

## **NUMERICAL ANALYSIS**

ROGER TEMAM

# NUMERICAL ANALYSIS



D. REIDEL PUBLISHING COMPANY

DORDRECHT-HOLLAND / BOSTON-U.S.A.

**ANALYSE NUMÉRIQUE**

*First published by Presses Universitaires de France, Paris, 1970*

*Translated from the French by Dr J. W. Nienhuys*

**Library of Congress Catalog Card Number 73-75643**

**ISBN-13: 978-94-010-2567-6    e-ISBN-13: 978-94-010-2565-2**

**DOI: 10.1007/978-94-010-2565-2**

**Published by D. Reidel Publishing Company,  
P.O. Box 17, Dordrecht, Holland**

**Sold and distributed in the U.S.A., Canada, and Mexico  
by D. Reidel Publishing Company, Inc.  
306 Dartmouth Street, Boston,  
Mass. 02116, U.S.A.**

**All Rights Reserved  
Copyright © 1973 by D. Reidel Publishing Company,  
Dordrecht, Holland**

**Softcover reprint of the hardcover 1st edition 1973**

**No part of this book may be reproduced in any form, by print,  
photoprint, microfilm, or any other means,  
without written permission from the publisher**

## CONTENTS

|                                                                  |     |
|------------------------------------------------------------------|-----|
| Preface                                                          | VII |
| Approximate Solution of Functional Equations:<br>General Remarks | 1   |

### PART I: APPROXIMATE SOLUTION OF LINEAR ELLIPTIC EQUATIONS

|                                                 |    |
|-------------------------------------------------|----|
| 1. The Projection Theorem                       | 5  |
| 2. The Method of Galerkin                       | 12 |
| 3. The Approximation of Normed Spaces           | 18 |
| 4. Approximation of Linear Variational Problems | 31 |
| 5. A Method for Computation of the Error        | 40 |
| 6. The Method of Fractionary Steps              | 43 |

### PART II: APPLICATIONS

|                                                                         |    |
|-------------------------------------------------------------------------|----|
| 7. Spaces of Functions Associated with an Open<br>Set in $\mathbf{R}^n$ | 53 |
| 8. Approximation of Some Function Spaces by<br>Finite Differences (I)   | 59 |
| 9. Approximation of Some Function Spaces by<br>Finite Differences (II)  | 70 |

## **VI**

## **CONTENTS**

|                                                                                |            |
|--------------------------------------------------------------------------------|------------|
| <b>10. Approximation of Some Function Spaces by<br/>Finite Element Methods</b> | <b>80</b>  |
| <b>11. Example I: The Dirichlet Problem</b>                                    | <b>105</b> |
| <b>12. Example II: The Neumann Problem</b>                                     | <b>121</b> |

### **PART III: A NONLINEAR ELLIPTIC EQUATION**

|                                            |            |
|--------------------------------------------|------------|
| <b>13. The Exact Problem</b>               | <b>131</b> |
| <b>14. Approximate Problems</b>            | <b>143</b> |
| <b>Appendix: Topological Vector Spaces</b> | <b>156</b> |
| <b>Bibliography</b>                        | <b>161</b> |
| <b>Index</b>                               | <b>165</b> |

## PREFACE

This book is an introduction to one of the important aspects of Numerical Analysis, namely the approximate solution of functional equations. We intend to show, by a few brief examples, the different theoretical and practical problems related to the numerical approximation of boundary value problems. We have chosen for this the approximate solution of certain linear elliptic partial differential equations (the first two parts of the book) and the approximate solution of a nonlinear elliptic differential equation.

This book is not a systematic study of the subject, but the methods developed here can be applied to large classes of linear and nonlinear elliptic problems.

The book assumes that the reader's knowledge of Analysis is comparable to what is taught in the first years of graduate studies. This means a good knowledge of Hilbert spaces, elements of measure theory and theory of distributions. The subject matter of the book covers the usual content of a first course on Numerical Analysis of partial differential equations.

In this English version the concept of approximation of a normed space has been weakened. In this way it becomes

more general and easier to handle. Since the finite element methods have been developed considerably in the last few years, a new chapter describing briefly these methods has been added.

## APPROXIMATE SOLUTION OF FUNCTIONAL EQUATIONS: GENERAL REMARKS

Every problem of mathematical physics leads naturally to solving one or more functional equations that we write in simplified form:

$$(0.1) \quad \mathcal{A}u = f,$$

where  $\mathcal{A}$  is an operator from some space  $X$  into some space  $Y$ ,  $f$  is given in  $Y$ , and  $u$  is the desired solution in  $X$  (examples: ordinary and partial differential equations and integral equations).

Generally speaking, it is not possible to determine the solution of (0.1) explicitly, or its explicit form may be so complicated that it is useless, so one is interested in the approximate solution of the equation. Now the idea is to replace the spaces  $X$  and  $Y$  by 'simpler' spaces  $X_h$  and  $Y_h$ , and to associate with Equation (0.1) a family of approximate equations (where  $h$  is a parameter):

$$(0.2) \quad \mathcal{A}_h u_h = f_h,$$

where  $\mathcal{A}_h$  approximates  $\mathcal{A}$ ,  $f_h \in Y_h$  approximates  $f$  and  $u_h \in X_h$  approximates  $u$  (at least, that is what we would like).

The problems that arise are the following:

(1) *Study of the exact equation.* We have to study ex-

istence and uniqueness of the solutions of (0.1). In the most favourable case, the equation has a unique solution. In case the equation has several solutions, we have to characterize the solution that we want to approximate, which is in general difficult.

(2) *Study of the approximate equations.* After the problems (0.2) have been suitably formulated, we have to study existence and uniqueness of the solutions of (0.2). Next we must develop algorithms for actually solving (0.2), because these equations are not altogether trivial (for example systems of linear or nonlinear algebraic equations).

(3) *Study of stability and convergence.* These questions are fundamental for applications. In the case of stability, the point is to ensure that the approximate solutions  $u_h$  stay bounded in a certain sense; for convergence the point to prove is that the  $u_h$  converge in a suitable sense to the solution  $u$  of (0.1).

The stability problem is trivial for the elliptic equations studied in the sequel, but such is not always the case.

(4) *Other problems.* We still must add the problem of computing the error, that is the computation of the error between  $u$  and  $u_h$  for fixed  $h$ . Then, there occur practical problems, which are also important, but more related to the use of computers, such as computation time, required memory space, etc.

We are now going to work out the above program for some types of equations.

## PART I

# APPROXIMATE SOLUTION OF LINEAR ELLIPTIC EQUATIONS

This part and Part II, in which the ‘theory’ and the examples respectively are developed, should be read at the same time. Suitable references in Part I indicate the chapters of Part II that contain the corresponding examples.

## CHAPTER 1

### THE PROJECTION THEOREM

In this chapter we prove a very simple theorem, known as the projection theorem or Lax-Milgram theorem. It implies existence and uniqueness of 'weak' solutions for certain classes of linear elliptic problems. In Part II, two applications are developed (cf. Chapter 7; Sections 11.1 and 12.1).

#### 1.1. Statement

Let a real Hilbert space  $V$  be given, with scalar product denoted by  $((.,.))$  and with associated norm  $\|.\|$ . The dual Hilbert space of  $V$  is denoted by  $V'$ , with norm  $\|.\|_*$ . If  $l \in V'$  and  $u \in V$ , then we denote the scalar product between  $l$  and  $u$  with respect to the duality between  $V$  and  $V'$  by  $\langle l, u \rangle$ .

We also consider a bilinear continuous form  $a(u, v)$  on  $V \times V$ , and we suppose that it satisfies the following condition:

$$(1.1) \quad \exists \alpha > 0 \quad \text{such that} \quad a(u, u) \geq \alpha \|u\|^2, \quad \forall u \in V.$$

Condition (1.1) is called coercivity condition; when it is satisfied, the form  $a(u, v)$  is said to be coercive.

**THEOREM 1.** *Let  $V$  be a real Hilbert space and  $a(u, v)$  a*

*continuous bilinear and coercive form on  $V$ . For each given  $l$  in  $V'$ , there exists one and only one  $u$  in  $V$  such that:*

$$(1.2) \quad a(u, v) = \langle l, v \rangle, \quad \forall v \in V;$$

*moreover, the map  $l \mapsto u$  is linear and continuous from  $V'$  into  $V$ .*

Before we prove this theorem, we reformulate Equation (1.2), by introducing the operator  $A$ .

### *The operator $A$*

For given  $u$  in  $V$ , the map  $v \mapsto a(u, v)$  is a continuous linear form on  $V$  and hence there exists an element of  $V'$ , which we denote by  $Au$  such that:

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

It is easy to verify that the map  $u \mapsto Au$  is a linear map from  $V$  into  $V'$ ; we are going to prove that it is an isomorphism from  $V$  onto  $V'$ , but first we note that Equation (1.2) is equivalent to

$$\langle Au, v \rangle = \langle l, v \rangle, \quad \forall v \in V;$$

in other words:

$$Au = l \quad (\text{equality in } V').$$

When we prove that the operator  $A$  is an isomorphism from  $V$  onto  $V'$ , Theorem 1 will easily follow.

**PROPOSITION 1.** *Under the hypotheses of Theorem 1 the operator  $A$  is an isomorphism from  $V$  onto  $V'$ .*

*Proof.* We already observed that  $A$  is linear. Let us now prove that it is continuous.<sup>1</sup> The form  $a(u, v)$  is continuous, so there exists a constant  $M$  such that:

$$|a(u, v)| \leq M \|u\| \|v\|, \quad \forall u, v \in V;$$

hence:

$$|\langle Au, v \rangle| \leq M \|u\| \|v\|,$$

and so:

$$\|Au\|_* \leq M \|u\| \quad \forall u \in V,$$

which proves the continuity of  $A$ .

By using (1.1) we obtain

$$\alpha \|u\|^2 \leq a(u, u) = \langle Au, u \rangle \leq \|Au\|_* \|u\|;$$

hence:

$$\alpha \|u\| \leq \|Au\|_*,$$

which proves that  $A$  is injective and that the (linear) operator  $A^{-1}$  is continuous from  $AV$  onto  $V$ .

Since  $A$  is an isomorphism from  $V$  onto  $AV$ , we see that  $AV$  is complete and hence closed in  $V'$ . We are going to finish by proving that  $AV = V'$ , in other words,  $A$  is surjective.

Let  $(AV)^0$  be the polar of  $AV$  in  $V$ ; so for  $u \in (AV)^0$  we

have:

$$\begin{aligned}\langle l, u \rangle &= 0 & \forall l \in AV \\ \langle Av, u \rangle &= 0 & \forall v \in V,\end{aligned}$$

and so in particular:

$$\langle Au, u \rangle = a(u, u) = 0,$$

which implies  $u=0$ , by (1.1). This implies in turn that the polar of  $AV$  is just  $\{0\}$ , and hence it follows that:

$$AV = \overline{AV} = (AV)^{00} = \{0\}^0 = V'.$$

The proof is completed.

## 1.2. A Special Case

A useful special case of the preceding is as follows. We have, together with the space  $V$ , another real Hilbert space  $H$  (with scalar product  $(.,.)$  and norm  $|\cdot|$ ), which contains  $V$  as a dense subspace, and such that the injection of  $V$  into  $H$  is continuous.

The canonical injection of  $V$  into  $H$  defines by transposition a linear continuous map of  $H'$  (the dual of  $H$ ) into  $V'$  (the dual of  $V$ ); one verifies easily that this map is injective, which allows us to identify  $H'$  with a subspace of  $V'$ , moreover this subspace is dense in  $V'$ :

$$V \subset H, \quad V \xrightarrow{i} H, \quad H' \xrightarrow{i'} V'.$$

Indeed, according to the Riesz theorem<sup>2</sup> there exists a

canonical isomorphism from  $H$  onto  $H'$ , and we may identify  $H$  and  $H'$  by this isomorphism. So:

$$V \subset H \equiv H' \subset V',$$

where the injections are continuous, and each space is dense in the following one.<sup>3</sup>

If  $f \in H$ , then  $f$  may be identified with an element of  $V'$ , which is just the continuous linear form  $v \in V \mapsto (f, v)$  (the scalar product between  $f$  and  $v$  taken in  $H$ ).

Theorem 1 implies in particular, that for given  $f$  in  $H$ , there exists a unique  $u \in V$  such that:

$$(1.4) \quad a(u, v) = (f, v), \quad \forall v \in V.$$

The set of  $u \in V$  such that  $Au \in H$  is called  $D(A)$ , or domain of  $A$  in  $H$ . The space  $D(A)$  is dense in  $V$  and  $H$ , because  $A$  is an isomorphism of normed spaces, and  $H$  is dense in  $V'$ . It is easily verified that  $D(A)$  is a Hilbert space for the scalar product:

$$(u, v) + (Au, Av),$$

and that  $A$  is an isomorphism from  $D(A)$  onto  $H$ .

### 1.3. Variational Properties

When the form  $a(u, v)$  is symmetric, problem (1.2) is equivalent with a problem from variational calculus.

**PROPOSITION 2.** *Hypotheses as in Theorem 1, moreover it*

is assumed that the form  $a(u, v)$  is symmetric. Then,  $u \in V$  is a solution of (1.2), if and only if the function  $E$  defined on  $V$  by:

$$(1.5) \quad E(v) = a(v, v) - 2 \langle l, v \rangle$$

takes its minimum for  $u$ . So there exists precisely one  $u \in V$  such that:

$$(1.6) \quad E(u) = \inf_{v \in V} E(v).$$

*Proof.* Let  $u$  be the solution of (1.2) and let  $v \in V$ ; then evidently:

$$a(u - v, u - v) \geq 0,$$

which gives, taking into account (1.2) and the symmetry of  $a$ , that  $E(u) \leq E(v)$ .

Conversely, suppose that the functional  $E(\cdot)$  takes its minimum in  $V$  for  $u$ , and let  $v \in V$ . Then:

$$E(u) \leq E(u + \lambda v)$$

for any  $\lambda \in \mathbb{R}$ . After working this out and simplifying, we find:

$$\lambda^2 a(v, v) - 2\lambda a(u, v) + 2\lambda \langle l, v \rangle \geq 0.$$

By considering successively the cases  $\lambda > 0$  and  $\lambda < 0$ , we divide by  $\lambda$  and let  $\lambda$  go to 0; we obtain:

$$a(u, v) = \langle l, v \rangle,$$

and since  $v \in V$  is arbitrary,  $u$  is indeed solution of (1.2).

## NOTES

<sup>1</sup> These two properties are independent of (1.1). Indeed, giving a bilinear continuous form on  $V$  amounts to giving a continuous linear operator from  $V$  to  $V'$ .

<sup>2</sup> The theorem that every linear functional on a Hilbert space is of the form  $x \rightarrow (x, y)$  for some  $y$ , proved by M. Fréchet for the case  $H = L^2([0, 1])$  and by F. Riesz in general.

<sup>3</sup> The injection which we have hereby defined from  $V$  into  $V'$  has nothing to do with the canonical isomorphism from  $V$  to  $V'$ , and one should be careful not to identify  $V$  and  $V'$  by this isomorphism.

## CHAPTER 2

### THE METHOD OF GALERKIN

The aim of the following chapters is to study the approximation of the solution of problem (1.2). To begin with, we present here the method of Galerkin and we will see how this leads to an approximate solution of Equation (1.2).

#### *The approximated problem*

The hypotheses are the same as in Chapter 1, but for simplicity we suppose that the Hilbert space  $V$  is separable, and we consider a sequence  $w_j$ ,  $j = 1, 2, \dots$ , which is free and total in  $V$  (the sequence is not necessarily orthonormal). Let  $V_m$  be the space spanned by  $w_1, \dots, w_m$ . Now, for each fixed integer  $m$ , we want to find an approximate solution of (1.2) in  $V_m$ , that is, a solution of type:

$$(2.1) \quad u_m = \sum_{i=1}^m \xi_{i,m} w_i, \quad \xi_{i,m} \in \mathbf{R}.$$

It would not be reasonable to demand that  $u_m$  satisfies

$$a(u_m, v) = \langle l, v \rangle$$

for all  $v \in V$ , but we can impose that this equality holds for all  $v$  of  $V_m$ . In fact, let us denote the restriction of  $a$  to

$V_m \times V_m$  by  $a_m$ , and the restriction of  $l$  to  $V_m$  by  $l_m$ ; the condition for  $u_m$  becomes then:

$$(2.2) \quad a_m(u_m, v) = \langle l_m, v \rangle, \quad \forall v \in V_m.$$

Clearly  $a_m$  is still a coercive continuous bilinear form on  $V_m \times V_m$  and also clearly  $l_m$  is a linear continuous form on  $V_m$ : hence the projection theorem gives us existence and uniqueness of an element  $u_m$  of  $V_m$  satisfying (2.2).

*Remark.* Equation (2.2) is evidently equivalent with the set of equations:

$$a_m(u_m, w_j) = \langle l_m, w_j \rangle, \quad j = 1, \dots, m;$$

in other words:

$$\sum_{i=1}^m a(w_i, w_j) \xi_{i,m} = \langle l, w_j \rangle, \quad j = 1, \dots, m.$$

This is a linear system of  $m$  equations with  $m$  unknowns  $\xi_{1,m}, \dots, \xi_{m,m}$ , and according to the preceding, this system has one and only one solution.

We are now going to prove the following convergence result.

**THEOREM.** *Let  $V$  be a real separable Hilbert space and let  $a(., .)$  be a coercive and continuous bilinear form on  $V$ .*

*The solution  $u_m \in V_m$  of (2.2) converges in  $V$  in the strong sense, for  $m \rightarrow \infty$ , to the solution  $u$  of (1.2).*

*Proof.* When we put  $v = u_m$  in (2.2), we obtain:

$$(2.3) \quad a(u_m, u_m) = \langle l, u_m \rangle,$$

from which, by (1.1), we get:

$$(2.4) \quad \alpha \|u_m\|^2 \leq a(u_m, u_m) = \langle l, u_m \rangle \leq \|l\|_* \|u_m\|$$

$$(2.4) \quad \|u_m\| \leq (1/\alpha) \|l\|_*;$$

this proves that the sequence  $u_m$  is bounded independently of  $m$  in  $V$ . Under these conditions (cf. Appendix) there exists an element  $u_*$  of  $V$  and a sequence  $u_{m'}$ ,  $m' \rightarrow \infty$ , extracted from  $u_m$  such that:

$$(2.5) \quad u_{m'} \rightarrow u_* \quad \text{in the weak topology of } V, \text{ for } m' \rightarrow \infty.$$

Let  $v$  be a finite linear combination of the elements  $w_i$ ; hence  $v \in V_j$  for fixed  $j$  and  $v \in V_{m'}$  as soon as  $m' \geq j$ . According to (2.2) we have:

$$a(u_{m'}, v) = \langle l, v \rangle;$$

by using (2.5) and (2.8) (the following lemma), we can take the limit in this equality, for  $m' \rightarrow \infty$ , and we obtain:

$$(2.6) \quad a(u_*, v) = \langle l, v \rangle.$$

Equality (2.6) holds for each finite linear combination of the  $w_i$ ; because the sequence  $w_i$  is total, the finite linear combinations of its elements are dense in  $V$ , so equality (2.6) still holds by continuity for all  $v$  of  $V$ . This proves that  $u_*$  is a solution of (1.2)<sup>1</sup> and because Equation (1.2) has a unique solution  $u$ ,  $u_* = u$ .

In the same way, it can be proved that from every subsequence of  $u_m$  a subsequence can be extracted, which converges weakly to  $u$ . This proves that the sequence  $u_m$  as a whole, and not just a subsequence, converges to  $u$  in the weak topology in  $V$  (argue by contradiction).

### *Strong convergence*

To prove that the sequence  $u_m$  converges strongly to  $u$ , we consider the expression:

$$X_m = a(u_m - u, u_m - u)$$

$$X_m = a(u_m, u_m) - a(u, u_m) - a(u_m, u) + a(u, u).$$

According to (2.3):

$$a(u_m, u_m) = \langle l, u_m \rangle,$$

and hence, since the sequence converges to  $u_m$  in the weak topology of  $V$ :

$$\lim_{m \rightarrow \infty} a(u_m, u_m) = \langle l, u \rangle.$$

Similarly, by using (2.7) and (2.8) (cf. below):

$$\lim_{m \rightarrow \infty} a(u, u_m) = a(u, u)$$

$$\lim_{m \rightarrow \infty} a(u_m, u) = a(u, u).$$

Hence we find:

$$\lim_{m \rightarrow \infty} X_m = -a(u, u) + \langle l, u \rangle,$$

and this expression is zero (put  $v=u$  in (1.2)). By (1.1) we get:

$$0 \leq \alpha \|u_m - u\|^2 \leq X_m \rightarrow 0,$$

which proves that  $u_m$  converges to  $u$  in the strong topology of  $V$ .

The proof of the theorem is now completed, except for the following lemma, that we have already used.

**LEMMA.** *Let  $\varphi_m$  be a sequence of elements of  $V$ , which converges to  $\varphi$  in the weak topology of  $V$ . Hence, for each fixed  $\psi$  of  $V$ :*

$$(2.7) \quad \lim_{m \rightarrow \infty} a(\psi, \varphi_m) = a(\psi, \varphi)$$

$$(2.8) \quad \lim_{m \rightarrow \infty} a(\varphi_m, \psi) = a(\varphi, \psi).$$

*Proof.* By definition of the operator  $A$ :

$$a(\psi, \varphi_m) = \langle A\psi, \varphi_m \rangle$$

$$a(\psi, \varphi) = \langle A\psi, \varphi \rangle,$$

and by definition of weak convergence in  $V$  (cf. Appendix), we have for each  $\sigma \in V'$ :

$$\lim_{m \rightarrow \infty} \langle \sigma, \varphi_m \rangle = \langle \sigma, \varphi \rangle,$$

from which (2.7) follows. For (2.8), we only have to apply

(2.7) to the bilinear form:

$$a^*(u, v) = a(v, u).$$

*Remark. The Ritz method.* In order to approximate the solution of the variational problem considered in Section 1.3 the Ritz method consists of finding an element  $\tilde{u}$  for which the minimum of  $E$  in a subset or subspace  $\tilde{V}$  of  $V$  is taken. If the set  $\tilde{V}$  equals one of the spaces  $V_m$  above, then  $\tilde{u} = u_m$ ; this follows from (2.2) and Proposition 2 of Chapter 1.

#### NOTE

<sup>1</sup> This proves, if that was not already known, the existence of a solution of (1.2). In this way the method of Galerkin can also be used to prove the existence of solutions for linear and nonlinear partial differential equations; cf. Chapter 13.

## CHAPTER 3

# THE APPROXIMATION OF NORMED SPACES

In the following sections we are going to consider the approximation of the solution  $u$  of Equation (1.2) by elements of a space different from  $V$ . This situation occurs in particular when (1.2) is a partial differential equation, that we discretize:  $V$  is then a certain function space, and we approximate  $u$  by functions that do not belong to  $V$ , for example, step functions, or piecewise polynomial functions.

With this in mind, we will first introduce here a few general concepts concerning the approximation of normed spaces. Examples are studied in Chapters 7, 8 and 9.

### 3.1. Introduction

The motivation of the following definitions is this: we want to approximate the elements of a normed space  $V$  by those of a family  $\{V_h\}_{h \in \mathcal{H}}$  of normed spaces. For each  $h \in \mathcal{H}$ , we must compare elements  $u$  of  $V$  with elements  $u_h$  of  $V_h$ ; if these spaces are different, one may conceive of three reasonable ways to do it:

(i) compare  $u$  with a certain image  $p_h u_h$  of  $u_h$  in  $V$  ( $p_h$  is an operator mapping  $V_h$  into  $V$ );

(ii) compare a certain image  $\bar{\omega}u$  of  $u$  in some other space  $F$ , with a certain image  $p_h u_h$  of  $u_h$  in  $F$  ( $\bar{\omega}$  is a mapping from  $V$  into  $F$  and  $p_h$  is a mapping from  $V_h$  into  $F$ );

(iii) compare  $u_h$  with a certain image  $r_h u$  of  $u$  in  $V_h$  ( $r_h$  is an operator mapping  $V$  into  $V_h$ ).

The first way leads to the concept of internal approximation of a space  $V$  (internal, because the comparison is made in  $V$ ). The second one leads to the concept of external approximation (external, because comparisons are made outside of  $V$ ). The third one will only be of auxiliary interest for both internal and external approximation, the reason being that it is not very useful to make comparisons in a space which is not fixed.

The family of approximating spaces  $V_h$  is indexed by  $h$ , which takes its values in some set  $\mathcal{H}$ . This set  $\mathcal{H}$  will be equipped with some filter and we will be concerned with taking the limit through this filter. Actually the main situations in numerical analysis are the following ones:

(a) *Galerkin and sequential approximation*:  $\mathcal{H} = \mathbb{N}$  and we take the limit as  $m \rightarrow \infty$ .

(b) *Finite differences*:

$$\mathcal{H} = \prod_{i=1}^n \{]0, h_i^0]\}, \quad h_i^0 > 0,$$

and we are interested in taking the limit for  $h \rightarrow 0$ .

(c) *Finite elements*:  $\mathcal{H}$  is a set of multidimensional triangulations of some open set  $\Omega$ , and we take the limit

$|h| \rightarrow 0$ , where  $|h|$  equals the upper bound of the diameters of the simplexes (see Chapter 10).

For the sake of simplicity we will only speak about the passage to the limit as ' $h \rightarrow 0$ '. Strictly speaking, this terminology is only correct for finite differences. But the definitions and results can be easily adapted to the other cases and more generally to the case where  $\mathcal{H}$  is a set with a filter with a countable basis.

### 3.2. Internal Approximation of a Normed Space

**DEFINITION 1.** *An internal approximation of a normed space  $V$  is a set consisting of a family of triples  $\{V_h, p_h, r_h\}$ ,  $h \in \mathcal{H}$ , where:*

- (i)  $V_h$  is a normed vector space;
- (ii)  $p_h$  is a linear continuous operator from  $V_h$  into  $V$ ;
- (iii)  $r_h$  is a linear continuous operator from  $V$  into  $V_h$ .

When  $V$  is a Hilbert space, and likewise the spaces  $V_h$ , the approximation is said to be a Hilbert approximation.

In the most interesting cases, the spaces  $V_h$  are finite dimensional; rather often, the operators  $p_h$  are injective and  $r_h$  surjective.

The operators  $p_h$  and  $r_h$  are called prolongation and restriction operators, respectively; the operators  $p_h \circ r_h$  of  $V$  into itself are called truncation operators.

We denote by  $\|\cdot\|$  (or  $\|\cdot\|_h$ ) the norm in  $V$  (or in  $V_h$ ).

**DEFINITION 2.** *For given  $h$ ,  $u \in V$  and  $u_h \in V_h$  we say that*

- (i)  $\|u - p_h u_h\|$  is the error between  $u$  and  $u_h$ ;
- (ii)  $\|u_h - r_h u\|_h$  is the discrete error between  $u$  and  $u_h$ ;
- (iii)  $\|u - p_h r_h u\|$  is the truncation error of  $u$ .

DEFINITION 3. *Stable approximations.* The prolongation operators  $p_h$  are called stable if their norms:

$$\|p_h\| = \sup_{\substack{u_h \in V_h \\ \|u_h\|_h \leq 1}} \|p_h u_h\|$$

can be majorized independently of  $h$ .

The restriction operators  $r_h$  are called stable if their norms:

$$\|r_h\| = \sup_{\substack{u \in V \\ \|u\| \leq 1}} \|r_h u\|_h$$

can be majorized independently of  $h$ .

Finally, we will say that the approximation  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  of the space  $V$  is stable if the prolongation and restriction operators are stable.

Let us now consider what happens when  $h \rightarrow 0$ .

DEFINITION 4. *Convergent approximations.* We will say that the approximation  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  of  $V$  is convergent if for all  $u \in V$ :

$$(3.2) \quad \lim_{h \rightarrow 0} p_h r_h u = u$$

in the strong topology of  $V$ .

*Remark.*<sup>1</sup> If the approximation of the space  $V$  is stable and convergent, then the operators  $p_h r_h$  form an equicontinuous set in  $\mathcal{L}(V, V)$ , which converges simply to the identity and hence (by the theorem of Banach–Steinhaus) uniformly on compact subsets of  $V$ .

It is useful to note the following:

**PROPOSITION 1.** *Let  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  be a stable approximation of the space  $V$ . For this approximation to be convergent, it is necessary and sufficient that:*

$$\lim_{h \rightarrow 0} \|u - p_h r_h u\| = 0$$

*for all  $u$  in a dense subspace  $\mathcal{V}$  of  $V$ .*

*Proof.* For all  $u \in V$  and each  $\varepsilon > 0$ , there exists an  $u_\varepsilon \in \mathcal{V}$  such that  $\|u - u_\varepsilon\| < \varepsilon$ . We have:

$$\begin{aligned} \|u - p_h r_h u\| &\leq \|u - u_\varepsilon\| + \\ &\quad + \|u_\varepsilon - p_h r_h u_\varepsilon\| + \|p_h r_h (u_\varepsilon - u)\| \\ \|u - p_h r_h u\| &\leq (1 + M) \varepsilon + \|u_\varepsilon - p_h r_h u_\varepsilon\|, \end{aligned}$$

in which  $M$  majorizes the norms of the  $p_h r_h$ . Taking the upper limit in the above inequality, we find:

$$\limsup_{h \rightarrow 0} \|u - p_h r_h u\| \leq (1 + M) \varepsilon.$$

Since  $\varepsilon$  is arbitrarily small,

$$\limsup_{h \rightarrow 0} \|u - p_h r_h u\| = 0$$

and the proposition follows.

Now we consider a family of elements  $u_h$ :

$$u_h \in V_h, \quad \forall h \in \mathcal{H}.$$

**DEFINITION 5.** *We will say that a family  $u_h$  converges strongly (or weakly) to  $u$  if  $p_h u_h$  converges strongly (or weakly) to  $u$  in  $V$ , when  $h \rightarrow 0$ .<sup>2</sup>*

*We will say that family  $u_h$  converges discretely to  $u$  if:*

$$\lim_{h \rightarrow 0} \|u_h - r_h u\|_h = 0.$$

The concepts of strong and discrete convergence are in general different, unless the approximation of the space  $V$  is stable and convergent, in which case we have:

**PROPOSITION 2.** *If the approximation  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  of the space  $V$  is stable and convergent, then discrete convergence implies strong convergence.*

*Proof.* This follows from the inequality:

$$\begin{aligned} \|u - p_h u_h\| &\leq \|u - p_h r_h u\| + \|p_h r_h u - p_h u_h\| \leq \\ &\leq \|u - p_h r_h u\| + M' \|r_h u - u_h\|_h, \end{aligned}$$

in which  $M'$  majorizes the norms of the  $p_h$ .

### *Galerkin approximation of a separable Hilbert space*

Suppose that  $V$  is a separable Hilbert space and let  $V_h$ ,  $h \in \mathbb{N}$ , be an increasing sequence of finite dimensional subspaces of  $V$ , whose union is dense in  $V$ ; for example (cf. Chapter 2), if  $w_1, w_2, \dots$ , is a free and total sequence in

$V$ , then one may take  $V_n$  equal to the subspace generated by  $w_1, \dots, w_n$ .

For each  $n \in \mathbb{N}$ , let  $p_n$  be the canonical injection of  $V_n$  into  $V$  and  $r_n$  the orthogonal projection of  $V$  onto  $V_n$ ;  $p_n$  is injective,  $r_n$  is surjective and  $\{V_n, p_n, r_n\}_{n \in \mathbb{N}}$  is an approximation of  $V$ . It can be easily verified that this approximation is stable and convergent; such an approximation is called Galerkin approximation of the space  $V$ ; it was implicitly used in Chapter 2.

### 3.3. External Approximation of a Normed Space

We consider now external approximations of normed spaces, as explained in Section 3.1.

**DEFINITION 6.** *An external approximation of a normed vector space  $V$  is a set consisting of:*

(i) *a normed vector space  $F$  and an isomorphism  $\bar{\omega}$  of  $V$  into  $F$ ;*

(ii) *a family of triples  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  in which, for each  $h$ ,  $V_h$  is a normed space,  $p_h$  a continuous linear map of  $V_h$  into  $F$ ,  $r_h$  a continuous linear map of  $V$  into  $V_h$ .*

When  $V$  and likewise  $F$  and  $V_h$  are Hilbert spaces, the approximation is said to be a Hilbert approximation.

$$\begin{array}{ccc} V & \xrightarrow{\omega} & F \\ r_h \searrow & & \nearrow p_h \\ & V_h & \end{array}$$

An internal approximation is evidently a particular case of an external approximation with  $F = V$  and  $\bar{\omega} = \text{identity}$ .

All definitions given in the case of internal approximations are easily extended to the case of external approximations (one replaces  $u$  by  $\bar{\omega}u$  and the norm of  $V$  by that of  $F$  in the pertaining case). However, the definition of convergent approximation does not work out well that way and it is more suitable to replace it by the following:

**DEFINITION 7.** *We will say an external approximation of a space  $V$  is convergent if:*

(i) *for all  $u \in V$ :*

$$\lim_{h \rightarrow 0} p_h r_h u = \bar{\omega}u$$

*in the strong topology of  $F$ ;*

(ii) *for every sequence  $u_h$  of elements of  $V_h$ , ( $h' \rightarrow 0$ ), such that  $p_h u_h$  converges to  $\varphi$  in the weak topology of  $F$ ,  $\varphi \in \bar{\omega}V$ .*

Evidently, we will say that the family  $u_h$  converges weakly to  $u \in V$  if  $p_h u_h$  converges weakly in  $F$  to  $\bar{\omega}u$  (and similar for strong convergence); we will say that the family  $u_h$  converges discretely to  $u$  if:

$$\lim_{h \rightarrow 0} \|u_h - r_h u\|_h = 0.$$

### 3.4. Generalized Approximation of a Normed Space

For many concrete situations the preceding definitions of approximations of a normed space are quite constraining

and it is useful to introduce a slightly more general concept of approximation in which the restriction operators are less important, and are allowed to be nonlinear.

**DEFINITION 8.** *A generalized external approximation of a normed space  $V$  is a set consisting of*

- (i) *a normed space  $F$  and an isomorphism  $\bar{\omega}$  of  $V$  into  $F$ ;*
- (ii) *a family of triples  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$ , in which for each  $h$ ,  $V_h$  is a normed space,  $p_h$  a continuous linear map of  $V_h$  into  $F$ ,  $r_h$  a (possibly nonlinear) map of  $V$  into  $V_h$ .*

When  $F=V$  and  $\bar{\omega}$  is the identity, we get of course a generalized internal approximation of  $V$ .

Definition 2 is easily extended:

**DEFINITION 9.** *For given  $h$ ,  $u \in V$  and  $u_h \in V_h$ , we say that*

- (i)  *$\|\bar{\omega}u - p_h u_h\|_F$  is the error between  $u$  and  $u_h$ ;*
- (ii)  *$\|u_h - r_h u\|_h$  is the discrete error between  $u$  and  $u_h$ ;*
- (iii)  *$\|\bar{\omega}u - p_h r_h u\|$  is the truncation error of  $u$ .*

Since the restriction operators  $r_h$  may be nonlinear, it is not possible to define their norm and we must relax the condition on the norm of  $r_h$  for the concept of stable approximation.

**DEFINITION 10.** *Stable approximation. The prolongation operators  $p_h$  are called stable if their norm*

$$\|p_h\| = \sup_{\substack{u_h \in V_h \\ \|u_h\|_h \leq 1}} \|p_h u_h\|_F$$

*can be majorized independently of  $h$ .*

*The approximation of the space  $V$  is said to be stable, if just the prolongation operation operators are stable.*

A convergent generalized approximation is defined exactly as in Definition 7.

**DEFINITION 11.** *We will say that a generalized external approximation of a normed space  $V$  is convergent if*

(i) *for all  $u \in V$*

$$\lim_{h \rightarrow 0} p_h r_h u = \bar{\omega} u$$

*in the strong topology of  $F$ .*

(ii) *for every sequence  $u_{h'}$  of elements of  $V_{h'}$  ( $h' \rightarrow 0$ ) such that  $p_{h'} u_{h'}$  converges to some element  $\varphi$  in the weak topology of  $F$ , we have  $\varphi \in V$ .<sup>3</sup>*

A family  $u_h$  will be said to converge weakly (or strongly or discretely) to some limit  $u \in V$ , if  $p_h u_h$  converges weakly in  $F$  to  $\bar{\omega} u$  (or strongly in  $F$  or if  $\lim_{h \rightarrow 0} \|u_h - r_h u\|_h = 0$ ).

As in Proposition 2,

**PROPOSITION 3.** *If a generalized external approximation of a space  $V$  is stable and convergent, then discrete convergence implies strong convergence.*

The concept of generalized approximation prevails for the following reason: *it is sufficient for such approximations to give the restriction operators on a dense subspace of  $V$ .*

**PROPOSITION 4.** *Let a generalized external approximation*

be given of a normed space  $V$ , the restriction operators being only defined on a dense subspace  $\mathcal{V}$  of  $V$ . Let us suppose that this approximation is stable and convergent, in the sense that property (i) of Definition 11 is only true for  $u \in \mathcal{V}$ .

Then it is possible to extend the domain of definition of the restriction operators to the whole space  $V$  in such a way that property (i) is true for each  $u \in V$ .<sup>4</sup>

*Proof.* Let  $u \in V$ ,  $u \notin \mathcal{V}$ ; we must in some way define  $r_h u_h \in V_h$ ,  $\forall h$ , such that  $p_h r_h u \rightarrow \bar{\omega} u$  as  $h \rightarrow 0$ .

For each integer  $n \geq 1$ , there exists  $u_n \in \mathcal{V}$  such that  $\|u_n - u\| \leq c_0/n$ ,  $c_0 > 0$ , and hence:

$$\|\bar{\omega} u_n - \bar{\omega} u\|_F \leq \|\bar{\omega}\| \|u_n - u\| \leq \|\bar{\omega}\| c_0/n = 1/n,$$

where  $c_0$  denotes the inverse of the norm of  $\bar{\omega}$ .

Now, for each fixed  $n$ ,  $p_h r_h u_n$  converges to  $\bar{\omega} u_n$  in  $F$  as  $h \rightarrow 0$ . Hence there exists an  $\eta_n > 0$  such that  $|h| \leq \eta_n$  implies

$$\|p_h r_h u_n - \bar{\omega} u_n\|_F \leq 1/n.$$

We can suppose that  $\eta_n$  is smaller than  $\eta_{n-1}$  so that the  $\eta_n$  form a strictly decreasing sequence:

$$0 < \dots < \eta_{n+1} < \eta_n < \dots < \eta_1.$$

Let us define  $r_h u$  by setting:

$$r_h u = r_h u_n \quad \text{for} \quad \eta_{n+1} < |h| \leq \eta_n.$$

It is clear that for  $\eta_{n+1} < |h| \leq \eta_n$

$$\|\bar{\omega}u - p_h r_h u\|_F \leq \|\bar{\omega}u - \bar{\omega}u_n\|_F + \|\bar{\omega}u_n - p_h r_h u_n\|_F$$

and consequently

$$\|\bar{\omega}u - p_h r_h u\|_F \leq 2/n,$$

for  $|h| \leq \eta_n$ .

This implies the convergence of  $p_h r_h u$  to  $\bar{\omega}u$  as  $h \rightarrow 0$ . This completes the proof.

### *Galerkin approximation of a normed space*

We defined in Section 3.2 the Galerkin approximation of a separable Hilbert space only. Using the generalized approximation, we can now define the Galerkin approximation of any separable normed space  $V$ .

Let  $V_n, n \in \mathbb{N}$  be an increasing sequence of finite dimensional subspaces of  $V$  whose union is dense in  $V$ . For each  $n \in \mathbb{N}$ , let  $p_n$  be the canonical injection of  $V_n$  into  $V$  and for any  $u \in V_{n_0}$ ,  $r_n u = 0$ , if  $n \leq n_0$ ,  $r_n u = u$ , if  $n \geq n_0$ . It is clear that  $p_n r_n u \rightarrow u$  as  $n \rightarrow \infty$ , for any  $u \in \bigcup_{n \in \mathbb{N}} V_n$ . The operator  $r_n$  is only defined on  $\mathcal{V} = \bigcup_{n \in \mathbb{N}} V_n$ , but since  $\mathcal{V}$  is dense in  $V$ , all the hypotheses of Proposition 4 are satisfied. We have thus defined a generalized internal approximation of  $V$ . It is a Galerkin approximation.

*In the sequel we will often use generalized approximations with restriction operators only defined on a dense subspace  $\mathcal{V}$  of  $V$ .*

## NOTES

- <sup>1</sup> This remark is not essential for the sequel.
- <sup>2</sup> The strong topology of  $V$  is the topology defined by the norm; the weak topology is the topology  $\sigma(V, V')$  (cf. Appendix).
- <sup>3</sup> Point (ii) is automatically satisfied for internal approximations.
- <sup>4</sup> All conditions for a generalized stable convergent approximation of  $V$  are then met. The situation here should not be compared with the situation in Proposition 1.

## CHAPTER 4

### APPROXIMATION OF LINEAR VARIATIONAL PROBLEMS

We study here the approximate solution of Equation (1.2) in a general setting, that covers the method of Galerkin, which we have already seen, as well as the finite difference and finite element methods. Examples are worked out in Sections 11.2, 11.3 and 12.2.

#### 4.1. Approximate Equations

Let  $V$  be a real Hilbert space,  $a(u, v)$  a continuous and coercive bilinear form on  $V \times V$  and let  $l \in V'$ ; we know that there exists a unique  $u \in V$  for which (1.2) holds.

With respect to the approximation of this element  $u$ , let a generalized internal<sup>1</sup>, stable and convergent Hilbert approximation be given of the space  $V$ , say  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$ . Likewise, for each  $h \in \mathcal{H}$  let be given:

(i) a continuous bilinear form  $a_h(u_h, v_h)$  on  $V_h \times V_h$ , which is coercive and which, more precisely, satisfies:

$$(4.1) \quad \begin{aligned} &\exists \alpha_0 > 0 \text{ independent of } h \text{ such that:} \\ &a_h(u_h, u_h) \geq \alpha_0 \|u_h\|_h^2, \quad \forall u_h \in V_h; \end{aligned}$$

(ii) a continuous linear form on  $V_h$ ,  $l_h \in V'_h$  such that:

$$(4.2) \quad \|l_h\|_{*h} \leq \beta,$$

in which  $\|\cdot\|_{*h}$  stands for the norm of  $V'_h$  and in which  $\beta$  is independent of  $h$ .

We associate now with Equation (1.2) the following family of approximate equations:

*For fixed  $h \in \mathcal{H}$ , to find a  $u_h \in V_h$  such that:*

$$(4.3) \quad a_h(u_h, v_h) = \langle l_h, v_h \rangle, \quad \forall v_h \in V_h.$$

By the preceding hypotheses, the projection theorem (in which  $V, V', a, l$  are replaced by  $V_h, V'_h, a_h, l_h$ ) says now that Equation (4.3) has a unique solution; we will say that  $u_h$  is the approximate solution of Equation (1.2).

*Remark.* One can introduce a linear operator  $A_h$  mapping  $V_h$  into  $V'_h$ , defined by

$$\langle A_h u_h, v_h \rangle = a_h(u_h, v_h), \quad \forall u_h, v_h \in V_h.$$

Then Equation (4.3) can be written (cf. Chapter 1) as:

$$(4.4) \quad A_h u_h = l_h.$$

## 4.2. Convergence Theorems

We want to study the convergence of  $u_h$  to  $u$ , so we must first make precise the way in which the forms  $a_h$  and  $l_h$  approximate the forms  $a$  and  $l$ . The corresponding hypotheses are usually called *consistency hypotheses*; we will make the following consistency hypotheses:

(4.5) If the family  $v_h$  converges weakly to  $v$  for  $h \rightarrow 0$ , then, for all  $w \in V$

$$\lim_{h \rightarrow 0} a_h(v_h, r_h w) = a(v, w).$$

(4.6) If the family  $v_h$  converges weakly to  $v$  for  $h \rightarrow 0$ , then for all  $w \in V$

$$\lim_{h \rightarrow 0} a_h(r_h w, v_h) = a(w, v).$$

(4.7) For all  $v \in V$

$$\lim_{h \rightarrow 0} a_h(r_h v, r_h v) = a(v, v).$$

(4.8) If the family  $v_h$  converges weakly to  $v$  for  $h \rightarrow 0$ , then:

$$\lim_{h \rightarrow 0} \langle l_h, v_h \rangle = \langle l, v \rangle.$$

Using these hypotheses we will prove the following results:

**THEOREM 1.** *Under hypotheses (1.1), (4.1), (4.2), (4.5) and (4.8), the solution  $u_h$  of (4.3) converges weakly to the solution  $u$  of (1.2), for  $h \rightarrow 0$ .*

**THEOREM 2.** *Under hypotheses (1.1), (4.1), (4.2) and (4.5) to (4.8) the solution  $u_h$  of (4.3) converges strongly and discretely to the solution  $u$  of (1.2), for  $h \rightarrow 0$ .*

*Proof of Theorem 1.* Putting  $v_h = u_h$  in (4.3) and using (4.1) and (4.2), we find:

$$(4.9) \quad \begin{aligned} a_h(u_h, u_h) &= \langle l_h, u_h \rangle \\ \alpha_0 \|u_h\|_h^2 &\leq \|l_h\|_{*h} \|u_h\|_h \leq \beta \|u_h\|_h; \end{aligned}$$

hence:

$$(4.10) \quad \|u_h\|_h \leq \beta/\alpha_0.$$

As the operators  $p_h$  are stable, there exists a constant  $c_0$  which majorizes the norm of these operators:

$$(4.11) \quad \|p_h\| \leq c_0,$$

and hence:

$$(4.12) \quad \|p_h u_h\| \leq c_0 \beta / \alpha_0.$$

Under these conditions, there exists an  $u_* \in V$  and a sequence  $h'$  converging to 0, such that:

$$(4.13) \quad \lim_{h' \rightarrow 0} p_{h'} u_{h'} = u$$

in the weak topology of  $V$ .

For a fixed  $v \in V$ , we write (4.3) with  $v_h = r_h v$  and we take the limit with the sequence  $h'$ ; by using (4.5), (4.8) and (4.13), we get:

$$\begin{aligned} a_h(u_h, r_h v) &= \langle l_h, r_h v \rangle \\ \lim_{h' \rightarrow 0} a_{h'}(u_{h'}, r_{h'} v) &= a(u_*, v) \\ \lim_{h' \rightarrow 0} \langle l_{h'}, r_{h'} v \rangle &= \langle l, v \rangle. \end{aligned}$$

Finally:

$$a(u_*, v) = \langle l, v \rangle,$$

and since  $v \in V$  is arbitrary,  $u_*$  is solution of (1.2),  $u_* = u$ .

It can be shown in the same way, that from every subsequence of  $p_h u_h$ , one can extract a subsequence which converges in the weak topology of  $V$  to  $u$ . This proves that the family  $p_h u_h$  as a whole converges to  $u$  in the weak topology, for  $h \rightarrow 0$ .

*Proof of Theorem 2.* We can apply Theorem 1, and this shows that

$$(4.14) \quad \lim_{h \rightarrow 0} p_h u_h = u \quad \text{in the weak topology of } V.$$

Let us now consider the expression:

$$\begin{aligned} X_h &= a_h(u_h - r_h u, u_h - r_h u) \\ X_h &= a_h(u_h, u_h) - a_h(u_h, r_h u) - a_h(r_h u, u_h) + \\ &\quad + a_h(r_h u, r_h u). \end{aligned}$$

By (4.6), (4.7) and (4.14):

$$\begin{aligned} \lim_{h \rightarrow 0} a_h(u_h, r_h u) &= a(u, u) \\ \lim_{h \rightarrow 0} a_h(r_h u, u_h) &= a(u, u). \end{aligned}$$

By (4.8)

$$\lim_{h \rightarrow 0} a_h(r_h u, r_h u) = a(u, u).$$

Finally, by (4.8), (4.9) and (4.14):

$$\begin{aligned} a_h(u_h, u_h) &= \langle l_h, u_h \rangle \\ \lim_{h \rightarrow 0} a_h(u_h, u_h) &= \langle l, u \rangle. \end{aligned}$$

After rearranging we find:

$$\lim_{h \rightarrow 0} X_h = \langle l, u \rangle - a(u, u) = 0.$$

With (4.1), we get now:

$$0 \leq \alpha_0 \|u_h - r_h u\|_h^2 \leq X_h \rightarrow 0,$$

which shows that  $u_h$  converges discretely to  $u$ . According to Proposition 3 of Chapter 3 this implies strong convergence of  $u_h$  to  $u$ .

### 4.3. Miscellaneous Remarks

#### *The case of external approximation*

The preceding is also valid if a generalized external Hilbert approximation of  $V$  is given (this is actually the case in the applications of Chapters 11 and 12). There are no modifications required in Section 4.1; for the proof of Theorem 1, we observe that the sequence  $p_h u_h$  is bounded in  $F$  and that hence there exist a  $\varphi \in F$  and a sequence  $h' \rightarrow 0$  such that:

$$\lim_{h' \rightarrow 0} p_{h'} u_{h'} = \varphi \quad \text{in the weak topology of } F.$$

By definition of a convergent external approximation  $\varphi$  is

of the form  $\bar{\omega}u_*$ ,  $u_* \in V$ ; so we prove that  $u_* = u$  and that the family  $p_h u_h$  converges as a whole to  $\bar{\omega}u$ , in the weak topology of  $F$ . Theorem 2 stays valid without modifications.

### *The Galerkin method*

The Galerkin method fits in the framework of Sections 4.1 and 4.2. If  $\{V_h, p_h, r_h\}_{h \in \mathbf{N}}$  is a Galerkin approximation of the space  $V$ , we put:

$$a_h(u_h, v_h) = a(u_h, v_h), \quad \forall u_h, v_h \in V_h,$$

and

$$\langle l_h, u_h \rangle = \langle l, u_h \rangle, \quad \forall u_h \in V_h (V_h \subset V).$$

Hypotheses (4.1), (4.2) and (4.5) to (4.8) are satisfied, and hence Theorem 2 is applicable.

If  $w_1, w_2, \dots$ , is a free and total sequence in  $V$ , and if  $V_h$  is the space generated by  $w_1, \dots, w_n$  (cf. Section 3.2), then one may verify that problem (4.2) is identical with problem (2.2), and so Theorem 2 also contains Theorem 1 of Chapter 2.

### *Restriction schemes*

Another particular case of the situation considered in Sections 4.1 and 4.2 is the following.

Suppose there is given a family of finite dimensional vector spaces  $V_h$ ,  $h \in \mathcal{H}$  (not necessarily *a priori* provided with a Hilbert structure) and, for each  $h \in \mathcal{H}$ , an *injective* linear operator  $p_h$  from  $V_h$  into  $V$  and a linear operator  $r_h$  from  $V$  into  $V_h$ . We suppose that

- (i)  $p_h r_h$  is a continuous linear operator in  $V$ ;
- (ii) for all  $v \in V$ :

$$\lim_{h \rightarrow 0} p_h r_h v = v \quad \text{in the strong topology of } V;$$

- (iii) the norms of the operators  $p_h r_h$  have an upper bound independent of  $h$ .<sup>2</sup>

We provide the space  $V_h$  with a scalar product:

$$((u_h, v_h))_h = ((p_h u, p_h v)),$$

and the  $V_h$  become Hilbert spaces. It is easy to see that  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  is a stable and convergent approximation of  $V$ .

Subsequently we define  $a_h$  and  $l_h$ :

$$\begin{aligned} a_h(u_h, v_h) &= a(p_h u_h, p_h v_h), \quad \forall u_h, v_h \in V_h \\ \langle l_h, v_h \rangle &= \langle l, p_h v_h \rangle, \quad \forall v_h \in V_h. \end{aligned}$$

The hypotheses of Theorem 2 are satisfied. This viewpoint has been developed systematically by Aubin [10].<sup>3</sup>

### *Resolution of (4.3)*

In most practical situations,  $V_h$  is a finite dimensional space with dimension  $J_h$ , and (4.3) is equivalent to a system of linear equations. If a basis of  $V_h$  is known, say  $\{w_{hj}\}_{1 \leq j \leq J_h}$  and if we set:

$$u_h = \sum_{j=1}^{J_h} \xi_{hj} w_{hj},$$

then (4.3) is equivalent to

$$a_h(u_h, w_{hk}) = \langle l_h, w_{hk} \rangle, \quad 1 \leq k \leq J_h,$$

or

$$(4.15) \quad \sum_{j=1}^{J_h} \xi_{hj} a_h(w_{hj}, w_{hk}) = \langle l_h, w_{hk} \rangle, \quad 1 \leq k \leq J_h.$$

This is an algebraic linear system for the components  $\xi_{hj}$  of  $u_h$ . The resolution of the system is made by direct or iterative methods adapted to its size and to the properties of its matrix  $\{a_h(w_{hj}, w_{hk})\}_{1 \leq j, k \leq J_h}$ . See, for example Varga [17].

## NOTES

<sup>1</sup> One can also suppose that a generalized external Hilbert approximation of  $V$  is given (cf. Section 4.3).

<sup>2</sup> This follows in fact from (i) and (ii) by the theorem of Banach-Steinhaus.

<sup>3</sup> Numbers in square brackets refer to the Bibliography on p. 161.

## CHAPTER 5

### A METHOD FOR COMPUTATION OF THE ERROR

We go back to the situation of Chapter 4 and more precisely to the framework of Theorem 2, and we want to estimate the error between  $u$  and  $u_h$ . This is not possible merely with the hypotheses of Chapter 4, and we are only going to indicate some methods for error computation, which subsequently must be developed in the examples.

#### 5.1. Error Computation

We set

$$(5.1) \quad \varepsilon_h(l) = \sup_{\substack{v_h \in V_h \\ \|v_h\|_h \leq 1}} \{ |\langle l, p_h v_h \rangle - \langle l_h, v_h \rangle| \}$$

$$(5.2) \quad \eta_h(u) = \sup_{\substack{v_h \in V_h \\ \|v_h\|_h \leq 1}} \{ |a(u, p_h v_h) - a_h(r_h u, v_h)| \}.$$

**THEOREM 1.** *Under hypotheses (1.1), (4.1) and (4.2), if  $u$  stands for the solution of (1.2) and  $u_h$  for the solution of (4.3), then we have:*

$$(5.3) \quad \|u_h - r_h u\|_h \leq \frac{1}{\alpha_0} \{ \varepsilon_h(l) + \eta_h(u) \}$$

$$(5.4) \quad \|u - p_h u_h\| \leq \|u - p_h r_h u\| + \frac{c_0}{\alpha_0} \{\varepsilon_h(l) + \eta_h(u)\},$$

where  $c_0$  is a constant that majorizes the norms of the operators  $p_h$ .

*Proof.* We have:

$$\begin{aligned} X_h &= a_h(u_h - r_h u, u_h - r_h u) = \\ &= a_h(u_h, u_h - r_h u) - a_h(r_h u, u_h - r_h u) = \\ &= \langle l_h, u_h - r_h u \rangle - a_h(r_h u, u_h - r_h u) = \\ &= \langle l_h, u_h - r_h u \rangle - \langle l, p_h(u_h - r_h u) \rangle + \\ &\quad + a(u, p_h(u_h - r_h u)) - a_h(r_h u, u_h - r_h u) \end{aligned}$$

(we have used (4.3) with  $v_h = u_h - r_h u$  and (1.2) with  $v = p_h(u_h - r_h u)$ )

According to (5.1) and (5.2):

$$\begin{aligned} |\langle l_h, u_h - r_h u \rangle - \langle l, p_h(u_h - r_h u) \rangle| &\leq \varepsilon_h(l) \|u_h - r_h u\|_h \\ |a(u, p_h(u_h - r_h u)) - a_h(r_h u, u_h - r_h u)| &\leq \\ &\leq \eta_h(u) \|u_h - r_h u\|_h, \end{aligned}$$

and hence:

$$X_h \leq \{\varepsilon_h(l) + \eta_h(u)\} \|u_h - r_h u\|_h.$$

Using (4.1), we derive (5.3). Obtaining (5.4) from (5.3) is evident, by Proposition 3 of Chapter 3.

## 5.2. The Case of Restriction Schemes

In the case of restriction schemes as considered in Section

4.3, we can give a better estimate of the error between  $u$  and  $u_h$ .

**THEOREM 2.** *In the case of restriction schemes, the error between the solution  $u$  of (1.2) and the solution  $u_h$  of (4.3) is majorized by the truncation error of  $u$ :*

$$(5.5) \quad \|u - p_h u_h\| \leq \frac{c_1}{\alpha} \|u - p_h r_h u\|,$$

where:

$$(5.6) \quad c_1 = \sup_{\|u\| \leq 1, \|v\| \leq 1} |a(u, v)|.$$

*Proof.* For all  $v_h \in V_h$ :

$$a(u - p_h u_h, p_h v_h) = 0,$$

and in particular:

$$a(u - p_h u_h, p_h(r_h u - u_h)) = 0.$$

Hence:

$$a(u - p_h u_h, u - p_h u_h) = a(u - p_h u_h, u - p_h r_h u)$$

$$\alpha \|u - p_h u_h\|^2 \leq c_1 \|u - p_h u_h\| \|u - p_h r_h u\|,$$

which proves the theorem.

## CHAPTER 6

### THE METHOD OF FRACTIONARY STEPS

We are now going to study a new method of solving Equation (1.2). In fact, the following can also be applied to the solution of the discretized Equation (4.3).

#### 6.1. Hypotheses

Let two Hilbert spaces  $V$  and  $H$  be given,  $V$  contained in  $H$ ,  $V$  dense in  $H$ , the injection of  $V$  into  $H$  being continuous. We denote the scalar product in  $V$  by  $((., .))$ , in  $H$  by  $(., .)$  and the norm by  $\|.\|$  and  $|\cdot|$  respectively.

Let likewise be given  $m$  Hilbert spaces  $V_1, \dots, V_m$  with scalar products  $((., .))_i$  and norms  $\|.\|_i$ , such that:

$$(6.1) \quad V \subset V_i \subset H,$$

$$(6.2) \quad \text{the injections of } V \text{ into } V_i \text{ and of } V_i \text{ into } H \text{ are continuous,}$$

$$(6.3) \quad V = \bigcap_{i=1}^m V_i.$$

As in Section 1.2, we may identify  $H'$  with a dense subspace of  $V'_i$  and  $V'_i$  with a dense subspace of  $V'$  ( $X'$

denotes the dual of  $X$ ):

$$H' \subset V'_i \subset V'.$$

If we moreover identify  $H$  and  $H'$ , we have

$$(6.4) \quad V \subset V_i \subset H \subset V'_i \subset V', \quad (1 \leq i \leq m)$$

where every space is dense in the following and in which the injections are continuous. We denote the duality between  $V$  and  $V'$  by  $\langle \cdot, \cdot \rangle$ .

We suppose, as usual, given a continuous and coercive bilinear form  $a(u, v)$  on  $V \times V$ :

$$(6.5) \quad a(u, u) \geq \alpha \|u\|^2 \quad \forall u \in V, \quad \alpha > 0,$$

and we suppose that  $a(u, v)$  can be decomposed as follows:

$$(6.6) \quad a(u, v) = \sum_{i=1}^m a_i(u, v) \quad \forall u, v \in V,$$

where  $a_i(u, v)$  is a continuous and coercive bilinear form on  $V \times V$ :

$$(6.7) \quad a_i(u, u) \geq \alpha_i \|u\|_i^2 \quad \forall u \in V, \quad \alpha_i > 0.$$

As the injection of  $V_i$  into  $H$  is continuous, (6.7) implies that there is a constant  $\gamma_i$  such that:

$$(6.8) \quad a_i(u, u) \geq \gamma_i |u|^2 \quad \forall u \in V_i, \quad \gamma_i > 0.$$

We denote by  $A \in \mathcal{L}(V, V')$  (or  $A_i \in \mathcal{L}(V_i, V'_i)$ ) the operator defined by:

$$(6.9) \quad a(u, v) = \langle Au, v \rangle \quad \forall u, v \in V$$

(or:  $a_i(u, v) = \langle A_i u, v \rangle \quad \forall u, v \in V_i$ ).

According to (6.6) and (6.9):

$$Au = \sum_{i=1}^m A_i u, \quad \forall u \in V.$$

We know (cf. Section 1.2) that for given  $f$  in  $H$ , there exists a unique element  $u$  of  $V$  such that:

$$(6.10) \quad a(u, v) = (f, v) \quad \forall v \in V,$$

that is

$$(6.11) \quad Au = f,$$

and this element  $u$  we want to approximate.

*We will make the following regularity hypothesis:*

$$(6.12) \quad A_i u \in H, \quad i = 1, \dots, m.$$

*Remark.* The situation becomes much more simple, if the space  $H$  is finite dimensional, which happens in particular when (6.10) is the discrete version of a certain problem (1.2). In this case the spaces  $V$  and  $V_i$  equal all  $H$ , and all carry the same or equivalent norms; in that case condition (6.12) is automatically satisfied.

## 6.2. Approximated Equations

We choose an arbitrary decomposition of  $f$ :

$$(6.13) \quad f = \sum_{i=1}^m f_i \quad f_i \in H$$

(for example,  $f_1 = f, f_i = 0; i = 2, \dots, m$ ).

For each integer  $N$  and each positive real number  $\tau$ , we will recursively define a family of elements of  $H$ ,  $u^{r+i/m}$ ,  $0 \leq r \leq N-1$ ,  $1 \leq i \leq m$ .

To begin with, we choose  $u^0$  arbitrary in  $H$ . When  $u^0, \dots, u^{r+(i-1)/m}$  are known,  $u^{r+i/m}$  is defined as the solution in  $V_i$  of:

$$(6.14) \quad (u^{r+i/m} - u^{r+(i-1)/m}, v) + \tau a_i(u^{r+i/m}, v) = \tau(f_i, v), \quad \forall v \in V_i.$$

This equation can also be written as:

$$(6.15) \quad u^{r+i/m} - u^{r+(i-1)/m} + \tau A_i u^{r+i/m} = \tau f_i$$

The existence and uniqueness of a solution for (6.14) or (6.15) follow directly from the projection theorem.

*Remark.* It is possible (and very useful in practice) to take parameters  $\tau$  that depend on  $r$  ( $\tau = \tau_r$ ). Theorems 1 and 2 hereunder can be adapted to this situation.

### 6.3. The Convergence Theorem and the Error Estimate

We want to compare the last element  $u^N = u^N(\tau)$  of the sequence  $u^{r+i/m}$  defined in Section 6.2 with the solution  $u$  of (6.10).

**THEOREM 1.** *Under the hypotheses (6.1) to (6.8) and (6.12), for each integer  $n$  ( $0 \leq n \leq N$ ) we have:*

$$(6.16) \quad |u^n - u|^2 \leq \varepsilon_n,$$

where

$$(6.17) \quad \varepsilon_n = \varepsilon_0(1 + \gamma\tau)^{-n} + \delta\tau^2 \sum_{j=0}^{n-1} (1 + \gamma\tau)^{j-n}$$

$$(6.18) \quad \varepsilon_0 = |u^0 - u|^2$$

$$(6.19) \quad \delta = m \left\{ \sum_{i=1}^m |f_i - A_i u| \right\}^2$$

$$(6.20) \quad \gamma = 2\gamma_m.$$

Then it immediately follows:

**THEOREM 2.** *Under the same hypotheses, if  $\tau \rightarrow 0$  and  $N \rightarrow \infty$  so that:*

$$(6.21) \quad \tau N \rightarrow \infty,$$

*then:*

$$(6.22) \quad \lim u^N(\tau) = u \quad \text{in the strong topology of } H.$$

*Proof of Theorem 1.* Let us put  $v^{r+i/m} = u^{r+i/m} - u$ . Hence, according to (6.15):

$$v^{r+i/m} - v^{r+(i-1)/m} + \tau A_i v^{r+i/m} = \tau(f_i - A_i u).$$

It follows that

$$\begin{aligned} (v^{r+i/m} - v^{r+(i-1)/m}, v^{r+i/m}) + \tau a_i(v^{r+i/m}, v^{r+i/m}) = \\ = \tau(f_i - A_i u, v^{r+i/m}). \end{aligned}$$

One easily verifies that:

$$\begin{aligned} 2(v^{r+i/m} - v^{r+(i-1)/m}, v^{r+i/m}) = \\ = |v^{r+i/m}|^2 - |v^{r+(i-1)/m}|^2 + |v^{r+i/m} - v^{r+(i-1)/m}|^2, \end{aligned}$$

and hence:

$$(6.23) \quad |v^{r+i/m}|^2 - |v^{r+(i-1)/m}|^2 + |v^{r+i/m} - v^{r+(i-1)/m}|^2 \\ + 2\tau a_i(v^{r+i/m}, v^{r+i/m}) = 2\tau(f_i - A_i u, v^r) \\ + 2\tau(f_i - A_i u, v^{r+i/m} - v^r).$$

For fixed  $r$ , we add Equations (6.23) corresponding to  $i = 1, \dots, m$ . Since:

$$\sum_{i=1}^m (f_i - A_i u) = f - Au = 0$$

we find:

$$(6.24) \quad |v^{r+1}|^2 - |v^r|^2 + \sum_{i=1}^m |v^{r+i/m} - v^{r+(i-1)/m}|^2 \\ + 2\tau \sum_{i=1}^m a_i(v^{r+i/m}, v^{r+i/m}) = 2\tau \sum_{i=1}^m (f_i - A_i u, v^{r+i/m} - v^r).$$

We write, using (6.8) and (6.20):

$$(6.25) \quad 2a_m(v^{r+1}, v^{r+1}) \geq \gamma |v^{r+1}|^2,$$

and we majorize the right-hand member of (6.24) by:

$$2\tau \sum_{i=1}^m |f_i - A_i u| |v^{r+i/m} - v^r| \leq \\ \leq 2\tau \sum_{i=1}^m |f_i - A_i u| \left\{ \sum_{j=1}^i |v^{r+j/m} - v^{r+(j-1)/m}| \right\} \\ \leq 2\tau \left\{ \sum_{i=1}^m |f_i - A_i u| \right\} \left\{ \sum_{j=1}^m |v^{r+j/m} - v^{r+(j-1)/m}| \right\}.$$

Using the Schwarz inequality:

$$\begin{aligned} \sum_{j=1}^m |v^{r+j/m} - v^{r+(j-1)/m}| &\leq \\ &\leq m^{1/2} \left\{ \sum_{j=1}^m |v^{r+j/m} - v^{r+(j-1)/m}|^2 \right\}^{1/2}, \end{aligned}$$

So we can majorize the right-hand member of (6.24) by:

$$\begin{aligned} &2\tau m^{1/2} \left\{ \sum_{i=1}^m |f_i - A_i u| \right\} \left\{ \sum_{j=1}^m |v^{r+j/m} - v^{r+(j-1)/m}|^2 \right\}^{1/2} \\ &\leq \tau^2 m \left\{ \sum_{i=1}^m |f_i - A_i u| \right\}^2 + \sum_{j=1}^m |v^{r+j/m} - v^{r+(j-1)/m}|^2. \end{aligned}$$

With this last inequality and (6.25), inequality (6.24) gives:

$$(6.26) \quad (1 + \gamma\tau) |v^{r+1}|^2 - |v^r|^2 \leq \delta\tau^2, \quad 0 \leq r \leq N-1.$$

Multiplying (6.26) by  $(1 + \gamma\tau)^r$  and adding these inequalities then for  $r=0, \dots, n-1$ , we obtain precisely (6.16)–(6.20).

## **PART II**

# **APPLICATIONS**

## CHAPTER 7

### SPACES OF FUNCTIONS ASSOCIATED WITH AN OPEN SET IN $\mathbf{R}^n$

Let  $\Omega$  be an open set of  $\mathbf{R}^n$  with boundary  $\Gamma$ ; we will associate with this open set several spaces of functions for later use.

#### 7.1. Definitions

##### $L^p(\Omega)$ spaces

For  $1 \leq p < \infty$ ,  $L^p(\Omega)$ , denotes the space of (classes of) real valued functions on  $\Omega$  whose  $p$ -th power is summable with respect to the Lebesgue measure  $dx$ ; it is known that this space is a Banach space when it is equipped with the norm:

$$\left\{ \int_{\Omega} |f(x)|^p dx \right\}^{1/p}.$$

For  $p=2$ , the space  $L^2(\Omega)$  is a Hilbert space for the scalar product:

$$(f, g) = \int_{\Omega} f(x) g(x) dx,$$

and we set:

$$|f| = (f, f)^{1/2}.$$

$L^\infty(\Omega)$  is the space of (classes of) real valued measurable functions in  $\Omega$  which are essentially bounded with respect to the Lebesgue measure. This is a Banach space for the norm:

$$\text{ess. sup.}_\Omega |f(x)|.$$

*Spaces  $\mathcal{D}(\Omega)$ ,  $\mathcal{D}(\bar{\Omega})$*

$\mathcal{D}(\Omega)$  (or  $\mathcal{D}(\bar{\Omega})$ ) is the space of indefinitely differentiable functions with compact support in  $\Omega$  (or  $\bar{\Omega}$ ).

It is known that  $\mathcal{D}(\Omega)$  is dense in each  $L^p(\Omega)$  space, for  $1 \leq p < \infty$ .

*Sobolev spaces*

We denote by  $D_i$  the differential operator  $\partial/\partial x_i$ , and if  $j$  is a multi-index,  $j = (j_1, \dots, j_n) \in \mathbb{N}^n$ , we denote:

$$|j| = j_1 + \dots + j_n$$

$$D^j = D_1^{j_1} \dots D_n^{j_n}.$$

$D_i^0$  and  $D_0$  denote the identity operator.

**DEFINITION.** For each integer  $m$ ,  $H^m(\Omega)$  is the space of functions of  $L^2(\Omega)$  whose derivatives (in the sense of distributions) of order less than or equal to  $m$  are in  $L^2(\Omega)$ .

Space  $H^m(\Omega)$  is evidently a linear subspace of  $L^2(\Omega)$ ; for  $u$  and  $v$  in  $H^m(\Omega)$ , we write:

$$(7.1) \quad ((u, v)) = \sum_{|j| \leq m} (D^j u, D^j v).$$

LEMMA. *The space  $H^m(\Omega)$  equipped with the scalar product (7.1) is a Hilbert space.*

*Proof.* It is clear that (7.1) defines a symmetric positive definite bilinear form on  $H^m(\Omega)$ . We show now that  $H^m(\Omega)$  is complete with respect to the norm  $\|\cdot\|$  associated with this scalar product.

Let  $\{u_r\}$  be a Cauchy sequence in  $H^m(\Omega)$ . For each multi-index  $j$ ,  $|j| \leq m$ , the sequence  $\{D^j u_r\}$  is a Cauchy sequence in  $L^2(\Omega)$  and hence convergent.

We write:

$$g = \lim_{r \rightarrow \infty} u_r, \quad g_j = \lim_{r \rightarrow \infty} D^j u_r.$$

So we have:

$$u_r \rightarrow g, \quad D^j u_r \rightarrow g_j$$

in the distribution sense in  $\Omega$ ; as  $D^j$  is a continuous operator in the distribution space, we have likewise:

$$D^j u_r \rightarrow D^j g$$

in this space, and hence  $D^j g = g_j \in L^2(\Omega)$  for all  $j$ ,  $|j| \leq m$ . This proves that  $g \in H^m(\Omega)$  and that the sequence  $u_r$  converges to  $g$  in  $H^m(\Omega)$ .

DEFINITION. *We denote by  $H_0^m(\Omega)$  the closure of  $\mathcal{D}(\Omega)$  in  $H^m(\Omega)$ .*

The space  $H_0^m(\Omega)$  is evidently provided with the Hilbert structure induced from  $H^m(\Omega)$ .

If  $\Omega = \mathbf{R}^n$ , it can be shown that:

$$(7.2) \quad H_0^m(\mathbf{R}^n) = H^m(\mathbf{R}^n),$$

but when  $\Omega$  is not the whole of  $\mathbf{R}^n$ ,  $H_0^m(\Omega)$  and  $H^m(\Omega)$  are in general different. The trace theorems below give, in case the boundary of  $\Omega$  is sufficiently smooth, a simple characterization of  $H_0^m(\Omega)$ .

## 7.2. Properties of Sobolev Spaces

We recall, without proof, some properties of Sobolev spaces. We will have to make some regularity hypotheses about the boundary  $\Gamma$  of  $\Omega$ , as, for example:

$$(7.3) \quad \Gamma \text{ is a manifold of dimension } n-1, \text{ of class } \mathcal{C}^r \text{ (} r \text{ to be specified) and } \Omega \text{ lies locally on one side of } \Gamma.$$

Suppose that (7.3) holds with  $r = m$ :

*Density*

$\mathcal{D}(\bar{\Omega})$  is dense in every space  $H^m(\Omega)$ .

*Prolongation operator*

There exists a linear and continuous operator  $\mathcal{P}$

$$\mathcal{P} \in \mathcal{L}(H^s(\Omega), H^s(\mathbf{R}^n)), \quad s = 0, \dots, m$$

such that, for all  $u$ , the restriction of  $\mathcal{P}u$  to  $\Omega$  is equal to  $u$ .

### *Trace theorem*

There exist linear continuous operators  $\gamma_0, \dots, \gamma_{m-1}$  from  $H^m(\Omega)$  into  $L^2(\Gamma)$  (the space of square integrable functions on  $\Gamma$ , with respect to the surface integral  $d\sigma$ ), such that for each  $m$  times continuously differentiable function  $u$  on  $\Omega$ , we have:

$$\begin{aligned}\gamma_0 u &= u|_{\Gamma} = \text{the restriction of } u \text{ to } \Gamma \\ \gamma_s u &= \frac{\partial^s u}{\partial \nu^s} \Big|_{\Gamma} \quad (s = 1, \dots, m-1),\end{aligned}$$

where  $\partial \nu^s$  is the unit normal vector on  $\Gamma$ , pointing outward with respect to  $\Omega$ .

*One can prove that  $H_0^m(\Omega)$  is the kernel of the map:*

$$\gamma_0 \times \gamma_1 \times \dots \times \gamma_{m-1}.$$

The space  $\gamma_j H^m(\Omega)$  is not the whole of  $L^2(\Gamma)$ , but a subspace of it. We will pay special attention to the space  $\gamma_0 H^1(\Omega)$ , which we denote by  $H^{1/2}(\Gamma)$ ; this space can be endowed with the Hilbert structure transported by  $\gamma_0$ , but there exist more direct procedures to study this space.

### *Generalized Green's formula*

We now suppose that (7.3) holds with  $r=2$ . If  $u$  and  $v$  are two twice continuously differentiable functions on  $\bar{\Omega}$ , the

classical Green's formula gives:

$$\sum_{i=1}^n (D_i u, D_i v) = -(\Delta u, v) + \int_{\Gamma} \frac{\partial u}{\partial \nu} v d\sigma.$$

If  $u$  and  $v$  are in  $H^2(\Omega)$ , one may show that a generalized Green's formula holds:

$$(7.4) \quad \sum_{i=1}^n (D_i u, D_i v) = -(\Delta u, v) + (\gamma_1 u, \gamma_0 v)_{L^2(\Gamma)},$$

where  $(., .)$  is the scalar product of  $L^2(\Omega)$  and  $(., .)_{L^2(\Gamma)}$  is the scalar product of  $L^2(\Gamma)$ .

If  $u \in H^1(\Omega)$  and  $\Delta u \in L^2(\Omega)$ , it can be shown that  $\gamma_1 u$  can also be defined as an element of the dual of  $H^{1/2}(\Gamma)$ , which we denote by  $H^{-1/2}(\Gamma)$ . For such a  $u$  and for  $v \in H^1(\Omega)$ , the generalized Green's formula becomes:

$$(7.5) \quad \sum_{i=1}^n (D_i u, D_i v) = -(\Delta u, v) + \langle \gamma_1 u, \gamma_0 v \rangle,$$

where  $\langle ., . \rangle$  stands for the scalar product in the duality between  $H^{-1/2}(\Gamma)$  and  $H^{1/2}(\Gamma)$ .

For a proof of the preceding results and for a systematic study of Sobolev spaces, the reader is referred to Lions and Magenes [8].

#### NOTE

<sup>1</sup> The variable point of  $\Omega$  is denoted by  $x = (x_1, \dots, x_n)$ ;  $dx = dx_1, \dots, dx_n$ .

## CHAPTER 8

### APPROXIMATION OF SOME FUNCTION SPACES BY FINITE DIFFERENCES (I)

In this chapter and in the two following ones, we will construct some approximations to the spaces  $L^2(\Omega)$ ,  $H^1(\Omega)$ ,  $H_0^1(\Omega)$ . These three chapters are rather technical and occasionally we will omit some details of proofs.

In this chapter we construct an internal approximation of  $L^2(\Omega)$  and  $H^1(\Omega)$  for the case  $\Omega = \mathbf{R}^n$ , using finite differences. In Chapter 9 we construct an internal approximation of  $L^2(\Omega)$ , an external approximation of  $H^1(\Omega)$ , and a generalized external approximation of  $H_0^1(\Omega)$  (finite differences). In Chapter 10 we construct two generalized external approximations of  $H_0^1(\Omega)$  using conforming and nonconforming finite elements.

Some results of Chapter 8 are needed in Chapter 9, but Chapter 10 is independent of Chapters 8 and 9, and could be read before these chapters.

#### 8.1. Notations

In this section and the next one  $\mathcal{H}$  is the set of vectors of  $\mathbf{R}^n$  of the form:

$$h = (h_1, \dots, h_n), \quad 0 < h_i \leq h_i^0.$$

For all  $h \in \mathcal{H}$ , we define:

(i)  $\alpha h (\alpha \in \mathbf{R}^n)$  is the point of  $\mathbf{R}^n$  with coordinates:

$$\alpha_1 h_1, \dots, \alpha_n h_n;$$

(ii)  $R_h$  is the set of points of  $\mathbf{R}^n$  of the form:

$$jh = (j_1 h_1, \dots, j_n h_n),$$

where the  $j_i$  are integers of arbitrary sign ( $j \in \mathbf{Z}^n$ ).

(iii)  $\sigma_h(M)$ ,  $M = (\mu_1, \dots, \mu_n)$  is called a *cell*, and it is the set:

$$\prod_{i=1}^n \left[ \mu_i - \frac{h_i}{2}, \mu_i + \frac{h_i}{2} \right];$$

(iv)  $\sigma_h(M, r)$  is the set

$$\bigcup_{\substack{j \in \mathbf{Z}^n \\ |j| \leq r}} \sigma_h(M + \tfrac{1}{2} jh),$$

where:

$$|j| = |j_1| + \dots + |j_n|;$$

(v)  $w_{hM}$  is the characteristic function of the cell  $\sigma_h(M)$ ;

(vi)  $\delta_{ih}$  (or  $\delta_i$  if no confusion can arise) is the finite difference operator:

$$(\delta_i \varphi)(x) = \frac{\varphi(x + \tfrac{1}{2} h_i e_i) - \varphi(x - \tfrac{1}{2} h_i e_i)}{h_i},$$

where  $e_1, \dots, e_n$  is the canonical basis of  $\mathbf{R}^n$ .

If  $j = (j_1, \dots, j_n)$  is a multi-index, then  $\delta_h^j$  (or more simply

$\delta^j$ ) will denote the operator:

$$\delta^j = \delta_1^{j_1}, \dots, \delta_n^{j_n}.$$

Finally, with each open set  $\Omega$  of  $\mathbf{R}^n$  and each non-negative integer  $r$ , we associate the following point sets:

$$\begin{aligned}\dot{\Omega}_h^r &= \{M \mid M \in R_h, \quad \sigma_h(M, r) \subset \Omega\} \\ \Omega_h^r &= \{M \mid M \in R_h, \quad \sigma_h(M, r) \cap \Omega \neq \emptyset\}.\end{aligned}$$

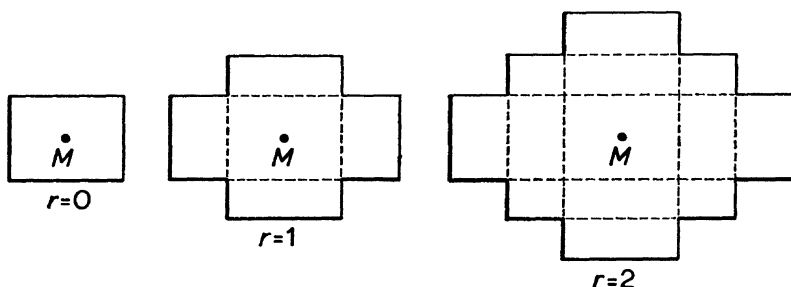


Fig. 1. Examples of sets  $\sigma_h(M, r)$  in the plane.

## 8.2. Approximation of the Space $L^2(\mathbf{R}^n)$

We are going to construct a stable and convergent internal approximation of the space  $V = L^2(\mathbf{R}^n)$ ; we denote the scalar product and norm in  $L^2(\mathbf{R}^n)$  by:

$$(f, g) = \int_{\mathbf{R}^n} f(x) g(x) dx, \quad |f| = (f, f)^{1/2}$$

respectively.

*Space  $V_h$* 

This will be the space of step functions  $u_h$ :

$$u_h(x) = \sum_{M \in R_h} u_h(M) w_{hM}(x),$$

where

$$u_h(M) \in \mathbf{R} \quad \text{and} \quad \int_{\mathbf{R}^n} (u_h(x))^2 dx < +\infty.$$

Evidently,  $V_h$  is just a subspace of  $L^2(\mathbf{R}^n)$  and one may verify that it is closed; we endow it with the induced Hilbert structure.

*Operator  $p_h$* 

The prolongation operator will simply be the identity; evidently  $p_h$  is injective and its norm is 1.

*Operator  $r_h$* 

The restriction operator will be defined as follows:

$$r_h u = u_h, \quad \forall u \in V$$

$$u_h(M) = (h_1 \dots h_n)^{-1} \int_{\sigma_h(M)} u(x) dx, \quad \forall M \in R_h$$

( $r_h$  is hence an operator which takes the average over the cells  $\sigma_h(M)$ ,  $M \in R_h$ ).

**PROPOSITION 1.** *The preceding internal approximation of  $L^2(\mathbf{R}^n)$  is stable and convergent.*

*Proof.* It is seen without difficulty that  $r_h$  is an orthogonal projection operator in  $V$  onto  $V_h$ :

$$(r_h u, v_h) = (u, v_h), \quad \forall v_h \in V_h.$$

The operator  $r_h$  is hence surjective, and its norm is 1:

$$(8.1) \quad |r_h u| \leq |u|, \quad \forall u \in L^2(\mathbf{R}^n).$$

Since the prolongations have a norm equal to 1, the approximation under consideration is stable.

To prove that it is convergent, it is sufficient, by Proposition 1 of Chapter 3, to prove that:

$$(8.2) \quad \lim_{h \rightarrow 0} p_h r_h u = u \quad \text{in the strong topology of } L^2(\mathbf{R}^n).$$

for all functions  $u$  of a dense subspace of  $L^2(\mathbf{R}^n)$ , for example, all continuous functions  $u$  with compact support in  $\mathbf{R}^n$ .<sup>1</sup> For such functions the proof of (8.2) is elementary.

*Remark.* This approximation is just a Galerkin approximation (cf. Section 3.2).

### 8.3. The Approximation of the Space $H^1(\mathbf{R}^n)$

We denote the scalar product and the norm in  $H^1(\mathbf{R}^n)$  by:

$$\begin{aligned} ((u, v)) &= \sum_{i=1}^n (D_i u, D_i v) + (u, v) \\ \|u\| &= ((u, v))^{1/2}. \end{aligned}$$

*Space  $V_h$*

This will be the space of step functions  $u_h$ :

$$u_h(x) = \sum_{M \in R_h} u_h(M) w_{hM}(x), \quad u_h(M) \in \mathbf{R}$$

such that:

$$\int_{\mathbf{R}^n} (u_h(x))^2 dx < +\infty$$

$$\int_{\mathbf{R}^n} (\delta_{ih} u_h(x))^2 dx < +\infty, \quad i = 1, \dots, n.$$

We provide this space with the scalar product:

$$((u_h, v_h))_h = \sum_{i=1}^n (\delta_{ih} u_h, \delta_{ih} v_h) + (u_h, v_h),$$

which makes it into a Hilbert space (proof analogous to the proof of the lemma in Section 7.1).

*Operator  $p_h$*

We put:

$$p_h u_h = w_h * u_h,$$

where  $*$  means the convolution product in  $\mathbf{R}^n$  and where  $w_h$  is the characteristic function of the cell:

$$\prod_{i=1}^n [-h_i/2, h_i/2[ ,$$

multiplied by the number  $(h_1, \dots, h_n)^{-1}$ .

# Operator $r_h$

Same as in Section 8.2.

**PROPOSITION 2.** *The preceding internal approximation of  $H^1(\mathbf{R}^n)$  is stable and convergent.*

*Proof.* Let  $\chi_h$  be the characteristic function in  $\mathbf{R}_{x_i}$  of the interval  $[-h_h/2, h_h/2[$ , and let  $\sigma_{i,h}$  and  $\theta_{i,h}$  be the following distributions:

$$\sigma_{i,h} = \left( \bigotimes_{j \neq i} \frac{1}{h_j} \chi_{h_j} \right) \otimes \delta_{(i)}$$

$$\theta_{i,h} = \left( \bigotimes_{j \neq i} \delta_{(j)} \right) \otimes \left( \frac{1}{h_i} \chi_{h_i} \right),$$

where  $\delta_{(j)}$  is the Dirac measure in 0 on  $\mathbf{R}_{x_j}$ .<sup>2</sup>

We mention without proof the following two results, which anyway follow quickly from the properties of the convolution product:

$$(8.3) \quad \left\{ \begin{array}{l} \text{For all: } \varphi \in L^2(\mathbf{R}^n): \\ w_h * \varphi \in L^2(\mathbf{R}^n), \quad \sigma_{i,h} * \varphi \in L^2(\mathbf{R}^n), \\ \theta_{i,h} * \varphi \in L^2(\mathbf{R}^n), \quad i = 1, \dots, n \\ \text{and:} \\ |w_h * \varphi| \leq |\varphi|, \quad |\sigma_{i,h} * \varphi| \leq |\varphi|, \\ |\theta_{i,h} * \varphi| \leq |\varphi|, \quad i = 1, \dots, n. \end{array} \right.$$

$$(8.4) \quad \left\{ \begin{array}{l} \text{When } h \rightarrow 0: \\ w_h * \varphi \rightarrow \varphi, \quad \sigma_{i,h} * \varphi \rightarrow \varphi, \quad \theta_{i,h} * \varphi \rightarrow \varphi \\ \text{in the strong topology of } L^2(\mathbf{R}^n), \quad i = 1, \dots, n. \end{array} \right.$$

The proposition follows now from the next three lemmas.

LEMMA 1. *The prolongation operators are stable:*

$$(8.5) \quad \|p_h u_h\| \leq \|u_h\|_h, \quad \forall u_h \in V_h.$$

*Proof.* According to (8.3):

$$|p_h u_h| = |w_h * u_h| \leq |u_h|.$$

On the other hand, we have:

$$(8.6) \quad D_i p_h u_h = (D_i w_h) * u_h = \sigma_{i,h} * \delta_{ih} u_h.$$

Using (8.3) once more, we obtain:

$$|D_i p_h u_h| \leq |\delta_{ih} u_h|,$$

and the lemma follows.

LEMMA 2. *The restriction operators are stable:*

$$(8.7) \quad \|r_h u\|_h \leq \|u\|, \quad \forall u \in V.$$

*Proof.* To begin with, we note that the translation operators whose vector of translation is  $\alpha h$ ,  $\alpha h \in R_h$ , commute with  $r_h$ :

$$(8.8) \quad \tau_{\alpha h} r_h u = r_h \tau_{\alpha h} u,$$

where  $\tau_{\alpha h} \varphi$  is defined by  $(\tau_{\alpha h} \varphi)(x) = \varphi(x - \alpha h)$ . In fact, for

$\beta h \in R_h$ :

$$\begin{aligned}
 (\tau_{\alpha h} r_h u)(\beta h) &= (r_h u)((\beta - \alpha)h) \\
 &= (h_1 \dots h_n)^{-1} \int_{\sigma_h((\beta - \alpha)h)} u(x) dx \\
 &= (h_1 \dots h_n)^{-1} \int_{\sigma_h(\beta h)} u(x - \alpha h) dx \\
 &= (r_h \tau_{\alpha h} u)(\beta h).
 \end{aligned}$$

Now, because:

$$\delta_i = \frac{1}{h_i} (\tau_{(-h_i e_i/2)} - \tau_{(h_i e_i/2)}) = \frac{1}{h_i} \tau_{(h_i e_i/2)} (\tau_{(-h_i e_i)} - \tau_{(0)}),$$

we have:

$$(8.9) \quad \delta_{ih} r_h u = \tau_{(h_i e_i/2)} r_h \tau_{(-h_i e_i/2)} \delta_{ih} u.$$

Because of (8.1) and the fact that translations are isometries in  $L^2(\mathbb{R}^n)$ , it follows that:

$$|\delta_{ih} r_h u| = |r_h \tau_{(-h_i e_i/2)} \delta_{ih} u| \leq |\tau_{(-h_i e_i/2)} \delta_{ih} u| = |\delta_{ih} u|.$$

But:

$$(8.10) \quad \delta_{ih} u = (D_i \theta_{i,h}) * u = \theta_{i,h} * (D_i u),$$

and by using (8.3), we obtain:

$$|\delta_{ih} u| \leq |D_i u|.$$

Hence:

$$|\delta_{ih} r_h u| \leq |D_i u|, \quad i = 1, \dots, n$$

and, by (8.1) we also have:

$$|r_h u| \leq |u|,$$

and the lemma follows.

**LEMMA 3.** *The approximation is convergent.*

*Proof.* According to (8.3):

$$\begin{aligned} p_h r_h u - u &= w_h * u - u + w_h * (r_h u - u) \\ |p_h r_h u - u| &\leq |w_h * u - u| + |r_h u - u|. \end{aligned}$$

Using Proposition 1 and (8.4), we find:

$$\lim_{h \rightarrow 0} p_h r_h u = u \quad \text{in the strong topology of } L^2(\mathbf{R}^n).$$

On the other hand, according (8.6), (8.9) and (8.10):

$$\begin{aligned} D_i p_h r_h u &= D_i (w_h * r_h u) = \\ &= \sigma_{i,h} * (\delta_{i,h} r_h u) = \\ &= \sigma_{i,h} * (\tau_{(h_i e_i/2)} r_h \tau_{(-h_i e_i/2)} \delta_{i,h} u) = \\ &= \sigma_{i,h} * (\tau_{(h_i e_i/2)} r_h \tau_{(-h_i e_i/2)} (\theta_{i,h} * D_i u)). \end{aligned}$$

Hence  $D_i p_h r_h u$  is of the form  $G_{5,h} \cdot \dots \cdot G_{1,h} (D_i u)$ , where:

$$\begin{aligned} G_{1,h} &= \theta_{i,h} *, \quad \text{of} \quad G_{2,h} = \tau_{(-h_i e_i/2)}, \quad G_{3,h} = r_h \\ G_{4,h} &= \tau_{(h_i e_i/2)}, \quad G_{5,h} = \sigma_{i,h} *; \end{aligned}$$

the operators  $G_{j,h}$  are linear continuous operators in  $L^2(\mathbf{R}^n)$ , of norm less than or equal 1, and for all  $\varphi \in L^2(\mathbf{R}^n)$  holds:

$$\lim_{h \rightarrow 0} G_{j,h} \varphi = \varphi \quad \text{in the strong topology of } L^2(\mathbf{R}^n);$$

for  $G_{1,h}$  and  $G_{5,h}$  these properties follow from (8.3) and

(8.4), for  $G_{3,h}$  this follows from (8.1) and Proposition 1, for  $G_{2,h}$  and  $G_{4,h}$  it follows from the fact that translations are isometries in  $L^2(\mathbb{R}^n)$  and from the Lebesgue theorem.

Hence it is straightforward, that for all  $\varphi$  in  $L^2(\mathbb{R}^n)$ :

$$\lim_{h \rightarrow 0} G_{5,h} \cdot \dots \cdot G_{1,h} \varphi = \varphi$$

in the strong topology of  $L^2(\mathbb{R}^n)$ , and in particular:

$$\lim_{h \rightarrow 0} D_i p_h r_h u = D_i u$$

in the strong topology of  $L^2(\mathbb{R}^n)$ , which completes the proof.

*Remark.* Similar procedures can be used to construct internal approximations of the spaces  $H_0^1(\Omega)$  and  $H^1(\Omega)$  when  $\Omega \neq \mathbb{R}^n$ ; for this the reader is referred to Aubin [10].

## NOTES

<sup>1</sup> We observed in Chapter 7 that  $\mathcal{D}(\Omega)$  is dense in  $L^2(\Omega)$ .

<sup>2</sup>  $\sigma_{i,h}$  and  $\theta_{i,h}$  can elementarily be defined in the following way:

$$\langle \sigma_{i,h}, \varphi \rangle = (h_1 \dots h_{i-1} h_{i+1} \dots h_n)^{-1}$$

$$\int_{-h_1/2}^{+h_1/2} \dots \int_{-h_{i-1}/2}^{+h_{i-1}/2} \dots \int_{-h_{i+1}/2}^{+h_{i+1}/2} \dots \int_{-h_n/2}^{+h_n/2} \varphi(x_1 \dots x_{i-1}, 0, x_i \dots x_n)$$

$$dx_1 dx_{i-1} dx_{i+1} \dots dx_n.$$

$$(\theta_{i,h}, \varphi) = \frac{1}{h_i} \int_{-h_i/2}^{h_i/2} \varphi(0, \dots, 0, x_i, 0, \dots, 0) dx_i \quad \forall \varphi \in \mathcal{D}(\mathbb{R}^n).$$

It will be observed that  $(D_i \theta_{i,h})^* = \delta_{i,h}$ .

## CHAPTER 9

### APPROXIMATION OF SOME FUNCTION SPACES BY FINITE DIFFERENCES (II)

In this chapter we construct an internal approximation of  $L^2(\Omega)$ , an external approximation of  $H^1(\Omega)$  and a generalized external approximation of  $H_0^1(\Omega)$ , using finite differences. We keep the notations introduced in Section 8.1, and the open set  $\Omega$  is assumed to be bounded.

#### 9.1. Internal Approximation of $L^2(\Omega)$

We denote the scalar product and the norm of  $L^2(\Omega)$  by:

$$(f, g) = \int_{\Omega} f(x) g(x) dx, \quad |f| = (f, f)^{1/2}.$$

For all  $h \in \mathcal{H}$ , let  $\Omega_h$  be a subset of  $R_h$  satisfying:

$$(9.1) \quad \text{for all } M \in \Omega_h, \quad \sigma_h(M) \subset \Omega$$

$$(9.2) \quad \text{for each compact set } K \subset \Omega:$$

$$K \subset \bigcup_{M \in \Omega_h} \sigma_h(M)$$

for  $h$  small enough.

It is easily seen that the sets  $\hat{\Omega}_h^r$ , as defined in Section 8.1, satisfy these conditions.

### Space $V_h$

This will be a subspace of  $L^2(\mathbf{R}^n)$ , consisting of step functions  $u_h$ :

$$u_h(x) = \sum_{M \in \Omega_h} u_h(M) w_{hM}(x), \quad u_h(M) \in \mathbf{R}.$$

We provide  $V_h$  with the Hilbert structure induced from  $L^2(\mathbf{R}^n)$ :

$$\begin{aligned} (u_h, v_h)_h &= \int_{\mathbf{R}^n} u_h(x) v_h(x) dx = \\ &= (h_1 \dots h_n) \sum_{M \in \Omega_h} u_h(M) v_h(M). \end{aligned}$$

Because  $\Omega$  is bounded,  $V_h$  is finite dimensional, and the dimension is just equal to the number  $n(h)$  of points of  $\Omega_h$  (observe that the set of functions  $w_{hM}$ ,  $M \in \Omega_h$ , constitutes a basis of  $V_h$ ).

### Operator $p_h$

This will be the operator of restriction to  $\Omega$ :

$$p_h u_h = u_h|_{\Omega}$$

is injective, and its norm is 1.

### Operator $r_h$

For  $u \in L^2(\Omega)$ ,  $r_h u = u_h$  is defined by:

$$u_h(M) = (h_1 \dots h_n)^{-1} \int_{\sigma_h(M)} u(x) dx, \quad \forall M \in \Omega_h.$$

Evidently,  $r_h$  is surjective.

PROPOSITION 1. *Under conditions (9.1) and (9.2) the above internal approximation of  $L^2(\Omega)$  is stable and convergent.*

*Proof.* For  $u \in V$ , we have:

$$\begin{aligned} |r_h u|_h^2 &= (h_1 \dots h_n) \sum_{M \in \Omega_h} |r_h u(M)|^2 = \\ &= (h_1 \dots h_n)^{-1} \sum_{M \in \Omega_h} \left( \int_{\sigma_h(M)} u(x) dx \right)^2. \end{aligned}$$

According to the Schwarz inequality:

$$\left( \int_{\sigma_h(M)} u(x) dx \right)^2 \leq (h_1 \dots h_n) \int_{\sigma_h(M)} (u(x))^2 dx;$$

because of (9.1):

$$\sum_{M \in \Omega_h} \int_{\sigma_h(M)} (u(x))^2 dx \leq \int_{\Omega} (u(x))^2 dx,$$

and hence:

$$(9.3) \quad |r_h u|_h \leq |u|.$$

Since the prolongations have a norm equal to 1, the approximation under consideration is stable; to prove that it is convergent, it is sufficient by Proposition 1 of Chapter 3 to prove that:

$$(9.4) \quad \lim_{h \rightarrow 0} p_h r_h u = u \quad \text{in the strong topology of } L^2(\Omega)$$

for all functions  $u$  of a dense subspace of  $L^2(\Omega)$ , for example, for all continuous functions  $u$  with compact

support in  $\Omega$ . If  $u$  is continuous and has compact support in  $\Omega$ , then  $u$  is uniformly continuous; if  $h$  is sufficiently small such that the support of  $u$  is included in  $\bigcup_{M \in \Omega_h} \sigma_h(M)$  (according to (9.2)), then the difference  $u - r_h u$  is uniformly bounded by  $\theta(|h|/2)$ , where  $|h|^2 = h_1^2 + \dots + h_n^2$  and  $\theta$  is the modulus of uniform continuity of  $u$ ; (9.4) follows easily.

## 9.2. External Approximation of $H^1(\Omega)$

Let  $V = H^1(\Omega)$  with the usual Hilbert structure, and let  $F = L^2(\Omega)^{n+1}$  with the product Hilbert structure; the map:

$$u \rightarrow \bar{\omega}u = (u, D_1 u, \dots, D_n u)$$

is an isomorphism of  $V$  into  $F$ .

*Space  $V_h$*

This is the space of step functions  $u_h$ :

$$u_h(x) = \sum_{M \in \Omega_h^1} u_h(M) w_{hM}(x), \quad u_h(M) \in \mathbb{R},$$

where  $\Omega_h^1$  is as defined in Section 8.1.  $V_h$  is provided with the following scalar product:

$$((u_h, v_h))_h = \int_{\mathbb{R}^n} u_h(x) v_h(x) dx + \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx.$$

$V_h$  is finite dimensional, and its dimension is just the number  $n(h)$  of points of  $\Omega_h^1$ .

*Operator  $p_h$* 

The prolongation operator will be:

$$p_h u_h = (u_h|_{\Omega}, \delta_1 u_h|_{\Omega}, \dots, \delta_n u_h|_{\Omega}),$$

where  $g|_{\Omega}$  is again the restriction to  $\Omega$  of a function  $g$  defined on  $\mathbf{R}^n$ .

The norm of  $p_h$  is less than or equal to 1.

*Operator  $r_h$* 

We assume that condition (7.3), with  $r=1$ , holds. Then, by Section 7.2, there exists a prolongation operator  $\mathcal{P}$ , which is a linear continuous map from  $L^2(\Omega)$  into  $L^2(\mathbf{R}^n)$  and from  $H^1(\Omega)$  into  $H^1(\mathbf{R}^n)$  satisfying:

$$\mathcal{P}u|_{\Omega} = u, \quad \forall u \in L^2(\Omega).$$

Now we put:

$$r_h u = u_h, \quad \forall u \in V$$

$$u_h(M) = (h_1 \dots h_n)^{-1} \int_{\sigma_h(M)} (\mathcal{P}u)(x) dx, \quad \forall M \in \Omega_h^1.$$

**PROPOSITION 2.** *The preceding external approximation of  $H^1(\Omega)$  is stable and convergent.*

*Proof.* By Proposition 2 of Chapter 8, we have:

$$\|r_h u\|_h \leq \|\mathcal{P}u\|_{H^1(\mathbf{R}^n)},$$

and by continuity of  $\mathcal{P}$ :

$$\|\mathcal{P}u\|_{H^1(\mathbf{R}^n)} \leq \|\mathcal{P}\| \|u\|,$$

which proves that the restrictions are stable. As the prolongations have norms smaller than 1, the stability of the approximation follows.

We prove now that it is convergent. We first must show that  $p_h r_h u \rightarrow \bar{\omega} u$ , in the strong topology of  $F$ , for  $h \rightarrow 0$ , in other words:

$$\lim_{h \rightarrow 0} r_h u \big|_{\Omega} = u$$

$$\lim_{h \rightarrow 0} \delta_i r_h u \big|_{\Omega} = D_i u \quad (i = 1, \dots, n)$$

in the strong topology of  $L^2(\Omega)$ . Because the approximation is stable, we need only prove this for functions of  $\mathcal{D}(\bar{\Omega})$ , by Proposition 1 of Chapter 3 and Section 7.1, and for these it is evident.<sup>1</sup>

We show now that property (ii) of the definition of convergence for external approximations of Chapter 3 holds. So suppose that  $h'$  is a sequence converging to 0, and that:

$$\lim_{h' \rightarrow 0} p_h u_h = \varphi \quad \text{in the weak topology of } F,$$

which means, for  $\varphi = (\varphi_0, \varphi_1, \dots, \varphi_n)$ :

$$\lim_{h' \rightarrow 0} u_{h'} \big|_{\Omega} = \varphi_0$$

$$\lim_{h' \rightarrow 0} \delta_i u_{h'} \big|_{\Omega} = \varphi_i \quad (i = 1, \dots, n)$$

in the weak topology of  $L^2(\Omega)$ . One verifies easily that

$\delta_i \varphi_h|_{\Omega}$  converges to  $D_i \varphi_0$  in the space of distributions on  $\Omega$ ; so  $\varphi_i = D_i \varphi_0$ , which proves that the limit  $\varphi$  is of the form:

$$\bar{\omega} \varphi_0 = (\varphi_0, D_1 \varphi_0, \dots, D_n \varphi_0), \quad \varphi_0 \in H^1(\Omega),$$

and we are ready with the proof of the proposition.

### 9.3. External Approximation of $H_0^1(\Omega)$

The space  $F$  and the map  $u \rightarrow \bar{\omega}u$  are as above, namely the space  $L^2(\Omega)^{n+1}$  and the isomorphism  $u \rightarrow \bar{\omega}u = (u, D_1 u, \dots, D_n u)$  from  $H_0^1(\Omega)$  into  $F$ .

*Space  $V_h$*

This is the space of step functions  $u_h$ :

$$u_h(x) = \sum_{M \in \Omega_h^1} u_h(M) w_{hM}(x), \quad u_h(M) \in \mathbb{R}$$

where  $\Omega_h^1$  is as in Section 8.1. Space  $V_h$  is finite dimensional, and its dimension is  $n(h)$ , the number of points of  $\Omega_h^1$ ; it is provided with the scalar product:

$$((u_h, v_h))_h = \int_{\Omega} u_h(x) v_h(x) dx + \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx.$$

*Operator  $p_h$*

As in Section 9.2.

*Operator  $r_h$* 

As a consequence of Proposition 4 of Chapter 3, we only define the operator  $r_h$  on  $\mathcal{V} = \mathcal{D}(\Omega)$ ; we take:

$$(9.5) \quad (r_h u)(M) = u(M), \quad \forall M \in \Omega_h^1, \forall u \in \mathcal{D}(\Omega).$$

**PROPOSITION 3.** *The preceding generalized external approximation of  $H_0^1(\Omega)$  is stable and convergent.*

*Proof.* The prolongation operators have norm 1 and are stable. Hence the approximation is stable.

According to Proposition 4 of Chapter 3 we have to prove that:

$$(9.6) \quad \lim_{h \rightarrow 0} p_h r_h u = \bar{\omega} u, \quad \text{in } F, \quad \forall u \in \mathcal{D}(\Omega),$$

and to check condition (ii) in Definition 10 of Chapter 3.

Function  $u \in \mathcal{V}$  and all its derivatives are uniformly continuous. Let  $h$  be sufficiently small such that the support of  $u$  is included in  $\bigcup_{M \in \Omega_h^1} \sigma_h(M)$ . Then the difference  $u - r_h u$  is uniformly bounded by  $\theta(|h|/2)$ , where  $\theta$  is the modulus of uniform continuity of  $u$ ; hence  $r_h u|_{\Omega}$  converges to  $u$  in  $L^2(\Omega)$  as  $h \rightarrow 0$ . By the Taylor formula,  $\delta_{ih} r_h u(M)$  equals the value of  $D_i u(x)$ , for some  $x \in \sigma_h(M)$ ,  $\forall M \in \Omega_h^1$ ; the difference  $D_i u - \delta_{ih} r_h u$  is then uniformly bounded by  $\theta_i(c|h|)$ , where  $c$  stands for some positive fixed constant and  $\theta_i$  is the modulus of continuity of  $D_i u$ . It follows that  $\delta_{ih} r_h u|_{\Omega}$  converges to  $D_i u$  in  $L^2(\Omega)$  as  $h \rightarrow 0$  and (9.6) is proved.

We now check condition (ii) of Definition 10 of a convergent external approximation of Chapter 3.

Let  $h'$  be a sequence converging to 0, and let  $u_{h'} \in V_{h'}$  be a sequence such that:

$$\lim_{h' \rightarrow 0} p_{h'} u_{h'} = \varphi \quad \text{in the weak topology of } F.$$

Then we have for  $\varphi = (\varphi_0, \varphi_1, \dots, \varphi_n)$ :

$$\lim_{h' \rightarrow 0} u_{h'}|_{\Omega} = \varphi_0$$

$$\lim_{h' \rightarrow 0} \delta_i u_{h'}|_{\Omega} = \varphi_i \quad (i = 1, \dots, n)$$

in the weak topology of  $L^2(\Omega)$ . As the functions  $u_{h'}$  and  $\delta_i u_{h'}$  have compact support in  $\Omega$ , we also have:

$$\lim_{h' \rightarrow 0} u_{h'} = \tilde{\varphi}_0$$

$$\lim_{h' \rightarrow 0} \delta_i u_{h'} = \tilde{\varphi}_i \quad (i = 1, \dots, n)$$

in the weak topology of  $L^2(\mathbf{R}^n)$ . Here  $\tilde{g}$  means the function equal to  $g$  in  $\Omega$  and equal to 0 in the complement of  $\Omega$ . Because  $\delta_i u_{h'}$  converges to  $D_i \tilde{\varphi}_0$  in the space of distributions on  $\mathbf{R}^n$ , we see that

$$\tilde{\varphi}_i = D_i \tilde{\varphi}_0 \quad (i = 1, \dots, n).$$

This proves that  $\tilde{\varphi} \in H^1(\mathbf{R}^n)$ ; as  $\tilde{\varphi}_0$  is zero outside of  $\Omega$ , clearly the trace of  $\tilde{\varphi}_0$  (and hence of  $\varphi_0$ ) on  $\Gamma$  is zero, and hence  $\varphi_0 = \tilde{\varphi}_0|_{\Omega} \in H_0^1(\Omega)$ . So the limit  $\varphi$  is indeed of the

form:

$$\bar{\omega}\varphi_0 = (\varphi_0, D_1\varphi_0, \dots, D_n\varphi_0), \quad \varphi_0 \in H_0^1(\Omega).$$

*Remark.* We could as well define an external approximation of  $H_0^1(\Omega)$ , where the restriction operator is defined by:

$$(r_h u)(M) = (h_1 \dots h_n)^{-1} \int_{\sigma_h(M)} u(x) \, dx,$$

$$\forall M \in \dot{\Omega}_h^1, \quad \forall u \in H_0^1(\Omega).$$

The proof of the stability of the  $r_h$  is then very technical and demands some special regularity properties of  $\Gamma$ .

#### NOTE

<sup>1</sup> The same argument can be used to show that  $\lim_{h \rightarrow 0} r_h u = \tilde{u}$  in the strong topology of  $L^2(\mathbb{R}^n)$ , where  $\tilde{u}$  is the function which is equal to  $u$  in  $\Omega$  and equal to 0 in the complement of  $\Omega$ .

## CHAPTER 10

### APPROXIMATION OF SOME FUNCTION SPACES BY FINITE ELEMENT METHODS

In this chapter we study the approximation of the space  $H_0^1(\Omega)$  by finite element methods. In finite element methods the approximating functions are not step functions, but functions defined on a set of  $n$ -simplices contained in  $\Omega$ . Inside the simplex the approximating function is a simple function, a linear function in the simplest case, or some type of polynomial function of restricted degree in more refined situations. We will only consider approximation by piecewise linear functions. In this case the approximation is said to be conforming or non-conforming according as the approximating functions are continuous or some kind of discontinuities are allowed at the boundaries of the simplices. We will give some preliminary results in Section 10.1 and then study these two approximations in Sections 10.2 and 10.3.

#### 10.1. Preliminary Results

Since we will have to work with linear functions defined on  $n$ -simplices, we introduce here some notations adapted to the situation.

*Barycentric coordinates*

Let there be given in  $\mathbf{R}^n$   $n+1$  points  $A_1, \dots, A_{n+1}$ , with coordinates  $a_{1,i}, \dots, a_{n,i}$ ,  $1 \leq i \leq n+1$ , which do not all lie in one hyperplane; this amounts to saying that the  $n$  vectors  $A_1 A_2, \dots, A_1 A_n$  are independent, or that the matrix:

$$\mathcal{A} = \begin{vmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,n+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,n+1} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \dots & a_{n,n+1} \\ 1 & 1 & \dots & 1 \end{vmatrix}$$

is regular. Given any point  $P \in \mathbf{R}^n$  with coordinates  $x_1, \dots, x_n$ , there exist  $n+1$  real numbers

$$\lambda_i = \lambda_i(P), \quad 1 \leq i \leq n+1,$$

such that:

$$(10.1) \quad OP = \sum_{i=1}^{n+1} \lambda_i OA_i,$$

with

$$(10.2) \quad \sum_{i=1}^{n+1} \lambda_i = 1.$$

To see this, it suffices to remark that (10.1) and (10.2) are equivalent to the linear system

$$(10.3) \quad \mathcal{A} \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_{n+1} \end{pmatrix} = \begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 1 \end{pmatrix},$$

which has a unique solution, since the matrix  $\mathcal{A}$  is regular, by hypothesis. Those numbers  $\lambda_i$  are called the *barycentric coordinates of  $P$* , with respect to the  $n+1$  points  $A_1, \dots, A_n$ . As a consequence of (10.3) the numbers  $\lambda_i$  appear as linear, in general, nonhomogeneous functions of the coordinates  $x_1, \dots, x_n$  of  $P$ :

$$(10.4) \quad \lambda_i = \sum_{j=1}^n b_{i,j} x_j + b_{i,n+1}, \quad 1 \leq i \leq n+1,$$

where the matrix  $\mathcal{B} = (b_{i,j})$  is the inverse of the matrix  $\mathcal{A}$ . It is easy to see that point  $O$  in (10.1) can be replaced by any other point of  $\mathbf{R}^n$  without changing the value of the barycentric coordinates; hence:

$$(10.5) \quad \sum_{i=1}^{n+1} \lambda_i P A_i = 0.$$

The barycentric coordinates are also independent of the choice of a basis in  $\mathbf{R}^n$ .

The convex hull of the  $n+1$  points  $A_i$  is the set of points of  $\mathbf{R}^n$  with barycentric coordinates satisfying the conditions:

$$(10.6) \quad 0 \leq \lambda_i \leq 1, \quad 1 \leq i \leq n+1.$$

This convex hull  $\mathcal{S}$  is the  *$n$ -simplex generated by the points  $A_i$ , which are called the vertices of the  $n$ -simplex*. The bary-

center  $G$  of  $\mathcal{S}$  is the point of  $\mathcal{S}$  whose barycentric coordinates are all equal, and hence equal to  $1/(n+1)$ . An  $m$ -dimensional face of  $\mathcal{S}$  is any  $m$ -simplex ( $1 \leq m \leq n-1$ ) generated by  $m+1$  of the vertices of  $\mathcal{S}$ . A one-dimensional face is an edge.

In the two-dimensional cases ( $n=2$ ), the two-simplices are triangles; the vertices and edges of the simplex are the vertices and edges of the triangle.

Let us show, using these coordinates, how simple the representation is of a linear function taking given values at the points  $A_1, \dots, A_{n+1}$ .

**PROPOSITION 1.** Let  $A_1, \dots, A_{n+1}$ , be  $n+1$  points of  $\mathbb{R}^n$  which are not included in a hyperplane. Given  $n+1$  real numbers  $\alpha_0, \dots, \alpha_n$ , there exists one and only one linear function  $u$  such that  $u(A_i) = \alpha_i$ ,  $1 \leq i \leq n+1$  and

$$(10.7) \quad u(P) = \sum_{i=1}^{n+1} \alpha_i \lambda_i(P), \quad \forall P \in \mathbb{R}^n,$$

where  $\lambda_i(P)$  are the barycentric coordinates of  $P$ .

*Proof.* Let

$$u(x) = \sum_{j=1}^n \beta_j x_j + \beta_{n+1},$$

be this function  $u$ . The unknowns are  $\beta_1, \dots, \beta_{n+1}$ , which satisfy the following equations which assert that  $u(A_i) = \alpha_i$ :

$$\sum_{j=1}^n \beta_j a_{j,i} + \beta_{n+1} = \alpha_i, \quad 1 \leq i \leq n+1.$$

The matrix of this system is the transposed  ${}^t\mathcal{A}$  of the matrix  $\mathcal{A}$  and hence the function  $u$  exists and is unique.

It remains to show that (10.7) is the required function; indeed

$$u(A_j) = \alpha_j, \quad 1 \leq j \leq n+1,$$

since  $\lambda_i(A_j) = \delta_{ij}$  = the Kronecker symbol, for each  $i$  and  $j$ .

### *Differential properties*

We give some differential properties of the  $\lambda_i$  considered as functions of the cartesian coordinates  $x_1, \dots, x_n$  of  $P$ ; let  $D$  denote the gradient operator  $(D_1, \dots, D_n)$ .

LEMMA 1.

$$(10.8) \quad \sum_{i=1}^{n+1} D\lambda_i = 0$$

$$(10.9) \quad D\lambda_i(P)PA_j = \delta_{ij} - \lambda_i(P), \quad 1 \leq i, j \leq n+1.$$

*Proof.* (10.8) follows immediately from identity (10.2). According to (10.4):

$$\partial\lambda_i/\partial x_k = b_{ik}, \quad 1 \leq i \leq n+1, \quad 1 \leq k \leq n,$$

and hence:

$$\begin{aligned}
D\lambda_i(P) PA_j &= D\lambda_i(P) OA_j - D\lambda_i(P) OP \\
&= \sum_{k=1}^n b_{i,k} a_{k,j} - \sum_{k=1}^n b_{i,k} x_k \\
&= \sum_{k=1}^n b_{i,k} a_{k,j} + b_{i,n+1} - \lambda_i \\
&= \delta_{ij} - \lambda_i.
\end{aligned}$$

For the last equality note that  $\mathcal{B} = \mathcal{A}^{-1}$ .

**LEMMA 2.** *Let  $\mathcal{S}$  be the  $n$ -simplex with edges  $A_1, \dots, A_{n+1}$  and let  $\rho'$  be the upper bound of the diameters of the balls contained in  $\mathcal{S}$ . Then:*

$$(10.10) \quad |D\lambda_i| \leq 1/\rho', \quad 1 \leq i \leq n+1,$$

where  $|D\lambda_i|$  is the norm of the (constant) vector  $D\lambda_i$ .

*Proof.* We have:

$$(10.11) \quad |D\lambda_i| = D\lambda_i \cdot x.$$

where  $x$  is the unit vector pointing in the same direction as  $D\lambda_i$ . There exist two points  $P$  and  $Q$  belonging to  $\mathcal{S}$  such that

$$x = \frac{1}{\rho'} PQ;$$

more precisely  $P$  and  $Q$  are two points of the ball of diameter  $\rho'$  included in  $\mathcal{S}$ . We denote the barycentric

coordinates of  $Q$  with respect to  $A_1, \dots, A_{n+1}$ , by  $\mu_1, \dots, \mu_{n+1}$ . Then we have, because of (10.1) and (10.2):

$$PQ = \sum_{j=1}^{n+1} \mu_j P A_j$$

$$\sum_{j=1}^{n+1} \mu_j = 1.$$

Then:

$$\begin{aligned} D\lambda_i \cdot x &= \frac{1}{\rho'} (D\lambda_i) \cdot \left( \sum_{j=1}^{n+1} \mu_j P A_j \right) = \\ &= \frac{1}{\rho'} \sum_{j=1}^{n+1} \mu_j D\lambda_i \cdot P A_j = \\ &= (\text{according to (10.9)}) = \\ &= \frac{1}{\rho'} \sum_{j=1}^{n+1} \mu_j (\delta_{ij} - \lambda_i) = \\ &= \frac{1}{\rho'} (\mu_i - \lambda_i). \end{aligned}$$

Since  $P$  and  $Q$  belong to  $\mathcal{S}$ ,  $0 \leq \lambda_i \leq 1$ ,  $0 \leq \mu_i \leq 1$ , for each  $i$ ,  $1 \leq i \leq n+1$ , and hence  $0 \leq \mu_i - \lambda_i \leq 1$ , so that:

$$|D\lambda_i \cdot x| \leq 1/\rho';$$

(10.10) follows.

## 10.2. Conforming Piecewise Linear Finite Elements for $H_0^1(\Omega)$

Let  $\Omega$  be an open bounded set in  $\mathbf{R}^n$ . Let  $\mathcal{T}_h$  be a family of  $n$ -simplices; such a family will be said to constitute an admissible triangulation of  $\Omega$  if the following conditions are satisfied:

$$(10.11) \quad \Omega_h = \bigcup_{\mathcal{S} \in \mathcal{T}_h} \mathcal{S} \subset \Omega.$$

$$(10.12) \quad \text{If } \mathcal{S} \text{ and } \mathcal{S}' \in \mathcal{T}_h, \text{ then either } \mathcal{S} = \mathcal{S}', \text{ or } \mathcal{S} \cap \mathcal{S}' \text{ is empty or } \mathcal{S} \cap \mathcal{S}' \text{ is exactly a complete } m\text{-face for } \mathcal{S} \text{ and } \mathcal{S}'.$$

We will denote by  $\{\mathcal{T}_h\}_{h \in \mathcal{H}}$  the family of all admissible triangulations of  $\Omega$ ; to each admissible triangulation  $\mathcal{T}_h$  we associate the following three numbers:

$$\rho(h) = \sup_{\mathcal{S} \in \mathcal{T}_h} \rho_{\mathcal{S}}, \quad \rho'(h) = \inf_{\mathcal{S} \in \mathcal{T}_h} \rho'_{\mathcal{S}},$$

$$\sigma(h) = \sup_{\mathcal{S} \in \mathcal{T}_h} \frac{\rho_{\mathcal{S}}}{\rho'_{\mathcal{S}}},$$

where  $\rho = \rho_{\mathcal{S}}$  is the diameter of  $\mathcal{S}$  and  $\rho' = \rho'_{\mathcal{S}}$  is the diameter of the greatest ball contained in  $\mathcal{S}$ .

We are concerned with the passage to the limit as  $\rho(h) \rightarrow 0$ . It will appear later that some restrictions are necessary on  $\sigma(h)$  so that the approximations converge.

We will suppose that  $\Omega_h$  converges to  $\Omega$  as  $\rho(h) \rightarrow 0$  in the sense:

(10.13) For each compact  $K \subset \Omega$ , there exists an  $\eta > 0$  such that  $\Omega_h \supset K$  for every  $h$  with  $\rho(h) \leq \eta$ .

Let us now define the approximation of the space  $H_0^1(\Omega)$ : it will be a *generalized internal approximation*.

### Space $V_h$

For each triangulation  $\mathcal{T}_h$ ,  $V_h$  is the space of continuous functions on  $\Omega$  which are linear on each simplex  $\mathcal{S} \in \mathcal{T}_h$ , and vanish outside  $\Omega_h$ .

As a consequence of Proposition 1 a function  $u_h$  which is linear on each simplex  $\mathcal{S}$  is completely characterized by its values on the vertices of the simplices. We remark that the continuity requirement is automatically satisfied for piecewise linear functions: if  $\mathcal{S}$  and  $\mathcal{S}'$  are two simplices of  $\mathcal{T}_h$  with a common  $m$ -dimensional face, then let  $\sigma$  and  $\sigma'$  be the restriction to this face of the linear functions equal to  $u_h$  on  $\mathcal{S}$  and  $\mathcal{S}'$  respectively;  $\sigma$  and  $\sigma'$  are linear on this  $m$ -dimensional face (i.e. a  $m$ -simplex) and equal on the  $m+1$  vertices of this face; because of Proposition 1,  $\sigma$  and  $\sigma'$  are equal everywhere on the face and this ensures that the function  $u_h$  is continuous.

Since  $u_h$  equals 0 in  $\Omega - \Omega_h$ ,  $u_h$  will be zero on the vertices belonging to the boundary  $\Gamma_h = \partial\Omega_h$  of  $\Omega_h$ ; this boundary consists of the union of some faces of the simplices. Then the function  $u_h \in V_h$  is completely characterized by its values on the family  $\mathcal{V}_h^0$  of vertices of the triangulation belonging to  $\mathring{\Omega}_h$ , the interior of  $\Omega_h$ .

LEMMA 3. *The functions of  $V_h$  equal to 1 on some vertex belonging to  $\mathcal{U}_h^0$  and equal to 0 on all the other vertices form a basis of  $V_h$ . Hence the dimension of  $V_h$  is the number of points in  $\mathcal{U}_h^0$ .*

*Proof.* Let for all  $M \in \mathcal{U}_h^0$ ,  $w_{hM}$  be the function of  $V_h$ , equal to 1 on  $M$  and to 0 on all the other vertices. These functions are obviously linearly independent and because of Proposition 1 any  $u_h \in V_h$  can be written:

$$(10.14) \quad u_h = \sum_{M \in \mathcal{U}_h^0} u_h(M) w_{hM}.$$

The lemma is proved.

It is clear that  $V_h \subset H_0^1(\Omega)$ ; if  $u_h \in V_h$  then  $u_h \in L^2(\Omega)$ , since it is continuous with compact support; its derivative  $D_i u_h$  is the bounded step function equal to  $D_i u_h$  on each simplex in  $\mathcal{T}_h$ . We equip  $V_h$  with the scalar product induced by  $H_0^1(\Omega)$ :

$$((u_h, v_h))_h = ((u_h, v_h))_{H_0^1(\Omega)}, \quad \forall u_h, v_h \in V_h.$$

*Operator  $p_h$*

This is the identity; its norm is one; these operators are then stable.

*Operator  $r_h$*

As allowed by Proposition 4 of Chapter 3, we only define  $r_h u$  for  $u \in \mathcal{V} = \mathcal{D}(\Omega)$ , which is a dense subspace of  $H_0^1(\Omega)$ . For such  $u$  we define  $r_h u$  by:

$$(10.15) \quad (r_h u)(M) = u(M), \quad \forall M \in \mathcal{U}_h^0.$$

PROPOSITION 2. *The above generalized internal approximation of  $H_0^1(\Omega)$  is stable and convergent, provided  $\sigma(h)$  remains bounded as  $\rho(h) \rightarrow 0$ :*

$$(10.16) \quad \sigma(h) \leq \alpha < +\infty.$$

*Proof.* The approximation is obviously stable. We only have to prove that for each  $u \in \mathcal{D}(\Omega)$ :

$$(10.17) \quad p_h r_h u \rightarrow u \quad \text{as} \quad \rho(h) \rightarrow 0.$$

This follows from the next lemma.

LEMMA 4. *If  $\rho(h)$  is small enough for  $\Omega_h$  to contain the support of  $u$ , then  $(u_h = r_h u)$ :*

$$(10.18) \quad \sup_{x \in \Omega} |u_h(x) - u(x)| \leq c_1(u) \rho(h)^2$$

$$(10.19) \quad \text{ess. sup}_{x \in \Omega} |D_i u_h(x) - D_i u(x)| \leq \\ \leq c_1(u) \rho(h) \sigma(h) \leq c_1(u) \alpha \rho(h),$$

where  $c_1(u)$  is a constant depending on the second derivatives of  $u$ .

*Proof.* It is sufficient to compare  $u$  and  $u_h$  on each  $n$ -simplex  $\mathcal{S} \in \mathcal{T}_h$ . Let  $A_1, \dots, A_{n+1}$  be the vertices of  $\mathcal{S}$ . Using the Taylor formula, we have:

$$(10.20) \quad u(A_i) = u(P) + D_j u(P) \varepsilon_{j,i} + \frac{1}{2} D_{jk} u(P_i) \varepsilon_{j,i} \varepsilon_{k,i},$$

where  $\varepsilon_i = (\varepsilon_{1,i}, \dots, \varepsilon_{n,i}) = PA_i$ ,  $P_i$  is some point of the segment  $PA_i$  and the summation convention of repeated indices is used.

Because of Proposition 1:

$$(10.21) \quad u_h(P) = \sum_{i=1}^{n+1} \lambda_i(P) u(A_i), \quad \forall P \in \mathcal{S},$$

and then using (10.1), (10.2) and (10.20):

$$\begin{aligned} u_h(P) &= u(P) + \frac{1}{2} D_{jk} u(P_i) \varepsilon_{j,i} \varepsilon_{k,i} \\ |u(P) - u_h(P)| &\leq \\ &\leq \frac{1}{2} \left( \sum_{j,k} |D_{jk}(u(P_i))|^2 \right)^{1/2} |PA_i| |PA_j|, \quad \forall P \in \mathcal{S}. \end{aligned}$$

We obtain (10.18) with

$$c_1(u)^2 = \frac{1}{4} \sup_{x \in \Omega} \left[ \sum_{j,k=1}^n |D_{jk} u(x)|^2 \right].$$

By differentiation of (10.21), and using (10.20):

$$\begin{aligned} Du_h(P) &= \sum_{i=1}^{n+1} u(A_i) D\lambda_i \\ Du_h(P) &= \sum_{i=1}^{n+1} u(P) D\lambda_i + \sum_{i=1}^{n+1} D_j u(P) \varepsilon_{j,i} D\lambda_i + \\ &\quad + \frac{1}{2} \sum_{i=1}^{n+1} D_{jk} u(P_i) \varepsilon_{j,i} \varepsilon_{k,i} D\lambda_i. \end{aligned}$$

One can check, that the first term of the right-hand member of this last equality is 0, because of (10.8), and the second term is  $Du(P)$ . Indeed, since the vectors  $PA_1, \dots, PA_{n+1}$  are linearly independent, it is sufficient to

show that

$$\begin{aligned} \left( \sum_{i=1}^{n+1} D_j u(P) \varepsilon_{j,i} D\lambda_i \right) PA_k &= \\ &= Du(P) PA_k, \quad k = 1, \dots, n+1. \end{aligned}$$

Because of (10.9) the left-hand side of this equality is:

$$\sum_{i=1}^{n+1} Du(P) PA_i (\delta_{ik} - \lambda_i(P)) = Du(P) PA_k,$$

since

$$\sum_{i=1}^{n+1} \lambda_i(P) PA_i = 0.$$

So we get ( $P \in \mathcal{S}$ ):

$$\begin{aligned} |Du_h(P) - Du(P)| &\leq \frac{1}{2} \sum_{i=1}^{n+1} |D\lambda_i| |D_{jk} u(P_i) \varepsilon_{j,i} \varepsilon_{j,k}| \leq \\ &\leq c_1(u) |D\lambda_i| \rho(h)^2 \leq \\ &\leq \text{(with (10.10))} \leq \\ &\leq c_1(u) \frac{\rho_{\mathcal{S}}^2}{\rho'_{\mathcal{S}}} \leq c_1(u) \rho(h) \sigma(h). \end{aligned}$$

*Remark.* In the two-dimensional case, the simplices  $\mathcal{S}$  are triangles, and if we denote the greatest side of  $\mathcal{S}$  by  $\rho$ , and the smallest angle of  $\mathcal{S}$  by  $\theta$ , the following inequality can be easily established:

$$\rho/\rho' \leq 2/\sin \theta.$$

Condition (10.16) is satisfied if all the angles remain bounded from below  $\theta \geq \theta_0 > 0$ ,  $\forall \mathcal{S} \in \mathcal{T}_h$ .

### 10.3. Non-conforming Piecewise Linear Finite Elements for $H_0^1(\Omega)$

Let  $\Omega$  be an open bounded set in  $\mathbf{R}^n$ , and again let  $\{\mathcal{T}_h\}_{h \in \mathcal{H}}$  be the family of all admissible triangulations of  $\Omega$ , as in Section 10.2. In some cases it can be more interesting to have piecewise linear functions which are not continuous (this situation arises for example in the approximation of some fluid-mechanic problems). We describe here an approximation of  $H_0^1(\Omega)$  by noncontinuous piecewise linear functions; this approximation method is called a nonconforming finite element method.

#### *Preliminary remarks*

Let  $\mathcal{S}$  be an  $n$ -simplex belonging to  $\mathcal{T}_h$  with vertices  $A_1, \dots, A_{n+1}$ . We denote by  $\mathcal{S}_i$  the  $(n-1)$ -dimensional face of  $\mathcal{S}$  spanned by the  $A_j, j \neq i$ , and by  $B_i$  the barycenter of  $\mathcal{S}_i$ .

Let  $G$  denote the barycenter of  $\mathcal{S}$ . Now, the barycentric coordinates of  $B_i$  with respect to the  $A_j, j \neq i$ , are equal to  $1/n$ . Hence:

$$(10.22) \quad GB_i = \sum_{j \neq i} \frac{GA_j}{n} = \sum_{j=1}^{n+1} \frac{GA_j}{n} - \frac{GA_i}{n},$$

or

$$(10.23) \quad GB_i = -\frac{1}{n} GA_i,$$

since

$$\sum_{j=1}^{n+1} GA_j = 0$$

(the barycentric coordinates of  $G$  with respect to  $A_1, \dots, A_{n+1}$  are equal to  $1/(n+1)$ ).

Then:

$$(10.24) \quad nB_iB_j = n(GB_j - GB_i) = GA_i - GA_j = -A_iA_j,$$

hence, as the vectors  $B_1B_j, j=2, \dots, n+1$ , are linearly independent, the vectors  $A_1A_j, j=2, \dots, n+1$  are such. Therefore the barycentric coordinates of a point  $P$  with respect to  $B_1, \dots, B_{n+1}$  make sense and we denote these coordinates by  $\mu_1, \dots, \mu_{n+1}$ . We remark also that given  $n+1$  numbers  $\beta_1, \dots, \beta_{n+1}$ , there exists one and only one linear function taking the values  $\beta_1, \dots, \beta_{n+1}$  at the points  $B_1, \dots, B_{n+1}$ , and this function  $u$  is

$$(10.25) \quad u(P) = \sum_{i=1}^{n+1} \mu_i(P) u(B_i).$$

*Space  $V_h$*

This is the space of functions  $u_h$  which are linear on each  $\mathcal{S} \in \mathcal{T}_h$ , vanish outside  $\Omega_h$ , and the value of  $u_h$  at the bary-

center  $B_i$  of some  $n-1$ -dimensional face  $\mathcal{S}_i$  of a simplex is zero if this face belongs to the boundary of  $\Omega_h$ ; if this face meets the interior of  $\Omega_h$ , then the values of  $u_h$  at  $B_i$  are the same whether  $B_i$  is considered as a point of one or another of two different simplices.

Let  $\mathcal{U}_h$  be the set of points  $B_i$  (=barycenter of an  $n-1$ -dimensional face of a simplex  $\mathcal{S} \in \mathcal{T}_h$ ), which belong to the interior of  $\Omega_h$ . A function  $u_h \in V_h$  is completely characterized by its values at the points  $B_i \in \mathcal{U}_h$ .

**LEMMA 5.** *The functions  $w_{hB}$  of  $V_h$  equal to 1 on some point of  $\mathcal{U}_h$  and to 0 on all the other points of  $\mathcal{U}_h$ , form a basis of  $V_h$ . Hence the dimension of  $V_h$  is the number of points in  $\mathcal{U}_h$ .*

*Proof.* It is clear that the functions  $w_{hB}$ ,  $B \in \mathcal{U}_h$  are linearly independent and because of Proposition 1 any

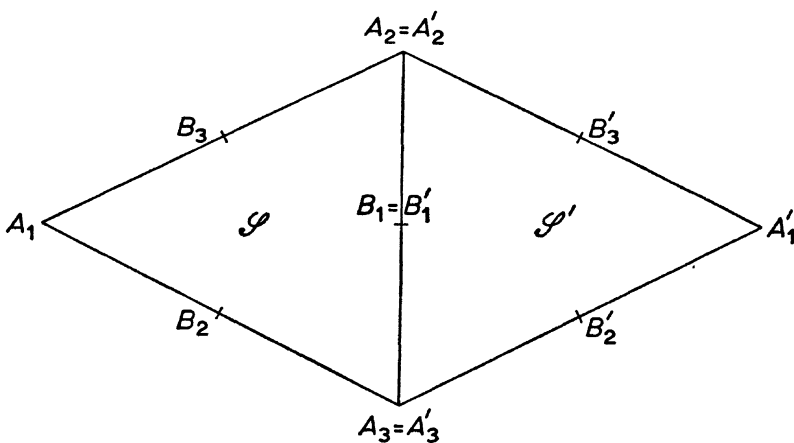


Fig. 2. Two adjacent triangles in the two-dimensional case.<sup>2</sup>

$u_h \in V_h$  can be written as:

$$(10.26) \quad u_h = \sum_{B \in \mathcal{U}_h} u_h(B) w_{hB}.$$

We equip  $V_h$  with the following scalar product:

$$(10.27) \quad ((u_h, v_h))_h = \int_{\Omega} u_h(x) v_h(x) dx + \\ + \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{i=1}^n \int_{\mathcal{S}} D_i^* u_h(x) D_i^* v_h(x) dx.$$

The space  $V_h$  is not contained in  $H_0^1(\Omega)$ , and the derivatives  $D_i u_h$  are sums of step functions  $D_i^* u_h$  and Dirac distributions located on the faces of the simplices connected with the jump of the  $u_h$ ; in (10.27) we integrate the constant function  $D_i^* u_h D_i^* v_h$  on  $\mathcal{S}$ . It is clear that  $V_h$  is a Hilbert space.

*Space  $F$ , operators  $\bar{\omega}$ ,  $p_h$*

As in Chapter 9 we take  $F = L^2(\Omega)^{n+1}$  and  $\bar{\omega}$  the isomorphism  $u \in H_0^1(\Omega) \mapsto \bar{\omega}u = (u, D_1 u, \dots, D_n u) \in F$ . The operator  $p_h$  is defined by:

$$u_h \in V_h \mapsto p_h u_h = (u_h, D_1^* u_h, \dots, D_n^* u_h) \in F.$$

The operators  $p_h$  have norms equal to 1, and they are stable.

*Operator  $r_h$*

As allowed by Proposition 4 of Chapter 3, we just define

$r_h u = u_h$ , for  $u \in \mathcal{V} = \mathcal{D}(\Omega)$ , a dense subspace of  $H_0^1(\Omega)$ ; we take:

$$(10.28) \quad u_h(B), \quad \forall B \in \mathcal{U}_h.$$

**PROPOSITION 3.** *The above generalized external approximation of  $H_0^1(\Omega)$  is stable and convergent, provided  $\sigma(h)$  remains bounded as  $\rho(h) \rightarrow 0$ ,*

$$(10.29) \quad \sigma(h) \leq \alpha < +\infty.$$

*Proof.* We have to check the conditions (i) and (ii) of Definition 10 of Chapter 3.

For condition (ii) we have to prove that, for each  $u \in \mathcal{D}(\Omega)$ :

$$(10.30) \quad u_h \rightarrow u \quad \text{in the norm of } L^2(\Omega) \text{ as } \rho(h) \rightarrow 0$$

$$(10.31) \quad D_{ih}^* u_h \rightarrow D_i u \quad \text{in the norm of } L^2(\Omega) \text{ as } \rho(h) \rightarrow 0.$$

But the conclusions of Lemma 4 are valid, (10.18) without any modification, and (10.19) modified by replacing  $D_i$  by  $D_{ih}^*$ :

$$(10.32) \quad \text{ess. sup.}_{x \in \Omega} |D_{ih}^* u_h(x) - D_i u(x)| \leq c_1(u) \alpha \rho(h).$$

The proofs of (10.18) and (10.32) are exactly the same; we only replace in (10.20) the points  $A_i$  by the points  $B_i$  and the barycentric coordinates  $\lambda_i$  by the barycentric coordinates  $\mu_i$ .

For proving condition (i), let us suppose that  $p_h u_h$  converges weakly to  $\varphi = (\varphi_0, \dots, \varphi_n)$ ; this means that

$$(10.33) \quad \begin{aligned} u_h &\rightarrow \varphi_0 \quad \text{in } L^2(\Omega) \text{ weakly} \\ D_{ih}^* u_h &\rightarrow \varphi_i \quad \text{in } L^2(\Omega) \text{ weakly, } 1 \leq i \leq n. \end{aligned}$$

Since the functions have compact supports in  $\Omega$ , (10.33) means also that

$$(10.34) \quad \begin{aligned} \tilde{u}_h &\rightarrow \tilde{\varphi}_0 \quad \text{in } L^2(\mathbf{R}^n) \text{ weakly} \\ D_{ih}^* \tilde{u}_h &\rightarrow \tilde{\varphi}_i \quad \text{in } L^2(\mathbf{R}^n) \text{ weakly, } 1 \leq i \leq n. \end{aligned}$$

As in Chapter 9, if we show that

$$(10.35) \quad \tilde{\varphi}_i = D_i \tilde{\varphi}_0, \quad 1 \leq i \leq n,$$

it will follow that  $\tilde{\varphi}_0 \in H^1(\mathbf{R}^n)$  and hence  $\varphi_0 \in H_0^1(\Omega)$ , with  $\varphi_i = D_i \varphi_0$ , which amounts to saying that  $\varphi = \bar{\omega} u$ ,  $u = \varphi_0$ .

Let  $\theta$  be any test function in  $\mathcal{D}(\mathbf{R}^n)$ :

$$\begin{aligned} \int_{\mathbf{R}^n} \tilde{u}_h D_i \theta \, dx &\rightarrow \int_{\mathbf{R}^n} \tilde{\varphi}_0 D_i \theta \, dx \\ \int_{\mathbf{R}^n} (D_{ih}^* \tilde{u}_h) \theta \, dx &\rightarrow \int_{\mathbf{R}^n} \tilde{\varphi}_i \theta \, dx. \end{aligned}$$

Equality (10.35) is proved if we show that

$$\int_{\mathbf{R}^n} \tilde{\varphi}_i \theta \, dx = - \int_{\mathbf{R}^n} \tilde{\varphi}_0 D_i \theta \, dx, \quad \forall \theta \in \mathcal{D}(\mathbf{R}^n),$$

or

$$(10.36) \quad \mathcal{J}_{h'} = \int_{\mathbb{R}^n} \tilde{u}_{h'} D_i \theta \, dx + \int_{\mathbb{R}^n} (D_{ih'}^* \tilde{u}_{h'}) \theta \, dx \rightarrow 0, \quad h' \rightarrow 0,$$

for each  $\theta \in \mathcal{D}(\mathbb{R}^n)$ .

We write now  $\mathcal{J}_h$  in the following way:

$$\begin{aligned} \mathcal{J}_h &= \int_{\Omega(h)} u_h D_i \theta \, dx + \int_{\Omega(h)} (D_{ih}^* u_h) \theta \, dx \\ &= \sum_{\mathcal{S} \in \mathcal{T}_h} \int_{\mathcal{S}} [u_h D_i \theta + (D_i u_h) \theta] \, dx \\ &= \sum_{\mathcal{S} \in \mathcal{T}_h} \int_{\mathcal{S}} D_i (u_h \theta) \, dx. \end{aligned}$$

By the Green–Stokes formula, this gives:

$$\int_{\mathcal{S}} D_i (u_h \theta) \, dx = \int_{\partial \mathcal{S}} u_h \theta v_{i, \mathcal{S}} \, d\Gamma,$$

where  $v_{i, \mathcal{S}}$  is the  $i$ -th component of the unit vector  $v_{\mathcal{S}}$ , normal to  $\partial \mathcal{S}$  and pointing outward with respect to  $\mathcal{S}$ ;  $\partial \mathcal{S}$  is composed of  $(n+1)$   $(n-1)$ -dimensional faces  $\mathcal{S}'$ ; we denote by  $\partial^+ \mathcal{S}$  the sets of these faces and then:

$$\int_{\mathcal{S}} D_i (u_h \theta) \, dx = \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} u_h \theta v_{i, \mathcal{S}'} \, d\Gamma.$$

Finally:

$$\mathcal{J}_h = \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} u_h \theta v_{i, \mathcal{S}'} d\Gamma.$$

The next two lemmas have the purpose of proving (10.36).

LEMMA 6

$$(10.37) \quad \mathcal{J}_h = \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} (u_h(x) - u_h(B)) \times \\ \times (\theta(x) - \theta(B)) v_{i, \mathcal{S}'} d\Gamma,$$

where  $B = B_{\mathcal{S}'}$  is the barycenter of  $\mathcal{S}'$ .

*Proof.* We show first that

$$(10.38) \quad \mathcal{J}_h = \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} (u_h(x) - \\ - u_h(B)) \theta(x) v_{i, \mathcal{S}'} d\Gamma.$$

To prove this equality we must have to show that

$$(10.39) \quad \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} u_h(B) \theta(x) v_{i, \mathcal{S}'} d\Gamma = 0.$$

But for a face  $\mathcal{S}'$  belonging to the boundary of  $\Omega(h)$ ,  $u_h(B) = 0$  and the contribution of this face in the sum is zero. If  $\mathcal{S}'$  belongs to the boundary of two adjacent simplices, then this face gives two opposite contributions to the sum:  $u_h(B)$  and  $\theta(x)$  are the same,  $v_{i, \mathcal{S}'}$  are equal but with opposite signs, depending on whether  $\mathcal{S}'$  is considered as part of the boundary of one or the other simplex. Hence the sum (10.39) is zero.

Equality (10.37) is then easily deduced from (10.38), if we prove that

$$(10.40) \quad \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{\mathcal{S}' \in \partial^+ \mathcal{S}} \int_{\mathcal{S}'} (u_h(x) - u_h(B)) \theta(B) v_{i, \mathcal{S}'} d\Gamma = 0.$$

But for (10.40) we remark more simply that

$$\int_{\mathcal{S}'} [u_h(x) - u_h(B)] \theta(B) v_{i, \mathcal{S}'} d\Gamma = 0,$$

since  $\theta(B)$  and  $v_{i, \mathcal{S}'}$  are constant on  $\mathcal{S}'$  and

$$(10.41) \quad \int_{\mathcal{S}'} u_h(x) dx = u_h(B) \int_{\mathcal{S}'} d\Gamma,$$

because  $u_h$  is linear on  $\mathcal{S}'$ , and  $B$  is the barycenter of  $\mathcal{S}'$ .

#### LEMMA 7

$$(10.42) \quad \mathcal{S}_h \rightarrow 0, \text{ as } \rho(h) \rightarrow 0 \text{ with } \sigma(h) \leq \alpha.$$

*Proof.* Since  $|v_{i, \mathcal{S}'}| \leq 1$  and

$$|\theta(x) - \theta(B)| \leq c(\theta) \rho_{\mathcal{S}}, \quad \forall x \in \mathcal{S}',$$

we can majorize as follows:

$$(10.43) \quad \left| \int_{\mathcal{S}'} (u_h(x) - u_h(B)) (\theta(x) - \theta(B)) v_{i, \mathcal{S}'} d\Gamma \right| \leq \\ \leq c(\theta) \rho_{\mathcal{S}} \int_{\mathcal{S}'} |u_h(x) - u_h(B)| d\Gamma.$$

Since  $u_h(x) - u_h(B)$  is a linear function vanishing at  $x=B$ , we can write it as:

$$u_h(x) - u_h(B) = \sum_{i=1}^n \frac{\partial u_h}{\partial x_i} \cdot (x_i - \beta_i),$$

where  $\beta_1, \dots, \beta_n$  are the coordinates of  $B$ . Therefore

$$|u_h(x) - u_h(B)| \leq \rho_{\mathcal{S}} \sum_{i=1}^n \left| \frac{\partial u_h}{\partial x_i} \right|, \quad \forall x \in \mathcal{S},$$

and

$$\begin{aligned} c(\theta) \rho_{\mathcal{S}} \int_{\mathcal{S}'} |u_h(x) - u_h(B)| \, d\Gamma &\leq \\ &\leq c(\theta) \rho_{\mathcal{S}}^2 \int_{\mathcal{S}'} \left( \sum_{i=1}^n \left| \frac{\partial u_h}{\partial x_i} \right| \right) d\Gamma = \\ &= c(\theta) \rho_{\mathcal{S}}^2 (\text{area } \mathcal{S}') \left\{ \sum_{i=1}^n \left| \frac{\partial u_h}{\partial x_i} \right| \right\}. \end{aligned}$$

Let us denote the vertices of  $\mathcal{S}$  by  $A_1, \dots, A_{n+1}$ , and let us suppose that  $\mathcal{S}'$  is the face containing  $A_2, \dots, A_{n+1}$ ; let  $\xi$  be the distance between  $A_1$  and  $\mathcal{S}'$ . The following elementary formula is easily checked:

$$\text{vol } \mathcal{S} = \frac{1}{n} \xi (\text{area } \mathcal{S}'),$$

and hence:

$$(10.44) \quad \text{area } \mathcal{S}' = \frac{n}{\xi} \text{vol } \mathcal{S} \leq \frac{n}{2\rho'_{\mathcal{S}}} \text{vol } \mathcal{S},$$

since

$$(10.45) \quad 2\rho'_{\mathcal{S}} \leq \xi.$$

The reason of (10.45) is that a ball included in  $\mathcal{S}$  is included in the set bounded by the hyperplane containing  $A_2, \dots, A_{n+1}$ , and the parallel hyperplane containing  $A_1$ .

We then have:

$$\begin{aligned} c(\theta) \rho_{\mathcal{S}} \int_{\mathcal{S}'} |u_h(x) - u_h(B)| \, d\Gamma &\leq \\ &\leq \frac{c(\theta) n \rho_{\mathcal{S}}^2}{2 \rho'_{\mathcal{S}}} (\text{vol } \mathcal{S}) \left\{ \sum_{i=1}^n \left| \frac{\partial u_h}{\partial x_i} \right| \right\} \leq \\ &\leq \frac{c(\theta) n \alpha}{2} \rho(h) \int_{\mathcal{S}'} \left( \sum_{i=1}^n \left| \frac{\partial u_h}{\partial x_i} \right| \right) dx. \end{aligned}$$

Combining this with (10.43) and (10.37) we obtain the following majorization of  $\mathcal{J}_h$ :

$$\begin{aligned} |\mathcal{J}_h| &\leq \frac{c(\theta) n(n+1)}{2} \alpha \rho(h) \sum_{\mathcal{S} \in \mathcal{T}_h} \sum_{i=1}^n \int_{\mathcal{S}} \left| \frac{\partial u_h}{\partial x_i} \right| dx \leq \\ &\leq c(\theta, n, \alpha) \rho(h) \int_{\Omega} |\text{grad } u_h| dx \leq \\ &\leq (\text{by the Schwarz inequality}) \\ &\leq c(\theta, n, \alpha, \Omega) \rho(h) \|u_h\|_h. \end{aligned}$$

Since  $\|u_h\|_h$  is bounded and  $\rho(h) \rightarrow 0$ , we see that  $\mathcal{J}_h \rightarrow 0$ .

*Remark.* We leave the study of similar approximations of the spaces  $L^2(\Omega)$  and  $H^1(\Omega)$  as an exercise to the reader.

### NOTES

<sup>1</sup> I.e. the points such that the barycentric coordinates with respect to the vertices of the simplex are  $\lambda_i$ ,  $0 < \lambda_i < 1$ ,  $1 \leq i \leq n+1$ .

<sup>2</sup>  $w_{hB}$ , for  $B = B_1 = B'_1$ , has its support contained in  $\mathcal{S} \cup \mathcal{S}'$ , and  $w_{hB}(B_2) = w_{hB}(B_3) = w_{hB}(B'_2) = w_{hB}(B'_3) = 0$ ,  $w_{hB}(B_1) = 1$ .

## CHAPTER 11

### EXAMPLE I: THE DIRICHLET PROBLEM

In this section and the following we apply the theory as developed in part one to two examples of elliptic boundary-value problems.

#### 11.1. Exact Problem

We are in the situation as discussed in Section 1.2;  $\Omega$  is an open bounded set in  $\mathbf{R}^n$  with boundary  $\Gamma$ ; we put  $H=L^2(\Omega)$  and  $V=H_0^1(\Omega)$  and these spaces carry the usual (cf. Chapter 7) Hilbert structure:

$$(f, g) = \int_{\Omega} f(x) g(x) \, dx$$

$$|f| = (f, f)^{1/2}, \quad \forall f, g \in H$$

$$((u, v)) = (u, v) + \sum_{i=1}^n (D_i u, D_i v)$$

$$\|u\| = ((u, u))^{1/2}, \quad \forall u, v \in V.$$

The injection of  $V$  into  $H$  is continuous ( $|u| \leq \|u\|$ ,  $\forall u \in V$ ), and  $V$  is dense in  $H$ , because  $V$  contains  $\mathcal{D}(\Omega)$ , which is dense in  $L^2(\Omega)$ .

The form  $a(u, v)$  is chosen as follows:

$$\sum_{i=1}^n (D_i u, D_i v) + \lambda(u, v), \lambda > 0.^1$$

Clearly  $a(u, v)$  is a bilinear form on  $V \times V$ , and this form is continuous because of the Schwarz inequality:

$$a(u, v) \leq \max(1, \lambda) \cdot \left\{ \sum_{i=1}^n |D_i u| |D_i v| + |u| |v| \right\}$$

$$a(u, v) \leq \max(1, \lambda) \cdot \|u\| \|v\|, \quad \forall u, v \in V.$$

The condition of coercivity (1.1) is satisfied with  $\alpha = \min(1, \lambda)$ .

The projection theorem gives in this case:

*For given  $f$  in  $L^2(\Omega)$ , there exists one and only one  $u$  in  $H_0^1(\Omega)$  such that:*

$$(11.1) \quad a(u, v) = (f, v), \quad \forall v \in H_0^1(\Omega).$$

We now interpret this solution. If we put  $v = \varphi \in \mathcal{D}(\Omega)$  in (10.1), we find:

$$\sum_{i=1}^n (D_i u, D_i \varphi) + \lambda(u, \varphi) = (f, \varphi), \quad \forall \varphi \in \mathcal{D}(\Omega)$$

which shows that:

$$-\Delta u + \lambda u = f$$

in the sense of distributions in  $\Omega$ . On the other hand,

because of the trace theorem mentioned in Section 7.2,  $u \in H_0^1(\Omega)$  means that the trace of  $u$  on the boundary  $\Gamma$  of  $\Omega$  is zero, at least when this boundary is smooth enough. So finally we have:

*The solution  $u$  of (11.1) solves the Dirichlet problem:*

$$(11.2) \quad \begin{aligned} -\Delta u + \lambda u &= f \\ \gamma_0 u &= 0 \quad (\text{or } "u|_L = 0"). \end{aligned}$$

*Remark.* According to Proposition 2 of Chapter 1, the solution  $u$  of (11.1) is the element of  $H_0^1(\Omega)$ , for which the following functional takes its minimum:

$$\begin{aligned} E(v) &= a(v, v) - 2(f, v) \\ &= \sum_{i=1}^n |D_i v|^2 + \lambda |v|^2 - 2(f, v). \end{aligned}$$

*Remark.* (11.1) is also called the variational form of the Dirichlet problem (11.2). The preceding remark justifies that name.

## 11.2. Approximation by Finite Differences

We are now going to study how the approximation by finite differences of problem (11.1), (11.2) fits into the framework of Chapter 4. We use the external approximation of  $H_0^1(\Omega)$  which we discussed in Section 9.3, and we define presently

the forms  $a_h$  and  $l_h$  (cf. Section 4.1):

$$(11.3) \quad a_h(u_h, v_h) = \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx + \\ + \lambda \int_{\Omega} u_h(x) v_h(x) dx$$

$$(11.4) \quad \langle l_h, v_h \rangle = \int_{\Omega} f(x) v_h(x) dx.$$

Clearly condition (4.1) is satisfied with  $\alpha_0 = \min(1, \lambda)$ , while condition (4.2) is satisfied with  $\beta = |f|$ .

We now show that the hypotheses (4.5) to (4.8) are satisfied; let us consider (4.5) first. Suppose the family  $v_h$  converges weakly to  $v \in H_0^1(\Omega)$ , for  $h \rightarrow 0$ . Then (4.5) means that:

$$(11.5) \quad \begin{cases} \lim_{h \rightarrow 0} v_h|_{\Omega} = v \\ \lim_{h \rightarrow 0} \delta_i v_h|_{\Omega} = D_i v \quad (i = 1, \dots, n) \end{cases}$$

in the weak topology of  $L^2(\Omega)$ . For all  $w \in V$ , we have according to Proposition 3 of Chapter 9:

$$(11.6) \quad \begin{cases} \lim_{h \rightarrow 0} r_h w|_{\Omega} = w \\ \lim_{h \rightarrow 0} \delta_i r_h w|_{\Omega} = D_i w \quad (i = 1, \dots, n) \end{cases}$$

in the strong topology of  $L^2(\Omega)$ ; hence it is clear

that:

$$\begin{aligned}
 \lim_{h \rightarrow 0} \int_{\Omega} v_h(x) r_h w(x) \, dx &= \\
 &= \int_{\Omega} v(x) w(x) \, dx \lim_{h \rightarrow 0} \int_{\Omega} \delta_i v_h(x) \delta_i r_h w(x) \, dx = \\
 &= \int_{\Omega} D_i v(x) D_i w(x) \, dx \quad (i = 1, \dots, n)
 \end{aligned}$$

and indeed:

$$\lim_{h \rightarrow 0} a_h(v_h, r_h w) = a(v, w).$$

Condition (4.6) is the same as (4.5), because  $a$  and  $a_h$  are symmetric. For all  $w \in V$  it follows from (11.6) that:

$$\begin{aligned}
 \lim_{h \rightarrow 0} |r_h w| &= |w| \\
 \lim_{h \rightarrow 0} |\delta_i r_h w| &= |D_i w| \quad (i = 1, \dots, n)
 \end{aligned}$$

and this shows that:

$$\lim_{h \rightarrow 0} a_h(r_h w, r_h w) = a(w, w).$$

Condition (4.7) follows. Finally, by (11.5), condition (4.8) follows immediately.

*The approximated problem*

The approximated problem (4.3) is

$$(11.7) \quad a_h(u_h, v_h) = \langle l_h, v_h \rangle, \quad \forall v_h \in V_h$$

can be made slightly more explicit. We write:

$$u_h = \sum_{M \in \Omega_h^1} u_h(M) w_{hM}.$$

The unknowns are the numbers  $u_h(M)$ ; as the set of functions  $w_{hP}$ ,  $P \in \Omega_h^1$  forms a basis of  $V_h$ , it is sufficient to write (11.7) with only the functions  $v_h = w_{hP}$ . In this way we find:

$$\begin{aligned} (11.8) \quad \sum_{M \in \Omega_h^1} u_h(M) a_h(w_{hM}, w_{hP}) &= \\ &= \int_{\Omega} f(x) w_{hP}(x) dx, \quad \forall P \in \Omega_h^1. \end{aligned}$$

*We have here a system of  $n(h)$  linear equations for the  $n(h)$  unknowns  $u_h(P)$  in which  $n(h)$  = the number of points of  $\Omega_h^1$ . The matrix of this system is symmetric (as  $a_h$  is symmetric) and positive definite: in fact, for any choice of numbers  $\xi_M \in R$ ,  $M \in \Omega_h^1$  we have:*

$$\begin{aligned} \sum_{M, P} \xi_M \xi_P a_h(w_{hM}, w_{hP}) &= a \left( \sum_M \xi_M w_{hM}, \sum_P \xi_P w_{hP} \right) \geq \\ &\geq \min(1, \lambda) \left\| \sum_M \xi_M w_{hM} \right\|_h^2. \end{aligned}$$

We can write (11.8) also in another form; this equality is equivalent to

$$a_h(u_h, w_{hP}) = \int_{\Omega} f(x) w_{hP}(x) dx;$$

can also be written as:

$$\sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i w_{hP}(x) dx + \lambda \int_{\Omega} u_h(x) w_{hP}(x) dx = \int_{\Omega} f(x) w_{hP}(x) dx.$$

Because the functions  $\delta_i u_h$  and  $\delta_i w_{hP}$  have compact support in  $\Omega$ , we have:

$$\int_{\Omega} \delta_i u_h(x) \delta_i w_{hP}(x) dx = \int_{\mathbb{R}^n} \delta_i u_h(x) \delta_i w_{hP}(x) dx.$$

The following ‘discrete’ integration by parts formula can easily be checked:

$$\begin{aligned} \int_{\mathbb{R}^n} \varphi_h(x) \delta_i \psi_h(x) dx &= \\ &= - \int_{\mathbb{R}^n} \delta_i \varphi_h(x) \psi_h(x) dx, \quad \forall \varphi_h, \psi_h \in V_h, \end{aligned}$$

and so we obtain:

$$\int_{\mathbb{R}^n} \left( - \sum_{i=1}^n \delta_i^2 u_h + \lambda u_h - f \right) w_{hP} dx = 0, \quad \forall P \in \tilde{\Omega}_h^1$$

or:

$$\begin{aligned}
 (11.9) \quad & - \sum_{i=1}^n \delta_i^2 u_h(M) + \lambda u_h(M) \\
 & = (h_1 \dots h_n)^{-1} \int_{\sigma_h(P)} f(x) \, dx, \quad \forall P \in \dot{\Omega}_h^1.
 \end{aligned}$$

### *Convergence result*

Theorem 2 of Chapter 4 can be applied; the most useful result is the strong convergence of  $u_h$  to  $u$ , and this means:

$$\begin{aligned}
 \lim_{h \rightarrow 0} u_h|_{\Omega} &= u \\
 \lim_{h \rightarrow 0} \delta_i u_h|_{\Omega} &= D_i u \quad (i = 1, \dots, n)
 \end{aligned}$$

in the strong topology of  $L^2(\Omega)$  ( $u_h$  is solution of (11.6) and  $u$  is solution of (11.1)).

### **11.3. Other Results**

*A priori estimate of the approximate solution in  $L^\infty(\Omega)$*

It can be shown that, if  $f \in L^\infty(\Omega)$ , then the solution of (11.1) belongs to  $L^\infty(\Omega)$  and satisfies:

$$(11.10) \quad \|u\|_{L^\infty(\Omega)} \leq \frac{1}{\lambda} \|f\|_{L^\infty(\Omega)}.$$

We want to show that the solution  $u_h$  of (11.6) also satisfies:

$$(11.11) \quad \|u_h\|_{L^\infty(\Omega)} \leq \frac{1}{\lambda} \|f\|_{L^\infty(\Omega)},$$

and from this we want to deduce a new result on the convergence of  $u_h$  to  $u$ .

We denote:

$$\mu = \|f\|_{L^\infty(\Omega)}, \nu = \mu/\lambda.$$

For each real valued function  $\varphi$ ,  $\varphi^+$  and  $\varphi^-$  are functions defined on the same set as  $\varphi$  and such that:

$$\begin{aligned} \varphi^+(x) &= \begin{cases} \varphi(x) & \text{if } \varphi(x) \geq 0 \\ 0 & \text{if } \varphi(x) \leq 0 \end{cases} \\ \varphi^-(x) &= \begin{cases} 0 & \text{if } \varphi(x) \geq 0 \\ -\varphi(x) & \text{if } \varphi(x) \leq 0. \end{cases} \end{aligned}$$

In particular, for  $v_h \in V_h$ ,  $v_h^+$  and  $v_h^-$  are in  $V_h$  too. For  $v_h \in V$ ,  $(v_h - \nu)^+ \in V_h$ .

In order to prove that:

$$(11.12) \quad u_h(x) \leq \nu, \quad \forall x \in \Omega,$$

we write (11.7) with  $v_h = (u_h - \nu)^+$ ; this is allowed according to the preceding. We then get:

$$\begin{aligned} (11.13) \quad & \sum_{i=1}^n \int_{\Omega} \delta_i \varphi_h(x) \delta_i \varphi_h^+(x) dx + \\ & + \lambda \int_{\Omega} \varphi_h(x) \varphi_h^+(x) dx = \int_{\Omega} (f(x) - \lambda \nu) \varphi_h^+(x) dx, \end{aligned}$$

where:

$$\varphi_h = u_h - \nu.$$

The right-hand member of (11.13) is negative or zero, because for almost all  $x$ :

$$f(x) \leq \lambda v = \mu \quad \text{and} \quad \varphi_h^+(x) \geq 0.$$

As  $\varphi_h = \varphi_h^+ - \varphi_h^-$  and  $\varphi_h^+ \cdot \varphi_h^- = 0$ , we have:

$$\int_{\Omega} \varphi_h(x) \varphi_h^+(x) \, dx = \int_{\Omega} (\varphi_h^+(x))^2 \, dx.$$

On the other hand, for all  $x \in \mathbf{R}^n$ :

$$(11.14) \quad (\delta_i \varphi_h^+(x))^2 \leq \delta_i \varphi_h(x) \delta_i \varphi_h^+(x).$$

To prove this, we remark that the functions  $\delta_i \varphi_h$  and  $\delta_i \varphi_h^+$  are constant on the cells:

$$\sigma_h \left( M + \frac{h_i e_i}{2} \right), \quad M \in R_h;$$

so it is sufficient to consider a point  $x = M + h_i e_i / 2$ ,  $M \in R_h$  and for such a point  $x$  there are four cases to consider, according to the sign of  $\varphi_h(M)$  and the sign of  $\varphi_h(M + h_i e_i)$ .

Finally, (11.13) gives:

$$\int_{\Omega} (\varphi_h^+(x))^2 \, dx \leq 0,$$

which proves that the function  $\varphi_h^+ = (u_h - v)^+$  is zero on  $\Omega$  and hence (11.12) follows.

Likewise, by putting  $v_h = (u_h + v)^-$  in (11.6) we prove

that:

$$u_h(x) \geq -v, \quad \forall x \in \Omega,$$

and (11.11) follows.

*Weak convergence in  $L^\infty(\Omega)$*

It is well known that  $L^\infty(\Omega)$  is the dual of  $L^1(\Omega)$ ; hence, by (11.11) (cf. Appendix), we see that there exists a  $u_* \in L^\infty(\Omega)$  and a sequence  $h' \rightarrow 0$  such that

$$u_{h'}|_\Omega \rightarrow u_* \quad \text{in the weak star topology of } L^\infty(\Omega).$$

This means that for each function  $\varphi \in L^1(\Omega)$ :

$$\lim_{h' \rightarrow 0} \int_{\Omega} u_{h'}(x) \varphi(x) dx = \int_{\Omega} u_*(x) \varphi(x) dx.$$

Hence  $u_{h'}$  converges to  $u_*$  in the sense of distributions and consequently  $u_* = u$ .<sup>1</sup> Likewise, from every subsequence of  $u_h$  we can extract a subsequence which converges to  $u$  in the weak star topology of  $L^\infty(\Omega)$ ; this means that the family  $u_h$  as a whole converges to  $u$  for the weak star topology of  $L^\infty(\Omega)$ .

#### 11.4. The Method of Fractionary Steps

We now want to investigate how the method of fractionary steps applies to the problem of solving (11.7). The space  $V_h$  is finite dimensional, so we are in the situation of the

remark in Section 6.1. The space  $V_h$  is now endowed with the following scalar products:

$$\begin{aligned}(u_h, v_h)_h &= \int_{\Omega} u_h(x) v_h(x) dx \\ ((u_h, v_h))_{ih} &= (u_h, v_h)_h + (\delta_i u_h, \delta_i v_h)_h \quad (i = 1, \dots, n) \\ ((u_h, v_h))_h &= (u_h, v_h)_h + \sum (\delta_i u_h, \delta_i v_h)_h.\end{aligned}$$

We put  $(i=1, \dots, n; m=n)$ :

$$a_{ih}(u_h, v_h) = (\delta_i u_h, \delta_i v_h)_h + \frac{\lambda}{n} (u_h, v_h)_h.$$

All conditions of Section 6.1 are satisfied:

$$\alpha_i = \gamma_i = \min(1, \lambda/n).$$

The approximation scheme outlined in (6.14) now takes the following form:

$$\begin{aligned}(11.15) \quad (u_h^{r+i/n} - u_h^{r+(i-1)/n}, v_h)_h + \tau a_{ih}(u_h^{r+i/n}, v_h) &= \\ &= \tau \int_{\Omega} f_i(x) v_h(x) dx, \quad \forall v_h \in V_h,\end{aligned}$$

where  $f = \sum_{i=1}^n f_i$  is an arbitrary decomposition of  $f$  into elements of  $L^2(\Omega)$ .

To render (11.15) more explicit, we follow the same procedure as for (11.8). We decompose  $u_h^{r+i/n}$  into components with respect to the basis consisting of the  $w_{hM}$ :

$$u_h^{r+i/n} = \sum_{M \in \Omega_h^1} \xi_M w_{hM}, \quad \xi_M \in \mathbf{R}$$

and we write (11.15) with  $v_h = w_{hP}$ ,  $P \in \hat{\Omega}_h^1$ :

$$(11.16) \quad \xi_P(h_1 \dots h_n) + \sum_{M \in \hat{\Omega}_h^1} \tau a_{ih}(w_{hM}, w_{hP}) = \\ = (u_h^{r+(i-1)/n}, w_{hP})_h + \tau \int_{\sigma_h(P)} f_i(x) dx.$$

It is easy to see, as in the case of (11.8), that the linear system (11.16) has a positive definite matrix. The important point in the present case is the fact that the system (11.16) uncouples into partial systems corresponding with sets of  $M \in \hat{\Omega}_h^1$  lying on the same line parallel to the coordinate axes. It is seen that these partial systems have a positive definite tridiagonal matrix, which makes their actual solution rather easy to compute, by Gauss elimination.

*Remark.* We can, as in the case of (11.9), write (11.15) in the form

$$(11.17) \quad -\tau \delta_i^2 u_h^{r+i/n}(P) + \left(1 + \frac{\tau \lambda}{n}\right) u_h^{r+i/n}(P) = \\ = u_h^{r+(i-1)/n}(P) + \tau (h_1 \dots h_n)^{-1} \int_{\sigma_h(P)} f_i(x) dx, \\ \forall P \in \hat{\Omega}_h^1 \quad (i = 1, \dots, n; \quad r = 0, \dots, N).$$

### *A numerical application*

The preceding considerations, and in particular those of

this section have been applied to the numerical solution of the equation:

$$-\Delta u + u = f, \quad u|_r = 0$$

in  $\Omega = ]1,2[ \times ]1,2[ \subset \mathbb{R}^2$ , with:

$$\begin{aligned} f &= f_1 + f_2 \\ f_1 &= x_2(x_1 - 1)(x_1 - 2)(x_2 - 3) \\ f_2 &= -2(x_2 - 1)(x_2 - 2). \end{aligned}$$

The exact solution, which we know in advance, and that we actually want to rediscover, is:

$$u = (x_1 - 1)(x_1 - 2)(x_2 - 1)(x_2 - 2).$$

The following table gives the normalized error:

$$e_h = \frac{|u_h - r_h u|_h}{|u_h|_h}$$

for the indicated values of  $\tau, h$  ( $=h_1=h_2$ =step length for the variables) and  $N$  (=number of iterations).

| $\tau$ | $h=1/10$               |     | $h=1/20$               |     | $h=1/30$               |     |
|--------|------------------------|-----|------------------------|-----|------------------------|-----|
|        | $e_h$                  | $N$ | $e_h$                  | $N$ | $e_h$                  | $N$ |
| 1/100  | $0.267 \times 10^{-2}$ | 49  | $0.181 \times 10^{-2}$ | 48  | $0.153 \times 10^{-2}$ | 48  |
| 1/200  | $0.696 \times 10^{-3}$ | 83  | $0.468 \times 10^{-3}$ | 81  | $0.394 \times 10^{-3}$ | 81  |
| 1/300  | $0.315 \times 10^{-3}$ | 113 | $0.211 \times 10^{-3}$ | 112 | $0.177 \times 10^{-3}$ | 112 |
| 1/400  | $0.179 \times 10^{-3}$ | 143 | $0.120 \times 10^{-3}$ | 142 | $0.101 \times 10^{-3}$ | 142 |

The computation time is approximately proportional

to the number of iterations and the number of cells used.

### 11.5. Miscellaneous Remarks

(i) All the preceding results can be extended to problems (11.1) and (11.2) with  $\lambda = 0$ . The coercivity of the exact problem is obtained using the so-called Poincaré Inequality; in the discrete case we can use a discrete Poincaré Inequality; some classical techniques and their discrete extensions are available to prove a maximum principle.

(ii) It is not difficult to study the approximation of problems (11.1) and (11.2), using the two approximations of the space  $H_0^1(\Omega)$  described in Chapter 10, and appropriate approximations of  $a$  and  $l$ . However, the particular techniques developed in Sections 11.3 and 11.4 (discrete maximum principle and fractionary step method) are not applicable in such a direct way. In the case of the conforming finite element approximation, the methods of Chapter 5 concerning estimation of the error work out easily. All these questions are left as exercises.

(iii) Using the preceding approximation of  $H_0^1(\Omega)$  one can extend the results of Sections 11.1 and 11.2 to more general second-order Dirichlet problems (involving, for example, mixed derivatives  $D_{ij}u$ ). Many other Dirichlet-type problems of higher order can be set in the framework of Chapter 1, and approximated by the methods of Chapters 3 and 4, using appropriate techniques, very close to the techniques used for problems (11.1) and (11.2).

## NOTES

<sup>1</sup> The hypothesis  $\lambda > 0$  is put here for simplification. The case  $\lambda = 0$  can be handled with suitable modifications. Some cases with  $\lambda > 0$  even can be treated too. (See Section 11.5(i).)

<sup>2</sup> Which shows (11.10).

## CHAPTER 12

### EXAMPLE II: THE NEUMANN PROBLEM

#### 12.1. The Exact Problem

We are in the situation discussed in Section 1.2;  $\Omega$  is a bounded open set in  $\mathbf{R}^n$  with boundary  $\Gamma$ , we put  $H=L^2(\Omega)$  and  $V=H^1(\Omega)$  and these spaces are provided with their usual Hilbert structure (cf. Chapter 7):

$$(f, g) = \int_{\Omega} f(x) g(x) dx$$

$$|f| = (f, f)^{1/2}, \quad \forall f, g \in H$$

$$((u, v)) = (u, v) + \sum_{i=1}^n (D_i u, D_i v)$$

$$\|u\| = ((u, u))^{1/2}, \quad \forall u, v \in V.$$

The injection of  $V$  into  $H$  is continuous ( $|u| \leq \|u\|, \forall u \in V$ ) and  $V$ , which contains  $\mathcal{D}(\Omega)$ , is dense in  $H$ .

The form  $a(u, v)$  is the same as in Chapter 11:

$$a(u, v) = \sum_{i=1}^n (D_i u, D_i v) + \lambda(u, v), \quad \lambda > 0,$$

so the only difference with Section 11.1 is the change of space  $V$ .

Theorem 1 of Chapter 1 gives in this case:

*For each given  $f$  in  $L^2(\Omega)$ , there exists one and only one  $u \in H^1(\Omega)$  such that:*

$$(12.1) \quad a(u, v) = f(v), \quad \forall v \in H^1(\Omega).$$

If we substitute  $v = \varphi \in \mathcal{D}(\Omega)$ , which is allowed, we see like in Section 11.1, that  $u$  satisfies the equation:

$$(12.2) \quad -\Delta u + \lambda u = f$$

in the sense of distributions on  $\Omega$ . It follows from (12.2) that  $\Delta u \in L^2(\Omega)$ .

We now take the scalar product of each of the members of (12.2) with  $v \in H^1(\Omega)$ , but in the  $L^2(\Omega)$  sense. We obtain:

$$-(\Delta u, v) + \lambda(u, v) = (f, v).$$

As  $u \in H^1(\Omega)$ ,  $\Delta u \in L^2(\Omega)$ ,  $v \in H^1(\Omega)$ , we can use the generalized Green's formula (7.5) and then we find:

$$\sum_{i=1}^n (D_i u, D_i v) + \lambda(u, v) = \langle f, v \rangle + \langle \gamma_1 u, \gamma_0 v \rangle.$$

Comparing with (12.1) gives us then:

$$(12.3) \quad \langle \gamma_1 u, \gamma_0 v \rangle = 0.$$

As  $v$  is arbitrary in  $H^1(\Omega)$  and  $\gamma_0$  is a surjection of  $H^1(\Omega)$  onto  $H^{1/2}(\Gamma)$  (cf. Section 7.2), equality (11.3) shows that:

$$\gamma_1 u = 0 \quad \text{in} \quad H^{-1/2}(\Gamma).$$

So finally:

*The solution of (12.1) is the solution of the Neumann problem:*

$$(12.4) \quad \begin{aligned} & -\Delta u + \lambda u = f \\ & \gamma_1 u = 0 \quad \left( \text{or} \quad \left. \frac{\partial u}{\partial \nu} \right|_r = 0 \right). \end{aligned}$$

## 12.2. Approximation by Finite Differences

We apply the results of Chapter 4 to the approximation of problem (12.1) or (12.4) by the method of finite differences. We use the external approximation of the space  $H^1(\Omega)$  from Section 9.2. The forms  $a_h$  and  $l_h$  are the following:

$$(12.5) \quad a_h(u_h, v_h) = \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx + \lambda \int_{\mathbb{R}^n} u_h(x) v_h(x) dx$$

$$(12.6) \quad \langle l_h, v_h \rangle = \int_{\Omega} f(x) v_h(x) dx.$$

Condition (4.1) holds with  $\alpha_0 = \min(1, \lambda)$  and condition (4.2) holds with  $\beta = |f|$ . Conditions (4.5) to (4.8) hold too, and this can be checked in the same way as in Section 11.4. One must use note 1 of Chapter 9.

*Approximated problem*

We have to find (cf. Section (4.3)) the  $u_h \in V_h$  such that:

$$(12.7) \quad a_h(u_h, v_h) = \langle l_h, v_h \rangle \quad \forall v_h \in V_h.$$

We write this in the form:

$$u_h = \sum_{M \in \Omega_h^1} u_h(M) w_{hM},$$

where the unknowns are the numbers  $u_h(M)$ , and the set of these numbers is the solution of the following system (where  $v_h = w_{hP}$ ):

$$(12.8) \quad \sum_{M \in \Omega_h^1} u_h(M) a_h(w_{hM}, w_{hP}) = \\ = \int_{\sigma_h(P)} f(x) dx, \quad \forall P \in \Omega_h^1.$$

We see, like in (11.8) that the *matrix of system (12.8) is symmetric and positive definite*.

*Remark.* The computation of the matrix of system (12.8) has to be done numerically; this matrix depends on the measure of the sets  $\sigma_h(M) \cap \Omega$  for points  $M$  on the boundary (that is to say for points  $M$  such that  $M \in \Omega_h^1$  and  $\sigma_h(M)$  has non empty intersection with  $\Omega$  and with the complement of  $\bar{\Omega}$ ). For the same reason (12.8) cannot be converted into an equation analogous to (11.9).

### *Convergence result*

Let  $u$  be the solution of (11.1) and let  $u_h$  be the solution of (12.7). Theorem 2 of Chapter 4 is applicable and  $u_h$  converges strongly to  $u$ ; this means in this case:

$$\begin{aligned}\lim_{h \rightarrow 0} u_h|_{\Omega} &= u \\ \lim_{h \rightarrow 0} \delta_i u_h|_{\Omega} &= D_i u \quad (i = 1, \dots, n)\end{aligned}$$

in the strong topology of  $L^2(\Omega)$ .

*Remark.* For  $f \in L^\infty(\Omega)$  it can be shown that the solution  $u$  of (12.1) belongs to  $L^\infty(\Omega)$  and satisfies:

$$\|u\|_{L^\infty(\Omega)} \leq \frac{1}{\lambda} \|f\|_{L^\infty(\Omega)}.$$

Just like in Section 11.3, it can be shown that the solution  $u_h$  of (12.7) satisfies:

$$\|u_h\|_{L^\infty(\Omega)} \leq \frac{1}{\lambda} \|f\|_{L^\infty(\Omega)},$$

and  $u_h$  converges to  $u$  for the weak star topology of  $L^\infty(\Omega)$ .

## 12.3. The Method of Fractional Steps

Now we will study the application of the method of fractional steps to the problem of solving (12.7). The space

$V_h$  is finite dimensional and we are in the situation of the remark of Section 6.1. The various scalar products on  $V_h$  are:

$$\begin{aligned}(u_h, v_h)_h &= \int_{\mathbb{R}^n} u_h(x) v_h(x) dx \\ ((u_h, v_h))_{ih} &= (u_h, v_h)_h + \\ &\quad + \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx \quad (i = 1, \dots, n) \\ ((u_h, v_h))_h &= (u_h, v_h)_h + \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx.\end{aligned}$$

We put  $(i=1, \dots, n; m=n)$ :

$$a_{ih}(u_h, v_h) = \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx + \frac{\lambda}{n} (u_h, v_h)_h.$$

All conditions of Section 6.1 are satisfied ( $\alpha_i = \gamma_i = \min(1, \lambda/n)$ ).

The approximation scheme (6.14) becomes now:

$$\begin{aligned}(12.9) \quad (u_h^{r+i/n} - u_h^{r+(i-1)/n}, v_h)_h + \tau a_{ih}(u_h^{r+i/n}, v_h) &= \\ &= \tau \int_{\Omega} f_i(x) v_h(x) dx, \quad \forall v_h \in V_h,\end{aligned}$$

where  $f = \sum_{i=1}^n f_i$  is an arbitrary decomposition of  $f$  into elements of  $L^2(\Omega)$ .

If we decompose  $u_h^{r+i/n}$  with respect to the basis consisting of the  $w_{hM}$ ,  $M \in \Omega_h^1$ , we obtain:

$$u_h^{r+i/n} = \sum_{M \in \Omega_h^1} \xi_M w_{hM},$$

and if we write (12.9) with  $v_P = w_{hP}$ ,  $P \in \Omega_h^1$ , we see that the set of  $\xi_M$  is the solution of the system:

$$\begin{aligned} (12.10) \quad \xi_P(h_1 \dots h_n) + \sum_{M \in \Omega_h^1} \tau a_{ih}(w_{hM}, w_{hP}) = \\ = (u_h^{r+(i-1)/n}, w_{hP})_h + \tau \int_{\sigma_h(P)} f_i(x) dx. \end{aligned}$$

System (12.10) is linear in the  $\xi_M$ , and its matrix is symmetric and positive definite. Just as in (11.16) this system uncouples into partial systems corresponding with sets of  $M \in \Omega_h^1$  lying on the lines parallel to the coordinate axes. Each partial system has a symmetric positive definite tridiagonal matrix. The inversion of such matrices is quite a standard procedure.

### *A numerical application*

The example in this case is:

$$\begin{aligned} -\Delta u + 2u &= f \quad \text{in } \Omega = ]0, 1[ \times ]0, 1[ \quad (\Omega \subset \mathbb{R}^2) \\ \frac{\partial u}{\partial \nu} \Big|_r &= 0 \end{aligned}$$

with:

$$f = 2(\pi^2 + 1) \cos \pi x_1 \cos \pi x_2$$

and exact solution:

$$u = \cos \pi x_1 \cos \pi x_2.$$

The table gives the normalized error:

$$e_h = \frac{|u_h - r_h u|_h}{|u_h|_h}$$

for different values of  $\tau$ ,  $h(=h_1=h_2)$  and  $N$ .

| $\tau$ | $h = 1/10$             |      | $h = 1/20$             |      |
|--------|------------------------|------|------------------------|------|
|        | $e_h$                  | $N$  | $e_h$                  | $N$  |
| 1/10   | $0.375 \times 10^{-2}$ | 5    | $0.932 \times 10^{-2}$ | 5    |
| 1/20   | $0.374 \times 10^{-2}$ | 10   | $0.927 \times 10^{-2}$ | 10   |
| 1/40   | $0.372 \times 10^{-2}$ | 18   | $0.907 \times 10^{-2}$ | 18   |
| 1/80   | $0.371 \times 10^{-2}$ | 36   | $0.897 \times 10^{-2}$ | 35   |
| 1/160  | $0.370 \times 10^{-2}$ | 70   | $0.891 \times 10^{-2}$ | 69   |
| 1/320  | $0.370 \times 10^{-2}$ | 198  | $0.888 \times 10^{-2}$ | 137  |
| 1/640  | $0.370 \times 10^{-2}$ | 275  | $0.887 \times 10^{-2}$ | 273  |
| 1/1280 | $0.370 \times 10^{-2}$ | 548  | $0.886 \times 10^{-2}$ | 545  |
| 1/2560 | $0.370 \times 10^{-2}$ | 1094 | $0.885 \times 10^{-2}$ | 1088 |

The computation time is approximately proportional to the number of iterations and the number of cells used.

*Remark.* The remarks made in Section 11.5 are also valid here. The methods are applicable to more general Neumann type problems, mixed Dirichlet-Neumann problems, transmission equations, elasticity equations, etc.

PART III

A NONLINEAR ELLIPTIC  
EQUATION

## CHAPTER 13

### THE EXACT PROBLEM

We will describe here the nonlinear elliptic problem that we want to study, and we prove existence and uniqueness of the solutions of this problem. Existence is shown by the Galerkin method, which at the same time gives first method for approximate solution of the equation.

We do not develop a general approach of the approximation of non-linear elliptic problems, although the methods in the sequel are applicable to a large class of such problems.

#### 13.1. Existence and Uniqueness Theorem

Let  $\Omega$  be an open bounded set in  $\mathbf{R}^n$  whose boundary  $\Gamma$  is sufficiently regular, for example, satisfies (7.3) with  $r=1$ . As before  $(., .)$  and  $|\cdot|$  stand for the scalar product and norm in  $L^2(\Omega)$ . We call  $V$  the space  $H_0^1(\Omega) \cap L^4(\Omega)$  provided with the norm:

$$\|u\| = \|u\|_{H^1(\Omega)} + \|u\|_{L^4(\Omega)}.$$

It is easy to see that  $V$  is a Banach space: indeed, if  $u_\nu$  is a Cauchy sequence in  $V$ , then it is a Cauchy sequence as well in  $H^1(\Omega)$  as in  $L^4(\Omega)$  and hence converges in these

spaces to limits which are necessarily equal to the same function  $u$  since convergence in both these spaces implies convergence in  $L^2(\Omega)$ ; hence  $u \in V$  and  $u_n$  converges to  $u$  in  $V$ .

Let  $a(u, v)$  and  $b(u, v)$  be the following forms defined on  $V \times V$ :

$$a(u, v) = \sum_{i=1}^n (D_i u, D_i v) + \lambda(u, v), \quad \lambda > 0,$$

$$b(u, v) = a(u, v) + \int_{\Omega} u^3(x) v(x) dx$$

Clearly  $a(u, v)$  is a continuous bilinear form on  $V \times V$ ; if  $u$  and  $v$  are in  $L^4(\Omega)$ , then  $u^3 \in L^{4/3}(\Omega)$  and  $u^3 v \in L^1(\Omega)$  so that  $b(u, v)$  is well defined for  $\forall u, v \in V$ .

**THEOREM. 1.** *For each given  $f$  in  $L^2(\Omega)$  there exists one and only one  $u$  in  $V$  such that:*

$$(13.1) \quad b(u, v) = (f, v), \quad \forall v \in V.$$

*This function satisfies:*

$$(13.2) \quad \begin{aligned} & -\Delta u + \lambda u + u^3 = f \\ & \gamma_0 u = 0 \quad (\text{or } "u|_r = 0"). \end{aligned}$$

*Proof.* (a) *Interpretation of (13.1):* If  $u$  satisfies (13.1), then we take  $v = \varphi \in \mathcal{D}(\Omega)$  in (13.1), and then we get:

$$\sum_{i=1}^n (D_i u, D_i \varphi) + \int_{\Omega} (\lambda u(x) + u^3(x) - f(x)) \times \varphi(x) dx = 0,$$

and because  $\varphi$  is arbitrary in  $\mathcal{D}(\Omega)$ , this means that:

$$-\Delta u + \lambda u + u^3 = f$$

in the sense of distributions in  $\Omega$ .

On the other hand,  $u \in H_0^1(\Omega)$  means that  $\gamma_0 u = 0$  or ' $u|_r = 0$ ' (cf. Section 7.2). Hence each function  $u$  of  $V$  that satisfies (13.1) is solution of (13.2).

(b) *Uniqueness of solutions of (13.1)*: Let  $u_1$  and  $u_2$  be two elements of  $V$  that satisfy (13.1) and let  $u = u_1 - u_2$ . Then we have:

$$b(u_1, v) - b(u_2, v) = 0, \quad \forall v \in V,$$

and if we put  $v = u$ , then we get

$$a(u, u) + \int_{\Omega} (u_1^3(x) - u_2^3(x)) (u_1(x) - u_2(x)) \, dx.$$

Since the function  $\xi \rightarrow \xi^3$  is increasing, we evidently have:

$$(u_1^3(x) - u_2^3(x)) (u_1(x) - u_2(x)) \geq 0 \quad \text{a.e.}$$

and hence:

$$\int_{\Omega} (u_1^3(x) - u_2^3(x)) (u_1(x) - u_2(x)) \, dx \geq 0,$$

so that indeed

$$\begin{aligned} a(u, u) &\leq 0 \\ \lambda |u|^2 &\leq 0, \end{aligned}$$

and hence  $u = 0$ ,  $u_1 = u_2$ .

(c) *Existence of a solution of (13.1)*: This is proved in Section 13.3. In 13.2 we give a few preliminary results.

*Remark.* It is proved in the same way as in (b) above, that:

$$(13.3) \quad b(u, u - v) - b(v, u - v) \geq 0, \quad \forall u, v \in V.$$

### 13.2. Lemmas

LEMMA 1. *The space  $V$  is separable* (in other words, there exists a dense, countable family in  $V$ ).

*Proof.* Consider the space:

$$F = L^4(\Omega) \times L^2(\Omega)^{n+1}$$

and the map  $\bar{\omega}$  from  $V$  into  $F$ :

$$u \mapsto \bar{\omega}u = (u, u, D_1u, \dots, D_nu).$$

The map  $\bar{\omega}$  is a nonsurjective isomorphism of normed spaces. We know that the spaces  $L^p(\Omega)$  are separable; hence  $F$  is separable and  $V$  too.

It follows from this lemma that in  $V$ , like in any other separable normed space, there exist total sequences consisting of linear independent elements.

LEMMA 2.  $\{u, v\} \rightarrow b(u, v)$  is a continuous map from  $V \times V$  into  $\mathbf{R}$ .

*Proof.* Because  $\{u, v\} \rightarrow a(u, v)$  is a continuous bilinear

form on  $V \times V$ , we only have to prove that the map:

$$(13.4) \quad \{u, v\} \rightarrow \int_{\Omega} u^3(x) v(x) dx$$

is continuous.

Now the map  $u \rightarrow u^3$  is continuous from  $L^4(\Omega)$  into  $L^{4/3}(\Omega)$ . For  $u, u_*, v, v_* \in V$  (or just in  $L^4(\Omega)$ ) the Hölder inequality gives:

$$\begin{aligned} & \left| \int_{\Omega} u^3(x) v(x) dx - \int_{\Omega} u_*^3(x) v_*(x) dx \right| \leq \\ & \leq \left| \int_{\Omega} u^3(x) (v(x) - v_*(x)) dx \right| + \\ & + \left| \int_{\Omega} (u^3(x) - u_*^3(x)) v_*(x) dx \right| \leq \\ & \leq \|u^3\|_{L^{4/3}(\Omega)} \|v - v_*\|_{L^4(\Omega)} + \\ & + \|u^3 - u_*^3\|_{L^{4/3}(\Omega)} \|v_*\|_{L^4(\Omega)}. \end{aligned}$$

The continuity of map (12.4) follows from this last inequality.

**LEMMA 3.** *Let  $X$  be a finite dimensional Hilbert space, with scalar product  $[\cdot, \cdot]$  and norm  $[\cdot]$ , and let  $P$  be a continuous map from  $X$  into itself such that:*

$$(13.5) \quad [P(\xi), \xi] > 0 \quad \text{for} \quad [\xi] = k > 0.$$

*Then there exists  $\xi \in X$ ,  $[\xi] \leq k$ , such that:*

$$P(\xi) = 0.$$

*Proof.* This is an easy consequence of a classical theorem, namely the Brouwer fixed point theorem. The proof of this theorem, which is rather long and technical, can be found in Dunford and Schwartz [1]. The statement of the theorem is: "Let  $D$  be a closed ball in a finite dimensional space and  $S$  a continuous map of  $D$  into itself. Then there exists at least one point  $\xi \in D$  such that  $S(\xi) = \xi$ ."

Suppose that  $P$  has no zero in the ball with center  $O$  and radius  $k$  in  $X$ . Then the following map of  $D$  into itself is continuous:

$$\xi \mapsto S(\xi) = -\frac{kP(\xi)}{[P(\xi)]}.$$

So there exists an  $\xi_0 \in D$  such that:

$$-\frac{kP(\xi_0)}{[P(\xi_0)]} = \xi_0.$$

If we take the norm of both sides, we see that  $[\xi_0] = k$ , and if we take the scalar products of both sides with  $P(\xi_0)$ , we find:

$$[P(\xi_0), \xi_0] = -k[P(\xi_0)] < 0.$$

and this contradicts (13.5).

### 13.3. The Galerkin Method

#### *Setting up the method*

Let  $w_1, w_2, \dots$ , be a free and total sequence in  $V$  (see Lemma 1). Let  $V_m$  be the space spanned by  $w_1, \dots, w_m$ . We now are going to find, just as in the linear case, for each fixed  $m$ , a  $u_m \in V_m$  such that:

$$(13.6) \quad b(u_m, v) = (f, v), \quad \forall v \in V_m.$$

We cannot use the projection lemma in this case; and we are going to use Lemma 3, to prove the existence of a  $u_m$  which satisfies (13.6). We apply Lemma 3 with the following conditions:

$X = V_m$  with an arbitrary scalar product;

for example, if:

$$u = \sum_{i=1}^m \xi_i w_i, \quad v = \sum_{i=1}^m \eta_i w_i,$$

we take

$$[u, v]_m = \sum_{i=1}^m \xi_i \eta_i.$$

Because  $V_m$  is finite dimensional, the norm  $[.]$  associated with this scalar product is equivalent to any other norm; hence there exists a constant  $c_m$ , depending on  $m$ , such that:

$$c_m |u| \leq [u]_m \leq \frac{1}{c_m} |u|, \quad \forall u \in V_m.$$

The map  $P$ , defined on  $V_m$  will be:

$$u \mapsto P(u) = \sum_{i=1}^m \{b(u, w_i) - (f, w_i)\} w_i.$$

The continuity of  $P$  follows easily from Lemma 2. On the other hand, for  $u = \sum_{i=1}^m \xi_i w_i$  we have:

$$\begin{aligned} [P(u), u]_m &= \sum_{i=1}^m \xi_i \{b(u, w_i) - (f, w_i)\} = \\ &= b(u, u) - (f, u) \geq \\ &\geq \lambda |u|^2 - |f| |u| \geq \\ &\geq \lambda c_m^2 [u]_m^2 - \frac{1}{c_m} |f| [u]_m. \end{aligned}$$

So it follows that  $[P(u), u]_m > 0$  for  $[u]_m = k$  and  $k$  large enough ( $k > |f|/\lambda c^3$ ). The hypotheses of Lemma 3 are satisfied, and hence there exists a  $u_m \in V_m$  such that  $P(u_m) = 0$ , which means:

$$b(u_m, w_i) = (f, w_i), \quad i = 1, \dots, m.$$

Hence  $u_m$  satisfies (13.6).

### *Passage to the limit*

We put  $v = u_m$  in (13.6); then we get:

$$\begin{aligned} (13.7) \quad b(u_m, u_m) &= (f, u_m) \\ b(u_m, u_m) &\leq |f| |u_m| \leq |f| \|u_m\|_{H^1(\Omega)} \\ \min(1, \lambda) \|u_m\|_{H^1(\Omega)}^2 &+ \|u_m\|_{L^4(\Omega)}^4 \leq |f| \|u_m\|_{H^1(\Omega)} \end{aligned}$$

$$(13.8) \quad \|u_m\|_{H^1(\Omega)} \leq \frac{1}{\min(1, \lambda)} |f|$$

$$(13.9) \quad \|u_m\|_{L^4(\Omega)}^4 \leq \frac{1}{\min(1, \lambda)} |f|^2.$$

Moreover, from (13.9) follows also:

$$(13.10) \quad \|u_m^3\|_{L^{4/3}(\Omega)} \leq \left\{ \frac{1}{\min(1, \lambda)} |f|^2 \right\}^{3/4}.$$

Since balls in  $H^1(\Omega)$ ,  $L^4(\Omega)$ ,  $L^{4/3}(\Omega)$  are weakly compact, we see (by successively extracting three subsequences) that there exists a subsequence  $m' \rightarrow \infty$  such that:

$$(13.11) \quad u_{m'} \rightarrow u \quad \text{in the weak topology of } H^1(\Omega)$$

$$(13.12) \quad u_{m'} \rightarrow u_* \quad \text{in the weak topology of } L^4(\Omega)$$

$$(13.13) \quad u_{m'}^3 \rightarrow \varphi \quad \text{in the weak topology of } L^{4/3}(\Omega).$$

It follows immediately that  $u = u_*$  because (13.11) and (13.12) imply that  $u_{m'}$  converges to  $u$  and to  $u_*$  respectively, in the weak topology of  $L^2(\Omega)$  or in the sense of distributions on  $\Omega$ .

*However, it is not at all evident that  $\varphi = u^3$  and this difficulty is the essential difference with the linear case of Chapter 2. To circumvent this difficulty we are going to use a method which can also be used for a large class of nonlinear equations.*

Let  $w \in \bigcup_{m \geq 2} V_m$  and let  $j$  be the smallest integer such

that  $w \in V_j$ ; for all  $m' > j$  we have according to (13.6):

$$b(u_{m'}, w - u_{m'}) = (f, w - u_{m'}),$$

and with (13.3) we get:

$$b(w, w - u_{m'}) \geq b(u_{m'}, w - u_{m'}) = (f, w - u_{m'}).$$

By (13.11) and (2.8):

$$\lim_{m' \rightarrow \infty} a(w, u_{m'}) = a(w, u);$$

according to (13.12) and the definition of weak convergence in  $L^4(\Omega)$  ( $w^3 \in L^{4/3}(\Omega)$ ) we have:

$$\lim_{m' \rightarrow \infty} \int_{\Omega} w^3(x) u_{m'}(x) dx = \int_{\Omega} w^3(x) u(x) dx.$$

Hence we get in the limit:

$$(13.14) \quad b(w, w - u) \geq (f, w - u).$$

According to Lemma 2 each of the members of (13.14) depends continuously on  $w$ , in the topology of  $V$ . Since  $\bigcup_{m \geq 1} V_m$  is dense in  $V$ , and (13.14) holds for all  $w \in \bigcup_{m \geq 1} V_m$ , the inequality still holds by continuity for all  $w \in V$ .

Now, for a fixed arbitrary element of  $V$ , we write (13.14) with:

$$w = u + \lambda v, \quad \lambda \in \mathbf{R};$$

and hence:

$$\lambda b(u + \lambda v, v) \geq \lambda(f, v).$$

If  $\lambda > 0$ , we find by division by  $\lambda$ :

$$b(u + \lambda v, v) \geq (f, v),$$

and when  $\lambda \rightarrow 0$  (cf. Lemma 2):

$$b(u, v) \geq (f, v).$$

By considering the case  $\lambda < 0$ , we get the opposite inequality, and by comparing the two inequalities, we find:

$$b(u, v) = (f, v).$$

This proves that  $u$  satisfies (13.1) and this completes the proof of Theorem 1.

### *Result of the approximation*

Let us work out the preceding, to obtain an approximation result concerning  $u$ .

We can prove exactly as above, that every sequence extracted from  $u_m$  contains a subsequence that converges (in the weak topology of  $H^1(\Omega)$  and in the weak topology of  $L^4(\Omega)$ ) to a solution of (13.1), which means to  $u$ . Under these conditions the sequence  $u_m$  as a whole converges (in the weak topology of  $H^1(\Omega)$  and of  $L^4(\Omega)$ ) to  $u$ .

Next we consider the expression:

$$X_m = b(u_m, u_m - u) - b(u, u_m - u).$$

With an argument similar to the one used in Chapter 2, we show that:

$$\lim_{m \rightarrow \infty} X_m = 0.$$

According to (13.3):

$$\min(1, \lambda) \|u_m - u\|_{H_0^1(\Omega)}^2 \leq a(u_m - u, u_m - u) \leq X_m,$$

and hence  $u_m$  converges to  $u$  in the strong topology of  $H_0^1(\Omega)$ .

We note on the other hand, that the approximate Equation (13.6) has at most one solution. This is proved in the same way as in the case of Equation (13.1).

The preceding remarks lead us to the following theorem:

**THEOREM 2.** *The approximate equation (13.6) has a unique solution  $u_m \in V_m$ .*

*For  $m \rightarrow \infty$ ,  $u_m$  converges in the strong topology of  $H_0^1(\Omega)$  to a solution of (13.1).*

*Remark.* It can be shown that  $u_m$  converges to  $u$  in the strong topology of  $L^4(\Omega)$  and hence in the sense of the norm of  $V$ .

*Remark.* Though Equation (13.6) has a unique solution, the actual computation of the solution (13.6), which is equivalent to solving a system of algebraic equations, is not easy. However, there exist algorithms for solving systems of nonlinear algebraic equations.

## CHAPTER 14

### APPROXIMATE PROBLEMS

We study the approximate solution of problem (13.1) by a finite difference method.

#### 14.1. External Approximation of the Space $V$

The space  $F$  and the operator  $\bar{\omega}$  are those as defined in Lemma 1 of Chapter 13:

$$F = L^4(\Omega) \times L^2(\Omega)^{n+1}$$
$$\bar{\omega}: u \mapsto \bar{\omega}u = (u, u, D_1u, \dots, D_nu).$$

We use the notations introduced in Section 7.1. For all  $h \in \mathcal{H}$ ,  $V_h$  is the space of step functions:

$$u_h(x) = \sum_{M \in \Omega_h^1} u_h(M) w_{hM}(x), \quad u_h(M) \in \mathbf{R}$$

provided with the norm:

$$\|u_h\|_h = \left\{ \int_{\Omega} \left[ \sum_{i=1}^n (\delta_i u_h(x))^2 + (u_h(x))^2 \right] dx \right\}^{1/2} +$$
$$+ \left\{ \int_{\Omega} (u_h(x))^4 dx \right\}^{1/4}.$$

$V_h$  is finite dimensional and the dimension is equal to the number  $n(h)$  of points of  $\hat{\Omega}_h^1$  (as the set of functions  $w_{hM}$ ,  $M \in \hat{\Omega}_h^1$  constitutes a basis of  $V_h^1$ ).

The operator  $p_h$  is defined as follows:

$$p_h u_h = (u_h|_{\Omega}, u_h|_{\Omega}, \delta_1 u_h|_{\Omega}, \dots, \delta_n u_h|_{\Omega}), \quad \forall u_h \in V_h,$$

where  $g|_{\Omega}$  stands for the restriction to  $\Omega$  of the function  $g$ .

The restriction operators are defined as in (9.5)

$$(14.1) \quad (r_h u)(M) = u(M), \quad \forall M \in \hat{\Omega}_h^1, \quad \forall u \in \mathcal{V} = \mathcal{D}(\Omega).$$

**PROPOSITION 1.**  $\{\bar{\omega}, F\}$ ,  $\{V_h, p_h, r_h\}_{h \in \mathcal{H}}$  define a stable and convergent generalized external approximation of the space  $V$ .

*Proof.* The operators  $p_h$  are linear, injective, continuous and have norm less than or equal to 1; hence they are stable, and the approximation too.

We now show that this approximation is convergent. According to Proposition 4 of Chapter 3, we have to show that for each  $u \in \mathcal{D}(\Omega)$ :

$$\lim_{h \rightarrow 0} r_h u|_{\Omega} = u \quad \text{in the strong topology of} \\ L^2(\Omega) \text{ and } L^4(\Omega)$$

$$\lim_{h \rightarrow 0} \delta_{ih} r_h u|_{\Omega} = D_i u \quad \text{in the strong topology of} \\ L^2(\Omega) \quad (i = 1, \dots, n).$$

( $\mathcal{D}(\Omega)$  is dense in  $V$ ). Actually, these results are already contained in the proof of Proposition 3 of Chapter 9.

Condition (ii) of the definition of a convergent external

approximation (cf. 3.4) is easily checked if one takes Proposition 3 of Chapter 9 into account. If  $u_h$  is a sequence of elements of  $V_h$ , ( $h' \rightarrow 0$ ) such that  $p_h \cdot u_h$  converges to:

$$\varphi = (\varphi_0, \varphi_1, \dots, \varphi_{n+1})$$

in the weak topology of  $F$ , then:

$$\lim_{h' \rightarrow 0} u_{h'}|_{\Omega} = \varphi_0 \quad \text{in the weak topology of } L^4(\Omega)$$

$$\lim_{h' \rightarrow 0} u_{h'}|_{\Omega} = \varphi_1 \quad \text{in the weak topology of } L^2(\Omega)$$

$$\lim_{h' \rightarrow 0} \delta_i u_{h'}|_{\Omega} = \varphi_{i+1} \quad \text{in the weak topology of } L^2(\Omega) \\ (i = 1, \dots, n).$$

It follows that  $u_h|_{\Omega}$  converges in the weak topology of  $L^2(\Omega)$  (or in the sense of distributions) to  $\varphi_0$  and to  $\varphi_1$ ; hence:

$$\varphi_0 = \varphi_1 \in L^4(\Omega).$$

By the already mentioned proposition, we see that  $\varphi_1 \in H_0^1(\Omega)$  and  $\varphi_{i+1} = D_i \varphi_1$  ( $i = 1, \dots, n$ ) so that  $\varphi_0 \in V$  and  $\varphi = \bar{\omega} \varphi_0$ .

## 14.2. Approximate Equations

For  $u_h, v_h \in V_h$ , we put:

$$(14.5) \quad (u_h, v_h)_h = \int_{\Omega} u_h(x) v_h(x) dx \\ |u_h|_h = (u_h, u_h)_h^{1/2}$$

$$a_h(u_h, v_h) = \sum_{i=1}^n \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx + \lambda (u_h, v_h)_h$$

$$b_h(u_h, v_h) = a_h(u_h, v_h) + \int_{\Omega} (u_h(x))^3 v_h(x) dx.$$

We observe, as in (13.3) that:

$$(14.6) \quad b_h(u_h, u_h - v_h) - b_h(v_h, u_h - v_h) \geq \\ \geq a_h(u_h - v_h, u_h - v_h) \geq 0.$$

With the function  $f$  we associate  $f_h \in V_h$ ,  $f_h = r_h f$ .

The approximate problem can now be formulated as follows:

*To find  $u_h \in V_h$  such that:*

$$(14.7) \quad b_h(u_h, v_h) = (f_h, v_h)_h, \quad \forall v_h \in V_h.$$

**PROPOSITION 2.** *Problem (14.7) has a unique solution.*

*Proof.* Uniqueness is shown as in (13.1); if  $u_{1h}$  and  $u_{2h}$  are two solutions of (14.7) and if  $u_h = u_{1h} - u_{2h}$ , then:

$$b_h(u_{1h}, v_h) - b_h(u_{2h}, v_h) = 0, \quad \forall v_h \in V_h;$$

in particular:

$$b_h(u_{1h}, u_h) - b_h(u_{2h}, u_h) = 0,$$

and by (14.6):

$$a_h(u_h, u_h) \leq 0$$

$$\lambda |u_h|_h^2 = 0,$$

which proves that  $u_h = 0$ ,  $u_{1h} = u_{2h}$ .

The existence of the solution follows, as in the case of (13.6), from Lemma 3 of Chapter 13, applied to the following situation:  $X = V_h$  with the scalar product of (14.5). The map  $P$  is defined as follows:

$$P(u_h) = \sum_{M \in \Omega_h^1} \{b_h(u_h, w_{hM}) - (f_h, w_{hM})_h\} w_{hM}.$$

The map  $b_h$  is continuous from  $V_h \times V_h$  into  $\mathbf{R}$ , which is checked as in Lemma 2 of Chapter 13. Hence  $P$  is continuous on  $V_h$ . On the other hand (the set of the  $w_{hM}$  forms an orthogonal basis of  $V_h$  with respect to the scalar product (14.5)):

$$\begin{aligned} (P(u_h), u_h)_h &= (h_1 \dots h_n) \times \\ &\quad \times \sum_{M \in \Omega_h^1} \{b_h(u_h, w_{hM}) - (f_h, w_{hM})_h\} u_h(M) \\ (h_1 \dots h_n)^{-1} (P(u_h), u_h)_h &= b_h(u_h, u_h) - (f_h, u_h)_h \geq \\ &\geq \lambda |u_h|_h^2 - |f_h|_h |u_h|_h, \end{aligned}$$

and this is strictly positive for  $|u_h|_h = k$  and  $k$  large enough ( $k > (1/\lambda) |f_h|_h$ ).

So the hypotheses of Lemma 3 of Chapter 13 are satisfied, and hence there exists a  $u_h \in V_h$  such that  $P(u_h) = 0$ , or:

$$b_h(u_h, w_{hM}) = (f_h, w_{hM})_h, \quad \forall M \in \Omega_h^1;$$

so  $u_h$  satisfies (14.7).

*Remark.* Equation (14.7) is equivalent to an algebraic system of equations for the set of the  $u_{hM}$ ; we will see in Section 14.3 how the method of fractionary steps can be

suitably adapted to get a practical method for the actual computation of the solution of this equation.

The convergence theorem is as follows:

**THEOREM 1.** *For  $h \rightarrow 0$ , the solution  $u_h$  of (14.7) converges weakly to the solution  $u$  of (12.1) and moreover:*

$$(14.8) \quad \lim_{h \rightarrow 0} u_h|_{\Omega} = u \quad \text{in the strong topology of } L^2(\Omega)$$

$$(14.9) \quad \lim_{h \rightarrow 0} \delta_i u_h|_{\Omega} = D_i u \quad \text{in the strong topology of } L^2(\Omega) \\ (i = 1, \dots, n).$$

*Proof.* When we put  $v_h = u_h$  in (14.7) we get:

$$b_h(u_h, u_h) = (f_h, u_h)_h$$

$$b_h(u_h, u_h) \leq |f_h|_h |u_h|_h.$$

Because of (9.3),  $|f_h| \leq |f|$  ( $f_h = r_h f$ ), and hence

$$\sum_{i=1}^n \int_{\Omega} (\delta_i u_h(x))^2 dx + \lambda |u_h|_h^2 + \\ + \int_{\Omega} (u_h(x))^4 dx \leq |f| |u_h|_h$$

$$(14.10) \quad |u_h|_h \leq \frac{1}{\lambda} |f|$$

$$(14.11) \quad \int_{\Omega} (\delta_i u_h(x))^2 dx \leq \frac{1}{\lambda} |f|^2 \quad (i = 1, \dots, n);$$

$$(14.12) \quad \int_{\Omega} (u_h(x))^4 dx \leq \frac{1}{\lambda} |f|^2.$$

Under these conditions, there exists a sequence  $h' \rightarrow 0$  such that:

$$(14.13) \quad \lim_{h' \rightarrow 0} u_{h'}|_{\Omega} = u_*$$

$$(14.14) \quad \lim_{h' \rightarrow 0} u_{h'}|_{\Omega} = u_{**}$$

$$(14.15) \quad \lim_{h' \rightarrow 0} \delta_i u_{h'}|_{\Omega} = u_i \quad (i = 1, \dots, n).$$

This means that  $p_h u_{h'}$  converges in the weak topology of  $F$  to  $\varphi = (u_*, u_{**}, u_1, \dots, u_n)$  and according to Proposition 1,  $u_* \in V$  and  $\varphi = \bar{\omega} u_*$ , that is to say:

$$(14.16) \quad u_{**} = u_*, \quad u_i = D_i u_* \quad (i = 1, \dots, n).$$

Let  $w$  be an arbitrary element of  $V$  and  $w_h = r_h w$ . If we write (14.7) with  $v_h = w_h - u_h$ , and use (14.6), we see that:

$$(14.17) \quad \begin{aligned} b_h(u_h, w_h - u_h) &= (f_h, w_h - u_h)_h \\ b_h(w_h, w_h - u_h) &\geq (f_h, w_h - u_h)_h. \end{aligned}$$

It is easy, by Proposition 1 and (14.13) to (14.16), to pass to the limit in this inequality, by letting  $h' \rightarrow 0$ . Since  $p_h w_h$  converges to  $\bar{\omega} w$  in the strong topology of  $F$ :

$$\lim_{h \rightarrow 0} w_h|_{\Omega} = w \quad \text{in the strong topologies of} \\ L^2(\Omega) \text{ and } L^4(\Omega)$$

$$\lim_{h \rightarrow 0} \delta_i w_h|_{\Omega} = D_i w \quad \text{in the strong topology of } L^2(\Omega) \quad (i = 1, \dots, n).$$

Hence:

$$\begin{aligned} \lim_{h' \rightarrow 0} a_{h'}(w_{h'}, w_{h'} - u_{h'}) &= a(w, w - u_*) \\ \lim_{h' \rightarrow 0} \int_{\Omega} (w_{h'}(x))^3 (w_{h'}(x) - u_{h'}(x)) \, dx &= \\ &= \int_{\Omega} (w(x))^3 (w(x) - u_*(x)) \, dx \end{aligned}$$

$$\lim_{h' \rightarrow 0} (f_{h'}, w_{h'} - u_{h'})_{h'} = (f, w - u_*)$$

$(f_h|_{\Omega})$  converges to  $f$  in the strong topology of  $L^2(\Omega)$ .  
Finally:

$$b(w, w - u_*) \geq (f, w - u_*), \quad \forall w \in V.$$

This last inequality shows, just like in the case of (13.14) that  $u_*$  is the solution of (13.1) and hence  $u_* = u$ .

From every sequence extracted from  $u_h$  we can extract a subsequence for which (14.13) to (14.16) hold, with  $u_* = u$ : This proves that convergence holds for the sequence  $u_h$  as a whole;  $p_h u_h$  converges to  $\bar{w}u$  in  $F$ , or, in the terminology of Chapter 3,  $u_h$  converges weakly to  $u$ .

To establish (14.8) and (14.9), we consider the expression:

$$X_h = b_h(u_h, u_h - r_h u) - b_h(r_h u, u_h - r_h u).$$

According to (14.7):

$$b_h(u_h, u_h - r_h u) = (f_h, u_h - r_h u)_h,$$

and

$$\lim_{h \rightarrow 0} (f_h, u_h - r_h u)_h = 0.$$

By Proposition 1, it is easily verified that:

$$\lim_{h \rightarrow 0} b_h(r_h u, u_h - r_h u) = 0.$$

Hence  $\lim_{h \rightarrow 0} X_h = 0$ , and by (14.6):

$$\lim_{h \rightarrow 0} a_h(u_h - r_h u, u_h - r_h u) = 0;$$

which means:

$$\lim_{h \rightarrow 0} \int_{\Omega} (u_h(x) - r_h u(x))^2 dx = 0$$

$$\lim_{h \rightarrow 0} \int_{\Omega} (\delta_i u_h(x) - \delta_i r_h u(x))^2 dx = 0 \quad (i = 1, \dots, n)$$

and because of Proposition 1:

$$\lim_{h \rightarrow 0} \int_{\Omega} (r_h u(x) - u(x))^2 dx = 0$$

$$\lim_{h \rightarrow 0} \int_{\Omega} (\delta_i r_h u(x) - D_i u(x))^2 dx = 0 \quad (i = 1, \dots, n).$$

whereby (14.8) and (14.9) are established. Theorem 1 is proved.

*Remark.* One may prove that  $u_h|_{\Omega}$  converges to  $u$  in the strong topology of  $L^4(\Omega)$ , and then conclude that  $u_h$  converges strongly and discretely to  $u$ .

### 14.3. The Method of Fractionary Steps

We are going to adapt the method of fractionary steps to the solution of (14.7) for fixed  $h$ .

We introduce the forms:

$$a_{ih}(u_h, v_h) = \int_{\Omega} \delta_i u_h(x) \delta_i v_h(x) dx + \frac{\lambda}{n} (u_h, v_h)_h$$

$$\forall u_h, v_h \in V_h, \quad i = 1, \dots, n.$$

We associate with these continuous linear operators  $A_{ih}$  in  $V_h$ , defined by:

$$(A_{ih}u_h, v_h)_h = a_{ih}(u_h, v_h), \quad \forall u_h, v_h \in V_h.$$

Let  $f_h = \sum_{i=1}^n f_{ih}$  be an arbitrary decomposition of  $f_h$  into elements of  $V_h$ .

For each integer  $N$  and for each real positive number  $\tau$ , we define recursively a family of elements of  $V_h$  (cf. Chapter 6). These we denote by:

$$u_h^{r+i/m}, \quad 0 \leq r \leq N-1, \quad 1 \leq i \leq m \quad (m = n+1).$$

We start from an element  $u_h^0$ , which is arbitrarily chosen in  $V_h$ . When  $u_h^0, \dots, u_h^{r+(i-1)/m}$  are known,  $u_h^{r+i/m}$  is defined as follows: *For  $i = 1$ ,  $u_h^{r+1/m}$  is the solution in  $V_h$  of:*

$$(14.18) \quad (u_h^{r+1/m} - u_h^r, v_h)_h + \int_{\Omega} (u_h^{r+1/m}(x))^3 v_h(x) dx = 0, \quad \forall v_h \in V_h.$$

The uniqueness of the solution of (14.18) is immediately clear. We can conclude existence from Lemma 3 of Chapter 13, but it is simpler to remark that (14.18) is equivalent to the following uncoupled algebraic system:

$$(14.19) \quad \xi_M + \tau \xi_M^3 = \eta_M, \quad \forall M \in \Omega_h^1,$$

where  $\eta_M = u_h^r(M)$  is known, and where  $\xi_M = u_h^{r+1/m}$  has to be found.

*For  $2 \leq i \leq m$ ,  $u_h^{r+i/m}$  is defined as the solution in  $V_h$  of:*

$$(14.20) \quad (u_h^{r+(i+1)/m} - u_h^{r+(i-1)/m}, v_h)_h + \tau a_{ih}(u_h^{r+i/m}, v_h) = \tau (f_{ih}, v_h)_h, \quad \forall v_h \in V_h.$$

Existence and uniqueness of solutions of (14.20) follow from the projection theorem.

We observe that Equations (14.20) are linear and furthermore identical to Equations (11.16). Equations (14.19) are totally uncoupled algebraic equations of the third degree.

We can estimate the difference between the last element

$u_h^N$  and the solution  $u_h$  of (14.7) by the following proposition:

**PROPOSITION 3.** *For all  $n, n=0, \dots, N$ :*

$$\left| u_h^n - u_h \right|_h^2 \leq \varepsilon_n$$

where:

$$\varepsilon_n = \varepsilon_0 (1 + \gamma\tau) + \delta\tau^2 \sum_{j=1}^n (1 + \gamma\tau)^{j-n}$$

$$\varepsilon_0 = \left| u_h^0 - u_h \right|_h^2$$

$$\delta = n \sum_{i=1}^n \left| f_{ih} - A_{ih} u_h \right|_h^2$$

$$\gamma = \min\left(1, \frac{\lambda}{n}\right).$$

**PROPOSITION 4.** *For  $\tau \rightarrow 0$  and  $N \rightarrow \infty$  such that  $\tau N \rightarrow \infty$ , we have:*

$$\lim |u_h^N - u_h|_h^2 = 0.$$

The proofs of these propositions are quite similar to those of Theorems 1 and 2 of Chapter 6. We will not discuss them further.

### *A numerical application*

We illustrate the preceding method by the following example:

$$\begin{aligned}
 -\Delta u + \lambda u + u^3 &= f \text{ in } \Omega = ]0, 1[ \times ]0, 1[ \quad (\Omega \subset \mathbb{R}^2) \\
 u|_r &= 0
 \end{aligned}$$

with:

$$\begin{aligned}
 f &= -200x_1(x_1 - 1) - 200x_2(x_2 - 1) + \\
 &+ 100\lambda x_1(x_1 - 1)x_2(x_2 - 1) + \\
 &+ 10^6 x_1^3(x_1 - 1)^3 x_2^3(x_2 - 1)^3.
 \end{aligned}$$

The exact solution is:

$$u = 100x_1(x_1 - 1)x_2(x_2 - 1).$$

| $\tau$ | $h = 1/10$             |     |                        |     | $h = 1/20$             |     |                        |     |
|--------|------------------------|-----|------------------------|-----|------------------------|-----|------------------------|-----|
|        | $\lambda = 1$          |     | $\lambda = 100$        |     | $\lambda = 1$          |     | $\lambda = 100$        |     |
|        | $e_h$                  | $N$ | $e_h$                  | $N$ | $e_h$                  | $N$ | $e_h$                  | $N$ |
| 1/100  | $0.199 \times 10^{-2}$ | 20  | $0.662 \times 10^{-1}$ | 13  | $0.220 \times 10^{-2}$ | 19  | $0.664 \times 10^{-1}$ | 13  |
| 1/200  | $0.537 \times 10^{-3}$ | 31  | $0.201 \times 10^{-1}$ | 20  | $0.591 \times 10^{-3}$ | 31  | $0.202 \times 10^{-1}$ | 20  |
| 1/300  | $0.245 \times 10^{-3}$ | 42  | $0.961 \times 10^{-2}$ | 27  | $0.270 \times 10^{-3}$ | 42  | $0.963 \times 10^{-2}$ | 27  |
| 1/400  | $0.140 \times 10^{-3}$ | 51  | $0.560 \times 10^{-2}$ | 33  | $0.154 \times 10^{-3}$ | 51  | $0.561 \times 10^{-2}$ | 33  |
| 1/500  | $0.907 \times 10^{-4}$ | 60  | $0.366 \times 10^{-2}$ | 39  | $0.997 \times 10^{-4}$ | 60  | $0.367 \times 10^{-2}$ | 39  |

These results are quite sensitive to changes in the values of  $\lambda$ . The computation time is approximately proportional to the number of iterations and the number of cells used.

## APPENDIX

### TOPOLOGICAL VECTOR SPACES

We recall without proof a few results concerning topological vector spaces. We only consider spaces over  $\mathbf{K}$ ,  $\mathbf{K} = \mathbf{R}$  or  $\mathbf{C}$ .

#### *Topological vector spaces*

Let  $E$  be a set endowed with a vector space structure over  $\mathbf{K}$  together with a topology. We say that these two structures are compatible if:

- (i) the map  $\{x, y\} \mapsto x + y$  from  $E \times E$  into  $E$  is continuous;
- (ii) the map  $\{\lambda, x\} \mapsto \lambda x$  from  $\mathbf{K} \times E$  into  $E$  is continuous.

The set  $E$  endowed with these two compatible structures is called a topological vector space, real if  $\mathbf{K} = \mathbf{R}$ , complex if  $\mathbf{K} = \mathbf{C}$ .

A normed space, and in particular a prehilbert space, endowed with its vector space structure, and with the topology associated with the norm, is a simple example of a topological vector space.

#### *Seminorm*

Let  $E$  be a vector space over  $\mathbf{K}$ . A seminorm on  $E$  is a map

$p$  from  $E$  into  $\mathbf{R}$  such that:

- (i)  $p(x) \geq 0, \quad \forall x \in E;$
- (ii)  $p(\lambda x) = |\lambda|p(x), \quad \forall x \in E, \quad \forall \lambda \in \mathbf{K}$
- (iii)  $p(x+y) \leq p(x) + p(y), \quad \forall x, y \in E.$

We say that  $p$  is a norm if moreover  $p(x) \neq 0$  for all  $x \neq 0$ . We call the set  $\{x \mid p(x-a) < \rho\}$  an open  $p$ -ball with center  $a$  and radius  $\rho$ .

*$\mathcal{P}$ -topology on a vector space*

Let  $\mathcal{P}$  be a family of seminorms on  $E$ . We call any finite intersection of open  $p$ -balls with center  $a$ ,  $p \in \mathcal{P}$ , an open  $\mathcal{P}$ -ball.

The  $\mathcal{P}$ -topology on  $E$  will be the topology whose open sets consist of arbitrary unions of open  $\mathcal{P}$ -balls. Such a topology can be shown to be compatible with the vector space structure of  $E$ .

Among the properties of  $\mathcal{P}$ -topologies we mention the following ones:

- (i) the  $\mathcal{P}$ -topology is Hausdorff if and only if:

$$\forall x \in E, \quad x \neq 0, \quad \exists p \in \mathcal{P}, \quad p(x) > 0;$$

- (ii) a sequence  $x_n$  converges to  $x$  in the  $\mathcal{P}$ -topology if and only if:

$$\forall p \in \mathcal{P}, \quad \lim_{n \rightarrow \infty} p(x_n - x) = 0.$$

*The weak topology of a Hilbert space*

Let  $E$  be a Hilbert space with scalar product  $((., .))$  and norm  $\|.\|$ .

For each  $y$  of  $E$ , we define the seminorm  $p_y$ :

$$p_y(x) = |((x, y))|, \quad \forall x \in E.$$

Let  $\mathcal{P}$  be the family of seminorms  $\{p_y\}_{y \in E}$ . The corresponding  $\mathcal{P}$ -topology is called the *weak* topology of the Hilbert space  $E$ . By contrast, the topology associated with the norm is called the strong topology of  $E$ .

The weak topology of  $E$  is Hausdorff (for  $x \neq 0$ ,  $p_x(x) = \|x\|^2 > 0$ ). A sequence  $x_n$  converges to  $x$  in this topology if and only if:

$$\forall y \in E, \quad \lim_{n \rightarrow \infty} ((x_n, y)) = ((x, y)).$$

We observe the following: A sequence  $x_n$  converges strongly to  $x$  if and only if  $x_n$  converges weakly to  $x$  and  $\|x_n\|$  converges to  $\|x\|$ .

These conditions are obviously necessary, in order that  $x_n$  converges strongly to  $x$ . Conversely, let us assume that  $x_n$  converges weakly to  $x$  and  $\|x_n\|$  converges to  $\|x\|$ ; then:

$$\|x_n - x\|^2 = \|x_n\|^2 + \|x\|^2 - 2\operatorname{Re}((x_n, x))$$

converges to 0, which proves the statement.

*Weak topologies of a normed space and its dual*

Let  $E$  be a normed space and  $E'$  its dual (the space of

linear continuous maps from  $E$  into  $K$ ), equipped with the dual norm.

The weak topology of  $E$ , denoted by  $\sigma(E, E')$ , is the topology associated with the following family of seminorms  $\{p_y\}_{y \in E'}$ :

$$p_y(x) = |\langle x, y \rangle|, \quad \forall x \in E,$$

where  $\langle \cdot, \cdot \rangle$  stands for the scalar product of the duality between  $E$  and  $E'$ .

The weak star topology of  $E'$ , denoted by  $\sigma(E', E)$ , is the topology associated with the family of seminorms  $\{p_x\}_{x \in E}$ :

$$p_x(y) = |\langle x, y \rangle|, \quad \forall y \in E'.$$

By contrast, we call the topology of  $E$  (or  $E'$ ) defined by the norm the strong topology of  $E$  (or  $E'$ ).

It is easily verified that the weak star topology of  $E'$  is Hausdorff. It is also true, but less obvious, that the weak topology of  $E$  is Hausdorff.

If  $E$  is a Hilbert space, then the two weak topologies which are defined on  $E$  are actually the same, as the defining families of seminorms are the same, by the Riesz theorem, which states that there exists a canonical isomorphism of  $E$  onto its dual.

### *Compactness theorem*

An important result, which we often used in the text, and

which we will now state, without proof, is the following (cf. Dunford and Schwartz [1]):

*Let  $E$  be a normed space and  $E'$  its dual. The closed balls of  $E'$  are compact and sequentially compact for the weak star topology  $\sigma(E', E)$ .*

In particular, if  $y_n$  is a bounded sequence of  $E'$ , then there exists a subsequence  $y_{n_i}$  and an element  $y$  of  $E'$  such that  $y_{n_i}$  converges to  $y$  in the weak star topology of  $E'$ , in other words:

$$\forall x \in E, \quad \lim_{i \rightarrow \infty} \langle x, y_{n_i} \rangle = \langle x, y \rangle.$$

Likewise, for a bounded sequence  $x_n$  in a Hilbert space  $E$ , there exists a subsequence  $x_{n_i}$  and an element  $x$  of  $E$  such that  $x_{n_i}$  converges to  $x$  in the weak topology of  $E$ , in other words:

$$\forall y \in E, \quad \lim_{i \rightarrow \infty} ((x_{n_i}, y)) = ((x, y)).$$

## BIBLIOGRAPHY

A few books and papers only are mentioned here. The bibliography must be completed by the references cited in the following works.

### *Analysis*

- [1] N. Dunford, J. T. Schwartz, *Linear Operators*, Vol. 1, Pure and Applied Mathematics, Academic Press, New York, 1958.

### *Topological vector spaces, distributions*

- [2] N. Bourbaki, *Espaces vectoriels topologiques*, Actualités Scientifiques et Industrielles, Hermann, Paris, 1964.
- [3] L. Schwartz, *Théorie des distributions*, Actualités Scientifiques et Industrielles, Hermann, Paris, 1957.
- [4] F. Trèves, *Topological vector spaces, Distributions and Kernels*, Pure and Applied Mathematics, Academic Press, New York, 1967.

### *Partial differential equations*

- [5] L. Hörmander, *Linear Partial Differential Operators*, Springer Verlag, Heidelberg, 1969.
- [6] J. L. Lions, *Equations différentielles et problèmes aux limites*, Springer Verlag, Heidelberg, 1969.
- [7] J. L. Lions, *Quelques méthodes de résolution des problèmes aux limites non linéaires*, Dunod-Gauthier-Villars, Paris, 1969; English translation in press.
- [8] J. L. Lions and E. Magenes, *Non-homogeneous Boundary-Value Problems and Applications* (Vols 1 and 2), Springer Verlag, Heidelberg, 1972.
- [9] F. Trèves, *Linear Partial Differential Operators with Constant Coefficients*, Gordon and Breach, New York, 1966.

*Approximate solutions of partial differential equations*

- [10] J.-P. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Wiley-Interscience, New York, 1972.
- [11] P. Ciarlet and P. A. Raviart, *Finite Element Methods*, Lecture Notes in Mathematics, Springer Verlag, Heidelberg, 1973 (in press).
- [12] G. I. Marchuk, *Mathematical Methods of Computing Sciences*, Novosibirsk, 1971; English translation in press.
- [13] K. W. Morton and R. D. Richtmyer, *Difference Methods for Initial Value Problems*, Interscience, New York, 1967.
- [14] G. Strang and G. Fix, *An Analysis of the Finite Element Method*, 1973 (in press).
- [15] R. Temam, 'Sur la stabilité et la convergence de la méthode des pas fractionnaires' *Ann. Mat. pura appl.* **79** (1968) 191–380.
- [16] R. Temam, *Approximation of Navier–Stokes Equations*, Lecture Notes, University of Maryland, 1973.
- [17] R. S. Varga, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, New Jersey, 1962.
- [18] N. N. Yanenko, *The Method of Fractional Steps*, Springer Verlag, Heidelberg, 1971.

*Proceedings*

- [19] J. Leray (ed.), *Proceedings of the International Conference of Mathematicians*, held at Nice, 1970, Gauthier-Villars, Paris, 1971.
- [20] J. H. Bramble (ed.), *Numerical Solution of Partial Differential Equations I*, Symposium held at the University of Maryland, 1965, Academic Press, New York, 1966.
- [21] B. Hubbard (ed.), *Numerical Solution of Partial Differential Equations II*, Symposium held at the University of Maryland, 1970, Academic Press, New York, 1971.
- [22] J. Babushka (ed.), *Numerical Solution of Partial Differential Equations III*, Symposium held at the University of Maryland, 1972, Academic Press, New York, 1973 (in press).
- [23] J. L. Morris (ed.), *Conference on Applications of Numerical Analysis*, Dundee, Scotland, 1971, Lecture Notes in Mathematics, Springer Verlag, 1971.

- [24] J. L. Morris (ed.), *Symposium on the Theory of Numerical Analysis*, Dundee, Scotland, 1970, Lecture Notes in Mathematics, Springer Verlag, 1971.
- [25] J. L. Morris (ed.), *Symposium on the Theory of Numerical Analysis*, Dundee, Scotland, 1972, Lecture Notes in Mathematics, Springer Verlag (in press).
- [26] E. Magenes (ed.), *Convegno su problema analisi numerica*, Rome, 1972, Zanichelli-Academic Press.

## INDEX

- admissible triangulation, 87
- approximate equations, 31
  - consistency hypotheses for –, 32
- approximation of
  - $L^2(\mathbb{R}^n)$ , 61
  - $H^1(\mathbb{R}^n)$ , 63
  - $L^2(\Omega)$ , 70
  - $H^1(\Omega)$ , 73
  - $H_0^1(\Omega)$ , 76
  - $H_0^1(\Omega) \cap L^4(\Omega)$ , 143  
(=  $V$  of Part III)
- approximation,
  - convergent, 21
  - external, 24
  - Galerkin, 29
  - generalized, 25
  - Hilbert, 20, 24
  - internal, 20
  - stable, 21
  - stable generalized, 26
- barycenter of a simplex, 83
- barycentric coordinates, 81
- boundaries, smoothness
  - conditions for, 56
- Brouwer fixed point theorem,
  - application of, 136
- cells, 60
- coercive bilinear form, 5
- compactness theorem, 160
- computation of error, 40
- conforming finite element
  - method, 87
- consistency hypotheses for
  - approximate equations, 32
- convergence
  - of an approximation, 21
  - of a family  $u_h$ , 23
  - theorems for approximate equations, 32
  - strong – of Galerkin method, 15
  - weak – in  $L^\infty$  of approximate solutions of the Dirichlet problem, 115
- convergent approximation
  - (external), 25
  - (generalized), 27
  - (internal), 21
- Dirichlet problem, 165
  - $L^\infty$  estimate of solution of –, 112
- discrete convergence of a
  - family  $u_h$ , 23
- discrete error, 21, 26

- dual of a topological vector space, 158
- error
  - computation of, 40
  - discrete, 21, 26
  - estimate of, 46
  - truncation, 21, 26
- estimate in  $L^\infty(\Omega)$  of solution of Dirichlet problem, 112
- estimate of error in case of fractionary step method, 46
- existence and uniqueness of solutions of a nonlinear equation, 131
- existence of solutions of partial differential equations, 14 note
- external approximation, 24
  - convergent – , 25
- finite difference method, 59
- finite difference operator, 60
- finite element method, 80
- fractionary steps method, 43
- function spaces  $LP$ , 53
  - Sobolev spaces
    - $H^m$ , 54
    - $H_0^m$ , 55
    - $H^{1/2}$ , 58
    - $H^{-1/2}$ , 58
- Galerkin approximation of separable Hilbert spaces, 23
- Galerkin method, 12, 37
  - strong convergence of –, 15
- generalized approximation of a space, 25
- Green's formula, 57
- Hilbert approximation, 20
- internal approximation, 20
- Lax-Milgram theorem, 5
- method, finite element, 80
- method of fractionary steps, 43
  - of Galerkin, 12, 37
  - of Ritz, 17
- Neumann problem, 121
- nonconforming finite element method, 93
- norm, 157
- numerical applications for Dirichlet problems, 117
  - Neumann problem, 127
  - nonlinear problem, 154
- operator,
  - finite difference, 60
  - prolongation, 20
  - prolongation – for Sobolev spaces, 56
  - restriction – , 20
  - translation – , 66
  - truncation – , 20
- piecewise linear finite elements,
  - conforming, 87
  - nonconforming, 93

- problem,
  - Dirichlet, 105
  - Neumann, 121
  - nonlinear, 131
- projection theorem, 5
  - special case of  $-$ , 8
  - variational formulation of  $-$ , 9
- prolongation operator, 20
  - for Sobolev spaces, 56
- restriction operator, 20
- restriction schemes, 37, 42
- Riesz theorem, 8
- Ritz method, 17
- seminorm, 156
- simplex, 82
  - barycenter of, 83
- smoothness conditions about
  - boundary, 56
- Sobolev spaces, *see* function spaces
  - prolongation operator for  $-$ , 56
- stable approximation, 21
- stable generalized
  - approximation, 26
- strong convergence of a
  - family, 23
- strong topology, 159
- topological vector space, 113
  - dual of  $-$ , 158
- trace theorem, 57
- translation operators, 66
- triangulation, admissible, 87
- truncation error, 21, 26
  - operator, 20
- variational formulation of
  - projection theorem, 9
- variational problems,
  - approximation of, 31
- weak convergence of a family
  - $u_h$ , 23
- weak convergence in  $L^\infty$  of
  - approximative solutions of the Dirichlet problem, 115
- weak star topology, 159
- weak topology, 158, 159