Lectures on

Optimization – Theory and Algorithms

By

Jean Cea

Tata Institute of Fundamental Research, Bombay
1978

Lectures on

Optimization – Theory and Algorithms

By

John Cea

Notes by

M. K. V. Murthy

Published for the

Tata Institute of Fundamental Research, Bombay
1978

ISBN 3-540-08850-4 Springer-Verlag Berlin, Heidelberg. New York
ISBN 0-387-08850-4 Springer-Verlag New York, Heidelberg, Berlin

Printed by K. P. Puthran at the Tata Press Limited,
414 Veer Savarkar Marg, Bombay 400 025 and published by H. Goetze,
Springer-Verlag, Heidelberg, West Germany

PRINTED IN INDIA

Contents

Chapter 1

Differential Calculus in Normed Linear Spaces

We shall recall in this chapter the notions of differentiability in the sense of Gateaux and Frechet for mappings between normed linear spaces and some of the properties of derivatives in relation to convexity and weak lower semi-continuity of functionals on normed linear spaces. We shall use these concepts throughout our discussions. 1

In the following all the vector spaces considered will be over the field of *real numbers* $\mathbb{R}$.

If V is a normed (vector) space we shall denote by $|| \cdot ||_V$ the norm in V, by V' its (strong) dual with $|| \cdot ||_{V'}$ as the norm and by $\langle \cdot, \cdot \rangle_{V' \times V}$ the duality pairing between V and V'. If V is a Hilbert space then $(\cdot, \cdot)_V$ will denote the inner product in V. If V and H are two normed spaces then $\mathscr{L}(V, H)$ denotes the vector space of all continuous linear mappings from V into H provided with the norm $A \to ||A||_{\mathscr{L}(V,H)} = \sup\{||Av||_H / ||v||_V, v \epsilon V\}$.

1 Gateaux Derivatives

Let V, H be normed spaces and $A : U \subset V \to H$ be a mapping of an open subset U of V into H. We shall often call a vector $\varphi \epsilon V$, $\varphi \neq 0$ a direction in V.

Definition 1.1. The mapping A is said to be differentiable in the sense of Gateaux or simply G-differentiable at a point $u\epsilon U$ in the direction φ if the difference quotient

$$(A(u + \theta\varphi) - A(u))/\theta$$

2 has a limit $A'(u, \varphi)$ in H as $\theta \to 0$ in $\mathbb{R}$. The (unique) limit $A'(u, \varphi)$ is called the Gateaux derivative of A at u in the direction φ.

A is said to be G-differentiable in a direction φ in a subset of U if it is G-differentiable at every point of the subset in the direction φ.

We shall simply call $A'(u, \varphi)$ the G-derivative of A at u since the dependence on φ is clear from the notation.

Remark 1.1. The operator $V \ni \varphi \mapsto A'(u, \varphi)\epsilon H$ is homogeneous:

$$A'(u, \alpha, \varphi) = \alpha A'(u, \varphi) \text{ for } \alpha > 0.$$

In fact,

$$A'(u, \alpha, \varphi) = \lim_{\theta\to 0}(A(u+\alpha\theta\varphi)-A(u))/\theta = \alpha \lim_{\lambda\to 0}(A(u+\lambda\varphi))/\lambda = \alpha A'(u, \varphi).$$

However, this operator is not, in general, linear as can be seen immediatly from Example 1.2 below.

We shall often denote a functional on U by J.

Remark 1.2. Every lineary functional $L : V \to \mathbb{R}$ is G-differentiable everywhere in V in all directions and its G-derivative is

$$L'(u, \varphi) = L(\varphi)$$

since $(L(u + \theta\varphi) - L(u))/\theta = L(\varphi)$. It is a constant functional (i.e. independent of u in V).

If a $(u, v) : V \times V \to \mathbb{R}$ is a bilinear functional on V then the functional $J : V \ni v \mapsto J(v) = a(v, v)\epsilon\mathbb{R}$ is G-differentiable everywhere in all direction and

$$J'(u, \varphi) = a(u, \varphi) + a(\varphi, u).$$

3 If further $a(u,v)$ is symmetric (i.e. $a(u,v) = a(v,u)$ for all $u, v\epsilon V$) then $J'(u,\varphi) = 2a(u,\varphi)$. This follows immediately from bilinearity :

$$a(u+\theta, u+\theta\varphi) = a(u,u) + \theta(a(u,\varphi) + a(\varphi,u)) + \theta^2 a(\varphi,\varphi)$$

so that

$$J'(u,\varphi) = \lim_{\theta\to 0}(J(u+\theta\varphi) - J(u))/\theta = a(u,\varphi) + a(\varphi,u).$$

The following example will be a model case of linear problems in many of our discussions in the following chapters.

Example 1.1. Let $(u,v) \mapsto a(u,v)$ be a symmetric bi-linear form on a Hilbert space V and $v \mapsto L(v)a$ linear form on V. Define the functional $J : V \to \mathbb{R}$ by

$$J(v) = \frac{1}{2}a(v,v) - L(v).$$

It follows from the above Remark that J is G-differentiable everywhere in V in all directions φ and

$$J'(u,\varphi) = a(u,\varphi) - L(\varphi).$$

In many of the questions we shall assume:

(i) $a(.,.)$ is $(bi-)$ continuous: there exists a constant $M > 0$ such that

$$a(u,v) \leq M||u||_V||v||_V \text{ for all } u, v\epsilon V;$$

(ii) $a(\cdot,\cdot)$ is V-coercive; There exists a constant $\alpha > 0$ such that

$$a(v,v) \geq \alpha||v||_V^2 \text{ for all } v\epsilon V$$

and

(iii) L is continuous: there exists a constant $N > 0$ such that

$$L(v) \leq N||v||_V \text{ for all } v\epsilon V.$$

4

Example 1.2. The function $f : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$f(x,y) = \begin{cases} 0 & \text{if } (x,y) = (0,0) \\ x^5/((x-y)^2 + x^4) & \text{if } (x,y) \neq (0,0) \end{cases}$$

is G-differentiable everywhere and in all directions. In fact, if $u = (0,0)\epsilon\mathbb{R}^2$ then given a direction $\varphi = (X,Y)\epsilon\mathbb{R}^2(\varphi \neq 0)$ we have

$$(f(\theta X, \theta Y) - f(0,0))/\theta = \theta^2 X^5/((X-Y)^2 + \theta^2 X^4)$$

which has a limit as $\theta \to 0$ and we have

$$f'(u,\varphi) = f'((0,0),(X,Y)) = \begin{cases} 0 & \text{if } X \neq Y \\ X & \text{if } X = Y \end{cases}$$

One can also check easily that f is G-differentiable in $\mathbb{R}^2$.

The following will be the general abstract form of functionals in amy of the non-linear problems that we shall consider.

Example 1.3. Let Ω be an open set in $\mathbb{R}^n$ and $V = L^p(\Omega)$, $p > 1$. Suppose $g : \mathbb{R}^1 \ni t \mapsto g(t)\epsilon\mathbb{R}^1$ be a C^1-function such that

$$(i) \quad |g(t)| \leq C|t|^p \text{ and } (ii) \quad |g'(t)| \leq C|t|^{p-1}$$

for some constant $C > 0$. Then

$$u \mapsto J(u) = \int_\Omega g(u(x))dx$$

defines a functional J on $L^p(\Omega) = V$ which is G-differentiable everywhere in all directions and we have

$$J'(u,\varphi) = \int_\Omega g'(u(x))\varphi(x)dx.$$

5

(The right hand side here exists for any u, $\varphi\epsilon L^p(\Omega)$).

In fact, since $u\epsilon L^p(\Omega)$ and since g satisfies (i) we have

$$|J(u)| \leq \int_\Omega |g(u)|dx \leq C\int_\Omega |u|^p dx < +\infty$$

which means J is well defined on $L^p(\Omega)$. On the other hand, for any $u\epsilon L^p(\Omega)$ since g' satisfies (ii), $g'(u)\epsilon L^{p'}(\Omega)$ where $p^{-1} + p'^{-1} = 1$. For, we have

$$\int_\omega |g'(u)|^{p'} dx \leq C\int_\Omega |u|^{(p-1)p'} dx = C\int_\Omega |u|^p dx < +\infty.$$

Hence, for any $u, \varphi\epsilon L^p(\Omega)$, we have by Hölder's inequality

$$\left|\int_\omega g'(u)\varphi dx\right| \leq ||g'(u)||_{L^p(\Omega)}||\varphi||_{L^p(\Omega)} \leq C||u||^{p/p'}_{L^p}||\varphi||_{L^p(\Omega)} < +\infty.$$

To compute $J'(u, \varphi)$, if $\theta\epsilon\mathbb{R}$ we define $h : [0, 1] \mapsto \mathbb{R}$ by setting

$$h(t) = g(u + t\theta\varphi).$$

Then $h\epsilon C^1(0, 1)$ and

$$h(1) - h(0) = \int_0^1 h'(t)dt = \theta\varphi(x)\int_0^1 g'(u + t\theta\varphi)dt$$

$(t = t(x))$, $|t(x)| \leq 1$ so that

$$(J(u + \theta\varphi) - J(u))/\theta = \int_\Omega \varphi(x)\int_0^1 g'(u(x) + t\theta\varphi(x))dtdx.$$

One can easily check as above that the function

$$(x, t) \mapsto \varphi(x)g'(u(x) + t\theta\varphi(x))$$

belongs to $L^1(\Omega \times [0, 1])$ and hence by Fubini's theorem

$$(J(u + \theta\varphi) - J(u))/\theta = \int_0^1 dt\int_\Omega \varphi(x)g'(u(x) + t\theta\varphi(x))dx.$$

Here the continuity of g' implies that

$$g'(u + t\theta\varphi) \to g'(u) \text{ as } \theta \to 0 \text{ (and hence as } t\theta \to 0)$$

uniformly for $t\epsilon[0, 1]$. Morever, the condition (ii) together with triangle inequality implies that, for $0 < \theta \leq 1$.

$$|\varphi(x)g'(u(x) + t\theta\varphi(x))| \leq C|\varphi(x)|(|u(x)| + |\varphi(x)|)^{p-1}$$

and the right side is integrable by Hölder's inequality. Then by dominated convergence theorem we conclude that

$$J'(u, \varphi) = \int_\Omega g'(u)\varphi dx.$$

Definition 1.2. An operator $A : U \subset V \to H$ (U being an open set in V) is said to be twice differentiable in the sense of Gateaux at a point $u\epsilon V$ in the directions φ, $\psi(\varphi, \psi\epsilon V, \varphi \neq 0, \psi \neq 0$ given) if the operator $u \mapsto A'(u, \varphi); U \subset V \to H$ is once G-differentiable at u in the direction ψ. The G-derivative of $u \mapsto A'(u, \varphi)$ is called the second G-derivative of A and is denoted by $A''(u, \varphi, \psi)\epsilon H$.

$$\text{i.e. } A''(u; \varphi, \psi) = \lim_{\theta\to 0}(A'(u + \theta\psi, \varphi) - A'(u, \varphi))/\theta.$$

Remark 1.3. Derivatives of higher orders in the sense of Gateaux can be defined in the same way. As we shall not use derivatives of higher orders in the following we shall not consider their properties.

Now let $J : U \subset V \to \mathbb{R}$ be a functional on an open set of a normed linear space V which is once G-differentiable at a point $u\epsilon U$. If the functional $\varphi \mapsto J'(u, \varphi)$ is continuous linear on V then there exists a (unique) element $G(u)\epsilon V'$ such that **7**

$$J'(u, \varphi) = \langle G(u), \varphi\rangle_{V'\times V} \text{ for all } \varphi\epsilon V.$$

Similarly, if J is twice G-differentiable at a point $u\epsilon U$ and if the form $(\varphi, \psi) \mapsto J''(u : \varphi, \psi)$ is a bilinear (bi-)continuous form on $V \times V$ then there exists a (unique) element $H(u)\epsilon\mathscr{L}(V, V')$ such that

$$J''(u; \varphi, \psi) = \langle H(u)\varphi, \psi\rangle_{V'\times V}.$$

Definition 1.3. $G(u)\epsilon V'$ is called the gradient of J at u and $H(u)\epsilon\mathscr{L}$ (V, V') is called the Hessian of J at u.

2 Taylor's Formula

We shall next deduce the mean value theorem and Taylor's formula of second order for a mapping $A : U \subset V \to H$ (U open subset of a normed linear space V) in terms of the G-derivatives of A. We shall begin case of functionals on a normed linear space V.

Let J be a functional defined on an open set U in a normed linear space V and $u, \varphi\epsilon V, \varphi \neq 0$ be given. Throughout this section we assume that the set $\{u + \theta\varphi; \theta\epsilon[0, 1]\}$ is contained in U. It is convenient to introduce the function $f : [0, 1] \to \mathbb{R}$ by setting

$$\theta \to f(\theta) = J(u + \theta\varphi).$$

We observe that if $J'(u + \theta\varphi, \varphi)$ exists then f is once differentiable in $]0, 1[$ and, as one can check immediately

$$f'(\theta) = J'(u + \theta\varphi, \varphi).$$

8

Similarly if $J''(u + \theta\varphi, \varphi, \varphi)$ exists then f is twice differentiable and

$$f''(\theta) = J''(u + \theta\varphi; \varphi, \varphi).$$

Proposition 2.1. *Let J be a functional on an open set U of a normed space V and $u\epsilon U$, $\varphi\epsilon V$ be given. If $\{u + \theta\varphi; \theta\epsilon[0, 1]\}\epsilon U$ and J is once G-differentiable on this set in the direction φ then there exists a $\theta_0\epsilon]0, 1[$ such that*

$$J(u + \varphi) = J(u) + J'(u + \theta_0\varphi, \varphi) \tag{2.1}$$

Proof. This follows immediately from the classical mean value theorem applied to the function f on $[0, 1]$: thete exists a $\theta_0\epsilon]0, 1[$ such that

$$f(1) = f(0) + 1 - f'(\theta_0)$$

which is noting nut (2.1). □

Proposition 2.2. *Let U be as in Proposition 2.1. If J is twice G - differentiable on the set $\{u+\theta\varphi; \theta\epsilon[0,1]\}$ in the directions φ,φ then there exists a $\theta_0\epsilon]0,1[$ such that*

$$J(u+\varphi) = J(u) + J'(u,\varphi) + \frac{1}{2}J''(u+\theta_0\varphi;\varphi,\varphi). \tag{2.2}$$

This again follows from the classical Taylor's formula applied to the function f on $[0,1]$.

Remark 2.1. If $L : V \to \mathbb{R}$ is a linear functional on V then by Remark 1.1 is G-differentiable everywhere in all directions and we find that the formula (2.1) reads

$$L(u+\varphi) = L(u) + L(\varphi)$$

9 which is noting but additivity of L.

Similarly, if $a(\cdot,\cdot)$ is a bi-linear form on V then the functional $J(v) = a(v,v)$ on V is twice G-differentiable in all pairs directions (φ,ψ) and

$$J'(u,\varphi) = a(u,\varphi) + a(\varphi,u), J''(u,\varphi,\psi) = a(\psi,\varphi) + a(\varphi,\psi).$$

Then the Taylor's formula (2.2) in this case reads

$$a(u+\varphi,u+\varphi) = a(u,u) + a(u,\varphi) + a(\varphi,u) + a(\varphi,\varphi)$$

which is noting but the bilinearity of a.

These two facts together imply that the functional

$$J(v) = \frac{1}{2}a(v,v) - L(v)$$

of Example 1.1 admits a Taylor expansion of the form (Proposition 2.2)

$$J(u+\varphi) = J(u) + a(u,\varphi) - L(\varphi) + \frac{1}{2}a(\varphi,\varphi).$$

We shall now pass to the case of general operators between normed spaces. We remark first of all that the Taylor's formula in the form (2.1) is not in general valid in this case. However, we have

Proposition 2.3. *Let V, H be two normed spaces, U an open subset of V and let $\varphi\epsilon V$ be given. If the set $\{u+\theta\varphi;\theta\epsilon[0,1]\}\subset U$ and $A:U\subset V\to H$ is a mapping which is G-differentiable everywhere on the set $\{u+\theta\varphi;\theta\epsilon[0,1]\}$ in the direction φ then, for any $g\epsilon H'$, there exists a $\theta_g\epsilon]0,1[$ such that*

$$\langle g, A(u+\varphi)\rangle_{H'\times H} = \langle g, A(u)\rangle_{H'\times H} + \langle g, A'(u+\theta_g\varphi,\varphi)\rangle_{H'\times H} \tag{2.3}$$

Proof. We define a function $f:[0,1]\to\mathbb{R}$ by setting

$$\theta' \mapsto f(\theta) = \langle g, A(u+\theta\varphi)\rangle_{H'\times H}.$$

□ **10**

Then $f'(\theta)$ exists in $]0,1[$ and

$$f'(\theta) = \langle g, A'(u+\theta\varphi,\varphi)\rangle_{H'\times H} \text{ for } \theta\epsilon]0,1[$$

Now (2.3) follows immediatly on applying the classical mean value theorem to the function f.

Proposition 2.4. *Let V, H, u, φ and U be as in Proposition 2.4. If $A:U\subset V\to H$ is G-differentiable in the set $\{u+\theta\varphi;\theta\epsilon[0,1]\}$ in the direction φ then there exists a $\theta_0\epsilon]0,1[$ such that*

$$\|A(u+\varphi)-A(u)\|_H \le \|A'(u+\theta_0\varphi,\varphi)\|_H. \tag{2.4}$$

The proof of this proposition uses the following Lemma which is a corollary to Hahn-Banach theorem.

Lemma 2.1. *If H is normed space then for any $v\in H$ there exists a $g\epsilon H'$ such that*

$$\|g\|_{H'} = 1 \text{ and } \|v\|_H = \langle g, v\rangle_{H\times H}. \tag{2.5}$$

For a proof see [34].

Proof of Proposition 2.4 The element $v = A(u+\varphi) - A(u)$ belongs to H and let $g \epsilon H'$ be an element given by the Lemma 2.1 satisfying (2.5) i.e.

$$||g||_{H'} = 1, ||A(u+\varphi) - A(u)||_H =< g, A(u+\varphi) - A(u) >_{H' \times H} .$$

Since A satisfies the assumptions of Proposition 2.3 it follows that there exists a $\theta_0 = \theta_g \epsilon]0, 1[$ such that

$$\begin{aligned} ||A(u+\varphi) - A(u)||_H &=< g, A(u+\varphi) - A(u) >_{H' \times H} \\ &=< g, A'(u+\theta_0\varphi, \varphi) >_{H' \times H} \\ &\leq ||g||_{H'} ||A'(u+\theta_0\varphi, \varphi)||_H = ||A'(u+\theta_0\varphi, \varphi)||_H . \end{aligned}$$

11 proving (2.4).

Proposition 2.5. *Suppose a functional $J : V \to \mathbb{R}$ has a gradient $G(u)$ for all $u \epsilon V$ which is bounded i.e. there exists a constant $M > 0$ such that $||G(u)|| \leq M$ for all $u \epsilon V$, then we have*

(2.6) $|J(u) - J(v)| \leq M||u - v||_V$ *for all $u, v \epsilon V$.*

Proof. If $u, v, \epsilon V$ then taking $\varphi = v - u$ in Proposition 2.1 we can write, with some $\theta_0 \epsilon]0, 1[$,

$$\begin{aligned} J(v) - J(u) &= J'(u + \theta_0(v-u), v-u) \\ &=< G(u + \theta_0(v-u)), v-u >_{V' \times V} \end{aligned}$$

and hence

$$|J(v) - J(u)| \leq ||G(u + \theta_0(v-u))||_{V'} ||v-u||_V \leq M||v-u||_V .$$

□

3 Convexity and Gateaux Differentiability

A subset U of a vector space V is convex if whenever $u, v \epsilon U$ the segment $\{(1-\theta)u + \theta v, \theta \epsilon [0, 1]\}$ joining u and v lies in U.

Definition 3.1. A functional $J : U \subset V \to \mathbb{R}$ on a convex set U of a vector space V is said to be convex if

(3.1) $J((1-\theta)u+\theta v) \leq (1-\theta)J(u)+\theta J(v)$ for all $u, v\epsilon U$ and $\theta\epsilon[0,1]$.

J is said to be strictly convex if strict inequality holds for all $u, v\epsilon V$ with $u \neq v$ and $\theta\epsilon]0,1[$.

We can write the inequality (3.1) in the above definition in the equivalent form

(3.1)′ $J(u+\theta(v-u)) \leq J(u)+\theta(J(v)-J(u))$ for all $u, v\epsilon U$ and $\theta\epsilon[0,1]$.

12

The following propositions relate the convexity of functionals with the properties of their G-differentiability

Proposition 3.1. *If a function $J : U \subset V \to \mathbb{R}$ on an open convex set is G-differentiable everywhere in U in all directions then*

(1) J is convex if and only if

$$J(v) \geq J(u) + J'(u, v-u) \textit{ for all } u, v\epsilon U.$$

(2) J is strictly convex if and only if

$$J(v) > J(u) + J'(u, v-u) \textit{for all } u, v\epsilon U \textit{ with } u \neq v.$$

Proof. (1) If J is convex then we can write

$$J(v) - J(u) \geq (J(u+\theta(v-u)) - J(u))/\theta \text{ for all } \theta\epsilon[0,1].$$

Now since $J'(u, v-u)$ exists the right side tends to $J'(u, v-u)$ as $\theta \to 0$. Thus taking limits as $\theta \to 0$ in this inequality the required inequality is obtained.

The proof of the converse assertion follows the usual proof in the case of functions. Let $u, v\epsilon V$ and $\theta\epsilon[0,1]$. We have

$$J(u) \geq J(u+\theta(v-u)) + J'(u+\theta(v-u)), u(u+\theta(v-u))$$

$$= J(u+\theta(v-u)) - \theta J'(u+\theta(v-u), v-u)$$

by the homogeneity of the mapping $\varphi \mapsto J'(w,\varphi)$ and

$$\begin{aligned} J(v) &\geq J(u+\theta(v-u)) + J'(u+\theta(v-u), v-(u+\theta(v-u))) \\ &= J(u+\theta(v-u)) + (1-\theta)J'(u+\theta(v-u), v-u). \end{aligned}$$

Multiplying the two inequalities respectively by $(1-\theta)$ and θ, and adding we obtain

$$(1-\theta)J(u) + \theta J(v) \geq J(u+\theta(v-u)),$$

13 thus proving the convexity of J.

(2) If J is strictly convex we can, first of all, write

$$J(v) - J(u) > \theta^{-1}[J(u+\theta(v-u)) - J(u)].$$

(Here we have used the inequality ((3.1)′)). On the other hand, using part (1) of the proposition we have

$$J(u+\theta(v-u)) - J(u) = J'(u, \theta(v-u)).$$

Since, by Remark 1.1 of Chapter 1, J is homogeneous in its second argument: i.e.

$$J'(u,\theta(v-u)) = \theta J'(u, v-u).$$

□

This together with the first inequality implies (2). The converse implication is proved exactly in the same way as in the first part.

Proposition 3.2. *If a functional $J : U \subset V \to \mathbb{R}$ on an open convex set of a normed space V is twice G-differentiable everywhere in U and in all directions and if the form $(\varphi,\psi) \mapsto J''(u;\varphi,\psi)$ is positive semi-definite t. e. if*

$$J''(u:\varphi,\varphi) \geq 0 \textit{ for all } u\epsilon U \textit{ and } \varphi\epsilon V \textit{ with } \varphi \neq 0$$

then J is convex.

If the form $(\varphi, \psi) \mapsto J''(u : \varphi, \psi)$ is positive definite i.e. if

$$J''(u; \varphi, \varphi) > 0 \text{ for all } u \epsilon U \text{ and } \varphi \epsilon V \text{ with } \varphi \neq 0$$

then J is strictly convex.

Proof. Since U is convex the set $\{u + \theta(v - u), \theta \epsilon [0, 1]\}$ is contained in U whenever $u, v \epsilon U$. Then by Taylor's formula (Proposition 2.2) we have, with $\varphi = v - u$.

$$J(v) = J(u) + J'(u, v - u) + \frac{1}{2} J''(u + \theta_0(v - u), v - u, v - u)$$

for some $\theta_0 \epsilon]0, 1[$. Then the positive semi-definitensess of J'' implies

$$J(v) \geq J(u) + J'(u, v - u)$$

from which convexity of J follows from (1) of Proposition 3.1. Similarly the strict convexity of J from positive definiteness of J'' follows on application of (2) Proposition 3.1. □ 14

Now consider the function $J : V \to \mathbb{R}$:

$$J(v) = \frac{1}{2} a(v.v) - L(v)$$

of Example 1.1. We have seen that J twice G-differentiable and $J''(u : \varphi.\varphi) = a(\varphi, \varphi)$. Applying Proposition 3.2 we get the

Corollary 3.1. *Under the assumptions of Example 1.1 J is convex (resp. strictly convex) if a (φ, ψ) is positive semi-definite (resp. positive definite). i.e.*

J is convex if $a(\varphi, \varphi) \geq 0$ for all $\varphi \epsilon V$ (resp. J is strictly convex if $a(\varphi, \varphi) > 0$ for all $\varphi \epsilon V$ with $\varphi \neq 0$).

In particular, if $a(\cdot, \cdot)$ is V-coercive then J is strictly convex.

4 Gateaux Differentiability and Weak Lower Semi-Continuity

Let V be a normed vector space. We use the standard notation "$v_n \rightharpoonup u''$ to denote weak convergence of a sequence v_n in V to u. i.e. For any $g \epsilon V'$ we have

$$< g, v_n >_{V' \times V} \to < g, u >_{V' \times V.}$$

Definition 4.1. A functional $J : V \to \mathbb{R}$ is said to be weakly lower semi-continuous if for every sequence $v_n \rightharpoonup u$ in V we have

$$\liminf_{n \to \infty} J(v_n) \geq J(u).$$

Remark 4.1. The notion of weak lower semi-continuity is a local property. The Definition 4.1 and the propositions below can be stated for functionals J defined on an open subset U of V with minor changes. We shall leave these to the reader. 15

Proposition 4.1. *If a functional $J : V \to \mathbb{R}$ is convex and admits a gradient $G(u) \epsilon V'$ at every point $u \epsilon V$ then J is weakly lower semi-continuous.*

Proof. Let v_n be a sequence in V such that $v_n \rightharpoonup u$ in V. Then $< G(u), v_n - u >_{V' \times V} \to 0$. On the other hand, since J is convex we have, by Proposition 3.1,

$$J(v_n) \geq J(u) + < G(u), v_n - u >$$

from which on taking limits we obtain

$$\liminf_{n \to \infty} . J(v_n) \geq J(u).$$

□

Proposition 4.2. *If a functional $J : V \to \mathbb{R}$ is twice G-differentiable everywhere in V in all directions and satisfies*

(i) J has a gradient $G(u) \epsilon V'$ at all points $u \epsilon V$.

(ii) $(\varphi,\psi) \mapsto J''(u;\varphi,\psi)$ *is positive semi-definite, i.e.* $J''(u;\varphi,\varphi) \geq 0$ *for all* $u, \varphi \epsilon V$ *with* $\varphi \neq 0$*,*

then J is weakly lower semi-continuous.

Proof. By Proposition 3.2 the condition (ii) implies that J is convex. Then the assertion follows from Proposition 4.1. □

We now apply Proposition 4.2 to the functional

$$v \mapsto J(v) = \frac{1}{2}a(v,v) - L(v)$$

of Example 1.1. We know that it has a gradient

$$G(u) : \varphi \mapsto < G(u), \varphi >= a(u,\varphi) - L(\varphi)$$

and $J''(u;\varphi,\varphi) = a(\varphi,\varphi)$ for all $u, \varphi \epsilon V$. 16

If further we assume that $a(\cdot,\cdot)$ is V-coercive, i.e. there exists an $\alpha > 0$ such that

$$(J''(u;\varphi,\varphi) =)a(\varphi,\varphi) \geq \alpha \|\varphi\|_V^2 (\geq 0) \text{ for all } \varphi \epsilon V$$

then by Proposition 4.2 we conclude that J is weakly lower semi - continuous.

5 Commutation of Derivations

We shall admit without proof the following useful result on commutativity of the order of derivations.

Theorem 5.1. *Let U be an open set in a normed vector space V and* $J : U \subset V \to \mathbb{R}$ *be a functional on U. If*

(i) $J''(u;\varphi,\psi)$ *exists everywhere in U in all directions* $\varphi, \psi \epsilon V$*, and*

(ii) for every pair $\varphi, \psi \epsilon V$ *the form* $u \mapsto J''(u,\varphi,\psi)$ *is continuous*

then we have

$$J''(u,\varphi,\psi) = J''(u;\psi,\varphi) \textit{ for all } \varphi,\psi\epsilon V.$$

For a proof we refer to [12].

As a consequence we deduce the

Corollary 5.1. *If a functional $J : U \subset V \to \mathbb{R}$ on an open set of a normed vector space V admits a Hessian $H(u) \in \mathscr{L}(V,V')$ at every points $u \in U$ and if the mapping $U \ni u \mapsto H(u) \in \mathscr{L}(V,V')$ is continuous then $H(u)$ is self adjoint.*

$$\textit{i.e.} \; < H(u)\varphi,\psi >_{V'\times V} = < H(u)\psi,\varphi >_{V'\times V} \; \textit{for all } \varphi,\psi \in V.$$

6 Frechet Derivatives

Let V and H be two normed vector spaces.

Definition 6.1. A mapping $A : U \subset V \to H$ from an open set U in V to H is said to be Fŕechet differentiable (or simply F-differentiable) at a point $u \in U$ if there exists a continuous linear mapping $A'(u) : V \to H$, i.e. $A'(u) \in \mathscr{L}(V,H)$ such that

$$\lim_{\varphi\to 0} \|A(u+\varphi) - A(u) - A'(u)\varphi\|/\|\varphi\| = 0. \tag{6.1}$$

17 Clearly, $A'(u)$, if it exists, is unique and is called the Fréchet derivative (F-derivative) of A at u.

We can, equivalently, sat that a mapping $A : U \subset V \to H$ is F-differentiable at a point $u \in U$ if there exists an element $A'(u) \in \mathscr{L}(V;H)$ such that

$$A(u+\varphi) = A(u) + A'(u)\varphi + \|\varphi\|_V \in (u,\varphi) \text{ where } \in (u,\varphi) \in H \text{ and}$$

(6.2)

$$\in (u,\varphi) \to 0 \text{ in } H \text{ as } \varphi \to 0 \text{ in } V.$$

Example 6.1. If f is a function defined in an open set $U \subset \mathbb{R}^2$, i.e. $f : U \to \mathbb{R}$, then it is F-differentiable if it is once differentiable in the usual sense and

$$f'(u) = grad\, f(u) = (\partial f/\partial x_1(u), \partial f/\partial x_2(u)) \in \mathscr{L}(\mathbb{R}^2, \mathbb{R}).$$

Example 6.2. In the case of the functional

$$v \mapsto J(v) = \frac{1}{2}a(v, v) - L(v)$$

Of Example 1.1 where (i) and (iii) are satisfied on a Hilbert space V we easily check that J is F-differentiable everywhere in V and its F-derivative isgiven by

$$\varphi \mapsto J'(u)\varphi = a(u, \varphi) - L(\varphi).$$

In fact, by (i) and (iii) of Example 1.1 $J'(u) \in V'$ since $\varphi \mapsto a(u, \varphi)$ **18** and $\varphi \mapsto L(\varphi)$ are continuous linear and we have

$$J(u + \varphi) - J(u) - [a(u, \varphi) - L(\varphi)] = a(\varphi, \varphi) = ||\varphi||_V \in (u, \varphi)$$

where $\in (u, \varphi) = ||\varphi||_V^{-1} a(\varphi, \varphi)$ and

$$0 \leq \in (u, \varphi) \leq M||\varphi||_V$$

so that $\in (u, \varphi) \to 0$ as $\varphi \to 0$ in V.

We observe that in this case the F-derivative of J is the same as the gradient of J.

Remark 6.1. If an operator $A : U \subset V \to H$ is F-differentiable then it is also G-differentiable and its G-derivative coincides with its F-derivative. In fact, let A be F-differentiable with $A'(u)$ as its F-derivative. Then, for $u \in U, \varphi \in V, \varphi \neq 0$, writting $\psi = \rho\varphi$ we have $\psi \to 0$ in V as $\rho \to 0$ and

$$\begin{aligned}
&\rho^{-1}(A(u + \rho\varphi) - A(u) - A'(u)\varphi) \\
&= \rho^{-1}(A(u + \psi) - A(u) - A'(u)\psi) \text{ since } A'(u) \text{ is linear} \\
&= \rho^{-1}||\psi|| \in (u, \psi) = ||\varphi|| \in (u, \psi) \to 0 \text{ in } H \text{ as } \psi \to 0 \text{ in } H \text{ i.e. as} \rho \to 0.
\end{aligned}$$

Remark 6.2. However, in general, the converse is not true. Example 1.2 shows that the function f has a G-derivative but not F-differentiable. We also note that the G-derivative need not be a linear map of V into H (as in Example 1.2) while the F-derivative is necessarily linear by definition and belongs to $\mathscr{L}(V, H)$.

Remark 6.3. The notions of F-differentiability of higher orders and the corresponding F-derivatives can be defined in an obvious manner. Since, whenever we have F-differentiability we also have G - differentiability the Taylor's formula and hence all its consequences remain **19** valid under the assumption of F-differentiability. We shall not therefore mention these facts again.

7 Model Problem

We shall collect here all the results we have obtained for the case of the functional

$$v \mapsto J(v) = \frac{1}{2}a(v, v) - L(v)$$

on a Hilbert space V satisfying conditions (i), (ii) and (iii) of Example 1.1. This contains, as the abstract formulation, most of the linear elliptic problems that we shall consider except for the case of non-symmetric elliptic operators.

(1) J is twice Fréchet differentiable (in fact, F-differentiable of all orders) and hence is also Gateaux differentiable.

$$J'(u, \varphi) = a(u, \varphi) - L(\varphi) \text{ and } J''(u; \varphi, \psi) = a(\varphi, \psi).$$

J has a gradient and a Hessian at every point $u \in V$

$$G(u) = (gradJ)(u) : \varphi \mapsto a(u, \varphi) - L(\varphi).$$

Moreover, $H(u)$ is self-adjoint since $a(\varphi, \psi) = a(\psi, \varphi)$ for all φ, $\psi \in V$.

(2) *Taylor's formula for* J If $u, v, \in V$ then

$$J(v) = J(u) + \{a(u, v - u) - L(v - u)\} + \frac{1}{2}a(v - u, v - u)$$

(3) Since the mapping $v \mapsto a(u, v)$ for any $u \in V$ is continuous linear and $L \in V'$, by the theorem of Fréchet-Riesz on Hilbert spaces there exist (unique elements $Au,\ f \in V$ such that

$$a(u, v) = (Au, v)_V \text{ and } L(v) = (f, v)_V \text{ for all } v \in V$$

Clearly $A : V \to V$ is a continuous linear map. Moreever we have **20**

$$\begin{aligned} ||A||_{\mathscr{L}(V,V)} &\leq M \text{ by } (i), \\ (Av, v)_V \geq \alpha ||v||_V^2 &\text{ for all } v \in V \text{ by (ii) and} \\ ||f||_V &\leq N. \end{aligned}$$

(4) The functional J is strictly convex in V.

(5) J is weakly lower semi-continuous in V.

Chapter 2

Minimisation of Functionals - Theory

In this chapter we shall discuss the local and global minima of functionals on Banach spaces and give some sufficient conditions for their existence, relate them to conditions on their G-derivatives (when they exist) and convexity properties. Then we shall show that the problem of minimisation applied to suitable functionals on Sobolev spaces lead to and equivalent to some of the standard examples of linear and non-linear elliptic boundary value problems. 21

1 Minimisation Without Convexity

Let $\mathcal{U}$ be a subset of a normed vector space V and $J : \mathcal{U} \subset V \to \mathbb{R}$ be a functional.

Definition 1.1. A funvtional $J : \mathcal{U} \subset V \to \mathbb{R}$ is said to have a local minimum at a point $u \epsilon \mathcal{U}$ if there exists a neighbourhood $\mathscr{V}(u)$ of u in V such that

$$J(u) \leq J(v) \text{ for all } v \epsilon \mathcal{U} \cap \mathscr{V}(u)$$

Definition 1.2. A functional J on $\mathcal{U}$ is said to have a global minimum (or an absolute minimum) in $\mathcal{U}$ if there exist a $u \epsilon \mathcal{U}$ such that

$$J(u) \leq J(v) \text{ for all } v \epsilon \mathcal{U}.$$

We have the following existence result.

Theorem 1.1. *Suppose $V, \mathcal{U}$ and $J : \mathcal{U} \to \mathbb{R}$ satisfy the following hypothesis :*

(H1) V is a reflexive Banach space,

(H2) $\mathcal{U}$ is weakly closed.

22 *(H3) $\mathcal{U}$ is bounded and*

(H4) $J : \mathcal{U} \subset V \to \mathbb{R}$ is weakly lower semi-continuous.

Then J has a global minimum in $\mathcal{U}$.

Proof. Let ℓ denote $\inf\limits_{v \epsilon \mathcal{U}} J(v)$. If v_n is a minimising sequence for J, i.e. $\ell = \inf\limits_{v \epsilon \mathcal{U}} J(v) = \lim\limits_{n \to \infty} J(v_n)$, then by the boundedness of $\mathcal{U}$ (i.e. by H3) v_n is a bounded sequence in V i.e. there exists a constant $C > 0$ such that $||v_n|| \leq C$ for all n. By the reflexivity of $V(H1)$ this bounded sequence is weakly relatively compact. So there is a subsequence $v_{n'}$ of v_n such that $v_{n'} \rightharpoonup u$ in V. $\mathcal{U}$ being weakly closed $(H2)$ $u \epsilon \mathcal{U}$. Finally, since $v_{n'} \rightharpoonup u$ and J is weakly lower semi-continuous

$$J(u) \leq \liminf_{n \to \infty} J(v_{n'})$$

which implies that

$$J(u) \leq \lim_{n \to \infty} J(v_{n'}) = \ell \leq J(v) \text{ for all } v \epsilon \mathcal{U}.$$

□

Theorem 1.2. *If $V, \mathcal{U}$ and J satisfy the Hypothesis $(H1)$, $(H2)$, $(H4)$ and J satisfies*

$(H3)'$
$$\lim_{||v||_V \to +\infty} J(v) = +\infty$$

then J admits a global minimum in $\mathcal{U}$.

Proof. We shall reduce the problem to the previous case. Let $w \in \mathcal{U}$ be arbitrary fixed. Consider the subset $\mathcal{U}_0$ of $\mathcal{U}$:

$$\mathcal{U}_0 = \{v; v\epsilon\mathcal{U} \text{ such that } J(v) \leq J(w)\}.$$

□

It is immediatly seen that the existence of a minimum in $\mathcal{U}_0$ is equivalent to that in $\mathcal{U}$. We claim that $\mathcal{U}_0$ is bounded and weakly closed in V. i.e. hypothesis $(H2)$ and $(H3)$ hold for $\mathcal{U}_0$. In fact, suppose $\mathcal{U}_0$ is not bounded then we can find a sequence $v_n \in \mathcal{U}_0$ with $||v_n||_V \to +\infty$. 23 The, by $(H3)', J(v_n) \to +\infty$ which is impossible since $v_n \in \mathcal{U}_0$ implies that $J(v_n) \leq J(w)$. Hence $\mathcal{U}_0$ is bounded. To prove that $\mathcal{U}_0$ is weakly closed, let $u_n \in \mathcal{U}_0$ be a sequence that $u_n \rightharpoonup u$ in V. Since is weakly closed $u \in \mathcal{U}$. On the other end, since J is weakly lower semi-continuous $u_n \rightharpoonup u$ in V implies that

$$J(u) \leq \liminf J(u_n) \leq J(w)$$

proving that $u \in \mathcal{U}_0$. Now $\mathcal{U}_0$ and J satisfy all the hypothesis of Theorem 1.1 and hence J has a global minimum in $\mathcal{U}_0$ and hence in $\mathcal{U}$.

Next we give a necessary condition for the existence of a local minimum in items of the first G-derivative (when it exists) of the functional J. For this we need the following concept of admissible (or feasible) directions at a points u for a domian $\mathcal{U}$ in V. It $u, v \in V$ $u \neq v$ then the nonzero vector $v - u$ can be consider as a direction in V.

Definition 1.3. (1) A direction $v-u$ in V is said to be a strongly admissible direction at the points u for the domian $\mathcal{U}$ if there exists a sequence $\epsilon_n > 0$ such that

$$\epsilon_n \to 0 \text{ as } n \to \infty \text{ and } u + \epsilon_n(v - u) \in \mathcal{U} \text{ for each } n.$$

(2) A direction $v - u$ in V is said to be weakly admissible at the points u for the domian $\mathcal{U}$ if there exist sequence $\epsilon_n > 0$ and $w_n \in V$ such that

$$\epsilon_n \to 0 \text{ and } w_n \to 0 \text{ in } V, u_n + \epsilon_n(v - u) + \epsilon_n w_n \in \mathcal{U} \text{ for each } n.$$

We shall mainly use the notion of strongly admissible direction. But some results on minimisation of functionals are known which make use of the notion of weakly admissible directions.

24 We have the following necessary condition for the existence of a local minimum.

Theorem 1.3. *Suppose a functional $J : \mathcal{U} \subset V \to \mathbb{R}$ has a local minimum at a point $u \in \mathcal{U}$ and is G-differentiable at u in all directions then $J'(u, v-u) \geq 0$ for every $v \in V$ such that $v-u$ is a strongly admissible direction.*

Furtheremore, if $\mathcal{U}$ is an open set then

$$J'(u, \varphi) = 0 \textit{ for all } \varphi \in V.$$

Proof. If $u \in \mathcal{U}$ is local minimum for J then there exists a neighbourhood $\mathscr{V}(u)$ of u in V such that

$$J(u) \leq J(w) \text{ for all } w \in \mathcal{U} \cap \mathscr{V}(u).$$

□

If $v \in V$ and $v-u$ is a strongly admissible direction then, for n large enough,

$$u + \epsilon_n(v-u) \in \mathcal{U} \cap \mathscr{V}(u)$$

so that

$$J(u) \leq J(u + \epsilon_n(v-u)).$$

Hence

$$J'(u, v-u) = \lim_{\epsilon_n \to 0} (J(u + \epsilon_n(viu)) - J(u))/\epsilon_n \geq 0.$$

Finally, if $\mathcal{U}$ is an open set in V then $\mathcal{U}$ contains an open ball in V of centre u and hence every direction is strongly admissible at u for $\mathcal{U}$. Taking $v = u \pm \varphi$, $\varphi \in V$ it follows from the first part that

$$J'(u, \pm\varphi) \geq 0 \text{ or equivalently } J'(u, \varphi) = 0 \text{ for all } \varphi \in V.$$

25 In particualr, if $\mathcal{U}$ is open and J has a gradient $G(u) \in V'$ at $u \in \mathcal{U}$

and if u is a local minimum then

$$J'(u, \varphi) =< G(u), \varphi >_{V' \times V} = 0 \text{ for all } \varphi \in V; \text{ i.e. } G(u) = 0 \in V'.$$

This result is thus in conformity with the classical case of differentiable functions.

Remark 1.1. The converse of Theorem 1.3 requires convexity assumptions as we shall see in the following section.

2 Minimistion with Convexity conditions

We shall show that under convexity assumptions on the domian $\mathcal{U}$ and the functional J the notions of local and global minima coincide. We also give another sufficient condition for the existence of minima.

Lemma 2.1. *If $\mathcal{U}$ is a convex subset of a normed vector space V and $J : \mathcal{U} \subset V \to \mathbb{R}$ is a convex functional then any local minimum is also a global minimum.*

Proof. Suppose $u \in \mathcal{U}$ is a local minimum of J. Then there is a neighbourhood $\mathscr{V}(u)$ of u in V such that

$$J(u) \leq J(v) \text{ for all } v \in \mathscr{V}(u) \cap \mathcal{U}.$$

On the other hand, if $v \in \mathcal{U}$ then $u + \theta(v - u) \in \mathcal{U}$ for all $\theta \in [0, 1]$ by convexity of $\mathcal{U}$. □

Moreover, if θ is small enough, say $0 \leq \theta \leq \theta_v$ then $u + \theta(v - u) \in \mathscr{V}(u)$. Hence

$$\begin{aligned} J(u) &\leq J(u + \theta(v - u)) \text{ for all } 0 \leq \theta \leq \theta_v \\ &\leq J(u) + \theta(J(v) - J(u)) \text{ by convexity of J, for all } 0 \leq \theta \leq \theta_v, \end{aligned}$$

which implies that

$$J(u) \leq J(v) \text{ for all } v \in \mathcal{U}.$$

26

Whenever the assumptions of Lemma 2.1 are satisfied we shall call a minimum without reference to local or global. Next lemma concerns the uniqueness of such a minimum.

Lemma 2.2. *If $\mathcal{U}$ is a convex subset of a normed vector space and $J : \mathcal{U} \subset V \to \mathbb{R}$ is strictly convex then there exixts a unique minimum $u \in \mathcal{U}$ for J.*

Proof. The existence is proved in Lemma 2.1. To prove the uniqueness, if $u_1 \neq u_2$ are two minima for J in $\mathcal{U}$ then

$$J(u_1) = J(u_2) \leq J(v) \text{ for all } v \in \mathcal{U}$$

and, in particular, this holds for $v = \frac{1}{2}u_1 + \frac{1}{2}u_2$ which belongs to $\mathcal{U}$ since $\mathcal{U}$ is convex. On the other hand, since J is strictly convex

$$J(\frac{1}{2}u_1 + \frac{1}{2}u_2) < \frac{1}{2}J(u_1) + \frac{1}{2}J(u_2) = J(u_1 \leq J(v))$$

which is impossible if we take $v = \frac{1}{2}(u_1+u_2)$. This proves the uniqueness of the minimum. □

We shall now pass to a sufficient condition for the existence of minima of functionals which is the exact analogue of the case of twice differentiable functions.

Theorem 2.1. *Let $J : V \to \mathbb{R}$ be a functional on $V, \mathcal{U}$ a subset of V satisfying the following hypothesis :*

(H1) V is a relexive Banach space;

(H2) J has a gradient $G(u) \in V'$ everywhere in $\mathcal{U}$;

(H3) J is twice G-differentiable in all directions $\varphi, \psi \in V$ and satisfies the condition

$$J''(u; \varphi, \varphi) \geq ||\varphi||_V \chi(||\varphi||_V) \textit{ for all } \varphi \in V,$$

27 *where $t \mapsto \chi(t)$ is afunction on $\{t \in \mathbb{R}; t \geq 0\}$ such that*

$$\chi(t) \geq 0 \textit{ and } \lim_{t\to+\infty} \chi(t) = +\infty;$$

(H4) *$\mathcal{U}$ is a closed convex set.*

Then there exists at least one minimum $u \in \mathcal{U}$ of J. Furthermore, if in $(H3)$

(H5)

$$\chi(t) > 0 \textit{ for } t > 0$$

is satisfied by χ then there exists a unique minimu of J in $\mathcal{U}$.

Remark 2.1. We note that a convex set $\mathcal{U}$ is weakly if and only if it is strongly closed and thus in $(H4)$ above $\mathcal{U}$ may be assumed weakly closed.

Proof of Theorem 2.1. First of all by $(H3)$, $J''(u; \varphi, \varphi) \geq 0$ and hence J is convex by Proposition 1.3.2. Similarly $(H5)$ implies that J is strictly convex again by Proposition 1. 3.2. Then, by Proposition 1. 4.2 $(H2)$ and $(H3)$ together imply that J is weakly lower semi-continuous. We next show that J satisfies condition $(H3)'$ of Theorem 1.2: namely $J(v) \to +\infty$ as $||v||_V \to +\infty$. For this let $w \in \mathcal{U}$ be arbitrarily fixed. Then, because of $(H2)$ and $(H3)$ we can apply Taylor's formula to get, for $v \in V$.

$$J(v) = J(w) + < G(w), v - w >_{V' \times V} + \frac{1}{2} J''(w + \theta_0(v - w), v - w; v - w)$$

for some $\theta_0 \in]0, 1[$. Using $(H3)$ and estimating the second and third terms on the right side we have

$$| < G(w), v - w >_{V' \times V} | \leq ||G(w)||_{V'} ||v - w||'_V$$

$$J''(w + \theta_0(v - w), v - w, v - w) \geq ||v - w||_V \times (||v - w||_V) \text{ and hence}$$

$$J(v) \geq J(w) + ||v - w||_V [\frac{1}{2} \times (||v - w||_V) - ||G(w)||_{V'}].$$

Here, since $w \in \mathcal{U}$ is fixed, as $||v||_V \to +\infty$ **28**

$$||v - w||_V \to +\infty,$$

$J(w)$ and $\|G(w)\|_{V'}$ are constants and

$$\chi(\|v-w\|_V) \to +\infty \text{ by } (H3)$$

which implies that $J(v) \to +\infty$ as $\|v\|_V \to +\infty$. The theorem then follows on application of Theorem 1.2.

Theorem 2.2. *Suppose $\mathcal{U}$ is a convex subset of a Banach space and $J : \mathcal{U} \subset V \to \mathbb{R}$ is a G-differentiable (in all directions) convex functional. Then*

$u \in \mathcal{U}$ is a minimum for J (i.e. $J(u) \leq J(v)$ for all $v \in V$) if and only if

$$u \in \mathcal{U} \textit{ and } J'(u, v-u) \geq 0 \textit{ for all } v \in \mathcal{U}.$$

Proof. Let $u \in \mathcal{U}$ be a minimum for J. Then, since $\mathcal{U}$ is convex, $v-u$ is a strongly admissible direction at u for $\mathcal{U}$ for any v. Then, by Theorem 1.3, $J'(u, v-u) \geq 0$ for any $v\epsilon\mathcal{U}$. Conversely, since J is convex and G-differentiable, by part (1) of Proposition 1. 3.1, we find that

$$J(v) \geq J(u) + J'(u, v-u) \text{ for any } v\epsilon\mathcal{U}.$$

□

Then using the assumption that $J'(u; v-u) \geq 0$ it follows that $J(u) \leq J(v)$ i.e. u is a minimum for J in $\mathcal{U}$.

Our next result concerns minima of convex functionals in the non-differentaible case.

Theorem 2.3. *Let $\mathcal{U}$ be a convex subset of a Banach space V. Suppose $J : \mathcal{U} \subset V \to \mathbb{R}$ is a functional of the form $J = J_1 + J_2$ where J_1, J_2 are convex functionals and J_2 is G-differentiable in $\mathcal{U}$ in all directions. Then $u\epsilon\mathcal{U}$ is a minimum for J if and only if* 29

$$u\epsilon\mathcal{U}, J_1(v) - J_1(u) + J_2'(u, v-u) \geq 0 \textit{ for all } v\epsilon\mathcal{U}$$

Proof. Suppose $u\epsilon\mathcal{U}$ is a minimum of J then

$$J(u) = J_1(u) + J_2(u) \leq J_1(u + \theta(v-u)) + J_2(u + \theta(v-u))$$

since $u + \theta(v - u)\epsilon\mathcal{U}$. Here, by convexity of J_1, we have

$$J_1(u + \theta(v - u)) \leq J_1(u) + \theta(J_1(v) - J_1(u))$$

so that

$$J_2(u) \leq \theta(J_1(v) - J_1(u)) + J_2(u + \theta(v - u)).$$

That is

$$J_1(v) - J_1(u) + (J_2(u + \theta(v - u)) - J_2(u))/\theta \geq 0.$$

□

Taking limits as $\theta \to 0$ we get the required assertion. Conversely, since J_2 is convex and is G-differentiable we have, from part (1) of Proposition 1. 3.1,

$$J_2(v) - J_2(u) \geq J_2'(u, v - u) \text{ for all } u, v\epsilon\mathcal{U}.$$

Now we can write, for any $v\epsilon\mathcal{U}$,

$$\begin{aligned} J(v) - J(u) &= J_1(v) - J_1(u) + J_2(v) - J_u \\ &\geq J_1(v) - J_1(u) + J_2'(u, v - u) \geq 0 \end{aligned}$$

by assumption which proves that $u\epsilon\mathcal{U}$ is a minimum for J.

3 Applications to the Model Problem and Reduction to variational Inequality

We shall apply the results pf Section 2 to the functional J of Example 1. 1.1 on a Hilbert space. More precisely, let V be a Hilbert space and $J : V \to \mathbb{R}$ be the functional 30

$$v \mapsto J(v) = \frac{1}{2}a(v, v) - L(v)$$

where $a(\cdot, \cdot)$ is a symmetric bilinear, bicontinuous, coercive form on V and $L\epsilon V'$. Further, let K be a closed convex subset of V. Consider the following

Problem 3.1. To find

$$u \epsilon K; J(u) \leq J(v) \text{ for all } v \epsilon K.$$

i.e. to find a $u \epsilon K$ which minimizes J on K. We have seen in Chapter 1 (Section 7) that J is twice F-(and hence also G-) differentiable and that

$$J'(u, \varphi) =< G(u), \varphi >_{V' \times V} = a(u, \varphi) - L(\varphi)$$

$$J''(u; \varphi, \psi) =< H(u)\varphi, \psi >_{V' \times V} = a(\varphi, \psi)$$

Moreover, the coercivity of $a(\cdot, \cdot)$ implies that

$$J''(u; \varphi, \varphi) = a(\varphi, \varphi) \geq \alpha ||\varphi||^2_V.$$

If we choose $\chi(t) = \alpha t$ then all the assumptions of Theorem 2.1 are satisfied by V, J and K so that the Problem 3.1 has a unique solution. Also, by Theorem 2.2, the problem 3.1 is equivalent to

Problem 3.2. To find

$$u \epsilon K; a(u, v - u) \geq L(v - u) \text{ for all } v \epsilon K.$$

We can summarise these facts as

31 **Theorem 3.1.** *(1) There exists a unique solution $u \epsilon K$ of the Problem 3.1 and*

(2) Problem 3.1 is equivalent to problem 3.2.

The problem 3.2 is called a variational inequality associted to the closed convex set K and the bilinear form $a(\cdot, \cdot)$. As we shall see in the following section the variational inequality (3.2) arises as generalizations of elliptic boundary value problems for suitable elliptic operators. It turns out that in many of the problems solving (numerically) the minimisation problem 3.1 is much easier and faster than solving the equivalent variational inequality (3.2).

In the particular case where $K = V$ the Problme 3.1 is nothing but the Problem

(3.3) to find $u \epsilon V; J(u) \leq J(v)$ for all $v \epsilon V$

which is equivalent to the Problem

(3.4) to find $u\epsilon V; a(u,\varphi) = L(\varphi)$ for a $\varphi\epsilon V$. As we have seen in Chapter 1, (3.4) is equivalent to (3.2) : if $\varphi\epsilon V$ we take $v = u \pm \varphi\epsilon K = V$ in (3.2) to get (3.4) and the converse is trivial.

The following result is a generalization of Theorem 3.1 to non-symmetric case and is due to G-Stampacchia. This generalizes and includes the classical Lax-Milgram theorem. (See [43]).

Theorem 3.2. *(Stampacchia). Let K be a closed convex subset of a Hilbert space V and $a(\cdot,\cdot)$ be a bilinear bicontinuous coercive form on V. Then for any given $L\epsilon V'$ the variational inequality (3.2) has a unique solution $u\epsilon K$.*

Proof. Since, for any $u, v \mapsto a(u,v)$ is continuous linear on V and $L\epsilon V'$ there exist unique elements Au, $f\epsilon V$ by Fréchet-Riesz theorem such that

$$a(u,v) = (Au,v)_V \text{ and } L(v) = (f,v)_V.$$

□ 32

Moreover $A\epsilon\mathscr{L}(V,V')$ with $||A||_{\mathscr{L}(V,V')} \leq M$ and $||f||_V \leq N$ where $M > 0$, $N > 0$ are constants such that

$$|a(u,v)| \leq M||u||_V||v||_V \text{ for all } u, v\epsilon V,$$
$$|L(v)| \leq N||v||_V \text{ for all } v\epsilon V.$$

Let $\alpha > 0$ be the constant of V-coercivity of $a(\cdot,\cdot)$ i.e.

$$a(v,v) \geq \alpha||v||_V^2 \text{ for all } v\epsilon V.$$

Since K is a closed convex set there exists a projection mapping $P : V \to K$ with $||P||_{\mathscr{L}(V,V)} \leq 1$. Let $\gamma > 0$ be a constant which we shall choose suitably later on. Consider the mapping

$$V \ni v \mapsto v - \gamma(Av - f) = T_\gamma(v)\epsilon V.$$

For γ sufficiently small T_γ is a contraction mapping. In fact, if $v_1, v_2\epsilon V$ then

$$T_\gamma v_1 - T_\gamma v_2 - (I - \gamma A)(v_1 - v_2).$$

Setting $w = v_1 - v_2$ we have

$$\begin{aligned}
||(I-\gamma A)w||_V^2 &= (w-\gamma Aw, w-\gamma Aw)_V \\
&= ||w||_V^2 - \gamma[(w, Aw)_V + (Aw, w)_V] + \gamma^2 ||Aw||_V^2 \\
&\leq ||w||_V^2 - 2\gamma\alpha||w||_V^2 + \gamma^2 M^2 ||w||_V^2 \\
&= (1-2\gamma\alpha+\gamma^2 M^2)||w||_V^2
\end{aligned}$$

by V-coercivity and continuity of the operator A. It is easy to see that if $0 < \gamma < 2\alpha/M^2$ then $1 - 2\gamma\alpha + \gamma^2 M^2 < 1$ and hence T_γ becomes a contraction mapping. Then the mapping $PT_\gamma|_K : K \to K$ is a contraction mapping and hence has a unique fixed point $u\epsilon K$ by contraction mapping theorem i.e. **33**

$$u\epsilon K \text{ and } u = P(u - \gamma(Au - f)).$$

This is the required solution of the variational inequality (3.2) as can easily be checked.

4 Some Functional Spaces

We shall briefly recall some important Sobolev spaces of distributions on an open set in $\mathbb{R}^n$ and some of their properties. These spaces play an important role in the weak (or variational) formulation of elliptic problems which we shall consider in the following. All our functionals in the examples will be defined on these spaces. For details we refer to the book of Lions and Magenes [32].

Let Ω be a bounded open subset in $\mathbb{R}^n$ and Γ denote its boundary. We shall assume Γ to be sufficiently "regular" which we shall make precise whenever necessary.

Sobolev spaces. We introduce the Sobolev space $H^1(\Omega)$:

$$H^1(\Omega) = \{v | v\epsilon L^2(\Omega), \partial x_j \epsilon L^2(\Omega), j = 1, \cdots, n\} \tag{4.1}$$

where $D_j v = \partial v/\partial x_j$ are taken in the sense of distributions

$$\text{i.e.} \qquad < D_j v, \varphi >= - < v, D_j\varphi > \text{ for all } \varphi\epsilon\mathscr{D}(\Omega)$$

Here $\mathscr{D}(\Omega)$ denotes the space of all C^∞ -functions with compact support in Ω and $< \cdot, \cdot >$ denotes the duality between $\mathscr{D}(\Omega)$ and the space of distributions $\mathscr{D}'(\Omega)$ on Ω. $H^1(\Omega)$ is provided with the inner-product

$$((u, v))(u, v)_{L^2(\Omega)} + \sum_{j=1}^{n}(D_j u, D_j v)_{L^2(\Omega)} \tag{4.2}$$
$$= \int_\Omega \{uv + \sum_{j=1}^{n}(D_j u)(D_j v)\} dx$$

for which becomes a Hilbert space. The following inclusions are obvious (and are continuous) $\mathscr{D}(\Omega) \subset C^1(\overline{\Omega}) \subset H^1(\Omega)$. 34

We also introduce the space

$$H^1_0(\Omega) = \text{ the closure of } \mathscr{D}(\Omega) \text{ in } H^1(\Omega). \tag{4.3}$$

We ahve the following well-known results.

(4.4) **Theorem of Density:** If Γ is "regular" (for instance, Γ is a C^1 (or C^∞)-mainfold of dimension $n - 1$) then $C^1(\overline{\Omega})$ (resp. $C^\infty(\overline{\Omega})$) is dense in $H^1(\Omega)$.

(4.5) **Theorem of Trace.** If Γ is "regular" then the linear mapping $v \mapsto v/\Gamma$ of $C^1(\overline{\Omega}) \to C^1(\Gamma)$ (resp pf $C^\infty(\overline{\Omega}) \to C^\infty(\Gamma)$) extends to a continuous linear map of $H^1(\Omega)$ into $L^2(\Gamma)$ denoted by γ and for any $v \epsilon H^1(\Omega)$ γv is called the trace of v on Γ. Moreover, $H^1_0(\Omega) = \{v \epsilon H^1(\omega) \gamma v = 0\}$. We shall more often use this characterization of $H^1_0(\Omega)$. The trace map is not surjective. For a characterization of the image of $H^1(\omega)$ by γ (which is proper subspace, denoted by $H^{\frac{1}{2}}(\Gamma)$) we refer to the book of Lions and Magenes [32]. We can also define spaces $H^m(\Omega)$ and $H^m_0(\Omega)$ in the same way for any $m > 1$.

Remark 4.1. The Theorem of trace is slightly more precise than our statement above. For this and also for a proof we refer to the book of Lions and Magenes [32].

For some non-linear problems we shall also need spaces of the form

$$V = H^1_0(\Omega) \cap L^p(\Omega) \text{ where } p \geq 2. \tag{4.6}$$

The space V is provided with the norm

$$v \mapsto ||v||_V = ||v||_{H^1(\Omega)} + ||v||_{L^p(\Omega)}$$

for which it becomes a Banach space. If $2 \leq p < +\infty$ then V is a 35 reflexive Banach space.

In order to given an interpretation of the solutions of weak formulations of the problems as solutions of certain differential equations with boundary conditions we shall need an extension of the classical Green's formula which we recall here.

(4.8) **Green's formula for Sobolev spaces.** Let Ω be a bounded open set with sufficiently "regular" boundary Γ. Then there exists a unique outer normal vector $\underline{n}(x)$ at each point x on Γ. Let $(\underline{n}_1(x), \cdots, \underline{n}_n(x))$ denote the direction cosines of $\underline{n}(x)$. We define the operator of exterior normal derivation formally as

$$\partial/\partial\underline{n} = \sum_{j=1}^{n} n_j(x) D_j. \tag{4.9}$$

Now if $u, v \epsilon C^1(\Omega)$ then by the classical Green's formula we have

$$\int_\Omega (D_j u) v dx = -\int_\Omega u (D_j v) dx + \int_\Gamma u v n_j d\sigma$$

where $d\sigma$ is the area element on Γ. This formual remains valid also if $u, v \epsilon H^1(\Omega)$ in view of the trace theorem and density theorem as can be seen using convergence theorems.

Next if $u, v \epsilon C^2(\overline{\Omega})$, then applying the above formula to $D_j u, D_j v$ and summing over $j = 1, \cdots .n$ we get

$$\sum_{j=1}^{n} (D_j u, D_j v)_{L^2(\Omega)} = -\sum_{j=1}^{n} \int_\Omega (D_j^2 u) v dx + \int_\Gamma \partial u/\partial\underline{n}.v d\sigma$$

$$\text{i.e. } \sum_{j=1}^{n} (D_j u, D_j v)_{L^2(\Omega)} = -\int_\Omega (\triangle u) v dx + \int_\Gamma \partial u/\partial\underline{n}.v d\sigma. \tag{4.10}$$

Once again this formula remains valid if ; for instance, $u\epsilon H^2(\Omega)$, $v\epsilon H^1(\Omega)$ using the density and trace theorems. In fact, $u\epsilon H^2(\Omega)$ implies that $\triangle u\epsilon L^2(\Omega)$ and since $D_j u\epsilon H^1(\Omega)$, $\gamma(D_j u)$ exists and belong to $L^2(\Gamma)$ so that $\partial u/\partial \underline{n} = \sum_{j=1}^{n} n_j\gamma(D_j u)\epsilon L^2(\Gamma)$.

5 Examples

In this section we shall apply results of the previous sections to some concrete example of functionals on Sobolev spaces and we interprete the corresponding variational inequalities as boundary value problems for differential operators. 36

Throughout this section Ω will be a bounded open set with sufficiently "regular" boundary Γ. We shall not make precise the exact regularity conditions on Γ except to say that it is such that the trace, density and Green's formula are valid.

We begin with the following abstract linear problem.

Example 5.1. Let $\Gamma = \overline{\Gamma}_1 \cup \overline{\Gamma}_2$ where Γ_j are open subsets of Γ such that $\Gamma_1 \cap \Gamma_2 = \phi$ Consider the space

$$V = \{v | v\epsilon H^1(\Omega); \gamma v = 0 \text{ on } \Gamma_1\}. \tag{5.1}$$

V is clearly a closed subspace of $H^1(\Omega)$ and is provided with the inner product induced from that in $H^1(\Omega)$ and hence it is a Hilbert space. Moreover,

$$H_0^1(\Omega) \subset V \subset H^1(\Omega) \tag{5.2}$$

and the inclusions are continuous linear. If $f\epsilon L^2(\Omega)$ we consider the functional

$$J(v) = \frac{1}{2}((u, v)) - (f, v)_{L^2(\Omega)} \tag{5.3}$$

i.e. $a(u, v) = ((u, v))$ and $L(v) = (f, v)_{L^2(\Omega)}$. Then $a(\cdot, \cdot)$ is bilinear, bicontinuous and V-coercive :

$$|a(u, v)| \leq ||u||_V||v||_V = ||u||_{H^1(\Omega)}||v||_{H^1(\Omega)} \text{ for } u, v\epsilon V,$$

$$a(v,v) = ||v||^2_{H^1(\Omega)} \text{ for } v\epsilon V$$

$$\text{and } |L(v)| \leq ||f||_{L^2(\Omega)}||v||_{L^2(\Omega)} \leq ||f||_{L^2(\Omega)}||v||_{H^1(\Omega)} \text{ for } v\epsilon V.$$

37 Then the problems (3.3) and (3.4) respectively become

(5.4) to find $u\epsilon V, J(u) \leq J(v)$ for all $v\epsilon V$ and

(5.5) to find $u\epsilon V, ((u,\varphi)) = (f,\varphi)_{L^2(\Omega)}$ for all $\varphi\epsilon V.$

From what we have seen in Section 3 these two equivalent problems have unique solutions.

The Problem (5.5) is the weak (or variational) formulation of the Dirichlet problem (if $\Gamma_2 = \phi$), Neumann problem if $\Gamma_1 = \phi$ and the mixed boundery value problem in the general case.

We now interprete the solutions of Problems (5.2) when they are sufficiently regular as solutions of the classical Dirichlet (resp. Neumann of mixed) problems.

Suppose we assume $u\epsilon C^2(\overline{\Omega}) \cap V$ and $v\epsilon C^1(\overline{\Omega}) \cap V$. We can write using the Green's formula (4.10)

$$a(u,v) = ((u,v)) = \int_\Omega (-\triangle u + u)vdx + \int_\Gamma \partial u/\partial \underline{n}.vd\sigma = \int_\Omega fvdx$$

$$\text{(5.6)} \qquad \text{i.e. } \int_\Omega (-\triangle u + u - f)vdx + \int_\Gamma \partial u/\partial \underline{n}.vd\sigma = 0.$$

We note that this formula remains valide if $u\epsilon H^2(\Omega) \cap V$ for any $v\epsilon V$.

First we choose $v\epsilon\mathscr{D}(\Omega) \subset V$ (enough to take $v\epsilon C^1_0(\Omega)(\Omega) \subset V$) then the boundary integral vanishes so that we get

$$\int_\Omega (-\triangle u + u - f)vdx = 0 \ \forall v\epsilon\mathscr{D}(\Omega).$$

Since $\mathscr{D}(\Omega)$ is dense in $L^2(\Omega)$ this implies that (if $u\epsilon H^2(\Omega)$) u is a solution of the differential equation

(5.7) $-\triangle u + u - f = o$ in Ω (in the sense of $L^2(\Omega)$).

38

More generally, without the strong regularity assumption as above, u is a solution of the differential equation

$$(5.8) \qquad -\triangle u + u - f = 0 \text{ in the sense of distributions in } \Omega.$$

Next we choose $v\epsilon V$ arbitrary. Since u satisfies the equation (5.8) in Ω we find from (5.6) that

$$(5.9) \qquad \int_{\Gamma_2} \partial u/\partial \underline{n} v d\sigma = 0 \ \forall v\epsilon V,$$

whcih means that $\partial u/\partial \underline{n} = 0$ on Γ in some generalized sense. In fact, by trace theorem $\gamma v\epsilon H^{\frac{1}{2}}(\Gamma)$ and hence $\partial u/\partial \underline{n} = 0$ in $H^{-\frac{1}{2}}(\Gamma)$ (see Lions and Magenese [32]). Thus, if the Problem (5.2) has a regular solution then it is the solution of the classical problem

$$(5.10) \qquad \begin{cases} -\triangle u + u & = f \text{ in } \Omega \\ u & = 0 \text{ on } \Gamma_1 \\ \partial u/\partial \underline{n} & = 0 \text{ on } \Gamma_2 \end{cases}$$

The Problem (5.10) is the classical Dirichlet (resp. Neumann, or mixed) problem for the elliptic differential operator $-\triangle u + u$ if $\Gamma_2 = \phi$ (resp. $\Gamma_1 = \phi$ or general Γ_1, Γ_2).

Remark 5.1. The variational formualtion (5.5) of the problem (5.5) is very much used in the Finite elements method.

Example 5.1 is a special case of the following more general problem.

Example 5.1′. *Let $\Omega, \Gamma = \Gamma_1 \cup \Gamma_2$ and V be as in Example 5.1. Suppose given an integro-differentail bilinear form ;*

$$(5.11) \qquad a(u, v) = \int_\Omega \sum_{i,j=1}^{n} a_{ij}(x)(D_i u)(D_j v) dx + \int_\Omega a_0(x) uv dx,$$

where the coefficients satisfy the following conditions: 39

(5.12)
$$\begin{cases} a_{ij}\epsilon L^{\infty}(\Omega),\ a_{\circ}\epsilon L^{\infty}(\Omega); \\ \text{ condition of ellipticity there exists a constant } \alpha > 0 \text{ such that} \\ \sum_{i,j} a_{ij}(x)\overline{\xi}_i\overline{\xi}_j \geq \alpha \sum_i \overline{\xi}_i^2 \text{ for } \overline{\xi} = (\overline{\xi}_1, \cdots, \overline{\xi}_n)\epsilon\mathbb{R}^n a.e. \text{ in } \Omega; \\ a_{\circ}(x) \geq \alpha > 0. \end{cases}$$

It follows by a simple application of Cauchy-Schwarz inequality that the bi-linear form is well defined and bi-continuous on V: for all $u, v\epsilon V$,

$$|a(u,v)| \leq \max(||a_{ij}||_{L^{\infty}(\Omega)}, ||a_{\circ}||_{L^{\infty}(\Omega)})||u||_V||v||_V$$

$a(\cdot,\cdot)$ is also coercive ; by the ellipticity and the last condition on $a_{\circ}$

$$a(v,v) \geq \alpha \int_{\Omega}(\sum_i |D_i v|^2 + |v|^2)dx = \alpha||v||_V^2, v\epsilon V.$$

Suppose given $f\epsilon L^2(\Omega)$ and $g\epsilon L^2(\Gamma_2)$. Then the linear functional

$$v \mapsto L(v) = \int_{\Omega} fvdx + \int_{\Gamma} gvf\sigma \tag{5.13}$$

on V is continuous and we have again by Cauchy-Schwarz inequality

$$\begin{aligned} |L(v)| &\leq ||f||_{L^2(\Omega)}||v||_{L^2(\Omega)} + ||g||_{L^2(\Omega)}||v||_{L^2(\Gamma)} \\ &\leq (||f||_{L^2(\Omega)} + ||g||_{L^2(\Gamma)})||v||_V \text{ by trace theorem.} \end{aligned}$$

We introduce the functional

$$v \mapsto J(v) = a(v,v) - L(v).$$

For the Problem (5.4) of minimising H on V we further assume

$$a_{ij} = a_{ji}, 1 \leq i, \leq n.$$

If $a_{i,j}$ are smooth functions in Ω and u is a smooth solution of the Problem (5.5) we can interprete u as a solution of a classical problem using the Green's formula as we did in the earlier case. We shall indicate 40 only the essential facts. We introduce the formula differential operator

$$Au = -\sum_{i,j=1}^{n} D_j(a_{ij}D_i u) + a_\circ u. \tag{5.14}$$

If a_{ij} are smooth (for instance, $a_{ij}\epsilon C^1(\Omega)$) then A is a differential operator in the usual sense. By Green's formula we find that
(5.15)

$$a(u,v) = -\sum_{i,j}\int_\Omega D_j(a_{ij}D_i u) + \int_\Gamma \sum_{i,j} a_{ij}(D_{iu})n_j(x)vd\sigma + \int_\Omega a_\circ uvdx$$

where $(n_1(x),\cdots,n_n(x))$ are the direction cosines of the exterior normal to Γ at x. The operator

$$\sum_{i,j} a_{ij}(D_i u)n_j(x) = \partial u/\partial n_A \tag{5.16}$$

is called the co-normal derivatives of u respect to the form $a(\cdot,\cdot)$. Thus we can write (5.15) as

$$a(u,v) = \int_\Omega (Au)vdx + \int_\Gamma \partial u/\partial n_A vd\sigma \tag{5.15$'$}$$

and hence the Problem (5.2) becomes

$$\int_\Omega (Au - f)vdx + \int_\Gamma (\partial u/\partial n_A - g)vd\sigma = 0.$$

Proceeding exactly as in the previous case we can conclude that the Problem (5.5) is equivalent to the classical problem.

$$\begin{cases} Au = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma_1 \\ \partial u/\partial n_A = g & \text{on } \Gamma_2 \end{cases} \tag{5.17}$$

Example 5.2. Let $V = H^1_\circ(\Omega) = \{v|v\epsilon H^1(\Omega), \gamma v = 0\}$, and J be the functional on V:

$$v \mapsto J(v) = \frac{1}{2}||v||_V^2 - (f,v)_{L^2(\Omega)}$$

where $f \epsilon L^2(\Omega)$ is a given function. Suppose 41

$$K = \{v | v \epsilon V, v(x) \geq 0 \text{ a. e. in } \Omega\} \tag{5.19}$$

It is clear that K is convex and it is easily checked that K is also closed in V.

In fact, if $v_n \epsilon K$ and $v_n \to v$ in V then, for any $\varphi \epsilon \mathscr{D}(\Omega)$ such that $\varphi > 0$ in Ω we have

$$\int_\Omega v\varphi dx = \lim_{n\to\infty} \int_\Omega v_n \varphi dx \geq 0$$

(the first equality is an immediate consequence of Cauchy-Schwarz inequality since $v, \varphi \epsilon L^2(\Omega)$). This immediately implies that $v \geq 0$ a. e. in Ω and hence $v \epsilon K$.

We know from Section 3 that the minimising problem.

$$u \epsilon K; J(u) \leq J(v), \ \forall v \epsilon K \tag{5.20}$$

is equivalent to the variational inequality:

$$u \epsilon K; a(u, v - u) \geq L(v - u) = (f, v - u)_{L^2(\Omega)}, \forall v \epsilon K \tag{5.21}$$

and both have unique solutions. In order to interprete this latter problem we find on applying the Green's formula.

$$\int_\Omega (-\triangle u + u - f)(v - u)dx + \int_\Gamma \partial u/\partial n(u - v)d\sigma \geq 0, \forall v \epsilon K. \tag{5.22}$$

Since $v \epsilon K \subset V = H^1_\circ(\Omega)$ the boundary integral vanishes and so

$$\int_\Omega (-\triangle u + u - f)(v - u)dx \geq 0, \forall v \epsilon K. \tag{5.23}$$

If $\varphi \epsilon K$, taking $v = u + \varphi \epsilon K$ we get

$$\int_\Omega (-\triangle u + u - f)\varphi dx \geq 0, \varphi \epsilon K$$

42 from which we conclude that $-\triangle u + u - f \geq 0$ a.e. in Ω. For, if ω is an

open sub-set of Ω where $-\triangle u + u - f > 0$ we take a $\varphi\epsilon\mathscr{D}(\Omega)$ with $\varphi \geq 0$ and supp $\varphi \subset \omega$. Such a φ clearly belongs to K and we would arrive at a contradiction. In particular, this argument also shows that on the subset of Ω where $u > 0$ is satisfies the equation $-\triangle u + u = f$.

Next if we choose $v = 2u\epsilon K$ in (5.23) we find

$$\int_\Omega (-\triangle u + u - f) u dx \geq 0$$

and if we choose $v = \frac{1}{2} u\epsilon K$ we find

$$\int_\Omega (-\triangle u + u - f) u dx \leq 0.$$

These two together imply that

$$(-\triangle u + u - f) u = 0 \tag{5.24}$$

Thus the solution of the variational inequality can be interpreted (when it is sufficiently smooth) as the (unique) solution of the problem :

$$\begin{cases} (-\triangle u + u - f) u & = 0 \text{ in } \Omega \\ -\triangle u + u - f & \geq 0 \text{ a. e. in } \Omega \\ u & \geq 0 \text{ a. e. in } \Omega \\ u & = 0 \text{ on } \Gamma. \end{cases} \tag{5.25}$$

Remark 5.2. The equivalent minimisation problem can be solved numerically (for example, by Gauss-Seidel method). (See Chapter 4 § 4.1).

Exercise 5.2. Let Ω be a bounded open set in $\mathbb{R}^n$ with smooth boundary Γ. Let $V = H^1(\Omega)$ and K be the subset

$$K = \{v | v\epsilon H^1(\Omega); \gamma v \geq 0 \text{ a. e. on } \Gamma\} \tag{5.26}$$

Once again K is a closed convex set. To see that it is closed, if $v_n \epsilon K$ is a sequence such that $v_n \to v$ in V then since $\gamma : H^1(\Omega) = V \to L^2(\Gamma)$ is continuous linear $\gamma v_n \to \gamma v$ in $L^2(\Gamma)$. Now, if $\varphi \epsilon L^2(\Gamma)$ is such that $\varphi > 0$ a. e. on Γ then

$$\int_\Gamma (\gamma v)\varphi d\sigma = \lim_{n\to\infty} \int_\Gamma (\gamma v_n)\varphi d \geq 0 \text{ since } v_n \epsilon K.$$

from which we deduce as in Example 5.1 that $\gamma v \geq 0$.

Let $f \epsilon L^2(\Omega)$ be given

The problem of minimising the functional

$$v \mapsto J(v) = \frac{1}{2}((v, v))_V - (f, v)_{L^2(\Omega)} \tag{5.27}$$

on the closed convex set K is equivalent to the variational inequality

$$u \epsilon K : a(u, v - u) \equiv ((u, v - u))_V \geq (f, v - u)_{L^2(\Omega)}, \forall v \epsilon K. \tag{5.28}$$

Assumig the solution u (which exists and is unique from section 3) is sufficiently regular we can interprete u as follows. By Green's formula we have

$$\int_\Omega (-\triangle u - f)(u - v)dx + \int_\Gamma \frac{\partial u}{\partial \underline{n}}(v - u)d\sigma \geq 0, \forall v \epsilon K. \tag{5.29}$$

If $\varphi \epsilon \mathscr{D}(\Omega)$ the boundary intergal vanishes for $v = u \pm \varphi$ which belongs to K and

$$\int_\Omega (-\triangle u - f)\varphi dx = 0$$

which implies that $-\triangle u = f$ in Ω.

Next since $v = 2u$ and $v = \frac{1}{2}u$ also belong to K we find that

$$\int_\Gamma \frac{\partial u}{\partial \underline{n}} u d\sigma = 0$$

44 which implies that $\dfrac{\partial u}{\partial \underline{n}} u = 0$ a.e. on Γ.

Thus the variational inequality (5.28) is equivalent to the following Problem:

$$\begin{cases} -\triangle u & = f \text{ in } \Omega \\ \partial u/\partial \underline{n} & u = 0 \text{ on } \Gamma \\ \partial u/\partial \underline{n} & \geq 0 \text{ on } \Gamma \\ u \geq 0 & \text{ on } \Gamma \end{cases} \tag{5.30}$$

One can also deduce from (5.30) that on the subset of Γ where $u > 0$, u satisfies the homogeneous Neumann condition

$$\frac{\partial u}{\partial \underline{n}} = 0.$$

Example 5.3. Let Ω be a bounded open set in $\mathbb{R}^n$ with smooth boundary Γ and $1 \leq p < +\infty$. We introduce the space

$$V = \{v | v \epsilon L^2(\Omega); D_j v \epsilon L^{2p}(\Omega), j = 1, \cdots, n\} \tag{5.31}$$

provided with its natural norm

$$v \mapsto ||v||_V = ||v||_{L^2(\Omega)} + \sum_{j=1}^{n} ||D_j v||_{L^{2p}(\Omega)}. \tag{5.32}$$

Then V becomes a reflexive Banach space. Consider the functional $J : V \to \mathbb{R}$:

$$v \mapsto J(v) = \frac{1}{2p} \sum_{j=1}^{n} \int_\Omega |D_j v|^{2p} dx + \frac{1}{2} \int_\Omega |v|^2 dx - \int_\Omega f v dx \tag{5.34}$$

where $f \epsilon L^2(\Omega)$ is given. If we set, $g_j(t) = \frac{1}{2p}|t|^{2p}$ we get a C^1 -function $g_j : \mathbb{R}^1 \to \mathbb{R}^1$ and we have $g'_j(t) = |t|^{2p-2} t$ for all $j = 1, \cdots, n$. Then from Exerices I. 1.1, the functional

$$v \mapsto \sum_j \int_\Omega g_j(v) dx = \frac{1}{2p} \sum_j \int_\Omega |D_j v|^{2p} dx$$

is once G-differentiable in all directions and its G-derivative in any direction φ is given by 45

$$\sum_j \int_\varphi g'_j(u)\varphi dx, \ \forall\varphi\epsilon V.$$

Hence we obtain, in our case,

$$J'(u,\varphi) = \sum_j \int_\Omega |D_j u|^{2p-2}(D_j u)(D_j\varphi)dx + \int_\Omega u\varphi dx - \int_\Omega f\varphi dx. \tag{5.35}$$

Then the minimisation problem

$$u\epsilon V; J(u) \leq J(v), \forall v\epsilon V, \tag{5.36}$$

is equivalent by Theorem 3.1 to the problem

$$u\epsilon V; J'(u,\varphi) = 0, \forall\varphi\epsilon V. \tag{5.37}$$

We can verify that J is strictly convex; for instance, we can compute $J''(u;\varphi,\varphi)$ for any $\varphi\epsilon V$ and find

$$J''(u;\varphi,\varphi) = (2p-1)\sum_j \int_\Omega (|D_j u|^{2(p-1)}|D_j\varphi|^2 + \frac{1}{2}\varphi^2)dx > 0 \tag{5.38}$$

for any $\varphi\epsilon V$ with $\varphi \neq 0$. Then Proposition 1. 3.2 implies the strict convexity of J.

We claim that

$$J(v) \to +\infty \text{ as } ||v||_V \to +\infty.$$

In fact, first of all by Cauchy-Schwarz inequality we have

$$\left|\int_\Omega fvdx\right| \leq ||f||_{L^2(\Omega)}||v||_{L^2(\Omega)}$$

and hence

$$\frac{1}{2}\int_\Omega |v|^2dx - \int_\Omega fvdx \geq \frac{1}{2}||v||_{L^2(\Omega)}(||v||_{L^2(\Omega)} - 2||v||_{L^2(\Omega)})$$

so that

$$J(v) \geq \frac{1}{2p} \sum_j ||D_j V||_{2p}^{2p} + \frac{1}{2} ||v||_{L^2(\Omega)} (||v||_{L^2(\Omega)} - f||_{L^2(\Omega)})$$

46 which tends to $+\infty$ as $||v||_V \to +\infty$.

Then by Theorem 1.2 the minimisation problem (5.36) has a unique solution.

Finally, if we take $\varphi \epsilon \mathscr{D}(\Omega) \subset V$ in the equation (5.35) we get

$$\int_\Omega (\sum_j |D_j u|^{2p-2} (D_j u)(D_j \varphi) + u\varphi - f\varphi) dx = 0.$$

On integration by parts this becomes

$$\int_\Omega (\sum_j -D_j (|D_j u|^{2p-2} D_j u) + u - f)\varphi dx = 0,$$

Thus the solution of the minimising problem (5.36) for J in V can interpreted as the solution of the non-linear problem

(5.39) $$u \epsilon V, -\sum_j D_j (|D_j u|^{2p-2} D_j u) + u = f \text{ in } \Omega.$$

We have used the fact $\mathscr{D}(\Omega)$ is dense in $L^{p'}(\Omega)$ where $\dfrac{1}{p} + \dfrac{1}{p'} = 1$.

The problem (5.39) is a generalized Neumann problem for the non-linear (Laplacian) operator

(5.40) $$-\sum_j D_j (|D_j u|^{2p-2} D_j u) + u.$$

Example 5.4. Let Ω and Γ be as in the previous example and

(5.41) $$V = H^1_\circ(\Omega) \cap L^4(\Omega).$$

We have seen in Section 4 that V is a reflexive Banach space for its natural norm

(5.42) $$v \mapsto ||v||_{H^1(\Omega)} + ||v||_{L^4(\Omega)} = ||v||_V.$$

Consider the functional J on V given by

$$v \mapsto J(v) = \frac{1}{2}||v||^2_{H^1(\Omega)} + \frac{1}{4}||v||^4_{L^4(\Omega)} - (f, v)_{L^2(\Omega)}, \tag{5.43}$$

47 where $f \epsilon L^2(\Omega)$ is given. It is easily verified that J is twice G - differentiable and

$$J'(u, \varphi) = ((u, \varphi))_{H^1(\Omega)} + \int_\Omega (u^3 - f)\varphi dx, \ \forall \varphi, \ u \epsilon V.$$

(Hence J has a gradient)

$$J''(u; \varphi, \psi) = ((\psi, \varphi))_{H^1(\Omega)} + 3\int_\Omega u^2 \psi\varphi dx, \ \forall u, \varphi, \psi \epsilon V.$$

Thus $J''(u; \varphi, \varphi) > 0$ for $u \epsilon V, \varphi \epsilon V$ with $\varphi \neq 0$ which implies that J is strictly convex by Proposition 1. 3.2. As in the previous example we can show using Cauchy-Schwarz inequatliy (for the term $(f, v)_{L^2(\Omega)}$), that

$$J(v) \to +\infty \text{ as } ||v||_V \to +\infty.$$

Then by Theorem 1.2 the minimisation problem for J on V has a unique solution. An application of Green's formula shows that this unique solution (when it is regular) is the solution of the non-linear problem :

$$\begin{cases} -\triangle u + u + u^3 = f & \text{in } \Omega \\ u = 0 & \text{on } \Gamma \end{cases} \tag{5.44}$$

Remark 5.3. It is, ingeneral, difficult to solve the non-linear problem (5.43) numerically and it is easier to solve the equivalent minimisation problem for J given by (5.44).

Remark 5.4. All the functionals considered in the examples discussed in this section are strictly convex and they give rise to strongly monotone operators. We recall the following

48 **Definition 5.1.** An operator $A : U \subset V \to V'$ on a subset U of a normed vector space into its dual is called monotone if

$$< Au - Av, u - v >_{V' \times V} \geq 0 \text{ for all } u, v \epsilon U.$$

A is said to be strictly monotone if $< Au - Av, u - v >_{V' \times V} > 0$ for any pair of distinct elements $u, v \epsilon V$ (i.e. if $u \neq v$). (See, for instance, [44]).

Chapter 3

Minimisation Without Constraints - Algorithms

We have considered in the previous chapter results of theoretical nature 49 on the existence and uniqueness of solutions to minimisation problems and the solutions were characterized with the aid of the convexity and differ entiability properties of the given functional. Here we shall be concerned with the constructive aspects of the minimisation problem, namely the description of algorithms for the construction of sequences approximating the solution. We give in this chapter some algorithms for the minimisation problem in the absence of constraints and we shall discuss the convergence of the sequences thus constructed.

The algorithms (i.e. the methods for constructing the minimizing sequences) described below will make use of the differential calculus of functionals on Banach spaces developed in Chapter 1. We shall be mainly concerned with the following classes of algorithms:

(1) the method of descent and

(2) generalized Newton's method.

We shall mention the conjugate gradient method only briefly. The first class of methods mainly make use of the calculus of first order derivatives while the generalized Newton's method relies heavily on the calculus involving second order derivatives in Banach spaces.

Suppose V is a Banach space and $J : V \to \mathbb{R}$ is a functional on it. The algorithms consist in giving an interative procedure to solve the minimisation problem:

$$\text{to find } u\epsilon V, J(u) = \inf_{v\epsilon V} J(v).$$

50

Suppose J has unique global minimum u in V. We are interested in constructing a sequence u_k, starting from an arbitrary $u_\circ \epsilon V$, such that under suitable hypothesis on the functional J, u_k converges to u in V. First of all, since u is the unique global minimum the sequence $J(u_k)$ is bounded below by $J(u)$. It is therefore natural to construct u_k such that

(i) $J(u_k)$ is monotone decreasing

This will imply that $J(u_k)$ converge to $J(u)$. Further, if J admits a gradient G then we necessarily have $G(u) = 0$ so much so that the sequence u_k constructed should satisfy also the natural requirement that

(ii) $G(u_k) \to 0$ in V as $k \to \infty$

Our method can roughly be described as follows: If, for some k, u_k is already known then the next iterate u_{k+1} is determined by choosing suitably a parameter $\rho_k > 0$ and a direction $w_k (w_k \epsilon V$, $w_k \neq 0)$ and then taking

$$u_{k+1} = u_k - \rho_k w_k.$$

We shall describe, in the sequel, certain choices of ρ_k and w_k which will imply (i), (ii) which in turn to convergence of u_k to u. We shall call such choices of ρ_k, w_k convergent choices.

To simplify our discussion we shall restrict ourselves to the case of a Hilbert space V. However, all our considerations of this chapter remain valid for any reflexive Banach space with very minor changes and we shall not go into the details of this. As there will be no possibility of confusion we shall write $(\cdot, \cdot)$ and $\|\cdot\|$ for the inner product $(\cdot, \cdot)_V$ and $\|\cdot\|_V$ respectively. 51

1 Method of Descent

This method includes a class of algorithms for the construction of minimising sequences u_k. We shall begin with the following generalities in order to motivate and explain the principle involved in this method.

Let $J : V \to \mathbb{R}$ be a functional on a Hilbert space V.

1.1 Generalities

Starting from an initial value $u_\circ \epsilon V$ we construct u_k iteratively with the properties described in the introduction. Suppose u_k is constructed then to construct u_{k+1} we make two choices:

(1) a direction w_k in V called the "direction of descent"

(2) a real parameter $\rho = \rho_k$, and set $u_{k+1} = u_k - \rho_k w_k$ so that the sequence thus constructed has the required properties. The main idea in the choices of w_k and ρ_k can be motivated as follows:

Choice of w_k. We find $w_k \epsilon V$ with $||w_k|| = 1$ such that the restriction of J to the line in V passing through u_k and parallel to the direction w_k is decreasing in a neighbourhood of u_k: i.e. the function $\mathbb{R} \ni \rho \to J(u_k + \rho w_k) \epsilon \mathbb{R}$ is decreasing for $|\rho|$ sufficiently small. 52

If J is G-differentiable then we have by Taylor's formula

$$\begin{aligned} J(u_k + \rho w_k) &= J(u_k) + J'(u_k, \rho w_k) + \dots \\ &= J(u_k) + \rho J'(u_k, w_k) + \dots \end{aligned}$$

(by homogeneity of $\varphi \mapsto J'(u, \varphi)$). For $|\rho|$ small since the dominant term in this expansion is $\rho J'(u_k, w_k)$ and since we want $J(u_k + \rho w_k) \leq J(u_k)$ the best choice of w_k (at least locally) should be such that

$$\rho J'(u_k, w_k) \leq 0 \text{ and is largest in magnitude.}$$

If J has a gradient G then

$$\rho J'(u_k, w_k) = \rho(G(u_k), w_k) \leq 0$$

and our requirement will be satisfied if w_k is chosen proportional to $G(u_k)$ and opposite in direction. We note that, this may not be the best choice of w_k from global point of view. We shall therefore write

$$J(u_k - \rho w_k) \text{ with } \rho > 0$$

so that $J(u_k - \rho w_k) \searrow$ as k increases for $\rho > 0$ small enough.

Choice of $\rho(= \rho_k)$. Once the direction of descent w_k is chosen then the iterative procedure can be done with a constant $\rho > 0$. It is however more suitable to do this with a variable ρ. We shall therefore choose $\rho = \rho_k > 0$ in a small interval with the property $J(u_k - \rho_k w_k) < J(u_k)$ and set

$$u_{k+1} = u_k - \rho_k w_k.$$

We do this in several steps. Since,

$$j = \inf_{v \in V} J(v) \leq J(u_{k+1}) \leq J(u_k)$$

53 we have

$$J(u_k) - J(u_{k+1}) \geq 0 \text{ and } \lim_{k \to +\infty} (J(u_k) - J(u_{k+1})) = 0$$

because $J(u_k)$ is decreasing and bounded below. If J is differentiable then Taylor's formula implies that

$$J(u_k) - J(u_{k+1}) \text{ behaves like } J'(u_k, u_{k+1} - u_k) = \rho_k J'(u_k, w_k)$$

so that it is natural to require that

$$\rho_k > 0, \rho_k J'(u_k, w_k) \to 0 \text{ as } k \to +\infty.$$

Roughly speaking, we shall say that the choice of ρ_k is a "convergent choice" if this condition implies $J'(u_k, w_k) \to 0$ as $k \to +\infty$. If, moreover, J has a gradient G then choice of the direction of descent w_k is a "convergent choice" if $J'(u_k, w_k) = (G(u_k), w_k) \to 0$ implies that $||G(u_k)|| \to 0$ as $k \to +\infty$.

The above considerations lead us to the following definitions which we shall use in all our algorithms and all our proofs of convergence.

Definition 1.1. The choice of ρ_k is said to be convergent if the conditions

$$\begin{cases} \rho_k > 0, u_{k+1} = u_k - \rho_k w_k \\ J(u_k) - J(u_{k+1}) > 0, \lim_{k\to+\infty}(J(u_k) - J(u_{k+1})) = 0 \end{cases}$$

imply that

$$\lim_{k\to+\infty} J'(u_k, w_k) = 0.$$

Suppose J has a gradient G in V.

Definition 1.2. The choice of the direction w_k is said to be convergent if the conditions

$$w_k \epsilon V, J'(u_k, w_k) > 0. \lim_{k\to+\infty} J'(u_k, w_k) = 0$$

imply that 54

$$\lim_{k\to+\infty} ||G(u_k)|| = 0.$$

1.2 Convergent choice of the direction of descent w_k

This section is devoted to some algorithms for convergent choices of w_k. In each case we show that the choice of w_k described is convergent in the sense of Definition 1.2

w-Algorithm 1. We assume that J has a gradient G in V. Let a real number α be given with $0 < \alpha \leq 1$. We choose $w_k \epsilon V$ such that

$$\begin{cases} (G(u_k)/||G(u_k)||, w_k) \geq \alpha > 0. \\ ||w_k|| = 1. \end{cases} \tag{1.1}$$

Proposition 1.1. *w-Algorithm 1 gives a convergent choice of* w_k.

Proof. We can write

$$J'(u_k, w_k) = (G(u_k), w_k)$$

so that by (1.1)

$$J'(u_k, w_k) \geq \alpha ||G(u_k)|| > 0$$

and hence

$$J'(u_k, w_k) \to 0 \text{ implies that } ||G(u_k)|| \to 0 \text{ as } k \to +\infty.$$

□

We note that (1.1) means that the angle between w_k and $G(u_k)$ lies in $]-\pi/2, \pi/2[$ and the cosine of this angle is bounded away from 0 by α.

w-Algorithm 2 - Auxiliary operatoe method.

This algorithm is a particular case of w-algorithm 1 but very much more used in practice.

55 Assume that J has a gradient G in V.

Let, for each k, $B_k \epsilon \mathscr{L}(V, V)$ be an such that

$$(1.2)\quad \begin{cases} B_k \text{ are uniformly bounded: there exists a constant } \gamma > 0 \\ \text{ such that } \quad ||B_k\psi|| \leq \gamma||\psi|| : \psi\epsilon V. \\ B_k \text{ are uniformly } V\text{-coercive: there exists a constant } \alpha > 0 \\ \text{ such thaat } \quad (B_k\psi, \psi) \geq \alpha||\psi||^2, \psi\epsilon V. \end{cases}$$

Let us choose

$$(1.3)\qquad w_k = B_k G(u_k)/||B_k G(u_k)$$

Proposition 1.2. *The choice (1.3) of w_k is convergent.*

Proof. As before we calculate

$$J'(u_k, w_k) = (G(u_k), w_k) = (G(u_k), B_k G(u_k)/||B_k G(u_k)||)$$

which, by uniform coercivity of B_k, is

$$\begin{aligned} &\geq \alpha||G(u_k)||^2/||B_k G(u_k)|| \\ &\geq \alpha\gamma^{-1} G(u_k) \text{ by uniform boundedness of } B_k. \end{aligned}$$

□

This immediatly implies that

$$J'(u_k, w_k) > 0 \text{ and if } J'(u_k, w_k) \to 0 \text{ then } \|G(u_k)\| \to 0$$

and hence the choice of w_k is convergent.

Moreover, again by (1.3), we get

$$(G(u_k)/\|G(u_k)\|, w_k) = (G(u_k)/\|G(u_k)\|, B_k G(u_k)/\|B_k G(u_k)\|) \geq \alpha\gamma^{-1} > 0,$$

which means that this algorithm is a particular case of w-Algorithm 1.

Remark 1.1. In certain (for example, when B_k are symmetric operators satisfying (1.2)) this method is equivalent to making a change of variables and taking as the direction of descent the direction of the gradient of J in the new variables and then choosing w_k as the inverse image of this direction in the original coordinates. 56

Consider the functional $J : V = \mathbb{R}^2 \to \mathbb{R}$ of our model problem of Chapter 1, §7:

$$\mathbb{R}^2 \ni v \mapsto J(v) = \frac{1}{2}a(v, v) - L(v) = \frac{1}{2}(Av, v)_{\mathbb{R}^2} - (f, v)_{\mathbb{R}^2} \epsilon \mathbb{R}.$$

Since $a(\cdot, \cdot)$ is a positive definite quadratic form, $\{v \epsilon \mathbb{R}^2, J(v) =$ constant $\}$ represents an ellipse. B_k can be chosen such that the change of variable effected by B_k transforms such an ellipse into a circle where the gradient direction is well-known i.e. the direction of the radial vector through u_k (in the new coordinates).

w-Algorithm 3 - Conjugate gradient method

There are several algorithms known in the literature under the name of conjugate gradient method. We shall, however, describe only one of the algorithms which generalizes the conjugate gradient method in the finite dimensional spaces. (See [20] [22] and [24]).

Suppose the functional J admits a gradient $G(u)$ and a Hessian $H(u)$ everywhere in V. Let $u_\circ \epsilon V$ be arbitrary. We choose $w_\circ = G(u_\circ)/\|G(u_\circ)\|$ (We observe that we may assume $G(u_\circ) \neq 0$ unless $u_\circ$ itself happens to be the required minimum). If u_{k-1}, w_{k-1} are already known then we choose $\rho_{k-1} > 0$ to be a points of minimum of the real valued function

$$\mathbb{R}_+ \ni \rho \mapsto J(u_{k-1} - \rho w_{k-1}) \epsilon \mathbb{R}$$

$$\text{i.e. } \rho_{k-1} > 0 \text{ and } J(u_{k-1} - \rho_{k-1} w_{k-1}) = \inf_{\rho>0} J(u_{k-1} - \rho w_{k-1}).$$

57 Since J is G-differentiable this real valued function of ρ is differentiable everywhere in $\mathbb{R}_+$ and

$$\frac{d}{d\rho} J(u_{k-1} - \rho w_{k-1})|_{\rho=\rho_{k-1}} = 0,$$

which means that, if we set

$$(1.4)_1 \qquad u_k = u_{k-1} - \rho_{k-1} w_{k-1}$$

then we have

$$(1.5) \qquad (G(u_k), w_{k-1}) = 0.$$

Now we define a vector $\widetilde{w}_k \epsilon V$ by

$$\widetilde{w}_k = G(u_k) + \lambda_k w_{k-1}$$

where $\lambda_k \epsilon \mathbb{R}$ is chosen such that

$$(H(u_k)\widetilde{w}_k, w_{k-1}) = 0$$

Hence w_k is given by

$$(1.4)_2 \qquad \lambda_k = -\frac{(H(u_k)G(u_k), w_{k-1})}{(H(u_k)w_{k-1}, w_{k-1})}.$$

We remark that in applications we usually assume that $H(u)$ (for any $u \epsilon V$) defines a positive operator and hence the denominator in $(1.4)_2$ above i non-zero (see Remark 1.2 below). Then the vector

$$(1.4)_3 \qquad w_k = \widetilde{w}_k / ||\widetilde{w}_k||$$

defines the direction of descent at the k-th stage of the algorithm.

This algorithm is called conjugate gradient method because of the following remark.

58 **Remark 1.2.** Two directions φ and ψ are said to be conjugate with respect to a positive definite quadratic form $a(\cdot,\cdot)$ on V if $a(\varphi,\psi) = 0$. In this sense, if $H(u_k)$ defines positive definite quadratic form (i.e. $H(u_k)$ is a symmetric positive operator on V) two consecutive choices of directions of descent w_{k-1}, w_k are conjugate with respect to the quadric $(H(u_k)w, w) = 1$. We recall that in the plane $\mathbb{R}^2$ such a quadric represents an ellipse and two directions φ, ψ in the plane are said to be conjugate with respect to such an ellipse if $(H(u_k)\varphi, \psi) = 0$.

Now we have the following

Proposition 1.3. *Suppose that the functional H admits a gradient $G(u)$ and a Hessian $H(u)$ everywhere in V and suppose further that there exist two constants $C_\circ > 0, C_1 > 0$ such that*

(i) $(H(u)\varphi,\varphi) \geq C_\circ||\varphi||^2$ *for all* $u, \varphi\epsilon V$ *and*

(ii) $|(H(u)\varphi,\psi)| \leq C_1||\varphi||||\psi||$ *for all* $u, \varphi, \psi\epsilon V$.

Then the w-Algorithm 3 defines a convergent choice of the w_k.

Proof. It is enough to verify that w_k satisfies the condition (1.1). First of all, in view of the definition of $\widetilde{w}_k$ and (1.5) we have

$$(G(u), \widetilde{w}_k) = ||G(u_k)||^2$$

so that

$$(G(u_k)/||G(u_k)||, w_k) = ||G(u_k)||||\widetilde{w}_k||^{-1}.$$

□

We shall show that this is bounded below by a constant $\alpha > 0$ (independent of k).

For this, we get, again using the definition of $\widetilde{w}_k$, $(1.4)_2$ and (1.5)

$$||\widetilde{w}_k||^2 = ||G(u_k)||^2 + \lambda_k^2||w_{k-1}||^2.$$

Here, in view of the assumptions (i) and (ii) we find that

$$\lambda_k^2||w_{k-1}||^2 = \frac{(H(u_k)G(u_k), w_{k-1})^2}{(H(u_k)w_{k-1}, w_{k-1})^2}||w_{k-1}||^2$$

$$\leq (C_{\circ}^{-1}C_1||G(u_k)||)^2$$

so that 59

$$||\widetilde{w}_k||^2 \leq ||G(u_k)||^2(1 + C_{\circ}^{-2}C_1^2).$$

Hence, taking the constant $\alpha > 0$ to be $(1 + C_{\circ}^{-2}C_1^2)^{-\frac{1}{2}}$ we get

$$||G(u_k)||||\widetilde{w}_k||^{-1} > \alpha > 0$$

which proves the assertion.

1.3 Convergent Choices of ρ_k

We shall describe in this section some algorithms for the choice of the parameter ρ_k and we shall prove that these choices are convergent in the sense of our Definition 1.1.

Given the idrection w_k of descent at the k^{th} stage we are interested in points of the type

$$u_k - \rho w_k, \rho > 0,$$

and therefore all out discussions of this section are as if we have functions of a single real variable ρ defined in $\mathbb{R}_+$.

We shall use the following notation throughout this and the next sections in order to simplify our writing:

Notation

$$\begin{cases} J(u_k - \rho w_k) & = J_{\rho}^{k} \text{ for } \rho > 0, \\ J(u_k) & = J_{\circ}^{k}, \end{cases}$$

$$J(u_k) - J(u_k - \rho w_k) = J_{\circ}^{k} - J_{\rho}^{k} = \triangle J_{\rho}^{k}, \rho > 0.$$

$$\begin{cases} J'(u_k - \rho w_k, w_k) & = J'^{k}_{\rho} \text{ for } \rho > 0. \\ J'(u_k, w_k) & = J'^{k}_{\circ}. \end{cases}$$

Smilarly, when J has gradient $G(u)$ and a hessian $H(u)$ at every points u in V, we write

$$\begin{cases} G(u_k - \rho w_k) & = G_{\rho}^{k} \text{ for } \rho > 0. \\ G(u_k) & = G_{\circ}^{k} \end{cases}$$

60 and

$$\begin{cases} H(u_k - \rho w_k) & = H^k_\rho \text{ for } \rho > 0, \\ H(u_k) & H^k_\circ \end{cases}$$

We shall make the following two hypothesis throughout this section.

Hypothesis $(H1)$: $\lim\limits_{||v||\to\infty} J(v) = +\infty$.

Hypothesis $(H2)$: J has a gradient $G(u)$ everywhere in V and satisfies a (uniform) Lipschitz condition on every bounded subset of V: for every bounded set K of V there exists a constant $M_K > 0$ such that

$$||G(u) - G(v)|| \leq M_K||u - v|| \text{ for all } u, v\epsilon K.$$

In particular, if J has a Hessian $H(u)$ everywhere in V and if $H(u)$ is bounded on bounded sets of V then an application of Tayler's formula to the mapping $V \ni u \mapsto G(u)\epsilon V' = V$ shows that J satisfies the hypothesis $(H2)$. In fact, if $u, v\epsilon V$ then

$$\begin{aligned} ||G(u) - G(v)|| &= \sup_\varphi |(G(u) - G(v), \varphi)|/||\varphi|| \\ &= \sup_\varphi |(H(u + \theta(u - v))(u - v), \varphi)|/||\varphi|| \leq const.||u - v||, \end{aligned}$$

since $u, v\epsilon K$ and $\theta\epsilon]0, 1[$ imply that $v + \theta(u - v)$ is also bounded and hence $H(v + \theta(u - v))$ is bounded uniformly for all $\theta\epsilon]0, 1[$.

Now suppose given a $u_\circ\epsilon V$ at the beginning of the algorithm. Starting from $u_\circ$ we shall construct a sequence u_k such that $J(u_k)$ is decreasing and so we have $J(u_k) \leq J(u_\circ)$. We are interested in points of the type $u_k - \rho w_k$ such that $J(u_k - \rho w_k) \leq J(u_k)$.

We shall now deduce some immediate consequences of the hypothesis H1 and H2, which will be constantly used to prove the convergence of the choice of ρ_k given by the algorithms of this section.

Let us denote by U the subset of V:

$$U = \{v|v\epsilon V; J(v) \leq J(u_\circ)\}.$$

The set U is bounded in V. In fact, if U is not bounded then we can find a sequence $v_j\epsilon U$ such that $||v_j|| \to +\infty$. Then $J(v_j) \to +\infty$ by 61 Hupothesis H1 and this is impossible since $v_j\epsilon U$.

We are thus interested in constructing a sequence u_k such that

$$u_k \epsilon U \text{ and } J(u_k) \searrow .$$

Also since by requirement $J(u_k - \rho w_k) \leq J(u_k)$ it follows that $u_k - \rho w_k \epsilon U$ and then ρ will be bounded by diam U; for, we find using triangle inequality:

$$0 < \rho = ||\rho w_k|| = ||u_k - (u_k - \rho w_k)|| \leq diamU.$$

Let us denote the constant $M_U > 0$ given by Hypothesis H2 for the bounded set U by M.

Now the points $u_k - \rho w_k, u_k - \mu w_k$ belongs to U if $\rho, \mu \geq 0$ are chosen sufficiently small. Then

$$\begin{aligned} ||G^k_\rho - G^k_\mu|| &= ||G(u_k - \rho w_k) - G(u_k - \mu w_k)|| \\ &\leq M|\rho - u|||w_k|| = M|\rho - \mu|; \end{aligned}$$

i.e. we have,

$$(1.6) \qquad \begin{cases} ||G^k_\rho - G^k_\mu|| & \leq M|\rho - \mu| \\ ||G^k_\rho - G^k_\circ|| & \leq M\rho \end{cases}$$

Since $J'^k_\rho = J'(u_k - \rho w_k, w_k) = (G(u_k - \rho w_k), w_k) = (G^k_\rho, w_k)$ we also find from (1.6) that

$$(1.7) \qquad \begin{cases} |J'^k_\rho - J'^k_\mu| & \leq M|\rho - \mu| \\ |J'^k_\rho - J'^k_\circ| & \leq M\rho. \end{cases}$$

We shall suppress the index K when there is no possibility of confusion and simply write G_ρ, J_ρ, J'_ρ etc. respectively for $G^k_\rho, J^k_\rho, J'^k_\rho$ etc.

62 By Taylor's expansion we can write

$$J_\rho = J(u - \rho w) = J(u) - \rho J'(u - \overline{\rho} w, w)$$

for some $\overline{\rho}$ such that $0 < \overline{\rho} < \rho$. i.e. we can write

$$(1.8) \qquad J_\rho = J_\circ - \rho J'_{\overline{\rho}}.$$

We can rewrite (1.8) also as

$$J_\rho = J_\circ - \rho J'_\circ + \rho(J'_\circ - J'_{\overline{\rho}}),$$

which together with (1.7) gives

$$J_\rho \leq J_\circ - \rho J'_\circ + M\rho\overline{\rho},$$

that is, since $0 < \overline{\rho} < \rho$

$$J_\rho \leq J_\circ - \rho J'_\circ + M\rho^2. \tag{1.9}$$

We shall use (1.8) and (1.9) in the following form

$$\triangle J_\rho = \rho J'_{\overline{\rho}}, \tag*{(1.8)$'$}$$

$$\rho J'_\circ - M\rho^2 \leq \triangle J_\rho. \tag*{(1.9)$'$}$$

We are now in a position to describe the algorithms for convergent choices of the parameter ρ_k.

ρ- Algorithm 1. Consider the two functions of $\rho > 0$ given by

$$J_\rho = J(u_k - \rho w_k) \text{ and } T(\rho) = J_\circ - \rho J'_\circ + M\rho^2.$$

Then $J_\circ = T(0)$ and (1.9) says that $J_\rho \leq T(\rho)$ for all $\rho > 0$. Geometrically the curve $y = J_\rho$ lies below the parabola $y = T(\rho)$ for $\rho > 0$ in the (ρ, y) -plane. Let $\hat{\rho} > 0$ be the points at which the function $T(\rho)$ has a minimum. Then $\dfrac{dT}{d\rho}|_{\rho=\hat{\rho}} = 0$ implies $-J'_\circ + 2M\hat{\rho} = 0$ so that we have

$$\hat{\rho} = J'_\circ/2M, T(\hat{\rho}) = \inf_{\rho>0} T(\rho). \tag{1.10}$$

63

Let C be a real number such that

$$0 < C \leq 1. \tag{1.11}$$

We choose $\rho = \rho_k$ in the interval $[C\hat{\rho}, (2 - C)\hat{\rho}]$, i.e.

$$C \leq \rho/\hat{\rho} \leq (2 - C). \tag{1.12}$$

Then we have the

Proposition 1.4. *Under the hypothesis* $(H1)$, $(H2)$ *the choice (1.12) of* $\rho = \rho_k$ *is a convergent choice.*

Proof. Since T has its minimum at the points $\rho = \hat{\rho}$ we have by (1.11) $C\hat{\rho} \leq \hat{\rho} \leq (2-C)\hat{\rho}$. Moreover $T(\rho)$ decreases in the interval $[0, \hat{\rho}]$ while it increases in the interval $[\hat{\rho}, (2-C)\hat{\rho}]$ as can easily be checked. Hence, if ρ satisfies (1.12) then we have two cases:

$$\begin{cases} T_\rho \leq T_{C\hat{\rho}} \text{ if } C\hat{\rho} \leq \rho \leq \hat{\rho} \text{ and} \\ T_\rho \leq T_{(2-C)\hat{\rho}} \text{ if } \hat{\rho} \leq \rho \leq (2-C)\hat{\rho}. \end{cases}$$

□

Since $T_{C\hat{\rho}} = J_\circ - CJ'_\circ/2M.J'_\circ + M(CJ'_\circ/2M)^2 = J_\circ - (2-C)$ $C(J'_\circ)^2/4M,)$

$$T_{(2-C)\hat{\rho}} = J_\circ - (2-C)J'_\circ/2MJ'_\circ + M((2-C)J'_\circ/2M)^2 = J_\circ - (2-C)C(J'_\circ)^2/4M$$

using the value of $\hat{\rho}$ given by (1.10) and since $J_p \leq T_p$ for all $\rho > 0$ we find that (in either of the above cases)

$$J_\rho \leq T_\rho \leq J_\circ - (2-C)C(J'^2_\circ)/4M.$$

This immediately implies that

$$C(2-C)(J'_\circ)^2/4M \leq \triangle J_{\rho.} \tag{1.13}$$

64 In order to show that the choice (1.12) is convergent we see that (1.13) is nothing but

$$C(2-C)/4M(J'(u_k, w_k))^2 \leq J(u_k) - J(u_k - \rho w_k) \leq J(u_k) - J(u_{k+1})$$

since $J(u_{k+1}) = J(u_k - \rho_k w_k) = \inf_{\rho>0} J(u_k - \rho w_k)$ i.e. $J(u_{k+1}) \leq J^k_\rho$. Hence if $J(u_k) - J(u_{k+1}) \to 0$ then $J'(u_k, w_k) \to 0$ as $k \to +\infty$, which proves that the choice of ρ_k such that

$$C \leq \rho_k \hat{\rho}_k^{-1} \leq 2 - C \text{ where } \hat{\rho}_k = J'(u_k, w_k)/2M$$

is a convergent choice.

ρ-Algorithm 2. The constant M in the ρ-Alogorithm 1 is not in general known a priori. This fact may cause difficulties in the sense that if we start with an arbitrarily large $M > 0$ then by (1.12) ρ_k will be very small and so the scheme may not converge sufficiently fast. We can get over this difficulty as described in the following algorithm, which does not directly involve the constant M and which can be considered as a special case of ρ-Algorithm 1. But for this algorithm we need the additional assumption that J is convex.

Hypothesis H3. The functional J is convex.

We suppose that, for some fixed $h > 0$, we have

$$\begin{cases} J_\circ > J_h > J_{2h} > J_{2h} > \cdots > J_{mh}, \\ J_{mh} < J_{(m+1)h}, \text{ for some integer } m \geq 2. \end{cases} \tag{1.14}$$

Since J is convex and has its minimum in $\rho > 0$ such an $m \geq 2$ always exists.

Proposition 1.5. *If J satisfies the hypothesis H1, H2, H3 then any choice of $\rho(= \rho_k)$ such that*

$$(m-1)h \leq \rho \leq mh \tag{1.15}$$

is a convergent choice. 65

Proof. Let $\widetilde{\rho} > 0$ be a point where J_ρ attains its minimum. Then $J'_{\widetilde{\rho}} = 0$, $J_{\widetilde{\rho}} \leq J_\rho$ for all $\rho > 0$ and by (1.14) we should have

$$(m-1)h \leq \widetilde{\rho} \leq (m+1)h. \tag{1.16}$$

Then (1.7) will imply

$$0 < J'_\circ = |J'_{\widetilde{\rho}} - J'_\circ| \leq M$$

and thus we find

$$2\hat{\rho} = J'_\circ/M \leq \widetilde{\rho} \tag{1.17}$$

and

$$2\hat{\rho}/(m+1) \leq h. \tag{1.18}$$

□

This, together with the fact that $m \geq 2$, will in turn imply

$$2\hat{\rho}/3 \leq (m-1)h.$$

As J_ρ decreasesd in $0 \leq \rho < mh$ we get

$$\triangle J_{(m-1)h} = J_\circ - J_{(m-1)h} \geq J_\circ - J_{2\hat{\rho}/3} = \triangle J_{(2\hat{\rho}/3)}.$$

If we now apply the ρ-Algorithm 1 with $C = 2/3$ in (1.12) and in (1.13) then we obtain, from the above inequality,

$$\triangle J_{(m-1)h} \geq 2/9M(J'{}_\circ)^2, \tag{1.19}$$

which proves that $\rho = (m-1)h$ is a convergent choice. Similarly, if $\rho \epsilon [(m-1)h, mh]$ (i.e. (1.15)) then the same argument shows that

$$\triangle J_\rho \geq \triangle J_{(m-1)h} \geq 2/9M(J'{}_\circ)^2, \tag{1.20}$$

66 and hence any $\rho_k = \rho$ satisfying (1.15) is again a convergent choice.

Some Generalizations of ρ-Algorithm 2.

In the above algorithm a suitable initial choice of $h > 0$ has to be made. But such an h can be either too large or too small and if for example h is too small then the procedure may become very long to use numerically. In order to over come such diffeculties we can generalize ρ-Algorithm 2 as follows.

If the initial value of $h > 0$ is too small we can repeat our arguments above with (1.14) replaced by

$$(1.14)' \qquad \begin{cases} J_\circ > J_{ph} > J_{p^2h} > J_{p^3h} > \cdots > J_{p^mh}, \\ J_{p^mh} < J_{p^{(m+1)}h}, \text{ for some integer } m \geq 2 \end{cases}$$

and if the initial value of h is too large we can compute J at the points $\dfrac{h}{\rho}, \dfrac{h}{\rho^2}, \dfrac{h}{\rho^3}, \cdots$ where p is an integer ≥ 2. Every such procedure gives a new algorithm for a convergent choice of $\rho_k = \rho$.

ρ-Algorithm 3. We have the following

Proposition 1.6. *Assume that J satisfies the hypothesis H1 - H3. If $h > 0$ is such that*

$$\begin{cases} \triangle J_h/h \geq (1-C)J'_\circ, \\ \triangle J_{2h}/2h < (1-C)J'_\circ \end{cases} \tag{1.21}$$

with some constant C, $0 < C < 1$ then $(\rho_k =)\rho = h$ is a convergent choice.

Proof. From the inequality ((1.9)$'$ and the second inequality in (1.21) we get

$$2hJ'_\circ - (2h)^2M \leq \triangle J_{2h} < (1-C)2hJ'_\circ$$

and hence

$$C\hat{\rho} = CJ'_\circ/2M \leq h.$$

□ 67

Now the first inequality in (1.21) implies

$$\triangle J_h \geq h(1-C)J'_\circ \geq C(1-C)(J'_\circ)^2/2M, \tag{1.22}$$

which proves that $\rho = h$ is a convergent choice since $\triangle J_h = J(u_k) - J(u_k - hw_k) \to 0$ implies that $J'_\circ = J'(u_k, w_k) \to 0$ as $k \to \infty$.

We shall now show that there exists an $h > 0$ satisfying (1.21). We consider the real valued function

$$\psi(\rho) = \triangle J_\rho/\rho - (1-C)J'_\circ$$

of ρ on $\mathbb{R}_+$ and observe the following two facts:

(1) $\psi(\rho) \geq 0$ for $\rho > 0$ sufficiently small. In fact, since $\triangle J_\rho/\rho \to J'_\circ > 0$ we have $|\triangle J_\rho/\rho - J'_\circ| < CJ'_\circ$ for $\rho > 0$ sufficiently small, which, in particular, implies the assertion.

(2) $\psi(\rho) < 0$ for $\rho > 0$ sufficiently large. For this, since u_k, w_k are already determined (at the $(k+1)$th stage of the algorithm) we see that $\|\rho w_k\| \to +\infty$ and hence $\|u_k - \rho w_k\| \to +\infty$. Then, by hypothesis $(H1)$,

$$J(u_k - \rho w_k) \to +\infty \text{ as } \rho \to +\infty$$

so much so that

$$\triangle J_\rho \leq 0 < \rho(1-C)J'_\circ \text{ for } \rho > 0$$

sufficiently large, which implies the assertion.

Thus the sign of ψ changes from positive to negative, say at some $\rho = h_\circ > 0$. Then, for instance, $h = 3h_\circ/4$ will satifsy our requirement (1.21).

More precisely, we can find h satisfying (1.21) in the following iterative manner. Assume that $0 < C < 1$ is given.

First of all we shall choose a τ arbitrarily (> 0) and we compute the difference quotient $\triangle J_\tau/\tau$. This is possible since all the quantities are known. Then there are two possible cases that can arise namely, either

(a) $\triangle J_\tau/\tau \geq (1-C)J'_\circ$

or(b) $\triangle J_\tau/\tau < (1-C)J'_\circ$.

68 Suppose (a) holds. Then we compute $\triangle J_{2\tau/2\tau}$ and we will have to consider again two possibilities:

either$(a)_1$ $\triangle J_{2\tau/2\tau} < (1-C)J'_\circ$,

or$(a)_2$ $\triangle J_{2\tau/2\tau} \geq (1-C)J'_\circ$.

If we have the first possibility $(a)_1$ then we are through we can choose $h = \tau$ itself. If on the order hand $(a)_2$ holds then we repeat this argument with τ replaced by 2τ.

Next suppose (b) holds. We can consider two possible cases:

either$(b)_1$ $\triangle J_{\tau/2}|\tau/2 \geq (1-C)J'_\circ$,

or$(b)_2$ $\triangle J_{\tau/2}|\tau/2 < (1-C)J'_\circ$.

Once again, in case $(b)_1$ holds we are through and we can choose $h = \tau/2$. In case $(b)_2$ holds we repeat this argument with τ replaced by $\tau/2$.

Remark 1.2. It was proposed by Goldstein (see [21]) that the initial value of τ can be taken to be taken to be $\tau = J'_\circ$.

ρ-**Algorithm 4.** We have the following

Proposition 1.7. *If there is a $\widetilde{\rho}$ such that*

$$\begin{cases} \widetilde{\rho} & > 0, \\ J\widetilde{\rho} & \leq J_\rho \quad \rho \epsilon [0, \widetilde{\rho}] \\ J'_{\widetilde{\rho}} & = 0 \end{cases} \tag{1.23}$$

then $\rho = \widetilde{\rho}$ is a convergent choice.

Proof. We have, by the last condition in (1.23) together with the estimate (1.7).

$$J'_\circ = |J'_{\widetilde{\rho}} - J'_\circ| \leq M\widetilde{\rho}$$

and hence $\hat{\rho} \leq 2\hat{\rho} = J'_\circ/M \leq \widetilde{\rho}$ using the value of $\hat{\rho}$ given by (1.10). The condition (1.23) that $J_{\widetilde{\rho}}$ is a minimum in $[0, \widetilde{\rho}]$ implies $J_{\widetilde{\rho}} \leq J_{\hat{\rho}}$ and therefore

$$\triangle J_{\hat{\rho}} = J_\circ - J_{\hat{\rho}} \leq J_\circ - J_{\widetilde{\rho}} = \triangle J_{\widetilde{\rho}}.$$

□

On the other hand, taking $C = 1$ in (1.22) we find that

$$J'^2_\circ/2M \leq \triangle J_{\hat{\rho}} \leq \triangle J_{\widetilde{\rho}} \tag{1.24}$$

which proves that $\rho = \widetilde{\rho}$ is a convergent choice.

We shall conclude the discussion of convergent choices of ρ_k for ρ **69** by observing that other algorithms for convergent choices of ρ can be obtained making use of the following remarks.

Remark 1.3. We recall that in ρ-Algorithm 1 we obtained convergent choices of ρ to be close to $\hat{\rho}$ (i.e. $C \leq \rho/\hat{\rho} \leq 2 - C$) where $\hat{\rho}$ is the points of minimum of the curve $y = T(\rho)$, which is a polynomial of degree 2. This method can be generalised to get other algorithms as follows:

Starting from u_0 if we have found u_k and the direction of descent w_k then $J_\circ = J(u_k)$, $J'_\circ = J'(u_k, w_k) = (G(u_k), w_k)$ are known. Now if we are given two more points (say h and $2h$) we know the values of J at these points also. Thus we know values at 3 points and the initial slope (i.e. $J'_\circ$). By interpolation we can find a polynomial of degree 3 from these. To get an algorithm for a convergent choice of ρ we can choose ρ to be close to the point where such a polynomial has a minimum. Similar method works also polynomial of higher degress if we are given more number of points by using interpolation.

Remark 1.4. In all our proofs for convergent choices of ρ we obtained an estimate of the type:

$$\gamma(J'_\circ)^2 \leq \triangle J_\rho$$

where γ is a constant > 0. For instance $\gamma = 2/9M$ in (1.20).

1.4 Convergence of Algorithms

In the previous we have given some algorithms to construct a minimising sequence for the solution of the minimisation problem:

Problem ***P*****.** to find $u \epsilon V$, $J(u) \leq J(v)$, $\forall v \epsilon V$.

In this section we shall prove that under some reasonable assumptions on the functional J any combination of w-algorithms and ρ - algorithms yield a convergent algorithm for the construction of the minimising sequence u_k and such a sequence converges to a solution of the problem P. 70

Let $J : V \to \mathbb{R}$ be a functional on a Banach space V. The following will be the assumptions that we shall make on J:

(H0) J is bounded below: there exists a real number j such that $-\infty < j \leq J(v), \forall v \epsilon V$.

(H1) $J(v) \to +\infty$ as $||v|| \to +\infty$.

(H2) J has a gradient $G(u)$ everywhere in V and $G(u)$ is bounded on every bounded subset of V: if K is a bounded set in V then there exists a constant $M_K > 0$ such that $||G(u)|| \leq M_K$ for all $u \epsilon K$.

(H3) J is convex.

(H4) V is a reflexive Banach space

(H5) J is strictly convex

(H6) J admits a hessian $H(u)$ everywhere in V which is V-coercive: there exists a constant $\alpha > 0$ such that

$$< H(u)\varphi, \varphi >_{V' \times V} \geq \alpha ||\varphi||_V^2, \forall u \epsilon V \text{ and } \forall \varphi \epsilon V.$$

As in the previous sections we shall restrict ourselves to the case of a Hilbert space V and all our arguments remain valid with almost no changes. We have the following result.

Theorem 1.1. *(1) If the hypothesis H0, H1, H2 are satisfied and if u_k isa sequence constructed using any of the algorithms:*

$$w - Algorithm\ i, i = 1, 2$$
$$\rho - Algorithm\ \ j, j = 1, 3, 4$$

then

$$||G(u_k)|\ \ \rightarrow 0 \text{ as } k \rightarrow +\infty.$$

(2) If the hypothesis H0 - H4 hold and if u_k are constructed using the algorithm $i = 1, 2$, $j = 1, 2, 3, 4$ then all algorithm have the following property: 71

(a) the sequence u_k has a weak cluster point;

(b) any weak cluster point is a solution of the problem P.

(3) If the hypothesis H0 - H5 are satisfied then

(a) the Problem P has a unique solution $u \epsilon V$,

(b) If u_k is constructed using any of the algorithms $i = 1, 2$, $j = 1, 2, 3, 4$ then

$$u_k \rightharpoonup u \text{ as } k \rightarrow +\infty.$$

(4) Under the hypothesis H0 - H6 we have

(a) the Problem P has a unique solution $u \in V$,

(b) if the sequence u_k is constructed using any of the algorithms $i = 1, 2, 3,\ j = 1, 2, 3, 4$ then

$$u_k \to u \text{ and moreover } ||u_k - u|| \leq 2/\alpha ||G(u_k)||\ \forall k.$$

Proof. (1) Since by $(H0)$, $J(u_k)$ is a decreasing sequence bounded below: $j \leq J(u_{k+1}) \leq J(u_k) \leq J(u_\circ), \forall k$ it follows that

$$\lim_{k\to+\infty} (J(u_k) - J(u_{k+1})) = 0.$$

Since by the ρ-Algorithms $j (j = 1, 3, 4)$ the choice of $\rho = \rho_k$ in $u_{k+1} = u_k - \rho w_k$ is a convergent choice we see that

$$J'(u_k, w_k) \to 0,\ \text{as } k \to +\infty.$$

Now since the choine (i) w_k is convergent $(i = 1, 2)$ this implies that

$$||G(u_k)|| \to 0 \text{ as } k \to +\infty.$$

(2) As we have seen in the previous section, if $u_\circ \epsilon V$ then the set $U = \{v | v\epsilon V, J(v) \leq J(u_\circ)\}$ is bounded by $(H1)$ and since

$$J(u_{k+1}) \leq J(u_k) \leq \cdots \leq J(u_\circ)\ \forall k$$

72 all the $u_k \epsilon U$ and thus u_k is a bounded sequence. Then $(H4)$ implies that u_k has a weak cluster points which proves (a) i.e. $\exists a$ subsequence $u_{k'}$ such that $u_{k'} \to u$ in V as $k' \to +\infty$. Now by $(H3)$ and by Proposition 1. 3.1 on convex functionals

(1.25) $J(v) \geq J(u_{k'}) + J'(u_{k'}, v - u_{k'})$ for any $v\epsilon V$ and any k'.

Then, by $(H2)$, $J'(u_{k'}, v - u_{k'}) = (G(u_{k'}), v - u_{k'})$. But here $v - u_{k'}$ is a bounded sequence and since all the assumptions of Part 1 of the theorem are satisfies $||G(u_{k'})|| \to 0$ i.e. $G(u_{k'}) \to 0$ strongly in V. Hence

$$|(G(u_{k'}), v - u_{k'})| \leq const. ||G(u_{k'})|| \to 0 \text{ as } k' \to +\infty$$

and so we find from (1.25) that

$$J(v) \geq \liminf_{k' \to +\infty} J(u_{k'})$$

or what is the same as saying $J(v) \geq J(u)\ \forall v \epsilon V$. Thus u is a solution of the Problem P which proves (b).

(3) The strong convexity of J implies the convexity of J (i.e. H5 implies H3) and hence by (b) of Part 2 of the theorem the Problem P has a solution $u \epsilon V$. Moreover, by Proposition 1. 3.1 this solution is unique since J is strictly convex.

Again by (2)(a) of the theorem u_k is bounded sequence and has a weak cluster points u which is unique and hence $u_k \rightharpoonup u$ as $k \to +\infty$.

(4) Since coercivity of $H(u)$ implies that J is strictly convex (a) is just the same as (3)(a). To prove (b) we expand $J(u)$ by Taylor's formula: there is a θ in $0 < \theta < 1$ such that

$$\begin{aligned} J(u) &= J(u_k) + J'(u_k, u - u_k) + \frac{1}{2} J''(u_k + \theta(u - u_k); u - u_k, u - u_k) \\ &= J(u_k) + (G(u_k), u - u_k) + \frac{1}{2}(H(u_k + \theta(u - u_k))(u - u_k), u - u_k). \end{aligned}$$

73

Here

$$|(G(u_k), u - u_k)| \leq ||G(u_k)||||u - u_k||\ \forall k$$

and

$$(H(u_k + \theta(u - u_k))(u - u_k), u - u_k) \geq \alpha ||u - u_k||^2\ \forall k.$$

These two together with the fact that u is a solution of the Problem P imply that

$$J(u) \geq J(u) - ||G(u_k)||||u - u_k|| + \alpha/2||u - u_k||\ \forall k$$

which gives

$$||u - u_k|| \leq 2/\alpha ||G(u_k)||\ \forall k.$$

But, by Part 1 of the theorem the right hand side here $\to 0$ as $k \to 0$ and this proves that $u_k \to u$ as $k \to +\infty$.

□

2 Generalized Newton's Method

In this section we give another algorithm for the construction of approximating sequences for the minimisation problem for functionals J on a Banach space V using first and second order G-derivatives of J. This algorithm generalizes the method of Newton-Rophson which consists in giving approximations to determine points of V where a given operator vanishes. The method we describe is a refinement of a method by R. Fages [54].

We can describe our approach to the algorithm as follows: Suppose $J : V \to \mathbb{R}$ is a very regular functional on a Banach space V; for instance, J has a gradient $G(u)$ and a Hessian $H(u)$ everywhere in V. Let 74 $u \epsilon V$ be a point where J attains its minimum i.e. $J(u) \leq J(v)\ \forall v \epsilon V$. We have seen in Chapter 2. 1 (Theorem 2. 1.3) that $G(u) = 0$ is a necessary condition and we have also discussed the question of when this condition is also sufficient in Chapter 2, §2. Thus finding a minimising sequence for J at u is reduced to the equivalent problem of finding an algorithm to construct a sequence u_k approximating a solution of the equation:

$$(*) \qquad u \epsilon V, G(u) = 0.$$

In this sense this is an extension of the classical Newton method fot the determination of zeros of a real valued function on the real line.

As in the previous section we shall restrict ourselves to the case of a Hilbert space V.

Starting from an initial point $u_\circ \epsilon V$ suppose we have constructed u_k, If u_k is sufficiently near the solution u of the equation $G(u) = 0$ then by expanding $G(u)$ using Taylor's formula we find:

$$0 = (G(u), \varphi) = (G(u_k)) + H(u_k + \theta(u - u_k))(u - u_k), \varphi).$$

The Newton-Raphson method consists in taking u_{k+1} as a solution of the equation

$$G(u_k) + H(u_k)(u_{k+1} - u_k) = 0 \text{ for } k \geq 0.$$

Roughly speaking, if the operator $H(u_k)\epsilon\mathscr{L}(V, V') \equiv \mathscr{L}(V, V)$ is invertible and if $H(u_k)^{-1}\epsilon\mathscr{L}(V, V)$ then the equation is equivalent to

$$u_{k+1} = u_k - H(u_k)^{-1}G(u_k).$$

Then one can show that under suitable assumptions on G and H that this is a convergent algorithm provided that the initial points $u_\circ$ is sufficiently close to the required solution u of the problem $(*)$. However, in practice, u and then a good neighbourhood of u where $u_\circ$ is to be taken is not known a priori and difficult to find. 75

The algorithm we give in the following avoids such a difficulty for the choice of the initial point $u_\circ$ in the algorithm.

Let V be a Hilbert space and $J : V \to \mathbb{R}$ be a functional on V. Throughout this section we make the following hypothesis on J:

(H1) $J(v) \to +\infty$ as $||v|| \to +\infty$.

(H2) J is regular: J is twice G-differentiable and has a gradient $G(u)$ and a hessian $H(u)$ everywhere in V.

(H3) H is uniformly V-coercive on bounded sets of F: for every bounded set K of V there exists a constant $\alpha_K > 0$ such that

$$(H(v)\varphi, \varphi) \geq \alpha_k ||\varphi||^2, \forall v \epsilon K \text{ and } \forall \varphi \epsilon V.$$

(H4) H satisfies a uniform Lipschitz condition on bounded sets of V: for every bounded subset K of V there exists a constant $\beta_K > 0$ such that

$$||H(u) - H(v)|| \leq \beta_K ||u - v||, \forall u, v \epsilon K.$$

We are interested in finding an algorithm starting from a $u_\circ \epsilon V$ to find u_k iteratively. Suppose we have determined u_k for some $k \geq 0$. In order to determine u_{k+1} we introduce a bi-linear bicontinuous form $b_k : V \times V \ni (\varphi, \psi) \mapsto b_k(\varphi, \psi)\epsilon\mathbb{R}$ satisfying either one of the following two hypothesis:

(H5) There exist two constants $\lambda_\circ > 0, \mu_\circ > 0$ independent of $k, \lambda_\circ$ large enough (see (2.12)), such that

$$b_k(\varphi, \varphi) \geq \lambda_\circ (G(u_k), \varphi)^2, \ \forall \varphi \in V,$$

and

$$|b_k(\varphi, \psi)| \leq \mu_\circ ||G(u_k)|| ||\varphi|| ||\psi||, \ \forall \varphi, \psi \in V.$$

76

(H6) There exist two constant $\lambda_1 > 0, \mu_1 > 0$ independent of k, λ_1 large enough see (2.14), such that

$$b_k(\varphi, \varphi) \geq \lambda_1 ||G(u_k)||^{1+\in} ||\varphi||^2, \forall \varphi \in V$$

and

$$|b_k(\varphi, \psi)| \leq \mu_1 ||G(u_k)||^{1+\in} ||\varphi|| ||\psi||, \forall \varphi, \psi \in V,$$

where $\epsilon \geq 0$.

It is easy to see that there does always exist such a bilinear form as can be seen from the following example.

Example 2.1. $b_k(\varphi, \psi) = \lambda^k (G_k, \varphi)(G_k, \psi), 0 < \lambda_\circ \leq \lambda^k \leq \mu_0 < +\infty, \lambda_\circ$ large enough.

Example 2.2. $b_k(\varphi, \psi) = \lambda^k ||G_k||^2 (\varphi, \psi), 0 < \lambda_\circ \leq \lambda^k \leq \mu_\circ < +\infty$. Cauchy-Schwarz inequality shows that $(H5)$ is satisfied by this and $(H6)$ is satisfied with $\in = 1$.

Example 2.3. Let $\lambda_k > 0$ be a number in a fixed interval $0 < \lambda_1 \leq \lambda^k \leq \mu_1 < +\infty$ then the bi-linear form

$$b_k(\varphi, \psi) = \lambda^k ||G(u_k)||^{1+c} (\varphi, \psi)$$

satisfies $(H6)$.

We are now in a position to describe our algorithm.

Algorithm. Suppose we choose an initial point $u_\circ$ in the algorithm arbitrarily and that we have determined u_k for some $k \geq 0$. Consider the linear problem:

(2.1)
$$\begin{cases} \text{to find } \triangle_k \epsilon V \text{ satisfying the linear equation} \\ (H(u_k)\triangle_k, \varphi) + b_k(\triangle_k, \varphi) = -(G(u_k), \varphi) = -(G(u_k), \varphi), \forall \varphi \epsilon V \end{cases}$$

Here since $H(u_k)$ is V-coercive and b_k is positive semi-definite on V:

$$\text{i.e.} \qquad (H(u_k)\varphi, \varphi) \geq \alpha ||\varphi||^2, \forall \varphi \epsilon V \quad \text{(by } (H3))$$

(with $\alpha = \alpha(u_k) > 0$, a constant) and 77

$$b_k(\varphi, \varphi) \geq 0 \quad \text{(by } (H5) \text{ or } (H6))$$

the linear problem (2.1) has a unique solution $\triangle_k \epsilon V$.

Now we set

$$u_{k+1} = u_k + \triangle_k$$

where $\triangle_k$ is the unique solution of the problem (2.1). Clearly, our algorithm depends on the choice of the bilinear form $b_k(\varphi, \psi)$. We also see that if $b_k \equiv 0$ our algorithm is nothing but the classical Newton method as we have described in the introduction to this section.

We have now the main result of this section.

Theorem 2.1. *Suppose J satisfies the hypothesis $(H1)$ - $(H4)$ and b_k satisfy either the hypothesis $(H5)$ or $(H6)$ for each $k \geq 0$. Then we have:*

(1) The minimization problem:

to find $u \epsilon V$, $J(u) \leq J(v), \forall v \epsilon V$ has a unique solution.

(2) The sequence u_k is well defined by the algorithm.

(3) The sequence u_k converges to the solution u of the minimization problem: $||u_k - u|| \to 0$ as $k \to +\infty$.

(4) There exist constants $\gamma_1 > 0, \gamma_2 > 0$ such that

$$\gamma_1 ||u_{k+1} - u_k|| \leq ||u_k - u|| \leq \gamma_2 ||u_{k+1} - u_k||, \forall k.$$

(5) *The convergence of u_k to u is quadratic: there exists a constant $\gamma_3 > 0$ such that*

$$||u_{k+1} - u|| \leq \gamma_3 ||u_k - u||^2, \forall k.$$

78 In the course of the proof we shall use the notation introduced in the previous section: $J_k, G_k, H_k, \triangle J_k, \cdots$ respectively denote $J(u_k), G(u_k)$, $H(u_k)$, $J(u_k) - J(u_{k+1}), \cdots$

Proof. We shall carry out the proof in several steps.

Step 1. Let U be the subset of V:

$$U = \{v | v \epsilon V; J(v) \leq J(u_\circ)\}.$$

If there exists a solution u of the minimization problem then u necessarily belongs to this set U (irrespective of the choice of $u_\circ$). The set U is bounded in V. In fact, if it is not bounded then there exists a sequence u_j such that $u_j \epsilon U$, $||u_j|| \to +\infty$ and hence by $(H2)$ and $(H3)$ J has a Hessian which is positive definite everywhere. Hence J is strictly convex.

The set U is also weakly closed. In fact, if $v_j \epsilon U$ and $v_j \to v$ in V then (strict) convexity of J implies by Proposition (1.3.1) that we have

$$J(u_\circ) \geq J(v_j) \geq J(v) + (G(v), v_j - v)$$

and hence passing to the limit (since $G(v)$ is bounded for all j) it follows that $J(v) \leq J(u_\circ)$ proving that $v \epsilon U$, i.e. U is closed (and hence also weakly).

Now J and U satisfy all the hypothesis of Theorem 2. 2.1 with $\chi(t) = \alpha_U t$ and hence it follows that there exists a unique $u \epsilon U$ solution of the minimizing problem for J. We have already remarked that u is unique in V. This proves assertion (1) of the statement.

We have also remarked before the statement of the theorem that the linear problem (2.1) has a unique solution $\triangle_k$ which implies that u_{k+1} is well defined and hence we have the assertion (2) of the statement.

Step 2. $J(v), G(v)$ and $H(v)$ are bounded on any bounded subset K of V:
79 There exists a constant $\gamma_k > 0$ such that

$$|J(v)| + ||G(v)|| + ||H(v)|| \leq \gamma_K, \forall v \epsilon K.$$

In fact let $d_k = diam K$ and let $w \epsilon K$ be any fixed point. By $(H4)$ we have

$$H(v) \leq ||H(v) - H(u)|| + ||H(u)|| \leq \beta_K d_K + ||H(u)||$$

which proves that H is bounded on K. Then by Taylor's formula applies to G gives

$$||G(v) - G(u)|| \leq ||H(u + \theta(v - u))|| ||v - u||.$$

for some $0 < \theta < 1$. Now if $u, v \epsilon K$ then $u + \theta(v - u)$ is also in a bounded set $K_1 = \{w | w \epsilon V, d(w, K) \leq 2d_K\}$ (for, if $w = u + \theta(v - u)$ and $u \epsilon K$ then $||w - a|| = ||u - a + \theta(v - u)|| \leq ||u - a|| + ||v - u|| \leq 2d_K$). Since H is bounded on K_1 it follows that G is uniformly Lipschitz on K and as above G is also bounded on K. A similar argument proves J is also bounded on K.

For the sake of simplicity we shall write

$$\alpha = \alpha_U, \gamma = \gamma_U.$$

Step 3. Suppose $u_k \epsilon U$ for some $k \geq 0$. (This is trivial for $k = 0$ by the definition of the set U). Then u_{k+1} is also bounded.

For this, taking $\varphi = \triangle_k$ in (2.1) we get

$$(H_k \triangle_k, \triangle_k) + b_k(\triangle_k, \triangle_k) = -(G_k, \triangle_k). \tag{2.3}$$

By using the coercivity of $H_k = H(u_k)$ (hypothesis $(H3)$) and the fact that $b_k(\triangle_k, \triangle_k) \geq 0$ we get

$$\alpha ||\triangle_k||^2 \leq -(G_k, \triangle_k). \tag{2.4}$$

Then the Cauchy-Schwarz inequality applied to the right hand side of (2.4) gives

Suppose $0 < \ell < +\infty$ be such that $\sup_{u \epsilon U} ||G(u)||/\alpha \leq \ell$ (for example we can take $\ell = \gamma/\alpha$) and suppose U_1 is the set 80

$$U_1 = \{v | v \epsilon V; \exists w \epsilon U \text{ such that } ||v - w|| \leq \ell\}. \tag{2.5}$$

Then U_1 is bounded and $u_{k+1} = u_k + \triangle_k \epsilon U_1$.

$$u_{k+1} \epsilon U_1. \tag{2.6}$$

We shall in fact show later that $u_{k+1} \epsilon U$ itself.

Step 4. Estimate for $\triangle J_k$ from below. By Taylor's formula we have

$$\begin{cases} J_{k+1} = J_k + (G_k, \triangle_k) + \frac{1}{2}(\overline{H}\triangle_k, \triangle_k), \\ \text{where} \\ \overline{H}_k = H(u_k + \theta\triangle_k) \text{ for some } \theta \text{ in } 0 < \theta < 1. \end{cases} \tag{2.7}$$

Replacing $(G_k, \triangle_k)$ in (2.7) by (2.3) we have

$$\begin{aligned} J_{k+1} &= J_k - (H_k\triangle_k, \triangle_k) - b_k(\triangle_k, \triangle_k) + \frac{1}{2}(\overline{H}\triangle_k, \triangle_k) \\ &= J_k - \frac{1}{2}(H_k\triangle_k, \triangle_k) - b_k(\triangle_k, \triangle_k) + \frac{1}{2}((\overline{H}_k - H_k)\triangle_k, \triangle_k). \end{aligned}$$

Now using V-coercivity of H_k (hypothesis $(H3)$) and the Lipschitz continuity (hypothesis $(H4)$) of H on the bounded set U_1 we find (since $u_k + \theta\triangle_k \epsilon U_1$):

$$J_{k+1} \leq J_k - \alpha/2\|\triangle_k\|^2 - b_k(\triangle_k, \triangle_k) + \frac{1}{2}\beta_{U_1}\|\triangle_k\|^3.$$

Thus setting

$$\beta = \beta_{U_1} \tag{2.8}$$

we obtain

$$\alpha/2\|\triangle_k\|^2 + b_k(\triangle_k, \triangle_k) - \frac{1}{2}\beta\|\triangle_k\|^3 \leq \triangle J_k (= J_k - J_{k+1}). \tag{2.9}$$

In particular, since b_k is positive (semi -) definite,

$$\alpha/2\|\triangle_k\|^2(1 - \beta/\alpha\|\triangle_k\|) \leq \triangle J_k \tag{2.10}$$

81 In the methos of Newton-Rophson we have only (2.10).

Step 5. $\triangle J_k$ is bounded below by a positive number: if $0 < C < 1$ is any number then we have

$$\alpha C/2\|\triangle_k\|^2 \leq \triangle J_k. \tag{2.11}$$

To prove this we consider two cases:

(i) $\|\triangle_k\|$ is sufficiently small, i.e. $\|\triangle_k\| \leq (1-C)\alpha/\beta$, and

(ii) $\|\triangle_k\|$ large, i.e. $\|\triangle_k\| > (1-C)\alpha/\beta$.

If (i) holds then (2.11) is immediate from (2.10). Suppose that (ii) holds. By hypothesis $(H5)$ and by (2.5):

$$b_k(\triangle_k, \triangle_k) \geq \lambda_\circ(G_k, \triangle_k)^2 \geq \lambda_\circ \alpha^2\|\triangle_k\|^4$$

Then from (2.9) we can get

$$\begin{aligned} &\alpha/2\|\triangle_k\|^2 + \lambda_\circ\alpha^2\|\triangle_k\|^4 - \beta/2\|\triangle_k\|^3 \leq \triangle J_k \\ \text{i.e.}\quad &\alpha/2\|\triangle_k\|^2 + \lambda_\circ\alpha^2\|\triangle_k\|^3(\|\triangle_k\| - \beta/(2\lambda_\circ)\alpha^2) \leq \triangle J_k. \end{aligned}$$

If we take

$$\lambda_\circ \geq \beta^2/(2\alpha^3(1-C)) \tag{2.12}$$

then we find that $\|\triangle_k\| > (1-C)\alpha/\beta > \beta/(2\lambda_\circ\alpha^2)$ and hence

$$\alpha/2\|\triangle_k\|^2 \leq \triangle J_k. \tag{2.13}$$

Since $0 < C < 1$ we again get (2.11) from (2.13). Suppose on the other hand (ii) holds and b_k satisfies $(H6)$ with a λ_1 to be determined. Again from (2.9), (2.5) and hypothesis $(H6)$ we have

$$\begin{aligned} &\alpha/2\|\triangle_k\|^2 + \lambda_1\|G_k\|^{1+\epsilon}\|\triangle_k\|^2 - \beta/(2\alpha)\|\triangle_k\|^2\|G_k\| \leq \triangle J_k \\ \text{i.e.}\quad &\alpha/2\|\triangle_k\|^2 + \lambda_1\|G_k\|\|\triangle_k\|^2(\|G_k\|^\epsilon - \beta/(2\alpha\lambda)) \leq \triangle J_k \end{aligned}$$

Using (ii) together with (2.5) we get 82

$$\frac{\alpha^\in(1-C)^\epsilon}{\beta^\epsilon}\alpha^3 \leq \alpha^\epsilon\|\triangle_k\|^\epsilon \leq \|G_k\|^\epsilon$$

so that if $\alpha^{2\epsilon}(1-C)^{\epsilon}/\beta^{\epsilon} > \beta/2\alpha\lambda_1$ then we can conclude that

$$\alpha/2\|\triangle_k\|^2 \leq \triangle J_k.$$

This is possible if λ_1 is large enough: i.e. if

$$\lambda_1 = \beta^{1+\epsilon}/2\alpha^{1+2\epsilon}(1-C)^{\epsilon}. \tag{2.14}$$

As before since $0 < C < 1$ we find the estimate (2.11) also in this case.

Step 6. $J_k = J(u_k)$ is decreasing, $u_{k+1}\epsilon U$ and $\|\triangle_k\| \to 0$ as $k \to +\infty$. The estimate (2.11) shows that

$$J_k - J_{k+1} = \triangle J_k \geq 0,$$

which implies that J_k is decreasing. On the other hand, since u is the solution of the minimization problem we have

$$J(u) \leq J_{k+1} \leq J_k,$$

which shows that $u_{k+1}\epsilon U$ since $J(u_{k+1}) \leq J(u_k) \leq J(u_\circ)$ since $u_k\epsilon U$. Thus J_k is a decreasing sequence bounded below (by $J(u)$) and hence converges as $k \to +\infty$.

In particular

$$\triangle J_k = J_k - J_{k+1} \geq 0 \text{ and } \triangle J_k \to 0 \text{ as } k \to +\infty.$$

Then, by (2.11)

$$\|\triangle_k\| \to 0 \text{ as } k \to +\infty \tag{2.15}$$

83

Step 7. The sequence u_k converges (strongly) to u, the solution of the minimization problem. In fact, we can write by applying Taylor's formula to (G, φ), for $\varphi\epsilon V$,

$$(G_k, \varphi) = (G(u), \varphi) + (\hat{H}_k(u_k - u), \varphi)$$

where

$$H_k = H(u+\theta(u_k-u)) \text{ for some } \theta_\varphi \text{ in } 0<\theta<1.$$

But here $G(u)=0$. Now replacing (G_k,φ) by using (2.1) defining $\triangle_k$ we obtain

$$(2.16) \qquad (H_k\triangle_k,\varphi)+b_k(\triangle_k,\varphi) = -(\hat{H}_k(u_k-u),\varphi),\ \forall\varphi\epsilon V.$$

We take $\varphi = u_k - u$ in (2.16). Since U is convex and since $u, u_k\epsilon U$ it follows that $u+\theta(u_k-u)\epsilon U$. By the uniform V-coercivity of H we know that

$$(\hat{H}_k(u_k-u), u_k-u) \geq \alpha\|u_k-u\|^2, \alpha=\alpha_u.$$

Applying Cauchy-Schwarz inequality to the term $-(H_k\triangle_k, u_k-u)$ and using the fact that H_k is bounded we get

$$|(H_k\triangle_k, u_k-u)| \leq \gamma_u\|\triangle_k\|\|u_k-u\|.$$

Then (2.16) will give

$$\alpha\|u_k-u\|^2 \leq \gamma\|\triangle_k\|\|u_k-u\| + |b_k(\triangle_k, u_k-u)|.$$

On the other hand, $\|G(u_k)\|$ is bounded since $u_k\epsilon U$. Let $d = \max(\mu_\circ \|G(u_k)\|^2, \mu_1\|G(u_k)\|^{1+\epsilon}) < +\infty$. The hypothesis $(H5)$ or $(H6)$ together with the last inequality imply

$$\alpha\|u_k-u\|^2 \leq (\gamma+d)\|\triangle_k\|\|u_k-u\|,$$

i.e.

$$(2.17) \qquad \|u_k-u\| \leq (\gamma+d)/\alpha\|\triangle_k\|$$

Since $\|\triangle_k\| \to 0$ as $k\to+\infty$ by (2.15) we conclude from (2.17) that $u_k \to u$ as $k\to+\infty$. Next if we take $\varphi=\triangle_k$ in (2.16) we get 84

$$(H_k\triangle_k,\triangle_k)+b_k(\triangle_k,\triangle_k) = -(\hat{H}_k(u_k-u),\triangle_k).$$

Once again using the facts that b_k is positive semi-definite by $(H5)$ or $(H6)$ and that H_k is V-coercive by (H_3) we see that

$$\alpha\|\triangle_k\|^2 \leq \|u_k-u\|\|\triangle_k\|$$

since $\hat{H}_k$ is bounded because $u + \theta(u_k - u) \epsilon U$ for any θ in $0 < \theta < 1$ i.e. we have

$$\alpha/\gamma ||\triangle_k|| \leq ||u_k - u||. \tag{2.18}$$

(2.17) and (2.18) together give the inequalities in the assertion (4) of the statement with $\gamma_1 = \alpha/\gamma, \gamma_2 = (\gamma + d)/\alpha$.

Step 8. Finally we prove that the convergence $u_k \to u$ is quadratic. If we set $\delta_k = u_k - u$ then $\triangle_k = \delta_{k+1} - \delta_k$ and (2.16) can now be written as

$$\begin{aligned}(H_k\delta_{k+1}, \varphi) + b_k(\delta_{k+1}, \varphi) &= (H_k\delta_k, \varphi) + b_k(\delta_k, \varphi) - (\hat{H}_k\delta_k, \varphi)\\ &= ((H_k - \hat{H}_k)\delta_k, \varphi) + b_k(\delta_k, \varphi).\end{aligned}$$

Here we take $\varphi = \delta_{k+1}$. Applying V-coercivity of H_k (hypothesis H3), using positive semi-definiteness of b_k on the left side and applying Cauchy-Schwarz inequality to the two terms on the right side together with the hypothesis ($H4$) to estimate $||H_k - \hat{H}_k||$ we obtain

$$\begin{aligned}\alpha||\delta_{k+1}||^2 &\leq ||H_k - H_k|| ||\delta_{k+1}|| + |b_k(\delta_k, \delta_{k+1})|\\ &\leq \beta||\delta_k||^2 ||\delta_{k+1}|| + |b_k(\delta_k, \delta_{k+1})|.\end{aligned} \tag{2.19}$$

85 But, by ($H5$).

$$|b_k(\delta_k, \delta_{k+1})| \leq \mu_\circ ||G_k||^2 ||\delta_k|| ||\delta_{k+1}||. \tag{2.20}$$

On the other hand, by mean-value property applied G we have

$$||G_k - G(u)|| \leq \gamma ||u_k - u||$$

since for any $w \epsilon U, ||U(w)|| \leq \gamma$. As $G(u) = 0$ this implies that

$$||G_k|| \leq \gamma||u_k - u|| = \gamma||\delta_k||. \tag{2.21}$$

Substituting this in the above inequality (2.19)

$$\alpha||\delta_{k+1}||^2 \leq \beta||\delta_k|| ||\delta_{k+1}|| + \mu_\circ\gamma^2 ||\delta_k||^3 ||\delta_{k+1}||.$$

Now dividing by $||\delta_{k+1}||$ and using the fact that $||\delta_k|| = ||u_k - u|| \leq diamU$ we get

$$\begin{aligned}||\delta_{k+1}|| &\leq \alpha^{-1}(\beta + \mu_\circ\gamma^2||\delta_k||)||\delta_k||^2\\ &\leq \alpha^{-1}(\beta + \mu_\circ\gamma^2 diamU)||\delta_k||^2\end{aligned}$$

which is the required assertion (5) of the statement with $\gamma_3 = \alpha^{-1}(\beta + \mu_\circ\gamma^2 diamU)$.

If we had used hypothesis $(H6)$ instead of $(H5)$ to estimate $|b_k(\delta_k, \delta_{k+1})|$ we would get

$$(2.20)' \qquad |b_k(\delta_k, \delta_{k+1})| \leq \mu_1||G_k||^{1+\epsilon}||\delta_k||||\delta_{k+1}||$$

in place of (2.20). Now by (2.19) together with (2.21) gives (exactly by the same arguments as in the earlier case)

$$||\delta_{k+1}|| \leq \alpha^{-1}(\beta + \mu_1\gamma^{1+\epsilon}(diamU)^\epsilon)||\delta_k||^2.$$

In this case, we can take $\gamma_3 = \alpha^{-1}(\beta + \mu_1\gamma^{1+\epsilon}(diamU)^\epsilon)$.

This completely proves the theorem. □

We shall conclude this section with remarks. **86**

Remark 2.1. In the course of our proof all the hypothesis $(H1)$ - $(H5)$ or $(H6)$ except $(H4)$ have been used only for elements v in the bigger bounded set U while the hypothesis $(H4)$ has been used also for elements in the bigger bounded set U_1.

Remark 2.2. As we have mentioned earlier the proof of Theorem 2.1 given above includes the proof of the classical Newton-Rophson method if we make the additional hypothesis that $u_\circ$ is close enough to u such that $\forall v \epsilon U$ we have

$$\frac{1}{\alpha}||G(u)|| \leq \frac{\alpha}{\beta}d,$$

d given in $]0, 1[$. Then using (2.5), (2.10) becomes

$$(1 - d)\frac{\alpha}{3}||\triangle_k||^2 \leq \triangle J_k.$$

Remark 2.3.

Example 2.4. Let $V = \mathbb{R}^n$. Then $G_k \epsilon (\mathbb{R}^n)' = \mathbb{R}^n$. If we represent an element $\varphi \epsilon \mathbb{R}^n$ as a column matrix

$$\varphi = \begin{pmatrix} \varphi_1 \\ \vdots \\ \varphi_n \end{pmatrix} \epsilon \mathbb{R}^n$$

trhen $\varphi\varphi^t$ (with matrix multiplication) is a square matrix of order n. In particular $G_k G_k^t$ is an $(n \times n)$ square matrix. Moreover under the hypothesis we have made $H_k + \lambda G_k G_k^t$ is a positive definite matrix for $\lambda > 0$. This corresponds to $b_k(\varphi, \psi) = \lambda(G_k^t \varphi, G_k^t \psi)' = \lambda(G_k G_k^t \varphi, \psi)$ and our linear problem (2.1) is nothing but the system of n-linear equations

$$(H_k + \lambda G_k G_k^t) \triangle_k = -G_k$$

in n-unknowns $\triangle_k$.

Example 2.5. Simiarly we can take $b_k(\varphi, \psi) = \lambda ||G_k||^2 (\varphi, \psi)$, and we get

$$(H_k + \lambda ||G_k||^2 I) \triangle_k = -G_k.$$

Example 2.6. We can take $b_k(\varphi, \psi) = \lambda ||G_k||^{1+\epsilon} (\varphi, \psi)$ and we get

$$(H_k + \lambda ||G_k||^{1+\epsilon} I) \triangle_k = -G_k$$

as the corresponding system of linear equations.

87

Remark 2.4. The other algorithms given in this chapter do make use only of the calculation of the first G-derivative of J while the Newton method uses the calculation of the second order derivatives (Hessian) of J. Hence Newton's method is longer, more expensive economically than the methods based on algorithms given earlier.

3 Other Methods

The following are some of the other interesting methods known in the literature to construct algorithms to approximate solutions of the minimization problems. We shall only mention these.

(a) *Conjugate gradient method:* One of the algorithms in the class of these methods is known as Devidon-Fletcher-Powell method. Here we need to compute the G-derivatives of first order of the functional to be minimized. This is a very good and very much used method for any problems. (See [11] and [15]).

(b) *Relaxation method:* In this method it is not necessary to compute the derivatives of the functionals. Later on in the next chapter we shall give relaxation method also when there are constraints. (See Chapter 4. §4.5).

(c) Rosenbrock method. (See, for instantce, [30]).

(d) Hooke and Jeeves method. (See for instance [30])

Also for these two methods we need not compute the derivatives of functionals. They use suitable local variations.

Chapter 4

Minimization with Constraints - Algorithms

We have discussded the existence and uniqueness results for solutions 88 of the minimization problems for convex functionals on closed convex subsets of a Hilbert space. This chapter will be devoted to give algorithm for the construction of minimizing sequences for solutions of this problem. We shall describe only a few methods in this direction and we prove that such an algorithm is convergent.

1 Linearization Method

The problem of minimization of a functional on a convex set is also some-times referred as the problem of (non-linear) programming. If the functional is convex the programming problem is call convex programming.

The main idea of the method we shall describe in this section consists in reducing at each stage of iteration the problem of non-linear convex programming to one of linear programming in one more variable i.e. to a problem of minimizing a linear functional on a convex set defined by linear constraints. However, when we reduce to this case we may not have coercivity. However, if we know that the convex set defined this way by linear constraints is bounded then we have seen in

Chapter 2 that the linear programming problem has a solution (which is not necessarily unique).

Then the solution of such a linear programming problem is used to obtain convergent choices of w and ρ.

Let V be a Hilbert space and K a closed subset of V. We shall prescribe some of the constraints of the problem by giving a finite number of convex functionals

$$J_i : V \ni v \mapsto J_i(v) \epsilon \mathbb{R}, i = 1, \cdots, k,$$

89 and we define a subset U of K by

$$U = \{v | v \epsilon K, J_i(v) \leq 0, i = 1, \cdots, \}$$

Then U is again a convex set in V. If $v, v' \epsilon U$ then $v, v' \epsilon K$ and $(1 - \theta)v + \theta v' \epsilon K$ for any $0 \leq \theta \leq 1$ since K is convex. Now $J_i(i = 1, \cdots, k)$ being convex we have

$$J_i((1 - \theta)v + \theta v') \leq (1 - \theta)J_i(v) + \theta J_i(v') \leq 0, i = 1, \cdots, k.$$

We note that in practice, the convex set K contains (i.e. is defined by) all the constraints which need not be linearized and the constraints to be linearized asre the $J_i(i = 1, \cdots, k)$.

Suppose now

$$J_\circ : v \ni V \to J_\circ(v) \epsilon \mathbb{R}$$

is a convex functional on V. We consider the minimization problem:

Problem 1.1. To find $u \epsilon U, J_\circ(u) \leq J_\circ(v), \forall v \epsilon U$. We assume that $J_\circ, J_1, \ldots, J_k$ satisfy the following hypothesis:

Hypothesis on $J_\circ : (HJ)_\circ$.

(1) $J_\circ(v) \to +\infty$ as $||v|| \to +\infty$

(2) $J_\circ$ is regular: $J_\circ$ is twice differentiable everywhere

in V and has a gradient $G_\circ$ and a hessian $H_\circ$ everywhere in V which are bounded on bounded subsets: for every bounded set U_1 of V there exists a constant $M_{U_1} > 0$ such that

$$||G_\circ(v)|| + ||H_\circ(v)|| \leq M_{U_1} \forall v \epsilon U_1.$$

(3)$_\circ$ $H_\circ$ is uniformly V-coercive on bounded subsets of V: for every bounded subset U_1 of V there exists a constant $\alpha_{U_1} > 0$ such that

$$(H_\circ(v)\varphi, \varphi) \geq \alpha_{U_1} ||\varphi||^2 \ \forall \varphi \epsilon V \text{ and } \forall v \epsilon U_1.$$

90

Hypothesis on $J_i.(HJ)_i$**:**

(1)$_i$ J_i is regular : J_i is twice G-differentiable everywhere in V and has a gradient G_i and a hessian H_i bounded on bounded sets of V: for every bounded set U_1 of V there exists a constant $M_{U_1} > 0$ such that

$$||G_i(v)|| + ||H_i(v)|| \leq M_{U_1} \quad \forall v \epsilon U_1, i = 1, \cdots, k.$$

(2)$_i$ $H_i(v)$ is positive semi-definite:

$$(H_i(v)\varphi, \varphi) \geq 0 \quad \forall \varphi \epsilon V (\forall v \epsilon U_1).$$

Hypothesis on $K.(HK)$**:** There exists and element $Z \epsilon K$ such that $J_i(Z) < 0$ for all $i = 1, \cdots, k$.

The hypothesis (HK) in particular implies that $U \neq \phi$.

In order to describe the algorithm let $u_\circ \epsilon U$ be the initial point (arbitrarily fixed) of the algorithm. In view of the hypothesis $(HJ)_\circ(1)$ we may, without loss of generality, assume that U is bounded since otherwise we can restrict ourselves to the set

$$\{v \epsilon U; J_\circ(v) \leq J_\circ(u)\}$$

which is bounded by $(HJ)_i(1)$. So in the rest of our discussion we assume U to be bounded.

Next, by hypothesis $(HJ)_i(1)$, the bounded convex set U is also closed. In fact, if $v_n \epsilon U$ and $v_n \to v$ then since K is closed, $v \epsilon K$. Moreover, by the mean value properly applied to $J_i (i = 1, \cdots, k)$ we have

$$|J_i(v_n) - J_i(v)| \leq ||G_i|| ||v_n - v||$$

so that $J_i(v_n) \to J_i(v)$ and hence $J_i(v) \leq 0$ for $i = 1, \cdots, k$ i.e. $v \epsilon U$.

Let $\mathscr{V}$ be a bounded closed convex subset of V which satisfies the 91 condition: there exist two numbers $r > 0$ and $d > 0$, ehich will be chosen suitably later on, such that

$$B(0,r) \subset \mathscr{V} \subset B(0,d)$$

where $B(0,t)$ denotes the ball $\{v\epsilon V|||v|| < t\}$ in $V(t = r,d)$. Consider the set

$$U_1 = \{v|v\epsilon V; \exists w\epsilon U \text{ such that } ||v-w|| \leq d\}.$$

Since U is bounded the set U_1 is also bounded and $U_1 \supset U$. In the hypothesis $(HJ)_\circ$ and $(HJ)_i$ we shall use only the bounded set U_1.

We shall use the following notation : $J_i(u_m), H_i(u_m)$ will be respectively denoted by J_i^m, G_i^m, H_i^m for $i = 0, 1, \cdots, k$ and all $m \geq 0$.

Now suppose that starting from $u_\circ\epsilon U$ we have constructed u_m. We wish to give an algorithm to obtain u_{m+1}. For this purpose we consider a linear programming problem.

A linear programming problem : Let U_m denote subset of $U\times\mathbb{R}$ defined as the set of all $(z,\sigma)\epsilon U\times\mathbb{R}$ satisfying

$$\begin{cases} z-u_m\epsilon\mathscr{V}, \\ (G_\circ^m, z-u_m)+\sigma \leq 0, \text{ and} \\ J_i^m + (G_i^m, z-u_m)+\sigma \leq 0 \quad \text{for } i = 1,\cdots,k. \end{cases}$$

It is easy to see that U_m is a nonempty closed convex bounded set: In fact, $(z,\sigma) = (u_m, 0)\epsilon U_m$ so that $U_m \neq \phi$. If $(z,\sigma)\epsilon U_m$ then since $z-u_m\epsilon V$, which is a bounded set it follows that z is bounded. Then using the other two inequalities in (1.1) it follows that σ is also bounded. If $(z_j,\sigma_j)\epsilon U_m$ and $(z_j,\sigma_j) \to (z,\sigma)$ in $U\times\mathbb{R}$ then since U is closed $z\epsilon U$ and hence $(z,\sigma)\epsilon U\times\mathbb{R}$. Again since $\mathscr{V}$ is closed $(z-u_m)\epsilon\mathscr{V}$. By the continuity of the (affine) functions

$$(z,\sigma) \mapsto J_i^m + (G_i^m, z-u_m)+\sigma$$
$$(z,\sigma) \mapsto (G_\circ^m, z-u_m)+\sigma$$

92 we find that

$$J_i^m + (G_i^m, z-u_m)+\sigma \leq 0, (G_\circ^m, z-u_m)+\sigma \leq 0.$$

Finally to prove the convexity, let $(z, \sigma), (z', \sigma')\epsilon U_m$. Then, for any, $0 \leq \theta \leq 1$,

$$(1-\theta)z + \theta z' - u_m = (1-\theta)(z-u_m) + \theta(z'-u_m)\epsilon\mathscr{V}$$

since $\mathscr{V}$ is convex. Moreover, we also have

$$\begin{aligned}&(G^m_\circ, (1-\theta)z + \theta z' - u_m) + (1-\theta)\sigma + \theta\sigma'\\ &= (1-\theta)[(G^m_\circ, z-u_m) + \sigma] + \theta[(G^m_\circ, z'-u_m) + \sigma'] \leq 0\end{aligned}$$

and similarly

$$J^m_i + (G^m_i, (-\theta)z + \theta z' - u_m) + (1-\theta)\sigma + \theta\sigma' \leq 0.$$

Next we consider the functional $g : V\times\mathbb{R} \to \mathbb{R}$ given by $g(z, \sigma) = \sigma$ and the linear programming problem :

(P_m) : to find $(z_m, \sigma_m)\epsilon U_m$ such that $g(z_m, \sigma_m) \geq g(z, \sigma), \forall (z, \sigma)\epsilon U_m$. i.e.

Problem P_m: To find $(z_m, \sigma_m)\epsilon U_m$ such that

(1.2) $$\sigma \leq \sigma_m \text{ for all } (z, \sigma)\epsilon U_m.$$

By the results of Chapter 2 we know that the Problem P_m has a solution (not necessarily unique).

We are now in a position to formulate our algorithm for the construction of u_{m+1}.

Algorithm. Suppose we have determined u_m starting from $u_\circ$. Then we take a solution (z_m, σ_m) of the linear programming Problem (P_m). We set **93**

(1.3) $$w_m = (z_m - u_m)/||z_m - u_m||$$

and

(1.4) $$\rho^\ell_m = \max\{\rho\epsilon\mathbb{R}, u_m + \rho w_m \epsilon U\}.$$

We shall prove later on that w_m is a direction of descent. We can define the notions of convergent choices of w_m and ρ in the same way as in Chapter 3, Section 1 for the functional $J_\circ$. We shall therefore not repeat these definitions here.

Let ρ^c_m be a convergent choice of ρ for the construction of the minimizing sequence for $J_\circ$ without constraints. We define

$$\rho_m = \min(\rho^c_m, \rho^\ell_m) \tag{1.5}$$

and we set

$$u_{m+1} = u_m + \rho_m w_m. \tag{1.6}$$

The following is the main result of this section.

Theorem 1.1. *Suppose that convex set K and the functionals $J_\circ, J_1, \ldots, J_k$ satisfy the hypothesis (HK) and $(HJ)_i, i = 0, 1, \cdots, k$. Suppose (1) the Problem (1.1) has a unique solution and (2) $u_m \to u$ as $m \to +\infty$.*

Then the algorithm described above to determine u_{m+1} from u_m is convergent.

i.e. If $u\epsilon U$ is the unique solution of the Problem (1.1) and if u_m is a sequence given by the above algorithm then $J(u_m) \to J(u)$ as $m \to +\infty$.

For this it will be necessary to prove that w_m is a direction of descent and w_m, ρ_m are convergent choices.

94 The following two lemmas are crucial for our proof of the Theorem 1.1.

Let $u\epsilon U$ be the unique solution of the Problem 1.1

Lemma 1.1. *Let the hypothesis of Theorem 1.1 be satisfied. If, for some $m \geq 0$ we have $J_\circ(u) < J_\circ(u_m)$ then there exists an element $(y_m, \epsilon_m) \in U_m$ such that $\epsilon_m > 0$.*

Proof. Let $u_m \in U$ be such that $J_\circ(u) < J_\circ(u_m)$. We first consider the case where $Z \neq u, Z$ being the point of K given in hypothesis (HK). We introduce two real numbers ℓ_m, ℓ'_m such that

$$J_\circ(u) < \ell'_m < \ell_m \leq J_\circ(u_m) \text{ and } \ell'_m < J_\circ(Z).$$

□

Let $I \equiv I(u, Z)$ denote the segment in V joining u and Z, i.e.

$$I = \{w | w \epsilon V; w = (1 - \theta)u + \theta Z, 0 \leq \theta \leq 1\}$$

Since u, Z belong to the convex set U we have $I \subset U$.

On the other hand, if $c \epsilon \mathbb{R}$ is any constant then the set

$$J_{\circ c} = \{v \epsilon U; J_\circ(v) \leq c\}$$

is convex and closed. For, if $v, v' \epsilon J_{\circ c}$ then for any, $0 \leq \lambda \leq 1$,

$$J_\circ((1 - \lambda)v + \lambda v') \leq (1 - \lambda)J_\circ(v) + \lambda J_\circ(v') \leq c$$

by the convexity of $J_\circ$ and $(1 - \lambda)v + \lambda v' \epsilon U$ since U is convex. To see that it is closed, let $v_j \in J_{\circ c}$ be a sequence such that $v_j \to v$ in V. Since U is closed $v \in U$. Moreover, by mean value property for $J_\circ$

$$|J_\circ(v_j) - J_\circ(v)| \leq M_U \|v_j - v\| \leq M_{U_1} \|v_j - v\|$$

by Hypothesis $(HJ)_\circ(2)$ so that $J_\circ(v_j) \to J_\circ(v)$ as $j \to +\infty$. Hence $J_\circ(v) \leq c$ i.e. $v \in J_{\circ c}$.

Now by the choice of $\ell]_m$, $u \epsilon I \cap J_{\circ \ell'_m}$ and hence $I_\circ \equiv I \cap J_{\circ \ell'_m} \neq \phi$. 95 It is clearly closed and bounded. $I_\circ$ being a closed bounded subset of a compact set I is itself compact.

Now the function $g : I_\circ \to \mathbb{R}$ defined by $g = J_\circ / I_\circ$ is continuous: In fact, if $w, w' \epsilon I_\circ$ then by the mean value property applies to $J_\circ$ gives

$$|g(w) - g(w')| = |J_\circ(w) - J_\circ(w')| \leq M_{U_1} \|w - w'\|$$

by hypothesis $(HJ)_\circ(2)$. Moreover, by the very definition of the set $I_\circ \subset J_{\circ, \ell'_m}$ we have

$$|g(w)| \leq \ell'_m.$$

Hence g attains its maximum in $I_\circ$ i.e. There exists a point $y_m \epsilon I_\circ$ such that $g(y_m) = J_\circ(y_m) = \ell'_m$. i.e. there exists a $\theta_m, 0 \leq \theta < 1$ such that

$$y_m = (1 - \theta_m)u + \theta_m Z, J_\circ(y_m) = \ell'_m.$$

Since $J_\circ(u) < \ell'_m$ we see that $y_m \neq u$ and therefore $\theta_m \neq 0$. i.e. $0 < \theta_m < 1$.

Next we show that $J_i(y_m) < 0$ for all $i = 1, \cdots, k$. In fact, since J_i is convex and has a gradient G_i we know from Proposition 3.1 of Chapter 1 that

$$J_i(y_m) \geq J_i^m + (G_i^m, y_m - u_m)$$

and we also have

$$J_i(y_m) \leq (1 - \theta_m) J_i(u) + \theta_m J_i(Z) < 0$$

since $0 < \theta_m < 1$ and $J_i(Z) < 0$.

Similarly, by convexity of $J_\circ$ we get

$$\ell'_m = J_\circ(y_m) \geq J_\circ^m + (G_\circ^m, y_m - u_m) \geq \ell_m + (G_\circ^m, y_m - u_m)$$
$$\text{i.e. } (G_\circ^m, y_m - u_m) \leq \ell'_m - \ell_m < 0 \text{ by the choice of } \ell_m, \ell'_m$$

96 We can now take

$$\in_m = \min\{\ell_m - \ell'_m, -J_1(y_m), \cdots, -J_k(y_m)\} > 0.$$

Then it follows immediately that $(y_m, \epsilon_m) \in U_m$ and $\epsilon_m > 0$.

We now consider the case $u = Z$. Then we can take $y_m = Z = u$ and hence $J_i(y_m) = J_i(u) = J_i(Z) < 0$. It is enough to take

$$\in_m = \min\{J_\circ(u_m) - J_\circ(u), -J_t(Z), \cdots, -J_k(Z)\} > 0.$$

If we now take $r > 0$ sufficiently large then $y_m - u_m \in \mathscr{V}$. This is possible since both y_m and u_m are in bounded sets:

$$||y_m|| \leq (1 - \theta_m)||u|| + \theta_m ||Z|| \leq ||u|| + ||Z||$$

so that

$$||y_m - u_m|| \leq ||y_m|| + ||u_m|| \leq ||u|| + ||Z|| + ||u_m||.$$

It is enough to take $r > ||u|| + ||Z|| + ||u_m|| > 0$. Thus $(y_m, c_m) \in \mathscr{V}$.

Corollary 1.1. *Under the assumptions of Lemma 1.1 there exists a strongly admissible direction of descent at U_m for the domain U.*

Proof. By Lemma 1.1 there exists an element $(y_m, \epsilon_m) \in U_m$ such that $\epsilon_m > 0$. On the other hand, let (z_m, σ_m) be a solution in U_m of the Linear programming problem (P_m). Then necessarily $\sigma_m \geq \epsilon_m > 0$ and we can write

$$(1.7) \quad \begin{cases} z_m - u_m \epsilon \mathscr{V}, z_m \epsilon U \\ J_i^m + (G_i^m, z_m - u_m) + \epsilon_m \leq J_i^m + (G_i^m, z_m - u_m) + \sigma_m \leq 0 \\ (G_\circ^m, z_m - u_m) + \epsilon_m \leq (G_\circ^m, z_m - u_m) + \sigma_m \leq 0 \end{cases}$$

Thus we have

$$(1.8) \qquad (G_\circ^m, z_m - u_m) \leq -\epsilon_m < 0,$$

and hence 97

$$(1.9) \qquad w_m = (z_m - u_m)/\|z_m - u_m\|$$

is a direction of descent. It is strongly admissible since U is convex and we can take any sequence of numbers $\epsilon_j > 0$, $\epsilon_j \to 0$. □

Lemma 1.2. *Let the hypothesis of Theorem 1.1 hold and, for some* $m \geq 0$, $J_\circ(u) < J_\circ(u_m)$. *If* $(z_m, \sigma_m)\epsilon U_m$ *is a solution of the linear programming problem* (P_m) *then there exists a number* $\mu_m > 0$ *depending only on* ϵ_m *of Lemma 1.1 such that*

$$(1.10) \qquad u_m + \rho(z_m - u_m)\epsilon U \textit{ for all } 0 \leq \rho \leq \mu_m.$$

Furthermore,

$$(G_\circ^m, z_m - u_m) < 0.$$

Proof. We have alredy shown the last assertion in the Corollary 1.1 and therefore we have to prove the existence of μ_m such that (1.10) holds. For this purpose, if $\rho > 0$, we get on applying Taylor's formula to each $J_i (i = 1, \cdots, k)$:

(1.11)

$$J_i(u_m + \rho(z_m - u_m)) = J_i^m + \rho(G_i^m, z_m - u_m) + \frac{1}{2}\rho^2(\overline{H}_i^m(z_m - u_m), z_m - u_m)$$

where

$$\overline{H}_i^m = H_i^m(u_m + \rho'(z_m - u_m)) \text{ for some } 0 < \rho' < \rho.$$

□

Here, since $z_m - u_m \epsilon \mathscr{V}$, $\|z_m - u_m\| < d$ and hence $u_m + \rho'(z_m - u_m)$, $(0 < \rho' < \rho)$ belongs to U_1 if we assume $\rho \leq 1$. $\|\overline{H}_i^m\|$ is bounded by M_{U_1} and so we get

$$J_i(u_m + \rho(z_m - u_m)) \leq J_i^m + \rho(G_i^m, z_m - u_m) + \frac{1}{2} M\rho^2 d^2. \tag{1.12}$$

Thus if we find a $\mu_m > 0$ such that $0 < \rho < \mu_m$ implies the right hand side of this last inequality is ≤ 0 forall $i = 1, \cdots, k$ then $u_m + \rho(z_m - u_m) \epsilon U$. 98

Using the first inequality (1.7) to replace the term $(G_i^m, z_m - u_m)$ in (1.12) we get

$$J_i(u_m + \rho(z_m - u_m)) \leq J_i^m + \rho(-J_i^m - \epsilon_m) + \frac{1}{2}\rho^2 M d^2. \tag{1.13}$$

The second degree polynomial on the right side of (1.13) vanishes for

$$\rho = \rho_i^m = [(J_i^m + \epsilon_m) + \{(J_i^m + \epsilon_m)^2 - 2Md^2 J_i^m\}^{\frac{1}{2}}]/Md^2. \tag{1.14}$$

Moreover the right side of (1.13) is smaller than

$$J_i^m + \rho(-J_i^m) + \frac{1}{2}\rho^2 M d^2$$

since $\epsilon_m > 0, \rho > 0$ and this last expression decreases as $\rho > 0$ decreases as $-J_i^m = -J_i(u_m) \leq 0$. Then it follows that, if $0 < \rho \leq \rho_i^m$, we have

$$J_i(u_m + \rho(z_m - u_m)) \leq 0.$$

We can now take $\mu_m = \min(\rho_1^m, \cdots, \rho_k^m)$ also that we will have

$$J_i(u_m + \rho(z_m - u_m)) \leq 0 \text{ for all } 0 < \rho \leq \mu_m \text{ and } i = 1, \cdots, k$$

But each of the ρ_i^m gives by (1.14) depend on J_i^m and hence on u_m. In order to get a $\mu > 0$ independent of u_m and dependent only on ϵ_m we can proceed as follows. If we set

$$\varphi(y) = [(y + \epsilon_m) + \{(y + \epsilon_m)^2 - 2Md^2 y\}^{\frac{1}{2}}]/Md^2 \tag{1.15}$$

for $y \leq 0$ then, since $y = J_i(u_m) = J_i^m \leq 0$, we can write

$$\rho_i^m = \varphi(J_i^m).$$

It is easily checked that the function $\varphi :]-\infty, 0] \to \mathbb{R}$ is continuous, $\varphi(y) > 0$ for all $y \leq 0$ and $\lim\limits_{y \to -\infty} \varphi(y) = 1$. Hence $\inf\limits_{y \leq 0} \varphi(y) = \eta(\epsilon_m)$ exists and $\eta(\epsilon_m) > 0$. 99

We choose $\mu_m = \eta(\epsilon_m)$. Then, if $0 < \rho \leq \mu_m \leq \rho_i^m$ for each $i = 1, \cdots, k$ given by (1.14) and consequently, for any such $\rho > 0$, $u_m + \rho(x_m - u_m) \epsilon U$.

We are niw in a position to prove Theorem 1.1

Proof of Theorem 1.1. We recall that $(z_m, \sigma_m) \epsilon U_m$ is a solution of the linear programming problem (P_m) and

$$\begin{aligned} w_m &= (z_m - u_m)/||z_m - u_m||, \\ \rho_m &= \min(\rho_m^\ell, \rho_m^c), \\ u_{m+1} &= u_m + \rho_m w_m. \end{aligned}$$

Then $J_\circ(u_m)$ is a decreasing sequence. In fact, if $\rho_m = \rho_m^c$ then by definition of ρ_m^c we have $J_\circ(u_{m+1}) \leq J_\circ(u_m)$. Suppose $\rho_m = \rho_m^\ell < \rho_m^c$. If $J_\circ(u_m + \rho_m^c w_m) \leq J_\circ(u_m + \rho_m^c w_m)$ there is nothing to prove. So we assume $J_\circ^{\rho_m} > J_\circ^{\rho_m}$. Consider the convex function $\rho \mapsto J(u_m + \rho w_m)$ in $[0, \rho_m^c]$. It attains its minimum at $\rho = \rho_{\min} \epsilon]0, \rho_m^c[$. Then $0 \leq \rho_m \leq \rho_{\min}$. In fact, if $\rho_{\min} < \rho_m < \rho_m^c$ then since $J_\circ$, being convex, is increasing in $[\rho_{\min}, \rho_m^c]$ we have $J_\circ^{\rho_m^c} \leq J_m^{\rho^c}$ contradicting our assumption. Once again since $J_\circ$ is convex $J_\circ$ is decreasing in $[0, \rho_{\min}]$. Hence $J_\circ^m = J_\circ(u_m) \geq J_\circ^{\rho_m} = J_\circ(u_{m+1})$. Since we know that there exists a (unique) solution u of the minimizing problem 1.1 we have $J_\circ(u_m) \geq J_\circ(u), \forall m \geq 0$. Thus $J_\circ(u_m)$, being a decreasing sequence bounded below, is convergent. Let $\ell = \lim\limits_{m \to +\infty} J_\circ(u_m)$. Clearly $\ell \geq J_\circ(u)$. Then there are two possible cases:

(1) $\ell = J_\circ(u)$ and

(2) $\ell > J_\circ(u)$.

Case (1). Suppose $J_\circ(u_m) \to \ell = J_\circ(u)$. Then, for any $m \geq 0$, we have by Taylor's formula :

$$J_\circ(u_m) = J_\circ(u) + (G_\circ(u), u_m - u) + \frac{1}{2}(\overline{H}_m(u_m - u), u_m - u).$$

100 where

$$\overline{H}_m = H_\circ(u + \theta(u_m - u)) \text{ for some } 0 < \theta < 1$$

Since $u, u_m \in U$ (which is convex), $u + \theta(u_m - u) \in U$ ofr any $0 < \theta < 1$ and hence by hypothesis $(HJ)_\circ(3)$

$$(\overline{H}_m(u_m - u), u_m - u) \geq \alpha||u_m - u||^2, \alpha = \alpha_{U_1} > 0.$$

Moreover, since $J_\circ$ is convex, we have by Theorem 2.2 of Chapter 2

$$(G_\circ(u), u_m - u) \geq 0.$$

Thus we find that

$$J_\circ(u_m) \geq J_\circ + \frac{1}{2}\alpha||u_m - u||^2$$

$$\text{i.e.} \qquad ||u_m - u||^2 \leq 2/\alpha(J_\circ(u_m) - J_\circ(u)).$$

Since $J_\circ(u_m) \to J_\circ(u)$ as $m \to +\infty$ it then follows that $u_m \to u$ as $m \to +\infty$.

Case(2). We shall prove that this case cannot occur. Suppose, if possible, let $J_\circ(u) < \ell \leq J_\circ(u_m), \forall m \geq 0$. We shall show that the choices of w_m and ρ_m are convergent for the problem of minimization of $J_\circ$ without constraints. i.e. the sequence u_m constructed using our algorithm tends to an absolute minimum of $J_\circ$ in V which will be a contradiction to our assumption.

w_m is a convergent choice. For this we introduce, as in the proof of Lemma 1.1 another real number ℓ' such that

$$J_\circ(u) < \ell' < \ell \leq J_\circ(u_m), \forall m \geq 0.$$

Then the proof of Lemma 1.1 gives the existence of $(y, \epsilon) \in U_m$ with
101 $\epsilon_m =\in> 0 \ \forall m \geq 0$. On the other hand, $(z_m, \sigma_m) \in U_m$ being a solution of

the linear programming problem (P_m) we have $\sigma_m \geq \in > 0$. Hence from (1.7) we get

$$(1.16) \qquad \begin{cases} (G_\circ^m, z_m - u_m) + \epsilon \leq 0, \\ J_i^m + (G_i^m, z_m - u_m) + \epsilon \leq 0. \end{cases}$$

From the first inequality here together with the Cauchy-Schwarz inequality gives

$$-\|G_\circ^m\| \|z_m - u_m\| \leq (G_\circ^m, z_m - u_m) \leq -\epsilon$$

$$\text{i.e.} \qquad \epsilon \leq \|G_\circ^m\| \|z_m - u_m\| \leq M\|z_m - u_m\|, M = M_{U_1},$$

using hypothesis $(HJ)_\circ(2)$. So we have

$$(1.17) \qquad \|z_m - u_m\| \geq \epsilon/M > 0.$$

By Lemma 1.2 there exists a $\mu = \eta(\epsilon) > 0$ such that

$$(1.10) \qquad u_m + \rho(z_m - u_m) \in U \text{ if } 0 \leq \rho < \eta(\epsilon).$$

If we denote by $\overline{\rho}, \overline{\rho} = \rho\|(z_m - u_m)\|$ then this is equivalent to saying that

$$u_m + \overline{\rho} w_m \in U \text{ if } 0 \leq \overline{\rho} < \eta(\epsilon)\|z_m - u_m\|.$$

Then, in view of (1.17), $0 \leq \overline{\rho} < \epsilon\eta(c)/M$ implies $0 \leq \overline{\rho} < \eta(c)\|z_m - u_m\|$ and hence

$$u_m + \overline{\rho} w_m \in U \text{ for all } 0 \leq \overline{\rho}\epsilon\eta(\epsilon)/M,$$

which means that

$$\rho_m^\ell \geq \epsilon\eta(c)/M.$$

Once again from (1.16) we have

$$(G_\circ^m, w_m) \leq -\epsilon/\|z_m - u_m\| \leq -\epsilon/d$$

because $z_m - u_m \in \mathscr{V}$ by (1.1) meancs that $\|z_m - u_m\| \leq d$. Since $\|G_\circ^m\| \leq M$ we obtain

$$(G_\circ^m/\|G_\circ^m\|, w_m) \leq -\epsilon/d\|G_\circ^m\|(\leq -\epsilon/Md).$$

Taking $\epsilon > 0$ small enough we conclude that 102

$(G^m_\circ/||G^m_\circ||, w_m) \leq -C_1 < 0, 1 \geq C_1 > 0$ being a constant. This is nothig but saying that the choice of w_m is convergent for the minimization problem without constraints by w-Algorithm 1 of Section 1.2 of Chapter 3.

ρ_m is a convergent choice. Since $\rho_m = \min(\rho^\ell_m, \rho^c_m)$ we consider two possible cases

(a) If $\rho_m = \rho^c_m$ then there is nothing to prove.

(b) Suppose $\rho_m = \rho^\ell_m$. We shall that this choice of ρ_m is also a convergent choice. For this let c_2 be a constant such that $0 < c_2 \leq \rho_m = \rho^\ell_m \leq \rho^c_m$.

Then $0 < \rho_m/\rho^c_m \leq 1$ and we can write

$$u_{m+1} = u_m + \rho_m w_m = (1 - \rho_m/\rho^c_m)u_m + \rho_m/\rho^c_m(u_m + \rho^c_m w_m).$$

The convexity of $J_\circ$ then implies that

$$J_\circ(u_{m+1}) \leq (1 - \rho_m/\rho^c_m)J(u_m) + \rho_m/\rho^c_m J_\circ(u_m + \rho^c_m w_m).$$

Hence we obtain

$$\begin{aligned}\triangle J^{\rho_m}_\circ &= J_\circ(u_m) - J_\circ(u_m + \rho_m w_m) = J_\circ(u_m) - J_\circ(u_{m+1})\\ &\geq \rho_m/\rho^c_m(J_\circ(u_m) - J_\circ(u_m + \rho^c_m w_m))\end{aligned}$$

i.e.

$$\triangle J^{\rho_m}_\circ \geq \rho_m/\rho^c_m \triangle J^{\rho^c_m}_\circ \tag{1.18}$$

We note that ρ^c_m is necessarily bounded above for any $m \geq 0$. For otherwise since, we find from triangle ineuality that

$$||u_m + \rho^c_m w_m|| \geq \rho^c_m||w_m|| - ||u_m|| = \rho^c_m - ||u_m||.$$

103 $u_m + \rho^c_m w_m$ would be unbounded. Then by Hypothesis $(HJ_\circ)(1) J_\circ(u_m +$

$\rho^c_m w_m)$ would also be unbounded. This is not possible by the definition of convergent choice of ρ^c_m.

Let C_3 be a constant such that $0 < \rho^c_m \leq C_3$ for all $m \geq 0$. Then (1.18) will give

$$\triangle J_\circ^{\rho_m} \geq C_2/C_3 \triangle J_\circ^{\rho^c_m} \tag{1.19}$$

Hence if $\triangle J_\circ^{\rho_m} \to 0$ then $\triangle J_\circ^{\rho^c_m} \to 0$ by (1.19). By the definition of ρ^c_m (as a convergent choice of ρ) we have

$$(G_m, w_m) \to 0 \text{ as } m \to +\infty$$

which means that ρ_m is also a convergent choice of ρ.

Finally, since the choices of ρ_m, w_m are both convergent for the minimization problem without constraints for $J_\circ$ we conclude using the results of Chapter 3 that $u_m \to \widetilde{u}$ where $\widetilde{u}$ is the global minimum for $J_\circ$ (which exists and is unique by results of Chapter 2, Theorem 2.1 of Section 2). Thus we have

$$J_\circ(\widetilde{u}) \leq J_\circ(u) < \ell \leq J_\circ(u_m)$$

$$\text{and } J_\circ(u_m) \to J_\circ(\widetilde{u})$$

which is impossible and hence the case (2) cannot therefore occur.

This proves the theorem completely.

We shall conclude this section with some remarks.

Remark 1.1. A special case of our algorithm was given a long time ago by Franck and Wolfe [17] in the absence of the constraints J_i which we have linearized. More precisely they considered the following problem:

Let $J_\circ$ be a convex quadratic functional on a Hilbert space V and K be a closed convex subset with non-empty interior. Then the problem is 104 to give an algorithm for finding a minimizing sequence u_m for

$$u \epsilon K, J_\circ(u) = \inf_{v \epsilon K} J_\circ(v).$$

The corresponding linear programming problem in this case will be the following:

$$\begin{cases} U_m = K_m = \{(z,\sigma)\epsilon K \times \mathbb{R}(G^m_\circ, z-u_m)+\sigma \leq 0\}, \\ \text{To find } (z_m,\sigma_m)\epsilon K_m \text{ such that } \sigma_m = \max_{(z,\sigma)\epsilon K_m} \sigma. \end{cases}$$

Since K itself can be assumed bounded using hypothesis $(HJ)_\circ(1)$ there is no need to introduce the bounded set V. When $z = z_m$ we have

$$(G^m_\circ, z_m - u_m) + \sigma \leq (G^m_\circ, z_m - u_m) + \sigma_m \leq 0 \quad \forall \sigma \epsilon \mathbb{R}$$

$$\text{i.e. } \min(G^m_\circ, z_m - u_m) + \sigma < 0.$$

The algorithm given by Franck and Wolfe was the first convex programming algorithm in the literature.

Remark 1.2. Our algorithm is a special case of a more general method known as Feasible direction method found by Zoutendjik [52].

Remark 1.3. We can repeat our method to give a slightly different algorithm in the choice of z_m as follows. We modify the set U_m used in the linear programming problem (P_m) by introducing certain parameters $\gamma_\circ, \gamma_1, \cdots, \gamma_k$ with σ. More precisely, we replace (1.1) by

$$(1.1)' \qquad \begin{cases} z - u_m \epsilon \mathscr{V} \\ (G^m_\circ, z-u_m) + \gamma_\circ \sigma \leq 0, \text{ and} \\ J^m_i + (G^m_i, z-u_m) + \gamma_i \sigma \leq 0 \text{ for } i = 1, \cdots, k, \end{cases}$$

where $\gamma_\circ, \gamma_1, \cdots, \gamma_k$ are certain suitably chosen parameters. This modified algorithm is useful when the curvature of the set U is small. 105

Remark 1.4. Suppose, in pur problem 1.1, some contraint J_i is such that $J_i(u_m) = J^m_i$ is "sufficiently negative" at some stage of the iteration (i.e. for some $m \geq 0$). Since J_i is regular then $J_i(v) \leq 0$ in a sufficiently small" ball with centre at u_m. This can be seen explicitely using Taylor's formula. Thus we can ignore the constraint J_i in the formulation of our problem i.e. in the definition of the set U.

Remark 1.5. The algorithm described in this section is not often used for minimizing problems arising from partial differential equation because the linear programming problem to be solved at each stage will be very large in this case. Hence our method will be expensive for numerical calculations for problems in partial diffeerential equation.

2 Centre Method

In this section we shall briefly sketch another algorithm to construct minimizing sequences for the minimizing problem for convex functionals on a finite dimensional space under constraints defined by a finite number of concave functionals. However we shall not prove the convergence of this algorithm. The main idea here is that at each step of the iteration we reduce the problem with constraints to one of a non-linear programming without contraints. An advantage with this method is that we do not use any regularity properties (i.e. existence of derivatives) of the functionals involved.

Let $V = \mathbb{R}^r$ and let

$$J_i : \mathbb{R}^r \to \mathbb{R}, i = 1, \cdots, k,$$

be continuous concave functionals (i.e. $-J_i$ are convex functionals). We define a set U by

$$U = \{v | v\epsilon\mathbb{R}^r, J_i(v) \geq 0 \text{ for all } i = 1, \cdots, k\}.$$

106

Since $-J_i$ are convex as in the previous section we see immediatly that U is a convex set.

Suppose given a functional $J_\circ : \mathbb{R}^r \to \mathbb{R}$ satisfying:

(1) $J_\circ$ is continuous,

(2) $J_\circ$ is strictly convex and

(3) $J_\circ(v) \to +\infty$ as $||v|| \to +\infty$.

We consider the following

Problem 2.1. To find $u \epsilon U$ such that

$$J_\circ(u) \leq J_\circ(v) \text{ for all } v \epsilon U.$$

As usual, in view of the hypothesis (3) on $J_\circ$, we may without loss of generality assume that U is bounded. We can describe the algorithm as follows.

Let $u_\circ \epsilon U$ be an initial point, arbitrarily fixed in U.

We shall find in our algorithm a sequence of triplets (u_m, u'_m, ℓ_m) where for each $m \geq 0, u_m, u'_m \epsilon U$ and ℓ_m is a sequence of real numbers such that $\ell_m \geq \ell_{m+1}$ $\forall_m$ and $\ell_m \geq J_\circ(u'_m)$.

We take at the beginning of the algorithm the triple $(u_\circ, u'_\circ, \ell_\circ)$ where $u'_\circ = u_\circ, \ell_\circ = J_\circ(u_\circ)$

Suppose we have determined (u_m, u'_m, ℓ_m). To determine the next triplet $(u_{m+1}, u'_{m+1}, \ell_{m+1})$ we proceed in the following manner.

Consider the subset U_m of U given by

$$(2.1) \qquad U_m = \{v | V \epsilon U, J_\circ(v) \leq \ell_m\}.$$

Since $J_\circ$ is convex and continuous it follows immediately that U_m is a bounded convex closed set in $\mathbb{R}^r$. Hence U_m is a compact convex set in $\mathbb{R}^r$.

107 We define a function $\varphi_m : \mathbb{R}^r \to \mathbb{R}$ by setting.

$$(2.2) \qquad \varphi_m(v) = (\ell_m - J_\circ(v)) \prod_{i=1}^{k} J_i(v).$$

The continuity of the functionals $J_\circ, J_1, \cdots, J_k$ immediatly imply that φ_m is also a continuous function. Moreover, φ_m has the properties of distance from the boundary of U_m. i.e.

(i) $\varphi_m(v) \geq 0$ for $v \epsilon U_m$.

(ii) $\varphi_m(v) = 0$ if v belongs to the boundary of U_m. i.e. For any v on any one of the $(k+1)$ -level surfaces defined by the equations

$$J_\circ(v) = \ell_m, J_1(v) = 0, \cdots, J_k(v) = 0$$

we have

$$\varphi_m(v) = 0.$$

Now since U_m is a compact convex set in $\mathbb{R}^r$ and φ_m is continuous it attains a maximum in U_m. $J_\circ$ being strictly convex this maximum is unique as can easily be checked.

We take u_{m+1} as the solution of the maximizing problem:

Problem 2.2$_m$. $u_{m+1} \epsilon U_m$ such that $\varphi_m(u_{m+1}) \geq \varphi_m(v), \forall v \epsilon U_m$.

Now suppose $u'_m \epsilon U_m$ so that $J_\circ(u'_m) \leq \ell_m$. This is true by assumption at the beginning of the algorithm (i.e. when $m = 0$). Hence $\varphi_m(u'_m) \geq 0$. We take a point u'_{m+1} such that

(2.3) $$u'_{m+1} \epsilon U_m \text{ and } J_\circ(u'_{m+1}) \leq J_\circ(u_{m+1}).$$

It is clear that such a point exists since we can take $u'_{m+1} = u_{m+1}$. However we shall choose u_{m+1} as follows: Consider the line $\Lambda(u'_m, u_{m+1})$ joining u'_m and u_{m+1}. We take for u'_{m+1} the point in U_m such that

(2.4) $$\begin{cases} u'_{m+1} \epsilon \lambda(u'_m . u_{m+1}) \cap \partial U_m, \\ \text{and } J_\circ(u'_{m+1}) \leq J_\circ(u_{m+1}). \end{cases}$$

108

Now we have onlyu to choose ℓ_{m+1}. For this, let r_m be a sequence of real numbers such that

(2.5) $$0 < \alpha \leq r_m \leq 1, \text{ where } \alpha > 0 \text{ is a fixed constant.}$$

We fix such a sequence arbitrarily in the beginning of the algorithm. We define ℓ_{m+1} by

(2.6) $$\ell_{m+1} = \ell_m - r_m(\ell_m - J_\circ(u'_{m+1})).$$

It is clear that $\ell_{m+1} \leq \ell_m$ and that $\ell_{m+1} \geq J_\circ(u'_{m+1})$. Thus we can state our algotrithm as follows:

Algorithm. Let $u_\circ \epsilon U$ be an arbitrarily fixed initial point. We determine a sequence of triplets (u_m, u'_m, ℓ_m) starting from $(u_\circ, u_\circ, J_\circ(u_\circ))$ as follows: Let (u_m, u'_m, ℓ_m) be given. Than $(u_{m+1}, u'_{m+1}, \ell_{m+1})$ is given by

(a) $u_{m+1} \epsilon U_m$ is the unique solution of the Problem 2.2_m.

(b) $u'_{m+1} \epsilon U_m$ is given by (2.4).

(c) ℓ_{m+1} is determined by (2.6).

Once again we can prove the convergence of this algorithm.

Remark 2.1. The maximization problem 2.2_m at each step of the iteration is a non-linear programming problem without constraints. For the soultion of such a problem we can use any of the algorithms described in Chapter 3.

Remark 2.2. Since the function φ_m which is maximized at each step has the properties of a distance function from the boundary of the domian U_m and is ≥ 0 in $U_m, \varphi_m > 0$ in $\overset{\circ}{U}_m$ and $\varphi_m = 0$ on U_m the maximum is
109 attained in the interior $\overset{\circ}{U}_m$ of U_m. This is the reason for the nomenclature of the algorithm as the Centre method. (See also [45]).

Remark 2.3. The algorithm of the centre method was first given by Huard [25] and it was improved later on, in particular, by Trémoliéres [45].

Remark 2.4. This method is once again not usded for functionals $J_\circ$ arising from problems for partial differential equations.

3 Method of Gradient and Prohection

We shall describe here a fairly simple type of algorithm for the minimization probelm for a regular convex functional on a closed convex subset of a Hilbert space. In this method we suppose that it is easy to find numerically projections onto closed convex subsets. At each step to construct the next iterate first we use a gradient method, as developed in Chapter 3, for the minimization problem without constraints and then we project on to the given convex set. “In the dual problem” which we shall study in Chapter 5 it is numerically easy to compute projections onto closed convex subsets and hence this method will be used there for a probelm for which the convex set is defined by certain constraints which we shall call dual constraints.

Let K be a closed convex subset of a Hilbert space V and $J : V \to \mathbb{R}$ be a functional on V. We make the following hypothesis on K and J.

(H1) K is a bounded closed convex set in V.

(H2) J is regular in V: J is twice G-differentiable everywhere in V and has a gradient $G(u)$ and hessian $H(u)$ everywhere in V. Moreover, there exists a constant $M > 0$ such that

$$||H(u)|| \leq M, \forall u \epsilon K.$$

(H3) H is uniformly coercive on K: there exists a constant $\alpha > 0$ such that

$$(H(u)\varphi, \varphi) \geq \alpha ||\varphi||^2, \forall \varphi \epsilon V \text{ and } u \epsilon K.$$

110

We note that the hypothesis of bounededness in $(H1)$ can be replaced by

$$(H1)' \qquad J(v) \to +\infty \text{ as } ||v|| \to +\infty.$$

Then we can fix a $u_\circ \epsilon K$ arbitrarily and restict our attention to the bounded closed convex set

$$K \cap \{v | v \epsilon V; J(v) \leq J(u_\circ)\}.$$

The hypothesis $(H3)$ implies that J is strongly convex. The hypothesis $(H2)$ implies that the gradient $G(u)$ is uniformly Lipschitz continuous on K and we have

$$(3.1) \qquad ||G(u) - G(v)|| \leq M||u - v||, \forall u, v \epsilon K.$$

We now consider the problem :

Problem 3.1. To find $u \epsilon K$ such that $J(u) \leq J(v)$, $\forall v \epsilon K$.

Algorithm. Let $u_\circ \in K$ be an arbitrarily fixed initial point of the algorithm and let $P : V \to K$ be the projection of V onto the bounded closed convex set K.

Suppose u_m is determined in the algorithm. The we define, for $\rho > 0$,

$$u_{m+1} = P(u_m - \rho G(u_m)). \tag{3.2}$$

Then we have the following

Theorem 3.1. *Under the hypothesis* $(H1) - (H3)$ *the Problem (3.1) has a unique solution u and $u_m \to u$ as $m \to +\infty$.*

This follows by a simple application of contraction mapping theorem.

Proof. Consider the mapping of K into itself defined by

$$T_\rho : K \ni u \mapsto P(u - \rho G(u) \epsilon K, \rho > 0. \tag{3.3}$$

□

111 Suppose this mapping T_ρ has a fixed point w. i.e.

$$w \epsilon K \text{ and satisfies } w = P(w - \rho G(w)).$$

Then we have seen that such a w is characterized as a solution of the variational inequality :

$$w \epsilon K; (w - (w - \rho G(w)), v - w) \geq 0, \forall v \epsilon K. \tag{3.4}$$

Then (3.4) is nothing but saying that

$$w \epsilon K; (G(w), v - w) \geq 0, \forall v \epsilon K. \tag{3.4$'$}$$

Then by Theorem 2.2 of Section 2, Chapter 2 w is a solution of the minimization Problem 3.1 and conversely. In other words, Problem 3.1 is equivalent to the following

Problem 3.1$'$. *To find a fixed points of the mapping $T_\rho : K \to K$. i.e. To find $w \in K$ such that $w = P(w - \rho G(w))$.*

We shall now show that this Problem (3.1)$'$ has a unique solution for $\rho > 0$ sufficiently small. For this we show that T_ρ is a strict contraction for $\rho > 0$ sufficiently small: there exists a constant $\gamma, 0 < \gamma < 1$ such that, for $\rho > 0$ small enough,

$$||P(u - \rho G(u)) - P(v - \rho G(u))|| \leq \gamma ||u - v||, \forall u, v \epsilon K.$$

In fact, if $\rho > 0$ is any number then we have

$$||P(u - \rho G(u)) - P(v - \rho G(v))||^2 \leq ||(u - \rho G(u)) - (v - \rho G(v))||^2$$

since $||P|| \leq 1$. The right hand side here is equal to

$$||u - v - \rho(G(u) - G(v))||^2 = ||u - v||^2 - 2\rho(G(u) - G(v), u - v) + \rho^2 ||G(u) - G(v)||^2$$

Here we can write by Taylor's formula

$$(G(u) - G(v), u - v) = (\overline{H}(u - v), u - v)$$

where $\overline{H} = H(v + \theta(u - v))$ for some $0 < \theta < 1$. Since K is convex, $u, v \epsilon K$, $v + \theta(u - v) \epsilon K$ and then by uniform coercivity of H on K (i.e by H3) 112

$$(H(u - v), u - v) \geq \alpha ||u - v||^2 \forall u, v \epsilon K.$$

This together with the Lipschitz continuity (3.1) of G gives

$$\begin{aligned} ||P(u - \rho G(u)) - P(v - \rho G(v))||^2 &\leq ||u - v||^2 - 2\rho\alpha ||u - v||^2 + M^2 \rho^2 ||u - v||^2. \\ &= ||u - v||^2 (1 - 2\rho\alpha + M^2 \rho^2). \end{aligned}$$

Now if we choose ρ such that

$$0 < \rho < 2\alpha / M^2 \tag{3.5}$$

it follows that $(1 - 2\rho\alpha + M^2 \rho^2) = \gamma^2 < 1$.

Then by contraction mapping theorem applied to T_ρ proves that there is a unique solution of the Problem (3.1)$'$.

Finally to show that $u_m \to u$ as $m \to +\infty$, we take such a $\rho > 0$ sufficiently small i.e. $\rho > 0$ satisfying (3.5). Now if u_{m+1} is defined iteratively by the algorithm (3.2) and u is the unique solution of the Problem 3.1 (or equivalently of the Problem (3.1)′) then,

$$\begin{aligned} ||u_{m+1} - u|| &= ||P(u_m - \rho G(u_m)) - P(u - \rho G(u))|| \\ &=\leq \gamma ||u_m - u|| \end{aligned}$$

so that we get

$$||u_{m+1} - u|| \leq \gamma^m ||u_\circ - u||.$$

Since $0 < \gamma < 1$ it follows immediatly from this that $u_m \to u$ as $m \to +\infty$.

This proves the theorem completely.

113 Now the convergence of the algorithm can be proved using the results of Chapter 3. (See Rosen [39], [40]).

We also remark that if $V = K$ and hypothesis $(H1)'$, $(H2)$ and $(H3)$ are satisfied for bounded sets of V then we get the gradirnt method of Chapter 3.

4 Minimization in Product Spaces

In this section we shall be concerned with the probelm of optimization with or without constraints by Gauss-Seidel or more generally, by relaxation methods. The classical Gauss-Seidel method is used for solutions of linear equations in finite dimensional spaces. The main idea of optimization described here is to reduce by an iterative procedure the problem of minimizing a functional on a product space (with or without constraints) to a sequence of minimization problems in the factor spaces. Thus the methods of earlier sections can be used to obtain approximations to the solution of the problem on the product space.

The method described here follows that of the paper of Céa and Glowinski [9], and generalizes earlier methods due to various authors.

We shall given algorithms for the construction of approximating sequences and prove that they converge to the solution of the optimization problem. One important feature is that we do not necessarily assume that the functionals to be minimized are G-differentiable.

4.1 Statement of the problem

The optimization problem in a product space can be formulated as follows: Let

(i) $V_i(i = 1, \cdots, N)$ be vector spaces over $\mathbb{R}$ and let

$$V = \prod_{i=1}^{N} V_i$$

(dim V_i are arbitrary). 114

(ii) K be a convex subset of V of the form $K = \prod_{i=1}^{N} K_i$ where each K_i is a (non-empty) convex subset of $V_i(i = 1, \cdots, N)$. Suppose given a functional $J : V \to \mathbb{R}$. Consider the optimization problem:

$$(4.1) \qquad \begin{cases} \text{To find } u \epsilon K \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon K. \end{cases}$$

For this problem we describe two algorithms which reduce the problem to a sequence of N problems at each step, each of which is a minimization problem successively in $K_i(i = 1, \cdots, N)$. Let us denote a point $v \epsilon V$ by its coordinates as

$$v = (v_1, \cdots, v_N), v_i \epsilon V_i.$$

Algorithm 4.1. (Gauss-Seidel method with constraints).

(1) Let $u^\circ = (u_1^\circ, \cdots, u_N^\circ)$ be an arbitrary point in K.

(2) Suppose $u^n \epsilon K$ is already determined. Then we shall determine u^{n+1} in N steps by successively computing its components u_i^{n+1} $(i = 1), \cdots, N$.

Assume $u_j^{n+1} \epsilon K_j$ is determined for all $j < i$. Then we determine u_i^{n+1} as the solution of the minimization problem:

$$(4.2) \qquad \begin{cases} u_i^{n+1} \epsilon K_i \text{ such that} \\ J(u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^n, \cdots, u_N^n) \\ \leq J(u_1^{n+1}, \cdots, u_{i-1}^{n+1}, v_i, u_{i+1}^n, \cdots, u_N^n) \text{ for all } v_i \epsilon K_i \end{cases}$$

In order to simplify the writing it is convenient to introduce the following notation.

115 **Notation.** Denote by $K_i^{n+1}(i = 1, \cdots, N)$ the subset of K:

$$K_i^{n+1} = \{v \epsilon K | v = (u_1^{n+1}, \cdots, u_{i-1}^{n+1}, v_i, u_{i+1}^n, \cdots, u_N^n), v_i \epsilon K_i\}. \tag{4.3}$$

and

$$\begin{cases} \widetilde{u}_o^{n+1} = u^n \\ \widetilde{u}_i^{n+1} = (u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^n, \cdots, u_N^n). \end{cases} \tag{4.4}$$

With this notation we can write (4.2) as follows:

$$\begin{cases} \text{To find } \widetilde{u}_i^{n+1} \epsilon K_i^{n+1} \text{ such that} \\ J(\widetilde{u}_i^{n+1}) \leq J(v) \text{ for all } v \epsilon K_i^{n+1}. \end{cases} \tag{4.2$'$}$$

Algorithm (4.2) (Relaxation method by blocks). We introduce numbers w_i with $0 < w_i < 2 (i = 1, 2, \cdots, N)$.

(1) Let $u^\circ \epsilon K$ be arbitrarily chosen.

(2) Assume $u^n \epsilon K$ is known. Then $u^{n+1} \epsilon K$ is determined in N successive steps as follows: Suppose $u_j^{n+1} \epsilon K_j$ is determined for all $j < i$. Then u_i^{n+1} is determined in two substeps:

$$\begin{cases} \text{To find } u_i^{n+\frac{1}{2}} \epsilon V_i \text{ such that} \\ J(u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i^{n+\frac{1}{2}}, u_{i+1}^n, \cdots, u_N^n) \\ \leq J(u_1^{n+1}, \cdots, u_{i-1}^{n+1}, v_i, u_{i+1}^n, \cdots, u_N^n) \text{ for all } v_i \epsilon V_i. \end{cases} \tag{4.5}$$

Then we define

$$u_i^{n+1} = P_i(u_i^n + w_i(u_i^{n+\frac{1}{2}} - u_i^n)) \tag{4.6}$$

where

(4.7) $P_i : V_i \to K_i$ is the projection onto K_i with respect to a suitable inner product which we shall specify later.

Remark 4.1. The numbers $w_i\epsilon(0,2)$ are called parameteres of relaxation. In the classical relaxation method each $w_i = w$, a fixed number $\epsilon(0,2)$ and $V_i = K_i$. Hence for the classical relaxation method

$$u_i^{n+1} = u_i^n + w(u_i^{n+\frac{1}{2}} - u_i^n). \tag{4.8}$$

116

4.2 Minimization with Constraints of Convex Functionals on Products of Reflexive Banach Spaces

Here we shall introduce all the necessary hypothesis on the functional J to be minimized. We consider J to consist of a differentiable part $J_\circ$ and a non-differentiable part J_1 and we make separate hypothesis on $J_\circ$ and J_1.

Let $V_i(i = 1,\cdots,N)$ be reflexive Banach spaces and $V = \prod_{i=1}^N V_i$. The duality pairing $(\cdot,\cdot)_{V'\times V}$ will simply be denoted by $(\cdot,\cdot)$, then norm in V by $\|\cdot\|$ and the dual norm in V' by $\|\cdot\|_*$. Let K_i be nonempty closed convex subsets of V_i and $K = \prod_{i=1}^N K_i$. Then clearly K is also a noneempty closed convex subset of V.

Let $J_\circ : V \to \mathbb{R}$ be a functional satisfying the following hypothesis:

(H1) $J_\circ$ is G-differentiable and admits a gradient $G_\circ$.

(H2) $J_\circ$ is convex in the following sense: If, for any $M > 0$, B_M denotes the ball $\{v\epsilon V; \|v\| \leq M\}$, then there exists a mapping

$$T_M : B_M \times B_M \to \mathbb{R}$$

such that (4.9) and (4.10) hold:

$$\begin{cases} J_\circ(v) \geq J_\circ(u) + (G_\circ(u), v-u) + T_M(u,v), \\ T_M(u,v) \geq 0 \text{ for all } u,v\epsilon B_M, \\ T_M(u,v) > 0 \text{ for all } u,v\epsilon B_M \text{ with } u \neq v. \end{cases} \tag{4.9}$$

$$\begin{cases} \text{If } (u_n,v_n)_n \text{ is a sequence in } B_M \times B_M \text{ such that} \\ T_M(u_n,v_n) \to 0 \text{ as } n \to +\infty \text{ ther} \\ \|u_n - v_n\| \to 0 \text{ as } n \to +\infty. \end{cases} \tag{4.10}$$

117

Remark 4.2. If $J_\circ$ is twice G-diffferentiable then we have

$$T_M(u,v) = \frac{1}{2} J''_\circ(u+\theta(v-u), v-u, v-u) \text{ for some } 0<\theta<1.$$

Then the hypothesis (4.9) and (4.10) can be restated in terms of $J''_\circ$. In particular, if $J_\circ$ admits a Hessian H and if for every $M>0$ there exists a constant $\alpha_M > 0$ such that

$$(H(u)\varphi,\varphi) \geq \alpha_M ||\varphi||^2 \text{ for all } \varphi \epsilon V \text{ and } u \epsilon B_M$$

then the two conditions (4.9) and (4.10) are satisfied.

(H3) *Continuity of the gradient $G_\circ$ of $J_\circ$.*

(4.11) $$\begin{cases} \text{If } (u_n,v_n)_n \text{ is a sequence in } B_M \times B_M \text{ such that} \\ ||u_n - v_n|| \to \text{ as } n \to +\infty \text{ then} \\ ||G(u_n) - G(v_n)||_* \to 0 \text{ as } n \to +\infty. \end{cases}$$

Next we consider the non-differentiable part J_1 of J. Let $J_1 : V \to \mathbb{R}$ be a functional of the form

(4.12) $$J_1(v) = \sum_{i=1}^{N} J_{1,i}(v_i), v = (v_1, \cdots, v_n) \epsilon V$$

where the functionals

$$J_{1,i} : V_i \to \mathbb{R} (i = 1, \cdots, N)$$

satisfy the hypothesis:

(H4) $J_{1,i}$ is a weakly lower semi-continuous convex functional on V_i.

118 We define

(4.13) $$J = J_\circ + J_1.$$

Finally we assume that J satisfies the hypothesis:

(H5) $J(v) \to +\infty$ as $||v|| \to +\infty$. We now consider the minimization problem:

(4.14) $$\begin{cases} \text{To find } u \epsilon K \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon K. \end{cases}$$

4.3 Main Results

The main theorem of this section can now be stated as:

Theorem 4.1. *Under the hypothesis* $(H1), \cdots, (H5)$ *we have the following:*

(1) The problem (4.14) has a unique solution $u \epsilon K$ *and the unique soultion is characterized by*

$$(4.15) \qquad \begin{cases} u \epsilon K \text{ such that} \\ G_\circ(u), v-u) + J_1(v) - J_1(u) \geq 0 \text{ for all } v \epsilon K. \end{cases}$$

(2) The sequence u^n *determined by the algorithm (4.1) converges strongly to* u *in* V.

Proof. We shall divide the proof into several steps.

Step 1. (Proof of (1)). The first part of the theorem is an immediate consequence of the Theorem (1.1) and (2.3) of Chapter 2. In fact, K is a closed non-empty convex subset of a reflexive Banach space V. By Hypothesis $(H2)$, J is strictly convex since, for any $v, u \epsilon V$, we have

$$\begin{aligned} J_\circ(v) &\geq J_\circ(u) + (G_\circ(u), v-u) + T_M(v,u) \\ &> J_\circ(u) + (G_\circ(u), v-u) \qquad \text{if } v \neq u, \end{aligned}$$

and hence strictly convex, while $J_1(v)$ is convex so that for any $v_1, v_2 \epsilon V$ **119** and $\theta \epsilon [0,1]$ we have

$$\begin{aligned} J(\theta v_1 + (1-\theta)v_2) &= J_\circ(\theta v_1 + (1-\theta)v_2) + J_1(\theta v_1 + (1-\theta)v_2) \\ &< \theta J_\circ(v_1) + (1-\theta)J_\circ(v_2) + \theta J_1(v_1) + (1-\theta)J_1(v_2) \\ &= \theta J(v_1) + (1-\theta)J(v_2). \end{aligned}$$

Next J is weakly lower semi-continuous in V: In fact, since $J_\circ$ has a gradient $G_\circ$ the mapping

$$\varphi \mapsto J'_\circ(u, \varphi) = (G_\circ(u), \varphi)$$

is continuous linear and hence, by Proposition 4.1 of Chapter 1, $J_\circ$ is weakly lower semi-continuous. On the other hand, by $(H4)$ J_1 is weakly lower semi-continuous which proves the assertion. Then Theorem (1.1) of Chapter 2 implies that states that u is characterized by (4.15).

We have therefore onlu to prove (2) of the statement. We shall prove the convergence of the algorithm in the following sequence of steps.

Step 2. At each stage of the algorithm the subproblem of determining $\widetilde{u}_i^{n+1}$ has a solution. In fact K_i^{n+1} is againd a non-empty closed convex subset of V. Moreover, again as in Step 1, J satisfies all the hypothesis of Theorem (1.1) of Chapter 2 and (2.3) of Chapter 2. Hence there exists a unique solution of the problem (4.14) and this soution $\widetilde{u}_i^{n+1}$ is characterized by

$$(4.16) \qquad \begin{cases} \widetilde{u}_i^{n+1} \epsilon K, \\ (G_\circ(\widetilde{u}_i^{n+1}), v - \widetilde{u}_i^{n+1}) + J_{1,i}(v_i) - J_{1,i}(\widetilde{u}_i^{n+1}) \geq 0 \end{cases}$$

since

$$J_1(v) - J_1(\widetilde{u}_i^{n+1}) = \sum_{j=1}^{N} (J_{1,j}(v_j) - J_{1,j}(\widetilde{u}_{i,j}^{n+1})) = J_{1,i}(v_i) - J_{1,i}(u_i^{n+1}).$$

120

Step 3. $J(u^n)$ is decresing. We know that $\widetilde{u}_{i-1}^{n+1} \epsilon K_i^{n+1}$ for $i = 1, \cdots, N$ and on taking $v = \widetilde{u}_{i-1}^{n+1}$ in $(4.2)'$ we get

$$J(\widetilde{u}_i^{n+1}) \leq J(\widetilde{u}_{i-1}^{n+1}).$$

using this successively we find that

$$J(\widetilde{u}_i^{n+1}) \leq J(\widetilde{u}_{i-1}^{n+1}) \leq \cdots \leq J(\widetilde{u}_\circ^{n+1}) = J(u^n)$$

and similarly

$$J(u^{n+1}) = J(\widetilde{u}_N^{n+1}) \leq \cdots \leq J(\widetilde{u}_i^{n+1}).$$

These two togrther imply that

$$J(u^{n+1}) \leq J(u^n) \text{ ofr all } n = 0, 1, 2, \cdots$$

which proves that the sequence $J(u^n)$ is decreasing. In particular it is bounded above:

$$J(u^n) \leq J(u^\circ) \text{ ofr all } n \geq 1.$$

Since $u \epsilon K$ is the unique absolute minimum for J given by Step (1) we have

$$J(u) \leq J(u^n) \leq J(u^\circ) \text{ for all } n \geq 1.$$

On the other hand, by Hypothesis $(H5)$ we see that $\|u^n\|$, $\|\widetilde{u}_i^{n+1}\|$ form bounded sequences. Thus there exists a constant $M > 0$ such that

$$(4.17) \qquad \|u^n\| + \|\widetilde{u}_i^{n+1}\| + \|u\| \leq M \text{ for all } n \geq 1 \text{ and all } 1 \leq i \leq N.$$

Since

$$J(u) \leq J(u^{n+1}) \leq J(u^n)$$

it also follows that 121

$$(4.18) \qquad J(u^n) - J(u^{n+1}) \to 0 \text{ as } n \to +\infty.$$

Step 4. We shall that $u^n - u^{n+1} \to 0$ as $n \to +\infty$. For this, by the-convexity hypothesis $(H2)$ of $J_\circ$ applied to $u = \widetilde{u}_i^{n+1}$ and $v = \widetilde{u}_{i-1}^{n+1}$ we get

$$J_\circ(\widetilde{u}_{i-1}^{n+1}) \geq J_\circ(\widetilde{u}_i^{n+1}) + (G_\circ(\widetilde{u}_i^{n+1}), \widetilde{u}_{i-1}^{n+1} - \widetilde{u}_i^{n+1}) + T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1})$$

where $M > 0$ is determined by (4.17) in Step (3). From this we find

$$J(\widetilde{u}_{i-1}^{n+1}) \geq J(\widetilde{u}_i^{n+1}) + [(G_\circ(\widetilde{u}_i^{n+1}), \widetilde{u}_{i-1}^{n+1} - \widetilde{u}_i^{n+1}) + J_1(\widetilde{u}_{i-1}^{n+1}) - J_1(\widetilde{u}_i^{n+1})]$$
$$+ T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1}).$$

Here by the characterization (4.16) of $\widetilde{u}_i^{n+1} \epsilon K_i^{n+1}$ as the solution sub-problem we see that the terms in the brackets $[\cdots] \geq 0$ and hence

$$J(\widetilde{u}_{i-1}^{n+1}) \geq J(\widetilde{u}_i^{n+1}) + T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1}) \text{ for all } i = 1, \cdots, N.$$

Adding there inequalities for $i = 1, \cdots, N$ we obtain

$$J(\widetilde{u}_\circ^{n+1}) = J(u^n) \geq J(\widetilde{u}_N^{n+1}) + \sum_i T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1})$$

$$= J(u^{n+1}) + \sum_i T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1}),$$

that is,

$$J(u^n) - J(u^{n+1}) \geq \sum_i T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1}).$$

Here the left side tends to 0 as $n \to \infty$ and each term in the sum on the right side is non-negative by (4.9) of Hypothesis ($H2$) so that

$$T_M(\widetilde{u}_i^{n+1}, \widetilde{u}_{i-1}^{n+1}) \to 0 \text{ as } n \to +\infty \text{ for all } i = 1, \cdots, N.$$

In view of (4.10) of Hypothesis ($H2$) it follows that

$$(4.19) \quad \begin{cases} \|\widetilde{u}_i^{n+1} - \widetilde{u}_{i-1}^{n+1}\| \to 0 \text{ as } n \to +\infty \text{ for all } i = 1, \cdots, N \text{ and} \\ \|u^{n+1} - u^n| \to 0 \text{ as } n \to +\infty \end{cases}$$

122 which proves the required assertion.

Step 5. Convergence of the algorithm. Using the convexity Hypothesis ($H2$) of $J_\circ$ with u and v interchanged we get

$$J_\circ(v) \geq J_\circ(u) + (G_\circ(u), v - u) + T_M(u, v)$$
$$J_\circ(u) \geq J_\circ(v) + (G_\circ(v), v - u) + T_M(v, u)$$

which on adding give

$$(4.20) \quad \begin{cases} (G_\circ(v) - G_\circ(u), v - u) \geq R_M(v, u) \\ \text{where} \\ R_M(v, u) = T_M(u, v) + T_M(v, u). \end{cases}$$

Taking for u the unique solution of the problem (4.14) and $v = u^{n+1}$ we obtain

$$(G_\circ(u^{n+1}) - G_\circ(u), u^{n+1} - u) \geq R_M(u, u^{n+1})$$

from which we get

$$(4.21) \quad \begin{cases} (G_\circ(u^{n+1}), u^{n+1} - u) + J_1(u^{n+1}) - J_1(u) \\ \geq [(G_\circ(u), u^{n+1} - u) + J_1(u^{n+1}) - J_1(u)] + R_M(u, u^{n+1}) \\ \geq R_M(u, u^{n+1}) \end{cases}$$

since u is characterized by (4.15). Introducting the notation

$$w_i^{n+1} = \widetilde{u}_i^{n+1} + (0, \cdots, 0, u_i - u_i^{n+1}, 0, \cdots, 0)$$

we have

$$\text{(4.22)} \qquad \begin{cases} w_i^{n+1} = (u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i, u_{i+1}^n, \cdots, u_N^n) \epsilon K_i^{n+1} \\ \sum_i (w_i^{n+1} - \widetilde{u}_i^{n+1}) = (u - u^{n+1}). \end{cases}$$

Now we use the fact that $J_1(v) = \sum_i J_{1,i}(v_i)$ to get 123

$$J_1(u^{n+1}) - J_1(u) = \sum_i (J_{1,i}(u_i^{n+1}) - J_{1,i}(u_i)),$$

which is the same as

$$\text{(4.23)} \qquad J_1(u^{n+1}) - J_1(u) = \sum_i (J_1(\widetilde{u}_i^{n+1}) - J_1(w_i^{n++1})).$$

Substituting (4.22) and (4.23) in (4.21) we have

$$\begin{cases} \sum_i [(G_\circ(u^{n+1}), \widetilde{u}_i^{n+1} - w_i^{n+1}) + J_1(\widetilde{u}_i^{n+1}) - J_1(w_i^{n+1})] \\ \geq R_M(u, u^{n+1}). \end{cases}$$

This can be rewritten as

$$\sum_i (G_\circ(u^{n+1}) - G_\circ(\widetilde{u}_i^{n+1}), \widetilde{u}_i^{n+1} - w_i^{n+1})$$
$$\sum_i [(G_\circ(\widetilde{u}_i^{n+1}), w_i^{n+1} - \widetilde{u}_i^{n+1}) + J_1(w_i^{n+1}) - J_1(\widetilde{u}_i^{n+1})] + R_M(u, u^{n+1}).$$

But again by the characterization (4.16) of the solution $\widetilde{u}_i^{n+1} \epsilon K_i^{n+1}$ of the sub-problem (4.14) the terms in the square brackets and hence their sum is non negative (to see this we take $v = w_i^{n+1} \epsilon K_i^{n+1}$). Thus

$$\text{(4.24)} \qquad \sum_i (G_\circ(u^{n+1}) - G_\circ(\widetilde{u}_i^{n+1}), \widetilde{u}_i^{n+1} - w_i^{n+1}) \geq R_M(u, u^{n+1}).$$

Here we have

$$\|\widetilde{u}_i^{n+1} - w_i^{n+1}\|_V = \|u_i - u_i^{n+1}\|_{V_i} \leq \|u\| + \|\widetilde{u}_i^{n+1}\| \leq M.$$

By Cauchy-Schwarz inequality we have

$$|(G_\circ(u^{n+1}) - G_\circ(\widetilde{u}_i^{n+1}), \widetilde{u}_i^{n+1} - w_i^{n+1})| \leq M\|G_\circ(u^{n+1}) - G_\circ(\widetilde{u}_i^{n+1})\|_*.$$

Now since

$$\|u^{n+1} - \widetilde{u}_i^{n+1}\| = \|\widetilde{u}_N^{n+1} - \widetilde{u}_i^{n+1}\| \leq \sum_{j=i+1}^{N} \|\widetilde{u}_j^{n+1} - \widetilde{u}_{j-1}^{n+1}\|$$

124 which tends to 0 by (4.19) and since $G_\circ$ satisfies the continuity hypothesis (4.11) of $(H3)$ it follows that

$$R_M(u, u^{n+1}) \to 0 \text{ as } n \to \infty.$$

This by the definition of $R_M(u, v)$ implies that

$$T_M(u, u^{n+1}) \to 0 \text{ as } n \to \infty.$$

Finally, by the property (4.10) to $T_M(u, v)$ in Hypothesis $(H2)$ we conclude that

$$\|u - u^{n+1}\| \to 0 \text{ as } n \to \infty.$$

This completes the proof of the theorem. □

Remark 4.3. If the convex set K is bounded then the Hypothesis $(H5)$ is superfluous since the existence of the constant $M > 0$ in (4.17) is then automatically assured since $u, u^n, \widetilde{u}_i^{n+1} \epsilon K$ for all $n \geq 1$ and $i = 1, \cdots, N$.

4.4 Some Applications : Differentiable and Non-Differatiable Functionals in Finite Dimensions

We shall conclude this section with a few examples as applications of our main result (Theorem 4.1) without going into the details of the proofs. To begin with have the following:

Theorem 4.2. *(Case of differentaible functionals on the finite dimensional spaces).*

Let $J_\circ : V = \mathbb{R}^p \to \mathbb{R}$ be a functional satisfying the Hypothesis:

(K1) $J_\circ \epsilon C^1(\mathbb{R}^p, \mathbb{R})$

(K2) $J_\circ$ *is strictly convex*

(K3) $J_\circ(v) \to +\infty$ *as* $||v|| \to +\infty$.

Then the assertion of the Theorem (4.1) hold with $J = J_\circ$.

It is immediate that the Hypothesis $(H1)$ and $(H3)$ are satisfied. Since $J_1 \equiv 0$, $(H4)$ and $(H5)$ are also satisfied. There remains only to prove that the Hypothesis $(H2)$ of the convexity of $J_\circ$ holds. For a proof of this we refer to the paper of Céa and Glowinski [9]. (See also Glowinski [18], [19]). 125

Remark 4.4. Suppose $p = \sum\limits_{i}^{N} p_i$ be a partition of p. Then in the above theorem we can take $V_i = \mathbb{R}^{p_i}$ so that $V = \prod\limits_{i=1}^{N} V_i$. We also have the

Theorem 4.3. *(Case of non-differentiable functions on finite dimensional spaces - Cea and Glowinski). Let* $V_i = \mathbb{R}^{p_i}(i = 1, \cdots, N)$ *and* $V = \mathbb{R}^p(p = \sum\limits_{i=1}^{N} p_i)$. *Suppose* $J_\circ : V \to \mathbb{R}$ *satisfies the hypothesis (K1), (K2) and (K3) pf Theorem (4.2) above and* $J_1 : V \to \mathbb{R}$ *be another functional of the form* $J_1(v) = \sum\limits_{i=1}^{N} J_{1,i}(v_i)$ *where the functionals* $J_{1,i} : V_i \to \mathbb{R}$ *satisfy the Hypothesis below:*

$(K4)J_{1,i}$ *is a non-negative, convex and continuous functional on* $\mathbb{R}^{p_i} = V_i(i = 1, \cdots, N)$.

Then the functional

$$J = J_\circ + J_1$$

satisfies all the Hypothesis of Theorem (4.1) and hence the algorithm (4.1) is (strongly) convergent in $V = \mathbb{R}^p$.

We shall now give a few examples of functional J_1 which satisfy (K4).

Example 4.1. We take $J_{1,i}(v_i) = \alpha_i|\ell_i(v_i)|$ where

(i) $\alpha_i \geq 0$ are fixed numbers

(ii) $\ell_i : V_i = \mathbb{R}^{p_i} \to \mathbb{R}$ is a continuous linear functional for each $i = 1, \cdots, N$.

In particular, if $p_i = 1(i = 1, \cdots, N)$ and hence $p = N$ we can take

$$J_{1,i}(v_i) = \alpha_i |v_i|,$$

and

$$J_1(v) = \sum_{i=1}^{N} \alpha_i |v_i|.$$

126 This case was treated earlier by Auslander [53] who proved that the algorithm for u^n converges to the solution of the minimization problem in this case.

Example 4.2. We take

$$J_{1,i}(v_i) = \alpha_i [\ell_i(v_i)^+]$$

where

(i) $\alpha_i \geq 0$ are fixed numbers,

(ii) $\ell_i : V_i \to \mathbb{R}$ are continuous linear forms on $\mathbb{R}^{p_i}$, and we have used the standard notation:

$$\ell_i(v_i)^+ = \begin{cases} \ell_i(v_i) \text{ when } \ell_i(v_i) \geq 0 \\ 0 \text{ when } \ell_i(v_i) < 0. \end{cases}$$

Example 4.3. We take

$$J_{1,i}(v_i) = \alpha_i ||v_i||_{\mathbb{R}^{p_i}}$$

where

$$||v_i||_{\mathbb{R}^{p_i}} = \left(\sum_{j=1}^{p_i} |v_{i,j}|^2 \right)^{\frac{1}{2}}.$$

4.5 Minimization of Quadratic Functionals on Hilbert Spaces-Relaxation Method by Blocks

Here we shall be concerned with the problem of minimization of quadratic funcitonals on convex subsets of a product of Hilbert spaces. This is one of the most used methods for problems associated with partial differential equations. We shall describe an algorithm and prove the convergence of the approximations (obtained by this algorithm) to the solution of the minimization problem under consideration.

Statement of the problem. Let $V_i(i = 1, 2, \cdots N)$ be Hilbert spaces, the inner products and the norms are respectively denoted by $((\cdot))_i$ and $||\cdot||_i$. On the product space we define the natural inner product and norm by 127

$$(4.25) \qquad \begin{cases} ((u, v)) = \sum\limits_{i=1}^{N} ((u_i, v_i))_i, \\ ||u|| = \left(\sum\limits_{i=1}^{N} ||u_i||_i^2\right)^{\frac{1}{2}}, \\ u = (u_1, \cdots, u_n), v = (v_1, \cdots, v_n)\epsilon V, \end{cases}$$

for which V becomes a Hilbert space. Let K be a closed convex subset of V of the form

$$(4.26) \qquad \begin{cases} K = \prod_{i=1}^{N} K_i \text{ where} \\ K_i \text{ is a closed convex nonempty subset of } V_i (1 \leq i \leq N). \end{cases}$$

Let $J : V \to \mathbb{R}$ be a functional of the form

$$(4.27) \qquad J(v) = \frac{1}{2} a(v, v) - L(v)$$

where $a(\cdot, \cdot)$ is a bilinear, symmetric, bicontinuous, V-coercive form on V:

There exist constants $M > 0$ and $\alpha > 0$ such that

$$(4.28) \qquad \begin{cases} |a(u, v)| \leq M||u||_V||v||_V & \text{for all } u, v\epsilon V, \\ a(u, u) \geq \alpha||u||_V^2 & \text{for all } u\epsilon V, \text{ and} \\ a(u, v) = a(v, u) \end{cases}$$

Moreover, $L : V \to \mathbb{R}$ is a continuous linear functional on V. Consider the optimization problem :

$$(4.29) \qquad \begin{cases} \text{To find } u \epsilon K \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon K. \end{cases}$$

Then we know by Theorem 3.1 of Chapter 2 that under the assumptions made on V, K and J the optimization problem (4.29) has a unique solution whihc is characterized by the variational inequality

$$(4.30) \qquad \begin{cases} u \epsilon K. \\ a(u, v-u) - L(v-u) \geq 0 \text{ for all } v \epsilon K. \end{cases}$$

128

4.6 Algorithm (4.2) of the Relaxation Method - Details

In order to give an algorithm for the solution of the problem (4.29) we obtain the following in view of the product Hilbert space structure of V. First of all, we observe that the bilinear form $a(\cdot, \cdot)$ give rise to bilinear forms

$$(4.31) \qquad a_{ij} : V_i \times V_j \to \mathbb{R}$$

such that

$$a(u, v) = \sum_{i,j=1}^{N} a_{ij}(v_i, v_j).$$

In fact, for any $v_i \epsilon V_i$ if we set v^i to be the element of V having components $(v^i)_j = 0$ for $j \neq 1$ and $(v^i)_i = v_i$, we define

$$(4.33) \qquad a_{ij}(v_i, v_j) = a(v^i, v^j).$$

It is the clear that the properties (4.28) of $a(\cdot . \cdot)$ immediately imply the following properties of $a_{ij}(\cdot, \cdot)$:

$$(4.34) \qquad \begin{cases} a_{ij} \text{ is bicontinuous } : |a_{ij}(v_i, v_j)| \leq M \|v_i\|_i \|v_j\|_j. \\ a_{ij}(v_i, v_j) = a_{ji}(v_j, v_i) \\ a_{ii} \text{ is } V_i - \text{coercive } : a_{ii}(v_i, v_i) = a(v^i, v^i) \geq \alpha \|v^i\|^2 = \alpha \|v^2_{i\|_i} \\ \text{for all } v_i \in V_i, \; v_j \epsilon V_j \end{cases}$$

Using the bicontinuity of the bilinear forms $a_{ij}(\cdot,\cdot)$ together with Riesz-representation theorem, we can find

$$A_{ij}\epsilon\mathscr{L}(V_i,V_j) \text{ suich that}$$
$$a_{ij}(v_i,v_j)=(A_{ij}v_i,v_j)_{V'_j\times V_j} \tag{4.35}$$

where $(\cdot,\cdot)_{V'_j\times V_j}$ denotes the duality pairinig between V_j and its dual V'_j (which is canonically isomorphic to V_j). The properties (4.34) can equivalently be stated in the following form: 129

$$\text{(4.34)}' \qquad \begin{cases} ||A_{ij}||_{\mathscr{L}(V_i,V_j)} \leq M, \\ A_{ij}=A^*_{ij}, A_{ii} \text{ are self adjoint} \\ (A_{ii}v_i,v_i)_{V'_i\times V_i} \geq \alpha||v_i||^2_i \text{ for all } v_i\epsilon V_i. \end{cases}$$

By lax-Milgram lemma A_{ii} are invertible and $A^{-1}_{ii}\epsilon\mathscr{L}(V_i,V_i)$.

In a similar way, we find the forms L defines continuous linear functionals $L_i: V_i\to\mathbb{R}$ such that

$$\begin{cases} L_i(v_i)=L(v^i) \text{ for all } v_i\epsilon V_i \\ L(v)=\sum_{i=1}^N L_i(v_i) \text{ for all } v\epsilon V. \end{cases}$$

Again by Riesz-representation theorem there exist $F_i\epsilon V_i$ such that

$$L_i(v_i)=((F_i,v_i))_i \text{ for all } v_i\epsilon V_i$$

so that we can write

$$L(v)=\sum_{i=1}^N ((F_i,v_i))_i. \tag{4.36}$$

As an immediate consequence of the properties of the bilinear forms $a_{ii}(\cdot,\cdot)$ on V_i we can introduce a new inner product on V_i by

$$[u_i,v_i]_{V_i}=a_{ii}(u_i,v_i). \tag{4.37}$$

which defines an equivalent norm which we shall denote by $|||\cdot|||_i$ (we can use Lax-Milgram lemma) on V_i.

We shall denote by P_i the projection of V_i onto the closed convex subset K_i with respect to the inner product $[\cdot,\cdot]_i$.

We are now in a position to describe the algorithm for the relaxation method with projection. (See also [19]). 130

Algorithm 4.2. - Relaxation with Projection by Blocks.

Let $w_i(i = 1, \cdots, N)$ be a fixed set of real numbers such that $0 < w_i < 2$.

(1) Let $u^\circ = (u_1^\circ, \cdots, u_N^\circ)\epsilon K$ be arbitrary.

(2) Suppose $u^n \epsilon K$ is already determined. We determine $u^{n+1}\epsilon K$ in N successive steps as follows: Suppose, $u_j^{n+1}\epsilon K$ are already found for $j < i$.

Then we take

$$\text{(4.38)} \qquad \begin{cases} u_i^{n+1} = P_i(u_i^n - w_i A_{ii}^i(\sum\limits_{j<i} A_{ij}u_j^{n+1} + \sum\limits_{j\geq i} A_{ij}u_j^n - F_i)) \\ i = 1, \cdots, N. \end{cases}$$

Remark 4.5. In applications, the boundary value problems associated with elliptic partial differential operators will be set in appropriate Sobolev spaces $H^m(\Omega)$ on some (bounded) open set Ω in Euclidean space. After discretization (say, by suitable finite elemnt approximations) we are led to problems in finite dimensional subspaces of $H^m(\Omega)$ which increase to $H^m(\Omega)$. In such a discretization A_{ii} and A_{ij} will be matrices with the properties (4.34)$'$ described above.

4.7 Convergence of the Algorithm

As usual we shall prove that the algorithm converges to the solution of the minimization problem (4.29) in a sequence of steps in the following. We shall begin with

Step 1. $J(u^n)$ **is a decreasing sequence.** For this we write

$$\text{(4.39)} \qquad J(u^n) - J(u^{n+1}) = J(\widetilde{u}_\circ^{n+1}) - J(\widetilde{u}_N^{n+1})$$

$$= \sum_{i=1}^{N} (J(\widetilde{u}_{i-1}^{n+1}) - J(\widetilde{u}_i^{n+1}))$$

and show that each term in tha last sum is non-negqtive. We observe here that 131

$$(4.40) \qquad \begin{cases} \widetilde{u}_{i-1}^{n+1} = (u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i^n, u_{i+1}^n, \cdots, u_N^n) \\ \widetilde{u}_i^{n+1} = (u_1^{n+1}, \cdots, u_{i-1}^{n+1}, u_i^{n+1}, u_{i+1}^n, \cdots, u_N^n). \end{cases}$$

Setting, for each $i = 1, \cdots, N$,

$$(4.41) \qquad \begin{cases} g_i & = -\frac{1}{2} \sum_{j<i} A_{ij} u_j^{n+1} + \frac{1}{2} \sum_{j>i} A_{ij} u_j^n + f_i \\ j_i(v_i) & = \frac{1}{2}(A_{ii} v_i, v_i) - (g_i, v_i) \end{cases}$$

we immediately see that

$$(4.42) \qquad J(\widetilde{u}_{i-1}^{n+1}) - J(\widetilde{u}_i^{n+1}) = j_i(u_i^n) - j_i(u_i^{n+1}).$$

Hence it is enough to show that the right hand side of (4.42) is non-negative. In fact, we shall prove the following

Proposition 4.1. *For each $i, 1 \leq i \leq N$, we have*

$$(4.43) \qquad j_i(u_i^n) - j_i(u_i^{n+1}) \geq \frac{2 - w_i}{2w_i} |||u_i^n - u_i^{n+1}|||.$$

The proof will be based on some simple lemmas:

Step 2. Two lemmas. Let H be a Hilbert space and C be a non-empty closed convex subset of H. Consider a quadratic functional $j : H \to \mathbb{R}$ of the form

$$(4.44) \qquad j(v) = \frac{1}{2} b(v, v) - (g, v)$$

where

$$(4.45) \qquad \begin{cases} b(\cdot, \cdot) \text{ is a symmetric, bicontinuous, } H\text{-coercive} \\ \text{bilinear form on } H \text{ and } g \epsilon H. \end{cases}$$

Then we know by Theorem 3.1 of Chapter 2 that the minimization problem

$$(4.46) \qquad \begin{cases} \text{To find } u \epsilon C \text{ such that} \\ j(u) \leq j(v) \text{ for all } v \epsilon C \end{cases}$$

132 has a unique solution. On the other hand, the hypothesis on $b(\cdot, \cdot)$ imply that we can write

$$\begin{cases} b(u, v) = v(Bv) \text{ for all } u, v \epsilon H \\ \text{and} \\ B \epsilon \mathscr{L}(H, H), B = B^* \text{ exists and belongs to } (H, H) \end{cases}$$

Moreover,

$$(4.48) \qquad [u, v] = b(u, v) = (u, Bv)$$

defines an inner product on H such that

$$(4.49) \qquad u \mapsto u = [u, u]^{\frac{1}{2}}$$

is an equivalent norm in H. Then we have the

Lemma 4.1. *If $u \epsilon C$ is the unique solution of the problem (4.46) and if $P : H \to C$ denotes the projection onto C with respect to the inner product $[\cdot, \cdot]$ then*

$$(4.50) \qquad u = P(B^{-1} g).$$

Proof. We also know that the solution of the problem (4.46) is characterized by the variational inequality

$$(4.51) \qquad \begin{cases} u \epsilon C, \\ b(u, v - u) \geq (g, v - u) \text{ for all } v \epsilon C. \end{cases}$$

□

Since we can write

$$(g, v-u) = (BB^{-1}g.v-u) = b(B^{-1}g, v-u) \tag{4.52}$$

this variational inequality can be rewritten in the form

$$\begin{cases} u\epsilon C, \\ [u-B^{-1}g, v-u] = b(u-B^{-1}g, v-u) \geq 0 \text{ for all } v\epsilon C. \end{cases} \tag{4.51$'$}$$

133

But it is a well known fact that this new variational inequality characterizes the projection $P(B^{-1}g)$ with respect to the inner product $[\cdot,\cdot]$ (For a proof, see for instance Stampacchia [44]).

Lemma 4.2. *Let $u_\circ\epsilon C$. If u_1 is defined by*

$$u_1 = P(u_\circ + w(B^{-1}g - u_\circ)), w > 0. \tag{4.53}$$

where P is the projection $H \to C$ with respect to $[\cdot,\cdot]$ then

$$j(u_\circ) - j(u_1) \geq \frac{2-w}{2w} |||u_\circ - u_1|||^2. \tag{4.54}$$

Proof. If $v_1, v_2 \epsilon H$ then we have

$$\begin{aligned} j(v_1) - j(v_2) &= \frac{1}{2}\{b(v_1, v_1) - b(v_2, v_2)\} - \{(g, v_1) - (g, v_2)\} \\ &= \frac{1}{2}\{b(v_1, v_1) - b(v_2, v_2)\} - (BB^{-1}g, v_1 - v_2) \\ &= \frac{1}{2}\{b(v_1, v_1) - b(v_2, v_2)\} - b(B^{-1}g, v_1 - v_2) \\ &= \frac{1}{2}\{b(v_1 - B^{-1}g, v_1 - B^{-1}g) - b(v_2 - B^{-1}g, v_2 - B^{-1}g)\} \\ &= \frac{1}{2}(|||v_1 - B^{-1}g|||^2 - |||v_2 - B^{-1}g|||^2). \end{aligned}$$

□

Since we can write

$$u_1 - B^{-1}g = (u_\circ - B^{-1}g) + (u_1 - u_\circ)$$

we find

$$(4.55)\quad |||u_\circ - B^{-1}g|||^2 = |||u_1 - B^{-1}g|||^2 - |||u_1 - u_\circ|||^2 + [u_\circ - B^{-1}g, u_1 - u_0]$$

But on the other hand, by definition of u_1 as the projection it follows that

$$[u_\circ + w(B^{-1}g - u_0) - u_1, u_\circ - u_1] \leq 0$$

and hence

$$|||u_\circ - u_1|||^2 \leq w[u_\circ - B^{-1}g, u_\circ - u_1].$$

134

Substituting this in the above identity (4.55) we get

$$\begin{aligned} |||u_\circ - B^{-1}g|||^2 - |||u_1 - B^{-1}g|||^2 &\geq (2-w)[u_\circ - B^{-1}g, u_\circ - u_1] \\ &\geq \frac{2-w}{2w}|||u_\circ - u_1|||^2, \end{aligned}$$

which is precisely the required estimate (4.54).

Step 3. Proof of the Proposition (4.1). It is enough to take

$$H = V_i, C = K_i, b(\cdot,\cdot) = a_{ii}(\cdot,\cdot), P = P_i = Proj\{V_i \to K_i\}$$

and

$$u_i^n = u_\circ, u_i^{n+1} = u_1$$

in Lemma 4.2.

Corollary 4.1. *We have, for each $n \geq 0$,*

$$(4.56)\qquad J(u^n) - J(u^{n+1}) \geq \sum_{i=1}^{N} \frac{2-w_i}{2w_i}|||u_i^{n+1} - u_i^n|||_i^2.$$

Proposition 4.2. *If $0 < w_i < 2$ for all $i = 1, \cdots, N$ then*

$$(4.57)\qquad \begin{cases} J(u^n) \geq J(u^{n+1}) \text{ for all } n \text{ and} \\ u^n - u^{n+1} \to 0 \text{ strongly in } V \text{ as } n \to \infty. \end{cases}$$

Proof. The fact that $J(u^n)$ is a decreasing sequence follows immediately from the Corollary (4.1). Moreover, $J(u^n) \geq J(u)$. for all n, where u is the (unique) absolute minimum of J in K. Hence,

$$J(u^n) - J(u^{n+1}) \to 0 \text{ as } n \to \infty.$$

□

Once again using the Corollary (4.1) and the fact that $2 - w_i > 0$ for each i it follows that

$$|||u_i^{n+1} - u_i^n|||_i \to 0 \text{ as } n \to \infty.$$

135

Since $|||\cdot|||_i$ and $||\cdot||_i$ are equivalent norms on V_i we find that

$$||u_i^n - u_i^{n+1}||_i \to 0 \text{ as } n \to \infty$$

and therefore

$$||u^n - u^{n+1}|| = \left(\sum ||u_i^n - u_i^{n+1}||_i^2\right)^{\frac{1}{2}} \to 0$$

which proves the assertion.

Step 4. Convergence of u^n. We hve the following result.

Theorem 4.4. *If $0 < w_i < 2$ for all $i = 1, \cdots, N$ and if u^n is the sequence defined by the Algotihm (4.2) then*

(4.58) $$u^n \to u \textit{ strongly in } V.$$

Proof. By V-coercivity of the bilinear form $a(\cdot,\cdot)$ we have

$$\begin{aligned}\alpha||u^{n+1} - u||^2 &\leq a(u^{n+1} - u, u^{n+1} - u)\\ &= a(u^{n+1}, u^{n+1} - u) - (f, u^{n+1} - u)\\ &- \{a(u, u^{n+1} - u) - (f, u^{n+1} - u)\}.\end{aligned}$$

□

Here $u^{n+1} - u\epsilon K$ and u is characterized by the variational inequality (4.30) so that

$$a(u, u^{n+1} - u) - (f, u^{n+1} - u) \geq 0$$

and we obtain

$$\alpha||u^{n+1} - u||^2 \leq a(u^{n+1}, u^{n+1} - u) - (f, u^{n+1} - u), \tag{4.59}$$

We can also wirte (4.59) in terms of the operators A_{ij} as

$$\alpha||u^{n+1} - u||^2 \leq \sum_i ((\sum_j A_{ij} u_j^{n+1} - f_i, u_i^{n+1} - u_i))_i. \tag{4.59$'$}$$

136 Consider the minimization problem

$$\begin{cases} \overline{u}_i^{n+1} \epsilon K_i \text{ such that} \\ j_i(\overline{u}_i^{n+1}) \leq j_i(v_i) \text{ for all } v_i \epsilon K_i \text{ where} \\ j_i(v_i) = J(u_1^{n+1}, \cdots, u_{i-1}^{n+1}, v_i, u_{i+1}^n, \cdots, u_N^n). \end{cases} \tag{4.60}$$

We notice that the definition of the functional $v_i \mapsto j_i(v_i)$ coincides with the definition (4.41). The unique solution of the problem (4.60) (which exists by Theorem 3.1 of Chapter 2) is characterized (in view of the Lemma (4.1)) by

$$\overline{u}_i^{n+1} = P_i(A_{ii}^{-1} g_i) = P_i(A_{ii}^{-1}(f_i - \sum_{j<i} A_{ij} u_j^{n+1} - \sum_{j>1} A_{ij} u_j^n)) \tag{4.61}$$

or equivalent by the variational inequality:

$$\begin{cases} (A_{ii}\overline{u}_i^{n+1} - g_i, v_i - \overline{u}_i^{n+1}) \geq 0 \text{ for all } v_i \epsilon K_i \\ \overline{u}_i^{n+1} \epsilon K_i. \end{cases}$$

This is, we have

(4.62)

$$\begin{cases} (A_{ii}\overline{u}_i^{n+1} + \sum\limits_{j<1} A_{ij} u_j^{n+1} + \sum\limits_{j>i} A_{ij} u_j^n - f_i, v_i - \overline{u}_i^{n+1}) \geq 0 \text{ for all } v_i \epsilon K_i \\ \overline{u}_i^{n+1} \epsilon K_i. \end{cases}$$

We can now write the right hand side of (4.59)′ as a sum

(4.59)″ $$I_1 + I_2 + I_2 + I_4$$

where

(4.63)
$$\begin{cases} I_1 = \sum_i((A_{ii}(u_i^{n+1} - \overline{u}_i^{n+1}), u_i^{n+1} - u_i))_i, \\ I_2 = \sum_i((\sum_{j>1} A_{ij}(u_j^{n+1} - u_j^n), u_i^{n+1} - u_i))_i, \\ I_3 = \sum_i((\sum_{j<i} A_{ij}u_j^{n+1} + A_{ii}\overline{u}_i^{n+1} + \sum_{j>1} A_{ij}u_j^n - f_i, u_i^{n+1} - \overline{u}_i^{n+1}))_i, \\ I_4 = \sum_i((\sum_{j<i} A_{ij}u_j^{n+1} + A_{ii}\overline{u}_i^{n+1} + \sum_{j>i} A_{ij}u_j^n - f_i, \overline{u}_i^{n+1} - u_i))_i. \end{cases}$$

First of all, (by 4.62), $I_4 \leq 0$ and hence 137

(4.64) $$\alpha \| u^{n+1} - u \|^2 \leq I_1 + I_2 + I_3.$$

We shall estimate each one of I_1, I_2, I_3 as follows: Since $A_{ij} \epsilon \mathscr{L}$ (V_i, V_j) we set

(4.65) $$M_1 = \max_{1 \leq i,j \leq N} \| A_{ij} \|_{\mathscr{L}(V_i, V_j)}$$

We also know that $\| u_i^n \|, \| \overline{u}_i^n \|$ and hence $\| u^n \|, \| \overline{u}^n \|$ are bounded sequences. For otherwise, $j_i(u_i^n)$ and $j_i(\overline{u}_i^n)$ would tend to $+\infty$ as $n \to \infty$. But we know that they are bounded above by $J(u^\circ)$. So let

(4.66) $$M_2 = \max_{1 \leq i \leq N}(\sup_n \| u_i^n \|, \sup_n \| \overline{u}_i^n \|).$$

The, by Cauchy-Schwarz inequality, we get

$$\begin{aligned} |I_1| &\leq (\sum_i \| u_i^{n+1} - u_i \|_i^2)^{\frac{1}{2}} (\sum_i \| A_{ii} \|^2_{\mathscr{L}(V_i, V_i)} \| u_i^{n+1} - \overline{u}_i^{n+1} \|^2)^{\frac{1}{2}} \\ &= M_1(M_2 + \| u \|) \| u^{n+1} - \overline{u}^{n+1} \| \end{aligned}$$

and similarly we have

$$|I_2| \leq M_1(M_2 + \| u \|) \| u^{n+1} - \overline{u}^{n+1} \|$$

$$|I_3| \leq M_1(M_2 + \|f\|)\|u^{n+1} - \overline{u}^{n+1}\|.$$

These estimates together with (4.64) give

$$\alpha\|u^{n+1} - u\|^2 \leq 3M_1(M_2 + \|u\| + \|f\|)\|u^{n+1} - \overline{u}^{n+1}\| \tag{4.67}$$

and hence it is enough to prove that

$$\|u^{n+1} - \overline{u}^{n+1}\| \to 0 \text{ as } n \to \infty. \tag{4.68}$$

For this purpose, since $w_i > 0$ we can multiply the variational inequality (4.62) by w_i and then we can rewrite it as

(4.62)′

$$((A_{ii}\overline{u}_i^{n+1} - \{A_{ii}\overline{u}_i^{n+1} - w_i(\sum_{j<i} A_{ij}u_j^{n+1} + A_{ii}\overline{u}_i^{n+1} + \sum_{j>i} A_{ij}u_j^n - f_i)\}, v_i - \overline{u}_i^{n+1})) \geq 0.$$

138 Once again using the fact that this variational inequality characterizes the projection $P_i : V_i \to K_i$ we see that

$$\overline{u}_i^{n+1} = P_i\{(1 - w_i)\overline{u}_i^{n+1} - A_{ii}^{-1}(\sum_{j<i} A_{ij}u_j^{n+1} + \sum_{j>i} A_{ij}u_j^n - f_i)\}. \tag{4.69}$$

By (4.38) we also have

$$u_i^{n+1} = P_i\{(1 - w_i)u_i^n - A_{ii}^{-1}(\sum_{j<1} A_{ij}u_j^{n+1} + \sum_{j>1} A_{ij}u_j^n - f_i)\}.$$

Substracting one from the other and using the fact that the projection are contractions we obtain

$$|||\overline{u}_i^{n+1} - u_i^{n+1}|||_i \leq |1 - w_i| |||\overline{u}_i^{n+1} - u_i^n||| \leq |||\overline{u}_i^{n+1} - u_i^n|||_i \tag{4.70}$$

since $0 < w_i < 2$ if and only if $0 < |1 - w_i| < 1$. Now by triangle inequality we have

$$\begin{aligned} |||u_i^n - u_i^{n+1}||| &\geq |||u_i^n - \overline{u}_i^{n+1}|||_i - |||\overline{u}_i^{n+1} - u_i^{n+1}|||_i \\ &\geq (1 - |1 - w_i|)|||\overline{u}_i^{n+1} - u_i^n|||_i \\ &\geq (1 - |1 - w_i|)|||\overline{u}_i^{n+1} - u_i^{n+1}|||_i. \end{aligned}$$

But here, by (4.57), we know that

$$|||u_i^n - u_i^{n+1}|||_i \to 0 \text{ as } n \to \infty.$$

and since $1 - |1 - w_i| > 0$ it follwos that

$$|||\overline{u}_i^{n+1} - u_i^{n+1}|||_i \to 0$$

which is the required assertion.

Remark 4.6. The Theorem (4.4) above on convergence of the relaxation method generalizes a result of Cryer [10] and of a classical result of Varge [50] in finite dimensional case but withour constraints.

Remark 4.7. In this section we have introduced the parameters w_i of relaxation. The algorithm described is said to be of over relaxation type (resp. relaxation, or under relaxation) with projection when $w_i > 1$ (resp. $w_i = 1$ or $0 < w_i < 1$) for all $i = 1, \cdots, N$. 139

4.8 Some Examples - Relaxation Method in Finite Dimensional Spaces

Let $V_i = \mathbb{R}(i = 1, \cdots, N)$ and $V = \prod_{i=1}^N V_i = \mathbb{R}^N$. Let A be a symmetric, positive definite $(n \times n)$ -matrix such that there is a constant $\alpha > 0$ with

$$(Av, v)_{\mathbb{R}^N} \geq \alpha ||v||^2_{\mathbb{R}^N} \text{ for all } v \epsilon \mathbb{R}^N. \tag{4.71}$$

Consider the quadratic functional $J : \mathbb{R}^N \to \mathbb{R}$ of the form

$$J(v) = \frac{1}{2}(Av, v)_{\mathbb{R}^N} - (f, v)_{\mathbb{R}^N}, f \epsilon \mathbb{R}^N. \tag{4.72}$$

We consider the optimization probel for J.

Example 4.4. (Optimization without constraints).

$$\begin{cases} \text{To find } u \epsilon \mathbb{R}^N \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon \mathbb{R}^N \end{cases} \tag{4.73}$$

If we write the matrix A as $A = (a_{ij})$ then

$$J(v) = \frac{1}{2}\sum_{i,j=1}^{N} a_{ij}v_jv_i - \sum_{i=1}^{N} f_iv_i, v = (v_1, \cdots, v_N)\epsilon\mathbb{R}^N. \tag{4.74}$$

We find then that the components of grad j are

$$(gradJ(v))_i = (Av - f)_i = (\sum_{j=1}^{N} a_{ij}v_j - f_i), i = 1, \cdots, N.$$

If $u\epsilon\mathbb{R}^N$ is the (unique) solution of (4.73) then grad $J(u) = 0$. That is,

$$\begin{cases} u = (u_1 \cdots, u_n) \\ \sum_{j=1}^{N} a_{ij}u_j = f_i, i = 1, \cdots, N. \end{cases}$$

140 To describe the algorithm (if we take $w_i = 1$ for all $i = 1, \cdots, N$) to construct u^{k+1} from u^k we find u_i^{k+1} as the solution of the equation

$$\sum_{j<1} a_{ij}u_j^{k+1} + a_{ii}u_i^{k+1} + \sum_{j>1} a_{ij}u_j^k = f_i.$$

Since $a_{ii} > \alpha > 0$ we have

$$u_i^{k+1} = a_{ii}^{-1}[f_i - \sum_{j<i} a_{ij}u_j^{k+1} - \sum_{j>i} a_{ij}u_j^k], \tag{4.75}$$

and thus we obtain the algorithm of the classical Gauss-Seidel methods in finite dimensional spaces.

More generally, introducting a parameter $w(0 < w < 2)$ of relaxation we obtain the following algorithm:

$$\begin{cases} u_i^{k+\frac{1}{2}} = a_{ii}^{-1}[f_i - \sum\limits_{j<i} a_{ij}u_j^{k+1} - \sum\limits_{j>i} a_{ij}u_j^k] \\ u_i^{k+1} = u_i^k - w(u_i^{k+\frac{1}{2}} - u_i^k) \end{cases} \tag{4.76}$$

Example 4.5. (Optimization with constraints in finite dimensional spaces).

Let V_i, V and J be as in Exampl (4.4). We take for the convex set K the following set: Let $I_\circ, I_1$ be a partition of the set $\{1, 2, \cdots, N\}$. That is

$$I_\circ \cap I_1 = \phi \text{ and } \{1, 2, \cdots, N\} = I_\circ \cup I_1.$$

Define

$$(4.77) \qquad \begin{cases} K_i = \{v_i \epsilon \mathbb{R}; v_i \geq 0\} \text{ for all } i \epsilon I_\circ \text{ and} \\ K_i = \mathbb{R} \text{ for all } i \epsilon I_1 \end{cases}$$

and hence

$$(4.78) \qquad K = \{v \epsilon \mathbb{R}^N; v = (v_1, \cdots, v_N) \text{ such that } v_i \geq 0 \text{ for } i \epsilon I_\circ\}$$

As in the previous case, suppose u^k are known, Assume that u_j^{k+1} are found for all $j < i$. We find u_i^{k+1} in there substeps as follows: We 141 define $u_i^{K+1/3}$ as the unique solution of the linear equation obtained by requiring the gradient to vanish at the minimum : more precisely,

$$(4.79) \qquad u_i^{k+1/3} = a_{ii}^{-1}[f_i - \sum_{j<i} a_{ij} u_j^{k+1} - \sum_{j>i} a_{ij} u_j^k].$$

The we set

$$(4.80) \qquad \begin{cases} u_i^{k+2/3} = u_i^k - w(u_i^{k+1/3} - u_i^k) \\ u_i^{k+1} = P_i(u_i^{k+2/3}) \end{cases}$$

where P_i is the projection of V_i onto K_i with respect to the inner product

$$[u_i, v_i] = a_{ii}(u_i, v_i) = a_{ii} u_i v - i.$$

Since $a_{ii} > 0$ and K_i are defined by (4.74) P_i coincides with the projection of V_i onto K_i with respect to the standard inner product on $\mathbb{R}$. Hence we have

$$(4.81) \qquad P_i(u_i^{k+2/3}) = \begin{cases} 0 \text{ if } u_i^{k+2/3} \leq 0 \text{ and } i \epsilon I_\circ \\ u_i^{k+2/3} \text{ in all other cases.} \end{cases}$$

Example 4.6. Let $V = \mathbb{R}^N = \mathbb{R}^1 \times \mathbb{R}^{N-1}, K = K_1 \times K_2$ with $K_1 = \mathbb{R}^1$ and

$$K_2 = \{v \epsilon \mathbb{R}^{N-1}; g(v) \leq 0\},$$

where $g : \mathbb{R}^{N-1} \rightarrow \mathbb{R}$ is a given smooth functional on $\mathbb{R}^{N-1}$. Let $J : V \rightarrow \mathbb{R}$ be a functional of the form (4.74). We can use again an algorithm of the above type. In order to give an algorithm for the construction of the projection P_2 of $V = \mathbb{R}^{N-1}$ onto K_2 we can use any one of the standard methods described in earlier section as, for instance, the method of descent.

4.9 Example in Infinite Dimensional Hilbert Spaces - Optimization with Constraints in Sobolev Spaces

142 We shall only mention briefly a few examples, without going into any details, of optimization problems in the typical saces of infinite dimensions which are of interest to linear partial differential equation, namely the Sobolev spaces $H^1(\Omega), H^1_\circ(\Omega)$ which occur naturally in various variational elliptic problems of second order.

Example 4.7. Let Ω be a bounded open set in $\mathbb{R}^n$ with smoth boundary Γ.

Consider the closed convex subset $K_\circ$ in $H^1(\Omega)$ given by

$$K_\circ = \{v; v \epsilon H^1(\Omega), \gamma_\circ v \geq 0 \text{ a. e. on } \Gamma\}, \tag{4.82}$$

and the quadratic functional $J_\circ : H^1(\Omega) \rightarrow \mathbb{R}$ defined by

$$J_\circ(v) = \frac{1}{2}||v||^2_{H^1(\Omega)} - (f, v)_{L^2(\Omega)}. \tag{4.83}$$

Then we have the optimization problem

$$\begin{cases} \text{To find } u \epsilon K_\circ \text{ such that} \\ J_\circ(u) \leq J_\circ(v) \text{ for all } v \epsilon K_\circ \end{cases} \tag{4.84}$$

Usually we use the method of over relaxation for this problem.

Example 4.8. Let Ω be a simply connected bounded open set in the plane $\mathbb{R}^2$.

Consider

$$K_1 = \{v \epsilon H^1_\circ(\Omega); |\text{grad } v(x)| \leq 1 \text{ a. e. in } \Omega\} \text{ and} \tag{4.85}$$

$$\begin{cases} J(v) = \frac{1}{2} \int_\Omega |grad v|^2 dx - C \int_\Omega v dx \\ \text{where } C \text{ is a constant} > 0. \end{cases} \tag{4.86}$$

The existence and uniqueness of the solution to the minimization problem:

$$\begin{cases} \text{To find } u \epsilon K_1 \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon K_1 \end{cases} \tag{4.87}$$

is classical and its properties have been studied in the paper of Brezis and Stampacchia [4] and some others. It was also shown by Brezis and Sibony [2] that the solution of (4.87) is also the solution of the problem 143

$$\begin{cases} \text{To fing } u \epsilon K_2 \text{ such that} \\ J(u) \leq J(v) \text{ for all } v \epsilon K_2, \text{ where} \\ K_2 = \{v \epsilon H^1_\circ(\Omega); |v(x)| \leq d(x, \Gamma) \text{ a.e. in } \Omega\}, \\ d(x, \Gamma) \text{ being the distance of } x \in \Omega \text{ to the boundary } \Gamma \text{ of } \Omega. \end{cases} \tag{4.88}$$

The method of relaxation described earlier has been used to solve the problem (4.88) numerically by Céa and Glowinski [8, 9]. We also remark that the problem (4.87) is a problem of elasto-palsticity where Ω denotes the cross section of a cylindrical bar whose boundary is Γ and which is made of an elastic material which is perfectly plastic. For details of the numerical analysis of this probel we refer the reader to the paper of Cea and Glowinski quoted above.

Chapter 5

Duality and Its Applications

We shall introduce in this chapter another method to solve the problem of minimization with constraints of functionals $J_\circ$ on a Hilbert space V. This method in turn permits us to construct new algorithm for finding minimizing sequences to the solution of our problem. In this chapter we shall refer to the minimization problem: 144

$$(P) \qquad \text{To find } u \epsilon U, J_\circ(u) = \inf_{v \epsilon U} J_\circ(v)$$

where the constraints are imposed by the set U as the "Primal problem". In the previous chapter U was defined by means of a finite number of functionals $J_1, \cdots, J_k$ on V :

$$U = \{v | v \epsilon V; J_i(v) \leq 0, i = 1, \cdots, k\}.$$

The main idea of the method used in this chapter can be described as follows: We shall describe the condition that an element v belongs to the constraint set U by means of an inequality condition for a suitable functional of two arguments. For this purpose, we introduce a cone A in a suitable topological vector space and a functional φ on $V \times \Lambda$ in such a way that $\varphi(v, \mu) \leq 0$ is equivalent to the fact that v belongs to U. Of course, the choices of Λ and φ are not unique. Then the primal problem (P) will be transformed to a mini-max problem for the functional $\mathscr{L}(v, \mu) = J(v) + \varphi(v, \mu)$ on $V \times \Lambda$. The new functional $\mathscr{L}$ is called a Lagrangain associated to the problem (P).

We shall show that the primal problem is equivalent the minimax problem for the Lagrangain (which is a functional in two arguments $\epsilon V \times \Lambda$). The interest of this method is that under suitable hypothesis, if (u, λ) is a minimax point for the Lagrangian then u will be a solution of the primal problem while λ will be a solution of the so called "dual max-mini problem" which is defined in a natural way by the Lagrangian in this method. Thus under certain hypothesis a minimax point characterizes a solution of the primal problem. 145

Results on the existence of minimax points are known in the literature. We shall show that when V is of finite dimension, under certain assumptions, the existence of a minimax point follows from the classical Hahn-Banach theorem. In the infinite dimensional case we shall illustrate our method which makes use of aresult of *Ky* Fan [29] and Sion [41], [42]. However our arguments are very general and extend easily to the general problem.

1 Preliminaries

We shall begin by recalling the above mentioned two results in the form we shall use in this chapter.

Theorem 1.1. *(Hahn-Banach). Let V be a topological vector space. Suppose M and N are two convex sets in V such that M has atleast one interior point and N does not have any interior point of M (i.e. $IntM \neq \phi, N \cap IntM = \phi$). Then there exist an $F \epsilon V'$, $F \neq 0$ and an $\alpha \epsilon \mathbb{R}$ such that*

$$(1.1) \qquad < F, m >_{V' \times V} = F(m) \leq \alpha \leq F(n), \forall m \epsilon M, \forall n \epsilon N.$$

In order to state the next result it is necessary to introduce the notion of minimax point or sometimes also called saddle point.

Let V and E be two sets and

$$\mathscr{L} : V \times E \to \mathbb{R}$$

be a functional on $V \times E$.

Definition. A point $(u, \lambda)\epsilon V \times E$ is said to be a minimax point or saddle point of $\mathscr{L}$ if

$$(1.2) \qquad \mathscr{L}(u, \mu) \leq \mathscr{L}(u, \lambda) \leq \mathscr{L}(v, \lambda), \qquad \forall (v, \mu)\epsilon V \times E.$$

In other words, $(u, \lambda)\epsilon V \times E$ is a saddle point of $\mathscr{L}$ if the point u is a minimum for the functional $\mathscr{L}(\cdot, \lambda) : V \ni v \mapsto \mathscr{L}(v, \lambda)\epsilon\mathbb{R}$, and if the point λ is a maximum for the functional 146

$$\mathscr{L}(u, \cdot) : E \ni \mu \mapsto \mathscr{L}(u, \mu)\epsilon\mathbb{R}.$$
$$\text{i.e. } \sup_{\mu\epsilon E} \mathscr{L}(u, \mu) = \mathscr{L}(u, \lambda) = \inf_{v\epsilon V} \mathscr{L}(v, \lambda).$$

Theorem 1.2. *(Ky Fan and Sion). Let V and E be two Hausdorff topological vector spaces, U be a convex compact subset of V and Λ be a convex compact subset of E. Suppose*

$$\mathscr{L} : U \times \Lambda \to \mathbb{R}.$$

be a functional such that

(i) For every $v\epsilon U$ the functional $\mathscr{L}(v, \cdot) : \Lambda \ni \mu \mapsto \mathscr{L}(v, \mu)\epsilon\mathbb{R}$ is upper-semi continuous and concave,

(ii) for every $\mu\epsilon\Lambda$ the functional $\mathscr{L}(\cdot, \mu) : U \ni v \mapsto \mathscr{L}(v, \mu)\epsilon\mathbb{R}$ is lower-semi continuous and convex. Then there exists a saddle point $(u, \lambda)\epsilon U \times \Lambda$ for $\mathscr{L}$.

Lagrangian and Lagrange Multipliers

First of all we need a method of describing a set of constraints by means of a functional.

Suppose V is a Hilbert space and U be a given subset of V. In all our applications U will be the set of constraints.

Let E be a vector space. We recall that a cone with vertex at 0 in E is a subset Λ of E which is left invariant by the action of $\mathbb{R}_+$, the set of non-negative real numbers: i.e. If $\lambda\epsilon\Lambda$ and if $\alpha\epsilon\mathbb{R}$ with $\alpha \geq 0$ then $\alpha\lambda$ also belogs to Λ.

We assume that there exists a vector space E, a cone Λ with vertex at 0 in E and a mapping 147

$$\Phi : V \times \Lambda \to \mathbb{R}$$

such that

(i) the mapping $\Lambda \ni \mu \mapsto \Phi(v, \mu) \epsilon \mathbb{R}$ is homogeneous of degree one

$$\text{i.e.} \qquad \Phi(v, \rho\mu) = \rho\Phi(v, \mu), \forall \rho \geq 0,$$

(ii) a point $v \epsilon V$ belongs to U if and only if

$$\Phi(v, \mu) \leq 0, \ \forall \mu \epsilon \Lambda.$$

The choice of the cone Λ and the mapping Φ with the two properties above is not unique in general.

The vector space E often is a topological vector space.

We illustrate the choice of Λ and Φ with the following example.

Example 1.1. Suppose U is a subset of $\mathbb{R}^n$ defined by

$$\begin{aligned} U = \{v | v \epsilon \mathbb{R}^n, \\ g(v) = (g_1(v), \cdots, g_m(v)) \epsilon \mathbb{R}^m \text{ such that } g_i(v) \leq 0 \ \forall i = 1, \cdots, m\}, \end{aligned}$$

i.e. g is a mapping of $\mathbb{R}^n \to \mathbb{R}^m$ and $g_i(v) \leq 0 \ \forall i$. We take

$$\lambda = \{\mu \epsilon \mathbb{R}^m | \mu = (\mu_1, \cdots, \mu_m) \text{ with } \mu_i \geq 0\}$$

Clearly Λ is a (convex) cone with vertex at $0 \epsilon \mathbb{R}^m$. Then we define

$$\Phi : \mathbb{R}^n \times \Lambda \to \mathbb{R}$$

$$\text{by} \qquad \Phi(v, \mu) = (\mu, g(v))_{\mathbb{R}^m} = \sum_{i=1}^{m} \mu_i g_i(v).$$

One can immediatly check that Φ has the properties (i) and (ii) and $U = \{v \epsilon \mathbb{R}^n; \Phi(v, \mu) = (\mu, g(v))_{\mathbb{R}^m} \leq 0\}$.

More generally if U is defined by a mapping $g : \mathbb{R}^n \to H$ where H is any vector space in which we have a notion of positivity then we can take

$$\Lambda = \{\mu | \mu \epsilon H, \mu \geq 0\}$$

148 and

$$\Phi(v, \mu) =< \mu, g(v) >_{H' \times H} .$$

Example 1.2. Let U be a convex closed subset of a Banach space V. We define a function $h : V' \to \mathbb{R}$ by

$$h(\mu) = \sup_{v \epsilon U} < \mu, v >'_{V' \times V}$$

Then clearly $h \geq 0$.

We take for the cone Λ:

$$\Lambda = \{\mu | \mu \epsilon V', h(\mu) < +\infty\}$$

and define $\Phi : V \times \Lambda \to \mathbb{R}$ by

$$\Phi(v, \mu) =< \mu, v > -h(\mu).$$

It is clear from the very definition that if $v \epsilon V$ and $\Phi(v, \mu) \leq 0$ then $v \epsilon U$. In fact,if $v \notin U$ then, since U is a closed convex set in V, by Hahn-Banach theorem there exists an element $\mu \epsilon V'$ such that $\mu(u) = 0 \ \ \forall u \epsilon U$ and $\mu(v) = 1$. Then for this $\mu, h(\mu) = 0$ so that $\mu \in \Lambda$ and $\Phi(v, \mu) =< \mu, v >= 1$ which contradicts the fact that $\Phi(v, \mu) \leq 0$. Hence $v \epsilon U$.

The arguments of Exercise 1.1 can be used to formulate the general problem of non-linear programming considered in Chapter 4: Given $(k + 1)$ functionals $J_\circ, J_1, \cdots, J_k$ on a Hilbert space V to find

$$\begin{aligned} & u \epsilon U = \{v | v \epsilon V; J_i(v) \leq 0 \text{ for } i = 1, \cdots, K\}, \\ & J_\circ(u) = \inf_{v \epsilon U} J_\circ(v). \end{aligned}$$

We note that $v \mapsto (J_1(v), \cdots, J_k(v))$ defines a mapping of V into $\mathbb{R}^k$. We take as E the space $(\mathbb{R}^k)' = \mathbb{R}^k$ and

$$\lambda = \{\mu | \mu \epsilon \mathbb{R}^k, \mu_i \geq 0, i = 1, \cdots, k\}$$

$$\Phi(v, \mu) = \sum_{i=1}^{k} \mu_i J_i(v).$$

149 It is immediately seen that Φ satisfies (i) and (ii), and that an element $v \epsilon V$ belongs to U if and onlu if $\Phi(v, \mu) \leq 0, \forall \mu \epsilon \Lambda$. So our problem can be reformulated equivalenty as follows:

To find $u \epsilon V$ such that $\sup_{\mu \epsilon \Lambda} \Phi(u, \mu) \leq 0$ and

$$J_\circ(u) = \inf_{\{\Phi(v,\mu) \leq 0,\ \forall \mu \epsilon \Lambda\}} J_\circ(v).$$

These considerations are very general and we have the following simple proposition.

Proposition 1.1. *Let V be a normed space and U be a subset of V such that we can find a cone Λ with vertex at 0 (in a suitable vector space) and a function $\Phi : V \times \Lambda \to \mathbb{R}$ satisfying (i) and (ii). Then the following two problems are equivalent: Let $J : V \to \mathbb{R}$ be a given functional*

Primal problem: *To find $u \epsilon U$ such that $J(u) = \inf_{v \epsilon U} J(v)$.*

Minimax problem: *To find a point $(u, \lambda) \epsilon V \times \Lambda$ such that*

$$J(u) + \Phi(u, \mu) = \inf_{v \epsilon V} \sup_{\mu \epsilon \Lambda} (J(v) + \Phi(v, \mu)). \tag{1.3}$$

Proof. First of all we show that

$$\sup_{\mu \epsilon \Lambda} \phi(v, \mu) = \begin{cases} 0 \text{ if } v \epsilon U \\ +\infty \text{ if } v \notin U. \end{cases}$$

□

In fact, if $u \epsilon U$ then by (ii) $\Phi(v, \mu) \leq 0 \quad \forall \mu \epsilon \Lambda$. Since $0 \epsilon \Lambda$ we get by homogeneity (i); $\Phi(v, 0) = 0$ and hence

$$\sup_{\mu \epsilon \Lambda} \Phi(v, \mu) = 0.$$

Suppose now $v \notin U$. Then there exists an element $\mu\epsilon\Lambda$ such that $\Phi(v, \mu) > 0$. But for any $\rho > 0$, $\rho\mu\epsilon\Lambda$ and by homogeneity

$$\Phi(v, \rho\mu) = \rho\Phi(v, \mu) > 0$$

so that $\Phi(v, \rho\mu) \to +\infty$ as $\rho \to +\infty$. This means that 150

$$\sup_{\mu\epsilon\Lambda} \Phi(v, \mu) = +\infty \text{ if } v \notin U.$$

Next we can write

$$\begin{aligned}\sup_{\mu\epsilon\Lambda}(J(v) + \Phi(v, \mu)) &= J(v) + \sup_{\mu\epsilon\Lambda} \Phi(v, \mu)\\ &= \begin{cases} J(v) \text{ if } v\epsilon U \\ +\infty \text{ if } v \notin U \end{cases}\end{aligned}$$

and we therefore find

$$\inf_{v\epsilon V} \sup_{\mu\epsilon\Lambda}(J(v) + \Phi(v, \mu)) = \inf_{v\epsilon U} J(v).$$

This proves the equivalence of the two problems.

Suppose given a functional $J : V \to \mathbb{R}$ on a Hilbert space V and U a subset V for which there exists a cone Λ and a function $\Phi : V \times \Lambda \to \mathbb{R}$ satisfying the conditons (i) and (ii).

Definition 1.1. The Lagrangain associated to the primal problem for J (with constraints defined by the set U) is the functional $\mathscr{L} : V \times \Lambda \to \mathbb{R}$ defined by

$$\mathscr{L}(v, \mu) = J(v) + \Phi(v, \mu). \tag{1.4}$$

$\mu\epsilon\Lambda$ is called a Lagrange multiplier.

The relation between the minimax problem and the saddle point for the Lagrangian is expressed by the following proposition. This proposition is true for any functional $\mathscr{L}$ on $V \times \Lambda$.

Proposition 1.2. *If (u, λ) is a saddle point for $\mathscr{L}$ then we have*

$$\sup_{\mu\epsilon\Lambda} \inf_{v\epsilon V} \mathscr{L}(v, \mu) = \mathscr{L}(u, \lambda) = \inf_{v\epsilon V} \sup_{\mu\epsilon\Lambda} \mathscr{L}(v, \mu). \tag{1.5}$$

Proof. First of all for any functional $\mathscr{L}$ on $V \times \Lambda$ we have the inequality

$$\sup_{\mu\epsilon\Lambda} \inf_{v\epsilon V} \mathscr{L}(v, \mu) \leq \inf_{v\epsilon V} \sup_{\mu\epsilon\Lambda} \mathscr{L}(v, \mu).$$

□

151 In fact, for any point $(v, \mu)\epsilon V \times \Lambda$, we have

$$\inf_{v\epsilon V} \mathscr{L}(v, \mu) \leq \mathscr{L}(v, \mu) \leq \sup_{\mu\epsilon\Lambda} \mathscr{L}(v, \mu).$$

But, there the first term $\inf\limits_{v\epsilon V} \mathscr{L}(v, \mu)$ is only a function of μ while $\sup_{\mu\epsilon\Lambda} \mathscr{L}(v, \mu)$ is a function only of v. Hence we get the required inequality.

Next, if (u, λ) is a saddle point for $\mathscr{L}$ then by definition

$$\begin{aligned} \inf_{v\epsilon V} \sup_{\mu\epsilon\Lambda} \mathscr{L}(v, \mu) \leq \sup_{\mu\epsilon\Lambda} \mathscr{L}(u, \mu) &= \mathscr{L}(u, \lambda) \\ &= \inf_{v\epsilon V} \mathscr{L}(v, \mu) \leq \sup_{\mu\epsilon\Lambda} \inf_{v\epsilon V} \mathscr{L}(v, \mu). \end{aligned}$$

The two inequalities together given the equalities in the assertion of the proposition.

Definition. The problem of finding $(w, \lambda)\epsilon V \times \Lambda$ such that

$$\mathscr{L}(w, \lambda) = \sup_{\mu\epsilon\Lambda} \inf_{v\epsilon V} \mathscr{L}(v, \mu) \tag{1.6}$$

is called the "dual problem" associated to the primal problem.

i.e.

$$\tag*{(1.6)$'$} \begin{cases} (w, \lambda)\epsilon V \times \Lambda \text{ such that} \\ J(w) + \Phi(w, \lambda) = \sup_{\mu\epsilon\Lambda} \inf_{v\epsilon V}(J(v) + \Phi(v, \mu)). \end{cases}$$

Remark. Since the choice of the cone λ and the function $\Phi : V \times \Lambda \to \mathbb{R}$ are not unique there are may ways of defining the dual problem for a given minimization problem.

In the following example we shall determine the dual problem of a linear programming problem.

Suppose given a linear functional $J : \mathbb{R}^n \to \mathbb{R}$ of the form $J(v) = (c, v)_{\mathbb{R}^n}$ where $c \epsilon \mathbb{R}^n$ is a fixed vector, a linear mapping $A : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $b \epsilon \mathbb{R}^m$. Let U be the set in $\mathbb{R}^n$.

$$U = \{v \epsilon \mathbb{R}^n ; Av - b = ((Av - b)_j, \cdots, (Av - b)_m) \epsilon \mathbb{R}^m$$

(1.7) $$\text{such that } (Av - b)_i \leq 0 \text{ for all } i = 1, \cdots, m\}.$$

Consider the linear programming problem: 152

(1.8) $$\text{To find } u \epsilon U \text{ such that } J(u) = \inf_{v \epsilon U} J(v).$$

i.e. To find $u \epsilon \mathbb{R}^n$ such that

(1.8)
$$Au - b \leq 0 \text{ and } (c, u)_{\mathbb{R}^n} \leq (c, v)_{\mathbb{R}^n} \text{ for all } v \epsilon \mathbb{R}^n \text{ satisfying } Av - b \leq 0.$$

We consider another linear programming problem defined as follows.

Let $J^* : \mathbb{R}^m \to \mathbb{R}$ be the functional $J^*(\mu) = (b, \mu)_{\mathbb{R}^m}$ and U^* be the subset of $\mathbb{R}^m$ given by

(1.9)
$$U^* = \{w | w \epsilon \mathbb{R}^m, A^* w + c \epsilon \mathbb{R}^n \text{ such that } (A^* w + c)_j \geq 0 \text{ for all } j = 1, \cdots, n\}.$$

where $A^* : \mathbb{R}^m \to \mathbb{R}^n$ is the adjoint of A.

(1.10) $$\text{To find } \mu \epsilon u^* \text{ such that} J^*(\mu) = \inf_{w \epsilon U^*} J^*(w)$$

i.e. To find $\mu \epsilon \mathbb{R}^m$ such that

(1.10)$'$
$$A^* \mu + c \geq 0 \text{ and } (b, \mu)_{\mathbb{R}^m} \leq (b, w)_{\mathbb{R}^m} \text{ for all } w \epsilon \mathbb{R}^m \text{ such that } A^* w + c \geq 0.$$

Proposition 1.3. *The linear programming problem ((1.10)$'$) is the dual of the linear programming problem ((1.8)).*

Proof. We have $V = \mathbb{R}^n, E = R^m$. Take the cone in $\mathbb{R}^m$ defined by

$\Lambda = \{\mu | \mu \epsilon (\mathbb{R}^m)' = \mathbb{R}^m, \mu = (\mu_1, \cdots, \mu_m)$ with $\mu_i \geq 0$ for all $i = 1, \cdots, m\}$

and the function

$$\Phi(v, \mu) = (Av - b, \mu)_{\mathbb{R}^m}.$$

□

By the very definitions we have $U = \{v \epsilon \mathbb{R}^n | \Phi(v, \mu) \leq 0\}$. The Lagrangian $\mathscr{L}(v, \mu)$ is given by

$$\mathscr{L}(v, \mu) = (c, v)_{\mathbb{R}^\propto} + (Av - b, \mu)_{\mathbb{R}^m}.$$

153 Hence by Definition ((1.6)′) the dual problem is the following: To find $(w, \lambda) \epsilon \mathbb{R}^n \times \Lambda$ such that

$$\begin{aligned}\mathscr{L}(w, \lambda) &= \sup_{\mu \epsilon \Lambda} \inf_{v \epsilon V = \mathbb{R}^n} \mathscr{L}(v, \mu)\\ &= \sup_{\mu \epsilon \Lambda} \inf_{v \epsilon \mathbb{R}^n} ((c, v)_{\mathbb{R}^n} + (Av - b, \mu)_{\mathbb{R}^m}).\end{aligned}$$

We can write

$$\mathscr{L}(v, \mu) = ((A^* \mu + c), v)_{\mathbb{R}^n} - (b, \mu)_{\mathbb{R}^m}$$

and hence

$$\inf_{v \epsilon \mathbb{R}^n} \mathscr{L}(v, \mu) = \inf_{v \epsilon \mathbb{R}^n} ((A^* \mu + c), v)_{\mathbb{R}^n} - (b, \mu)_{\mathbb{R}^m}.$$

If $A^* + \mu + c \neq 0$ then by Cauchy-Schwarz inequality we have

$$-||v||_{\mathbb{R}^n} ||A^* \mu + c||_{\mathbb{R}^n} \leq (A^* \mu + c, v)_{\mathbb{R}^n}$$

and so

$$((A^* \mu + c), v)_{\mathbb{R}^n} \to -\infty \text{ as } ||v|| \to +\infty$$

i.e.

$$\inf_{v \epsilon \mathbb{R}^n} (A^* \mu + c, v)_{\mathbb{R}^n} = -\infty \text{ if } A^* \mu + c \neq 0.$$

But if $A^*\mu + c = 0$ then $\inf_{v\epsilon\mathbb{R}^n}(A^*\mu + c, v)_{\mathbb{R}^n} = 0$. Thus our dual problem becomes

$$\sup_{\mu\epsilon\Lambda} \inf_{v\epsilon\mathbb{R}^n} \mathscr{L}(v,\mu) = \sup_{\mu\epsilon\Lambda} -(b,\mu)_{\mathbb{R}^m} = -\inf_{\mu\epsilon\Lambda}(b,\mu)_{\mathbb{R}^m}.$$

In other words the dual problem is nothing but ((1.10)′)

We conclude this section with the following

Proposition 1.4. *If $(u,\lambda)\epsilon V\times\Lambda$ is a saddle point for the Lagrangian associated to the primal problem then u is a solution of the primal problem and λ is a solution of the dual problem.*

Proof. (u,λ) is a saddle point for the Lagrangian $\mathscr{L}$ is equivalent to saying that

$$(1.11)\quad J(u) + \Phi(u,\mu) \leq J(u) + \Phi(u,\lambda) \leq J(v) + \Phi(v,\lambda), \forall(v,\mu)\epsilon V\times\Lambda.$$

□ 154

Form the first inequality we have

$$(1.12)\qquad \Phi(u,\mu) \leq \Phi(u,\lambda), \forall\mu\epsilon\Lambda.$$

Taking $\mu = 0$ in this inequality we get $\Phi(u,0) \leq \Phi(u,\lambda)$ which means by homogeneity $\Phi(u,\lambda) \geq 0$. Similarly taking $u = 2\lambda$ and using homogeneity we get

$$2\Phi(u,\lambda) = \Phi(u,2\lambda) \leq \Phi(u,\lambda)$$
$$\text{i.e.}\qquad \Phi(u,\lambda) \leq 0.$$

Hence we find that $\Phi(u,\lambda) = 0$. Then it follows from (1.12) that

$$\Phi(u,\mu) \leq 0, \forall\mu\epsilon\Lambda$$

and therefore $u\epsilon U$ by definition of Λ and Φ. Thus we have

$$(1.13)\qquad \begin{cases} u\epsilon U, \lambda\epsilon\Lambda, \Phi(u,\lambda) = 0 \text{ and} \\ J(u) + \Phi(u,\lambda) \leq J(v) + \Phi(v,\lambda)\ \forall v\epsilon V \end{cases}$$

Conversely, it is immediate to see that (1.13) implies (1.11). It is enough to observe that $\Phi(u,\mu) \leq 0 = \Phi(u,\lambda)$ $\forall \mu \epsilon \Lambda$ since $u \epsilon U$ so that we have the inequality

$$J(u) + \Phi(u,\mu) \leq J(u) + \Phi(u,\lambda).$$

Now in (1.13) we take $v \epsilon U$ so that $\Phi(v,\mu) \leq 0, \forall \mu \epsilon \Lambda$ and (1.13) will imply

$$(1.14) \qquad \begin{cases} u \epsilon U, \lambda \epsilon \Lambda, \Phi(u,\lambda) = 0 \text{ and} \\ J(u) \leq J(v) \ \forall v \epsilon U. \end{cases}$$

which proves that u is a solution of the primal problem. We have already seen in Proposition 1.1 the implication that if u is a solution of the problem then

$$\mathscr{L}(u,\lambda) = \inf_{v \epsilon V} \sup_{\mu \epsilon \Lambda} \mathscr{L}(v,\mu).$$

155

On the other hand, if we use proposition 1.2 it follows that Λ is a solution of the dual problem.

2 Duality in Finite Dimensional Spaces Via Hahn - Banach Theorem

In this section we describe a duality method based on the classical Hahn-Banach theorem for convex programming problem in finite dimensional spaces i.e. our primal problem is that of minimizing a convex functional on a finite dimensional vector space subject to constraints defined by convex functionals.

We introduce a condition on the constraints which is of fundamental importance called the Qualifying hypothesis. Under this hypothesis we prove that if the primal problem has a solution then there exists a saddle point for the Lagrangian associated to it. We shall also give sufficient conditions in order that the Qualifying hypothesis on the constraints are satisfied.

Let $J_i : \mathbb{R}^n \to \mathbb{R}(i = 0, 1, \cdots, k)$ be $(k + 1)$ convex functionals on $\mathbb{R}^n$ and K be the set defined by

$$K = \{v | v \epsilon \mathbb{R}^n; J_i(v) \leq 0 \text{ for } i = 1, \cdots, k\}.$$

Our primal problem then is

Problem 2.1. To find $u \epsilon K$ such that $J_\circ(u) = \inf_{v \epsilon K} J_\circ(v)$.

It is clear that K is a convex set.

Let

$$j = \inf_{v \epsilon K} J_\circ(v) \tag{2.1}$$

We introduce the Lagrangian associated to the problem (2.1) as described in the previous section. More precisely, let

$$\Lambda = \{\mu | \mu = (\mu_1, \cdots, \mu_k) \epsilon \mathbb{R}^k \text{ such that } \mu_i \geq 0\}$$

which is clearly a cone with vertex as 0 in $\mathbb{R}^k$ and let 156

$$\Phi : \mathbb{R}^n \times \Lambda \to \mathbb{R}$$

be defined by

$$\Phi(v, \mu) = \sum_{i=1}^{k} \mu_i J_i(v).$$

Then the Lagrangian associated to the problem (2.1) is

$$\mathscr{L}(v, \mu) = J_\circ(v) + \sum_{i=1}^{k} \mu_i J_i(v).$$

Suppose that the problem (2.1) has a solution. Then we wish to find conditions on the constraints J_i in order that there exists a saddle point for $\mathscr{L}$. For this purpose we proceed as follows:

Suppose S and T are two subsets of $\mathbb{R}^{k+1}$ defines in the following way:

S is the set of all points

$$\begin{cases} (J_\circ(v) - j + s_\circ, J_1(v) + s_1, \cdots, J_k(v) + s_k) \epsilon \mathbb{R}^{k+1}, \\ \text{ where } v \epsilon \mathbb{R}^n \text{ and} \\ s = (s_\circ, s_1, \cdots, s_k) \epsilon \mathbb{R}^{k+1} \text{ such that } s_i \geq 0 \ \forall i. \end{cases}$$

T is the set of all points

$$(-t_\circ, -t_i, \cdots, -t_k) \epsilon \mathbb{R}^{k+1} \text{ where } t_i \geq 0 \ \forall i.$$

It is obvious that T is convex. In fact T is nothing but the negative cone in $\mathbb{R}^{k+1}$. On the other hand, since $J_\circ, J_1, \cdots, J_k$ are convex and $s_i \geq 0 \ \forall i$ it follows that S is also convex. It is also clear that $\text{Int}\, T \neq \phi$. In fact any point $(-t_\circ, -t_1, \cdots, -t_k) \epsilon \mathbb{R}^{k+1}$ with $t_i > 0 \ \forall i$ is an interior point.

Next we claim that $S \cap (\text{Int}\, T) = \phi$. In fact, if $S \cap (\text{Int}\, T) \neq \phi$ then there exist

$$\text{some } t \epsilon \mathbb{R}^{k+1} \text{ with } t = (t_\circ, t_1, \cdots, t_k), t_i > 0 \ \forall i,$$

157

$$\text{some } v \epsilon \mathbb{R}^n, \text{ and an } s \epsilon \mathbb{R}^{k+1} \text{ with } s = (s_\circ, s_1, \cdots, s_k), s_i > 0 \ \forall i$$

such that

$$J_\circ(v) - j + s_\circ = -t_\circ, J_1(v) + s_1 = -t_1, \cdots, J_k(v) + s_k = -t_k$$

Now we have form this

$$J_i(v) = -t_i - s_i < 0 \text{ since } s_i \geq 0 \text{ for any} i = 1, \cdots, k$$

This means that $v \epsilon K$. On the other hand,

$$J_\circ(v) = -t_\circ - s_\circ + j < j = \inf_{w \epsilon K} J_\circ(w)$$

which is impossible since $v \epsilon K$.

We can now apply Hahn-Banach theorem to the sets S and T in the form we have recalled in Section 1. There exist an $F\epsilon(\mathbb{R}^{k+1})' = (\mathbb{R}^{k+1})$ and an $\alpha\epsilon\mathbb{R}$ such that $F \neq 0, F(x) \geq \alpha \geq F(y)$ where $x\epsilon S$ and $y\epsilon T$. More precisely we can write this as follows:

$$\exists F = (\alpha_\circ, \alpha_1, \cdots, \alpha_k)\epsilon\mathbb{R}^{k+1} \text{ such that } \sum_{i=0}^{k} |\alpha_i| > 0 \text{ and } \exists\alpha\epsilon\mathbb{R}$$

such that

$$(2.2) \quad \begin{cases} \alpha_\circ(J_\circ(v) - j + s_\circ) + \sum_{i=1}^{k} \alpha_i(J_i(v) + s_i) \geq \alpha \geq -\sum_{i=0}^{k} \alpha_i t_i, \\ \forall v\epsilon V, s = (s_\circ, s_1, \cdots, s_k) \text{ with } s_i \geq 0\ \forall i \\ \qquad\qquad \text{and } t = (t_\circ, t_1, \cdots, t_k) \text{ with } t_i \geq 0\ \forall i \end{cases}$$

We next show from (2.2) that we have

$$(2.3) \qquad \alpha = 0, \alpha_i \geq 0\ \forall i \text{ and } \sum_{i=0}^{k} \alpha_i > 0.$$

In fact, if we take $t_1 = \cdots = t_k = 0$ then we get, from the second inequality in (2.2).

$$\alpha \geq -\alpha_\circ t_\circ = (-\alpha_\circ)t_\circ\ \forall t_\circ \geq 0.$$

158

If $\alpha_\circ < 0$ then $(-\alpha_\circ)t_\circ \to +\infty$ as $t_\circ \to +\infty$ and therefore we necessarily have $\alpha_\circ \geq 0$. Similarly we can show that $\alpha_i \geq 0\ \forall i = 0, 1, \cdots, k$. Then

$$\sum_{i=0}^{k} |\alpha_i| = \sum_{i=0}^{k} \alpha_i > 0 \text{ since } F \neq 0.$$

If we take $t_\circ = t_1 = \cdots = t_k = 0$ we also find, from the second inequalities in (2.2) that $\alpha \geq 0$.

We have therefore only to show that $\alpha \leq 0$. For this, taking $s_\circ = \cdots = s_k = 0$ in the first inequality of (2.2) we get

$$(2.4) \qquad \alpha_\circ(J_\circ(v) - j) + \sum_{i=1}^{k} \alpha_i J_i(v) \geq \alpha.$$

Suppose v^m is a minimizing sequence for the problem (2.1)

$$\text{i.e} \qquad v^m \epsilon K \text{ and } J_\circ(v^m) \to j = \inf_{v\epsilon K} J_\circ(v).$$

This means that $J_i(v^m) \leq 0$ for $i = 1, \cdots, k$ and $J_\circ(v^m) \to j$. Hence (2.4) will imply, since $\alpha_i \geq 0$

$$\alpha_\circ(J_\circ(v^m) - j) \geq \alpha_\circ(J_\circ(v^m) - j) + \sum_{i=1}^{k} \alpha_i J_i(v) \geq \alpha.$$

Now taking limits as $m \to +\infty$ it follows that $\alpha \leq 0$. Thus we have

(2.5) $$\begin{cases} \alpha_i \geq 0, \text{ for } i = 0, 1, \cdots, k \text{ and } \sum_{i=0}^{k} \alpha_i > 0, \\ \alpha_\circ(J_\circ(v) - j) + \sum_{i=1}^{k} \alpha_i J_i(v) \geq 0, \forall v\epsilon\mathbb{R}^n \end{cases}$$

We now make the fundamental hypothesis that

(2.6) $$\alpha_\circ > 0.$$

Under the hypothesis (2.6) if we write $\lambda_i = \alpha_i/\alpha_\circ$ then (2.5) can be written in the form

(2.7) $$\begin{cases} \lambda_i \geq 0 \text{ for } i = 1, \cdots, k \text{ and} \\ j \leq j_\circ(v) + \sum_{i=1}^{k} \lambda_i J_i(v). \forall v\epsilon\mathbb{R}^n \end{cases}$$

159 i.e. $\lambda\epsilon\Lambda$ and $\mathscr{L}(v, \lambda) \geq j \ \forall v\epsilon\mathbb{R}^n$.

The condition (2.6) is well known in the literature on optimization. We introduce the following definition.

Definition 2.1. Any hypothesis on the constraints J_i which implies (2.6) is called a Qualifying hypothesis.

We shall see a little later some examples of Qualifying hypothesis. (See [26], [27], [28]).

We have thus proved the

Theorem 2.1. *If all the functionals $J_i (i = 0, 1, \cdots, k)$ are convex and if the Qualifying hypothesis is satisfied then there exists a $\lambda \epsilon \Lambda$ such that $\mathscr{L}(v, \lambda) \geq j \ \forall v \epsilon \mathbb{R}^n$.*

i.e. there exists a$\lambda = (\lambda_1, \cdots, \lambda_k) \epsilon \mathbb{R}^k$ with $\lambda_i \geq 0 \ \forall i$ such that

$$J_\circ(v) + \sum_{i=1}^{k} \lambda_i J_i(v) \geq j, \forall v \epsilon \mathbb{R}^n.$$

We can also deduce from (2.7) the following result.

Theorem 2.2. *Suppose all the functionals $J_\circ, J_1, \cdots, J_k$ are convex and the Qualifying hypothesis holds. If the problem (2.1) has a solution, i.e.*

(2.8) *there exists a $u \epsilon K$ such that $J_\circ(u) = j = \inf_{v \epsilon K} J_\circ(v)$*

then the lagrangian $\mathscr{L}$ has a saddle point.

Proof. We can write (2.7) as

$$\lambda_i \geq 0 \text{ for } i = 1, \cdots, k \text{ and}$$

(2.9) $$J_\circ(u) \leq J_\circ(v) + \sum_{i=1}^{k} \lambda_i J_i(v) = \mathscr{L}(v, \lambda), \forall v \epsilon \mathbb{R}^n.$$

Choosing $v = u$ in (2.9) we find that

$$\sum_{i=1}^{k} \lambda_i J_i(u) \geq 0.$$

But here $\lambda_i \geq 0$ and $J_i(u) \leq 0$ since $u \epsilon K$ so that $\lambda_i J_i(u) \leq 0$ for all $i = 1, \cdots, k$ and hence $\sum_{i=1}^{k} \lambda_i J_i(u) \leq 0$. Thus we necessarily have 160

$$\sum_{i=1}^{k} \lambda_i J_i(u) = 0$$

and, further more, it follows immediately from this that

$$\lambda_i J_i(u) = 0 \text{ for } i = 1, \cdots, k.$$

Thus we can rewrite (2.9) once again as :

$$(2.10)\quad \begin{cases} \lambda_i \geq 0, i = 1, \cdots, k. \\ u \epsilon K, \sum_{i=1}^{k} \lambda_i J_i(u) = 0 \text{ and} \\ \mathscr{L}(u, \lambda) = J_\circ(u) + \sum_{i=1}^{k} \lambda_i J_i(u) \leq J_\circ(v) + \sum_{i=1}^{k} \lambda_i J_i(v) \\ \qquad = \mathscr{L}(v, \lambda)\ \forall v \epsilon \mathbb{R}^n. \end{cases}$$

But, since $u \epsilon K$, $J_i(u) \leq 0$ and we also have

(2.11)

$$\begin{cases} \mathscr{L}(u, \mu) = J_\circ(u) + \sum_{i=1}^{k} \mu_i J_i(u0 \leq J_\circ(u) = J_\circ(u) + \sum_{i=1}^{k} \lambda_i J_i(u) \\ \qquad = \mathscr{L}(u, \lambda) \\ \forall \mu \epsilon \mathbb{R}^k \text{ with } \mu = (\mu_1, \cdots, \mu_k), \mu_i \geq 0. \end{cases}$$

(2.10) asnd (2.11) together means that

$$\mathscr{L}(u, \mu) \leq \mathscr{L}(u, \lambda) \leq \mathscr{L}(v, \lambda), \forall v \epsilon \mathbb{R}^n \text{ and } \forall \mu \epsilon \Lambda.$$

This proves the theorem. □

Some examples of Qualifying hypothesis. We recall that if all the functionals $J_\circ, J_1, \cdots, J_k$ are convex then we always have (2.5) $\forall v \epsilon \mathbb{R}^n$. If suppose $\alpha_\circ = 0$ in (2.5) then we get

$$(2.12)\quad \begin{cases} \alpha_i \geq 0 \text{ for } i = 1, \cdots, k, \sum\limits_{i=1}^{k} \alpha_i > 0 \text{ and} \\ \sum\limits_{i=1}^{k} \alpha_i J_i(v) \geq 0, \forall v \epsilon \mathbb{R}^n \end{cases}$$

In all the examples we give below we state the Qualifying hypothesis in the following form. The given hypothesis together with the fact **161** that $\alpha_\circ = 0$ will imply that it is impossible that (2.5) holds. i.e. The hypothesis will imply that (2.12) cannot hold. Hence if (2.5) should hold we necessarily have $\alpha_\circ > 0$, i.e. (2.6) holds.

Qualifying hypothesis (1). There exists a vector $Z \epsilon \mathbb{R}^n$ such that $J_i(Z) < 0$ for $i = 1, \cdots, k$.

This condition is due to Slater (See for instance [6]).

Suppose the Qualifying hypothesis (1) is satisfied. Let $c\epsilon\mathbb{R}$ be such that $J_i(Z) \leq c < 0$ for all $i = 1, \cdots, k$. Obviously such a constant c exists since we can take $c = \max\limits_{1\leq i\leq k} J_i(Z)$. Now if $\alpha_i \geq 0 (i = 1, \cdots, k)$ are such that $\sum\limits_{i=1}^{k} \alpha_i > 0$ then

$$\sum_{i=1}^{k} \alpha_i J_i(Z) \leq c \sum_{i=1}^{k} \alpha_i < 0.$$

This means that (2.12) does not hold for the vector $Z\epsilon\mathbb{R}^n$. Hence $\alpha_\circ > 0$ necessarily so that (2.5) holds $\forall v\epsilon\mathbb{R}^n$ and in particular for Z.

Qualifying hypothesis (2). There do not exist real numbers

$$(2.13) \quad \begin{cases} \alpha_i (i = 1, \cdots, k) \text{ with } \alpha_i \geq 0 \text{ and } \sum\limits_{i=1}^{k} \alpha_i > 0 \text{ such that} \\ \sum\limits_{i=1}^{k} \alpha_i J_i(v) = 0, \forall v\epsilon K. \end{cases}$$

Suppose this hypothesis holds and $\alpha_\circ = 0$. Then we have (2.12) for all $v\epsilon\mathbb{R}^n$.

In particulas, we have

$$\sum_{i=1}^{k} \alpha_i J_i(v) \geq 0, \forall v\epsilon K.$$

But $v\epsilon K$ and $\alpha_i \geq 0$ imply that $\alpha_i J_i(v) \leq 0$ for $i = 1, \cdots, k$ and so $\sum\limits_{i=1}^{k} \alpha_i J_i(v) \leq 0$. The two inequalities together imply that $\exists \alpha_i \geq 0$ with $\sum\limits_{i=1}^{k} \alpha_i > 0$ such that $\sum\limits_{i=1}^{k} \alpha_i J_i(v) = 0$, contrary to the hypothesis. Hence $\alpha_\circ > 0$.

Qualifying hypothesis (3). Suppose $J_i (i = 1, \cdots, k)$ further have gradi- **162**

ents $G_i(i = 1, \cdots, k)$.

$$(2.14) \quad \begin{cases} \text{There do not exist real numbers } \alpha_i \text{ with} \\ \alpha_i \geq 0, i = 1, \cdots, k, \sum\limits_{i=1}^{k} \alpha_i > 0 \text{ such that} \\ \sum\limits_{i=1}^{k} \alpha_i G_i(v) = 0, \forall v \epsilon K. \end{cases}$$

The condition (2.14) seems to be due to to Kuhn and Tucker [28]

It is enough to show that Qualifying hypothesis (3) implies Qualifying hypothesis (2). Suppose there exist $\alpha_i \geq 0, i = 1, \cdots, k$, with $\sum\limits_{i=1}^{k} \alpha_i J_i(v) = 0 \ \forall v \epsilon K$. Then taking derivatives it will imply the existence of $\alpha_i \geq 0 (i = 1, \cdots, k)$ with $\sum\limits_{i=1}^{k} \alpha_i > 0$ such that $\sum\limits_{i=1}^{k} \alpha_i G_i(v) = 0 \ \forall v \epsilon K$. This contradicts the given hypothesis. Hence $\alpha_\circ > 0$.

Finally we remark that the existence of a saddle point can also be proved using the minimax theorem of Ky Fan and Sion. We refer for this to the book of Cea [6].

3 Duality in Infinite Dimensional Spaces Via Ky Fan - Sion Theorem

This section will be concerned with the duality theory for the minimisation problem with constraints for functionals on infinite dimensional Hilbert spaces. We confine ourselves to illustrate the method in the special example of a quadratic form (see the model problem considered in Chapter 1, Section 7) in which case we have proved the existence of a unique solution for our probelm (see Section 2 of Chapter 2). As we have already mentioned this example includes a large class of variational inequalities associated to second order elliptic differential operators and conversely. Our main tool in this will be the theorem of *Ky* Fan and Sion. However we remark that our method is very general and is applicable but for some minor details to the case of general convex programming problems in infinite dimesional spaces.

3.1 Duality in the Case of a Quadratic Form

163

We take for the Hilbert space V the Sobolev space $H^1(\Omega)$ where Ω is a bounded open set with smooth boundary Γ in $\mathbb{R}^n$. Let $a(\cdot,\cdot)$ be a continuous quadratic form on V (i.e. it is a symmetric bilinear bicontinuous mapping: $V \times V \to \mathbb{R}$) and $L(\cdot)$ be a continuous linear functional on V (i.e. $L\epsilon V'$). We assume that $a(\cdot,\cdot)$ is $H^1(\Omega)$ - coercive. Let $J : H^1(\Omega) \to \mathbb{R}$ be the (strictly) convex continuous functional on $H^1(\Omega)$ defined by

$$J(v) = \frac{1}{2}a(v, v) - L(v). \tag{3.1}$$

We denote by $|||\cdot|||$ the norm $||\cdot||_{H^1(\Omega)}$ and by $||\cdot||$ the norm $||\cdot||_{L^2(\Omega)}$. Let us consider the set

$$K\{v|v\epsilon H^1(\Omega), ||v|| \leq 1\}. \tag{3.2}$$

We check immediately that K is a closed convex set in $H^1(\Omega)$. We are interested in the following minimisation problem :

Problem 3.1. To find $u\epsilon K$ such that $J(u) \leq J(v), \forall v\epsilon K$.

Since J is $H^1(\Omega)$ -coercive (hence strictly convex) and since J has a gradient and a hessian everywhere in V we know from Theorem 2. 2.1 that the problem 3.1 has unique solution.

In order to illustrate our method we shall consider a simple case and take

$$\Lambda = \{\mu|\mu\epsilon\mathbb{R}, \mu \geq 0\} \tag{3.3}$$

and

$$\Phi(v, \mu) = \frac{1}{2}\mu(||v||^2 - 1)\ \forall v\epsilon V = H^1(\Omega) \text{ and } \mu\epsilon\Lambda. \tag{3.4}$$

Thus K is nothing but the set $\{v|v\epsilon V, \Phi(v, \mu) \leq 0\}$. We define the associated Lagrangian by

$$\mathscr{L}(v, \mu) = J(v) + \Phi(v, \mu)$$

i.e. 164

$$\mathscr{L}(v,\mu) = \frac{1}{2}a(v,v) - L(v) + \frac{1}{2}\mu(||v|| - 1). \tag{3.5}$$

We observe that

(i) the mapping $\mu \mapsto \mathscr{L}(v,\mu)$ is continuous linear and hence, in particular, it is concave and upper-semi-continuous and

(ii) the mapping $v \mapsto \mathscr{L}(v,\mu)$ is continuous and convex and hence in particualr, it is convex and lower semi-continuous.

We are now in a position to prove the first result of this section using the theorem of Ky Fan and Sion. This can be stated as follows:

Theorem 3.1. *Suppose the functional J on $V = H^1(\Omega)$ is given by (3.1) and the closed convex set K of V is given by (3.2). Then the Lagrangian (3.5) associated to the primal problem 3.1 has a saddle point. Moreover, if (u,λ) is a saddle point of $\mathscr{L}$ then u is a solution of the generalized Neumann problem*

$$\begin{aligned} +Au + u\lambda u &= f \textit{ in } \Omega \\ \partial/\partial n_A u &= 0 \textit{ on } \Gamma \end{aligned} \tag{3.6}$$

We note that here u and λ are subjected to the constraints

$$\lambda \geq 0, ||u|| \leq 1 \textit{ but } \lambda(||u||^2 - 1) = 0. \tag{3.7}$$

Here the formal (differential) operator A is defined in the following manner. For any fixed $v\epsilon V = H^1(\Omega)$ the linear mapping $\varphi \mapsto a(v,\varphi)$ is a continuous linear functional Av i.e. $Av\epsilon V'$. Moreover $v \mapsto Av$ belongs to $\mathscr{L}(V, V')$ and we have

$$(Av,\varphi)_V = a(v,\varphi), \forall\varphi\epsilon H^1(\Omega) = V.$$

Similarly $f\epsilon L^2(\Omega)$ is defined by $L(\varphi) = (f,\varphi)_{L^2(\Omega)}, \forall\varphi\epsilon V$. Further $\partial u/\partial n_A$ is the co-normal derivative of u associated to A and is defined by the Green's formula:

$$a(u,\varphi) = (Au,\varphi)_V + \int_\Gamma \partial u \partial n_A \varphi d\sigma, \forall\varphi\epsilon V,$$

as in Section 4 of Chapter 2.

165 In particular, if we take $a(v, v) = |||v|||^2$, then $A = -\triangle$ and the problem is nothing but the classical Neumann problem

$$(3.6)' \qquad \begin{cases} -\triangle u + u + \lambda u = f \text{ in } \Omega, \\ \partial u/\partial \underline{n} = 0 \text{ on } \Gamma \end{cases}$$

Of course, we again have (3.7).

Proof of Theorem 3.1. Let $\ell > 0$ be any real number. We consider the subsets K_ℓ and Λ_ℓ of $H(\Omega)$ and Λ respectively defined by

$$K_\ell = \{v | v \epsilon H^1(\Omega), |||v||| \leq \ell\}$$
$$\Lambda_\ell = \{\mu | \mu \epsilon \mathbb{R}, 0 \leq \mu \leq \ell\}$$

It is immediately verified that K_ℓ and Λ_ℓ are convex sets, and that Λ_ℓ is a compact set in $\mathbb{R}$. Since K_ℓ is a closed bounded set in the Hilbert space $H^1(\Omega)$, K_ℓ is weakly compact. We consider $H^1(\Omega)$ with its weak topoligy.

Now $H^1(\Omega) = V$ with the weak topology is a Hausdorff topological vector space. All the hypothesis of the theorem of Ky Fan and Sion are satisfied by K_ℓ, Λ_ℓ and $\mathscr{L}$ in view of (i) and (ii). Hence $\mathscr{L} : K_\ell \times \Lambda_\ell \to \mathbb{R}$ has a saddle point (u_ℓ, λ_ℓ). i.e.

$$(3.8) \qquad \begin{cases} \text{There exist } (u_\ell, \lambda_\ell) \epsilon K_\ell \times \Lambda_\ell \text{ such that} \\ J(u_\ell) + \frac{1}{2}\mu(|||u_\ell|||^2 - 1) \leq J(u_\ell) + \frac{1}{2}\lambda_\ell(|||u_\ell|||^2 - 1) \\ \qquad \leq J(v) + \frac{1}{2}\lambda_\ell(|||v|||^2 - 1). \\ \forall (v, \mu) \epsilon K_\ell \times \Lambda_\ell. \end{cases}$$

We shall show that if we choose $\ell > 0$ sufficiently large then such a saddle point can be obtained independent of ℓ and this would prove the first part of the assertion. For this we shall first prove that $||u_\ell||$ and λ_ℓ are bounded by constants independent of ℓ.

If we take $\mu = 0 \epsilon \Lambda_\ell$ in (3.8) we get

$$(3.9) \qquad J(u_\ell) \leq J(v) + \frac{1}{2}\lambda_\ell(||v||^2 - 1), \forall v \in K_\ell$$

and, in particular, we also get 166

(3.10) $$J(u_\ell) \leq J(v), \forall v \in K \cap K_\ell.$$

Taking $v = 0 \in K \cap K_\ell$ in (3.10) we see that $J(u_\ell) \leq J(0)(= 0)$. On the other hand, since $a(u_\ell, u_\ell) \geq 0$ and since $u_\ell \in K_\ell$

$$L(u_\ell) \leq ||L||_{V'} |||u_\ell||| \leq \ell ||L||_{V'}$$

we see that

$$J(u_\ell) = \frac{1}{2} a(u_\ell, u_\ell) - L(u_\ell) \geq -\ell ||L||_{V'}$$

which proves that $J(u_\ell)$ is also bounded below. Thus we have

(3.11) $$\ell ||L||_{V'} \leq J(u_\ell) \leq J(0).$$

Now by coercivity of $a(\cdot, \cdot)$ and (3.11) we find

$$\alpha |||u_\ell|||^2 \leq a(u_\ell, u_\ell) = 2(J(u_\ell) + L(u_\ell)) \leq 2(J(0) + ||L||_{V'} |||u_\ell|||).$$

with a constant $\alpha > 0$ (independent of ℓ). Here we use the trivial inequality

$$||L||_{V'} |||u_\ell||| \leq \epsilon |||u_\ell|||^2 + 1/\epsilon ||L||^2_{V'}. \text{ for any } \epsilon > 0.$$

with $\epsilon = \alpha/4 > 0$ and we obtain

$$|||u_\ell|||^2 \leq 4/\alpha (J(0) + 4/\alpha ||L||^2_{V'})$$

This proves that there exists a constant $c_1 > 0$ such that

(3.12) $$|||u_\ell||| \leq c_1, \forall \ell.$$

To prove that λ_ℓ is also bounded by a constant $c_2 > 0$ independent of ℓ, we observe that since J satisfies all the assumptions of Theorem 167 2.3.1 of Chapter 2, (Section 3) there exists a unique global minimum in $V = H^1(\Omega)$ i.e.

(3.13) There exists unique a $\widetilde{u} \epsilon H^1(\Omega)$ such that $J(\widetilde{u}) \leq J(v), \forall v \epsilon V$.

Hence we have

$$J(\widetilde{u}) + \lambda_\ell/2 \leq J(u_\ell) + \lambda_\ell/2.$$

But, if we take $v = 0\epsilon K_\ell$ in the second inequality in (3.9) we get

$$J(u_\ell) + \lambda_\ell/2 \leq J(0).$$

These two inequalities together imply that

$$\lambda_\ell/2 \leq J(0) - J(\widetilde{u}).$$

i.e.

(3.14) $$0 \leq \lambda_\ell \leq 2(J(0) - J(\widetilde{u})) = c_2$$

which proves that λ_ℓ is also bounded.

(3.15) We choose $\ell > \max(c_1, 2c_2) > 0$.

Next we show that (3.8) holds for any $\mu\epsilon\Lambda$. For this, we use the first inequality in (3.8) in the form

$$\mu(||u_\ell||^2 - 1) \leq \lambda_\ell(||u_\ell||^2|| - 1).$$

This implies (i) taking $\mu = 0$, $\lambda_\ell(||u_\ell||^2 - 1) \geq 0$ and
(ii) taking $\mu = 2\lambda_\ell \leq 2c_2 < \ell, \lambda_\ell, \lambda_\ell(||u_\ell||^2 - 1) \leq 0$. Thus we have

$$\lambda_\ell(||u_\ell||^2 - 1) = 0 \text{ and } \mu(||u_\ell||^2 - 1) \leq 0, \forall\mu\epsilon\Lambda_\ell.$$

In particular, $\mu = \ell\epsilon\Lambda_\ell$ and so $\ell(||u_\ell||^2 - 1) \leq 0$. Thus we have

$$\lambda_\ell(||u_\ell||^2 - 1) = 0 \text{ and } \mu(||u_\ell||^2 - 1) \leq 0, \forall\mu\epsilon\Lambda_\ell$$

In particular, $\mu = \ell\epsilon\Lambda_\ell$ and so $\ell(||u_\ell||^2 - 1) \leq 0$ which means that $||u_\ell||^2 - 1 \leq 0$.

Hence we also have

$$\mu(||u_\ell||^2 - 1) \leq 0 \text{ for any } \mu \geq 0.$$

and therefore 168

$$(3.16)\qquad \mathscr{L}(u_\ell,\mu) \leq \mathscr{L}(u_\ell,\lambda_\ell) \leq \mathscr{L}(v,\lambda_\ell), \forall\mu \geq 0 \text{ and } v\epsilon K_\ell$$

where $\ell \geq \max(c_1, 2c_2)$.

We have now only to show that we have (3.16) for any $v\epsilon H^1(\Omega) = V$. For this we note that $|||u_\ell||| \leq c_1 < \ell$ and hence we can find an $r > 0$ such that the ball

$$B(u_\ell, r) = \{v|v\epsilon H^1(\Omega); |||v - u_\ell||| < r\}$$

is contained in the ball $B(0,\ell) = \{v|v\epsilon H^1(\Omega), |||v||| < \ell\}$. In fact, it is enough to take $0 < r < (\ell - c_1)/2$. Now the functional $\mathscr{L}(\cdot,\lambda_\ell)$: $v \mapsto \mathscr{L}(v,\lambda_\ell) = J(v) + \lambda_\ell/2(||v||^2 - 1)$ has a local minimum in $B(u_\ell, r)$. But since this functional is convex such a minimum is also a global minimum. This means that

$$\inf_{v\epsilon R(u_\ell r)} \mathscr{L}(v,\lambda_\ell) = \inf_{v\epsilon V} \mathscr{L}(v,\lambda_\ell).$$

On the other hand, since $B(u_\ell, r) \subset K_\ell$ we see from (3.16) that

$$\mathscr{L}(u_\ell,\mu) \leq \mathscr{L}(u_\ell,\lambda_\ell) \leq \inf_{v\epsilon K_\ell} \mathscr{L}(v,\lambda_\ell) \leq \inf_{v\epsilon B(u_\ell,r)} \mathscr{L}(v,\lambda_\ell) = \inf_{v\epsilon V} \mathscr{L}(v,\lambda_\ell).$$

In other words, we have

$$\mathscr{L}(u_\ell,\mu) \leq \mathscr{L}(u_\ell,\lambda_\ell) \leq \mathscr{L}(v,\lambda_\ell), \forall v\epsilon V \text{ and } \forall\mu \geq 0$$

which means that $\mathscr{L}$ has a saddle point.

Finally we prove that $(u,\lambda) = (u_\ell,\lambda_\ell)(\ell > \max(c_1, 2c_2))$ satisfies (3.6). First of all the functional $v \mapsto \mathscr{L}(v,\lambda)$ is G-differentiable and has a gradient everywhere in V. In fact, we have

$$(3.17)\qquad ((grad\mathscr{L})(v),\varphi)_V = a(v,\varphi) - L(\varphi) + \lambda(v,\varphi)_V.$$

169 We know by Theorem 2.1.3 (Chapter 2, Section 1) that at the point u where $v \mapsto \mathscr{L}(v,\lambda)$ has a minimum we should have

$$(3.18)\qquad ((grad\mathscr{L}(\cdot,\lambda))u,\varphi)_V = 0.$$

Now, if we use (3.17), (3.18) and the definition of Au, f and $\partial u/\partial n_A$ we obtain (3.6).

This proves the theorem completely.

Remark 3.1. The above argument using the theorem of Ky Fan and Sion can be carried out for the functional J given again by (3.1) but the convex set K of (3.2) replaced by any one of the following sets

$$\begin{aligned} K_1 &= \{v | v \epsilon H^1_\circ(\Omega), v \geq 0 \text{ a. e. in } \Omega\},\\ K_2 &= \{v | v \epsilon H^1(\Omega), \gamma_\circ v \geq 0 \text{ a. e. on } \Gamma\} \text{ and}\\ K_3 &= \{v | v \epsilon H^1(\Omega), 1 - grad^2 u(x) \geq 0 \text{ a. e. in } \Omega\}. \end{aligned}$$

Since $v \epsilon H^1(\Omega), \gamma_\circ v \epsilon H^{\frac{1}{2}}(\Gamma), 1 - grad^2 u(x) \epsilon L^1(\Omega)$ and since

$$(H^1_\circ(\Omega))' = H^{-1}(\Omega), (H^{\frac{1}{2}}(\Gamma))' = H^{-\frac{1}{2}}(\Gamma), (L^{-1}(\Omega))' = L^\infty(\Omega)$$

we will have to choose the cone Λ respectively in these spaces.

We recall that if E is a vector space in which we have a notion of positivity then we can define in a natural way a notion of positivity in its dual space E' by requiring an element $\mu \epsilon E'$ is positive (i.e. $\mu \geq 0$ in E') if and only if $< \mu, \varphi >_{E' \times E} \geq 0$, $\forall \varphi \epsilon E$ with $\varphi \geq 0$. For the above examples we can take for E the spaces $H^1_\circ(\Omega), H^{\frac{1}{2}}(\Gamma)$ and $L^1(\Omega)$ respectively and we have notions of positivity for their dual spaces.

We can now take

$$\begin{aligned} \Lambda_1 &= \{\mu \epsilon H^{-1}(\Omega) | \mu \geq 0 \text{ in } \Omega\},\\ \Lambda_2 &= \{\mu | \mu \epsilon H^{-\frac{1}{2}}(\Gamma), \mu \geq 0 \text{ on } \Gamma\} \text{ and}\\ \Lambda_3 &= \{\mu | \mu \epsilon L^\infty(\Omega), \mu \geq 0 \text{ in } \Omega\}. \end{aligned}$$

and correspondingly the Lagrangians 170

$$\begin{aligned} \mathscr{L}_1(v, \mu) &= J(v) + < \mu, v >_{H^1(\Omega) \times H^1_\circ(\Omega)},\\ \mathscr{L}_2(v, \mu) &= J(v) + < \mu, \gamma_\circ v >_{H^{-\frac{1}{2}}(\Gamma) \times H^{\frac{1}{2}}(\Gamma)} \text{ and}\\ \mathscr{L}_3(v, \mu) &= J(v) + < \mu, v >_{L^\infty(\Omega) \times L^1(\Omega)} . \end{aligned}$$

We leave other details of the proof to the reader except to remark that Λ_i being cones in infinite dimensional Banach spaces the sets $\Lambda_{i,\ell}(i = 1, 2, 3)$ for any $\ell > 0$ will only be convex sets which are compact in the weak topologies of $H^{-1}(\Omega)$ and $H^{-\frac{1}{2}}(\Gamma)$ for $i = 1, 2$ and in the weak $*$ topology of $L^\infty(\Omega)$ for $i = 3$.

3.2 Dual Problem

We once again restrict ourselves to the problem considerer in 3.1 i.e. J is a quadratic form on $V = H^{-1}(\Omega)$ given by (3.1) and the closed convex set K is given by (3.2). We shall study the dual problem in this case. We take Λ and Φ as before.

We recall that the dual problem is the following:

To find $(u,\lambda)\epsilon V \times \Lambda$ such that

$$\begin{aligned}\mathscr{L}(u,\lambda) &= \sup_{\mu\geq 0}\inf_{v\epsilon V}\mathscr{L}(v,\mu)\\ &= \sup_{\mu\geq 0}\inf_{v\epsilon V}\{\frac{1}{2}a(v,v) - L(v) + \frac{1}{2}\mu(||v||^2 - 1)\}.\end{aligned}$$

We fix a $\mu \geq 0$.

First of all we consider the minimization problem without constrains for the functional

$$\mathscr{L}(\cdot,\mu) : v \mapsto \frac{1}{2}a(v,v) - L(v) + \frac{1}{2}\mu(||v||^2 - 1)$$

171 on the space $V = H^1(\Omega)$. We know from Chapter 2 (Theorem 2. 2.1) that it has a unique minimum $u_\mu \epsilon V$ since $\mathscr{L}(\cdot,\mu)$ has a gradient and a hessian (which is coercive) everywhere. Moreover, $(grad\mathscr{L}(\cdot,\mu))(u_\mu) = 0$ i.e. we have

$$\text{(3.19)} \qquad a(u_\mu,\varphi) - L(\varphi) + \mu(u_\mu,\varphi) = 0, \quad \forall\varphi\epsilon V.$$

We can write using Fréchet-Riesz theorem

$$a(u,\varphi) = ((Au,\varphi)), L(\varphi) = ((F,\varphi)), (u,\varphi) = ((Bu,\varphi))$$

where $((\cdot,\cdot))$ denotes the inner product in $H^1(\Omega)$ and $Au, F, Bu\epsilon H^1(\Omega)$. Then (3.19) can be rewritten as

$$\text{(3.20)} \qquad Au_\mu - F + \mu Bu_\mu = 0.$$

Hence the unique solution $u_\mu\epsilon V$ of the minimizing problem without constrainer for $\mathscr{L}(\cdot,\mu)$ is given by

$$\text{(3.21)} \qquad u_\mu = (A + \mu B)^{-1}F.$$

We can now formulate our next result as follows.

Theorem 3.2. *Under the assumptions of Theorem 3.1 the dual of the primal Problem 3.1 is the following:*

To find $\lambda \epsilon \Lambda$ such that $J^*(\Lambda) = \inf_{\mu\epsilon\lambda} J^*(\mu)$, where

$$J^*(\mu) = ((F, u_\mu)) + \mu. \text{ i.e.} \tag{3.22}$$

Dual Problem (3.2). To find $\lambda \geq 0$ such that $J^*(\lambda) = \inf_{\mu\geq 0} J^*(\mu)$.

Proof. Consider

$$\begin{aligned}\mathscr{L}(u_\mu, \mu) &= \frac{1}{2}((Au_\mu, u_\mu)) - ((F, u_\mu)) + \frac{1}{2}\mu(\|u_\mu\|^2 - 1)\\ &= \frac{1}{2}((Au_\mu, u_\mu)) - ((F, u_\mu)) + \frac{1}{2}\mu(((Bu_\mu, u_\mu)) - 1)\\ &\frac{1}{2}(((A + \mu B)u_\mu, u_\mu) - (F, u_\mu)) - \mu/2.\end{aligned}$$

□ 172

Now using (3.20) we can write

$$\mathscr{L}(u_\mu, \mu) = -\frac{1}{2}((F, u_\mu)) - \mu/2 = -\frac{1}{2}\{((F, u_\mu)) + \mu\}$$

Thus we see that

$$\begin{aligned}\sup_{\mu\geq 0}\inf_{v\epsilon V} \mathscr{L}(v, \mu) &= \sup_{\mu\geq 0}(-\frac{1}{2})\{((F, u_\mu)) + \mu\}\\ &= -\frac{1}{2}\inf_{\mu\geq 0} J^*(\mu)\end{aligned}$$

which proves the assertion.

We wish to construct an algorithm for the solution of the dual problem (3.2). We observe that in this case the constraint set $\Lambda = \{\mu | \mu\epsilon\mathbb{R}, \mu \geq 0\}$ is a cone with vertex at $0\epsilon\mathbb{R}$ and that numerically it is easy to compute the projection on a cone. In face, in our special case we have

$$P_\Lambda(\mu) = \begin{cases}\mu & \text{if } \mu \geq 0\\ 0 & \text{otherwise .}\end{cases}$$

Hence we can use the algorithm given by the method of gradient with projection. This we shall discuss a little later. We shall need, for this method, to calculate the gradient of the cost function J^* for the dual problem.

Form (3.22) we have

$$J^*(\mu) = ((F, u_\mu)) + \mu.$$

Taking G-derivatives on both sides we get

$$\text{(3.23)} \qquad (\text{grad } J^*)(\mu) = J^*(\mu) = ((F, u'_\mu)) + 1$$

where u'_μ is the derivative of u_μ with respect to μ. In order to compute u'_μ we differentiate with equation (3.20) with respect to μ to get.

$$Au'_\mu + \mu Bu'_\mu + Bu_\mu = 0$$

173 and so

$$\text{(3.24)} \qquad u'_\mu = -(A + \mu B)^{-1} Bu_\mu.$$

Substituting (3.24) in (3.23) we see that

$$J^*(\mu) = -((F, (A + \mu B)^{-1} Bu_\mu)) + 1.$$

Since $a(\cdot, \cdot)$ is symmetric A is self adjoint and since $(\cdot, \cdot)$ is symmetric B is also self adjoint. Then $(A + \mu B)^{-1}$ is also self adjoint. This fact together with (3.21) will imply

$$J^*(\mu) = -((A + \mu B)^{-1} F, Bu_\mu) + 1 = -(u_\mu, Bu_\mu) + 1$$

This nothing but saying

$$\text{(3.25)} \qquad J^*(\mu) = 1 - ||u_\mu||^2$$

Remark 3.2. In our discussion above the functional Φ is defined by (3.4) and we found the gradient of the dual cost function is given by 3.25. More generally, if $\Phi(v, \mu) = (g(v), \mu)$ then the gradient of the dual cost function can be shown to be $J^*(\mu) = -g(u_\mu)$. We leave the straight forward verification of this fact to the reader.

3.3 Method of Uzawa

The method of Uzawa that we shall study in this section gives an algorithm to construct a minimizing sequence for the dual problem and also an algorithm for the primal problem itself (see [6], [49]). The important idea used is that since the dual problem is one of minimization over a cone in a suitable space it is easy to compute the projection numerically onto such a cone. The algorithm we give is nothing but the method of 174 gradient with projection for the dual problem (see Section 3 of Chapter 2). We shall show that this method provides a strong convergence of the minimizing sequence obtained for the primal problem while we have only a very weak result on the convergence of the algorithm for the dual problem.

In general the algorithm for the dual problem may not converge. The interest of the method is mainly the convergence of the minimizing sequence for the primal problem.

We shall once again restrict ourselves only to the situation considered earlier i.e. J, K, Λ, Φ and $\mathscr{L}$ are defined by (3.1) - (3.5) respectively.

Algorithm. Let $\lambda_\circ$ be an arbitrarily fixed point and suppose λ_m is determined.

We define λ_{m+1} by

$$\lambda_{m+1} = P_\Lambda(\lambda_m - \rho J^*(\lambda_m)). \tag{3.26}$$

where P_Λ denotes the projection on to the cone Λ and $\rho > 0$.

In our special case we get, using (3.25).

$$\lambda_{m+1} = P_\Lambda(\lambda_m - \rho(1 - ||u_m||^2)) \tag*{(3.26)$'$}$$

where $u_m = u_{\lambda_m}$ is the unique solution of the problem

$$Au_m + \lambda_m Bu_m = F. \tag*{(3.20)$'$}$$

i.e.

$$u_m = (A + \lambda_m B)^{-1}F. \tag*{(3.21)$'$}$$

We remark that (3.21) is equivalent to solving a Neumann problem. In the special case where $a(v, v) = |||v|||^2$ we have to solve the Neumann problem

$$(3.20)'' \qquad \begin{cases} \triangle u_m + (1 + \lambda_m) u_m & = F \text{ in } \Omega, \\ \partial u_m / \partial \underline{n} & = 0 \text{ on } \Gamma \end{cases}$$

i.e. At each stage of the iteration we need to solve a Neumann problem in order to determine the next iterate λ_{m+1}. 175

We shall prove the following main result of this section.

Theorem 3.3. *Suppose the hypothesis of Theorem 3.1 are satisfied. Then we have the following assertions.*

(a) The sequence $u_m = u_{\lambda_m}$ determined by (3.20)′ *converges strongly to the (unique) solution of the primal Problem 3.1.*

(b) Any cluster point of the sequence λ_m determined by (3.26)′ *is a solution of the dual Problem 3.2.*

The proof of the theorem is in several steps. For this we shall need a Taylor's formula for the dual cost function J^* (i.e. the functional (3.22)) and an inequality which is a consequence of Taylor's formula.

Taylor's formula for J^*. Let $\lambda, \mu \epsilon \Lambda$ and we consider the problem

$$(3.27) \qquad (A + \lambda B)u = F \text{ and } (A + \mu B)v = F$$

where we have written $u_\mu = v$ and $u_\lambda = u$. We can also write the first equation as

$$(A + \lambda B)v = F - (\mu - \lambda)Bv = (A + \lambda B)u - (\mu - \lambda)Bv$$

i.e.

$$(A + \lambda B)(v - u) = -(\mu - \lambda)Bv.$$

Similarly we have

$$(A + \mu B)(v - u) = -(\mu - \lambda)Bu.$$

which implies that

$$u_\mu - u_\lambda = v - u = -(\mu - \lambda)(A + \mu B)^{-1} B u_\lambda \tag{3.28}$$

Then (3.22) together with (3.28) gives

$$\begin{aligned} J^*(\mu) &= J^*(\lambda) + ((F, u_\mu - u_\lambda) + (\mu - \lambda)) \\ &= J^*(\lambda) - (\mu - \lambda)((F, (A + \mu B)^{-1} b u_\lambda)) + \mu - \lambda \\ &= J^*(\lambda) - (\mu - \lambda)(((A + \mu B)^{-1} F, B u_\lambda)) + \mu - \lambda \end{aligned}$$

since $(A + \mu B)^{-1}$ is self adjoint because $a(\cdot, \cdot)$ is symmetric and $(\cdot, \cdot)$ is 176
symmetric. Once again using the second equation in (3.27) we get

$$\begin{aligned} J^*(\mu) &= J^*(\lambda) - (\mu - \lambda)((u_\mu, B u_\lambda)) + (\mu - \lambda) \\ &= J^*(\lambda) - (\mu - \lambda)(u_\lambda, u_\lambda) + (\mu - \lambda) - (\mu - \lambda)(u_\lambda - u_\mu, u_\lambda) \end{aligned}$$

where we have used $((\cdot, B\cdot)) = (\cdot, \cdot)$. i.e. We have

$$J^*(\mu) = J^*(\lambda) + (\mu - \lambda)[1 - \|u_\lambda\|^2] - (\mu - \lambda)(u_\lambda - u_\mu, u_\lambda). \tag{3.29}$$

We shall now get an estimate for the last term of (3.29). From (3.28) we can write

$$(((A + \mu V)(v - u), v - u)) = -(\mu - \lambda)((Bu, v - u))$$

which is nothing but

$$a(v - u, v - u) + \mu(v - u, v - u) = -(\mu - \lambda)(u, v - u).$$

Using coercivity of $a(\cdot, \cdot), \mu(v - u, v - u) \geq 0$ on the left side and Cauchy-Schwarz inequality on the side we get

$$\alpha |||v - u|||^2 \leq |\mu - \lambda| |||u||| |||v - u|||$$

i.e.

$$|||v - u||| \leq |\mu - \lambda| / \alpha |||u||| \tag{3.30}$$

On the other hand, since u is a solution of (3.20), we also have

$$a(u,u) + \lambda(u,u) = L(u)$$

from which we get again using coercivity on the left

$$\alpha |||u|||^2 \leq ||L||_{V'} |||u||| \leq N|||u|||, \text{ for some constant } N > 0.$$

i.e. 177

$$|||u||| \leq N/\alpha.$$

On substituting this in (3.30) we get the estimate

$$|||v - u||| \leq N|\mu - \lambda|/\alpha^2$$

which is the same thing as

$$|||u_\mu - u_\lambda||| \leq N|\mu - \lambda|/\alpha^2. \tag{3.31}$$

Finally (3.29) together with this estimate (3.31) implies

$$J^*(\mu) \leq J^*(\lambda) + (\mu - \lambda)(1 - ||u_\lambda||^2) + N^2|\mu - \lambda|^2/\alpha^3. \tag{3.32}$$

Proof of Theorem 3.3.

Step 6. $J^*(\lambda_m)$ is a decreasing sequence and is bounded below if the parameter $\rho > 0$ is sufficiently small. We recall that λ_{m+1} is bounded as

$$\lambda_{m+1} = P_\Lambda(\lambda_m - \rho(1 - ||u_m||^2)).$$

We know that in the Hilbert space $\mathbb{R}$ the projection P onto the closed convex set Λ is characterized by the variational inequality

$$(\lambda_m - \rho(1 - ||u_m||^2) - \lambda_{m+1}, \mu - \lambda_{m+1})_{\mathbb{R}} \leq 0, \forall \mu \epsilon \Lambda.$$

i.e. we have

$$(\lambda_m - \rho(1 - ||u_m||^2) - \lambda_{m+1})(\mu - \lambda_{m+1}) \leq 0, \forall \mu \epsilon \Lambda. \tag{3.33}$$

Putting $\mu = \lambda_m$ in this variational inequality we find

$$|\lambda_m - \lambda_{m+1}|^2 \leq \rho(1 - ||u_m||^2)(\lambda_m - \lambda_{m+1}) \tag{3.34}$$

178 On the other hand (3.32) with $\mu = \lambda_{m+1}, \lambda = \lambda_m, u_\lambda = u_m (= u_{\lambda_m})$, becomes

$$J^*(\lambda_{m+1}) \leq J^*(\lambda_m) + (\lambda_{m+1} - \lambda_m)(1 - ||u_m||^2) + M|\lambda_{m+1} - \lambda_m|^2$$

where M is the constant $N^2/\alpha^3 > 0$. If we use (3.34) on the right side of this inequality we get

$$J^*(\lambda_{m+1}) \leq J^*(\lambda_m) - 1/\rho|\lambda_{m+1} - \lambda_m|^2 + M|\lambda_{m+1} - \lambda_m|^2$$

i.e.

$$J^*(\lambda_{m+1}) + (1/\rho - M)|\lambda_{m+1} - \lambda_m|^2 \leq J^*(\lambda_m). \tag{3.35}$$

Here, $1/\rho - M$ would be > 0 if we take $0 < \rho < 1/M = \alpha^3/N^2$, a fixed constant independent of λ. We therefore take $\rho \epsilon]0, 1/M[$ in the definition of λ_{m+1} so that we have

$$J^*(\lambda_{m+1}) \leq J^*(\lambda_m),$$

which proves that the sequence $J^*(\lambda_m)$ is decreasing for $0 < \rho < 1/M$. To prove that it is bounded below we use the definition of $J^*(\lambda)$ and Cauchy-Schwarz inequality: From (3.22)

$$J^*(\lambda) = ((F, u_\lambda)) + \lambda \geq -|||F||||||u_\lambda||| \geq -N/\alpha|||F|||$$

since $|||u_\lambda||| \leq N/\alpha$. This proves that $J^*(\lambda_m)$ is bounded below by $-N/\alpha|||F|||$, a known constant.

Step 7. By step 1 it follows that $J^*(\lambda_m)$ converges to a limit as $m \to +\infty$. Moreover, (3.35) will then imply that

$$|\lambda_{m+1} - \lambda_m|^2 \to 0 \text{ as } m \to +\infty. \tag{3.36}$$

Step 8. The sequence λ_m has a cluster point in $\mathbb{R}$. For this, since $J^*(\lambda_m)$ is decreasing we have $J^*(\lambda_{m+1}) \leq J^*(\lambda_\circ)$ i.e. we have

$$((F, u_{m+1})) + \lambda_{m+1} \leq ((F, u_\circ)) + \lambda_\circ$$

and the right hand side is a constant independent of m. So, by Cauchy-Schwarz inequality, 179

$$\lambda_{m+1} \leq ((F, u_\circ - u_{m+1})) + \lambda_\circ \leq ((F, u_\circ)) + \lambda_\circ + |||u_{m+1}||||||F|||.$$

But $|||u_{m+1}|||$ is bounded by a constant $(= N/\alpha)$ and hence

$$0 \leq \lambda_{m+1} \leq ((F, u_\circ)) + \lambda_\circ + N|||F|||/\alpha.$$

i.e. The sequence λ_m is bounded. We can then extract a subsequence which converges.

Similarly, since u_m is a bounded sequence in $H^1(\Omega)$ there exists a sub-sequence which converges weakly in $H^1(\Omega)$. Let $\{m'\}$ be a subsequence of the positive integers such that

$$\lambda'_m \to \lambda^* \text{ in } \mathbb{R} \text{ and } u_{m'} = u_{\lambda_{m'}} \rightharpoonup u^* \text{ in } H^1(\Omega).$$

Step 9. Any cluster point λ^* of the sequence λ_m is a solution of the dual problem 3.2.

Let $\lambda_{m'}$ be a subsequence which converges to λ^*. We may assume, if necessary by extracting a subsequence that $u_{m'} \rightharpoonup u^*$ in $H^1(\Omega)$. By Rellich's lemma the inclusion of $H^1(\Omega)$ in $L^2(\Omega)$ is compact (since Ω is bounded) and hence $u_{m'} \to u^*$ in $L^2(\Omega)$. Then u^* satisfies the equation

$$u^* \epsilon H^1(\Omega), Au^* + \lambda^* Bu^* = F. \tag{3.37}$$

To see this, since $u_{m'}$ is a solution of ((3.20)′) we have

$$((Au_{m'}, \varphi)) + \lambda_{m'}((Bu_{m'}, \varphi)) = ((F, \varphi)), \forall \varphi \epsilon H^1(\Omega).$$
$$\text{i.e.} \quad ((Au_{m'}, \varphi)) + \lambda_{m'}(u_{m'}, \varphi) = ((F, \varphi)), \forall \varphi \epsilon H^1(\Omega).$$

Taking limits as $m' \to +\infty$ we have

$$((Au^*, \varphi)) + \lambda^*(u^*, \varphi) = ((F, \varphi)), \forall \varphi \epsilon H^1(\Omega)$$

180 which is the same thing as (3.37).

On the other hand, (3.33) for the subsequence becomes

$$1/\rho(\lambda_{m'}-\lambda_{m'+1})(\mu-\lambda_{m'+1}) \leq (1-||u_{m'}||^2)(\mu-\lambda_{m'+1}), \forall\mu\epsilon\Lambda.$$

Here on the left side $\mu-\lambda_{m'+1}$ is bounded indepedent of m' and $\lambda_{m'}-\lambda_{m'+1} \to 0$ as $m' \to +\infty$ by (3.36). On the right side again by (3.36), $\mu-\lambda_{m'+1} \to \mu-\lambda^*$ and $(1-||u_{m'}||^2) \to (1-||u^*||^2)$ as $m' \to +\infty$. Thus we get on passing to the limits

$$\lambda^*\epsilon\Lambda, (1-||u^*||^2)(\mu-\lambda^*) \geq 0, \forall\mu\epsilon\Lambda. \tag{3.38}$$

Since u^* is a solution of (3.37), we know on using (3.25), that

$$(grad J^*)(\lambda^*) = J^*(\lambda^*) = (1-||u^*||^2).$$

Then (3.38) is the same thing as

$$\lambda^*\epsilon\Lambda, J^*(\lambda^*).(\mu-\lambda^*) \geq 0, \forall\mu\epsilon\Lambda.$$

By the results of Chapter 2 (Theorem 2. 2.2) this last variational inequality characterizes a solution of the dual Problem (3.2). Thus λ^* is a solution of the dual problem.

Step 10. The sequence u_m converges weakly in $H^1(\Omega)$ to the unique solution u of the primal problem.

As in the earlier steps since the sequence u_m is bounded in $H^1(\Omega)$ and λ_m is bounded in $\mathbb{R}$ we can find a subsequence m' of integers such that

$$u_{m'} \rightharpoonup u^* \text{ in } H^1(\Omega) \text{ and } \lambda_{m'} \to \lambda^* \text{ in } \mathbb{R}.$$

We shall prove that (u^*, λ^*) is a saddle point for the Lagrangian. It is easily verified that 181

$$(grad_v\mathscr{L}(\cdot,\lambda^*))(u^*) = a(u^*,u^*) + \lambda^*(u^*,u^*) - L(u^*).$$

But the right hand side vanishes because u^* is the solution of the equation

$$Au^* + \lambda^*Bu^* = F$$

as can be proved exactly as in Step 4. Moreover $\mathscr{L}(\cdot,\lambda^*)$ is convex (strongly convex). Hence by Theorem 2.2.2.

$$\mathscr{L}(u^*,\lambda^*) \leq \mathscr{L}(v,\lambda^*), \forall v \epsilon H^1(\Omega). \tag{3.39}$$

Next we see similarly that

$$(grad_\mu \mathscr{L}(u^*,\cdot))(\lambda^*) = \frac{1}{2}(||u||^2 - 1)$$

and $\mathscr{L}(u^*,\cdot)$ is concave. One again using (3.38) and the Theorem 2.2.2 we conclude that

$$\mathscr{L}(u^*,\mu) \leq \mathscr{L}(u^*,\lambda^*), \forall \mu \epsilon \Lambda. \tag{3.40}$$

The two inequalities (3.39) and (3.40) together mean that (u^*,λ^*) is a saddle point for $\mathscr{L}$. Hence u^* is a solution of the Primal problem and λ^* is a solution of the dual problem. But since J is strictly convex it has unique minimum in $H^1(\Omega)$. Hence $u = u^*$ and u is the unique weak-cluster point of the sequence u_m in $H^1(\Omega)$. This implies that the entire sequence u_m converges weakly to u in $H^1(\Omega)$.

Step 11. The sequence u_m converges strongly in $H^1(\Omega)$ to the unique solution of the primal problem.

We can write using the definition of the functional J:

$$J(u) = J(u_m) + a(u_m, u - u_m) - L(u - u_m) + \frac{1}{2}a(u - u_m, u - u_m).$$

By the coercivity of $a(\cdot,\cdot)$ applied to the last terms on the right side

$$\begin{aligned} J(u_m) + \alpha/2|||u - u_m|||^2 &\leq J(u) - \{a(u_m, u - u_m) - L(u - u_m)\} \\ &= J(u) + ((Au_m - F, u - u_m)) \\ &= J(u) + \lambda_m((Bu_m, u - u_m)) \end{aligned}$$

182 since u_m satisfies the equation ((3.20)′). i.e. we have

$$J(u_m) + \alpha/2|||u - u_m|||^2 \leq J(u) + \lambda_m(u_m, u - u_m).$$

On the left hand side we know that $J(u_m) \to J(u)$ and on the right hand side we know that $|\lambda_m|$ and u_m are bounded while by Step 5, $u - u_m \rightharpoonup 0$ (weakly)in $H^1(\Omega)$.

Hence taking limits as $m \to +\infty$ we see that

$$|||u - u_m||| \to 0 \text{ as } m \to +\infty.$$

This completely proves the theorem.

In conclusion we make some remarks on the method of Uzawa.

Remark 3.3. In the example we have considered to describe the method of Uzawa Λ is a cone in $\mathbb{R}$. But, in general, the cone Λ will be a subset of an infinite dimensional (Banach) space. We can still use our argument of Step 3 of the proof to show that λ_m has a weak cluster point and that of Step 4 to show that a weak cluster point gives a solution of the dual problem.

Remark 3.4. We can also use the method of Frank and Wolfe since also in this case the dual problem is one of minimization on a cone on which it is easy to compute projections numerically.

Remark 3.5. While the method of Uzawa gives strong convergence results for the algorithm to the primal the result the dual problem is very weak.

Remark 3.7. Suppose we consider a more general type of the primal problem for the same functional J defined by (3.1) of the form:

$$\text{to find } u \epsilon K, J(u) = \inf_{v \epsilon K} J(v)$$

where K is a closed convex by set in $V = H^1(\Omega)$ is defined by 183

$$K = \{v | v \epsilon H^1(\Omega), g(v) \leq 0\}.$$

with g a mapping of $H^1(\Omega)$ into a suitable topological vector space E (in fact a Banach space) in which we have a notion of positivity. Then we take a cone Λ in E as in Remark 3.2 and $\Phi(v, \mu) =< \mu, g(v) >_{E' \times E}$. In order to carry over the same kind of algorithm as we have given above in the special case we proceed as follows: Suppose Λ_m is determined starting from a $\lambda_\circ \epsilon \Lambda$. We firstsolve the minimization problem

$$\begin{aligned} \text{to find } u_m \text{ such that } \mathscr{L}(u_m, \lambda_m) &= \inf_v \mathscr{L}(v, \lambda_m) \\ grad J^*(\lambda_m) &= -g(u_m) \end{aligned}$$

Then we can use Remark 3.2 to determine λ_{m+1} :

$$\lambda_{m+1} = P_\Lambda(\lambda_m - \rho J^*(\lambda_m)) = P_\Lambda(\lambda_m + \rho g(u_\lambda)).$$

We can now check that the rest of our argument goes through easily in this case also except that we keep in view our earlier remarks about taking weak topologies in E'. For instance, we can use this procedure in the cases of convex sets K_1, K_2, K_3 of Remark 3.1. We leave the details of these to the reader.

4 Minimization of Non-Differentiable Functionals Using Duality

In this section we apply the duality method using Ky Fan and Sion Theorem to the case of a minimization problem for a functional which is not G-differentiable. The main idea is to transform the minimization problem into one of determining a saddle point for a suitable functional
184 on the product of the given space with a suitable cone. This functional of two variables behaves very much like the Lagrangian (considered in Section 3) for the regular part of the given functional. In fact we choose the cone Λ and the function Φ in such a way that the non-differentiable part of the given functional can be written as $-\sup_{\mu\epsilon\Lambda}\Phi(v,\mu)$. It turns out that in this case the dual cost function will be G-regular and hence we can apply, for instance, the method of gradient with projection. This in its turn enables us to give an algorithm to determine a minimizing sequence for the original minimization problem. The proof of convergence is on lines similar to the one we have given for the convergence of the algorithm in the method of Uzawa.

We shall however begin our discussion assuming that we are given the cone Λ and the function Φ in a special form and thus we start in fact with a saddle point problem.

Let V and E be two Hilbert spaces and let $J_\circ : V \to \mathbb{R}$ be a functional on V of the form

$$V \ni v \mapsto J(v) = \frac{1}{2}a(v, v) - L(v)\epsilon\mathbb{R} \tag{4.1}$$

where as usual we assume:

$$(i) a(\cdot, \cdot) \text{ is a bilinear bicontinuous coercive form on } V \text{ and}$$
$$ii) L\epsilon V' \tag{4.2}$$

Suppose we also have

$$(iii) \text{ a closed convex bounded set } \Lambda \text{ in } E \text{ with } 0\epsilon\Lambda, \text{ and}$$
$$(iv) \text{ and operator } B\epsilon\mathscr{L}(V, E). \tag{4.3}$$

We set

$$J_1(v) = \sup_{\mu\epsilon\Lambda}(-(Bv, \mu)_E) \tag{4.4}$$

and 185

$$J(v) = J_\circ(v) + J_1(v). \tag{4.5}$$

Consider now the minimization problem:

Primal Problem (4.6). To find $u\epsilon V$ such that $J(u) = \inf_{v\epsilon V} J(v)$. We introduce the functional $\mathscr{L}$ on $V \times \Lambda$ by

$$\mathscr{L}(v, \mu) = J_\circ(v) - (Bv, \mu)_E. \tag{4.7}$$

It is clear that if we define $\Phi(v, \mu) = -(Bv, \mu)_E$ then $\mathscr{L}$ can be considered a Lagrangian associated to the functional $J_\circ$ and the cone generated by Λ. Since $v\epsilon V$ the condition that $\Phi(v, \mu) \leq 0$ implies $v\epsilon V$ is automatically satisfied and more over, we also have

$$\Phi(v, \rho\mu) = -(Bv, \rho\mu)_E = -\rho(Bv, \mu)_E = \rho\Phi(v, \mu), \forall\rho > 0.$$

On the other hand we see that the minimax problem for the functional $\mathscr{L}$ is nothing but our primal problem. In fact, we have

$$\text{(4.8)} \qquad \inf_{v \epsilon V} \sup_{\mu \epsilon \Lambda} \mathscr{L}(v,\mu) = \inf_{v \epsilon V} (J_\circ(v) + \sup_{\mu \epsilon \lambda} (-(Bv,\mu)_E)) = \inf_{v \epsilon V} J(v).$$

We are thus led to the problem of finding a saddle point for $\mathscr{L}$.

Remark 4.1. In practice, we are given J_1, the non- G-differentiable part of the functional J to be minimized and hence it will be necessary to choose the hilbert space E, a closed convex bounded set λ in F (with $0 \epsilon \Lambda$) and an operator $B \epsilon \mathscr{L}(V,E)$ suitably so that $J_1(v) = \sup_{\mu \epsilon \Lambda} - (Bv,\mu)_E = -\inf_{\mu \epsilon \Lambda} (Bv,\mu)_E$.

We shall now examine a few examples of the functionals J_1 and the correspond E, Λ, and the operator B. In all the following examples we take

$$V = \mathbb{R}^n, E = \mathbb{R}^m \text{ and } B \epsilon \mathscr{L}(V,E) \text{ an } (m \times n) - \text{ matrix .}$$

186

We also use the following satandard norms in the Euclidean space $\mathbb{R}^m$. If $1 \leq p < +\infty$ then we define the norms:

$$|\mu|_p = (\sum_{i=1}^{m} |\mu_i|^p)^{1/p}$$

and

$$|\mu|_\infty = \sup_{1 \leq i \leq m} |\mu_i|.$$

Example 4.1. Let $\Lambda_1 = \{\mu \epsilon \mathbb{R}^m : |\mu|_2 \leq 1\}$. Then

$$J_1(v) = \sup_{\mu \epsilon \Lambda} (-(Bv,\mu)_E) = |Bv|_2.$$

Example 4.2. Let $\Lambda_2 = \{\mu \epsilon \mathbb{R}^m : |\mu|_1 \leq 1\}$. Then $J_1(v) = |Bv|_\infty$. If we denote the elements of the matrix B by b_{ij} then $b_i = (b_{i1}, \cdots, b_{in})$ is a vector in $\mathbb{R}^n$ and $Bv = ((Bv)_1, \cdots, (Bv)_m)$:

$$(Bv)_i = (b_i, v)_{\mathbb{R}^n} = \sum_{j=1}^{n} b_{ij} v_j.$$

Hence

$$J_1(v) = \max_{1 \leq i \leq m} |(Bv)_i| = \max_{1 \leq i \leq m} |\sum_{j=1}^{n} b_{ij} v_j|.$$

Example 4.3. If we take $\Lambda_3 = \{\mu \epsilon \mathbb{R}^m ; |\mu|_\infty \leq 1\}$ then we will find $J_1(v) = |Bv|_1$ and hence

$$J_1(v) = \sum_{i=1}^{m} |\sum_{j=1}^{n} b_{ij} v_j|$$

Example 4.4. If we take $\Lambda_4 = \{\mu \epsilon \mathbb{R}^m ; |\mu|_\infty \leq 1, \mu \geq 0\}$ then we find

$$J_1(v) = |(Bv)^+|_1 \text{ where } ((Bv)^+)_i = \begin{cases} (Bv)_i \text{ when } (Bv)_i \geq 0 \\ 0 \text{ when } (Bv)_i < 0. \end{cases}$$

Hence

$$J_1(v) = \sum_{i=1}^{m} |\sum_{j=1}^{n} (b_{ij} v_j)^+| = \sum_{i=1}^{m} \sum_{j=1}^{n} (b_{ij} v_j)^+.$$

187

Proposition 4.1. *Under the assumptions made on $J_\circ, \Lambda$ and B there exists a saddle point for $\mathscr{L}$ in $V \times \Lambda$.*

Proof. The mapping $v \mapsto \mathscr{L}(v, \mu)$ of $V \to \mathbb{R}$ is convex (in fact strictly convex since $a(\cdot, \cdot)$ is coercive) and continuous and in particular lower semi-continuous. The mapping $\Lambda \ni \mu \mapsto (v, \mu)$ is concave and continuous and hence is upper semi-continuous. Let $\ell > 0$ be a constant which we shall choose suitably later on and let us consider the set

$$U_\ell = \{v | v \epsilon V, ||v||_V \leq \ell\}.$$

□

The set U_ℓ is a closed convex bounded set in V and hence is weakly compact. Similarly Λ is also weakly in E. Thus taking weak topologies on V and E we have two Hausdorff topological vector spaces. We can now apply the theorem of Ky Fan and Sion to sets U_ℓ and Λ. We see that there exists a saddle point $(u_\ell, \lambda_\ell)\epsilon U_\ell \times \Lambda$ for $\mathscr{L}$. i.e. We have

(4.9)

$$(u_\ell, \lambda_\ell)\epsilon U_\ell \times \lambda, \mathscr{L}(u_\ell, \mu) \leq \mathscr{L}(u_\ell, \lambda_\ell) \leq \mathscr{L}(v, \lambda_\ell),\ \forall (v, \mu)\epsilon U_\ell \times \Lambda.$$

Choosing $\mu = 0$ in the first inequality of (4.9) we get $0 \leq -(Bu_\ell, \lambda_\ell)_E$ i.e. $(Bu_\ell, \lambda_\ell)_E \leq 0$ and

$$J_\circ(u_\ell) \leq J_\circ(u_\ell) - (Bu_\ell, \lambda_\ell)_E \leq J_\circ(v) - (Bv, \lambda_\ell)_E.$$

Next, if we take $v = 0\epsilon U_\ell$ we get

$$J_\circ(u_\ell) \leq J_\circ(v)(= 0). \tag{4.10}$$

From this we can show that $||u_\ell||_V$ is bounded. In fact, the inequality (4.10) is nothing but

$$\frac{1}{2}a(u_\ell, u_\ell) - L(u_\ell) \leq 0.$$

188 Using the coercivity of $a(\cdot, \cdot)$ (with the constant of coercivity $\alpha > 0$)

$$\alpha ||u_\ell||_V^2 \leq a(u_\ell, u_\ell) \leq 2L(u_\ell) \leq 2||L||_{V'}||u_\ell||_V$$

$$\text{i.e. } ||u_\ell||_V \leq 2||L||_{V'}/\alpha. \tag{4.11}$$

In other words, $||u_\ell||_V$ is bounded by a constant $c = 2||L||_{V'}/\alpha$ independent of ℓ.

Now we take $\ell > c$. Then we can find a ball $B(u_\ell, r) = \{v\epsilon V|||v - u_\ell||_V < r\}$ contained in the ball $B(0, \ell)$. It is enough to take $r\epsilon]0, \frac{\ell-c}{2}[$. The functional $J_\circ$ attains a local minimum in such a ball. Now $J_\circ$ being (strictly) convex it is the unique global minimum. Thus we have proved that if we choose $\ell > c > 0$ where $c = 2||L||_{V'}/\alpha$ there exists

(4.12)

$$(u, \lambda)\epsilon V \times \Lambda \text{ such that } \mathscr{L}(u, \mu) \leq \mathscr{L}(u, \lambda) \leq \mathscr{L}(v, \lambda)\forall (v, \mu)\epsilon V \times \Lambda$$

which means that (u, λ) is a saddle point for $\mathscr{L}$ in $V \times \Lambda$.

Dual problem. By definition the dual problem is characterized by considering the problem:

$$(4.13) \qquad \begin{cases} \text{to find } (u, \lambda) \epsilon U \times \Lambda \text{ such that} \\ \sup_{\mu \epsilon \Lambda} \inf_{v \epsilon V} \mathscr{L}(v, \mu) = \mathscr{L}(u, \lambda). \end{cases}$$

We write $\mathscr{L}(v, \mu)$ in the following form: Since the mapping $v \mapsto a(u, v)$ is continuous linear there exists an element $Au \epsilon V$ such that

$$a(u, v) = (Au, v)_V, \forall v \epsilon V.$$

Moreover, $A \epsilon \mathscr{L}(V, V)$. Also by Frechet-Riesz theorem there exists an $F \epsilon V$ such that

$$L(v) = (F, v)_V, \quad \forall \epsilon V.$$

Thus we have **189**

$$\begin{aligned} \mathscr{L}(v, \mu) &= \frac{1}{2}(Av, v)_V - (F, v)_V - (Bv, \mu)_E \\ &= \frac{1}{2}(Av, v)_V - (v, F + B^* \mu)_V. \end{aligned}$$

For any $\mu \epsilon \Lambda$ fixed we consider the minimization problem

$$(4.14) \qquad \text{to find } u_\mu \epsilon \lambda \text{ such that } \mathscr{L}(u_\mu, \mu) = \inf_{v \epsilon V} \mathscr{L}(v, \mu).$$

Once again $v \mapsto \mathscr{L}(v, \mu)$ is twice G-differentiable and has a gradient and a hessian everywhere in V. In fact,

$$(4.15) \qquad (grad_v \mathscr{L}(\cdot, \mu))(\varphi) = (Av, \varphi)_V - (F, \varphi)_V - (B^* \mu, \varphi)$$

and

$$(Hess_v \mathscr{L}(\cdot, \mu))(\varphi, \psi) = (A\psi, \varphi)_V.$$

Hence, the coercivity of $a(\cdot, \cdot)$ implies that

$$(Av, v)_V = a(v, v) \geq \alpha ||v||_V^2, \forall v \epsilon V$$

which then implies that $v \mapsto \mathscr{L}(v, \mu)$ is strictly convex. Then by Theorem 2.2.2 there exists a unique solution u_μ of the problem (4.14) and u_μ satisfies the equation

$$[grad_v \mathscr{L}(\cdot, \mu)]_{v=u_\mu} = 0.$$

i.e. There exists a unique $u_\mu \epsilon V$ such that

$$\mathscr{L}(u_\mu, \mu) = \inf_{v \epsilon V} \mathscr{L}(v, \mu)$$

and moreover u_μ satisfies the equation

(4.16) $$(Au_\mu, \varphi)_V - (B^*\mu, \varphi)_V - (F, \varphi)_V = 0, \forall \varphi \epsilon V.$$

i.e.

(4.16) $$Au_\mu = F + B^*\mu.$$

190 Thus we have

(4.17) $$u_\mu = A^{-1}(F + B^*\mu)$$

and taking $\varphi = u_\mu$ in (4.16) we also find that

(4.18) $$(Au_\mu, u_\mu)_V = (F + B^*\mu, u_\mu)_V.$$

using (4.17) and (4.18) we can write

$$\begin{aligned}
\mathscr{L}(u_\mu, \mu) &= \frac{1}{2}\{(Au_\mu, u_\mu)_V - 2(F, u_\mu)_V - 2(B^*\mu, u_\mu)_V\} \\
&= -\frac{1}{2}\{(F, u_\mu)_V + (B^*\mu, u_\mu)_V\} \\
&= -\frac{1}{2}\{(F, A^{-1}(F + B^*\mu))_V + (B^*\mu, A^{-1}(F + B^*\mu))_V\} \\
&= -\frac{1}{2}\{(BA^{-1}B^*\mu, \mu)_E + 2(BA^{-1}F, \mu)_E + (F, A^{-1}F)_E\}
\end{aligned}$$

since A is symmetric implies A^{-1} is also self adjoint. Thus we see that

$$\sup_{\mu \epsilon \Lambda} \inf_{v \epsilon V} \mathscr{L}(v, \mu) = \sup_{\mu \epsilon \Lambda} \mathscr{L}(u_\mu, \mu)$$

$$= \sup_{\mu\epsilon\Lambda} -\frac{1}{2}\{(BA^{-1}B^{*}\mu,\mu)_E + 2(BA^{-1}F,\mu)_E + (F,A^{-1}F)_E\}.$$

If we set

$$\mathscr{A} = BA^{-1}B^{*} \text{ and } \mathscr{F} = -BA^{-1}F \tag{4.19}$$

then $\mathscr{A}\epsilon\mathscr{L}(E.E)$ and $\mathscr{F}\epsilon E$ and moreover

$$\sup_{\mu\epsilon\Lambda}\inf_{v\epsilon V}\mathscr{L}(v,\mu) = -\frac{1}{2}\inf_{\mu\epsilon\Lambda}\{(\mathscr{A}\mu,\mu)_E - 2(\mathscr{F},\mu)_E + (F,A^{-1}F)_E\}. \tag{4.20}$$

Here the functional

$$\mu \mapsto \frac{1}{2}(\mathscr{A}\mu,\mu)_E - (\mathscr{F},\mu)_E \tag{4.21}$$

is quadratic on the convex set λ. It is twice G-differentiale with respect 191 to μ in all directions in L and has a gradient $G^{*}(\mu)$ and a Hessian $H^{*}(\mu)$ every where in Λ. In fact, we can easily see that

$$G^{*}(\mu) = \mathscr{A}\mu - \mathscr{F}. \tag{4.22}$$

Thus we have provd the following

Proposition 4.2. *Under the assumptions made on $J_\circ,\Lambda$ and B the dual of the primal problem (4.6) is the following problem:*

Dual Problem.

$$\textit{To find } \lambda\epsilon\Lambda \textit{ such that } J^{*}(\Lambda) = \inf_{\mu\epsilon\Lambda} J^{*}(\mu), \tag{4.23}$$

where

$$\begin{cases} J^{*}(\mu) & = \frac{1}{2}(\mathscr{A}\mu,\mu)_E - (\mathscr{F},\mu)_E, \\ \mathscr{A} & = BA^{-1}B^{*}, \mathscr{F} = -BA^{-1}F. \end{cases} \tag{4.24}$$

Remark 4.2. In view of the Remark (3.2) and the fact that $g(v) = -Bv$ in our case we know that the gradient of J^{*} is given by $G^{*}(\mu) = +Bu_\mu$.

We see easily that this is also the case in pur present problem. In fact, by (4.24)

$$G^*\mu = \mathscr{A}\mu - \mathscr{F} = BA^{-1}B^*\mu + BA^{-1}F = BA^{-1}(B^*\mu + F).$$

On the other hand, by (4.17) $u_\mu = A^{-1}(B^*\mu + F)$ so that

$$Bu_\mu = BA^{-1}(B^*\mu + F) = G^*(\mu).$$

Algorithm. To determine a minimizing sequence for our primal proble we can use the same algorithm as in the method of Uzawa.

Suppose $\lambda_\circ$ is an arbitrarily fixed point in Λ. We determine $u_\circ$ by solving the equation

$$(4.25) \qquad u_\circ \epsilon V, Au_\circ = F + B^*\lambda_\circ.$$

192

If we have determined λ_m (and u_{m-1}) iteratively we determine u_m as the unique solution of the functional (differential in most of the applications) equation

$$(4.26) \qquad u_m \epsilon V, Au_m = F + B^*\lambda_m$$

i.e. u_m is the solution of the equation

$$(4.26)' \qquad a(u_m, \varphi) = (F + B^*\lambda_m, \varphi)_V = (F, \varphi)_V + (\lambda_m, B\varphi)_E, \forall\varphi\epsilon V.$$

Then we define

$$(4.27) \qquad \lambda_{m+1} = P_\Lambda(\lambda_m - \rho Bu_m)$$

where P_Λ is the projection of E onto the closed convex set Λ and $\rho > 0$ is a sufficiently small parameter.

The convergence of the algorithm to a solution of the minimizing problem for the (non-differentiable) functional $J, J = J_\circ + J_1$, can be proved exactly as in the proof of convergence in the method of Uzawa. However, we shall omit the details of this proof.

Remark 4.3. If we choose the Hilbert space E, the convex set Λ in E and the operator $B\epsilon\mathscr{L}(V, E)$ properly this method provides a good algorithm to solve the minimization problem for many of the known non-differentiable functionals.

Remark 4.4. In the above algorithm (4.26) is a linear system if V is finite dimensional, and if V is an infinite dimensional (Hilbert) space then (4.26) can be interpreted as a Neumann type problem.

Remark 4.5. We can also give an algorithm using the method of Franck and Wolfe to solve the dual problem instead of the method of gradient with projection. Here we can take $\rho > 0$ to be a fixed constant which is sufficiently small.

Chapter 6

Elements of the Theory of Control and Elements of Optimal Design

This chapter will be concerned with two problem which can be treated 193
can be using the techniques developed in the previous chapters, namely,

(1) the optimal control problem,

(2) the problem of optimal design.

These two problems are somewhat similar. We shall reduce the problems to suitable minimization problems so that we can use the algorithms discussed in earlier chapters to obtain approximations to the solution of the two problems considered here.

1 Optimal Control Theory

We shall give an abstract formulation of the problem of optimal control and this can be considered as a problem of optimization for a functional on a convex set of functions. By using the duality method for example via the theorem of *Ky* Fan and Sion we reduce our control problem to a system consisting of the state equation, the adjoint state equation, and a variational inequality for the solution of the original problem. The variational inequality can be considered as Pontrjagin maximum principle well known in control theory. Inorder to obtain an algorithm we eliminate at least formally the state and obtain a pure minimization problem for which we can use the appropriate algorithms described in earlier chapters:

The theory of optimal control can roughly be described starting from the following data. We are given

194 (i) A control u, which belogs to a given convex set K of functions K is called the set of controls.

(ii) The state (of the system to be controled) $y(u) \equiv y_u$ is, for a given $u \epsilon K$, a solution of a functional equation. This equation is called the state equation governing the problem of control.

(iii) A functional $J(y, u)$ - called the cost function - defined by means of certain non-negative functionals of u and y.

If we set

$$j(u) = J(y_u, u)$$

then the problem of optimal control consists in finding a solution of the minimization problem:

$$\begin{cases} u \epsilon K \text{ such that} \\ j(v) = \inf_{v \epsilon K} j(v). \end{cases}$$

Usually the state equations governing the system to be controled are ordinary or partial differential equation.

The main object of the theory is to find necessary (and sufficient) conditions for the existence and uniqueness of the solution of the above problem and to obtain algorithm for determining approximations to the solutions of the problem. We shall restrict ourselves to the optimal control problem governed by partial differential equaiton of elliptic type, more precisely, by linear homogeneous variational elliptic boundary value problems. One can also consider, in a similar way, the problems governed by partial differential equation of evolution type. (See, for instance, the book of Lions [31].)

1.1 Formulation of the Problem of Optimal Control

Let Ω be a bounded open set in the Euclidean space $\mathbb{R}^n$ with smooth boundary Γ. We shall denote the inner product and the corresponding norm in the Hilbert space $L^2(\Omega)$ by $(\cdot, \cdot)$ and $||\cdot||$ while those in the Sobolev space $V = H^1(\Omega)$ by $((\cdot, \cdot))$ and $|||\cdot|||$ respectively. 195

We suppose given the following:

Set of controls. A nonempty closed convex subset K of $L^2(\Omega)$, called the set of controls, and we denote the elements of K by u, which we call controls.

State equation. A continuous, bilinear, coercive form $a(\cdot, \cdot)$ on V i.e. there exists contants $\alpha_a > 0$ and $M_a > 0$ such that

$$(1.1) \qquad \begin{cases} |a(\varphi, \psi)| \leq M_a|||\varphi||||||\psi||| & \text{for all } \varphi, \psi \epsilon V \\ a(\varphi, \varphi) \geq \alpha_a|||\varphi|||^2 & \text{for all } \varphi \epsilon V. \end{cases}$$

Let $f \epsilon L^2(\Omega)$ be given.

For any $u \epsilon K$ a solution of the functional equation

$$(1.2) \qquad \begin{cases} y_u \epsilon V, \\ a(y_u, \varphi) = (f, \varphi) + (u, \varphi) \quad \text{for all } \varphi \epsilon V \end{cases}$$

is said to define a state. The system to be governed is said to be governed by the state equation (1.2). We know, by the results of Chapter 2, that for any $u \epsilon K (\subset L^2(\Omega) \subset V')$ there exists a unique solution y_u of (1.2). Thus for a given f and a given control $u \epsilon K$ there exists a unique state y_u governing the system.

Cost function. Let $b(\cdot, \cdot)$ be a symmetric, continuous and positive semidefinite form on V. i.e. There exists a constant $M_b > 0$ such that

$$(1.3) \qquad \begin{cases} b(\varphi, \psi) = b(\psi, \varphi) & \text{for all } \varphi.\psi \epsilon V \\ |b(\varphi, \psi)| \leq M_b|||\varphi||||||\psi||| & \text{for all } \varphi, \psi \epsilon V \\ B(\varphi, \varphi) \geq 0. \end{cases}$$

196

Further let $C\epsilon\mathscr{L}(L^2(\Omega), L^2(\Omega))$ be an operator the following conditions: there exist positive constants $\alpha_C \geq 0$ and $M_C > 0$ such that

$$\begin{cases} (Cv, v) \geq \alpha_C ||v||^2, \text{ for all } v\epsilon L^2(\Omega) \\ ||C|| \leq M_C \end{cases} \tag{1.4}$$

Let $y_g \epsilon V$ be given. We now define the functional

$$J(y, u) = \frac{1}{2} b(y - y_g, y - y_g) + \frac{1}{2}(Cu, u) \tag{1.5}$$

Proof of control. This consists in finding a solution of the minimization problem:

$$\begin{cases} u\epsilon K \text{ such that} \\ J(y_u, u) = \inf_{v\epsilon K} J(y_v, v) \end{cases} \tag{1.6}$$

We shall show in the next section that the problem (1.6) has a unique solution. However, we remark that one can also prove that a solution of (1.6) u exists and is unique directly using the differential calculus of Chapter 1 and the results of Chapter 2 on the existence and uniqueness of minima of convex functionals.

Definition 1.1. The unique solution $u\epsilon K$ of the problem (1.6) is called the optimale control.

Remark 1.1. If the control set K is a convex set described by a set of functions defined over the whole of Ω and the constraint conditions are imposed on the whole of Ω then the problem (1.6) is said to be one of distributed control. This is the case we have considered here. However, we can also consider in a similar way the problem when K consists of functions defined over the boundary Γ of Ω and satisfying constraint 197 conditions on Γ. In this case the problem is said to be one of boundary control - For example, we can consider

$$\varphi \mapsto \int_\Gamma u\varphi d\sigma$$

defined on a suiteble class of functions φ on Γ.

Remark 1.2. If we set

$$j(u) = J(y_u, u)$$

then the problem of control is a minimization problem for the functional $u \mapsto j(u)$ on K.

Remark 1.3. Usually the state equation governing the system to be controled are ordinary differential equations or partial differential equation or linear equations. (See the book of Lions [31]).

Remark 1.4. We have restricted ourselevs to systems governed by a linear homogeneous boundary problem of Neumann type with distributed control. One can treat in a similar way the systems governed by other homogeneous or inhomogeneous boundery calue problems; for instance, problems of Dirichlet type, mixed case we necessarily have inhomogeneous problems.

Remark 1.5. In practice, the operator C is of the form αI where $\alpha > 0$ is a small number.

1.2 Duality and Existence

We shall show that there exists a unique of the optimal control problem (1.6). We make use the existence of saddle point via the theorem of *Ky* Fan and Sion (Theorem 1.2 of Chapter 5) for this purpose. This 198 also enables us to characterize the solution of the optimal control problem (1.6). As in the earlier chapters we also obtain the dual problem govergned by the adjoint state equation.

We consider the optimal control problem as a minimization problem for this purpose and we duality in the vaiable y keeping u fixed in K.

We take for the cone Λ the space $V = H^1(\Omega)$ it self define the functional

$$\Phi : V \times \Lambda \to \mathbb{R} \tag{1.7}$$

by setting

$$\Phi(y, u, q) = a(y, q) - (f + u, q). \tag*{(1.7)'}$$

It is clear that Φ is homogeneous of degree in q:

$$\Phi(y, u = \lambda q) = \lambda\Phi(y, u[q) \text{ for all } \lambda > 0.$$

Next $\Phi(y, u; q) \leq 0$ for all $q\epsilon\Lambda$ if and only if $u \in K$ and y, u are related by the state equation (1.2). In fact, the state equation implies that $\Phi(y, u; q) = 0$. Conversely, $\Phi(y, u; q) \leq 0$ implies that $u\epsilon K$ and y, u are related by the state equation. For, we have

$$a(y, q) - (f + u, q) \leq \text{ for all } q\epsilon\Lambda$$

and since, for any $q\epsilon\Lambda$, $-q\epsilon\Lambda$ also we have

$$a(y, -q) - (f + u, -q) \leq 0.$$

The two inequalities together imply that

$$a(y, q) = (f + u, 1) \text{ for all } q\epsilon\Lambda = V = H^1(\Lambda),$$

199 which means that $y = y_u = u(u)$. We introduce the Lagrangian $\mathscr{L}$ associated to the minimization problem by setting

$$\mathscr{L}(z, v; q) = J(z, v) + \Phi(z, v; q). \tag{1.8}$$

More explicitly we have

(1.8)′
$$\begin{cases} \mathscr{L}(z, v; q) = \frac{1}{2}b(z - y_g, z - y_g) + \frac{1}{2}(cz, z) + a(z, q) - (f + v, q) \\ \text{for } z\epsilon V, v\epsilon K \text{ and } q\epsilon\Lambda = V. \end{cases}$$

We shall now prove the following theorem:

Theorem 1.1. *There exists a saddle point for $\mathscr{L}(z, v; q)$ in $V \times K \times V$.*

In other words,

(1.9)
$$\begin{cases} \textit{Theorem exists } (y, u; p)\epsilon V \times K \times V \textit{ such that} \\ \mathscr{L}(y, u; q) \leq \mathscr{L}(y, u; p) \leq \mathscr{L}(z, v; p) \textit{ for all } (z, v; q)\epsilon V \times K \times V. \end{cases}$$

Proof. The proof will be carried out in several steps.

Step 1. (Application of the theorem of Ky Fan and Sion). Let $\ell > 0$ be a constant which we shall choose suitably later. Consider the two sets

$$(1.10) \qquad \begin{cases} \Lambda_\ell = U_\ell = \{z | z \epsilon V = H^1(\Omega); |||z||| \leq \ell\} \text{ and} \\ K_\ell = \{v | v \epsilon K : ||v|| \leq \ell\}. \end{cases}$$

It is clear that $\Lambda_\ell = U_\ell$ is a closed convex and bounded set in V. Since K is closed and convex K_ℓ is also a closed convex subset of $L^2(\Omega)$. Hence, for the weak topologies V and $L^2(\Omega)$ are Hausdorff topological vector spaces in which U_ℓ, (respectively K_ℓ) is compact.

On the other hand, since, for every $(z, v) \epsilon U_\ell \times K_\ell$, the functional 200

$$U_\ell \ni q \mapsto \mathscr{L}(z, v : q) \epsilon \mathbb{R}$$

is linear and strongly (and hence also for the weak topology on V) continuous it is concave and upper semi-continuous (for the weak topology on V). The mapping

$$U_\ell \times K_\ell \ni (z, v) \mapsto \mathscr{L}(z, v; q) \epsilon \mathbb{R}$$

is strongly continuous and hence, in particular, (weakly) lower semi-continuous for every fixed $q \epsilon \Lambda_\ell = K_\ell$. Since the bilinear forms $a(\cdot, \cdot)$, $b(\cdot, \cdot)$ on V and $(C\cdot, \cdot)$ on $L^2(\Omega)$ are positive semi-definite and $v \mapsto (v, q)$ is linear it follows from the results of Chapter 1 § 3 that the mapping

$$(z, v) \mapsto \mathscr{L}(, v; q)$$

is convex.

Thus all the hypothesis of the theorem of Ky Fan and Sion (Theorem 1.2 of Chapter 5) are satisfied. Hence there exists a saddle point $(y_\ell, u_\ell; p_\ell) \epsilon U_\ell \times K_\ell \times U_\ell$ This is the same as saying

$$(1.11)' \qquad \begin{cases} \text{there exists } (y_\ell, u_\ell; p_\ell) \epsilon U_\ell \times K_\ell \times \Lambda_\ell \text{ such that} \\ J(y_\ell, u_\ell) + \Phi(y_\ell, u_\ell; q) \leq J(y_\ell, u_\ell) + \Phi(y_\ell, u_\ell; p_\ell) \\ \qquad \leq J(z, v) + \Phi(z, v : p_\ell) \\ \text{for all } (z, v; q) \epsilon U_\ell \times K_\ell \times \Lambda_\ell. \end{cases}$$

Choosing $\ell > 0$ sufficiently large we shall show, in the following steps that y_ℓ, u_ℓ, p_ℓ are bounded independent of the choice of such an ℓ.

Step 2. u_ℓ is bounded. In fact, the second inequality in ((1.11)$'$) means that the functional

$$(z, v) \mapsto \mathscr{L}(z, v; p_\ell)$$

201 on $U_\ell \times K_\ell$ attains a local minimum at (y_ℓ, u_ℓ). But since this functional is convex, by Lamma 2.1 of Chapter 2, it is also a global minimum. i.e. We have

$$(1.12) \qquad \begin{cases} \mathscr{L}(y_\ell, u_\ell, q_\ell) \leq \mathscr{L}(y_\ell, u_\ell; p_\ell) \leq \mathscr{L}(z, v; p_\ell) \\ \text{for all } z\epsilon V, v\epsilon K \text{ and } q\epsilon\Lambda_\ell = U_\ell. \end{cases}$$

Now we fix a $v\epsilon K$ arbitrarily and take $q = 0, z = y_v$ in $(1.11)'$ and we obtain

$$J(y_\ell, u_\ell) \leq J(y_\ell, u_\ell) + \Phi(y_\ell, u_\ell; p_\ell) \leq J(y_v, v) \equiv j(v).$$

It follows from this that, for any fixed $v\epsilon K$, we have

$$(1.13) \qquad \Phi(y_\ell, u_\ell, p_\ell) \geq 0 \text{ and } J(y_\ell, u_\ell) \leq J(v).$$

But by (1.3), (1.4) the latter inequality in (1.13) implies that

$$\frac{1}{2}\alpha_C \|u_\ell\|^2 \leq J(y_\ell, u_\ell) \leq j(v).$$

which means that u_ℓ is bounded:

$$(1.14) \qquad \|u_\ell\| \leq c_1, c_1^2 = 2\alpha_C^{-1} j(v).$$

Step 3. y_ℓ is bounded. As before we fix a $v\epsilon K$ and take $z = y_v$, $q = \ell |||y_\ell|||^{-1} \epsilon U_\ell = \Lambda_\ell$ in $(1.11)'$. (We may assume that $y_\ell \neq 0$, for otherwise there is nothing to prove). We get

$$J(y_\ell, u_\ell) + \ell |||y_\ell|||^{-1} \Phi(y_\ell, u_\ell, y_\ell) \leq j(v)$$

because of the homogeneity of Φ in the last argument. Here $J(y_\ell, u_\ell) \geq 0$ because of (1.3), (1.4) and (1.5) so that we get

$$\ell |||y_\ell|||^{-1} \Phi(y_\ell, u_\ell; y_\ell) \leq j(v).$$

$$\text{i.e.} \qquad \ell |||y_\ell|||^{-1}\{a(y_\ell, y_\ell) - (f + u_\ell, y_\ell)\} \leq j(v).$$

202 By the coercivity (1.1) of $a(\cdot,\cdot)$ on V we have

$$\alpha_a |||y_\ell|||^2 \leq a(y_\ell, y_\ell)$$

and by the Cauchy-Schwarz inequality we have

$$|(f + u_\ell, y_\ell)| \leq ||f + u_\ell|| |y_\ell|| \leq (||f|| + ||u_\ell||) |||y_\ell|||.$$

Hence using (1.14)

$$\begin{aligned} \ell \alpha_a |||y_\ell||| &\leq j(v) + \ell |||y_\ell|||^{-1}(f + u_\ell, y_\ell) \\ &\leq j(v) + \ell(||f|| + ||u_\ell||) \\ &\leq j(v) + \ell(||f|| + C_1) \end{aligned}$$

so that, first by dividing by ℓ, we see that if $\ell > 1$ then

$$(1.15) \qquad |||y_\ell||| \leq \alpha_a^{-1}(j(v) + ||f|| + C_1) \equiv C_2.$$

Step 4. p_ℓ is bounded. For this we recall that, as has already been observed, (y_ℓ, u_ℓ) is a global minimum for the convex functional

$$g : (z, v) \mapsto \mathscr{L}(z, v; p_\ell)$$

on $V \times K$. Hence, by Theorem 2. 1.3, the G-derivative of g at (y_ℓ, u_ℓ) should vanish:

$$g'(y_\ell, u_\ell; \varphi, v) = 0 \text{ for all } (\varphi, v) \epsilon V \times K.$$

This on calculation of the derivative gives

$$\begin{cases} b(y_\ell - y_g, \varphi) + (Cu_\ell, v) + a(\varphi, p_\ell) - (f + u_\ell, \varphi) = 0 \\ \text{for all } (\varphi, v) \epsilon V \times K. \end{cases}$$

Taking $\varphi = p_\ell$ and $v = u_\ell$ we get 203

$$(Cu_\ell, u_\ell) + a(p_\ell, p_\ell) = (f + u_\ell, p_\ell) - b(y_\ell - y_g, p_\ell).$$

Using the coercivity of the terms on the left side and Cauchy - Schwarz inequality for the first term on the right side together with the continuity of $b(\cdot, \cdot)$ we find that

$$\begin{aligned}\alpha_a|||p_\ell|||^2 \leq \alpha_C||u_\ell||^2 + \alpha_a|||p_\ell|||^2 &\leq ||f + u_\ell||||p_\ell|| + M_b|||y_\ell - y_g||||||p_\ell|||\\ &\leq (||f|| + ||u_\ell|| + M_b|||y_\ell - y_g|||)|||p_\ell|||\\ &(||f|| + c_1 + M_b c_2 + M_b|||y_g|||)|||p_\ell|||\end{aligned}$$

which implies that there exists a constant $c_3 > 0$ such that

$$|||p_\ell||| \leq c_3. \tag{1.16}$$

Step 5. We now choose $\ell > \max(c_1, c_2, 2c_3, 1)$ and use the sets U_ℓ and K_ℓ for the application of the theorem of Ky Fan and Sion.

Step 6. To show that $y_\ell = y_{u_\ell}$ (i.e. y_ℓ is the solution of the state equation corresponding to the control $u_\ell \epsilon K$.) For this purpose we have to show that

$$\Phi(y_\ell, u_\ell; q) = 0 \text{ for all } q\epsilon\Lambda = V \tag{1.17}$$

We already know from (1.13) that $\Phi(y_\ell, u_\ell; p_\ell) \geq 0$. Since $q = 2p_\ell\epsilon\Lambda = V$ satisfies

$$|||q||| = 2|||p_\ell||| \leq 2c_3 < \ell$$

we can take $q = 2p_\ell$ in the first inequality of (1.11)′ and get

$$2\Phi(y_\ell, u_\ell; p_\ell) \leq \Phi(y_\ell, u_\ell, p_\ell).$$

so that we also have

$$\Phi(y_\ell, u_\ell; p_\ell) \leq 0 \tag{1.18}$$

204

Then it follows once again from the first inequality of (1.11)′ that

$$\Phi(y_\ell, u_\ell; q) \leq 0 \text{ for all } q\epsilon\Lambda_\ell = U_\ell \tag{1.19}$$

If $q \notin U_\ell$ then $\ell |||q|||^{-1} q \epsilon U_\ell$ which on substituting in (1.19) gives (1.17).

Finally, combining the facts (1.12) and (1.17) together with the definition of $\mathscr{L}(z, v; q)$ we conclude the there exists a saddle point $(y, u; p)$ in $V \times K \times V$. This completes the proof of the theorem.

The theoem (1.1) implies that (y, u) is the solution of the primal problem and p is the solution of the dual problem. The equation (1.17) is nothing but the fact that y is the solution y_u of the state equation.

From the above theorem we obtain the main result on existence (and uniqueness) of the solution to the optimal control problem and also a characterization of this solution. For this purpose, if we choose $v = u$ in the second inequality of (1.9) we find that $y \epsilon V$ is the minimum of the convex functional

$$h : V \ni z \mapsto \mathscr{L}(z, u; p) \epsilon \mathbb{R}.$$

Hence taking the G-derivative of h we should have

$$h'(u, \psi) = b(y - y_g, \psi) + a(\psi, p) = 0 \text{ for all } \psi \epsilon V.$$

Thus we see that p satisfies the equation

$$(1.20) \qquad a(\psi, p) = -b(y - y_g, \psi) \text{ for all } \psi \epsilon V.$$

The equation (1.20) is thus the adjoint state equation in the present problem. Again, in view of the hypothesis (1.1) and (1.3) it follows (by the Lax-Milgram lemma) that, for any given $y \epsilon V$, there exists a unique solution $p \epsilon V$ of the wquation (1.20). 205

Now consider the functional

$$k : K \ni v \mapsto \mathscr{L}(y, v; p) \epsilon \mathbb{R}.$$

The secone inequality in (1.9) with $z = y$ implies that this functional k is minimum at $v = u$. Again taking G-derivatives we have

$$k'(v, w) = (Cv, w) - (w, p) \text{ for all } w \epsilon K.$$

The solution of the minimization problem for k on K is, by theorem 2.2 of Chapter 2, characterized by

$$\begin{cases} u\epsilon K \text{ such that} \\ k'(u, v-u) \geq 0 \text{ for all } v\epsilon K, \end{cases}$$

which is the same as the variational inequality

$$(1.21) \qquad \begin{cases} \qquad\qquad\qquad\qquad u\epsilon K \text{ such that} \\ (Cu, v-u) - (p, v-u) \geq 0 \text{ for all } v\epsilon K. \end{cases}$$

□

The above facts can now be summarized as follows:

Theorem 1.2. *Suppose given the set K of controls, the state equation (1.2) and the cost function J defined by (1.5) such that the hypothesis (1.1), (1.3) and (1.4) are satisfied. Then we have the following:*

(i) The optimal control problem (1.6) has a unique solution $u\epsilon K$.

(ii) The unique solution u of the optimal control problem us characterized by the coupled system consisting of the pair of equations (1.2) and (1.20) defining the state y and the adjoint state p governing the system together with the variational inequality (1.21).

(iii) A solution $(y, u; p)$ to (1.2), (1.20) and (1.21) exists (and is unique) and is the unique saddle point of the Lagrangian $\mathscr{L}$ defined by 206 *(1.8).*

Remark 1.6. The variational inequality (1.21) is nothing but the well known maximum principle of Pontrjagin in the classical theory of controls.

1.3 Elimination of State

In order to obtain algorithm for the construction of approximations to the solution of the optimal control problem (1.6) we use the characterization given by Theorem (1.2) (ii) to obtain a pure minimization problem with constraints. This is achieved by eliminating the state y_u which occurs explicitly in the above characterization.

We can rewrite the problem of control (1.6) in terms of the operators defined on V by the bilinear forms $a(\cdot,\cdot)$ and $b(\cdot,\cdot)$ and the operator defined by the inclusion mapping of $V = H^1(\Omega)$ in $L^2(\Omega)$.

In fact, for any fixed $y\epsilon V$, the linear form

$$\varphi \mapsto a(y,\varphi)$$

is continuous linear on V by (1.1) and hence by Riesz-representation theorem there exists a unique element $Ay\epsilon V$ such that

$$a(y,\varphi) = ((Ay,\varphi)) \text{ for all } \varphi\epsilon V. \tag{1.22}$$

Once again from (1.1) the mapping $y \mapsto Ay$ is a continuous linear operator on V. Similarly, by (1.3) there exists a continuous linear operator B on V such that

$$b(y,\varphi) = ((By,\varphi)) \text{ for all } \varphi\epsilon V. \tag{1.23}$$

Finally since the inclusion mapping of V in $L^2(\Omega)$ is continuous linear it follows that for any $u\epsilon L^2(\Omega)$ the linear mapping $v \mapsto (u,v)$ on V 207 is a continuous linear functional. Hence there exists a continuous linear operator $D : L^2(\Omega) \to V$ such that

$$(u,v) = ((Du,v)) \text{ for all } u\epsilon L^2(\Omega), v\epsilon V. \tag{1.24}$$

The state equation can now be written as

$$((Ay,\varphi)) = ((Df + Du,\varphi)) \text{ for all } \varphi\epsilon V.$$

which is the same as the operational equation in V:

$$Ay = Df + Du. \tag{1.25}$$

In view of the well known result of Lax and Miligram we have

Theorem 1.3. *Under the hypothesis (1.1) the state equation (1.2) (or equilvalently (1.25)) has a unique solution $y_u\epsilon V$ for any given $u\epsilon L^2(\Omega)$ and there exists constant $c > 0$ such that*

$$|||y||| \leq c(|||Df||| + |||Du|||). \tag{1.26}$$

This is equivalent to saying that the operator A is invertible, A^{-1} is a continuous linear operator on V and (1.26) gives an estimate for the norm of A^{-1}. Hence we can write

$$y_u = A^{-1}(Df + Du) \tag{1.27}$$

as the solution of the state equation.

Next we shall reduce the optimal control problem (1.6) to a minimization problem as follows. We substitute y_u given by (1.27) in the cost function (1.5) and thus we eliminate the state from the functional to minimize. Using (1.23) together with (1.27) we can write

$$\begin{aligned} b(y_u - y_g, y_u - y_g) &= ((B(y_u - y_g), y_u - y_g)) \\ &= ((B[A^{-1}(Df + Du) - y_g], A^{-1}(Df + Du) - y_g)) \\ &= ((BA^{-1}Du, A^{-1}Du)) + 2((B(A^{-1}Df - y_g), A^{-1}Du)) \\ &\quad + ((B(A^{-1}Df - y_g), A^{-1}Df - y_g)) \\ &= ((A^{-1*}BA^{-1}Du, Du)) + 2((A^{-1*}B(A^{-1}Df - y_g), \\ &\qquad Du)) + G(f, y_g) \end{aligned}$$

208 where A^{-1*} is the adjoint of the operator A^{-1} and $G(f, y_g)$ denoted the functional

$$G(f, y_g) = ((BA^{-1}Df - By_g, A^{-1}Df - y_g))$$

which is independent of u. Once again using (1.24) we can write

$$b(y_u - y_g, y_u - y_g) = (A^{-1*}BA^{-1}Du, u) + (A^{-1*}B(A^{-1}Df - y_g), u) + G(f, y_g)$$

and hence the cost function can be written in the form

$$j(u) = \frac{1}{2}(A^{-1*}BA^{-1}Du, u) + (A^{-1*}B(A^{-1}Df - y_g), u) + G(f, y_g) + \frac{1}{2}(Cu, u).$$

Setting

$$\begin{cases} \mathscr{A} = A^{-1*}BA^{-1}D + C \text{ and} \\ \mathscr{F} = A^{-1*}B(A^{-1}Df - y_g) \end{cases} \tag{1.28}$$

We have the following

Proposition 1.1. *The optimal control problem (1.6) is equivalent to the minimization problem:*

$$(1.29)\qquad \begin{cases} \textit{to find } u\epsilon K \textit{ such that} \\ j(u) = \inf_{v\epsilon K} j(v) \textit{ where} \\ j(v) = \frac{1}{2}(\mathscr{A}v, v) - (\mathscr{F}, v) + G(f, y_g). \end{cases}$$

We observe that, since the last term in the expression for the quadratic functional $j(v)$ is a constant (independent of v), $u\epsilon K$ is a solution of (1.29) if and only if u is a solution of the minimization problem: 209

$$(1.30)\qquad \begin{cases} \text{to find } u\epsilon K \text{ such that} \\ k(u) = \inf_{v\epsilon K} k(v) \text{ where} \\ k(v) = \frac{1}{2}(\mathscr{A}v, v) - (\mathscr{F}, v). \end{cases}$$

We know by the results of Chapter 2 § 3 (Theorem 3.1) that the problem (1.30) has a unique solution and it is characterized by the condition

$$k'(u, v-u) \geq 0 \text{ for all } v\epsilon K,$$

where $k(\cdot, \varphi)$ denotes the G-derivative of $k(\cdot)$. This is nothing but the variational inequality

$$(1.31)\qquad \begin{cases} \text{To find } u\epsilon K \text{ such that} \\ (\mathscr{A}u - \mathscr{F}, v-u) \geq 0 \text{ for all } v\epsilon K. \end{cases}$$

This variational inequality (1.31) together with the state equation is an equivalent formulation of the characterization of the optimal control problem given by Theorem (1.2) (ii). More precisely, we have the following

Theorem 1.4. *The solution of the optimal control problem (1.6) is characterized by the variational inequality:*

$$(1.32)\qquad \begin{cases} \textit{To find } u\epsilon K \textit{ such that} \\ (Cu - p_u, v-u) \geq 0 \textit{ for all } v\epsilon K \end{cases}$$

where p_u is the adjoint state.

Proof. We have by the definitions (1.28) of $\mathscr{A}$ and $\mathscr{F}$ 210

$$\mathscr{A}u - \mathscr{F} = A^{-1*}B(A^{-1}Du + A^{-1}Df - y_g) + Cu$$

which on using the state equation (1.25) becomes

$$\mathscr{A}u - \mathscr{F} = A^{-1*}B(y_u - y_g) + Cu. \tag{1.33}$$

□

If we now define p_u by setting

$$-p_u = A^{-1*}B(y_u - y_g) \tag{1.34}$$

then we see that p_u satisfies the functional equation

$$((A^* p_u, \psi)) = -((B(y_u - y_g), \psi)) \text{ for all } \psi \epsilon V.$$

We notice that this is nothing but the adjoint state equation:

$$a(\psi, p_u) = -b(y_u - y_g, \psi) \text{ for all } \psi \epsilon V.$$

Thus if, for a given control $u \epsilon K, y_u$ is the solution of the state equation then p_u defined by (1.34) is the solution of the adjoint state equation. Moreover, we have

$$\mathscr{A}u - \mathscr{F} = Cu - p_u. \tag{1.33$'$}$$

Substituting (1.33)′ in the variational inequality (1.31) we obtain the assertion of the theorem.

We are thus reduced to a pure minimization problem in K for which we have known algorithms.

1.4 Approximation

The formulation of the optimal control problem as a pure minimization problem given above in Section (1.3) together with the algorithms 211 described in earlier chapters for the minimization problem will immediately lead to algorithm to determine approximations to the solution of the optimal control problem (1.6). Hence we shall only mention this briefly in the following.

We observe first of all that the operator $\mathscr{A}$ is $L^2(\Omega)$-coercive and bounded. In fact, in view of (1.24) and (1.23) we can write

$$\begin{aligned}(A^{-1*}BA^{-1}Du, u) &= ((A^{-1*}BA^{-1}Du, Du))\\ &= (BA^{-1}Du, A^{-1}Du) = b(A^{-1}Du, A^{-1}Du) \geq 0.\end{aligned}$$

Since we also have $(Cu, u) \geq \alpha_C \|u\|^2$ we find that $\mathscr{A}$ is $L^2(\Omega)$-coercive and

$$(\mathscr{A}u, u) = (A^{-1*}Ba^{-1}Du6Cu, u) \geq \alpha_C \|u\|^2, u \epsilon V. \tag{1.35}$$

To prove that is bounded we note that A^{-1} is the operator

$$L^2(\Omega) \ni f + u \mapsto y_u \epsilon L^2(\Omega)$$

defining the solution of the state equation:

$$\begin{cases} y_u \epsilon V \text{ such that} \\ a(y_u, \varphi) = ((Ay_u, \varphi)) = ((D(f+u), \varphi)) \text{ for all } \varphi \epsilon V. \end{cases}$$

Here taking $\varphi = y_u$ and using the coercivity of the bilinear form $a(\cdot, \cdot)$ we see that

$$\alpha_a |||y_u|||^2 \leq |||Df + Du||||||y_u|||$$

and hence

$$|||y_u||| \leq |||A^{-1}(Df + Du)||| \leq \alpha_a^{-1} |||Df + Du|||.$$

which implies that A^{-1} is bounded and in fact, we have

$$\|A^{-1}\|_{\mathscr{L}(V,V)} \leq \alpha_a^{-1}. \tag{1.36}$$

Now since all the operators involved in the definition of $\mathscr{A}$ are linear and bounded it follows that $\mathscr{A}$ is also bounded. Moreover, we also have

$$\begin{aligned} \|\mathscr{A}\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))} &= \|A^{-1*}BA^{-1}D + C\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))} \\ \leq \|A^{-1}\|^2_{\mathscr{L}(V,V)}\|B\|_{\mathscr{L}(V,V)}\|\|\|D\|_{\mathscr{L}(L^2(\Omega),V)} &+ \|C\|_{\mathscr{L}(L^2\|(\Omega),L^2(\Omega))} \end{aligned}$$

and hence (since $\|D\|_{\mathscr{L}(L^2(\Omega),V)} = 1$)

$$\|\mathscr{A}\|_{\mathscr{L}(L^2(\Omega),L^2(\Omega))} \leq \alpha_a^{-2}M_b + M_c. \tag{1.37}$$

212 We are now in a position to describe the algorithms.

Method of contraction. We recall the the solution of the optimal control problem is equivalent to the solution of the minimization problem (1.29) and that the solution of this is characterized by the variational inequality (1.31):

$$\begin{cases} u \epsilon K \text{ such that} \\ (\mathscr{A}u - \mathscr{F}, v - u) \geq 0 \text{ for all } v \epsilon K. \end{cases}$$

We can now use the method of contraction mapping (as is standard in the proof of existence of solutions of variational inequality - see, for instance, Lions and Stampacchia [ ]) to describe an algorithm for the solution of the variational inequality (1.31).

Algorithm. Suppose we know an algorithm to calculate numerically the projection P of $L^2(\Omega)$ onto K. Let ρ be a constant (which we fix) such that

$$0 < \rho < 2\alpha_C^{-1}/(\alpha_a^{-2}M_b + M_C) = 2\alpha_a^2/\alpha_C(M_b + \alpha_a^2 M_C). \tag{1.38}$$

Let $u_\circ \epsilon K$ be arbitrarily chosen. Suppose $u_\circ, \cdots, u_m$ are determined starting from $u_\circ$. We define u_{m+1} by setting

$$u_{m+1} = P\Phi(u_m) \tag{1.39}$$

where

$$\Phi(u_m) = u_m - \rho\mathscr{A}u_m + \rho^2\mathscr{F} \tag*{(1.39)$'$}$$

We can express $\Phi(u_m)$ in terms of the operators A, B, C and the data f and y_g as follows:

$$\Phi(u_m) = u_m - \rho(A^{-1*}BA^{-1}Du_m + Cu_m) + \rho^2 A^{-1*}B(A^{-1}Df - y_g). \tag{1.40}$$

The choice (1.38) of ρ implies that the mapping

$$T : K \ni w \mapsto P\Phi(w)\epsilon K \tag{1.41}$$

is a contraction, so that T has a fixed point u in K to which the sequence u_m converges. 213

Method of gradient with projection. We consider the minimization problem for the quadratic functional

$$v \mapsto \mathscr{G}(v) = \frac{1}{2}(\mathscr{A}v, v) - (\mathscr{F}, v) \tag{1.42}$$

on K. Since $\mathscr{A}$ is coercive, we can use the method of Chapter 4, Section 3 and we can show that we can choose as convergent choices for $\rho > 0$ a constant and for the direction of descent

$$w_m = grad\mathscr{G}(u_m)/||grad\mathscr{G}(u_m)||. \tag{1.43}$$

Thus starting from an arbitrary $u_\circ \epsilon K$, we define

$$u_{m+1} = P_K(u_m - \rho grad\mathscr{G}(u_m)/||grad\mathscr{G}(u_m)||) \tag{1.44}$$

where P_K is the projection of $L^2(\Omega)$ onto K.

This method, however, requires the computation of $\mathscr{G}(u_m)$ and its gradient at each step. For this purpose, knowing $u_m \epsilon K$ we have to solve the state equation:

$$\begin{cases} y_m \epsilon V \text{ such that} \\ a(y_m, \varphi) = (f + u_m, \varphi) \text{ for all } \varphi\epsilon V \end{cases}$$

to obtain y_m and the adjoint state equation:

$$\begin{cases} p_m \epsilon V \text{ such that} \\ a(\varphi, p_m) = -b(y_m - y_g, \varphi) \text{ for all } \varphi\epsilon V \end{cases}$$

to obtain the adjoint state p_m. We can then calculate grad $\mathscr{G}(u_m)$ by using 214

(1.45) $$grad\mathscr{G}(u_m) = Cu_m - p_m.$$

We shall not go into details of the algorithm which we shall leave to the reader.

Remark 1.7. This method is rather long as it involves several steps for each of which we have sub-algorithms for computations. Hence this procedure may not be very economical.

1.46

As an illustration of the methods described in this section we consider the following two-dimensional optimal control problem: Let Ω be a bounded open set in $\mathbb{R}^2$ with smooth boundary Γ. We consider the following optimal control problem

$$\text{State equation :} \begin{cases} -\triangle y_u + y_u = f + u \text{ in } \Omega \\ \partial y_u/\partial \underline{n} = 0 \text{ on } \Gamma \end{cases}$$

where $\underline{n}$ denotes the exterior normal vector field to Γ

Controal set : $K = \{u\epsilon L^2(\Omega)|0 \leq u(x) \leq 1 \text{ a.e. on } \Omega\}$

Cost function : $J(y,u) = \int_\Omega(|y_u - y_g|^2 + |u|^2)dx.$

We shall leave the description of the algorithm to this problem on the lines suggested in this section as an exercise to the reader.

2 Theory of Optimal Design

In this section we shall be concerned with the problem of optimal design. We shall show that certain free boundary problems can be considered as special cases of this type of optimal design problem. We shall consider a special case of one-dimensional problem and explain a very general method to obtain a solution to the problem, which also enables us to give algorithms to obtain approximations for the solution. This method can be seen to be readily applicable to the higher dimensional problems also 215 except for some technical details. Though there is a certain similarity with the problem of optimal control we cannot use the duality method earlier used in the case as we shall see later.

2.0 Optimal Design

In this section we shall give a general formulation of the problem of optimal design. Once again this problem will be considered as a minimization problem for a suitable class of functionals. As in the case of optimal control problem these functionals are defined through a family of state equations. We shall consider here the states governing the system to be determined by variational elliptic boundary value problems. Though there is some analogy with the optimal control problem studied in the previous section there is an important difference because of the fact in the present case the convex set L (in our case the set K will be the whole of an Hilbert space), on which the given functional is to be minimized, itself is in some sense to be determined, as it is a set of functions on the optimal domian to be determined by the problem. Therefore this problem cannot be treated as an optimal control problem and requires somewhat different techniques than the ones used before.

Roughly speaking the problem of optimal design can be described as follows: Suppose given

(1) A family of possible domians Ω (bounded open sets in the Euclidean space) having certain minimum regularity properties.

(2) A family of elliptic boundary value problems describing the states, one each on a Ω of the family in (1).

(3) A cost function j (described in terms of the state determined by (2) considered as a functional of the domian Ω in the family).

Then the problem consists in finding a domian Ω^* in the given family 216 for which $j(\Omega^*)$ is a minimum.

We shall describe a fairly general theory to obtain a solution to the optimal design problem. In order to simplify the details we shall, however, describe our general method in the special case of one dimension.

Thus the states governing the problem is described by solutions of a two point boundary value problem for a linear second order ordinary differential equation. We shall first describe the main formal steps involved in the reduction of the problem to one of minimization in a fixed domian. We shall then make the necessary hypothesis and show that this formal procedure is justified.

2.1 Formulation of the Problem of Optimal Design

Let $\mathscr{A}$ be a family of bounded open sets Ω in $\mathbb{R}^n$ and let Γ denote the boundary of $\Omega, \Omega\epsilon\mathscr{A}$. We assume that every $\Omega\epsilon\mathscr{A}$ satisfies some regularity properties. For instance, every $\Omega\epsilon\mathscr{A}$ satisfies a cone condition or every $\Omega\epsilon\mathscr{A}$ has a locally Lipschitz boundary etc.

We suppose the following data:

(1) For each $\Omega\epsilon\mathscr{A}$ we are given a bilinear form

$$V \times V \ni (y, \varphi) \mapsto a(\Omega; y, \varphi)\epsilon\mathbb{R}$$

on $V = V_\Omega = H^1(\Omega)$ such that

(i) it is continuous ; i.e. there exists a constant $M_\Omega > 0$ such that

$$(2.1) \quad a(\Omega; \varphi, \psi) \leq M_\Omega ||\varphi||_V ||\psi||_V \text{ for all } \varphi, \psi\epsilon V = H^1(\Omega), \Omega\epsilon\mathscr{A}.$$

(ii) it is $H^1(\Omega)$ -coercive : there exists a constant $C_\Omega > 0$ such that

$$(2.2) \qquad a(\Omega; \varphi, \varphi) \geq C_\Omega ||\varphi||^2_{H^1(\Omega)}, \text{ for all } \Omega\epsilon H^1(\Omega), \Omega\epsilon\mathscr{A}.$$

217 **Example 2.1.** Let $\Omega\epsilon\mathscr{A}$ and

$$a(\Omega; \varphi, \psi) = (\varphi, \psi)_{H^1(\Omega)} = \int_\Omega \left(\sum_j \frac{\partial\varphi}{\partial x_j} \frac{\partial\psi}{\partial x_j} + \varphi\psi \right) dx.$$

(2) For each $\Omega\epsilon\mathscr{A}$ we are given a continuous linear functional $\varphi \mapsto L(\Omega; \varphi)$ on $H^1(\Omega), \Omega\epsilon\mathscr{A}$.

Example 2.2. Let $F \epsilon L^2(\mathbb{R}^n)$ and $f = F|_\Omega$ = restriction of F to Ω.

$$L(\Omega; \varphi) = \int_\Omega f\varphi dx. \text{ for all } \varphi \epsilon H^1(\Omega), \Omega \epsilon \mathscr{A}.$$

Consider the variational elliptic boundary value problem:

$$(2.3) \qquad \begin{cases} \text{To find } y = y_\Omega \epsilon H^1(\Omega) \text{ such that} \\ a(\Omega; y, \varphi) = L(\Omega; \varphi), (\text{ for all } \varphi \epsilon H^1(\Omega)). \end{cases}$$

We know by Lax-Milgram lemma that under the assumptions (1) and (2) there exists a unique solution $y_\Omega \epsilon H^1(\Omega)$ for this problem (2.3). We observe that since f is given as $F|_\Omega$ this solution y_Ω depends only on the geometry of $\Omega, \Omega \epsilon \mathscr{A}$.

(3) Cost function. For each $\Omega \epsilon \mathscr{A}$ we are given a functional on $H^1(\Omega)$:

$$(2.4) \qquad H^1(\Omega) \ni z \mapsto J(\Omega; z) \epsilon \mathbb{R}$$

Example 2.3.

$$\begin{cases} J(\Omega; z) = \int_\Gamma |z - g|^2 d\sigma, \text{ where} \\ g \epsilon \gamma_\circ G = G|\Gamma, G \epsilon H^1(\Omega), \Omega \epsilon \mathscr{A}. \end{cases}$$

Example 2.4.

$$\begin{cases} J(\Omega;) = \int_\Omega |z - g|^2 dx, \text{ where} \\ G \epsilon L^2(\mathbb{R}^n) \text{ and } g = G|\Omega, \Omega \epsilon \mathscr{A}. \end{cases}$$

2.5 Example of a family $\mathscr{A}$ of domains.

Suppose B and ω are two fixed open subsets of $\mathbb{R}^n$ such that $\overline{\omega} \subset B$. Let A be the family of open sets Ω in $\mathbb{R}^n$ such that $\omega \subset \Omega \subset B$ and Ω satisfies some regularity property (say, for instance, Ω satisfies a cone **218** condition).

Define

$$(2.5) \qquad j(\Omega) = J(\Omega; y_\Omega), \Omega \epsilon \mathscr{A}$$

where y_Ω is the (unique) solution of the homogeneous boundary value problem (2.3).

The problem of optimal design consists in minimizing $j(\Omega)$ over $\mathscr{A}$:

$$\text{(2.6)} \qquad \begin{cases} \text{To find } \Omega^* \epsilon \mathscr{A} \text{ such that} \\ j(\omega^*) = \inf_{\Omega \epsilon \mathscr{A}} j(\Omega). \end{cases}$$

Optimal design and free boundary problem. Certain free boundary problems can be considered as a problem of optimal design as is illustrated by the following example in two dimensions.

Let $\Gamma_\circ$ be a smooth curve in the plane $\mathbb{R}^2$ defined by an equation of the form

$$\text{(2.7)} \qquad z(x) = x_1 - \varphi(x_2) = 0,$$

where $\varphi : I = [0,1] \ni x_2 \mapsto \varphi(x_2) \epsilon \mathbb{R}_+$ is a smooth function. Let Q denote the (open) strip in $\mathbb{R}^2$:

$$\text{(2.8)} \qquad Q = \{x = (x_1, x_2) \epsilon \mathbb{R}^2 | x_1 > 0, 0 < x_2 < 1\}.$$

Consider the open set Ω given by

$$\text{(2.9)} \qquad \Omega = \{x \epsilon Q | z(x) < 0\} \equiv \{x = (x_1, x_2) \epsilon Q | x_1 < \varphi(x_2)\}.$$

The boundary Γ of Ω decomposes into a union $\sum \cup \Gamma_\circ$ with $\sum^\circ \cap \Gamma_\circ^\circ = \phi$.

There exists a one-one correspondence between Ω and the function z, Thus the family $\mathscr{A}$ is determined by the family of smooth functions

$$z : Q \to \mathbb{R}.$$

219 Let us consider the optimal design problem:

(2.10)

$$\begin{cases} a(\Omega; y, \varphi) = (y, \varphi)_{H^1(\Omega)}, \text{ for } y, \varphi \epsilon H^1(\Omega); \\ L(\Omega; \varphi) = (f, \varphi)_{L^2(\Omega)}, \text{ for } \varphi \epsilon H^1(\Omega) \text{ where } f = F|\Omega, F \epsilon L^2(\mathbb{R}^2) \\ J(\Omega; z) = \int_{\Gamma_\circ} |z(x)|^2 d\sigma, \text{ where } d\sigma \text{ is the line element on } \Gamma_\circ. \end{cases}$$

Then $y = y_\Omega$ is the unique solution of the Neumann problem:

$$(2.11) \qquad \begin{cases} y_\Omega \epsilon H^1(\Omega) \\ (y_\Omega, \varphi)_{H^1(\Omega)} = (f, \varphi)_{L^2(\Omega)} \text{ for all } \varphi \epsilon H^1(\Omega) \end{cases}$$

and

$$(2.12) \qquad j(\Omega) = J(\Omega; y_\Omega) = \int_{\Gamma_\circ} |y_\Omega(x)|^2 d\sigma.$$

The optimal design problem then becomes

(2.13)

To find Ω^* such that $j(\Omega^*) \leq j(\Omega)$ for all $\Omega \epsilon \mathscr{A}$ In other words,

$$(2.13)' \qquad \begin{cases} \text{To fin } y_{\Omega^*} \epsilon H^1(\Omega^*) \text{ such that} \\ \int_{\Gamma^*} |y_{\Omega^*}(x)|^2 d\sigma \text{ is minimum} \end{cases}$$

Suppose now that $\inf_{\Omega \epsilon \mathscr{A}} j(\Omega) = j(\Omega^*) = 0$. The it follows that

$$(2.14) \qquad y_{\Omega^*} = 0 \text{ a.e. on } \Gamma^*_\circ$$

In this case, the optimal design problem reduces to the following so called "free boundary problem" :

To find a domian $\Omega^* \epsilon \mathscr{A}$ whose boundary is of the form $\Gamma^* = \sum \cup \Gamma^*_\circ$ where $\sum$ is a fixed curve while $\Gamma^*_\circ$ is a curve determined by the solution of the homogeneous boundary value problem

$$(2.13)'' \qquad \begin{cases} -\triangle y + y = f \text{ in } \Omega^* \\ \partial y / \partial \underline{n} = 0 \text{ on } \sum \\ \partial y / \partial \underline{n} = 0, y = 0 \text{ on } \Gamma^*_\circ. \end{cases}$$

220

This equivalent formulation is obtained in the standard manner from the state equation (2.3) using the Green's formula together with the condition (2.14). Free boundary problems occur naturally in many contexts - for example in theorey of gas dynamics.

2.2 A Simple Example

We shall illustrate our general method to obtain approximations to the solution of the optimal design problem for the following one dimensional problem.

Let $\mathscr{A}$ denote the family of open intervals

$$\Omega_a = (0, a), a \geq 1 \tag{2.15}$$

on the real line.

State equation. Assume that an $f \epsilon L^2(\mathbb{R}^1)$ is given. The state governing the system is a solution of the following problem:

$$\left\{\begin{array}{l} \text{To find } y_{\Omega_a} \epsilon H^1(\Omega_a) \equiv H^1(0, a) \text{ such that} \\ a(\Omega_a; y_{\Omega_a}, \varphi) \equiv \int\limits_0^a \left(\dfrac{dy_{\Omega_a}}{dx}\dfrac{d\varphi}{dx} + y_{\Omega_a}\varphi\right) dx \\ = \int\limits_0^a f\varphi dx \equiv L(\Omega_a; \varphi), \text{ for all } \varphi \epsilon H^1(\Omega_a). \end{array}\right. \tag{2.16}$$

On integration by parts (or more generally, using the Green's formula) we see that this is nothing but the variational formulation of the two pointy boundary value problem (of Neumann type boundary value problem):

$$\left\{\begin{array}{l} \text{To find } y_{\Omega_a} \epsilon H^1(\Omega_a) \text{ satisfying} \\ \dfrac{d^2 y_{\Omega_a}}{dx^2} + y_{\Omega_a} = f \text{ in } \Omega_a \\ \dfrac{dy_{\Omega_a}}{dx}(0) = 0 = \dfrac{dy_{\Omega_a}}{dx}(a) \end{array}\right. \tag{2.16$'$}$$

221 **Cost function.** Suppose given a $g \epsilon L^2(0, 1)$. Define

$$j(a) = \frac{1}{2}\int_0^1 |y_{\Omega_a} - g|^2 dx. \tag{2.17}$$

Problem of optimal design.

$$\left\{\begin{array}{l} \text{To find } a^* \geq 1 (\text{ i.e. to find } \Omega^* = \Omega_{a^*}) \text{ such that} \\ j(a^*) \leq j(a) \text{ for all } a \geq 1. \end{array}\right. \tag{2.18}$$

Remark 2.1. It appears natural to consider a as the control variable and use the duality argument as we did in the case of the optimal control problem. However, since the space $V = H^1(\Omega_a)$ varies with a the duality method may not be useful to device algorithms.

In what follows, we shall adopt the following notation to simplify the writing:

$$\text{(2.19)} \qquad \begin{cases} y_{\Omega_a}(x) = y(a,x) \\ \partial y/\partial x(a,x) = y'(a,x) \\ \partial y/\partial a(a,x) = y_a(a,x) \end{cases}$$

2.3 Computation of the Derivative of j.

We shall use the method of gradient to obtain algorithms to construct approximations converging to the required solution of the problem (2.18). In order to be able to apply the gradient method we make the formal 222 computation of the gradient of j (in the present case, the derivative of j) with respect to a in this section. We justify the various steps involved under suitable hypothesis in the next section.

Settinf for $\varphi \epsilon H^1(\Omega_a)$

$$\text{(2.20)} \qquad F(a,x) = y'(a,x)\varphi'(a,x) + y(a,x)\varphi(a,x) - f(a,x)\varphi(a,x)$$

we can write the state equation (2.16) as

$$\text{(2.21)} \qquad K(a) = \int_0^a F(a,x)dx = 0$$

Here since we have a Neumann type boundary value problem for a second order ordinary differential operator the test function φ belongs to $H^1(\Omega_a)$ and so φ is defined in a variable domian $\Omega_a = (0,a)$. This may cause certain inconveniences, which however can easily be overcome be overcome as follows:

(1) We can take φ to be the restriction to Ω_a of a function $\psi \epsilon H^1$ $(0,+\infty)$ and write the state equation as

$$\int_0^a \{y'(a,x)\psi'(x) + y(a,x)\psi(x) - f(a,x)\psi(x)\} = 0, \text{ for } \psi \epsilon H^1(0,+\infty).$$

Such a choice for the test functions $\varphi \epsilon H^1(\Omega_a)$ would suffice when the state is described by a Neumann type problem (as we have in the present case.) But if the boundary conditions are of Dirichlet type this choice is not suitable since the restrictions of functions in $H^1(0, +\infty)$ to Ω_a do not necessarily give functions in the space of test functions $H^1_\circ(\Omega_a)$. We can use another method in which such a problem do not arise and we shall use this method.

(2) Suppose $\psi \epsilon H^m(\Omega_1), \Omega_1(0, 1)$ and $m \geq 2$. Then the function $x \mapsto \varphi(a, x)$ defined by

$$\varphi(a, x) = \psi(x/a) \tag{2.22}$$

223 is well defined in Ω_a and belongs to $H^m(\Omega_a) \hookrightarrow H^1(\Omega_a)$. (This inclusion, we note is a dense inclusion.) We also note that, in this case, if $\Psi \epsilon H^m_\circ(\Omega_1)$ then $\varphi \epsilon H^m_\circ(\Omega_a)$ and conversely.

Thus we set

$$(2.20)' \quad F(a, x) = y'(a, x)\psi(x/a) + (y(a, x) - f(x))\psi(x/a) \text{ for } \psi \epsilon H^m(\Omega_1)$$

and we can write the state equation with this F as

$$K(a) = \int_0^a F(a, x) dx = 0 \tag{2.21}$$

We shall make use of the following classical result to calculate the derivative dK/da.

Let Λ denote the closed subset of the (x, a)-plane:

$$\Lambda = \{(x, a) \epsilon \mathbb{R}^2; a \geq 1 \text{ and } 0 \leq x \leq a\}. \tag{2.23}$$

Suppose $F : \Lambda \to \mathbb{R}$ be a function satisfying:

Hypothesis (1). For every $a \geq 1$, the real valued function

$$x \mapsto F(a, x)$$

is continuous in $0 \leq x \leq a$.

Hypothesis (2). For every $x\epsilon[0, a]$, the function

$$a \mapsto F(a, x)$$

is differentiable and $\partial F/\partial a : \Lambda \to \mathbb{R}$ is continuous. Then the integral

$$K(a) = \int_0^a F(a, x) dx$$

exists, $a \mapsto K(a)$ belongs to $C^1(1 \leq a < +\infty)$ and we have

$$\frac{dK}{da}(a) = \int_0^a \partial F/\partial a(a, x) dx + F(a, a) \tag{2.24}$$

224

Remark 2.2. We observe that this classical result has a complete analogue also in higher dimensions and we have a similar identity for $grad_a K$ (with respect to a) in place of dK/da.

Now differentiating the equation (2.20)′ with respect to a and using the above result we get

$$\begin{aligned} dK/da(a) &= \int_0^a \partial F/\partial a(a, x) dx + F(a, a) \\ &= \int_0^a [\{y'_a(a, x)\psi'(x/a) + y_a(a, x)\psi(x/a)\} + \\ &\quad + \{y'(a, x)(\psi'(x/a))_a + y(a, x)(\psi(x/a))_a - f(x)(\psi(x/a))_a\}]dx \\ &\quad + [y'(a, x)\Psi'(x/a) + y(a, x)\psi(x/a) - f(x)\psi(x/a)]_{x=a} = 0. \end{aligned}$$

We observe that, if $m \geq 2$ then $x \mapsto (\psi(x/a))_a \epsilon H^1(0, a)$. In fact,

$$\begin{aligned} &(\psi(x/a))_a = (-x/a^2)\psi'(x/a)\epsilon L^2(\Omega_a), \\ &(\psi(x/a))'_a = (-1/a^2)\psi'(x/a) + (-x/a^3)\psi''(x/a)\epsilon L^2(\Omega_a). \end{aligned}$$

where ψ' and ψ'' are (strong) L^2 -derivatives of ψ, which exist since $\psi\epsilon H^2(0, 1)$.

Hence by the state equation (2.16) we find that

$$\int_0^a \{y'(a, x)(\psi'(x/a))_a + y(a, x)(\psi(x/a))_a - f(x)(\psi(x/a))_a\} dx$$

$$= a(\Omega_a; y_{\Omega_a})(\psi(x/a))_a - L(\Omega_a; (\psi(x/a))_a) = 0$$

Thus we conclude that

$$\int_0^a \{y'_a(a,x)\Psi'(x/a) + y_a(a,x)\psi(x/a)\}dx$$

(2.25) $$= -[y'(a,x)\psi'(x/a) + y(a,x)\psi(x/a) - f(x)\psi(x/a)]_{x=a},$$

for all $\psi \epsilon H^m(0,1)$ with $m \geq 2$.

225 **Remark 2.3.** It is obvious that the above argument easily carries over to dimensions ≥ 2 of rhte computation of $grad_a K(a)$.

Finally, we calculate the derivative of the cost function j with respect to a and we have

$$dj/da = \frac{1}{2} d/da \int_0^1 |y(a,x) - g(x)|^2 dx$$

(2.26) $$= \int_0^1 (y(a,x) - g(x)) y_a(a,x) dx.$$

In (2.26) we eliminate the derivative y_a of the state y_{Ω_a} using the adjoint state equation. The adjoint state $p_{\Omega_a} = p(a,x)$ is the solution of the equation:

(2.27)
$$\begin{cases} \int_0^a \{\varphi'(x)p'(a,x) + \varphi(x)p(a,x)\}dx = \int_0^1 (y(a,x) - g(x))\varphi(x)dx, \\ \text{for all } \varphi \epsilon H^1(0,a). \end{cases}$$

If we know that $y(a,x)$ is sufficiently regular, for instance say, $y_a \epsilon H^1$ (Ω_a) then taking $\varphi = y_a(a,x)$ in the adjoint state equation (2.27) above we obtain

$$dj/da = \int_0^1 (y(a,x) - g(x)) y_a(a,x) \qquad \text{bf (2.26)}$$

$$= \int_0^a \{y'_a(a,x)p'(a,x) + y_a(a,x)p(a,x)\}dx \qquad \text{bf (2.27)}$$

This together with (2.25) for $\psi = p$ gives

(2.28) $$dj/da = -[y'(a,x)p'(a,x) + y(a,x)p(a,x) - f(x)p(a,x)]_{x=a}$$

2.4 Hypothesis and Results

In the calculation of the derivatives of the cost function $j(a)$ in the previous section we have made use of the regularity properties of the state $y_{\Omega_a} = y(a,x)$ as well as that of the adjoint state $p_{\Omega_a} = p(a,x)$ with respect to both the variables x and a. This in turn implies the regularity 226
of the function $F(a,x)$ define by (2.20)′ which is required for the validity of the theorem on differentiation of the integral $K(a)$ of $F(a,x)$. The regularity of $y(a,x)$. The regularity of $y(a,x)$ and $p(a,x)$ are again necessary in order that the expression on the right side of (2.28) for the derivative value problmes for (ordinary) differential equation, the regularity of y and p as a consequence of suitable hypothesis on the data f and g.

We begin with the following assumptions on the data:

Hypothesis (3). For all $a \geq 1, t \mapsto f(at) \epsilon H^1(0,1)$.

Hypothesis (4). $g \epsilon H^1(0,1)$.

Then we have the following

Proposition 2.1. (Existence of the derivatives y_a and y'_a). *Under the hypothesis (3) on f, if $y(a,x)$ is the solution of the state equation (2.16) then*

(i) $x \mapsto y(a,x) \epsilon H^3(0,a)$

(ii) y_a *exists and* $x \mapsto y_a(a,x) \epsilon H^2(0,a)$ *and as a consequence we have*

(iii) $x \mapsto y(a,x) \epsilon C^2([0,a])$ *and*

$x \mapsto y_a(a,x) \epsilon C^1([0,a])$.

Proof. By a change of variable of the form

$$x = at, x \epsilon (0,a) \text{ and } t \epsilon (0,1) \tag{2.29}$$

we can transform the state equation (2.16) to a two point boundary value problem in the fixed domain $\Omega_1 = (0,1)$. Under the transformation (2.29) we have the one-one corresponding between y and u given by

$$y(a,at) = u(a,t), u(a,x/a) = y(a,x) \tag{2.30}$$

and for $m \geq 1$ we have:

$$x \mapsto y(a, x) \epsilon H^m(0, a) \text{ if and only if } t \mapsto u(a, t) \epsilon H^m(0, 1) \tag{2.31}$$

227

Similarly if $\varphi \epsilon H^m(0, 1)$ then

$$x \mapsto \psi(a, x) = \varphi(x/a) = \varphi(t) \epsilon H^m(0, a) \tag{2.32}$$

and conversely. Moreover, we also have

$$\begin{cases} y'(a, x) = a^{-1} \partial u/\partial t(a, x/a) = a^{-1} u_t(a, x/a) \\ \psi'(a, x) = a^{-1} \varphi_t(x/a), \end{cases} \tag{2.33}$$

so that the state equation can now be written as
(2.34)

$$\begin{cases} \int_0^a \{a^{-2} u_t(a, x/a) \varphi_t(x/a) + u(a, x/a) \varphi(x/a) - f(x) \varphi(x/a)\} dx = 0 \\ \text{for all } \varphi \epsilon H^m(0, 1). \end{cases}$$

□

By the transfomation (2.29) this becomes

$$\begin{cases} \int_0^1 \{a^{-2} u_t(a, t) \varphi_t(t) - (u(a, t) + f(at)) \varphi(t)\} dt = 0 \\ \text{for all } \varphi \epsilon H^m(0, 1). \end{cases} \tag{2.34)'$$

Since $h^m(0, 1)$ is dense in $H^1(0, 1)$ (for any $m \geq 1$) it follows that (2.34)′ is valid also for any $\varphi \epsilon H^1(0, 1)$. This means that $t \mapsto u(a, t)$ is a solution of the two point boundary value problem

$$\begin{cases} u = u(a, t) \\ d^2u/dt^2 + u = f(at) \\ u_t(a, 0) = 0 = u_t(a, 1) \end{cases} \tag{2.34)''$$

Since $t \mapsto f(at) \epsilon H^1(0, 1)$ by Hypothesis (3) we know, form the regularity theorey for (ordinary) differentail equation, that

$$t \mapsto u(a, t) \epsilon H^3(0, 1)$$

228 which proves (i). Then by Sobolev's lemma $t \mapsto u(a,t)\epsilon C^2([0,1])$. It follows then that

$$x \mapsto y(a,x) = u(a, x/a)\epsilon C^2([0,1]). \tag{2.35}$$

which proves the second part of (iii).

In order to prove that y_a exists and is regular it is enough to prove the same for u_a. For this purpose, we shall show the u_a satisfies a second order (elliptic) variational boundary value problem.

We note that, by the theorem of dependence on parameters, the solution of (2.34)″ as a functiona of the variable a is differentiable since the Hypothesis (3) implies that

$$(df/da)(at) = tf_t(at)\epsilon L^2(0,1). \tag{2.36}$$

Now if we differentiate (2.34)′ with respect to a we get

$$\begin{cases} \int_0^1 \{a^{-2}u_{t,a}(a,t)\varphi_t(t) + u_a(a,t)\varphi(t)\}dt \\ = 2a^{-3}\int_0^1 u_t(a,t)\varphi_t(t)dt + \int_0^1 f_t(at)t\varphi(t)dt. \\ \text{for all } \varphi\epsilon H^m(0,1). \end{cases} \tag{2.37}$$

Here on the right side the first term exists since $t \mapsto u_t(a,t)\epsilon L^2(0,1)$ while the second term exists since $t \mapsto f_t(at)\epsilon L^2(0,1)$ by Hypothesis (3). Now $t \mapsto u(a,t)\epsilon H^3(0,1)$ implies that $u_{t,t}\epsilon H^1(0,1) \subset L^2(0,1)$ and so on integrating by parts we find that

$$\int_0^1 u_t(a,t)\varphi_t dt = -\int_0^1 u_{t,t}(a,t)\varphi(t)dt + [u_t(a,t)\varphi(t)]_{t=0}^{t=1}.$$

Since $u_t(a,t) = ay'(a,x)$ the boundary conditions in (2.34)″ on y 229 imply that

$$[u_t(a,t)\varphi(t)]_{t=0}^{t=1} = [ay'(a,x)\varphi(x/a)]_{x=0}^{x=a} = 0.$$

Hence the right side of (2.37) can be written as

$$\int_0^1 \{-2a^{-3}u_{t,t}(a,t) + tf_t(at)\}\varphi(t)dt. \tag{2.38}$$

Since $-2a^{-3}u_{t,t}(a,t) + tf_t(a,t)\epsilon L^2(0,1)$ we conclude that $u_a(a,t)$ satisfies a variational second order (elliptic) boundary value problem (2.37) with the right hand side (2.38) data in $L^2(0,1)$. Then by the regularity theory of solutions of (ordinary) differential equation it follows that

$$t \mapsto u_a(a,t)\epsilon H^2(0,1) \tag{2.39}$$

Then

$$y_a(a,x) = u_a(a,x/a) + (-x/a^2)u_t(a,x/a)\epsilon H^2(0,a) \tag*{(2.39)$'$}$$

which proves the assertion (ii). Again, applying Sobolev's lemma to y_a, the second part of (iii) is also proved. This proved the proposition completely.

We also have the following regularity property for the adjoint state $p(a,x)$.

Proposition 2.2. *If satisfies the Hypothesis (3) and g the Hypothesis (4) then the adjoint state $x \mapsto p(a,x)$ belongs to $H^3(0,a)$ and consequently $x \mapsto p(a,x)\epsilon C^2([0,a])$.*

Proof. The adjoint state equation (2.27) is transformed by (2.29) as follows:

$$p(a,at) = q(a,t) \text{ and } \psi(a,x) = \varphi(x/a)$$

$$\begin{cases} \int_0^a \{a^{-2}q_t(a.x/a)\varphi_t(x/a) + a(a,x/a)\varphi(x/a)\}dx \\ = \int_0^a (y(a,x/a) - g(x/a))\varphi(x/a)dx, \text{ for all } \varphi\epsilon H^1(0,a) \end{cases}$$

230 That is, we have

$$\begin{cases} \int_0^1 \{a^{-2}q_t(a,t)\varphi_t(t) + q(a,t)\varphi(t)\}dt = \int_0^1 (u(a,t) - g(t))\varphi(t)dt. \\ \text{for all } \varphi\epsilon H^1(0,1). \end{cases} \tag{2.40}$$

□

Since on the right hand side $t \mapsto u(a,t) - g(t)\epsilon H^1(0,1)$ by Proposition (2.1) above it follows, again by the regularity theory for ordinary differential equations, that

$$t \mapsto q(a,t)\epsilon H^3(0,1) \tag{2.41}$$

This is equivalent to saying that

$$x \mapsto p(a,x)\epsilon H^3(0,a). \tag{2.41$'$}$$

By Sobolev's lemma it follows that $x \mapsto p(a,x)\epsilon C^2([0,1])$, completing the proof of the proposition.

Next we verify that F defined by (2.20)$'$ satisfies the required Hypothesis (1) and (2) for the validity of the calculation of dj/da.

If we assume that $\varphi\epsilon H^3(0,1)$ then $x \mapsto \varphi(x/a)\epsilon H^3(0,a)$ and then by Sobolev's lemma, $x \mapsto \varphi(x/a)\epsilon C^2([0,1])$ and $\varphi'(x/a)\epsilon H^2(0,a) \subset C^1([0,1])$. Hence we find, on using Proposition (2.1) (i) and (iii), that

$$x \mapsto F(x,a) = y'(a,x)\varphi'(x/a) + (y(a,x) - f(x))\varphi(x/a)\epsilon C^\circ([0,a]) \tag{2.42}$$

since we know that $f\epsilon H^1(0,a) \subset C^\circ([0,a])$ by Hypothesis (3) and Sobolev's lemma. Moreover, differentiating the expression for F with respect to a using Proposition (2.1) (ii) and (iii) we see that

$$\begin{aligned} x \mapsto y'_a(a,x)\varphi'(x/a) + y_a(a,x)\varphi(x/a) + y'(a,x)(\varphi'(x/a))_a \\ + (y(a,x) - f(x))(\varphi(x/a))_a \epsilon C^\circ([0,a]) \end{aligned} \tag{2.43}$$

which proves that $F : \Lambda \to \mathbb{R}$ satisfies the Hypothesis (1) and (2).This the expression on the right hand side of (2.28) has a meaning since 231

$$y'(a,x)p'(a,x) + (y(a,x) - f(x))p(a,x)\epsilon C^\circ([0,a]) \tag{2.44}$$

and we obtain

$$dj/da = -[y'(a,a)p'(a,a) + (y(a,a) - f(a))p(a,a)]. \tag{2.28$'$}$$

Thus we have proved the following main result of this section:

Theorem 2.1. *Under the Hypothesis (3) and (4) on the data f and g the cost function $a \mapsto j(a)$ is differentiable and dj/da is given by* (2.28)$'$ *where $y(a,x)$ and $p(a,x)$ represent the direct and adjoint state respectively governing the problem of optimal design (2.18).*

Remark 2.4. The genral method described in this section is not, in general, used for one-dimensional problems since it is not economical to compute dj/da which in turn involves computations of y and p, and their derivativex (see $(2.28)'$. In the case of one dimensional problems other more efficient and simper methods are known in literature. The importance of our method consists in its usefulness in higher dimensions to device algorithms using, for instance, the gradient method.

Bibliography

[1] Brezis, H. Multiplicateus de Lagrange en torsion élasto - plastique, Archive Rat. Mech. Anal. 49 (1972), 32-40. 232

[2] Brezis, H and Sibony M., Equivalence de deux inéquations variationnelles et applications, Archive Rat. Mech. Anal. 41 (1971), 254-265.

[3] Brezis, H and Sibony M, Méthodes d'approximation et d'itération pourles opérateurs monotones, Archive Rat. Mech. Anal. 28(1968),

[4] Brezis, H. and Stampacchia, G., Sur la régularité de la solution d'inéquations elliptiques, Bull. Soc. Mathématique de France, 96(1968), 153-180.

[5] Brezis, H. and Stampacchia, G, Une nouvelles methode pour l'étude d'écoulement stationnaires, C. R. Acad. Sci Paris. 276(1973),

[6] Céa, J., Optimisation, Théorie et algorithmes, Dunod, Gauthier - Villars Paris (1971).

[7] Céa, J, Approximation variationelle des problémes aux limites Annales de l'Institut Fourier, 14(1964), 345-344.

[8] Céa, J and Glowinski, R., Méthodes numériques pour l'écoulement lamminaire d'une fluide rigide visco-plastique incompressible, Int. Jr. of comp. Math. B, 3(1972), 225-255.

[9] Céa, J and Glowinski, R, Sur des méthodes d'optimisation par relaxation, Revue Francaise d'Automatique, Informatique, Recherche Opérationelle R-3 (1973), 5-32.

[10] Cryer, C. W., The solution of a quadratic programming problem using systematic over-relaxation. SIAM J. on control 9(1971).

[11] Davidon, W. D., Variable metric method for minimization. A. G. G. Res and Dev. Report No. ANL - 5990 (1959).

[12] Dieudonné, J., Foundations of modern analysis, Academic Press. Newyork (1960).

[13] Ekeland, I. and Temam R., Analyse convexe et problémes variationelles, Dunod, Gauthier-Villars, Paris (1974).

233 [14] Fletcher, R., Optimization, Academic Press, Landon (1969).

[15] Fletcher, R. and Powell, M., A rapidly convergent method for minimization, Comp. J. 6(1963), 163-168.

[16] Fletcher R. and Reeves C. M., Functional minimization by conjugate gradients, Comp. J. 7(1964), 149-153.

[17] Frank, M, and Wolfe, P., An algorithm for quadratic programming, Naval Res. Log. Quart. 3(1956), 95-110.

[18] Glowinski, R., La méthode de relaxation, Rendiconti di Matematica, Universitá di Rome (1971).

[19] Glowinski, R, Sur la minimization, par surrelaxation avec projection de fontionneles quadratiques dans les espaces de Hilbert, C. R. Acad. Sci. Paris 276(1973). 1421-1423.

[20] Glowinski, R., Lions J. L. and Trémoliéres, R., Analyse numérique des inéquations variationelles. Volumes 1and 2, Dunod Fauthier-Villars, Paris (1976).

[21] Glodstein, A. A., Convex programming in Hilbert spaces, Bull. Amer. Math. Soc. 70(1964), 709-710.

[22] Hestenes, M. R., The conjugate gradient method for solving lenear systems, Proc. Symposium in Appl. Math. 6 - Numerical Analysis Amer. Math. Soc. and McGraw Hill, New york (1956), 83-102.

[23] Hestenes, M. R, Multiplier and gradient methods J.O.T.A. 4(1969), 303-320.

[24] Hestenes, M. R. and Stiefel, E., Methods of conjugate gradient for solving linear systems, J. Res. Nat. Bureau of Standarada 49(1952), 409-436.

[25] Huard, P., Resolution of mathematical programming with non-linear constraints by the method of centres, Non-linear Programming (Edited by Abadie J.) North Holland (1967).

[26] Kuhn, H. W., On a pair of dual non-linear programs, Non-linear programming, (Edited by E.M.L. Beale) NATO Summer School. Menton (1964), North Holland, Amsterdam (1967), 37-54.

[27] Kuhn, H. W., An a logrithm for equilibrium points in bimatrix games, Proc. Nat Acad, Sci, U.S.A. 47(1961), 1657-1662.

[28] Kuhn, H. W. and Tucker, A. W., Non-linear programming, Proc. **234** Second Berkeley Symposium on math. Statistic and Probability, University of California Pres (1951), 481-492.

[29] Ky - Fan, Sur un théoréme minimax, C. R. Acade. Sci. Paris 259(1964), 3925-3928.

[30] Lions, J. L., Cours d'analyse numérique, Ecole Polytechnique, Paris (1972).

[31] Lions, J. L, Contole optimal des systémes gouvernés par des équations aux dérivées partielles, Dunod, Gauthier-Villars, Paris (1968).

[32] Lions, J. L. and Magenes E., Problémes aux limites non-homogénes, Vol.1, Dunod, Gautheir-Villars, Paris (1968); (English translation: Non-homogeneous boundary value problems Vol.1, Springer-Verlag(1972)).

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

[34] Loomis, L. H., An introduction to abstract harmonic analysis, Van Nostrand, New Yorl (1953).

[35] Powell, M.J.D., A method for non-linear optimization in minimization problems, optimization (Edited by R. Fletcher) Academic Press, New York (1969).

[36] Rockafellat, A.T., The multiplier method of Hestenses and Powell applies to convex programming, J. Opt. Th. and Appl. 12(1973).

[37] Rockafellat, A.T., Augmented Lagrange multiplier functions and duality in non-convex programming SIAM J. on control 12(1974), 268-285.

[38] Rockafellat, A.T., Dualtiy and stability in extremum problems involving convex functions, Pacific J. Math., 21(1962), 167-187.

[39] Rosen, J, B., The gradient projection method for non-linear programming, Part I, Linear constraints, SIAM J. 8(1960), 181-217.

[40] Rosen, J, B., The gradient projection method for non-linear programming, Part II, Non-linear constraints SIAM J. 9(1961). 514-532.

235 [41] Sion, M., Existence des cols pour les functions quasi-convexex at sémi-continues, C. R. Acad. Sci. paris 244(1954). 2120-2123.

[42] Sion, M., On general minimax theorems, Pacific J. Math. 8(1958), 171-175.

[43] Stampacchia G., Formes bilinéaires coercitives sur les ensembles convexes, C. R. Acad Sci. Paris 258(1964), 4413-4416.

[44] Stampacchia G., Variational inequalities, Theorey and Applications of Monotone, operators, Proc. NATO Advanced Study Institute, Venice (1968), 101-192.

[45] Trémoliéres, A, La méthode des Centers a troncature variable, Thése 3^e Cycle, Paris (1968).

[46] Trémoliéres, A, Optimisation non-linéaire avec containtes, Rapport IRIA.

[47] Tucker, A. W., Duality theory of linear programs, A constructive approach with applications, SIAM Revue 11(1969), 347-377.

[48] Tucker, A. W., Solving a matrix game by linear programming, IBM J. Res. and Dev. 4(1960), 507-517.

[49] Uzawa, H., Iterative methods for concave programming, Studies in linear and non-linear programming (Edited by Arrow, K. J., Hurwitz, L. and Uzawa, H.) Stanford University Press (1958), 154-165.

[50] Varge, R. S., Matrix iterative analysis, Prentice Hall, Englewood Clifs, New Jersey (1962).

[51] Wolfe, P., Method of non-linear programming, Recent advances in mathematical programming (Edited by Graves, R. L. and Wolfe, P.) McGraw Hill, New York (1963), 67-86.

[52] Zoutendijk, G., Non-linear programming, A numerical survey, SIAM J. on Control, 4(1966), 194-210.

[53] Céa, J. On the problem of optimal design.

[54] Fages, R. A generalized Newton method (to appear).

[55] Auslander, A., Méthodes numériques pour la décomposition et la minimisation de fonctions non différentiables, Numer. Math, 18(1972), 213-223.