the best choice of ω, denoted by $\omega^*(\rho)$, is given, if $\rho \le \rho_{opt}$, by

$$\omega^*(\rho) = \omega_m^*(\rho) = 1 - \left(\frac{1 + (r-\rho)\lambda_m}{1 + (r+\rho)\lambda_m}\right)^2 , \tag{4.25}$$

giving the convergence rate

$$\theta^*(\rho) = \theta_m^*(\rho) = \left|\frac{1 + (r-\rho)\lambda_m}{1 + (r+\rho)\lambda_m}\right| , \tag{4.26}$$

and for $\rho \geq \rho_{opt}$, by

$$\omega^*(\rho) = \omega_M^*(\rho) = 1 - \left(\frac{1 + (r-\rho)\lambda_M}{1 + (r+\rho)\lambda_M}\right)^2 , \tag{4.27}$$

giving the convergence rate

$$\theta^*(\rho) = \theta_M^*(\rho) = \left|\frac{1 + (r-\rho)\lambda_M}{1 + (r-\rho)\lambda_M}\right| . \tag{4.28}$$

In all cases, the maximal value of ω (i.e. the value above which the algorithm diverges) is given by

$$\omega^{Max}(\rho) = \omega_M^{Max}(\rho) = 2\,\frac{1 + r\lambda_M}{1 + (r+\frac{\rho}{2})\lambda_M} . \tag{4.29}$$

Figure 4.4 shows the graphs of $\omega^*(\rho)$ and of $\omega^{Max}(\rho)$.

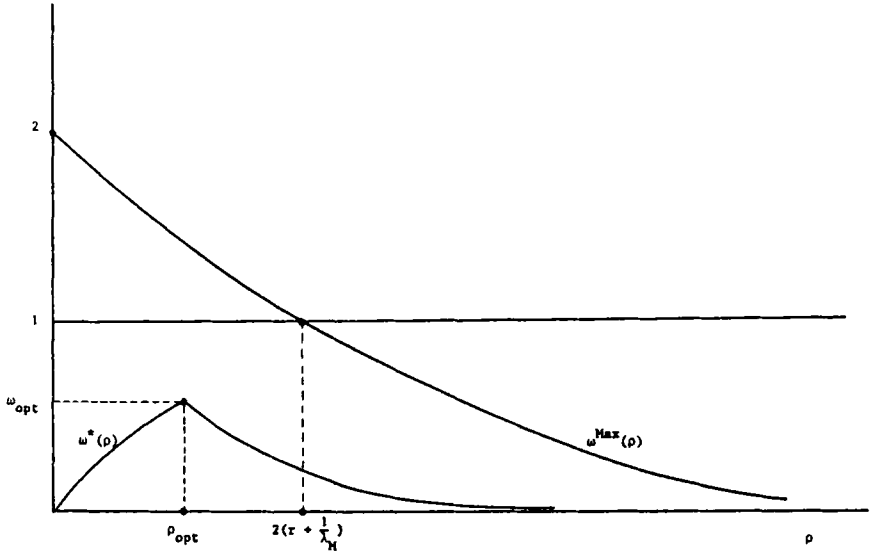


Figure 4.4

We note that if $\rho > 2(r + \frac{1}{\lambda_M})$ we have $\omega^{Max}(\rho) < 1$. This corresponds to the fact that for $\omega = 1$ the maximal value of ρ is precisely $2(r + \frac{1}{\lambda_M})$ (see Section 2.2). Further, if $3r < \frac{1}{\lambda_M} - \frac{4}{\lambda_M}$, we have $\rho_{opt} > 2(r + \frac{1}{\lambda_M})$.

This will be the case, in particular, for $r = 0$, when $\frac{\lambda_M}{\lambda_m} > 4$. This point clearly illustrates the fact that the introduction of the parameter ω guarantees the convergence of (4.1)-(4.4) for values of ρ greater than the bounds determined in Section 2. In fact, we can obtain convergence of (4.1)-(4.4) for any positive value of ρ, as long as we take ω sufficiently small. ∎

Remark 4.3: Algorithm (4.1)-(4.4) can be written in the form

$$u^{n+1/2} = A_r^{-1}(b - B^t p^n), \tag{4.30}$$

$$p^{n+1} = p^n + \rho(\omega B u^{n+1/2} + (1-\omega) B u^n). \tag{4.31}$$

Observing that, with the notation of Section 3, $-Bu^{n+\frac{1}{2}}$ is in fact the gradient g_n of J^* at p^n, we deduce from (4.31) that:

$$p^{n+1} = p^n - \rho\omega(g_n - \frac{(1-\omega)}{\rho\omega}(p^n - p^{n-1})) \tag{4.32}$$

since $Bu^n = \frac{1}{\rho}(p^n - p^{n-1})$.

Writing $\tilde{\rho} = \rho\omega$ and $\lambda = \frac{(1-\omega)}{\rho\omega}$, we then obtain

$$p^{n+1} = p^n + \tilde{\rho}(g_n + \lambda(p^n - p^{n-1})). \tag{4.33}$$

We thus obtain, for parameters suitably chosen and *depending on* n, the conjugate-gradient method of Section 3. We could also use in this algorithm variable parameters corresponding to a semi-iterative Chebychev method. ∎

Remark 4.4: It is possible to once more re-write (4.31), in the form

$$p^{n+1} = p^{n-1} - (2-\omega)(\frac{\rho\omega}{2-\omega} g_n - p^n + p^{n-1}). \tag{4.34}$$

Writing $\tilde{\omega} = (2-\omega)$ and $\alpha = \frac{\rho\omega}{2-\omega}$, we obtain

$$p^{n+1} = p^{n-1} - \tilde{\omega}(\alpha g_n - p^n + p^{n-1}). \tag{4.35}$$

Algorithm (4.1)-(4.4) is thus equivalent to the *two-step Richardson algorithm* applied to the minimisation of the dual functional J^*.

It can be shown that, subject to making the appropriate changes of notation, the optimal parameters given in (4.22)-(4.23) correspond exactly to the standard optimal parameters in Richardson's method (see GOLUB [1]). ∎

4.3 Study of algorithm (4.6)-(4.8)

4.3.1 *General notes*

In this section we shall finish off Remark 2.5, since in fact algorithm (4.6)-(4.8) corresponds to the variant of algorithm (2.1)-(2.3) obtained when, in the calculation of u^{n+1}, we use only *a single* iteration of a gradient-type algorithm (with the auxiliary operator S_r), starting from u^n. We recall also (see Section 4.2) that (4.1)-(4.4) is a particular case of (4.6)-(4.8) corresponding to $S_r = A_r$. In the case where $S_r = I$, we get back to the standard method, called the *Arrow-Hurwicz* method, introduced in ARROW-HURWICZ-UZAWA [1]. In the case where $\omega_n/\rho_n = \beta$, $\forall n$, we recall that this method consists of searching for the saddle-points of the Lagrangian $\mathcal{L}_r$ *via the approximate integration of the differential system*

$$\text{(4.36)} \qquad \begin{cases} S_r \dfrac{du}{dt} + \beta \dfrac{\partial \mathcal{L}_r}{\partial u}(u,p) = 0, \\ \dfrac{dp}{dt} = \dfrac{\partial \mathcal{L}_r}{\partial p}(u,p). \end{cases}$$

In the case where we discretise using an *Euler type* scheme, we obtain a scheme which is slightly different from (4.6)-(4.8), namely

$$\text{(4.37)} \qquad u^{n+1} = u^n - \omega_n S_r^{-1}\left((A+rB^tB)u^n + B^tp^n - b\right),$$

$$\text{(4.38)} \qquad p^{n+1} = p^n + \rho_n Bu^n.$$

The scheme (4.6)-(4.8) can thus be considered as a "semi-implicit" variant of the Euler scheme (4.37), (4.38).

4.3.2 *Reduction of* (4.6)-(4.8) *to the discrete form of a second-order differential system*

We shall assume for simplicity that $\omega_n = \omega$, $\rho_n = \rho$, $\forall n$. We then deduce from (4.7), by subtraction,

$$u^{n+1}-2u^n+u^{n-1} + \omega S_r^{-1}\left(A_r(u^n-u^{n-1})+B^t(p^n-p^{n-1})\right) = 0 \quad ,$$

then from (4.8)

$$p^n-p^{n-1} = \rho B u^n \quad ,$$

hence

$$u^{n+1}-2u^n+u^{n-1} + \omega S_r^{-1} A_r(u^n-u^{n-1})+\rho\omega S_r^{-1}B^tBu^n = 0. \tag{4.39}$$

Again writing $\bar{u}^n = u^n - u$ and noting that $Bu = 0$, we deduce from (4.39) that

$$\bar{u}^{n+1}-2\bar{u}^n+\bar{u}^{n-1} + \omega S_r^{-1}A_r(\bar{u}^n-\bar{u}^{n-1})+\rho\omega S_r^{-1}B^tB\bar{u}^n = 0. \tag{4.40}$$

We observe that (4.39) is a discretised form, with $\Delta t = 1$ and using an explicit scheme, of the second-order differential system

$$\frac{d^2u}{dt^2} + \omega S_r^{-1}A_r\frac{du}{dt} + \rho\omega S_r^{-1}B^tBu = 0 \quad . \tag{4.41}$$

We note the presence of the damping term $\omega S_r^{-1}A_r\frac{du}{dt}$ and furthermore that, other things being equal, the natural frequencies of the undamped system grow with $\rho\omega$. Looking at (4.40) and (4.41), we may expect that the behaviour of $\bar{u}^n = u^n - u$, as a function of n, will depend in complicated fashion on the parameters ω and ρ. We recall that even in the case of the scalar equation

$$\ddot{x} + 2k\dot{x} + \omega^2 x = 0 \quad ,$$

the behaviour of the solution, as $t \to +\infty$, brings in the notion of *critical damping*. A priori, when $S_r^{-1}A_r$ and $S_r^{-1}B^tB$ do not commute the study of the behaviour of $\bar{u}^n$ as a function of n, by *spectral methods*, (see Section 4.2) would seem impracticable for the moment; therefore in the following section, we shall use *energy methods* to study the convergence to zero of $\bar{u}^n$.

4.3.3 *Convergence of algorithm* (4.6)-(4.8).

We shall now seek conditions on ρ and ω which assure the convergence of (4.6)-(4.8); we first define some notation.

With the standard Euclidian inner product still written as $(.,.)$, we associate with the symmetric positive-definite operator S_r the norm

$$|v|_{S_r}^2 = (S_r v, v) \ . \tag{4.42}$$

We likewise write

$$\|v\|_r^2 = (A_r v, v). \tag{4.43}$$

We then have, by the equivalence of the norms on $\mathbb{R}^N$, the existence of a constant α_r such that

$$\|v\|_r^2 \leq \alpha_r |v|_{S_r}^2 \ . \tag{4.44}$$

The operator B being continuous, we have furthermore (with $|q|^2 = (q,q)$)

$$|Bv|^2 \leq \beta_r \|v\|_r^2 \tag{4.45}$$

and

$$|Bv|^2 \leq \gamma_r |v|_{S_r}^2 \ . \tag{4.46}$$

We next define $\bar{u}^n$ and $\bar{p}^n$ as in (4.9) and we establish that $\bar{u}^n$, $\bar{p}^n$ satisfy the equations

$$(\bar{u}^{n+1}-\bar{u}^n, S_r v) + \omega(A_r \bar{u}^n, v) + \omega(\bar{p}^n, Bv) = 0 \quad \forall v \in \mathbb{R}^N \ , \tag{4.47}$$

$$(\bar{p}^{n+1}-\bar{p}^n, q) - \rho(B\bar{u}^{n+1}, q) = 0 \quad \forall q \in \mathbb{R}^M \ . \tag{4.48}$$

In the following text, we shall for simplicity denote $\bar{u}^n$ by u^n and $\bar{p}^n$ by p^n.

We then set $v = u^n$ in (4.47) and we multiply by 2ρ; we obtain

$$\rho|u^{n+1}|_{S_r}^2 - \rho|u^n|_{S_r}^2 - \rho|u^{n+1}-u^n|_{S_r}^2 + 2\rho\omega\|u^n\|_r^2 + 2\rho\omega(p^n, Bu^n) = 0 \ . \tag{4.49}$$

Furthermore, from (4.48), we have

$$2\rho\omega(p^n, Bu^n) = 2\omega(p^n, p^n - p^{n-1}) = \omega|p^n|^2 - \omega|p^{n-1}|^2 + \omega|p^n - p^{n-1}|^2,$$

and by substituting in (4.49) we thereby deduce

$$(4.50)\quad \left\{ \begin{array}{l} \rho|u^{n+1}|^2_{S_r} - \rho|u^n|^2_{S_r} + 2\rho\omega \|u^n\|^2_r + \omega|p^n|^2 - \omega|p^{n-1}|^2 + \omega|p^n - p^{n-1}|^2 = \\ = \rho|u^{n+1} - u^n|^2_{S_r} . \end{array} \right.$$

We now have to *obtain an appropriate upper bound for the right-hand side; from* (4.47) we have

$$(4.51)\qquad \rho|u^{n+1} - u^n|^2_{S_r} = -\rho\omega(A_r u^n, u^{n+1} - u^n) - \rho\omega(p^n, Bu^{n+1} - Bu^n) ,$$

then, by a variant of the Cauchy-Schwarz inequality and from (4.44), we have

$$(4.52)\quad |\rho\omega(A_r u^n, u^{n+1} - u^n)| \le \rho\varepsilon|u^{n+1} - u^n|^2_{S_r} + \frac{\rho\omega^2}{4\varepsilon}\alpha_r \|u^n\|^2_r , \quad \forall \varepsilon > 0.$$

We further deduce from (4.48) that

$$(4.53)\quad \left\{ \begin{array}{l} -\rho\omega(p^n, Bu^{n+1} - Bu^n) = -\omega(p^n, p^{n+1} - p^n) + \omega(p^n, p^n - p^{n-1}) = \\ = -\frac{\omega}{2}(|p^{n+1}|^2 - 2|p^n|^2 + |p^{n-1}|^2) + \frac{\omega}{2}(|p^{n+1} - p^n|^2 + |p^n - p^{n-1}|^2) . \end{array} \right.$$

Substituting (4.52), (4.53) into (4.51) we obtain

$$(4.54)\quad \left\{ \begin{array}{l} \rho|u^{n+1} - u^n|^2_{S_r} \le \frac{\rho\omega^2\alpha_r}{4\varepsilon(1-\varepsilon)} \|u^n\|^2_r - \frac{\omega}{2(1-\varepsilon)}(|p^{n+1}|^2 - 2|p^n|^2 + |p^{n-1}|^2) + \\ + \frac{\omega}{2(1-\varepsilon)}(|p^{n+1} - p^n|^2 + |p^n - p^{n-1}|^2) , \ \forall\ 0 < \varepsilon < 1. \end{array} \right.$$

Next, replacing the right-hand side of (4.50) by the upper bound obtained in (4.54) and noting from (4.48) that $|p^{n+1} - p^n|^2 = \rho^2|Bu^n|^2$, we obtain, after re-grouping various terms

$$(4.55)\quad \left\{ \begin{array}{l} \rho|u^{n+1}|^2_{S_r} - \rho|u^n|^2_{S_r} + (2\rho\omega - \frac{\rho\omega^2\alpha_r}{4\varepsilon(1-\varepsilon)}) \|u^n\|^2_r + \omega|p^n|^2 - \omega|p^{n-1}|^2 + \\ + \frac{\omega}{2(1-\varepsilon)}(|p^{n+1}|^2 - 2|p^n|^2 + |p^{n-1}|^2) - \frac{\omega}{2(1-\varepsilon)}(|p^{n+1} - p^n|^2 - |p^n - p^{n-1}|^2) - \\ - \frac{\omega\varepsilon\rho^2}{(1-\varepsilon)}|Bu^n|^2 \le 0. \end{array} \right.$$

Utilising (4.45) we then have

$$(4.56)\quad \left\{ \begin{array}{l} \rho|u^{n+1}|^2_{S_r} - \rho|u^n|^2_{S_r} + (2\rho\omega - \dfrac{\rho\omega^2\alpha_r}{4\varepsilon(1-\varepsilon)} - \dfrac{\beta_r\omega\varepsilon\rho^2}{(1-\varepsilon)})\,\|u^n\|^2_r + \\ \\ + \omega\,|p^n|^2 - \omega|p^{n-1}|^2 + \dfrac{\omega}{2(1-\varepsilon)}\,(|p^{n+1}|^2 - 2|p^n|^2 + |p^{n-1}|^2) - \\ \\ - \dfrac{\omega}{2(1-\varepsilon)}\,|p^{n+1}-p^n|^2 + \dfrac{\omega}{2(1-\varepsilon)}\,|p^n - p^{n-1}|^2 \le 0 \quad . \end{array} \right.$$

The coefficient of $\|u^n\|^2_r$ will be positive if for some ε, which remains to be determined, we have

$$(4.57)\quad \frac{\omega\alpha_r}{4\varepsilon(1-\varepsilon)} + \rho\beta_r\,\frac{\varepsilon}{1-\varepsilon} < 2 \quad .$$

To define this condition precisely, we now try to find, for given ρ, the value of ε in $]0,1[$ giving the best bound for ω (i.e. the largest possible). An elementary calculation shows us that the optimal choice for ε is,

$$(4.58)\quad \varepsilon^* = \frac{1}{2+\rho\beta_r} \quad ,$$

which together with (4.57) implies

$$(4.59)\quad \omega\alpha_r < \frac{2}{(1+\frac{\rho}{2}\,\beta_r)} \quad ,$$

or further

$$(4.60)\quad \omega < \frac{2}{\alpha_r(1+\frac{\rho}{2}\,\beta_r)} = \omega_{Max}(\rho) \, .$$

If condition (4.59) is satisfied, we have

$$(4.61)\quad C = 2\rho\omega - \frac{\rho\omega^2\alpha_r}{4\varepsilon^*(1-\varepsilon^*)} - \frac{\beta_r\omega\varepsilon^*\rho^2}{1-\varepsilon^*} > 0 \quad .$$

Summing the inequalities (4.56) for $n = 1,\ldots,\bar{N}$, we obtain

$$(4.62)\quad \left\{ \begin{array}{l} \rho|u^{\bar N+1}|^2_{S_r} + C\sum_{n=1}^{\bar N}\|u^n\|^2_r + \omega|p^{\bar N}|^2 + \frac{\omega}{2(1-\varepsilon)}\,(|p^{\bar N+1}|^2 - |p^{\bar N}|^2) - \\ \\ - \frac{\omega}{2(1-\varepsilon)}\,|p^{\bar N+1}-p^{\bar N}|^2 \le \rho|u^1|^2_{S_r} + \omega|p^o|^2 + \frac{\omega}{2(1-\varepsilon)}\,(|p^1|^2 - |p^o|^2) - \\ \\ - \frac{\omega}{2(1-\varepsilon)}\,|p^1-p^o|^2 \le \text{Const.} \end{array} \right.$$

The left-hand side contains the term

$$(4.63)\qquad \frac{\omega}{2(1-\varepsilon)}\,(|p^{\bar N+1}|^2 - |p^{\bar N+1}-p^{\bar N}|^2)$$

for which we shall now obtain a lower bound. To do this we write

$$(4.64)\qquad |p^{\bar N+1}|^2 - |p^{\bar N+1}-p^{\bar N}|^2 = -|p^{\bar N}|^2 + 2(p^{\bar N+1},p^{\bar N}) = -|p^{\bar N}|^2 + 2(p^{\bar N}+\rho Bu^{\bar N+1},p^{\bar N})\ ,$$

hence

$$(4.65)\quad \left\{ \begin{array}{l} |p^{\bar N+1}|^2 - |p^{\bar N+1}-p^{\bar N}|^2 = |p^{\bar N}|^2 + 2\rho(Bu^{\bar N+1},p^{\bar N}) \ge (1-\frac{1}{\delta})|p^{\bar N}|^2 - \rho^2\delta|Bu^{\bar N+1}|^2 \\ \\ \forall\delta \in\ ]0,1[\,. \end{array} \right.$$

We then substitute this lower bound into (4.62), regrouping the terms in $|p^{\bar N}|^2$ and utilising (4.46), i.e. $\rho^2\delta|Bu^{\bar N+1}|^2 \le \rho^2\delta\gamma_r|u^{\bar N+1}|^2_{S_r}$. We then have, with δ still to be chosen,

$$(4.66)\qquad \rho\Big(1 - \frac{\rho\omega\delta\gamma_r}{2(1-\varepsilon)}\Big)\,|u^{\bar N+1}|^2_{S_r} + C\sum_{n=1}^{\bar N}\|u^n\|^2_r + \omega\Big(1 - \frac{1}{2\delta(1-\varepsilon)}\Big)|p^{\bar N}|^2 \le \text{Const.}$$

If we then suppose $\varepsilon = \varepsilon^*$, the coefficient of $|p^{\bar N}|^2$ must be positive, which implies

$$(4.67)\qquad \delta \ge \frac{1}{2(1-\varepsilon^*)} = \frac{2+\rho\beta_r}{2+2\rho\beta_r}\ .$$

On the other hand, in order for the coefficient of $|u^{\bar N+1}|^2$ to remain positive, we must have

$$(4.68)\qquad \omega \le \frac{2(1-\varepsilon^*)}{\rho\delta\gamma_r} = \frac{2(1+\rho\beta_r)}{\rho\delta\gamma_r(2+\rho\beta_r)} = \frac{\alpha_r(1+\rho\beta_r)}{2\rho\delta\gamma_r}\ \omega_{\text{Max}}(\rho).$$

Consequently, if we can choose δ so that

$$(4.69)\qquad \frac{\alpha_r(1+\rho\beta_r)}{2\rho\delta\gamma_r} \ge 1$$

then condition (4.68) will in fact be a consequence of (4.60). We therefore have to seek δ satisfying simultaneously (4.67) and (4.69), i.e.

$$(4.70)\qquad \frac{2+\rho\beta_r}{2+2\rho\beta_r} \le \delta \le \frac{\alpha_r(1+\rho\beta_r)}{2\rho\gamma_r}\ .$$

We thus need to show that the interval defined by (4.70) is not empty, i.e. that

$$(4.71)\qquad 4\rho\gamma_r + 2\rho^2\beta_r\gamma_r \le 2\alpha_r(1+\rho\beta_r)^2,$$

or, regrouping the terms, that

$$(4.72)\qquad \beta_r(\gamma_r-\alpha_r\beta_r)\rho^2+2(\gamma_r-\alpha_r\beta_r)\rho - \alpha_r \le 0.$$

It can easily be seen that (4.72) will be satisfied for any ρ if we have

$$(4.73)\qquad \gamma_r \le \alpha_r\beta_r\ .$$

Before confirming this point, we shall conclude the proof of convergence; we have seen in fact, assuming (4.73) is true for the moment, that subject to obeying the condition (4.60) we can choose δ so as to have in the left-hand side of (4.66) only positive terms. It then follows in particular that we have

$$(4.74)\qquad C\sum_{n=1}^{\bar{N}}\|u^n\|_r^2 \le \text{constant, } \textit{for any}\ \bar{N},$$

which implies the convergence of the series with general term $\|u^n\|_r^2$ and therefore that

$$\lim_{n\to\infty} \|u^n\|_r^2 = 0. \tag{4.75}$$

Likewise, it follows from (4.66) that $|p^{\bar{N}}|^2$ is bounded; we can thus extract from p^n a convergent sub-sequence. We shall in fact prove, as in Section 2, that p^n converges to zero (recall that u^n and p^n actually denote $\bar{u}^n$ and $\bar{p}^n$ here). ∎

It now remains for us to prove the inequality (4.73). To do this we must first determine the constants α_r, β_r and γ_r. It can easily be shown that these are respectively the largest eigenvalues of $S_r^{-1}A_r$, $A_r^{-1}B^tB$ and $S_r^{-1}B^tB$. The inequality (4.73) can thus be written, if $\rho(M)$ denotes the spectral radius of the matrix M, as

$$\rho(S_r^{-1}B^tB) \leq \rho(S_r^{-1}A_r)\rho(A_r^{-1}B^tB). \tag{4.76}$$

This inequality is only one particular case of the following general result:

LEMMA 4.1: *If* A *and* B *are two symmetric matrices and* C *is a symmetric positive-definite matrix, we have*

$$\rho(AB) \leq \rho(AC)\rho(C^{-1}B). \tag{4.77}$$

Proof: We in fact have $\rho(AB) = \rho(C^{1/2}ABC^{-1/2})$, the spectral radius being invariant under a change of basis; furthermore

$$\rho(C^{1/2}ABC^{-1/2}) \leq \|C^{1/2}ABC^{-1/2}\| \leq \|C^{1/2}AC^{1/2}\| \, \|C^{-1/2}BC^{-1/2}\|.$$

But $C^{\frac{1}{2}}AC^{\frac{1}{2}}$ is symmetric and we have $\|C^{1/2}AC^{1/2}\| = \rho(C^{1/2}AC^{1/2}) = \rho(AC)$. Similarly $\|C^{-1/2}BC^{-1/2}\| = \rho(C^{-1}B)$; hence the result. ∎

Remark 4.5: In the case (which was studied in Section 4.2) where $S_r = A_r$, the condition (4.59) clearly covers the condition (4.29); thus it is probable that the best possible result has been obtained. ∎

Remark 4.6: It should be noted that (4.61) allows ρ to be taken as large as desired, as long as we take ω sufficiently small. This result may seem paradoxical because the algorithm may be considered as a version of Uzawa's algorithm with an *incomplete* solution. But, in the complete-solution case we have seen in Section 2 that ρ must be taken sufficiently small. ∎

Remark 4.7: It is clear that the speed of the convergence will depend essentially on the choice of the auxiliary operator S_r. The simplest choices are of course $S_r = I$ and $S_r = A_r$. We can also show that in the case where the system $A_r u^n = -B^t p^n + b$ is solved by a method of symmetric successive overrelaxation (SSOR), performing one double pass (forward and back) is equivalent to (4.6) for an operator S_r which can be constructed explicitly and which is clearly symmetric positive-definite. For more details O. AXELSSON [1] can be consulted. ∎

Remark 4.8: The choice of the *optimal parameters* for the algorithm (4.6)-(4.8) is, to the best of our knowledge, an *open problem* (except in the case $S_r = A_r$, see Section 4.2). We refer the reader to Chapter II in which we shall give some information (of *experimental* origin) on this subject, in connection with the numerical solution of the Stokes and the Navier-Stokes equations. ∎

5. MISCELLANEOUS REMARKS AND DISCUSSION

Remark 5.1: All the discussion in Sections 1,2,3 and 4 is still valid if we consider, instead of problem (1.2), the problem

$$\left\{ \begin{array}{l} J(u) \leq J(v) \quad \forall\, v \in K, \\ u \in K \end{array} \right. \tag{5.1}$$

with

$$K = \{v \in \mathbb{R}^N,\ Bv = c\} \quad , c \in \mathrm{Im}B. \tag{5.2}$$

It is actually sufficient to replace

$$\left\{ \begin{array}{l} (A+rB^tB)u^n = b - B^t p^n, \\ p^{n+1} = p^n + \rho_n B u^n, \end{array} \right. \tag{5.3}$$

by

$$\left\{ \begin{array}{l} (A+rB^tB)u^n = b - B^t p^n + rB^t c, \\ p^{n+1} = p^n + \rho_n (Bu^n - c). \end{array} \right. \tag{5.4}$$

∎

Remark 5.2: Suppose $A \in \mathcal{L}(\mathbb{R}^n, \mathbb{R}^n)$ is positive definite, *not necessarily symmetric*, and suppose that K is defined by (5.2). It

can be shown that the variational problem[6]

$$\left\{ \begin{array}{l} (Au,v-u) \geq (b,v-u) \quad \forall v \in K, \\ u \in K \end{array} \right. \tag{5.5}$$

admits one and only one solution characterised by the existence of $p \in \mathbb{R}^M$ such that

$$\left\{ \begin{array}{l} Au + B^t p = b \ , \\ Bu = c. \end{array} \right. \tag{5.6}$$

In view of (5.6), we can apply to the solution of (5.5) the algorithm

$$p^o \in \mathbb{R}^M \quad \textit{chosen arbitrarily;} \tag{5.7}$$

with p^n *known, calculate* u^n, *then* p^{n+1}, *by*

$$(A+rB^tB)u^n = b-B^tp^n + rB^tc, \tag{5.8}$$

$$p^{n+1} = p^n + \rho_n(Bu^n-c) \ , \quad \rho_n \geq 0 \ . \tag{5.9}$$

By proceeding as for Theorem 2.1, it can easily be shown that algorithm (5.7)-(5.9) converges, whatever the value of p^o, subject to the condition that

$$0 < \alpha_o \leq \rho_n \leq \alpha_1 < 2(r + 1/\beta^2) \ , \tag{5.10}$$

where β^2 is defined by

$$\beta^2 = \underset{v \neq 0}{\text{Max}} \ \frac{|Bv|^2}{(A_\sigma v,v)} \ , \tag{5.11}$$

and where in (5.11), A_σ is the symmetric component of A, i.e. $A_\sigma = \frac{1}{2}(A+A^t)$. By contrast, a "finely detailed" study of convergence rates seems much more difficult, since the spectral methods of Section 2.2 cannot then be used. Likewise, the extension to problem (5.5) of the variable step-length and conjugate-gradient methods of Section 3 may pose difficulties; this applies particularly to the conjugate-gradient method.

The proof of the convergence of algorithm (4.6)-(4.8), based on energy equalities and inequalities, extends without too many extra difficulties to the case where in (4.7) A is positive-definite,

---

[6] This is an elementary case of a variational inequality (see for example G.L.T. [1], [2]).

non-symmetric.

In Chapter II we shall utilise, for the solution of the Navier-Stokes equations, algorithms of the type (5.7)-(5.9) and (4.6)-(4.8) with A *non-symmetric*. ∎

Remark 5.3: In certain problems it may be advantageous, as suggested by FLETCHER [1], to use instead of the penalisation term $r|Bv|^2$ a term of the form (RBv,Bv) where the matrix R is symmetric, positive-definite, and "appropriately chosen". It is clearly apparent that the introduction of R, however, will complicate the study of convergence and of the choice of optimal parameters for the algorithms of Sections 2,3, and 4. ∎

Remark 5.4: The augmented Lagrangian methods introduced by HESTENES [1] and POWELL [1] have given rise to a large number of works, which it would be quite impossible to record individually. We thus refer the reader to the titles given below and to the corresponding bibliographies.

We find in ROCKAFELLAR [1], [2], [3], a study of augmented Lagrangian methods, applied to the minimisation of convex and non-convex functionals, with convex inequality constraints (possibly nonlinear). This study introduces the augmented Lagrangian method within the framework of the theory of duality in convex analysis.

We find also in BERTSEKAS [1], [2] a study of the convergence of algorithms closely related to those considered in Section 2, but within a rather more general framework. In KORT-BERTSEKAS [1] there are also to be found other penalisation procedures applied to the construction of augmented Lagrangians of a type different from those considered in this volume. Finally, the reader who desires a general view of solution methods for constrained optimisation problems, including augmented Lagrangian methods, can profitably refer to GILL and MURRAY [1]. ∎

Remark 5.5: In FORTIN [1],[2], G.L.T. [1], [2] and in Chapter II of the present volume there can be found studies and applications of algorithms of the Uzawa and Arrow-Hurwicz types, to the solution of problems which are much more complex than those considered in this chapter. ∎

Remark 5.6: We shall round off Remark 5.1 by considering the case where in (5.2) *we no longer have* $c \in \mathrm{Im}B$. We therefore have $K = \emptyset$ in this case, and problem (5.1) becomes ill-posed since it has no solution. Consider, however, the algorithm

(5.12) $p^0 \in \mathbb{R}^M$ *chosen arbitrarily;*
then for $n \geq 0$, *with* p^n *known, define* u^n *and* p^{n+1} *by*

$$(A+rB^tB)u^n = b+rB^tc - B^tp^n , \tag{5.13}$$

$$p^{n+1} = p^n + \rho_n(Bu^n-c). \tag{5.14}$$

We can show that under the condition

$$0 < \alpha_o \leq \rho_n \leq \alpha_1 < 2r, \tag{5.15}$$

we have

$$\lim_{n\to+\infty} u^n = u^*, \tag{5.16}$$

where u^* is the solution of the problem,

$$\begin{cases} u^* \in K^* = \{v \mid v \in \mathbb{R}^N,\ B^t(Bv-c) = 0\} \\ J(u^*) \leq J(v) \quad \forall v \in K^*, \end{cases} \tag{5.17}$$

where we once again have $J(v) = \frac{1}{2}(Av,v)-(b,v)$.

We know that K^* $(\neq \emptyset)$ is the *set of the solutions of the normal equation*

$$B^tBz = B^tc. \tag{5.18}$$

We can likewise show that the convergence of u^n to u^* is linear (i.e. at least as rapid as that of a geometric sequence with ratio less than 1). As regards the sequence $\{p^n\}_{n \geq 0}$, it follows from (5.14), and from the fact that $c \notin \mathrm{Im}B$, that this diverges like an arithmetic progression. This divergence is "much less rapid" than the convergence of u^n, which means that in practice there will be no risk of "overflow".

The convergence result stated above holds only for r *strictly positive;* the use of a *strictly* augmented Lagrangian is therefore necessary. This also shows the robustness of the methods described

in this chapter, in the presence particularly of rounding errors. In actual fact the condition $c \in \mathrm{Im}B$ can no longer be satisfied exactly because of these errors; nonetheless the above convergence results show that the augmented Lagrangian method remains usable and provides the *best possible result (in the least squares sense)* in this "noisy" environment.

CHAPTER II

APPLICATION TO THE STOKES AND NAVIER-STOKES EQUATIONS

M. Fortin, F. Thomasset

1. INTRODUCTION

1.1 Motivation

The objectives of this chapter are twofold. First, we shall show that the *augmented Lagrangian* method can be applied directly to the solution of certain problems in *Hydrodynamics*. Secondly, we shall illustrate by numerical examples the results of Chapter I, comparing the properties of the different algorithms on specific examples. In particular we shall see, through concrete cases, the importance of the choice of the parameters; on this topic we shall give experimental results for an Arrow-Hurwicz type algorithm (see Chapter I, Section 4.3). These experimental results will be given in the case of the *linearised Stokes equations;* these equations can be written as the optimality conditions of a quadratic programming problem in the sense of Chapter I.

We shall then show that the algorithms used can be extended to the case of the *steady-state nonlinear Navier-Stokes equations.* Finally we shall indicate briefly how similar techniques can be applied in the time-dependent case.

1.2 Statement of the problem

We consider an open domain Ω in $\mathbb{R}^2$ or in $\mathbb{R}^3$, *bounded*, with regular boundary Γ (for example, Lipschitz continuous). The coordinates in $\mathbb{R}^N$ (N = 2 or 3) will be denoted by $x = \{x_1, x_2\}$ or $x = \{x_1, x_2, x_3\}$. We seek to determine in Ω the characteristics of the flow of an *incompressible viscous fluid.*

Thus let

(1.1) $\underset{\sim}{u}(x,t) = \{u_1(x,t), u_2(x,t), u_3(x,t)\}$

denote the *velocity* of the flow and let p be the hydrostatic *pressure*.

We shall now try to calculate solutions of the Navier-Stokes equations which in standard nondimensional form are written:

(1.2) $$\frac{\partial \underset{\sim}{u}}{\partial t} - \frac{1}{Re}\,\Delta \underset{\sim}{u} + (\underset{\sim}{u}\cdot\underset{\sim}{\nabla})\underset{\sim}{u} + \underset{\sim}{\nabla} p = \underset{\sim}{f} \qquad in \quad \Omega \quad (\Delta = \underset{\sim}{\nabla}^2),$$

(1.3) $$\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0 \qquad in \quad \Omega,$$

(1.4) $$\underset{\sim}{u}\big|_{\Gamma} = \underset{\sim}{0},$$

(1.5) $$\underset{\sim}{u}(x,o) = \underset{\sim}{u}_o(x) \qquad in \quad \Omega$$

We consider here the (not very realistic, physically, but simpler to handle) case where the boundary conditions are of homogeneous Dirichlet type, and where the fluid is driven by a distributed external force $\underset{\sim}{f} = \{f_1, f_2, f_3\}$. The transition to more realistic cases poses no problem with regard to the numerical treatment. In the greater part of this chapter we shall concern ourselves with the steady-state case ($\frac{\partial \underset{\sim}{u}}{\partial t} = \underset{\sim}{0}$) and we shall not therefore have to specify an initial condition of the type (1.5).

The *Reynolds number* Re, the reciprocal of which appears in (1.2) in front of the viscosity term $\Delta \underset{\sim}{u}$, plays, as we know, a critical role in determining the behaviour of the solutions. It is written, in general, in the form

(1.6) $$Re = \frac{Vd}{\nu},$$

where V is a reference velocity, d is a reference length and ν is the kinematic viscosity. The constants V and d are chosen in such a way that, for example, the diameter of Ω and the maximum velocity of the flow are of order unity. Let us say immediately that the methods which will be presented here are valid for flows with small or intermediate values of Re, and that the calculation of solutions with large Reynolds numbers presents considerable difficulties which it would take too long to describe here (see, for example, FORTIN [2], FORTIN-THOMASSET [1] for further details).

In order to convey quite clearly the approach which will allow us

to utilise here the results of Chapter I, we shall first consider in detail the case of the steady-state linearised Stokes equations which we write as follows:

$$-\mu\Delta\underset{\sim}{u} + \underset{\sim}{\nabla}p = \underset{\sim}{f} \quad in \quad \Omega, \tag{1.7}$$

$$\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0 \quad in \quad \Omega, \tag{1.8}$$

$$\underset{\sim}{u}\big|_{\Gamma} = \underset{\sim}{0}, \tag{1.9}$$

which can be deduced from (1.2)-(1.4) by neglecting the nonlinear terms $\underset{\sim}{u}\cdot\underset{\sim}{\nabla}\underset{\sim}{u}$ and by taking $\frac{\partial \underset{\sim}{u}}{\partial t} = \underset{\sim}{0}$. This approximation will be valid if Re is very small, i.e., from (1.6), for a flow with low velocity, or for a very viscous fluid.

1.3 Stokes problem and quadratic programming

We shall now show that equations (1.7)-(1.9) are the optimality conditions of a quadratic programming problem, similar to those studied in Chapter I. The essential point will be to consider the zero-divergence condition (1.8) as a *linear constraint* on the solution $\underset{\sim}{u}$, the *pressure* then appearing as a *Lagrange multiplier*. The problems which we consider will be formulated in Hilbert spaces, of infinite dimension, which we first define.

Suppose then that $L^2(\Omega)$ is the space of square-summable functions on Ω, equipped with the usual norm and inner product, i.e.

$$|v|_o = \left(\int_\Omega |v|^2 dx\right)^{1/2}, \quad (u,v) = \int_\Omega uv \, dx. \tag{1.10}$$

We define, in standard fashion,

$$H^1(\Omega) = \{v \mid v \in L^2(\Omega), \ \frac{\partial v}{\partial x_i} \in L^2(\Omega), \ i=1,\ldots,N\}, \tag{1.11}$$

this *Sobolev space* being equipped with the norm,

$$\|v\|_1 = \left(|v|_o^2 + \sum_{i=1}^{N} \left|\frac{\partial v}{\partial x_i}\right|_o^2\right)^{1/2}. \tag{1.12}$$

It can be shown (see, for example, LIONS-MAGENES [1]) that the trace at the boundary, $v|_\Gamma$, of a function v from $H^1(\Omega)$ has a meaning and we put

$$(1.13) \qquad H_o^1(\Omega) = \{v | v \in H^1(\Omega) \ , \ v|_\Gamma = 0\} \ .$$

The space $H_o^1(\Omega)$ will be equipped with the norm (which is only a semi-norm on $H^1(\Omega)$)

$$(1.14) \qquad |v|_1 = \Big(\sum_{i=1}^{N} \Big| \frac{\partial v}{\partial x_i} \Big|_o^2 \Big)^{1/2} .$$

The norms (1.12) and (1.14) are *equivalent* on $H_o^1(\Omega)$, this result being a direct consequence of Poincare's inequality,

$$(1.15) \qquad |v|_o \leq C(\Omega) \ \Big| \frac{\partial v}{\partial x_i} \Big|_o , \ i=1,\ldots,N \ ,$$

for functions v which are zero at the boundary of Ω, a *bounded* open subset of $\mathbb{R}^N$.

Suppose then that

$$(1.16) \qquad V = \{\underset{\sim}{v} | \underset{\sim}{v} \in (H_o^1(\Omega))^N, \ \underset{\sim}{\nabla} \cdot \underset{\sim}{v} = 0 \ \textit{in} \ \ \Omega\}.$$

For $\underset{\sim}{v} = \{v_1,\ldots v_N\} \in V$, we write

$$(1.17) \qquad \|\underset{\sim}{v}\|_1^2 = \sum_{i=1}^{N} \|v_i\|_1^2 .$$

We now consider the functional, defined for $\underset{\sim}{f} \in (L^2(\Omega))^N$ and $\underset{\sim}{v} \in (H_o^1(\Omega))^N$ by

$$(1.18) \qquad J(\underset{\sim}{v}) = \frac{\mu}{2} a(\underset{\sim}{v},\underset{\sim}{v}) - \int_\Omega \underset{\sim}{f} \cdot \underset{\sim}{v} \, dx = \frac{\mu}{2} a(\underset{\sim}{v},\underset{\sim}{v}) - (\underset{\sim}{f},\underset{\sim}{v}) \ ^1$$

where we have

$$(1.19) \qquad a(\underset{\sim}{u},\underset{\sim}{v}) = \sum_{i,j=1}^{N} \int_\Omega \frac{\partial u_i}{\partial x_j} \ \frac{\partial v_i}{\partial x_j} \, dx.$$

Having defined the quadratic functional $J(\underset{\sim}{v})$, we can seek a solution of the problem

---

[1] $\int_\Omega \underset{\sim}{f} \cdot \underset{\sim}{v} \, dx = (\underset{\sim}{f},\underset{\sim}{v}) = \sum_{i=1}^{N} \int_\Omega f_i v_i dx.$

$$(1.20)\qquad \begin{cases} J(\underset{\sim}{u}) \le J(\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in V, \\ \underset{\sim}{u} \in V. \end{cases}$$

The existence of a *unique solution* is an immediate consequence of the Lax-Milgram theorem (see, for example, EKELAND-TEMAM [1], LIONS [1]). Problem (1.20) clearly consists of minimising a quadratic functional under a linear constraint ($\underset{\sim}{u} \in V$, i.e. $\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0$) ; it is therefore natural to seek to impose this constraint by means of a Lagrange multiplier, thereby transforming (1.20) into a saddle-point problem. We thus define, for $\underset{\sim}{v} \in (H_o^1(\Omega))^N$ and $q \in L^2(\Omega)$, the Lagrangian

$$(1.21)\qquad \mathcal{L}(\underset{\sim}{v},q) = J(\underset{\sim}{v}) - (q, \underset{\sim}{\nabla}\cdot\underset{\sim}{v}) = \frac{\mu}{2} a(\underset{\sim}{v},\underset{\sim}{v}) - (\underset{\sim}{f},\underset{\sim}{v}) - (q, \underset{\sim}{\nabla}\cdot\underset{\sim}{v}),$$

and we seek a pair $\{\underset{\sim}{u},p\}$ defining a saddle-point of $\mathcal{L}$ on $(H_o^1(\Omega))^N \times L^2(\Omega)$, i.e. a solution of the problem

$$(1.22)\qquad \begin{cases} \mathcal{L}(\underset{\sim}{u},q) \le \mathcal{L}(\underset{\sim}{u},p) \le \mathcal{L}(\underset{\sim}{v},p) \quad \forall \underset{\sim}{v} \in (H_o^1(\Omega))^N , \ \forall q \in L^2(\Omega), \\ \underset{\sim}{u} \in (H_o^1(\Omega))^N, \ q \in L^2(\Omega). \end{cases}$$

The existence of a saddle-point, (actually of the pressure p), is a more subtle problem here than in the finite-dimensional case of Chapter I. In the case of the Stokes equations, the result is in fact a standard one (see TEMAM [1], EKELAND-TEMAM [1]) and it can be deduced directly from the *Hahn-Banach* theorem, subject to certain regularity conditions on the boundary of Ω (see FORTIN [3], TARTAR [1]).

We now give an interpretation of problem (1.22) in order to verify that the pair $\{\underset{\sim}{u}, p\}$, actually is a solution of problem (1.7), (1.8). The optimality conditions for (1.22) are respectively

$$(1.23)\qquad \mu a(\underset{\sim}{u},\underset{\sim}{v}) - (p, \underset{\sim}{\nabla}\cdot\underset{\sim}{v}) = (\underset{\sim}{f},\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in (H_o^1(\Omega))^N,$$

and

$$(1.24)\qquad (q, \underset{\sim}{\nabla}\cdot\underset{\sim}{u}) = 0 \quad \forall q \in L^2(\Omega).$$

Condition (1.23) is true in particular for $\underset{\sim}{v} = \underset{\sim}{\phi} \in (\mathcal{D}(\Omega))^N$ [2] and we thus have, *in the distributional sense*,

$$(1.25) \qquad -\mu\Delta\underset{\sim}{u} + \nabla p = \underset{\sim}{f} \quad \text{in} \quad \Omega.$$

Condition (1.24) is clearly equivalent to (1.8).

In view of the results of Chapter I, we are thus led, for the solution of (1.22), to use in place of (1.21) the *augmented Lagrangian*

$$(1.26) \qquad \mathcal{L}_r(v,q) = \frac{\mu}{2} a(\underset{\sim}{v},\underset{\sim}{v}) - (\underset{\sim}{f},\underset{\sim}{v}) - (q, \nabla\cdot\underset{\sim}{v}) + \frac{r}{2} |\nabla\cdot\underset{\sim}{v}|_0^2.$$

For this problem in infinite dimensions, we could consider directly algorithms similar to those of Chapter I. In particular, Uzawa's algorithm converges in this case under the same conditions as in Chapter I, Section 2.

It need hardly be said that, in practice, we first of all have to try to find a discretised version of (1.26); this takes us to a finite-dimensional problem which then falls exactly within the framework of Chapter I.

Remark 1.1: The fact that we can prove the convergence of the algorithm in infinite dimensions allows us to anticipate that the convergence rate will to a certain extent be independent of the discretisation employed. In particular, the fact of refining the mesh in a finite-element (or finite difference) approximation of the problem should not in itself bring about a drastic diminution of the speed of convergence of the algorithm. A study of the choice of the optimal parameters has been carried out (in the case of $r = 0$) for the infinite-dimensional problem, by CROUZEIX [1].

2. DISCRETISATION OF THE STOKES PROBLEM

The use of the methods of Chapter I, for the solution of incompressible viscous fluid flow problems, falls naturally within the framework of finite-element methods where the incompressibility constraint is treated by penalisation. These methods have recently enjoyed considerable popularity and the corresponding

---

[2] $\mathcal{D}(\Omega) = \{\phi \mid \phi \in C^\infty(\overline{\Omega}),\ \phi \text{ has compact support in } \Omega\}$

theoretical developments have enabled certain of these whose operation is reliable and efficient to be picked out. It is difficult, within the inevitably restricted scope of this volume, to quote even a small part of the works devoted to penalisation methods, applied to the numerical treatment of the Navier-Stokes equations; we shall therefore make do with referring to BERCOVIER [1], ODEN [1], TAYLOR-ZIENKIEWICZ [1], MALKUS-HUGHES [1], and to the bibliographies in these works.

In order to highlight clearly the usefulness of the augmented Lagrangian method within the field of penalisation methods, it is worth recalling here several results. This leads us firstly to introduce a velocity-pressure mixed variational formulation discretising (1.23), (1.24), then to introduce a supplementary penalisation term (see GIRAULT-RAVIART [1] for various mixed formulations of the Stokes and Navier-Stokes problems).

Suppose, then, that $\mathcal{T}_h$ is a triangulation of Ω ; we associate with $\mathcal{T}_h$ an approximation W_h of $(H_o^1(\Omega))^N$ generated by conforming or nonconforming finite elements (for simplicity we limit ourselves to the case $N = 2$); likewise, Q_h will be an approximation of $L^2(\Omega)$. It is not necessary to impose matching conditions for the elements of Q_h. In the case of *nonconforming* elements we have in general, if $\underset{\sim}{u}_h = \{u_{1h}, u_{2h}\} \in W_h$, $\frac{\partial u_{ih}}{\partial x_j} \notin L^2(\Omega)$ (in fact $\frac{\partial u_{ih}}{\partial x_j}$ is a *measure*); we cannot therefore utilise directly the bilinear form $a(.,.)$ defined in (1.19), so we "approximate" $a(.,.)$ by $a_h(.,.)$ defined by

$$(2.1) \qquad a_h(\underset{\sim}{u}_h, \underset{\sim}{v}_h) = \sum_{i,j=1}^{2} \sum_{K \in \mathcal{T}_h} \int_K \frac{\partial u_{ih}}{\partial x_j} \frac{\partial v_{ih}}{\partial x_j} \, dx \qquad \forall \underset{\sim}{u}_h, \underset{\sim}{v}_h \in W_h ;$$

we of course have

$$a_h(\underset{\sim}{u}_h, \underset{\sim}{v}_h) = a(\underset{\sim}{u}_h, \underset{\sim}{v}_h) \quad \forall \underset{\sim}{u}_h, \underset{\sim}{v}_h \in W_h \cap (H^1(\Omega))^2.$$

We next define on W_h a linear operator div_h, of *discrete divergence*, with values in Q_h, by

$$(2.2)\qquad \begin{cases} \operatorname{div}_h \underset{\sim}{v}_h \in Q_h \quad \forall \underset{\sim}{v}_h \in W_h \\ (\operatorname{div}_h \underset{\sim}{v}_h, q_h) = \sum_{K \in \mathcal{C}_h} \int_K \nabla \cdot \underset{\sim}{v}_h \; q_h \, dx \quad \forall q_h \in Q_h \; ; \end{cases}$$

in the following discussion we shall use the notation

$$\operatorname{div}_h \underset{\sim}{v}_h = \nabla_h \cdot \underset{\sim}{v}_h .$$

We then put

$$(2.3)\qquad V_h = \{ \underset{\sim}{v}_h \in W_h \mid \nabla_h \cdot \underset{\sim}{v}_h = 0 \} ,$$

and we approximate the Stokes problem (1.7)-(1.9) by the problem

$$(2.4)\qquad \begin{cases} \textit{Find } \underset{\sim}{u}_h \in V_h \textit{ such that} \\ \mu \, a_h(\underset{\sim}{u}_h, \underset{\sim}{v}_h) = (\underset{\sim}{f}, \underset{\sim}{v}_h) \quad \forall \underset{\sim}{v}_h \in V_h \; ; \end{cases}$$

problem (2.4) is equivalent to the minimisation problem

$$(2.5)\qquad \begin{cases} \textit{Find } \underset{\sim}{u}_h \in V_h \textit{ such that} \\ J_h(\underset{\sim}{u}_h) \le J_h(\underset{\sim}{v}_h) \quad \forall \underset{\sim}{v}_h \in V_h , \end{cases}$$

where

$$J_h(\underset{\sim}{v}_h) = \frac{\mu}{2} a_h(\underset{\sim}{v}_h, \underset{\sim}{v}_h) - (\underset{\sim}{f}, \underset{\sim}{v}_h) .$$

Introducing the multiplier $p_h \in Q_h$ to impose the approximate condition of zero divergence, namely $\nabla_h \cdot \underset{\sim}{v}_h = 0$, we get down to the discrete *mixed* problem

Find $\{u_h, p_h\} \in W_h \times Q_h$ *such that*

$$(2.6)\qquad \mu \, a_h(\underset{\sim}{u}_h, \underset{\sim}{v}_h) - (p_h, \nabla_h \cdot \underset{\sim}{v}_h) = (\underset{\sim}{f}, \underset{\sim}{v}_h) \quad \forall \underset{\sim}{v}_h \in W_h ,$$

$$(2.7)\qquad \nabla_h \cdot \underset{\sim}{u}_h = 0,$$

which is clearly a discrete analogue of (1.23), (1.24). This problem is equivalent to finding a saddle-point of the Lagrangian

$$(2.8)\qquad \mathcal{L}_h(\underset{\sim}{v}_h, q_h) = \frac{\mu}{2} a_h(\underset{\sim}{v}_h, \underset{\sim}{v}_h) - (q_h, \nabla_h \cdot \underset{\sim}{v}_h) - (\underset{\sim}{f}, \underset{\sim}{v}_h) ,$$

and of course of the *augmented Lagrangian*

$$(2.9)\qquad \mathcal{L}_{rh}(\underset{\sim}{v}_h, q_h) = \frac{\mu}{2} a_h(\underset{\sim}{u}_h, \underset{\sim}{v}_h) - (q_h, \nabla_h \cdot \underset{\sim}{v}_h) + \frac{r}{2} |\nabla_h \cdot \underset{\sim}{v}_h|_o^2 - (f, v_h) .$$

Putting $q_h = 0$, we obtain the penalisation problem

$$(2.10)\qquad \begin{cases} \textit{Minimise on } W_h \textit{ the functional} \\ J_{rh}(\underset{\sim}{v}_h) = \frac{\mu}{2} a_h(\underset{\sim}{u}_h,\underset{\sim}{v}_h) + \frac{r}{2} |\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h|_o^2 - (\underset{\sim}{f},\underset{\sim}{v}_h). \end{cases}$$

We observe that the penalty term contains the discrete divergence $\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h$ and not the exact divergence $\underset{\sim}{\nabla}\cdot\underset{\sim}{v}_h$. The reason for this is simple: when r is *large*, the solution of (2.10) approaches that of the mixed problem (2.6), (2.7) and $-r(\underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h)$ converges to p_h; we can therefore only obtain a correct solution of the penalised problem if the mixed problem is well posed. It is well known that the approximations W_h and Q_h cannot be chosen independently; in order to obtain a convergent approximation, the *Babuska-Brezzi* condition must be satisfied (see BREZZI [1], BABUSKA [1], FORTIN [4], GIRAULT-RAVIART [1]) which is here written:

$$(2.11)\qquad \sup_{\underset{\sim}{v}_h\in W_h-\{\underset{\sim}{0}\}} \frac{(q_h,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)}{\|\underset{\sim}{v}_h\|_1} \geq k \|q_h\|_{L^2(\Omega)/\mathbb{R}} ,$$

where the constant k *is independent of* h.

We shall not dwell here on the meaning of this condition, but solely on its consequences. One of these is that in general we cannot define Q_h by $Q_h = \underset{\sim}{\nabla}\cdot(W_h)$, even for conforming elements. It therefore follows that, except in special cases for which we refer to e.g. MERCIER [1] and GIRAULT-RAVIART [1], we cannot in the discretised problem make the divergence vanish completely, hence the impossibility of penalising with $|\underset{\sim}{\nabla}\cdot\underset{\sim}{v}_h|_o^2$; this fact rapidly became apparent to users of penalisation methods, and one solution which has been adopted has been to evaluate $|\underset{\sim}{\nabla}\cdot\underset{\sim}{v}_h|_o^2 = \int_\Omega |\underset{\sim}{\nabla}\cdot\underset{\sim}{v}_h|^2 \, dx$ *by an inexact quadrature formula.* This method of procedure has become known under the name of *reduced integration;* but it must be underlined (see HUGHES-MALKUS [1]) that this procedure implicitly defines an operator div_h and a space Q_h for which $\int_\Omega p_h q_h \, dx$ is in fact evaluated exactly, $\forall p_h,q_h \in Q_h$, by the quadrature formula used. In summary, the penalisation is indissociable from a mixed (velocity-pressure) method, and must be considered as a solution technique for this latter method, and not as an approximation technique in itself. In this sense the use of augmented Lagrangian

methods is quite natural and the techniques of Chapter I provide some advance on the more usual methods, since several iterations actually enable the error due to the penalisation to be eliminated. We do not therefore have to choose values of r as large as in a pure penalisation method. This possibility allows an improvement in the conditioning of the problems in $\underset{\sim}{u}_h$, and this is particularly useful if one is unable to use double precision, or if the problem in $\underset{\sim}{u}_h$ is to be solved by an iterative method.

Regarding the numerical experiments which we are about to discuss, let it be said immediately that our objective here is to check the efficiency of the algorithms of Chapter I, rather than to obtain precise solutions to a specific hydrodynamic problem. We are therefore satisfied with quite a coarse approximation in which, nonetheless, all the difficulties inherent in the problem in question are still present.

We therefore use a (nonconforming) approximation of $H_0^1(\Omega)$, defined on a triangulation of the open domain Ω. This discrete space W_h is made up of functions whose restriction to each triangle is a polynomial of degree 1, and which are continuous at the midpoints of the sides of the triangles. This is therefore a space of nonconforming finite elements in the usual sense (see CIARLET [1], CROUZEIX-RAVIART [1], STRANG-FIX [1]). We thus define

$$V_h = \{\underset{\sim}{v}_h = \{v_{1h}, v_{2h}\} \mid \underset{\sim}{v}_h \in W_h \ , \ \underset{\sim}{\nabla}_h \cdot \underset{\sim}{v}_h = 0\} \ . \tag{2.12}$$

The discrete operator $\underset{\sim}{\nabla}_h$ represents discrete differentiation in the sense of nonconforming elements, i.e. restricted to the interior of each triangle.

In this case the discrete divergence $\underset{\sim}{\nabla}_h \cdot \underset{\sim}{v}_h$ is constant over each triangle and the zero-divergence condition is thus expressed by a linear constraint associated with each of the triangles. This being so it is natural to choose for the space Q_h of the discrete pressures, the space of functions which are *constant* on each triangle of $\mathcal{C}_h$.

It is shown in CROUZEIX-RAVIART [1], that a pair $\{\underset{\sim}{u}_h, p_h\}$, this being a saddle-point of the $\mathcal{L}_h$, defined in (2.8), is an approximation to order h (h being the longest side belonging to

the triangles in the triangulation) of the solution $\{\underset{\sim}{u},p\}$ of the Stokes problem. Recall that p and p_h are defined to within an additive constant.

Reference may be made to THOMASSET [1] for a complete discussion of the implementation of this approximation.

We have considered in our experiments a model problem, namely the (two-dimensional) flow between two non-concentric cylinders, the inner cylinder being fixed and the outer cylinder rotating with a uniform angular velocity ω. In detail, we have taken for Ω the region of $\mathbb{R}^2$ whose boundaries are

$$\left\{\begin{array}{l} C_1\text{: } \textit{circle of radius } 5 \textit{ and centre } (0,0), \\ \\ C_2\text{: } \textit{circle of radius } 2 \textit{ and centre } (1,0). \end{array}\right. \tag{2.13}$$

The discretisation employed used 126 triangles, the number of interior midpoints being 172. For each midpoint we have to determine the components u_1 and u_2 of the flow velocity, and for each triangle the value of the pressure.

Our discrete problem is therefore a *quadratic programming* problem with 344 variables, related by 126 linear constraints (of which 125 are linearly independent). Looking back at the notation of Chapter I, the matrices A, B and B^t correspond (save possibly for a sign) respectively to the Laplacian, to the divergence and to the discrete gradient.

To illustrate the concepts, we have displayed on Figure 2.1 the domain and the triangulation used, and on Figure 2.2 the streamlines of the solution $\underset{\sim}{u}_h$ obtained.

3. ALGORITHMS AND DISCUSSION OF RESULTS

3.1 Explicit formulation of the algorithms

We shall now, for the case of the Stokes problem, give an explicit description of the algorithms of Chapter I; we will then be in a position, by inspecting the numerical results obtained, to compare their efficiency and their ease of implementation. We have of course used in our experiments the augmented Lagrangian

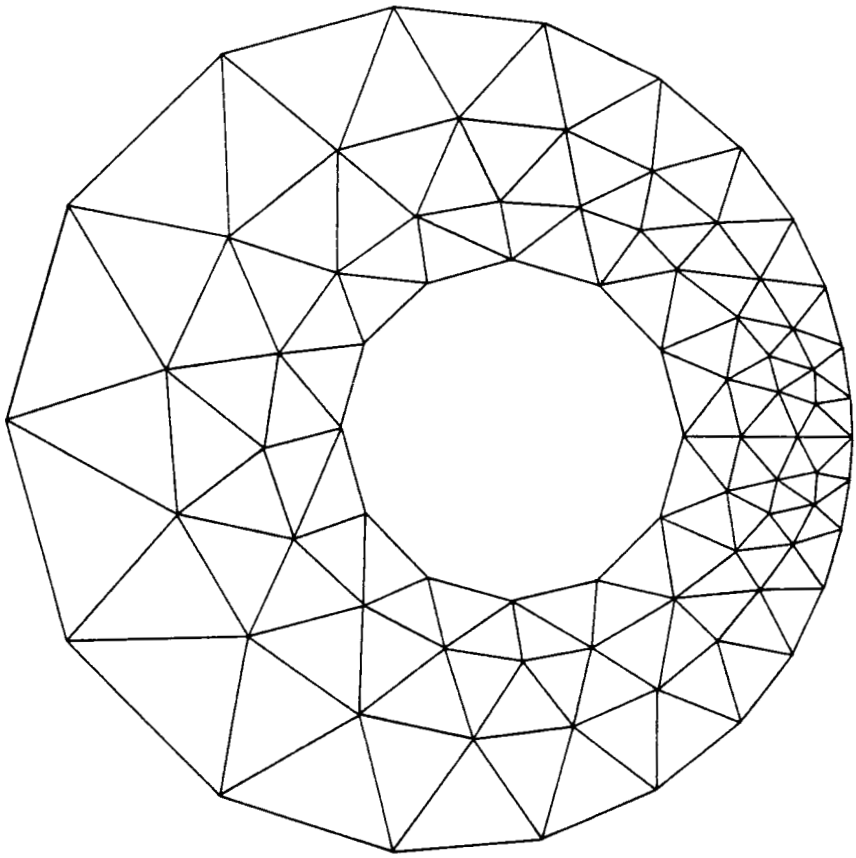


Figure 2.1

Triangulation of Ω

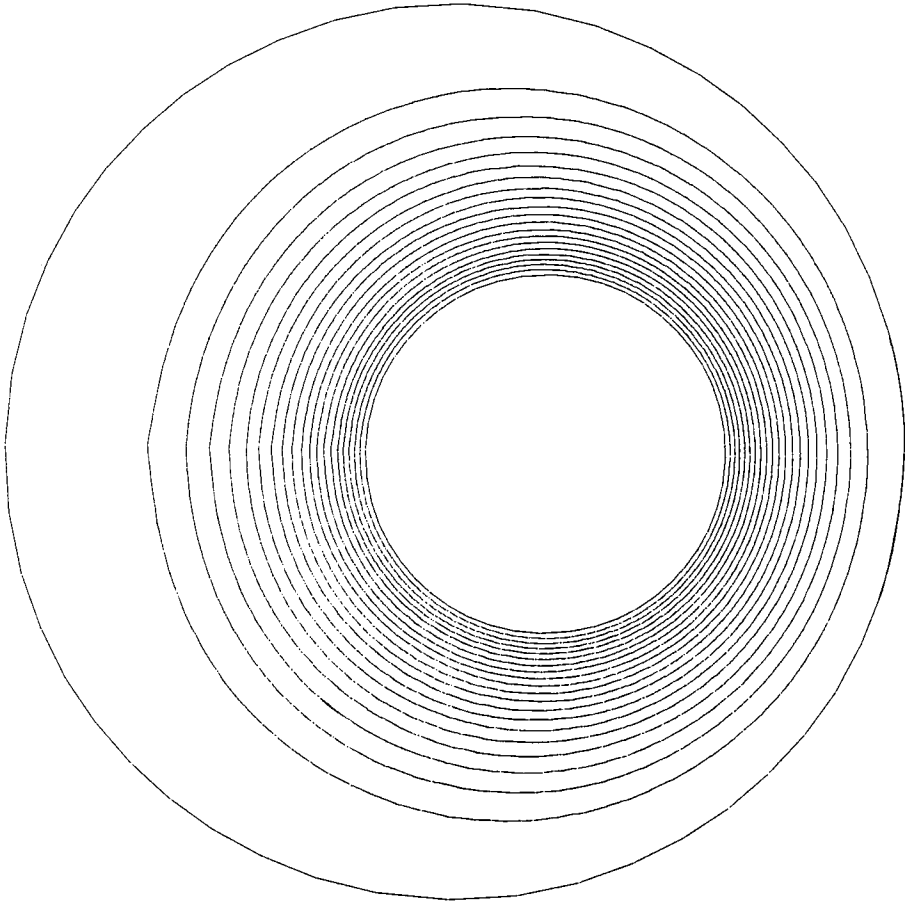


Figure 2.2

Streamlines

$$(3.1)\qquad \mathcal{L}_{rh}(\underset{\sim}{v}_h,q_h) = \frac{\mu}{2}\, a_h(\underset{\sim}{v}_h,\underset{\sim}{v}_h)-(\underset{\sim}{f},\underset{\sim}{v}_h)-(q_h,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h) + \frac{r}{2}\,|\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h|_o^2 .$$

The simplest algorithm for the solution of our problem is Uzawa's algorithm of Chapter I, Section 2, which we write here as:

$$(3.2)\qquad p_h^o \quad \textit{specified arbitrarily;}$$

with p_h^n *known, calculate a solution* $\underset{\sim}{u}_h^n$ *of*

$$(3.3)\qquad \begin{cases} \mu\, a_h(\underset{\sim}{u}_h^n,\underset{\sim}{v}_h)-(\underset{\sim}{f},\underset{\sim}{v}_h)-(p_h^n,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)+r(\underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^n,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h) = 0 \quad \forall \underset{\sim}{v}_h \in W_h, \\ \underset{\sim}{u}_h^n \in W_h , \end{cases}$$

then p_h^{n+1} *by*

$$(3.4)\qquad p_h^{n+1} = p_h^n - \rho\underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^n.$$

We have also used the variable steplength methods of Chapter I, Section 3 and the conjugate-gradient method. We shall now briefly review the operation of these algorithms by going through them in the particular case of the Lagrangian (3.1). We present within a single algorithm the *variable-step gradient* methods and the *conjugate-gradient* method which differ only through the choice of the descent direction. Thus suppose we have :

$$(3.5)\qquad p_h^o \quad \textit{specified arbitrarily;}$$

and $\underset{\sim}{u}_h^o$ *is satisfying*

$$(3.6)\qquad \begin{cases} \mu a_h(\underset{\sim}{u}_h^o,\underset{\sim}{v}_h)+r(\underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^o,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)-(p_h^o,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)-(\underset{\sim}{f},\underset{\sim}{v}_h) = 0 \quad \forall \underset{\sim}{v}_h \in W_h , \\ \underset{\sim}{u}_h^o \in W_h. \end{cases}$$

On iteration n, *calculate the descent direction by*

$$(3.7)\qquad w_n = \underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^n$$

in the methods of *steepest descent* and of *minimum residual*. *In the conjugate-gradient method, we proceed as follows:*

$$(3.8)\qquad w_o = \underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^o \quad \textit{if} \quad n=0,$$

$$(3.9)\qquad w_n = \underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^n + \lambda_n w_{n-1} \quad \textit{if} \quad n \geq 1,$$

$$(3.10)\qquad \lambda_n = \frac{|\nabla_h \cdot \underset{\sim}{u}_h^n|_o^2}{|\nabla_h \cdot \underset{\sim}{u}_h^{n-1}|_o^2} .$$

Knowing now the direction w_n, *solve in* W_h *the problem*

$$(3.11)\qquad \mu a_h(\underset{\sim}{z}_h^n, \underset{\sim}{v}_h) + r(\nabla_h \cdot \underset{\sim}{z}_h^n, \nabla_h \cdot \underset{\sim}{v}_h) - (w_n, \nabla_h \cdot \underset{\sim}{v}_h) = 0 \quad \forall \underset{\sim}{v}_h \in W_h .$$

Next calculate

$$(3.12)\qquad \rho_n = - \frac{|\nabla_h \cdot \underset{\sim}{u}_h^n|_o^2}{(\nabla_h \cdot \underset{\sim}{u}_h^n, \nabla_h \cdot \underset{\sim}{z}_h^n)} ,$$

$$(3.13)\qquad p_n^{n+1} = p_h^n - \rho_n w_n ,$$

$$(3.14)\qquad \underset{\sim}{u}_h^{n+1} = \underset{\sim}{u}_h^n + \rho_n \underset{\sim}{z}_h^n .$$

In the minimum residual method, replace (3.12) *by*

$$(3.12')\qquad \rho_n = - \frac{(\nabla_h \cdot \underset{\sim}{u}_h^n, \nabla_h \cdot \underset{\sim}{z}_h^n)}{|\nabla_h \cdot \underset{\sim}{z}_h^n|_o^2} .$$

We may recollect that the purpose of introducing the intermediate vector $\underset{\sim}{z}_h^n$ is so that only a single linear system needs to be solved at each iteration for the calculation of $\underset{\sim}{u}_h^n, \rho_n, p_h^{n+1}$. This method of procedure slightly increases the memory requirements, but this increase would only become a limiting factor for very large systems and hardly ever poses a problem with modern computers.

Finally we have considered the algorithms of Chapter I, Section 4. In a general way these can be written, in the case of the Stokes problem, in the form

$$(3.15)\qquad \begin{cases} (S_{rh}(\underset{\sim}{u}_h^{n+1} - \underset{\sim}{u}_h^n), \underset{\sim}{v}_h) + \omega_n \{\mu a_h(\underset{\sim}{u}_h^n, \underset{\sim}{v}_h) + r(\nabla_h \cdot \underset{\sim}{u}_h^n, \nabla_h \cdot \underset{\sim}{v}_h) - (p^n, \nabla_h \cdot \underset{\sim}{v}_h) - \\ - (\underset{\sim}{f}, \underset{\sim}{v}_h)\} = 0 \quad \forall \underset{\sim}{v}_h \in W_h, \ \underset{\sim}{u}_h^{n+1} \in W_h , \end{cases}$$

$$(3.16)\qquad p_h^{n+1} = p_h^n - \rho_n \nabla_h \cdot \underset{\sim}{u}_h^n .$$

The auxiliary operator S_{rh} must be *positive definite*. We have of course considered the canonical choice $(S_r = A_r)$ of Chapter I, Section 4.2, in which case (3.15) can be written in the form

$$(3.17)\quad \begin{cases} \mu a_h(\underset{\sim}{u}_h^{n+1/2},\underset{\sim}{v}_h)+r(\underset{\sim}{\nabla}_h\cdot\underset{\sim}{u}_h^{n+1/2},\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)-(p_h^n,\underset{\sim}{\nabla}_h\cdot\underset{\sim}{v}_h)-(\underset{\sim}{f},\underset{\sim}{v}_h) = 0 \quad \forall \underset{\sim}{v}_h \in W_h\ , \\ \underset{\sim}{u}_h^{n+1/2} \in W_h\ , \end{cases}$$

$$(3.18)\quad \underset{\sim}{u}_h^{n+1} = \underset{\sim}{u}_h^n + \omega_n(\underset{\sim}{u}_h^{n+1/2}-\underset{\sim}{u}_h^n),$$

which is very similar to (3.3). The use of this algorithm thus requires the solution of a linear system at each iteration.

We have, however, also used another choice of S_{rh}, in which this matrix is not given explicitly. The idea consists of solving (3.3) by a symmetric successive overrelaxation method (SSOR), carrying out, at each iteration of the Uzawa algorithm, only one double pass (forward and back) in the overrelaxation method. It can easily be shown that this procedure is equivalent to (3.15) for an auxiliary operator S_{rh} which can be constructed explicitly if necessary (see AXELSSON [1], [2]). Any other iterative method could similarly be used for the solution of (3.3) and an expression of the form (3.15) could be obtained (implicitly) by taking just a single iteration at each stage of the Uzawa method. The auxiliary operator thus introduced is not in general symmetric; this is the case, in particular, for the usual overrelaxation method (SOR). It would be possible to envisage a complete family of intermediate algorithms by carrying out at each step a fixed number of iterations for the solution of (3.3). However, as we shall see from the experimental results, the optimal number of iterations seems to be one (or at any rate small). Another procedure might be to solve (3.3) with an accuracy which is low in the initial steps and which becomes higher and higher as we approach the solution. Supplementary details and proofs of convergence for such techniques can be found in BERTSEKAS [2] and KORT-BERTSEKAS [1]. We have retained the limitation on the number of iterations because of its simplicity of implementation and from the fact that it leads to algorithms of the Arrow-Hurwicz type studied in Chapter I, Section 4.3.

3.2 Results and Discussion

3.2.1 *Fixed-step UZAWA Algorithms*

The implementation of this type of algorithm is very simple, and

MAXIMUM VALUES OF $|\nabla_h \cdot u_h^n|$ AFTER "n" ITERATIONS

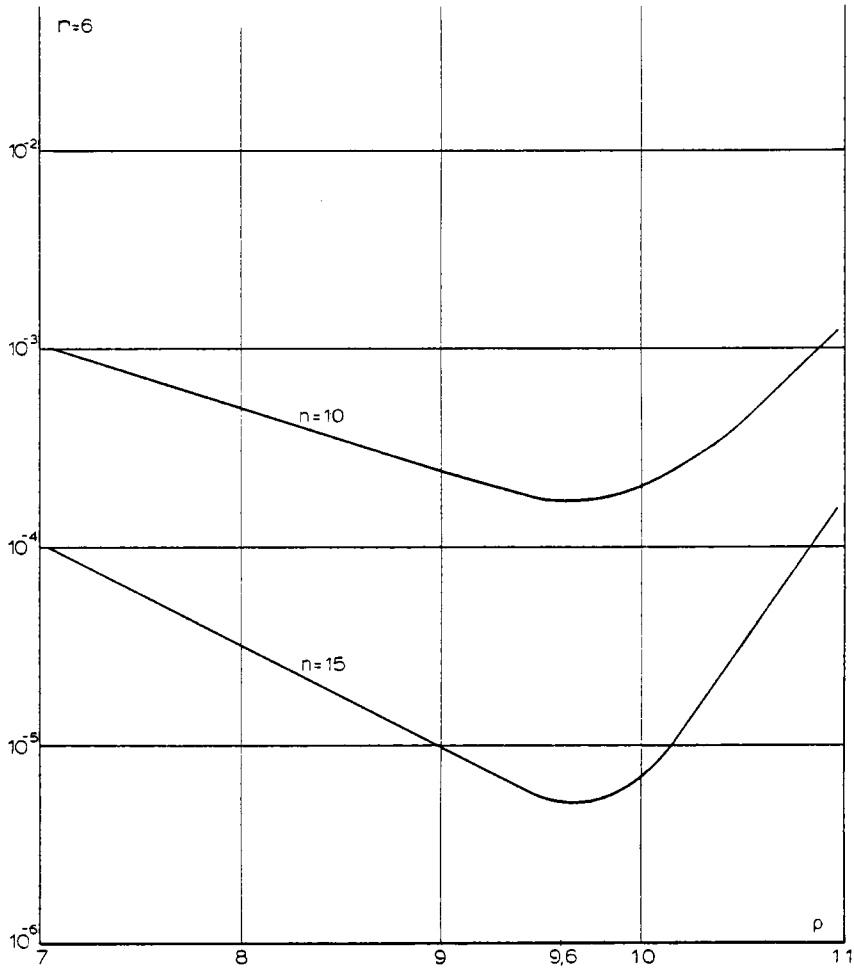


Figure 3.1

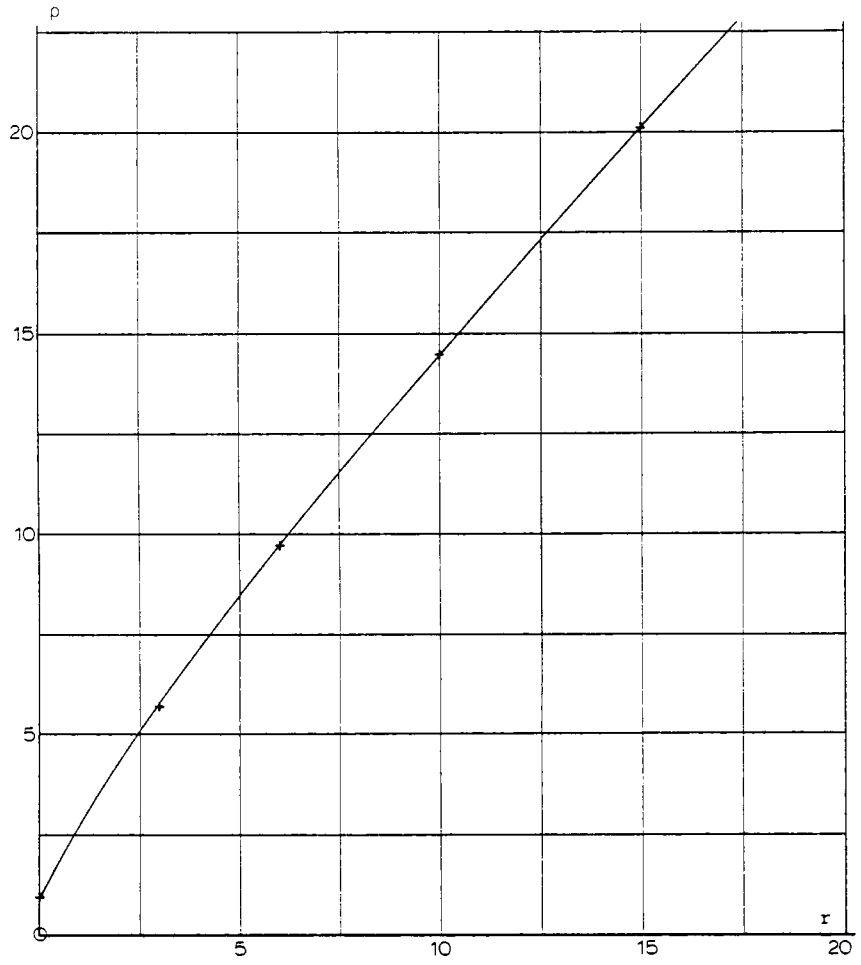


Figure 3.2

optimal ρ , theoretical and experimental, as a function of r

—— theoretical

\+ experimental

the essential difficulty as regards algorithm (3.2)-(3.4) is to determine the optimal parameter. To illustrate the importance of this choice, Figure 3.1 shows, as a function of ρ, the accuracy obtained in respect of the constraint $\underset{\sim}{\nabla}_h \cdot \underset{\sim}{u}_h = 0$, after n iterations. We actually show the logarithm (to base 10) of the maximum, over the triangles, of the absolute value of the discrete divergence (which is here constant over each triangle). The value of r used here is equal to 6, though similar curves would be obtained for any other value of r. As we might expect from the results of Chapter I, a very "sharp" optimal value is obtained. The a priori determination of this optimal value requires knowledge of the smallest and largest *eigenvalues* of $A_r^{-1} B^t B$. It would be possible to determine these values by the power method. We have actually estimated these values by carrying out, with $r = 0$, one pass of the algorithm for a small value of ρ (hence smaller than the optimal ρ), then for a value very close to the limit value. Then knowing the relation linking λ_M and λ_m to the convergence rate and having been able to obtain an estimate of the latter, we can thus calculate the desired eigenvalues. In our case these values are

$$\lambda_M = 2 \ , \ \lambda_m = .07425 \ . \tag{3.19}$$

The value $\lambda_M = 2$ coincides with a theoretical bound, which can be obtained through an energy inequality $(|\underset{\sim}{\nabla}_h \cdot \underset{\sim}{v}_h|_0^2 \leq 2 \, \|\underset{\sim}{v}_h\|_1^2)$ in the discrete spaces used.

We have thus been able to confirm experimentally the agreement between the observed optimal ρ values and the theoretical values given by formula (2.48) of Chapter I. Figure 3.2 presents the results of this comparison; this shows a perfect agreement between the predicted value and the experiment. In practice it is often possible, as is the case here, to obtain an a priori estimate for the largest eigenvalue. In contrast the determination of λ_m is costly and, in fact, if one employs a technique analogous to that which we have used, requires the solution of the Stokes problem. This calculation only becomes profitable if calculations with the same discretisation (and therefore with the same matrices) have to be carried out many times. An approximate rule might be:

. *For* r *large take* ρ *slightly larger than* r.

We note that from (3.19), the condition number of the dual problem is here of the order of 27. We thus have here a relatively well conditioned problem. We have seen furthermore, in Chapter I, that the conditioning improves as r increases, an improvement which ought normally to show itself through an acceleration of the convergence of the algorithm. In Table 3.1 we show the number of iterations of algorithm (3.2)-(3.4) which were necessary (with the optimal ρ) to obtain $|\nabla_h \cdot \underset{\sim}{u}_h^n| \leq 10^{-5}$ on every triangle. This positively establishes that the situation improves rapidly as r increases.

r	1	2	3	5	10	30
n	42	28	22	19	10	6

Table 3.1

Consequently, if (3.3) is solved by a direct method, A_r being for example factorised once and for all, it seems clear that the optimal strategy consists of taking r as large as possible, as long as we maintain good accuracy in the factorisation. For a problem of the size of our model problem, values of r of the order of 10^4 still appear quite reasonable if the calculations are performed in double precision. The number of iterations is then of the order of 2 or 3, depending on the accuracy desired. For further examples, reference may be made to the article by SEGAL [1]. For larger problems, in particular in three dimensions, it is probable that for large r the ill-conditioning of A_r would be a greater constraint.

3.2.2 *Effect of the incomplete solution of* (3.3)

We have specifically considered the case where problem (3.3) is solved by an *iterative* method, in this case in the shape of an *overrelaxation* method. Since it seems pointless, a priori, to carry out the solution fully and completely in the initial stages, it is natural to limit the number of overrelaxation iterations to a value which may be quite small. In our model problem, the determination of $\underset{\sim}{u}_h^n$ in the initial steps required approximately 50 iterations. Figure 3.3 shows the number of iterations of

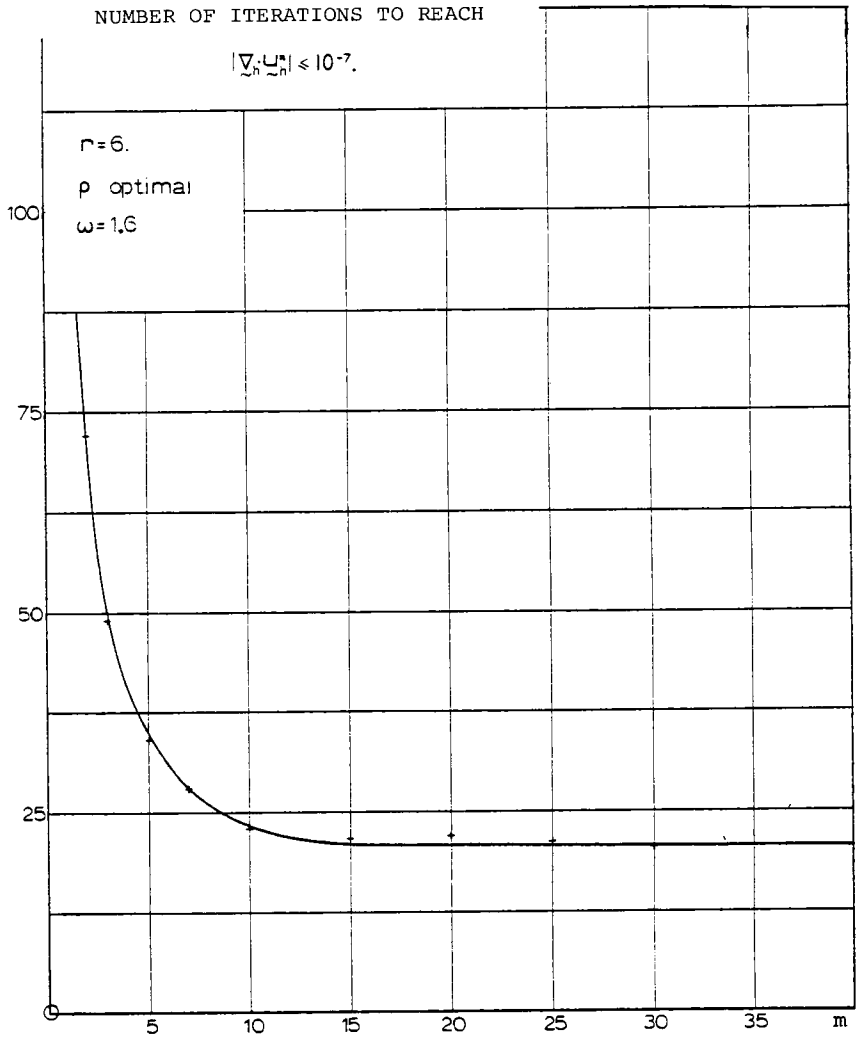

NUMBER OF ITERATIONS TO REACH
$|\nabla_h \underset{\sim}{U}_h^n| \leq 10^{-7}$.
r = 6.
ρ optimal
ω = 1.6
100
75
50
25
0
5
10
15
20
25
30
35
m

Figure 3.3

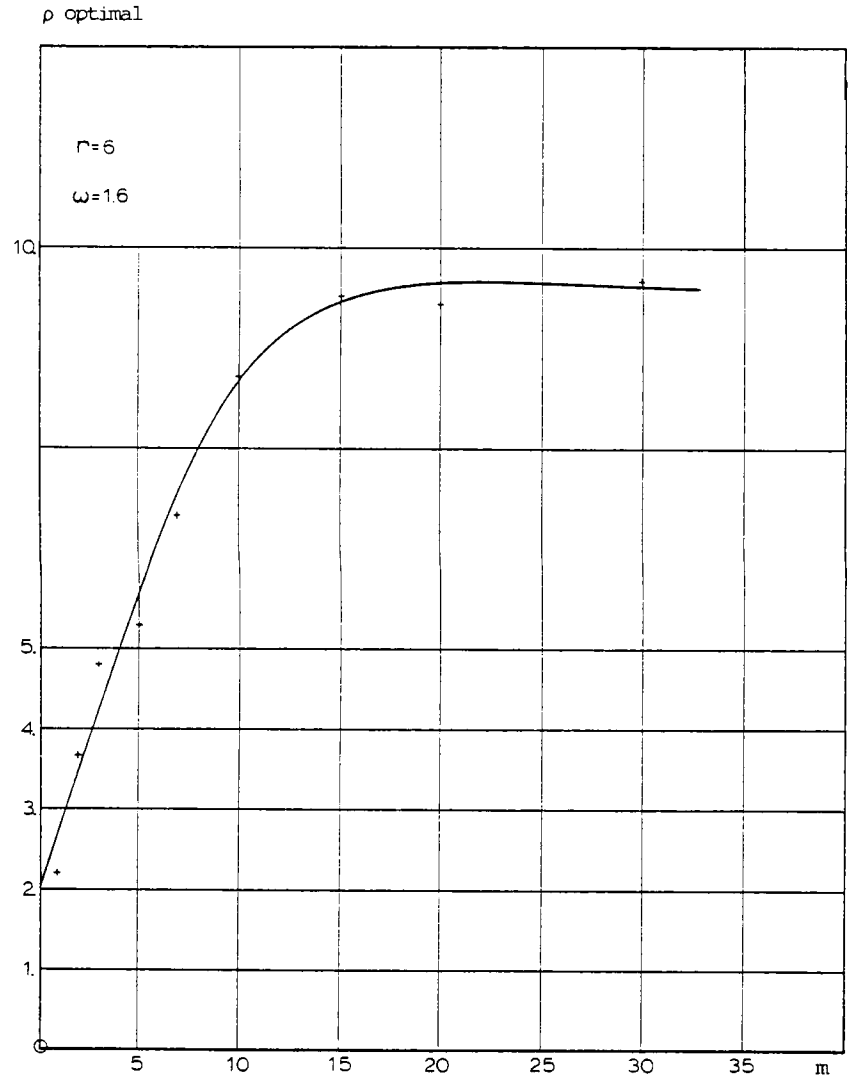


Figure 3.4

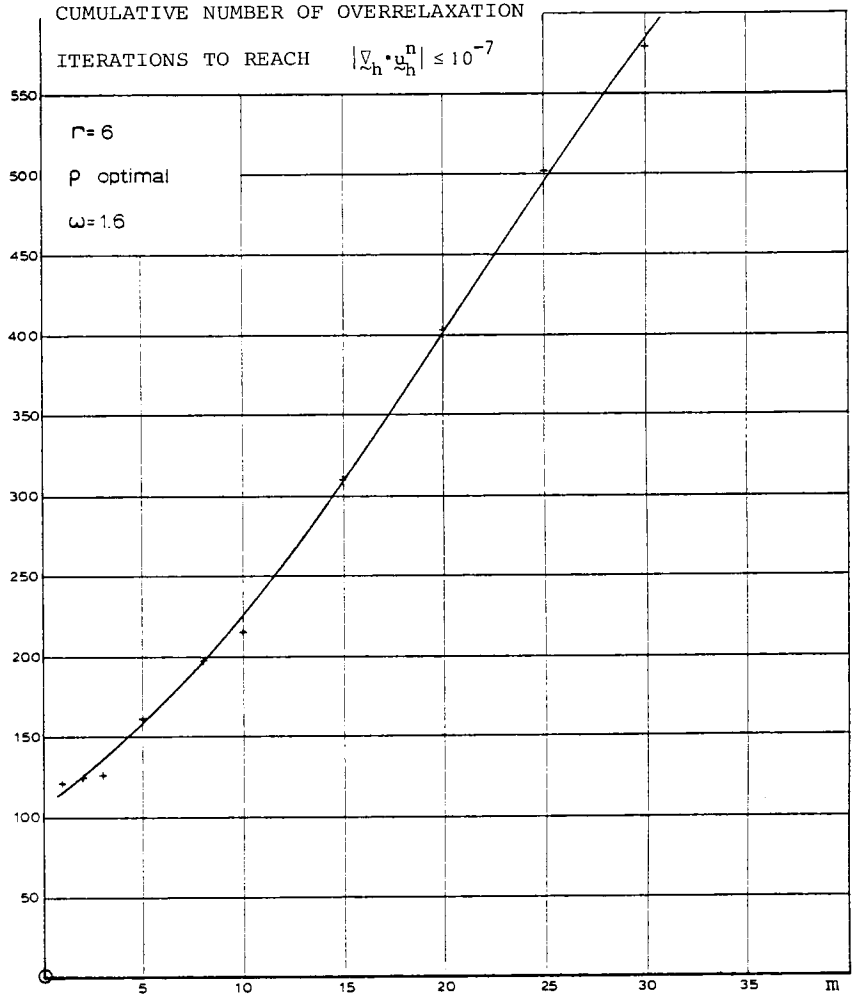

CUMULATIVE NUMBER OF OVERRELAXATION
ITERATIONS TO REACH $|\nabla_h \cdot \underset{\sim}{u}_h^n| \leq 10^{-7}$
r = 6
p optimal
ω = 1.6
550
500
450
400
350
300
250
200
150
100
50
0
5
10
15
20
25
30
35
m

Figure 3.5

(3.2)-(3.4) which were required to obtain $|\underset{\sim}{\nabla}_h \cdot \underset{\sim}{u}_h^n| \leq 10^{-7}$ on every triangle, as a function of the number m of overrelaxation iterations allowed for the solution of (3.3). It shows that for $m \geq 10$ the method undergoes practically no further change, the poor solution of (3.3) in the initial steps having very little effect on the overall process. For $m < 10$, however, the convergence of the algorithm is considerably retarded.

Figure 3.4 shows the variation of the (experimental) optimal value of ρ as a function of m. This demonstrates that ρ ought to diminish if m is small (in our case if $m < 10$). On the other hand, we obtain a very different picture if we try to visualize the effect of m on the total number of iterations necessary for convergence. To illustrate this fact, Figure 3.5 shows as a function of m the cumulative number of overrelaxation iterations necessary to achieve convergence, this number being to a first approximation proportional to the computation time. This figure shows that this cumulative number decreases almost linearly with m. In the case where (3.3) is solved by an iterative method, it thus seems that a good strategy is to carry out only one or two passes at each stage. This observation may be important in the cases where the sub-problems are no longer quadratic. The problems of the choice of r and of ρ are in this case open questions and may be compared with the problem of the choice of parameters in the Arrow-Hurwicz algorithm. In particular, there certainly exists an optimal value for r, because for r large the condition number of A_r increases prohibitively, causing a deterioration in the convergence of the internal iterative method.

3.2.3 *Variable steplength and conjugate-gradient methods*

These methods have proved to be extremely efficient for our model problem; this is particularly true of the conjugate-gradient method, the implementation of which is no more difficult than that of the methods of steepest descent or of minimum residual. The principal advantage of this type of technique lies of course in its ability to be used as a "black box" with no need for user-intervention for the choice of parameters, and without this choice being linked to the solution of a spectral problem.

To illustrate the ideas and to make clearer the comparison between the various methods, we present in Figure 3.6 the decrease of the maximum value of $|\underset{\sim}{\nabla}_h \cdot \underset{\sim}{u}_h^n|$ over the triangles as a function of the number of iterations. Since the scale is logarithmic, the slope corresponds to the *convergence rate of a first-order method.*

As might be expected, the conjugate-gradient method, which is of second order, appears as the most efficient technique. It requires, however, the exact solution of problem (3.11), and this makes it particularly attractive in cases where the solution is accomplished by a direct method.

As regards the other algorithms, the convergence of the variable-steplength gradient methods is better than that of the method with ρ fixed, for the optimal choice of the parameter. This is explained by the fact that in our tests the ρ_n values settle down to a limit cycle between two parameters situated either side of the optimal ρ. In the limit we thus have a cyclic variation of ρ_n, a strategy which has already been suggested in CROUZEIX [1], the cycle being determined automatically by the algorithm. As might have been anticipated, the steepest descent method is superior to the minimum residual method, the latter being more sensitive to the condition number.

3.2.4 *Algorithm with relaxation parameter*

Figure 3.6 shows also the convergence of the algorithm (3.15)-(3.18), which was deduced from Uzawa's algorithm through the introduction of a relaxation parameter. The optimal parameters have been estimated using formulas (4.22), (4.23) of Chapter I. This figure shows that the introduction of the parameter ω accelerates the convergence appreciably, and this is for a dual problem which is quite well conditioned. The acceleration effect should be felt still more when the conditioning is worse. In our model problem, this method is comparable to the variable-steplength gradient algorithms. We have already noted in Chapter I that in the case of variable parameters, this method encompasses the conjugate-gradient method. Consequently, since the determination of the optimal parameters requires the solution of a spectral problem, this algorithm offers little attraction in the case of our model problem.

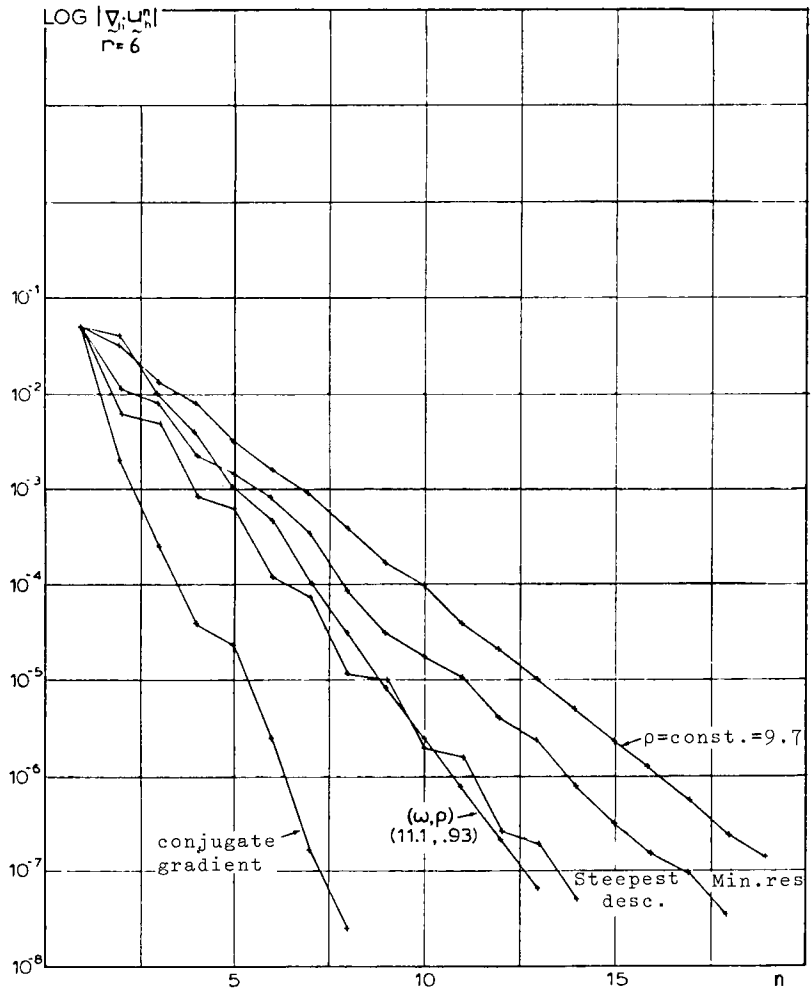

LOG $|\nabla_h \underset{\sim}{U}_h^n|$
r= 6
10^{-1}
10^{-2}
10^{-3}
10^{-4}
10^{-5}
10^{-6}
10^{-7}
10^{-8}
5
10
15
n
ρ=const.=9.7
(ω,ρ)
(11.1, .93)
conjugate
gradient
Steepest
desc.
Min.res

Figure 3.6

Its attraction would be more obvious in the case of the nonlinear Navier-Stokes equations.

3.2.5 *Algorithms of Arrow-Hurwicz type*

As we have already mentioned in Section 3.1, if we carry out the solution of (3.3) by a symmetric overrelaxation (SSOR) method, we are actually using an algorithm of the type (3.15)-(3.16). This algorithm has the advantage, compared with the preceding ones, of not requiring the solution of a linear system at each iteration, the calculation being entirely explicit, and consequently rapid. The essential problem here of course is to determine the optimal parameters ω^* and ρ^*, which minimise the number of iterations required to reach a specified level of accuracy. Since we are not in a position to give a theoretical analysis of this question, we instead present here some experimental results, which, we hope, will allow the reader to obtain an intuitive view of the behaviour of the algorithm.

First we recall the results of Chapter I, Section 4.3, where we pointed out the analogy between this algorithm and a wave equation with damping present. In this analogy ω has the appearance of a damping coefficient and $\rho\omega$ of the square of a propagation speed. Intuitively we can therefore expect, for a specified value of ρ, an under-damping phenomenon for the small values of ω, and an over-damping beyond some critical value which will be a function (presently unknown) of the eigenvalues of the various operators which occur in the algorithm.

Such a phenomenon can in fact be verified; for $r = 0$ the critical value appears to be $\omega = 1$. This threshold seems to increase for larger values of r. For $r = 6$, for example, a transitory phase of under-damping was observed for $\omega = 1.1$. Unfortunately it is not possible to estimate the optimal value of ω by making use of the critical damping value. In actual fact this value may well be different for each of the "components" of the solution.

The main experimental findings are as follows:

- The optimal speed of convergence is almost the same for $r = 0$ and $r = 1$ and subsequently diminishes rapidly as r increases because

of the ill-conditioning of A_r.

- For a given value of ω, the optimal value of ρ obeys (to a first approximation) the law

$$\rho_{opt} = 2r+1.$$

This estimate slightly underestimates the experimental value, but the speed of convergence is not very sensitive to such an underestimate .

- The choice of ω is more difficult. It is necessary to "under-relax" for r large. To indicate the trends, Table 3.2 summarises a few experimental values.

r	0	1	2	6
ρ_{opt}	1.03	3.3	5.4	13.6
ω_{opt}	1.06	1.2	1.1	.9
n	94	94	120	200

Table 3.2

In summary, the benefit of the augmented Lagrangian is by no means apparent for such an algorithm: the optimal value of r is, if not precisely 0, at least in the neighbourhood of 0. By contrast, the computational effort is competitive with the preceding methods in which every iteration requires the solution of a linear system. The problem of the optimal choice of the parameters, or of an automatic choice similar to that made in the method of steepest descent, merits further study in spite of the difficulty of the question.

3.2.6 *Conclusions*

The methods based on the use of the augmented Lagrangian are apparently very efficient for solving quadratic problems with linear constraints, and it is clear that this efficiency carries across to a more general setting. Of all the algorithms tested, the most attractive are those in which the parameters are chosen automatically; this applies particularly for the conjugate-gradient

algorithm in association with a direct method for the solution of the linear sub-problems.

The methods of Arrow-Hurwicz type deserve a more thorough study as regards the choice of parameters. They could prove to be excellent for cases in which the sub-problems are nonlinear. We thus have here an open problem, difficult most certainly, but of great practical interest.

Regarding, more especially, the Stokes problem, the results obtained show that the augmented Lagrangian approach allows us to reduce the problem to solving several problems of linear-elasticity type (3 or 4 for r sufficiently large), which can themselves be solved using a factorisation carried out once and for all. We believe that this method of procedure may prove to be less costly than currently-used methods which consist of directly solving by factorisation the global linear system in $\{\underset{\sim}{u},p\}$. In this case the systems being dealt with are actually larger in size, as well as being singular, and their bandwidth is considerably larger. We can therefore expect that for large problems the augmented Lagrangian method will be faster and less sensitive to rounding errors. In conclusion, we remark that the method has also been used for problems which are analogous to the Stokes problem but which arise from the field of soil mechanics, by M. SOULIE [1].

4. NAVIER-STOKES EQUATIONS, STEADY-STATE NONLINEAR CASE

4.1 Statement of the problem

We now consider the case of the nonlinear Navier-Stokes equations; these describe numerous problems of great practical importance. Just as in the preceding sections, we consider for simplicity the case of a flow with homogeneous boundary conditions in a domain Ω of $\mathbb{R}^2$ (or of $\mathbb{R}^3$), bounded and with regular boundary. With the notation used earlier for the velocity vector, the pressure and the external forces, the problem to be solved is now

$$(4.1) \qquad -\nu\Delta\underset{\sim}{u} + (\underset{\sim}{u}\cdot\nabla)\underset{\sim}{u} + \nabla p = \underset{\sim}{f} \quad \text{in } \Omega,$$

$$(4.2) \qquad \nabla\cdot\underset{\sim}{u} = 0 \quad \text{in } \Omega,$$

$$(4.3) \qquad \underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0}.$$

These equations are nonlinear and *their solution does not correspond to searching for the minimum of a functional.* We can however write them in the form of a variational equation in the following manner. Again putting, as in Section 2,

$$(4.4) \qquad a(\underset{\sim}{u},\underset{\sim}{v}) = \sum_{i,j} \int_\Omega \frac{\partial u_i}{\partial x_j} \frac{\partial v_i}{\partial x_j} \, dx ,$$

and

$$(4.5) \qquad b(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}) = \sum_{i,j} \int_\Omega u_i \frac{\partial v_j}{\partial x_i} w_j dx ,$$

it can be shown (see LIONS [2], TEMAM [1]) that problem (4.1)-(4.3) is equivalent to seeking $\underset{\sim}{u} \in V$ (see (1.16) for the definition of V) satisfying

$$(4.6) \qquad \nu a(\underset{\sim}{u},\underset{\sim}{v}) + b(\underset{\sim}{u},\underset{\sim}{u},\underset{\sim}{v}) = (\underset{\sim}{f},\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in V.$$

Equation (4.6) is deduced from (4.1) by multiplying by a function with zero divergence and then integrating by parts. The pressure term vanishes because we have

$$\int_\Omega \underset{\sim}{\nabla} p \cdot \underset{\sim}{v} \, dx = - \int_\Omega p \, \underset{\sim}{\nabla} \cdot \underset{\sim}{v} \, dx = 0$$

if $\underset{\sim}{v}$ has zero divergence and is zero at the boundary.
Conversely, if $\underset{\sim}{u}$ is a solution of (4.6) it can be shown that there exists $p \in L^2(\Omega)$ such that we have

$$(4.7) \qquad \nu a(\underset{\sim}{u},\underset{\sim}{v}) + b(\underset{\sim}{u},\underset{\sim}{u},\underset{\sim}{v}) - (p, \underset{\sim}{\nabla} \cdot \underset{\sim}{v}) = (\underset{\sim}{f},\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in (H_0^1(\Omega))^N.$$

Thus taking $\underset{\sim}{v} \in (\mathcal{D}(\Omega))^N$, we again obtain (4.1) in the distributional sense.
Note that (4.7) and

$$(4.8) \qquad (q, \underset{\sim}{\nabla} \cdot \underset{\sim}{u}) = 0 \quad \forall q \in L^2(\Omega)$$

are the analogue of the optimality conditions for the Lagrangian (1.21) in the Stokes problem. We can therefore consider (4.7), (4.8) as a "variational Lagrange problem" and we say that a pair $\{\underset{\sim}{u},p\}$ satisfying these conditions constitutes an *equilibrium point* for this problem. We shall now attempt to extend to this new type of case the algorithms used in the preceding sections for the

solution of the Stokes problem.

The underlying idea is closely related to the techniques used in the proofs of existence of a solution, which are based on the search for a fixed point for the mapping T from V into V defined by

$$(4.9)\qquad \begin{cases} \forall \underset{\sim}{v} \in V\ ,\ \underset{\sim}{u} = T(\underset{\sim}{v}) \quad \textit{is a solution in } V \textit{ of} \\ \nu a(\underset{\sim}{u},\underset{\sim}{w}) + b(\underset{\sim}{v},\underset{\sim}{u},\underset{\sim}{w}) = (\underset{\sim}{f},\underset{\sim}{w}) \quad \forall \underset{\sim}{w} \in V, \end{cases}$$

or alternatively

$$(4.10)\qquad -\nu\Delta\underset{\sim}{u} + (\underset{\sim}{v}\cdot\nabla)\underset{\sim}{u} + \nabla p = f \quad \textit{in} \quad \Omega$$

$$(4.11)\qquad \nabla\cdot\underset{\sim}{u} = 0 \quad \textit{in} \quad \Omega$$

$$(4.12)\qquad \underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0}.$$

Problem (4.9) is a linear, nonsymmetric variational problem. We denote by $A(\underset{\sim}{v})$ the linear operator from V into V' (V': dual of V) defined by

$$(4.13)\qquad \langle A(\underset{\sim}{v})\underset{\sim}{u},\underset{\sim}{w}\rangle = \nu a(\underset{\sim}{u},\underset{\sim}{w}) + b(\underset{\sim}{v},\underset{\sim}{u},\underset{\sim}{w}) \quad \forall \underset{\sim}{w} \in V,$$

where $\langle .,. \rangle$ denotes the bilinear form of the duality between V' and V. It can easily be shown (see LIONS [2]) that the trilinear form b is *anti-symmetric* for the last two variables if $\underset{\sim}{u} \in V$, i.e. if $\underset{\sim}{u}$ has zero divergence and is zero at the boundary. Thus if $\underset{\sim}{v} \in V$, we have $b(\underset{\sim}{v},\underset{\sim}{u},\underset{\sim}{u}) = 0$. Then, putting $\underset{\sim}{w} = \underset{\sim}{u}$ in (4.13), we thus have

$$(4.14)\qquad \langle A(\underset{\sim}{v})\underset{\sim}{u},\underset{\sim}{u}\rangle \geq \nu \|\underset{\sim}{u}\|_1^2 .$$

The operator $A(\underset{\sim}{v})$ is therefore V-*elliptic* (see LIONS [2]) and the Lax-Milgram Theorem allows us to infer the existence of a *unique solution* of problem (4.9). This result implies that the operator T is *well defined*. We further deduce from (4.9) the a priori upper bound

$$(4.15)\qquad \|\underset{\sim}{u}\|_1 = \|T(\underset{\sim}{v})\|_1 \leq \frac{\|f\|_{V'}}{\nu} .$$

The operator T thus maps a closed ball (hence a weakly compact set) of V into itself.

Using the *compactness* of the injection of $H_0^1(\Omega)$ into $L^4(\Omega)$, it

can be proved that the operator T is *continuous* for the *weak topology* of V [3]. *Schauder's Fixed-Point Theorem* can thus be applied. Furthermore it is clear that any fixed point of T is a solution of the Navier-Stokes equations.

Uniqueness can be proved for ν "large" or f "small" (see LIONS [2]). Knowing that the existence of a solution can be deduced from a fixed-point theorem, it is natural to try to solve the problem by trying to find this fixed point numerically, by means of an iterative method.

4.2 Basic Algorithm

In the light of what has gone before, we thus consider the following algorithm:

(4.16) $\underset{\sim}{u}^0$ *specified arbitrarily;*
then for $n \geq 0$, *with* $\underset{\sim}{u}^n$ *known, calculate* $\underset{\sim}{u}^{n+\frac{1}{2}}$, p^{n+1} *satisfying*

$$(4.17) \qquad -\nu\Delta\underset{\sim}{u}^{n+1/2} + (\underset{\sim}{u}^n \cdot \underset{\sim}{\nabla})\underset{\sim}{u}^{n+1/2} + \underset{\sim}{\nabla} p^{n+1} = \underset{\sim}{f} \quad \text{in } \Omega$$

$$(4.18) \qquad \underset{\sim}{\nabla} \cdot \underset{\sim}{u}^{n+1/2} = 0 \quad \text{in } \Omega,$$

$$(4.19) \qquad \underset{\sim}{u}^{n+1/2}|_{\Gamma} = \underset{\sim}{0},$$

and then $\underset{\sim}{u}^{n+1}$ *by*

$$(4.20) \qquad \underset{\sim}{u}^{n+1} = \omega\, \underset{\sim}{u}^{n+1/2} + (1-\omega)\underset{\sim}{u}^n.$$

Actually, the case $\omega = 1$ defines $\underset{\sim}{u}^{n+1}$ by

$$(4.21) \qquad \underset{\sim}{u}^{n+1} = T(\underset{\sim}{u}^n).$$

The general case can thus be written

$$(4.22) \qquad \underset{\sim}{u}^{n+1} = \omega\, T(\underset{\sim}{u}^n) + (1-\omega)\underset{\sim}{u}^n.$$

The choice of $\omega \neq 1$ is clearly designed to accelerate the convergence, if possible. It is shown in CROUZEIX [1] that this algorithm converges whenever the solution is unique. In cases of non-uniqueness we may hope to converge to the stable solution "nearest" to $\underset{\sim}{u}^0$,

---

[3] hence *compact*

Remark 4.1: From the Lax-Milgram Theorem and (4.14), the solution $\underset{\sim}{u}^{n+\frac{1}{2}}$ of (4.17)-(4.19) exists whatever the values of υ and $\underset{\sim}{u}^{n}$. By contrast, the convergence is only assured if $\underset{\sim}{u}^{o}$ is sufficiently near to a solution $\underset{\sim}{u}$ that the spectral radius $\rho(T'(\underset{\sim}{u}))$, (where $T'(\underset{\sim}{u})$ is the *Frechet derivative* of T at $\underset{\sim}{u}$) is strictly less than 1. The investigation of these "stable" solutions is a difficult problem which again is largely left open. In practice, the user must be satisfied with proving convergence experimentally. ∎

Remark 4.2: The implementation of algorithm (4.16)-(4.20) requires problem (4.17)-(4.19) to be solved. This problem is in every respect analogous to the Stokes problem, the operator $A(\underset{\sim}{u}^{n}) = (-\nu\Delta+\underset{\sim}{u}^{n}\cdot\nabla)$ being linear and V-elliptic (see (4.14)) though nonsymmetric. We have already mentioned in Chapter I, Section 5, that the convergence of the UZAWA algorithm can in this case be proved by the energy methods of Chapter I, Section 2.1. In order to state this algorithm in explicit fashion we put[4]

$$A_r(\underset{\sim}{u}^{n}) = -\nu\Delta + \underset{\sim}{u}^{n}\cdot\nabla - r\nabla(\nabla\cdot). \tag{4.23}$$

The algorithm is then written

$$p^{n+1,o} \text{ } \textit{specified arbitrarily;} \tag{4.24}$$

then for $s \geq 0$, *with* $p^{n+1,s}$ *known, compute* $u^{n+\frac{1}{2},s}$ *satisfying*

$$\begin{cases} A_r(\underset{\sim}{u}^{n})\underset{\sim}{u}^{n+1/2,s} + \nabla p^{n+1,s} = \underset{\sim}{f} \text{ } \textit{in} \text{ } \Omega, \\ \\ \underset{\sim}{u}^{n+1/2,s}\big|_{\Gamma} = \underset{\sim}{0}, \end{cases} \tag{4.25}$$

and

$$p^{n+1,s+1} = p^{n+1,s} - \rho\nabla\cdot\underset{\sim}{u}^{n+1/2,s}. \tag{4.26}$$

It can be shown that the condition for convergence is in this case[5]

---

[4] $\nabla(\nabla\cdot\underset{\sim}{v}) = \text{grad (div } \underset{\sim}{v})$.

[5] For certain approximations this condition would become

$$0 < \rho < 2(r + \tfrac{\nu}{N}).$$

(4.27) $\quad 0 < \rho < 2(r+\nu)$.

We thus see that the advantage of the penalty term will become greater as ν gets smaller. We shall return to this remark in Section 5. ∎

Remark 4.3: It can immediately be deduced from (4.17)-(4.19) that p^{n+1} is a solution of a nonsymmetric "dual problem" which does not in general correspond to the minimisation of a functional and which can be written

$$(4.28) \qquad \underset{\sim}{\nabla}\cdot(A_r^{-1}(\underset{\sim}{u}^n)\underset{\sim}{\nabla}p^{n+1}) = \underset{\sim}{\nabla}\cdot(A_r^{-1}(\underset{\sim}{u}^n)\underset{\sim}{f}),$$

an expression which is clearly analogous to that obtained for the Stokes problem. Likewise we see from (4.25), (4.26) that $p^{n+1,s+1}$ can be written

$$(4.29) \qquad p^{n+1,s+1} = p^{n+1,s} + \rho\{\underset{\sim}{\nabla}\cdot(A_r^{-1}(\underset{\sim}{u}^n)\underset{\sim}{\nabla}p^{n+1,s}) - \underset{\sim}{\nabla}\cdot(A_r^{-1}(\underset{\sim}{u}^n)\underset{\sim}{f})\} \quad .$$

The convergence of the algorithm depends on the spectrum of the operator (from $L^2(\Omega) \to L^2(\Omega)$)

$$(4.30) \qquad q \to q + \rho\underset{\sim}{\nabla}\cdot(A_r^{-1}(\underset{\sim}{u}^n)\underset{\sim}{\nabla}q).$$

We refer the reader to CROUZEIX [1] for a study of this spectrum which is necessarily somewhat complex. Even in the *finite-dimensional* case, a spectral analysis of the convergence is scarcely possible in view of the non-symmetry of the operator. ∎

Remark 4.4: It is again possible to apply here certain of the variable-step methods of Chapter I, Section 3. As a result of the non-symmetry of the problem, the method of steepest descent loses its meaning and this leads us to favour using the method of minimum residuals, i.e. to determine ρ_s so as to minimise $|\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+1,s+1}|_0^2$, this procedure remaining perfectly justifiable here. Following the notation already introduced and following also Section 3 of this chapter, the algorithm can be written

(4.31) $\quad p^{n+1,0}$ *specified arbitrarily;*
calculate the solution $\underset{\sim}{u}^{n+\frac{1}{2},0}$ *of*

$$(4.32) \qquad A_r(\underset{\sim}{u}^n)\underset{\sim}{u}^{n+1/2,0} + \underset{\sim}{\nabla}p^{n+1,0} = \underset{\sim}{f} .$$

On iteration s, *define the descent direction* w_s *by*

$$w_s = \underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+1/2,s}. \tag{4.33}$$

Then solve the problem

$$A_r(\underset{\sim}{u}^n)\underset{\sim}{z}_s + \underset{\sim}{\nabla} w_s = \underset{\sim}{0}\,, \tag{4.34}$$

and calculate ρ_s *by*

$$\rho_s = -\frac{(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+1/2,s}, \underset{\sim}{\nabla}\cdot\underset{\sim}{z}_s)}{|\underset{\sim}{\nabla}\cdot\underset{\sim}{z}_s|_0^2}\,, \tag{4.35}$$

then $\underset{\sim}{u}^{n+\frac{1}{2},s+1}$ *and* $p^{n+1,s+1}$ *by*

$$p^{n+1,s+1} = p^{n+1,s} - \rho_s w_s\,, \tag{4.36}$$

$$\underset{\sim}{u}^{n+1/2,s+1} = \underset{\sim}{u}^{n+1/2,s} + \rho_s \underset{\sim}{z}_s\,. \tag{4.37}$$

The advantage of such an algorithm is of course that the choice of the parameter ρ, always a rather difficult matter, is carried out in an automatic manner. ∎

Remark 4.5: In the light of the numerical results of Section 3, it would seem effective to employ here a conjugate-gradient method, even though at first sight this may appear difficult in view of the non-symmetry of the problem. Following PAIGE [1] and GOLUB [1] it is, however, possible to construct algorithms of conjugate-gradient type for a non-symmetric problem, by using a least-squares formulation. The algorithms thus constructed require at each iteration the solution of two linear systems (instead of only one in the symmetric case of the Stokes problem). The complexity of the calculation is therefore doubled. An investigation of the application to (4.17)-(4.19) of PAIGE's algorithm, or of a variant better adapted to the particular case we are presently considering, thus appears to be a direction of research which ought to be explored. As a particular reference, we may cite WIDLUND [1]. ∎

Algorithm (4.16)-(4.20), combined with (4.31)-(4.37), thus appears as a method which is simple and *completely automatic* for calculations involving the Navier-Stokes equations. Unfortunately, convergence is not assured and the choice of the parameter ω is an open problem. The use of the *"augmented" operator* $A_r(\underset{\sim}{u}^n)$ in place of $A(\underset{\sim}{u}^n)$ is aimed at accelerating the convergence of

(4.31)-(4.37) at high Reynolds numbers, i.e. when ν is small. Nevertheless the global convergence of (4.16)-(4.20) is unchanged by this procedure and is certainly slower for ν small. One possible strategy is clearly to calculate a sequence of solutions at increasing Reynolds numbers, each time initialising (4.16)-(4.20) with the last solution obtained. In practice, this approach does appear to give good results in the majority of problems. To conclude this section, and before we enter upon the study of variants of the above algorithm, we should point out that certain authors, (OLSON [1], ROACHE [1]) suggest the use of the *Newton-Raphson* method, for solving the nonlinear Navier-Stokes problem. This algorithm is written

(4.38) $\underset{\sim}{u}^0$ *specified arbitrarily*, $\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^0 = 0$:
then for $n \geq 0$, *put*

$$\underset{\sim}{u}^{n+1} = \underset{\sim}{u}^n + \delta\underset{\sim}{u}^n \tag{4.39}$$

where $\delta\underset{\sim}{u}^n, p^n$ *are solutions of*

$$-\nu\Delta\delta\underset{\sim}{u}^n + (\underset{\sim}{u}^n\cdot\underset{\sim}{\nabla})\delta\underset{\sim}{u}^n + (\delta\underset{\sim}{u}^n\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^n + \underset{\sim}{\nabla}p^n = \underset{\sim}{f} + \nu\Delta\underset{\sim}{u}^n - (\underset{\sim}{u}^n\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^n \ , \tag{4.40}$$

$$\underset{\sim}{\nabla}\cdot\delta\underset{\sim}{u}^n = 0 \ , \tag{4.41}$$

$$\delta\underset{\sim}{u}^n|_\Gamma = \underset{\sim}{0} \ . \tag{4.42}$$

Problem (4.40)-(4.42) is very similar to (4.17)-(4.19) and we can adapt to this case the algorithm of Remark 4.4. It is natural to expect that Newton's method, if used, will converge more rapidly than the fixed-point algorithm. However, the operator appearing in (4.40) is not a priori V-elliptic and there is nothing to allow us to assert the *existence* of a solution $\delta\underset{\sim}{u}^n$. The use of Newton's method thus appears more questionable than that of (4.16)-(4.20) and in certain cases we may expect to encounter numerical difficulties in the solution of (4.40)-(4.42). ∎

Remark 4.6: The basic algorithm described in the preceding section requires the solution of a *nonsymmetric* linear system in which the matrix varies from one iteration to another. This latter consideration makes the use of such an algorithm relatively expensive; it therefore seems attractive to look for a method in which the linear part will have a fixed matrix and will thus be solvable in an efficient manner. The methods described in the

preceding sections allow us to obtain, very efficiently, solutions of Stokes problems and more generally of problems of the type given below (with $\alpha \geq 0$)

$$\alpha \underset{\sim}{u} - \nu \Delta \underset{\sim}{u} + \underset{\sim}{\nabla} p = \underset{\sim}{f} \quad in \quad \Omega \ , \tag{4.43}$$

$$\underset{\sim}{\nabla} \cdot \underset{\sim}{u} = 0 \quad in \quad \Omega, \tag{4.44}$$

$$\underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0} \ ; \tag{4.45}$$

it is therefore natural to look for iterative methods which exploit these possibilities to the maximum. Algorithms of this kind are described in GLOWINSKI-MANTEL-PERIAUX [1] and GLOWINSKI [2, Chapter 7]; in the above references the solution of the Navier-Stokes problem is reduced to that of a sequence of problems of the type (4.43)-(4.45) (therefore soluble by the methods of the preceding sections) and of problems of nonlinear Dirichlet type which are solved by least-squares preconditioned conjugate-gradient methods. In particular it should be mentioned that the *alternating direction* methods allow the decoupling, in a simple and efficient manner, of the numerical difficulties associated with the incompressibility, and those associated with the nonlinearity; we direct the reader to the two references given above for further details concerning these methods.

5. VARIANTS AND APPROXIMATIONS OF THE BASIC ALGORITHM OF SECTION 4

The algorithm proposed in Section 4.2 requires, within any one overall fixed-point iteration, the solution of nonsymmetric linear sub-problems by methods analogous to those used in the case of the Stokes problem. A natural idea at this point would be to carry out initially only an incomplete solution of the sub-problems; we have already seen in Section 3 that this method of procedure may in certain cases be advantageous and may accelerate the overall process. To implement this incomplete solution we can proceed in two ways, either by employing a termination test which becomes progressively more stringent as the algorithm proceeds, or by fixing an upper bound to the number of internal iterations.

We shall be adopting this latter method here, and we shall be discussing, within a context where we restrict ourselves to a

single iteration, algorithms which may be viewed as nonlinear variants of the algorithm of Chapter I. We shall content ourselves with the simplest cases, which will nonetheless illustrate the fundamental approach.

5.1 Variants of UZAWA type

In accordance with the general principle stated above, we here consider algorithm (4.16)-(4.20) in which at each step we carry out only a single iteration of (4.24)-(4.26) for the solution of (4.17)-(4.19). In this case the various stages can be condensed into a simplified algorithm, as follows:

(5.1) $\underset{\sim}{u}^0$ *and* p^0 *chosen arbitrarily;*
with $\underset{\sim}{u}^n$ *and* p^n *known, calculate a solution* $\underset{\sim}{u}^{n+\frac{1}{2}}$ *of*

(5.2) $$A_r(\underset{\sim}{u}^n)\underset{\sim}{u}^{n+1/2} + \underset{\sim}{\nabla} p^n = \underset{\sim}{f} ,$$

(5.3) $$\underset{\sim}{u}^{n+1/2}\big|_\Gamma = \underset{\sim}{0},$$

then knowing $\underset{\sim}{u}^{n+\frac{1}{2}}$, *calculate* $\underset{\sim}{u}^{n+1}$ *and* p^{n+1} *by*

(5.4) $$\underset{\sim}{u}^{n+1} = \omega\, \underset{\sim}{u}^{n+1/2} + (1-\omega)\underset{\sim}{u}^n ,$$

(5.5) $$p^{n+1} = p^n - \rho_n \underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{n+1} .$$

This algorithm can immediately be seen as the adaptation to our nonlinear problem of the algorithm of Chapter I, Section 4, in which we had introduced a relaxation parameter into the Uzawa algorithm. Note here, however, that even for $\omega = 1$ an initial value $\underset{\sim}{u}^0$ of the velocity has to be specified; this corresponds to the fact that this algorithm is an approximation of the fixed-point algorithm of Section 4.2.

The choice of ρ_n can be made using the *method of minimum residuals* described in Remark 4.4. One could also consider using a carefully-chosen fixed value of ρ. We should emphasise that the choice of parameters here is a very tricky open problem.

Remark 5.1: The *existence* of a solution of problem (5.2) is assured if the operator $A_r(\underset{\sim}{u}^n)$ is V-elliptic (see (4.14)). We have already shown that such will be the case if $\underset{\sim}{u}^n$ has zero divergence. This result arises from the fact that the trilinear

form $b(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$ defined in (4.5) is anti-symmetric with respect to $\underset{\sim}{v}$ and $\underset{\sim}{w}$ if $\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0$. Unfortunately this condition is not satisfied by $\underset{\sim}{u}^n$. To circumvent this difficulty, we can use (see TEMAM [1]) the anti-symmetrised form

$$\tilde{b}(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}) = b(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w}) + \frac{1}{2}\sum_{i=1}^{N}\int_\Omega (\underset{\sim}{\nabla}\cdot\underset{\sim}{u})\, \underset{\sim}{v}_i \underset{\sim}{w}_i \, dx. \tag{5.6}$$

We observe that we clearly have in this case

$$\tilde{b}(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{v}) = 0 \quad \forall \underset{\sim}{u},\ \forall \underset{\sim}{v} \in (H_0^1(\Omega))^N. \tag{5.7}$$

Furthermore the supplementary term does not perturb the equations because it vanishes in the limit when $\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0$. The existence of $\underset{\sim}{u}^{n+\frac{1}{2}}$ will therefore be assured if we replace $A_r(\underset{\sim}{u}^n)$ by $\tilde{A}_r(\underset{\sim}{u}^n)$ where

$$\langle \tilde{A}_r(\underset{\sim}{u}^n)\underset{\sim}{u},\underset{\sim}{v}\rangle = \nu a(\underset{\sim}{u},\underset{\sim}{v}) + \tilde{b}(\underset{\sim}{u}^n,\underset{\sim}{u},\underset{\sim}{v}). \blacksquare \tag{5.8}$$

We shall not attempt here to prove the convergence of these algorithms. The available proofs (see for example CROUZEIX [1]) are modelled on the proof of the linear case and all culminate in imposing upon the viscosity parameter a condition (ν "sufficiently large") which would appear, in the light of the experimental results, to be much too restrictive and which for all practical purposes limits us to cases where the linear Stokes approximation itself would be adequate. We thus have here an unresolved technical difficulty, associated with our inability to obtain an appropriate upper bound for the nonlinear term. This problem can also be related to that of the non-uniqueness of the solutions. However, from TEMAM [1] we can obtain a result concerning weak convergence in the mean, i.e. $\frac{1}{M}\sum_{m=1}^{M} \underset{\sim}{u}^m$ converges weakly to a solution when $M \to +\infty$. Thus, lacking a proof, we shall have to make do with a heuristic discussion. We first advocate the principle that any convergence can only come about from a dissipative term, this principle being justifiable both from the mathematical point of view and from the physical point of view. In the case $r = 0$, the only dissipative term in algorithm (5.1)-(5.5) is the term $-\nu\Delta\underset{\sim}{u}^{n+1/2}$ which corresponds in (4.1) to the term $-\nu\Delta\underset{\sim}{u}$. The nonlinear term itself is conservative and would be totally incapable of producing any convergence whatsoever towards a steady state.

Now, in the physical applications of most interest the viscosity

parameter is very small, which means that the dissipation contained in our algorithm is very weak. For ν small (or for a large Reynolds number, in the terminology of fluid mechanics) we may therefore expect the convergence to be slow. We thus see the advantage of the penalty term of Hestenes, which for $r \neq 0$ introduces a supplementary dissipation into the algorithm. This supplementary dissipation, which does not change the solution, forces the convergence of the pressure term and adds to the natural dissipation which, however, is alone in being able to ensure the convergence of $\underset{\sim}{u}^n$. The advantage of the method of Hestenes, which we propose here, ought therefore to become much more apparent for small values of the viscosity. It must however be noted that in this case other problems arise, such as the appearance of boundary layers and possibly turbulence; such problems require an appropriate treatment, chiefly with regard to the discretisation, and we shall not enter upon a discussion of such topics here. The numerical solution of the Navier-Stokes equations at large Reynolds numbers is in our opinion still an unresolved problem, and the methods which we describe here cover only one aspect of the question.

5.2 Variants of ARROW-HURWICZ type

We have seen in Chapter I that the analogue (in the linear case) of algorithm (5.1)-(5.5) forms part of a family of algorithms of the ARROW-HURWICZ type, with auxiliary operator. As we have already mentioned in Section 3 of the present chapter, these methods can in practice be implemented as a version of (5.1)-(5.5) in which the solution of (5.2) is carried out incompletely. The advantage, in certain cases, is that the calculations can be carried out in a fully explicit manner. The most general form is written

$$S_r(\underset{\sim}{u}^{n+1}-\underset{\sim}{u}^n) - \omega(A_r(\underset{\sim}{u}^n)\underset{\sim}{u}^n + \underset{\sim}{\nabla} p^n - \underset{\sim}{f}) = \underset{\sim}{0}, \tag{5.9}$$

$$p^{n+1} = p^n - \rho \underset{\sim}{\nabla}\cdot \underset{\sim}{u}^{n+1}. \tag{5.10}$$

We observe that in the case of $S_r = A_r(\underset{\sim}{u}^n)$ and $\omega = 1$, the algorithm reduces to (5.1)-(5.5) with $\omega = 1$. The simplest case is of course $S_r = I$, which gives a fully explicit algorithm. The use of relaxation methods in the solution of (5.2) leads to

nonsymmetric auxiliary operators, given the particular form of $A_r(\underset{\sim}{u}^n)$.

The convergence of this type of algorithm is evidently even more difficult to establish than that of the preceding case, and it would be out of the question to attempt to study it here. Intuitively, the behaviour at low Reynolds numbers ought to be similar to that of the corresponding algorithm in the linear case.

An algorithm of this type has been used (in the case where $S_r = I$, $r = 0$) in FORTIN-PEYRET-TEMAM [1], up to Reynolds numbers of the order of 10^3. The speed of convergence diminishes considerably as ν decreases, and the choice of the parameters becomes critical.

5.3 Numerical results

The tests carried out are fragmentary but nonetheless enable some insight to be gained into the effect of the penalty term on the convergence of the algorithm. The numerical results we have available are mainly for algorithm (5.1)-(5.5).

The first experimental finding is that the study carried out in the linear case still remains valid qualitatively. The optimal value of ω is less than 1, i.e. it is preferable to *"under-relax"*. The problem taken was analogous to the one introduced for the numerical experiments in the linear case, and the values of ν tested were 1 and 1/10. The optimal value of ω was, for these two values of ν, around 0.7, and that of ρ was around 1.5 for $r = 0$, and around 7 for $r = 5$. The speed of convergence diminished with ν but much less for $r = 5$ than for $r = 0$. The value of ρ seems to be more or less independent of ν. By contrast, the solution of (5.2) is clearly more difficult for ν small.

In the case of algorithm (5.9)-(5.10), the results we have available are very incomplete. The speed of convergence diminishes with ν but it seems here to be a real advantage to take $r > 0$ for ν small, contrary to what happened in the linear case. The reason is undoubtedly that the dissipation is increased, at least in a sub-space of the space of "admissible" solutions.

6. NAVIER-STOKES EQUATIONS. TIME-DEPENDENT CASE.

This section does not pretend to be exhaustive. The aim is simply to show that the techniques developed in the steady-state case remain usable in the time-dependent case, and to illustrate this assertion through a simple example. The methods proposed can easily be adapted to more complex situations and to more elaborate schemes .

6.1 Statement of the problem

We consider here the nonlinear Navier-Stokes equations in the time-dependent case. We thus seek a solution $\underset{\sim}{u}(x,t)$ of

$$\frac{\partial \underset{\sim}{u}}{\partial t} - \nu\Delta\underset{\sim}{u} + (\underset{\sim}{u}\cdot\nabla)\underset{\sim}{u} + \nabla p = \underset{\sim}{f}, \tag{6.1}$$

$$\nabla\cdot\underset{\sim}{u} = 0, \tag{6.2}$$

$$\underset{\sim}{u}(x,o) = \underset{\sim}{u}_o(x) \quad \textit{given}, \tag{6.3}$$

$$\underset{\sim}{u}|_\Gamma = \underset{\sim}{0}. \tag{6.4}$$

We consider a *discretisation* with respect to time by a very simple scheme of *implicit* type ($k = \Delta t$):

$$\frac{\underset{\sim}{u}^{n+1} - \underset{\sim}{u}^{n}}{k} - \nu\Delta\underset{\sim}{u}^{n+1} + (\underset{\sim}{u}^{n+1}\cdot\nabla)\underset{\sim}{u}^{n+1} + \nabla p^{n+1} = \underset{\sim}{f}^{n+1}, \tag{6.5}$$

with $\underset{\sim}{f}^{n+1} = \underset{\sim}{f}(x,(n+1)k)$,

$$\nabla\cdot\underset{\sim}{u}^{n+1} = 0, \tag{6.6}$$

$$\underset{\sim}{u}^{o} = \underset{\sim}{u}_o, \tag{6.7}$$

$$\underset{\sim}{u}^{n+1}|_\Gamma = \underset{\sim}{0}. \tag{6.8}$$

In practice we shall obviously be attempting to solve a discretised form of this problem. We could, for example, use the finite-element method described in Section 2 of this chapter.

The important point to consider is that, $\underset{\sim}{u}^{n}$ being known, the solution of (6.5)-(6.8) is a problem of the same type as those described in Section 5; the operators involved are merely modified slightly. More specifically, we have to solve a problem of the type

$$\underset{\sim}{u} - k\nu\Delta\underset{\sim}{u} + k(\underset{\sim}{u}\cdot\nabla)\underset{\sim}{u} + k\nabla p = \underset{\sim}{F}, \tag{6.9}$$

$$\nabla\cdot\underset{\sim}{u} = 0 \quad in \quad \Omega, \tag{6.10}$$

$$\underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0}. \tag{6.11}$$

Remark 6.1: Instead of (6.5) we could consider a semi-implicit scheme, replacing (6.5) by

$$\frac{\underset{\sim}{u}^{n+1}-\underset{\sim}{u}^{n}}{k} - \nu\Delta\underset{\sim}{u}^{n+1} + (\underset{\sim}{u}^{n}\cdot\nabla)\underset{\sim}{u}^{n+1} + \nabla p^{n+1} = \underset{\sim}{f}. \tag{6.12}$$ ■

Remark 6.2: (6.5)-(6.8) can be written in variational form . With the notation of Section 4, we obtain

$$\begin{cases} (\frac{\underset{\sim}{u}^{n+1}-\underset{\sim}{u}^{n}}{k}, \underset{\sim}{v}) + \nu a(\underset{\sim}{u}^{n+1},\underset{\sim}{v}) + b(\underset{\sim}{u}^{n+1},\underset{\sim}{u}^{n+1},v) = (\underset{\sim}{f}^{n+1},\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in V, \\ \underset{\sim}{u}^{n+1} \in V, \end{cases} \tag{6.13}$$

$$\underset{\sim}{u}^{0} \in V \quad given. \tag{6.14}$$

By virtue of the anti-symmetry of the form $b(\underset{\sim}{u},\underset{\sim}{v},\underset{\sim}{w})$, it can easily be shown that such implicit schemes are unconditionally stable. ■

Remark 6.3: The existence and uniqueness of $\underset{\sim}{u}^{n+1}$ in the *semi-implicit* scheme defined by (6.12) pose no problems. In fact, $\underset{\sim}{u}^{n+1}$ is the solution of a linear problem relating to the operator $I+k(-\nu\Delta+\underset{\sim}{u}^{n}\cdot\nabla)$ which is V-elliptic. ■

By using, as in Section 4, a fixed-point theorem, the existence of a solution $\underset{\sim}{u}$ of problem (6.9)-(6.11) can be demonstrated without difficulty. We shall now show that if we consider an approximation of (6.9)-(6.11) in a space of *finite dimensions* (for example by means of finite elements as in Section 2), we have uniqueness of the solution for k sufficiently small.

As in Section 4, we denote by $|.|_0$ and $\|\cdot\|_1$ the respective norms of $(L^2(\Omega))^N$ and V. We shall in fact be working within a space $V_h \subset V$, where in the majority of cases the parameter h represents the size of the mesh used for the approximation. There then exists a *mesh-dependent* constant $S(h)$ such that

$$\|\underset{\sim}{v}\|_1 \leq S(h)\,|\underset{\sim}{v}|_0, \quad \forall \underset{\sim}{v} \in V_h. \tag{6.15}$$

We can now prove the following:

LEMMA 6.1: *In finite dimensions, and for the spatial dimension* N = 2, *the solution of the problem* (6.9)-(6.11) *is unique if* k *is sufficiently small.*

Proof: Suppose $\underset{\sim}{u}_1$ and $\underset{\sim}{u}_2$ are two solutions of the problem. In variational form we can thus write

$$(6.16) \qquad (\underset{\sim}{u}_1,\underset{\sim}{v}) + k\nu a(\underset{\sim}{u}_1,\underset{\sim}{v}) + kb(\underset{\sim}{u}_1,\underset{\sim}{u}_1,\underset{\sim}{v}) = (\underset{\sim}{f},\underset{\sim}{v}) \quad \forall\, \underset{\sim}{v} \in V,$$

$$(6.17) \qquad (\underset{\sim}{u}_2,\underset{\sim}{v}) + k\nu a(\underset{\sim}{u}_2,\underset{\sim}{v}) + kb(\underset{\sim}{u}_2,\underset{\sim}{u}_2,\underset{\sim}{v}) = (\underset{\sim}{f},\underset{\sim}{v}) \ \forall \underset{\sim}{v} \in V\ ;$$

subtracting (6.17) from (6.16) and putting $\underset{\sim}{v} = \underset{\sim}{u}_1 - \underset{\sim}{u}_2$, we obtain, after various manipulations on the nonlinear terms and taking advantage of the anti-symmetry of b (see Section 4)

$$(6.18) \qquad |\underset{\sim}{u}_1 - \underset{\sim}{u}_2|_o^2 + k\nu \|\underset{\sim}{u}_1 - \underset{\sim}{u}_2\|_1^2 + kb(\underset{\sim}{u}_1 - \underset{\sim}{u}_2, \underset{\sim}{u}_2, \underset{\sim}{u}_1 - \underset{\sim}{u}_2) = 0.$$

Utilising the inequality of Cagliardo (see LIONS [2]), we can, in two dimensions, obtain the following upper bound for the nonlinear term:

$$(6.19) \qquad |b(\underset{\sim}{u}_1 - \underset{\sim}{u}_2, \underset{\sim}{u}_2, \underset{\sim}{u}_1 - \underset{\sim}{u}_2)| \le C |\underset{\sim}{u}_1 - \underset{\sim}{u}_2|_o \|\underset{\sim}{u}_2\|_1 \|\underset{\sim}{u}_1 - \underset{\sim}{u}_2\|_1,$$

hence, by (6.15),

$$(6.20) \qquad |b(\underset{\sim}{u}_1 - \underset{\sim}{u}_2, \underset{\sim}{u}_2, \underset{\sim}{u}_1 - \underset{\sim}{u}_2)| \le CS(h)\ \|\underset{\sim}{u}_2\|_1\ |\underset{\sim}{u}_1 - \underset{\sim}{u}_2|_o^2.$$

Substituting this upper bound into (6.18), we obtain

$$(6.21) \qquad (1 - kCS(h)\ \|\underset{\sim}{u}_2\|_1)|\underset{\sim}{u}_1 - \underset{\sim}{u}_2|_o^2 + k\nu\ \|\underset{\sim}{u}_1 - \underset{\sim}{u}_2\|_1^2 \le 0.$$

We shall therefore have $\underset{\sim}{u}_1 = \underset{\sim}{u}_2$ if we can choose k such that we have

$$(6.22) \qquad (1 - kCS(h)\ \|\underset{\sim}{u}_2\|_1) \ge 0.$$

Still using (6.15), we can transform (6.22) into

$$(6.23) \qquad kCS^2(h)\ |\underset{\sim}{u}_2|_o \le 1\ .$$

But setting $\underset{\sim}{v} = \underset{\sim}{u}_2$ in (6.17), it can easily be seen that $|\underset{\sim}{u}_2|_o$ is bounded above by a constant depending only on $\underset{\sim}{f}$. It is thus possible to verify (6.23), and hence the result. ∎

6.2 Solution algorithms

We have already noted the analogy between problem (6.9)-(6.11) and the cases discussed in Sections 4 and 5 of this chapter. We can therefore use, for its solution, algorithms closely related to those described for the steady-state Navier-Stokes equations. To illustrate the ideas, we shall discuss here an algorithm of the UZAWA type.

UZAWA algorithm, semi-implicit case

Suppose we have $\bar{\underset{\sim}{u}} \in V$ given. We wish to solve

$$\underset{\sim}{u} - k\nu\Delta\underset{\sim}{u} + k(\bar{\underset{\sim}{u}}\cdot\underset{\sim}{\nabla})\underset{\sim}{u} + k\underset{\sim}{\nabla}p = \underset{\sim}{F}\,, \tag{6.24}$$

$$\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0\,, \tag{6.25}$$

$$\underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0}\,. \tag{6.26}$$

Remark 6.4: In practice, $\bar{\underset{\sim}{u}} = \underset{\sim}{u}^n$, $\underset{\sim}{F} = \underset{\sim}{u}^n + k\underset{\sim}{f}$. ■

In order to solve (6.9)-(6.11), we can therefore use the following algorithm:

(6.27) p^0 *specified arbitrarily;*
with p^s *known, calculate the solution* $\underset{\sim}{u}^s$ *of*

$$\begin{cases} \underset{\sim}{u}^s - k\nu\Delta\underset{\sim}{u}^s - kr\underset{\sim}{\nabla}(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^s) + k(\bar{\underset{\sim}{u}}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^s + k\underset{\sim}{\nabla}p^s = \underset{\sim}{F}\,, \\ \underset{\sim}{u}^s|_{\Gamma} = \underset{\sim}{0}\,, \end{cases} \tag{6.28}$$

then p^{s+1} *by*

$$p^{s+1} = p^s - \rho_s\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^s. \tag{6.29}$$

We note that problem (6.28) is linear (nonsymmetric). The convergence of this algorithm is a direct consequence of the results of Chapter I. In particular, we have for ρ_s the condition

$$0 < \rho_s < 2(r+\nu). \tag{6.30}$$

It is also possible, as in Section 4, to utilise the method of minimum residuals for the determination of ρ_s.

Remark 6.5: The use of such an iterative method in a time-dependent scheme would clearly be prohibitive if the convergence at

each time-step were not extremely rapid. In actual fact the situation here is much more favourable than in the steady-state case. In practice

- the initialisation can be effected using p^n which will in general be close to p^{n+1},
- the problem is well conditioned for k sufficiently small and r sufficiently large.

In practice, we may expect to carry out only a very small number of iterations at each time step (2 or 3). ∎

We are now going to consider the fully implicit case and show that the UZAWA algorithm is still applicable.

UZAWA algorithm in the implicit scheme case.

We have already shown in Section 6.1 that the implicit scheme (6.5)-(6.8) is well posed for k sufficiently small. We shall now see that, under the same conditions, we can calculate $\underset{\sim}{u}^{n+1}$ by a nonlinear UZAWA algorithm. Since the calculation of $\underset{\sim}{u}^{n+1}$ is equivalent to the solution of (6.9)-(6.11), we shall consider the solution of this latter problem. We shall utilise here the anti-symmetrised form of the nonlinear term introduced earlier in Section 5.1. The algorithm considered is then written:

(6.31) $\underset{\sim}{u}^0$, p^1 *chosen arbitrarily;*
with $\underset{\sim}{u}^{s-1}$ *and* p^s *known, calculate the solution* $\underset{\sim}{u}^s$ *of*

$$(6.32)\qquad \begin{cases} \underset{\sim}{u}^s - k\nu\Delta\underset{\sim}{u}^s - kr\underset{\sim}{\nabla}(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^s) + k(\underset{\sim}{u}^{s-1}\cdot\underset{\sim}{\nabla})\underset{\sim}{u}^s + \frac{k}{2}(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^{s-1})\,\underset{\sim}{u}^s + k\underset{\sim}{\nabla}p^s = \underset{\sim}{F}\,, \\ \underset{\sim}{u}^s = \underset{\sim}{0} \quad \textit{on } \Gamma, \end{cases}$$

then p^{s+1} *by*

$$(6.33)\qquad p^{s+1} = p^s - \rho(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^s).$$

In *variational form*, we can write (6.32) as follows:

$$(6.34)\qquad \begin{cases} (\underset{\sim}{u}^s,\underset{\sim}{v}) + k\nu a(\underset{\sim}{u}^s,\underset{\sim}{v}) + kr(\underset{\sim}{\nabla}\cdot\underset{\sim}{u}^s,\underset{\sim}{\nabla}\cdot\underset{\sim}{v}) + k\tilde{b}(\underset{\sim}{u}^{s-1},\underset{\sim}{u}^s,\underset{\sim}{v}) - \\ - k(p^s,\underset{\sim}{\nabla}\cdot\underset{\sim}{v}) = (\underset{\sim}{F},\underset{\sim}{v}) \quad \forall \underset{\sim}{v} \in (H_0^1(\Omega))^N\,,\ \underset{\sim}{u}^s \in (H_0^1(\Omega))^N. \end{cases}$$

Remark 6.5: In the context of the implicit scheme (6.5)-(6.8), a natural choice would be to take $\underset{\sim}{u}^0 = \underset{\sim}{u}^n$, $p^1 = p^n$. ∎

Remark 6.6: Problem (6.32) is linear ($\underset{\sim}{u}^{s-1}$ being known) and non-symmetric. Use of the anti-symmetrised form $\tilde{b}$ allows us to state the existence of a unique solution $\underset{\sim}{u}^s$, the operator being $(H_0^1(\Omega))^N$ - elliptic. ∎

In the following discussion, we consider algorithm (6.31)-(6.33) in *finite dimensions*, i.e. for the solution of a discretised version of (6.5)-(6.8). We have already proved a uniqueness result for this case in Lemma 6.1. We shall now prove the following result:

PROPOSITION 6.1: *In finite dimensions, and for the spatial dimension* N = 2, *the sequence* $\underset{\sim}{u}^s$ *converges for* k *sufficiently small, to the solution* $\underset{\sim}{u}$ *of problem* (6.9)-(6.11). *The sequence* p^s *is bounded and for any cluster point* p^* *of this sequence the pair* $\{\underset{\sim}{u}, p^*\}$ *satisfies* (6.9)-(6.11).

Proof: We write (6.9) in variational form. Putting $\bar{\underset{\sim}{u}}^s = \underset{\sim}{u}^s - \underset{\sim}{u}$ and $\bar{p}^s = p^s - p$, we obtain by subtracting (6.9) from (6.32)

$$|\bar{\underset{\sim}{u}}^s|_o^2 + k\nu \|\bar{\underset{\sim}{u}}^s\|_1^2 + kr|\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s|_o^2 + kb\tilde{}(\bar{\underset{\sim}{u}}^{s-1}, \underset{\sim}{u}, \bar{\underset{\sim}{u}}^s) - k(\bar{p}^s, \underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s) = 0 . \tag{6.35}$$

Taking the inner product of (6.33) with p^{s+1} and using (6.10) we obtain

$$\frac{1}{2}|\bar{p}^{s+1}|_o^2 - \frac{1}{2}|\bar{p}^s|_o^2 + \frac{1}{2}|\bar{p}^{s+1} - \bar{p}^s|_o^2 + \rho(\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s, \bar{p}^{s+1}) = 0. \tag{6.36}$$

We can write the last term of (6.36) in the form

$$\rho(\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s, \bar{p}^{s+1}) = \rho(\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s, \bar{p}^s) + \rho(\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s, \bar{p}^{s+1} - \bar{p}^s). \tag{6.37}$$

Furthermore, by the Cauchy-Schwarz inequality we have

$$|\rho(\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s, \bar{p}^{s+1} - \bar{p}^s)| \le \frac{\varepsilon}{2}|\bar{p}^{s+1} - \bar{p}^s|_o^2 + \frac{1}{2\varepsilon}\rho^2|\underset{\sim}{\nabla}\cdot\bar{\underset{\sim}{u}}^s|_o^2. \tag{6.38}$$

We multiply (6.35) by ρ and (6.36) by k. Taking the sum and using (6.37) and (6.38), we see that

$$(6.39)\qquad \begin{cases} \rho|\bar{\underset{\sim}{u}}^s|_o^2 + \rho k\nu \|\bar{\underset{\sim}{u}}^s\|_1^2 + \rho k(r-\frac{\rho}{2\varepsilon})|\nabla\cdot\bar{\underset{\sim}{u}}^s|_o^2 + \rho k\tilde{b}(\bar{\underset{\sim}{u}}^{s-1},\underset{\sim}{u},\bar{\underset{\sim}{u}}^s) + \\ + \frac{k}{2}|\bar{p}^{s+1}|_o^2 - \frac{k}{2}|\bar{p}^s|_o^2 + \frac{k}{2}(1-\varepsilon)|\bar{p}^{s+1}-\bar{p}^s|_o^2 \le 0. \end{cases}$$

As in Lemma 6.1, we obtain an upper bound for the nonlinear term by using Cagliardo's inequality

$$(6.40)\qquad |\rho k\tilde{b}(\bar{\underset{\sim}{u}}^{s-1},\underset{\sim}{u},\bar{\underset{\sim}{u}}^s)| \le \rho kC|\bar{\underset{\sim}{u}}^{s-1}|_o^{\frac{1}{2}}\|\bar{\underset{\sim}{u}}^{s-1}\|_1^{\frac{1}{2}}\|\underset{\sim}{u}\|_1|\bar{\underset{\sim}{u}}^s|_o^{\frac{1}{2}}\|\bar{\underset{\sim}{u}}^s\|_1^{\frac{1}{2}}.$$

By the equivalence of norms (6.15) and the Cauchy-Schwarz inequality, we obtain

$$(6.41)\qquad |\rho k\tilde{b}(\bar{\underset{\sim}{u}}^{s-1},\underset{\sim}{u},\bar{\underset{\sim}{u}}^s)| \le \frac{\rho kCS(h)\|\underset{\sim}{u}\|_1}{2}(|\bar{\underset{\sim}{u}}^s|_o^2+|\bar{\underset{\sim}{u}}^{s-1}|_o^2).$$

Substituting into (6.39) we obtain

$$(6.42)\qquad \begin{cases} \rho(1-\frac{kCS(h)}{2}\|\underset{\sim}{u}\|_1)|\bar{\underset{\sim}{u}}^s|_o^2 - \rho\frac{kCS(h)}{2}\|\underset{\sim}{u}\|_1|\bar{\underset{\sim}{u}}^{s-1}|_o^2 + \rho k\nu\|\bar{\underset{\sim}{u}}^s\|_1^2 \\ + \rho k(r-\frac{\rho}{2\varepsilon})|\nabla\cdot\bar{\underset{\sim}{u}}^s|_o^2 + \frac{k}{2}|\bar{p}^{s+1}|_o^2 - \frac{k}{2}|\bar{p}^s|_o^2 + \frac{k}{2}(1-\varepsilon)|\bar{p}^{s+1}-\bar{p}^s|_o^2 \le 0. \end{cases}$$

We will be able to apply the argument of Theorem 2.1 of Chapter I if we have

$$(6.43)\qquad 1 - \frac{kCS(h)}{2}\|\underset{\sim}{u}\|_1 \ge \frac{kCS(h)}{2}\|\underset{\sim}{u}\|_1,$$

i.e.

$$(6.44)\qquad \frac{kCS(h)}{2}\|\underset{\sim}{u}\|_1 \le 1 .$$

Thus we again have the *uniqueness condition* of Lemma 6.1. The statement of the proposition can be deduced from (6.42) by the usual procedures. ∎

The principles behind this proof could in fact be extended to problems of the same type associated with other implicit schemes which are more accurate than (6.5)-(6.8). This algorithm is very similar to those of Section 5.1 and it is of course possible to

introduce into it variants such as the use of a relaxation parameter. The fact that we have been able to prove convergence here is clearly linked with the existence of a uniqueness result. ■

7. GENERAL DISCUSSION ON CHAPTER II

The results obtained for the solution of the Stokes problem show that the use of an augmented Lagrangian can be an efficient method of approach for this problem. Its attraction ought to become even more pronounced in three-dimensional problems where the size of the matrices involved makes the use of direct methods difficult. The extension to the nonlinear case is often found to be efficient. The analogy between the ARROW-HURWICZ type algorithm of Section 5.2 and the method of CHORIN [1] should be mentioned. A similar approach has been used in FORTIN-PEYRET-TEMAM [1] and in BEGIS [1], this latter work relating to the calculation of non-Newtonian fluids.

In the nonlinear case, the benefit of the penalty term lies in the fact that it improves the dissipation of the system without perturbing the solution. We shall see in the following chapters that it can sometimes actually bring about the convergence of algorithms which would otherwise be ill-posed.

This Page Intentionally Left Blank

CHAPTER III

ON DECOMPOSITION-COORDINATION METHODS USING AN AUGMENTED LAGRANGIAN

M. Fortin, R. Glowinski

1. INTRODUCTION

1.1 Motivation. Examples.

A large number of problems in Mechanics, in Physics, in Economics, etc... (see Chapter IV, V, VI, VII, VIII) can be stated in the form

$$(1.1) \qquad (P) \quad \operatorname{Min}_{v \in V} \{F(Bv) + G(v)\} ,$$

where

* V, H are normed vector spaces (real for simplicity) of finite or infinite dimension,
* $B \in \mathcal{L}(V,H)$,
* F, G are functionals which are convex, proper, and lower semi-continuous (l.s.c.) on, respectively, H and V.

The formulation (1.1) is quite general, since as will be seen in the Chapters which follow (see also ROCKAFELLER [4], EKELAND-TEMAM [1]), such a formulation encompasses the minimisation of functionals which are possibly nondifferentiable over convex sets, the nondifferentiability or the constraint relating to v and/or Bv. We illustrate this by two examples:

Example 1: *Flow of a Bingham fluid in a cylindrical duct.*

Let Ω be a bounded open domain in $\mathbb{R}^2$ with regular boundary Γ (Ω is the *cross section* of the duct). We define V, H by

$$(1.2) \qquad V = H_0^1(\Omega),$$

(1.3) $\quad H = (L^2(\Omega))^2.$

Let ν and g be *two positive constants* (refer to Chapter V, Section 5 for the physical meaning of ν and g and of the example considered). Now suppose we have the problem

(1.4) $$\operatorname*{Min}_{v\in V} \left[\frac{\nu}{2} \int_\Omega |\nabla v|^2 dx + g \int_\Omega |\nabla v| dx - \int_\Omega fv dx \right],$$

where, in (1.4), $f \in L^2(\Omega)$ (actually f = constant in the applications considered).

Problem (1.4) is obviously a particular form of problem (P) obtained by putting

(1.5) $\quad B = \nabla$

and by defining F, G by

(1.6) $$F(q) = \frac{\nu}{2} \int_\Omega |q|^2 dx + g \int_\Omega |q| dx ,$$

(1.7) $$G(v) = - \int_\Omega fv dx.$$

In (1.6) we have put $|q| = \sqrt{q_1^2+q_2^2}$.

An alternative choice for F, G is given by

(1.8) $$F(q) = g \int_\Omega |q| dx ,$$

(1.9) $$G(v) = \frac{\nu}{2} \int_\Omega |\nabla v|^2 dx - \int_\Omega fv dx .$$

The above functions F, G are convex, and continuous, and F is *nondifferentiable* on H because of the presence of the term $\int_\Omega |q| dx$.

The choices (1.6), (1.7) or (1.8), (1.9) will lead to slightly different algorithms for the solution of (1.4), by the methods described in Section 3. Incidentally this possibility of choice in the decomposition allows us to predict a certain degree of versatility in the use of the methods studied in this chapter.

Example 2: *Elastoplastic torsion of a cylindrical bar.*

Let Ω be a bounded open domain in $\mathbb{R}^2$, with regular boundary Γ (Ω is the *cross section* of the bar). We define V, H by (1.2), (1.3) and we consider the problem

$$\text{(1.10)} \qquad \min_{v \in K} \left[\frac{1}{2} \int_\Omega |\nabla v|^2 dx - \int_\Omega fv dx \right] ,$$

where $f \in L^2(\Omega)$ (f = constant in the fluid mechanics applications considered), and

$$\text{(1.11)} \qquad K = \{v \mid v \in H_0^1(\Omega),\ |\nabla v| \leq 1 \text{ a.e.}\} .$$

We refer to Chapter V, Section 5, for the physical meaning of (1.10), (1.11). Problem (1.10) is also a particular form of problem (P) obtained by putting

$$\text{(1.12)} \qquad B = \nabla ,$$

with F, G defined by

$$\text{(1.13)} \qquad F(q) = \frac{1}{2} \int_\Omega |q|^2 dx + I_{\hat{K}}(q) ,$$

$$\text{(1.14)} \qquad G(v) = - \int_\Omega fv dx .$$

In (1.13), $I_{\hat{K}}$ denotes the *indicator function* of the convex set

$$\text{(1.15)} \qquad \hat{K} = \{q \mid q \in H,\ |q| \leq 1 \text{ a.e.}\} .$$

We thus have, by definition of the indicator function,

$$\text{(1.16)} \qquad \begin{cases} I_{\hat{K}}(q) = 0 \text{ *if* } q \in \hat{K}, \\ I_{\hat{K}}(q) = +\infty \text{ *if* } q \notin \hat{K}. \end{cases}$$

Since the convex set $\hat{K}$ is *nonempty* ($0 \in \hat{K}$) and *closed* in H, $I_{\hat{K}}$ is *convex, proper, and l.s.c.* on H; it thus follows that F satisfies the same properties. Furthermore G is clearly convex and continuous on V.

An alternative choice for F, G is given by

$$\text{(1.17)} \qquad F = I_{\hat{K}}$$

$$(1.18) \qquad G(v) = \frac{1}{2}\int_\Omega |\nabla v|^2 dx - \int_\Omega fv dx.$$

The various remarks made in relation to Example 1 apply equally to Example 2. ∎

1.2 Principle of the method

The essential idea in the whole of the following discussion is based on the fact that there is trivially an equivalence between (P) and

$$(1.19) \qquad (\Pi) \quad \operatorname*{Min}_{\{v,q\} \in W} \{F(q)+G(v)\},$$

with

$$(1.20) \qquad W = \{\{v,q\} \in V \times H,\ Bv-q = 0\}.$$

We have thus introduced a supplementary variable q, linked to v through the *linear equality relation* $Bv = q$. To handle this constraint we shall, as in Chapter I, utilise a *Lagrange multiplier* and reduce the problem (Π) (thus also (P)) to a *saddle-point* problem.

In the following discussion we shall assume that the spaces V and H are *Hilbert*[1] spaces; H is identified with its dual and we denote by $(.,.)$ the inner product in H, and by $|.|$ the associated norm (in certain cases the results will apply to the case where H is a reflexive Banach space.) We then define, for $v \in V$, $q \in H$, $\mu \in H$ the Lagrangian

$$(1.21) \qquad \mathcal{L}(v,q,\mu) = F(q) + G(v) + (\mu, Bv-q),$$

and then for $r \geq 0$, the *augmented Lagrangian*

$$(1.22) \qquad \mathcal{L}_r(v,q,\mu) = \mathcal{L}(v,q,\mu) + \frac{r}{2}|Bv-q|^2.$$

In Section 2 we shall study the problem of the existence of saddle-points for $\mathcal{L}$ and $\mathcal{L}_r$ and the relations between the saddle-points of $\mathcal{L}$ and those of $\mathcal{L}_r$. ∎

The approach followed up to now may seem somewhat contrived *and* complicated, since we have introduced a supplementary variable q and a supplementary constraint $Bv - q = 0$. By doing this we have in actual fact *simplified the nonlinear structure* of problem (P) by *decoupling* F and B. To illustrate this fact we shall apply the

[1] There will thus be no difficulties in finite dimensions.

above principle to the two examples of Section 1.1; we remark that the numerical utilisation of this type of method seems to have been first introduced by GLOWINSKI-MARROCCO [1] for the numerical solution of the particular problem $(1 < p < \infty)$

$$(1.23)\qquad \begin{cases} -\nabla\cdot(|\nabla u|^{p-2}\nabla u) = f \quad \textit{in } \Omega, \\ u|_{\Gamma} = 0 \end{cases}$$

to which we shall return in Chapter V.

Application to Example 1 *of Section* 1.1 *(Flow of a Bingham fluid)*:

With the functionals F and G defined by (1.6),(1.7), the Lagrangian is written

$$(1.24)\qquad \mathcal{L}(v,q,\mu) = \frac{\nu}{2}\int_{\Omega}|q|^2dx + g\int_{\Omega}|q|dx - \int_{\Omega}fvdx + \int_{\Omega}\mu\cdot(\nabla v - q)dx,$$

so that the augmented Lagrangian is

$$(1.25)\qquad \mathcal{L}_r(v,q,\mu) = \mathcal{L}(v,q,\mu) + \frac{r}{2}\int_{\Omega}|\nabla v - q|^2dx.$$

We note immediately that $\mathcal{L}$ is *linear* with respect to v and hence that $\mathcal{L}_r$ is *quadratic* (with positive-definite quadratic part). This implies that, with μ and q fixed, we can minimise $\mathcal{L}_r$ with respect to v on V, whereas this operation is impossible with $\mathcal{L}$ (this drawback disappears if F and G are defined by (1.8),(1.9)). It follows from this that certain algorithms (see Section 3) will be applicable to the calculation of the saddle-points of $\mathcal{L}_r$ but not for those of $\mathcal{L}$. We shall now give the optimality conditions characterising any saddle-point $\{u,p,\lambda\}$ of $\mathcal{L}_r$ on $H_0^1(\Omega)\times(L^2(\Omega))^2\times(L^2(\Omega))^2$. We obtain the system of equations and (variational) inequalities

$$(1.26)\qquad r\int_{\Omega}\nabla u\cdot\nabla v dx - r\int_{\Omega}p\cdot\nabla v dx + \int_{\Omega}\lambda\cdot\nabla v dx - \int_{\Omega}fvdx = 0,\ \forall v\in H_0^1(\Omega),$$

$$(1.27)\qquad \begin{cases} (\nu+r)\displaystyle\int_{\Omega}p\cdot(q-p)dx + g\int_{\Omega}|q|dx - g\int_{\Omega}|p|dx - \int_{\Omega}(\lambda+r\nabla u)\cdot(q-p)dx \geq 0, \\ \forall q\in(L^2(\Omega))^2, \end{cases}$$

$$(1.28)\qquad \nabla u - p = 0.$$

Note that problem (1.26) is, with p and λ fixed, a problem which is linear with respect to u, and of a completely standard type. *The variational inequality* (1.27) expresses the fact that,

for specified u and λ, p gives the minimum of $\mathcal{L}(u,q,\lambda)$. We have here a minimisation problem relating to a nondifferentiable functional for which the optimality condition is expressed by an inequality (see G.L.T. [1], [2] and LIONS [1]). Problem (1.27) offers a considerable advantage over the initial problem (1.4). In fact we can reduce it to a family of point problems of the type

$$(1.29)\qquad \begin{cases} \xi \in \mathbb{R}^2 \\ \inf\limits_{\xi} \left[\frac{(\nu+r)}{2} |\xi|^2 + g|\xi| - (d(x),\xi) \right] \end{cases}$$

where $|.|$ and $(.,.)$ denote the norm and the inner product on $\mathbb{R}^2$ and where we have put, at the point $x \in \Omega$,

$$(1.30)\qquad d(x) = \lambda(x) + r\nabla u(x).$$

Denoting the solution of (1.29) by $\bar{\xi}(x)$, we then have

$$(1.31)\qquad \begin{cases} \bar{\xi}(x) = 0 \text{ if } g \geq |d(x)| , \\ \bar{\xi}(x) = \frac{1}{\nu+r} \left(d(x) - g \frac{d(x)}{|d(x)|} \right) \text{ if } g < |d(x)| . \end{cases}$$

It can easily be verified that p defined by $p(x) = \bar{\xi}(x)$ is a solution of (1.27).

The attraction of the decomposition carried out is thus that it transforms the "global" nondifferentiable problem (1.4) into a family of local problems, coordinated via the Lagrange multiplier λ. This transformation of course only becomes helpful if we can solve the system (1.26)-(1.28) by an appropriate algorithm.

Application to Example 2 of Section 1.1 (Elastoplastic torsion):

With the functionals F and G now defined by (1.13), (1.14), the Lagrangian is written:

$$(1.32)\qquad \mathcal{L}(v,q,\mu) = \frac{1}{2} \int_\Omega |q|^2 dx + I_{\hat{K}}(q) - \int_\Omega fv dx + \int_\Omega \mu \cdot (\nabla v - q) dx ,$$

and as in (1.25), we have the augmented Lagrangian

$$(1.33)\qquad \mathcal{L}_r(v,q,\mu) = \mathcal{L}(v,q,\mu) + \frac{r}{2} |\nabla v - q|^2 .$$

The remarks made in the previous example are wholly applicable again here. Any saddle-point $\{u,p,\lambda\}$ of $\mathcal{L}_r$ on $H_0^1(\Omega) \times (L^2(\Omega))^2 \times (L^2(\Omega))^2$ is characterised by a system analogous to (1.26)-(1.28), only the

inequality (1.27) being altered into

$$(1.34)\qquad \begin{cases} (1+r)\int_\Omega p\cdot(q-p)dx - \int_\Omega (\lambda+r\nabla u)\cdot(q-p)dx \geq 0 \quad \forall\, q \in \hat{K}, \\ p \in \hat{K}. \end{cases}$$

We can recognise in (1.34) the *characterisation of the projection* of $\frac{\lambda+r\nabla u}{1+r}$ on $\hat{K}$ into $(L^2(\Omega))^2$, i.e.

$$(1.35)\qquad p = P_{\hat{K}}\left(\frac{\lambda+r\nabla u}{1+r}\right) = \frac{\lambda+r\nabla u}{\mathrm{Sup}(1+r, |\lambda+r\nabla u|)} .$$

This projection into $(L^2(\Omega))^2$ can thus be carried out point by point. As in the preceding example, the decomposition has allowed a problem which included a constraint which was difficult to manipulate, to be reduced to a family of local problems, coordinated via a Lagrange multiplier. ∎

2. INVESTIGATION OF PROBLEM (P) AND OF THE SADDLE-POINTS OF $\mathcal{L}$ AND $\mathcal{L}_r$.

This section is devoted to review material of a theoretical nature; thus the reader who is primarily interested in the algorithmic aspects of this volume can at a first reading go forward to Section 3 of this chapter.

2.1 Existence and uniqueness properties for problem (P).

From standard results of Convex Analysis in infinite dimensions (see for example LIONS [1], EKELAND-TEMAM [1]) the problem (P) will admit a solution if the functional $J(v) = F(Bv) + G(v)$ satisfies

$$(2.1)\qquad \lim J(v) = +\infty \quad \textit{when} \quad \|v\|_V \to +\infty ,$$

a sufficient condition for *uniqueness* being the *strict convexity* of J. We assume of course that $J \not\equiv +\infty$.

Remark 2.1: In the following discussion we shall denote by $\|\cdot\|$ or $\|\cdot\|_V$ the norm on V, the norm on H being everywhere denoted by $|\cdot|$.

Remark 2.2: Suppose that we have

$$(2.2)\qquad \begin{cases} \lim F(q) \to +\infty \\ \textit{if } |q| \to +\infty. \end{cases}$$

If the operator B is *injective* with $\operatorname{Im} B$ *closed* in H, then $|Bv|$ defines on V a norm equivalent to $\|v\|$ such that we have

$$(2.3)\qquad \begin{cases} \lim F(Bv) \to +\infty \\ \textit{if } \|v\| \to +\infty \end{cases}$$

■

2.2 Properties of the saddle-points of $\mathcal{L}$ and of $\mathcal{L}_r$

The use of the algorithms of Sections 3 and 4 is essentially justified by the following:

THEOREM 2.1: *Suppose* $\{u,p,\lambda\}$ *is a saddle-point of* $\mathcal{L}$ *on* $V \times H \times H$; *then* $\{u,p,\lambda\}$ *is a saddle-point of* $\mathcal{L}_r$, $\forall r > 0$, *and vice versa. Furthermore* u *is a solution of* (P), *and we have* $p = Bu$.

Proof: Let $\{u,p,\lambda\}$ be a saddle-point of $\mathcal{L}$ on $V \times H \times H$. We thus have

$$(2.4)\qquad \mathcal{L}(u,p,\mu) \le \mathcal{L}(u,p,\lambda) \le \mathcal{L}(v,q,\lambda), \quad \forall \{v,q\} \in V\times H, \quad \forall \mu \in H.$$

From the first inequality in (2.4) we deduce

$$(\mu, Bu-p) \le (\lambda, Bu-p) \quad \forall \mu \in H,$$

hence

$$(2.5)\qquad Bu = p.$$

From the second inequality in (2.4) we then deduce

$$\begin{cases} F(Bu) + G(u) = \mathcal{L}(u,p,\lambda) \le F(q) + G(v) + (\lambda, Bv-q), \\ \forall \{v,q\} \in V\times H, \end{cases}$$

hence a fortiori

$$F(Bu) + G(u) \le F(Bv) + G(v) \quad \forall v \in V,$$

which proves that u is a solution of (P).

In view of (2.5) we immediately have

$$\begin{cases} \mathcal{L}(u,p,\mu) = \mathcal{L}_r(u,p,\mu) = \mathcal{L}(u,p,\lambda) = \mathcal{L}_r(u,p,\lambda) \le \mathcal{L}(v,q,\lambda) \le \mathcal{L}_r(v,q,\lambda), \\ \forall \{v,q\} \in V\times H\ ,\ \forall \mu \in H. \end{cases}$$

Suppose, conversely, that $\{u,p,\lambda\}$ is a saddle-point of $\mathcal{L}_r$; hence

$$\mathcal{L}_r(u,p,\mu) \le \mathcal{L}_r(u,p,\lambda) \le \mathcal{L}_r(v,q,\lambda)\ \forall \{v,q\} \in V\times H,\ \forall \mu \in H. \tag{2.6}$$

Proceeding as above, we deduce from the first inequality in (2.6) that $p = Bu$.
Furthermore we have

$$\mathcal{L}_r(u,p,\lambda) \le \mathcal{L}_r(v,q,\lambda)\quad \forall \{v,q\} \in V\times H. \tag{2.7}$$

Taking into account the convexity of $\mathcal{L}_r$ with respect to $\{v,q\}$, the pair $\{u,p\}$ is, for given λ, characterized by

$$\begin{cases} G(v) - G(u) + (\lambda, B(v-u)) + r(Bu-p, B(v-u)) \ge 0\ \forall v \in V, \\ F(q) - F(p) - (\lambda, q-p) + r(p-Bu, q-p) \ge 0\quad \forall q \in H, \end{cases}$$

thus, since $p = Bu$, we further have

$$G(v)-G(u) + (\lambda, B(v-u)) \ge 0\quad \forall v \in V, \tag{2.8}$$

$$F(q)-F(p) - (\lambda, q-p) \ge 0\quad \forall q \in H. \tag{2.9}$$

The relations (2.8), (2.9) are equivalent to

$$\begin{cases} \mathcal{L}(u,p,\lambda) \le \mathcal{L}(v,q,\lambda)\quad \forall \{v,q\} \in V\times H, \\ u \in V,\ p \in H. \end{cases}$$

Furthermore

$$\mathcal{L}(u,p,\mu) \le \mathcal{L}(u,p,\lambda)$$

follows trivially from $Bu = p$.
Theorem 2.1 is therefore completely proved. ∎

Remark 2.3: In *infinite dimensions*, the question of the possible

existence of a saddle-point, i.e. of a Lagrange multiplier λ, is problematical and is always dependent on the possibility of using in some form or other the Hahn-Banach Theorem. Sufficient conditions to assure the existence of the multiplier would be

(2.10) (P) *admits a solution,*

(2.11) $\exists u_o \in V$ *where* $F \circ B+G$ *is finite,*

(2.12) F *is continuous with respect to* Bu_o.

A proof of this result can be found in EKELAND-TEMAM [1] and an analogous result appears in GABAY-MERCIER [1]. In practice condition (2.12) is the most difficult to satisfy.

In finite dimensions the existence of a saddle-point is assured since we minimise under a *linear equality constraint;* in this case it is sufficient that problem (P) admit a solution.

In problems which are discretised from a problem in infinite dimensions we could thus have u_h which approaches u as $h \to 0$ whereas the multiplier λ_h does not converge in H. ■

2.3 Relations with perturbation theory in Convex Analysis

The approach followed in Section 1.2 consisted of "dualising" the original problem (P) by a procedure which at first sight may seem somewhat contrived. We now show that this approach is no more than a different way of expressing the duality in the sense of Fenchel-Rockafellar.

Consider, then, problem (P), namely:

$$\inf_{v \in V} \{F(Bv)+G(v)\} . \tag{2.13}$$

In convex analysis (see ROCKAFELLAR [4], EKELAND-TEMAM [1]) the standard procedure for associating a Lagrangian with (P) is to consider a *perturbed functional* on $V \times H$, defined by

$$\phi(v,z) = F(Bv-z)+G(v). \tag{2.14}$$

The Lagrangian associated with (2.14) is then defined classically by

$$L(v,\mu) = \inf_z \{(z,\mu) + \phi(v,z)\} = \inf_z \{(z,\mu) + F(Bv-z)\} + G(v). \tag{2.15}$$

Suppose $q = Bv - z$; we can then write (2.15) in the form

$$(2.16) \qquad L(v,\mu) = \inf_{q} \{(Bv-q,\mu) + F(q)\} + G(v).$$

If $L(.,.)$ admits a saddle-point $\{u,\lambda\}$ on $V \times H$, then u is a solution of (P) and we have

$$(2.17) \qquad \inf_{v\in V} \sup_{\mu\in H} L(v,\mu) = \sup_{\mu\in H} \inf_{v\in V} L(v,\mu).$$

We then describe as the *dual problem* of (P) the *maximisation* problem on H designated (P^*) and defined by

$$(2.18) \qquad \sup_{\mu} \{\inf_{v} L(v,\mu)\} .$$

If we express $L(.,.)$ explicitly we can write

$$(2.19) \qquad \sup_{\mu} \inf_{v} \inf_{q} \{F(q) + G(v) + (\mu,Bv-q)\} = \sup_{\mu} \inf_{\{v,q\}} \mathcal{L}(v,q,\mu).$$

The approach followed in Section 1.2 thus consisted of utilising the Lagrangian $L(.,.)$ in an equivalent but "more explicit" form.

This remark applies equally well to the augmented Lagrangian. In fact the standard definition (ROCKAFELLAR [1], FORTIN [1]) is obtained by considering

$$(2.20) \qquad \phi_r(v,z) = \phi(v,z) + \frac{r}{2} |z|^2,$$

$$(2.21) \qquad L_r(v,\mu) = \inf_{z} \{(z,\mu) + \phi_r(v,z)\} .$$

By the same argument as above we can derive the Lagrangian $\mathcal{L}_r$ of Section 1.2.

In FORTIN [1], $L_r(v,\mu)$ is calculated by elimination of q in (2.21). For the solution of certain problems which are nonlinear with respect to v, this leads to algorithms very similar to those studied in Sections 3 and 4. The search for a saddle-point of $L_r(v,\mu)$ using an algorithm of the UZAWA type (see Chapter I) requires at each stage the solution of a problem which is *nonlinear* with respect to v (see FORTIN [1] where the case of Example 1 of Section 1.1 is dealt with). The algorithms described in the remainder of this chapter can be considered as methods of solution by decomposition of this nonlinear problem (see Remark 3.1 below). ∎

Remark 2.4: Suppose $r \geq 0$; by proceeding as in Theorem 2.1, it can easily be shown that if $\{u,\lambda\}$ is a saddle-point of L_r on $V \times H$ then $\{u,Bu,\lambda\}$ is a saddle-point of $\mathcal{L}_r$ on $V \times H \times H$, and vice versa. ∎

Remark 2.5: An important property concerning the use of the augmented Lagrangian is that, for any $r > 0$, the Lagrangians L_r and $\mathcal{L}_r$ together with the associated dual problems are always *differentiable* with respect to μ, which is not the case in general for $r = 0$. It is in fact proved in FORTIN [1] that for $r > 0$ the dual problem is a regularisation by an inf-convolution of the dual problem for $r = 0$. This property is especially useful for the construction of algorithms using the gradient of the functional of the dual problem. ∎

3. DESCRIPTION OF THE ALGORITHMS

In this section we shall describe two iterative methods of solution of (P) which are in fact methods for calculating the saddle-points of $\mathcal{L}_r$, this approach being justified by Theorem 2.1.

3.1 First algorithm (ALG1)

In view of Theorem 2.1 it is natural, for calculating the saddle-points of $\mathcal{L}_r$ on $V \times H \times H$, to utilise the algorithm *of the Uzawa type* given below (see Chapter I, Section 2.1):

(3.1) $\lambda^0 \in H$ *specified arbitrarily;*
with λ^n *known, determine* $\{u^n,p^n\}$, *then* λ^{n+1} *by*

$$\text{(3.2)} \quad \begin{cases} \mathcal{L}_r(u^n,p^n,\lambda^n) \leq \mathcal{L}_r(v,q,\lambda^n) \ \forall \{v,q\} \in V \times H, \\ \{u^n,p^n\} \in V \times H, \end{cases}$$

$$\text{(3.3)} \quad \lambda^{n+1} = \lambda^n + \rho_n(Bu^n - p^n).$$

In the following text algorithm (3.1)-(3.3) will often be referred to as ALG1. The convergence of ALG1 will be studied in Section 4. It follows from (3.2) that the pair $\{u^n,p^n\}$ is characterized by the coupled system

$$(3.4)\qquad \begin{cases} G(v)-G(u^n)+(\lambda^n,B(v-u^n))+r(Bu^n-p^n,B(v-u^n)) \geq 0 \quad \forall\, v \in V, \\ u^n \in V, \end{cases}$$

$$(3.5)\qquad \begin{cases} F(q)-F(p^n)-(\lambda^n,q-p^n)+r(p^n-Bu^n,q-p^n) \geq 0 \quad \forall q \in H, \\ p^n \in H. \end{cases}$$

Remark 3.1: The above algorithm can also be written as a saddle-point calculation algorithm for L_r (see (2.15)). In fact we have seen in Section 2.3 that we actually have

$$(3.6)\qquad L_r(v,\mu) = \inf_q \mathcal{L}_r(v,q,\mu).$$

We thereby deduce that if $\{u^n,p^n\}$ is a solution of (3.2) we also have

$$(3.7)\qquad \begin{cases} L_r(u^n,\lambda^n) \leq L_r(v,\lambda^n) \quad \forall\, v \in V, \\ u^n \in V. \end{cases}$$

It is shown in FORTIN [1] that problem (3.7) is in general nonlinear with respect to v and that its *direct* solution is difficult. Through the introduction of q, then of $\mathcal{L}_r$, this problem (3.7) has been decomposed into the system which is the equivalent of the two inequalities (3.4), (3.5). Expressing the problem in the form of a system will lead to efficient procedures for solving (3.7) (and (3.2)). ∎

Remark 3.2: The reader may verify that algorithm ALG1 can be interpreted as a *subgradient* algorithm for the dual functional

$$h_r(\mu) = \inf_{\{v,q\}\in V\times H} \{\mathcal{L}_r(v,q,\mu)\}.$$

It can actually be shown that, whatever the values of $\lambda,\mu, \in H$, we have

$$h_r(\mu) \leq h_r(\lambda)+(Bu_\lambda-p_\lambda,\mu-\lambda),$$

where $\{u_\lambda,p_\lambda\}$ is the solution of the problem

$$\begin{cases} \mathcal{L}_r(u_\lambda, p_\lambda, \lambda) \le \mathcal{L}_r(v,q,\lambda) & \forall \{v,q\} \in V\times H, \\ \{u_\lambda, p_\lambda\} \in V\times H. \end{cases}$$

Additionally, we have noted in Remark 2.5 that, for $r > 0$, $h_r(\lambda)$ is always differentiable. Thus in this case we in fact have a standard gradient algorithm when r is strictly positive. ∎

3.2 Second algorithm (ALG2)

In the implementation of ALG1 the essential difficulty is clearly the solution *at each iteration* of the system (3.4), (3.5). A natural solution procedure consists of using the *block relaxation* method given below *(where* n *is fixed):*

$$p^{n,0} = p^{n-1}, \tag{3.8}$$

then for $k \ge 1$

$$\begin{cases} G(v)-G(u^{n,k})+(\lambda^n, B(v-u^{n,k}))+r(Bu^{n,k}-p^{n,k-1}, B(v-u^{n,k})) \ge 0 \quad \forall v \in V, \\ u^{n,k} \in V, \end{cases} \tag{3.9}$$

$$\begin{cases} F(q)-F(p^{n,k})-(\lambda^n, q-p^{n,k})+r(p^{n,k}-Bu^{n,k}, q-p^{n,k}) \ge 0 \ \forall q \in H, \\ p^{n,k} \in H. \end{cases} \tag{3.10}$$

The algorithm (3.8)-(3.10) is convergent under quite general assumptions on F and G (see for example, CEA-GLOWINSKI [1]). Taking into account the results of Chapters I and II concerning *incomplete* minimisation in the UZAWA algorithm (see Chapter I, Section 2.3, Remark 2.5 and Chapter II, Sections 3.2 and 5.2) we obtain natural variants of ALG1 by restricting ourselves when solving (3.4), (3.5) by (3.8)-(3.10) to a *fixed* number of block relaxation passes. In the limiting case of a *single* pass we obtain the algorithm

$$\{p^0, \lambda^1\} \in H \times H \ \textit{arbitrarily specified;} \tag{3.11}$$

with $\{p^{n-1}, \lambda^n\}$ *known, determine successively* u^n, p^n, λ^{n+1} *by*

$$\begin{cases} G(v)-G(u^n)+(\lambda^n, B(v-u^n))+r(Bu^n-p^{n-1}, B(v-u^n)) \ge 0 \quad \forall v \in V, \\ u^n \in V, \end{cases} \tag{3.12}$$

$$(3.13)\qquad \begin{cases} F(q)-F(p^n)-(\lambda^n,q-p^n)+r(p^n-Bu^n,q-p^n) \geq 0 \quad \forall\, q \in H, \\ p^n \in H, \end{cases}$$

$$(3.14)\qquad \lambda^{n+1} = \lambda^n + \rho_n(Bu^n-p^n).$$

In the following text algorithm (3.11)-(3.14) will often be referred to as ALG2. The convergence of ALG2 will be studied in Section 5.

Remark 3.3: The algorithm ALG2 seems to have been first introduced by GLOWINSKI-MARROCCO [1] in connection with the numerical solution of problem (1.23). The convergence of ALG2 for $\rho_n = \rho = r$ was demonstrated in MERCIER [2] in relation to the *nonlinear elasticity* problem of Chapter VI, and then extended by GABAY-MERCIER [1] to the case where $0 < \rho_n = \rho < 2r$, under quite general assumptions on F (G being linear).

3.3 Application to the examples of Section 1

In order to clarify the concepts introduced in Sections 3.1 and 3.2, we shall now set out ALG1 and ALG2 explicitly for the two model problems of Section 1.1.

Case of Example 1 *of Section* 1.1 *(Flow of a Bingham fluid)*

With the Lagrangian $\mathcal{L}_r$ defined by (1.24), (1.25), and taking into account (1.26)-(1.31) and (3.4), (3.5), ALG1 takes the following form:

(3.15) λ^0 *arbitrarily chosen in* $(L^2(\Omega))^2$;
then, λ^n *being known, determine* $\{u^n,p^n\}$ *by solving the coupled system*

$$(3.16)\qquad \begin{cases} -r\Delta u^n + r\nabla\cdot p^n - \nabla\cdot\lambda^n = f \ \text{ on } \Omega, \\ u^n|_\Gamma = 0, \end{cases}$$

$$(3.17)\qquad \begin{cases} p^n(x) = 0 \ \text{ if } g \geq |\lambda^n(x)+r\nabla u^n(x)|, \\ p^n(x) = \dfrac{\lambda^n(x)+r\nabla u^n(x)}{\nu+r}\left(1 - \dfrac{g}{|\lambda^n(x)+r\nabla u^n(x)|}\right) \ \text{ otherwise,} \end{cases}$$

and λ^{n+1} *by*

$$\lambda^{n+1} = \lambda^n + \rho_n(\nabla u^n - p^n). \tag{3.18}$$

Remark 3.4: In practice (3.15)-(3.18) will be applied to an approximation (by finite differences or finite elements) of the problem (1.4). Looking at (3.16), (3.17), it appears that the implementation of the block relaxation method (3.8)-(3.10) will present no practical difficulties provided that an efficient program is available for solving the Dirichlet problem for $-\Delta$. ∎

To pass from ALG1 to ALG2 it suffices to replace (3.15) by

$$\{p^0, \lambda^1\} \quad \textit{arbitrarily chosen in} \quad (L^2(\Omega))^2 \times (L^2(\Omega))^2; \tag{3.19}$$

and (3.16) by

$$\begin{cases} -r\Delta u^n + r\nabla \cdot p^{n-1} - \nabla \cdot \lambda^n = f \ \textit{on} \ \Omega, \\ u^n|_\Gamma = 0 ; \end{cases} \tag{3.20}$$

with this algorithm the determination of u^n, p^n is sequential.

Case of Example 2 *of Section* 1.1 (*Elastoplastic torsion*)

With the Lagrangian $\mathcal{L}_r$ defined by (1.32), (1.33) and taking account of (1.34), (1.35) and of (3.4), (3.5), ALG1 takes the following form:

$$\lambda^0 \quad \textit{arbitrarily chosen in} \quad (L^2(\Omega))^2; \tag{3.21}$$

then, λ^n *being known, determine* $\{u^n, p^n\}$ *by*

$$\begin{cases} -r\Delta u^n + r\nabla \cdot p^n - \nabla \cdot \lambda^n = f \ \textit{on} \ \Omega, \\ u^n|_\Gamma = 0, \end{cases} \tag{3.22}$$

$$p^n = \frac{\lambda^n + r\nabla u^n}{\sup(1+r, |\lambda^n + r\nabla u^n|)} , \tag{3.23}$$

$$\lambda^{n+1} = \lambda^n + \rho_n(\nabla u^n - p^n). \tag{3.24}$$

This algorithm is very closely related to (3.15)-(3.18) and Remark 3.4 is equally valid for (3.21)-(3.24).

To pass from ALG1 to ALG2, it suffices to replace (3.21), (3.22) by (3.19), (3.20).

Remark 3.5: All other things being equal, the "cost" of an iteration is higher for ALG1 than for ALG2, and in a large number of problems it is preferable to use this latter algorithm. Nonetheless we should point out that in certain *very stiff* problems, for example (1.4) with g "large" or alternatively (1.23) with p "close" to 1 or "large", ALG1 is faster than ALG2 both in the number of iterations and in the computation time. All these points will be illustrated through various examples in the following chapters. ∎

4. CONVERGENCE OF ALG1

In section 4.1, we shall study the convergence of ALG1 when V and H are Hilbert spaces of *arbitrary dimension;* then in Section 4.2 we shall examine the extent to which the assumptions of Section 4.1 can be weakened when V and H are of *finite dimension.*

4.1 General case

To study the convergence of ALG1, we shall be making several supplementary assumptions. We shall first assume that we have

(4.1) B *is injective and* ImB *is closed in* H.

In addition, we shall make the following assumption on F concerning growth at infinity

$$\lim_{|q| \to +\infty} \frac{F(q)}{|q|} = +\infty . \tag{4.2}$$

Taking into account the properties of G, (4.1) and (4.2) imply that (P) admits a unique solution u [2]. In the following text we shall write p = Bu.

We next assume that $F = F_0 + F_1$, where F_1 is convex, proper and l.s.c. on H, and where F_0 is convex, differentiable on H, and

---

[2] Provided we assume that $\mathrm{Dom}(F \circ B) \cap \mathrm{Dom}(G) \neq \emptyset$ (we recall that if $j: X \to \overline{\mathbb{R}}$ then $\mathrm{Dom}(j) = \{x \in X,\ j(x) \in \mathbb{R}\}$).

uniformly convex over the bounded subsets of H in the following sense:

For any $M > 0$, *there exists a continuous function* $\delta_M: [0,2M] \to \mathbb{R}$, *strictly increasing, with* $\delta_M(0) = 0$, *such that for all* $p,q \in H$, $|p| \le M$, $|q| \le M$, *we have:*

$$(F_o'(q)-F_o'(p),q-p) \ge \delta_M(|q-p|), \tag{4.3}$$

where, in (4.3), F_o' *denotes the gradient of* F_o.

THEOREM 4.1: *Suppose that* $\mathcal{L}_r$ *admits a saddle-point* $\{u,p,\lambda\}$ *on* $V \times H \times H$. *Under the above assumptions on* B, F *and* G, *and if* ρ_n *satisfies*

$$0 < \alpha_o \le \rho_n \le \alpha_1 < 2r , \tag{4.4}$$

we have for ALG1 *the following convergence results:*

$$u^n \to u \quad \textit{strongly in} \quad V, \tag{4.5}$$

$$p^n \to p \quad \textit{strongly in} \quad H, \tag{4.6}$$

$$\lambda^{n+1}-\lambda^n \to 0 \quad \textit{strongly in} \quad H, \tag{4.7}$$

$$\lambda^n \quad \textit{is bounded in} \quad H. \tag{4.8}$$

Furthermore if λ^* *is a (weak) cluster point of* λ^n *in* H, $\{u,p,\lambda^*\}$ *is a saddle-point of* $\mathcal{L}_r(v,q,\mu)$ *on* $V \times H \times H$.

Proof: In the following text we shall write

$$\begin{cases} \bar{u}^n = u^n-u , \\ \bar{p}^n = p^n-p , \\ \bar{\lambda}^n = \lambda^n-\lambda . \end{cases} \tag{4.9}$$

We thus have to show that $\bar{u}^n \to 0$, $\bar{p}^n \to 0$, and that $\bar{\lambda}^n$ remains bounded; $\{u,p,\lambda\}$ being a saddle-point of $\mathcal{L}_r$, we have

$$G(v)-G(u)+(\lambda,B(v-u))+r(Bu-p,B(v-u)) \ge 0, \ \forall\, v \in V, \tag{4.10}$$

$$(F_o'(p),q-p)+F_1(q)-F_1(p)-(\lambda,q-p)+r(p-Bu,q-p) \ge 0, \ \forall\, q \in H, \tag{4.11}$$

(4.12) $\lambda = \lambda + \rho_n(Bu-p), \forall n$.

Furthermore (3.2)-(3.5) imply

(4.13) $G(v)-G(u^n)+(\lambda^n, B(v-u^n))+r(Bu^n-p^n, B(v-u^n)) \geq 0$, $\forall v \in V$,

(4.14) $(F_o'(p^n), q-p^n)+F_1(q)-F_1(p^n)-(\lambda^n, q-p^n)+r(p^n-Bu^n, q-p^n) \geq 0$, $\forall q \in H$,

(4.15) $\lambda^{n+1} = \lambda^n+\rho_n(Bu^n-p^n)$.

We thus set $v = u^n$ in (4.10), $q = p^n$ in (4.11), then $v = u$ in (4.13) and $q = p$ in (4.14), so that, by addition, we have

(4.16) $r|B\bar{u}^n|^2-r(\bar{p}^n, B\bar{u}^n)+(\bar{\lambda}^n, B\bar{u}^n) \leq 0$,

(4.17) $(F_o'(p^n)-F_o'(p), p^n-p)+r|\bar{p}^n|^2-r(\bar{p}^n, B\bar{u}^n)-(\bar{\lambda}^n, \bar{p}^n) \leq 0$.

Summing (4.16) and (4.17) and regrouping the terms, we obtain

(4.18) $(F_o'(p^n)-F_o'(p), p^n-p)+r|B\bar{u}^n-\bar{p}^n|^2+(\bar{\lambda}^n, B\bar{u}^n-\bar{p}^n) \leq 0$.

Also, subtracting (4.12) from (4.15) and taking the scalar square in H, we obtain

(4.19) $|\bar{\lambda}^n|^2-|\bar{\lambda}^{n+1}|^2 = -2\rho_n(B\bar{u}^n-\bar{p}^n, \bar{\lambda}^n)-\rho_n^2|B\bar{u}^n-\bar{p}^n|^2$.

It then follows from (4.18) and (4.19) that

(4.20) $|\bar{\lambda}^n|^2-|\bar{\lambda}^{n+1}|^2 \geq 2\rho_n(F_o'(p^n)-F_o'(p), p^n-p)+\rho_n(2r-\rho_n)|B\bar{u}^n-\bar{p}^n|^2$.

In view of the assumption (4.4) on ρ_n we thus have

(4.21) $\lim_{n\to+\infty} |Bu^n-p^n| = 0$ (*since* $Bu=p$),

(4.22) $\lim_{n\to+\infty} (F_o'(p^n)-F_o'(p), p^n-p) = 0$,

and we have λ_n bounded in H.

We now show that p^n is bounded in H; in fact, since the functional F is proper, there exists $\hat{p} \in H$, such that $-\infty < F(\hat{p}) < +\infty$. Then putting $q = \hat{p}$ in (3.5), we obtain

(4.23) $F(\hat{p})-(\lambda^n, \hat{p})+r(p^n-Bu^n, \hat{p}) \geq F(p^n)-(\lambda^n, p^n)+r(p^n-Bu^n, p^n)$.

Since λ^n is bounded in H and knowing that $p^n - Bu^n$ tends to zero, we thereby deduce that there exist positive constants β_o, β_1,

independent of n, such that

$$\beta_o \geq F(p^n) - \beta_1 |p^n| \quad . \tag{4.24}$$

Taking into account assumption (4.2), it follows from (4.24) that p^n is bounded, i.e. that there exists $M > 0$, such that

$$|p^n| \leq M, \ \forall n. \tag{4.25}$$

Furthermore, for M sufficiently large, we also have

$$|p| \leq M. \tag{4.26}$$

It then follows from (4.25), (4.26) and from the uniform convexity of F_o over the bounded subsets of H, that we have

$$(F_o'(p^n) - F_o'(p), p^n - p) \geq \delta_M(|p^n - p|) \ \forall n. \tag{4.27}$$

We thus have

$$\lim_{n \to +\infty} \delta_M(|p^n - p|) = 0. \tag{4.28}$$

It then follows from the properties of δ_M that

$$\lim_{n \to +\infty} |p^n - p| = 0, \tag{4.29}$$

and from (4.21) that

$$\lim_{n \to +\infty} Bu^n = p \ (=Bu), \ \textit{strongly in} \ H. \tag{4.30}$$

Since the operator B is injective with ImB closed in H, this implies (see Remark 2.2) that

$$\lim_{n \to +\infty} u^n = u \ , \quad \textit{strongly in} \ V. \tag{4.31}$$

The properties of the sequence λ^n, given in the statement of the Theorem, are then immediate consequences of the convergence properties of u^n and of p^n and are obtained by passing to the limit in (3.3)-(3.5). ∎

Remark 4.1: We have proved the convergence of the algorithm by making use of the assumption of coercivity (4.3) on $F_o'(.)$. It is easy to show that an analogous result is obtained by assuming that $G(\cdot)$ is differentiable, its derivative satisfying a condition

similar to (4.3). In fact the convergence proof simplifies and it is no longer necessary to assume B to be injective. We shall encounter in Chapter IV a case where this variant will be useful.

4.2 The finite-dimensional case

If V and H are of *finite dimension*, convergence of ALG1 can be obtained under weaker assumptions than those given in Section 4.1.

Firstly, for $\mathcal{L}_r$ to admit a saddle-point *it will be sufficient for* (P) *to admit a solution* (see Remark 2.3). Furthermore, Im B is still closed. As regards F, it follows from CEA-GLOWINSKI [1, Section 2.2] that the uniform convexity property of F_o given in Section 4.1 is satisfied if

(4.32) F_o *is strictly convex and of class* C^1.

In fact if F_o satisfies (4.32) and *if we assume that* (P) *admits a solution* we can dispense with assumption (4.2). This follows from the fact that the *strict convexity* of F_o implies the *strict monotonicity*[3] of F_o', and from the following:

LEMMA 4.1: *Suppose* H *is of finite dimension; let* $A: H \to H$ *be a continuous and strictly monotone operator,* p *an element of* H *and* $(p^n)_n$ *a sequence of elements of* H *such that*

$$\lim_{n \to +\infty} (A(p^n)-A(p),\ p^n-p) = 0. \tag{4.33}$$

We then have

$$\lim_{n \to +\infty} p^n = p. \tag{4.34}$$

Proof: Suppose that (4.34) is not true; in this case there exists $\delta > 0$ and a sub-sequence extracted from $(p^n)_n$, $(p^m)_m$ say, such that

$$|p^m-p| \geq \delta \quad \forall m\ . \tag{4.35}$$

Let $S(p;\frac{\delta}{2})$ be the sphere with centre p and radius $\frac{\delta}{2}$. We define $z^m \in S(p;\frac{\delta}{2})$ by

---

[3] We recall that $A: H \to H$ is said to be *strictly monotone* if $(A(q)-A(p),\ q-p) > 0 \quad \forall p,q \in H,\ p \neq q$.

$$z^m = p + \frac{\delta}{2}\frac{p^m-p}{|p^m-p|} \quad ; \tag{4.36}$$

z^m thus belongs to the "open interval" $]p,p^m[$ of the space H (see Figure 4.1).

We shall denote by t^m the quantity $\dfrac{\delta}{2|p^m-p|}$; we thus have

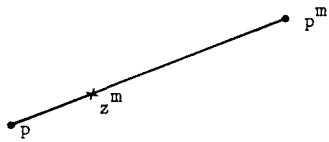


Figure 4.1

$$z^m = p + t^m(p^m-p), \tag{4.37}$$

and, from (4.35),

$$0 < t^m \le \frac{1}{2} . \tag{4.38}$$

Since the operator A is strictly monotone we have

$$(A(p^m)-A(p),p^m-p) > (A(p+t(p^m-p))-A(p),p^m-p) > 0 \quad \forall\, t \in]0,1[, \tag{4.39}$$

hence in particular (we set $t = t^m$ in (4.39))

$$(A(p^m)-A(p),p^m-p) > (A(z^m)-(A(p),p^m-p) > 0. \tag{4.40}$$

In view of (4.37), (4.38) it follows from (4.40) that

$$\left\{ \begin{array}{l} (A(p^m)-A(p),p^m-p) > \dfrac{1}{t^m}\,(A(z^m)-(A(p),z^m-p) \ge 2(A(z^m)-A(p),z^m-p) > \\ \qquad\qquad > (A(z^m)-A(p),\ z^m-p) > 0. \end{array} \right. \tag{4.41}$$

Since the sphere $S(p;\frac{\delta}{2})$ is *compact* we can extract from $(z^m)_m$ a sub-sequence, itself also denoted by $(z^m)_m$, such that

$$\lim_{m \to +\infty} z^m = z \ , \ z \in S(p;\tfrac{\delta}{2}) \ . \tag{4.42}$$

Since the operator A is *continuous* we deduce from (4.33), (4.41), (4.42) that

$$(A(z)-A(p),z-p) = 0. \tag{4.43}$$

The operator A being *strictly monotone*, (4.33) implies that $z = p$, which is absurd since $|p-z| = \frac{\delta}{2} > 0$. We cannot therefore have (4.35); hence p^n converges to p. ∎

In view of these various remarks, one can easily prove the following variant of Theorem 4.1:

THEOREM 4.2: *Suppose that* V *and* H *are of finite dimension and that* (P) *admits a solution* u. *We make the following assumptions on* B, G, F

- B *is injective,*
- G *is convex, proper and l.s.c. on* V,
- *We have* $F = F_0 + F_1$ *with* F_1 *convex, proper and l.s.c. on* H, *and* F_0 *strictly convex and of class* C^1 *on* H.

The solution of (P) *is then unique, and under the condition*

$$0 < \alpha_0 \leq \rho_n \leq \alpha_1 < 2r$$

we have for ALG1 *the following convergence results*

$$\lim_{n \to +\infty} u^n = u,$$

$$\lim_{n \to +\infty} p^n = Bu,$$

λ^n *is bounded in* H.

Moreover, if λ *is a cluster point of* λ^n *in* H, *then* $\{u, Bu, \lambda\}$ *is a saddle-point of* $\mathcal{L}_r$ *on* $V \times H \times H$.

Remark 4.2: We shall meet in Chapter V, in connection with the *minimum surfaces* problem, a situation in which the assumption (4.2) on F is not satisfied, either for the continuous problem or for the approximate problem, and in which the convergence of ALG1 will follow (for the approximate problem) from Theorem 4.2. ∎

4.3 On the choice of r and of $\{\rho_n\}_n$.

In the general case the determination of the optimal parameters is a complicated matter, as a consequence of the nonlinearity of problem (3.4), (3.5); furthermore, the convergence properties of the

sequences $\{u^n\}_n$ and $\{p^n\}_n$, $\{\lambda^n\}_n$ may be different, as we shall see below.

We consider the case where

$$F(q) = \frac{1}{2}|q|^2 , \tag{4.44}$$

$$G(v) = -((f,v)), \tag{4.45}$$

the operator B still satisfying (4.1).
The problem (P) is then equivalent to

$$B^tBu = f, \tag{4.46}$$

and $\mathcal{L}_r$ admits as a saddle-point $\{u,Bu,Bu\}$; we thus have $\lambda = p = Bu$. It is apparent that the use of ALG1 is of no practical interest for solving (4.46) since, at each iteration, it will be necessary to solve linear problems relating to B^tB. Nonetheless, this trivial case offers a certain theoretical interest for the study of the convergence of the algorithm. The latter is written, (limiting ourselves to the case where $\rho_n = \rho$),

(4.47) *λ^0 arbitrarily specified in* H;
then, λ^n *being known, calculate* u^n,p^n,λ^{n+1} *by*

$$B^t\lambda^n + rB^t(Bu^n-p^n) = f, \tag{4.48}$$

$$p^n = \lambda^n + r(Bu^n-p^n), \tag{4.49}$$

$$\lambda^{n+1} = \lambda^n + \rho(Bu^n-p^n). \tag{4.50}$$

Again using the notation $\bar{u}^n = u^n-u$, $\bar{p}^n-p$, $\bar{\lambda}^n = \lambda^n-\lambda$, we have

$$B^t\bar{\lambda}^n+rB^t(B\bar{u}^n-\bar{p}^n) = 0 \quad \forall n \geq 0, \tag{4.51}$$

$$\bar{p}^n = \bar{\lambda}^n+r(B\bar{u}^n-\bar{p}^n) \quad \forall n \geq 0, \tag{4.52}$$

$$\bar{\lambda}^{n+1} = \bar{\lambda}^n+\rho(B\bar{u}^n-\bar{p}^n) \ \forall n \geq 0. \tag{4.53}$$

Multiplying (4.52) by B^t, we deduce from (4.51) that

$$B^t\bar{p}^n = 0 \quad \forall n \geq 0. \tag{4.54}$$

We then deduce from (4.51) and from (4.54) that

$$B^t\bar{\lambda}^n = -rB^tB\bar{u}^n \quad \forall n \geq 0. \tag{4.55}$$

Multiplying (4.53) by B^t and taking account of (4.54) and (4.55), we obtain

$$(4.56) \qquad B^t\bar{Bu}^{n+1} = (\frac{r-\rho}{r})\, B^t\bar{Bu}^n = (1-\frac{\rho}{r})\, B^t\bar{Bu}^n \quad \forall\, n \geq 0.$$

We thereby deduce that if $r = \rho$, we have[4]

$$(4.57) \qquad u^n = u, \forall\, n \geq 1 \ ;$$

we thus have convergence of u^n *in two iterations*, whatever the value of $\lambda^o \in H$. For $\rho \neq r$ the rate of convergence of the sequence u^n depends on the ratio $\frac{\rho}{r}$. ∎

We now consider the convergence of the sequence λ^n. In view of (4.55) we have $\bar{Bu}^n = -\frac{1}{r} B(B^tB)^{-1}B^t\bar{\lambda}^n$,

or alternatively

$$(4.58) \qquad \bar{Bu}^n = -\frac{1}{r} P\bar{\lambda}^n$$

where $P = (B^tB)^{-1}B^t$ is the projection operator from H onto ImB. By substituting (4.58) into (4.53), and by using (4.52), we obtain

$$(4.59) \qquad \bar{\lambda}^{n+1} = (1 - \frac{\rho}{1+r})\, \bar{\lambda}^n + \rho(\frac{1}{1+r} - \frac{1}{r})P\bar{\lambda}^n, \forall\, n \geq 0.$$

By projection of (4.59) onto ImB, then onto the orthogonal subspace Ker B^t we thus deduce

$$(4.60) \qquad P\bar{\lambda}^{n+1} = (1 - \frac{\rho}{r})P\bar{\lambda}^n, \quad \forall\, n \geq 0 \ ,$$

$$(4.61) \qquad (I-P)\bar{\lambda}^{n+1} = (1- \frac{\rho}{1+r})(I-P)\bar{\lambda}^n \ , \forall\, n \geq 0.$$

In the case where $\rho = r$ we thus have convergence of $P\lambda^n$ *in two iterations*. The projection of $\lambda^n - \lambda$ onto Ker B^t "decreases" by the factor $(1- \frac{\rho}{1+r})$, i.e. in the case where $\rho = r$, a factor of $\frac{1}{1+r}$. If we choose λ^o in ImB we thus have, for $\rho = r$, convergence of λ^n *in two iterations*. If λ^o is chosen arbitrarily in H the convergence of λ^n to $\lambda = Bu$ will be faster the larger the value of r, when $\rho = r$. As regards the sequence p^n it follows from the preceding relations that

$$\bar{p}^n = \frac{1}{1+r}\,(I-P)\bar{\lambda}^{n-1} \ ;$$

---

[4] We recall that B *injective* with ImB *closed* in H implies that B^tB is an *isomorphism* of V onto V.

the sequence p^n thus behaves like the sequence $(I-P)\bar{\lambda}^n$. ∎

The preceding analysis indicates that in certain cases we may expect a faster convergence for u^n than for λ^n or p^n. Such a phenomenon has in fact been established experimentally (see GABAY-MERCIER [1]) in the case of the elastoplastic torsion problem (see Section 1.1, Example 2). It also appears, in the light of numerous numerical experiments, that the choice of $\rho_n = \rho = r$ is "quasi-optimal".

It is also easy to show that the convergence of ALG1 will be faster the larger the value of r. From a practical point of view, however, the situation is rather more complex: in fact, the conditioning of the system (3.4), (3.5) gets worse as r increases, so that the speed of convergence of the relaxation algorithm (3.8)-(3.10) decreases. Moreover the choice of the termination test for the internal iterations (3.8)-(3.10) and the effect of rounding errors also play a part. Experimentally the combined effect of these factors - namely, with an increase of r an acceleration of ALG1 but a slowing down of the internal relaxation algorithm - leads in many cases to an algorithm whose overall speed of convergence (in terms of computation time) depends relatively little on the choice of r; this fact will be illustrated by the various examples considered in the following chapters.

5. CONVERGENCE OF ALG2

In this section we shall show that under quite general assumptions on F and G we have convergence of ALG2 under the condition $0 < \rho_n = \rho < (\frac{1+\sqrt{5}}{2})\, r$; we do not know whether this result is optimal, since in certain particular cases (G linear, for example) the upper bound of the interval of convergence can be replaced by $2r$. In fact this question becomes somewhat academic in character (in our opinion) since in the various applications of ALG2 which have been undertaken the optimal choice for ρ seems to be $\rho = r$.

5.1 General case

We are now going to consider the convergence of ALG2 under the same assumptions as those used in Section 4.1 for ALG1.

We have

THEOREM 5.1: *We assume that* $\mathcal{L}_r$ *admits a saddle-point* $\{u,p,\lambda\}$ *on* $V \times H \times H$. *Under the assumptions on* B, F, G *used in Section* 4.1, *and if* ρ_n *satisfies*

$$0 < \rho_n = \rho < \left(\frac{1+\sqrt{5}}{2}\right) r, \tag{5.1}$$

we have for ALG2 *the following convergence results*

$$u^n \to u \quad \textit{strongly in } V, \tag{5.2}$$

$$p^n \to p \quad \textit{strongly in } H, \tag{5.3}$$

$$\lambda^{n+1}-\lambda^n \to 0 \quad \textit{strongly in } H, \tag{5.4}$$

$$\lambda^n \quad \textit{is bounded in } H. \tag{5.5}$$

Moreover if λ^* *is a (weak) cluster point of* λ^n *in* H, $\{u,p,\lambda\}$ *is a saddle-point of* $\mathcal{L}_r$ *on* $V \times H \times H$.

Proof: We again write $\bar{u}^n = u^n-u$, $\bar{p}^n = p^n-p$, $\bar{\lambda}^n = \lambda^n-\lambda$; $\{u,p,\lambda\}$ being a saddle-point of $\mathcal{L}_r$ on $V \times H \times H$, we have

$$G(v)-G(u)+(\lambda,B(v-u))+r(Bu-p,B(v-u)) \geq 0, \ \forall\, v \in V, \tag{5.6}$$

$$(F_0'(p),q-p)+F_1(q)-F_1(p)-(\lambda,q-p)+r(p-Bu,q-p) \geq 0, \ \forall\, q \in H, \tag{5.7}$$

$$\lambda = \lambda+\rho(Bu-p), \ \forall\, n. \tag{5.8}$$

In addition, (3.12)-(3.14) imply

$$G(v)-G(u^n)+(\lambda^n,B(v-u^n))+r(Bu^n-p^{n-1},B(v-u^n)) \geq 0, \ \forall\, v \in V, \tag{5.9}$$

$$(F_0'(p^n),q-p^n)+F_1(q)-F_1(p^n)-(\lambda^n,q-p^n)+r(p^n-Bu^n,q-p^n) \geq 0, \ \forall\, q \in H, \tag{5.10}$$

$$\lambda^{n+1} = \lambda^n+\rho(Bu^n-p^n). \tag{5.11}$$

We set $v = u^n$ in (5.6), $q = p^n$ in (5.7), then $v = u$ in (5.9) and $q = p$ in (5.10); hence by addition

$$r|B\bar{u}^n|^2-r(\bar{p}^{n-1},B\bar{u}^n)+(\bar{\lambda}^n,B\bar{u}^n) \leq 0, \tag{5.12}$$

$$(F_0'(p^n)-F_0'(p),p^n-p)+r|\bar{p}^n|^2-r(\bar{p}^n,B\bar{u}^n)-(\bar{\lambda}^n,\bar{p}^n) \leq 0. \tag{5.13}$$

Adding (5.12), (5.13) and regrouping the terms, we obtain

$$(F_0'(p^n)-F_0'(p),p^n-p)+r|B\bar{u}^n-\bar{p}^n|^2+(\bar{\lambda}^n,B\bar{u}^n-\bar{p}^n)+r(\bar{p}^n-\bar{p}^{n-1},B\bar{u}^n) \leq 0. \tag{5.14}$$

Also, subtracting (5.8) from (5.11) and taking the scalar square in H, we obtain

$$(5.15)\qquad |\bar{\lambda}^n|^2-|\bar{\lambda}^{n+1}|^2 = -2\rho(B\bar{u}^n-\bar{p}^n,\bar{\lambda}^n)-\rho^2|B\bar{u}^n-\bar{p}^n|^2.$$

It then follows from (5.14), (5.15) that

$$(5.16)\qquad |\bar{\lambda}^n|^2-|\bar{\lambda}^{n+1}|^2 \geq 2\rho(F_o'(p^n)-F_o'(p),p^n-p)+\rho\ (2r-\rho)\ |B\bar{u}^n-\bar{p}^n|^2+2\rho r(\bar{p}^n-\bar{p}^{n-1},B\bar{u}^n).$$

We now try to modify the final term on the right-hand side of (5.16).

Starting from

$$B\bar{u}^n = (B\bar{u}^n-B\bar{u}^{n-1})+(B\bar{u}^{n-1}-\bar{p}^{n-1})+\bar{p}^{n-1}$$

we thence deduce

$$(5.17)\qquad (B\bar{u}^n,\bar{p}^n-\bar{p}^{n-1}) = (B\bar{u}^n-B\bar{u}^{n-1},\bar{p}^n-\bar{p}^{n-1})+(B\bar{u}^{n-1}-\bar{p}^{n-1},\bar{p}^n-\bar{p}^{n-1})+(\bar{p}^{n-1},\bar{p}^n-\bar{p}^{n-1}).$$

From (5.17) and from

$$(\bar{p}^{n-1},\bar{p}^n-\bar{p}^{n-1}) = \frac{1}{2}\ (|\bar{p}^n|^2-|\bar{p}^{n-1}|^2-|\bar{p}^n-\bar{p}^{n-1}|^2)$$

we deduce that

$$(5.18)\qquad \begin{cases} 2\rho r(B\bar{u}^n,\bar{p}^n-\bar{p}^{n-1}) = 2\rho r(B\bar{u}^n-B\bar{u}^{n-1},\bar{p}^n-\bar{p}^{n-1}) + \\ +2\rho r(B\bar{u}^{n-1}-\bar{p}^{n-1},\bar{p}^n-\bar{p}^{n-1})+\rho r(|\bar{p}^n|^2-|\bar{p}^{n-1}|^2-|\bar{p}^n-\bar{p}^{n-1}|^2). \end{cases}$$

Considering (5.10) on iteration $n-1$ instead of n, we have

$$(5.19)\qquad (F_o'(p^{n-1}),q-p^{n-1})+F_1(q)-F_1(p^{n-1})-(\lambda^{n-1},q-p^{n-1})+r(p^{n-1}-Bu^{n-1},q-p^{n-1}) \geq 0.$$

Taking $q = p^{n-1}$ in (5.10) and $q = p^n$ in (5.19) we obtain by addition

$$(5.20)\qquad \begin{cases} (F_o'(p^n)-F_o'(p^{n-1}),p^n-p^{n-1})-(\bar{\lambda}^n-\bar{\lambda}^{n-1},\bar{p}^n-\bar{p}^{n-1})+r|\bar{p}^n-\bar{p}^{n-1}|^2 - \\ - r(B\bar{u}^n-B\bar{u}^{n-1},\bar{p}^n-\bar{p}^{n-1}) \leq 0. \end{cases}$$

It follows from the monotonicity of F_o', and from (5.20), that

$$(5.21)\qquad r|\bar{p}^n-\bar{p}^{n-1}|^2-(\bar{\lambda}^n-\bar{\lambda}^{n-1},\bar{p}^n-\bar{p}^{n-1})-r(B\bar{u}^n-B\bar{u}^{n-1},\bar{p}^n-\bar{p}^{n-1}) \leq 0.$$

We have (from (3.14))

$$\bar{\lambda}^n = \bar{\lambda}^{n-1} + \rho(B\bar{u}^{n-1} - \bar{p}^{n-1}). \tag{5.22}$$

It then follows from (5.21) and (5.22) that

$$r|\bar{p}^n - \bar{p}^{n-1}|^2 - \rho(B\bar{u}^{n-1} - \bar{p}^{n-1}, \bar{p}^n - \bar{p}^{n-1}) - r(B\bar{u}^n - B\bar{u}^{n-1}, \bar{p}^n - \bar{p}^{n-1}) \leq 0 ,$$

i.e.

$$r(B\bar{u}^n - B\bar{u}^{n-1}, \bar{p}^n - \bar{p}^{n-1}) \geq r|\bar{p}^n - \bar{p}^{n-1}|^2 - \rho(B\bar{u}^{n-1} - \bar{p}^{n-1}, \bar{p}^n - \bar{p}^{n-1}). \tag{5.23}$$

It then follows from (5.18), (5.23) that

$$\left\{ \begin{array}{l} 2\rho r(B\bar{u}^n, \bar{p}^n - \bar{p}^{n-1}) \geq \rho r(|\bar{p}^n|^2 - |\bar{p}^{n-1}|^2) + \rho r|\bar{p}^n - \bar{p}^{n-1}|^2 + \\ \\ + 2\rho(r-\rho)(B\bar{u}^{n-1} - \bar{p}^{n-1}, \bar{p}^n - \bar{p}^{n-1}). \end{array} \right. \tag{5.24}$$

Finally, combining (5.16), (5.24), we obtain

$$\left\{ \begin{array}{l} (|\bar{\lambda}^n|^2 + \rho r|\bar{p}^{n-1}|^2) - (|\bar{\lambda}^{n+1}|^2 + \rho r|\bar{p}^n|^2) \geq 2\rho(F_o'(p^n) - F_o'(p), p^n - p) + \\ \\ + \rho(2r-\rho)|B\bar{u}^n - \bar{p}^n|^2 + \rho r|\bar{p}^n - \bar{p}^{n-1}|^2 + 2\rho(r-\rho)(B\bar{u}^{n-1} - \bar{p}^{n-1}, \bar{p}^n - \bar{p}^{n-1}). \end{array} \right. \tag{5.25}$$

Then using the Cauchy-Schwarz inequality it follows from (5.25) that $\forall \alpha > 0$ we have

$$\left\{ \begin{array}{l} (|\bar{\lambda}^n|^2 + \rho r|\bar{p}^{n-1}|^2) - (|\bar{\lambda}^{n+1}|^2 + \rho r|\bar{p}^n|^2) \geq 2\rho(F_o'(p^n) - F_o'(p), \bar{p}^n) + \\ \\ + \rho(2r-\rho)|B\bar{u}^n - \bar{p}^n|^2 + \rho r|\bar{p}^n - \bar{p}^{n-1}|^2 - \rho|r-\rho|\left(\frac{1}{\alpha}|B\bar{u}^{n-1} - \bar{p}^{n-1}|^2 + \alpha|\bar{p}^n - \bar{p}^{n-1}|^2\right). \end{array} \right. \tag{5.26}$$

If $\rho = r$ it is clear that by utilising the same method as in the proof of Theorem 4.1 we have (5.2)-(5.5).

If $0 < \rho < r$, taking $\alpha = 1$ and observing that $|r-\rho| = r-\rho$, and taking into account (5.26), we have

$$\left\{ \begin{array}{l} (|\bar{\lambda}^n|^2 + \rho r|\bar{p}^{n-1}|^2 + \rho(r-\rho)|B\bar{u}^{n-1} - \bar{p}^{n-1}|^2) - (|\bar{\lambda}^{n+1}|^2 + \rho r|\bar{p}^n|^2 + \rho(r-\rho)|B\bar{u}^n - \bar{p}^n|^2) \geq \\ \\ \geq 2\rho(F_o'(p^n) - F_o'(p), \bar{p}^n) + \rho r|B\bar{u}^n - \bar{p}^n|^2 + \rho^2|\bar{p}^n - \bar{p}^{n-1}|^2 , \end{array} \right.$$

which implies, as in Theorem 4.1, (5.2)-(5.5). If $\rho > r$ we have $|r - \rho| = \rho - r$; it then follows from (5.26) that the convergence

results (5.2)-(5.5) will apply if we have $\rho < \rho_M$, where

$$(5.27)\quad \begin{cases} \rho_M(2r-\rho_M) = \frac{1}{\alpha}\rho_M(\rho_M-r), \\ \rho_M r = \alpha\rho_M(\rho_M-r). \end{cases}$$

By elimination of α we deduce from (5.27) that

$$\rho_M^2 - r\rho_M - r^2 = 0$$

i.e. (since $\rho_M > 0$)

$$\rho_M = \frac{1+\sqrt{5}}{2}\, r.$$

Taking into account the convergence results (5.2)-(5.5), the weak cluster point property of $\{\lambda^n\}_n$ in the statement of the Theorem can easily be deduced, by proceeding to the limit in (3.12)-(3.14). ∎

5.2 Finite-dimensional case

Using a variant of the proof of Theorem 5.1, together with Lemma 4.1, we can easily prove the following:

THEOREM 5.2: *Suppose that the assumptions on* V, H, F, B, G *are those in the statement of Theorem* 4.2. *Then if*

$$0 < \rho_n = \rho < \frac{1+\sqrt{5}}{2}\, r$$

the conclusions in the statement of Theorem 4.2 *are still valid.*

5.3 Discussion. Choice of ρ and of r.

We begin with some remarks:

Remark 5.1: If G is *linear* it has been proved by GABAY-MERCIER [1] that ALG2 converges if

$$0 < \rho_n = \rho < 2r.$$

The proof of this result is rather technical and does not seem to be extendable to the more general cases considered in this book. ∎

Remark 5.2: In the case where G is linear we note that the stage (3.12) of ALG2 is a linear problem relative to the symmetric operator B^tB. Consequently, in finite dimensions (and assuming B injective) it will be convenient to *factorise* once and for all (for example by Cholesky's method) the matrix B^tB which in this case is symmetric and positive definite. ∎

Remark 5.3: Here again, as we have mentioned in Remark 5.1 for ALG1, we can replace the coercivity assumptions on $F_o'(.)$ by an assumption on $G'(.)$. ∎

On the choice of ρ *and* r.

We saw in Section 4.3 that if $F(q) = \frac{1}{2}|q|^2$ and if G is linear, then the sequence $\{u^n\}_n$ relating to ALG1 converges in two iterations at the most if we use $\rho_n = \rho = r$. In the case of ALG2, with the same assumptions on F and G, we also have convergence of $\{u^n\}_n$ in two iterations at the most if $\rho_n = \rho = r = 1$ (for any choice of $\{p^o,\lambda^1\}$). This fact would appear to indicate a rather greater degree of robustness for ALG1.

In general, for given r, experience indicates that the best choice is $\rho = r$. The choice of r is more problematical and in this respect ALG2 is more sensitive than ALG1. Also, for very "stiff" problems ALG1 would appear to be more robust than ALG2; by this we mean that the choice of r is less critical and that the computation time with ALG1 may be much shorter for a given problem than with ALG2.

6. APPLICATION TO SOME NONLINEAR PROBLEMS

6.1 Introduction

The purpose of this section is to show via several examples that the methods developed previously are applicable to a variety of different types of problem. In certain cases, we shall re-encounter classical methods for which the algorithms ALG1 and ALG2 will provide solution procedures which are often simpler than the usual techniques. In other cases we shall obtain decomposition procedures which appear to be new and which may possess certain advantages from the algorithmic point of view. Generally, it will be established

that the decomposition methods of this chapter can extend to a very diverse range of applications, and that they constitute a flexible means of approach to optimisation problems, in a manner which is sometimes rather non-standard.

6.2. Case of the examples of Section 1.1 (Flow of a Bingham fluid and elastoplastic torsion)

The application of algorithms ALG1 and ALG2 to these problems has already been described in Section 3.3 and we shall not go over this ground again. We shall content ourselves here with a few remarks. As regards the convergence of the algorithms, we recall that we have

$$F(q) = \frac{\nu}{2} \int_\Omega |q|^2 dx + g \int_\Omega |q| dx \tag{6.1}$$

in the case of a Bingham fluid, and

$$F(q) = \frac{1}{2} \int_\Omega |q|^2 \, dx + I_{\hat{K}}(q) \tag{6.2}$$

in the case of elastoplastic torsion. In both cases, we can write[5] $F_o(q) = \frac{1}{2} \int_\Omega |q|^2 dx$, this functional being differentiable and strictly convex. It can be shown without difficulty that all the assumptions of Theorem 4.1 and 5.1 are satisfied.

As regards the choice of ρ and of r, we note that these problems are very closely related to the case dealt with in Section 4.3. Experimentally, (GABAY-MERCIER [1]), it has been established that for ALG2 the convergence of u^n is faster than that of p^n or λ^n for $\rho = r = 1$. It would certainly appear that the choice of $r = 1$ is optimal in this case[5]. We shall return to the above topic in more detail in Chapter V.

6.3 A nonlinear Dirichlet problem

This description follows the approach of GLOWINSKI-MARROCCO [1]; the problem described will be considered in detail in Chapter V. For $1 < s < +\infty$, and with Ω a bounded open subset of $\mathbb{R}^N$, we

---

[5] This assumes $\nu = 1$ in (6.1).

consider

$$W_o^{1,s}(\Omega) = \{v \mid v \in L^s(\Omega),\ \nabla v \in (L^s(\Omega))^N,\ v|_\Gamma = 0\}\ . \tag{6.3}$$

We wish to solve in Ω the following nonlinear Dirichlet problem

$$\begin{cases} -\nabla(|\nabla u|^{s-2}\nabla u) = f \ \textit{in}\ \Omega, \\ u|_\Gamma = 0 \end{cases} \tag{6.4}$$

where $f \in W^{-1,s'}(\Omega)$ $(\frac{1}{s} + \frac{1}{s'} = 1)$.
It can be shown by standard techniques, (see LIONS [2]), that (6.4) possesses a unique solution which is also a solution of the minimisation problem

$$\operatorname*{Min}_{v \in W_o^{1,s}(\Omega)} \left[\frac{1}{s}\int_\Omega |\nabla v|^s dx - \langle f, v\rangle\right]\ . \tag{6.5}$$

We note that $W_o^{1,s}(\Omega)$ *is not a Hilbert space*[6]. We cannot therefore, in the infinite-dimensional case, apply the convergence theorems of ALG1 and ALG2. One way of surmounting this difficulty is to consider a discretised problem. We shall then be in a space of finite dimension and it will be possible, the conditions of application being satisfied, to utilise ALG1 and ALG2 for the solution of the approximate problems. To avoid introducing new notation which will needlessly encumber the discussion, we shall be satisfied with a *formal* argument relating to the continuous problem.

Thus suppose

$$V = W_o^{1,s}(\Omega),\ H = (L^s(\Omega))^N,\ H' = (L^{s'}(\Omega))^N,$$
$$B = \nabla.$$

We then put

$$\begin{cases} F(q) = F_o(q) = \frac{1}{s}\int_\Omega |q|^s dx, \\ G(v) = -\langle f, v\rangle\ , \end{cases} \tag{6.6}$$

hence

[6] Except if $s = 2$.

$$F'(q) = q|q|^{s-2} . \tag{6.7}$$

First we note that we clearly have

$$\lim_{|q|_s \to +\infty} \frac{F(q)}{|q|_s} = +\infty , \tag{6.8}$$

where $|q|_s$ is the norm of q in $(L^s(\Omega))^N$, i.e.

$$|q|_s = \left(\int_\Omega |q|^s dx\right)^{1/s}. \tag{6.9}$$

Moreover, for any $p,q \in H$ (see for example GLOWINSKI-MARROCCO [1], CIARLET [1]), we have

$$(F'(q)-F'(p),q-p) \geq \alpha|q-p|_s^s \quad \textit{if} \quad s \geq 2, \tag{6.10}$$

$$(F'(q)-F'(p),q-p) \geq \frac{\alpha|q-p|_s^2}{(|p|_s+|q|_s)^{2-s}} \quad \textit{if} \quad 1 < s \leq 2. \tag{6.11}$$

In addition we can show that

$$|F'(q)-F'(p)|_{s'} \leq \beta(|p|_s+|q|_s)^{s-2}|q-p|_s \quad \textit{if} \quad s \geq 2, \tag{6.12}$$

and

$$|F'(q)-F'(p))|_{s'} \leq \beta|q-p|_s^{s-1} \quad \textit{if} \quad 1 \leq s \leq 2. \tag{6.13}$$

The constants α and β are *independent* of p and of q and are *strictly positive.*

For the solution of (6.4) or (6.5) we shall use the augmented Lagrangian:

$$\mathcal{L}_r(v,q,\mu) = \frac{1}{s}\int_\Omega |q|^s dx - \langle f,v\rangle + \frac{r}{2}\int_\Omega |\nabla v - q|^2 dx + \int_\Omega \mu\cdot(\nabla v - q)dx. \tag{6.14}$$

The algorithm ALG1 can then be written:

$$\lambda^0 \in (L^{s'}(\Omega))^N, \quad \textit{arbitrarily chosen;} \tag{6.15}$$

for $n \geq 0$, λ^n *being known, calculate the solution* u^n, p^n *of*

$$\begin{cases} -r\Delta u^n = f + \nabla\cdot\lambda^n - r\nabla\cdot p^n \ \textit{in} \ \Omega, \\ u^n|_\Gamma = 0, \end{cases} \tag{6.16}$$

$$|p^n|^{s-2}p^n + rp^n = r\nabla u^n + \lambda^n, \tag{6.17}$$

then

$$\lambda^{n+1} = \lambda^n + \rho_n(\nabla u^n - p^n). \tag{6.18}$$

The system (6.16), (6.17) can be solved by the relaxation method (3.8)-(3.10). We note that, by means of an appropriate discretisation, the nonlinear problem (6.17) decomposes into a family of two-dimensional problems which are easy to solve. To obtain ALG2, we replace (6.15) by

$$\{p^0, \lambda^1\} \in H \times H' \quad \textit{arbitrarily chosen;} \tag{6.19}$$

and (6.16) by

$$\begin{cases} -r\Delta u^n = f + \nabla\cdot\lambda^n - r\nabla\cdot p^{n-1} & \textit{in } \Omega, \\ u^n|_\Gamma = 0. \end{cases} \tag{6.20}$$

The properties (6.8) and (6.10)-(6.13) allow, in finite dimensions, the Theorems 4.1 and 5.1 to be applied and the convergence of the above algorithms to be thereby deduced. Numerical experiments show that the direct solution of (6.4) for s in the neighbourhood of 1 or s large (e.g. $s < 1.5$ or $s > 5$) by standard iterative methods (conjugate gradient, Newton-Raphson , nonlinear overrelaxation, etc...) is very difficult. To our knowledge, the only really efficient methods in this case are ALG1 and ALG2. For more details see Chapter V.

6.4 Application to the solution of mildly nonlinear systems and relationship with alternating direction methods

The algorithms ALG1 and ALG2 can also be used (assuming the introduction of an appropriate augmented Lagrangian) for the solution of *mildly nonlinear* problems of the type

$$Au + \varphi(u) = f, \tag{6.21}$$

where, limiting ourselves to finite dimensions and writing $v = \{v_1, \ldots v_N\}$:

- A is an $N \times N$ symmetric positive definite matrix,
- $\varphi : \mathbb{R}^N \to \mathbb{R}^N$ is a nonlinear mapping of diagonal type, i.e.

$$(\varphi(v))_i = \varphi_i(v_i) \ , \ i=1,\ldots N,$$

with $\varphi_i : \mathbb{R} \to \mathbb{R}$, continuous and increasing (we can always assume that $\varphi_i(0) = 0$),

- $f \in \mathbb{R}^N$.

The discretisation by *finite differences* or *finite elements* of certain mildly nonlinear elliptic or parabolic problems leads to problems of the type (6.21) (some examples will be given in Chapter IV).

Remark 6.1: The algorithms which we shall be describing for solution of (6.21) extend to the case where A is *non-symmetric, and positive definite.* ∎

We define $\forall \ i = 1,\ldots\ldots N,$

$$\phi_i(t) = \int_0^t \varphi_i(\tau)d\tau \ .$$

Since the function φ_i is continuous and increasing it follows that ϕ_i is C^1 and *convex*. Since the operator A is symmetric it follows that solving (6.21) is equivalent to solving the minimisation problem

$$(6.22) \quad \begin{cases} J(u) \le J(v) \quad \forall\, v \in \mathbf{R}^N, \\ u \in \mathbf{R}^N, \end{cases}$$

with, in (6.22)

$$(6.23) \qquad J(v) = \frac{1}{2}\,(Av,v) + \sum_{i=1}^{N} \phi_i(v_i) - (f,v)$$

where $(.,.)$ denotes the canonical Euclidian inner product on $\mathbb{R}^N$ and $\|\cdot\|$ denotes the associated norm.

The above properties of A and φ_i imply that (6.21), (6.22) admits a *unique solution*.

Remark 6.2: If A is non-symmetric and positive definite, it can easily be shown that (6.21) still admits a unique solution. ∎

The problem (6.22) is a particular problem (P) in which, with the notation of Section 1.1, we can take

$$V = H = \mathbb{R}^N, \quad B = I, \tag{6.24}$$

$$G(v) = \sum_{i=1}^{N} \phi_i(v_i) - (f,v), \tag{6.25}$$

$$F(q) = F_o(q) = \frac{1}{2}(Aq,q) \Longrightarrow F_o'(q) = Aq. \tag{6.26}$$

By virtue of this decomposition we can solve (6.21), (6.22) by ALG1 and ALG2 (note that in the present case G is nonlinear).

Remark 6.3: Instead of defining G and F by (6.25), (6.26), we could use

$$G(v) = \sum_{i=1}^{N} \phi_i(v_i),$$

$$F(q) = \frac{1}{2}(Aq,q) - (f,q). \quad \blacksquare$$

We naturally associate with (6.24)-(6.26) the augmented Lagrangian

$$\mathcal{L}_r(v,q,\mu) = \frac{1}{2}(Aq,q) + \sum_{i=1}^{N} \phi_i(v_i) - (f,v) + \frac{r}{2}\|v-q\|^2 + (\mu,v-q). \tag{6.27}$$

Since the constraint $v - q = 0$ is linear, $\mathcal{L}_r$ admits a saddle-point on $\mathbb{R}^N \times \mathbb{R}^N \times \mathbb{R}^N$. This saddle-point is in fact unique and equal to $\{u,u,Au\}$.

Solution of (6.21), (6.22) by ALG1:

It follows from (3.4), (3.5), (6.27) that the application of ALG1 to the solution of (6.21), (6.22) leads to the following algorithm

$$\lambda^o \in \mathbb{R}^N, \tag{6.28}$$

then for $n \geq 0$,

$$ru^n + \varphi(u^n) = f + rp^n - \lambda^n, \tag{6.29}$$

$$(6.30)\qquad (rI+A)p^n = ru^n+\lambda^n,$$

$$(6.31)\qquad \lambda^{n+1} = \lambda^n+\rho_n(u^n-p^n).$$

The nonlinear system (6.29), (6.30) can be solved by the *block relaxation* method of Section 3.2 and we note that, p^n and λ^n being known, the calculation of u^n in (6.29) reduces to the solution of N nonlinear equations, each in a single variable, which are independent of each other and of the form

$$(6.32)\qquad r\xi+\varphi_i(\xi) = b\ ,\ i=1,\dots N.$$

The parameter r being >0 and φ_i being C^0 and increasing, equation (6.32) admits a unique solution which can be calculated by various methods (see HOUSEHOLDER [1], BRENT [1]).

Similarly if u^n and λ^n are known in (6.30), we obtain p^n by solving a linear system with matrix rI+A. Assuming r independent of n, it is thus convenient to factorise rI+A once and for all (by a Gauss or Cholesky method).

Solution of (6.21), (6.22) *by* ALG2:

It suffices to replace (6.28) by

$$(6.33)\qquad \{p^0,\lambda^1\} \in \mathbf{R}^N\times\mathbf{R}^N,$$

and (6.29) by

$$(6.34)\qquad ru^n+\varphi(u^n) = f+rp^{n-1}-\lambda^n.$$

It then follows from Theorem 5.2 that we have convergence of (6.33), (6.34), (6.30), (6.31) if $0<\rho_n=\rho<\frac{1+\sqrt{5}}{2}r$.

Remark 6.4: Suppose that $\rho_n=\rho=r$ in ALG2: we then have

$$(6.35)\qquad \begin{cases} ru^n+\varphi(u^n) = f+rp^{n-1}-\lambda^n, \\ rp^n+Ap^n = ru^n+\lambda^n, \\ \lambda^{n+1} = \lambda^n+r(u^n-p^n). \end{cases}$$

It then follows from (6.35) that

$$(6.36)\qquad \lambda^{n+1} = Ap^n.$$

We then deduce from (6.35), (6.36) that

$$(6.37)\qquad ru^n+\varphi(u^n)+Ap^{n-1} = f+rp^{n-1}$$

$$(6.38) \qquad rp^n + Ap^n + \varphi(u^n) = f + rp^{n-1}.$$

Consequently therefore, if $\rho_n = \rho = r$, ALG2 reduces (putting $u^n = p^{n-\frac{1}{2}}$) to an *alternating direction method* applied to the solution of (6.21) (we shall return to the above topic in Chapters IV and IX).

Remark 6.5: We shall see in Chapter IV that ALG1 combined with the block relaxation method of Section 3.2 is more robust and more efficient than ALG2 if φ is of "low" differentiability; this will be the case if for example we use a finite-element or finite-difference approximation of the non-linear elliptic problem

$$(6.39) \qquad \begin{cases} -\Delta u + u|u|^{s-2} = f \ \textit{in} \ \Omega, \\ u|_\Gamma = 0, \end{cases}$$

with $1 < s < 2$. Chapter IV will give a number of results of numerical experiments relating to problems of the type (6.39). ∎

7. APPLICATIONS TO NONLINEAR PROGRAMMING PROBLEMS

In this section we shall be presenting two families of applications of the methods of this chapter to nonlinear programming. The first, described in Section 7.1, is linked to the augmented Lagrangian methods introduced by ROCKAFELLAR [1] for the case of *inequality constraints*. This involves a natural extension of the method of *Hestenes and Powell* studied in Chapter I. In this respect, algorithm ALG2 can be considered as a new technique for obtaining the required saddle-point. We shall then present in Section 7.2 another approach which, in somewhat more tangible fashion, is based on a decomposition principle which results in a decoupling of the constraints. Finally, in Section 7.3 we shall describe an application of the methods of this chapter to the solution of the so-called Weber problem, in connection with which some numerical results will be presented.

7.1 An augmented Lagrangian in the case of inequality constraints

We consider here in $\mathbb{R}^N$ a classical nonlinear programming problem subject to inequality constraints:

$$(7.1)\qquad \begin{cases} \text{Inf } F(v) \ , \ v \in \mathbf{R}^N, \\ g_i(v) \le 0, \ i=1,\dots,M. \end{cases}$$

We shall assume that the M functions g_i are *convex* and we write

$$(7.2)\qquad \begin{cases} G : \mathbf{R}^N \to \mathbf{R}^M, \\ (G(v))_i = g_i(v) \ , \ i=1,\dots,M. \end{cases}$$

The problem (7.1) can clearly be written in the form:

$$(7.3)\qquad \begin{cases} \underset{\{v,q\}}{\text{Inf}} \ F(v) \ \textit{under the constraints} \\ q \le 0 \ , \ G(v) = q \quad (\textit{with } q \le 0 \Leftrightarrow q_i \le 0 \ \ \forall i=1,\dots M \textit{ if } q = \{q_i\}_{i=1}^M), \end{cases}$$

which leads to the augmented Lagrangian

$$(7.4)\qquad \mathcal{L}_r(v,q,\mu) = F(v)+\frac{r}{2}\,|G(v)-q|^2+(\mu,G(v)-q) \ ,$$

where $|q| = \left(\sum_{i=1}^{M} q_i^2\right)^{1/2}$.

The minimisation with respect to q can be carried out directly (see Section 2.3); in fact, for fixed v and μ, it can easily be shown that for $q \le 0$ the minimum is attained at $q = p$, where, for $i = 1,\dots,M$, we have

$$(7.5)\qquad p_i = P_{\le 0}\left(\frac{\mu_i}{r} + g_i(v)\right) = \min\{0, \frac{\mu_i}{r} + g_i(v)\} \ .$$

Substituting (7.5) into (7.4), we obtain with the notation of Section 2.3

$$(7.6)\qquad L_r(v,\mu) = F(v) + \sum_{i=1}^{M} \{ \frac{r}{2} \,|\max(0,\frac{\mu_i}{r} + g_i(v))|^2 - \frac{\mu_i^2}{2r} \} \ ,$$

which corresponds directly with the form given by ROCKAFELLAR [1]. A UZAWA-type algorithm for $L_r(v,\mu)$ requires the minimisation with respect to v, with μ fixed, of $L_r(v,\mu)$. This requires the solution of a highly nonlinear problem.

We now consider algorithm ALG2 for $\mathcal{L}_r(v,q,\mu)$. If we assume that F and G are differentiable, for given λ^1 and p^0, we can determine u^n then p^n by

$$F'(u^n)+(G'(u^n))^t[rG(u^n)-rp^{n-1}+\lambda^n] = 0, \tag{7.7}$$

and

$$p_i^n = \min \{0, g_i(u^n)+\frac{\lambda_i^n}{r}\}, \quad i=1,\ldots M. \tag{7.8}$$

Finally, we can calculate λ^{n+1} by

$$\lambda^{n+1} = \lambda^n+\rho(G(u^n)-p^n). \tag{7.9}$$

In the case of *affine constraints*, Theorem 5.1 assures the convergence of this algorithm under quite general conditions on F. We have not tried to extend this proof to the general case. Similarly the use of ALG1 and of the relaxation method (3.8)-(3.10) can be considered as being equivalent (apart from the solution method) to the application of a UZAWA-type algorithm to $L_r(v,\mu)$. Viewed in this way, the use of ALG2 is a variant in which the minimisation of $L_r(v,\mu)$ with respect to v is carried out in an incomplete manner.

7.2 Minimisation of a functional over an intersection of convex sets

7.2.1 *Statement of the problem*

Let V be a real Hilbert space. We consider on V a functional F, which is convex, proper and l.s.c. We wish to minimise F over the (non-empty) closed convex set K, where we have

$$K = \bigcap_{i=1}^{M} K_i \tag{7.10}$$

where for $i = 1,\ldots,M$, each of the K_i is itself convex and closed.

Remark 7.1: This situation obviously encompasses the classical case of Section 7.1. It suffices to put $K_i = \{v | g_i(v) \leq 0\}$. We shall present here another procedure for associating an augmented Lagrangian with this problem. ∎

Remark 7.2: An important particular case is that where F is *quadratic*, i.e.

$$(7.11) \qquad F(v) = \frac{1}{2}\, a(v,v) - \langle f,v\rangle_{V'\times V} = \frac{1}{2}\langle Av,v\rangle_{V'\times V} - \langle f,v\rangle_{V'\times V} \quad .$$

In (7.11), $a(.,.)$ is bilinear, continuous, symmetric and V-elliptic (i.e. $a(v,v) \geq \alpha \|v\|_V^2 \quad \forall v \in V,\ \alpha > 0$) and the operator $A \in \mathcal{L}(V,V')$ is defined by $a(u,v) = \langle Au,v\rangle \quad \forall u,v \in V$.

The problem to be solved can then be written in the form of a variational inequality:

$$(7.12) \qquad \begin{cases} a(u,v-u) \geq \langle f,v-u\rangle \quad , \forall v \in K, \\ u \in K. \end{cases}$$

This formulation can be extended to the case where $a(u,v)$ is not symmetric. In the latter case (7.12) is no longer equivalent to a minimisation problem. ∎

We shall now introduce for the solution of this problem a decomposition principle whose aim is to obtain a family of optimisation problems, coordinated via a Lagrange multiplier.

We thus put

$$(7.13) \qquad W = \{\{v,q\} \in V \times V^M,\ v - q_i = 0 \ ,\ \forall\, i=1,\ldots,M\} \ ,$$

and

$$(7.14) \qquad \chi = \{\{v,q\} \in W \ ,\ q_i \in K_i, \forall\, i=1,\ldots,M\}.$$

It is clear that the original problem is equivalent to

$$(7.15) \qquad \inf_{\{v,q\} \in \chi} f_o(q) \ ,$$

where we have written,

$$(7.16) \qquad f_o(q) = \frac{1}{M} \sum_{i=1}^{M} F(q_i).$$

Suppose I_{K_i} is the indicator function of the convex set K_i. We write

$$(7.17) \qquad f_1(q) = \sum_{i=1}^{M} I_{K_i}(q_i).$$

It is then natural to consider the augmented Lagrangian

$$(7.18)\qquad \mathcal{L}_r(v,q,\mu) = f_0(q)+f_1(q)+\frac{r}{2M}\sum_{i=1}^{M}|v-q_i|_V^2+\frac{1}{M}\sum_{i=1}^{M}(\mu_i,v-q_i)_V.$$

We thus look for a saddle-point of $\mathcal{L}_r$ on $V \times V^M \times V^M$. The existence of such a saddle-point poses no problem in the finite dimensional case.

Remark 7.3: The formulation (7.18) will be simpler to work with if V is identified with its dual. This poses no problem in *finite dimensions*. It will suffice to equip $\mathbb{R}^N$ with the Euclidian metric. If another metric is used, we shall indicate at that time the modifications which need to be made to the algorithms.

7.2.2. *Solution of the problem by* ALG1 *and* ALG2

In accordance with the general results, algorithm ALG1 is here written as follows:

$$(7.19)\qquad \lambda^0 \in V^M \text{ specified arbitrarily;}$$

for $n \geq 0$, *and with* λ^n *known, calculate the solution* $\{u^n,p^n\}$ *of the system*

$$(7.20)\qquad u^n = \frac{1}{M}\sum_{i=1}^{M} p_i^n - \frac{1}{rM}\sum_{i=1}^{M}\lambda_i^n,$$

$$(7.21)\qquad \begin{cases} F(q_i)-F(p_i^n)+r(p_i^n,q_i-p_i^n)_V \geq (ru^n+\lambda_i^n,q_i-p_i^n)_V \quad \forall\, q_i \in K_i, \\ p_i^n \in K_i\ , \quad i=1,\ldots,M, \end{cases}$$

then λ^{n+1} *by*

$$(7.22)\qquad \lambda_i^{n+1} = \lambda_i^n + \rho_n(u^n-p_i^n),\ i=1,\ldots,M.$$

We note that (7.21) is a system of variational inequalities, each of these inequalities involving only *a single* constraint $p_i^n \in K_i$. In many cases each of the problems decoupled in this way will be easier to solve than the original problem. For example if we use an algorithm based upon a projection onto K, it is in general much easier to project onto each of the K_i independently than onto their intersection. The same remark also applies for algorithms requiring the construction of an admissible solution. Also, such a process is well adapted to *parallel computation*, the possibility

of which can be anticipated on future computers. We can clearly pass from algorithm ALG1 to ALG2 by replacing (7.19) and (7.20) by:

(7.23) $\{p^0,\lambda^1\} \in V^M \times V^M$ *chosen arbitrarily;*

and

$$u^n = \frac{1}{M}\sum_{i=1}^{M} p_i^{n-1} - \frac{1}{rM}\sum_{i=1}^{M} \lambda_i^n \ . \tag{7.24}$$

The calculation of u^n and p^n has become *sequential* and no longer *simultaneous*, but the M components of p^n can be calculated *independently* of one another, and in particular can be calculated in *parallel*.

By way of an example, we consider ALG2 in the particular case where F is of the form (7.11). In order to fully describe the algorithm we introduce the operator S of isomorphism between V and V'. (We have $S = I$ if V is identified with its dual). We clearly have

$$(u,v)_V = \langle Su,v\rangle_{V'\times V} \ . \tag{7.25}$$

Using this notation, the algorithm ALG2 can be written:

(7.26) $\{p^0,\lambda^1\} \in V^M \times V^M$ *chosen arbitrarily;*

$$u^n = \frac{1}{M}\sum_{i=1}^{M} p_i^{n-1} - \frac{1}{rM}\sum_{i=1}^{M} \lambda_i^n \ , \tag{7.27}$$

$$\begin{cases} a(p_i^n,q_i-p_i^n)+r(p_i^n,q_i-p_i^n)_V \geq \langle f,q_i-p_i^n\rangle + \langle S(ru^n+\lambda_i^n),q_i-p_i^n\rangle \\ \\ \forall q_i \in K_i,\ p_i^n \in K_i \ , \ i=1,\ldots,M, \end{cases} \tag{7.28}$$

$$\lambda_i^{n+1} = \lambda_i^n + \rho_n(u^n-p_i^n) \ , \ i=1,\ldots M. \tag{7.29}$$

Such an algorithm can also be applied in the case where the bilinear form $a(u,v)$ is V-elliptic, and non-symmetric. If, in the symmetric case, we equip V with the norm $\|v\|^2 = a(v,v) = \langle Av,v\rangle$, (7.28) becomes

$$\begin{cases} (1+r)\ \langle Ap_i^n,q_i-p_i^n\rangle \geq \langle f,q_i-p_i^n\rangle + \langle A(ru^n+\mu_i^n),\ q_i-p_i^n\rangle \\ \\ \forall q_i \in K_i \ , \ p_i^n \in K_i \ , \ i=1,\ldots,M. \end{cases} \tag{7.30}$$

Remark 7.4: The algorithm ALG2 has been deduced from ALG1 via a solution method based on a block relaxation. The reader can easily deduce an algorithm in which problems of the type (7.21) or (7.28) are solved sequentially and no longer in parallel fashion. ■

Finally, we note that the augmented Lagrangian which we have used is only one possible example. We could for example in (7.13) have defined alternatively

$$W = \{\{v,q\} \in V\times V^M \ ; \ v=q_1 \ ; \ q_i=q_{i-1} \ , \ i=1,\ldots,M\} \tag{7.31}$$

which would of course have led to quite different algorithms.

Remark 7.5: The methods described in the present section 7.2 can be viewed as fractional-step methods with multiplier, which in a certain sense generalise the methods described in BENSOUSSAN-LIONS-TEMAM [1,Chapter 2], and which,amongst other things, allow us to avoid the use of the divergent series utilised in the above reference. ■

7.3. Application to the solution of the Weber problem

7.3.1. *Statement of the problem*

Certain authors (see COOPER-KATZ [1], for example) use the designation *Weber problem* for the following *nondifferentiable minimisation* problem:

$$\underset{\underset{\sim}{y}\in \mathbb{R}^N}{\text{Min}} \ J(\underset{\sim}{y}), \tag{7.32}$$

where

$$J(\underset{\sim}{y}) = \sum_{i=1}^{M} \alpha_i \, \|\underset{\sim}{y}-\underset{\sim}{x}_i\| \ , \tag{7.33}$$

with, in (7.33),

$$\alpha_i > 0 \ \ \forall i=1,\ldots M \ ; \ \underset{\sim}{x}_i \in \mathbb{R}^N \ \ \forall i=1,\ldots M, \tag{7.34}$$

and with $\|\cdot\|$ defined by $\|\underset{\sim}{y}\| = (\sum_{j=1}^{N} y_j^2)^{1/2}$ (if $\underset{\sim}{y} = \{y_j\}_{j=1}^N$).

Problem (7.32) *admits at least one solution.*

7.3.2. *Introduction of an augmented Lagrangian for the solution of problem* (7.32)

Problem (7.32) is clearly equivalent to the problem

$$\text{(7.35)} \qquad \min_{\{\underset{\sim}{q},\underset{\sim}{y}\}\in W} \{ \sum_{i=1}^{M} \alpha_i \|\underset{\sim}{q}_i\| \},$$

where

$$\text{(7.36)} \qquad W = \{\{\underset{\sim}{q},\underset{\sim}{y}\} \in \mathbb{R}^{NM}\times\mathbb{R}^{N} \mid \underset{\sim}{q} = \{\underset{\sim}{q}_i\}_{i=1}^{M},\ \underset{\sim}{q}_i \in \mathbb{R}^{N}\ \forall i,\ \underset{\sim}{q}_i = \underset{\sim}{y}-\underset{\sim}{x}_i\}.$$

It follows from (7.35), (7.36), and from the preceding sections, that an augmented Lagrangian naturally associated with problem (7.32) is given by:

$$\text{(7.37)} \qquad \mathscr{L}_r(\underset{\sim}{y},\underset{\sim}{q},\underset{\sim}{\mu}) = \sum_{i=1}^{M} \alpha_i \|\underset{\sim}{q}_i\| + \frac{r}{2} \sum_{i=1}^{M} \|(\underset{\sim}{y}-\underset{\sim}{x}_i)-\underset{\sim}{q}_i\|^2 + \sum_{i=1}^{M} (\underset{\sim}{\mu}_i,(\underset{\sim}{y}-\underset{\sim}{x}_i)-\underset{\sim}{q}_i),$$

where $(.,.)$ denotes, in (7.37), the ordinary Euclidian inner product on $\mathbb{R}^N$ (i.e. that associated with $\|\cdot\|$) and $\underset{\sim}{\mu} = \{\underset{\sim}{\mu}_i\}_{i=1}^{M} (\in \mathbb{R}^{NM})$.

Remark 7.6: In the case where $\dfrac{\max_i \alpha_i}{\min_i \alpha_i} \gg 1$, it would be appropriate to use, instead of (7.37), the augmented Lagrangian defined by

$$\text{(7.38)} \qquad \mathscr{L}_r(\underset{\sim}{y},\underset{\sim}{q},\underset{\sim}{\mu}) = \sum_{i=1}^{M} \alpha_i \|\underset{\sim}{q}_i\| + \frac{r}{2} \sum_{i=1}^{M} \alpha_i \|(\underset{\sim}{y}-\underset{\sim}{x}_i)-\underset{\sim}{q}_i\|^2 + \sum_{i=1}^{M} \alpha_i (\underset{\sim}{\mu}_i,(\underset{\sim}{y}-\underset{\sim}{x}_i)-\underset{\sim}{q}_i). \blacksquare$$

7.3.3. *Application of* ALG2 *to the solution of* (7.32)

The application of ALG2 to the solution of (7.32), via the determination of the saddle-points, in $\mathbb{R}^N \times \mathbb{R}^{NM} \times \mathbb{R}^{NM}$, of the augmented Lagrangian (7.37), leads to the following algorithm:

$$\text{(7.39)} \qquad \{\underset{\sim}{x}^0,\underset{\sim}{\lambda}^1\} \in \mathbb{R}^N \times \mathbb{R}^{NM}, \quad \textit{specified arbitrarily;}$$

then, for $n \geq 1$, *assuming* $\{\underset{\sim}{x}^{n-1},\underset{\sim}{\lambda}^n\}$ *known, determine successively* $\underset{\sim}{p}^n,\underset{\sim}{x}^n$ *and* $\underset{\sim}{\lambda}^{n+1}$ *by*

$$\text{(7.40)} \qquad \underset{\sim}{a}_i^n = r(\underset{\sim}{x}^{n-1}-\underset{\sim}{x}_i) + \underset{\sim}{\lambda}_i^n,\ i=1,\ldots M,$$

$$(7.41)\qquad \begin{cases} \underset{\sim}{p}_i^n = \dfrac{\|\underset{\sim}{a}_i^n\| - \alpha_i}{r\,\|\underset{\sim}{a}_i^n\|}\ \underset{\sim}{a}_i^n \ \textit{if}\ \|\underset{\sim}{a}_i^n\| \geq \alpha_i\ , \\ \underset{\sim}{p}_i^n = \underset{\sim}{0} \quad \textit{otherwise} \quad ;\ i=1,\ldots M, \end{cases}$$

$$(7.42)\qquad \underset{\sim}{x}^n = \frac{1}{M}\sum_{i=1}^{M}(\underset{\sim}{x}_i+\underset{\sim}{p}_i^n) - \frac{1}{rM}\sum_{i=1}^{M}\underset{\sim}{\lambda}_i^n,$$

$$(7.43)\qquad \underset{\sim}{\lambda}_i^{n+1} = \underset{\sim}{\lambda}_i^n + \rho(\underset{\sim}{x}^n - \underset{\sim}{x}_i - \underset{\sim}{p}_i^n).\ \blacksquare$$

It should be noted that we are not satisfying here the conditions of application of Theorems 5.1 and 5.2 of Section 5.4. In fact, having regard to the choice made for $\mathcal{L}_r$ (see (7.37)), we have

$$F(\underset{\sim}{q}) = \sum_{i=1}^{M}\alpha_i\,\|\underset{\sim}{q}_i\|$$

and $G \equiv 0$; F is therefore *nondifferentiable* and *not strictly convex*.

7.3.4. *Numerical applications.*

We shall now apply algorithm (7.39)-(7.43) to the solution of a particular Weber problem; the problem in question (considered in COOPER-KATZ, loc. cit.) is defined in $\mathbb{R}^2$ by the α_i and $\underset{\sim}{x}_i$ (i = 1,...10) in Table 7.1:

i	α_i	$\underset{\sim}{x}_i$
1	3	{89,73}
2	8	{36,89}
3	3	{39,9}
4	7	{14,5}
5	1	{46,12}
6	3	{55,1}
7	9	{53,64}
8	6	{32,57}
9	7	{68,42}
10	5	{63,92}

Table 7.1

We have used algorithm (7.39)-(7.43) with

(7.44) $\underset{\sim}{x}^0 = \underset{\sim}{0}, \ \underset{\sim}{\lambda}^1 = \underset{\sim}{0},$

(7.45) $\rho = r,$

and the *termination test*

(7.46) $$R^n = \frac{\|\underset{\sim}{x}^{n+1} - \underset{\sim}{x}^n\|_1}{\|\underset{\sim}{x}^n\|_1} \leq 10^{-6}$$

(where $\|\underset{\sim}{y}\|_1 = |y_1| + |y_2|$ if $\underset{\sim}{y} = \{y_1, y_2\}$) . In Table 7.2 we have indicated, for several values of r, the number of iterations necessary for convergence (under the conditions (7.44)-(7.46)) and the corresponding calculated solutions.

r	number of iterations	calculated soln.
0.1	41	{51.670,62.159}
1	168	{51.669,62.159}
5	710	{51.666,62.154}

Table 7.2

It will be noted that $\underset{\sim}{x}^0 = \{0,0\}$ is "rather" far away from the calculated solutions. The results obtained by means of (7.39)-(7.43) coincide, to very high accuracy, with those obtained in COOPER-KATZ, loc. cit., by a steepest descent method; in actual fact the convergence of algorithm (7.39)-(7.43) is very fast (for r = 0.1) since (see Table 7.3), as early as the fifth iteration, we already have a very good approximate solution to the Weber problem considered.

If instead of initialising algorithm (7.39)-(7.43) by (7.44) we use $\underset{\sim}{\lambda}^1 = \underset{\sim}{0}$ and (as in COOPER-KATZ, loc. cit.)

$$\underset{\sim}{x}^0 = \frac{\sum_{i=1}^{10} \alpha_i \underset{\sim}{x}_i}{\sum_{i=1}^{10} \alpha_i} \quad (\text{ i.e. the } \textit{barycentre} \text{ of the } \underset{\sim}{x}_i),$$

we have convergence, for r = 0.1, in 25 iterations (instead of 41).

n	R^n	$\underset{\sim}{x}^n$
0		{0,0}
5	0.950×10^{-3}	{51.773,62.166}
10	0.117×10^{-3}	{51.699,62.098}
15	0.682×10^{-3}	{51.687,62.144}
20	0.185×10^{-4}	{51.679,62.154}
25	0.957×10^{-5}	{51.674,62.156}
30	0.512×10^{-5}	{51.672,62.158}
35	0.234×10^{-5}	{51.671,62.159}
40	0.109×10^{-5}	{51.670,62.159}
41	0.944×10^{-6}	{51.670,62.159}

Table 7.3 (r = 0.1)

8. GENERAL DISCUSSION ON CHAPTER III

As we have mentioned several times in the preceding text, the methods of this chapter can be extended to variational-inequality problems which are not equivalent to optimisation problems. They can also be used, as in BEGIS [2], for the solution of nonlinear problems of order 4, corresponding to a problem involving the flow of a Bingham fluid which is more general than the case of Example 1 of Section 1.1. We shall return to the above topic in Chapter VII.

The decomposition-coordination method which we have presented can be related to the methods described in BENSOUSSAN-LIONS-TEMAM [1]. Historically speaking, it would appear that the use of an augmented Lagrangian for solving nonlinear variational problems[7] via ALG1 and ALG2 is due to GLOWINSKI-MARROCCO [1], [2], [3]. The first proof of convergence of ALG2 (in the case where G is linear) is due to GABAY-MERCIER [1].

It should also be pointed out that, depending on the type of problem considered, natural variations of the algorithms described may lead to more rapid convergence.

To conclude this chapter, it should be mentioned that by making use of the results of OPIAL [1], we in fact obtain in Theorems 4.1 and 5.1 (respectively 4.2 and 5.2) the convergence of the whole sequence $\{\lambda^n\}$ to λ^*, such that $\{u,p,\lambda^*\}$ is a saddle-point of

[7] Of boundary value type.

$\mathcal{L}$ (and of $\mathcal{L}_r$) on $V \times H \times H$. We refer to G.L.T. [2, Appendix 2] for a proof of this result in a more general context.

CHAPTER IV

NUMERICAL SOLUTION OF MILDLY NONLINEAR PROBLEMS BY AUGMENTED LAGRANGIAN METHODS

M. Fortin, R. Glowinski, T.F. Chan

1. INTRODUCTION

This chapter partly carries on the work of CHAN-GLOWINSKI [1], [2] and extends the algorithmic part of it, in particular the part dealing with approximation by *finite element* methods and with the use of *quadrature formulas*. We shall also see how, by a judicious choice of the functional spaces and of the decomposition, we can obtain *hybrid finite element* methods and solve the corresponding approximate problems by augmented Lagrangian methods.

We shall present some numerical results illustrating the potentialities of the methods described below and we shall show the close links which exist between these algorithms and the *alternating direction methods* of *Peaceman-Rachford* and *Douglas-Rachford*.

In the remainder of this chapter we shall thus be considering the numerical solution of *mildly nonlinear* problems of the following type, on a domain Ω of $\mathbb{R}^N$ with boundary $\partial\Omega = \Gamma$

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

where, in (1.1), we have:

(i) A is a second-order elliptic operator, possibly non-symmetric,

(ii) ϕ is an increasing mapping (in the wide sense) continuous from $\mathbb{R}$ into $\mathbb{R}$,

(iii) f is a function defined over Ω.

As we shall see later, the results obtained can be extended to

multivalued equations (*) of the type

$$f \in Au + \partial j(u), \tag{1.2}$$

where $\partial j(u)$ denotes the *sub-differential*[1] with respect to u of a *convex functional* $j(.)$.

We shall first briefly review the results of CHAN-GLOWINSKI [1],[2] concerning the *existence* and the *uniqueness* of a solution of problem (1.1), then we shall next describe a procedure for approximating this problem by a finite element method. Finally we shall show how - using the methods of Chapter III - we can decompose problem (1.1) through the use of a suitable augmented Lagrangian so as to obtain the classical alternating direction methods.

2. A CLASS OF MILDLY NONLINEAR ELLIPTIC PROBLEMS

2.1. Formulation of the problem

We consider a *bounded* domain Ω in $\mathbb{R}^N$, with sufficiently regular boundary Γ (say, Lipschitz continuous in the sense of NECAS [1]), also (see Chapters II and III for the notation)

(i) $V = H_0^1(\Omega)$,

(ii) a *continuous linear form* $L : V \to \mathbb{R}$, i.e. $L(v) = \langle f, v \rangle$, where $f \in V' = H^{-1}(\Omega)$ and where $\langle \cdot, \cdot \rangle$ is the bilinear form of the duality between V' and V,

(iii) $a: V \times V \to \mathbb{R}$, *a continuous bilinear form*, which is V-elliptic (i.e. $\exists\, \alpha > 0$ such that

$$a(v,v) \geq \alpha |v|_1^2 \quad \forall v \in V, \tag{2.1}$$

where we write

(*) *Translator's Note:* The term "multivalued equation" is used to denote an equation associated with a multivalued operator; such an equation is sometimes known under the French name *multivoque equation.*

[1] See, for example, ROCKAFELLAR [4], EKELAND-TEMAM [1] for the definition of subdifferentials.

(2.2) $$|v|_1 = \left(\int_\Omega |\nabla v|^2 \, dx\right)^{1/2}$$

the usual norm[2] on $H_0^1(\Omega)$).

We do not assume a priori that $a(.,.)$ is symmetric.

(iv) a continuous function $\phi : \mathbb{R} \to \mathbb{R}$, increasing in the wide sense and satisfying $\phi(0) = 0$; we put

(2.3) $$\Phi(t) = \int_0^t \phi(\tau) \, d\tau,$$

(2.4) $$j(v) = \begin{cases} \int_\Omega \Phi(v) dx \ \textit{if} \ \Phi(v) \in L^1(\Omega), \\ +\infty \quad \textit{otherwise} \ ; \end{cases}$$

the function Φ is then *convex*, C^1 and *non-negative* with $\Phi(0) = 0$; it can be shown that $j(.)$ is *convex, proper* and *l.s.c.* on $L^1(\Omega)$ (thus a fortiori on $V = H_0^1(\Omega)$).

Remark 2.1: The continuity of $\phi(.)$ is essential for obtaining certain of the results of CHAN-GLOWINSKI [1], [2]; formally, at least, it is in no way necessary for the implementation of the algorithms which we describe in the remainder of this chapter. ∎

Consider then the *nonlinear variational equality* problem

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

we associate with (2.5) the *variational inequality* problem

(2.6) $$\begin{cases} u \in V, \\ a(u, v-u) + j(v) - j(u) \geq \langle f, v-u \rangle \quad \forall v \in V. \end{cases}$$

Under the above assumptions on f, $a(\cdot,\cdot)$, $\phi(\cdot)$, it can be shown that problems (2.5) and (2.6) are equivalent and admit a *unique solution;* nonetheless problem (2.6) remains meaningful (see LIONS [1],

---

[2] At least, when Ω is *bounded*.

G.L.T. [1], [2]) even when $j(.)$ *is not differentiable;* this is for example the case with

$$j(v) = \int_\Omega |v| dx \ . \tag{2.7}$$

The inequality (2.6) corresponds, in general, to a multivalued equation of the type (1.2). In the case where the bilinear form $a(.,.)$ is *symmetric*, problem (2.6) is equivalent to the following problem in the *Calculus of Variations:*

$$(2.8)_1 \quad \left\{ \begin{array}{l} \textit{Find} \quad u \in V \quad \textit{such that} \\ \\ J(u) \leq J(v) \quad \forall v \in V \ , \end{array} \right.$$

with

$$(2.8)_2 \qquad J(v) = \frac{1}{2} a(v,v) + j(v) - \langle f,v \rangle \ ;$$

under the preceding assumptions on f, $a(\cdot,\cdot)$, $\phi(\cdot)$, the minimisation problem (2.8) possesses a unique solution; this comes from the fact that, $j(.)$ being convex, proper and l.s.c. on V, we can apply to the problems (2.6) and (2.8) a number of general results concerning variational inequalities and the minimisation of convex functions; these results are established in e.g. LIONS-STAMPACCHIA [1], LIONS [1], GLOWINSKI [1], [2], EKELAND-TEMAM [1]. It is shown in CHAN-GLOWINSKI [1] and GLOWINSKI [1] that the sufficient conditions of application are fulfilled; it is further shown that $\phi(u) \in L^1(\Omega) \cap V'$, and that there is equivalence between (2.5), (2.6) (and (2.8) if $a(.,.)$ is symmetric). We shall not dwell any further on these theoretical questions concerning problem (1.1) and its several variational formulations; in the following sections we shall discuss the approximation of problem (1.1), and its *iterative solution via decomposition methods* of the same type as those in Chapter III.

2.2. Approximation of problem (2.5), (2.6) by finite element methods

In the following we consider a case where problem (2.5), or one of its equivalent formulations (2.6) or (2.8), is approximated by a method of conforming finite elements of the most frequently used type. The terminology and the notation used in relation to the method of finite elements are the same as in Chapter II; thus let

$\mathcal{T}_h$ be a triangulation of the *two-dimensional* domain Ω which for simplicity we shall assume to be *polygonal*. We then consider a space of conforming finite elements of degree k (≥ 1), namely

$$W_{kh} = \{v_h | v_h \in C^0(\bar{\Omega}),\ v_h|_K \in P_k(K)\ \ \forall K \in \mathcal{T}_h,\ v_h|_\Gamma = 0\}\ , \tag{2.9}$$

where, in (2.9), $P_k(K)$ denotes the space of polynomials of degree $\leq k$ on the element K. Next we consider the approximate problem

$$\begin{cases} \textit{Find} \quad u_h \in W_{kh} \quad \textit{such that} \\ a(u_h, v_h - u_h) + j(v_h) - j(u_h) \geq \langle f, v_h - u_h \rangle \quad \forall v_h \in W_{kh}\ ; \end{cases} \tag{2.10}$$

we have here used the formulation (2.6), but we could equally well have used the equivalent formulation (2.5) and (2.8). In general it is not possible to use (2.10) from a practical point of view; in reality it is not possible to obtain an exact analytical evaluation of the integrals defining $j(v_h)$, and in order to obtain a numerically tractable problem it is necessary to use *numerical integration* formulas in order to approximate $j(v_h)$. We thus consider in the element K of the triangulation $\mathcal{T}_h$, s *numerical integration points* x_{iK}, $i = 1,\ldots s$, each being assigned a *weight* ω_i, such that we have

$$\int_K f(x)ds \simeq \text{Area } (K) \sum_{i=1}^{s} \omega_i\, f(x_{iK}). \tag{2.11}$$

We then put

$$j_h(v_h) = \sum_{K \in \mathcal{T}_h} \text{Area } (K) \sum_{i=1}^{s} \omega_i\, \Phi(v_h(x_{iK})), \tag{2.12}$$

and we consider the approximate problem

$$\begin{cases} \textit{Find} \quad u_h \in W_{kh} \quad \textit{such that} \\ a(u_h, v_h - u_h) + j_h(v_h) - j_h(u_h) \geq \langle f, v_h - u_h \rangle \quad \forall v_h \in W_{kh}. \end{cases} \tag{2.13}$$

It is clear that the accuracy of our approximation will be influenced by the accuracy of the numerical integration formula, the choice of which will be guided by the properties of the function Φ. For reasons which will become apparent in the following section it

will be especially desirable to use quadrature formulas which are capable of integrating *exactly* the inner product $L^2(\Omega)$ in W_{kh}, hence

$$\left\{ \begin{array}{l} \displaystyle\int_\Omega u_h v_h \, dx = \sum_{K \in \mathfrak{C}_h} \text{Area}\,(K) \sum_{i=1}^{s} \omega_i u_h(x_{iK}) v_h(x_{iK}) \\ \forall u_h, v_h \in W_{kh}. \end{array} \right. \tag{2.14}$$

This requirement is not mandatory however; in particular it is not satisfied in CHAN-GLOWINSKI [1] and GLOWINSKI [1]. However, it will be essential that the set of points x_{iK} should be P_k-*unisolvent* on K, that is to say (see CIARLET [1]) that knowing the s values of a polynomial of degree $\leq k$, at the quadrature points, determines this polynomial *uniquely*.

Example 2.1: In CHAN-GLOWINSKI [1], GLOWINSKI [1, Chapter 4] an approximation of the problem by conforming finite elements of *degree one* on triangles is used; the integration points are the vertices of the triangles, assigned weights 1/3, and they correspond to the degrees of freedom of the approximation. This quadrature formula is of order one and does not satisfy condition (2.14); nonetheless it does lead to convergent approximations.

Example 2.2: We consider, still, an approximation by finite elements of *degree one* and we use as integration points the midpoints of the sides of the triangles, these being assigned weights 1/3. This quadrature formula is *exact* for polynomials of *degree two* and satisfies condition (2.14). It will be noted that the numerical integration points correspond to the degrees of freedom of an approximation by *nonconforming* finite elements of degree one (in fact that utilised in Chapter II for the solution of the Stokes and Navier-Stokes problems).

Example 2.3: We consider an approximation by finite elements of degree two. It is known (see for example LYNESS-JESPERSEN [1]) that it is possible to construct a quadrature formula on a triangle, which is exact for polynomials of degree 4, and which uses six integration points. We can moreover use these six points to define uniquely a polynomial of degree two.

Examples 2.2 and 2.3 show that condition (2.14) can be satisfied by the *triangular* finite elements most usually employed. The case of quadrilateral elements is even simpler because the corresponding quadrature formulas (deduced from the Gaussian formulas) are well known and easily obtained, at least on the reference rectangle. It is shown in CHAN-GLOWINSKI [1], GLOWINSKI [1, Chapter 4] that the approximation obtained by using the functional $j_h(v_h)$ defined by the quadrature formula in Example 2.1 leads to an approximate solution which converges, as $h \to 0$, to the exact solution (i.e. that of problem (2.5), (2.6)). The proof generalises without difficulty and it is even possible to obtain estimates of the approximation error (these bring in the nonlinearity of the problem). We shall not dwell any further on these points because our objective in the present work is rather to describe iterative methods of solution.

3. AUGMENTED LAGRANGIAN AND DECOMPOSITION OF THE PROBLEM (2.5), (2.6)

We shall assume in this section, although this is not in fact essential, that the bilinear form $a(.,.)$ is *symmetric*.

3.1 Construction of the augmented Lagrangian. (I) Continuous case.

In accordance with the general principles introduced in Chapter III, we shall first try to decompose problem (2.5), (2.6) by introducing a supplementary artificial variable. The coordination is then achieved by means of a Lagrange multiplier and a penalisation term. As we shall see at a later stage the choice of the decomposition is not unique, and the one which we shall consider in this section is the one which to us appears to be the simplest of the various possible choices.

Referring back to the notation of Chapter III, we put

$$V = H_0^1(\Omega),\ H = L^2(\Omega) \tag{3.1}$$

and we take as the operator B the *canonical injection* of V into H; we then define $G(.)$ and $F(.)$ by

(3.2) $$G(v) = \frac{1}{2}\, a(v,v) - \langle f,v\rangle$$

and

(3.3) $$F(q) = j(q),$$

respectively.

We then consider the *augmented Lagrangian* $\mathcal{L}_r : V \times H \times H \to \mathbb{R}$, defined by

(3.4) $$\mathcal{L}_r(v,q,\mu) = \frac{1}{2}\, a(v,v) + j(q) - \langle f,v\rangle + (\mu, v-q) + \frac{r}{2}\, |v-q|_o^2$$

(where $|\cdot|_o = \|\cdot\|_{L^2(\Omega)}$ and $(\cdot,\cdot) = (\cdot,\cdot)_{L^2(\Omega)}$).

It is clear that if $\{u,p,\lambda\}$ is a saddle-point on $V \times H \times H$ of the augmented Lagrangian $\mathcal{L}_r$, we then have $u = p$, where u is the solution of problem (2.5), (2.6), (2.8); we have obtained (3.4) by introducing the artificial variable q and by imposing the constraint $v - q = 0$ in $L^2(\Omega)$. Proving the existence of a Lagrange multiplier λ poses no difficulties in this particular case.

3.2 Construction of the augmented Lagrangian. (II) The discrete case

To approximate the augmented Lagrangian (3.4) via a finite element method, it is necessary to define an approximation of $L^2(\Omega)$ in order to approximate the functions q and μ appearing in (3.4); moreover we need to keep in mind that our objective is to obtain algorithms in which the treatment of the nonlinear part is purely local.

In regard to the approximation of $L^2(\Omega)$ a natural choice is to consider

(3.5) $$Q_{kh} = \{ q_h \mid q_h|_K \in P_k(K) \quad \forall K \in \mathcal{C}_h \},$$

that is, to use the same finite elements as in the construction of W_{kh}, but suppressing the matching conditions at the interfaces of these elements. We then approximate the functional $j(.)$ using (2.12) and we consider for $v_h \in W_{kh}$, $q_h \in Q_{kh}$, $\mu_h \in Q_{kh}$ the discrete augmented Lagrangian defined by

$$(3.6)\qquad \mathcal{L}_{rh}(v_h,q_h,\mu_h) = \frac{1}{2}\,a(v_h,v_h)+j_h(q_h) - \langle f,v_h\rangle + (\mu_h,v_h-q_h)+\frac{r}{2}|v_h-q_h|_o^2.$$

Now consider the case where the quadrature formula defining $j_h(.)$ (via 2.12)) satisfies the condition (2.14), i.e. allows the inner product in $L^2(\Omega)$ of two functions from Q_{kh} to be calculated exactly; in this case we then have

$$(3.7)\qquad \left\{\begin{array}{l} \mathcal{L}_{rh}(v_h,q_h,\mu_h) = \frac{1}{2}\,a(v_h,v_h)+j_h(q_h) - \langle f,v_h\rangle + \\ + \sum_{K\in\mathcal{T}_h} \mathrm{Area}\,(K) \sum_{i=1}^{s} \omega_i[\Phi(q_h(x_{iK}))+\mu_h(x_{iK})(v_h(x_{iK})-q_h(x_{iK})) + \\ + \frac{r}{2}\,|v_h(x_{iK})-q_h(x_{iK})|^2]. \end{array}\right.$$

If condition (2.14) is not satisfied, the Lagrangians defined by (3.6) and (3.7) are distinct. However if $v_h = q_h$, whenever these two functions *coincide* at the quadrature points, it is permissible to use (3.7) for the numerical solution of the approximate problem (2.13); this is the case in particular for the quadrature formula in Example 2.1.

Before describing the algorithms which will enable the saddle-points of the augmented Lagrangian (3.7) to be calculated, we will first introduce a certain amount of notation; it will also be useful to identify the optimality conditions of this saddle-point problem.

The space W_{kh} defined by (2.9) is a standard space for approximation of $H_o^1(\Omega)$ by the method of finite elements; we shall assume that functions from W_{kh} are characterised by N_h scalars, the *degrees of freedom*; for example in the case of *conforming* finite elements of degree *one* (resp. *two*) we shall use the values taken at the *vertices* (resp. at the *vertices* and the *midpoints of the sides*) of the triangles in the triangulation $\mathcal{T}_h$ (*not situated* on Γ) to completely define a function from W_{1h} (resp. W_{2h}). As regards Q_{kh}, the natural choice of degrees of freedom will be to consider the values taken by the functions from Q_{kh} at the *quadrature points;* we denote by P_h the total number of quadrature points (not situated on Γ). In order not to over-complicate the notation unnecessarily, we shall henceforth write $N = N_h$, $P = P_h$; in addition we shall denote by $\underset{\sim}{v},\underset{\sim}{q},\underset{\sim}{\mu}$ the vectors in $\mathbb{R}^N$, $\mathbb{R}^P$, $\mathbb{R}^P$, whose components correspond to the degrees of freedom associated with

respectively, $v_h \in W_{kh}$, $q_h \in Q_{kh}$, $\mu_h \in Q_{kh}$. We then consider the *linear* operator $\underset{\sim}{S}$ from $\mathbb{R}^N$ into $\mathbb{R}^P$ defined by

$$\underset{\sim\sim}{S}\underset{\sim}{v} = \{v_h(x_{iK})\} \quad \textit{for } 1 \le i \le s,\ K \in \mathcal{T}_h,\ x_{iK} \notin \Gamma, \tag{3.8}$$

i.e. we associate with $\underset{\sim}{v}$ the values taken by the function $v_h \in W_{kh}$ at the quadrature points not located on Γ.

We also define the *linear* operator $\underset{\sim}{M}$ from $\mathbb{R}^P$ into $\mathbb{R}^P$, associated with the approximate inner product on $L^2(\Omega)$, by

$$(\underset{\sim\sim}{M}\underset{\sim}{p},\underset{\sim}{q})_{\mathbb{R}^P} = \sum_{K \in \mathcal{T}_h} \text{Area}(K) \sum_{i=1}^{s} \omega_i p_h(x_{iK}) q_h(x_{iK}) \tag{3.9}$$

where, in (3.9), $p_h(x_{iK}) = q_h(x_{iK}) = 0$ if $x_{iK} \in \Gamma$; we have $\underset{\sim}{M} = \underset{\sim}{M}^t$.

If condition (2.14) is satisfied we have

$$(\underset{\sim\sim}{M}\underset{\sim}{p},\underset{\sim}{q})_{\mathbb{R}^P} = \int_\Omega p_h q_h \, dx \quad \forall p_h, q_h \in Q_{kh}\ . \tag{3.10}$$

Finally we denote by $\underset{\sim}{A}$ the linear operator from $\mathbb{R}^N$ into $\mathbb{R}^N$ defined by

$$(\underset{\sim\sim}{A}\underset{\sim}{u},\underset{\sim}{v})_{\mathbb{R}^N} = a(u_h,v_h) \quad \forall u_h, v_h \in W_{kh}. \tag{3.11}$$

With regard to the *nonlinearity*, we denote by $\phi(\underset{\sim}{q})$ (resp. $\Phi(\underset{\sim}{q})$) the vector obtained by applying ϕ (resp. Φ) to each of the components of $\underset{\sim}{q}$.

Taking account of the above notation, the augmented Lagrangian (3.7) can be written in the form

$$\begin{cases} \mathcal{L}_{rh}(\underset{\sim}{v},\underset{\sim}{q},\underset{\sim}{\mu}) = \frac{1}{2}(\underset{\sim\sim}{A}\underset{\sim}{v},\underset{\sim}{v})_{\mathbb{R}^N} - (\underset{\sim}{F},\underset{\sim}{v})_{\mathbb{R}^N} + (\underset{\sim}{M}\Phi(\underset{\sim}{q}),\underset{\sim}{1})_{\mathbb{R}^P} + \\ + (\underset{\sim\sim}{M}\underset{\sim}{\mu},\underset{\sim\sim}{S}\underset{\sim}{v}-\underset{\sim}{q})_{\mathbb{R}^P} + \frac{r}{2}(\underset{\sim}{M}(\underset{\sim\sim}{S}\underset{\sim}{v}-\underset{\sim}{q}),\underset{\sim\sim}{S}\underset{\sim}{v}-\underset{\sim}{q})_{\mathbb{R}^P} \end{cases} \tag{3.12}$$

where $\underset{\sim}{1} = \{1,\ldots 1\}$ $(\in \mathbb{R}^P)$.

The optimality conditions of our problem can then be written

$$\underset{\sim\sim}{A}\underset{\sim}{u}-\underset{\sim}{F} + \underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{\lambda} + r\underset{\sim}{S}^t\underset{\sim\sim\sim}{M}\underset{\sim}{S}\underset{\sim}{u} - r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{p} = \underset{\sim}{0}, \tag{3.13}$$

$$(3.13)_2 \qquad \phi(\underset{\sim}{p}) - \underset{\sim}{\lambda} - r\underset{\sim\sim}{S}\underset{\sim}{u} + r\underset{\sim}{p} = \underset{\sim}{0},$$

$$(3.13)_3 \qquad \underset{\sim\sim}{S}\underset{\sim}{u} = \underset{\sim}{p}.$$

In cases where Φ is not differentiable, $(3.13)_2$ would have to be replaced by the variational inequality

$$(3.14) \qquad \begin{cases} \underset{\sim}{p} \in \mathbf{R}^P, \textit{ and } \forall \underset{\sim}{q} \in \mathbf{R}^P \textit{ we have} \\ r(\underset{\sim\sim}{M}(\underset{\sim}{p}-\underset{\sim\sim}{S}\underset{\sim}{u}),\underset{\sim}{q}-\underset{\sim}{p})_{\mathbf{R}^P} - (\underset{\sim\sim}{M}\underset{\sim}{\lambda},\underset{\sim}{q}-\underset{\sim}{p}) + (\underset{\sim}{M}\Phi(\underset{\sim}{q}),\underset{\sim}{1}) - (\underset{\sim}{M}\Phi(\underset{\sim}{p}),\underset{\sim}{1}) \geq 0 \ ; \end{cases}$$

in practice this inequality has to be solved *pointwise* at each of the quadrature points, which in general creates no difficulties ($\underset{\sim}{\lambda}$ and $\underset{\sim}{u}$ being known).

We are now in a position to describe the algorithms for solving the approximate problem (2.13) (of the same type as those considered in Chapter III) associated with the augmented Lagrangian (3.12).

4. ALGORITHMS FOR SOLUTION OF THE APPROXIMATE PROBLEM (2.13). DISCUSSION

Bearing in mind the equivalence between the approximate problem (2.13) and the system (3.13), we shall be applying, for the solution of the latter, the algorithms of Chapter III. In the following discussion $\underset{\sim}{u}$, $\underset{\sim}{p}$, $\underset{\sim}{\lambda}$ again denote the vectors in $\mathbb{R}^N$, $\mathbb{R}^P$, $\mathbb{R}^P$ associated, respectively, with u_h, p_h, λ_h. Thus, having regard to (3.12), (3.13), we have the following algorithms:

ALG1:

(4.1) $\underset{\sim}{\lambda}^0 \in \mathbf{R}^P$, *chosen arbitrarily;*
then for $n \geq 0$, $\underset{\sim}{\lambda}^n \in \mathbf{R}^P$ *being known, determine* $\underset{\sim}{u}^n$, $\underset{\sim}{p}^n$ *and* $\underset{\sim}{\lambda}^{n+1}$ *by*

$$(4.2) \qquad \begin{cases} \phi(\underset{\sim}{p}^n) - \underset{\sim}{\lambda}^n - r\underset{\sim\sim}{S}\underset{\sim}{u}^n + r\underset{\sim}{p}^n = 0, \\ \underset{\sim\sim}{A}\underset{\sim}{u}^n + r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{S}\underset{\sim}{u}^n + \underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{\lambda}^n - r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{p}^n = \underset{\sim}{F}, \end{cases}$$

$$(4.3) \qquad \underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^n + \rho_n(\underset{\sim\sim}{S}\underset{\sim}{u}^n - \underset{\sim}{p}^n).$$

Remark 4.1: In the case described in CHAN-GLOWINSKI [1], [2], we have $\underset{\sim}{S} = \underset{\sim}{S}^t = \underset{\sim}{M} = \underset{\sim}{I}$. ■

Remark 4.2: It should be noted that by multiplying the first relation in (4.2) by $\underset{\sim}{S}^t\underset{\sim}{M}$, then by adding the result obtained to the second relation in (4.2), we obtain

$$\underset{\sim\sim}{A}\underset{\sim}{u}^n + \underset{\sim}{S}^t\underset{\sim}{M}\,\phi(\underset{\sim}{p}^n) = \underset{\sim}{F}. \tag{4.4}$$

The iterative *relaxation* method described in Chapter III is applicable for the solution of system (4.2); hence for $n \geq 0$ we have:

Given the vector $\underset{\sim}{\lambda}^n$, *choose* $\underset{\sim}{u}^{n,0}$ *arbitrarily, (for example* $\underset{\sim}{u}^{n,0} = \underset{\sim}{u}^{n-1}$), *then for* $k \geq 0$, $\underset{\sim}{u}^{n,k}$ *being known, solve successively*

$$\phi(\underset{\sim}{p}^{n,k+1}) - \underset{\sim}{\lambda}^n - r\underset{\sim}{S}\underset{\sim}{u}^{n,k} + r\underset{\sim}{p}^{n,k+1} = \underset{\sim}{0}, \tag{4.5}$$

$$\underset{\sim\sim}{A}\underset{\sim}{u}^{n,k+1} + r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{S}\underset{\sim}{u}^{n,k+1} - r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{p}^{n,k+1} + \underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{\lambda}^n = \underset{\sim}{F}. \tag{4.6}$$

The results of CEA-GLOWINSKI [1], GLOWINSKI [2, Chapter 5] apply to (4.5), (4.6) and, using the assumptions already made, we can prove the convergence of (4.5), (4.6), to $\{\underset{\sim}{u}^n, \underset{\sim}{p}^n\}$. In the implementation of (4.5), (4.6) two strategies can be used:

(1) Continue to iterate until the difference between two successive iterates is smaller than some threshold ε, chosen in advance, before proceeding to update $\underset{\sim}{\lambda}^n$ via (4.3); this corresponds exactly to ALG1 in so far as ε is sufficiently small for system (4.2) to be solved to high accuracy.

(2) Limit the number of relaxation iterations (4.5), (4.6) to a "small" number k_{max}, then update $\underset{\sim}{\lambda}^n$ via (4.3); the limiting case $k_{max} = 1$ obviously gives the algorithm ALG2 described below.

ALG2:

$$\underset{\sim}{u}^0, \underset{\sim}{\lambda}^1 \text{ chosen arbitrarily;} \tag{4.7}$$

for $n \geq 1$, $\underset{\sim}{u}^{n-1}$ *and* $\underset{\sim}{\lambda}^n$ *being known, determine successively* $\underset{\sim}{p}^n$, $\underset{\sim}{u}^n$ *and* $\underset{\sim}{\lambda}^{n+1}$ *by*

$$r\underset{\sim}{p}^n + \phi(\underset{\sim}{p}^n) = r\underset{\sim}{S}\underset{\sim}{u}^{n-1} + \underset{\sim}{\lambda}^n, \tag{4.8}$$

$$(\underset{\sim}{A} + r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{S})\underset{\sim}{u}^n = r\underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{p}^n - \underset{\sim}{S}^t\underset{\sim\sim}{M}\underset{\sim}{\lambda}^n + \underset{\sim}{F}, \tag{4.9}$$

$$\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^{n} + \rho_n(S\underset{\sim}{u}^n - \underset{\sim}{p}^n). \tag{4.10}$$

This latter algorithm is worthy of further attention because it contains as particular cases a number of the classical *alternating direction* methods. In fact combining (4.9) and (4.10) we obtain the relation

$$S^t M\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{F} - A\underset{\sim}{u}^n + (\rho_n - r)S^t M(S\underset{\sim}{u}^n - \underset{\sim}{p}^n). \tag{4.11}$$

The relation (4.11) enables us to eliminate $\underset{\sim}{\lambda}^n$ from (4.8) and (4.9); first we consider the case $\rho^n = \rho = r$; we then obtain

$$r(S^t M\underset{\sim}{p}^n - S^t MS\underset{\sim}{u}^{n-1}) + A\underset{\sim}{u}^{n-1} + S^t M\phi(\underset{\sim}{p}^n) = \underset{\sim}{F}, \tag{4.12}$$

$$r(S^t MS\underset{\sim}{u}^n - S^t MS\underset{\sim}{u}^{n-1}) + S^t M\phi(\underset{\sim}{p}^n) + A\underset{\sim}{u}^n = \underset{\sim}{F}. \tag{4.13}$$

Suppose we put $\underset{\sim}{p}^n = \underset{\sim}{u}^{n-\frac{1}{2}}$; if $S = M = I$ as is the case in CHAN-GLOWINSKI [1], [2] (where the quadrature formula of Example 2.1 is used) we obtain from (4.12), (4.13) an *alternating direction* method of the *Douglas-Rachford* type (see DOUGLAS-RACHFORD [1]).

In the general case ($\rho_n \neq r$) (4.12) would be replaced by

$$r(S^t M\underset{\sim}{p}^n - S^t M\hat{\underset{\sim}{p}}^{n-1}) + A\underset{\sim}{u}^{n-1} + S^t M\phi(\underset{\sim}{p}^n) = \underset{\sim}{F}, \tag{4.14}$$

where we have put

$$\hat{\underset{\sim}{p}}^{n-1} = S\underset{\sim}{u}^{n-1} + \frac{(r-\rho_n)}{\rho_n}\underset{\sim}{p}^{n-1}. \tag{4.15}$$

In practice it is easier to work with (4.7)-(4.10) than (4.12),(4.13). It is also interesting to observe that we can derive an alternating direction method of the *Peaceman-Rachford* type (see PEACEMAN-RACHFORD [1]) through a variant of ALG2; indeed, consider the algorithm

ALG3:

(4.16) $\underset{\sim}{u}^0$, $\underset{\sim}{\lambda}^1$ *chosen arbitrarily;*

for $n \geq 1$, $\underset{\sim}{u}^{n-1}$ *and* $\underset{\sim}{\lambda}^n$ *being known, determine successively* $\underset{\sim}{p}^n$, $\underset{\sim}{\lambda}^{n+\frac{1}{2}}$, $\underset{\sim}{u}^n$, $\underset{\sim}{\lambda}^{n+1}$ *by*

$$r\underset{\sim}{p}^n + \phi(\underset{\sim}{p}^n) = rS\underset{\sim}{u}^{n-1} + \underset{\sim}{\lambda}^n, \tag{4.17}$$

$$\underset{\sim}{\lambda}^{n+1/2} = \underset{\sim}{\lambda}^{n} + \rho(\underset{\sim\sim}{S}\underset{\sim}{u}^{n-1} - \underset{\sim}{p}^{n}), \tag{4.18}$$

$$(\underset{\sim\sim}{A} + r\underset{\sim\sim}{S}^{t}\underset{\sim\sim}{M}\underset{\sim\sim}{S})\underset{\sim}{u}^{n} = r\underset{\sim\sim}{S}^{t}\underset{\sim\sim}{M}\underset{\sim}{p}^{n} - \underset{\sim\sim}{S}^{t}\underset{\sim\sim}{M}\,\underset{\sim}{\lambda}^{n+1/2} + \underset{\sim}{F}, \tag{4.19}$$

$$\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^{n+1/2} + \rho(\underset{\sim\sim}{S}\underset{\sim}{u}^{n} - \underset{\sim}{p}^{n}). \tag{4.20}$$

We thus carry out a first update of $\underset{\sim}{\lambda}^{n}$ (by (4.18)) following the solution for $\underset{\sim}{p}^{n}$, then a second (by (4.20)) following the solution for $\underset{\sim}{u}^{n}$; it will also be noted that in ALG3 the $\underset{\sim}{p}^{n}$ and the $\underset{\sim}{u}^{n}$ play roles which are symmetric, which is not the case in ALG2. From the point of view of the search for a saddle-point and in relation to ALG1, this algorithm is "less implicit" than ALG2, and in fact for "stiff" problems ALG3 is usually less robust than ALG2. If $\rho = r$ we deduce from (4.17)-(4.20)

$$\underset{\sim}{\lambda}^{n+1/2} = \phi(\underset{\sim}{p}^{n}) \quad \textit{and} \quad \underset{\sim}{S}^{t}\underset{\sim}{M}\,\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{F} - \underset{\sim\sim}{A}\underset{\sim}{u}^{n}.$$

We then obtain

$$r(\underset{\sim}{S}^{t}\underset{\sim\sim}{M}\underset{\sim}{p}^{n} - \underset{\sim}{S}^{t}\underset{\sim\sim\sim}{M}\underset{\sim\sim}{S}\underset{\sim}{u}^{n-1}) + \underset{\sim\sim}{A}\underset{\sim}{u}^{n-1} + \underset{\sim}{S}^{t}\underset{\sim}{M}\phi(\underset{\sim}{p}^{n}) = \underset{\sim}{F}, \tag{4.21}$$

$$r(\underset{\sim}{S}^{t}\underset{\sim\sim\sim}{M}\underset{\sim\sim}{S}\underset{\sim}{u}^{n} - \underset{\sim}{S}^{t}\underset{\sim\sim}{M}\underset{\sim}{p}^{n}) + \underset{\sim}{S}^{t}\underset{\sim}{M}\phi(\underset{\sim}{p}^{n}) + \underset{\sim\sim}{A}\underset{\sim}{u}^{n} = \underset{\sim}{F}\ ; \tag{4.22}$$

putting $\underset{\sim}{p}^{n} = \underset{\sim}{u}^{n-\frac{1}{2}}$, we indeed obtain the method of *alternating directions* of *Peaceman-Rachford*.

It is interesting to note that we have been able to locate some alternating direction methods within a more general framework, namely the several possible variants of the algorithm ALG1.

With regard to the convergence of the algorithms, the results of Chapter III can be applied without difficulty. We have in fact to distinguish two cases, depending on whether or not the quadrature formula used integrates exactly the inner product on $L^2(\Omega)$ restricted to the space Q_{kh}. If it does, the results of Chapter III (Theorem 4.1 and Remark 4.1) are applicable in their entirety. Otherwise they apply to the system in finite dimensions; however the equivalence of the norms plays no overriding role in the proof of convergence and we can expect an overall rate of convergence which is almost independent of the discretisation.

5. NUMERICAL EXPERIMENTS

5.1 Formulation of a model problem. General notes.

In order to illustrate the results of the preceding sections, we shall now summarise the numerical experiments of CHAN-GLOWINSKI [1] relating to the convergence of the algorithms ALG1 and ALG2 applied to the solution of a particular problem (1.1).

We have therefore considered the model problem

$$(5.1)\qquad \begin{cases} -\Delta u + \phi(u) = f \ \ in \ \ \Omega =]0,1[\times]0,1[, \\ u = 0 \ \ on \ \ \Gamma, \end{cases}$$

the function $\phi(.)$ being defined, for $\ell > 0$, by

$$(5.2)\qquad \phi(t) = \mathrm{sgn}(t)|t|^{\ell} = t|t|^{\ell-1}.$$

We have shown in Figure 5.1 the form of the function ϕ for three values of ℓ. In the majority of our tests we took $\ell = 0.1$, which leads to a problem which is quite difficult numerically; we in fact have $\phi'(0) = +\infty$, so that we can expect some difficulties in the regions where $u(x_1,x_2) = 0$. We took (for $x = \{x_1,x_2\}$)

$$(5.3)\qquad \begin{cases} u(x) = \sin 2\pi x_1 \ \sin 2\pi x_2, \\ f = 8\pi^2 u + |u|^{\ell-1}u, \end{cases}$$

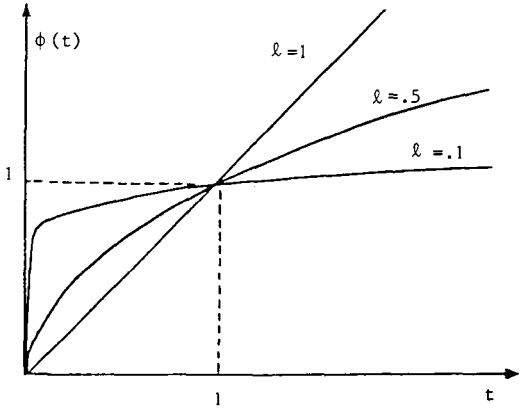


Figure 5.1

and we attempted to solve a discretised version of problem (5.1), (5.2).

Application of the algorithms described in Section 4 leads to having to solve one-dimensional problems of the form

$$r\xi + \phi(\xi) = b, \tag{5.4}$$

where b is given. Note that the singularity of ϕ' at 0 obliges us to take certain precautions during the numerical solution of (5.4), even for such an elementary problem.

We shall first make a number of general comments on the behaviour of the various algorithms tested; then we shall go into rather more detail on certain points which seem to us to possess some importance in relation to the algorithms.

5.2 Comments on the implementation and the convergence of ALG1 and ALG2

As we have seen in Section 4, algorithm ALG2 is a special case of ALG1 in which the number of internal iterations has been limited to *one*. In practice a greater degree of generality would be offered by incorporating ALG1 in a program allowing the number of internal iterations to be limited by a termination test based either on the decrease of some residual, or on a maximum number of internal iterations. As regards the variant ALG3, this would involve only a minor modification.

Concerning the speed of convergence, the main difficulties encountered related to those regions where the function ϕ' is singular, that is, where $u(x) = 0$; it is in fact observed that the asymptotic rate of convergence of ALG2 is very slow at these points, and this fact is clearly illustrated in Figure 5.2 which corresponds to a numerical test in which we have initialised ALG2 first with $\underset{\sim}{u}^0 = \underset{\sim}{0}$, $\underset{\sim}{\lambda}^1 = \underset{\sim}{0}$, and then $\underset{\sim}{u}^0 = \underset{\sim}{1}0$ (= {10,...10}), $\underset{\sim}{\lambda}^1 = \underset{\sim}{0}$. In the first case the solution has already been attained at the singular points because it can easily be seen that the algorithm will leave the values unchanged at these points; in the second case the solution has to reach the value zero at the above points. Figure 5.2 shows that for $\underset{\sim}{u}^0 = \underset{\sim}{0}$ we have a very fast *linear* convergence; in the second case (i.e. $\underset{\sim}{u}^0 = \underset{\sim}{1}0$) we have a *sublinear*

convergence. A close examination of the results reveals that the slowness of the convergence is localised at the points where $u(x) = 0$, the other values having already been obtained to a reasonable accuracy. It is therefore clear that the choice of good initial values can lead to a very significant improvement; however it is still very important to make the algorithm more robust because in practice it is unlikely that it will be possible to find initial values which will enable the difficulty associated with the singular points of ϕ' to be circumvented.

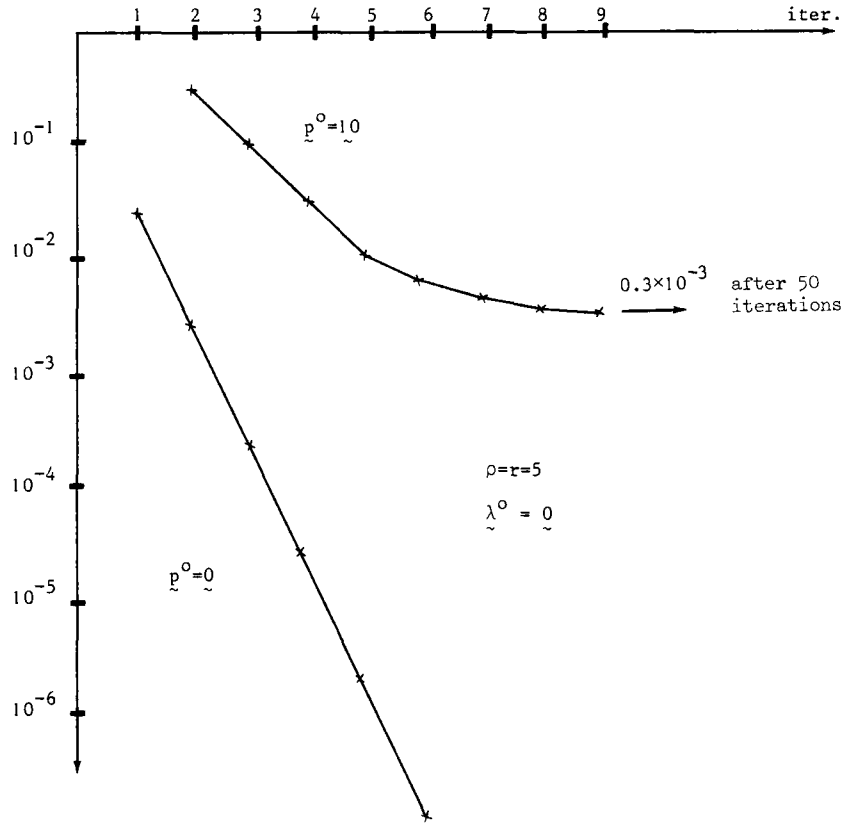


Figure 5.2

Convergence of ALG2

(the vertical axis represents the L^2 error in $\underset{\sim}{u}^n$)

In practice two remedies are possible:

(1) Increase r during the course of the calculation ; this necessitates the refactorisation of the matrix $\underset{\sim}{A}_r = \underset{\sim}{A} + r\underset{\sim}{S}^t\underset{\sim}{M}\underset{\sim}{S}$, which may be expensive.

(2) Carry out further iterations within the internal relaxation loop (4.5), (4.6); this in fact means using ALG1 with a more or less complete internal solution.

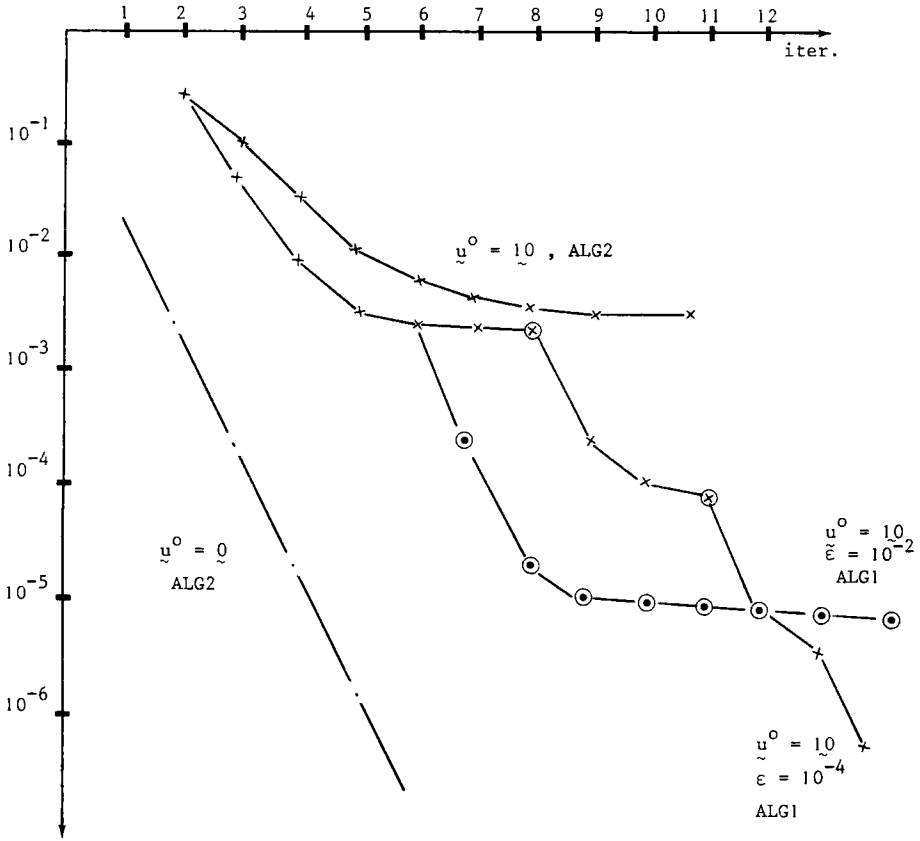


Figure 5.3

Figure 5.3 illustrates the results obtained by using the second strategy; in the corresponding numerical tests we stopped the internal iterations (4.5), (4.6) when the difference between two successive iterates was smaller in norm than ε. Figure 5.3 shows the convergence results obtained with $\varepsilon = 10^{-2}$ and 10^{-4}; the horizontal axis shows the total number of internal iterations of (4.5), (4.6) needed to attain the accuracy shown on the vertical axis (this being the L^2 norm of the error corresponding to u^n). The circled points indicate that there has been an update of $\underset{\sim}{\lambda}^n$, via (4.3), at that iteration. By way of comparison we have replotted on this figure the results for ALG2 already presented in Figure 5.2. For $\varepsilon = 10^{-2}$, algorithm ALG1 degenerates rapidly into ALG2, since for n_o of the order of ten to forty or so we have convergence of (4.5), (4.6) *in a single iteration* as soon as $n \geq n_o$; it can be seen that convergence is attained rapidly at points for which $u_h(x) \neq 0$.

For $\varepsilon = 10^{-4}$ a larger number of internal iterations (4.5), (4.6) was carried out before updating $\underset{\sim}{\lambda}_n$ via (4.3); *the mean rate of convergence* is, however, greatly improved in comparison to $\varepsilon = 10^{-2}$, and is in fact close to that observed for ALG2 with $\underset{\sim}{u}_o = \underset{\sim}{0}$; thus we have in large part eliminated the effect of the initial conditions.

Remark 5.1: In view of the above numerical tests the algorithm ALG1 is apparently more robust than ALG2 which is in fact equivalent to an alternating direction method. We can therefore consider ALG1, obtained by introducing an augmented Lagrangian, as a method which enables the robustness and the speed of convergence of these alternating direction methods to be increased.

5.3 Discussion on the choice of the parameters r and ρ.

We shall conclude Section 5 by discussing the choice of parameters r and ρ. As far as ρ is concerned we have systematically used $\rho = r$ all the time; this value is always a very good one even if it is not absolutely optimal. The numerical tests showed that in the case of ALG2 a value of ρ rather larger than r ($\rho \simeq 1.1\ r$) accelerated the convergence very slightly; in the absence of a precise method for determining the optimal value of ρ,

however, we recommend the choice $\rho = r$. With regard to the choice of r, Figure 5.4 indicates the number of iterations needed for convergence of ALG1 for different values of r. An optimum was recorded near $r = 5$, but this optimum is not clearly defined and the choice of r is not critical within a rather wide interval; this is due to an effect of partial cancellation between two phenomena with opposite actions. What happens is that increasing r increases the speed of convergence of $\underset{\sim}{\lambda}^n$ in ALG1, but decreases that of the internal iterations (4.5), (4.6); the combined effect is highly complex but the result is an algorithm which is not very sensitive to the choice of r.

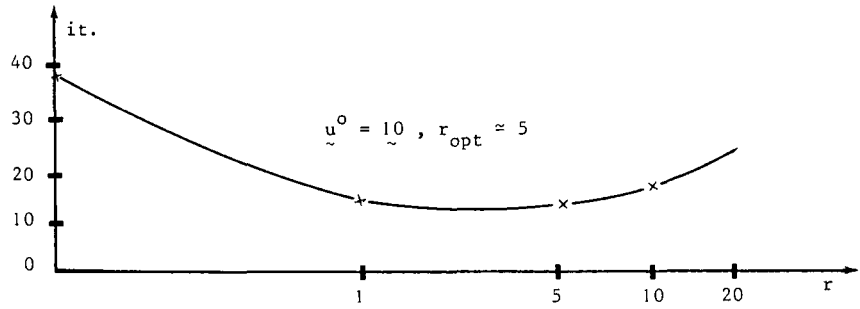


Figure 5.4

Effect of the choice of r on the convergence of ALG1

The reader can refer to CHAN-GLOWINSKI [1] for further details relating to the convergence of ALG1 and ALG2 applied to the solution of problem (5.1), (5.2), together with a number of comparisons with other iterative methods.

6. SOME REMARKS ON HYBRID METHODS

In this section we consider a variant of the preceding methods which may be useful for certain problems and which is linked with the *hybrid-primal* finite element methods (*hybrides primaux* in the original French terminology of THOMAS [1]). The starting point will once again be the methodology of Chapter III, the generality of which will again be further illustrated here.

In *hybrid* finite element methods it is standard practice to *decouple* a boundary-value problem into a family of local problems

defined over each element of the triangulation. The recoupling of these local problems is generally carried out by means of a Lagrange multiplier associated with the matching constraint at the interface of the elements. In the light of what we have presented earlier, the advantage of such a method is apparent in the case of nonlinear problems; in fact the solution at the element level involves a problem having a small number of variables, which is much easier than dealing with a global nonlinear problem. However two obstacles arise, in regard to the algorithms, for the efficient exploitation of hybrid methods:

(i) The local problems are often ill posed and cannot be solved independently of one another.

(ii) Coordination algorithms based on the convergence of the multipliers are slow, and their efficiency deteriorates rapidly as the number of elements increases.

We shall now show that the use of a suitably defined augmented Lagrangian enables the above difficulties to be overcome, and allows efficient algorithms to be constructed.

We shall thus consider a model problem of the form (2.5), namely:

$$(6.1)\qquad \begin{cases} \textit{Find}\quad u \in H_o^1(\Omega) \quad \textit{such that}\quad \phi(u) \in L^1(\Omega) \cap H^{-1}(\Omega), \quad \textit{satisfying} \\ a(u,v) + \langle \phi(u), v \rangle = \langle f, v \rangle \quad \forall v \in H_o^1(\Omega). \end{cases}$$

When $a(.,.)$ is *symmetric*, (6.1) is equivalent to the minimisation on $H_o^1(\Omega)$ of the functional defined by $(2.8)_2$, i.e.

$$(6.2)\qquad J(v) = \frac{1}{2}\, a(v,v) + j(v) - \langle f, v \rangle .$$

Thus, let $\mathcal{T}_h$ be a triangulation of Ω ; the principle of hybrid methods consists of defining the problem on a larger space, considering membership of $H_o^1(\Omega)$ as a *constraint*. To do this we put

$$(6.3)\qquad H = \prod_{K \in \mathcal{T}_h} H^1(K),$$

this space being equipped with the product topology, i.e. that associated with the inner product

$$(6.4)\qquad (p,q)_H = \sum_{K \in \mathcal{T}_h} \int_K (pq + \nabla p \cdot \nabla q)\, dx.$$

Suppose that $f \in L^2(\Omega)$; the functional (6.2) then extends naturally onto H, and $V = H_o^1(\Omega)$ is a *closed* subspace of H (we retain the notation $a(.,.)$ and $j(.)$). The standard approach would consist of introducing a Lagrange multiplier on the interfaces of the triangulation $\mathcal{T}_h$ in order to enforce matching. This approach, however, does not lend itself well to the use of an augmented Lagrangian because the natural penalisation term utilises the norm on $H^{\frac{1}{2}}(\partial K)$, the manipulation of which is somewhat awkward. It is in fact necessary to use a *lift* onto the element to obtain a calculable expression, and this is what we shall do indirectly in the work which follows.

We shall utilise the methodology of Chapter III with $G \equiv 0$. Thus for $v \in V$, $q \in H$, $\mu \in H$, we define the augmented Lagrangian

$$\mathcal{L}_r(v,q,\mu) = \frac{1}{2}\, a(q,q)+j(q)-(f,q)+(\mu,v-q)_H + \frac{r}{2}\,|v-q|_H^2 \,. \tag{6.5}$$

Remark 6.1: For $r = 0$, the Lagrangian (6.5) is the standard Lagrangian of hybrid methods with a multiplier on the interfaces of the triangulation. Consider, in fact, a particular finite element K and $g \in H^{-\frac{1}{2}}(\partial K)$; we can solve

$$\begin{cases} -\Delta\mu + \mu = 0 \quad \textit{in } K, \\ \dfrac{\partial \mu}{\partial n} = g \quad \textit{on } \partial K. \end{cases} \tag{6.6}$$

Suppose that g takes on a single value along each of the interfaces (apart from a sign change to take account of the orientation of the normal); it can easily be seen that $(\mu,v)_H = 0$ and that the term $(\mu,q)_H$ reduces to boundary terms of the form $\int_{\partial K} gq \, d\gamma$. ■

It is of interest to state the optimality conditions of the problem: in *variational* form, this amounts to finding $\{u,p,\lambda\} \in V \times H \times H$ such that

$$a(p,q) + (\phi(p),q) - (f,q) - (\lambda,q)_H + r(p-u,q)_H = 0 \quad \forall q \in H, \tag{6.7}$$

$$r(u-p,v)_H + (\lambda,v)_H = 0 \quad \forall v \in V \ (= H_o^1(\Omega)), \tag{6.8}$$

$$(u-p,\mu)_H = 0 \quad \forall \mu \in H. \tag{6.9}$$

For given λ and u, the problems (6.7) are decoupled and are solved element by element; the recoupling is achieved via the multiplier and via the linear problem (6.8) which is of the form

$$(6.10)\qquad \begin{cases} -\Delta u + u = F(p,\lambda) \quad in\ \Omega, \\ u|_{\Gamma} = 0, \end{cases}$$

where $F(q,\lambda)$ denotes a right-hand side depending on p and on λ. It is very simple to adapt algorithms ALG1 and ALG2 to this case; note that the updating of λ^n is carried out using

$$(6.11)\qquad (\lambda^{n+1}-\lambda^n,\mu)_H = \rho_n(u^n-p^n,\mu)_H \quad \forall \mu \in H.$$

Taking into account (6.4) and putting $A\mu = -\Delta\mu + \mu$, we then have

$$(6.12)\qquad \begin{cases} A\lambda^{n+1} = A\lambda^n + \rho_n A(u^n-p^n) \quad in\ K, \\ \dfrac{\partial \lambda^{n+1}}{\partial n_K} = \dfrac{\partial \lambda^n}{\partial n_K} + \rho_n \dfrac{\partial}{\partial n_K}(u^n-p^n), \end{cases}$$

where n_K denotes the outward normal to K. The Neumann problems (6.12) are in fact local and are solved element by element.

Remark 6.2: In contrast to the preceding sections, the decoupling method no longer necessitates splitting the operator into a linear part and a nonlinear part; it is therefore very general and can be extended to cases more complex than those considered above. The solution of the nonlinear problem (6.7) may then become more complicated, and the advantage of the hybrid finite element methods then accrues essentially from the fact that the nonlinear problems contain only a small number of variables.

Remark 6.3: The formulation presented above encompasses the usual hybrid methods. In regard to the discretisation of the problem by finite elements, however, it is more general in that it allows the use of conforming finite elements for the approximation of u. In fact the awkward problem of the discretisation of the multiplier on the interfaces is no longer present. ■

On the question of the convergence of algorithms of the type ALG1, ALG2 applied to the solution of (6.1), via the augmented Lagrangian (6.5), the results of Chapter III are applicable in

their entirety. It should also be noted that the recoupling through the linear problem (6.10), which is global on Ω, enables us to anticipate convergence properties more or less independent of the number of elements used.

7. GENERAL DISCUSSION ON CHAPTER IV

The algorithms developed in this chapter enable us to solve efficiently a large number of nonlinear problems. In bringing the nonlinearity down to a local level we simplify the problem considerably, and the recoupling through the solution of linear systems, with matrices which are fixed during the iterative process, is a considerable advantage. Finally, these methods are well adapted to *Parallel Computation* and should prove particularly efficient for future generations of computers.

The fact that we have been able to obtain as particular cases a number of alternating direction methods, which have proved their efficiency elsewhere, is thus a reliable guide for implementation. In Chapter IX we shall give more information concerning alternating direction methods and the links which exist between these methods and the augmented Lagrangian methods (see also LIONS-MERCIER [1]).

CHAPTER V

APPLICATION TO THE SOLUTION OF STRONGLY NONLINEAR SECOND-ORDER BOUNDARY-VALUE PROBLEMS

M. Fortin, R. Glowinski, A. Marrocco

1. INTRODUCTION

The aim of this chapter is to give a relatively detailed account of a number of applications of Augmented-Lagrangian methods to the solution of nonlinear second-order boundary-value problems, in which the nonlinearity relates to the gradient of the unknown function. The general setting is the same as that in Chapter III, and we shall show how the various problems we describe can be fitted into this framework. We shall also discuss the approximation of these problems by means of finite-element methods and the consequences with regard to the iterative solution algorithms. The results of various numerical tests will serve to illustrate the behaviour of the algorithms used, and will demonstrate their efficiency.

We shall show, in a relatively general context, that the framework presented permits easy implementation of the so-called '*interior penalty*' finite-element methods.[1]

2. GENERAL FRAMEWORK OF THE PROBLEMS IN CHAPTER V

This second section aims to provide a common general framework for the examples which follow; the reader should therefore not be surprised at the formal character of this framework, which is justified by our desire to unify a number of extremely diverse problems. We ought also to point out that in this chapter we shall be giving much more attention to the algorithmic aspects than to the precise mathematical formulation or to the convergence of the solutions of the approximate problems.

---

[1] In the sense of DOUGLAS-DUPONT [1], WHEELER [1], ...

In this chapter, we shall thus be considering the problem of the minimisation, in appropriate functional spaces, of functionals J of the form

$$J(v) = F(\nabla v) + G(v). \tag{2.1}$$

In (2.1), v denotes a function defined on a domain Ω in $\mathbb{R}^N$, and ∇v denotes its gradient. The term $F(\nabla v)$ will be written more specifically in the form[2]

$$F(\nabla v) = \int_\Omega \Phi(x,|\nabla v|)dx. \tag{2.2}$$

In many cases the function Φ will not depend explicitly on x, and in general it will be *convex* with respect to $|\nabla v|$. The solutio of such a problem clearly lends itself very well to the introduction of an *augmented Lagrangian* of the same type as those considered in Chapter III. Thus, using the notation of Chapter III, the following form suggests itself:

$$\mathcal{L}_r(v,q,\mu) = \int_\Omega \Phi(x,|q|)dx + G(v) + (\mu,\nabla v-q)_H + \frac{r}{2}|\nabla v-q|_H^2 . \tag{2.3}$$

In our examples, we shall have $H = (L^2(\Omega))^N$ for the continuous problem, with a discrete version of this space for the approximate problems. It should be pointed out immediately that in some cases it will be possible to give a rigorous definition of $\mathcal{L}_r$ only in the case of the approximate problems; in particular, this will be the case for problem (3.2) of Section 3, if $1 < s < 2$, and for the *minimal surfaces* problem of Section 6.3.

In the numerical tests which will be presented, the discretisation will be performed by means of conforming finite elements of degree one, on triangles. The discretisation of the variable q which represents the gradient will thus be particularly simple since in this case the components of q will be constant over each element. Just as in Chapter IV, the discussion can be generalised to more complicated cases: quadratic finite elements and piecewise-linear gradients; with these, it will once again be necessary to make judicious use of numerical integration techniques, as we did in Chapter IV. Our approach will be centred, primarily, on the use of algorithms ALG1 and ALG2 of Chapter III, and the results of the numerical tests will serve to illustrate their convergence properties.

---

[2] $|q| = \left(\sum_{i=1}^{N} q_i^2\right)^{\frac{1}{2}}$ if $q = \{q_i\}_{i=1}^N$.

3. A CLASS OF NONLINEAR DIRICHLET PROBLEMS

3.1 Formulation of the problems. Augmented Lagrangians

3.1.1 *The continuous case*

In this section we shall be considering the *monotone nonlinear operator* A, defined (with $1 < s < +\infty$) by

$$(3.1) \qquad Av = -\nabla\cdot(|\nabla v|^{s-2})\nabla v.$$

This operator appears in certain mathematical models describing the mechanical deformation of ice (see for example PELISSIER [1] and the associated bibliography). The solution, in a domain Ω with boundary Γ, of the nonlinear Dirichlet problem

$$(3.2) \qquad \begin{cases} Au = f \ \textit{in} \ \Omega, \\ u = 0 \ \textit{on} \ \Gamma, \end{cases}$$

considered earlier in Chapter III, Section 6.3, can be reduced to the solution of the following problem in the *Calculus of Variations* :

$$(3.3)_1 \qquad \begin{cases} \textit{Find} \quad u \in V \\ J(u) \le J(v) \quad \forall v \in V, \end{cases}$$

where

$$(3.3)_2 \qquad V = W_o^{1,s}(\Omega),$$

$$(3.3)_3 \qquad J(v) = \frac{1}{s}\int_\Omega |\nabla v|^s \, dx - \int_\Omega fv \, dx \ ;$$

the space $W_o^{1,s}(\Omega)$ (described earlier in Chapter III, Section 6.3) is, for $1 < s < +\infty$, a reflexive Banach space when it is equipped with the norm[3]

$$(3.4) \qquad |v|_s = \left(\int_\Omega |\nabla v|^s \, dx\right)^{1/s}.$$

We therefore have here, using the notation of (2.2), $\Phi(x,z) = \frac{1}{s} z^s$.

---

[3] $|\cdot|_s$ is a norm on $W_o^{1,s}(\Omega)$ if Ω is *bounded* in at least one direction in $\mathbb{R}^N$.

3.1.2 *The approximate problems*

Except for $s = 2$, the space V is not a Hilbert space and we can not therefore apply the results of Chapter III directly to problem (3.2),(3.3); fortunately, however, this can be done for the *approximations* of this problem.

Suppose we have a triangulation[4] $\mathcal{T}_h$ of Ω; we consider on $\mathcal{T}_h$ an approximation of V using conforming finite elements of degree *one* or *two*. More precisely, for $k = 1$ or 2, we define

$$V_{kh} = \{v_h | v_h \in C^0(\bar{\Omega}),\ v_h|_K \in P_k(K)\ \ \forall K \in \mathcal{T}_h,\ v_h|_\Gamma = 0\}\ , \tag{3.5}$$

where $P_k(K)$ is the space of polynomials of degree $\leq k$ on K. Likewise, we define

$$L_{kh} = \{q_h | q_h = \{q_{1h}, q_{2h}\}\ ,\ q_{ih}|_K \in P_{k-1}(K)\ \ \forall i=1,2,\ \ \forall K \in \mathcal{T}_h\}\ ; \tag{3.6}$$

clearly, we have

$$\nabla v_h \in L_{kh} \quad \forall v_h \in V_{kh}. \tag{3.7}$$

The discretised version of (3.2), (3.3) thus consists of minimising $J(\cdot)$ over V_{kh}, leading to the augmented Lagrangian[5]

$$\begin{cases} \mathcal{L}_r(v_h, q_h, \mu_h) = \dfrac{1}{s}\displaystyle\int_\Omega |q_h|^s \, dx - \int_\Omega f v_h \, dx + \frac{r}{2}\int_\Omega |\nabla v_h - q_h|^2 \, dx \\ \qquad\qquad + \displaystyle\int_\Omega \mu_h \cdot (\nabla v_h - q_h) \, dx \end{cases} \tag{3.8}$$

defined on $V_{kh} \times L_{kh} \times L_{kh}$.

For $k = 1$ the calculation of $\mathcal{L}_r$ does not present any problems since q_h is constant over each triangle of $\mathcal{T}_h$. If $k = 2$, we shall use a quadrature formula to evaluate $\int_\Omega |q_h|^s \, dx$; as in Chapter IV, we shall thus consider, in the element K, ℓ *numerical integration* points x_{iK}, $i = 1, \ldots, \ell$, each being assigned a weight

---

[4] This assumes $\Omega \subset \mathbb{R}^2$ and *bounded*; however, the following discussion may readily be extended to the case $\Omega \subset \mathbb{R}^N$, $N \geq 3$. We shall also assume that Ω is polygonal.

[5] $q \cdot q' = \sum_{i=1}^{2} q_i q_i'$ if $q = \{q_1, q_2\}$, $q' = \{q_1', q_2'\}$.

ω_i such that we have

$$\int_K f(x)\,dx \simeq \text{Area}\,(K) \sum_{i=1}^{\ell} \omega_i\, f(x_{iK}). \tag{3.9}$$

For $k = 2$, the most appropriate choice would be to take the points x_{iK} as the midpoints of the sides of the triangle K (so that $\ell = 3$), each being assigned a weight $1/3$ (so that $\omega_i = 1/3 \quad \forall i = 1,2,3$); this quadrature formula is in fact exact for polynomials of degree 2, and this allows us to calculate the terms $\frac{r}{2}\int_\Omega |\nabla v_h - q_h|^2\,dx$ and $\int_\Omega \mu_h \cdot (\nabla v_h - q_h)\,dx$ *exactly*. Furthermore, again if $k = 2$, knowing the value of a piecewise-linear function at the midpoints of the sides of a triangle fixes this function uniquely; it then follows that the values of q_{1h} and q_{2h} at the midpoints of the sides of $\mathcal{C}_h$ can be used as degrees of freedom relative to q_h (it would be equally straightforward, if we so desired, to enforce matching of certain components of q_h at the midpoints of the sides). Proceeding as in Chapter IV, Section 3, this leads to the discrete augmented Lagrangian

$$\left\{\begin{aligned} \mathcal{L}_{rh}(v_h,q_h,\mu_h) = \sum_{K\in\mathcal{C}_h} \text{Area}(K) \sum_{i=1}^{\ell} \omega_i \Big[&\frac{1}{s}\,|q_h(x_{iK})|^s + \\ &\frac{r}{2}\,|\nabla v_h(x_{iK}) - q_h(x_{iK})|^2 + \mu_h(x_{iK})\cdot\Big(\nabla v_h(x_{iK}) - q_h(x_{iK})\Big) - f(x_{iK})v_h(x_{iK})\Big]. \end{aligned}\right. \tag{3.10}$$

This Lagrangian is defined on $V_{kh} \times L_{kh} \times L_{kh}$ and it coincides with $\mathcal{L}_r(v_h,q_h,\mu_h)$ if $k = 1$ (to within the term $\int_\Omega f v_h\,dx$). We have assumed for simplicity that $\int_\Omega f v_h\,dx$ has been approximated in (3.10) by the same quadrature formula as used for the other terms; however, this is not necessary.

In order to gain a better understanding of the algorithms which follow, it is important to consider the numerical solution of the variational problem equivalent to the minimisation with respect to q_h of $\mathcal{L}_r$ and $\mathcal{L}_{rh}$, with v_h and μ_h *fixed*. In the case of $\mathcal{L}_r$ defined by (3.8) we obtain

$$\left\{\begin{aligned} &p_h \in L_{kh}, \\ &\int_\Omega \left(|p_h|^{s-2}p_h + r p_h\right)\cdot q_h\,dx = \int_\Omega (r\nabla v_h + \mu_h)\cdot q_h\,dx \quad \forall q_h \in L_{kh}. \end{aligned}\right. \tag{3.11}$$

For $k = 1$ the components of p_h are constant over each triangle. For $k = 2$ we would use numerical integration via the Lagrangian $\mathcal{L}_{rh}$; in this case we have to determine the values of p_h at the quadrature

points. In both cases, we find that we have to solve, triangle-by-triangle or point-by-point, the nonlinear system

$$(3.12)\qquad \begin{cases} |p_h|^{s-2} p_{1h} + r p_{1h} = r \dfrac{\partial v_h}{\partial x_1} + \mu_{1h}\,, \\ |p_h|^{s-2} p_{2h} + r p_{2h} = r \dfrac{\partial v_h}{\partial x_2} + \mu_{2h}\,. \end{cases}$$

Once $|p_h|$ is known, p_{1h} and p_{2h} can immediately be deduced from (3.12); now (3.12) implies

$$(3.13)\qquad |p_h|^{s-1} + r|p_h| = |r\nabla v_h + \mu_h|\,,$$

or, defining $g(\cdot)$ by $g(q) = q^{s-1} + rq$,

$$(3.14)\qquad g(q) = |r\nabla v_h + \mu_h|\ ;$$

since the function g is a homeomorphism of $\mathbb{R}_+$ onto $\mathbb{R}_+$, (3.14) admits a unique solution which can be calculated by standard methods for nonlinear equations in one variable, although certain precautions have to be taken for extreme values of s (i.e. s close to 1, or s large)

3.2 The basic algorithm and its convergence properties

The basic algorithm for implementation of the augmented Lagrangian $\mathcal{L}_{rh}$ is once again algorithm ALG1 of Chapter III. The following is a brief summary of the principles of this algorithm.

ALGORITHM ALG1 (*General form*)

$$(3.15)\qquad \lambda_h^0 \ \textit{chosen arbitrarily in}\ L_{kh}$$

then for $n \geq 0$, λ_h^n *being known, calculate* u_h^n, p_h^n *and* λ_h^{n+1} *by*

$$(3.16)\qquad \begin{cases} \{u_h^n, p_h^n\} \in V_{kh} \times L_{kh}\,, \\ \mathcal{L}_{rh}(u_h^n, p_h^n, \lambda_h^n) \leq \mathcal{L}_{rh}(v_h, q_h, \lambda_h^n) \quad \forall\ \{v_h, q_h\} \in V_{kh} \times L_{kh}\,, \end{cases}$$

and

$$(3.17)\qquad \lambda_h^{n+1} = \lambda_h^n + \rho(\nabla u_h^n - p_h^n)\,;$$

test for convergence and return to (3.16) *if necessary.* ∎

In practice, it is necessary to have a method for solving problem (3.16); this leads us to a more 'operational' version of algorithm (3.15) - (3.17) :

ALGORITHM ALG1 (*Practical form*)

(i) *Choose :*
- *an accuracy tolerance* ε,
- *a maximum number of inner iterations* k_{max}
- *the values by which the outer iterations are initialised :*

$$\lambda_h^0 \in L_{kh} \ , \ p_h^{-1} \in L_{kh}. \tag{3.18}$$

(ii) *For* $n \geq 0$ *and with* λ_h^n *and* p_h^{n-1} *known, put* $z^0 = p_h^{n-1}$; *then for* $m \geq 0$ *and with* z^m *known, solve :*

$$\begin{cases} v^m \in V_{kh}, \\ r\int_\Omega \nabla v^m \cdot \nabla v_h \, dx = \int_\Omega f v_h \, dx + \int_\Omega (rz^m - \lambda_h^n)\cdot \nabla v_h \, dx \quad \forall v_h \in V_{kh} \end{cases} \tag{3.19}$$

(note that the solution v^m of (3.19) minimises $v_h \to \mathcal{L}_{rh}(v_h, z^m, \lambda_h^n)$ over V_{kh}).

Next, knowing the solution v^m *of* (3.19), *solve :*

$$\begin{cases} z^{m+1} \in L_{kh}, \\ \int_\Omega (|z^{m+1}|^{s-2} z^{m+1} + rz^{m+1})\cdot q_h \, dx = \int_\Omega (r\nabla v^m + \lambda_h^n)\cdot q_h \, dx \quad \forall q_h \in L_{kh} \end{cases} \tag{3.20}$$

(actually, the 'point' version of these equations; z^{m+1} thus calculated minimises $q_h \to \mathcal{L}_{rh}(v^m, q_h, \lambda_h^n)$ over L_{kh}).

If $z^{m+1} - z^m$ *is less than* ε (*in the* L^2 *norm, for example*), *or if* $m + 1 = k_{max}$, *put*

$$u_h^n = v^m, \ p_h^n = z^{m+1} \ ; \tag{3.21}$$

otherwise set $m = m + 1$ *and return to* (3.19).

(iii) *With* u_h^n *and* p_h^n *known, calculate* (*triangle-by-triangle if* $k = 1$, *or at the quadrature points if* $k = 2$)

$$\lambda_h^{n+1} = \lambda_h^n + \rho(\nabla u_h^n - p_h^n). \tag{3.22}$$

If $\nabla u_h^n - p_h^n$ *is sufficiently small, terminate* ALG1; *otherwise set* $n = n+1$ *and return to* (ii). ■

The determination of $\{u_h^n, p_h^n\}$ is thus carried out by a *relaxation* method; if we take $k_{max} = 1$, we obtain algorithm ALG2 of Chapter III.

Remark 3.1: For reasons which will be given in Remark 3.2, it is recommended that a termination test be used which relates to z^m rather than to v^m ; as is pointed out in MARROCCO [1], a termination test on v^m can lead to premature termination ('sticking' or pseudo-convergence) of the iterative process. ■

In order to gain a better understanding of the behaviour of ALG1, it is instructive to study this algorithm when $\rho = r$ and $s = 2$; for this value of s the problem to be solved is *linear*. The interesting fact is that in this case we obtain convergence in two iterations at most; this was demonstrated earlier in Chapter III, Section 4.3, but we shall nonetheless prove it again for the particular case in question here (the proof which follows is valid for ALG1; a similar one can be given for ALG2 if $\rho = r = 1$, and this value $r = 1$ will also be encountered once more when we examine the experimental convergence results).

Just as in Chapter III, it is sufficient to consider the case $f = 0$ (so that $u_h = 0$); additionally, we write

$$(3.23) \qquad Q = \{q_h \mid q_h \in L_{kh},\ q_h = \nabla v_h,\ v_h \in V_{kh}\}\ (= \nabla V_{kh}),$$

and we denote by P_Q the operator of projection from L_{kh} onto Q for the norm $L^2(\Omega)$.

It then follows from (3.19), (3.21) that

$$(3.24) \qquad P_Q \lambda_h^n = r(P_Q p_h^n - \nabla u_h^n),$$

and then from (3.20) that

$$(3.25) \qquad p_h^n + r(p_h^n - \nabla u_h^n) = \lambda_h^n.$$

By projection of (3.22) onto Q and its orthogonal space, and taking account of the fact that $\rho = r$, we obtain

$$(3.26)_1 \qquad P_Q \lambda_h^{n+1} = P_Q \lambda_h^n - r(P_Q p_h^n - \nabla u_h^n)\ ,$$

$$(3.26)_2 \qquad (I - P_Q)\lambda_h^{n+1} = (I - P_Q)\lambda_h^n - r(I - P_Q)p_h^n.$$

We deduce from (3.24) and $(3.26)_1$ that

$$P_Q\lambda_h^{n+1} = 0 \quad \forall n \geq 0, \tag{3.27}$$

so that from (3.24) we have

$$P_Q p_h^n - \nabla u_h^n = 0 \quad \forall n \geq 1. \tag{3.28}$$

By projection of (3.25) onto Q, we then obtain

$$P_Q\lambda_h^n = P_Q p_h^n = 0 \quad \forall n \geq 1.$$

We thus have $\nabla u_h^n = 0$ for $n \geq 1$, so that $u_h^n = 0$. On the other hand, we have

$$(I-P_Q)\lambda_h^{n+1} = \frac{1}{1+r}(I-P_Q)\lambda_h^n. \quad \blacksquare \tag{3.29}$$

There is thus decoupling between the convergence of $\{u_h^n\}_{n\geq 0}$ and that of $\{\lambda_h^n\}_{n\geq 0}$ if $s = 2$. For $s \neq 2$, a variant of the above proof would show that $P_Q\lambda_h^n$ always converges in a single iteration, but since it is then no longer possible to decompose (3.20) onto Q and its orthogonal in L_{kh}, we no longer have convergence of u_h^n in a finite number of iterations. We can nonetheless expect very different rates of convergence for $P_Q\lambda_h^n$, $P_Q p_h^n$ and $(I - P_Q)\lambda_h^n$, $(I - P_Q)p_h^n$.

Remark 3.2: In view of these quite different convergence properties of the sequences $\{u_h^n\}_{n\geq 0}$ and $\{p_h^n\}_{n\geq 0}$, $\{\lambda_h^n\}_{n\geq 0}$, it is necessary to use a termination test somewhat more sophisticated than one which relates solely to the sequence $\{u_h^n\}_{n\geq 0}$; it is for this reason that in the case of ALG1 discussed earlier, we suggested terminating the outer iterations when $\nabla u_h^n - p_h^n$ became sufficiently small in norm.

Remark 3.3: Another case, the practical importance of which will become apparent in Section 4, is the one in which the problem to be solved is linear but with a variable coefficient, i.e. of the form (with $a(x) \geq \alpha > 0$ a.e.)

$$\inf_{v\in H_0^1(\Omega)} \left\{\frac{1}{2}\int_\Omega a(x)|\nabla v|^2\, dx - \int_\Omega fv\, dx\right\}. \tag{3.30}$$

We shall assume $a(x)$ to be such that the *minimisation problem* (3.30) admits a unique solution. If for the solution of this problem we consider the augmented Lagrangian

$$\mathcal{L}_r(v,q,\mu) = \frac{1}{2}\int_\Omega a(x)|q|^2 dx - \int_\Omega fv\, dx + \int_\Omega \mu\cdot(\nabla v - q)\, dx + \frac{r}{2}\int_\Omega |\nabla v - q|^2 dx, \tag{3.31}$$

then algorithm ALG1 no longer converges in two iterations; however, this property is recovered if instead of (3.31) we use the augmented Lagrangian

$$(3.32)\qquad \mathcal{L}_r(v,q,\mu) = \frac{1}{2}\int_\Omega a(x)|q|^2dx-\int_\Omega fv\,dx+\int_\Omega a(x)\mu\cdot(\nabla v-q)dx+\frac{r}{2}\int_\Omega a(x)|\nabla v-q|^2dx.$$

It is reasonable to suppose that this property can be used to improve the convergence of ALG1 in the nonlinear case. In fact, let us assume that we know an estimate $\eta(x)$ of $|p_h|^{s-2}$; we could then consider an augmented Lagrangian of the form

$$(3.33)\qquad \begin{cases} \mathcal{L}_r(v_h,q_h,\mu_h)= \dfrac{1}{s}\displaystyle\int_\Omega |q_h|^s dx - \int_\Omega f v_h dx + \int_\Omega \eta(x)\mu_h\cdot(\nabla v_h - q_h)dx \\ \qquad\qquad + \dfrac{r}{2}\displaystyle\int_\Omega \eta(x)\,|\nabla v_h - q_h|^2 dx. \end{cases}$$

The ideal procedure would undoubtedly be to update $\eta(x)$ at each iteration of ALG1, using p_h^n, thereby obtaining a sequence $\{\eta^n\}_{n\geq 0}$ (we could even consider an update during the inner iterations (3.19), (3.20)); however, this update modifies, at each iteration, the matrix of the linear problem corresponding to (3.19); this problem would be of the form

$$(3.34)\qquad \begin{cases} \textit{Find}\quad v^m \in V_{kh} \textit{ such that } \forall v_h \in V_{kh} \textit{ we have} \\ r\displaystyle\int_\Omega \eta^n(x)\nabla v^m\cdot\nabla v_h\,dx = \int_\Omega f v_h dx + \int_\Omega \eta^n(x)(rz^m-\lambda_h^n)\cdot\nabla v_h\,dx. \end{cases}$$

It is appropriate to assess whether the acceleration of convergence thereby obtained justifies the cost of the new Cholesky factorisation if (3.34) is solved by a direct method; it is clear that the use of an effective iterative method (preconditioned conjugate gradient, multigrid method, etc.) for solving the linear problem (3.34) would be an attractive alternative to a direct method which requires at each iteration a refactorisation of the matrix associated with the linear problem (3.34).

Note the obvious links which exist between this method using $\{\eta^n\}_{n\geq 0}$, and a Newton method.

We shall encounter an application of this present remark later on in Section 4 of this chapter. ∎

Remark 3.4: It follows from Remark 3.3 that, all other things being equal, the optimal choice of the parameter r (in ALG2 in particular) will vary, in order of magnitude, as $|p_h|^{s-2}$; we can in fact interpret (3.33) as defining an augmented Lagrangian with a *variable* penalisation parameter equal to $\frac{r}{2}\eta(x)$, the optimal value of which would be (asymptotically in n at least) equal to $\frac{r}{2}|p_h|^{s-2}$ if p_h were known. Similarly, it may be conjectured that in the case of the standard penalty term (i.e. $\frac{r}{2}\int_\Omega |\nabla v_h - q_h|^2 \, dx$) the optimal parameter r will vary from one problem to another as $|p_h|^{s-2}$.

3.3 Numerical experiments

In this section we shall be discussing some of the many numerical results obtained by GLOWINSKI-MARROCCO [1],[4] and MARROCCO [1], giving particular emphasis to certain points which we consider to be especially important. In the numerical experiments which we shall describe, the domain Ω will be a *disc* and the right-hand side f of (3.2) will be a *positive constant* denoted by C ; under these assumptions, the exact solution of (3.2) is known, and in the case of a disc of radius R centred on the origin, this solution is given by

$$u(x) = \frac{s-1}{s}\left(\frac{C}{2}\right)^{\frac{s}{s-1}}\left(R^{\frac{s}{s-1}} - |x|^{\frac{s}{s-1}}\right), \tag{3.35}$$

where $|x| = (x_1^2 + x_2^2)^{\frac{1}{2}}$ if $x = \{x_1, x_2\}$. We have taken $R = 0.5$ and $C = 0.1$ in the numerical tests which will be described. The disc Ω has been triangulated by means of a triangulation $\mathcal{T}_h$ containing 256 triangles, the finite elements used being C^0-conforming and of degree *one*; the components of ∇v_h and of q_h are therefore constant over each triangle.

In order to evaluate the performance of ALG1, we have to remember that it is not sufficient merely to evaluate the number of updates of λ_h^n via (3.17); the sequence $\{\lambda_h^n\}_{n\geq 0}$ will in fact converge more quickly (in terms of the number of outer iterations) as r becomes larger; however, the increase in r slows down the convergence of the relaxation method used to solve system (3.16); there is therefore a balance to be struck between two opposing effects, and this implies the existence of an optimal value of r . Furthermore, in order to have a proper evaluation of the results, it is important to consider the total computational effort required for the solution of the

approximate problem; we shall therefore compare the *cumulative numbers* of relaxation iterations (3.19), (3.20) required for the convergence of the algorithm to within a given level of precision. In the case of ALG2 this number coincides with the number of updates of λ_h^n. Figures 3.1, 3.2 and 3.3 illustrate the behaviour of ALG1 for three values of s, namely $s = 1.1$, 3 and 10. Figures 3.4, 3.5 and 3.6 illustrate the corresponding results for ALG2 (used in exactly the same way as ALG1 with $\rho = r$).

In Figures 3.1 to 3.3 (i.e. relating to ALG1) two curves appear; the curve labelled C_1 shows the cumulative number of relaxation iterations as a function of r; curve C_2 indicates the number of updates of λ_h^n, this number tending towards unity as r increases. In Figures 3.4 to 3.6, curves C_1 and C_2 coincide. For the three values of s considered, the optimal r is to all intents and purposes the same for ALG1 and ALG2. Furthermore, even for $s = 1.1$, algorithm ALG2 appears to be more efficient than ALG1; for this family of problems, therefore, it does not appear worthwhile to solve system (3.16) by the relaxation method (3.19), (3.20) to very high accuracy before updating λ_h^n; the same also applies for values of s only slightly greater than unity. This observation concerning the superiority of ALG2 over ALG1 seems to contradict what was established in Chapter IV; however, this can be explained by the fact that the termination test used in MARROCCO [1], from which we have taken our results, relates to the difference between u_h^{n+1} and u_h^n; now it would once again appear that there exists a certain *decoupling* between the convergence of $\{u_h^n\}_{n\geq 0}$ and that of $\{p_h^n\}_{n\geq 0}$ or $\{\lambda_h^n\}_{n\geq 0}$, this decoupling becoming complete for $s = 2$ as we saw in Section 3.2. This decoupling property was not present in Chapter IV where, in fact, u_h^n and p_h^n are equal in the limit as $n \to +\infty$.

Another observation which at first sight is somewhat baffling concerns the high sensitivity of the optimal value of r to the problem being considered. The optimal value of r depends on s and also on the right-hand side f of equation (3.2).

Let us first consider the case in which s is fixed, and in which $f = \text{const.} = C$; we recall the empirical formula of GLOWINSKI-MARROCCO [4] for the optimal value of r: if r_o (resp. r_1) is the optimal value of r for $f = C_o$ (resp. $f = C_1$), then we have (approximately)

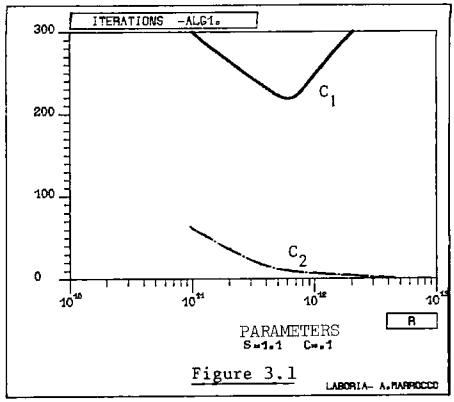


Figure 3.1

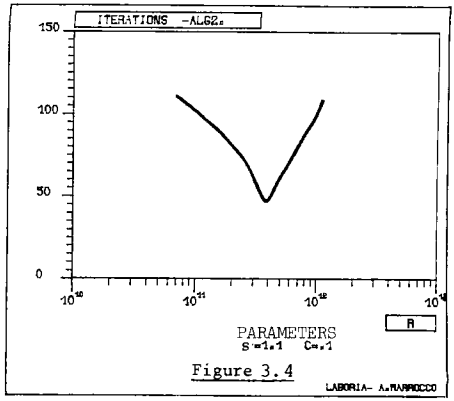


Figure 3.4

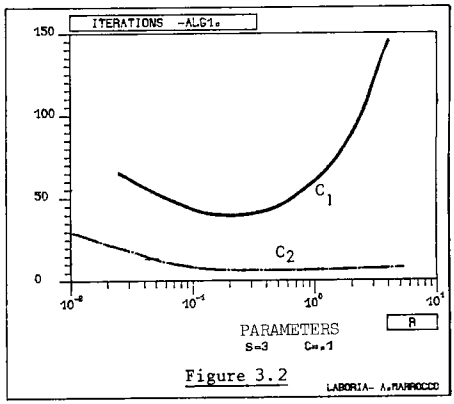


Figure 3.2

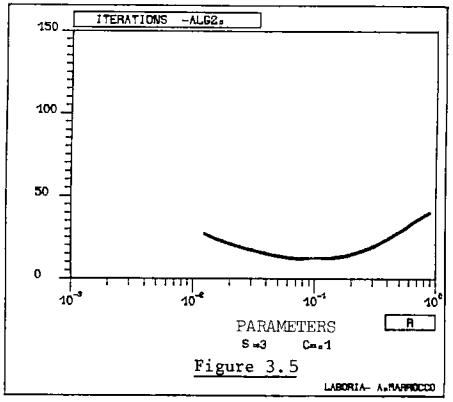


Figure 3.5

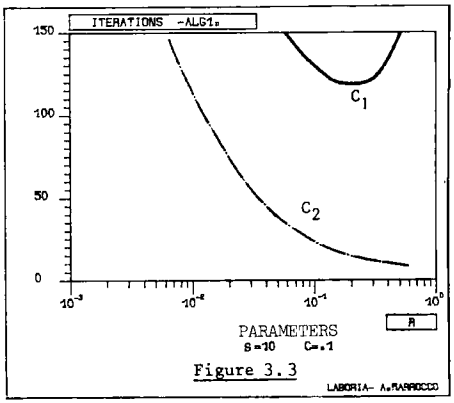


Figure 3.3

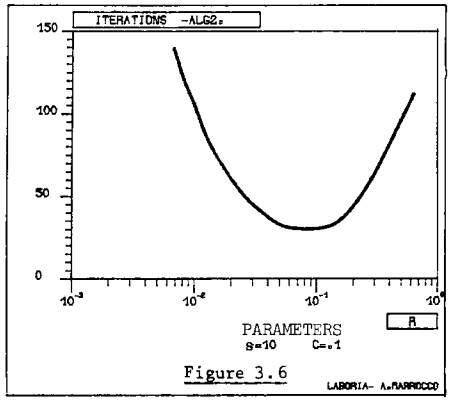


Figure 3.6

$$\frac{r_o}{r_1} = \left(\frac{C_1}{C_o}\right)^{\frac{2-s}{s-1}} . \tag{3.36}$$

This formula is to be viewed alongside Remark 3.4, where it was conjectured that the optimal r should be proportional to $|p_h|^{s-2}$; now it follows from (3.35) that when we pass from $f = C_o$ to $f = C_1$, $|p|$ $(= |\nabla u|)$ is multiplied precisely by the factor $(C_1/C_o)^{s-2}$; this is consistent with (3.36) and confirms the conjecture of Remark 3.4.

The effect of a variation of s is more complicated to investigate since we are then no longer dealing with a simple factor of proportionality; in the particular case of the test problem considered, which is in fact one-dimensional (being axisymmetric), it is nonetheless easy to verify that if $\hat{p} = \nabla \hat{u}$, where $\hat{u}$ is the solution of the problem for $s = 2$, then we have $|\nabla u| = |\hat{p}|^{1/(s-1)}$ where u is the solution corresponding to an arbitrary value of s in the range $1 < s < +\infty$; now, for $s = 2$, $|\hat{p}|$ is proportional to $|x|$ $(= (x_1^2 + x_2^2)^{\frac{1}{2}})$ and $|\nabla u|^{s-2}$ will thus be proportional to $|x|^{\frac{s-2}{s-1}}$, so that we have

$$\begin{cases} |p_h|^{s-2} \simeq \dfrac{1}{|x|^9} \ \textit{if}\ s = 1.1, \\ |p_h|^{s-2} \simeq |x|^{1/2} \ \textit{if}\ s=3, \\ |p_h|^{s-2} \simeq |x|^{8/9} \ \textit{if}\ s=10. \end{cases} \tag{3.37}$$

Now an optimal r value 3×10^{11} is found for $s = 1.1$; 2×10^{-1} for $s = 3$; and 0.9×10^{-2} for $s = 10$; these numbers correspond exactly to the values predicted by (3.37). It can therefore be seen that once again the use of an augmented Lagrangian of the type (3.33), where $\eta(x)$ is an estimate of $|p_h|^{s-2}$, could very probably stabilise the choice of r. We shall encounter another application of this principle in Section 6.3.

Finally, we must point out that the results of GLOWINSKI-MARROCCO [1],[4] show the importance of working in double precision (on IBM computers at least), in particular for extreme values of s (i.e. s close to 1 and $>>2$). In similar vein, comparisons between ALG1, ALG2 and other methods of solving problem (3.2) (in particular, nonlinear overrelaxation methods) may be found in GLOWINSKI-MARROCCO [4]; it would appear that the algorithms described above are the most efficient for solving (3.2), particularly for s close to 1 or $>>2$.

3.4 Continuity constraints on the gradient; Interior penalty methods

We consider (for simplicity) the case of a *linear* problem ($s = 2$) of the type (3.2), solved by finite-element methods of degree at least *two*. In a standard approximation, the *normal* component of the gradient of the approximate solution is not continuous across inter-element boundaries. Now, for certain applications, it may be desirable (and useful) to enforce this continuity. One possible approach for achieving this consists of considering this continuity condition as a supplementary constraint which can be treated either by means of a *Lagrange multiplier* or by means of a *penalty* term; this latter approach is known by the name of the '*interior penalty method*', the penalisation being applied to any jump in the normal derivative occurring at the element interfaces, in the interior of the domain.

The framework which we have just described does in fact adapt quite readily to such methods: it is sufficient to impose matching conditions on the functions from L_{kh} and to use a variant of the augmented Lagrangian defined earlier.

A particularly favourable case is that in which, if $k = 2$, we enforce matching of the components of the gradient at the midpoints of the element sides (L_{2h} then comprises nonconforming elements of degree one); we will thus be substituting the constraint $\nabla v_h - q_h = 0$ for the constraint of matching the normal derivatives. Since the midpoints of the sides coincide with the quadrature points which integrate polynomials of degree two exactly, the basis of L_{2h} formed by associating an interpolation function with each node is orthogonal. It therefore follows that the calculation of p_h^n remains a point calculation despite the inter-element matching conditions. However, despite the linearity of the problem, we no longer have convergence of $\{u_h^n\}_{n\geq 0}$ in a finite number of iterations; in fact, the matching constraint imposed in L_{2h} invalidates the proof of Section 3.2. Nonetheless, since the calculation of p_h^n via the analogue of (3.20) is *linear* (since $s = 2$) and *local*, we can determine p_h^n as a function of u_h^n and λ_h^n and insert the results obtained into the analogue of (3.19), and thereby eliminate the inner relaxation iterations at the cost of having to assemble a matrix which is only slightly more complicated than that in the original scheme. Then we need only choose r as large as possible, i.e. at the limit of the machine precision, in order to obtain a correct solution of the linear problem obtained

by elimination of p_h.

Remark 3.5: The convergence, as $h \to 0$, of such approximations with constraints on the normal derivative is often quite difficult to prove (but see DOUGLAS-DUPONT [1], WHEELER [1] and the corresponding bibliography, as well as FORTIN-SOULIE [1] in which a nonconforming element of degree two is considered, for which convergence can be proved).

4. A MAGNETO-STATIC PROBLEM

4.1 Formulation of the problem

In this section we shall be discussing some of the results obtained by GLOWINSKI-MARROCCO [5] and MARROCCO [1] in connection with the calculation of the magnetic state of rotating machines (motors, alternators, ...) or static machines (transformers). Here, we are concerned with problems in which even the linearised case has a highly variable coefficient, and in which it will be necessary to use an augmented Lagrangian of the kind introduced in Remark 3.3 of Section 3.2.

With $\vec{\nabla} = \{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\}$, the Maxwell equations of magnetostatics are written:

$$\vec{\nabla} \times \vec{H} = \vec{j} \tag{4.1}$$

$$\vec{B} = \mu \vec{H}, \tag{4.2}$$

$$\vec{\nabla} \cdot \vec{B} = 0 \ ; \tag{4.3}$$

in (4.1) - (4.3), $\vec{H}$ is the *magnetic field* vector, $\vec{j}$ is the *current density* vector, $\vec{B}$ the *magnetic induction* vector and μ the *magnetic permeability* of the medium. In view of (4.3) there exists a vector potential $\vec{A}$ such that $\vec{B} = \vec{\nabla} \times \vec{A}$; the above equations therefore lead to the equation

$$\vec{\nabla} \times (\nu \vec{\nabla} \times \vec{A}) = \vec{j}, \tag{4.4}$$

with $\nu = 1/\mu$. We have $\nu = \nu_r \nu_o$, where ν_r is the relative magnetic reluctivity and where ν_o is the magnetic reluctivity in vacuo; we have $\nu_o = 10^7/4\pi$ (M.K.S.A.). In the following, we shall be restricting our attention to two-dimensional cases for which

(4.5) $\vec{A} = \{0,0,A\}$, $\vec{j} = \{0,0,j\}$,

so that (4.4) reduces (with $\nabla = \{\frac{\partial}{\partial x_1}, \frac{\partial}{\partial x_2}, \frac{\partial}{\partial x_3}\}$) to:

(4.6) $-\nabla\cdot(\nu\nabla A) = j.$

The coefficient ν appearing in (4.6) depends on the material; in *copper* and in *air* we have $\nu = \nu_o$; in *ferro-magnetic media* we shall assume - neglecting hysteresis effects - that ν is an increasing function (in the wide sense) of $|\vec{B}|$; it is actually more convenient to consider ν to be an increasing function of $|\vec{B}|^2$ and this is what we shall do in the following discussion. In the two-dimensional case we have

(4.7) $|\vec{B}| = |\vec{\nabla}\times\vec{A}| = |\nabla A|$

and we can then rewrite (4.6) in the more explicit form

(4.8) $-\nabla\cdot(\nu(x,|\nabla A|^2)\nabla A) = j.$

Finally, denoting by $\psi(x,\sigma)$ the function which satisfies

(4.9) $\frac{\partial\psi}{\partial\sigma}(x,\sigma) = \nu(x,\sigma)$, $\psi(x,0) = 0,$

we see that (4.6) is the Euler equation, in an appropriate functional space V, of a problem in the Calculus of Variations, namely the minimisation of the energy functional $\mathfrak{J}$ defined by

(4.10) $\mathfrak{J}(v) = \frac{1}{2}\int_\Omega \psi(x,|\nabla v|^2)dx - \int_\Omega jv\,dx,$

where Ω is usually the median cross-section of the rotating machine. The problem

(4.11) $$\begin{cases} A \in V, \\ \mathfrak{J}(A) \le \mathfrak{J}(v) \quad \forall v \in V \end{cases}$$

is clearly of the form considered in Section 2. The choice of V depends on the boundary conditions; in the following we shall assume that $V = H_o^1(\Omega)$, which corresponds to boundary conditions of *homogeneous Dirichlet* type. On the basis of laboratory measurements we have taken for ν, in the ferromagnetic parts, a function of the form

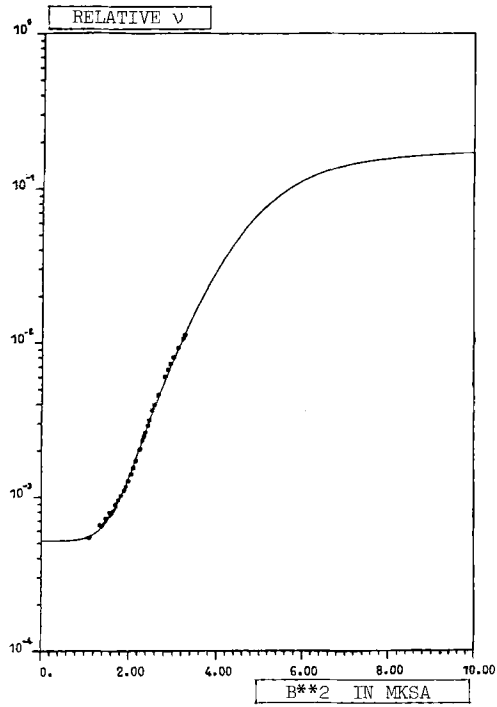
$$(4.12) \qquad \nu(\varepsilon,\alpha,C,T;\sigma) = \varepsilon + (C-\varepsilon)\,\frac{\sigma^{\alpha}}{\sigma^{\alpha}+T}\,,$$

the parameters ε,α,C,T being positive with $C > \varepsilon$. The function thus obtained[6] is monotone; Figure 4.1 represents the results of this smoothing for ν_r. For reluctivities of the type (4.12), the problem possesses a unique solution if $V = H_o^1(\Omega)$.

Figure 4.1

STATOR RESULT OF SMOOTHING

ALPHA	0.54192E01
C	0.17577E00
T	0.87589E04
EPS	0.51636E-03



4.2 Formulation using an augmented Lagrangian

Even if we neglect the nonlinearity, which would be permissible for *low currents*, problem (4.8) has variable coefficients since the reluctivity is not the same in air as in iron. To reduce the problem to the form of the general case, referring back to what was said in Remark 3.3 of Section 3.2, we would use in the general case the augmented Lagrangian[7] defined as follows :

---

[6] Using a *least squares* smoothing technique.

[7] The μ in (4.13) has nothing to do with the magnetic permeability, a quantity which will not appear again explicitly hereinafter.

$$(4.13)\qquad \begin{cases} \mathcal{L}_r(v,q,\mu) = \frac{1}{2}\int_\Omega \psi(x,|q|^2)dx - \int_\Omega jv\ dx + \\ \frac{r}{2}\int_\Omega \eta(x)|\nabla v - q|^2\ dx + \int_\Omega \eta(x)\mu\cdot(\nabla v - q)dx. \end{cases}$$

The optimal choice for η is the *unknown* function ν ; in practice, one attempts to obtain a suitable estimate for this function.

To solve (4.11) we shall again use algorithms ALG1 and ALG2 which we shall not describe explicitly here; for information on these algorithms we refer the reader to the preceding sections and chapters, and to GLOWINSKI-MARROCCO [5] and MARROCCO [1]. In the case of a linear problem, we shall have convergence of ALG1 (resp. ALG2) in at most two iterations (for $\{u_h^n\}_{n\geq 0}$) if $\eta = \nu$ and if $\rho = r$ (resp. $\rho = r = 1$).

The execution of ALG1 and ALG2 involves the simultaneous solution for ALG1, and the sequential solution for ALG2, of the equations (omitting the iteration superscripts):

$$(4.14)\qquad \begin{cases} r\int_\Omega \eta(x)\nabla A\cdot\nabla v\ dx = \int_\Omega jv\ dx + \int_\Omega \eta(x)(rp-\lambda)\cdot\nabla v\ dx \quad \forall v \in V, \\ A \in V, \end{cases}$$

which is *linear*, and

$$(4.15)\qquad \begin{cases} \int_\Omega (\nu(x,|p|^2)p + r\eta p)\cdot q\ dx = \int_\Omega \eta(r\nabla A+\lambda)\cdot q\ dx \quad \forall q \in (L^2(\Omega))^2, \\ p \in (L^2(\Omega))^2, \end{cases}$$

which is *nonlinear*, but which can be solved locally. In fact, to solve (4.15) we proceed as in Section 3 by introducing $z = |p|$, which takes us from solving (4.15) to solving

$$(4.16)\qquad (\nu(x,z^2)+r\eta(x))z = \eta(x)|r\nabla A+\lambda|.$$

We put $g(x;z) = (\nu(x,z^2)+r\eta(x))z$; the dependence on x means that we have several types of equations depending on whether the point x lies in air, copper or iron. In air and copper we have $\nu=\nu_0$ and (4.16) gives z explicitly; in iron $g(x;\cdot)$ is a strictly

increasing function possessing a point of inflection which depends neither on r nor on η. If we take care with the initialisation, we can use Newton's method for solving (4.16). For the discrete problems, the solution takes place at the quadrature points for finite elements of order ≥ 2, and triangle-by-triangle for elements of order 1. Most of the comments made in Section 3 remain valid for the present problem.

4.3 Numerical experiments

The results presented here concern a four-pole alternator; the corresponding domain Ω is therefore a disc and the triangulation $\mathcal{T}_h$ used (actually a quarter of it) is shown in Figure 4.2; this triangulation $\mathcal{T}_h$ contains 812 triangles and 384 interior vertices. The finite elements used are C^0-conforming of order 1. Figure 4.3 shows the contours (which are actually the lines of magnetic induction) for the solution corresponding to a current of 5 A/mm^2.

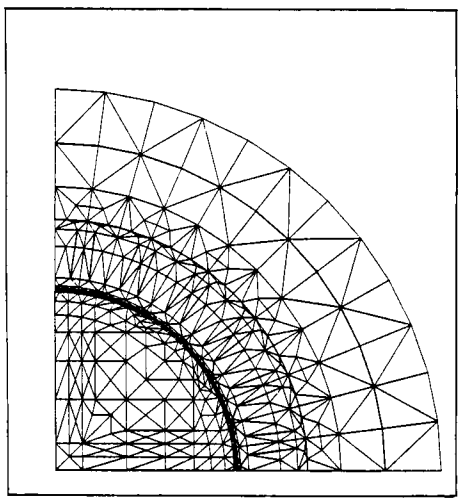


Figure 4.2

Triangulation used.

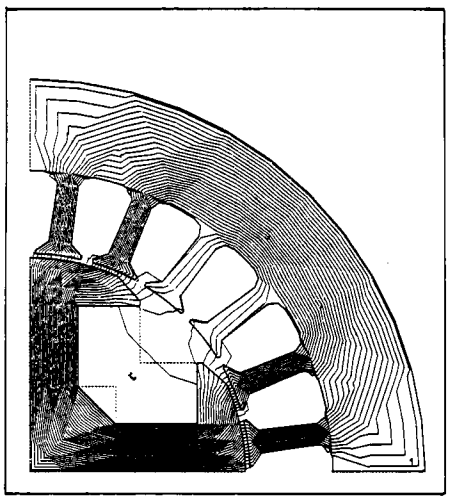


Figure 4.3

Lines of magnetic induction.

On this figure the rotor, the stator and the air-gap of the alternator can clearly be seen. The choice of the function $\eta(x)$ requires an initial estimate for ν; for the results presented here we have used two methods, the second of which is actually a more sophisticated version of the first:

Method 1: We know the minimal values ε_S and ε_R of ν in the stator and the rotor, respectively. For $k > 0$, chosen arbitrarily, we put

$$(4.17)\qquad \begin{cases} \eta_k = \nu_o & \textit{in the air and the copper} \\ \eta_k = k\,\varepsilon_R & \textit{in the rotor} \\ \eta_k = k\,\varepsilon_S & \textit{in the stator.} \end{cases}$$

Method 2: Having defined η_k by (4.17) we solve the following linear problem (actually its discrete version)

$$(4.18)\qquad \begin{cases} A_k \in V, \\ \int_\Omega \eta_k\, \nabla A_k \cdot \nabla v\, dx = \int_\Omega jv\, dx \quad \forall v \in V. \end{cases}$$

We then put

$$(4.19)\qquad \begin{cases} \eta = \nu_o & \textit{in the air} \\ \eta = \nu(|\nabla A_k|^2) & \textit{in the rotor or the stator.} \end{cases}$$

We could consider using a systematic procedure for updating $\eta(x)$ (after a specified number of iterations of ALG1 or ALG2, for example). The major drawback of such an update lies in the fact that the bi-linear form in problem (4.14) would be modified very frequently so that for the approximate problem a new matrix would have to be factorised if the discrete analogue of (4.14) is solved by a direct method.

In practice, a reliable method (actually a *'continuation'*-type method) consists of progressively increasing the current density j (which increases the nonlinearity of the problem) and using the results of the preceding calculation for the estimation of ν. The

results of MARROCCO [1], which will be briefly repeated here, were all obtained using algorithm ALG2. We list below what we consider to be the most important features of these results.

For a low value of the current ($j = 0.5$ A/mm^2) the problem is practically linear and ν, both in the rotor and in the stator, stays close to the minimal values ε_R and ε_S respectively. For $\rho=r=1$ we obtain the approximate solution with excellent precision in two iterations for $\eta(x)$ chosen in accordance with (4.17) with $k = 1$. For $k > 1$, still with $j = 0.5$ A/mm^2, the value of η is over-estimated, and this slows down the convergence; hence for $k = 10$ approximately 25 iterations of ALG2 are required for convergence.

For a larger value of j ($j = 2$ A/mm^2), the importance of the non-linearity increases and it is still the estimation of $\eta(x)$ which has the predominant effect on the rate of convergence. Suppose that $j = 2$ A/mm^2 and consider an estimate of η obtained by using (4.17); by making k vary from 1 to 20, we observe that the optimal value of r decreases from 10 to 1 and that the corresponding number of iterations decreases from 90 to 50. Figure 4.4 represents a typical result obtained with $k = 20$ and $j = 2$ A/mm^2.

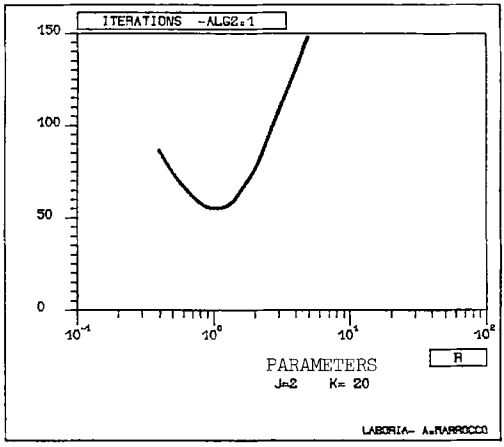


Figure 4.4

Method 2 (i.e. (4.18), (4.19)) is very efficient if we start with the estimate obtained by using (4.17) with $k = 1$. We see the optimal value of r go to 1 and the number of iterations decrease to about 25. For larger values of k the use of (4.18), (4.19) does not improve, and in fact actually decreases, the rate of convergence. For $j = 2$ A/mm^2, which gives a small nonlinearity, taking $k = 1$ amounts to estimating $\eta(x)$ by solving the linear problem, and this is a natural choice to make.

In contrast, for $j = 7.5$ A/mm^2, which in this example corresponds to the normal régime, the effect of the nonlinearities is larger. If we choose $\eta(x)$ by method (4.17) we obtain very comparable results for k varying from 10 to 50: approximately 50 iterations to obtain the solution, with an optimal value of r going from 10 to 1. In the case of method (4.18), (4.19) we obtain a significant improvement for $k = 10$, which reduces the number of iterations to about 25. For higher values, method (4.18), (4.19) can become disastrous.

We can deduce from all this that the use of the augmented Lagrangian method will be effective if we succeed in obtaining a good initial estimate of $\eta(x)$, which allows us to work (in ALG2) with $r = 1$. The most rational choice would be to make the fullest possible use of the information provided by earlier calculations: for example, continuously increasing the current enables the nonlinear effects to be introduced gradually. It should be noted, however, that the algorithm itself is very robust; a poor choice of $\eta(x)$ or of r slows down, but does not prevent, convergence. This algorithm can therefore be used with every confidence, and its performance can be improved by experiment.

5. CALCULATION OF SUBSONIC AND TRANSONIC POTENTIAL FLOWS OF COMPRESSIBLE IDEAL FLUIDS

In this section we attempt to present a succinct examination of the possibilities offered by augmented Lagrangian methods for the numerical simulation of *subsonic* and *transonic* flows of compressible ideal fluids; in the *transonic* case the problem considered is difficult and has given rise to numerous publications; for further details on the numerical solution of such problems by *finite element* methods, we refer to BRISTEAU-GLOWINSKI-PERIAUX-PERRIER-PIRONNEAU-

POIRIER [1],[2], GLOWINSKI [2, Chapter 7], AMARA-JOLY-THOMAS [1], and to their associated bibliographies.

5.1 Formulation of the problem

In this section we consider the numerical simulation of the *potential* (and therefore *irrotational*) flow of a compressible ideal fluid. Within the limited scope of this section it would be out of the question to attempt to give a detailed description of the physical assumptions underlying the model described below, and we advise the reader who is not familiar with such problems to consult one of the treatises on Fluid Mechanics, such as LANDAU-LIFCHITZ [1].

Let ϕ denote the *velocity potential*, with $u = \nabla\phi$ the velocity field; with ρ denoting the density of the fluid, the mass conservation equation is then written

$$\nabla\cdot\rho u = \nabla\cdot\rho\nabla\phi = 0. \tag{5.1}$$

By introducing certain assumptions relating to the equation of state of the fluid, we deduce from the conservation of momentum equation that

$$\rho = \rho(u) = \rho_o\left(1 - \frac{\gamma-1}{\gamma+1}\frac{|u|^2}{c_\star^2}\right)^{\frac{1}{\gamma-1}}, \tag{5.2}$$

where, in (5.2):

(i) ρ_o is the *density of the fluid at rest*

(ii) γ is the *ratio of specific heats;* $\gamma = 1.4$ in the case of air

(iii) $c_\star$ is the *critical speed of sound*.

In the case of air we thus have

$$\rho = \rho_o(1-k|u|^2)^{2.5} \quad (\textit{if } |u|^2 \le \tfrac{1}{k}); \tag{5.3}$$

in a suitably chosen system of units we have $\rho_o = 1$ and $k = \frac{1}{6}$. Combining (5.1), (5.2) and (5.3) we obtain a nonlinear problem which is formally very similar to the magnetostatic problem studied above in Section 4. It is therefore quite natural to follow the same line of approach as in Section 4 in order to attempt to obtain a numerical

solution of the flow problems in question here.

We therefore define ψ by

$$(5.4)\qquad \begin{cases} \dfrac{d\psi}{d\sigma} = \rho_o(1-k\sigma)^{5/2}, \quad 0 \le \sigma \le \dfrac{1}{k}, \\ \psi(0) = 0. \end{cases}$$

The function $\xi \to \psi(\xi^2)$ is increasing, convex on $[0,1]$ and concave on $[1,1/k]$. The functional $\mathfrak{F}$ defined by

$$(5.5)\qquad \mathfrak{F}(v) = \frac{1}{2}\int_\Omega \psi(|\nabla v|^2)dx - L(v),$$

where Ω is the flow domain and $L(\cdot)$ is a linear form which takes into account certain of the boundary conditions, is *non-convex* over its domain of definition; equation (5.1) in conjunction with an adequate set of boundary conditions characterises the *stationary points* of $\mathfrak{F}$, which in the present case are not necessarily minimum points. In the case where the stationary point ϕ satisfies $|\nabla\phi| \le 1$, which corresponds to a *subsonic* flow, we are in the *convex* 'part' of the functional and ϕ is the *unique* subsonic solution (possibly to within an additive constant). If there exists a region in which $|\nabla\phi| > 1$, then we have a genuine transonic flow and in order to ensure uniqueness it is necessary to impose a supplementary condition; this condition is provided by the *Second Law of Thermodynamics*, and may be formulated as follows:

$$(5.6)\qquad \left\{ \begin{array}{l} \textit{The entropy is non-decreasing along a streamline} \\ \textit{orientated in the direction of the flow.} \end{array} \right.$$

In the case of the present problem of potential flows, the entropy condition (5.6) can be formulated as an *inequality constraint* on the possible discontinuities in $\nabla\phi$, as follows. Suppose that S is a surface (or line) of discontinuity of the flow velocity (i.e. a shock); we denote by u_- the velocity immediately upstream of the shock, and by u_+ the velocity immediately downstream; we then have

$$(5.7)\qquad |n\cdot u_-| \ge |n\cdot u_+|, \ \textit{that is} \ |n\cdot(\nabla\phi)_-| \ge |n\cdot(\nabla\phi)_+|,$$

where n is the *normal* to S. The condition (5.7) eliminates *expansion shocks* (across which there would be a sudden *drop* in pressure and *decrease* in entropy); it thus admits only *compression* shocks.

It should be noted that the condition (5.7) depends on the solution itself (through the medium of S and n) and that the problem thus posed falls into the context of *quasi-variational inequalities*. The description of an *interior penalty* method for the numerical treatment of (5.7) and its application to the calculation of transonic flows may be found in BRISTEAU-GLOWINSKI-PERIAUX-PERRIER-PIRONNEAU-POIRIER [2], GLOWINSKI [2, Chapter 7], and GLOWINSKI-LIONS-TREMOLIERES [2, Appendix 4] .

5.2 Formulation via an augmented Lagrangian. Solution algorithms

Since the functional $\mathfrak{F}$ (see (5.5)) is not convex, the problem under consideration does not (except in the purely subsonic case) fall within the general framework of Chapter III; nonetheless, we can - formally at least - still introduce an augmented Lagrangian and consider seeking its stationary points, possibly taking account of condition (5.7) (by means of an interior penalty method, for example). We therefore introduce the augmented Lagrangian

$$\left\{\begin{array}{l} \mathcal{L}_r(v,q,\mu) = \frac{1}{2}\int_\Omega \psi(|q|^2)dx - L(v) + \frac{r}{2}\int_\Omega \eta(x)|\nabla v - q|^2\, dx \\ \quad + \int_\Omega \eta(x)\mu\cdot(\nabla v - q)\, dx, \end{array}\right. \tag{5.8}$$

where, in (5.8), the linear form $L(\cdot)$ depends on the boundary conditions of the problem. Just as in the magnetostatic problems in Section 4, we have introduced a renormalisation function η, the optimal value of which is $\rho(\nabla\phi)$ where ϕ is the required solution; the precise value of this factor, however, is not so critical in the present problem, since within the range of validity of the transonic model (which we may here take to be $|\nabla\phi| \leq 1.5$, say) the density finally varies only slightly (in the sense that it remains of the order of ρ_o). In order to find the stationary points of $\mathcal{L}_r$ we shall use algorithms ALG1 and ALG2; it will therefore be necessary to solve, at each iteration of ALG1, a system in $\{\phi^n, p^n\}$, λ^n being known, namely

$$\left\{\begin{array}{l} \int_\Omega \eta(x)\nabla\phi^n\cdot\nabla v\, dx = \int_\Omega \eta(x)(p^n - \frac{1}{r}\lambda^n)\cdot\nabla v\, dx + \frac{1}{r}L(v) \quad \forall v \in V, \\ \phi^n \in V, \end{array}\right. \tag{5.9}$$

$$(5.10)\qquad \begin{cases} \displaystyle\int_\Omega [\rho(|p^n|^2)p^n + r\eta p^n]\cdot q \, dx = \int_\Omega \eta(r\nabla\phi^n + \lambda^n)\cdot q \, dx \\ \forall q \in H,\ p^n \in H, \end{cases}$$

where V is a subspace of $H^1(\Omega)$, which takes into account the boundary conditions, and where $H = (L^2(\Omega))^N$ ($N = 1, 2, 3$ in applications). If we solve the system (5.9), (5.10) by relaxation, then (5.10) will involve solving an equation in one variable of the type (if $\eta \equiv 1$) :

$$(5.11)\qquad (\rho(z^2)+r)z = b.$$

We put $g_r(z) = (\rho(z^2)+r)z$. Figure 5.1 shows g_r for various values of r; for r small there will be two solutions to the equation $g_r(z) = \text{Const}$.

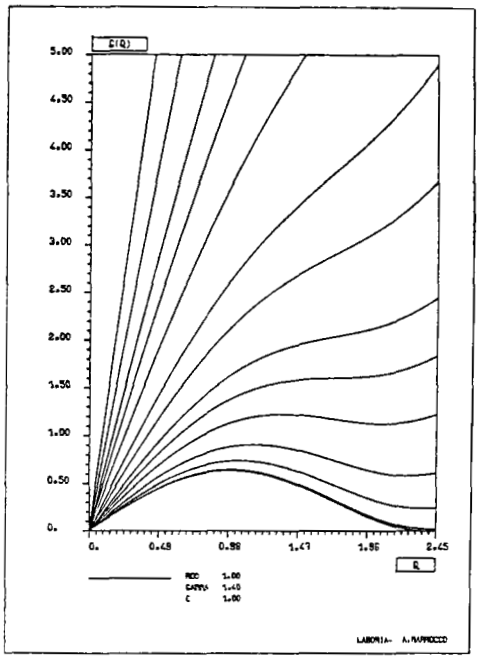


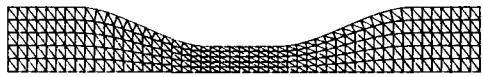
Figure 5.1

If in addition to λ^n we suppose that ϕ^n is known in (5.10), then for r sufficiently large, (5.10) admits a unique solution p^n; in contrast, the system (5.9), (5.10) with a given λ^n does not in general

have a unique solution.

In the case of *subsonic* flows, for which we remain within the domain of convexity of the functional $\mathfrak{F}$ in (5.5), we can use ALG1 and ALG2 without taking any particular precautions; some numerical tests for $\Omega \subset \mathbb{R}^2$ are given in MARROCCO [1], using ALG2 with the Lagrangian $\mathcal{L}_r$ from (5.8), in which $\eta \equiv 1$ has been taken; the spaces V and H in (5.9), (5.10) are approximated using C^0-conforming finite elements of order 1 (i.e. piecewise affine) over a triangulation $\mathcal{T}_h$ of Ω. In the case of the nozzle shown in Figure 5.2 we have convergence of ALG2, for subsonic flows, in at most 20 iterations. The optimal value of r is close to 1.

In the case of *genuine transonic flows*, the solution calculated by the above algorithms can depend on r and may contain expansion shocks (i.e. non-physical shocks); it is therefore necessary to incorporate into the mathematical model the condition (5.7). There are several possible ways of doing this; however, we think it useful to point out one in particular, which results directly from the decomposition associated with the relation $\nabla v - q = 0$. We have already mentioned in Section 3.4 that it is possible to impose continuity constraints on the gradient of the solution; now, the conditions (5.7), which are of inequality type, are very similar to the constraints on $\nabla\phi$ considered in Section 3.4. A natural approach would therefore be to approximate ϕ (and the corresponding test functions v) by piecewise functions of degree $k \geq 2$, and $p = \nabla\phi$ by piecewise functions of degree $k - 1$; if $k = 2$, we can impose condition (5.7), that is $|p_- \cdot n| \geq |p_+ \cdot n|$, at the midpoints of the element sides. It remains to investigate solution algorithms which are suitable for such a treatment of condition (5.7).



NUMBER OF NODES 300
NUMBER OF ELEMENTS 490

(The figure depicts a half nozzle)

Figure 5.2

6. FURTHER APPLICATIONS

In this section we consider three examples of applications which come within the framework of Section 2 of this chapter; two of these examples have already been touched upon in Chapter III. Since the numerical treatment of these problems is very similar to that of the problems in the preceding sections and since the numerical results obtained only serve to confirm those obtained earlier, it will suffice to give only a relatively brief outline of them; nonetheless, we shall attempt to shed some light on the particular qualities of the problems considered, and their impact on the numerical treatment. Sections 6.1 and 6.2 will thus discuss, respectively, *the flow of a Bingham fluid in a cylindrical duct* and *the elastoplastic torsion of a cylindrical bar*. Section 6.3 will treat a *minimal surface* problem.

6.1 Flow of a viscoplastic Bingham fluid in a cylindrical duct

6.1.1 *Formulation of the problem*

We consider the flow of a viscoplastic material of Bingham type in a cylindrical duct with cross-section Ω. This problem involves a *plasticity threshold* g; if the mechanical stresses remain below this threshold value, the material stays rigid, whereas beyond the threshold it behaves like an *incompressible fluid with viscosity* ν. The flow in the duct is induced by a linear pressure drop f which, in practice, is constant over the cross-sections of the duct.

Let $u \in H_0^1(\Omega)$ denote the required velocity; we obtain this by minimising in $H_0^1(\Omega)$ the functional

$$J_g(v) = \frac{\nu}{2}\int_\Omega |\nabla v|^2 dx + g\int_\Omega |\nabla v|\, dx - \int_\Omega fv\, dx \tag{6.1}$$

with $\nu > 0$, $g > 0$, $f \in L^2(\Omega)$.[8]

We note that if $\nu = 0$ we again have the problem of Section 3 with $s = 1$, the term $\int_\Omega |\nabla v| dx$ being non-differentiable. It is essential to assume $\nu > 0$ if we are seeking a solution in $H_0^1(\Omega)$ for the above problem; this problem in the Calculus of Variations admits a unique solution, characterised by the *variational inequality*

[8] We have assumed $f \in L^2(\Omega)$ in order to slightly widen the generality.

$$\text{(6.2)}\qquad \begin{cases} \nu\int_\Omega \nabla u\cdot\nabla(v-u)\,dx + g\int_\Omega |\nabla v|\,dx - g\int_\Omega |\nabla u|\,dx \geq \int_\Omega f(v-u)\,dx \quad \forall v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega) \end{cases}$$

a detailed numerical analysis of which may be found in GLOWINSKI-LIONS-TREMOLIERES [1], [2] and GLOWINSKI [1], [2].

6.1.2 *Solution by augmented Lagrangian methods*

An augmented Lagrangian associated naturally with the flow problem defined in Section 6.1.1 is $\mathcal{L}_r : H_0^1(\Omega) \times (L^2(\Omega))^2 \times (L^2(\Omega))^2 \to \mathbb{R}$ given by

$$\text{(6.3)}\qquad \begin{cases} \mathcal{L}_r(v,q,\mu) = \frac{\nu}{2}\int_\Omega |q|^2\,dx + g\int_\Omega |q|\,dx - \int_\Omega fv\,dx \\ + \frac{r}{2}\int_\Omega |\nabla v - q|^2 dx + \int_\Omega \mu\cdot(\nabla v - q)\,dx. \end{cases}$$

The corresponding algorithms ALG1 and ALG2 have already been described in Chapter III, Section 3.3; however it will be worth making the effort to write out the system corresponding to the minimisation of $\mathcal{L}_r$ with respect to q and v, with μ *fixed*. The optimality condition with respect to v, at the point $\{u,p,\lambda\}$ leads to

$$\text{(6.4)}\qquad \begin{cases} r\int_\Omega \nabla u\cdot\nabla v\,dx = \int_\Omega fv\,dx + \int_\Omega (rp-\lambda)\cdot\nabla v\,dx \quad \forall v \in H_0^1(\Omega), \\ u \in H_0^1(\Omega). \end{cases}$$

It is important to note that problem (6.4) is in fact independent of r, in the sense that, after division by r, the bilinear form in (6.4) is $\{v,w\} \to \int_\Omega \nabla v\cdot\nabla w\,dx$, i.e. that associated with the homogeneous Dirichlet problem for the operator $-\Delta$. Further, minimising over $(L^2(\Omega))^2$ the functional $q \to \mathcal{L}_r(u,q,\lambda)$ reduces to the minimisation over $(L^2(\Omega))^2$ of the functional

$$\text{(6.5)}\qquad \frac{\nu+r}{2}\int_\Omega |q|^2\,dx + g\int_\Omega |q|\,dx - \int_\Omega (r\nabla u+\lambda)\cdot q\,dx,$$

and this reduces to solving a.e. in $x \in \Omega$ and with ξ describing $\mathbb{R}^2$, for the functional

$$\text{(6.6)}\qquad \frac{\nu+r}{2}|\xi|^2 + g|\xi| - (r\nabla u(x)+\lambda(x))\cdot\xi.$$

In the discrete case the solution is performed triangle by triangle, or at the quadrature points used for evaluating $\int_\Omega |\nabla v| dx$ or $\int_\Omega |q|\ dx$. Putting $d(x) = r\nabla u(x) + \lambda(x)$, the solution $\bar{\xi}$ of the above minimisation problem is given by

$$(6.7) \qquad \begin{cases} \bar{\xi} = 0 \ \textit{if}\ g \geq |d|, \\ \bar{\xi} = \frac{1}{\nu+r}\left(d - g\frac{d}{|d|}\right) \ \textit{if}\ g < |d|. \end{cases}$$

The decomposition associated with the constraint $q - \nabla v = 0$ and with the augmented Lagrangian (6.3) has thus allowed us to eliminate any difficulty associated with the nondifferentiable term $\int_\Omega |\nabla v|\ dx$.

Remark 6.1: If we refer back to Remark 3.3 of Section 3, it would be advisable, in order to improve the convergence of ALG1 and ALG2, to use a penalty term of the form $\frac{r}{2}\int_\Omega \eta(x)|\nabla v - q|^2\ dx$, where, *formally*, η represents an estimate of $\nu + \frac{g}{|\nabla u|}$. This expression is not applicable here, since it becomes *infinite*; nonetheless this leads us to think that the penalty term should be very large in the rigid zones, i.e. those in which $\nabla u = 0$. Since the rigid zones increase with g, we should expect to see[9] the optimal value of r (in ALG2 particularly) increase with g.

Remark 6.2: In the zones where $g \geq |d|$, which after convergence correspond to the rigid zones of the problem, we have, from (6.7), $|p| = 0$. Inserting this result into (6.4) it can readily be seen that in these regions we are in fact solving

$$(6.8) \qquad -\Delta u = \frac{1}{r}(f + \nabla\cdot\lambda)\ ;$$

taking r to be large in these regions in fact amounts to forcing $-\Delta u$ to have the value zero.

6.1.3 *Numerical results*

The numerical results of MARROCCO [1], which we shall discuss

[9] All this assumes that we are using $\rho = r$ in ALG1, ALG2.

very briefly in this section, were obtained in a very simple case for which the exact solution is known; with the domain Ω as the disc of radius 1 centred at the origin, the solution of problem (6.2) is given, for $f = C$ (> 0), by

$$(6.9)\qquad \begin{cases} u \equiv 0 \ \textit{if}\ g > g_c = \frac{C}{2} \\ u(x) = (\frac{1-|x|}{2})(\frac{C}{2}(1+|x|)-2g) \ \textit{if}\ R' \le |x| \le 1, \\ u(x) = (\frac{1-R'}{2})(\frac{C}{2}(1+R')-2g) \ \textit{if}\ 0 \le |x| \le R', \end{cases}$$

where, in (6.9), $R' = \frac{2g}{C}$ (if $g \le g_c$) and $|x| = \sqrt{x_1^2 + x_2^2}$ if $x = \{x_1, x_2\}$. An approximation by C^0-conforming finite elements of order 1 was used, the corresponding triangulation $\mathcal{T}_h$ comprising 256 triangles.

The calculations were performed for $g = 2,5,8$, the rigid zone then being the circle of radius $R' = 0.2$, 0.5, 0.8, respectively.

If we consider ALG2 (with $\rho = r$), the optimal value of r is 5×10^{-2} (resp. 1.0, 7.0) for $g = 2$ (resp. 5, 8), the corresponding numbers of iterations being respectively 10, 25, 50 for a termination test which relates solely to the convergence of the sequence $\{u_h^n\}_{n\ge 0}$.

In all the cases considered, ALG1 (with $\rho = r$) performs less effectively than ALG2, and this is true even for problems in which the nonlinearity is very large, i.e. g is large ($g = 8$ for example). The calculations confirm that, for ALG1, the convergence of the relaxation iterations is slow in the rigid zones. Decoupling between the convergence of $\{u_h^n\}_{n\ge 0}$ and that of $\{p_h^n\}_{n\ge 0}$ and $\{\lambda_h^n\}_{n\ge 0}$ is also evident; the value of u_h^n depends only on the components of p_h^n and λ_h^n in the space ∇V_h, where $V_h = \{v_h | v_h \in C^0(\bar{\Omega}_h), v_h|_K \in P_1 \ \forall K \in \mathcal{T}_h,\ v_h = 0 \text{ on } \partial\Omega_h\}$ [10], and it would certainly appear that these components converge rapidly and do not require an accurate solution for $\{u_h^n, p_h^n\}$ to be obtained at each iteration of ALG1.

A powerful algorithm (in terms of the number of iterations) would undoubtedly be ALG2 with r made to increase during the course of the calculation so as to accelerate convergence in the rigid zones.

---

[10] $\Omega_h = \bigcup_{K \in \mathcal{T}_h} \mathring{K}$ ($\mathring{X}$ = interior of X).

Remark 6.3: In Chapter VII of this book we describe the application of the augmented Lagrangian methods of Chapter III to the solution of problems involving the flow of Bingham fluids which are much more complicated than those considered in this section; in fact, by switching to the stream function, we obtain *variational inequalities of order* 4[11], whereas problem (6.2) is a *variational inequality of order* 2.

6.2 Elastoplastic torsion of a cylindrical bar

6.2.1 *Formulation of the problem*

The physical motivation of the problem is as follows:

We consider a cylindrical bar of infinite length and with cross section Ω, made of an isotropic elastic/perfectly-plastic material, the threshold of plasticity (i.e. the yield stress) being given by the von Mises criterion. Starting from an unstressed initial state, an increasing torsional couple is applied to the bar, the torsion being characterised by the angle of twist per unit length, denoted by C in the following notes.

We can then reduce this problem (see GERMAIN [1] for a detailed analysis) to seeking a function u, the so-called *stress potential* (defined to within an additive constant). For slightly greater generality we shall assume that Ω is ℓ-connected (if $\ell = 0$, Ω is simply connected); Figure 6.1 illustrates a situation in which $\ell = 3$. We denote by Ω^* the domain obtained by the union of Ω and the ω_i, $i = 1,\dots\ell$. We next define

$$\tilde{K} = \{q \mid q \in (L^2(\Omega^*))^2,\ |q| \leq 1 \text{ a.e., } q=0 \text{ on } \omega_i,\ i=1,\dots\ell\}, \tag{6.10}$$

then

$$K = \{v \mid v \in H_0^1(\Omega^*),\ \nabla v \in \tilde{K}\}\ , \tag{6.11}$$

and finally the functional $J : H_0^1(\Omega^*) \to \mathbb{R}$ by

$$J(v) = \frac{1}{2}\int_{\Omega^*} |\nabla v|^2\ dx - C\int_{\Omega^*} v\ dx. \tag{6.12}$$

---

[11] i.e. relative to an elliptic operator of order 4.

The stress potential u, mentioned above, is then the solution of the following problem in the Calculus of Variations (in some appropriate system of units):

$$(6.13)\qquad \begin{cases} u \in K, \\ J(u) \leq J(v) \quad \forall v \in K, \end{cases}$$

this itself being equivalent to the variational inequality problem

$$(6.14)\qquad \begin{cases} u \in K, \\ \int_{\Omega^*} \nabla u \cdot \nabla (v-u)\, dx \geq C \int_{\Omega^*} (v-u)\, dx \quad \forall v \in K. \end{cases}$$

The essential difficulty with (6.13), (6.14) stems from the constraint of belonging to K.

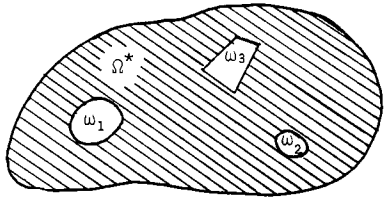


Figure 6.1

6.2.2 *Solution of* (6.13), (6.14) *by augmented Lagrangian methods*

We introduce the augmented Lagrangian $\mathcal{L}_r : H_o^1(\Omega^*) \times (L^2(\Omega^*))^2 \times (L^2(\Omega^*))^2 \to \mathbb{R}$ defined by

$$(6.15)\qquad \begin{cases} \mathcal{L}_r(v,q,\mu) = \frac{1}{2} \int_{\Omega^*} |q|^2 dx - C \int_{\Omega^*} v\, dx + \frac{r}{2} \int_{\Omega^*} |\nabla v - q|^2\, dx \\ + \int_{\Omega^*} \mu \cdot (\nabla v - q)\, dx. \end{cases}$$

We shall determine u by seeking the saddle points of $\mathcal{L}_r$ on $H_o^1(\Omega^*) \times \tilde{K} \times (L^2(\Omega^*))^2$; for this, we employ algorithms ALG1 or ALG2 of Chapter III whose implementation for the solution of the elasto-plastic torsion problem (in the case where Ω is simply connected) was described in Chapter III, Section 3.3.

We suppress the iteration indices; the implementation of ALG1 and ALG2 requires the solution (simultaneous or sequential, depending on the case in question) of the following equations and inequalities, with λ fixed :

$$(6.16)\qquad \begin{cases} r\int_{\Omega^*} \nabla u\cdot\nabla v \, dx = C\int_{\Omega^*} v \, dx + \int_{\Omega^*}(rp-\lambda)\cdot\nabla v \, dx \quad \forall v \in H_o^1(\Omega^*), \\ u \in H_o^1(\Omega^*), \end{cases}$$

$$(6.17)\qquad \begin{cases} (1+r)\int_{\Omega^*} p\cdot(q-p)\, dx \geq \int_{\Omega^*}(r\nabla u+\lambda)\cdot(q-p)dx \quad \forall q \in \tilde{K}, \\ p \in \tilde{K}. \end{cases}$$

Equation (6.17) is solved pointwise in explicit fashion since

$$(6.18)\qquad \begin{cases} p = 0 \quad in \ \omega_i, \ i=1,\dots \ell, \\ p = \dfrac{\lambda+r\nabla u}{\sup(1+r,|\lambda+r\nabla u|)} \quad in \ \Omega, \end{cases}$$

so that our decomposition method has eliminated the difficulties directly related to the von Mises criterion $|\nabla u| \leq 1$.

In practice, (6.18) is solved at a certain number of points depending on the discretisation used. For finite-element approximations in which the functions u and v in (6.13), (6.14) are approximated by piecewise-linear functions, (6.18) is solved triangle-by-triangle to obtain the two constant components of p . In the general case the points are chosen to correspond with a quadrature formula which is *exact* for integrating terms of the form $\int_{\Omega^*} p\cdot q \, dx$. This corresponds to introducing an approximate convex set $\tilde{K}_h$ whose *support function* (see EKELAND-TEMAM [1] for this concept) approximates the support function of K , i.e. $\int_\Omega |q| \, dx$ (in the case without any holes) by the use of the chosen quadrature method.

Remark 6.4: It is not possible to apply Remark 3.3 of Section 3 to this example; it is, however, easy to see that in the plastic zones (where $|p| = |\nabla u| = 1$), taking r to be large will, in view of (6.16), force $-\nabla u + \nabla\cdot p$ to vanish. We can therefore expect an optimal value of r which will increase with C , since increasing the twist angle causes an enlargement of the plastic zones.

6.2.3 *Numerical results*

The numerical results obtained by MARROCCO [1] (with $\rho = r$) confirm the results of Section 6.1.3 relating to the flow of a Bingham fluid in a cylindrical duct. Algorithm ALG2 in fact performs better than ALG1 and once again the regions of Ω where the convergence is slowest are those in which the nonlinear effects manifest themselves, that is, in the case of the torsion problem, the regions where $|\nabla u| = 1$. The numerical tests were performed with $\Omega =]0,1[\times]0,1[$ and $C = 10$ and these show that the convergence rate - measured by the number of iterations - is more or less independent of the discretisation; this is shown by Figure 6.2 (in relation to ALG2) in which the curves 1,2,3 correspond respectively to a triangulation $\mathcal{T}_h$ with 128, 512 and 2048 triangles. These curves indicate the number of iterations required for convergence, as a function of r.

One of the consequences of the extremely weak dependence of ALG1 and ALG2 on the choice of h is that it is possible to determine the optimal r on a coarse mesh, and then to use the optimal r thus obtained for calculations on a much finer mesh.

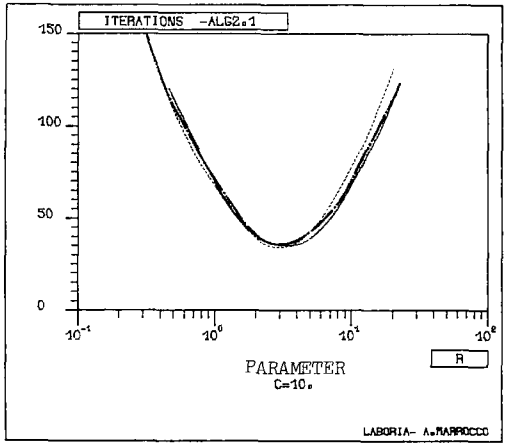


Figure 6.2

Remark 6.5: Chapter VI describes the application of ALG1 and ALG2 to the numerical solution of an elastoplasticity problem which is much more complicated than that discussed in the present section; nonetheless, the basic principles of solution using an augmented Lagrangian remain the same, and are once again based on the general concepts developed in Chapter III.

6.3 Application to the solution of the minimal surfaces problem

6.3.1 *Formulation of the problem*

In this we will be considering the application of the general methods of Chapter III to the solution of a problem which once again falls - formally, at least - within the general framework defined in Section 2 of the present chapter; this is a particularly simple (as far as its formulation is concerned) *minimal surfaces* problem. We thus consider the contour C in $\mathbb{R}^3$, defined using a domain Ω of $\mathbb{R}^2$ (with boundary Γ), by

$$C = \{\{x, g(x)\} \in \mathbb{R}^3,\ x \in \Gamma,\ g(x) \in \mathbb{R}\}\ , \tag{6.19}$$

and the functional

$$J(v) = \int_\Omega \sqrt{1+|\nabla v|^2}\ dx. \tag{6.20}$$

The minimal surfaces problem is then defined by

$$\begin{cases} u \in V_g \\ J(u) \leq J(v) \quad \forall v \in V_g \end{cases} \tag{6.21}$$

where

$$V_g = \{v \in W^{1,1}(\Omega),\ v|_\Gamma = g\}. \tag{6.22}$$

We are here dealing with a nontrivial problem since - among other difficulties - the space $W^{1,1}(\Omega)$ is not reflexive; we have to consider (see EKELAND-TEMAM [1]) *generalised solutions*, and the condition $u|_\Gamma = g$ cannot be satisfied in the usual sense, even for very regular boundaries Γ and functions g. The treatment which follows is therefore formal, and is totally justified only for *discretised* problems (which then fall within the context of Theorems 4.2 and 5.2 of Chapter III, Sections 4 and 5, respectively).

6.3.2 *Solution of problem (6.21) by augmented Lagrangian algorithms*

We introduce the augmented Lagrangian

$$\mathcal{L}_r(v,q,\mu) = \int_\Omega \sqrt{1+|q|^2}\, dx + \frac{r}{2}\int_\Omega |\nabla v - q|^2\, dx + \int_\Omega \mu\cdot(\nabla v - q)\, dx. \tag{6.23}$$

To determine u we shall thus seek (formally in infinite dimensions) the saddle points of $\mathcal{L}_r$ on $(V_g \cap H^1(\Omega)) \times (L^2(\Omega))^2 \times (L^2(\Omega))^2$ by algorithms of the type ALG1, ALG2. We shall therefore be led to solve, at each iteration, *simultaneously* for ALG1 and *sequentially* for ALG2, the following nonlinear system (we omit the iteration indices), with λ fixed :

$$\begin{cases} r\displaystyle\int_\Omega \nabla u\cdot\nabla v\, dx = \int_\Omega (rp-\lambda)\cdot\nabla v\, dx \quad \forall v \in H_o^1(\Omega), \\ u \in V_g \cap H^1(\Omega), \end{cases} \tag{6.24}$$

$$\int_\Omega \left(\frac{p}{\sqrt{1+|p^2|}} + rp\right)\cdot q\, dx = \int_\Omega (r\nabla u+\lambda)\cdot q\, dx \quad \forall q \in (L^2(\Omega))^2,\ p \in (L^2(\Omega))^2. \tag{6.25}$$

The nonlinear equation can be solved point by point; putting $z = |p|$, we first have to solve the following nonlinear equation in one variable :

$$\left(\frac{1}{\sqrt{1+z^2}} + r\right)z = |r\nabla u+\lambda|, \tag{6.26}$$

for which Newton's method may be applied without difficulty. Depending on the type of approximation used, we solve (6.26) either element-by-element, or at quadrature points, as for the nonlinear problems described in the preceding sections of the present chapter.

Remark 6.6: On the basis of Remark 3.3 of Section 3, we should in this case use a penalty term of the form $\int_\Omega \eta(x)\,|\nabla v - q|^2\, dx$, with $\eta(x)$ an estimate of $(1+|\nabla u|^2)^{-\frac{1}{2}}$; this term is small when $|\nabla u|$ is large. We must therefore expect that the optimal value of r for ALG2 (with $\rho = r$), and in the case of the Lagrangian (6.23), will be less than unity.

6.3.3 *Numerical results*

The results of MARROCCO [1], which we summarise briefly here, relate to the case where Ω is the circular corona defined by :

(6.27) $$\Omega = \{x | x = \{x_1, x_2\} \in \mathbb{R}^2,\ 1 < \sqrt{x_1^2+x_2^2} < 4\} \ ;$$

the boundary conditions are $g(x) = 0$ on the circle of radius 4 and $g(x) = B$ (= const.) on the circle of radius 1. Since the solution is axisymmetric it is easily calculated, and with $|x| = (x_1^2 + x_2^2)^{\frac{1}{2}}$ it is given by

(6.28) $$u(x) = A(\text{Arg ch } \frac{|x|}{A} - \text{Arg ch } \frac{4}{A}),$$

the constant A having to be determined from the value of B (knowing that $u(x) = B$ if $|x| = 1$). This classical solution exists only for B less than a critical value $B_c \simeq 2.07$. If $B > B_c$, then the solution 'breaks down' in the sense that the condition $u = B$ can no longer be satisfied on the circle of radius 1. Numerically, if we discretise by means of C^0-conforming finite elements, this shows itself (see Figure 6.3 where cross-sections of an approximate solution are shown) as a very large gradient near the boundary where the aforementioned 'breakdown' phenomenon occurs. In view of Remark 6.6 we should expect, in the case of ALG2 (with $\rho = r$), to see the optimal value of r decrease as B increases; this is confirmed by the numerical tests, since for $B = 1$ (resp. 2.07, 4) the optimal value of r is close to 1 (resp. 0.2, 0.1); the corresponding numbers of iterations are 20, 30 and 50, respectively. We thus see that Remark 3.3 of Section 3 has enabled this phenomenon to be predicted even though it runs counter to the numerical experiments of the earlier sections of the present chapter where, other things being equal, the optimal value of r increased when the nonlinear effects became more significant.

As far as ALG1 is concerned, it turns out once again to be more expensive than ALG2.

Remark 6.7: JOURON [1] gives a detailed account of the approximation of the minimal-surfaces problem by means of methods using conforming finite elements of order 1, and of their iterative solution by nonlinear overrelaxation methods (see also JOHNSON-THOMEE [1] and CIARLET [1, Chapter 5] for finite-element approximations of the minimal-surfaces problem).

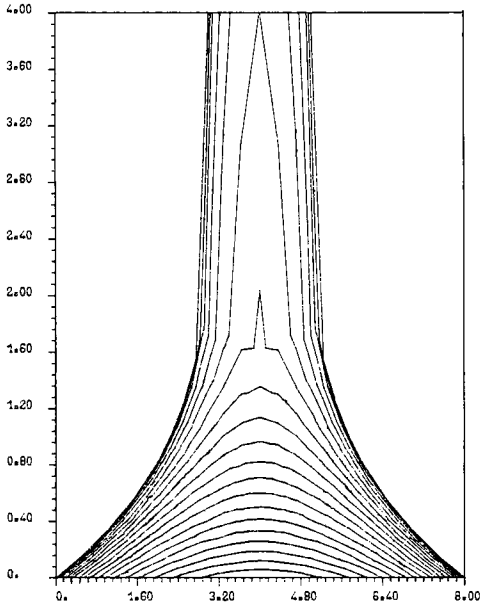


Figure 6.3

7. DISCUSSION ON CHAPTER V

In this chapter we have applied the methods of Chapter III to the solution of problems with various physical origins. We have thereby been able to demonstrate the fact that the decomposition of a non-linear problem through the introduction of an augmented Lagrangian is a robust method, readily adaptable to numerous situations (we shall be seeing further examples of this in Chapters VI, VII & VIII). This robustness and this generality enable the augmented-Lagrangian methods to be used for the efficient solution of numerous types of problems.

Two phenomena worthy of our attention have become apparent: the first of these concerns the partial decoupling between the convergence of the sequence $\{u^n\}_{n\geq 0}$ approximating the unknown function u, and that of the sequence $\{p^n\}_{n\geq 0}$ approximating ∇u; this decoupling is total in the linear case if we put $\rho = r$. This is one aspect of algorithms ALG1 and ALG2 which would merit a more detailed investigation : it is this decoupling which partially explains the superior-

ity, for this type of problem at least, of ALG2 over ALG1; this is true even for problems in which the nonlinearity is very strong. We had in fact observed the opposite situation in Chapter IV; the difference obviously relates to the fact that the image under ∇ of the space V in which u is sought is a strict, closed subspace of $(L^2(\Omega))^N$ and we can have very rapid convergence of the component of p^n which belongs to this space and yet slow convergence of the sequence $\{p^n\}_{n\geq 0}$. It would be interesting to analyse this phenomenon in more detail, with a view to developing algorithms which exploit this feature as far as possible.

The second phenomenon is related to the rôle which could be played by a penalty term, with variable coefficients, of the form $\int_\Omega \eta(x)\,|\nabla v - q|^2\,dx$, for accelerating convergence. A considerable advantage would lie in the fact that for ALG2, with $\rho = r$ and η suitably chosen, the optimal r would be close to 1. In reality, the choice of η requires an *a priori* knowledge of the solution. Thus, as one possibility, we can consider algorithms which involve updating $\eta(x)$ during the course of the calculation; if the linear systems are solved by *direct methods*, however, such an update would require the factorisation of a new matrix, which is a relatively expensive operation. This drawback would disappear if *powerful iterative methods* could be used to solve these linear systems; amongst the methods which can be considered, we may list preconditioned conjugate-gradient methods, multigrid methods, etc. Secondly, there exist situations for which a family of similar problems has to be solved, differing only through the values of a few parameters. In such cases, it would be possible to use a function η derived from a mean solution, or to employ a strategy of gradually increasing the parameters, with an update of η when the solution has changed sufficiently. We have also been able to use the existence of an optimal coefficient η having a certain form - related to the behaviour of ∇u - to predict, at least qualitatively, the corresponding behaviour of the optimal parameter r with the coefficient η taken equal to 1.

Finally, we should point out that for the problems treated in the present chapter, the choice of $\mathcal{L}_r$ makes it possible to make r vary without having to refactorise the matrix of the linear system which occurs during the calculation; in certain cases it would certainly be helpful if we could make r vary in an effective manner, though the precise means of doing this remains to be determined.

The magnetostatic problems investigated in Section 4 are of great industrial importance (transformers, rotating machines, electromagnets in particle accelerators, read/write heads for disks and magnetic tapes, etc.); since the formulation used in Section 4 is by no means the only one possible, we consider it necessary, in view of the importance of the subject, to indicate a few other formulations and to make a number of observations on the associated augmented-Lagrangian algorithms.

Following, for example, MUNRO [1] we can, in magnetostatics, define the functions U_c and U by, respectively,

$$\left\{ \begin{array}{l} U_c(\vec{H}) = \int_{\vec{0}}^{\vec{H}} \vec{B}(\vec{h}) \cdot d\vec{h}, \\ (U_c : \textit{complementary magnetic energy per unit volume}), \end{array} \right. \tag{7.1}$$

$$\left\{ \begin{array}{l} U(\vec{B}) = \int_{\vec{0}}^{\vec{B}} \vec{H}(\vec{b}) \cdot d\vec{b}, \\ (U : \textit{stored magnetic energy per unit volume}) \end{array} \right. \tag{7.2}$$

Suppose that $\vec{j} = \vec{0}$ in (4.1); then there exists ϕ such that

$$\vec{H} = \vec{\nabla}\phi, \tag{7.3}$$

i.e. $\vec{H}$ derives from a *scalar potential*; it is convenient in this case to use U_c, which gives the energy functional

$$\mathcal{J}_s = \int_\Omega U_c \, dx. \tag{7.4}$$

If $\vec{j} \neq \vec{0}$, then $\vec{H}$ no longer derives from a scalar potential, but $\vec{\nabla} \cdot \vec{B} = 0$ implies the existence of a vector potential $\vec{A}$ such that $\vec{\nabla} \times \vec{A} = \vec{B}$; it is then more convenient to use $\vec{A}$ (this is what was done in Section 4 in the case where $\vec{A} = \{0,0,A\}$), the energy functional to be used being defined by

$$\mathcal{J}_v = \int_\Omega U \, dx - \int_\Omega \vec{j} \cdot \vec{A} \, dx. \tag{7.5}$$

The augmented Lagrangians associated with the above two situations are defined as follows:

(i) When $\vec{j} = \vec{0}$ and when the *scalar potential* ϕ is used, we obviously put $\vec{H} = \vec{\nabla}\phi$ and we *penalise* and *dualise* this *linear* constraint so as to obtain the augmented Lagrangian $\mathcal{L}_r^s$ defined by

$$(7.6)\qquad \mathcal{L}_r^s(\phi,\vec{H},\vec{\mu}) = \int_\Omega U_c(\vec{H})dx + \frac{r}{2}\int_\Omega \eta(x)\,|\vec{\nabla}\phi-\vec{H}|^2\,dx + \int_\Omega \eta(x)\vec{\mu}\cdot(\vec{\nabla}\phi-\vec{H})dx.$$

In implementing the above algorithms ALG1, ALG2 we obtain the following equations (the iteration indices have been omitted) :

$$(7.7)\qquad \begin{cases} \phi \in V, \\ r\int_\Omega \eta(x)\vec{\nabla}\phi\cdot\vec{\nabla}v\,dx = \int_\Omega \eta(x)(r\vec{H}-\vec{\lambda})\cdot\vec{\nabla}v\,dx \quad \forall v \in V_o \end{cases}$$

(V is a subspace of $H^1(\Omega)$ which takes into account the boundary conditions, and V_o is the associated test-function space, corresponding to homogeneous boundary conditions),

$$(7.8)\qquad \begin{cases} \int_\Omega (\vec{B}(\vec{H})+r\eta\vec{H})\cdot\vec{q}\,dx = \int_\Omega \eta(r\vec{\nabla}\phi+\vec{\lambda})\cdot\vec{q}\,dx \quad \forall \vec{q} \in (L^2(\Omega))^N, \\ \vec{H} \in (L^2(\Omega))^N. \end{cases}$$

At the numerical integration points or triangle-by-triangle, depending on the approximation used, (7.8) leads to the following vector equation (in $\mathbb{R}^2$ or $\mathbb{R}^3$ depending on the dimension N of the problem) :

$$(7.9)\qquad \begin{cases} \vec{H} \in \mathbb{R}^N, \\ \vec{B}(\vec{H}) + r\eta\vec{H} = \vec{C}, \end{cases}$$

where $\vec{C}$ is a known vector. In the case where the material is isotropic, we can reduce the solution of (7.9) to that of a nonlinear equation in $\mathbb{R}_+$ giving $|\vec{H}|$ (like (4.16) in Section 4).

(ii) When $\vec{j} \neq \vec{0}$ and when the *vector potential* $\vec{A}$ is used, we penalise and dualise $\vec{B} = \vec{\nabla}\times\vec{A}$; this gives the augmented Lagrangian

$$(7.10)\qquad \begin{cases} \mathcal{L}_r^v(\vec{A},\vec{B},\vec{\mu}) = \int_\Omega U(\vec{B})dx + \frac{r}{2}\int_\Omega \eta(x)\,|\vec{\nabla}\times\vec{A}-\vec{B}|^2\,dx \\ \qquad\qquad + \int_\Omega \eta(x)\vec{\mu}\cdot(\vec{\nabla}\times\vec{A}-\vec{B})dx - \int_\Omega \vec{j}\cdot\vec{A}\,dx. \end{cases}$$

which includes the augmented Lagrangian defined by (4.13) in Section 4.2 as a particular case. The equations corresponding to (7.7),(7.8) are then, respectively,

$$(7.11)\qquad \begin{cases} \vec{A} \in \vec{V}, \\ r\int_\Omega \eta(\vec{\nabla}\times\vec{A})\cdot(\vec{\nabla}\times\vec{v})dx = \int_\Omega \vec{j}\cdot\vec{v}\,dx + \int_\Omega \eta(r\vec{B}-\vec{\lambda})\cdot(\vec{\nabla}\times\vec{v})dx \quad \forall \vec{v} \in \vec{V}_o, \end{cases}$$

$$(7.12)\qquad \begin{cases} \vec{B} \in (L^2(\Omega))^N, \\ \int_\Omega (\vec{H}(\vec{B})+r\eta\vec{B})\cdot\vec{q}\,dx = \int_\Omega \eta(r\vec{\nabla}\times\vec{A}+\vec{\lambda})\cdot\vec{q}\,dx \quad \forall \vec{q} \in (L^2(\Omega))^N. \end{cases}$$

As before, (7.12) is solved triangle-by-triangle or at the numerical integration points, depending on the approximation chosen, by solving the following in $\mathbb{R}^N$:

$$(7.13)\qquad \begin{cases} \vec{B} \in \mathbb{R}^N, \\ \vec{H}(\vec{B}) + r\eta\vec{B} = \vec{C}. \end{cases}$$

The solution of (7.11), on the other hand, can pose a number of difficulties and it is convenient to distinguish the cases N = 2 and N = 3; if $\vec{A} = \{0,0,A\}$ this leads to a problem in $\mathbb{R}^2$ and, as we saw in Section 4, the relation

$$(7.14)\qquad (\vec{\nabla}\times\vec{A})\cdot(\vec{\nabla}\times\vec{v}) = \nabla A\cdot\nabla v \quad (if\ \vec{v} = \{0,0,v\})$$

reduces the solution of (7.11) to that of a linear elliptic problem of second order and of standard type. For N = 3, problem (7.11) is in general ill-posed since the semi-norm

$$\vec{v} \to \|\vec{\nabla}\times\vec{v}\|_{(L^2(\Omega))^3}$$

is not a norm on $(H^1(\Omega)/\mathbb{R})^3$; this is due to the fact that

$$\vec{\nabla}\times(\vec{v}+\vec{\nabla}\phi) = \vec{\nabla}\times\vec{v} \quad \forall\phi,$$

which means that the vector potential is in general defined only to within a gradient (see DURAND [1]). It then follows that the functional $\mathcal{J}_V$ given by (7.4) is not coercive in $(H^1(\Omega)/\mathbb{R})^3$. It is shown in MARROCCO [2] that a functional space adapted to 3-dimensional magnetostatic problems is the following:

$$(7.15)\qquad W = \{\vec{v} \mid \vec{v} \in (H^1(\Omega))^3,\ \vec{v}\times\vec{n} = \vec{0}\ on\ \partial\Omega\},$$

and that it is sufficient to add to the function $\mathfrak{J}_v$ a term of the type $\frac{1}{2}\int_\Omega \alpha(x)|\nabla\cdot\vec{A}|^2 dx$ (with $\alpha(x) \geq \alpha_o > 0$) in order to make it coercive on W. The functional $\mathfrak{J}_v$ thus corrected admits a unique minimum on W, the corresponding vector potential $\vec{A}$ satisfying the Maxwell equations of magnetostatics, as well as the condition $\vec{\nabla}\cdot\vec{A} = 0$. If we add the above term to the Lagrangian $\mathcal{L}_r^v$ defined by (7.10), we obtain in place of (7.11)

$$(7.16)\qquad \left\{\begin{array}{l} \vec{A}\in W, \\ r\int_\Omega \eta(\vec{\nabla}\times\vec{A})\cdot(\vec{\nabla}\times\vec{v})dx + \int_\Omega \alpha(\vec{\nabla}\cdot\vec{A})(\vec{\nabla}\cdot\vec{v})dx = \\ \int_\Omega \vec{j}\cdot\vec{v}\ dx + \int_\Omega \eta(r\vec{B}-\vec{\lambda})\cdot(\vec{\nabla}\times\vec{v})dx \quad \forall\vec{v}\in W\ ; \end{array}\right.$$

it is reasonable to take $\alpha = r\eta$, in which case, for certain geometries (if Ω is a parallelepiped, for example), (7.16) can be decomposed into three problems of Dirichlet type (one for each component of $\vec{A}$); equation (7.12) remains unchanged.

The addition of the term $\frac{1}{2}\int_\Omega \alpha|\vec{\nabla}\cdot\vec{A}|^2\ dx$ to $\mathfrak{J}_v$ and $\mathcal{L}_r^v$ may be considered as a penalisation of the condition $\vec{\nabla}\cdot\vec{A} = 0$; it is therefore natural to think of associating a Lagrange multiplier with this constraint; this leads to the augmented Lagrangian $\mathcal{L}_r^v$ defined (if $\alpha = r\eta$) by

$$(7.17)\qquad \mathcal{L}_r^v(\vec{A},\vec{B},\vec{\mu},q) = \mathcal{L}_r^v(\vec{A},\vec{B},\vec{\mu}) + \frac{r}{2}\int_\Omega \eta|\vec{\nabla}\cdot\vec{A}|^2\ dx + \int_\Omega \eta q\ \vec{\nabla}\cdot\vec{A}\ dx.$$

All the above reminds us of the Stokes problem in Chapter II, the function ηq playing the rôle of a *pressure*. In the implementation of algorithms ALG1, ALG2 in relation to the Lagrangian (7.17), equation (7.12) remains unchanged; as regards the equation in $\vec{A}$, this becomes

$$(7.18)\qquad \left\{\begin{array}{l} \vec{A}\in W, \\ r\int_\Omega \eta(\vec{\nabla}\times\vec{A})\cdot(\vec{\nabla}\times\vec{v})\ dx + r\int_\Omega \eta(\vec{\nabla}\cdot\vec{A})(\vec{\nabla}\cdot\vec{v})dx = \\ \int_\Omega \vec{j}\cdot\vec{v}\ dx + \int_\Omega \eta(r\vec{B}-\vec{\lambda})\cdot(\vec{\nabla}\times\vec{v})dx - \int_\Omega \eta q\vec{\nabla}\cdot\vec{v}\ dx \quad \forall\vec{v}\in W. \end{array}\right.$$

From the point of view of approximation by finite elements, the aforementioned analogy with the Stokes problem suggests using, amongst others, *nonconforming* finite elements P_1 of the type used earlier in

Chapter II for approximating Stokes and Navier-Stokes problems (in 3 dimensions, $\mathcal{C}_h$ will be a family of *tetrahedra* and the associated *degrees of freedom* will be the values taken at the centres of the faces of these tetrahedra by the approximations $\vec{A}_h$, $\vec{v}_h$ of $\vec{A}$ and $\vec{v}$). (See MARROCCO [2] for more details on these nonconforming approximations and on the corresponding numerical experiments).

We shall conclude this chapter with a few bibliographic comments. We have already pointed out that the nonlinear operator

$$v \to -\nabla\cdot(|\nabla v|^{s-2}\nabla v)$$

of Section 3, has appeared in mathematical models in *glaciology* and we refer the reader to PELISSIER [1] and the associated bibliography; the numerical augmented-Lagrangian treatment was introduced by GLOWINSKI-MARROCCO [1] and developed in MARROCCO [1]. For the magnetostatic problem of Section 4, the reader may refer to GLOWINSKI-MARROCCO [5]. The potential-flow problems of Section 5 are classical, and the reader interested in the fluid-mechanical aspects of these problems may refer to LANDAU-LIFCHITZ [1]. The viscoplasticity and elastoplasticity problems of Section 6 are treated, in particular, in DUVAUT-LIONS [1], and the numerical treatment of the corresponding variational inequalities is described in detail in GLOWINSKI-LIONS-TREMOLIERES [1],[2] and GLOWINSKI [1],[2]; the case of the torsion of a cylindrical bar with multi-connected cross-section is treated in GLOWINSKI-LANCHON [1]. In connection with minimal-surface problems we have already cited EKELAND-TEMAM [1] in which the concept of a *generalised solution* is discussed; once again, this is a classical problem which has given rise to numerous works.

CHAPTER VI

APPLICATION OF ALGORITHM ALG2 TO A TWO-DIMENSIONAL ELASTOPLASTICITY PROBLEM

B. Mercier

INTRODUCTION

We shall now consider a new example arising from the Mechanics of Continuous Media. In comparison to the preceding examples, in particular Examples 1 and 2 of Chapter III, the situation will be somewhat different: in fact the functional F will be *noncoercive* (but differentiable); as G is linear, the problem (P) will be *'noncoercive'* and we will not be able to prove the existence of a solution in infinite dimensions. In contrast, the dual of (P), which involves F*, the conjugate of F, is *well posed* as F* is coercive in the example which we are considering, and this also implies the differentiability of F. Furthermore, as we saw in Chapter V in connection with other problems, in order to improve the rate of convergence of algorithm ALG2, there is an advantage to be gained in this example by choosing a *penalty term* in the augmented Lagrangian $\mathcal{L}_r$ which is not equal to the square of the natural norm on H, but which is associated with *another quadratic form*.

1. THE CONTINUOUS PROBLEM

We consider a continuous elastoplastic medium held fixed on one part of its boundary. We seek the stress field σ and the displacement field u which are set up in the continuous medium when it is subjected to external forces (the above situation is illustrated in Figure 1.1).

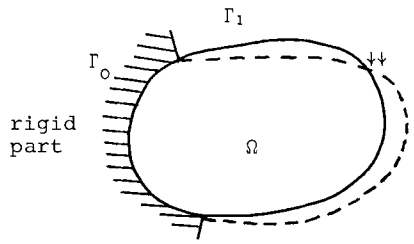


Figure 1.1 - The continuous medium before and after application of the external loads

Let $\Omega \subset \mathbb{R}^d$ (d = 2 or 3 in applications) be the open bounded domain with sufficiently regular boundary, representing the continuous medium; we seek the stress field σ in the space

$$H = \{ \tau = (\tau_{ij}),\ \tau_{ij} \in L^2(\Omega),\ \tau_{ij} = \tau_{ji},\ 1 \le i,j \le d \} .$$

We denote by Γ_o the part of the boundary where the continuous medium is *fixed* and by

$$V = \{ v \in (H^1(\Omega))^d ,\ v = 0 \ \textit{on}\ \Gamma_o \}$$

the space of *admissible displacements*.

The operator B : V → H is defined as follows:

$$(Bv)_{ij} = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i} \right) ,\ 1 \le i,j \le d \ ;$$

Bv represents the *linearised strain tensor*. From *Korn's inequality*, which is proved in DUVAUT-LIONS [1], B is of *closed image*.

We introduce a symmetric automorphism Λ of $\mathbb{R}^{d^2}$ satisfying

$$(1.1) \qquad \begin{cases} (\Lambda\phi)\cdot\tau = \phi\cdot(\Lambda\tau) & ,\ \forall\, \phi,\tau \in \mathbb{R}^{d^2} , \\ |\Lambda\tau|^2 \ge \alpha|\tau|^2 & ,\ \forall\ \tau \in \mathbb{R}^{d^2} , \end{cases}$$

where $\alpha > 0$; $|\cdot|$ and $\cdot$ denote the norm and the Euclidian inner product on $\mathbb{R}^d$. This automorphism takes into account the elasticity coefficients, so that the *energy* of the continuous medium is written

$$\frac{1}{2}\int_\Omega (\Lambda\sigma).\sigma \ dx .$$

We denote by G(v) the function equal and opposite to the work

done by the external forces in a displacement $v \in V$ of the continuous medium ($G \in V'$, dual of V), and we define

$$E = \{\tau \in H \ , \int_\Omega \tau \cdot Bv \ dx + G(v) = 0 \ , \forall v \in V\} \ ,$$

which is termed the set of *statically-admissible* stress fields.

Finally, we denote by $C \subset \mathbb{R}^{d^2}$ the (closed) *plasticity convex set* and by

$$K = \{\tau \in H \ , \tau(x) \in C \ a.e. \ x \in \Omega\} \tag{1.2}$$

the set of *plastically admissible* stress fields. Hencky's law then states that the stress field σ is the solution of the optimisation problem

$$\min_{\tau \in K \cap E} \frac{1}{2} \int_\Omega (\Lambda\tau) \cdot \tau \ dx \ . \tag{1.3}$$

If the condition $K \cap E \neq 0$ is satisfied, that is, if we are *'below' the limit load, this problem admits a unique solution*, from (1.1). (Note that K and E are *closed*). Since the set E depends linearly on G, if the origin belongs to the interior of C ($0 \in \text{Int } C$), which we shall assume to be the case here, and if the external forces (and hence G) are sufficiently small, then the condition $K \cap E \neq 0$ will be realised.

2. THE PROBLEM (P)

As we have stated above, we shall equip H, not with its natural inner product, but with an inner product related to the energy, namely:

$$(p,q) = \int_\Omega (\Lambda^{-1}p) \cdot q \ dx \ ,$$

and we denote by $\|\cdot\|$ the associated norm. With this notation, the energy to be minimised in (1.3) is written as $\frac{1}{2}\|\Lambda\tau\|^2$ and the set E as:

$$E = \{\tau \in H \ , \ (\Lambda\tau, Bv) + G(v) = 0 \ , \forall v \in V\} \ ,$$

so that

$$\sup_{v\in V}\{-G(v) - (\Lambda\tau, Bv)\} = \begin{cases} 0 \text{ if } \tau \in E, \\ +\infty \text{ otherwise,} \end{cases}$$

and problem (1.3) is equivalent to

$$\min_{\tau\in K}\sup_{v\in V}\{\tfrac{1}{2}\|\Lambda\tau\|^2 - (\Lambda\tau, Bv) - G(v)\}.$$

Its *dual* (obtained by permutation of the minimum and the supremum) is written, after a change of sign, as

$$\inf_{v\in V}\{F(Bv) + G(v)\} \tag{2.1}$$

with

$$F(\mu) = \sup_{q\in K_o}\{(\mu, q) - \tfrac{1}{2}\|q\|^2\} \tag{2.2}$$

and

$$K_o = \{q \in H,\ \Lambda^{-1}q \in K\}. \tag{2.3}$$

This dual is clearly a problem of the form investigated in Chapter III; the function F is the conjugate of the functional $\frac{1}{2}\|q\|^2 + I_{K_o}(q)$, and consequently it is *differentiable*, but *noncoercive* in general. Problem (2.1) therefore does not always admit a solution in infinite dimensions, even if $K \cap E$ is nonempty (see the counter-example in MERCIER [3]).

3. APPROXIMATION BY FINITE ELEMENTS

In practice, we are obliged to reduce the problem to finite dimensions in order to solve (1.3) (or (2.1)). To this end we introduce a family of triangulations $\{\mathcal{C}_h\}_h$, indexed by a parameter $h > 0$; with h given, $\mathcal{C}_h$ is a set of triangles covering Ω [1], satisfying the following properties: let $T, T' \in \mathcal{C}_h$ be two distinct triangles of $\mathcal{C}_h$; then we have

[1] To simplify the description we assume that Ω is a polygon in $\mathbb{R}^2$.

$$\begin{cases} T \cap T' = \emptyset \ , \text{ or} \\ T \cap T' = 1 \text{ one complete common edge} \\ T \cap T' = 1 \text{ one common vertex.} \end{cases}$$

In short, the situation shown in Figure 3.1 is forbidden.

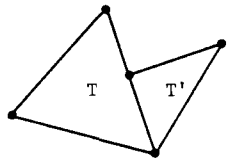


Figure 3.1 : Forbidden situation

The parameter h denotes, for example, the diameter of the largest triangle in $\mathcal{C}_h$. We then denote by $V_h \subset V$ the space of finite elements constituted by *piecewise affine and continuous displacement fields over each triangle of* $\mathcal{C}_h$. We denote by $H_h \subset H$ the subspace of H_h composed of *piecewise-constant tensors over each triangle of* $\mathcal{C}_h$, so that the operator B maps V_h into a part of H_h . We then put

$$E_h = \{\tau_h \in H_h \ , \ (\Lambda\tau_h, Bv_h) + G(v_h) = 0 \ , \forall\, v_h \in V_h\} \ ,$$

and to approximate (1.3) we choose the finite-dimensional problem

$$(3.1) \qquad \operatorname*{Min}_{\tau_h \in K \cap E_h} \frac{1}{2} \|\Lambda\tau_h\|^2$$

which admits a unique solution σ_h . It can be shown that $\sigma_h \to \sigma$ when $h \to 0$ (see MERCIER [3]).

The definition of the dual of (3.1) again depends on considering (on $K \times V$) the Lagrangian

$$(3.2) \qquad \frac{1}{2} \|\Lambda\tau\|^2 - (\Lambda\tau, Bv) - G(v) \ ,$$

where the dual variable u (u_h in finite dimensions) is in this case the *Lagrange multiplier* of the stress $\tau \in E$ ($\tau \in E_h$ in finite dimensions). As the interior of the convex set K is empty in H , it is not possible to deduce the existence of u from this remark. On the other hand, $K \cap H_h$ has a nonempty interior in H_h , thus showing the existence of u_h , from ROCKAFELLAR [4, Section 28], as

long as $E_h \cap (\text{int } K)$ is nonempty (which is a stronger condition than that $K \cap E_h$ be nonempty but which is true if the external forces are *sufficiently small*). Incidentally, u_h also satisfies

$$\text{(3.3)} \qquad \inf_{v_h \in V_h} \{F(Bv_h) + G(v_h)\}$$

which is the dual of (3.1) and which itself is evidently a problem of the type investigated in Chapter III (note that (3.3) is clearly a discretised form of (2.1), but the convergence of u_h to u, even if u exists, is improbable).

4. APPLICATION OF ALGORITHM ALG2

The augmented Lagrangian introduced in Chapter III is of the form:

$$\mathcal{L}_r(v,q,\mu) = F(q) + G(v) + (\mu, Bv-q) + \frac{r}{2}\|Bv-q\|^2 .$$

We note that in the present case G is linear.

LEMMA 4.1: *Let* $\{u,p,\lambda\}$ *be a saddle point of* $\mathcal{L}_r$; *then* u *is a solution of* (2.1), $p = Bu$ *and furthermore*

$$\text{(4.1)} \qquad \lambda = \Lambda\sigma ,$$

where σ *is a solution of the initial problem* (1.1).

Proof. The first part follows directly from Theorem 2.1 of Chapter III. From Section 2.3 of Chapter III, it also follows that $\{u,\lambda\}$ is a saddle point of the Lagrangian (see (2.16), Chapter III)

$$L(v,\mu) \equiv \frac{1}{2}\|\mu\|^2 + I_{K_o}(\mu) - (\mu, Bv) - G(v) ,$$

which again gives the Lagrangian (3.2) after the simple change of variable $\mu = \Lambda\tau$. ∎

Obviously, the existence of such a saddle point, like that of u, is doubtful in the infinite-dimensional case; however, in practice, the problem will be solved in finite dimensions.

In view of the linearity of G, algorithm ALG2 may be written:

$\{p^o,\lambda^o\} \in H\times H$ *and* $\rho > 0$ *are given;*

$\{p^n,\lambda^n\} \in H\times H$ *being given by recurrence, calculate*

$(4.2)_1$ *the solution* u^{n+1} *of* $r(Bu^{n+1},Bv)+(Bv,\lambda^n-rp^n)+ G(v) = 0 \quad \forall v \in V$,

$(4.2)_2$ *the solution* p^{n+1} *of* $F'(p^{n+1})+rp^{n+1} = \lambda^n + rBu^{n+1}$,

$(4.2)_3$ $\lambda^{n+1} = \lambda^n + \rho(Bu^{n+1}-p^{n+1})$.

The calculation of p^{n+1} in stage $(4.2)_2$ of algorithm (4.2) can be written out explicitly. In fact, in view of the definition of F, we have

(4.3) $F'(q) = \pi_o q$

where $\pi_o : H \to K_o$ is the projection onto K_o. From the definitions (2.3) and (1.2) of K_o and of K, π_o is local, and the nonlinear equation $(4.2)_2$ can therefore be solved almost everywhere. It decomposes triangle by triangle for the approximate problem, since we have taken the precaution of choosing a space of piecewise-constant functions for H_h. We can even solve $(4.2)_2$ explicitly with (4.3).

LEMMA 4.2 *Let* $\phi^n = \frac{1}{1+r}(\lambda^n+rBu^{n+1})$; *then we have* $p^{n+1} = \frac{1}{r}((1+r)\phi^n - \pi_o\phi^n)$.

Proof. We show that ϕ^n is a convex combination of p^{n+1} and $\pi_o p^{n+1}$. Consequently we have $\pi_o\phi^n = \pi_o p^{n+1}$, which then gives the result. ∎

Stage $(4.2)_2$ of algorithm (4.2) is linear since G is linear. In finite dimensions, this consists of solving a linear system with matrix B^tSB, where the matrix S is symmetric and positive-definite relative to the inner product $(\cdot,\cdot)$.

Synopsis: We shall now prove the convergence of algorithm (4.2) in the case where $\rho = r$. Here we are in a situation which is the *reverse* of that in Chapter III: the problem (P) is noncoercive and, in contrast, its dual ((1.3) in this case) is coercive. In Chapter III, it was (P) which was coercive and its dual (in Examples

1 and 2 of Section 1.1 at least) which was not. Consequently, we proved that $\{u^n,p^n\} \to \{u,p\}$ and could prove only a weak convergence property for λ^n. Here, in contrast, we shall show that $\lambda^n \to \lambda$ and that we have only a weak convergence property for $\{u^n,p^n\}$. We have a somewhat analogous situation for the approximation, since $\sigma_h \to \sigma$ (isomorphic to λ) and since we cannot prove anything with regard to u_h.

5. CONVERGENCE OF ALGORITHM ALG2

THEOREM 5.1 *If there exists a saddle point* $\{u,p,\lambda\}$ *of the Lagrangian* $\mathcal{L}_r$, *then algorithm* (4.2) *converges for* $\rho = r$ *in the following sense :* $\{u^n,p^n\}$ *remains bounded and* $\lambda^n \to \lambda$ *when* $n \to +\infty$.

Proof. By subtracting $(4.2)_3$ from $(4.2)_2$ we obtain (since $\rho = r$):

$$\lambda^{n+1} = F'(p^{n+1})$$

so that λ^n and p^n are linked by a simple relation. Furthermore, since $\{u,p,\lambda\}$ is a saddle point of $\mathcal{L}_r$, we have

$$(5.1)_1 \qquad r(Bu,Bv) + (Bv,\lambda-rp)+G(v) = 0 ,$$

$$(5.1)_2 \qquad F'(p)+rp = \lambda+rBu ,$$

$$(5.1)_3 \qquad p-Bu = 0 .$$

Similarly we have $\lambda = F'(p)$ in $(5.1)_2$ in view of $(5.1)_3$. Putting $\bar{\lambda}^n = \lambda^n - \lambda$, $\bar{p}^n = p^n - p$ and $\bar{u}^n = u^n - u$, we obtain

$$(5.2) \qquad \bar{\lambda}^{n+1} + r\bar{p}^{n+1} = \bar{\lambda}^n + r\,B\bar{u}^{n+1}.$$

By subtracting $(5.1)_1$ from $(4.2)_1$ we also obtain

$$(5.3) \qquad r(B\bar{u}^{n+1},Bv) = -(Bv,\bar{\lambda}^n - r\bar{p}^n),$$

i.e. by introducing the operator $P : H \to \mathrm{Im}B$, which projects onto the image of B:

$$(5.4) \qquad rB\bar{u}^{n+1} = -P(\bar{\lambda}^n - r\bar{p}^n) ;$$

therefore (5.2) becomes

$$\bar{\lambda}^{n+1}+r\bar{p}^{n+1} = (I-P)\bar{\lambda}^{n}+r\ P\ \bar{p}^{n}$$

and squaring (since P and I-P are 2 orthogonal projectors)

$$\|\bar{\lambda}^{n+1}\|^2 + r(\bar{p}^{n+1},\bar{\lambda}^{n+1}) + r^2\|\bar{p}^{n+1}\|^2 \leq \|\bar{\lambda}^{n}\|^2 + r^2\|\bar{p}^{n}\|^2$$

which proves, since by the monotonicity of F', $(\bar{p}^{n+1}, \bar{\lambda}^{n+1}) \geq 0$, that λ^n and p^n remain bounded, and therefore u^n also remains bounded in view of (5.4).

Furthermore $(\bar{p}^{n+1}, \bar{\lambda}^{n+1}) \to 0$ when $n \to +\infty$, which can be written, since $F'(q) = \pi_o q$, in the form

$$(\pi_o p^{n+1} - \pi_o p,\ p^{n+1}-p) \to 0\ .$$

Now

$$\|\lambda^{n+1}-\lambda\|^2 = \|\pi_o p^{n+1}-\pi_o p\|^2 \leq (\pi_o p^{n+1}-\pi_o p\ ,p^{n+1}-p)$$

in view of the properties of projection onto a convex set; we have thus proved the convergence of λ^n to λ when $n \to +\infty$. ∎

6. NUMERICAL APPLICATION

6.1 Description of the mechanical problem

We have considered the problem of the bending of an encastred beam of length ℓ and of thickness $2a$, subjected to a shear force F_o (see Figure 6.1).

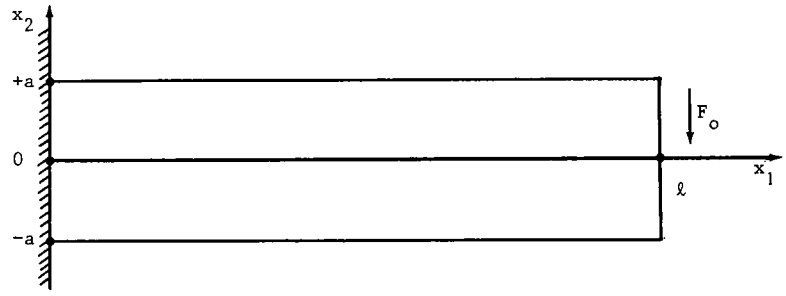


Figure 6.1 - Encastred beam subject to a shear force F_o.

We assume that the width of the beam (in the direction Ox_3 orthogonal to the plane of the figure) is sufficiently large and we will then be justified in studying the plane-strain problem instead of the three-dimensional problem, by assuming that the displacement field depends only on x_1 and x_2 and satisfies $u_3 = 0$. The strain tensor $\varepsilon = Bu$ can then be written:

$$\varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{12} & 0 \\ \varepsilon_{21} & \varepsilon_{22} & 0 \\ 0 & 0 & 0 \end{pmatrix} .$$

We shall see that the same does not apply for the stress tensor. Assuming the medium to be isotropic, we have

$$A^{-1}e \equiv \lambda \ \mathrm{tr}(e)\ \delta + 2\mu e$$

where $\mathrm{tr}(e)$ denotes the trace of the tensor e, λ and μ are the Lamé constants, and δ is the Kronecker tensor. We have adopted the von Mises plasticity criterion, and the plasticity convex set is therefore written

$$C = \{\tau \in \mathbb{R}^9 : |\tau^D| \leq \zeta \sqrt{2}\}$$

where ζ is a coefficient characteristic of the material, τ^D is the deviator of the tensor τ ($\tau^D \equiv \tau - \frac{1}{d}\,\mathrm{tr}(\tau)\delta$), and $|\cdot|$ is the Euclidian norm of $\mathbb{R}^9$. An explicit calculation shows that

$$F(e) = \int_\Omega \psi(e(x))dx$$

where

$$\psi(e) \equiv K_o \mathrm{tr}(e) + \begin{cases} \mu |e^D|^2 & \text{if } |e^D| \leq \frac{k\sqrt{2}}{2\mu} \\ k\sqrt{2}\,(|e^D| - \frac{k\sqrt{2}}{4\mu}) & \text{otherwise.} \end{cases}$$

As may be expected, F is differentiable: we have

$$F'(e) = \psi'(e(x)) \ , \ a.e.\ x \in \Omega \ , \ \text{where}$$

$$\psi'(e) = K_o\ \mathrm{tr}(e)\delta + \min(2\mu, \frac{k\sqrt{2}}{|e^D|})e^D \ ,$$

where $K_o = \lambda + \frac{2\mu}{3}$.

Since in the equilibrium state we have $\sigma = \phi'(\varepsilon(u))$, we see that in general $\sigma_{33} \neq 0$. The problem is nonetheless two-dimensional, and the third component of the displacement is zero.

6.2 Choice of constants

We have chosen $\mu = \frac{1.33}{30}$, $\lambda = \frac{0.33}{15}$ (in Imperial units which we shall not define here), $a = 2$ and $\ell = 20$. The triangulations chosen were uniform, with mesh intervals $\frac{\ell}{m_1}$ and $\frac{2a}{m_2}$ (see Figure 6.2). We have chosen $m_1 = 10$ and $m_2 = 6$ for an initial mesh (this gives dim $V_h = 140$) and $m_1 = 14$ and $m_2 = 8$ for a second mesh (dim $V_h = 252$). With these data the matrix $B^t SB$ of the linear system to be solved in the elastic case ($\zeta = +\infty$) is badly conditioned and it is necessary to perform the calculations in double precision.

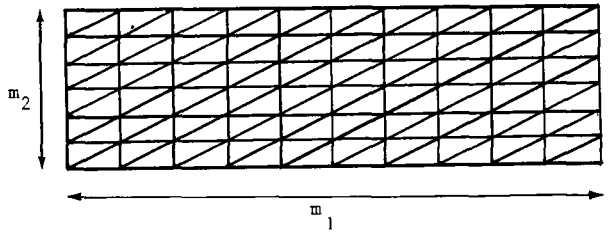


Figure 6.2 - First triangulation used ($m_1 = 10$, $m_2 = 6$).

The majority of iterative methods are inefficient for solving this linear system, with the exception of the conjugate-gradient method and its variants. Even so, this latter method only still converges in a number of iterations close to the number of variables, and this means that it is not competitive with direct methods. It is quite possible that this conclusion would need to be reconsidered if a suitable change of the inner product were made, since this may have a preconditioning effect.

For the elastoplastic case, we have chosen $k = \mu\sqrt{3}.10^{-3}$ and have compared algorithm ALG2 with two other standard algorithms.

6.3 Gradient method

We put $\phi(v) = F(Bv) + G(v)$; solving (P_h) is equivalent to minimising ϕ which is differentiable over V_h; using ψ and denoting by $\langle .,. \rangle$ the inner product chosen on V_h, we have

$$(6.1) \qquad \langle \phi'(v), w \rangle = \int_\Omega \psi'(Bv) \cdot Bw \, dx + G(w) \quad , \forall w \in V_h \ .$$

The gradient method (for minimising ϕ on V_h) is then written

$u^o \in V_h$ *and* $\rho>0$ *are given;*

$u^n \in V_h$ *being given by recurrence, calculate*

$u^{n+1} = u^n - \rho\, \phi'(u^n)$.

If we choose as inner product $< , >$ the natural inner product on $\mathbb{R}^N$ where N is the dimension of V_h, the calculation of the gradient ϕ' from formula (6.1) is immediate. However, as already pointed out, even in the quadratic case this procedure does not produce good results because of the ill conditioning of the matrix B^tSB. It is also possible to choose any other inner product on V_h, but then we would have to solve, at each iteration, a linear system with matrix R, where R is the matrix obtained from the chosen inner product $< , >$. We say that the matrix R acts as an *auxiliary operator* (or as a *preconditioning* operator). We have tested the following inner product

$$<u,v> = \int_\Omega (\Lambda Bu)\, Bv \, dx ,$$

and the matrix R obtained is then the stiffness matrix of the elastic problem, which would appear a natural choice for solving the elastoplastic problem.

6.4 Conjugate-gradient method

To minimise ϕ on V_h, we can also use the conjugate-gradient method:

$u^o \in V_h$, $w^o = \phi'(v_o)$ *are given; by recurrence, knowing* u^n *and* $w^n (\in V_h)$, *calculate*

$u^{n+1} = u^n - \rho_n w^n$, *where* ρ_n *minimises* $g(\rho) = \phi(u^n - \rho w^n)$,

$r^{n+1} = \phi'(u^{n+1})$,

$w^{n+1} = r^{n+1} + \lambda_n w^n$, *where* $\lambda_n = \dfrac{<r_{n+1}, r_{n+1} - r_n>}{<r^n, r^n>}$.

Let us say straight away that the results obtained with this method were less than excellent. We shall now describe the way in which the method was applied. In practice it is adequate to use an approximation of ρ_n, and this is what was done. It is possible to determine a value, possibly non-unique, of the second derivative of ϕ, since

ψ is differentiable almost everywhere. For ρ_n we have chosen the first iterate of Newton's method:

$$\rho_n = - \frac{g'(o)}{g''(o)}$$

which actually gives the exact solution in the elastic case. This choice is quite adequate; a more accurate calculation of ρ_n would be more expensive to compute and would not bring about any significant improvement in the convergence. We reinitialise $w^n = r^n$ every 2N or 3N iterations; once again no significant improvement would be evident if a reinitialisation were performed each time the inner product

$$<w^{n+1}, u^{n+1}-u^n>$$

becomes positive or smaller than a sufficiently-small positive constant. The problem is fundamentally ill-conditioned: in the (elastic) linear case we can successfully construct N conjugate directions and solve the problem; however, we are unable to do this in the (elastoplastic) nonlinear case.

In the light of recent results (see AXELSSON [1], CONCUS-GOLUB-O'LEARY [1]) it would appear that the idea of changing the inner product $<.,.>$ on V_h could in this case lead to a significant improvement. This is the basic idea behind so-called *preconditioning* methods. Two options appear available: either to take as the matrix R (arising from the inner product $<,>$) that of the elastic problem, or to use an incomplete Cholesky decomposition of this matrix following an idea due to MEIJERINK-VAN DER VORST [1], which would significantly reduce the cost of each iteration. The use of some preconditioning is in our opinion essential if the performance of the conjugate-gradient method is to be improved. It should be noted that the use of an auxiliary operator in the gradient method, just as in the penalisation-duality methods studied in the present book, has an analogous effect which amounts to changing the metric of the space.

6.5 Choice of the parameters for algorithm ALG2

We have chosen for the *termination test* for algorithm ALG2:

$$\sum_{T \in \mathcal{T}_h} |Bu^{n+1}-p^{n+1}| \leq 10^{-8} .$$

When the force F_o is sufficiently small, the problem is purely elastic and the choice $r = 1$ is optimal. In the elastoplastic domain, the

choice $r < 1$ accelerates the convergence: a good choice would seem to be $r = \frac{1}{2}$ or $r = \frac{1}{3}$. In fact, in this case, from the very first iteration we arrive at a solution two or three times larger than that of the elastic problem, and which is approximately of the order of magnitude of the elastoplastic solution. However, the choice of r is not crucial. The following tables summarise the results obtained:

F_o	U	n_1	t_1	n_2	t_2
0.2	0.101	133	(elastic)		
0.4	0.203	245	-	-	-
0.5	0.270	411	-	-	-
0.55	0.322	547	42	35	3
0.6	0.404	574	-	-	-

Table 6.1 - *Results for* $m_1 = 6$, $m_2 = 10$ (140 *variables*)

U : vertical displacement observed at $x_1=\ell$, $x_2=-a$,

n_1 : number of iterations of the conjugate-gradient method

n_2 : number of iterations of ALG2

t_1, t_2 : respective machine times (in seconds on IBM 370/168).

		Conjugate gradient		ALG2			Gradient (with auxiliary operator)	
F_o	U	n_1	t_1	r	n_2	t_2	n_3	t_3
0.2	0.116	181	30	1	1	-	1	-
0.4	0.236	377	58	-	-	-	-	-
0.5	0.346	645	80	0.5	51	6	82	10
0.55	0.460	800	107	0.5	57	7	-	-
0.6	0.680	1200	150	0.4	79	9	215	24
0.65	1.00	-	-	0.33	88	10	250	28

Table 6.2 - *Results for* $m_1 = 8$, $m_2 = 14$ (252 *variables*) *(compared with Table 6.1 we have added the values of* r *used in* ALG2 *and* n_3 *(the number of iterations) and* t_3 *(machine time relative to the gradient method with auxiliary operator)).*

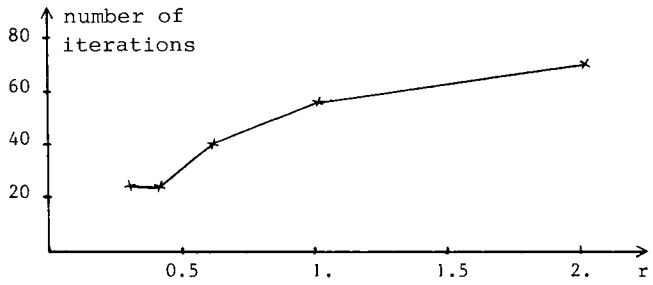


Figure 6.3 - *Variation of the number of iterations of* ALG2 *as a function of* r (*case* $m_1 = m_2 = 3$).

7. DISCUSSION

For the problem considered here, algorithm ALG2 is two to three times faster than the gradient method with auxiliary operator and ten to twenty times faster than the conjugate-gradient method without preconditioning. The good performance of both of the first two methods may be attributed to the fact that, even though we had a linear system with matrix B^tSB to solve at each iteration, since this matrix was *fixed* and of banded structure, it was factorised *once and for all* (by Cholesky's method) at the start of the algorithm into a product LL^t, L being lower triangular and of banded structure. At each iteration we therefore have to solve only two linear systems with matrix L, and this is extremely rapid. As regards stage $(4.2)_2$ of algorithm (4.2), this can also be performed very rapidly since it *decomposes triangle by triangle*.

It is now appropriate to explain the importance of the choice of the inner product adopted in Section 2. In fact, in the augmented Lagrangian $\mathcal{L}_r$ the penalty term is the square of the norm on H. At any rate, it was under such an assumption that the proofs in Chapter III were performed. However, if we take for H the 'natural' inner product on H:

$$(p,q) = \int_\Omega p \cdot q \, dx$$

then the performance of algorithm ALG2 deteriorates significantly. In this case the matrix B^tSB will not, in fact, be the matrix of the underlying elastic problem, but that of an elastic problem with different coefficients, which does not bear such a close relation to

the (nonlinear) problem being treated. The same applies also for the gradient algorithm with auxiliary operator.

We cannot over-emphasise for such problems the importance of the inner product of the space on which we are working - that is, in the vocabulary of the gradient method, of choosing a good auxiliary operator or, in the vocabulary of the conjugate-gradient method, of choosing a good 'preconditioning'. This concept is also important for the penalisation-duality algorithms: the choice of the penalty term is at our disposal and it is necessary to take the one closest to the nonlinear problem being considered, as has already been pointed out in Chapter V.

CHAPTER VII

APPLICATION TO THE NUMERICAL SOLUTION OF THE TWO-DIMENSIONAL FLOW OF INCOMPRESSIBLE VISCOPLASTIC FLUIDS

D. Begis, R. Glowinski

1. GENERAL NOTES. SYNOPSIS

The present chapter is based largely on BEGIS [2] and GLOWINSKI-LIONS-TREMOLIERES [2 , Appendix 6]. It extends Section 6.1 of Chapter V relating to the flow of a Bingham fluid in a cylindrical duct. Here we shall be considering the much more complicated problem of the unsteady flow of a fluid of the above type in a bounded two-dimensional cavity. We shall see that the introduction of a stream function enables the problem considered to be reduced to a *parabolic variational inequality of order* 4 *with respect to the space variables*. We shall then examine the approximation of the above problem by methods using *mixed finite elements* (for the *spatial approximation*) and *finite differences* (for the *approximation in time*). We shall then show that these approximate problems can be solved by the augmented Lagrangian methods of Chapter III, the algorithms thereby obtained generalising those of Chapter V, Section 6.1, relating to the flow of a Bingham fluid in a cylindrical duct. Finally, we shall present some numerical results obtained by the above methods, and this will demonstrate some of the properties of Bingham fluids.

2. FORMULATION OF BINGHAM FLOWS USING THE VELOCITY AND THE PRESSURE

Let Ω be a bounded domain in $\mathbb{R}^2$ with regular boundary Γ. With $\underset{\sim}{v} = \{v_1, v_2\}$ denoting an $\mathbb{R}^2$-valued function, we put

$$D_{ij}(\underset{\sim}{v}) = \frac{1}{2}\left(\frac{\partial v_i}{\partial x_j} + \frac{\partial v_j}{\partial x_i}\right), \tag{2.1}$$

$$D_{II}(\underset{\sim}{v}) = \frac{1}{2}\sum_{i,j=1}^{2}(D_{ij}(\underset{\sim}{v}))^2, \tag{2.2}$$

$$a(\underset{\sim}{v},\underset{\sim}{w}) = 2 \sum_{i,j=1}^{2} D_{ij}(\underset{\sim}{v}) D_{ij}(\underset{\sim}{w}) dx, \tag{2.3}$$

$$j(\underset{\sim}{v}) = \int_{\Omega} (D_{II}(\underset{\sim}{v}))^{1/2} dx, \tag{2.4}$$

$$V_o = \{\underset{\sim}{v} \mid \underset{\sim}{v} \in H_o^1(\Omega) \times H_o^1(\Omega),\ \nabla \cdot \underset{\sim}{v} = 0\}, \tag{2.5}$$

$$H = \{\underset{\sim}{v} \mid \underset{\sim}{v} \in L^2(\Omega) \times L^2(\Omega),\ \nabla \cdot \underset{\sim}{v} = 0,\ \underset{\sim}{v} \cdot \underset{\sim}{n} = 0 \text{ on } \Gamma\}. \tag{2.6}$$

If, furthermore, $\underset{\sim}{z} \in H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)$, with [1]

$$\int_{\Gamma} \underset{\sim}{z} \cdot \underset{\sim}{n}\, d\Gamma = 0, \tag{2.7}$$

we associate with $\underset{\sim}{z}$ the *(nonempty)* affine space

$$V_z = \{\underset{\sim}{v} \mid \underset{\sim}{v} \in H^1(\Omega) \times H^1(\Omega),\ \nabla \cdot \underset{\sim}{v} = 0,\ \underset{\sim}{v} = \underset{\sim}{z} \text{ on } \Gamma\}. \tag{2.8}$$

In the following discussion, we shall *neglect the effects of inertia* (associated with the trilinear form $(\cdot,\cdot,\cdot)$ of Chapter II, Section 4.1, relation (4.5)); this leads us (see DUVAUT-LIONS [1, Chapter VI]) to model the *unsteady* flow in Ω of a Bingham fluid satisfying $\underset{\sim}{u} = \underset{\sim}{z}$ on Γ, by

$$\left\{\begin{array}{l} \text{Find } \underset{\sim}{u} \in L^2(0,T;V_z) \cap L^\infty(0,T;H),\ \underset{\sim}{u}' \in L^2(0,T;V_o') \text{ such that} \\ (\underset{\sim}{u}'(t), \underset{\sim}{v} - \underset{\sim}{u}(t)) + \nu a(\underset{\sim}{u}(t), \underset{\sim}{v} - \underset{\sim}{u}(t)) + gj(\underset{\sim}{v}) - gj(\underset{\sim}{u}) \geq (\underset{\sim}{f}(t), \underset{\sim}{v} - \underset{\sim}{u}(t)) \\ \forall \underset{\sim}{v} \in V_z, \text{ a.e. in } t, \\ \underset{\sim}{u}(0) = \underset{\sim}{u}_o \in H,\ \underset{\sim}{f} \in L^2(0,T;V_o'). \end{array}\right. \tag{2.9}$$

We recall that in (2.9)

ν is the *viscosity* of the fluid,
g is the *threshold of plasticity (yield stress)*
$\underset{\sim}{f}$ is a *density of external forces.*

DUVAUT-LIONS [1, Chapter 6] proves (for $\underset{\sim}{z} = \underset{\sim}{0}$) the existence and

---

[1] $\underset{\sim}{n}$: unit normal vector on Γ, pointing outwards from Ω.

uniqueness of a solution of problem (2.9); the case $\underset{\sim}{z} \neq \underset{\sim}{0}$, with $\underset{\sim}{z}$ satisfying (2.7), may be treated analogously.

Remark 2.1: We have assumed in the above that $\underset{\sim}{z}$ $(= \underset{\sim}{u}|_{\Gamma})$ is independent of t; there are no further difficulties in treating this case numerically by the methods to be discussed subsequently.

3. FORMULATION OF BINGHAM FLOWS USING A STREAM FUNCTION

In this section we shall adopt the following two simplifying assumptions:

(i) Ω is *simply connected*

(ii) $\underset{\sim}{z} = \underset{\sim}{u}|_{\Gamma} = \underset{\sim}{0}$;

nonetheless Remark 2.1 still holds for situations in which (i) and/or (ii) are not satisfied. If we confine our attention to *two-dimensional* flows, we can eliminate the condition $\underset{\sim}{\nabla}\cdot\underset{\sim}{u} = 0$ in a natural manner by introducing a *stream function* defined (to within an additive constant) by

$$u_1 = \frac{\partial \psi}{\partial x_2}, \quad u_2 = -\frac{\partial \psi}{\partial x_1}. \tag{3.1}$$

The condition $\underset{\sim}{u} = \underset{\sim}{0}$ on Γ implies

$$\psi = \text{Const. } on\ \Gamma, \tag{3.2}$$

$$\frac{\partial \psi}{\partial n} = 0 \ on\ \Gamma. \tag{3.3}$$

We shall take $\psi = 0$ on Γ, which fixes the constant mentioned above.

Let $\underset{\sim}{v} \in V_0$; we associate with $\underset{\sim}{v}$ the function $\phi \in H_0^2(\Omega)$ defined uniquely by

$$v_1 = \frac{\partial \phi}{\partial x_2}, \quad v_2 = -\frac{\partial \phi}{\partial x_1}, \tag{3.4}$$

$$\phi = \frac{\partial \phi}{\partial n} = 0 \ on\ \Gamma. \tag{3.5}$$

We recall that

$$H_o^2(\Omega) = \{\phi \mid \phi \in H^2(\Omega),\ \phi = \frac{\partial \phi}{\partial n} = 0 \ \textit{on}\ \Gamma\}. \tag{3.6}$$

In view of (3.1), (3.4) we can reduce (2.9) to the *following parabolic variational inequality* (of order 4 with respect to the space variables):

$$\left\{\begin{array}{l} \textit{Find}\ \psi \in L^2(0,T;H_o^2(\Omega)) \cap L^\infty(0,T;H_o^1(\Omega))\ \textit{such that} \\ -\int_\Omega \psi'(t)\Delta(\phi-\psi(t))dx + \nu\tilde{a}(\psi(t),\phi-\psi(t)) + g\tilde{j}(\phi) - g\tilde{j}(\psi(t)) \\ \geq (\tilde{f}(t),\phi-\psi(t))^{(2)}\ \forall \phi \in H_o^2(\Omega), \textit{a.e. in}\ t, \\ \psi(0) = \psi_o \in H_o^1(\Omega), \end{array}\right. \tag{3.7}$$

where

$$\left\{\begin{array}{l} \tilde{a}(\phi_1,\phi_2) = \int_\Omega \left[(2\,\frac{\partial^2\phi_1}{\partial x_1 \partial x_2})(2\,\frac{\partial^2\phi_2}{\partial x_1 \partial x_2}) \right. \\ \qquad\qquad \left. + (\frac{\partial^2\phi_1}{\partial x_2^2} - \frac{\partial^2\phi_1}{\partial x_1^2})(\frac{\partial^2\phi_2}{\partial x_2^2} - \frac{\partial^2\phi_2}{\partial x_1^2})\right] dx\ , \\ \forall\ \phi_1,\phi_2 \in H_o^2(\Omega), \end{array}\right. \tag{3.8}$$

$$\left\{\begin{array}{l} \tilde{j}(\phi) = \int_\Omega \left[(2\,\frac{\partial^2\phi}{\partial x_1 \partial x_2})^2 + (\frac{\partial^2\phi}{\partial x_2^2} - \frac{\partial^2\phi}{\partial x_1^2})^2\right]^{1/2} dx \\ \forall \phi \in H_o^2(\Omega). \end{array}\right. \tag{3.9}$$

Remark 3.1: In fact we have

$$\tilde{a}(\phi_1,\phi_2) = \int_\Omega \Delta\phi_1 \Delta\phi_2\ dx \quad \forall \phi_1,\phi_2 \in H_o^2(\Omega)\ . \tag{3.10}$$

In the following we shall be using (3.8) and (3.10) simultaneously.

---

[2] If, in (2.9), $(f,v) = \int_\Omega fv\ dx$ then $\tilde{f} = \frac{\partial f_2}{\partial x_1} - \frac{\partial f_1}{\partial x_2}$.

4. APPROXIMATION OF THE STEADY-STATE PROBLEM

4.1 Synopsis. Formulation of the steady-state problem

Before approximating (3.7) - by means of a mixed finite-element method - we shall first study the approximation of the corresponding steady-state problem, i.e. the following elliptic variational inequality of order 4: [3]

$$(4.1)\qquad \left\{ \begin{array}{l} \textit{Find } \psi \in H_o^2(\Omega) \textit{ such that} \\ \nu a(\psi,\phi-\psi)+gj(\phi)-gj(\psi) \geq (f,\phi-\psi) \quad \forall\, \phi \in H_o^2(\Omega), \end{array} \right.$$

where $a(\cdot,\cdot)$ and $j(\cdot)$ are defined by (3.8), (3.9); we note that (4.1) is equivalent to the minimisation problem

$$(4.2)\qquad \left\{ \begin{array}{l} \textit{Find } \psi \in H_o^2(\Omega) \textit{ such that} \\ J(\psi) \leq J(\phi) \quad \forall \phi \in H_o^2(\Omega) \end{array} \right.$$

where, in (4.2), we have

$$(4.3)\qquad J(\phi) = \frac{\nu}{2}\, a(\phi,\phi) + gj(\phi) - (f,\phi).$$

We shall assume in the following that $f \in H^{-1}(\Omega)$; in fact there would be no difficulty in treating the case in which

$$(f,\phi) = \int_\Omega f\, \Delta\phi \, dx \quad \forall \phi \in H_o^2(\Omega) \; ; \; f \in L^2(\Omega).$$

Since the bilinear form $a(\cdot,\cdot)$ is $H_o^2(\Omega)$-*elliptic* (i.e. *coercive*), and the functional $j(\cdot)$ is *convex* and *continuous* on $H_o^2(\Omega)$, with $\phi \to (f,\phi)$ *linear and continuous*, then it is a classical result (see, for example, LIONS [1]) that (4.1), (4.2) *admits a unique solution*.

4.2 Approximation of (4.1), (4.2) by a mixed finite-element method

We shall approximate (4.1), (4.2) here by a mixed finite-element method suggested by MIYOSHI [1]. The objective is to reduce the approximation to that of a problem in which we only have to perform the discretisation of $H^1(\Omega)$ and $L^2(\Omega)$ instead of discretising $H^2(\Omega)$

[3] We shall henceforth omit the symbol $\sim$.

which is a much more complicated task. To do this, we first introduce a weakened variational formulation of our problem. The new variational problem thus obtained possesses a unique solution which coincides with that of (4.1), (4.2) under fairly unrestrictive conditions. For a general presentation of this approach, the reader may refer to GIRAULT-RAVIART [1].

Thus, suppose $\phi \in H_o^2(\Omega)$ and put for $1 \le i,j \le 2$,

$$z_{ij} = z_{ji} = \frac{\partial^2 \phi}{\partial x_i \partial x_j} \in L^2(\Omega). \tag{4.4}$$

We then have, for all $v \in H^1(\Omega)$,

$$\int_\Omega z_{ij}\, v\, dx + \frac{1}{2}\int_\Omega \left(\frac{\partial \phi}{\partial x_i}\frac{\partial v}{\partial x_j} + \frac{\partial \phi}{\partial x_j}\frac{\partial v}{\partial x_i}\right) dx = 0. \tag{4.5}$$

Conversely if $\phi \in H_o^1(\Omega)$ and $z = \{z_{ij}\}_{1\le i,j\le 2}$ satisfy (4.5) then $\phi \in H_o^2(\Omega)$ and z and ϕ are related by (4.4). Thus writing

$$\begin{cases} J(\phi,z) = \dfrac{\alpha\nu}{2}\displaystyle\int_\Omega (z_{11}+z_{22})^2 dx + \frac{\beta\nu}{2}\int_\Omega \left[(2z_{12})^2+(z_{22}-z_{11})^2\right] dx \\ \qquad\qquad + gj(z) - (f,\phi), \end{cases} \tag{4.6}$$

where we have $\alpha,\beta \in]0,1[$ with $\alpha + \beta = 1$, then

$$j(z) = \int_\Omega \left[(2z_{12})^2+(z_{22}-z_{11})^2\right]^{1/2} dx \tag{4.7}$$

and putting

$$W = \{\{q,z\} \mid \phi \in H_o^1(\Omega),\ z \in (L^2(\Omega))^4,\ \phi \text{ and } z \text{ satisfy (4.5)}\} \tag{4.8}$$

this leads us to replace problem (4.1), (4.2) by the following problem:

$$\begin{cases} \textit{Find } \{\psi,s\} \in W \textit{ such that} \\ J(\psi,s) \le J(\phi,z) \quad \forall\ \{\phi,z\} \in W. \end{cases} \tag{4.9}$$

This problem, which is equivalent to the initial problem, offers a considerable advantage as far as the discretisation is concerned, since it requires only the approximation of the spaces $H^1(\Omega)$ and $L^2(\Omega)$. The discrete variables are then related by a weak form of (4.4).

We shall assume in the following that Ω is a convex polygon in $\mathbb{R}^2$; let $\{\mathcal{T}_h\}_h$ be a standard family of triangulations of Ω. We then put

$$
(4.10)\quad \left\{
\begin{array}{l}
V_h = \{v_h \in C^0(\overline{\Omega}),\ v_h|_K \in P_k \quad \forall K \in \mathcal{T}_h\}\ , \\
V_{oh} = \{v_h \in V_h,\ v_h = 0 \ \textit{on}\ \Gamma\} = V_h \cap H_o^1(\Omega), \\
W_h = \{\{\phi_h, z_h\} | \phi_h \in V_{oh},\ z_h \in (V_h)^4, \\
2\displaystyle\int_\Omega z_{ijh} v_h dx + \int_\Omega \left(\frac{\partial \phi_h}{\partial x_i}\frac{\partial v_h}{\partial x_j} + \frac{\partial \phi_h}{\partial x_j}\frac{\partial v_h}{\partial x_i}\right) dx = 0 \quad \forall v_h \in V_h,\ 1 \le i,j \le 2\}\ .
\end{array}
\right.
$$

It may be noted that the approximations of $H^1(\Omega)$ and $L^2(\Omega)$ are performed here using the same space of finite elements. This procedure is well adapted to the present situation, but it is not the only means possible. Finally the approximate problem will obviously be:

$$
(4.11)\quad \left\{
\begin{array}{l}
\textit{Find}\ \ \{\psi_h, s_h\} \in W_h,\ \ \textit{such that} \\
J(\psi_h, s_h) \le J(\phi_h, z_h) \quad \forall\ \{\phi_h, z_h\} \in W_h\ .
\end{array}
\right.
$$

To conclude, we note that the fact that we are using $z_{ij} = \dfrac{\partial^2 \phi}{\partial x_i \partial x_j}$ *as an auxiliary variable means that the process is* particularly well adapted to the treatment of the nondifferentiable term appearing in the functional to be minimised; it is for this reason also that the above mixed method has been chosen.

4.3 Solvability of problem (4.11)

The following theorem is proved in GLOWINSKI-LIONS-TREMOLIERES [2, Appendix 6, Section 4.4.3]:

THEOREM 4.1: *The approximate problem* (4.11) *admits one and only one solution.*

The solution of (4.11) by algorithms of the ALG1 or ALG2 type will form the subject of Section 6 later in the present chapter.

4.4 Convergence of the approximate solutions

We shall restrict our attention to the cases k = 1, 2 (see Remark 4.1 below for $k \ge 3$); concerning the convergence of the approximate solutions when $h \to 0$, we have:

THEOREM 4.2: *Suppose that when* $h \to 0$ *the angles of* $\mathcal{T}_h$ *remain bounded below, uniformly in* h, *by* $\theta_0 > 0$; *suppose also that the condition*

$$\frac{\underset{K \in \mathcal{T}_h}{\text{Max}}\ h(K)}{\underset{K \in \mathcal{T}_h}{\text{Min}}\ h(K)} \leq \tau \quad \forall\, \mathcal{T}_h, \tau \text{ independent of } h, \tag{4.12}$$

(where $h(K)$ = *length of the longest side of* K*) is satisfied. We then have*

$$\lim_{h \to 0} \{\psi_h, s_h\} = \{\psi, s\} \text{ strongly in } H_0^1(\Omega) \times (L^2(\Omega))^3, \tag{4.13}$$

where $\{\psi_h, s_h\}$ *is the solution of the approximate problem* (4.11), ψ *is that of the continuous problem* (4.1), (4.2) *and where* $s = \{s_{ij}\}_{1 \leq i,j \leq 2}$ *with* $s_{ij} = \dfrac{\partial^2 \psi}{\partial x_i \partial x_j}$.

We refer the reader to GLOWINSKI-LIONS-TREMOLIERES [2, Appendix 6, Section 4.4.4] for the proof of Theorem 4.2.

Remark 4.1: We have assumed above that $k = 1, 2$; in fact, similar convergence results could be obtained for approximations based on finite elements of order $k \geq 3$, but given the *limited regularity* of the solutions ($\psi \notin H^4(\Omega) \times H_0^2(\Omega)$ in general) the use of elements of such a high order is not justified.

4.5 Approximation using numerical integration

From a practical point of view it is necessary to use a numerical integration procedure in order to approximate the functional $J(\cdot,\cdot)$ in (4.9), (4.11); we shall restrict our attention to the case $k = 1$. Let Σ_h denote the set of the vertices of $\mathcal{T}_h$; we approximate on V_h the inner product induced by $L^2(\Omega)$, i.e. $\{\lambda_h, \mu_h\} \to \int_\Omega \lambda_h \mu_h \, dx$, by

$$(\lambda_h, \mu_h)_h = \frac{1}{3} \sum_{P \in \Sigma_h} m(P) \lambda_h(P) \mu_h(P), \tag{4.14}$$

where, in (4.14), $m(P)$ is the sum of the areas of the triangles which have P as a common vertex. In view of (4.14) we shall in fact use in (4.11) the functional $J_h(\cdot,\cdot)$ defined (if $k = 1$) by

$$
(4.15)\qquad \left\{\begin{array}{l} J_h(\phi_h,z_h) = \frac{\alpha\nu}{2}\,(z_{11h}+z_{22h},z_{11h}+z_{22h})_h \\ + \frac{\beta\nu}{2}\,[\,(2z_{12h},2z_{12h})_h + (z_{22h}-z_{11h},z_{22h}-z_{11h})_h\,] \\ + \frac{g}{3}\sum_{P\in\Sigma_h} m(P)\,[\,(2z_{12h}(P))^2 + (z_{22h}(P)-z_{11h}(P))^2\,]^{1/2} - (f_h,\phi_h), \end{array}\right.
$$

where f_h is an approximation of f. Similarly, instead of using W_h defined by (4.8), we shall, if $k = 1$, use W_h defined by

$$
(4.16)\qquad \left\{\begin{array}{l} W_h = \{\{\phi_h,z_h\} \in V_{oh} \times (V_h)^3,\ 2(z_{ijh},\mu_h)_h \\ = -\int_\Omega \left(\frac{\partial\phi_h}{\partial x_i}\frac{\partial\mu_h}{\partial x_j} + \frac{\partial\phi_h}{\partial x_j}\frac{\partial\mu_h}{\partial x_i}\right)dx \quad \forall\mu_h \in V_h,\quad 1\le i,j\le 2\}. \end{array}\right.
$$

Using the relations (4.16) it is easy to express $z_{ijh}(P)$, $\forall P \in \Sigma_h$ explicitly as a function of the values taken by ψ_h on Σ_h; in fact the matrix associated with the discrete inner product $(\cdot,\cdot)_h$ in V_h is diagonal. In the numerical solution, it is therefore possible to eliminate the variable z_h; we refer the reader to BEGIS [2] for further details.

5. APPROXIMATION OF THE EVOLUTION PROBLEM (3.7)

5.1 Semi-discretisation with respect to time

Let $k = \Delta t$ (> 0) denote one step in the time discretisation; we then approximate (3.7) by the following *implicit scheme* (where $\psi^n \simeq \psi(nk)$ and where the $\sim$ have been omitted):

$$
(5.1)\qquad \left\{\begin{array}{l} \textit{for } \psi^n \textit{ known, determine } \psi^{n+1} \textit{ by solving} \\ \int_\Omega \nabla\left(\frac{\psi^{n+1}-\psi^n}{k}\right)\cdot\nabla(\phi-\psi^{n+1})\,dx + \nu a(\psi^{n+1},\phi-\psi^{n+1}) \\ + gj(\phi) - gj(\psi^{n+1}) \ge (f((n+1)k),\phi-\psi^{n+1}) \\ \forall\phi \in H_o^2(\Omega),\ \psi^{n+1} \in H_o^2(\Omega),\ n=0,1,\ldots\ ;\ \psi^o = \psi(0) = \psi_o. \end{array}\right.
$$

The use of the above semi-discrete scheme has thus enabled us to reduce the solution of the evolution problem (3.7) to that of a sequence of elliptic variational inequalities, equivalent to the

following sequence of minimisation problems (with $n \geq 0$):

$$\text{(5.2)} \qquad \begin{cases} \textit{Find } \psi^{n+1} \in H_o^2(\Omega) \textit{ such that} \\ J_k^{n+1}(\psi^{n+1}) \leq J_k^{n+1}(\phi) \quad \forall \phi \in H_o^2(\Omega) \end{cases}$$

where

$$\text{(5.3)} \qquad \begin{cases} J_k^{n+1}(\phi) = \frac{1}{2k} \int_\Omega |\nabla\phi|^2 \, dx + \frac{\nu}{2} a(\phi,\phi) + gj(\phi) \\ -(f((n+1)k),\phi) - \frac{1}{k} \int_\Omega \nabla\psi^n \cdot \nabla\phi \, dx. \end{cases}$$

The discretisation of (5.2), (5.3) by the mixed finite-element method of Section 4 is treated in Section 5.2 below.

5.2 Complete discretisation of (3.7)

The notation is the same as that in Section 4.2; we approximate $\psi^o = \psi_o$ by $\psi_h^o \in V_{oh}$ and the semi-discrete scheme (5.1) by the following:

With the function $\psi_h^n \in V_{oh}$ *known, obtain* $\{\psi_h^{n+1}, s_h^{n+1}\}$ *by solving, for* $n = 0,1,\ldots,$ *the minimisation problem*

$$\text{(5.4)} \qquad \begin{cases} \textit{Find } \{\psi_h^{n+1}, s_h^{n+1}\} \in W_h \textit{ such that} \\ J_{kh}(\psi_h^{n+1}, s_h^{n+1}) \leq J_{kh}(\phi_h, z_h) \quad \forall \{\phi_h, z_h\} \in W_h \end{cases}$$

where ($j(\cdot)$ still defined by (4.7)):

$$\text{(5.5)} \qquad \begin{cases} J_{kh}^{n+1}(\phi_h, z_h) = \frac{1}{2k} \int_\Omega |\nabla\phi_h|^2 \, dx + \frac{\alpha\nu}{2} \int_\Omega (z_{11h}+z_{12h})^2 \, dx \\ + \frac{\beta\nu}{2} \int_\Omega [(2z_{12h})^2 + (z_{22h}-z_{11h})^2] \, dx + gj(z_h) \\ - (f((n+1)k),\phi_h) - \frac{1}{k} \int_\Omega \nabla\psi_h^n \cdot \nabla\phi_h \, dx. \end{cases}$$

It can easily be shown that problem (5.4), (5.5) admits a unique solution; furthermore, the comments in Section 4.5 concerning the use of *numerical integration* are still valid for problem (5.4), (5.5). With regard to the convergence, as $h, k \to 0$, of the above approximate solutions to the solution of problem (3.7), we refer the reader to

GLOWINSKI-LIONS-TREMOLIERES [2].

6. SOLUTION OF (4.1) (5.2) BY AUGMENTED-LAGRANGIAN METHODS

6.1 Synopsis

In this section we shall show that it is possible to solve the *steady state problem* (4.1), or the sequence of problems (5.2) (obtained by the *semi-discretisation in time* of problem (3.9)), by means of augmented Lagrangian methods which fall within the general framework defined in Chapter III. We shall confine our attention to the case which is *continuous with respect to the space variables*, but the generalisation to problems which are approximate in space and time does not present any particular difficulty (apart from the fact that the formalism which has to be constructed is extremely cumbersome).

6.2 The model problem. Introduction of an augmented Lagrangian

Problems (4.1) and (5.2) lead us to consider the minimisation problem

$$(6.1)\qquad \left\{ \begin{array}{l} \textit{Find } \psi \in H_o^2(\Omega) \textit{ such that} \\ \\ J(\psi) \leq J(\phi) \quad \forall \phi \in H_o^2(\Omega), \end{array} \right.$$

with

$$(6.2)\qquad \left\{ \begin{array}{l} J(\phi) = \frac{\gamma}{2} \int_\Omega |\nabla \phi|^2 \, dx + \frac{\nu}{2} \int_\Omega |\Delta \phi|^2 \, dx \\ \\ + g \int_\Omega \left[\left(2 \frac{\partial^2 \phi}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 \phi}{\partial x_2^2} - \frac{\partial^2 \phi}{\partial x_1^2} \right)^2 \right]^{1/2} dx - (f, \phi) \end{array} \right.$$

and $\gamma \geq 0$ ($\gamma = 0$ for the steady-state problem, $\gamma = 1/k$ if (6.1) arises from problem (5.2)). The principal difficulty in the solution of (6.1), (6.2) arises from the *nondifferentiable* functional

$$\phi \to \int_\Omega \left[\left(2 \frac{\partial^2 \phi}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 \phi}{\partial x_2^2} - \frac{\partial^2 \phi}{\partial x_1^2} \right)^2 \right]^{1/2} dx.$$

To get round this difficulty (as well as to simplify the discretisation of the problem) we shall adopt the framework of Section (4.2) and consider a mixed variational formulation of problem (6.1), (6.2).

With $j(\cdot)$ still defined by (4.7) we again put

$$\begin{cases} W = \{\{\phi,z\} \in H_0^1(\Omega) \times (L^2(\Omega))^4,\ 2\int_\Omega z_{ij} v\, dx = -\int_\Omega (\frac{\partial \phi}{\partial x_i}\frac{\partial v}{\partial x_j} + \frac{\partial \phi}{\partial x_j}\frac{\partial v}{\partial x_i})dx \\ \forall v \in H^1(\Omega),\ 1 \le i,j \le 2\} \end{cases}$$

and

$$\begin{cases} J(\phi,z) = \frac{\gamma}{2}\int_\Omega |\nabla\phi|^2 dx + \frac{\alpha\nu}{2}\int_\Omega (z_{11}+z_{22})^2 dx + \frac{\beta\nu}{2}\int_\Omega \left[(2z_{12})^2 + (z_{22}-z_{11})^2\right] dx \\ \qquad + gj(z) - (f,\phi), \end{cases}$$

so that it is clear that (6.1), (6.2) can be written

$$(6.3) \quad \begin{cases} \textit{Find} \quad \{\psi,s\} \in W \ \textit{such that} \\ J(\psi,s) \le J(\phi,z) \quad \forall\ \{\phi,z\} \in W. \end{cases}$$

In order to adapt the general method of Chapter III to this case, it is natural to introduce here a supplementary variable $q = \{q_i\}_{i=1}^2 \in (L^2(\Omega))^2$ related to z by the linear equations

$$(6.4) \quad \begin{cases} q_1 = 2z_{12}, \\ q_2 = z_{22}-z_{11}. \end{cases}$$

It is this constraint (6.4) that we shall be treating by penalisation and duality, via the introduction of an *augmented Lagrangian*. So as to allow the notation of Chapter III to be used here, we put:

$$(6.5) \quad \begin{cases} V = W \\ H = (L^2(\Omega))^2, \\ B \in \mathcal{L}(V,H) \quad \textit{defined by} \quad B\{\phi,z\} = \{2z_{12}, z_{22}-z_{11}\}, \\ F(q) = gj(q), \\ G(\phi,z) = \frac{\gamma}{2}\int_\Omega |\nabla\phi|^2 dx + \frac{\alpha\nu}{2}\int_\Omega (z_{11}+z_{22})^2 dx + \frac{\beta\nu}{2}\int_\Omega [(2z_{12})^2 + (z_{22}-z_{11})^2] dx - (f,\phi). \end{cases}$$

We then define for $r > 0$, $\{\phi,z\} \in V$, $q \in H$, $\mu \in H$ the augmented Lagrangian $\mathcal{L}_r$: $V \times H \times H \to \mathbb{R}$ by

$$(6.6) \quad \begin{cases} \mathcal{L}_r(\{\phi,z\},q,\mu) = G(\phi,z) + F(q) + \int_\Omega (2z_{12}-q_1)\mu_1\, dx \\ + \int_\Omega (z_{22}-z_{11}-q_2)\mu_2 dx + \frac{r}{2}\int_\Omega |2z_{12}-q_1|^2\, dx + \frac{r}{2}\int_\Omega |z_{22}-z_{11}-q_2|^2 dx. \end{cases}$$

The solution of problem (6.1), (6.2) then reduces to seeking a

saddle point of $\mathcal{L}_r$ on $V \times H \times H$. We could also have considered in the above the decomposition associated with

$$\begin{cases} F(q) = \frac{\beta\nu}{2}\int_\Omega |q|^2\,dx + g\int_\Omega |q|\,dx, \\ G(\phi,z) = \frac{\gamma}{2}\int_\Omega |\nabla\phi|^2 dx + \frac{\alpha\nu}{2}\int_\Omega (z_{11}+z_{22})^2 dx. \end{cases}$$

In the following sections we shall have to solve problems corresponding to the minimisation of $\mathcal{L}_r$ on V, z and μ being fixed. This minimisation leads to solving a *linear mixed problem* in ϕ,z. The remarks made earlier relating to the discretisation and the use of numerical integration still apply, and we can consider the solution of such a problem as being standard.

6.3 Application of ALG1 to seeking a saddle point of $\mathcal{L}_r$

In view of Section 6.2, it is natural to solve problem (6.1), (6.2) by using algorithm ALG1 of Chapter III; we then obtain the following:

(6.7) $\lambda^0 \in H = (L^2(\Omega))^2$ *given,*

then, for $n \geq 0$, $\lambda^n \in H$ *being known, determine* $\{\psi^n, s^n\} \in V$ *and* $p^n \in H$ *then* λ^{n+1} *by*

(6.8) $$\begin{cases} \{\psi^n,s^n\} \in V,\ p^n \in H, \\ \mathcal{L}_r(\{\psi^n,s^n\},p^n,\lambda^n) \leq \mathcal{L}_r(\{\phi,z\},q,\lambda^n) \quad \forall\ \{\phi,z\} \in V,\ \forall q \in H \end{cases}$$

(6.9) $$\lambda^{n+1} = \lambda^n + \rho(B\{\psi^n,s^n\} - p^n).$$

In view of the convergence results established in Chapter III, Section 4, we have:

THEOREM 6.1: *Suppose that* $\mathcal{L}_r$ *admits a saddle point* $\{\{\psi,s\},p,\lambda\}$ *on* $V \times H \times H$; *then if*

(6.10) $$0 < \rho < 2r,$$

we have for all $\lambda^0 \in H$

(6.11) $$\lim_{n\to\infty} \{\psi^n,s^n\} = \{\psi,s\} \text{ strongly in } H_0^1(\Omega) \times (L^2(\Omega))^4,$$

(6.12) $$\lim_{n\to\infty} p^n = p \text{ strongly in } (L^2(\Omega))^2,$$

(6.13) $\lim_{n\to\infty} \lambda^n = \lambda^*$ *weakly in* $(L^2(\Omega))^2$,

where λ^* *is such that* $\{\{\psi,s\}, p,\lambda^*\}$ is a saddle point of $\mathcal{L}_r$ on $V \times H \times H$. ∎

It is clear that, once again, the essential difficulty with this approach lies in the fact that system (6.8) has to be solved at each iteration; in view of the structure of $\mathcal{L}_r$, this problem can be solved by a block over-relaxation method like that described in Chapter III, Section 3.2. As far as the choice of ρ is concerned, numerical experiments indicate once again that the optimal value lies close to $\rho = r$.

6.4 On variants of algorithm (6.6) - (6.8)

The first of these variants is algorithm ALG2 which has already been studied in some detail in Chapter III and used extensively in Chapters IV, V and VI. In the case under consideration here, we obtain the following:

(6.14) $\{\psi^{-1},s^{-1}\} \in V$, $\lambda^0 \in H$ *are given,*

then, for $n \geq 0$, $\{\psi^{n-1}, s^{n-1}\} \in V$, $\lambda^n \in H$ *being known, we determine* $p^n,\{\psi^n,s^n\}$ *and* λ^{n+1} *successively by*

(6.15)
$$\begin{cases} p^n \in H, \\ \mathcal{L}_r(\{\psi^{n-1},s^{n-1}\},p^n,\lambda^n) \leq \mathcal{L}_r(\{\psi^{n-1},s^{n-1}\},q,\lambda^n) \quad \forall q \in H, \end{cases}$$

(6.16)
$$\begin{cases} \{\psi^n,s^n\} \in V, \\ \mathcal{L}_r(\{\psi^n,s^n\},p^n,\lambda^n\} \leq \mathcal{L}_r(\{\phi,z\},p^n,\lambda^n) \quad \forall\{\phi,z\} \in V, \end{cases}$$

(6.17) $\lambda^{n+1} = \lambda^n + \rho(B\{\psi^n,s^n\}-p^n)$.

It follows from Chapter III that the convergence results of Theorem 6.1 still hold if instead of (6.10) we have

(6.18) $0 < \rho < \frac{1+\sqrt{5}}{2}\, r$.

We shall see that the solution of (6.15) and (6.16) does not pose any difficulties; problem (6.16) is in fact a problem of *linear* biharmonic type written in mixed formulation, as follows:

$$
\left\{
\begin{array}{l}
\textit{Find } \{\psi^n, s^n\} \in W \textit{ such that} \\
\gamma \int_\Omega \nabla\psi^n \cdot \nabla\phi \, dx + \alpha\nu \int_\Omega (s_{11}^n + s_{22}^n)(z_{11} + z_{22}) dx \\
+ (\beta\nu + r) \int_\Omega [(2s_{12}^n)(2z_{12}) + (s_{22}^n - s_{11}^n)(z_{22} - z_{11})] dx = (f, \phi) \\
+ \int_\Omega (\lambda_1^n - r p_1^n)(2z_{12}) + \int_\Omega (\lambda_2^n - r p_2^n)(z_{22} - z_{11}) dx \\
\forall \{\phi, z\} \in W.
\end{array}
\right. \tag{6.19}
$$

This problem admits one and only one solution. As regards (6.15), it can easily be shown that this problem admits a unique solution given explicitly by

$$
p^n = \frac{1}{r} (|X^n| - g)^+ \frac{X^n}{|X^n|} , \tag{6.20}
$$

where $X^n = \{X_1^n, X_2^n\}$ is defined by

$$
X_1^n = \lambda_1^n + 2r s_{12}^{n-1} , \tag{6.21}
$$

$$
X_2^n = \lambda_2^n + r(s_{22}^{n-1} - s_{11}^{n-1}) , \tag{6.22}
$$

and where $|X| = \sqrt{X_1^2 + X_2^2}$.

Once again the optimal choice for ρ appears to lie close to $\rho = r$; the optimal choice of r is a much more complicated problem since this optimal value appears to depend on ν and g; nonetheless, the comments made in Remark 6.1 in Chapter V, Section 6.1.2, again hold for the present problem.

To conclude this section on variants of ALG1, we should mention the following variant due to GABAY [1], which we have already met in Chapter IV. We shall return in Chapter IX to the properties of this algorithm which we shall refer to as ALG3:

With (6.14), (6.15) *as before, determine* $\lambda^{n+1/2}$ *by*

$$
\lambda^{n+1/2} = \lambda^n + \rho(B\{\psi^{n-1}, s^{n-1}\} - p^n) , \tag{6.23}
$$

then $\{\psi^n, s^n\} \in V$ *by*

$$\mathcal{L}_r(\{\psi^n,s^n\},p^n,\lambda^{n+1/2}) \leq \mathcal{L}_r(\{\phi,z\},p^n,\lambda^{n+1/2}) \quad \forall\, \{\phi,z\} \in V \tag{6.24}$$

and finally λ^{n+1} *by*

$$\lambda^{n+1} = \lambda^{n+1/2} + \rho(B\{\psi^n,s^n\}-p^n). \quad \blacksquare \tag{6.25}$$

It may be noted that in the above algorithm the variables $\{\phi,z\}$ and q play a symmetric role; this was not the case in ALG2, where the order chosen is governed by considerations of ellipticity, for the details of which we refer the reader to Chapter III.

Section 7 below gives the results of a number of numerical experiments in which the above algorithms (actually, finite-dimensional variants of these) have been applied to the solution of problems (3.7) and (4.1).

7. NUMERICAL EXPERIMENTS

In this section we shall describe some of the numerical results obtained by BEGIS [2] using the methods of approximation by finite elements, and of iterative solution by augmented Lagrangian, of the preceding sections.

7.1 Formulation of the test problem

Let $\Omega =]0,1[\times]0,1[$; we consider a family of Bingham flows for which (using the notation of Section 2) we have

$$\nu = 1, \tag{7.1}$$

$$\underset{\sim}{f} = \underset{\sim}{0}, \tag{7.2}$$

$$\begin{cases} \underset{\sim}{u}|_\Gamma = \underset{\sim}{b} = \{b_1,b_2\} \textit{ with } b_2 = 0 \textit{ on } \Gamma, \\ b_1(0,x_2) = b_1(1,x_2) = 0 \quad \forall x_2 \in]0,1[, \\ b_1(x_1,1) = 0 \quad \forall x_1 \in]0,1[\ , \ b_1(x_1,0) = 1 \quad \forall x_1 \in]0,1[. \end{cases} \tag{7.3}$$

We are thus dealing with problems which can be classed as flow in a cavity with a sliding wall; we note that $\int_\Gamma \underset{\sim}{b}\cdot\underset{\sim}{n}\, d\Gamma = 0$, but that

$\underset{\sim}{b} \notin H^{1/2}(\Gamma) \times H^{1/2}(\Gamma)$ (we have $\underset{\sim}{b} \in H^{s}(\Gamma) \times H^{s}(\Gamma)$ for all $s < \frac{1}{2}$).

7.2 Numerical results

Details of the practical implementation of the methods of approximation and iterative solution described in the previous sections may be found in BEGIS [2].

Steady-State Cases: For various values of g, we have shown in Figures 7.1 to 7.5 the rigid regions (shown hatched) and the visco-plastic regions (in white), as well as the streamlines (these are the equipotentials of ψ); we observe that the zones of rigidity increase with g (the asymmetries which can be seen are due to the asymmetries of the triangulation employed).

Unsteady Cases: For various values of g, we have considered the case in which the material filling the cavity Ω is initially at rest (so $\underset{\sim}{u}(x,0) = \underset{\sim}{0} \quad \forall\, x \in \Omega \Leftrightarrow \psi(x,0) = 0 \ \forall x \in \Omega$), and is then set in motion by the sliding (defined by (7.3)) of the lower wall of Ω. When the steady state has been attained (to within a certain precision), the motion of the lower wall is arrested, with the aim of observing the return to the initial state $\underset{\sim}{u} = \underset{\sim}{0}$ (i.e. $\psi = 0$) in Ω.

For various values of g (including $g = 0$) we have shown in Figure 7.6 the behaviour of

$$t \to \int_{\Omega} |\psi(x,t)|^2 \, dx \ ;$$

it may be noted that for $g > 0$ the *rigid state is attained in a finite time* which grows progressively smaller as g becomes larger; this accords with physical intuition and can be justified theoretically.

Methods for the numerical simulation of the two-dimensional flow of Bingham fluids may be found in FORTIN [2], BEGIS [1], these being based on different principles (including the direct use of the velocity-pressure formulation of Section 2); numerous numerical results are also given in these references, which are in agreement with those presented here (see also Chapter VI of GLOWINSKI-LIONS-TREMOLIERES [1], [2]).

BINGHAM FLUID

PARAMETERS:

Threshold of plasticity	0.5
Viscosity	1.0
External force: $f_1(x_1,x_2,t) = 0.0$, $f_2(x_1,x_2,t) = 0.0$	

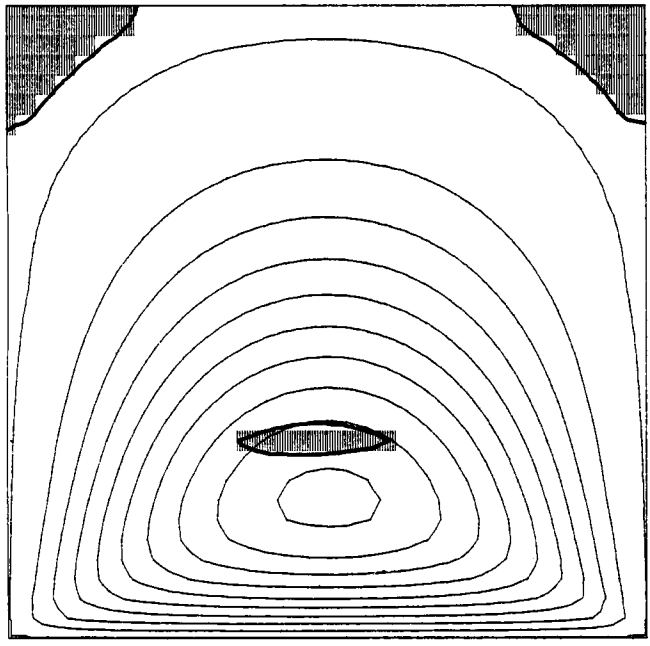
BOUNDARY CONDITIONS:

Normal component of velocity	0.0
Tangential component of velocity	0.0 at $x_1 = 0, 1; x_2 = 1$ 1.0 at $x_2 = 0$

INITIAL CONDITION:

Initial velocity	0.0

STEADY STATE



Max. value of stream function	0.928E-1
Value of stream function on line 1	0.510E-3
Difference between successive lines	0.100E-1

|||||| Rigid zones

LABORIA D.BEGIS

Figure 7.1 (g = 0.5)

BINGHAM FLUID

PARAMETERS:

Threshold of plasticity	1.0
Viscosity	1.0
External force: $f_1(x_1,x_2,t) = 0.0$,	$f_2(x_1,x_2,t) = 0.0$

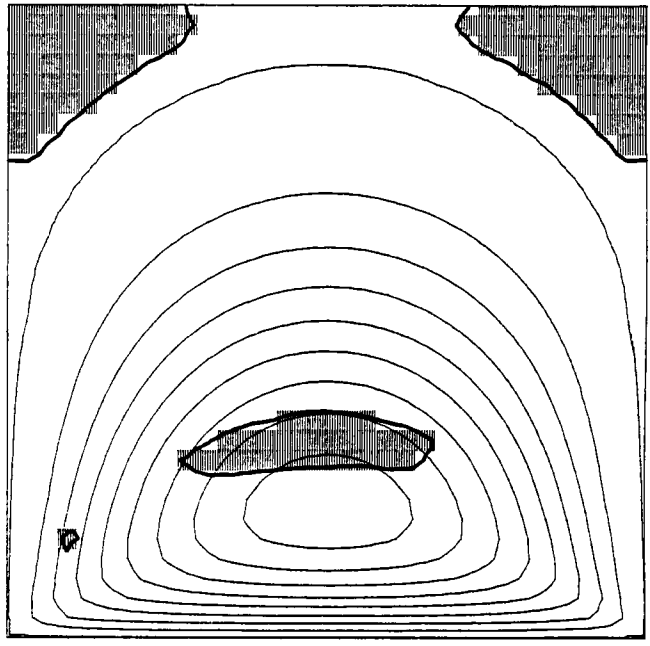
BOUNDARY CONDITIONS:

Normal component of velocity	0.0
Tangential component of velocity	0.0 at $x_1 = 0, 1; x_2 = 1$ 1.0 at $x_2 = 0$

INITIAL CONDITION:

Initial velocity	0.0

STEADY STATE



Max. value of stream function	0.864E-1
Value of stream function on line 1	0.510E-3
Difference between successive lines	0.100E-1

||||| Rigid zones

LABORIA D.BEGIS

Figure 7.2 (g = 1)

BINGHAM FLUID

PARAMETERS:

Threshold of plasticity	2.5
Viscosity	1.0
External force: $f_1(x_1,x_2,t) = 0.0$,	$f_2(x_1,x_2,t) = 0.0$

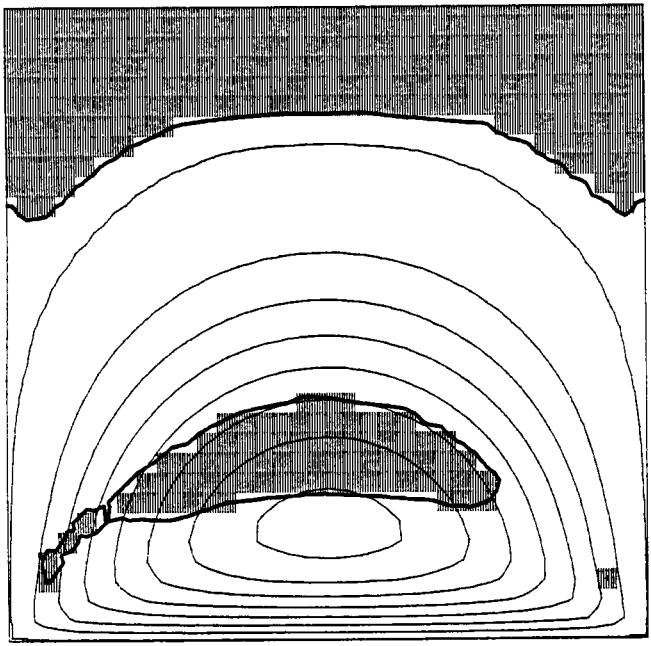
BOUNDARY CONDITIONS:

Normal component of velocity	0.0
Tangential component of velocity	0.0 at $x_1 = 0, 1; x_2 = 1$ 1.0 at $x_2 = 0$

INITIAL CONDITION:

Initial velocity	0.0

STEADY STATE



Max. value of stream function	0.737E-1
Value of stream function on line 1	0.510E-3
Difference between successive lines	0.100E-1

||||| Rigid zones

LABORIA D.BEGIS

Figure 7.3 (g = 2.5)

BINGHAM FLUID

PARAMETERS:

Threshold of plasticity	5.0
Viscosity	1.0
External force: $f_1(x_1,x_2,t) = 0.0$,	$f_2(x_1,x_2,t) = 0.0$

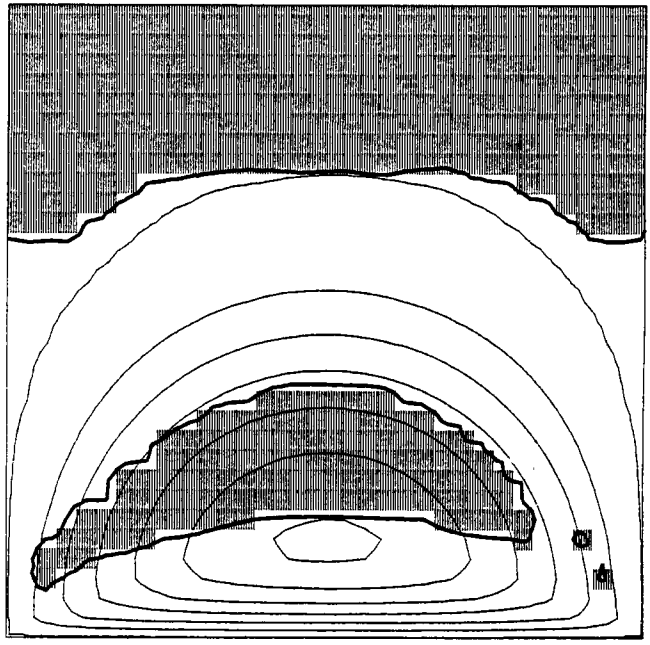
BOUNDARY CONDITIONS:

Normal component of velocity	0.0
Tangential component of velocity	0.0 at $x_1 = 0, 1; x_2 = 1$ 1.0 at $x_2 = 0$

INITIAL CONDITION:

Initial velocity	0.0

STEADY STATE



Max. value of stream function	0.616E-1
Value of stream function on line 1	0.510E-3
Difference between successive lines	0.100E-1

||||| Rigid zones

LABORIA D.BEGIS

Figure 7.4 (g = 5)

BINGHAM FLUID

PARAMETERS:

Threshold of plasticity	10.0
Viscosity	1.0
External force: $f_1(x_1,x_2,t) = 0.0$,	$f_2(x_1,x_2,t) = 0.0$

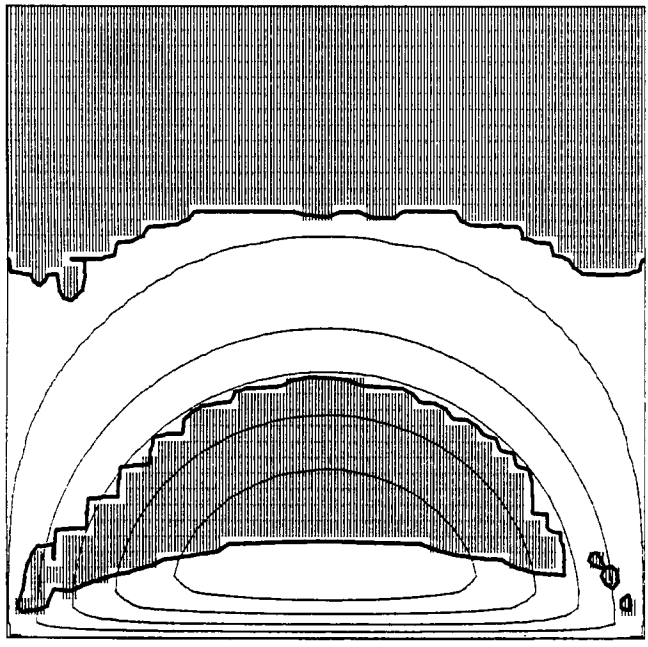
BOUNDARY CONDITIONS:

Normal component of velocity	0.0
Tangential component of velocity	0.0 at $x_1 = 0, 1; x_2 = 1$ 1.0 at $x_2 = 0$

INITIAL CONDITION:

Initial velocity	0.0

STEADY STATE



Max. value of stream function	0.500E-1
Value of stream function on line 1	0.500E-3
Difference between successive lines	0.100E-1

|||||| Rigid zones

LABORIA D.BEGIS

Figure 7.5 (g = 10)

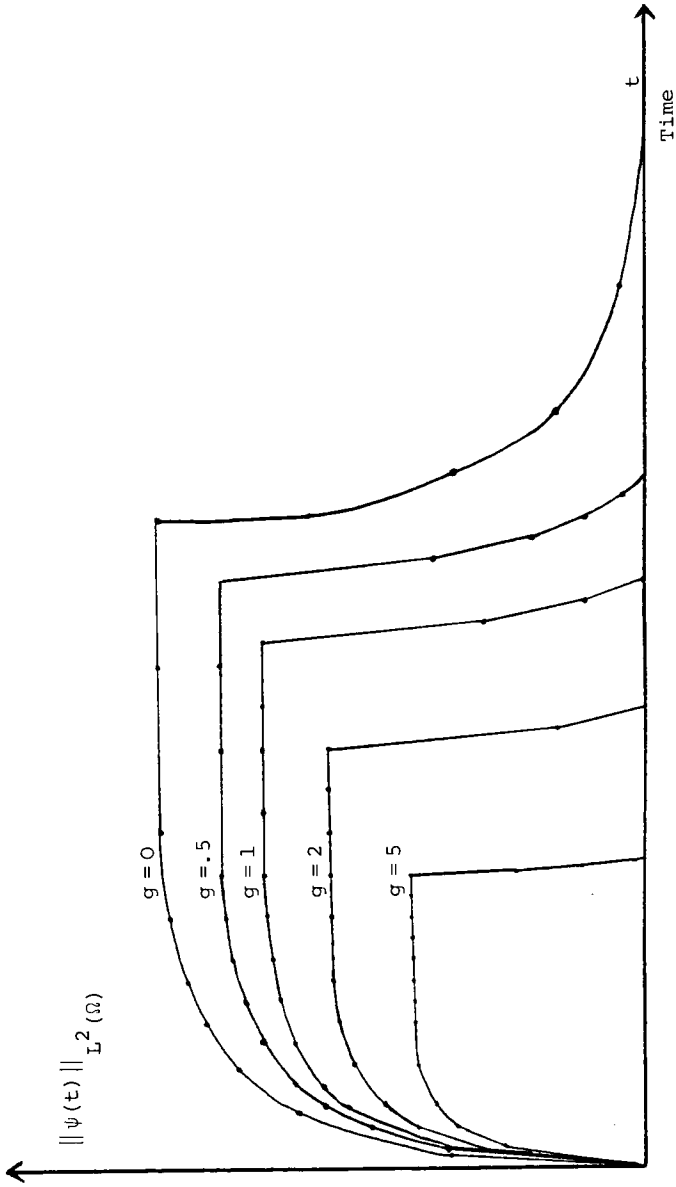


Figure 7.6

This Page Intentionally Left Blank

CHAPTER VIII

APPLICATION TO THE SOLUTION OF FINITE NONLINEAR ELASTICITY PROBLEMS

J.F. Bourgat, R. Glowinski, P. Le Tallec

1. GENERAL NOTES. SYNOPSIS

The aim of this chapter is to show that the general principles of *decomposition-coordination* studied in Chapter III have a range of application considerably wider than that considered in Chapter III, which arose from *Convex Analysis* and *Monotone Operators;* in fact, we shall show in this chapter that these principles and the associated algorithms will lead to the development of iterative methods, still related to ALG1, ALG2 (and possibly ALG3), which are extremely efficient for solving *non-convex variational problems* arising from *Nonlinear Elasticity*, in which the *displacements* and/or the *strains* are large relative to the more usual context of Linear Elasticity, where they are assumed to be 'very small'; this *finite* character of the displacements and/or the strains justifies the title of the present chapter which must necessarily be considered as merely an introduction to a vast and difficult subject which is as yet still relatively undeveloped in terms of numerical methodology.

The principles and methods mentioned above will be applied to the solution of two types of problem arising from *Finite Nonlinear Elasticity*; namely:

(i) In Section 3, the *large-displacement* calculation of equilibrium configurations for a class of *inextensible and flexible pipelines.*

(ii) In Section 4, the mechanical behaviour of *incompressible, elastic* materials of Mooney-Rivlin type.

The results of numerical experiments will be presented and discussed for both the above cases.

Although the methods used have their formal origin in Chapter III,

it seems desirable to repeat (without proof) the essential points relating to the principles and methods of Chapter III, so as to improve the readability of the present chapter; we do this in Section 2, below.

To conclude this introduction we would like to point out the following selection of works in which various aspects of the above nonlinear elasticity problems are developed:

BOURGAT-DUMAY-GLOWINSKI [1], GLOWINSKI-LE TALLEC [1], [2], LE TALLEC [1], [2], GLOWINSKI-LE TALLEC-RUAS [1], RUAS [1] and especially BOURGAT-GLOWINSKI-LE TALLEC [1] on which the contents of this chapter are heavily based.

2. DECOMPOSITION OF VARIATIONAL PROBLEMS. ASSOCIATED ALGORITHMS.

In this section we shall briefly summarise the various considerations which were developed in detail in Chapter III; this will allow the reader who is more particularly interested in the applications treated in this section to tackle it directly without first having to read Chapter III (which can therefore be postponed to a second reading).

2.1 A family of variational problems

In the following we shall restrict our attention to real Hilbert spaces; we thus let V and H be two such spaces, equipped with the norms and inner products

$$\|\cdot\|,\ ((\cdot,\cdot)) \quad \text{and} \quad |\cdot|,\ (\cdot,\cdot),$$

respectively. Let $B \in \mathcal{L}(V,H)$ and let F and G be two *convex, proper, lower semi-continuous* functionals from H and V into $\mathbb{R} \cup \{+\infty\}$, respectively; we assume that

$$\text{dom } (G) \cap \text{dom } (F \circ B) \neq \emptyset, \tag{2.1}$$

where

$$\text{dom } (G) = \{v \mid v \in V,\ -\infty < G(v) < +\infty\},$$

with a similar definition for dom (FoB). We associate with V, H, B, F, G, above, the *minimisation problem:*

$$(P) \quad \begin{cases} \textit{Find } u \in V \textit{ such that} \\ \\ J(u) \leq J(v) \quad \forall v \in V, \end{cases}$$

where $J : V \to \overline{\mathbb{R}}$ is defined by

$$J(v) = F(Bv) + G(v). \tag{2.2}$$

The functional $J(\cdot)$ and problem (P) have a very special structure; thus it is natural to think in terms of using methods which take advantage of this structure.

Remark 2.1: The majority of the considerations which follow can be applied to variational problems of the form

$$f \in B'A_1(Bu) + A_2(u), \tag{2.3}$$

where $f \in V'$ (the *dual* space of V) and where A_1 (resp. A_2) are *monotone operators* (possibly multivalued) from H into H' (the dual of H) (resp. from V into V'); the operator $A = B' \circ A_1 \circ B + A_2$ from V into V' is not in general the gradient (or subgradient [1]) of a functional J (B' denotes the *transpose* of the operator B).

For numerical results relating to these generalisations we refer the reader to LIONS-MERCIER [1], GABAY [1] (see also Chapter IX of the present book and GLOWINSKI-LIONS-TREMOLIERES [2, Appendix 2]).

If we assume that in addition to (2.1) we also have

$$\lim_{\|v\| \to +\infty} J(v) = +\infty \tag{2.4}$$

then (P) admits a solution which is *unique* if J is *strictly convex*.

Remark 2.2: The applications to Nonlinear Elasticity in Sections 3 and 4 actually relate to *non-convex* minimisation problems.

2.2 A decomposition principle

We shall now briefly summarise the developments of Chapter III;

[1] See EKELAND-TEMAM [1] for this concept.

we thus define $W \subset V \times H$ by

$$W = \{\{v,q\} \in V \times H,\ Bv-q = 0\}\ . \tag{2.5}$$

Problem (P) is equivalent to

$$(\pi) \quad \begin{cases} \textit{Find} \quad \{u,p\} \in W \ \textit{such that} \\ j(u,p) \leq j(v,q) \quad \forall\ \{v,q\} \in W \end{cases}$$

with

$$j(v,q) = F(q)+G(v). \tag{2.6}$$

Remark 2.3: The new problem (π) clearly resembles *mixed formulations*, to the extent that the relation $Bv - q = 0$ suggests the introduction of a *Lagrange multiplier*.

Remark 2.4: Problems (P) and (π) are equivalent, but by considering (π) we have in some ways simplified the nonlinear structure of (P) though at the cost of a new variable q and of the relation

$$Bv-q = 0\ ; \tag{2.7}$$

in fact, since relation (2.7) is *linear*, some very efficient techniques exist for treating it; we shall treat it in the following work by making *simultaneous* use of *penalisation* and *Lagrange multiplier* methods, through the medium of a suitably-chosen *augmented Lagrangian*.

2.3 An augmented Lagrangian associated with (π)

Let $r > 0$; we define $\mathcal{L}_r$: $V \times H \times H \to \overline{\mathbb{R}}$ by

$$\mathcal{L}_r(v,q,\mu) = F(q)+G(v)+ \frac{r}{2} |Bv-q|^2 + (\mu,Bv-q). \tag{2.8}$$

It is shown in Chapter III, Section 2.2, that if $\{u,p,\lambda\}$ is a saddle point of $\mathcal{L}_r$ on $V \times H \times H$ (i.e.

$$\begin{cases} \{u,p,\lambda\} \in V \times H \times H \ \textit{and} \\ \mathcal{L}_r(u,p,\mu) \leq \mathcal{L}_r(u,p,\lambda) \leq \mathcal{L}_r(v,q,\mu) \quad \forall\ \{v,q,\mu\} \in V \times H \times H), \end{cases} \tag{2.9}$$

then $\{u,p\}$ is a solution of (π), i.e. u is a solution of (P) (with $p = Bu$).

2.4 A first algorithm for solving (P)

To solve (P) and (π) we shall determine the saddle points of $\mathcal{L}_r$ by a duality algorithm of the type considered in GLOWINSKI-LIONS-TREMOLIERES [1, Chapter 2],[2, Chapter 2 and Appendix 2]. Such an algorithm applied to the solution of (2.9) is algorithm ALG1, introduced in Chapter III, Section 3.1; that is:

(2.10) $\lambda^0 \in H$, *given*

then for $n \geq 0$, λ^n *being known, determine* u^n, p^n, λ^{n+1} *by*

(2.11)
$$\left\{ \begin{array}{l} \textit{Find } \{u^n,p^n\} \in V \times H \textit{ such that} \\ \\ \mathcal{L}_r(u^n,p^n,\lambda^n) \leq \mathcal{L}_r(v,q,\lambda^n) \quad \forall \{v,q\} \in V \times H, \end{array} \right.$$

(2.12) $$\lambda^{n+1} = \lambda^n + \rho(Bu^n - p^n).$$

As regards the *convergence* of (2.10) - (2.12), it is shown in Chapter III, Section 4, that under very reasonable assumptions on F, B, G and if

(2.13) $$0 < \rho < 2r$$

we have, when $n \to +\infty$:

(2.14) $u^n \to u$ *strongly in* V

(2.15) $p^n \to p = Bu$ *strongly in* H

(2.16) $\lambda^n \to \lambda$ *weakly in* H

where u is the solution of (P), and where λ is such that $\{u,p,\lambda\}$ is a saddle point of $\mathcal{L}_r$ on $V \times H \times H$.

Remark 2.5: The only nontrivial stage in the above algorithm is the solution of problem (2.11); in fact to solve (2.11), taking into consideration its very special structure, it is very convenient

to use block-relaxation algorithms such as those considered in CEA-GLOWINSKI [1], CEA [1], [2], GLOWINSKI [2, Chapter 5] and which are also used in Chapter III, Section 3.2. If we use these relaxation methods, and if in the calculation of $\{u^n,p^n\}$ we perform only a single *relaxation iteration, starting from* $\{u^{n-1},p^{n-1}\}$, we then obtain the algorithm described below in Section 2.5.

2.5 A second algorithm for solving (P)

In this case we are in fact dealing with algorithm ALG2 of Chapter III, Section 3.2, namely:

(2.17) u^{-1} *given in* V, λ^o *given in* H,

then for $n \geq 0$, u^{n-1}, λ^n *known, determine* p^n,u^n,λ^{n+1} *successively by*

$$(2.18)\qquad \begin{cases} \mathcal{L}_r(u^{n-1},p^n,\lambda^n) \leq \mathcal{L}_r(u^{n-1},q,\lambda^n) \quad \forall q \in H, \\ p^n \in H, \end{cases}$$

$$(2.19)\qquad \begin{cases} \mathcal{L}_r(u^n,p^n,\lambda^n) \leq \mathcal{L}_r(v,p^n,\lambda^n) \quad \forall v \in V, \\ u^n \in V. \end{cases}$$

$$(2.20)\qquad \lambda^{n+1} = \lambda^n + \rho(Bu^n - p^n).$$

Remark 2.6: Different variants of (2.17) - (2.20) are possible; we can for example

(i) interchange the roles of q and v (see also Remark 2.7),

(ii) update λ^n between stages (2.18), (2.19); by doing this we then obtain the following variant (due to GABAY [1]), of (2.17) - (2.20), and to which we shall return in more detail in Chapter IX:

(2.21) u^{-1} *given in* V, λ^o *given in* H,

then for $n \geq 0$, u^{n-1} *and* λ^n *given, calculate* $p^n,\lambda^{n+1/2},u^n,\lambda^{n+1}$ *by*

$$(2.22)\qquad \begin{cases} \mathcal{L}_r(u^{n-1},p^n,\lambda^n) \leq \mathcal{L}_r(u^{n-1},q,\lambda^n) \quad \forall q \in H, \\ p^n \in H, \end{cases}$$

$$\lambda^{n+1/2} = \lambda^n + \rho(Bu^{n-1} - p^n), \tag{2.23}$$

$$\begin{cases} \mathcal{L}_r(u^n, p^n, \lambda^{n+1/2}) \leq \mathcal{L}_r(v, p^n, \lambda^{n+1/2}) \quad \forall v \in V, \\ u^n \in V, \end{cases} \tag{2.24}$$

$$\lambda^{n+1} = \lambda^{n+1/2} + \rho(Bu^n - p^n) \ ; \tag{2.25}$$

q and v play a much more symmetric role in (2.21) - (2.25) than in (2.17) - (2.20). ∎

Remark 2.7: If algorithm (2.17) - (2.20) is used, it is recommended that in the second stage the problem solved should be the one which has the better ellipticity properties (see Chapter III for the motivation for such a choice); if algorithm (2.21) - (2.25) is being used, this is not important since q and v then play a symmetric role. ∎

As regards the convergence of (2.17) - (2.20), it is proved in Chapter III, Section 5, that under very reasonable assumptions on F, B, G we again have the convergence properties (2.14) - (2.16) if

$$0 < \rho < \frac{1+\sqrt{5}}{2} r. \tag{2.26}$$

2.6 Remarks on the choice of ρ and r.

For given r, the optimal choice for ρ is very close to $\rho = r$, in the light of the large number of numerical experiments performed with the algorithms in Sections 2.4 and 2.5. The choice of r is a much trickier problem; theoretically, the convergence gets progressively faster as r increases; in practice, for large values of r the system (2.11) is ill-conditioned, and to solve it accurately becomes a costly operation, whilst, for very large values of r, the rounding errors generated can be disastrous. It can thus be seen that when r increases, two phenomena with contradictory effects arise; the combined effect of these two phenomena on (2.10)-(2.12) produces an algorithm which is not very sensitive to the choice of r and which is very robust. If we use algorithms (2.17)-(2.20) and (2.21) - (2.35), with $\rho = r$, the optimal choice of r is in general a very difficult problem to analyse.

2.7 Relations with alternating-direction methods. Further discussion.

2.7.1 *Relations between algorithms (2.17) - (2.20), (2.21) - (2.25) and certain alternating-direction methods.*

We shall assume for simplicity that $V = H$ and $B = I$; we shall also assume that F and G admit as differentials (or subdifferentials) A_1 and A_2, respectively; A_1 and A_2 are necessarily *monotone operators* (possibly multivalued). Problem (P) is then equivalent to

$$0 = A_1(u)+A_2(u) \tag{2.27}$$

(where $=$ must be replaced by ϵ if A_1 and/or A_2 are multivalued).

Suppose that $\rho = r$; we then obtain by elimination of λ^n in (2.17)-(2.20):

$$u^{-1} \text{ given;} \tag{2.28}$$

then for $n \geq 0$,

$$rp^n+A_1(p^n) = ru^{n-1}-A_2(u^{n-1}), \tag{2.29}$$

$$ru^n+A_2(u^n) = ru^{n-1}-A_1(p^n). \tag{2.30}$$

Putting $u^{n+1/2} = p^{n+1}$, *we finally obtain*

$$ru^{n+1/2}+A_1(u^{n+1/2}) = ru^n-A_2(u^n), \tag{2.31}$$

$$ru^{n+1}+A_2(u^{n+1}) = ru^n-A_1(u^{n+1/2}). \tag{2.32}$$

In (2.31), (2.32) we can recognise an *alternating-direction method of Douglas-Rachford type* (see DOUGLAS-RACHFORD [1]).

Similarly by elimination of λ^n and $\lambda^{n+1/2}$ in (2.21) - (2.25) we obtain, (still assuming $\rho = r$)

$$ru^{n+1/2}+A_1(u^{n+1/2}) = ru^n-A_2(u^n), \tag{2.33}$$

$$ru^{n+1}+A_2(u^{n+1}) = ru^{n+1/2}-A_1(u^{n+1/2}) \tag{2.34}$$

which is in fact an *alternating-direction method of Peaceman-Rachford type* (see PEACEMAN-RACHFORD [1]).

An investigation of the convergence of the above alternating-direction methods when A_1 and A_2 are monotone operators from H into H, possibly multivalued, and not necessarily the gradients of convex functionals, may be found in LIONS-MERCIER [1], GABAY [1] (see also Chapter IX).

2.7.2 *Interpretation of algorithms (2.17) - (2.20) and (2.21) - (2.25) in terms of the numerical integration of evolution equations*

It follows from Section 2.7.1 that if $\rho = r$, $V = H$ and $B = I$, then (2.17) - (2.20) and (2.21) - (2.25) can be considered as *implicit schemes*, using the decomposition $A = A_1 + A_2$ for the *numerical integration with respect to time* of the *Cauchy problem*

$$\text{(2.35)} \qquad \begin{cases} \dfrac{du}{dt} + A(u) = 0, \\ \\ u(0) = u_0. \end{cases}$$

In view of this interpretation, r appears as the inverse of a time step Δt, with $\Delta t = \frac{1}{r}$ (resp. $\Delta t = \frac{2}{r}$) for algorithm (2.17) - (2.20) (resp. (2.21) - (2.25)). As is shown in BOURGAT-DUMAY-GLOWINSKI [1], this interpretation of the above algorithms, related to the numerical integration with respect to t of the evolution problem (2.35), can be very useful for giving a qualitative understanding of the behaviour of these algorithms; for example, it is clear that the above algorithms will become progressively more reliable as r increases (i.e. as Δt becomes smaller).

The numerical integration of (2.35) using alternating-direction methods is studied in LIONS-MERCIER [1] under relatively general assumptions on A_1 and A_2.

2.7.3 *Further discussion*

The solution of problems such as (P) by decomposition-coordination methods using augmented Lagrangians seems to be due (in the context of boundary-value problems at least) to GLOWINSKI-MARROCCO [1]-[3] (see also POLYAK [1] for applications in Nonlinear Programming). For

further details and various applications we refer the reader to the other chapters of the present book and to LIONS-MERCIER [1] , GABAY [1], GLOWINSKI [1], [2] , GABAY-MERCIER [1], BOURGAT-DUMAY-GLOWINSKI [1], GLOWINSKI-MARROCCO [4], [5], MARROCCO [1], MERCIER [1], CHAN-GLOWINSKI [1] and BEGIS [2] .

As far as we are aware, the aforementioned relationships between the algorithms of Section 2.5 and alternating-direction methods were observed for the first time by CHAN-GLOWINSKI [1], [2]. We conclude this section by pointing out that the majority of the ideas considered are developed further in Chapters III and IX.

3. APPLICATIONS IN FINITE NONLINEAR ELASTICITY. (I) LARGE-DISPLACEMENT CALCULATION OF THE EQUILIBRIUM POSITIONS OF INEXTENSIBLE, FLEXIBLE PIPELINES

This section is based on the article by BOURGAT-DUMAY-GLOWINSKI [1].

3.1 Formulation of the problem

3.1.1 *General discussion*

Flexible pipelines play an important role in the exploitation of off-shore oil fields ; the engineers concerned are interested in the static and dynamic behaviour of these pipelines, in the effects of currents and tidal swells, in the problems of contact at the sea-bed and with other obstacles (for example the pipeline itself), etc.. Figure 3.1 below illustrates this type of situation and shows some of the notation which will be used in the following discussion.

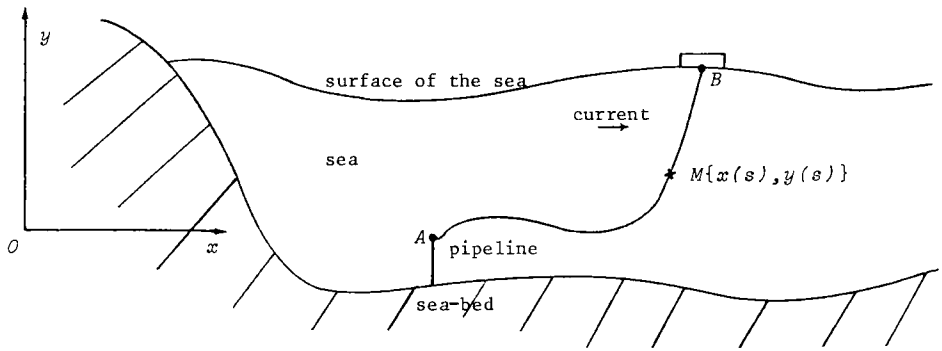


Figure 3.1

We have:

A, B : the ends of the pipeline

s : the curvilinear coordinate; s(A) = 0, s(B) = L (L: length of the pipeline)

M(s) : generic point of the pipeline, with coordinates x(s), y(s).

In this chapter we shall confine our attention to the static displacements of pipelines and we shall neglect the effect of water currents; for calculations involving currents, and for dynamic behaviour we refer to BOURGAT-DUMAY-GLOWINSKI [1].

3.1.2 *Simplifying assumptions*

We shall assume for simplicity (but also because this still allows a number of interesting preliminary results to be obtained regarding the behaviour of these pipelines), that

- *torsional effects are neglected,*
- *the pipeline is inextensible,*
- *the diameter of the pipeline is small relative to its length* L,
- *we confine ourselves to two-dimensional displacements,*
- *the pipeline is flexible and, consequently, can support large displacements whilst still obeying a linear constitutive law between the stresses and the strains.*

3.1.3 *Modelling of static problems*

The mathematical formulation of the *inextensibility condition* is given by

$$x'^2+y'^2 = 1 \quad \textit{on } [0,L] \tag{3.1}$$

(where $x' = \frac{dx}{ds}$, $y' = \frac{dy}{ds}$, $x'' = \frac{d^2x}{ds^2}$, $y'' = \frac{d^2y}{ds^2}$, etc...).

Considering the pipeline as a *beam,* and also assuming that the only external forces are those due to gravity, it is reasonable to assume that the displacement fields corresponding to *stable* equilibrium positions are solutions of the following *local minimisation* problem:

$$\underset{\{x,y\}\in\mathcal{B}}{\text{Loc min}} \left\{\frac{EI}{2}\int_0^L (x''^2+y''^2)ds + \rho g\int_0^L y\, ds\right\}, \tag{3.2}$$

where in (3.2):

(i) EI (> 0) is the flexural rigidity of the pipeline.

(ii) g is the acceleration due to gravity.

(iii) ρ is the linear density (i.e. mass per unit length) of the pipeline (if the pipeline is immersed in water, in order to take account of buoyancy effects we take

$$\rho = \rho_o - \sigma\rho_w,$$

where ρ_o and σ respectively are the *intrinsic linear density* and the *cross-sectional area* of the pipeline, and where ρ_w is the volumetric density of the water). ∎

Finally, the local minima which we are seeking are to be found in the *nonconvex set* $\mathcal{S}$ defined by

$$(3.3) \qquad \begin{cases} \mathcal{S} = \{\{x,y\} \mid x,y \in C^1[0,L],\ x'',y'' \in L^2(0,L), \\ x'^2+y'^2 = 1 \ \textit{on}\ [0,L],\ \textit{together with appropriate boundary conditions}\}. \end{cases}$$

Remark 3.1: Problem (3.2) can clearly be seen to be a *non-convex, non-quadratic, nonlinear programming problem*, which is thus inherently difficult.

Remark 3.2: By using the numerical methods which will be described below, we shall easily be able to treat cases for which EI and/or ρ are functions of s, and in which external forces and energy terms, more complicated than those in (3.2), are present in the functional to be minimised.

3.2 Results on the existence of solutions for the static problem

The mathematical study of problems such as (3.2) goes back at least as far as Euler (the problem of the "elastica", or elastic rod). Suppose that the boundary conditions are given by

$$(3.4) \qquad \begin{cases} x(0) = x_A,\ y(0) = y_A, \\ x(L) = x_B,\ y(L) = y_B \end{cases}$$

with x_A, y_A, x_B, y_B given, or by

$$(3.5) \qquad \begin{cases} x(0) = x_A,\ y(0) = y_A,\ x'(0) = \alpha_o,\ y'(0) = \beta_o, \\ x(L) = x_B,\ y(L) = y_B,\ x'(L) = \alpha_L,\ y'(L) = \beta_L, \end{cases}$$

where, in (3.5), x_A, y_A, x_B, y_B, α_o, β_o, α_L, β_L are given with $\alpha_o^2 + \beta_o^2 = 1$, $\alpha_L^2 + \beta_L^2 = 1$; the following theorem is then proved by *compactness* arguments in BOURGAT-DUMAY-GLOWINSKI [1].

THEOREM 3.1: *Suppose that* $|\vec{AB}| < L$ *and that the boundary conditions* (3.4) *or* (3.5) *are satisfied; then problem* (3.2) *admits at least one solution.*

For a detailed mathematical investigation of problems of the above type (in particular for non-uniqueness properties) we refer to ANTMAN [1] and ANTMAN-ROSENFELD [1].

3.3 Numerical solution of the static problem. (I) General Notes

The numerical solution of problems similar to (3.2) has been considered by several authors; amongst these we should mention HIBBIT-BECKER-TAYLOR [1] and MAIER-ANDREUZZI-GIANESSI-JURINA-TADDEI [1].

Problem (3.2) is in fact nontrivial from the numerical point of view; it can be dealt with by introducing a *Lagrangian* associated with the functional

$$J(x,y) = \frac{EI}{2}\int_0^L (x''^2+y''^2)ds + \rho g\int_0^L y\, ds \tag{3.6}$$

and with the (nonlinear) inextensibility condition (3.1), as follows:

$$\mathcal{L}(x,y,\mu) = J(x,y) + \frac{1}{2}\int_0^L \mu(x'^2+y'^2-1)\, ds. \tag{3.7}$$

Let λ be a *Lagrange multiplier* associated with a local minimum $\{\bar{x},\bar{y}\} \in \mathcal{B}$; it then follows from the fact that $\mathcal{L}$ is stationary that $\{\bar{x},\bar{y},\lambda\}$ must satisfy

$$\begin{cases} EI\dfrac{d^4\bar{x}}{ds^4} - \dfrac{d}{ds}\left(\lambda\dfrac{d\bar{x}}{ds}\right) = 0 \ \textit{on}\ ]0,L[\ , \\ + \textit{boundary conditions}, \end{cases} \tag{3.8}$$

$$\begin{cases} EI\dfrac{d^4\bar{y}}{ds^4} - \dfrac{d}{ds}\left(\lambda\dfrac{d\bar{y}}{ds}\right) = -g \ \textit{on}\ ]0,L[\ , \\ + \textit{boundary conditions}, \end{cases} \tag{3.9}$$

(3.10) $\bar{x}'^2+\bar{y}'^2 = 1$ *on* $]0,L[$.

It follows from (3.8) - (3.10) that λ can be viewed as a *generalised eigenvalue* to which there corresponds the *generalised eigenfunction* $\{x,y\}$.

Since the essential difficulty in problem (3.2) lies in the inextensibility constraint (3.1), it seems natural to overcome it by using the transformation $x' = \cos\phi$, $y' = \sin\phi$, where ϕ is in fact the angle between Ox and the directed tangent to the pipeline at $M = \{x,y\}$, and where $d\phi/ds$ is the curvature at M. This transformation, which is used in a number of similar problems as well as in problem (3.2) by *M.O. Bristeau* and the second author, leads to problems involving second-order differential equations (whereas the original problem is of fourth order) but in which the nonlinear structure is more complicated since it involves transcendental functions such as $t \to \sin t$ and $t \to \cos t$, the repetitive evaluation of which can be costly; furthermore the boundary conditions, of the type (3.4) (3.5), are expressed in terms of nonlinear integral relations on ϕ (see BOURGAT-DUMAY-GLOWINSKI [1, Section 3]).

A further argument for working directly with x, y instead of ϕ, is that this constitutes a starting point for the solution of much more complicated problems in Finite Nonlinear Elasticity, concerning *incompressible materials*, for which the geometric domain considered is two- or three-dimensional (a specific example of such a situation will be met in Section 4).

3.4 Numerical solution of the static problem. (II) Approximation

3.4.1. *Approximation of the space* $H^2(0,L)$ *and the functional J*

Since $\mathcal{A}$ is a subset of $H^2(0,L) \times H^2(0,L))$ an important step towards the numerical solution of (3.2) will be to define a suitable approximation of $H^2(0,L)$; to do this we introduce $\{s_i\}_{i=0}^{N}$ such that $s_i \in [0,L]$ $\forall i$, $s_0 = 0$, $s_N = L$ and $s_i < s_{i+1}$, $\forall i = 0,\ldots, N-1$. We then approximate $H^2(0,L)$ by

(3.11) $$V_h = \{v_h \in C^1[0,L],\ v_h|_{[s_i,s_{i+1}]} \in P_3,\ \forall i=0,\ldots,N-1\},$$

where, in general, P_k is the space of polynomials in a single variable of degree $\leq k$; we have $V_h \subset H^2(0,L)$ and dim $V_h = 2(N+1)$. We take $h = \max_i (s_{i+1} - s_i)$. If $v_h \in V_h$ it is convenient to define it by using

$$\{v_h(s_i)\}_{i=0}^N , \{\frac{dv_h}{ds}(s_i)\}_{i=0}^N .$$

In view of these degrees of freedom it is obvious that V_h corresponds to an *approximation by finite elements of Hermite type* (see Figure 3.2).

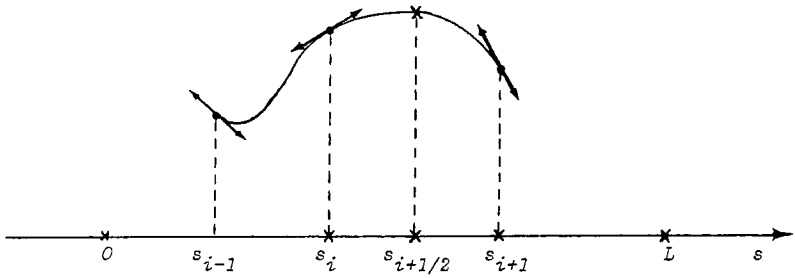


Figure 3.2

Since $V_h \times V_h \subset H^2(0,L) \times H^2(0,L)$, the functional J introduced in (3.6) is defined on $V_h \times V_h$; furthermore since $x_h, y_h, \in V_h$ implies that $(x_h''^2 + y_h''^2)|_{[s_i,s_{i+1}]} \in P_2$, the two integrals appearing in J can be calculated *exactly* by using *Simpson's rule* on each sub-interval $[s_i,s_{i+1}]$, i = 0,1,..., N - 1. By restriction of J to $V_h \times V_h$ we obtain a functional which depends on 4(N + 1) variables.

3.4.2 *Approximation of* $\mathcal{E}$

As we are using an approximation of *Hermite cubic* type, there is no difficulty in approximating boundary conditions such as (3.4) or (3.5). As regards the inextensibility condition (3.1), the obvious choice is to use

$$x_h'^2(s_i)+y_h'^2(s_i) = 1 \quad \forall i=0,1,\ldots,N. \tag{3.12}$$

Since $\{x_h'(s_i)\}_{i=0}^N$, $\{y_h'(s_i)\}_{i=0}^N$ constitute precisely a subset of the degrees of freedom used for defining x_h and y_h, the practical implementation of (3.12) does not present any difficulty. However, we have observed that for 'stiff' problems (i.e. with strong

variations in the curvature) relatively inaccurate results may be obtained by using (3.12), unless the discretisation is refined in those regions where the curvature varies strongly. Such a procedure can increase N considerably, and therefore, as far as accuracy is concerned, we have found it more convenient to approximate (3.1) by

$$(3.13)\qquad \begin{cases} x_h'^2(s_i)+y_h'^2(s_i) = 1 \quad \forall i=0,1,\ldots N, \\ x_h'^2(s_{i+1/2}) + y_h'^2(s_{i+1/2}) = 1 \quad \forall i=0,1,\ldots,N-1, \end{cases}$$

where $s_{i+1/2} = \frac{1}{2}(s_i + s_{i+1})$ (see Figure 3.2).

The numerical results presented in Section 3 were in fact obtained by using (3.13) to approximate (3.1). Using (3.12) (resp. (3.13)) to approximate (3.1) introduces around N (resp. $2N$) *quadratic equality constraints* (the exact number depends on the boundary conditions). In the following, the approximation of $\mathcal{E}$, obtained by approximating (3.1) by (3.12) or (3.13), will be denoted by $\mathcal{E}_h$; it is clear that $\mathcal{E}_h$ is a *closed* subset of $V_h \times V_h$.

3.4.3 *Approximation of problem* (3.2)

In the light of Sections 3.4.1 and 3.4.2, we approximate (3.2) by

$$(3.14)\qquad \underset{\{x_h,y_h\}\in\mathcal{E}_h}{\text{Loc min}}\ J(x_h,y_h),$$

where

$$(3.15)\qquad J(x_h,y_h) = \frac{EI}{2}\int_0^L (x_h''^2+y_h''^2)\,ds + \rho g \int_0^L y_h\,ds.$$

In relation to the existence of solutions for the discrete problem (3.14), the following variant of Theorem 3.1 is proved in BOURGAT-DUMAY-GLOWINSKI [1, Section 5.3]:

THEOREM 3.2: *Suppose that* $\mathcal{E}_h$ *is nonempty (for this it is sufficient that* $|\vec{AB}| < L$ *if the boundary conditions are given by* (3.4) *or* (3.5)); *problem* (3.14) *then admits at least one solution.*

3.4.4 *Convergence of the approximate solutions*

We shall now introduce the notion of an *isolated solution* for the local minimisation problem (3.2):

DEFINITION 3.1: *Let* $\{\bar{x},\bar{y}\}$ *be a solution of* (3.2); *we say that* $\{\bar{x},\bar{y}\}$ *is an isolated solution of* (3.2) *if there exists a neighbourhood* η *of* $\{\bar{x},\bar{y}\}$ *such that*

$$J(\bar{x},\bar{y}) < J(x,y) \quad \forall \{x,y\} \in \eta \cap \mathcal{K}, \ \{x,y\} \neq \{\bar{x},\bar{y}\}. \quad (3.16)$$ ■

The convergence results of the following theorem are proved in BOURGAT-DUMAY-GLOWINSKI [1, Section 5.4]:

THEOREM 3.3: *Suppose that the boundary conditions appearing in the definition of* $\mathcal{K}$ *are given by* (3.4) *or* (3.5); *suppose also that* $\mathcal{K}_h$ *is defined via* (3.12). *We then have: if* $\{\bar{x},\bar{y}\}$ *is an isolated solution of* (3.2), *then for* h *sufficiently small the approximate problem* (3.14) *admits a solution* $\{\bar{x}_h,\bar{y}_h\}$ *in the neighbourhood of* $\{\bar{x},\bar{y}\}$; *we also have*

$$\lim_{h\to 0} \{\bar{x}_h,\bar{y}_h\} = \{\bar{x},\bar{y}\} \text{ strongly in } H^2(0,L)\times H^2(0,L). \quad (3.17)$$

Similarly we may prove:

THEOREM 3.4: *Suppose that* $\mathcal{K}$ *and* $\mathcal{K}_h$ *are as in the statement of Theorem* 3.3. *Then if* $(\{\bar{x}_h,\bar{y}_h\})_h$ *is a family of global minima for* J *on* $\mathcal{K}_h$, *we have (at least for a subsequence)*

$$\lim_{h\to 0} \{\bar{x}_h,\bar{y}_h\} = \{\bar{x},\bar{y}\} \text{ strongly in } H^2(0,L)\times H^2(0,L),$$

where $\{\bar{x},\bar{y}\}$ *realises the global minimum of* J *on* $\mathcal{K}$.

Remark 3.3: Here, we have not considered the convergence of the approximate solutions when $\mathcal{K}_h$ is defined via (3.13); in such a case we are dealing with a much more difficult problem. Neither have we considered the behaviour of the approximate solutions in the neighbourhood of turning points or genuine bifurcation points; in this direction we mention the works of F. KIKUCHI [1], YAMAGUTI-FUJII [1], KESAVAN [1], BREZZI-RAPPAZ-RAVIART [1]-[3].

3.5 Numerical solution of the static problem. (III) Iterative methods of solution

In this section we follow BOURGAT-DUMAY-GLOWINSKI [1, Section 6].

3.5.1 *General notes and synopsis*

Although it is clear that the considerations discussed below should strictly have been developed for the approximate problems, we have nonetheless chosen to use the continuous problem since both these types of problem have the same nonlinear structure; furthermore, the formalism of the continuous problem is simpler. The nonlinear equality constraints (3.1) (or their finite-dimensional variants (3.12), (3.13)) constitute the major difficulty to be surmounted in the numerical solution of the local minimisation problem (3.2), when working directly with the displacements x, y. Schematically, for solving (3.2) and its finite-dimensional variants, two families of methods are available:

(i) Methods using multipliers and penalisation

As was seen in Section 3.3, we can associate a *Lagrange multiplier function* λ with the equality constraint (3.1); in doing this one has to solve, with respect to $\{\bar{x},\bar{y},\lambda\}$, the nonlinear differential system (3.8) - (3.10) (actually its finite-dimensional variants); in practice, the discrete variants of (3.8) - (3.10) can be solved by the variable metric methods developed by POWELL [2] which generalise the widely known method of *Davidon-Fletcher-Powell* (see also MATTHIES-STRANG [1] for some similar methods directed at the solution of problems in Nonlinear Mechanics). We have the impression that the methods mentioned above are trickier to implement than the methods described below in Section 3.5.2, and also that they have much greater computer memory requirements when large-scale problems are being treated.

It is natural to associate with the *Lagrangian* $\mathcal{L}$ defined by (3.7) the *augmented Lagrangian* $\mathcal{L}_r$ defined (with $r > 0$) by

$$\mathcal{L}_r(x,y,\mu) = \mathcal{L}(x,y,\mu) + \frac{r}{4}\int_0^L (x'^2+y'^2-1)^2\, ds. \tag{3.18}$$

If we replace $\mathcal{L}$ by $\mathcal{L}_r$, the conditions for $\mathcal{L}_r$ to be stationary lead to the following variant of (3.8) - (3.10):

$$(3.19)\qquad \begin{cases} EI\,\bar{x}^{(4)} - \frac{d}{ds}\left(\lambda \frac{d\bar{x}}{ds}\right) - r\frac{d}{ds}\left((\bar{x}'^2+\bar{y}'^2-1)\frac{d\bar{x}}{ds}\right) = 0 \ \textit{on}\ ]0,L[, \\ + \textit{boundary conditions} \end{cases}$$

$$(3.20)\qquad \begin{cases} EI\,\bar{y}^{(4)} - \frac{d}{ds}\left(\lambda \frac{d\bar{y}}{ds}\right) - r\frac{d}{ds}\left(\bar{x}'^2+\bar{y}'^2-1)\frac{d\bar{y}}{ds}\right) = -\rho g \ \textit{on}\ ]0,L[, \\ + \textit{boundary conditions} \end{cases}$$

$$(3.21)\qquad \bar{x}'^2+\bar{y}'^2-1 = 0 \ \textit{on}\ ]0,L[,$$

which is clearly equivalent to system (3.8) - (3.10).

It is clear that the above approach, using an augmented Lagrangian, further complicates a problem which is already complicated enough in its own right since (3.19) - (3.21) is even '*more nonlinear*' than (3.8) - (3.10); furthermore, (3.19) - (3.21) are 'more coupled' than (3.8) - (3.10). If we take $\lambda = 0$ in (3.19), (3.20), and if we do not consider (3.21), we obtain the *necessary conditions* of optimality for a problem deduced from (3.2) by *penalisation* of the condition $x'^2 + y'^2 - 1 = 0$. ■

(ii) Methods using direct minimisation on manifolds

Instead of 'relaxing' the constraint (3.1), i.e. $x'^2 + y'^2 - 1 = 0$, by *Lagrange multipliers* and/or *penalisation*, we can attempt to minimise J *directly* on the *manifold* defined by (3.1), as is done in GABAY [1] and LICHNEWSKY [1] (by the *optimal descent* or the *conjugate gradient* method). However, although these methods are extremely elegant in their underlying principles and are very efficient for certain problems in that they perform the minimisation on the *geodesics* of the manifold, they are in practice somewhat difficult to implement if the number of constraints is very large; this is certainly the case for the discrete variants of (3.2) described in Section 3.4. ■

The methods which we shall describe in Section 3.5.2 differ quite considerably from the two types of method mentioned above; nonetheless they do have a certain number of characteristics in common with them, in the sense that:

(1) They are also based on the use of an *augmented Lagrangian;* in the present case, however, the constraints to be treated by Lagrange multipliers and penalisation are *linear*, and this constitutes a substantial simplification.

(2) We retain the notion of *direct minimisation* on a manifold which (in a certain sense) is associated with the inextensibility condition (3.1).

3.5.2 *Solution of problem (3.2) by an augmented Lagrangian method*

In spite of the fact that problem (3.2) is non-convex, to solve it we shall apply the methodology developed in Chapter III and summarised in Section 2 of the present chapter. In the present context, problem (P) is problem (3.2), that is

$$(P) \qquad \underset{\{x,y\}\in \mathcal{E}}{\text{Loc min}} \left\{\frac{EI}{2}\int_0^L (x''^2+y''^2)ds + \rho g \int_0^L y\, ds\right\},$$

with $\mathcal{E}$ defined by (3.3). We then have the following very obvious proposition :

Proposition 3.1: *The problem* (P) *is equivalent to the problem*

$$(\pi) \qquad \underset{\{x,y,p,q\}\in \tilde{\mathcal{E}}}{\text{Loc min}} \left\{\frac{EI}{2}\int_0^L (x''^2+y''^2)ds + \rho g \int_0^L y\, ds\right\}$$

with

$$(3.22) \qquad \tilde{\mathcal{E}} = \{\{x,y,p,q\}\in Z\times (L^2(0,L))^2,\ x'=p,\ y'=q,\ p^2+q^2 = 1\},$$

where Z *is the subspace of* $H^2(0,L)\times H^2(0,L)$ *defined by the boundary conditions specified for* $\{x,y\}$ *in the definition of* $\mathcal{E}$.

The next step is to 'relax' the functional relation between $\{x,y\}$ and $\{p,q\}$ by introducing (with $r > 0$) the following *augmented Lagrangian:*

$$(3.23) \qquad \begin{cases} \mathcal{L}_r(x,y,p,q,\lambda,\mu) = \dfrac{EI}{2}\displaystyle\int_0^L (x''^2+y''^2)ds + \rho g \int_0^L y\, ds \\ + \displaystyle\int_0^L \lambda(x'-p)ds + \int_0^L \mu(y'-q)ds + \frac{r}{2}\int_0^L |x'-p|^2\, ds + \frac{r}{2}\int_0^L |y'-q|^2 ds. \end{cases}$$

By analogy with the convex situation described in Chapter III and in Section 2 of this chapter, we suppose that $\{\bar{x},\bar{y},\bar{p},\bar{q},\bar{\lambda},\bar{\mu}\}$ is a *(local) saddle point* for $\mathcal{L}_r$ on $Z \times S \times (L^2(0,L))^2$, where

$$S = \{\{p,q\} \in (L^2(0,L))^2,\ p^2+q^2 = 1 \ a.e.\}\ ; \tag{3.24}$$

it can then be proved that $\{\bar{x},\bar{y}\} \in \hat{\mathcal{E}}$, that $\bar{x}' = \bar{p}'$, $\bar{y}' = \bar{q}$, and that $\bar{\lambda},\bar{\mu}$ are Lagrange multipliers for the equality constraints $x' - p = 0$, $y'-q = 0$.

In view of these properties, it is thus natural to extend to the Lagrangian $\mathcal{L}_r$ defined by (3.23) the iterative methods of Chapter III, the description of which is repeated in Sections 2.4 and 2.5; the corresponding algorithms are described in Sections 3.5.3 and 3.5.4 below.

3.5.3 *A first iterative method using* $\mathcal{L}_r$

This is in fact algorithm ALG1 of Chapter III, Section 3.1, and of Section 2.4 of the present chapter. Using the notation of Section 3.5.2 above, this algorithm is written:

$$\lambda^0,\mu^0 \ \textit{are given;} \tag{3.25}$$

then for $n \geq 0$, *assuming that* λ^n *and* μ^n *are known, calculate* $x^n,y^n,p^n,q^n,\lambda^{n+1},\mu^{n+1}$ *by*

$$\begin{cases} \textit{Find } \{x^n,y^n,p^n,q^n\} \in Z \times S \ \textit{ such that } \ \forall \{x,y,p,q\} \in Z \times S \\ \mathcal{L}_r(x^n,y^n,p^n,q^n,\lambda^n,\mu^n) \leq \mathcal{L}_r(x,y,p,q,\lambda^n,\mu^n) \ \textit{(locally, at least)}, \end{cases} \tag{3.26}$$

$$\begin{cases} \lambda^{n+1} = \lambda^n + \tilde{\rho}\left(\frac{dx^n}{ds} - p^n\right), \\ \mu^{n+1} = \mu^n + \tilde{\rho}\left(\frac{dy^n}{ds} - q^n\right). \end{cases} \tag{3.27}$$

The non-trivial part of algorithm (3.25) - (3.27) is obviously the solving of problem (3.26); we can again proceed by *block relaxation* (see Section 2.4, Remark 2.5) by minimising alternately with respect to $\{x,y\}$ and $\{p,q\}$; if we confine this to a *single inner iteration* we obtain - with a suitable initialisation - the variant of algorithm

(3.25) - (3.27) described in Section 3.5.4 below.

3.5.4 *A second iterative method using* $\mathcal{L}_r$

In this case the algorithm we use is ALG2 of Chapter III, Section 3.2, and of Section 2.5 of the present chapter; that is:

(3.28) $\lambda^1, \mu^1, x^o, y^o$ *given;*

then for $n \geq 1$, *assuming that* x^{n-1}, y^{n-1}, λ^n, μ^n *are known, calculate* $\{p^n,q^n\}$, $\{x^n,y^n\}$ *and* $\{\lambda^{n+1}, \mu^{n+1}\}$ *by*

(3.29)
$$\begin{cases} \textit{Find} \quad \{p^n,q^n\} \in S \quad \textit{such that} \quad \forall \{q,p\} \in S \\ \mathcal{L}_r(x^{n-1},y^{n-1},p^n,q^n,\lambda^n,\mu^n) \leq \mathcal{L}_r(x^{n-1},y^{n-1},p,q,\lambda^n,\mu^n), \end{cases}$$

(3.30)
$$\begin{cases} \textit{Find} \quad \{x^n,y^n\} \in Z \quad \textit{such that} \quad \forall \{x,y\} \in Z, \\ \mathcal{L}_r(x^n,y^n,p^n,q^n,\lambda^n,\mu^n) \leq \mathcal{L}_r(x,y,p^n,q^n,\lambda^n,\mu^n) \ , \end{cases}$$

(3.31)
$$\begin{cases} \lambda^{n+1} = \lambda^n + \tilde{\rho}\left(\frac{dx^n}{ds} - p^n\right), \\ \mu^{n+1} = \mu^n + \tilde{\rho}\left(\frac{dy^n}{ds} - q^n\right). \end{cases}$$

Remark 3.4: A variant of algorithm (3.29) - (3.31) is given in BOURGAT-DUMAY-GLOWINSKI [1, Section 6.2.1] in which use is made of a relaxation parameter in the calculation of $\{x^n,y^n\}$. ■

Remark 3.5: We could also use, instead of (3.28) - (3.31), the variant deduced from algorithm (2.21) - (2.25) in Section 2.5, Remark 2.6. ■

From the practical point of view it is essential to have equivalent, more explicit, formulations for (3.29) and (3.30). In this direction, it may be noted that (3.30) is in fact equivalent to the following fourth-order boundary-value system:

(3.32)
$$\begin{cases} EI\frac{d^4x^n}{ds^4} - r\frac{d^2x^n}{ds^2} = \frac{d}{ds}(\lambda^n - rp^n) \ \textit{on} \ ]0,L[\ , \\ + \ \textit{boundary conditions}, \end{cases}$$

$$(3.33)\qquad \begin{cases} EI\,\dfrac{d^4y^n}{ds^4} - r\,\dfrac{d^2y^n}{ds^2} = \dfrac{d}{ds}\,(\mu^n - rq^n) - \rho g \quad \textit{on }]0,L[\ , \\ + \textit{ boundary conditions.} \end{cases}$$

If the boundary conditions are given by (3.4) or (3.5), we can then solve (3.32) and (3.33) *independently of each other*, and furthermore their discretised versions are *linear systems* with the *same matrix*; this matrix is *sparse, symmetric, positive definite and independent of* n *if* r *is fixed;* in this case we can perform a *Cholesky factorisation* once and for all, and at each iteration of (3.28) - (3.31) we shall have to solve only 4 *sparse, triangular, well-posed* systems to determine $\{x^n,y^n\}$.

We shall now study the solution of (3.29); to obtain $\{p^n,q^n\}$ it is necessary to solve, a.e. on [0,L] the *two-dimensional* minimisation problem

$$(3.34)\qquad \begin{cases} \underset{\{p(s),q(s)\}}{\text{Min}} \Big\{ \dfrac{r}{2}\,(p^2(s)+q^2(s)) - (\lambda^n(s) + r\,\dfrac{dx^{n-1}}{ds}\,(s))p(s) \\ - (\mu^n(s) + r\,\dfrac{dy^{n-1}}{ds}\,(s))q(s)\Big\}\ , \textit{ with } \{p(s),q(s)\} \in \mathbb{R}^2,\ p^2(s)+q^2(s)=1. \end{cases}$$

However, since $p^2(s) + q^2(s) = 1$, (3.34) reduces to

$$(3.35)\qquad \begin{cases} \underset{\{p(s),q(s)\}}{\text{Max}} \{X^n(s)p(s) + Y^n(s)q(s)\}\ , \\ \textit{with } \{p(s),q(s)\} \in \mathbb{R}^2,\ p^2(s)+q^2(s) = 1, \end{cases}$$

where, in (3.35), we have put

$$X^n(s) = \lambda^n(s) + r\,\frac{dx^{n-1}}{ds}\,(s),$$

$$Y^n(s) = \mu^n(s) + r\,\frac{dy^{n-1}}{ds}\,(s).$$

If $\{X^n(s),\ Y^n(s)\} \neq \{0,0\}$, we have

$$(3.36)\qquad \begin{cases} p^n(s) = \dfrac{X^n(s)}{\sqrt{|X^n(s)|^2 + |Y^n(s)|^2}}\ , \\ q^n(s) = \dfrac{Y^n(s)}{\sqrt{|X^n(s)|^2 + |Y^n(s)|^2}}\ . \end{cases}$$

Remark 3.6: We have just shown that (3.30) is a *well-posed* problem if the boundary conditions are given by (3.4) or (3.5). Problem (3.29) is also well posed if $\{X^n(s),Y^n(s)\} \neq \{0,0\}$; if $X^n(s) = Y^n(s) = 0$, the entire circle $p^2 + q^2 = 1$ is a solution. In actual fact, in *all* the numerical experiments which we have performed, we have noted that this problematical situation never arose if r was *sufficiently large*; it is possible to account for such behaviour.

Remark 3.7: In accordance with Remark 2.7 of Section 2.5, we solve the problem in {x,y} at the second step of algorithm (3.28) - (3.31), as this problem is associated with a *strongly elliptic* operator (in contrast to the problem in {p,q} which is associated with a *non-monotone, multivalued* operator).

3.6 Numerical experiments

In this section we shall describe and discuss the numerical results obtained in solving a number of test problems; we refer to BOURGAT-DUMAY-GLOWINSKI [1] for further numerical tests, and in particular for the numerical solution of problems in which there are water currents acting on the pipeline, and of dynamic problems (oscillations, for example) concerning this pipeline.

3.6.1 *Description of the test problem*

Mechanical parameters:

$EI = 7000\ Nm^2$, $\rho = 7.67$ Kg/m, $L = 32.6$ m.

Boundary conditions:

$x(0) = y(0)$, $x'(0) = 1$, $y'(0) = 0$,

$x(L) = 1,2,3,4,5,6,7,8$; $y(L) = 0$, $x'(L) = 1$, $y'(L) = 0$.

3.6.2 *Further information concerning the numerical solution*

For approximating (3.2) we used a *uniform* discretisation of [0,L] with h = L/50 and the approximation described in Section 3.4. The approximate problems were solved by a discretised variant of algorithm (3.28) - (3.31) with $\tilde{\rho} = r = 50\ 000$. For the termination test we took a discretised version of

$$(3.37)\qquad \frac{\int_0^L \{|x^n-x^{n-1}|+|y^n-y^{n-1}|+|x'^n-x'^{n-1}|+|y'^n-y'^{n-1}|\}\,ds}{\int_0^L \{|x^n|+|y^n|+|x'^n|+|y'^n|\}\,ds} \le 10^{-5}.$$

3.6.3 *Presentation of the numerical results*

(i) We show in Figure 3.3, for x(L) = 2,3,4,5,6 the numerical results which were obtained as follows:

We first calculated the solution corresponding to x(L) = 6 by initialising in (3.28) with

$$(3.38)\qquad \begin{cases} \lambda^1 = \mu^1 = 0, \\ x^o(s) = 3(1-\cos \pi \frac{s}{L}),\ y^o(s) = -3 \sin \pi \frac{s}{L}, \end{cases}$$

which corresponds to a *semicircle* with diameter AB; as the length of this semicircle is $3\pi = 9.424 \ldots$, we can see that the initial

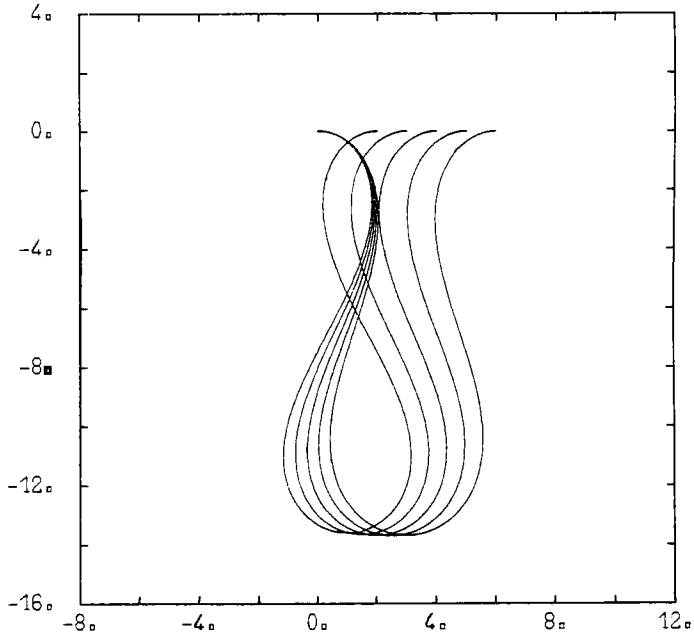


Figure 3.3 (x(L) = 2,3,4,5,6)

solution lies a long way from the solution required; convergence was reached in 166 iterations of algorithm (3.28) - (3.31). For x(L) = 5,4,3,2 (this was the order we actually followed) we used a kind of *incremental method*, the initialisation of (3.28) - (3.31) being performed by using the results obtained for the previous value of x(L).

For reasons of clarity the solutions corresponding to x(L) = 6,4,2, respectively, are pictured individually in Figures 3.4, 3.5 and 3.6.

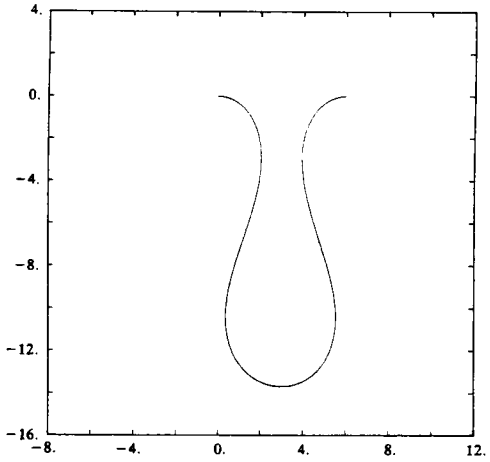


Figure 3.4: (x(L)=6)

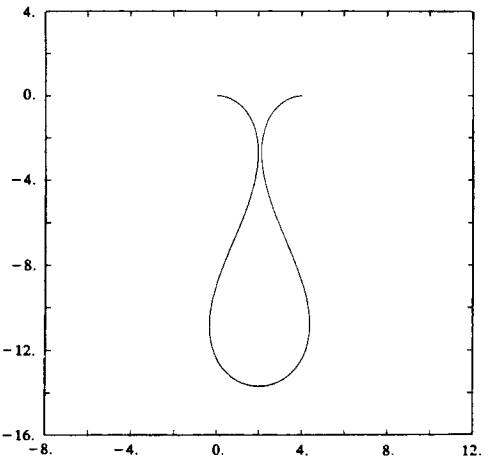


Figure 3.5: (x(L)=4)

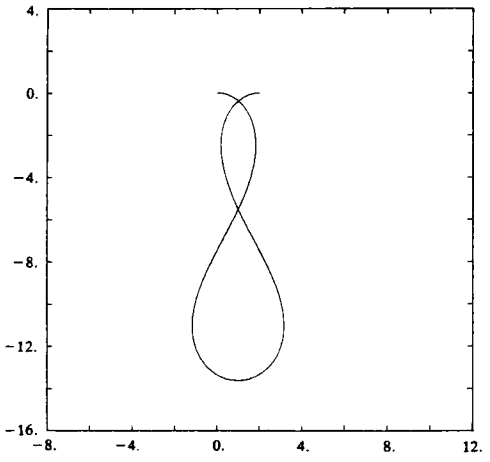


Figure 3.6: (x(L)=2)

Table 3.1 shows the number of iterations required for convergence, using the termination test (3.37):

x(L)	Number of iterations
6	166
5	105
4	105
3	107
2	105

Table 3.1

The above five calculations were performed in a *single* computer run, and required three minutes on a CII/IRIS 80 computer.

(ii) Figure 3.7 shows the numerical results obtained as follows for x(L) = 1,2,3,4,5,6,7,8 : *each calculation has been performed by* initialising algorithm (3.28) - (3.31) with $\lambda^1 = \mu^1 = 0$ and $\{x^o, y^o\}$ corresponding to the lower semicircle with diameter AB ; we are thus

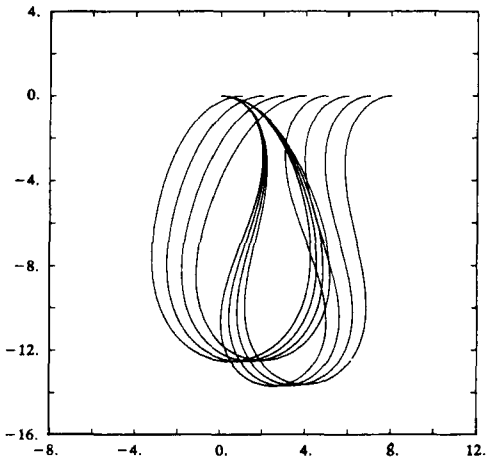


Figure 3.7: (x(L)=1,2,3,4,5,6,7,8)

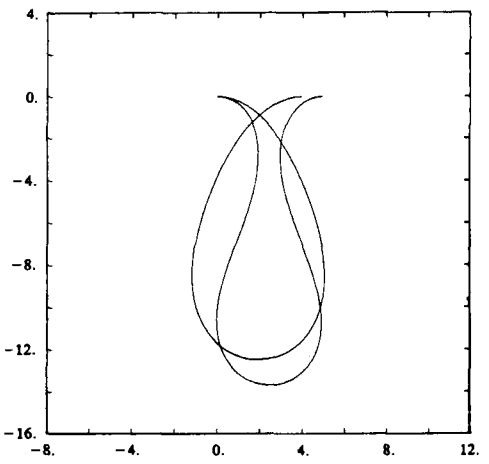


Figure 3.8: (x(L)=4, 5)

starting from a point far away from the required solution and we are not employing an incremental strategy. We observe in Figure 3.7 two types of form for the solutions calculated (this corresponds to distinct branches of solutions); it can also be seen that if x(L) is sufficiently small then the solutions in Figure 3.7 differ from those obtained in (i) using an incremental method. Since the *critical value* of x(L) for the above phenomenon seems to lie between 4 and 5, we have singled out the solutions for x(L) = 4 and 5 separately in Figure 3.8.

Table 3.2 below indicates the number of iterations required for convergence:

x(L)	Number of iterations
1	220
2	220
3	220
4	220
5	133
6	166
7	170
8	187

Table 3.2

The above eight calculations correspond to an overall execution time of 7 minutes on a CII/IRIS 80 computer.

3.6.4 *Further discussion*

Table 3.3 shows the values taken by the functional J (defined in (3.6)) for the solutions of (3.2) described in Section 3.6.3 above.

x(L)	8	7	6	5	4	3	2	1
Incremental strategy (case (i))	-8561	-8142	-7688	-7199	-6674	-6112	-5510	-4868
non-incremental strategy (case (ii))	-8561	-8142	-7688	-7199	-9434	-9702	-9932	-10124

Table 3.3

Table 3.3 demonstrates the following (hardly surprising) fact: by using an *incremental strategy* we have been able to follow one branch of solutions, despite the fact that more stable solutions exist for the same values of x(L).

4. APPLICATIONS IN FINITE NONLINEAR ELASTICITY. (II) TWO-DIMENSIONAL CALCULATIONS INVOLVING LARGE DISPLACEMENTS AND LARGE STRAINS FOR INCOMPRESSIBLE MATERIALS OF MOONEY-RIVLIN TYPE

4.1 Synopsis

The aim of this section is to assess the possibilities offered by the methods of Chapter III and of Section 2 of the present chapter, for the numerical solution of nonlinear problems arising in the field of multidimensional Finite Nonlinear Elasticity. In this section we shall be concentrating on a relatively 'simple' static problem, namely the mechanical behaviour of a *two-dimensional* body made from an *incompressible material of Mooney-Rivlin type*. The major difficulty in this problem is the *incompressibility condition* and we shall see how the *decomposition-coordination* methods of Chapter III, and of Section 2 of the present chapter, provide a simple and elegant means of overcoming this difficulty. The method described has in fact also been used successfully for the solution of the static equilibrium problem for three-dimensional bodies; problems of this kind are much more difficult, and for their numerical treatment by the methods of this book we refer the reader to GLOWINSKI-LE TALLEC [1], [2] and LE TALLEC [1], [2].

4.2 Formulation of the problem

4.2.1 *Notation. Mechanical assumptions*

A fundamental problem in Nonlinear Elasticity is the calculation of the deformations and displacements of a solid body consisting of a homogeneous, isotropic, *hyperelastic and incompressible* material, subjected to *volume forces* $\rho_o \underset{\sim}{f}$ (ρ_o is the density in the reference configuration) and to *surface forces* $\underset{\sim}{S}_o$. In a *Lagrangian formulation*, the *energy functional* corresponding to a *displacement field* $\underset{\sim}{v}$ is given by

$$(4.1) \qquad \pi(\underset{\sim}{v}) = \int_{\Omega} \rho_o(\sigma(\underset{\sim}{v}) - \underset{\sim}{f}\cdot\underset{\sim}{v})dx - \int_{\partial\Omega_2} \underset{\sim}{S}_o\cdot\underset{\sim}{v}\, d\Gamma,$$

where, in (4.1), Ω is a domain in $\mathbb{R}^N$ corresponding to the reference configuration; $\partial\Omega$ $(=\partial\Omega_1 \cup \partial\Omega_2)$ is the boundary of Ω, the body being *fixed* on $\partial\Omega_1$; we have denoted by $\sigma(\underset{\sim}{v})$ the *internal elastic energy* function (per unit mass). For a Mooney-Rivlin material we have

$$(4.2) \qquad \sigma(\underset{\sim}{v}) = E_1(I_1-2) \quad \textit{if } N=2,$$

$$(4.3) \qquad \sigma(\underset{\sim}{v}) = E_1(I_1-3)+E_2(I_2-3) \quad \textit{if } N=3,$$

where, in (4.2), (4.3), I_i is the i^{th} invariant of the tensor $\underset{\approx}{F}\underset{\approx}{F}^t$, with

$$(4.4) \qquad \underset{\approx}{F} = \underset{\approx}{I}+\underset{\sim\sim}{\nabla v},$$

and E_1, E_2 are positive coefficients which depend on the material. The displacement $\underset{\sim}{v}$ must also satisfy the *incompressibility condition*, which is expressed by

$$(4.5) \qquad \det \underset{\approx}{F}(\underset{\sim}{v}) = 1 \ \textit{a.e. on } \Omega.$$

Remark 4.1: We have assumed in (4.1) that $\underset{\sim}{S}_o$ is independent of $\underset{\sim}{v}$; this corresponds to a classical simplifying assumption known as the *dead load assumption*; this assumption facilitates the presentation of the problem without changing its fundamental nature, inasmuch as the essential difficulty lies in the incompressibility condition (4.5). We refer to GLOWINSKI-LE TALLEC [1] and LE TALLEC [1], [2] for the generalisation of the algorithms in Section 4.3 to the case where the dead load assumption is no longer satisfied; a number of numerical tests showing the efficiency of these generalised algorithms may also be found in the above references.

4.2.2 *Mathematical formulations*

In this section we shall describe various formulations for the

elastostatic problem; demonstrating their equivalence in the general case is still an open mathematical problem (we refer to LE TALLEC [1], [2], as well as to LE TALLEC-ODEN [1] for a discussion on these questions regarding equivalent formulations; see also the discussion in Section 4.2.2.1).

4.2.2.1 *Formulation by minimisation of the energy functional*

It is reasonable to assume that the displacements $\underset{\sim}{u}$ corresponding to the *stable equilibrium states* satisfy the following condition:

(4.6) $\underset{\sim}{u}$ *locally minimises on* K *the functional* $\underset{\sim}{v} \to \pi(\underset{\sim}{v})$,

where, for an incompressible Mooney-Rivlin material, we have

$$(4.7)\qquad \begin{cases} K = \{\underset{\sim}{v} \in (H^1(\Omega))^N,\ \underset{\sim}{v} = \underset{\sim}{0} \text{ on } \partial\Omega_1,\ \det \underset{\approx}{F}(\underset{\sim}{v}) = 1 \\ a.e.,\ \underset{\approx}{F}^{-1}(\underset{\sim}{v}) \in (L^2(\Omega))^{N\times N}\}\ , \end{cases}$$

and $\underset{\sim}{v} \to \pi(\underset{\sim}{v})$ defined by (4.1), (4.2), (4.3). The existence of solutions for (4.6), (4.7) is proved in BALL [1].

4.2.2.2 *Formulation by equilibrium equations*

The equilibrium positions (stable or unstable) correspond to the solutions of the system of nonlinear partial differential equations

$$(4.8)\qquad \begin{cases} \underset{\sim}{u} \in K, \\ (D\pi(\underset{\sim}{u}),\underset{\sim}{v}) + \int_\Omega p[\underset{\sim}{u},\underset{\sim}{v}]\,dx = 0 \quad \forall \underset{\sim}{v} \in X, \end{cases}$$

where $D\pi$ is the differential of π (on $H^1(\Omega))^N$) and where

$$(4.9)\qquad [\underset{\sim}{u},\underset{\sim}{v}] = \frac{\partial}{\partial u_{i,j}} (\det \underset{\approx}{F}(\underset{\sim}{u}))v_{i,j}\ ,$$

$$(4.10)\qquad X = \{\underset{\sim}{v} \in (H^1(\Omega))^N,\ \underset{\sim}{v} = \underset{\sim}{0} \text{ on } \partial\Omega_1\}$$

(in (4.9) we have used the classical notation of Mechanics with regard to summation and differentiation). The above function p is clearly a *Lagrange multiplier* associated with the incompressibility condition

(4.5) and is seen to be a *pressure*.

4.2.2.3 *Formulation by augmented Lagrangian*

We proceed as in Sections 2 and 3 (and as in Chapter III) by 'relaxing' the linear relation (4.4) simultaneously by a Lagrange multiplier and penalisation, giving the augmented Lagrangian (with $r > 0$):

$$\mathcal{L}_r(\underset{\sim}{v},\underset{\approx}{G},\underset{\approx}{\mu}) = \pi(\underset{\sim}{v}) + \frac{r}{2}\left\|\underset{\sim\sim}{\nabla v}+\underset{\approx}{I}-\underset{\approx}{G}\right\|_{L^2}^2 - \int_\Omega \underset{\approx}{\mu}\cdot(\underset{\sim\sim}{\nabla v}+\underset{\approx}{I}-\underset{\approx}{G})\,dx. \tag{4.11}$$

This leads to the following formulation of the elastostatic problem

$$\left\{\begin{array}{l} \textit{Find } \{\underset{\sim}{u},\underset{\approx}{F},\underset{\approx}{\lambda}\} \in W = X\times Y\times (L^2(\Omega))^{N\times N}, \textit{ the stationary point} \\ \textit{on } W \textit{ of the augmented Lagrangian } \mathcal{L}_r \end{array}\right. \tag{4.12}$$

where, in (4.12),

$$Y = \{\underset{\approx}{G}\,|\,\underset{\approx}{G}\in (L^2(\Omega))^{N\times N},\ \underset{\approx}{G}^{-1}\in (L^2(\Omega))^{N\times N},\ \det \underset{\approx}{G} = 1 \textit{ a.e.}\}\ .$$

4.2.2.4 *On some relations between formulations* (4.6), (4.8) *and* (4.12)

The following results are proved in LE TALLEC [1], [2]:

(i) There is equivalence between (4.8) and (4.12),

(ii) Any "regular" solution of (4.6) is a solution of (4.8) and (4.12).

(iii) If the functional π is *convex* (which is the case for a Mooney-Rivlin material if $N = 2$) then any solution $\{\underset{\sim}{u},\underset{\approx}{F},\underset{\approx}{\lambda}\}$ of (4.12) is such that $\underset{\sim}{u}$ (locally) minimises $\underset{\sim}{v} \to \mathcal{L}_r(\underset{\sim}{v},\underset{\approx}{F},\underset{\approx}{\lambda})$ on X ; likewise for r sufficiently large, any solution of (4.12) is such that $\underset{\approx}{F}$ minimises $\underset{\approx}{G} \to \mathcal{L}_r(\underset{\sim}{u},\underset{\approx}{G},\underset{\approx}{\lambda})$ (locally on Y).

Remark 4.2: If $\{\underset{\sim}{u},\underset{\approx}{F},\underset{\approx}{\lambda}\}$ is a solution of (4.12) the condition $\partial_{\underset{\sim}{u}}\mathcal{L}_r(\underset{\sim}{u},\underset{\approx}{F},\underset{\approx}{\lambda}) = 0$ implies that

$$(4.13)\qquad -\frac{\partial}{\partial x_j}\left(\frac{\partial}{\partial u_{i,j}}(\rho_o\sigma) - \lambda_{ij}\right) = \rho_o f_i .$$

In view of (4.13), $\underset{\approx}{\lambda}$ can be seen to be the part of the *first Piola-Kirchhoff tensor* corresponding to the *incompressibility*. It should also be noted that any algorithm solving (4.12) yields the stress field directly.

4.3 Solution of problem (4.12)

4.3.1 *A first algorithm for solving (4.12)*

Once again, we use algorithm ALG1 of Chapter III, Section 3.1, and of Section 2.4 of the present chapter; in the notation of Section 4.2 we obtain the following:

$$(4.14)\qquad \lambda^o \text{ given in } (L^2(\Omega))^{N\times N},$$

then for $n \geq 0$, $\underset{\approx}{\lambda}^n$ *being known, determine* $\underset{\sim}{u}^n, \underset{\approx}{F}^n$ *and* $\underset{\approx}{\lambda}^{n+1}$ *by*

$$(4.15)\qquad \begin{cases} \mathcal{L}_r(\underset{\sim}{u}^n,\underset{\approx}{F}^n,\underset{\approx}{\lambda}^n) \leq \mathcal{L}_r(\underset{\sim}{v},\underset{\approx}{G},\underset{\approx}{\lambda}^n) \quad \forall \{\underset{\sim}{v},\underset{\approx}{G}\} \in X\times Y , \\ \{\underset{\sim}{u}^n,\underset{\approx}{F}^n\} \in X\times Y , \end{cases}$$

$$(4.16)\qquad \underset{\approx}{\lambda}^{n+1} = \underset{\approx}{\lambda}^n - \rho(\underset{\sim\sim}{\nabla u}^n + \underset{\approx}{I} - \underset{\approx}{F}^n),\ \rho > 0.$$

Remark 4.3 Problem (4.15) is equivalent to the nonlinear system

$$(4.17)\qquad \mathcal{L}_r(\underset{\sim}{u}^n,\underset{\approx}{F}^n,\underset{\approx}{\lambda}^n) \leq \mathcal{L}_r(\underset{\sim}{u}^n,\underset{\approx}{G},\underset{\approx}{\lambda}^n) \quad \forall \underset{\approx}{G} \in Y,\ \underset{\approx}{F}^n \in Y,$$

$$(4.18)\qquad \partial_{\underset{\sim}{v}}\mathcal{L}_r(\underset{\sim}{u}^n,\underset{\approx}{F}^n,\underset{\approx}{\lambda}^n)\cdot\underset{\sim}{v} = 0 \quad \forall \underset{\sim}{v} \in X,\ \underset{\sim}{u}^n \in X,$$

which, when solved by block relaxation, leads to the algorithm described in Section 4.3.2 below.

4.3.2 *A second algorithm for solving (4.12)*

This time, we employ algorithm ALG2 of Chapter III, Section 3.2,

and of Section 2.5 of the present chapter, namely:

(4.19) $\underset{\sim}{u}^{-1}$ *given in* X, $\underset{\approx}{\lambda}^{o}$ *given in* $(L^2(\Omega))^{N\times N}$,

then for $n \geq 0$, $\underset{\sim}{u}^{n-1}$ *and* $\underset{\approx}{\lambda}^{n}$ *being known, determine* $\underset{\approx}{F}^{n}$, $\underset{\sim}{u}^{n}$ *and* $\underset{\approx}{\lambda}^{n+1}$ *by*

$$(4.20)\qquad \mathcal{L}_r(\underset{\sim}{u}^{n-1},\underset{\approx}{F}^{n},\underset{\approx}{\lambda}^{n}) \leq \mathcal{L}_r(\underset{\sim}{u}^{n-1},\underset{\approx}{G},\underset{\approx}{\lambda}^{n}) \quad \forall \underset{\approx}{G} \in Y,\ \underset{\approx}{F}^{n} \in Y,$$

$$(4.21)\qquad \partial_{\underset{\sim}{v}} \mathcal{L}_r(\underset{\sim}{u}^{n},\underset{\approx}{F}^{n},\underset{\approx}{\lambda}^{n})\cdot\underset{\sim}{v} = 0 \quad \forall \underset{\sim}{v} \in X,\ \underset{\sim}{u}^{n} \in X,$$

$$(4.22)\qquad \underset{\approx}{\lambda}^{n+1} = \underset{\approx}{\lambda}^{n} - \rho(\underset{\sim\sim}{\nabla u}^{n}+\underset{\approx}{I}-\underset{\approx}{F}^{n}) \ , \ \rho > 0.$$

Problem (4.21), which is equivalent to

$$(4.23)\qquad \begin{cases} \textit{Find } \underset{\sim}{u}^{n} \in X \textit{ such that} \\ \mathcal{L}_r(\underset{\sim}{u}^{n},\underset{\approx}{F}^{n},\underset{\approx}{\lambda}^{n}) \leq \mathcal{L}_r(\underset{\sim}{v},\underset{\approx}{F}^{n},\underset{\approx}{\lambda}^{n}) \quad \forall \underset{\sim}{v} \in X, \end{cases}$$

is in fact an *unconstrained* minimisation problem, the solution of which presents little difficulty, especially if r is sufficiently large; if N = 2, the functional in (4.23) is *quadratic*, and solving (4.21), (4.23) reduces to solving a *linear problem* relative to an operator with partial derivatives of second order (similar to the *Linear Elasticity* operator) which is *independent of* n, and whose finite-dimensional variants are linear systems associated with symmetric, positive-definite matrices which are independent of n (we then use a pre-factorisation of these matrices).

Problem (4.20) is not so straightforward (in appearance at least); if N = 2, (4.20) reduces (omitting the index n) to:

$$(4.24)\qquad \begin{cases} \textit{Find } \underset{\approx}{F} \in (L^2(\Omega))^4 \textit{ such that } F_{11}F_{22}-F_{12}F_{21} = 1 \textit{ a.e.} \\ \textit{and which minimises the functional } {}^{(1)} \\ \underset{\approx}{G} \to \int_\Omega [rG_{ij}^2-2(r(u_{i,j}+\delta_{ij})-\lambda_{ij})G_{ij}]\,dx \\ \textit{over the set of the } \underset{\approx}{G} \in (L^2(\Omega))^4 \textit{ such that } G_{11}G_{22}-G_{12}G_{21}=1 \textit{ a.e.} \end{cases}$$

[1] The δ_{ij} in (4.24) is the Kronecker delta.

In so far as there are no derivatives of $\underset{\approx}{G}$ and $\underset{\approx}{F}$ in (4.24), we can solve this latter problem point by point; it is thus necessary to solve an infinity (in theory at least) of problems in $\mathbb{R}^4$ of the type:

$$(4.25)\qquad \left\{\begin{array}{l} \textit{Find } \{F_{ij}\}\in\mathbb{R}^4 \textit{ such that } F_{11}F_{22}-F_{12}F_{21}=1 \textit{ and which} \\ \textit{minimises the functional } G_{ij}\to rG_{ij}^2-2a_{ij}G_{ij} \\ \textit{over } \{\{G_{ij}\}\in\mathbb{R}^4,\ G_{11}G_{22}-G_{12}G_{21}=1\}. \end{array}\right.$$

The above constraint is diagonalised with the aid of the new variables

$$(4.26)\qquad \left\{\begin{array}{l} b_1=(F_{11}+F_{22})/\sqrt{2}\ ,\ b_2=(F_{11}-F_{22})/\sqrt{2}\ , \\ b_3=(F_{12}+F_{21})/\sqrt{2}\ ,\ b_4=(F_{12}-F_{21})/\sqrt{2}. \end{array}\right.$$

Using $\underset{\sim}{b}=\{b_i\}_{i=1}^4$ defined by (4.26), problem (4.25) reduces to

$$(4.27)\qquad \left\{\begin{array}{l} \textit{Find } \underset{\sim}{b}\in\mathbb{R}^4 \textit{ such that } \varepsilon_i b_i^2=2,\ \varepsilon_1=\varepsilon_4=1, \\ \varepsilon_2=\varepsilon_3=-1 \textit{ and which minimises } \underset{\sim}{c}\to rc_i^2-2z_ic_i \\ \textit{over } \{\underset{\sim}{c}\,|\,\underset{\sim}{c}=\{c_i\}_{i=1}^4\ ,\ \varepsilon_i c_i^2=2\}\ . \end{array}\right.$$

The solutions of (4.27) are given by

$$(4.28)\qquad \{b_i\}\in\mathbb{R}^4\ ,\ b_i=z_i/(r+\varepsilon_i p),\quad \forall i=1,2,3,4,$$

where the scalar p (the Lagrange multiplier associated with $\varepsilon_i b_i^2=2$) satisfies

$$(4.29)\qquad (z_1^2+z_4^2)/(r+p)^2=2+(z_2^2+z_3^2)/(r-p)^2.$$

Suppose that $z_1^2+z_4^2\neq 0$; it then can easily be shown that (4.29) admits *just one solution* in $]-r,+r[$; furthermore, using the *Implicit Function Theorem* (see LE TALLEC [1], [2] GLOWINSKI-LE TALLEC [1] for further details) it can be shown that this solution of (4.29) belonging to $]-r,+r[$ is precisely that associated, via (4.28), with the *global*

minimum of the functional $\underset{\sim}{c} \to rc_i^2 - 2z_i c_i$ on $\varepsilon_i c_i^2 = 2$ and also that there are in fact no other local or global minima. Solving (4.29) on $]-r,+r[$ is a trivial problem; we then deduce $\underset{\sim}{b}$ from p, by using (4.28), and $\underset{\approx}{F}$ from $\underset{\sim}{b}$ by using (4.26). The multiplier p is interpreted mechanically as a pressure; in fact, it is shown in LE TALLEC [1], [2], GLOWINSKI-LE TALLEC [1] that this multiplier p is equal to the pressure p which appears in (4.8) (which therefore justifies our use of identical notation).

Remark 4.3: In the numerical tests we have performed, not once did we encounter the case $z_1^2 + z_4^2 = 0$; in fact we conjecture that for r sufficiently large this situation cannot arise, if N = 2, for problem (4.12). Furthermore, this condition of "r sufficiently large" is fundamental, as is shown in LE TALLEC [1], [2] and GLOWINSKI-LE TALLEC [1] (these references even go so far as to give a lower bound for r, this bound being related to certain norms of the pressure p).

4.4 Numerical tests

Suppose that N = 2; we reduce problem (4.6) (as well as problems (4.8), (4.12)) to a finite-dimensional problem by using a *finite-element approximation*. We have used *rectangular finite elements* $K \in \mathcal{Q}_h$, where $\mathcal{Q}_h$ is a *quadrangulation* of Ω. We then approximate the displacement $\underset{\sim}{v}$ by $\underset{\sim}{v}_h \in C^0(\bar{\Omega}) \times C^0(\bar{\Omega})$, such that

$$v_h|_K \in Q_1 \times Q_1 \quad \forall K \in \mathcal{Q}_h, \tag{4.30}$$

where

$$Q_1 = \{q | q(x_1,x_2) = a_{00}+a_{10}x_1+a_{01}x_2+a_{11}x_1x_2\} ; \tag{4.31}$$

the incompressibility condition (4.5) is imposed at the centre of each elementary rectangle $K \in \mathcal{Q}_h$ (which is equivalent to imposing it as an average over each rectangle).

The convergence of the approximate solutions when $h \to 0$ is a very difficult question; this topic is tackled in LE TALLEC [1], [2].

In the numerical tests which follow, Ω is a (two-dimensional)

bar containing a crack; this crack is assumed not to propagate any further. Figure 4.1 shows the right-hand portion of the bar, the crack and the quadrangulation Q_h (actually the right-hand half of Q_h). We suppose that in (4.1), (4.2) we have $\rho_o = 1$, $E_1 = 1$, $\partial\Omega_2 = \partial\Omega$ and that $\underset{\sim}{S}_o$ corresponds to horizontal forces applied to the ends of the bar and tending to elongate it, the density of these forces being 2 (in modulus). The bar thus stretches under the action of these forces, and Figure 4.2 shows the equilibrium position obtained; this was calculated by means of the discretised variant of algorithm (4.14) - (4.16), initialised with the configuration of Figure 4.1.

Using $\rho = r = 10$, convergence of (4.14) - (4.16) was attained in 20 iterations, corresponding to a computation time of 5 seconds on a CDC 6400. It is interesting to observe the behaviour of the crack.

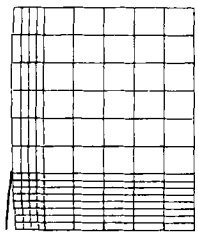


Figure 4.1

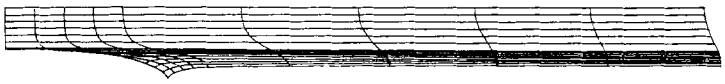


Figure 4.2

A number of numerical tests relating to other two-dimensional problems and to certain axisymmetric and three-dimensional problems may be found in LE TALLEC [1], [2], GLOWINSKI-LE TALLEC [1], [2].

5. SOME REMARKS ON THE APPLICATION OF THE ALGORITHMS OF SECTION 2 TO THE SOLUTION OF EIGENVALUE AND EIGENVECTOR PROBLEMS

The problems in Finite Nonlinear Elasticity considered in Sections 3 and 4 of this chapter are somewhat reminiscent of *eigenvalue and*

eigenvector problems: they in fact involve the minimisation of functionals (sometimes quadratic) over sets defined by *nonlinear equality constraints*. It is therefore natural to consider using the algorithms of Chapter III and of Section 2 of the present chapter for solving certain eigenvalue/eigenvector problems. In the following, we shall confine our attention to the determination of the *smallest eigenvalue* of a *symmetric positive-definite matrix* (and of an associated eigenvector); it is in fact possible to generalise the discussion below to the solution of certain nonlinear eigenvalue problems.

Let $\underset{\sim}{A}$ be an $N \times N$ *symmetric positive-definite* matrix; let Λ_m (>0) be *its smallest eigenvalue* and let $\underset{\sim}{x}_m$ ($\neq \underset{\sim}{0}$) be an associated eigenvector. We thus have

$$\underset{\sim}{A}\underset{\sim}{x}_m = \Lambda_m \underset{\sim}{x}_m , \tag{5.1}$$

and it is a classical result that $\underset{\sim}{x}_m$ is a (non-unique) solution of the minimisation problem

$$\begin{cases} J(\underset{\sim}{x}_m) \leq J(\underset{\sim}{y}) \quad \forall \underset{\sim}{y} \in S, \\ \\ \underset{\sim}{x}_m \in S \end{cases} \tag{5.2}$$

where (with $\|\underset{\sim}{y}\| = \left(\sum_{i=1}^{N} y_i^2\right)^{1/2}$ if $\underset{\sim}{y} = \{y_i\}_{i=1}^N$, and $(\underset{\sim}{x},\underset{\sim}{y}) = \sum_{i=1}^{N} x_i y_i$ if $\underset{\sim}{x},\underset{\sim}{y} \in \mathbb{R}^N$)

$$S = \{\underset{\sim}{y} | \underset{\sim}{y} \in \mathbb{R}^N, \|\underset{\sim}{y}\| = 1\} \tag{5.3}$$

and

$$J(\underset{\sim}{y}) = \frac{1}{2} (\underset{\sim}{A}\underset{\sim}{y}, \underset{\sim}{y}). \tag{5.4}$$

It is also known classically that if we associate with (5.2) - (5.4), the Lagrangian $L : \mathbb{R}^N \times \mathbb{R} \to \mathbb{R}$ defined by

$$L(\underset{\sim}{y},\mu) = J(\underset{\sim}{y}) - \frac{\mu}{2} (\|\underset{\sim}{y}\|^2 - 1), \tag{5.5}$$

then Λ_m is the Lagrange multiplier associated with the minimisation problem (5.2) and with the Lagrangian (5.5).

In order to apply the decomposition-coordination methods of Chapter III and of Section 2 of the present chapter, we first note that problem (5.2) is equivalent to

$$\text{(5.6)} \qquad \operatorname*{Min}_{\{\underset{\sim}{y},\underset{\sim}{q}\}\in W} \{\tfrac{1}{2}\,(\underset{\sim}{A}\underset{\sim}{y},\underset{\sim}{y})\}$$

where

$$\text{(5.7)} \qquad W = \{\{\underset{\sim}{y},\underset{\sim}{q}\}\in \mathbb{R}^N\times\mathbb{R}^N,\ \underset{\sim}{y}-\underset{\sim}{q} = \underset{\sim}{0},\ \underset{\sim}{q}\in S\}.$$

We then associate with the linear constraint $\underset{\sim}{y}-\underset{\sim}{q} = \underset{\sim}{0}$, the augmented Lagrangian $\mathcal{L}_r : \mathbb{R}^{3N} \to \mathbb{R}$ defined, with $r > 0$, by

$$\text{(5.8)} \qquad \mathcal{L}_r(\underset{\sim}{y},\underset{\sim}{q},\underset{\sim}{\mu}) = \tfrac{1}{2}\,(\underset{\sim}{A}\underset{\sim}{y},\underset{\sim}{y}) + \tfrac{r}{2}\|\underset{\sim}{y}-\underset{\sim}{q}\|^2 + (\underset{\sim}{\mu},\underset{\sim}{y}-\underset{\sim}{q}).$$

In order to solve (5.6) (and therefore (5.1), (5.2)) we are therefore led to determine (local) saddle points of $\mathcal{L}_r$ on $(\mathbb{R}^N \times S) \times \mathbb{R}^N$; the application of algorithm ALG1 of Section 2.4 (and of Chapter III, Section 3.1) leads to the algorithm

$$\text{(5.9)} \qquad \underset{\sim}{\lambda}^0\in\mathbb{R}^N, \quad \textit{given},$$

then, for $n \geq 0$, $\underset{\sim}{\lambda}^n$ *being known, determine* $\{\underset{\sim}{x}^n,\underset{\sim}{p}^n\}$ *then* $\underset{\sim}{\lambda}^{n+1}$, *by*

$$\text{(5.10)} \qquad \begin{cases} \{\underset{\sim}{x}^n,\underset{\sim}{p}^n\}\in\mathbb{R}^N\times S, \\ \mathcal{L}_r(\underset{\sim}{x}^n,\underset{\sim}{p}^n,\underset{\sim}{\lambda}^n) \leq \mathcal{L}_r(\underset{\sim}{y},\underset{\sim}{q},\underset{\sim}{\lambda}^n) \quad \forall\ \{\underset{\sim}{y},\underset{\sim}{q}\}\in\mathbb{R}^N\times S \end{cases}$$

and

$$\text{(5.11)} \qquad \underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^n + \rho(\underset{\sim}{x}^n-\underset{\sim}{p}^n),\ \rho>0. \ \blacksquare$$

Once again, we can solve (5.10) by a block-relaxation method and, as before, by restricting this to *a single* inner relaxation iteration, we deduce from (5.9) - (5.11) the following variant (of type ALG2 (cf. Section 2.5 and Chapter III, Section 3.2)):

$$\text{(5.12)} \qquad \underset{\sim}{x}^{-1} \textit{ and } \underset{\sim}{\lambda}^0 \textit{ given},$$

then, for $n \geq 0$, $\underset{\sim}{x}^{n-1}$ *and* $\underset{\sim}{\lambda}^{n}$ *being known, determine* $\underset{\sim}{p}^{n}$, $\underset{\sim}{x}^{n}$ *and* $\underset{\sim}{\lambda}^{n+1}$ *successively by*

$$\text{(5.13)} \qquad \begin{cases} \mathcal{L}_r(\underset{\sim}{x}^{n-1},\underset{\sim}{p}^{n},\underset{\sim}{\lambda}^{n}) \leq \mathcal{L}_r(\underset{\sim}{x}^{n-1},\underset{\sim}{q},\underset{\sim}{\lambda}^{n}) \quad \forall \underset{\sim}{q} \in S, \\ \underset{\sim}{p}^{n} \in S, \end{cases}$$

$$\text{(5.14)} \qquad \begin{cases} \mathcal{L}_r(\underset{\sim}{x}^{n},\underset{\sim}{p}^{n},\underset{\sim}{\lambda}^{n}) \leq \mathcal{L}_r(\underset{\sim}{y},\underset{\sim}{p}^{n},\underset{\sim}{\lambda}^{n}) \quad \forall \underset{\sim}{y} \in \mathbb{R}^N, \\ \underset{\sim}{x}^{n} \in \mathbb{R}^N, \end{cases}$$

$$\text{(5.15)} \qquad \underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^{n} + \rho(\underset{\sim}{x}^{n}-\underset{\sim}{p}^{n}), \ \rho > 0. \ \blacksquare$$

It is clear that the minimisation problem (5.14) is equivalent to the solution of the *linear system*

$$\text{(5.16)} \qquad (r\underset{\sim}{I}+\underset{\sim}{A})\underset{\sim}{x}^{n} = r\underset{\sim}{p}^{n}-\underset{\sim}{\lambda}^{n} .$$

Since the matrix $r\underset{\sim}{I} + \underset{\sim}{A}$ is *symmetric and positive-definite*, we can perform once and for all a Cholesky factorisation of this matrix; hence we solve, at each iteration, two linear systems of triangular matrices.

The solution of problem (5.13) does not present any further difficulties; in fact, taking account of the condition $\underset{\sim}{q} \in S$, problem (5.13) is equivalent to

$$\text{(5.17)} \qquad \underset{\underset{\sim}{q}\in S}{\text{Max}} \ (\underset{\sim}{q},\underset{\sim}{\lambda}^{n}+r\underset{\sim}{x}^{n-1}),$$

so that

$$\text{(5.18)} \qquad \underset{\sim}{p}^{n} = \frac{\underset{\sim}{\lambda}^{n} + r\underset{\sim}{x}^{n-1}}{\|\underset{\sim}{\lambda}^{n}+r\underset{\sim}{x}^{n-1}\|} .$$

Taking account of (5.16), (5.18), we can write algorithm (5.12) - (5.15) in the following, more practical, form:

$$\text{(5.19)} \qquad \underset{\sim}{x}^{-1} \text{ and } \underset{\sim}{\lambda}^{0} \text{ given,}$$

then, for $n \geq 0$, $\underset{\sim}{x}^{n-1}$ *and* $\underset{\sim}{\lambda}^{n}$ *being known*

$$\underset{\sim}{p}^{n} = \frac{\underset{\sim}{\lambda}^{n} + r\underset{\sim}{x}^{n-1}}{\|\underset{\sim}{\lambda}^{n}+r\underset{\sim}{x}^{n-1}\|} , \tag{5.20}$$

$$(r\underset{\sim}{I}+\underset{\sim}{A})\underset{\sim}{x}^{n} = r\underset{\sim}{p}^{n}-\underset{\sim}{\lambda}^{n} , \tag{5.21}$$

$$\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^{n} + \rho(\underset{\sim}{x}^{n}-\underset{\sim}{p}^{n}),\ \rho > 0. \blacksquare \tag{5.22}$$

Similarly, by updating $\underset{\sim}{\lambda}^{n}$ in (5.19) - (5.22) between the stages (5.20) and (5.21), we obtain the following algorithm (of type ALG3) in which $\underset{\sim}{y}$ and $\underset{\sim}{q}$ play symmetric roles:

$\underset{\sim}{x}^{-1}$ *and* $\underset{\sim}{\lambda}^{0}$ *given,* (5.23)

then, for $n \geq 0$, $\underset{\sim}{x}^{n-1}$ *and* $\underset{\sim}{\lambda}^{n}$ *being known,*

$$\underset{\sim}{p}^{n} = \frac{\underset{\sim}{\lambda}^{n} + r\underset{\sim}{x}^{n-1}}{\|\underset{\sim}{\lambda}^{n}+r\underset{\sim}{x}^{n-1}\|} , \tag{5.24}$$

$$\underset{\sim}{\lambda}^{n+1/2} = \underset{\sim}{\lambda}^{n} + \rho(\underset{\sim}{x}^{n-1}-\underset{\sim}{p}^{n}), \tag{5.25}$$

$$(r\underset{\sim}{I}+\underset{\sim}{A})\underset{\sim}{x}^{n} = r\underset{\sim}{p}^{n} - \underset{\sim}{\lambda}^{n+1/2}, \tag{5.26}$$

$$\underset{\sim}{\lambda}^{n+1} = \underset{\sim}{\lambda}^{n+1/2} + \rho(\underset{\sim}{x}^{n}-\underset{\sim}{p}^{n}). \blacksquare \tag{5.27}$$

The convergence of the above algorithms remains to be proved; nonetheless, if we suppose that

$$\underset{\sim}{x}_{m} = \lim_{n\to+\infty} \underset{\sim}{x}^{n} = \lim_{n\to+\infty} \underset{\sim}{p}^{n},$$

then we have

$$\Lambda_{m} = (\underset{\sim}{A}\underset{\sim}{x}_{m}, \underset{\sim}{x}_{m}).$$

We note (and it is sufficient to set $\rho = r$ to show this) that algorithms (5.19) - (5.22) and (5.23) - (5.27) are variants of the *power*

method, which is a standard method for the calculation of the eigenvalues and eigenvectors of matrices (see WILKINSON [1], WILKINSON-REINSCH [1], STEWART [1], PARLETT [1]). Numerical tests carried out by *M.O. Bristeau* at INRIA, have demonstrated the good convergence properties of the above algorithms *as long as* r *is taken sufficiently large* (if we put $\rho = r$, which once again seems to be the best choice).

We leave it as an exercise for the reader to derive variants of the above algorithms which enable the other eigenvalues and eigenvectors of $\underset{\sim}{A}$ to be calculated.

CHAPTER IX

APPLICATIONS OF THE METHOD OF MULTIPLIERS TO VARIATIONAL INEQUALITIES

D. Gabay

1. INTRODUCTION

This chapter extends and complements some of the remarks made in the previous chapters, in particular in Chapter III. We generalise the augmented Lagrangian method to the case of variational inequalities and we give to it the more appropriate name of the *method of multipliers* since these problems do not generally involve a Lagrangian. We shall also demonstrate the equivalence between algorithm ALG1 and a method of solution well-known in Nonlinear Analysis, namely the proximal-point algorithm. Finally, we reconsider in detail the ideas introduced earlier, in Chapter IV, on the subject of *alternating direction methods*, and we describe the relationship between these methods and ALG2 and ALG3 (see also Chapter VIII, Section 2). To facilitate a proper presentation of the problems, we first recall a number of definitions and results.

1.1 Monotone operators

Let X be a real Hilbert space equipped with the inner product $(\cdot,\cdot)$. We designate as a *monotone operator* a multi-valued mapping $T: X \to 2^X$ such that we have

$$(1.1) \qquad (z'-z,x'-x)_X \geq 0 \quad \forall x,x' \in X, \ \forall z \in T(x), \ \forall z' \in T(x').$$

We say that T is a *maximal monotone* operator if, in addition, the graph

$$(1.2) \qquad \text{graph}(T) = \{\{x,z\} \in X \times X \mid z \in T(x)\}$$

is not strictly included within the graph of any other monotone

operator on X. These operators occur in Convex Analysis and in the study of certain partial differential equations (see, for example, BREZIS [1]). In fact a very important special case arises from the field of Convex Analysis. Let $\phi : X \to]-\infty, +\infty]$ be a function which is proper, convex and lower semi-continuous on X and consider its *subgradient* $\partial\phi$, defined at $x \in X$ by

$$(1.3) \qquad \partial\phi(x) = \{z \in X \mid (z,y-x)_X \leq \phi(y)-\phi(x) \quad \forall y \in X\} .$$

The operator $x \to \partial\phi(x)$ is maximal monotone, and if $x \in X$ is a solution of the *multivalued equation* (*)

$$(1.4) \qquad 0 \in \partial\phi(x)$$

then we also have

$$(1.5) \qquad \phi(x) \leq \phi(y) \quad \forall y \in X .$$

The multivalued equation (1.4) is thus equivalent to the optimisation problem (1.5), (or convex programming problem), which is implicitly a constrained problem since the set of points where $\phi(y) = +\infty$ is obviously excluded from the set of admissible solutions.

However, it is not always possible (in particular in examples arising from the theory of partial differential equations) to associate an optimisation problem with a multivalued equation: we then consider directly a formulation using a variational inequality. Given a nonempty, closed, convex subset K of X and a maximal monotone operator A on X, not necessarily defined by a subgradient, we seek $x \in K$, satisfying the following variational inequality (see LIONS-STAMPACCHIA [1]):

$$(1.6) \qquad \textit{There exists } z \in A(x) \textit{ such that } (z,y-x)_X \geq 0 \quad \forall y \in K.$$

Let us now consider the multivalued equation

$$(1.7) \qquad 0 \in T(x).$$

(*) *Translator's note:* Sometimes known under the name *multivoque equation.*

Any solution of the inequality (1.6) also satisfies (1.7) with T defined by

$$T(x) = \begin{cases} A(x)+N_K(x) \text{ if } x \in K, \\ \emptyset \text{ otherwise,} \end{cases} \tag{1.8}$$

where $N_K(x)$ denotes the *cone normal to* K at x[1]; ROCKAFELLAR [6] has shown that T is maximal monotone. We note that if A is a maximal monotone *single-valued* operator and if K is a closed convex cone in X with vertex O, then the variational inequality (1.6) is equivalent to the *complementarity problem:*

$$\begin{cases} \text{Find } x \in K \text{ such that} \\ -A(x) \in K^O \text{ and } (A(x),x)_X = 0, \end{cases} \tag{1.9}$$

where K^O denotes the *polar cone* [2] of K.

The multivalued equation (1.7) thus unifies variational problems of the type (1.5), variational inequalities of the type (1.6) and complementarity problems of the type (1.9). We can therefore transpose algorithms designed for solving problems of one type to the solution of problems of the other two types.

1.2 The method of multipliers

The development of Operational Research during the last twenty years has promoted an increased emphasis on the investigation of optimisation problems; more advanced numerical methods and experiments therefore exist in the field of convex programming than in the fields of variational inequalities and complementarity problems. The unifying framework presented above would thus indicate one possible methodology for reducing the gap between these domains: namely, given a known algorithm for the convex programming problem (1.5), find the corresponding general algorithm for solving the multivalued equation (1.7) and deduce from it the corresponding algorithm for the variational inequality (1.6), which can be specialised to the

---

(1) $N_K(x) = \{z \mid (z,y-x)_X \leq 0 \quad \forall y \in K\}$

(2) $K^O = \{z \mid (z,y)_X \leq 0, \forall y \in K\}$

complementarity problem (1.9). This approach has been followed implicitly by GLOWINSKI-LIONS-TREMOLIERES [1] , [2] for defining relaxation algorithms, gradient algorithms (with auxiliary operator), conjugate-gradient algorithms, duality methods and penalisation methods, for the solution of variational inequalities.

The method of multipliers has for a number of years generated a considerable amount of interest for the solution of constrained optimisation problems, both for its simplicity of implementation and for the advantages it offers over penalisation methods (see BERTSEKAS [1]). Proposed initially by HESTENES [1] and POWELL [1] for minimisation problems with equality constraints, it can be interpreted as a gradient method for solving a dual problem associated with an augmented Lagrangian, this being obtained by adding to the ordinary Lagrangian a penalisation term depending on a parameter $r > 0$ (which need not tend to infinity); hence the alternative name: *penalisation-duality* method. ROCKAFELLAR [1], [7] defined the method for convex programming problems and demonstrated its global and linear (superlinear if $r \to +\infty$) convergence.

In this chapter, we propose to extend the method of multipliers to variational inequalities and then to propose approximations of this method which will enable a decomposition of the calculations to be effected. This work may be recognised as a generalisation of Chapter III and of GABAY-MERCIER [1], in which this objective was achieved for particular inequalities corresponding to *convex variational problems* of the form

$$\inf_{v \in V} \{F(Bv)+G(v)\}, \tag{1.10}$$

where F and G are functions with values in $]-\infty,+\infty]$, and which are convex, proper, lower semi-continuous and defined respectively on the real Hilbert spaces H and V, equipped with the inner products $(\cdot,\cdot)_H$ and $(\cdot,\cdot)_V$; it can be shown (see Chapter III, GABAY-MERCIER [1], FORTIN [1]) that we can associate with (1.10) the *regularised dual problem* (or augmented dual problem):

$$\sup_{\mu} \psi_r(\mu), \tag{1.11}$$

where, for $r \geq 0$, the concave functional $\psi_r : H \to [-\infty,+\infty[$ is defined by

$$(1.12)\qquad \psi_r(\mu) = \inf_{v\in V}\{G(v)+\inf_{q\in H}[F(q)+(\mu,Bv-q)_H+\frac{r}{2}|Bv-q|_H^2]\}.$$

The functional ψ_r is always *differentiable* for $r > 0$. We can solve problem (1.10) and its dual (1.11) by seeking on $V\times H\times H$ a saddle point of the *augmented Lagrangian* defined (as in Chapter III) by

$$(1.13)\qquad \mathcal{L}_r(v,q,\mu) = F(q)+G(v)+(\mu,Bv-q)_H+\frac{r}{2}|Bv-q|_H^2 .$$

It is possible to obtain $\mathcal{L}_r$ directly from the variational problem (1.10) by introducing the variable q and the artificial constraint $Bv-q = 0$, this constraint then being penalised and dualised along the lines of the original procedure of HESTENES [1]. For seeking saddle points of $\mathcal{L}_r$, we have made extensive use in this book of Uzawa's algorithm applied to $\mathcal{L}_r$, (ALG1), and a particular variant of this, (ALG2). In order to clarify the rest of the description, we give below a brief restatement of algorithm ALG1, the properties of which were studied in Chapter III.

ALG1: *λ^0 being chosen arbitrarily, we seek for $n = 0, 1, \ldots$, with λ^n known, solutions $u^n \in V$, $p^n \in H$, of*

$$(1.14)\qquad \mathcal{L}_r(u^n,p^n,\lambda^n) \le \mathcal{L}_r(v,q,\lambda^n) \quad \forall v \in V,\ \forall q \in H,$$

then we calculate λ^{n+1} by

$$(1.15)\qquad \lambda^{n+1} = \lambda^n+\rho(Bu^n-p^n). \ \blacksquare$$

We then consider the *augmented dual functional* ψ_r defined by

$$(1.16)\qquad \psi_r(\mu) = \inf_{\{v,q\}\in V\times H} \mathcal{L}_r(v,q,\mu).$$

We have, from (1.14), $\psi_r(\lambda^n) = \mathcal{L}_r(u^n,p^n,\lambda^n)$; furthermore, it can be shown (see ROCKAFELLAR [2], FORTIN [1]) that we have

$$(1.17)\qquad -r\psi_r(\mu) = \inf_{\nu\in H}\{\frac{1}{2}|\mu-\nu|^2 - r\psi_o(\nu)\}$$

where ψ_o is in fact the functional defining the dual of problem (1.10) in the sense of Fenchel (see ROCKAFELLAR [5], EKELAND-TEMAM [1]); that is

$$-\psi_o(\mu) = F^*(\mu)+G^*(-B^t\mu), \tag{1.18}$$

where F* and G* denote the conjugate functions of F and G[3], defined on H and V respectively, and where B^t is the operator from H into V defined by

$$(Bv,q)_H = (v,B^tq)_V \quad \forall v \in V, \quad \forall q \in H \text{ [4]}. \tag{1.19}$$

In the terminology of MOREAU [1], relation (1.17) states that $-r\psi_r$ is the *proximal point mapping** relative to $-r\psi_o$. We deduce from this (see ROCKAFELLAR [2] that the method of multipliers (algorithm ALG1 with $\rho = r$) generates the same sequence of iterates $\{\lambda^n\}_{n\geq 0}$ as the *proximal point algorithm* for solving the multivalued equation (1.7) with $T = -\partial\psi_o$. This algorithm was introduced by MARTINET [1], [2] and generalised by ROCKAFELLAR [8] for an arbitrary maximal monotone operator T on H; it constructs a sequence $\lambda^n \in H$ in accordance with the recurrence relation

$$\lambda^{n+1} = J_T^r(\lambda^n) \quad n = 0,1,\ldots, \tag{1.21}$$

where $J_T^r = (I+rT)^{-1}$ is a contracting single-valued operator called the *resolvent* of T (see BREZIS [1]). This observation will guide our approach to generalising the method of multipliers to variational inequalities.

After first recalling, in Section 2 below, the convergence properties of the proximal-point algorithm, we then define in Section 3 a variational inequality which generalises the variational problem (1.10) and which includes the inequality (1.6) as a particular case, and we associate with this a dual variational inequality. In Section 4 we apply the proximal-point algorithm to the representation of the

---

[3] By definition $F^*(\mu) = \sup_{q\in H} \{(\mu,q)_H - F(q)\}$, $G^*(\sigma) = \sup_{v\in V} \{(\sigma,v)_V - G(v)\}$.

[4] See Section 3 for the relation between B^t and the usual adjoint B' of B.

* *Translator's note: application de proximité* in the original French.

dual variational inequality, in multivalued (*"multivoque"*) form, and this defines a method of multipliers for the solution of variational inequalities by penalisation-duality, generalising algorithm ALG1. Observing that the multivalued operator associated with the dual inequality has the form of the sum of two maximal-monotone operators, we investigate in Section 5 some approximations of the proximal-point algorithm which take advantage of this structure. In particular, we apply two algorithms recently proposed by P.L. LIONS-MERCIER [1] which generalise the alternating-direction methods to the solution of the multivalued equation (1.7) and we thereby obtain two variants of the method of multipliers for the solution of variational inequalities; these variants provide a decomposition scheme coordinated via the multipliers (see BENSOUSSAN-LIONS-TEMAM [1]). One of these variants generalises to variational inequalities the algorithm ALG2 from Chapter III and from GABAY-MERCIER [1], which can therefore be interpreted as an alternating-direction method as was noted in CHAN-GLOWINSKI [1], for a particular class of problems, and as was pointed out in Chapters IV and VIII. The other variant comprises algorithm ALG3 mentioned in Chapter VIII. In Section 6 we study another approximation of the proximal-point algorithm which employs a splitting of the operator T. We then recover the point projection gradient method for convex programming (GOLDSTEIN [1]), and its generalisation to the variational inequality (1.6); applied to the dual inequality of (1.6) this may be interpreted as a method of multipliers with projection.

2. THE PROXIMAL-POINT ALGORITHM

Let X be a real Hilbert space equipped with the inner product $(\cdot,\cdot)_X$ and the corresponding norm $|\cdot|_X$, and let T be a maximal-monotone operator on X. We wish to solve the multivalued equation

$$\text{(2.1)} \qquad \left\{ \begin{array}{l} \textit{Find} \quad x \in X \ \textit{such that} \\ \\ 0 \in T(x). \end{array} \right.$$

For all $x \in X$ and all $r > 0$ there exists (see MINTY [1]) a unique $y \in X$ such that

$$\text{(2.2)} \qquad x \in (I+rT)(y).$$

The operator $J_T^r = (I+rT)^{-1}$ is thus single-valued and defined on the whole of X; we call this the *resolvent* of T. This is a *contraction* from X into X, i.e.

$$|J_T^r(x')-J_T^r(x)|_X \leq |x'-x|_X \quad \forall x,x' \in X.$$

We shall show that J_T^r is moreover a *firm contraction*.

PROPOSITION 2.1: *For all* $r > 0$ *the resolvent* J_T^r *of the maximal monotone operator* T *is a firm contraction, i.e. it satisfies*

$$|J_T^r(x')-J_T^r(x)|_X^2 \leq (J_T^r(x')-J_T^r(x),x'-x)_X \quad \forall x,x' \in X. \tag{2.3}$$

Proof: We put $y = J_T^r(x)$, $y' = J_T^r(x')$. From (2.2) we have

$x = y+rz$ where $z \in T(y)$,

$x' = y'+rz'$ where $z' \in T(y')$.

We can thus write

$$|y'-y|_X^2 = (y'-y,x'-x)_X - r(y'-y,z'-z)_X$$

$$\leq (y'-y,x'-x)_X$$

in view of the maximal monotonicity of T; hence the result.

Remark 2.1: If $T = \partial\phi$, the subgradient of a convex, proper, lower semi-continuous function $\phi : X \to]-\infty,+\infty]$, then

$$J_T^r(x) = \operatorname{Arg}\min_y \{\tfrac{1}{2}|y-x|_X^2 + r\phi(y)\} ; \tag{2.4}$$

in the terminology of MOREAU [1] J_T^r is the *proximal point mapping* relative to the functional $r\phi$. ■

We note that $0 \in T(x)$ is equivalent to $J_T^r(x) = x$. The solution of the multi-valued equation (2.1) thus reduces to seeking the fixed points of the contraction mapping J_T^r.

DEFINITION 2.2: (see ROCKAFELLAR [8]): *Given a nondecreasing sequence* $\{r_n\}$ *of positive numbers and an arbitrary point* $x^0 \in X$, *the proximal-point algorithm generates a sequence* $\{x^n\}$ *of points of* X *according to the recurrence relation*

$$x^{n+1} = J_T^{r_n}(x^n), \quad n=0,1,\ldots . \tag{2.5}$$

Remark 2.2: The proximal-point algorithm can be interpreted as an implicit discretisation scheme for the multivalued evolution equation

$$0 \in \frac{dx}{dt} + T(x) \text{ with } x(0) = x^o, \tag{2.6}$$

the parameters r_n representing the time steps for the discretisation

$$0 \in \frac{x^{n+1}-x^n}{r_n} + T(x^{n+1}). \quad \blacksquare$$

The following convergence result states that the sequence $\{x^n\}$ converges to a solution of the steady-state equation $0 \in T(x)$.

THEOREM 2.1: *Suppose that there exists at least one solution to equation* (2.1). *The proximal-point algorithm generates a sequence* $\{x^n\}$ *which converges weakly to* $x \in X$ *such that* $0 \in T(x)$ *and* $|x^{n+1}-x^n|_X \to 0$.

Proof: Using relation (2.3) which expresses the fact that $J_T^{r_n}$ is a firm contraction, we obtain for all n

$$|x^{n+1}-x|_X^2 \le (x^{n+1}-x, x^n-x)_X$$

$$= \frac{1}{2}\left(|x^{n+1}-x|_X^2 + |x^n-x|_X^2 - |x^{n+1}-x^n|_X^2\right).$$

By adding these inequalities from $n = 0$ to an arbitrary integer N, we arrive at

$$|x^{N+1}-x|_X^2 + \sum_{n=0}^{N} |x^{n+1}-x^n|_X^2 \le |x^o-x|_X^2 ,$$

which shows that $|x^{n+1}-x^n|_X$ converges to 0 and that the sequence $\{x^n\}$ is bounded.

By hypothesis, the set of the fixed points of J_T^r is nonempty (since there exists at least one solution of (2.1)); *Opial's Lemma*,

see OPIAL [1], thus enables the weak convergence of the sequence $\{x^n\}$ to be established (see also MARTINET [1], [2] and ROCKAFELLAR [1]).

Remark 2.3: We can define the *proximal point algorithm with relaxation* by the recurrence

$$x^{n+1} = (1-\omega)x^n + \omega J_T^{r_n}(x^n), \quad n=0,1,\ldots . \tag{2.7}$$

Theorem 2.1 remains valid for any relaxation parameter $0 < \omega < 2$. It can in fact be shown that if J is a firm contraction, then $J_\omega = (1-\omega)I+\omega J$ is also a firm contraction for $0 < \omega < 1$ from which we deduce that $|x^{n+1} - x^n|_X \to 0$ and the weak convergence. For $1 < \omega < 2$ we can establish the inequality

$$|J_\omega(x')-J_\omega(x)|_X^2 \le (2-\omega)(J_\omega(x')-J_\omega(x),x'-x)_X+(\omega-1)|x'-x|_X^2 \quad \forall x,x' \in X, \tag{2.8}$$

which allows us to prove that $|x^{n+1} - x^n|_X$ again converges to 0. ∎

If T *is coercive* (with modulus $\alpha > 0$), i.e. if

$$(z'-z,x'-x)_X \ge \alpha|x'-x|_X^2 \quad \forall x,x' \in X, \ \forall z \in T(x), \ \forall z' \in T(x'), \tag{2.9}$$

then it can easily be shown that

$$|J_T^{r_n}(x') - J_T^{r_n}(x)|_X \le (1+\alpha r_n)^{-1} |x'-x|_X \quad \forall x,x' \in H, \tag{2.10}$$

which implies that $J_T^{r_n}$ has a unique fixed point x which is the unique solution satisfying $0 \in T(x)$.

THEOREM 2.2: *If* T *is coercive, the proximal-point algorithm* (2.3) *generates a sequence* $\{x^n\}$ *which converges strongly and linearly to the unique solution* x *of the multivalued equation* (2.1). *If* $r_n \to +\infty$, *the convergence is superlinear.*

Proof: The inequality (2.10) can be written, with $x' = x^n$ and with x a solution of (2.1), as

$$|x^{n+1}-x|_X \le (1+\alpha r_n)^{-1} |x^n-x|_X, \quad n=0,1,\ldots \tag{2.11}$$

which implies the strong convergence and which gives an estimate of the rate of convergence. If $r_n \to +\infty$, the convergence is superlinear since

$$\lim_{n \to +\infty} \frac{|x^{n+1}-x|_X}{|x^n-x|_X} = 0 .$$

Remark 2.4: The introduction of the relaxation mentioned in Remark 2.3 enables the convergence of the proximal-point algorithm to be accelerated if the parameter ω in the recurrence (2.7) is suitably chosen (assuming for simplicity that r is fixed). Suppose that T is coercive (condition (2.9)) and also that it is uniformly Lipschitz-continuous, i.e. (T then being single-valued)

$$|T(x')-T(x)|_X \le M|x'-x|_X \quad \forall x,x' \in X, \tag{2.12}$$

with, naturally, $M \ge \alpha$. We now introduce the notation $x^{n+1/2} = J_T^r x^n$; by definition, we have $x^n = x^{n+1/2}+rz^{n+1/2}$, where $z^{n+1/2} = T(x^{n+1/2})$, and from (2.10) we have

$$|x^{n+1/2}-x|_X \le (1+\alpha r)^{-1}|x^n-x|_X .$$

We can write (2.7) in the form

$$x^{n+1} = (1-\omega)x^n+\omega x^{n+1/2} = x^{n+1/2}+r(1-\omega)z^{n+1/2}$$

and we obtain (for $\omega > 1$) the estimate

$$|x^{n+1}-x|_X^2 \le [1+(\omega-1)^2r^2M^2 - 2\alpha r(\omega-1)]|x^{n+1/2}-x|_X^2 .$$

The coefficient on the right is a minimum for $\omega^* = 1+\alpha/rM^2$ ($\omega^* < 2$ for r sufficiently large); it then follows that

$$|x^{n+1}-x|_X \le \left(1-\frac{\alpha^2}{M^2}\right)^{1/2} (1+\alpha r)^{-1} |x^n-x|_X . \tag{2.13}$$

This estimate is not very sharp but, compared with (2.11), it demonstrates the acceleration of the rate of convergence produced by overrelaxation. ∎

Before concluding this section we should mention a relation which exists between the resolvents of the operator T and its inverse T^{-1} which will be used later on in this Chapter (see Section 5).

PROPOSITION 2.2: *Given* $r > 0$, *we put* $\varepsilon = 1/r$; *we then have the relation*

$$(2.14) \qquad J_T^r x = r(I - J_{T^{-1}}^{\varepsilon})(\varepsilon x) \quad \forall x \in X.$$

Proof: We put $y = J_T^r x$; by definition we have $x = y + rz$ with $z \in T(y)$, from which we deduce that $y \in T^{-1}(\varepsilon(x-y))$. We therefore obtain

$$\varepsilon x \in (I + \varepsilon T^{-1})(\varepsilon(x-y)),$$

which then gives (2.14).

3. VARIATIONAL INEQUALITIES IN DUALITY

In the following, V and H denote real Hilbert spaces equipped respectively with the inner products $(\cdot,\cdot)$ and $(\cdot,\cdot)_H$; V' and H' denote the corresponding dual spaces. We denote by $\langle \cdot,\cdot \rangle_{V' \times V}$ and $\langle \cdot,\cdot \rangle_{H' \times H}$ the bilinear forms of duality between V' and V and between H' and H, and by Λ_V and Λ_H the isomorphisms of V onto V' and of H onto H' defined respectively by

$$\langle \Lambda_V u, v \rangle_{V' \times V} = (u,v)_V \quad \forall u,v \in V \text{ with } \Lambda_V u \in V',$$

$$\langle \Lambda_H p, q \rangle_{H' \times H} = (p,q)_H \quad \forall p,q \in H \text{ with } \Lambda_H p \in H' .$$

In many cases it it possible to identify H with its dual and we then have $\Lambda_H = I$.

Let $A : V \to 2^V$ be a (multivalued) maximal monotone operator on V, with domain

$$(3.1) \qquad \mathrm{dom}(A) = \{v \in V \mid A(v) \neq \emptyset\} .$$

Let $B : V \to H$ be a continuous linear operator from V into H, let B' be its adjoint ($B' \in \mathcal{L}(H',V')$) defined by $\langle B'q', v \rangle_{V' \times V} = \langle q', Bv \rangle_{H' \times H}$ $\forall v \in V$, $\forall q' \in H'$ and let $F : H \to]-\infty,+\infty]$ be a lower semi-continuous, proper, convex function with effective domain

$$\text{(3.2)} \qquad \mathrm{dom}(F) = \{q \mid q \in H,\ F(q) < +\infty\},$$

the interior of which we shall assume to be nonempty. We also assume that the following *qualification condition* is satisfied: there exists $v_o \in \mathrm{int}(\mathrm{dom}(A))$ such that $Bv_o \in \mathrm{dom}(F)$.

We consider a *variational inequality* in the following general form:

$$\text{(3.3)} \qquad \left\{ \begin{array}{l} \textit{Find } u \in V \textit{ such that} \\ \\ \exists\, w \in A(u) \quad \textit{such that} \quad (w, v-u)_V + F(Bv) - F(Bu) \geq 0 \quad \forall v \in V. \end{array} \right.$$

We recall that if $A = \partial G$, ∂G being the subgradient of a function $G : V \to]-\infty, +\infty]$ which is convex, proper and lower semi-continuous, then the inequality (3.3) is equivalent to the convex variational problem (1.10), i.e.

$$\inf_{v \in V} \{F(Bv) + G(v)\},$$

introduced by ROCKAFELLAR [5] to generalise *Fenchel's* duality theory (this problem includes in particular the ordinary convex programming problems (1.5)). The variational inequality (1.6) also constitutes a particular case of (3.3) with $V = H$, with B the identity mapping and with $F = I_K$, the indicator function of the closed convex set K with nonempty interior defined by

$$\text{(3.4)} \qquad I_K(v) = \left\{ \begin{array}{ll} 0 & \textit{if } v \in K, \\ +\infty & \textit{otherwise.} \end{array} \right.$$

Transposing the analysis of MOSCO [1] into this formalism, we associate with (3.3) the *dual variational inequality:*

$$\text{(3.5)} \qquad \left\{ \begin{array}{l} \textit{Find } \lambda \in H \textit{ such that} \\ \\ \exists\, p \in A_B^t(\lambda) \quad \textit{such that} \quad (p, \mu-\lambda)_H + F^*(\mu) - F^*(\lambda) \geq 0 \quad \forall \mu \in H\ ; \end{array} \right.$$

in (3.5) $F^* : H \to]-\infty, +\infty]$ denotes the convex function conjugate to F (see ROCKAFELLAR [4] EKELAND-TEMAM [1]) defined on H by

$$(3.6) \qquad F^*(\mu) = \sup_{q \in H} \{(\mu,q)_H - F(q)\}$$

and $A_B^t : H \to 2^H$ is the multivalued operator such that

$$(3.7) \qquad A_B^t(\mu) = \{q \in H \mid \exists v \in V \text{ such that } q = -Bv,\ -B^t\mu \in A(v)\},$$

where B^t denotes the operator from H into V defined by $B^t = \Lambda_V^{-1} \circ B' \circ \Lambda_H$.

The title of 'dual inequality' is justified by the following result:

THEOREM 3.1: *A vector* $u \in V$ *is a solution of the variational inequality* (3.3) *if and only if there exists a solution* λ *of the inequality* (3.5) *such that* $-B^t\lambda \in A(u)$. *Furthermore,* u *and* λ *are respectively solutions of* (3.3) *and* (3.5) *if and only if* $-B^t\lambda \in A(u)$ *and* $-Bu \in A_B^t(\lambda)$ *and we have the identity*

$$(3.8) \qquad F(Bu)+F^*(\lambda) = (Bu,\lambda)_H.$$

Proof: First, we note that the inequality (3.3) can be written

$$\exists w \in A(u) \text{ such that } -w \in \partial(F \circ B)(u).$$

Since the domain of F has a nonempty interior, there exists a point of H where F is finite and continuous; we therefore have (see EKELAND-TEMAM [1] Chapter 1, Proposition 5.7)

$$\partial(F \circ B)(u) = B^t \partial F(Bu)$$

since we have chosen to define the subgradient ∂F as a multivalued operator from H into subsets of H (whereas the subdifferential of F is an operator from H into subsets of H'). The subgradient ∂F^* of the conjugate function F^* defined in (3.6) is in fact identical to the inverse, in the sense of multivalued operators, of ∂F (see EKELAND-TEMAM [1]); it follows from this that $\mu \in \partial F(q)$ is equivalent to $q \in \partial F^*(\mu)$. The inequality (3.3) is therefore equivalent to

$$\exists w \in A(u), \exists \lambda \in H \text{ such that } -B^t\lambda = w \text{ with } Bu \in \partial F^*(\lambda) ;$$

from the definition (3.7) of A_B^t, we have $-Bu \in A_B^t(\lambda)$ and λ is such

that

$$\exists p \in A_B^t(\lambda) \ , \ -p \in \partial F^*(\lambda),$$

which is an equivalent formulation of the inequality (3.5). The identity (3.8) follows immediately from the property $Bu \in \partial F^*(\lambda)$. ∎

If $A = \partial G$, the dual operator $A_B^t = B \circ \partial G^{-1} \circ (-B^t)$ is in fact the subgradient of the function $G^* \circ (-B^t)$ (since the qualification condition is satisfied) and the dual variational inequality (3.5) is equivalent to the variational problem

$$\inf_{\mu \in H} \{G^*(-B^t\mu)+F^*(\mu)\}, \tag{3.9}$$

the *dual problem, in the sense of Fenchel,* of the variational problem (3.4).

4. THE METHOD OF MULTIPLIERS FOR VARIATIONAL INEQUALITIES

The variational inequality

$$\left\{ \begin{array}{l} \textit{Find } u \in V \textit{ such that} \\ \\ \exists w \in A(u),\ (w, v-u)_V + F(Bv) - F(Bu) \geq 0 \quad \forall v \in V \end{array} \right. \tag{4.1}$$

is equivalent to the multivalued equation

$$\left\{ \begin{array}{l} \textit{Find } u \in V \textit{ such that} \\ \\ 0 \in T(u), \end{array} \right. \tag{4.2}$$

where since dom(F) has nonempty interior, the operator $T : V \to 2^V$ is defined by

$$T = A + B^t \circ \partial F \circ B \, . \tag{4.3}$$

By hypothesis, A is maximal monotone. The same applies for $B^t \circ \partial F \circ B$; in fact, for all $v \in V$, $v' \in V$ and all $\mu \in \partial F(Bv)$ and $\mu' \in \partial F(Bv')$ we have $(B^t\mu - B^t\mu', v-v')_V = (\mu-\mu', Bv-Bv')_H \geq 0$ since F is convex and proper; the maximal monotonicity then

follows from the lower semi-continuity of F. The qualification condition and the property $\text{int}(\text{dom}(F)) = \text{int}(\text{dom}(\partial F))$ guarantee that T, the sum of two maximal monotone operators, is itself maximal monotone (see BREZIS [1], Corollary 2.7).

Similarly, the dual variational inequality;

$$(4.4)\qquad \left\{ \begin{array}{l} \textit{Find } \lambda \in H \textit{ such that} \\ \\ p \in A_B^t(\lambda),\ (p,\mu-\lambda)_H + F^*(\mu) - F^*(\lambda) \geq 0 \quad \forall \mu \in H \end{array} \right.$$

is equivalent to

$$(4.5)\qquad 0 \in U(\lambda)$$

where the operator $U : H \to 2^H$ is defined by the sum

$$(4.6)\qquad U = A_B^t + \partial F^*.$$

The subgradient $\partial F^* = (\partial F)^{-1}$ is maximal monotone. Furthermore, for μ and $\mu' \in H$ and for all $q \in A_B^t(\mu)$ and $q' \in A_B^t(\mu')$ we have

$$(q-q',\mu-\mu')_H = (v-v',w-w')_V,$$

with $w = -B^t\mu \in A(v)$, $w' = -B^t\mu' \in A(v')$, so that

$$(q-q',\mu-\mu')_H \geq 0,$$

which proves the monotonicity of A_B^t. In order to prove the maximal monotonicity of A_B^t we adopt a supplementary assumption; we shall assume henceforth either that

$$(4.7)\qquad A \textit{ is coercive} \text{ (in the sense of (2.9))},$$

or that

$$(4.8)\qquad B^tB \textit{ is an isomorphism of } V.$$

PROPOSITION 4.1: *Suppose that one of the assumptions* (4.7) *or* (4.8) *is satisfied. Then the operator* A_B^t *defined by* (3.7) *is maximal monotone.*

Proof: We shall now show that for all $r > 0$ the multivalued equation

$$y \in (I+rA_B^t)(\mu) \tag{4.9}$$

admits for all $y \in H$ a unique solution μ. The vector $\mu \in H$, if it exists, satisfies, by definition of A_B^t,

$$y = \mu - rBv,$$

with v such that $-B^t \mu \in A(v)$; v is thus a solution of the multivalued equation

$$-B^t y \in (A+rB^tB)(v) \tag{4.10}$$

which, for all $y \in H$, admits a unique solution if (4.7) or (4.8) is satisfied. We then deduce the unique solution of (4.9) to be

$$\mu = y+rB(A+rB^tB)^{-1}(-B^t y), \tag{4.11}$$

and we also deduce the maximal monotonicity of A_B^t (see BREZIS [1], Proposition 2.2). ∎

Suppose that the variational inequality (4.1) admits a solution $u \in \text{int}(\text{dom}(A))$; Theorem 3.1 indicates that the dual inequality (4.4) possesses a solution λ and $\lambda \in \text{int}[\text{dom}(A_B^t)] \cap \text{dom}(\partial F^*)$; hence the maximal monotonicity of U (see BREZIS [1]). We can therefore solve the "dual" multivalued equation (4.5) by using the proximal-point algorithm for U; given a nondecreasing sequence $\{r_n\}$ of positive numbers and an initial approximation $\lambda^0 \in H$, we generate the sequence $\{\lambda^n\}$ via the recurrence

$$\lambda^{n+1} = J_U^{r_n}(\lambda^n) \quad , \quad n=0,1,\ldots \quad . \tag{4.12}$$

By definition of the resolvent $J_U^{r_n} = (I+r_nU)^{-1}$, we have

$$\lambda^{n+1} = \lambda^n + r_n(Bu^{n+1}-p^{n+1}) \tag{4.13}$$

with u^{n+1}, p^{n+1} such that

$$-B^t\lambda^{n+1} \in A(u^{n+1}), \tag{4.14}$$

$$\lambda^{n+1} \in \partial F(p^{n+1}) ; \tag{4.15}$$

inserting (4.13) into (4.14) and (4.15), we characterise u^{n+1}, p^{n+1} as a solution of

$$0 \in A(u^{n+1}) + r_n B^t B u^{n+1} + B^t(\lambda^n - r_n p^{n+1}), \tag{4.16}$$

$$0 \in \partial F(p^{n+1}) + r_n p^{n+1} - \lambda^n - r_n B u^{n+1}. \tag{4.17}$$

We can therefore describe the algorithm in the following form, which *generalises to the variational inequality* (4.1) *Uzawa's algorithm* ALG1 *for the augmented Lagrangian* (1.13) associated with the variational problem (1.10).

Multiplier methods for variational inequalities (ALG1):

Given a nondecreasing sequence $\{r_n\}$ *of positive numbers, and an initial approximation* $\lambda^0 \in H$, *we deduce a sequence* $\{\lambda^n\}$ *via the following recurrence:*

(i) *Given* λ^n, *find* $\{u^{n+1}, p^{n+1}\} \in V \times H$ *satisfying*

$$\begin{cases} \exists w^{n+1} \in A(u^{n+1}) \ \textit{such that} \\ (w^{n+1}, v)_V + (\lambda^n - r_n p^{n+1} + r_n B u^{n+1}, Bv)_H = 0 \quad \forall v \in V, \end{cases} \tag{4.18}$$

$$\begin{cases} F(p^{n+1}) - (\lambda^n, p^{n+1})_H + \frac{r_n}{2} |Bu^{n+1} - p^{n+1}|_H^2 \\ \leq F(q) - (\lambda^n, q)_H + \frac{r_n}{2} |Bu^{n+1} - q|_H^2 \quad \forall q \in H. \end{cases} \tag{4.19}$$

(ii) *Update the multipliers:*

$$\lambda^{n+1} = \lambda^n + r_n(Bu^{n+1} - p^{n+1}). \quad \blacksquare \tag{4.20}$$

At each iteration we therefore have to solve a variational equation in v coupled with a minimisation problem in q. The convergence of the method follows from the convergence of the proximal-point algorithm for the maximal monotone operator U.

THEOREM 4.1: *Suppose that the variational inequality* (4.1) *admits at least one solution* $v \in \text{int}(\text{dom}(A))$ *and that one of the assumptions* (4.7), (4.8) *is satisfied. Then the method of multipliers* ALG1 *is well defined and generates a sequence* $\{\lambda^n\}$ *converging weakly to* λ, *a solution of the dual variational inequality* (4.4). *If, in addition,* A^{-1} *is coercive with modulus* γ, *which implies that* A *is Lipschitz-continuous with constant* γ^{-1}, *then the sequence* $\{\lambda^n\}$ *converges strongly to* λ, *the unique solution of* (4.4) *and*

$$|\lambda^{n+1}-\lambda|_H \leq (1+\gamma r_n)^{-1}|\lambda^n-\lambda|_H. \tag{4.21}$$

If $r_n \to +\infty$, *then the convergence of the sequence* $\{\lambda^n\}$ *is superlinear.*

Proof: The first part of the theorem is a corollary of Theorem 2.1, whilst the estimate of the rate of convergence follows from Theorem 2.2.

Remark 4.1: We can also apply the *proximal-point algorithm with relaxation*, (2.7), to the dual multivalued equation (4.5); we obtain a method of multipliers, in which the updating formula (4.20) is changed into

$$\lambda^{n+1} = \lambda^n+\rho_n(Bu^{n+1}-p^{n+1}). \tag{4.22}$$

Remark 2.3 allows us to conclude that the method converges for all ρ_n, $\rho_n = \omega r_n$, with $0 < \omega < 2$; we have thus generalised the convergence result of Chapter III. The analysis of Remark 2.4 indicates that the rate of convergence of the method of multipliers can be accelerated by suitably choosing ρ_n in (4.22); assuming r fixed, an approximation of the optimal parameter is given by $\rho^* = \omega^* r = r + \frac{\alpha}{M^2} > r$; this result should be compared with the analysis carried out in Chapter I in the simpler context of quadratic functionals.

Remark 4.2: For the solution of (4.18), (4.19), we can obviously use the usual successive relaxation method; by performing only a single inner iteration, we obtain algorithm ALG2, to which we shall return in Section 5 below.

Remark 4.3: We can also attempt to apply the proximal-point algorithm relative to the maximal monotone operator T defined by

(4.3), to solve the multivalued equation (4.2), equivalent to the inequality (4.1). Thus, given a nondecreasing sequence $\{r_n\}$ of positive numbers and an initial approximation u^0, we define a sequence $\{u^n\}$ in V by the following recurrence:

Given u^n *find* u^{n+1} *satisfying the variational inequality*

$$(4.23)\quad \begin{cases} \exists w^{n+1} \in A(u^{n+1}),\ p^{n+1} \in \partial F(Bu^{n+1}) \ \textit{such that} \\ (w^{n+1}+r_n^{-1}(u^{n+1}-u^n), v-u^{n+1})_V+(p^{n+1}, Bv-Bu^{n+1})_H \geq 0 \quad \forall v \in V. \end{cases}$$

Problem (4.23) is difficult to solve since it involves A, B and F all at the same time. The introduction of the duality and the method of multipliers enables us to get round this difficulty very effectively.

5. DECOMPOSITION BY MULTIPLIERS: (I) ALTERNATING-DIRECTION METHODS

The variational inequalities (4.1) and (4.4) are equivalent to the multivalued equations (4.2) and (4.5), which are of the form:

$$(5.1)\quad \begin{cases} \textit{Find}\ \ x \in X\ \ \textit{such that} \\ 0 \in T(x), \end{cases}$$

where T is a maximal monotone operator such that

$$(5.2)\qquad T = R+S,$$

where R and S denote maximal monotone operators on X.

The proximal-point algorithm requires the calculation of J_T^r which may be much more complicated than that of J_R^r and J_S^r. Recently, P.L. LIONS and B. MERCIER [1] have proposed two algorithms which generalise alternating-direction methods to the multivalued equation form (5.1) when T is given by (5.2), and which involve only the resolvents of R and S. In particular, they analyse an algorithm of the DOUGLAS-RACHFORD type [1]:

$$(5.3)\qquad t^{n+1} = J_R^r(2J_S^r-I)t^n + (I-J_S^r)t^n,\ n=0,1,\ldots,$$

and an algorithm of the PEACEMAN-RACHFORD type [1]:

$$t^{n+1} = (2J_R^r - I)(2J_S^r - I)t^n \ , \ n=0,1,\ldots \ ; \tag{5.4}$$

knowing an approximation x^o of a solution x of (5.1), these algorithms must be initialised at t^o such that

$$x^o = J_S^r(t^o), \tag{5.5}$$

that is

$$t^o = x^o + rs^o, \ s^o \in S(x^o). \tag{5.6}$$

We can apply these algorithms to the multivalued equation (4.5) equivalent to the dual variational inequality (4.4); this corresponds to a decomposition of type (5.2) with

$$R = A_B^t \ , \ S = \partial F^* = (\partial F)^{-1} \tag{5.7}$$

on the Hilbert space H. We thus obtain two *variants of the method of multipliers* (ALG1) *in which the problems in* v *and* q *are now decoupled.*

5.1 The Douglas-Rachford variant of the method of multipliers: algorithm ALG2

Given an approximation λ^o *of the solution of the dual inequality* (4.4) *we define* t^o *such that* $\lambda^o = J_S^r(t^o)$, *that is*

$$t^o = \lambda^o + rp^o \tag{5.8}$$

with $p^o \in \partial F^*(\lambda^o)$, and hence such that

$$\lambda^o \in \partial F(p^o). \tag{5.9}$$

In general we put

$$\lambda^n = J_S^r(t^n) \ ; \tag{5.10}$$

by definition of the resolvent J_S^r, there exists $p^n \in \partial F^*(\lambda^n)$ such that

(5.11) $\lambda^n = t^n - rp^n$

and

(5.12) $\lambda^n \in \partial F(p^n)$.

Using the formula (4.11) which characterises $J_R^r = (I+rA_B^t)^{-1}$ when one of the assumptions (4.7), (4.8) is satisfied, the recurrence (5.3) defines

(5.13) $t^{n+1} = \lambda^n + rBu^{n+1}$,

with

(5.14) $u^{n+1} = (A+rB^tB)^{-1}(rB^tp^n - B^t\lambda^n)$.

Inserting relation (5.11), for the index $n+1$, into (5.13) we obtain the recurrence formula for the sequence $\{\lambda^n\}$:

(5.15) $\lambda^{n+1} = \lambda^n + r(Bu^{n+1} - p^{n+1})$,

which is in fact identical to the formula (4.20) for updating the multipliers, when the parameter r is fixed. We note that by recurrence $p^{n+1} \in \partial F^*(\lambda^{n+1})$ and therefore satisfies

(5.16) $\lambda^n + rBu^{n+1} \in \partial F(p^{n+1}) + rp^{n+1}$,

this being a multivalued equation which has the unique solution

(5.17) $p^{n+1} = (\partial F + rI)^{-1}(\lambda^n + rBu^{n+1})$.

We can thus use (5.14), (5.15), (5.17) to describe algorithm (5.3) in a form similar to the method of multipliers, as below.

The D.R. variant of the method of multipliers (ALG2):

Given $r > 0$, *and initial approximations* $\lambda^0 \in H$, $p^0 \in H$ *such that* $\lambda^0 \in \partial F(p^0)$, *define the sequences* $\{u^n\},\{p^n\},\{\lambda^n\}$ *by the recurrence* :

(i) λ^n *and* p^n *being known, seek* $u^{n+1} \in V$ *satisfying the variational inequality*

$$(5.18) \qquad \exists w^{n+1} \in A(u^{n+1}) \text{ such that } (w^{n+1},v)_V + (\lambda^n - rp^n + rBu^{n+1}, Bv)_H = 0 \quad \forall v \in V ;$$

(ii) u^{n+1} *being known, seek a solution* p^{n+1} *of the minimisation problem:*

$$(5.19) \qquad F(p^{n+1}) - (\lambda^n, p^{n+1})_H + \frac{r}{2}|Bu^{n+1} - p^{n+1}|_H^2 \leq F(q) - (\lambda^n, q)_H + \frac{r}{2}|Bu^{n+1} - q|_H^2 \quad \forall q \in H ;$$

(iii) *update the multipliers by*

$$(5.20) \qquad \lambda^{n+1} = \lambda^n + r(Bu^{n+1} - p^{n+1}). \quad \blacksquare$$

In this form, the method can be seen to be a variant of algorithm ALG1 of Section 4 in which the problem (4.18), (4.19) is solved in approximate fashion by performing only a single relaxation step. This generalises algorithm ALG2 of Chapter III to the case of variational inequalities. Problem (5.18) is especially simple when A is an *affine single-valued operator* since, following an appropriate discretisation, it reduces to the solution of a *linear system with a matrix which is constant* during the course of the iterations. Problem (5.19) consists of the minimisation of a strongly convex function which is *independent* of B; we can therefore solve this easily using an iterative method, even if B is ill-conditioned; we have achieved a *decoupling of the difficulties relating to* F *and to* B. Finally, we note that if F has a *separable structure*, i.e. if H can be written as the cartesian product of m spaces H_i, i = 1,....m and if F is defined as the sum

$$(5.21) \qquad F = \sum_{i=1}^{m} F_i$$

of functions $F_i : H_i \to]-\infty,+\infty]$ which are lower semi-continuous, proper and convex, then *problem* (5.19) *decomposes into* m *independent problems on each of the* H_i of the form:

$$(5.22) \qquad p_i^{n+1} = \underset{q_i \in H_i}{\text{Arg min}} \{F_i(q_i) - (\lambda_i^n, q_i)_{H_i} + \frac{r}{2}|(Bu^{n+1})_i - q_i|_{H_i}^2\},$$

where the subscript i denotes the component pertaining to H_i.

THEOREM 5.1: *Suppose that the variational inequality* (4.1) *admits at least one solution* $u \in \text{int}(\text{dom}(A))$ *and that one of the assumptions* (4.7) - (4.8) *is satisfied. Then algorithm* ALG2 *is well defined and constructs a sequence* $\{t^n = \lambda^n + rp^n\}$ *which converges weakly to* t *such that* $\lambda = J_S^r(t)$ *satisfies the dual variational inequality* (4.4). *The sequence* $\{\lambda^n\}$ *is bounded and* $|\lambda^{n+1}-\lambda^n|_H \to 0$.

Proof: The theorem follows from Theorem 3.1 and from an analysis of the convergence of algorithm (5.3) (see LIONS-MERCIER [1], Section 1.3, Proposition 2). ∎

We can conclude that the sequence $\{\lambda^n\}$ converges weakly to λ if J_R^r or J_S^r is a compact mapping. Finally, we give an estimate of the rate of convergence in the special case where $(\partial F)^{-1}$ is both Lipschitz-continuous (with constant γ) and coercive (with modulus $\alpha \leq \gamma$), i.e. if for all $q,q' \in H$ and for all $\lambda \in \partial F(q)$, $\lambda' \in \partial F(q')$:

$$|q'-q|_H \leq \gamma|\lambda'-\lambda|_H , \tag{5.23}$$

$$(\lambda'-\lambda,q'-q)_H \geq \alpha|\lambda'-\lambda|_H^2. \tag{5.24}$$

THEOREM 5.2: *Suppose that the assumptions of* Theorem 5.1 *are satisfied and furthermore that* $(\partial F)^{-1}$ *is Lipschitz-continuous (with constant* γ*) and coercive (with modulus* $\alpha \leq \gamma$*). Then the sequence* $\{\lambda^n\}$ *defined by algorithm* ALG2 *converges strongly to* λ, *a solution of the dual variational inequality* (4.4), *and*

$$|\lambda^{n+1}-\lambda|_H \leq \left(1- \frac{2r\alpha}{(1+\gamma r)^2}\right)^{1/2}|\lambda^n-\lambda|_H . \tag{5.25}$$

In particular, there exists an optimal parameter r^* *for which we get the estimate*

$$|\lambda^{n+1}-\lambda|_H \leq \left(1- \frac{\alpha}{2\gamma}\right)^{1/2}|\lambda^n-\lambda|_H. \tag{5.26}$$

Proof: See LIONS-MERCIER [1], Section 1.3, Proposition 4.

Remark 5.1: Using the relationship established in Proposition 2.2

between the resolvents of ∂F^* and its inverse ∂F, we can re-express the recurrence (5.3) defining algorithm ALG2 in the form

$$t^{n+1} = J^r_{A^t_B}(r(I-2J^\varepsilon_{\partial F})(\varepsilon t^n)) + rJ^\varepsilon_{\partial F}(\varepsilon t^n) \ , \tag{5.27}$$

where $\varepsilon=1/r$. Using the equation (4.11) defining $J^r_{A^t_B}$, it can be shown that εt^n also satisfies the recurrence

$$\varepsilon t^{n+1} = \tilde{J}^\varepsilon_A((2J^\varepsilon_{\partial F}-I)(\varepsilon t^n))+(I-J^\varepsilon_{\partial F})(\varepsilon t^n), \tag{5.28}$$

where $\tilde{J}^\varepsilon_A$ is the contraction defined on H (when (4.7) or (4.8) is satisfied) by

$$\tilde{J}^\varepsilon_A = B\circ(B^tB+\varepsilon A)^{-1}\circ B^t. \tag{5.29}$$

The formulation (5.28) is interesting for two reasons: it expresses algorithm ALG2 in the terms of the *primal variational inequality* (4.1), and it introduces naturally the new operator $\tilde{J}^\varepsilon_A$.

In the special case where $V = H$ and $B = I$, relation (5.28) is in fact identical to the Douglas-Rachford algorithm (5.3) for solving the multivalued equation

$$0 \in Au + \partial F(u), \tag{5.30}$$

associated with the primal variational inequality; it thus generates the same iterates as when it is applied to the dual variational inequality.

5.2 The Peaceman-Rachford variant of the method of multipliers: algorithm ALG3

We use the same initialisation (5.8), (5.9) as before, and we again put

$$t^n = \lambda^n+rp^n \ \textit{with} \ \ \lambda^n = J^r_S(t^n) \ ; \tag{5.31}$$

we have

$$(2J^r_S-I)(t^n) = \lambda^n-rp^n. \tag{5.32}$$

It is now convenient to introduce

$$\lambda^{n+1/2} = J_R^r(\lambda^n - rp^n) = \lambda^n - rp^n + rBu^{n+1}, \tag{5.33}$$

where u^{n+1} is again defined by (5.14). The recurrence (5.4) is written

$$t^{n+1} = J_R^r(\lambda^n - rp^n) + (J_R^r - I)(\lambda^n - rp^n) = \lambda^{n+1/2} + rBu^{n+1}, \tag{5.34}$$

giving the formula

$$\lambda^{n+1} = \lambda^{n+1/2} + r(Bu^{n+1} - p^{n+1}), \tag{5.35}$$

with p^{n+1} now defined by

$$p^{n+1} = (\partial F + rI)^{-1}(\lambda^{n+1/2} + rBu^{n+1}). \tag{5.36}$$

We can therefore describe algorithm (5.4) by using (5.14), (5.33), (5.35), (5.36), as below.

The P.R. variant of the method of multipliers (ALG3):

Given $r > 0$, *and initial approximations* $\lambda^0 \in H$ *and* $p^0 \in H$ *such that* $\lambda^0 \in \partial F(p^0)$, *define the sequences* $\{u^n\},\{p^n\},\{\lambda^n\}$ *via the recurrence:*

(i) *Given* λ^n, p^n, *find* $u^{n+1} \in V$ *satisfying the variational equation*

$$\begin{cases} \exists w^{n+1} \in A(u^{n+1}) \text{ such that} \\ (w^{n+1}, v)_V + (\lambda^n - rp^n + rBu^{n+1}, Bv)_H = 0 \quad \forall v \in V ; \end{cases} \tag{5.37}$$

(ii) *Update the multipliers:*

$$\lambda^{n+1/2} = \lambda^n + r(Bu^{n+1} - p^n) ; \tag{5.38}$$

(iii) *Find* $p^{n+1} \in H$ *satisfying the minimisation problem*

$$(5.39)\qquad \begin{cases} F(p^{n+1})-(\lambda^{n+1/2},p^{n+1})_H+\frac{r}{2}|Bu^{n+1}-p^{n+1}|_H^2 \\ \leq F(q)-(\lambda^{n+1/2},q)_H+\frac{r}{2}|Bu^{n+1}-q|_H^2 \quad \forall q\in H\ ; \end{cases}$$

(iv) *Update the multipliers*

$$(5.40)\qquad \lambda^{n+1}=\lambda^{n+1/2}+r(Bu^{n+1}-p^{n+1}).\ \blacksquare$$

The Peaceman-Rachford variant (P.R.) differs from the Douglas-Rachford variant (D.R.) only through the addition of the intermediate update of the multipliers (5.38); it thus offers the same set of advantages. The Peaceman-Rachford variant is, however, less 'robust', in that it converges under more restrictive assumptions than the Douglas-Rachford variant; nonetheless, as we shall see, if it does converge, then its rate of convergence is faster.

THEOREM 5.3: *Suppose that the assumptions of Theorem* (5.1) *are satisfied. Then the algorithm* ALG3 *is well defined and the sequences* $\{Bu^n\},\{p^n\},\{\lambda^n\}$ *and* $\{t^n=\lambda^n+rp^n\}$ *are bounded in* H; *there exists an extracted subsequence of* $\{t^n\}$ *which converges weakly to* $t\in H$ *such that* $\lambda=J_S^r(t)$ *satisfies the dual variational inequality* (4.4).

Proof: The first part of the theorem follows from LIONS-MERCIER [1], Section 1.2, Proposition 1. We next note that, since J_R^r and J_S^r are firm contractions (in the sense of (2.3)), then $(2J_R^r-I)$ and $(2J_S^r-I)$ are themselves contractions, as is their product. Since the set of fixed points of $(2J_R^r-I)(2J_S^r-I)$ is nonempty (since by hypothesis there exists at least one solution of the equation $0\in R(p)+S(p)$), we can extract from $\{t^n\}$ a subsequence which converges weakly to one of these fixed points t. $\blacksquare$

We shall now give an estimate for the rate of convergence, under the same assumptions as for algorithm ALG2.

THEOREM 5.4: *Suppose that the assumptions of Theorem* 5.2 *are satisfied. Then the sequence* $\{\lambda^n\}$ *defined by algorithm* ALG3 *converges strongly to* λ, *a solution of the dual inequality* (4.4), *and we have*

$$|\lambda^{n+1}-\lambda|_H \le (1-\frac{4r\alpha}{(1+\gamma r)^2})^{1/2}|\lambda^n-\lambda|_H \ . \tag{5.41}$$

In particular, there exists an optimal parameter r^* *for which we have the estimate*

$$|\lambda^{n+1}-\lambda|_H \le (1-\frac{\alpha}{\gamma})^{1/2}|\lambda^n-\lambda|_H. \tag{5.42}$$

Proof: As condition (5.23) states that $(\partial F)^{-1}$ is Lipschitz-continuous, this implies that

$$|t^n-t|_H \le |\lambda^n-\lambda|_H + r|p^n-p|_H \le (1+\gamma r)|\lambda^n-\lambda|_H. \tag{5.43}$$

Since the mapping $(2J_R^r-I)$ is a contraction, we have

$$|t^{n+1}-t|_H \le |(2J_S^r-I)t^n - (2J_S^r-I)t|_H,$$

so that

$$|t^{n+1}-t|_H^2 \le 4|J_S^r(t^n)-J_S^r(t)|_H^2 - 4(J_S^r(t^n)-J_S^r(t),t^n-t)_H + |t^n-t|_H^2. \tag{5.44}$$

Using once again the argument in the proof of Proposition 2.1, and using condition (5.24), it can be shown that

$$|J_S^r(t^n)-J_S^r(t)|_H^2 \le (J_S^r(t^n)-J_S^r(t),t^n-t)_H - r\alpha|\lambda^n-\lambda|_H^2 \ ; \tag{5.45}$$

inserting (5.45) into (5.44) and using (5.43), the estimate (5.41) can be established. The constant $(1-\frac{4r\alpha}{(1+\gamma r)^2})^{1/2}$ is minimal for $r^* = \frac{1}{\gamma}$, i.e. the same value as for ALG2, and is thus equal to $(1-\frac{\alpha}{\gamma})^{1/2}$. ∎

Comparison of (5.26) and (5.42) would appear to indicate that ALG3 is faster than ALG2. Naturally, these estimates are not precise enough to allow us to reach such a conclusion definitively.

6. DECOMPOSITION BY MULTIPLIERS: (II) PROJECTION METHODS

Once again, we consider the multivalued equation

$$(6.1)\qquad \begin{cases} \textit{Find } x \in X \textit{ such that} \\ 0 \in T(x), \end{cases}$$

where T is defined as the sum of two maximal monotone operators:

$$(6.2)\qquad T = R+S.$$

We now assume that R is *single-valued* and Lipschitz-continuous with constant M, and we consider for $r > 0$ the algorithm

$$(6.3)\qquad x^{n+1} = J_S^r(I-rR)\,(x^n)\quad , n=0,1,\ldots .$$

Remark 6.1: Algorithm (6.3) can be interpreted as a discretisation scheme for the multivalued evolution equation

$$(6.4)\qquad 0 \in \frac{dx}{dt} + R(x) + S(x),$$

which is explicit relative to R and implicit relative to S; if r denotes the time step, then (6.3) defines the solution of the discretised equation

$$(6.5)\qquad 0 \in \frac{x^{n+1}-x^n}{r} + R(x^n) + S(x^{n+1}).$$

THEOREM 6.1: *If* R *is Lipschitz-continuous (with constant* M*) and coercive (with modulus* α*), then the sequence* $\{x^n\}$ *generated by* (6.3) *converges strongly to* $x \in H$, *satisfying* (6.1) *for all* $0 < r < 2\alpha M^{-2}$. *If* R^{-1} *is coercive (with modulus* M^{-1}*), then the sequence* $\{x^n\}$ *converges weakly to* x *for all* $0 < r < 2/M$.

Proof: If R is Lipschitz-continuous and coercive, then $(I-rR)$ is a strict contraction for all $0 < r < 2\alpha M^{-2}$ and we deduce the first part of the theorem by noting that J_S^r is also a contraction.

Let us now assume that R^{-1} is coercive (with modulus M^{-1}), i.e.

$$(6.6)\qquad (R(x')-R(x),\ x'-x)_X \geq M^{-1}\,|R(x')-R(x)|_X^2 \quad \forall x,x' \in X.$$

Then

$$(6.7)\qquad \begin{cases} |(I-rR)(x') - (I-rR)(x)|_X^2 \le |x'-x|_X^2 + (Mr^2-2r)(R(x')-R(x),x'-x)_X \\ = (2-Mr)((I-rR)(x')-(I-rR)(x),\ x'-x)_X + (Mr-1)|x'-x|_X^2. \end{cases}$$

For $1/M \le r < 2/M$, (6.7) implies, since J_S^r is a firm contraction, that $|x^{n+1}-x^n|_X \to 0$ and we then deduce, from Opial's lemma, the weak convergence of $\{x^n\}$ to a solution x of (6.1).

For $0 < r \le \frac{1}{M}$, the monotonicity of R leads to the upper bound

$$(6.8)\qquad |(I-rR)(x')-(I-rR)(x)|_X^2 \le ((I-rR)(x')-(I-rR)(x),x'-x)_X,$$

and $(I-rR)$ is a firm contraction; we therefore conclude that $|x^{n+1}-x^n|_X \to 0$ and that we have weak convergence of $\{x^n\}$ (see also MERCIER [1]). ∎

We illustrate this algorithm by applying it first to the *variational inequality* (1.5) in the particular case where A is a *single-valued* operator from V into V, which is monotone and semi-continuous (and therefore maximal monotone):

$$(6.9)\qquad \begin{cases} \textit{Find } u \in K \textit{ such that} \\ (A(u),v-u)_V \ge 0 \quad \forall v \in K\ ; \end{cases}$$

the variational inequality (6.9) is equivalent to equation (6.1) with

$$(6.10)\qquad R = A \quad \textit{and} \quad S = \partial I_K.$$

LEMMA 6.1: *The resolvent* J_S^r *of* $S = \partial I_K$, *the subgradient of the indicator function* I_K *of a convex set* K *which is a nonempty closed subset of* V, *is independent of* $r > 0$ *and equal to the operator of projection onto* K, i.e.

$$(6.11)\qquad J_S^r = (I+r\partial I_K)^{-1} = P_K.$$

Proof: From Remark 2.1, we have

$$J_S^r(w) = \underset{v \in V}{\text{Arg min}} \{ \frac{1}{2} |v-w|_V^2 + rI_K(v) \}$$

$$= \underset{v \in K}{\text{Arg min}} \frac{1}{2} |v-w|_V^2 = P_K(w). \blacksquare$$

Algorithm (6.3) thus leads to the recurrence

$$u^{n+1} = P_K(u^n - rA(u^n)) \quad n=0,1,\ldots ; \tag{6.12}$$

when A is the gradient $\nabla\phi$ of a convex and continuously differentiable functional $\phi : V \to \mathbb{R}$, we recognise in (6.12) the *gradient projection method* of GOLDSTEIN [1], the convergence properties of which are obtained as a corollary of Theorem 6.1.

Let us now consider the *dual variational inequality* of (6.9):

$$\left\{ \begin{array}{l} \textit{Find } \lambda \in V \textit{ such that} \\ \\ (-A^{-1}(-\lambda), \mu-\lambda)_V + \sigma_K(\mu) - \sigma_K(\lambda) \geq 0 \quad \forall \mu \in V ; \end{array} \right. \tag{6.13}$$

this is the particular case of (3.5) in which B is the identity mapping of V, and σ_K, the conjugate functional of I_K, is the *support function* [5] of the convex set K. The inequality (6.13) is equivalent to a multivalued equation on V of the form (6.1), (6.2) with, in this case,

$$R = -A^{-1} \circ (-I), \quad S = \partial\sigma_K. \tag{6.14}$$

We suppose that A is coercive (with modulus α) which implies that A^{-1} is single-valued and Lipschitz-continuous (with constant $1/\alpha$). It can be shown, as in Lemma 6.1, that the resolvent J_S^r of $\partial\sigma_K$ is independent of r:

$$J_S^r = (I+r\partial\sigma_K)^{-1} = I-P_K ; \tag{6.15}$$

algorithm (6.3) then leads to the recurrence

$$\lambda^{n+1} = (I-P_K)(\lambda^n + rA^{-1}(-\lambda^n)), \tag{6.16}$$

---

[5] $\sigma_K(\mu) = \underset{v \in K}{\text{Sup}} (\mu, v)_V$

which can be expressed in a form similar to that of the method of multipliers.

Method of multipliers with projection (algorithm ALG4):

Given $r > 0$ *and an initial approximation* $\lambda^0 \in V$, *define* $\{u^n\},\{\lambda^n\}$ *via the recurrence:*

(i) *Given* λ^n, *find* u^{n+1}, *satisfying the variational equation*

$$(A(u^{n+1})+\lambda^n,v)_V = 0 \quad \forall v \in V \ ; \tag{6.17}$$

(ii) *Update the multipliers:*

$$\lambda^{n+1} = (I-P_K)(\lambda^n+ru^{n+1}). \ \blacksquare \tag{6.18}$$

Theorem 6.1 gives the conditions for convergence of the method.

COROLLARY 6.1: *Suppose that the variational inequality* (6.9) *admits at least one solution and that the operator* A *is coercive (with modulus* α*). Then algorithm* ALG4 *is well defined and generates a sequence* $\{\lambda^n\}$ *which converges weakly to* $\lambda \in V$, *a solution of the dual variational inequality* (6.13), *for all* $0 < r < 2\alpha$.

Remark 6.2: If the convergence of $\{\lambda^n\}$ is strong (for example if V is of finite dimension), it can be shown that the sequence $\{\lambda^n\}$ converges linearly and that there exists an optimal step r^*.

Remark 6.3: Algorithm (6.17), (6.18) effects a decomposition of the variational inequality (6.9) wherein the problem relating to A and that relating to the constraint $v \in K$ become decoupled due to the introduction of the multipliers λ^n.

7. GENERAL DISCUSSION

This chapter, which is of a distinctly more abstract character than the preceding chapters, has primarily been aimed at demonstrating the links which exist between the augmented-Lagrangian method and some of the well-known methods of non-linear analysis. These

connections may serve to suggest some new approaches for the study of algorithms and for the optimisation of the parameters which control convergence. In particular, we believe that the techniques developed in this chapter should allow an investigation of the convergence of variants of algorithm ALG1 in which a *relaxation parameter is introduced to accelerate the convergence of the inner iterations*. This procedure for accelerating convergence has actually been used in Chapter I in the simpler context of quadratic functionals and linear constraints.

This Page Intentionally Left Blank

REFERENCES

AMARA M., JOLY P., THOMAS J.M. [1] A mixed finite-element method for solving transonic flow equations, *Comp. Meth. Appl. Mech. Eng.* (to appear).

ANTMAN S.S. [1] Kirchoff's problem for nonlinearly elastic rods, *Quart. Appl. Math.*, 32, (1974), pp. 221-240.

ANTMAN S.S., ROSENFELD [1] Global behaviour of buckled states of nonlinearly elastic rods, *SIAM Rev.*, 20, (1975), pp. 513-566.

ARROW K.J., HURWICZ., UZAWA H. [1] *Studies in nonlinear programming*, Stanford University Press, 1958.

AXELSSON O., [1] A class of iterative methods for finite-element equations, *Comp. Meth. Appl. Mech. Eng.*, 9, (1976), pp. 123-137.
[2] A generalized SSOR method, *BIT*, 13, (1972), pp. 443-457.

BABUSKA I. [1] Error bounds for finite-element methods, *Numer. Math.*, 16, (1971), pp. 322-333.

BALL J.M. [1] Convexity conditions and existence theorems in nonlinear elasticity, *Arch. Rat. Mech. Anal.*, 63, (1977), pp. 307-403.

BEGIS D. [1] *Etude numérique du comportement d'un fluide de Bingham*, IRIA-Laboria Report No. 42, December, 1973.
[2] *Etude numérique de l'écoulement d'un fluide visco-plastique de Bingham par une méthode de lagrangien augmenté*, IRIA-Laboria Report No. 355, June 1979.

BENSOUSSAN A., LIONS J.L., TEMAM R. [1] Sur les méthodes de décomposition, de décentralisation, de coordination et applications, in *Sur les méthodes numériques en sciences physiques et économiques*, J.L. Lions, G.I. Marchouk eds., Dunod-Bordas, Paris, 1974, pp. 133-257.

BERCOVIER M. [1] *Régularisation des problèmes variationnels mixtes. Application aux éléments finis mixtes et extension à quelques problèmes non linéaires*, Thèse d'Etat, Université de Rouen, 1976.

BERTSEKAS D.P., [1] Multiplier methods : A survey, *Automatica*, 12, (1976), pp. 133-145.
[2] Approximation procedures based on the method of multipliers, *J. Optimisation Th. Appl.*, 23, (1977), pp. 487-510.

BOURGAT J.F., DUMAY., GLOWINSKI R. [1] Large-displacement calculations of flexible pipelines by finite elements and nonlinear programming methods, *SIAM J. Sc. Stat. Comp.*, 1, (1980), pp. 34-81.

BOURGAT J.F., GLOWINSKI R., LE TALLEC P., [1] Decomposition of variational problems. Applications in Finite Elasticity, in *Partial Differential Equations in Engineering and Applied Sciences*, R.L. Sternberg ed., Marcel Dekker, New York, 1980, pp. 445-480.

BRENT R.P. [1] *Algorithms for minimization without derivatives*, Prentice-Hall, Englewood Cliffs, N.J., 1973.

BREZIS H., [1] *Opérateurs Maximaux Monotones*, North-Holland, Amsterdam, 1973.

BREZZI F. [1] On the existence, uniqueness and approximation of saddle-point problems arising from Lagrangian multipliers, *Revue Française d'Automatique Informatique Recherche Opérationnelle, Analyse Numérique*, 8, R-2, (1974), pp. 129-151.

BREZZI F., RAPPAZ J., RAVIART P.A. [1] Finite-dimensional approximation of nonlinear problems, Part I : Branches of nonsingular solutions, *Num. Math.*, 36, (1980), pp. 1-25.
[2] Finite-dimensional approximation of nonlinear problems, Part II: Limit points, *Num. Math.*, 37, (1981), pp. 1-28.
[3] Finite-dimensional approximation of nonlinear problems, Part III : Simple bifurcations, *Num. Math.*, 38, (1981), pp. 1-30.

BRISTEAU M.O., GLOWINSKI R., PERIAUX J., PERRIER P., PIRONNEAU O., POIRIER G., [1] Application of Optimal-Control and Finite-Element Methods to the calculation of Transonic Flows and Incompressible Viscous Flows, in *Numerical Methods in Applied Fluid Dynamics*, B. Hunt ed., Acad. Press, London (1980), pp. 203-312.
[2] Transonic flow simulations by finite-element and least-squares methods, in *Finite Elements in Fluids*, Vol. 4, R.H. Gallagher, D.H. Norrie, J.T. Oden, O.C. Zienkiewicz eds., John Wiley, Chichester (1982), pp. 453 - 482.

CEA J. [1] *Optimisation : Théorie et Algorithmes*, Dunod, Paris, 1971.
[2] *Optimisation : Theory and algorithms*, Lectures on Mathematics and Physics, Vol. 53, Tata Institute of Fundamental Research, Bombay, 1978 (distributed by Springer-Verlag, Berlin).

CEA J., GLOWINSKI R. [1] Sur des méthodes d'optimisation par relaxation, *Revue Française d'Automatique Informatique Recherche Opérationnelle, Analyse Numérique*, R-3, (1973), pp. 5-32.

CHAN T., GLOWINSKI R. [1] *Finite-Element approximation and iterative solution of a class of mildly non-linear elliptic equations*, STAN-CS-78-674, Computer Science Department, Stanford University, 1978.
[2] Numerical methods for a class of mildly nonlinear elliptic equations, in *Atos Do Decimo Primeiro Coloquio Brasileiro de Matematica*, Vol. I, Impa, Rio de Janeiro, 1978, pp. 279-318.

CHORIN A.J. [1] A numerical method for solving incompressible flow problems, *J. Comp. Physics*, 2, (1967), pp. 12-26.

CIARLET P.G. [1] *The finite-element method for elliptic problems*, North-Holland, Amsterdam, 1978.

CONCUS P., GOLUB G.H. [1] A generalized conjugate-gradient method for nonsymmetric systems of linear equations, in *Computing Methods in Applied Sciences and Engineering*, R. Glowinski, J.L. Lions eds., Lecture Notes in Economics and Mathematical Systems, Vol. 134, Springer-Verlag, Berlin, 1976, pp. 56-65.

CONCUS P., GOLUB G.H., O'LEARY D.P. [1] Numerical solution of nonlinear partial differential equations by a generalized conjugate-gradient method, *Computing*, 19, (1977), pp. 321-340.

COOPER L., KATZ I.N. [1] The Weber problem revisited, *Comp. and Maths. with Appls.*, 7, (1981), pp. 225-234.

CROUZEIX M. [1] Etude d'une méthode de linéarisation. Résolution numérique des équations de Stokes stationnaires. Applications aux équations de Navier-Stokes stationnaires, in *Approximations et Méthodes Itératives de Résolution d'Inéquations Variationnelles et de Problèmes Non Linéaires*, Cahier de l'IRIA, No. 12, 1974, pp. 139-244.

CROUZEIX M., RAVIART P.A. [1] Conforming and non-conforming finite-element methods for solving the stationary Stokes equations, *Revue Française d'Automatique Informatique Recherche Opérationnelle, Analyse Numérique*, 7, R-3, 1973, pp. 33-76.

DANIEL J. [1] *The approximate minimization of functionals*, Prentice Hall, Englewood Cliffs, N.J., 1970.

DOUGLAS J., DUPONT T. [1] Interior penalty procedures for elliptic and parabolic Galerkin methods, in *Computing Methods in Applied Sciences*, R. Glowinski, J.L. Lions, eds., Lecture Notes in Physics, Vol. 58, Springer-Verlag, Berlin, 1976, pp. 207-216.

DOUGLAS J., RACHFORD H.H. [1] On the numerical solution of heat-conduction problems in two or three space variables, *Trans. Math. Soc.*, 82, (1956), pp. 421-439.

DURAND E. [1] *Magnétostatique*, Masson, Paris, 1968.

DUVAUT G., LIONS J.L. [1] *Les inéquations en Mécanique et en Physique*, Dunod, Paris, 1972.

EKELAND I., TEMAM R. [1] *Analyse Convexe et Problèmes Variationnels*, Dunod, Gauthier Villars, Paris, 1974.

FLETCHER, R. [1] Methods related to Lagrangian functions, Chapter VIII of *Numerical Methods for Constrained Optimization*, P.E. Gill, W. Murray eds., Academic Press, London, 1974, pp. 219-239.

FORTIN M. [1] Minimization of some non-differentiable functions by the augmented Lagrangian method of Hestenes and Powell, *Appl. Math. Opt.*, 2, (1976), pp. 236-250.

[2]$_1$ Utilisation de la méthode des éléments finis en Mécanique des Fluides, I, *Calcolo*, 4, (1975), pp. 406-441.

[2]$_2$ Utilisation de la méthode des éléments finis en Mécanique des Fluides, II, *Calcolo*, 1, (1976), pp. 1-20.

[3] *Leçons sur l'Analyse Convexe et l'approximation des problèmes de points-selle*, Orsay Mathematical Publication, Université Paris-Sud, Department of Mathematics, 1978.

[4] An analysis of the convergence of mixed finite-element methods, *Revue Française d'Automatique Informatique Recherche Opérationnelle, Analyse Numérique*, 11, R-3, (1977), pp. 341-354.

FORTIN M., PEYRET R., TEMAM R. [1] Résolution numérique des équations de Navier-Stokes pour un fluide incompressible, *Journal de Mécanique*, 10, (1971), 3, pp. 357-390.

FORTIN M., SOULIE M. [1] A second-order non-conforming finite element on the triangle (to appear).

FORTIN M., THOMASSET F. [1] Mixed finite-element methods for incompressible flow problems, *J. of Comp. Physics*, 31, (1979), pp. 113-145.

GABAY D. [1] *Méthodes numériques pour l'optimisation non linéaire*, Thèse d'Etat, Université Pierre et Marie Curie, 1979.

GABAY D., MERCIER B. [1] A dual algorithm for the solution of non-linear variational problems via finite-element approximations, *Comp. and Math. with Applications*, 2, (1976), 1, pp. 17-40.

GERMAIN P. [1] *Mécanique des milieux continus*, Vol. 1, Masson, Paris, 1973.

GILL P.E., MURRAY W. [1] (eds.) *Numerical Methods for Constrained Optimization*, Acad. Press, London, 1974.

GIRAULT V., RAVIART P.A. [1] *Finite-element approximation of Navier-Stokes equations*, Lectures Notes in Math., Vol. 749, Springer-Verlag, Berlin, 1979.

GLOWINSKI R. [1] *Numerical methods for nonlinear variational problems*, Lectures in Math. and Physics, Vol. 65, Tata Institute of Fundamental Research, Bombay, 1980 (distributed by Springer-Verlag, Berlin).
[2] *Numerical methods for nonlinear variational problems* (second edition), Springer-Verlag, (in preparation).

GLOWINSKI R., LANCHON H. [1] Torsion élasto-plastique d'une barre cylindrique de section multiconnexe, *Journal de Mécanique*, Vol. 12, No. 1, (1973), pp. 151-171.

GLOWINSKI R., LE TALLEC P. [1] Numerical solution of problems in incompressible finite elasticity by augmented Lagrangian methods. (I) Two-dimensional and axisymmetric problems, *SIAM J. Appl. Math.* Vol. 42, No. 2, (1982), pp. 400 - 429.
[2] Numerical solution of problems in incompressible finite elasticity by augmented Lagrangian methods. (II) Three-dimensional problems (to appear).

GLOWINSKI R., LE TALLEC P., RUAS V. [1] Approximate solution of non-linear problems in incompressible finite elasticity, in *Nonlinear Finite Element Analysis in Structural Mechanics*, W. Wunderlich, E. Stein, K.J. Bathe eds., Springer-Verlag, Berlin, 1981, pp. 666-695.

GLOWINSKI R., LIONS J.L., TREMOLIERES R. [1] *Analyse Numérique des Inéquations Variationnelles*, Vol. 1 and 2, Dunod-Bordas, Paris, 1976.
[2] *Numerical Analysis of Variational Inequalities*, North-Holland, Amsterdam, 1981.

GLOWINSKI R., MANTEL B., PERIAUX J. [1] Numerical solution of the time-dependent Navier-Stokes equations for incompressible viscous fluids by finite-element and alternating-direction methods, in *Numerical Methods in Aeronautical Fluid Dynamics*, P.L. Roe ed., Academic Press, London (1982), pp. 309 - 336.

GLOWINSKI R., MARROCCO A. [1] Sur l'approximation par éléments finis d'ordre un et la résolution par penalisation-dualite d'une classe de problèmes de Dirichlet non linéaires, *Revue Française d'Automatique Informatique Recherche Opérationnelle, Analyse numérique*, R-2, (1975), pp. 41-76.
[2] Sur l'approximation par éléments finis d'ordre 1, et la résolution par penalisation-dualité, d'une classe de problèmes de Dirichlet non linéaires, *C.R. Acad, Sc. Paris*, 278A, (1974), pp. 1649-1652
[3] On the solution of a class of non-linear Dirichlet problems by a penalty-duality method and finite-elements of order one, in *Optimization Techniques, IFIP Technical Conference*, G.I. Marchouk ed., Lecture Notes in Computer Sciences, Vol. 27, Springer-Verlag, Berlin 1975, pp. 327-333.
[4] *Sur l'approximation par éléments finis d'ordre un et la résolution par pénalisation-dualité d'une classe de problèmes de Dirichlet non linéaires*, Laboria Report 115, 1975.

[5] Numerical solution of two-dimensional magnetostatic problems by augmented Lagrangian methods, *Comp. Meth. Appl. Mech. Eng.*, 12, (1977), pp. 33-46.

GOLDSTEIN A.A. [1] Convex Programming in Hilbert Spaces, *Bull. A.M.S.*, 70, (1964), pp. 709-710.

GOLUB G.H. [1] Sparse matrix computations : eigenvalues and linear equations, in *Analyse et Cônstrole des Systèmes*, IRIA Seminars, 1974, pp. 117-140.

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

HIBBITT H.D., BECKER E.B., TAYLOR L.M. [1] Nonlinear analysis of some slender pipelines, *Comp. Meth. Appl. Mech. Eng.* 17/18, Part I, (1979), pp. 203-225.

HOUSEHOLDER A.S. [1] *The numerical treatment of a single nonlinear equation*, McGraw-Hill, New York, 1970.

JOHNSON C., THOMEE V. [1] Error estimates for a finite-element approximation of a minimal surface, *Math. Comp.*, 29, (1975), pp. 343-349.

JOURON, C. [1] Résolution numérique du problème des surfaces minima, *Arch. Ration. Mech. Analysis*, Vol. 59, No. 4, (1975), pp. 311-342.

KESAVAN S. [1] La méthode de Kikuchi appliquée aux équations de Von Karmann, *Numer. Math.*, 32, (1979), pp. 209-232.

KIKUCHI F. [1] Finite-element approximation of bifurcation problems, in *Functional Analysis and Numerical Analysis, Japan-France Seminar, Tokyo and Kyoto 1976*, H. Fujita ed., Japan Society for the Promotion of Science, 1978, pp. 203-222.

KORT B.V., BERTSEKAS D.P. [1] Combined primal-dual and penalty methods for convex programming, *Siam J. Control and Optimization*, 14, (1976), pp. 268-294.

LANDAU L., LIFCHITZ E. [1] *Mécanique des fluides*, Editions de Moscou, 1953.

LE TALLEC P. [1] *Numerical Analysis of Equilibrium Problems in Incompressible Nonlinear Elasticity*, Ph.D. Thesis, The University of Texas at Austin, 1980.

[2] *Les problèmes d'équilibre d'un corps hyperélastique incompressible en grandes déformations*, Thèse d'Etat, Université Pierre et Marie Curie, 1981.

LE TALLEC P., ODEN J.T. [1] Existence and characterization of hydrostatic pressure in finite deformations of incompressible elastic bodies, *Journal of Elasticity*, 11, (1981), 4,

LICHNEWSKY A. [1] Une méthode de gradient conjugué sur des variétés. Application à certains problèmes de valeurs propres non linéaires, *Num. Funct. Analysis and Optimization*, 1, (1979), 5, pp. 515-560.

LIONS J.L. [1] *Contrôle optimal des systèmes gouvernés par des équations aux dérivées partielles*, Dunod-Gauthier Villars, Paris, 1968.

[2] *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod-Gauthier Villars, Paris, 1969.

LIONS J.L., MAGENES E. [1] *Non-homogeneous boundary-value problems*, I, Springer-Verlag, New York, 1972.

LIONS J.L., STAMPACCHIA G. [1] Variational Inequalities, *Comm. Pure*

Applied Math., 20, (1967), pp. 493-519.

LIONS P.L., MERCIER B. [1] Splitting Algorithms for the Sum of two Nonlinear Operators, *SIAM J. Num. Anal.*, 16, (1979), pp. 964-979.

LYNESS J.N., JESPERSEN D. [1] Moderate-Degree Symmetric Quadrature Rules for the triangle, *J. Inst. Maths. Applics.*, 15, (1975), pp. 19-32.

MAIER G., ANDREUZZI F., GIANESSI F., JURINA L., TADDEI F. [1] Unilateral contact, elasto-plasticity and complementarity with reference to off-shore pipeline design, *Comp. Meth. Appl. Mech. Eng.*, 17/18, Part II, (1979), pp. 469-496.

MALKUS D.S., HUGHES T.J.R. [1] Mixed finite-element methods. Reduced and selective integration techniques: A unification of concepts, *Comp. Meth. Appl. Mech. Eng.*, 15, (1978), pp. 63-81.

MARCHOUK G.I., KUZNETSOV Y.A. [1] Méthodes itératives et fonctionnelles quadratiques, in *Sur les Methodes Numériques en Sciences Physiques et Economiques*, J.L. Lions, G.I. Marchouk eds., Dunod, Paris, 1974, pp. 1-132.

MARROCCO A. [1] *Expériences numériques sur des problèmes non linéaires résolus par éléments finis et lagrangien augmenté*, Laboria Report No. 309, May 1978.
[2] Analyse Numérique de problèmes d'Electrotechnique, *Ann. Sc. Math. Québec*, Vol. 1, No. 2, (1977), pp. 271-296.

MARTINET B. [1] Régularisation d'inéquations variationnelles par approximations successives, *Rev. Française Inf. Rech. Oper.*, (1970), pp. 154-159.
[2] Détermination approchée d'un point fixe d'une application pseudo-contractante, *C.R. Acad. Sci. Paris*, 274A, (1972), pp. 163-165.

MATTHIES H., STRANG G. [1] The solution of nonlinear finite-element equations, *Int. J. Num. Meth. Eng.*, 14 (1979), pp. 1613-1626.

MEIJERINK J.A., VAN DER VORST H.A. [1] An iterative solution method for linear systems of which the coefficients matrix is a symmetric M-matrix, *Math. of Comp.*, 31, (1977), pp. 148-162.

MERCIER B. [1] *Topics in finite-element solution of elliptic problems*, Lectures in Math. and Physics, Vol. 63, Tata Institute of Fundamental Research, Bombay, 1979 (distributed by Springer-Verlag, Berlin).
[2] Approximation par éléments finis et résolution par un algorithme de pénalisation-dualité d'un problème d'élasto-plasticité, *C.R. Acad. Sc. Paris*, 280A, (1975), pp. 287-290.
[3] *Sur La Théorie et l'Analyse Numérique de Problèmes de Plasticité*, Thèse d'Etat, Université Paris VI, 1977.

MINTY G.J. [1] Monotone (nonlinear) operators in Hilbert space, *Duke Math. J.*, 29, (1962), pp. 341-346.

MIYOSHI R. [1] A finite-element method for the solution of fourth-order partial differential equations, *Kunamoto J. Sci (Math.)*, 9, (1973), pp. 87-116.

MOREAU J.J. [1] Proximité et dualité dans un espace Hilbertien, *Bull. Soc. Math. France*, 93, (1968), pp. 273-299.

MOSCO U. [1] Dual Variational Inequalities, *J. Math. Anal. Appl.*, 40, (1972), pp. 202-206.

MUNRO E. [1] Some techniques and applications of the finite-element method for solving magnetic-field problems, *Proceedings of COMPUMAG Conference*, April 1976, Oxford, pp. 35-46.

NECAS J. [1] *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris, 1967.

ODEN J.T. [1] RIP - Methods for Stokesian flows, in *Finite Elements in Fluids*, Vol. 4, R.H. Gallagher, D.H. Norrie, J.T. Oden, O.C. Zienkiewicz eds., John Wiley, Chichester, (1982) pages 305 - 318.

OLSON M.D., TUANN S.Y. [1] New finite-element results for the square cavity, *Comp. Fluids*, 7, (1979), pp. 123-135.

OPIAL, Z. [1] Weak convergence of the successive approximation for non-expansive mappings in Banach Spaces, *Bull. A.M.S.*, 73, (1967), pp. 591-597.

ORTEGA J., RHEINBOLDT W.C. [1] *Iterative solution of nonlinear equations in several variables*, Academic Press, New York, 1970.

PAIGE C.C. [1] Bidiagonalization of matrices and solution of linear equations, *SIAM J. Num. Anal.*, 11, (1974), 1, pp. 197-209.

PARLETT B.N. [1] *The Symmetric Eigenvalue Problem*, Prentice Hall, Englewood Cliffs, New-Jersey, 1980.

PEACEMAN D.H., RACHFORD H.H. [1] The numerical solution of parabolic and elliptic differential equations, *SIAM J. Appl. Math.*, 3, (1955), pp. 24-41.

PELISSIER M.C. [1] *Sur quelques problèmes non linéaires en glaciologie*, Orsay Mathematical Publication, Université Paris-Sud, Department of Mathematics, Fasc. 75-24, No. 110, 1975.

POLAK E. [1] *Computational methods in Optimization*, Academic Press, New York, 1971.

POLYAK B.T. [1] On the Bertsekas method for minimization of composite functions, in *International Symposium on Systems Optimization and Analysis*, A. Bensoussan, J.L. Lions eds., Lecture Notes in Control and Information Sciences, Vol. 14, Springer-Verlag, Berlin, 1979, pp. 179-186.

POWELL M.J.D. [1] A method for nonlinear constraints in minimization problems in *Optimization*, R. Fletcher ed., Academic Press, London, 1969, Chapter 19.

[2] Variable metric methods for constrained optimization, in *Computing Methods in Applied Sciences and Engineering*, 1977, I, R. Glowinski, J.L. Lions eds., Lecture Notes in Mathematics, Vol. 704, Springer-Verlag, Berlin, 1979, pp. 62-72.

ROACHE P.J. [1] *Computational Fluid Dynamics*, Hermosa Publishers, Albuquerque, N.M., 1972.

ROCKAFELLAR R.T. [1] The multiplier method of Hestenes and Powell applied to convex programming, *J. Opt. Theory Appl.*, 12, (1973), pp. 555-562.

[2] Augmented Lagrangians and applications to the proximal-point algorithm in convex programming, *Math. Op. Research*, 1, (1976), pp. 97-116.

[3] Augmented Lagrange multiplier functions and duality in non-convex programming, *Siam J. Control*, Vol. 12, No. 2, (1974), pp. 268-285.

[4] *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.

[5] Extension of Fenchel's duality theorem, *Duke Math. J.*, 33, (1966), pp. 81-89.

[6] On the Maximality of sums of nonlinear monotone operators, *Trans. Amer. Math. Soc.*, (1970), pp. 75-88.

[7] A dual approach to solving nonlinear programming problems by unconstrained optimization, *Math. Prog.*, 5, (1973), pp. 354-373.

[8] Monotone operators and the proximal-point algorithm, *SIAM J. Control and Optimization*, 14, (1976), pp. 877-898.

RUAS V. [1] A class of asymmetric simplicial finite-element methods for solving incompressible elasticity problems, *Comp. Meth. Appl. Mech. Eng.*, 27, (1981), 3, pp. 319-343.

SEGAL A. [1] On the numerical solution of Stokes equations using the finite element method, *Comp. Meth. Appl. Mech. Eng.*, 19, (1979), pp. 165-185.

SOULIE M. [1] Private communication.

STEWART G.W. [1] *Introduction to Matrix Computations*, Acad. Press, New York, 1973.

STRANG, G., FIX G. [1] *An analysis of the finite-element method*, Prentice Hall, Englewood Cliffs, N.J., 1973.

TARTAR L. [1] *Topics in Nonlinear Analysis*, Orsay Math. Publication, Université Paris-Sud, Department of Mathematics, 1978.

TAYLOR R.L., ZIENKIEWICZ O.C. [1] Complementarity Energy with Penalt' Functions in Finite-Element Analysis, Chapter 8 of *Energy Methods in Finite Element Analysis*, R. Glowinski, E.Y. Rodin, O.C. Zienkiewicz eds., J. Wiley and Sons, Chichester, 1979, pp. 153-174.

TEMAM R. [1] *Navier-Stokes equations*, North-Holland, Amsterdam, 1977

THOMAS J.M. [1] *Sur l'analyse numérique des méthodes d'éléments finis hybrides et mixtes*, Thèse d'Etat, Université Pierre et Marie Curie, 1977.

THOMASSET F. [1] *Implementation of Finite Element Methods for Navier-Stokes equations*, Springer-Verlag, New York, 1981.

WHEELER M.F. [1] An elliptic Collocation/Finite-Element Method with Interior Penalties, *SIAM J. Num. Anal.*, 15, (1978), pp. 152-161.

WIDLUND O. [1] A Lanczos method for a class of nonsymmetric systems of linear equations, *Siam J. Num. Anal.*, 15, (1978), 4, pp. 801-812.

WILKINSON J.H. [1] *The algebraic Eigenvalue Problem*, Oxford University Press, 1965.

WILKINSON J.H., REINSCH C. [1] *Handbook for Automatic Computation, Vol. II: Linear Algebra*, Springer-Verlag, Berlin, 1971.

YAMAGUTI M., FUJII H. [1] On numerical deformation of singularities in nonlinear elasticity, in *Computing Methods in Applied Sciences and Engineering*, Part I, R. Glowinski, J.L. Lions eds., 1977, Lecture Notes in Mathematics, Vol. 704, Springer-Verlag, Berlin, 1979, pp. 267-281.