NUMERICAL ANALYSIS FOR NONLINEAR AND BIFURCATION PROBLEMS

by

Gabriel Caloz
I.R.M.A.R.
Université de Rennes I
35042 Rennes Cedex

and

Jacques Rappaz
D.M.A.
École Polytechnique Fédérale de Lausanne
1015 Lausanne

GC/JR

February 1994

PREFACE

Computational applications generally involve nonlinear problems and often contain parameters. They may represent properties of the physical system they describe or quantities which can be varied. A basic problem in approximation consists in studying existence and convergence of approximated solutions for a given nonlinear problem, for instance when the parameters are fixed. Another problem is to represent the families or manifolds of solutions under variations of some parameters. Apart from a theoretical approach, such representations are computed and continuation methods are concerned with generating the solution manifolds. By varying one parameter, we can follow a path of solutions. Then to study the effects of change of parameters on a system, it is of prime interest to know the effects of numerical approximation on its behavior.

The goal of this article is to present a general framework in which approximations of nonlinear problems and approximations of solution manifolds can be studied. We will consider regular solutions, regular solution families, and singular solutions. Even though we will illustrate the general theory only with elementary finite element approximations of model boundary value problems, it can be applied to a much wider range of problems in connection with approximation methods. Our presentation is a remodelling of the one proposed by Crouzeix and Rappaz [1989] taking its origin in Descloux and Rappaz [1982] and in Brezzi, Rappaz, and Raviart [1980, 1981a, 1981b].

The general problem we will handle and which covers a lot of applications is the following: find $x \in X$ such that

$$F(x) = 0$$

where X and Z are Banach spaces, $F : X \to Z$ is a smooth nonlinear mapping. Of particular interest is the case where the space X has the form $\mathbb{R}^m \times Y$, where $\mathbb{R}^m$ with $m \geq 1$ is the parameter space and the Banach space Y is the state space. We will work under the assumption that the derivative of F is a Fredholm operator of index $n \geq 0$. Both cases $n = 0$ and $n \geq 1$ with a surjective derivative are studied separately. Note that when n is positive, the family of solutions to $F(x) = 0$ is a differentiable manifold. The singular situation with a not surjective derivative is also studied.

In the general setting, the approximation schemes are written in the form

$$F_h(x) = 0$$

where h is a parameter in $(0, 1]$ and $F_h : X \to Z$ is an approximation of F. The family $\{F_h\}_{0<h\leq 1}$ converges to F as h tends to 0. Our work then is to compare the solution set

of the problem $F_h(x) = 0$ to the one of the problem $F(x) = 0$. The error estimates will generally be deduced from the inverse function theorem. Our program is the following.

In Sections 1 and 2, we present the functional analysis material we use to establish approximation results. It covers mainly variants of the inverse and of the implicit function theorems.

The sections 3, 4, and 5 present model examples which are used later to illustrate the general theory. They do not reflect the range of applications of the theory, but have been chosen because they are simple or familiar to the reader. In each case we discuss the exact problem and give some standard approximations of it.

Then we introduce a general framework to study the approximation of a nonlinear problem without parameter (or with fixed parameters). The originality of this approach is a generalization of the well-known inf-sup conditions to nonlinear problems. Furthermore we are naturally led up to discuss some a posteriori estimates in the nonlinear case, which are effective in an adaptive mesh procedure. The general case is studied in Section 6, Galerkin approximations are analyzed in Section 7 while examples are detailed in Sections 8 through 10.

Sections 11 through 17 are devoted to the study of the approximation of parametrized nonlinear problems. We consider exclusively regular solutions, that is points $(\lambda, u) \in \mathbb{R}^m \times Y \equiv X$ where the derivative $DF(\lambda, u)$ is surjective, as it is described in Section 11. First we study the simplest case of regular solution families, the ones parametrized by λ. The approximation in an abstract setting is developed in Section 12, Galerkin approximations are studied in Section 13 and an example is presented in Section 14. In the remainder of the chapter devoted to the approximation of regular solutions, simple limit points and their Galerkin approximations are also discussed. By introducing a new equation we extend the system and transform the problem to apply the results of Section 12.

In Sections 18 through 23 we are concerned with singular solutions and their approximations. Section 18 is devoted to the classical Lyapunov-Schmidt decomposition procedure, which leads to a bifurcation equation. A situation frequently encountered in the applications is presented in Section 19 where the bifurcation equation is studied carefully. Then approximations are analyzed in Section 20 and we get estimates between the bifurcation equation and the approximate bifurcation equations. In Section 21 we compare thoroughly the bifurcation equation and the approximate bifurcation equations when these are defined in $\mathbb{R}^2$. A simple model example illustrates in Section 22 the general theory. In Section 23 we address bibliographical complements.

As mentionned previously the emphasis here is put on the analysis of the convergence and the error estimates of the approximate problems. The methods necessary to understand the approximation of a parametrized nonlinear problem are described in a simple way with a minimum of prerequisite. Our approach is based on a detailed analysis of the error estimates and gives a global insight on the perturbation of a solution set by an approximation procedure. We notice that the solution family of the exact problem or its approximation is under mild hypotheses a differentiable manifold. An approach based on modern differential geometry is proposed in Fink and Rheinboldt [1983a,

1983b] or in RHEINBOLDT [1986] to study regular points. Although of great importance, we do not emphasize on the wide variety of algorithms to numerically solve the problem $F_h(x) = 0$, such as the predictor-corrector continuation method or the piecewise linear method, see KELLER [1977], KUEPPER, MITTELMANN, and WEBER [1984], ALLGOWER and GEORG [1990] for instance and the literature therein.

CHAPTER I

INTRODUCTION

Among the tools commonly used to study parameter dependent problems are the inverse and the implicit function theorems, two basic results of the differential calculus in Banach spaces. It turns out that the approach of nonlinear problems based on contraction mappings is well-suited to study numerical perturbation.

We collect in the next two sections various notations and we recall the inverse and the implicit function theorems in a form convenient for our later use. Our presentation is voluntarily brief. We refer to the literature for further details.

1. Preliminaries

Let X and Z be two real Banach spaces with the norms $\|.\|_X$ and $\|.\|_Z$ respectively. The set $B(x,\delta) = \{y \in X; \|x-y\|_X < \delta\}$ is the open ball in X centered at x with radius δ, while $\overline{B}(x,\delta) = \{y \in X; \|y-x\|_X \leq \delta\}$ denotes the closed ball. The product space $X \times Z$ is equipped with the norm $\|(x,z)\|_{X\times Z} = \|x\|_X + \|z\|_Z$.

We note X' the dual space of the real Banach space X and $\langle .\,,.\rangle_{X'X}$ the duality pairing between X' and X. The space of continuous linear mappings from X into Z denoted by $\mathscr{L}(X;Z)$ and endowed with the norm

$$\|A\|_{X;Z} = \sup_{\substack{x\in X \\ \|x\|_X \leq 1}} \|Ax\|_Z$$

for all $A \in \mathscr{L}(X;Z)$, is a Banach space. The spaces of multilinear mappings are then defined by

$$\mathscr{L}^k(X;Z) = \mathscr{L}\left(X;\mathscr{L}^{k-1}(X;Z)\right), \quad k = 2,\ldots$$

Given an invertible mapping $A \in \mathscr{L}(X;Z)$ and a mapping $B \in \mathscr{L}(X;Z)$, the mapping $A+B \in \mathscr{L}(X;Z)$ is invertible if

$$\|A^{-1}B\|_{X;X} < 1. \tag{1.1}$$

Then the following bound holds

$$\|(A+B)^{-1}\|_{Z;X} \leq \frac{1}{1-\|A^{-1}B\|_{X;X}}\|A^{-1}\|_{Z;X}. \tag{1.2}$$

Let $\{A_n\}_{n\geq 1} \subset \mathscr{L}(X;Z)$ be a sequence of linear operators converging punctually to $A \in \mathscr{L}(X;Z)$, that is for all $x \in X$,

$$\lim_{n\to\infty} A_n x = Ax.$$

Let Y be a Banach space and $B \in \mathscr{L}(Y;X)$ be a compact operator. Then the sequence $\{A_nB\}_{n\geq 1} \subset \mathscr{L}(Y;Z)$ converges to AB in the operator norm, that is

$$\lim_{n\to\infty} \|A_nB - AB\|_{Y;Z} = 0. \tag{1.3}$$

A continuous mapping $G : X \to Z$ is Fréchet differentiable at a point $x \in X$ if there exists an operator $DG(x) \in \mathscr{L}(X;Z)$ such that

$$\|G(y) - G(x) - DG(x)(y-x)\|_Z = o(\|x-y\|_X).$$

The mapping $DG : X \to \mathscr{L}(X;Z)$ is the (Fréchet) derivative of G. The mapping G is said to be of class C^1 when the mapping DG is continuous on X. The mapping G is said to be of class C^2 when DG is of class C^1; the second derivative $D^2G(x)$ at the point x belongs to $\mathscr{L}^2(X;Z)$. We can repeat this procedure in order to define the class C^p with $p \geq 2$. Let $G : X \to Z$ be a C^p mapping with $p \geq 1$. Then the Taylor expansion of G at the point $x \in X$ reads

$$G(y) = G(x) + \sum_{k=1}^{p-1} \frac{1}{k!} D^k G(x)(\underbrace{y-x,\dots,y-x}_{k \text{ times}}) + \tag{1.4}$$

$$+\frac{1}{(p-1)!}\int_0^1 (1-t)^{p-1} D^p G(x+t(y-x))(y-x,\dots y-x)\,dt.$$

The literature in linear and nonlinear functional analysis is rich, for further details we refer for instance to Bonic [1969], Brezis [1983], Cartan [1967], Dieudonné [1968], Dunford and Schwartz [1958].

Let us finally introduce some notations useful in the applications. In the examples presented in the next chapter, the appropriate function spaces are the Sobolev spaces $W^{m,p}(\Omega)$, where Ω is an open set in $\mathbb{R}^2$ and the numbers m, p are non negative integers. These Sobolev spaces are equipped with the norms $\|.\|_{m,p,\Omega}$ and semi norms $|.|_{m,p,\Omega}$. The space $W_0^{m,p}(\Omega)$ is the closure of $\mathscr{D}(\Omega)$ in $W^{m,p}(\Omega)$ and $W^{-m,q}(\Omega)$ is the dual space of $W_0^{m,p}(\Omega)$ where $\frac{1}{p}+\frac{1}{q}=1$. For $p=2$, we shall write $H^m(\Omega)$ and $H_0^m(\Omega)$ instead of $W^{m,p}(\Omega)$ and $W_0^{m,p}(\Omega)$ respectively, $\|.\|_{m,\Omega}$ and $|.|_{m,\Omega}$ instead of $\|.\|_{m,2,\Omega}$ and $|.|_{m,2,\Omega}$. Numerous presentations of the Sobolev spaces can be found, as for example in the books of Adams [1975], Lions and Magenes [1968], Maz'ya [1985] or Nečas [1967].

If v_1, v_2 are two functions on Ω, $v_1, v_2 : (x_1,x_2) \in \Omega \to v_1(x_1,x_2), v_2(x_1,x_2) \in \mathbb{R}$ and $\mathbf{v} = (v_1, v_2)$, then

$$\mathbf{grad} v_1 = \left(\tfrac{\partial v_1}{\partial x_1}, \tfrac{\partial v_1}{\partial x_2}\right), \qquad \mathrm{div}\mathbf{v} = \frac{\partial v_1}{\partial x_1} + \frac{\partial v_2}{\partial x_2}$$

and

$$\mathbf{gradv} = \begin{pmatrix} \frac{\partial v_1}{\partial x_1} & \frac{\partial v_1}{\partial x_2} \\ \frac{\partial v_2}{\partial x_1} & \frac{\partial v_2}{\partial x_2} \end{pmatrix}, \qquad (\mathbf{v}\,\boldsymbol{\nabla})\mathbf{v} = \begin{pmatrix} v_1\frac{\partial v_1}{\partial x_1} + v_2\frac{\partial v_1}{\partial x_2} \\ v_1\frac{\partial v_2}{\partial x_1} + v_2\frac{\partial v_2}{\partial x_2} \end{pmatrix}.$$

2. Inverse and implicit function theorems

In our analysis, the inverse and implicit function theorems will play an essential role. We present here the results in a form convenient for our use. Let $G : X \to Z$ be a C^1 mapping and $v \in X$ be such that $DG(v) \in \mathscr{L}(X;Z)$ is an isomorphism. We introduce the notations

$$\begin{aligned} \epsilon &= \|G(v)\|_Z, \\ \gamma &= \|DG(v)^{-1}\|_{Z;X}, \\ L(\alpha) &= \sup_{x \in \overline{B}(v,\alpha)} \|DG(v) - DG(x)\|_{X;Z}, \end{aligned} \tag{2.1}$$

and we are interested in the problem to find $x \in X$ such that

$$G(x) = 0. \tag{2.2}$$

Theorem 2.1. *We assume that $DG(v) \in \mathscr{L}(X;Z)$ is an isomorphism and that $2\gamma L(2\gamma\epsilon) \leq 1$. Then problem (2.2) has a unique solution u in the ball $\overline{B}(v, 2\gamma\epsilon)$ and $DG(u) \in \mathscr{L}(X;Z)$ is invertible with*

$$\|DG(u)^{-1}\|_{Z;X} \leq 2\gamma. \tag{2.3}$$

Moreover for all $x \in \overline{B}(v, 2\gamma\epsilon)$

$$\|x - u\|_X \leq 2\gamma \|G(x)\|_Z. \tag{2.4}$$

Proof. The proof follows Brezzi, Rappaz, and Raviart [1980] or Crouzeix and Rappaz [1989]. Let $H : X \to X$ be the mapping defined by

$$H(x) = x - DG(v)^{-1} G(x). \tag{2.5}$$

Clearly x is a fixed point of H if and only if x is a zero of the mapping G.

For any $x \in \overline{B}(v, 2\gamma\epsilon)$ we can write

$$H(x) - v = DG(v)^{-1} \big[DG(v)(x - v) - (G(x) - G(v)) \big] - DG(v)^{-1} G(v).$$

With the Taylor expansion (1.4) with $p = 1$ we get

$$\begin{aligned} \|H(x) - v\|_X &\leq \gamma \left[\epsilon + \| \int_0^1 \big(DG(v) - DG(v + t(x - v)) \big)(x - v)\, dt \|_Z \right] \\ &\leq \gamma \left[\epsilon + L(2\gamma\epsilon) 2\gamma\epsilon \right] \leq 2\gamma\epsilon. \end{aligned}$$

Thus H maps the closed ball $\overline{B}(v, 2\gamma\epsilon)$ into itself.

Let now x, y be in $\overline{B}(v, 2\gamma\epsilon)$, then

$$H(x) - H(y) = DG(v)^{-1} \int_0^1 \big[DG(v) - DG(y + t(x-y))\big](x-y)\, dt$$

and

$$\|H(x) - H(y)\|_X \leq \gamma L(2\gamma\epsilon)\|x-y\|_X \leq \frac{1}{2}\|x-y\|_X.$$

We have proved that H is a strict contraction from the ball $\overline{B}(v, 2\gamma\epsilon)$ into itself. So H possesses a unique fixed point u in the ball $\overline{B}(v, 2\gamma\epsilon)$. Since

$$\|DG(v)^{-1}\,[DG(u) - DG(v)]\,\|_{X;X} \leq \gamma L(2\gamma\epsilon) \leq \frac{1}{2}$$

and

$$DG(u) = DG(v) + \big(DG(u) - DG(v)\big),$$

by (1.1) and (1.2), $DG(u)$ is invertible and (2.3) is true, i.e.

$$\|DG(u)^{-1}\,\|_{Z;X} \leq 2\gamma.$$

Finally to prove the estimate (2.4), we write for $\alpha > 0$ and $x \in \overline{B}(v, \alpha)$

$$\begin{aligned} u - x &= H(u) - x \\ &= DG(v)^{-1}\left[-G(x) + \int_0^1 \big[DG(v) - DG(u + t(x-u))\big](u-x)\, dt\right] \end{aligned}$$

and

$$\|u - x\|_X \leq \gamma\big[\|G(x)\|_Z + L(\alpha)\|u-x\|_X\big].$$

Thus we finally get

$$(1 - \gamma L(\alpha))\,\|u - x\|_X \leq \gamma\|G(x)\|_Z. \quad \square$$

Remark 2.1. The uniqueness result in Theorem 2.1 can be improved in the following way. Under the hypotheses of the theorem and for all $\alpha \geq 2\gamma\epsilon$ satisfying $\gamma L(\alpha) < 1$, problem (2.2) has a unique solution in the ball $\overline{B}(v, \alpha)$. Indeed if u_1 and u_2 are two different fixed points in $\overline{B}(v, \alpha)$ then

$$\begin{aligned} \|u_1 - u_2\|_X &= \|H(u_1) - H(u_2)\|_X \\ &\leq \gamma L(\alpha)\|u_1 - u_2\|_X < \|u_1 - u_2\|_X \end{aligned}$$

and we have a contradiction. $\square$

Remark 2.2. In Theorem 2.1 the hypothesis on the differentiability of G can be relaxed. If there exists an isomorphism $A \in \mathscr{L}(X;Z)$ and $v \in X$ such that $2\gamma L(2\gamma\epsilon) \leq 1$, with

$$\begin{aligned} \epsilon &= \|G(v)\|_Z, \\ \gamma &= \|A^{-1}\|_{Z;X}, \\ L(\alpha) &= \sup_{x,y\in\overline{B}(v,\alpha)} \frac{\|G(x) - G(y) - A(x-y)\|_Z}{\|x-y\|_X}, \end{aligned}$$

then problem (2.2) has a unique solution in $\overline{B}(v, 2\gamma\epsilon)$. Moreover the estimate (2.4) still holds.

Note that if $\lim_{\alpha\to 0} L(\alpha) = 0$, then G is strongly differentiable at v and satisfies a Lipschitz condition in a neighborhood of v, see NIJENHUIS [1974] for details. □

Remark 2.3. Theorem 2.1 is the basic result to start an error analysis when we consider approximations of nonlinear problems. Usually u will be a solution of the approximate problem and $x = v$ the solution of the exact problem. The estimate (2.4) is the main relation to analyze the error between exact and approximated solutions. □

The results stated in Theorem 2.1 will be sufficient for most applications. In fact under similar hypotheses we get the following variant of the inverse function theorem.

Theorem 2.2. *For $v \in X$ and the function $G : X \to Z$ of class C^p, $p \geq 1$, we assume that $DG(v) \in \mathscr{L}(X;Z)$ is an isomorphism and that α satisfies $2\gamma L(\alpha) \leq 1$, with $\gamma = \|DG(v)^{-1}\|_{Z;X}$. Then there exists a C^p mapping $F : B\left(G(v), \alpha/2\gamma\right) \to B(v,\alpha)$ such that for all $z \in B\left(G(v), \alpha/2\gamma\right)$ we have*

$$G(F(z)) = z \tag{2.6}$$

and

$$DF(z) = \left[DG(F(z))\right]^{-1}.$$

Moreover for all z_1, z_2 in $B\left(G(v), \alpha/2\gamma\right)$

$$\|F(z_1) - F(z_2)\|_X \leq 2\gamma\|z_1 - z_2\|_Z. \tag{2.7}$$

Proof. We refer to BREZZI, RAPPAZ, and RAVIART [1980] for a detailed proof. We will only sketch it here. For $z \in B\left(G(v), \alpha/2\gamma\right)$ we define the mapping $H : X \to X$ by

$$H(x) = x + DG(v)^{-1}(z - G(x)).$$

As in the proof of Theorem 2.1, we can prove that for x, $y \in \overline{B}(v,\alpha)$

$$\|H(x) - H(y)\|_X \leq \frac{1}{2}\|x - y\|_X$$

and for $x \in \overline{B}(v, \alpha_1)$ with $\alpha_1 = 2\gamma\|z - G(v)\|_Z \leq \alpha$

$$\|H(x) - v\|_X \leq \|H(x) - H(v)\|_X + \|H(v) - v\|_X \leq \alpha_1 .$$

Hence the mapping H maps the closed ball $\overline{B}(v, \alpha_1)$ into itself and is a contraction. So H has a unique fixed point in $\overline{B}(v, \alpha_1)$; that is for any $z \in B\left(G(v), \alpha/2\gamma\right)$ the equation $G(x) = z$ has a unique solution $x = F(z)$ in $\overline{B}(v, \alpha_1)$, which is unique in $B(v, \alpha)$ (in a similar way as in Remark 2.1). Whence the mapping $F : B\left(G(v), \alpha/2\gamma\right) \to B(v, \alpha)$ is well-defined. In a same way we proved (2.3), we get for $x \in B(v, \alpha)$ that $DG(x)$ is an isomorphism from X onto Z with $\|DG(x)^{-1}\| \leq 2\gamma$. Then in a standard way we prove that $DF(z) = \left[DG(F(z))\right]^{-1}$ and that F is of class C^p. Finally the estimate (2.7) is a direct consequence of (2.6) and of a Taylor expansion of F. $\square$

If we consider a C^1 mapping $G : \Lambda \times X \to Z$ where Λ is an other Banach space and if (λ_0, x_0) is an element of $\Lambda \times X$, then $DG(\lambda_0, x_0) \in \mathscr{L}(\Lambda \times X; Z)$, $D_\lambda G(\lambda_0, x_0) \in \mathscr{L}(\Lambda; Z)$, and $D_x G(\lambda_0, x_0) \in \mathscr{L}(X; Z)$ are respectively the total derivative of G, the derivative of G with respect to λ and the derivative of G with respect to x at the point (λ_0, x_0).

Finally we recall a version of the implicit function theorem. Let Λ, X, and Z be three Banach spaces and $G : \Lambda \times X \to Z$ be a C^p mapping with $p \geq 1$. For a given $(\lambda_0, x_0) \in \Lambda \times X$, we assume that $D_x G(\lambda_0, x_0) \in \mathscr{L}(X; Z)$ is an isomorphism and we introduce the notations

$$\begin{aligned} \epsilon &= \|G(\lambda_0, x_0)\|_Z , \\ \gamma_0 &= \|D_\lambda G(\lambda_0, x_0)\|_{\Lambda;Z} , \\ \gamma_1 &= \|D_x G(\lambda_0, x_0)^{-1}\|_{Z;X} , \\ L(\alpha) &= \sup_{(\lambda,x) \in \overline{B}((\lambda_0,x_0),\alpha)} \|DG(\lambda_0, x_0) - DG(\lambda, x)\|_{\Lambda \times X;Z} . \end{aligned} \tag{2.8}$$

Theorem 2.3. *Let α satisfy $2\gamma L(\alpha) \leq 1$ with $\gamma = \max(\gamma_1, 1 + \gamma_0\gamma_1)$. If $\epsilon < \alpha/4\gamma$, then there exists a unique C^p mapping $g : B(\lambda_0, \alpha/4\gamma) \subset \Lambda \to B(x_0, \alpha) \subset X$ satisfying for all $\lambda \in B(\lambda_0, \alpha/4\gamma)$*

$$G\left(\lambda, g(\lambda)\right) = 0$$

and

$$\|g(\lambda) - x_0\|_X \leq 2\gamma(\epsilon + \|\lambda - \lambda_0\|_\Lambda). \tag{2.9}$$

Proof. We consider the mapping $F : \Lambda \times X \to \Lambda \times Z$ given by

$$F(\lambda, x) = \left(\lambda, G(\lambda, x)\right) .$$

Clearly F is of class C^p and for $(\lambda, x) \in \Lambda \times X$ we have

$$DF(\lambda_0, x_0)(\lambda, x) = \big(\lambda, D_\lambda G(\lambda_0, x_0)\lambda + D_x G(\lambda_0, x_0)x\big).$$

Since $D_x G(\lambda_0, x_0)$ is an isomorphism from X onto Z, so is $DF(\lambda_0, x_0)$ from $\Lambda \times X$ onto $\Lambda \times Z$ and for all $(\lambda, z) \in \Lambda \times Z$ we get

$$\|DF(\lambda_0, x_0)^{-1}(\lambda, z)\|_{\Lambda\times X} \leq \max(1 + \gamma_0\gamma_1, \gamma_1)\,(\|\lambda\|_\Lambda + \|z\|_Z)\,.$$

So we get the estimate

$$\|DF(\lambda_0, x_0)^{-1}\|_{\Lambda\times Z;\Lambda\times X} \leq \gamma.$$

Moreover for all $(\lambda, x) \in \overline{B}\,((\lambda_0, x_0), \alpha)$, we have

$$\|DF(\lambda, x) - DF(\lambda_0, x_0)\|_{\Lambda\times X;\Lambda\times Z} \leq L(\alpha).$$

We are in a position to apply Theorem 2.2 with $G = F$ and $v = (\lambda_0, x_0)$. There exists a unique C^p mapping $\mathscr{F} : B\,(F(\lambda_0, x_0), \alpha/2\gamma) \to B\,((\lambda_0, x_0), \alpha)$ such that

$$F\,(\mathscr{F}(\lambda, z)) = (\lambda, z).$$

Since we assume $\epsilon < \alpha/4\gamma$, then $(\lambda, 0)$ is in the ball $B\,((\lambda_0, G(\lambda_0, x_0))\,, \alpha/2\gamma)$ provided $\|\lambda - \lambda_0\|_\Lambda \leq \alpha/4\gamma$. Hence the mapping $g : B(\lambda_0, \alpha/4\gamma) \to B(x_0, \alpha)$ given by

$$g(\lambda) = P\mathscr{F}(\lambda, 0)$$

with P the projection operator from $\Lambda \times X$ onto X, is well defined on $B(\lambda_0, \alpha/4\gamma)$, of class C^p, and satisfies

$$G\,(\lambda, g(\lambda)) = 0 \quad \text{for} \quad \lambda \in B(\lambda_0, \alpha/4\gamma).$$

Finally we can use the estimate (2.7) to get (2.9) in the following way. For $\lambda \in B(\lambda_0, \alpha/4\gamma)$

$$\begin{aligned}\|g(\lambda) - x_0\|_X &= \|P\mathscr{F}(\lambda, 0) - P\mathscr{F}\,(\lambda_0, G(\lambda_0, x_0))\,\|_X \\ &\leq 2\gamma\,(\|\lambda - \lambda_0\|_\Lambda + \|G(\lambda_0, x_0)\|_Z)\,. \quad \square\end{aligned}$$

Remark 2.4. The hypothesis that G is C^1 can be relaxed in a same way it is done in Remark 2.2. For an operator $A \in \mathscr{L}(\Lambda; Z)$ and an isomorphism $B \in \mathscr{L}(X; Z)$, we introduce the notations

$$\begin{aligned}\gamma_0 &= \|A\|_{\Lambda;Z}\,, \\ \gamma_1 &= \|B^{-1}\|_{Z;X}\,, \\ L(\alpha) &= \sup_{(\lambda,x),(\lambda^*,x^*)\in\overline{B}((\lambda_0,x_0),\alpha)} \frac{\|G(\lambda, x) - G(\lambda^*, x^*) - A(\lambda - \lambda^*) - B(x - x^*)\|_Z}{\|\lambda - \lambda^*\|_\Lambda + \|x - x^*\|_X}.\end{aligned}$$

We assume that α is such that $2\gamma L(\alpha) \leq 1$ and $\epsilon < \alpha/4\gamma$ with $\gamma = \max(\gamma_1, 1 + \gamma_0\gamma_1)$. Then Theorem 2.3 holds with $p = 0$. This generalization has several applications, see Girault and Raviart [1982], Rappaz [1984], Crouzeix and Rappaz [1989], Barrett and Elliott [1989], Caloz [1987] [1991] for instance. $\square$

CHAPTER II

MODEL EXAMPLES

In this chapter we present examples of nonlinear problems and introduce standard finite element approximations for all of them. It is not our purpose to present many applications where parametrized equations arise. We prefer to choose model problems and discuss thoroughly their finite element approximations. We shall consider successively the case of semilinear second order elliptic boundary value problem, the Navier-Stokes equations, and a stationary heat problem with convection.

3. Semilinear problems

A generic example to our theory is the case of semilinear boundary value problems. This class includes problems of the form

$$\begin{aligned} -\Delta u + g(\lambda, u) &= 0 \quad \text{in} \quad \Omega, \\ u &= 0 \quad \text{on} \quad \partial\Omega, \end{aligned} \tag{3.1}$$

where Ω is a bounded open set in $\mathbb{R}^2$ with smooth boundary $\partial\Omega$, Δ is the Laplacian operator, λ is a real parameter, and $g : \mathscr{D}(g) \subset \mathbb{R} \times H_0^1(\Omega) \to L^2(\Omega)$ is a given mapping. To study problem (3.1) in a convenient functional setting, we define the mapping $F : \mathscr{D}(g) \subset \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega)$ by

$$\text{for all}\,(\lambda, v) \in \mathscr{D}(g), \quad F(\lambda, v) = -\Delta v + g(\lambda, v)$$

or equivalently for all $(\lambda, v) \in \mathscr{D}(g)$ and for all $w \in H_0^1(\Omega)$

$$\langle F(\lambda, v), w \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx + \int_\Omega g(\lambda, v) w\, dx. \tag{3.2}$$

Then we look for solutions $(\lambda, u) \in \mathscr{D}(g)$ of the problem

$$F(\lambda, u) = 0. \tag{3.3}$$

The interesting point here is to study how the solutions u of (3.3) depend on λ and to develop numerical methods to compute them.

Given a function $f \in L^2(\Omega)$, we set $w \in H_0^1(\Omega)$ to be the unique solution to

$$\text{for all}\, v \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad} w\, \mathbf{grad} v\, dx = \int_\Omega f v\, dx. \tag{3.4}$$

Thus we have defined the continuous linear operator $T : f \in L^2(\Omega) \to Tf = w \in H_0^1(\Omega)$ which maps f onto the solution w of problem (3.4). It is useful but not always necessary for further development to assume that

$$T \in \mathscr{L}(L^2(\Omega); H_0^1(\Omega) \cap H^2(\Omega)), \tag{3.5}$$

which is a restriction on the regularity of $\partial\Omega$, see the standard elliptic regularity result in GILBARG and TRUDINGER [1977] or GRISVARD [1985] for instance. Then the operator T is compact from $L^2(\Omega)$ into $H_0^1(\Omega) \cap C^0(\overline{\Omega})$. Introducing the operator T, problem (3.3) can be written in the equivalent way

$$u + Tg(\lambda, u) = 0. \tag{3.6}$$

With some specific choices of the function g, the solution set to problem (3.3) or (3.6) can be described precisely. Several cases will be developed later.

Given a finite-dimensional space $V_h \subset H_0^1(\Omega)$, a Galerkin approximation to problem (3.3) consists in finding $(\lambda, u_h) \in (\mathbb{R} \times V_h) \cap \mathscr{D}(g)$ such that

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad} u_h \, \mathbf{grad} v_h \, dx + \int_\Omega g(\lambda, u_h) v_h \, dx = 0. \tag{3.7}$$

Usually problem (3.7) is not solved as it is but with numerical quadrature rules. We introduce a projector P_h onto V_h and consider the problem of finding $(\lambda, u_h) \in (\mathbb{R} \times V_h) \cap \mathscr{D}(g)$ such that

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad} u_h \, \mathbf{grad} v_h \, dx + \int_\Omega P_h \big[g(\lambda, u_h) v_h \big] \, dx = 0. \tag{3.8}$$

In order to write the approximate problem (3.7) in a way similar to (3.6), we set for a given $f \in L^2(\Omega)$, $w_h \in V_h$ to be the unique solution to

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad} w_h \, \mathbf{grad} v_h \, dx = \int_\Omega f v_h \, dx.$$

So we have defined a mapping $T_h \in \mathscr{L}(L^2(\Omega); V_h)$ and (3.7) reads

$$u_h + T_h g(\lambda, u_h) = 0. \tag{3.9}$$

Remark 3.1. Note that the two problems (3.6) and (3.9) have a structure identical to the general one studied by BREZZI, RAPPAZ, and RAVIART [1980, 1981a, 1981b]

$$u + Tg(\lambda, u) = 0, \qquad u_h + T_h g(\lambda, u_h) = 0,$$

where T is in $\mathscr{L}(W; V)$, W and V are two real Banach spaces, $g : \mathscr{D}(g) \subset \mathbb{R} \times V \to W$ is a regular mapping, $\{V_h\}_{0<h\leq 1}$ is a family of finite-dimensional subspaces of V and for $0 < h \leq 1$, T_h is a linear mapping in $\mathscr{L}(W; V_h)$. Under the main hypothesis

$$\lim_{h \to 0} \|T - T_h\|_{W;V} = 0$$

they have analyzed regular branches of solutions in BREZZI, RAPPAZ, and RAVIART [1980], limit points [1981a], and simple bifurcation points [1981b]. Bifurcations at multiple eigenvalues have been studied in RAPPAZ and RAUGEL [1982]. □

Two model choices of a function $g : \mathscr{D}(g) \subset \mathbb{R} \times H_0^1(\Omega) \to L^2(\Omega)$ are presented now. First for a given function $f \in L^2(\Omega)$, we consider problem (3.1) with $g(\lambda, u) = u^3 - f$ and the problem reads : find $u \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$

$$\langle F(u), v\rangle_{H^{-1}(\Omega)\ H_0^1(\Omega)} \equiv \int_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dx + \int_\Omega (u^3 - f) v\, dx = 0. \tag{3.10}$$

The mapping g is independent of λ and this example will illustrate our general approach applied to a problem not depending on a parameter. It is a classical matter to prove the existence of a solution to problem (3.10). The following result holds.

Theorem 3.1. *There exists a unique solution $u \in H_0^1(\Omega)$ to the problem (3.10) and u is in $H^2(\Omega)$. Moreover the solution u is regular, in the sense that the derivative $DF(u) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is an isomorphism.*

Proof. With the notation introduced in (3.6), our problem $F(u) = 0$ is equivalent to

$$u + T(u^3 - f) = 0. \tag{3.11}$$

Since the mapping $u \in H_0^1(\Omega) \to u^3 - f \in L^2(\Omega)$ is continuous and compact, the mapping $u \in H_0^1(\Omega) \to T(u^3 - f) \in H_0^1(\Omega)$ is compact. If the real number μ is in $[0, 1]$ and the function $w \in H_0^1(\Omega)$ satisfy

$$\text{for all } v \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad} w\, \mathbf{grad} v\, dx + \mu \int_\Omega (w^3 - f) v\, dx = 0,$$

by taking $v = w$ we get the estimate

$$|w|_{1,\Omega}^2 + \mu \|w\|_{0,4,\Omega}^4 \le \mu \|f\|_{0,\Omega} \|w\|_{0,\Omega}$$

which implies

$$|w|_{1,\Omega}^2 \le \|f\|_{0,\Omega} \|w\|_{0,\Omega}.$$

With Poincaré Inequality we obtain a bound for $|w|_{1,\Omega}$ independent of μ. As a consequence of the Leray-Schauder homotopy theorem, see LERAY and SCHAUDER [1934] p.64, we get the existence of a solution to (3.11). The regularity assumption implies that u is in $H^2(\Omega)$.

Let now u_1 and u_2 be two solutions to (3.11), then

$$\text{for all } v \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad}(u_1 - u_2)\, \mathbf{grad} v\, dx + \int_\Omega (u_1^3 - u_2^3) v\, dx = 0$$

or with $v = u_1 - u_2$

$$\int_\Omega |\mathbf{grad}(u_1 - u_2)|^2\, dx + \int_\Omega (u_1 - u_2)^2 \left[\frac{u_1^2 + u_2^2}{2} + \frac{(u_1 + u_2)^2}{2} \right] dx = 0,$$

which implies $u_1 = u_2$.

Let $u \in H_0^1(\Omega)$ be the solution to (3.11). For all $v, w \in H_0^1(\Omega)$ we have

$$\langle DF(u)v, w\rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx + 3\int_\Omega u^2 vw\, dx \equiv b(v,w).$$

The bilinear form $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ is coercive on $H_0^1(\Omega)$. So the derivative $DF(u) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is an isomorphism. □

We consider now finite element approximations to problem (3.10). For ease of exposition we assume that Ω is a convex polygonal domain in $\mathbb{R}^2$, so it makes sense to consider $\{\mathscr{T}_h\}_{0<h\leq 1}$ to be a regular family of triangulations of $\overline{\Omega}$, in the sense defined in CIARLET [1978], p.124 or in CIARLET [1991], p.128. The maximum diameter of the triangles $T \in \mathscr{T}_h$ is equal to h. The approximation space and the test space are chosen to be

$$V_h = \{v \in C^0(\overline{\Omega}); v_{|T} \in \mathscr{P}_1(T), \text{ for all } T \in \mathscr{T}_h\} \cap H_0^1(\Omega), \tag{3.12}$$

where $\mathscr{P}_1(T)$ represents the space of polynomials of degree ≤ 1 defined on T. The corresponding finite element approximation to (3.10) reads

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad} u_h\, \mathbf{grad} v_h\, dx + \int_\Omega (u_h^3 - f)v_h\, dx = 0. \tag{3.13}$$

Suppose that in the system (3.13), we would like to use numerical integration. For instance we could use the generalized trapezoidal quadrature rule which is exact for polynomials of degree up to 1 on each triangle. Then the problem reads: find $u_h \in V_h$ such that for all $v_h \in V_h$

$$\int_\Omega \mathbf{grad} u_h\, \mathbf{grad} v_h\, dx + \int_\Omega r_h\left[(u_h^3 - f)v_h\right] dx = 0, \tag{3.14}$$

where the V_h-interpolation operator $r_h \in \mathscr{L}(C^0(\overline{\Omega}); V_h)$ is defined for $v \in C^0(\overline{\Omega})$ by

$$r_h v \in V_h, \quad r_h v(x_i) = v(x_i) \quad \text{for all node } x_i \text{ of } \mathscr{T}_h.$$

Note that here the function f is supposed to be continuous in $\overline{\Omega}$. The main goal of the next chapter will be to develop a general formalism to study nonlinear problems. As an example we shall prove that the problem (3.13) or the (3.14) one has a unique solution u_h in some neighborhood of the solution u to problem (3.10). In particular we will show that u_h converges to u in $H_0^1(\Omega)$ when h tends to zero and we will establish a priori and a posteriori error estimates for $\|u - u_h\|_{1,\Omega}$.

The second example of a function $g : \mathscr{D}(g) \subset \mathbb{R} \times H_0^1(\Omega) \to L^2(\Omega)$ is

$$g(\lambda, u) = -\lambda f(u),$$

where f is the Nemytskii operator of some function $\phi : \Omega \times \mathbb{R} \to \mathbb{R}$, that is for every function $u : \Omega \to \mathbb{R}$ we have

$$\text{for all}\, x \in \Omega \quad f(u)(x) = \phi(x, u(x)).$$

The function $\phi : (x, z) \in \Omega \times \mathbb{R} \to \phi(x, z) \in \mathbb{R}$ is assumed to satisfy the following conditions:

$$\begin{cases} \text{the function } \phi \text{ is continuously differentiable with respect to } z \text{ uni-} \\ \text{formly in } \Omega \text{ and for all } z \in \mathbb{R},\ \phi(.,z) \text{ and } \dfrac{\partial \phi}{\partial z}(.,z) \text{ are in } L^\infty(\Omega), \end{cases} \tag{3.15}$$

$$\phi(.,0) > 0 \quad \text{in} \quad \Omega, \tag{3.16}$$

$$\frac{\partial \phi}{\partial z}(x,z) > 0 \quad \text{for} \quad x \in \Omega,\ z \geq 0. \tag{3.17}$$

Under these regularity and monotonicity hypotheses on ϕ, we shall study the problem: for $\lambda > 0$ find $u \in H_0^1(\Omega)$ nonnegative in Ω solution to

$$u - \lambda T f(u) = 0. \tag{3.18}$$

The hypothesis (3.15) implies that $f : H_0^1(\Omega) \cap C^0(\overline{\Omega}) \to L^2(\Omega)$ is well defined and C^1 in a neighborhood of a solution of (3.18).

We mention that the corresponding problems have physical motivations. For instance they concern the temperature distribution in an object heated by an electric current, in which case only the positive temperature u is of interest. It is known that a limit current exists beyond which steady state solutions do not exist, see JOSEPH [1965]. The choice $\lambda = 0$ represents the steady state problem with no current. The boundary condition is assumed to be homogeneous, only to subtract a harmonic function with a prescribed value on $\partial\Omega$.

To prove the existence of a value λ^* beyond which no solution exists, we shall use the following corollary of the maximum principle.

Theorem 3.2. *Let ρ be a nonnegative, not identically zero function in $L^\infty(\Omega)$ and μ_1 be the least (or principal) eigenvalue of*

$$\begin{aligned} -\Delta\zeta - \mu\rho\zeta &= 0 \quad \textit{in} \quad \Omega, \\ \zeta &= 0 \quad \textit{on} \quad \partial\Omega. \end{aligned} \tag{3.19}$$

Then the eigenspace E_1 corresponding to μ_1 is one-dimensional and any non zero function $\zeta \in E_1$ has a constant sign in Ω. Furthermore let $\psi \in H_0^1(\Omega) \cap H^2(\Omega)$ satisfy

$$-\Delta\psi - \lambda\rho\psi \gneqq 0 \quad \textit{in} \quad \Omega. \tag{3.20}$$

Then ψ is positive in Ω if and only if $\lambda < \mu_1$.

Proof. The theorem will be a consequence of the classical maximum principle. A function $v \in H_0^1(\Omega) \cap H^2(\Omega)$ with a non positive not identically zero Laplacian in Ω is strictly positive in Ω. The first point is a classical result.

The eigenvalue μ_1 is characterized by the Rayleigh quotient

$$\mu_1 = \min_{\substack{v\in H_0^1(\Omega)\\ v\neq 0}} \frac{|v|^2_{1,\Omega}}{\|\sqrt{\rho}v\|^2_{0,\Omega}},$$

and $E_1 = \{v \in H_0^1(\Omega); |v|^2_{1,\Omega} - \mu_1\|\sqrt{\rho}v\|^2_{0,\Omega} = 0\}$. Now let ζ be a non zero function in E_1. Then the function $\eta = |\zeta|$ is in $H_0^1(\Omega)$ and necessarily

$$|\eta|^2_{1,\Omega} - \mu_1\|\sqrt{\rho}\eta\|^2_{0,\Omega} = 0,$$

which means η is in E_1. Since $\zeta_+ \equiv \sup(\zeta, 0) = \frac{1}{2}(\zeta+\eta)$ and $\zeta_- \equiv -\inf(\zeta, 0) = \frac{1}{2}(\eta-\zeta)$, we conclude, E_1 being a vector space, that ζ_+ and ζ_- are also in E_1, that is ζ_+ and ζ_- are in $H_0^1(\Omega)$ and

$$-\Delta\zeta_+ = \mu_1\rho\zeta_+ \quad \text{in} \quad \Omega, \qquad -\Delta\zeta_- = \mu_1\rho\zeta_- \quad \text{in} \quad \Omega.$$

By the maximum principle and since $\zeta \not\equiv 0$, either $\zeta_+ > 0$ and then $\zeta_- \equiv 0$, or $\zeta_+ \equiv 0$ and then $\zeta_- > 0$.

Set $p = -\Delta\psi - \lambda\rho\psi$. Assume first that $\lambda < \mu_1$. Then let $\varphi \in H_0^1(\Omega)$ be the unique minimum of the functional

$$J(v) = \frac{1}{2}\int_\Omega \mathbf{grad}v\,\mathbf{grad}v\,dx - \frac{\lambda}{2}\int_\Omega \rho v^2\,dx - \int_\Omega pv\,dx$$

over $v \in H_0^1(\Omega)$. Such a function exists, is unique, and by standard elliptic regularity φ belongs to $H^2(\Omega)$, thus $\varphi = \psi$. Suppose that ψ is negative somewhere. Then we define the function $\eta = |\psi|$ which is in $H_0^1(\Omega)$. Since the function p is non negative, we get $J(\eta) \le J(\psi)$, which contradicts the fact that ψ is the unique minimum of J. By the maximum principle and the relation (3.20), ψ cannot be zero in any open subset of Ω.

We assume now that ψ is positive and let ζ_1 be the positive eigenvector corresponding to μ_1 with $\|\zeta_1\|_{0,\Omega} = 1$. Then for v_1, $v_2 \in H_0^1(\Omega)$, we have

$$\int_\Omega \mathbf{grad}\psi\,\mathbf{grad}v_1\,dx - \lambda\int_\Omega \rho\psi v_1\,dx = \int_\Omega pv_1\,dx,$$
$$\int_\Omega \mathbf{grad}\zeta_1\,\mathbf{grad}v_2\,dx - \mu_1\int_\Omega \rho\zeta_1 v_2\,dx = 0,$$

or setting $v_1 = \zeta_1$, $v_2 = \psi$ and subtracting

$$(-\lambda + \mu_1)\int_\Omega \rho\psi\zeta_1\,dx = \int_\Omega p\zeta_1\,dx,$$

so $-\lambda + \mu_1 > 0$. □

We first note that if ϕ satisfies (3.16), (3.17) then ϕ is positive for positive z and problem (3.18) can have a positive solution only for positive λ. The only solution of (3.18) with $\lambda = 0$ is $u \equiv 0$. With the implicit function theorem 2.3, it is not difficult to check that for λ in a right neighborhood of 0, (3.18) has a positive solution. So let $\lambda^* > 0$ be given by

$$\lambda^* = \sup\{\lambda \ge 0; (3.18)\,\text{has a positive solution}\}.$$

Theorem 3.3. *We assume that ϕ satisfies (3.15), (3.16), and (3.17). Then for all $\lambda \in (0, \lambda^*)$, problem (3.18) has a minimal positive solution $u(\lambda)$, that is if v is a positive solution then $u(\lambda) \leq v$ in Ω. Moreover $u(\lambda)$ is an increasing function of λ, that is if $0 < \lambda_1 < \lambda_2 < \lambda^*$ then*

$$0 \leq u(\lambda_1) < u(\lambda_2) \quad in \quad \Omega.$$

Proof. Let $\lambda \in (0, \lambda^*)$ be fixed. There exists $\overline{\lambda}$, $\lambda \leq \overline{\lambda} \leq \lambda^*$, such that $(\overline{\lambda}, \overline{u})$ is a solution of (3.18) with $\overline{u} \geq 0$. We define then the sequences

$$\begin{aligned} \underline{u}_0 &\equiv 0, \qquad \underline{u}_{i+1} = \lambda T f(\underline{u}_i), \\ \overline{u}_0 &= \overline{u}, \qquad \overline{u}_{i+1} = \lambda T f(\overline{u}_i). \end{aligned}$$

We prove that the sequence $\{\underline{u}_i\}_{i\geq 0}$ is monotonous increasing, that is

$$\text{for all}\, x \in \Omega \qquad \underline{u}_{i+1}(x) > \underline{u}_i(x).$$

With the assumption (3.16), the function $\underline{u}_1 \in H_0^1(\Omega)$ satisfies

$$-\Delta \underline{u}_1 = \lambda \phi(.,0) > 0 \quad \text{in} \quad \Omega,$$

and by the maximum principle $\underline{u}_1 > 0$. Suppose now we have $0 < \underline{u}_1 < \underline{u}_2 < \cdots < \underline{u}_i$. This implies that the difference $\underline{u}_{i+1} - \underline{u}_i$ is a solution to

$$\begin{aligned} -\Delta(\underline{u}_{i+1} - \underline{u}_i) &= \lambda \left[f(\underline{u}_i) - f(\underline{u}_{i-1}) \right] \quad \text{in} \quad \Omega, \\ \underline{u}_{i+1} - \underline{u}_i &= 0. \end{aligned}$$

So using (3.17) we get $\underline{u}_i < \underline{u}_{i+1}$. Finally with the same arguments we prove

$$0 \equiv \underline{u}_0 < \underline{u}_1 < \cdots < \underline{u}_n < \cdots < \overline{u}_n \cdots < \overline{u}_1 < \overline{u}_0 .$$

The sequence $\{\underline{u}_i\}_{i\geq 0}$ is bounded in $L^\infty(\Omega)$; since T is compact as an operator in $\mathscr{L}(L^2(\Omega); H_0^1(\Omega) \cap C^0(\overline{\Omega}))$, there exists a subsequence $\{\underline{u}_{n_i}\}_{i\geq 0}$ converging to u in $H_0^1(\Omega) \cap C^0(\overline{\Omega})$, with $u \leq \overline{u}$. In fact the whole sequence converges to $u \geq 0$, which is a solution of (3.18). Clearly from our construction u is minimal and we set $u = u(\lambda)$. Moreover if $\lambda < \overline{\lambda} < \lambda^*$, then

$$u(\lambda)(x) < u(\overline{\lambda})(x) \quad \text{for all}\, x \in \Omega. \quad \square$$

The method of proof used here is known as the monotonous iteration principle.

If we add some hypotheses on ϕ, then we can get more information on λ^*. Suppose for instance there exists a positive function $\varphi \in L^2(\Omega)$ such that

$$\text{for all } x \in \Omega \text{ and } z > 0 \qquad \phi(x,z) < \varphi(x). \tag{3.21}$$

Then we can start the monotonous iteration scheme with

$$\underline{u}_0 = 0 \qquad \text{and} \qquad \overline{u}_0 = \lambda T\varphi.$$

Therefore we get the existence of a positive solution for all $\lambda > 0$ and we have $\lambda^* = \infty$. Suppose on the contrary that ϕ satisfies

$$\phi(x,z) > \varphi(x) + \rho(x)z \qquad \text{for all } x \in \Omega \text{ and } z > 0 \tag{3.22}$$

with positive functions $\varphi, \rho \in L^\infty(\Omega)$. Then $\lambda^* \leq \mu_1$, where μ_1 is the least eigenvalue of (3.19). Indeed suppose that the problem (3.18) has a solution $u > 0$ for some $\lambda > \mu_1$. This implies

$$-\Delta u = \lambda\phi(x,u) > \lambda\rho u \quad \text{in} \quad \Omega.$$

We get a contradiction from Theorem 3.2. So under the hypothesis (3.22), there exists a $\lambda^* < \infty$ beyond which problem (3.18) has no positive solution.

So far we have used essentially the monotonicity of ϕ. We assume furthermore

$$\text{for all}\, x \in \Omega, \quad \frac{\partial\phi}{\partial z}(x,.) \quad \text{is strictly increasing in} \quad \mathbb{R}_+. \tag{3.23}$$

Then we can say more on $u(.)$ as a function of λ.

Theorem 3.4. *We assume that ϕ satisfies (3.15), (3.16), (3.17), and (3.23). Then λ^* is bounded by μ_1 the principal characteristic value of*

$$\psi = \mu T Df(0)\psi$$

and the mapping $\lambda \in [0,\lambda^*) \to u(\lambda) \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ where $u(\lambda)$ denotes the minimal positive solution is C^1 and $(\lambda, u(\lambda))$ is a regular point of (3.18), that is $I - \lambda T Df(u(\lambda)) \in \mathscr{L}\big(H_0^1(\Omega) \cap C^0(\overline{\Omega}); H_0^1(\Omega) \cap C^0(\overline{\Omega}))\big)$ is an isomorphism.

Proof. Since ϕ is convex in z, see (3.23), we note that

$$\phi(x,z) > \phi(x,0) + \frac{\partial}{\partial z}\phi(x,0)z \quad \text{for } x \in \Omega \text{ and } z > 0.$$

With the arguments developed in (3.22), we get that λ^* is bounded by μ_1.

The mapping $u : \lambda \in [0,\lambda^*) \to u(\lambda) \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ is monotone increasing, so the limits for fixed λ, $u(\lambda-)$ and $u(\lambda+)$ both exist, are positive, and are solution of (3.18). So

$$0 \leq u(\lambda-) \leq u(\lambda) \leq u(\lambda+).$$

Thus u is left continuous by the minimality of $u(\lambda)$, that is $u(\lambda-) = u(\lambda)$.

We introduce the operator $K(\lambda) : H_0^1(\Omega) \to H_0^1(\Omega)$ by

$$\text{for} \quad w \in H_0^1(\Omega), \quad K(\lambda)w = \lambda T Df(u(\lambda))w,$$

and the set

$$A = \{\lambda \in [0, \lambda^*); (K(\lambda)w, w)_{1,\Omega} < (w, w)_{1,\Omega} \quad \text{for all}\, w \in H_0^1(\Omega), w \neq 0\}$$

with $(w, v)_{1,\Omega} = \int_\Omega \mathbf{grad} w\, \mathbf{grad} v\, dx$. Since $Df(u(\lambda))$ is an increasing function of λ and

$$(K(\lambda)w, w)_{1,\Omega} = \lambda \int_\Omega Df(u(\lambda))w^2\, dx,$$

the set A is an interval containing zero. For $\lambda \in A$, $I - \lambda T Df(u(\lambda))$ is an isomorphism on $H_0^1(\Omega)$ (or $H_0^1(\Omega) \cap C^0(\overline{\Omega})$). By using the implicit function theorem, it is a simple matter to prove that $u : A \to H_0^1(\Omega) \cap C^0(\overline{\Omega})$ is a C^1 function. Suppose that $A \neq [0, \lambda^*)$ and let $\overline{\lambda} = \sup A$, $\overline{\lambda} < \lambda^*$. Since u is continuous on A and left continuous at $\overline{\lambda}$, we have $\lim_{\lambda \to \overline{\lambda}, \lambda < \overline{\lambda}} u(\lambda) = u(\overline{\lambda})$ and $\overline{\lambda} = \mu_1(\overline{\lambda})$ the principal characteristic value of $T Df(u(\overline{\lambda}))$. Since the function $u(.)$ is increasing, we have for $\lambda > \overline{\lambda}$

$$\text{for all}\, x \in \Omega \quad \phi(x, u(\lambda)(x)) > \phi(x, u(\overline{\lambda})(x)) + \frac{\partial}{\partial z}\phi(x, u(\overline{\lambda})(x))(u(\lambda)(x) - u(\overline{\lambda})(x)).$$

So we obtain

$$\begin{aligned} &-\Delta\big(u(\lambda) - u(\overline{\lambda})\big) - \overline{\lambda}\frac{\partial}{\partial z}\phi(x, u(\overline{\lambda}))\big(u(\lambda) - u(\overline{\lambda})\big) \\ &\qquad = (\lambda - \overline{\lambda})f(u(\lambda)) + \overline{\lambda}\big(f(u(\lambda)) - f(u(\overline{\lambda}))\big) - \overline{\lambda} Df(u(\overline{\lambda}))\big(u(\lambda) - u(\overline{\lambda})\big) > 0. \end{aligned}$$

We know that $u(\lambda) - u(\overline{\lambda})$ is positive and we apply Theorem 3.2 to get $\overline{\lambda} < \mu_1(\overline{\lambda})$ which is a contradiction. □

In Theorem 3.4, we prove that the minimal positive solutions of (3.18) form a regular solution path parametrized by λ, see Figure 3.1.

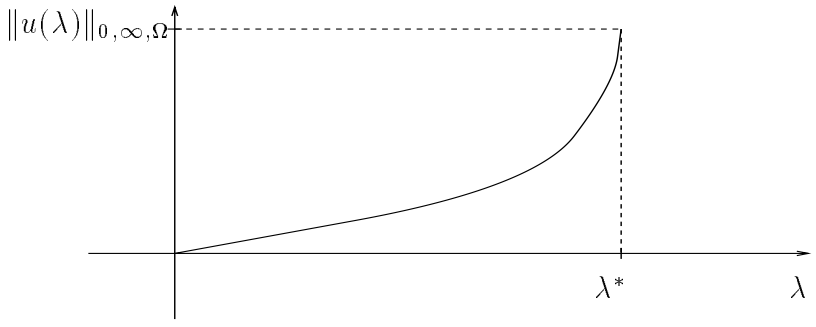

FIGURE 3.1.

The property $\mu_1(\lambda) > \lambda$ gives a stability result of the solution branch. We shall develop a more precise study in a left neighborhood of λ^* when illustrating the study of simple limit points.

Remark 3.2. If we consider the case

$$\phi(x,z) = z^2 + 1,$$

the hypotheses (3.15), (3.16), and (3.23) are satisfied. It is a simple matter to modify Theorem 3.4 when the hypothesis (3.17) is relaxed to

$$\frac{\partial \phi}{\partial z}(x,z) > 0 \quad \text{for} \quad x \in \Omega,\ z > 0.$$

We can apply the above theory and λ^* is bounded. Moreover one can prove that there exists $u^* \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ satisfying (3.18) with $\lambda = \lambda^*$, that is

$$u^* - \lambda^* T(u^{*2} + 1) = 0$$

and

$$\lim_{\lambda \to \lambda^*} u(\lambda) = u^*;$$

furthermore (λ^*, u^*) is a singular point of the solution branch, in the sense that $DF(u^*)$ is not an isomorphism, where the mapping $F(v) = v - \lambda^* T(v^2 + 1)$.

Suppose we can get a bound in the $L^2(\Omega)$ norm of the right-hand side, that is

$$\|u^2(\lambda) + 1\|_{0,\Omega} \le C \qquad \text{for all } \lambda \in [0, \lambda^*), \tag{3.24}$$

this implies that the increasing function $u^2(\lambda) + 1$ converges in $L^2(\Omega)$ as λ tends to λ^*, so $u(\lambda) = \lambda T(u^2(\lambda) + 1)$ converges to u^* in $H_0^1(\Omega) \cap C^0(\overline{\Omega})$. Clearly u^* is a solution of (3.18) with $\lambda = \lambda^*$, and the operator $I - 2\lambda^* T(u^*.)$ is singular. To prove (3.24) it is sufficient to prove

$$\|u(\lambda)\|_{0,4,\Omega} \le C \quad \text{for all } \lambda \in [0, \lambda^*) \tag{3.25}$$

since Ω is bounded.

For a given $\lambda \in [0, \lambda^*)$, we note $u = u(\lambda)$. We have

$$\text{for all } v \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dx = \lambda \int_\Omega (u^2 + 1) v\, dx$$

and

$$\text{for all } w \in H_0^1(\Omega) \quad 2\lambda \int_\Omega u w^2\, dx \le \int_\Omega |\mathbf{grad} w|^2\, dx.$$

We choose $w = u^{3/2}$ and $v = u^2$ to write successively

$$2\lambda \int_\Omega u^4\,dx \le \int_\Omega \frac{9}{4} u\,\mathbf{grad}u\,\mathbf{grad}u\,dx$$
$$= \frac{9}{8}\int_\Omega \mathbf{grad}u^2\,\mathbf{grad}u\,dx = \frac{9}{8}\lambda\int_\Omega (u^2+1)u^2\,dx$$

and so

$$\int_\Omega u^2\left(u^2 - \frac{9}{7}\right)\,dx \le 0. \tag{3.26}$$

Let us define the two sets

$$A = \{x\in\Omega; u(x) < \frac{3\sqrt{2}}{\sqrt{7}}\} \qquad \text{and} \qquad B = \{x\in\Omega; u(x) \ge \frac{3\sqrt{2}}{\sqrt{7}}\}.$$

Then for $x\in B$, $u^2(x) - 9/7 \ge u^2(x)/2$, and so we have with (3.26)

$$\int_A u^2(u^2 - \frac{9}{7})\,dx + \int_B \frac{u^4}{2}\,dx \le 0.$$

We are in position now to estimate

$$\int_\Omega u^4\,dx = \int_A u^4\,dx + \int_B u^4\,dx$$
$$\le \int_A \left[-u^4 + \frac{18}{7}u^2\right]\,dx \le \int_A u^2\left(\frac{18}{7} - u^2\right)\,dx \le 4\left(\frac{9}{7}\right)^2|\Omega|. \quad \square$$

Remark 3.3. If we consider the case

$$\phi(x,z) = e^z,$$

the hypotheses (3.15), (3.16), (3.17), and (3.23) are satisfied and we can apply the above theory. Moreover one can prove that there exists $u^*\in H_0^1(\Omega)\cap C^0(\overline{\Omega})$ satisfying (3.18), that is

$$u^* - \lambda^* T e^{u^*} = 0, \qquad \lim_{\lambda\to\lambda^*} u(\lambda) = u^*;$$

furthermore (λ^*, u^*) is a singular point of the solution branch. The proof of this result is analogous to the one developed in Remark 3.2, see Crouzeix and Rappaz [1989] for details or Crandall and Rabinowitz [1973]. $\square$

Remark 3.4. We consider both cases when $\phi(x,z) = z^2+1$ or $\phi(x,z) = e^z$. Then we know that the operator $I - \lambda^* TDf(u^*)$ is singular with a kernel of dimension 1, see Theorem 3.2.

In fact since T is compact, $I - \lambda^* TDf(u^*)$ is a self-adjoint Fredholm operator with index zero in $H_0^1(\Omega)$ and if $\varphi^* \in H_0^1(\Omega)$, $\varphi^* \neq 0$ is such that

$$\mathrm{Ker}\,(I - \lambda^* TDf(u^*)) = \mathrm{span}\{\varphi^*\}$$

then the range of $(I - \lambda^* TDf(u^*))$ is orthogonal to $\mathrm{span}\{\varphi^*\}$ in $H_0^1(\Omega)$ with respect to the scalar product $(.,.)_{1,\Omega}$ in $H_0^1(\Omega)$.

We define now the mapping $\mathscr{F} : (s, \lambda, u) \in \mathbb{R} \times \mathbb{R} \times H_0^1(\Omega) \cap C^0(\overline{\Omega}) \to \mathscr{F}(s, \lambda, u) \in \mathbb{R} \times H_0^1(\Omega) \cap C^0(\overline{\Omega})$ by the relation

$$\mathscr{F}(s, \lambda, u) = \begin{pmatrix} s - (u - u^*, \varphi^*)_{1,\Omega} \\ u - \lambda T f(u) \end{pmatrix}.$$

We have $\mathscr{F}(0, \lambda^*, u^*) = 0$ and since $Tf(u^*)$ is a positive function and φ^* has a sign we can check that $D^2_{\lambda u}\mathscr{F}(0, \lambda^*, u^*)$ is an isomorphim on $\mathbb{R} \times H_0^1(\Omega) \cap C^0(\overline{\Omega})$. As a consequence of the implicit function theorem, we can parametrize λ and u by s. There exists a mapping $(\lambda, u) : s \in (-\epsilon, \epsilon) \to (\lambda(s), u(s)) \in \mathbb{R} \times H_0^1(\Omega) \cap C^0(\overline{\Omega})$ satisfying

$$\lambda(0) = \lambda^*, \qquad u(0) = u^*, \qquad u(s) \neq u^* \quad \text{when} \quad s \neq 0,$$

and

$$u(s) - \lambda(s) T f(u(s)) = 0 \quad \text{for all}\, s \in (-\epsilon, \epsilon).$$

So we conclude that (λ^*, u^*) can be considered as a regular point of the mapping $\mathscr{F}$. A similar study is presented in a general setting in Section 15. □

Remark 3.5. There is an extensive literature concerning positive solutions of problems of the type (3.18). We mention the work of Keller and Cohen [1967] where positive minimal solutions are studied with the monotonous iteration principle. In Crandall and Rabinowitz [1973], the problem is studied by continuation and variational methods, with emphasis on the turning point λ^*. We can also refer to the review works of Amann [1976], Mignot and Puel [1980], and Lions [1982]. □

We consider now finite element approximations of problem (3.18). We introduce first a family of isoparametric triangulations with polynomials of degree 1 or 2, where $h = \max_{T \in \mathscr{T}_h} h_T$ (h_T is the diameter of T) and we denote by Ω_h the interior of $\overline{\Omega}_h = \cup_{T \in \mathscr{T}_h} T$, see Ciarlet [1978] p.224 or Ciarlet [1991] p.230 for further details.

Actually each element $T \in \mathscr{T}_h$ is the image $F_T(\hat{T})$, where $(\hat{T}, \hat{\mathscr{P}}_k, \hat{\Sigma}_k)$ is the Lagrange reference element of type (k) ($\hat{\mathscr{P}}_k$ is the space of polynomials of degree $\leq k$ on $\hat{T}$), and where the invertible mapping $F_T : \hat{T} \to T$ has its components to be a polynomial of degree 1 or 2 and

$$F_T(\hat{a}_i) = a_i \quad \text{for all nodes}\, \hat{a}_i \in \hat{\Sigma}_k,$$

see Figure 3.2 below (for $k = 2$).

The approximation space is given by

$$V_h = \{v \in C^0(\mathbb{R}^2); v(x) = 0 \text{ for all}\, x \notin \Omega_h, v_{|T} \in P_T \text{ for all}\, T \in \mathscr{T}_h\}, \tag{3.27}$$

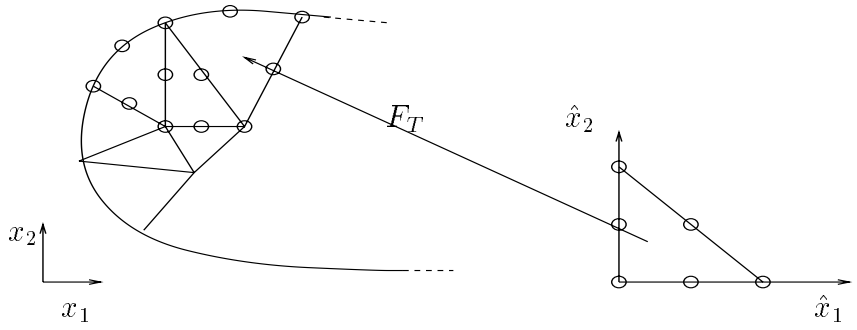

FIGURE 3.2.

with $P_T = \{p : T \to \mathbb{R}; p \circ F_T \in \hat{\mathscr{P}}_k\}$. The corresponding finite element approximation to (3.18) reads for all $v_h \in V_h$

$$a_h(u_h, v_h) = \int_{\Omega \cap \Omega_h} \lambda f(u_h) v_h \, dx, \tag{3.28}$$

where

$$a_h(u_h, v_h) = \int_{\Omega_h} \mathbf{grad} u_h \, \mathbf{grad} v_h \, dx.$$

Generally problem (3.28) is not solved as it is, but with numerical quadrature rules, in which case we suppose that for $v_h \in V_h$, $f(v_h)$ is continuous. In the case $k = 1$, the elements $F_T(\hat{T})$ are triangles and we consider the problem for all $v_h \in V_h$

$$a_h(u_h, v_h) = \sum_{T \in \mathscr{T}_h} \left[\frac{S_T}{3} \sum_{i=1}^{3} \big(f(u_h) v_h\big)(a_{i,T}) \right] \tag{3.29}$$

where S_T is the measure of T and $a_{1,T}$, $a_{2,T}$, $a_{3,T}$ the vertices of $T \in \mathscr{T}_h$.

To study the case $k = 2$, we introduce for u, $v \in V_h$ and $T \in \mathscr{T}_h$

$$a_T(u, v) = \sum_{i=1}^{3} \frac{1}{6} \det(DF_T(\hat{m}_i)) \mathbf{grad} u(m_{i,T}) \, \mathbf{grad} v(m_{i,T});$$

here $\hat{m}_i, i = 1, 2, 3$, are the middles of the sides of $\hat{T}$ and $m_{i,T}$ the images $F_T(\hat{m}_i)$. Then problem (3.28) with numerical quadrature rules reads

$$\text{for all } v_h \in V_h \quad \sum_{T \in \mathscr{T}_h} a_T(u_h, v_h) = \sum_{T \in \mathscr{T}_h} \sum_{i=1}^{3} \frac{1}{6} \det(DF_T(\hat{m}_i)) \big(f(u_h) v_h\big)(m_{i,T}). \tag{3.30}$$

As an illustration of our abstract theory developed in the next chapter, we shall analyze the problems (3.28), (3.29), and (3.30).

4. A discretization of the Navier–Stokes equations

Let Ω be a polygonal open domain in $\mathbb{R}^2$ with boundary $\partial\Omega$. The stationary incompressible Navier–Stokes problem consists in finding a velocity vector $\mathbf{u} : \Omega \to \mathbb{R}^2$ and a pressure $p : \Omega \to \mathbb{R}$ satisfying

$$-\nu\Delta\mathbf{u} + (\mathbf{u}\boldsymbol{\nabla})\mathbf{u} + \mathbf{grad}p = \mathbf{f} \quad \text{in} \quad \Omega, \tag{4.1}$$

$$\mathrm{div}\mathbf{u} = 0 \quad \text{in} \quad \Omega, \tag{4.2}$$

$$\mathbf{u} = 0 \quad \text{on} \quad \partial\Omega, \tag{4.3}$$

where the positive number ν and the function $\mathbf{f} \in (L^2(\Omega))^2$ are given. We introduce the spaces $L^2_0(\Omega) = \{f \in L^2(\Omega); \int_\Omega f\,dx = 0\}$ and $X = (H^1_0(\Omega))^2 \times L^2_0(\Omega)$. A weak formulation of the problem reads: find $(\mathbf{u}, p) \in X$ such that

$$\text{for all}\,\mathbf{v} \in (H^1_0(\Omega))^2 \quad \nu\int_\Omega \mathbf{gradu}\,\mathbf{gradv}\,dx + \int_\Omega (\mathbf{u}\,\boldsymbol{\nabla})\mathbf{u}\,\mathbf{v}\,dx - \int_\Omega p\,\mathrm{div}\mathbf{v}\,dx = \int_\Omega \mathbf{f}\mathbf{v}\,dx, \tag{4.4}$$

$$\text{for all}\,q \in L^2_0(\Omega) \quad \int_\Omega q\,\mathrm{div}\mathbf{u}\,dx = 0. \tag{4.5}$$

Note that if $\mathbf{u} \in (H^1_0(\Omega))^2$, then $\mathbf{u} \in (L^p(\Omega))^2$ for all $p \in [1, \infty[$ and $(\mathbf{u}\,\boldsymbol{\nabla})\mathbf{u}$ is in $(L^q(\Omega))^2$ for all $q \in [1, 2[$. Since $\mathbf{v} \in (H^1_0(\Omega))^2$, the term $(\mathbf{u}\,\boldsymbol{\nabla})\mathbf{u}\,\mathbf{v}$ is in $L^1(\Omega)$.

We introduce the multilinear forms $a : (H^1_0(\Omega))^2 \times (H^1_0(\Omega))^2 \to \mathbb{R}$ given by

$$a(\mathbf{u}, \mathbf{v}) = \nu\int_\Omega \mathbf{gradu}\,\mathbf{gradv}\,dx,$$

$c : (H^1_0(\Omega))^2 \times (H^1_0(\Omega))^2 \times (H^1_0(\Omega))^2 \to \mathbb{R}$ given by

$$c(\mathbf{w}; \mathbf{u}, \mathbf{v}) = \int_\Omega (\mathbf{w}\,\boldsymbol{\nabla})\mathbf{u}\,\mathbf{v}\,dx,$$

and $b : (H^1_0(\Omega))^2 \times L^2_0(\Omega) \to \mathbb{R}$ given by

$$b(\mathbf{u}, p) = -\int_\Omega p\,\mathrm{div}\mathbf{u}\,dx.$$

Then the problem (4.4)–(4.5) is equivalent to find $(\mathbf{u}, p) \in X$ such that

$$\text{for all}\,(\mathbf{v}, q) \in X \quad a(\mathbf{u}, \mathbf{v}) + c(\mathbf{u}; \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) - \int_\Omega \mathbf{f}\mathbf{v}\,dx + b(\mathbf{u}, q) = 0. \tag{4.6}$$

The Navier-Stokes problem (4.6) has received a lot of attention and has been widely studied. It is well-known that it has at least one solution. See Temam [1977], Girault and Raviart [1986] and the bibliography therein for further references.

Let $(\mathbf{u},p)\in X$ be a solution of (4.6); we assume that $\mathbf{u}\in(H^2(\Omega))^2$. We also want to assume that the solution $(\mathbf{u},p)\in X$ is regular, in the sense that there exists $\epsilon>0$ such that the only solution to (4.6) in the ball $B((\mathbf{u},p),\epsilon)$ is $(\mathbf{u},p)$. Recall that on X we introduce the norm $\|(\mathbf{w},r)\|_X=|\mathbf{w}|_{1,\Omega}+\|r\|_{0,\Omega}$. Referring to Theorem 2.1 the hypothesis is stated precisely in the following terms. The bilinear form $A:X\times X\to\mathbb{R}$ given by

$$\begin{aligned}A(\underline{w},\underline{v}) &= A\big((\mathbf{w},r),(\mathbf{v},q)\big)\\ &= a(\mathbf{w},\mathbf{v})+b(\mathbf{w},q)+b(\mathbf{v},r)+c(\mathbf{u};\mathbf{w},\mathbf{v})+c(\mathbf{w};\mathbf{u},\mathbf{v})\end{aligned}\tag{4.7}$$

satisfies the so called inf-sup condition, see Babuška and Aziz [1972], that is

$$\begin{aligned}&\inf_{\substack{\underline{w}\in X\\ \|\underline{w}\|_X=1}}\ \sup_{\substack{\underline{v}\in X\\ \|\underline{v}\|_X=1}} A(\underline{w},\underline{v})>0,\\ &\qquad\sup_{\substack{\underline{w}\in X\\ \|\underline{w}\|_X=1}} A(\underline{w},\underline{v})>0\quad\text{for all}\,\underline{v}\in X,\quad \underline{v}\neq 0.\end{aligned}\tag{4.8}$$

For simplicity we have assumed that Ω is polygonal, so we can consider $\mathscr{T}_h^c$ a regular triangulation of $\overline{\Omega}$ made up of triangles T of diameter inferior to h and $\mathscr{T}_h^f$ a finer triangulation obtained from $\mathscr{T}_h^c$ by dividing each triangle $T\in\mathscr{T}_h^c$ in four triangles as shown in Figure 4.1.

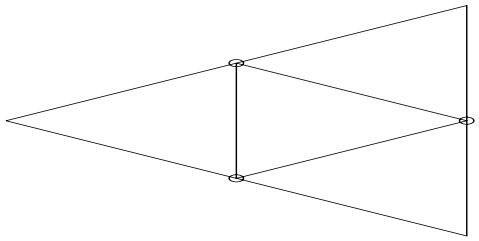

Figure 4.1.

The parameter h will tend to zero. We assume that the family $\{\mathscr{T}_h^c\}_{0<h\leq 1}$ is regular, see Ciarlet [1978] p.124. To define the approximation spaces X_h to X we introduce the finite-dimensional spaces

$$V_h=\{v\in C^0(\overline{\Omega});v_{|T'}\in\mathscr{P}_1(T')\ \text{for all}\,T'\in\mathscr{T}_h^f,\ v_{|\partial\Omega}=0\}$$

and

$$W_h=\{w\in C^0(\overline{\Omega});w_{|T}\in\mathscr{P}_1(T)\ \text{for all}\,T\in\mathscr{T}_h^c,\ \int_\Omega w\,dx=0\},$$

where $\mathscr{P}_1(T)$ is the space of polynomials of degree up to 1 on T. The velocity approximation space is then $V_h\times V_h$ while the pressure aprroximation space is W_h, so we set $X_h=V_h\times V_h\times W_h$. This is a variant of the Hood-Taylor method. A finite element

approximation of (4.6), which is called the $(4\mathscr{P}_1 - \mathscr{P}_1)$ finite element approximation, consists in finding $(\mathbf{u}_h, p_h) \in X_h$ such that

$$\text{for all } (\mathbf{v}_h, q_h) \in X_h \quad a(\mathbf{u}_h, \mathbf{v}_h) + c(\mathbf{u}_h; \mathbf{u}_h, \mathbf{v}_h) + b(\mathbf{v}_h, p_h) - \int_\Omega \mathbf{f}\mathbf{v}_h \, dx + b(\mathbf{u}_h, q_h) = 0. \tag{4.9}$$

On $\mathscr{T}_h^c$ we make the following hypothesis: there exists a set of interior nodes in Ω, P_1, $\ldots$, P_ℓ, such that the finite element stars S_1, $\ldots$, S_ℓ

$$S_i = \bigcup_{T \ni P_i} T$$

have the property

$$\text{for all } i \neq j, \quad \overset{\circ}{S}_i \cap \overset{\circ}{S}_j = \emptyset, \qquad \overline{\Omega} = \bigcup_{i=1}^{\ell} S_i, \tag{4.10}$$

where $\overset{\circ}{S}_i$ denotes the interior of S_i. Note that we always can realize such a triangulation only to refine some elements. The study of the $(4\mathscr{P}_1 - \mathscr{P}_1)$ finite element for the Stokes problem is well known, see GIRAULT and RAVIART [1986] p.181 for instance.

From the theory developed in the next chapter, we will be able to analyze the approximated Navier-Stokes problem (4.9) and get a priori and a posteriori error estimates. It is then essential to get the discrete analogue of the hypothesis (4.8). From the study of the $(4\mathscr{P}_1 - \mathscr{P}_1)$ finite element for the Stokes problem, we know that under the hypothesis (4.10) we have

$$\inf_{\substack{\underline{w}_h \in X_h \\ \|\underline{w}_h\|_X = 1}} \sup_{\substack{\underline{v}_h \in X_h \\ \|\underline{v}_h\|_X = 1}} S(\underline{w}_h, \underline{v}_h) \geq \beta > 0, \tag{4.11}$$

where the bilinear form corresponding to the Stokes problem $S : (H_0^1(\Omega))^2 \times L_0^2(\Omega) \to \mathbb{R}$ is given by

$$S(\underline{w}, \underline{v}) = S\big((\mathbf{w}, r), (\mathbf{v}, q)\big) = a(\mathbf{w}, \mathbf{v}) + b(\mathbf{w}, q) + b(\mathbf{v}, r). \tag{4.12}$$

Theorem 4.1. *Let $(\mathbf{u}, p) \in X$ be a solution of (4.6) with $\mathbf{u} \in (H^2(\Omega))^2$, such that (4.8) is valid. Then under the assumptions (4.10) there exists $\delta > 0$ such that*

$$\inf_{\substack{\underline{w}_h \in X_h \\ \|\underline{w}_h\|_X = 1}} \sup_{\substack{\underline{v}_h \in X_h \\ \|\underline{v}_h\|_X = 1}} A(\underline{w}_h, \underline{v}_h) \geq \delta > 0. \tag{4.13}$$

Proof. We define the linear operator $\mathscr{A} : X' \to X$ by: for $\underline{f} \in X'$, $\mathscr{A}\underline{f} \in X$ is the unique solution to

$$\text{for all } \underline{v} \in X \quad A(\mathscr{A}\underline{f}, \underline{v}) = \langle \underline{f}, \underline{v} \rangle_{X'X},$$

where $\langle .\,,.\rangle_{X'X}$ is the duality pairing between X' and X. From the hypothesis (4.8), we know that $\mathscr{A}$ is an isomorphism from X' onto X. In the same way from the Stokes form (4.12), we define the linear operator $\mathscr{S} : X' \to X$ by: for $\underline{f} \in X'$, $\mathscr{S}\underline{f} \in X$ is the unique solution to

$$\text{for all}\,\underline{v} \in X \quad S(\mathscr{S}\underline{f}, \underline{v}) = \langle \underline{f}, \underline{v}\rangle_{X'X},$$

which is an isomorphism. We also define the linear operator $N : X \to X'$ by for all $\underline{w} = (\mathbf{w}, r) \in X$, for all $\underline{v} = (\mathbf{v}, q) \in X$

$$\langle N\underline{w}, \underline{v}\rangle_{X'X} = \int_\Omega \left[(\mathbf{u}\boldsymbol{\nabla})\mathbf{w} + (\mathbf{w}\boldsymbol{\nabla})\mathbf{u}\right]\mathbf{v}\,dx.$$

For $\underline{f} \in X'$ we have the relation for all $\underline{v} \in X$

$$A(\mathscr{A}\underline{f}, \underline{v}) = S(\mathscr{A}\underline{f}, \underline{v}) + \langle N(\mathscr{A}\underline{f}), \underline{v}\rangle_{X'X} = \langle \underline{f}, \underline{v}\rangle_{X'X}$$

or

$$\langle \big(\mathscr{S}^{-1} + N\big)(\mathscr{A}\underline{f}), \underline{v}\rangle_{X'X} = \langle \underline{f}, \underline{v}\rangle_{X'X}. \tag{4.14}$$

So we have proved that

$$\mathscr{S}^{-1} + N = \mathscr{A}^{-1} \quad \text{or} \quad I + \mathscr{S}N = \mathscr{S}\mathscr{A}^{-1} \tag{4.15}$$

where I is the identity in X.

The discrete analogue of $\mathscr{S}$ is denoted by $\mathscr{S}_h : X' \to X_h$ and is given by: for $\underline{f} \in X'$

$$\text{for all}\,\underline{v}_h \in X_h \quad S(\mathscr{S}_h\underline{f}, \underline{v}_h) = \langle \underline{f}, \underline{v}_h\rangle_{X'X}.$$

As a consequence of the hypothesis (4.10), we get

$$\lim_{h\to 0} \|\mathscr{S}_h\underline{f} - \mathscr{S}\underline{f}\|_X = 0 \quad \text{for all}\,\underline{f} \in X'. \tag{4.16}$$

The convergence property (4.16) and the compactness of $N : X \to X'$ imply

$$\lim_{h\to 0} \|\mathscr{S}N - \mathscr{S}_h N\|_{X;X} = 0, \tag{4.17}$$

see (1.3).

Writing

$$I + \mathscr{S}_h N = (I + \mathscr{S}N)\left[I + (I + \mathscr{S}N)^{-1}(\mathscr{S}_h N - \mathscr{S}N)\right]$$

we deduce from (4.15), (4.17), and (1.2) that there exists an $h_0 > 0$ such that $I + \mathscr{S}_h N$ is an isomorphism from X onto X for $h \le h_0$ with a uniformly bounded inverse.

Let us prove now the discrete analogue of (4.15) or in other words

$$(I + \mathscr{S}_h N)_{|X_h} = \mathscr{S}_h \mathscr{A}^{-1}{}_{|X_h}. \tag{4.18}$$

We deduce from the relation (4.14) that for $\underline{w} \in X$ and $\underline{v}_h \in X_h$

$$S\left(\left[\mathscr{S}_h \mathscr{A}^{-1} - (I + \mathscr{S}_h N)\right]\underline{w}, \underline{v}_h\right) = 0.$$

In particular with the inf-sup relation (4.11) we get

$$\text{for all}\, \underline{w}_h \in X_h \quad \left[\mathscr{S}_h \mathscr{A}^{-1} - (I + \mathscr{S}_h N)\right]\underline{w}_h = 0.$$

Now since $I + \mathscr{S}_h N$ is an isomorphism from X onto X with a uniformly bounded inverse, we have

$$\text{for all}\, \underline{w}_h \in X_h \quad \|\mathscr{S}_h \mathscr{A}^{-1} \underline{w}_h\|_X = \|(I + \mathscr{S}_h N)\underline{w}_h\|_X \geq \gamma \|\underline{w}_h\|_X, \tag{4.19}$$

with $\gamma > 0$, γ independent on h.

We are in a position now to prove (4.13). Let $\underline{w}_h \in X_h$ be such that $\|\underline{w}_h\|_X = 1$. Then by using (4.11) and (4.19) we get

$$\sup_{\substack{\underline{v}_h \in X_h \\ \|\underline{v}_h\|_X = 1}} A(\underline{w}_h, \underline{v}_h) = \sup_{\substack{\underline{v}_h \in X_h \\ \|\underline{v}_h\|_X = 1}} \langle \mathscr{A}^{-1} \underline{w}_h, \underline{v}_h \rangle_{X'X} =$$

$$\sup_{\substack{\underline{v}_h \in X_h \\ \|\underline{v}_h\|_X = 1}} S(\mathscr{S}_h \mathscr{A}^{-1} \underline{w}_h, \underline{v}_h) \geq \beta \|\mathscr{S}_h \mathscr{A}^{-1} \underline{w}_h\|_X \geq \beta\gamma \|\underline{w}_h\|_X = \beta\gamma,$$

which proves (4.13) with $\delta = \beta\gamma$. □

Remark 4.1. We could have formulated an abstract version of the above theorem. The generalization with X a reflexive Banach space, bilinear forms $A(.,.)$ and $S(.,.)$, a compact operator $N : X \to X'$, and $\{X_h\}_{0<h\leq 1}$ a family of finite-dimensional subspaces in X satisfying

$$\lim_{h \to 0} \inf_{u_h \in X_h} \|u - u_h\|_X = 0 \quad \text{for all}\, u \in X,$$

is straighforward. □

In the next chapter we present how to use the above result to analyze the $(4\mathscr{P}_1 - \mathscr{P}_1)$ approximation of the Navier-Stokes problem.

5. A stationary heat problem with convection

The last example we develop is a convection diffusion type problem. Let $\Omega \subset \mathbb{R}^2$ be a bounded domain with a regular boundary $\partial\Omega$. Given the functions $f \in L^2(\Omega)$, $k \in L^\infty(\mathbb{R})$ and the vector field $\mathbf{c} \in (L^\infty(\Omega))^2$, we shall consider the problem to find $u \in H_0^1(\Omega)$ satisfying

$$-\text{div}(k(u)\mathbf{grad}u) + \mathbf{c}\,\mathbf{grad}u = f \quad \text{in} \quad \Omega. \tag{5.1}$$

To study it we introduce some further assumptions on the data. The function $k \in C^2(\mathbb{R})$ satisfies the hypotheses

$$\text{for all } s \in \mathbb{R} \quad k(s) \geq \alpha > 0, \tag{5.2}$$

$$\text{for all } s \in \mathbb{R} \quad |D^\ell k(s)| \leq \gamma_\ell, \quad \ell = 0, 1, 2, \tag{5.3}$$

with positive real numbers α, γ_0, γ_1, and γ_2; here $D^\ell k(s)$ stands for the ℓth derivative of k. The vector field $\mathbf{c}$ satisfies

$$\mathbf{c} \in (W^{1,\infty}(\Omega))^2 \qquad \text{and} \qquad \mathrm{div}\mathbf{c} = 0. \tag{5.4}$$

In the approximation schemes, we shall keep the nonlinear term in the principal part of (5.1), although to analyze the exact problem, we reduce it to a semilinear problem. Under the assumption (5.2) the primitive K,

$$\text{for all } s \in \mathbb{R} \quad K(s) = \int_0^s k(t)\, dt,$$

of k is a strictly increasing function. Let G be the inverse function of K and g be the derivative of G. Then (5.1) can be written in the form

$$-\Delta K(u) + \mathbf{cgrad}u = f \quad \text{in} \quad \Omega$$

or when setting $U = K(u) \in H_0^1(\Omega)$

$$-\Delta U + g(U)\mathbf{cgrad}U = f \quad \text{in} \quad \Omega, \tag{5.5}$$

and the weak form reads: find $U \in H_0^1(\Omega)$ such that

$$\text{for all } \varphi \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad}U\, \mathbf{grad}\varphi\, dx + \int_\Omega g(U)\mathbf{c}\, \mathbf{grad}U\varphi\, dx = \int_\Omega f\varphi\, dx. \tag{5.6}$$

Theorem 5.1. *In addition to the hypotheses (5.2), (5.3), and (5.4) we assume that $f \in L^\infty(\Omega)$. Then problem (5.6) has a unique solution U; moreover the solution U is in $W^{2,p}(\Omega)$ for all $1 \leq p < \infty$.*

Proof. To prove the existence of a solution we introduce first the operator $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$, the inverse of the minus Laplacian operator with homogeneous boundary conditions defined in (3.4), and write (5.6) in the equivalent way

$$U - T(f - g(U)\mathbf{cgrad}U) = 0. \tag{5.7}$$

The mapping $U \in H_0^1(\Omega) \to f - g(U)\mathbf{cgrad}U \in H^{-1}(\Omega)$ is continuous and compact. To apply the Leray-Schauder homotopy theorem, see LERAY and SCHAUDER [1934], we derive a a priori estimate for $V \in H_0^1(\Omega)$ with

$$V - \lambda T(f - g(V)\mathbf{cgrad}V) = 0,$$

which is independent of $\lambda \in [0,1]$. For $\varphi \in H_0^1(\Omega)$ we have

$$\int_\Omega \mathbf{grad} V \, \mathbf{grad}\varphi \, dx - \lambda \int_\Omega (f - g(V)\mathbf{cgrad} V)\, \varphi \, dx = 0$$

or with $\varphi = V$ and $\mathscr{G}(s) = \int_0^s g(t)t \, dt$

$$\int_\Omega |\mathbf{grad} V|^2 \, dx - \lambda \int_\Omega fV \, dx + \lambda \int_\Omega \mathbf{cgrad}\mathscr{G}(V)\, dx = 0.$$

With the hypothesis (5.4) we check that $\mathbf{cgrad}\mathscr{G}(V) = \operatorname{div}(\mathbf{c}\mathscr{G}(V))$ and since $\mathscr{G}(V)$ is in $H_0^1(\Omega)$ we get

$$\int_\Omega |\mathbf{grad} V|^2 \, dx = \lambda \int_\Omega fV \, dx$$

and so

$$|V|_{1,\Omega}^2 \le \|f\|_{0,\Omega}\|V\|_{0,\Omega} \le C\|f\|_{0,\Omega}|V|_{1,\Omega}.$$

Thus we conclude to the existence of a solution to (5.7).

Considering the relation (5.7) and the regularity of the domain, we get from the classical elliptic regularity result that the function U is in $H^2(\Omega)$, from which it follows that $f - g(U)\mathbf{cgrad} U \in L^p(\Omega)$, $1 \le p < \infty$, and then $U \in W^{2,p}(\Omega)$.

Let us finally check the uniqueness result. Suppose that U and V are two solutions of (5.5). Then

$$-\Delta(U - V) + g(U)\mathbf{cgrad} U - g(V)\mathbf{cgrad} V = 0 \quad \text{in} \quad \Omega. \tag{5.8}$$

Since G is the primitive of g with $G(0) = 0$ and since $\operatorname{div}\mathbf{c} = 0$, the relation (5.8) can be written in the form

$$-\Delta(U - V) + \operatorname{div}(\mathbf{c}(G(U) - G(V))) = 0 \quad \text{in} \quad \Omega \tag{5.9}$$

or with $W = U - V$

$$-\Delta W + \operatorname{div}\left(\mathbf{c}W \int_0^1 g(sU + (1-s)V)\, ds\right) = 0 \quad \text{in} \quad \Omega. \tag{5.10}$$

Our goal is to prove that $W \equiv 0$. The weak form of (5.10) reads with $\zeta = \int_0^1 g(sU + (1-s)V)\, ds$

$$\text{for all}\, \varphi \in H_0^1(\Omega) \quad \int_\Omega \mathbf{grad} W \, \mathbf{grad}\varphi \, dx - \int_\Omega \zeta W \mathbf{c}\, \mathbf{grad}\varphi \, dx = 0.$$

Given $\epsilon > 0$, we choose $\varphi = W^+/(W^+ + \epsilon)$ where W^+ stands for the positive part of W. Then we have since $\|\zeta\|_{0,\infty} \leq 1/\alpha$,

$$\int_\Omega |\mathbf{grad}\log(1+\frac{W^+}{\epsilon})|^2\,dx = \int_\Omega \frac{1}{(\epsilon+W^+)^2}|\mathbf{grad}W^+|^2\,dx$$
$$= \int_\Omega \frac{1}{\epsilon}\mathbf{grad}W\,\mathbf{grad}\frac{W^+}{\epsilon+W^+}\,dx = \int_\Omega \frac{1}{\epsilon}W^+\zeta\mathbf{c}\,\mathbf{grad}\frac{W^+}{\epsilon+W^+}\,dx$$
$$= \int_\Omega \frac{W^+}{\epsilon+W^+}\zeta\mathbf{c}\,\mathbf{grad}\log(1+\frac{W^+}{\epsilon})\,dx \leq \frac{\|\mathbf{c}\|_{0,\infty}}{\alpha}\int_\Omega |\mathbf{grad}\log(1+\frac{W^+}{\epsilon})|\,dx.$$

So with the Cauchy-Schwarz inequality we get

$$\int_\Omega |\mathbf{grad}\log(1+\frac{W^+}{\epsilon})|^2\,dx \leq C.$$

Since $\log(1+W^+/\epsilon)$ is in $H_0^1(\Omega)$, it follows from the Poincaré inequality that

$$\int_\Omega |\log(1+\frac{W^+}{\epsilon})|^2\,dx \leq C$$

with C independent of ϵ. Letting ϵ tend to 0, we get that W^+ must vanish in Ω. In the same way we prove that W^- must vanish in Ω. □

In effective computations problem (5.1) is most often discretized as it is, but not problem (5.6). For that reason we shall focus on the approximation of (5.1). The mapping $F : H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by: for $v \in H_0^1(\Omega)$

$$F(v) = -\mathrm{div}\,(k(v)\mathbf{grad}v) + \mathbf{c}\mathbf{grad}v - f.$$

Let U be the unique solution of (5.6) given in Theorem 5.1. Then $u \equiv G(U)$ is the unique zero of F,

$$F(u) = 0, \tag{5.11}$$

and $u \in W^{2,p}(\Omega)$, $1 \leq p < \infty$.

Generally the mapping F is not C^1, which would be a sufficient condition to apply the approximation theory developed in the next section. In fact the appropriate spaces to work with are $W^{1,p}(\Omega)$ with $2 < p < \infty$, see Pousin and Rappaz [1992]; the mapping F is then considered as $F : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ and has desired differentiability properties.

Theorem 5.2. *Under the hypotheses of Theorem 5.1, the mapping $F : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ is of class C^1 when $2 < p < \infty$ and moreover the derivative $DF(u) \in \mathscr{L}(W_0^{1,p}(\Omega); W^{-1,p}(\Omega))$ at the solution u is an isomorphism.*

Proof. Let q be the conjugate of p, $\frac{1}{p} + \frac{1}{q} = 1$. Then for all $v \in W_0^{1,p}(\Omega)$ and $\psi \in W_0^{1,q}(\Omega)$

$$\langle F(v), \psi\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = \int_\Omega k(v)\mathbf{grad}v\,\mathbf{grad}\psi\,dx + \int_\Omega \mathbf{c}\mathbf{grad}v\psi\,dx - \int_\Omega f\psi\,dx,$$

$\langle .\,. \rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)}$ denoting the duality bracket. Then for all v, w in $W_0^{1,p}(\Omega)$, ψ in $W_0^{1,q}(\Omega)$

$$\langle DF(v)w, \psi \rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = \int_\Omega \mathbf{grad}(k(v)w)\,\mathbf{grad}\psi\, dx + \int_\Omega \mathbf{cgrad} w\, \psi\, dx \tag{5.12}$$

and with the regularity assumption on k in (5.2), (5.3), we easily verify that F is C^1.

It is known that the Laplacian operator with homogeneous boundary conditions is an isomorphism from $W_0^{1,p}(\Omega)$ onto $W^{-1,p}(\Omega)$, see for instance in SIMADER [1972] or in DAUGE [1992] or the regularity result in DAUTRAY and LIONS [1987] p.538 and the characterization of $W^{-1,p}(\Omega)$ given in GRISVARD [1985], p.17. Relation (5.12) with $v = u$ also reads

$$TDF(u)w = k(u)w + T(\mathbf{cgrad} w) \tag{5.13}$$

where T is the inverse of the minus Laplacian operator. We introduce the mapping $R : \varphi \in W_0^{1,p}(\Omega) \to R\varphi = \varphi/k(u) \in W_0^{1,p}(\Omega)$, which is an isomorphism, in (5.13) with $w = R\varphi$ to get

$$TDF(u)R\varphi = \varphi + T(\mathbf{cgrad}(R\varphi)). \tag{5.14}$$

Since the mapping $T \in \mathscr{L}(L^p(\Omega); W_0^{1,p}(\Omega))$ is compact, so the mapping $TDF(u)R \in \mathscr{L}(W_0^{1,p}(\Omega); W_0^{1,p}(\Omega))$ is a Fredholm operator with index 0. To prove the theorem it is then sufficient to prove that $DF(u) \in \mathscr{L}(W_0^{1,p}(\Omega); W^{-1,p}(\Omega))$ is injective.

Let $\varphi \in \mathrm{Ker}(DF(u))$, then $\varphi \in W_0^{1,p}(\Omega)$ satisfies

$$-\Delta(k(u)\varphi) + \mathbf{cgrad}\varphi = 0 \quad \text{in} \quad \Omega. \tag{5.15}$$

With the notations $\psi = k(u)\varphi$, $\boldsymbol{\alpha} = \mathbf{c}/k(u)$, (5.15) takes the form

$$-\Delta\psi + \mathrm{div}(\boldsymbol{\alpha}\psi) = 0 \quad \text{in} \quad \Omega. \tag{5.16}$$

From the relations (5.15) and (5.16) we can conclude that $\psi \equiv \varphi \equiv 0$ following the proof of uniqueness in Theorem 5.1. □

We introduce now a finite element approximation of (5.11). For the sake of simplicity we suppose that Ω is a polygonal domain in $\mathbb{R}^2$. Given a regular family $\{\mathscr{T}_h\}_{0<h\leq 1}$ of triangulations of Ω, we associate to $\mathscr{T}_h$ the finite element space

$$V_h = \{\varphi \in C^0(\overline{\Omega}); \varphi_{|T} \in \mathscr{P}_1 \quad \text{for all} \, T \in \mathscr{T}_h \quad \text{and} \quad \varphi_{|\partial\Omega} = 0\},$$

where $\mathscr{P}_1$ denotes the space of polynomials of degree up to 1. A finite element approximation of (5.11) consists in finding $u_h \in V_h$ satisfying

$$\text{for all} \, v_h \in V_h \quad \int_\Omega k(u_h)\mathbf{grad}u_h\,\mathbf{grad}v_h\, dx + \int_\Omega \mathbf{c}\,\mathbf{grad}u_h\, v_h\, dx = \int_\Omega f v_h\, dx \tag{5.17}$$

or equivalently

$$\text{for all} \, v_h \in V_h \quad \langle F(u_h), v_h \rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = 0. \tag{5.18}$$

In the next chapter we will prove that (5.18) has a unique solution $u_h \in V_h$ and get a priori and a posteriori error estimates for $\|u - u_h\|_{1,p,\Omega}$.

CHAPTER III

NONLINEAR PROBLEMS AND THEIR NUMERICAL APPROXIMATION

It is the purpose of this chapter to develop an approximation method for nonlinear problems not depending on a parameter. We follow here the presentation of POUSIN and RAPPAZ [1991], [1992].

A general framework is first established. Let X and Z be two real (or complex) Banach spaces, $F : X \to Z$ be a C^1 mapping, and $\{F_h\}_{0<h\leq 1}$ be a family of C^1 mappings defined from X to Z. Under equicontinuity, consistency, and stability hypotheses we prove the convergence of a solution u_h of $F_h(u_h) = 0$ to a solution u of $F(u) = 0$.

When Z is the dual space of a Banach space, a particular attention is paid to the Galerkin approximations of the problem $F(u) = 0$. Our approach is presented as a generalization of the linear theory. We get the analogue of the inf-sup conditions and both a priori and a posteriori error estimates.

Finally the abstract results are applied to the model problem of Section 3, to the approximation of the Navier-Stokes problem given in Section 4, and to the stationary heat problem with convection of Section 5.

6. An abstract framework

The two real (or complex) Banach spaces X and Z are endowed with the norms $\|.\|_X$ and $\|.\|_Z$ respectively. Given a family $\{F_h\}_{0<h\leq 1}$ of C^1 mappings, $F_h : X \to Z$, we assume there exists a $u \in X$ such that the three following hypotheses hold. The mappings DF_h are Lipschitzian at u, uniformly in a neighborhood of u, with constant L_h, that is there exists $\eta > 0$ and for all $h \in (0, 1]$ there exists a constant L_h such that for all $v \in \overline{B}(u, \eta)$

$$\|DF_h(u) - DF_h(v)\|_{X;Z} \leq L_h \|u - v\|_X; \tag{6.1}$$

$$\lim_{h\to 0}(1 + L_h)\|F_h(u)\|_Z = 0; \tag{6.2}$$

$DF_h(u)$ is an isomorphism from X onto Z with an inverse uniformly bounded in h, that is there exists C_1 such that

$$\text{for all } h \in (0, 1] \quad \|DF_h(u)^{-1}\|_{Z;X} \leq C_1. \tag{6.3}$$

In the examples we shall develop, we will take $u \in X$ as a zero of a mapping $F : X \to Z$ and $\{F_h\}_{0<h\leq 1}$ as an approximation family of F. Most often in the applications L_h is bounded and the assumption (6.2) can be interpreted as a consistency hypothesis while (6.3) is a stability hypothesis. Note that with our presentation we can work with L_h unbounded but with $\|F_h(u)\|_Z$ tending to zero faster than L_h tends to ∞. Naturally the goal of this section is to prove that consistency, stability, and equicontinuity imply existence and convergence for the approximations.

Theorem 6.1. *Let $\{F_h\}_{0<h\leq 1}$ be a family of C^1 mappings from X into Z and $u \in X$ satisfy the hypotheses (6.1), (6.2), and (6.3). Then there exist $\delta_0 > 0$ and $h_0 > 0$ such that for all $h \leq h_0$, there is a unique u_h satisfying*

$$F_h(u_h) = 0 \qquad \text{and} \qquad \|u - u_h\|_X \leq \frac{\delta_0}{1 + L_h}. \tag{6.4}$$

Moreover for all $h \leq h_0$

$$\|u - u_h\|_X \leq 2\|DF_h(u)^{-1}\|_{Z;X}\|F_h(u)\|_Z. \tag{6.5}$$

Proof. We set

$$\begin{aligned} \epsilon_h &= \|F_h(u)\|_Z, \\ \gamma_h &= \|DF_h(u)^{-1}\|_{Z;X}, \\ \tilde{L}_h(\alpha) &= \sup_{v \in \overline{B}(u,\alpha)} \|DF_h(u) - DF_h(v)\|_{X;Z}. \end{aligned}$$

From the stability (6.3), we get $\gamma_h \leq C_1$. If η is the radius of the ball $\overline{B}(u, \eta)$ in (6.1) we set $\delta_0 = \min(\eta, 1/(2C_1))$ and $\delta_h = \delta_0/(1 + L_h)$. From (6.1) we have for all $h \in (0, 1]$

$$\tilde{L}_h(\delta_h) \leq (2C_1)^{-1}.$$

Now from the consistency (6.2), there exists h_0, $0 < h_0 \leq 1$ such that for all $h \in (0, h_0]$,

$$2C_1\epsilon_h \leq \delta_h.$$

Thus for $h \leq h_0$,

$$2\gamma_h\tilde{L}_h(2C_1\epsilon_h) \leq 2C_1\tilde{L}_h(\delta_h) \leq 1.$$

We apply Theorem 2.1 with $G = F_h$. So there exists a unique $u_h \in \overline{B}(u, 2\gamma_h\epsilon_h)$ with $F_h(u_h) = 0$. Moreover the following estimate holds

$$\|u - u_h\|_X \leq 2\gamma_h\|F_h(u)\|_Z.$$

Applying Remark 2.1, we get the uniqueness of u_h in $\overline{B}(u, \delta_h)$. Finally Theorem 6.1 is proved. □

Remark 6.1. In Theorem 6.1, if L_h is bounded with respect to h, we have proved the existence and uniqueness of a zero of F_h in a fixed neighborhood of u. □

We consider now a C^1 mapping $F : X \to Z$, $u \in X$ a zero of F, and a family $\{F_h\}_{0<h\leq 1}$ of mappings satisfying (6.1), (6.2), and (6.3). To relate u and u_h we have the following result.

Theorem 6.2. *If u is a zero of F such that $DF(u) \in \mathscr{L}(X;Z)$ is an isomorphism from X onto Z and if the family $\{F_h\}_{0<h\leq 1}$ satisfies (6.1), (6.2), and (6.3), then there exists $0 < \overline{h}_0 \leq h_0$ such that for all $h \leq \overline{h}_0$, $DF(u_h) \in \mathscr{L}(X;Z)$ is an isomorphism with a uniformly bounded inverse and*

$$\|u - u_h\|_X \leq 2\|DF(u_h)^{-1}\|_{Z;X}\|F(u_h)\|_Z, \tag{6.6}$$

where h_0 and u_h are given in Theorem 6.1.

Proof. To apply Theorem 2.1, we set

$$\begin{aligned} \epsilon_h &= \|F(u_h)\|_Z, \\ \gamma_h &= \|DF(u_h)^{-1}\|_{Z;X}, \\ \tilde{L}_h(\alpha) &= \sup_{v\in\overline{B}(u_h,\alpha)} \|DF(v) - DF(u_h)\|_{X;Z}. \end{aligned}$$

Let h_0, δ_0, and u_h be given in Theorem 6.1. A consequence of the consistency (6.2) and the estimate (6.5) is

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

Since $DF(u)$ is an isomorphism from X onto Z, we can choose $h_1 \leq h_0$ such that $DF(u_h)$ is an isomorphism with

$$\gamma_h = \|DF(u_h)^{-1}\|_{Z;X} \leq 2\|DF(u)^{-1}\|_{Z;X} \equiv \gamma.$$

There exist $0 < h_2 \leq h_1$ and $\delta > 0$ such that

$$\tilde{L}_h(\delta) \leq \frac{1}{2\gamma} \qquad \text{and} \qquad \|u - u_h\|_X < \delta.$$

Now let $h_3 \leq h_2$ be such that for all $h \leq h_3$

$$2\gamma\epsilon_h \leq \delta.$$

We are in position to apply Theorem 2.1 with $G = F$, $v = u_h$, $0 < h \leq h_3$. We conclude that there exists a unique $w \in \overline{B}(u_h, 2\gamma_h\epsilon_h)$ such that

$$F(w) = 0 \qquad \text{and} \qquad \|w - u_h\|_X \leq 2\|DF(u_h)^{-1}\|_{Z,X}\|F(u_h)\|_Z.$$

Applying Remark 2.1, we get the uniqueness of w in $\overline{B}(u_h, \delta)$. Since $F(u) = 0$, we have $w = u$ (actually we must choose $\overline{h}_0 < h_3$) and (6.6) is proved. □

The estimates (6.5) and (6.6) have a different meaning. From the first estimate we can deduce a a priori estimate bounded by the consistency error $\|F_h(u)\|_Z$. On the other hand we get from the second estimate a a posteriori estimate bounded by the residual $\|F(u_h)\|_Z$.

Theorem 6.3. *Let $F : X \to Z$ be a C^1 mapping and $\{F_h\}_{0<h\leq 1}$, $F_h : X \to Z$, be a family of C^2 mappings satisfying*

$$\text{(6.7)}\quad \begin{cases} \lim_{h\to 0} \|F(v) - F_h(v)\|_Z = 0 \text{ for all } v \in X, \\ \lim_{h\to 0} \|DF(v) - DF_h(v)\|_{X;Z} = 0 \text{ for all } v \in X, \\ D^2F_h \text{ is uniformly bounded with respect to } h \text{ on all bounded domain of } X. \end{cases}$$

Let $u \in X$ be such that

$$\text{(6.8)}\qquad F(u) = 0,$$

$$\text{(6.9)}\qquad DF(u) \quad \text{is an isomorphism from } X \text{ onto } Z.$$

Then there exist $h_0 > 0$, $\delta_0 > 0$ such that for all $h \leq h_0$ there exists a unique $u_h \in X$ satisfying

$$F_h(u_h) = 0 \qquad \text{and} \qquad \|u - u_h\|_X \leq \delta_0.$$

Moreover for all $h \leq h_0$ we have

$$\text{(6.10)}\qquad \|u - u_h\|_X \leq 2\|DF_h(u)^{-1}\|_{Z;X}\|F_h(u)\|_Z,$$

$$\text{(6.11)}\qquad \|u - u_h\|_X \leq 2\|DF(u_h)^{-1}\|_{Z;X}\|F(u_h)\|_Z,$$

Proof. It is easy to prove that the hypotheses (6.7), (6.8), and (6.9) imply the ones of Theorems 6.1 and 6.2 with the constants L_h bounded with respect to h. □

Remark 6.2. For simplicity, the function u in Theorem 6.1 is not supposed to depend on h. In Section 12, we assume a dependence on h. A convenient choice of a function depending on h may lead to error estimates in different norms, see Section 14. □

In the next result we present a simple way to derive error estimates in different norms, see Rappaz [1983] for a variant of this technique.

We assume there exist two Banach spaces $\mathscr{X}$ and $\mathscr{Z}$, with the norms $\|.\|_{\mathscr{X}}$ and $\|.\|_{\mathscr{Z}}$, satisfying the following assumptions.

$$\text{(6.12)}\qquad X \subset \mathscr{X} \text{ and } Z \subset \mathscr{Z} \text{ with continuous injection;}$$

there exist $\epsilon > 0$ and $\kappa > 0$ such that for all $v \in B_X(u,\epsilon) = \{w \in X; \|w - u\|_X < \epsilon\}$, the operators $DF_h(v)$ and $DF(v)$ can be extended as operators in $\mathscr{L}(\mathscr{X};\mathscr{Z})$ and

$$\text{(6.13)}\qquad \begin{cases} \text{for all } v \in B_X(u,\epsilon) & \|DF(v) - DF(u)\|_{\mathscr{X};\mathscr{Z}} \leq \kappa\|v - u\|_X, \\ \text{for all } v \in B_X(u,\epsilon) & \|DF_h(v) - DF_h(u)\|_{\mathscr{X};\mathscr{Z}} \leq \kappa\|v - u\|_X; \end{cases}$$

$$\text{(6.14)}\qquad \lim_{h\to 0} \|DF(u) - DF_h(u)\|_{\mathscr{X};\mathscr{Z}} = 0;$$

$$\text{(6.15)}\qquad DF(u) \text{ is an isomorphism from } \mathscr{X} \text{ onto } \mathscr{Z}.$$

The following result holds.

Theorem 6.4. *Under the assumptions of Theorem 6.3 and the (6.12), (6.13), (6.14), (6.15) ones, there exists $0 < h_1 \leq h_0$ such that for all $0 < h \leq h_1$, we have*

$$\|u - u_h\|_{\mathscr{X}} \leq 2\|DF_h(u)^{-1}\|_{\mathscr{Z};\mathscr{X}}\|F_h(u)\|_{\mathscr{Z}}, \tag{6.16}$$

$$\|u - u_h\|_{\mathscr{X}} \leq 2\|DF(u_h)^{-1}\|_{\mathscr{Z};\mathscr{X}}\|F(u_h)\|_{\mathscr{Z}}. \tag{6.17}$$

Proof. We first prove the estimate (6.16). Noticing that $F_h(u_h) = 0$, we use the Taylor expansion to write

$$u - u_h = DF_h(u)^{-1}\left[\int_0^1 \big(DF_h(u) - DF_h(su + (1-s)u_h)\big)(u - u_h)\,ds\right] + DF_h(u)^{-1}F_h(u). \tag{6.18}$$

Clearly the assumptions (6.14), (6.15), the result (1.1), and the estimate (6.10) imply the existence of $0 < \overline{h} \leq h_0$ and $\beta > 0$ such that for all $0 < h \leq \overline{h}$, the mapping $DF_h(u)$ is an isomorphism from $\mathscr{X}$ onto $\mathscr{Z}$ with

$$\|DF_h(u)^{-1}\|_{\mathscr{Z};\mathscr{X}} \leq \beta$$

and

$$\|u - u_h\|_X < \epsilon.$$

With (6.18), we get

$$\|u - u_h\|_{\mathscr{X}} \leq \|DF_h(u)^{-1}\|_{\mathscr{Z};\mathscr{X}}\left\{\kappa\|u - u_h\|_X\|u - u_h\|_{\mathscr{X}} + \|F_h(u)\|_{\mathscr{Z}}\right\}. \tag{6.19}$$

Let us choose now $0 < h_1 \leq \overline{h}$ such that for all $0 < h \leq h_1$ the following holds

$$\beta\kappa\|u - u_h\|_X \leq \frac{1}{2}.$$

Then from the estimate (6.19) we deduce the (6.16) one.

To prove the estimate (6.17), we proceed as before with the relation

$$u - u_h = DF(u)^{-1}\left[\int_0^1 \big(DF(u) - DF(su + (1-s)u_h)\big)(u - u_h)\,ds\right] - DF(u)^{-1}F(u_h). \quad \square$$

7. Galerkin approximation of nonlinear problems

In order to relate the results of the previous section with the ones relative to the Galerkin approximation of linear problems, we shall focus on a particular abstract framework. We suppose from now on that we have $Z = Y'$, Y' being the dual space of a real Banach space Y. We assume

$$\begin{cases} F : X \to Y' \text{ is a } C^1 \text{ mapping and } u \in X \text{ is such that } F(u) = 0 \\ \text{and } DF(u) \text{ is an isomorphism from } X \text{ onto } Y'. \end{cases} \tag{7.1}$$

Note that

$$F(u) = 0 \quad \Longleftrightarrow \quad \text{for all } y \in Y \quad \langle F(u), y \rangle_{Y'Y} = 0.$$

Let now $\{X_h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces of X and $\{Y_h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces of Y. A Galerkin approximation of the problem to find $u \in X$ such that

$$F(u) = 0 \tag{7.2}$$

reads: find $u_h \in X_h$ such that

$$\text{for all } y_h \in Y_h \quad \langle F(u_h), y_h \rangle_{Y'Y} = 0. \tag{7.3}$$

To introduce the analogue of the inf-sup conditions for the nonlinear case, we define the bilinear form $b : X \times Y \to \mathbb{R}$ by

$$\text{for all } x \in X, y \in Y \quad b(x, y) = \langle DF(u)x, y \rangle_{Y'Y}.$$

We assume there exists a constant $\beta_h > 0$ such that

$$\inf_{\substack{x \in X_h \\ \|x\|_X = 1}} \sup_{\substack{y \in Y_h \\ \|y\|_Y = 1}} b(x, y) \geq \beta_h \tag{7.4}$$

and

$$\dim X_h = \dim Y_h \tag{7.5}$$

for $0 < h \leq 1$. Finally we assume

$$\lim_{h \to 0} \inf_{x_h \in X_h} \beta_h^{-2} \|u - x_h\|_X = 0. \tag{7.6}$$

Remark that $\beta_h \leq \|DF(u)\|_{X;Y'}$ and if β_h is bounded from below by $\beta > 0$ for all h, then (7.4) is a standard stability condition for linear problems and (7.6) is a standard approximation property of X by X_h.

We are in a position to state the major result of the section.

Theorem 7.1. *Let us suppose that the hypotheses (7.1), (7.4), (7.5), and (7.6) are satisfied. Moreover we assume that DF is Lipschitzian at u, that is there exist $\epsilon_0 > 0$ and $L > 0$ such that for all $v \in X$ with $\|u - v\|_X \leq \epsilon_0$*

$$\|DF(u) - DF(v)\|_{X;Y'} \leq L\|u - v\|_X .$$

Then there exist two constants $h_0 > 0$, $\delta_0 > 0$, and for $h \leq h_0$ a unique solution u_h to problem (7.3) in the closed ball $\overline{B}(u, \delta_h)$ with $\delta_h = \delta_0\beta_h$. Moreover there exists a constant C independent of h such that

$$\|u - u_h\|_X \leq C\beta_h^{-1} \inf_{x_h \in X_h} \|u - x_h\|_X \tag{7.7}$$

and

$$\|u - u_h\|_X \leq 2\|DF(u_h)^{-1}\|_{Y';X}\|F(u_h)\|_{Y'} . \tag{7.8}$$

Proof. The proof is based on Theorems 6.1 and 6.2. To define the mapping F_h, we introduce the two projectors $\Pi_{X_h} \in \mathscr{L}(X; X_h)$ and $\Pi_{Y_h} \in \mathscr{L}(Y; Y_h)$ by setting

$$\text{for all} \, y \in Y_h \quad b(x - \Pi_{X_h} x, y) = 0, \tag{7.9}$$

$$\text{for all} \, x \in X_h \quad b(x, y - \Pi_{Y_h} y) = 0. \tag{7.10}$$

Under the hypotheses (7.4) and (7.5), both operators are well defined. Therefore we can construct a family $\{F_h\}_{0<h\leq 1}$ of mappings, $F_h : X \to Y'$, by

$$\text{for all} \, x \in X, y \in Y \quad \langle F_h(x), y\rangle_{Y'Y} = \langle F(x), \Pi_{Y_h} y\rangle_{Y'Y} + b(x, y - \Pi_{Y_h} y). \tag{7.11}$$

Then we can analyze the problem to find $x \in X$ such that

$$F_h(x) = 0. \tag{7.12}$$

Problems (7.3) and (7.12) are equivalent. Clearly if $u_h \in X_h$ is a solution to (7.3) then it is a solution to (7.12). Conversely let $u_h \in X$ be a solution to (7.12). To prove that u_h is a solution to (7.3) it is sufficient to prove that $u_h \in X_h$. By taking $y = y_1 - \Pi_{Y_h} y_1$ with $y_1 \in Y$ in (7.11) we get

$$b(u_h, y_1 - \Pi_{Y_h} y_1) = 0$$

and so

$$\text{for all} \, y_1 \in Y \quad b(u_h - \Pi_{X_h} u_h, y_1) = 0.$$

Since $DF(u)$ is an isomorphism, we conclude that

$$u_h = \Pi_{X_h} u_h \in X_h .$$

In order to conclude, we will apply Theorems 6.1 and 6.2 to the family $\{F_h\}_{0<h\leq 1}$ defined in (7.11). So let us check that the hypotheses (6.1), (6.2), and (6.3) are fulfilled.

The Fréchet derivative of F_h at $v \in X$ is,

$$\text{for all}\, x \in X, y \in Y \quad \langle DF_h(v)x, y\rangle_{Y'Y} = \langle DF(v)x, \Pi_{Y_h} y\rangle_{Y'Y} + b(x, y - \Pi_{Y_h} y).$$

It follows that for $x \in X, y \in Y$

$$\langle \big(DF_h(u) - DF_h(v)\big)x, y\rangle_{Y'Y} = \langle \big(DF(u) - DF(v)\big)x, \Pi_{Y_h} y\rangle_{Y'Y}$$

and consequently for $v \in X$ with $\|u - v\|_X \leq \epsilon_0$ we get

$$\|DF_h(u) - DF_h(v)\|_{X;Y'} \leq L\|u - v\|_X \|\Pi_{Y_h}\|_{Y;Y}. \tag{7.13}$$

With the notation $\|b\| = \|DF(u)\|_{X;Y'}$, we easily check with (7.4) that

$$\|\Pi_{X_h}\|_{X;X} \leq \frac{\|b\|}{\beta_h}. \tag{7.14}$$

In order to bound the norm of Π_{Y_h} we write for $y \in Y$

$$\begin{aligned}\|\Pi_{Y_h} y\|_Y &= \sup_{\substack{\varphi\in Y'\\ \|\varphi\|_{Y'}=1}} \langle \varphi, \Pi_{Y_h} y\rangle_{Y'Y} = \sup_{\substack{\varphi\in Y'\\ \|\varphi\|_{Y'}=1}} b(DF(u)^{-1}\varphi, \Pi_{Y_h} y)\\ &= \sup_{\substack{\varphi\in Y'\\ \|\varphi\|_{Y'}=1}} b(\Pi_{X_h} DF(u)^{-1}\varphi, y).\end{aligned}$$

By using (7.14) we finally obtain

$$\|\Pi_{Y_h}\|_{Y;Y} \leq \frac{\|b\|^2}{\beta_h}\|DF(u)^{-1}\|_{Y';X}. \tag{7.15}$$

We notice that $\beta_h \leq \|b\|$ and $\|DF(u)^{-1}\|^{-1}_{Y';X} \leq \|b\|$, so the inequalities (7.13) and (7.15) imply that the hypothesis (6.1) holds with

$$L_h = (1 + L)\frac{\|b\|^2}{\beta_h}\|DF(u)^{-1}\|_{Y';X} - 1.$$

Let us check the hypothesis (6.2). By the definition of F_h we have

$$\begin{aligned}\|F_h(u)\|_{Y'} &= \sup_{\substack{y\in Y\\ \|y\|_Y=1}} \langle F_h(u), y\rangle_{Y'Y} = \sup_{\substack{y\in Y\\ \|y\|_Y=1}} b(u, y - \Pi_{Y_h} y)\\ &= \sup_{\substack{y\in Y\\ \|y\|_Y=1}} b(u - \Pi_{X_h} u, y) \leq \|b\|\|u - \Pi_{X_h} u\|_X.\end{aligned} \tag{7.16}$$

Using the assumption (7.4), we get for $x_h \in X_h$

$$\begin{aligned}\beta_h \|x_h - \Pi_{X_h} u\|_X &\leq \sup_{\substack{y \in Y_h \\ \|y\|_Y = 1}} b(x_h - \Pi_{X_h} u, y) \\ &\leq \sup_{\substack{y \in Y_h \\ \|y\|_Y = 1}} b(x_h - u, y) \leq \|b\| \|u - x_h\|_X .\end{aligned}$$

So we deduce that for $x_h \in X_h$

$$\|u - \Pi_{X_h} u\|_X \leq (1 + \frac{\|b\|}{\beta_h}) \|u - x_h\|_X$$

and using (7.16) we have

$$\|F_h(u)\|_{Y'} \leq \frac{C}{\beta_h} \inf_{x_h \in X_h} \|u - x_h\|_X \tag{7.17}$$

where C is independent of h.

When taking into account that

$$L_h = (1 + L)\|b\|^2 \|DF(u)^{-1}\|_{Y'X} / \beta_h - 1$$

and using the hypothesis (7.6), we prove that the hypothesis (6.2) is fulfilled. Going back to the definition of F_h we have $DF_h(u) = DF(u)$, which is an isomorphism from X onto Y' with a uniformly bounded inverse. Then we conclude that hypothesis (6.3) is satisfied.

We conclude with Theorems 6.1 and 6.2. The error estimates (7.7) is a direct consequence of (6.5), (7.17), and of the relation $DF_h(u) = DF(u)$. □

Remark 7.1. Theorem 7.1 can be viewed as a generalization of the linear theory. If the mapping F is smooth enough and the solution regular in the sense of (7.1), then we have reduced the approximation study to a verification of an inf-sup condition, see (7.4), and of an approximability result, see (7.6). Then the error estimates are deduced via an interpolation result for the a priori estimate (7.7) and via a computation of the residual for the a posteriori estimate (7.8). □

Remark 7.2. The approach proposed by RHEINBOLDT [1986] to study the approximation of parameter dependent problems with projection methods is in some respect close to ours. □

Remark 7.3. Quite often in the applications the spaces X and Y are identical Hilbert spaces and there exists a Hilbert space W satisfying $X \subset W$ with continuous embedding and X dense in W. The space W takes the place of a pivot space in the embeddings

$$X \subset W \subset X' .$$

The scalar product in X (respectively W) is denoted by $(.,.)_X$ (respectively$(.,.)_W$).

We assume that $X_h = Y_h$, the hypotheses of Theorem 7.1 hold, and the mapping F is of class C^2. Then we can derive an estimate in the norm of W from the estimate in the norm of X using a standard duality argument.

Indeed we have

$$\|u - u_h\|_W = \sup_{\substack{g \in W \\ g \neq 0}} \frac{|(g, u - u_h)_W|}{\|g\|_W}.$$

For a given g in W, let w be the unique solution to

$$\text{for all } v \in X \quad b(v, w) = (v, g)_W.$$

Then there exists a constant C independent of g such that

$$\|w\|_X \leq C\|g\|_W.$$

So for a given g in W we get

$$|(g, u - u_h)_W| = |b(u - u_h, w)| \leq |b(u - u_h, w - w_h)| + |b(u - u_h, w_h)|$$

for any function $w_h \in X_h$. We can estimate the term $|b(u - u_h, w_h)|$ in the following way

$$\begin{aligned} b(u-u_h, w_h) = \langle DF(u)(u-u_h), w_h\rangle_{X'X} &= \langle -F(u)+F(u_h)-DF(u)(u_h-u), w_h\rangle_{X'X} \\ &= \langle \int_0^1 (1-t)D^2F(u+t(u_h-u))(u_h-u)^2, w_h\rangle_{X'X} \end{aligned}$$

and so

$$|b(u - u_h, w_h)| \leq C\|u - u_h\|_X^2 \|w_h\|_X.$$

For any g in W and $w_h \in X_h$, we get

$$|(u - u_h, g)_W| \leq C \left\{ \|u - u_h\|_X \|w - w_h\|_X + \|u - u_h\|_X^2 \|w_h\| \right\}.$$

We have proved that for any $g \in W$ and $w_h \in X_h$

$$|(u - u_h, g)_W| \leq C \left\{ \|u - u_h\|_X \left[1 + \|u - u_h\|_X\right] \|w - w_h\|_X + \|u - u_h\|_X^2 \|g\|_W \right\}.$$

An illustration of the method is given in Theorem 8.2. □

Remark 7.4. When the mapping $F : X \to Y'$ has the following structure

$$F(x) = Lx + G(x)$$

where $L \in \mathscr{L}(X; Y')$ is an isomorphism and $G : X \to Y'$ is such that $DG(x) \in \mathscr{L}(X; Y')$ is compact, then we can use the simple bilinear form $b : X \times Y \to \mathbb{R}$ given by

$$b(x, y) = \langle Lx, y\rangle_{Y'Y}.$$

In Section 16, this idea is used to simplify the study of the approximation of simple limit points. □

8. Application to a model problem

The general results of last section are applied to the semilinear problem (3.10). We shall study both approximation schemes (3.13) and (3.14).

In this particular case, the notations are the following: $X = Y = H_0^1(\Omega)$ equipped with the norm $\|v\|_X = \|v\|_Y = |v|_{1,\Omega} = (\int_\Omega |\mathbf{grad} v|^2\, dx)^{1/2}$ equivalent to $\|.\|_{1,\Omega}$, $Z = Y' = H^{-1}(\Omega)$, and $F : X \to Y'$ is given by

$$\text{(8.1)} \quad \text{for all } u, v \in H_0^1(\Omega) \quad \langle F(u), v\rangle_{H^{-1}(\Omega) H_0^1(\Omega)} = \int_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dx + \int_\Omega u^3 v\, dx - \int_\Omega f v\, dx,$$

where f is an $L^2(\Omega)$ function. Theorem 3.1 implies there exists a unique solution $u \in H_0^1(\Omega)$ to the problem

$$F(u) = 0; \tag{8.2}$$

furthermore u is in $H^2(\Omega)$ and the derivative $DF(u) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is an isomorphism, which means that (7.1) is satisfied; moreover DF is Lipschitzian at u.

The approximation space X_h and the test space Y_h are both V_h given in (3.12), the space of continuous piecewise polynomials of degree ≤ 1, so the hypothesis (7.5) is guaranteed. The bilinear form $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$,

$$b(v, w) = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx + 3 \int_\Omega u^2 v w\, dx,$$

is coercive on $H_0^1(\Omega)$ and this property ensures the hypothesis (7.4) with $\beta_h = 1$. Since the family of triangulations $\{\mathscr{T}_h\}_{0<h\leq 1}$ is regular and β_h is equal to 1, the assumption (7.6) is satisfied.

We can apply Theorem 7.1 to study the existence, the local uniqueness and the convergence of problem (3.13), that is find $u_h \in V_h$ such that

$$\text{for all } v_h \in V_h \quad \langle F(u_h), v_h\rangle_{H^{-1}(\Omega) H_0^1(\Omega)} = 0. \tag{8.3}$$

A definition useful in the next theorem is the following. Let w be a function in $W^{1,1}(\Omega)$ (so that the trace operator can be defined), T_i, T_j, with $i < j$, be two adjacent triangles, and let T' denote the common side. If w_i and w_j denote the restrictions of w on T_i and T_j respectively, we set

$$[w] = w_i - w_j \quad \text{on} \quad T',$$

$[w]$ is the jump of w through the edge T'. If T' is a side of T_i on the boundary of Ω, then $[w] = w_i$ on T'.

Theorem 8.1. *There exist positive constants h_0, δ_0, C, and for $h \le h_0$ a unique solution u_h to (8.3) in $\overline{B}(u, \delta_0)$. Moreover the following a priori and a posteriori estimates hold for $h \le h_0$:*

$$|u - u_h|_{1,\Omega} \le Ch\|f\|_{0,\Omega} \tag{8.4}$$

and

$$|u - u_h|_{1,\Omega} \le C\left(\sum_{T\in\mathscr{T}_h} \eta(T)^2\right)^{1/2}, \tag{8.5}$$

where $\eta(T)$ is the local estimator given by

$$\eta(T) = \left(h_T^2\| - \Delta u_h + u_h^3 - f\|_{0,T}^2 + \sum_{i=1}^{3} h_{t_i}\| \left[\frac{\partial u_h}{\partial n}\right] \|_{0,t_i}^2\right)^{1/2}$$

with h_T the diameter of $T \in \mathscr{T}_h$, t_i $i = 1,2,3$ the edges of T of length h_{t_i}, and $[.]$ the jump through an edge.

Proof. The hypotheses of Theorem 7.1 were verified before with $\beta_h = 1$, so there exist positive constants h_0, δ_0, and for $h \le h_0$ a unique solution u_h to (8.3) in $\overline{B}(u, \delta_0)$. From the estimates (7.7) and (7.8), we shall deduce (8.4) and (8.5) respectively.

Since the solution u to (8.2) is in $H^2(\Omega)$ and $\beta_h = 1$, we deduce from (7.7) that

$$|u - u_h|_{1,\Omega} \le Ch\|f\|_{0,\Omega}.$$

So we have proved (8.4).

The residual $\|F(u_h)\|_{-1,\Omega}$ is estimated following the method presented in Baranger and El Amri [1991]. For $v \in H_0^1(\Omega)$ we have

$$\begin{aligned}\langle F(u_h), v\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} &= \int_\Omega \mathbf{grad} u_h \,\mathbf{grad} v\, dx + \int_\Omega (u_h^3 - f)v\, dx \\ &= \int_\Omega \mathbf{grad} u_h \,\mathbf{grad}(v - \tilde{r}_h v)\, dx + \int_\Omega (u_h^3 - f)(v - \tilde{r}_h v)\, dx,\end{aligned}$$

where $\tilde{r}_h \in \mathscr{L}(H_0^1(\Omega), V_h)$ is the Clément interpolation operator onto V_h, see Clément [1975]. So for all $v \in H_0^1(\Omega)$

$$\begin{aligned}\langle F(u_h), v\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} = \sum_{T\in\mathscr{T}_h} \Big\{ &\int_T \mathbf{grad} u_h \,\mathbf{grad}(v - \tilde{r}_h v)\, dx \\ &+ \int_T (u_h^3 - f)(v - \tilde{r}_h v)\, dx\Big\}.\end{aligned} \tag{8.6}$$

For all $T \in \mathscr{T}_h$,

$$\int_T \mathbf{grad} u_h \, \mathbf{grad}(v - \tilde{r}_h v)\, dx = -\int_T \Delta u_h (v - \tilde{r}_h v)\, dx + \int_{\partial T} \frac{\partial u_h}{\partial n_T}(v - \tilde{r}_h v)\, ds,$$

where n_T is the outward unit normal along the sides of T. Since u_h is in the space $\mathscr{P}_1$ on T the above equality reads

$$\int_T \mathbf{grad} u_h \, \mathbf{grad}(v - \tilde{r}_h v)\, dx = \int_{\partial T} \frac{\partial u_h}{\partial n_T}(v - \tilde{r}_h v)\, ds.$$

Let S_h be the set of all sides of triangles in $\mathscr{T}_h$. For each $t \in S_h$ we choose the direction of a unit normal n and denote by h_t its length, then

$$\Big| \sum_{T \in \mathscr{T}_h} \int_T \mathbf{grad} u_h \, \mathbf{grad}(v - \tilde{r}_h v)\, dx \Big| \le \sum_{t \in S_h} \int_t \Big| \left[\frac{\partial u_h}{\partial n} \right] (v - \tilde{r}_h v) \Big|\, ds;$$

if t is a side on $\partial\Omega$, we naturally set to zero the value outside $\overline{\Omega}$. Then the Cauchy Schwarz inequality gives

$$\Big| \sum_{T \in \mathscr{T}_h} \int_T \mathbf{grad} u_h \, \mathbf{grad}(v - \tilde{r}_h v)\, dx \Big| \le \left(\sum_{t \in S_h} h_t \Big\| \left[\frac{\partial u_h}{\partial n} \right] \Big\|_{0,t}^2 \right)^{1/2} \left(\sum_{t \in S_h} h_t^{-1} \| v - \tilde{r}_h v \|_{0,t}^2 \right)^{1/2}. \tag{8.7}$$

Using the technique of the reference element, see CIARLET [1978] or CIARLET [1991], it is a simple matter to prove the existence of a constant C such that for all $T \in \mathscr{T}_h$ with edges t_1, t_2, and t_3, and for all $w \in H^1(T)$ we have

$$\sum_{j=1}^{3} h_{t_j}^{-1} \| w \|_{0,t_j}^2 \le C \left(h_T^{-2} \| w \|_{0,T}^2 + |w|_{1,T}^2 \right).$$

So the inequality (8.7) is transformed into

$$\Big| \sum_{T \in \mathscr{T}_h} \int_T \mathbf{grad} u_h \, \mathbf{grad}(v - \tilde{r}_h v)\, dx \Big| \le C \left(\sum_{t \in S_h} h_t \Big\| \left[\frac{\partial u_h}{\partial n} \right] \Big\|_{0,t}^2 \right)^{1/2} \left(\sum_{T \in \mathscr{T}_h} \left(h_T^{-2} \| v - \tilde{r}_h v \|_{0,T}^2 + |v - \tilde{r}_h v|_{1,T}^2 \right) \right)^{1/2}. \tag{8.8}$$

With Clément's interpolation estimates, see CLÉMENT [1975], for $v \in H^1(\Omega)$

$$\| v - \tilde{r}_h v \|_{0,T} \le Ch \sum_{T' \in S_T} |v|_{1,T'} \quad \text{and} \quad |v - \tilde{r}_h v|_{1,T} \le Ch \sum_{T' \in S_T} |v|_{1,T'},$$

where S_T is the set of triangles T' with $T' \cap T \neq \emptyset$, and with (8.8), the relation (8.6) is transformed by using the Cauchy Schwarz inequality into

$$\langle F(u_h), v\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} \leq C\Big[\left(\sum_{t\in S_h} h_t \| \left[\frac{\partial u_h}{\partial n}\right] \|_{0,t}^2\right)^{1/2} |v|_{1,\Omega} + \left(\sum_{T\in\mathcal{T}_h} h_T^2 \|u_h^3 - f\|_{0,T}^2\right)^{1/2} |v|_{1,\Omega}\Big]. \tag{8.9}$$

Finally the estimate (8.5) is a consequence of (7.8), the fact that $DF(u_h)$ for $h \leq h_0$ is an isomorphism from $H_0^1(\Omega)$ onto $H^{-1}(\Omega)$ with a uniformly bounded inverse because $\lim_{h\to 0} u_h = u$ and the estimate (8.9). □

Remark 8.1. Although the a posteriori estimate (8.5) is a crude one, it can justify an adaptive mesh refinement procedure, based on the minimization of the residual. The theory of a posteriori estimates for nonlinear problems has advanced considerably in recent years. Model problems in one space dimension are studied in BABUŠKA and RHEINBOLDT [1982] and in RHEINBOLDT [1981]. An other approach is presented in RHEINBOLDT [1985] which produces reliable a posteriori estimates. We refer to the review paper of VERFÜRTH [1993] for recent results on a posteriori estimates. □

It is also possible to obtain a priori estimates in other norms than the energy norm $|.|_{1,\Omega}$. For instance with duality arguments of Aubin-Nitsche, see Remark 7.3, we shall derive an error estimate in the $L^2(\Omega)$ norm.

Theorem 8.2. *Let h_0, δ_0, and for all $h \leq h_0$ $u_h \in \overline{B}(u, \delta_0)$ be given in Theorem 8.1. There exists a constant C such that*

$$\|u - u_h\|_{0,\Omega} \leq Ch^2. \tag{8.10}$$

Proof. Since the form $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$, $b(v,w) = \langle DF(u)v, w\rangle_{H^{-1}(\Omega)H_0^1(\Omega)}$, is coercive on $H_0^1(\Omega)$, there exists a unique solution w to the problem: find $w \in H_0^1(\Omega)$ satisfying

$$\text{for all } v \in H_0^1(\Omega) \quad b(v,w) = \int_\Omega (u - u_h)v\, dx. \tag{8.11}$$

By the standard elliptic regularity, we have $w \in H_0^1(\Omega) \cap H^2(\Omega)$ and

$$\|w\|_{2,\Omega} \leq C\|u - u_h\|_{0,\Omega}; \tag{8.12}$$

for ease of exposition we have assumed that Ω is a convex polygonal domain in $\mathbb{R}^2$, see the definition of V_h in (3.12). If we choose $w_h \in V_h$ to be the solution of

$$\text{for all } v \in V_h \quad b(v, w_h) = \int_\Omega (u - u_h)v\, dx,$$

the following estimate holds

$$\|w - w_h\|_{1,\Omega} \leq Ch\|w\|_{2,\Omega}$$

and with (8.12)

$$\|w - w_h\|_{1,\Omega} \leq Ch\|u - u_h\|_{0,\Omega}. \tag{8.13}$$

Starting from (8.11), we deduce from the inequality (8.13)

$$\|u - u_h\|_{0,\Omega}^2 = b(u - u_h, w) \leq Ch\|u - u_h\|_{0,\Omega}\|u - u_h\|_{1,\Omega} + |b(u - u_h, w_h)|. \tag{8.14}$$

Since

$$F(u) = 0 \quad \text{and} \quad \text{for all } v_h \in V_h \quad \langle F(u_h), v_h\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} = 0,$$

we get with the C^2 regularity of F and for $w_h \in V_h$

$$\begin{aligned} b(u - u_h, w_h) = \langle DF(u)(u - u_h), w_h\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} &= \langle F(u) - F(u_h), w_h\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} \\ &+ \langle \int_0^1 (1-t)D^2F(u + t(u_h - u))(u_h - u)^2\, dt, w_h\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} \end{aligned}$$

and so

$$|b(u - u_h, w_h) \leq C\|u_h - u\|_{1,\Omega}^2 \left[\|w_h - w\|_{1,\Omega} + \|w\|_{1,\Omega}\right]. \tag{8.15}$$

With (8.4), (8.12), (8.13), (8.14), and (8.15) we finally get (8.10). □

To study the approximation scheme with numerical integration (3.14), we cannot use directly the result of Section 7 relative to Galerkin approximation. We have to adapt the formalism and use the general results of Section 6. The elliptic projection operator $\pi_h \in \mathscr{L}(H_0^1(\Omega); V_h)$ is defined for $v \in H_0^1(\Omega)$ by

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad}(v - \pi_h v)\, \mathbf{grad} v_h\, dx = 0.$$

We assume that f is in $C^0(\overline{\Omega})$ and then we construct the mapping $F_h : H_0^1(\Omega) \to H^{-1}(\Omega)$ by the formula, for $v, w \in H_0^1(\Omega)$,

$$\langle F_h(v), w\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx + \int_\Omega r_h\left[\left((\pi_h v)^3 - f\right)\pi_h w\right] dx. \tag{8.16}$$

It is a simple matter to check that problem (3.14) is equivalent to find $u_h \in H_0^1(\Omega)$ satisfying

$$\text{for all } v \in H_0^1(\Omega) \quad \langle F_h(u_h), v\rangle_{H^{-1}(\Omega)H_0^1(\Omega)} = 0. \tag{8.17}$$

Indeed let $u_h \in V_h$ be a solution to (3.14). Then clearly u_h is also a solution to (8.17). Conversely let $u_h \in H_0^1(\Omega)$ be a solution to (8.17). If the test function v in (8.17) has the form $w - \pi_h w$ with $w \in H_0^1(\Omega)$, then the equation (8.17) reads

$$\int_\Omega \mathbf{grad} u_h \mathbf{grad}(w - \pi_h w)\, dx = 0,$$

which implies

$$u_h = \pi_h u_h \in V_h.$$

If we take $v_h \in V_h$ to be a test function in (8.17) then we get (3.14). Thus a solution $u_h \in H_0^1(\Omega)$ of (8.17) is a solution to (3.14).

Before studying problem (8.17), we present a technical result.

Theorem 8.3. *Let V_h be the space of continuous piecewise polynomials of degree ≤ 1 given in (3.12) and $r_h \in \mathscr{L}(C^0(\overline{\Omega}); V_h)$ be the Lagrange interpolation operator. Then for $M \in \mathbb{N}$, there exists a constant C such that for any $u_j \in V_h$ $j = 1, 2, \ldots, M$, we have*

$$\Big| \int_\Omega \Big(r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big)\, dx \Big| \leq C h \prod_{j=1}^M \|u_j\|_{1,\Omega}.$$

Proof. Since

$$\Big| \int_\Omega \Big(r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big)\, dx \Big| \leq \sum_{T \in \mathscr{T}_h} \int_T \Big| r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big|\, dx,$$

we derive an estimate for each $T \in \mathscr{T}_h$. On an element T with the reference element technique, we get

$$\int_T \Big| r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big|\, dx = |\det B| \int_{\hat{T}} \Big| \hat{r} \big(\prod_{j=1}^M \hat{u}_j \big) - \big(\prod_{j=1}^M \hat{u}_j \big) \Big|\, d\hat{x},$$

where $x \equiv F_T(\hat{x}) = B\hat{x} + b$ is the affine mapping from the reference element $\hat{T}$ onto T, $\hat{r}$ the Lagrange $\hat{\mathscr{P}}_1$-interpolant on $\hat{T}$, and $\hat{u}_j(\hat{x}) = u_j(x)$; $\hat{\mathscr{P}}_\ell$ denotes the space of polynomials of degree $\leq \ell$ on $\hat{T}$. For any $\hat{p} \in \hat{\mathscr{P}}_0$, we have

$$\int_T \Big| r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big|\, dx = |\det B| \int_{\hat{T}} \Big| (I - \hat{r}) \Big(\big(\prod_{j=1}^M \hat{u}_j \big) - \hat{p} \Big) \Big|\, d\hat{x}.$$

Clearly the mapping $\hat{r} : \hat{\mathscr{P}}_M \to \hat{\mathscr{P}}_1$ is linear and continuous. So

$$\int_T \Big| r_h \big(\prod_{j=1}^M u_j \big) - \big(\prod_{j=1}^M u_j \big) \Big|\, dx \leq C |\det B| \, \Big\| \big(\prod_{j=1}^M \hat{u}_j \big) - \hat{p} \Big\|_{1,1,\hat{T}}$$

for any $\hat{p} \in \hat{\mathscr{P}}_0$. Taking the infimum over $\hat{p}$ as in the Bramble-Hilbert Lemma and using the affine change of variables we get

$$\int_T |r_h(\prod_{j=1}^M u_j) - (\prod_{j=1}^M u_j)|\,dx \le C|\det B|\,|(\prod_{j=1}^M \hat{u}_j)|_{1,1,\hat{T}} \le C\|B\|\,|(\prod_{j=1}^M u_j)|_{1,1,T},$$

where $\|.\|$ is the matrix norm induced by the Euclidian vector norm. Noticing that

$$\|B\| \le Ch_T$$

and that

$$\int_T |\mathbf{grad}(\prod_{j=1}^M u_j)|\,dx \le \sum_{\sigma\in\Delta} \int_T |\mathbf{grad} u_{\sigma_1}|\,|\prod_{\ell=2}^M u_{\sigma_\ell}|\,dx$$
$$\le \sum_{\sigma\in\Delta} \big(\|\mathbf{grad} u_{\sigma_1}\|_{0,T} \prod_{\ell=2}^M \|u_{\sigma_\ell}\|_{0,2(M-1),T}\big)$$

with Δ the set of circular permutations of $\{1,2,\dots,M\}$, we write successively

$$\Big|\int_\Omega \Big(r_h(\prod_{j=1}^M u_j) - (\prod_{j=1}^M u_j)\Big)\,dx\Big|$$
$$\le Ch \sum_{T\in\mathscr{T}_h} \big(\sum_{\sigma\in\Delta} |u_{\sigma_1}|_{1,T} \prod_{\ell=2}^M \|u_{\sigma_\ell}\|_{0,2(M-1),T}\big) \le Ch \sum_{\sigma\in\Delta} |u_{\sigma_1}|_{1,\Omega} \prod_{\ell=2}^M \|u_{\sigma_\ell}\|_{0,2(M-1),\Omega}$$
$$\le \tilde{C}h \sum_{\sigma\in\Delta} \big(\prod_{j=1}^M \|u_{\sigma_j}\|_{1,\Omega}\big) = \tilde{C}hM \prod_{j=1}^M \|u_j\|_{1,\Omega}.$$

and Theorem 7.3 is proved. □

As a consequence of Theorem 6.1, we get the following result.

Theorem 8.4. *In the framework developed above, assume that f is in $W^{1,p}(\Omega)$ for some $p > 2$. Then there exist positive constants h_0, δ_0, C, and for $h \le h_0$ a unique solution $\tilde{u}_h$ to (8.17) (or equivalently to (3.14)) in $\overline{B}(u,\delta_0)$. Moreover the following a priori estimate holds*

$$\|u - \tilde{u}_h\|_{1,\Omega} \le Ch. \tag{8.18}$$

Proof. The result will be a consequence of Theorem 6.1. Thus we shall check the hypotheses (6.1), (6.2), and (6.3). Note that for $v \in H_0^1(\Omega)$, the derivative $DF_h(v) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is given by

$$\text{for all } w_1, w_2 \in H_0^1(\Omega) \quad \langle DF_h(v)w_1, w_2\rangle$$
$$= \int_\Omega \mathbf{grad} w_1\, \mathbf{grad} w_2\,dx + 3\int_\Omega r_h\left[(\pi_h v)^2 \pi_h w_1 \pi_h w_2\right]\,dx,$$

with the notation $\langle .\,,.\rangle = \langle .\,,.\rangle_{H^{-1}(\Omega)\,H_0^1(\Omega)}$.

Let us first check (6.1) with L_h =constant. By the definitions for $v \in H_0^1(\Omega)$ we have

$$\|DF_h(u) - DF_h(v)\|_{H_0^1(\Omega);H^{-1}(\Omega)} = \sup_{\substack{w_1 \in H_0^1(\Omega)\\ |w_1|_{1,\Omega}\le 1}} \sup_{\substack{w_2 \in H_0^1(\Omega)\\ |w_2|_{1,\Omega}\le 1}} \langle [DF_h(u) - DF_h(v)]\, w_1, w_2\rangle.$$

Let v, w_1, w_2 in $H_0^1(\Omega)$, then following the technique used in the proof of Theorem 8.3, we get

$$\begin{aligned}\langle [DF_h(u) - DF_h(v)]\, w_1, w_2\rangle &= 3\int_\Omega r_h\left[\left((\pi_h u)^2 - (\pi_h v)^2\right)\pi_h w_1 \pi_h w_2\right]\, dx\\ &\le C\|\pi_h(u-v)\|_{0,4,\Omega}\|\pi_h(u+v)\|_{0,4,\Omega}\|\pi_h w_1\|_{0,4,\Omega}\|\pi_h w_2\|_{0,4,\Omega}.\end{aligned}$$

Since the $L^4(\Omega)$ norm is bounded by the $H^1(\Omega)$ norm and $\pi_h \in \mathscr{L}(H_0^1(\Omega);H_0^1(\Omega))$ is bounded with respect to h, we finally get for $v \in \overline{B}(u,1)$

$$\|DF_h(u) - DF_h(v)\|_{H_0^1(\Omega);H^{-1}(\Omega)} \le C|u-v|_{1,\Omega}, \tag{8.19}$$

which implies (6.1) with $L_h = C$ and $\eta = 1$.

To check the hypothesis (6.3), we notice that, since $DF(u) \in \mathscr{L}(H_0^1(\Omega);H^{-1}(\Omega))$ is an isomorphism, it is sufficient to check that $\lim_{h\to 0} DF_h(u) = DF(u)$. For $w_1, w_2 \in H_0^1(\Omega)$ we have

$$\begin{aligned}\langle [DF_h(u) - DF(u)]\, w_1, w_2\rangle &= 3\int_\Omega r_h\left[(\pi_h u)^2 \pi_h w_1 \pi_h w_2\right]\, dx\\ &\quad - 3\int_\Omega (\pi_h u)^2 \pi_h w_1 \pi_h w_2\, dx + 3\int_\Omega (\pi_h u)^2 \pi_h w_1 \pi_h w_2\, dx - 3\int_\Omega u^2 w_1 w_2\, dx.\end{aligned}$$

Making use of Theorem 8.3 and of the fact that $\pi_h \in \mathscr{L}(H_0^1(\Omega);H_0^1(\Omega))$ is uniformly bounded with respect to h, we can prove that the first difference in the right hand side of the last equation tends to zero. The second difference reads

$$\begin{aligned}\int_\Omega (\pi_h u)^2 \pi_h w_1 \pi_h w_2\, dx - \int_\Omega u^2 w_1 w_2\, dx &= \int_\Omega (\pi_h u - u)(\pi_h u + u) w_1 w_2\, dx\\ &\quad + \int_\Omega (\pi_h u)^2 (\pi_h w_1 - w_1) w_2\, dx + \int_\Omega (\pi_h u)^2 \pi_h w_1 (\pi_h w_2 - w_2)\, dx\\ \le \|u - \pi_h u\|_{0,4,\Omega}\|u + \pi_h u\|_{0,4,\Omega}\|w_1\|_{0,4,\Omega}\|w_2\|_{0,4,\Omega} &+ \|\pi_h u\|_{0,4,\Omega}^2\|w_1 - \pi_h w_1\|_{0,4,\Omega}\|w_2\|_{0,4,\Omega}\\ &+ \|\pi_h u\|_{0,4,\Omega}^2\|\pi_h w_1\|_{0,4,\Omega}\|\pi_h w_2 - w_2\|_{0,4,\Omega}.\end{aligned}$$

The family of triangulations being regular, we can check that

$$\lim_{h\to 0} \sup_{\substack{w\in H_0^1(\Omega)\\ w\neq 0}} \frac{\|w - \pi_h w\|_{0,4,\Omega}}{|w|_{1,\Omega}} = 0.$$

So the hypothesis (6.3) has been verified.

Let us finally check the hypothesis (6.2). For $v \in H_0^1(\Omega)$ with $|v|_{1,\Omega} \leq 1$

$$\begin{aligned}\langle F_h(u), v\rangle &= \int_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dx + \int_\Omega r_h\left[\left((\pi_h u)^3 - f\right)\pi_h v\right]\, dx \\ &= \int_\Omega \left\{r_h\left[(\pi_h u)^3 \pi_h v\right] - (\pi_h u)^3 \pi_h v\right\}\, dx + \int_\Omega \left[(\pi_h u)^3 - u^3\right]\pi_h v\, dx \\ &\quad - \int_\Omega \left\{r_h\left[f\pi_h v\right] - f\pi_h v\right\}\, dx + \int_\Omega \mathbf{grad}(u - \pi_h u)\, \mathbf{grad} v\, dx.\end{aligned}$$

We study successively all four terms in the last equality. We have

$$\left|\int_\Omega \mathbf{grad}(u - \pi_h u)\, \mathbf{grad} v\, dx\right| \leq |u - \pi_h u|_{1,\Omega} \leq Ch. \tag{8.20}$$

For the third difference we notice that, with $v_h = \pi_h v$,

$$\begin{aligned}\int_\Omega \left[r_h(f v_h) - f v_h\right]\, dx &= \int_\Omega \left[r_h(r_h f v_h) - f v_h\right]\, dx \\ &= \int_\Omega \left[r_h(r_h f v_h) - r_h f v_h\right]\, dx + \int_\Omega (r_h f - f) v_h\, dx\end{aligned}$$

so with Theorem 8.3 we get the estimate

$$\begin{aligned}\left|\int_\Omega \left[r_h(f v_h) - f v_h\right]\, dx\right| &\leq Ch \|r_h f\|_{1,\Omega} \|v_h\|_{1,\Omega} + \|f - r_h f\|_{0,\Omega} \|v_h\|_{0,\Omega} \\ &\leq Ch |v_h|_{1,\Omega}.\end{aligned} \tag{8.21}$$

The second difference can be estimated in the following way

$$\left|\int_\Omega \left[(\pi_h u)^3 - u^3\right]\pi_h v\, dx\right| \leq Ch, \tag{8.22}$$

where C depends on u but is independent of h. As a consequence of Theorem 8.3 an estimate for the first difference reads

$$\left|\int_\Omega \left\{r_h\left[(\pi_h u)^3 \pi_h v\right] - (\pi_h u)^3 \pi_h v\right\}\, dx\right| \leq Ch. \tag{8.23}$$

As a consequence of (8.20) through (8.23), we get

$$\|F_h(u)\|_{-1,\Omega} \leq Ch, \tag{8.24}$$

where C is independent of h (but actually depends on u).

We apply Theorem 6.1 and the estimate (8.24) with (6.5) to conclude. □

For a model problem we have developed a method to get a priori and a posteriori error estimates in the cases of Galerkin approximation without and with numerical integration. The interest of the technique lies in its ability to be adapted to other cases.

9. Estimates for an approximation of the Navier-Stokes problem

The abstract results of Section 7 are applied now to the commonly called $(4\mathscr{P}_1 - \mathscr{P}_1)$ approximation of the Navier-Stokes problem introduced in Section 4.

In this case we are using the following notations: $L^2_0(\Omega) = \{f \in L^2(\Omega); \int_\Omega f\,dx = 0\}$, $X = Y = (H^1_0(\Omega))^2 \times L^2_0(\Omega)$, and $F : X \to X'$ is defined for $\underline{u} = (\mathbf{u}, p) \in X$,

$$\begin{aligned} &\text{for all } \underline{v} = (\mathbf{v}, q) \in X \quad \langle F(\mathbf{u}, p), (\mathbf{v}, q)\rangle_{X'X} = \nu \int_\Omega \mathbf{grad}\,\mathbf{u}\,\mathbf{grad}\mathbf{v}\,dx \\ &\quad + \int_\Omega (\mathbf{u}\,\boldsymbol{\nabla})\mathbf{u}\,\mathbf{v}\,dx - \int_\Omega p\,\mathrm{div}\mathbf{v}\,dx - \int_\Omega \mathbf{f}\mathbf{v}\,dx - \int_\Omega q\,\mathrm{div}\mathbf{u}\,dx \\ &\quad \equiv a(\mathbf{u}, \mathbf{v}) + c(\mathbf{u}; \mathbf{u}, \mathbf{v}) + b(\mathbf{v}, p) + b(\mathbf{u}, q) - \int_\Omega \mathbf{f}\mathbf{v}\,dx. \end{aligned} \tag{9.1}$$

Let $\underline{u} = (\mathbf{u}, p)$ denote a regular solution to $F(\underline{v}) = 0$, in the sense that (4.8) is satisfied.

The approximation spaces and the test spaces are $X_h \equiv Y_h = V_h^2 \times W_h$, where the finite element spaces V_h and W_h are defined in Section 4. A Galerkin approximation of the problem $F(\underline{v}) = 0$ reads (see (4.9)): find $\underline{u}_h \in X_h$ such that

$$\text{for all } \underline{v}_h \in X_h \quad \langle F(\underline{u}_h), \underline{v}_h\rangle_{X'X} = 0. \tag{9.2}$$

Theorem 9.1. *Let $\underline{u} = (\mathbf{u}, p) \in X$ be a solution to $F(\underline{v}) = 0$ with $\mathbf{u} \in (H^2(\Omega))^2$. We assume that the hypothesis (4.8) and the assumption (4.10) on the triangulations are satisfied. Then there exist constants $h_0 > 0$, $\delta_0 > 0$, $C > 0$, and for $h \le h_0$ a unique solution $\underline{u}_h = (\mathbf{u}_h, p_h) \in X_h$ to the problem (9.2) satisfying $\|\underline{u} - \underline{u}_h\|_X \le \delta_0$. Furthermore the following estimates hold*

$$|\mathbf{u} - \mathbf{u}_h|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \le C \inf_{(\mathbf{v}_h, q_h)\in X_h} (|\mathbf{u} - \mathbf{v}_h|_{1,\Omega} + \|p - q_h\|_{0,\Omega}) \tag{9.3}$$

and

$$|\mathbf{u} - \mathbf{u}_h|_{1,\Omega} + \|p - p_h\|_{0,\Omega} \le C \left(\sum_{T\in\mathscr{T}_h^f} \eta(T)^2 \right)^{1/2}, \tag{9.4}$$

where for any triangle $T \in \mathscr{T}_h^f$ the estimator $\eta(T)$ is expressed by

$$\eta(T)^2 = h_T^2 \| - \nu\Delta\mathbf{u}_h + (\mathbf{u}_h\,\boldsymbol{\nabla})\mathbf{u}_h + \mathbf{grad}p_h - \mathbf{f}\|^2_{0,T} + \sum_{i=1}^{3} h_{t_i} \| \left[\frac{\partial \mathbf{u}_h}{\partial n}\right] \|^2_{0,t_i} + \|\mathrm{div}\mathbf{u}_h\|^2_{0,T}$$

with h_T the diameter of T, t_i $i = 1, 2, 3$, the edges of T of length h_{t_i}, and [.] the jump through an edge.

Proof. We will first check that we are in a position to apply Theorem 7.1. Since we assume that $\underline{u} = (\mathbf{u}, p) \in X$ satisfies (4.8), so the hypothesis (7.1) is satisfied. The

assumption (7.4) is verified in Theorem 4.1 while the (7.5) and (7.6) ones are easy to check with β_h independent of h. We can apply Theorem 7.1 with β_h constant.

The estimate (9.3) is immediately deduced from (7.7) while from (7.8) we shall deduce the estimate (9.4). In our case (7.8) reads

$$|\mathbf{u}-\mathbf{u}_h|_{1,\Omega}+\|p-p_h\|_{0,\Omega}\le C\|F(\mathbf{u}_h,p_h)\|_{X'}. \tag{9.5}$$

For $\underline{v}=(\mathbf{v},q)\in X$ and $\underline{v}_h=(\mathbf{v}_h,q_h)\in X_h$ with $\mathbf{v}_h=\tilde{r}_h\mathbf{v}$, where $\tilde{r}_h\in\mathscr{L}((H_0^1(\Omega))^2;V_h^2)$ is the Clément interpolation operator for discontinuous functions, see CLÉMENT [1975], we have

$$\begin{aligned}
&\langle F(\underline{u}_h),\underline{v}\rangle_{X'X}=\langle F(\underline{u}_h),\underline{v}-\underline{v}_h\rangle_{X'X} \qquad (9.6)\\
&=a(\mathbf{u}_h,\mathbf{v}-\mathbf{v}_h)+c(\mathbf{u}_h;\mathbf{u}_h,\mathbf{v}-\mathbf{v}_h)+b(\mathbf{v}-\mathbf{v}_h,p_h)+b(\mathbf{u}_h,q-q_h)-\int_\Omega\mathbf{f}(\mathbf{v}-\mathbf{v}_h)\,dx\\
&=\sum_{T\in\mathscr{T}_h^f}\Big\{\int_T(-\nu\Delta\mathbf{u}_h+(\mathbf{u}_h\boldsymbol{\nabla})\mathbf{u}_h+\mathbf{grad}p_h-\mathbf{f})(\mathbf{v}-\mathbf{v}_h)\,dx\\
&\qquad-\int_T\mathrm{div}\mathbf{u}_h(q-q_h)\,dx+\nu\int_{\partial T}\frac{\partial\mathbf{u}_h}{\partial n}(\mathbf{v}-\mathbf{v}_h)\,ds\Big\}.
\end{aligned}$$

Let us bound each of these terms. Setting $\mathbf{w}=-\nu\Delta\mathbf{u}_h+(\mathbf{u}_h\boldsymbol{\nabla})\mathbf{u}_h+\mathbf{grad}p_h-\mathbf{f}$ on T, we get successively since $\mathbf{v}_h=\tilde{r}_h\mathbf{v}$

$$\int_T\mathbf{w}(\mathbf{v}-\mathbf{v}_h)\,dx\le\|\mathbf{w}\|_{0,T}\|\mathbf{v}-\mathbf{v}_h\|_{0,T}\le C\|\mathbf{w}\|_{0,T}h\sum_{T'\in S_T}\|\mathbf{v}\|_{1,T'}$$

where S_T is the set of triangles T' in $\mathscr{T}_h^f$ with $T'\cap T\neq\emptyset$, and summing over all the elements

$$\begin{aligned}
\Big|\sum_{T\in\mathscr{T}_h^f}\int_T\mathbf{w}(\mathbf{v}-\mathbf{v}_h)\,dx\Big|&\le C\left(\sum_{T\in\mathscr{T}_h^f}h_T^2\|\mathbf{w}\|_{0,T}^2\right)^{1/2}\left(\sum_{T\in\mathscr{T}_h^f}\|\mathbf{v}\|_{1,T}^2\right)^{1/2}\\
&\le\tilde{C}|\mathbf{v}|_{1,\Omega}\left(\sum_{T\in\mathscr{T}_h^f}h_T^2\|\mathbf{w}\|_{0,T}^2\right)^{1/2}.
\end{aligned} \tag{9.7}$$

We use the same techniques as the ones developed in the proof of Theorem 8.1 to estimate

$$\Big|\sum_{T\in\mathscr{T}_h^f}\nu\int_{\partial T}\frac{\partial\mathbf{u}_h}{\partial n}(\mathbf{v}-\tilde{r}_h\mathbf{v})\,ds\Big|\le C|\mathbf{v}|_{1,\Omega}\left(\sum_{t\in S_h}h_t\Big\|\left[\frac{\partial\mathbf{u}_h}{\partial n}\right]\Big\|_{0,t}^2\right)^{1/2} \tag{9.8}$$

where S_h is the set of all sides t of triangles in $\mathscr{T}_h^f$. Since q is only a $L_0^2(\Omega)$ function, we take $q_h = 0$ and so

$$(9.9) \qquad |\sum_{T\in\mathscr{T}_h^f}\int_T \mathrm{div}\mathbf{u}_h(q-q_h)\,dx| \le \|q\|_{0,\Omega}\left(\sum_{T\in\mathscr{T}_h^f}\|\mathrm{div}\mathbf{u}_h\|_{0,T}^2\right)^{1/2}.$$

From the estimates (9.7), (9.8), and (9.9) we get the following bound for (9.6) noticing that $\Delta\mathbf{u}_h = 0$ on T,

$$(9.10) \quad |\langle F(\underline{u}_h),\underline{v}\rangle_{X'X}| \le C\Big\{\sum_{T\in\mathscr{T}_h^f}\Big[h_T^2\|(\mathbf{u}_h\boldsymbol{\nabla})\mathbf{u}_h + \mathbf{grad}p_h - \mathbf{f}\|_{0,T}^2 + \sum_{i=1}^{3} h_{t_i}\|\left[\frac{\partial\mathbf{u}_h}{\partial n}\right]\|_{0,t_i}^2 + \|\mathrm{div}\mathbf{u}_h\|_{0,T}^2\Big]\Big\}^{1/2}\|(\mathbf{v},q)\|_X.$$

The estimates (9.10) and (9.5) imply the error estimate (9.4). □

10. Error estimates for the stationary heat problem

The abstract results of Section 7 are applied now to the approximation (5.18) of the stationary heat problem with convection. Our framework is well-suited to study this effective Galerkin approximation.

We are using the following notations: $X = W_0^{1,p}(\Omega)$, $Y = W_0^{1,q}(\Omega)$ with p, q in $\mathbb{N}$, $p > 2$ and $\frac{1}{p} + \frac{1}{q} = 1$, and $F : X \to Y'$ defined for $v \in W_0^{1,p}(\Omega)$ by for all $w \in W_0^{1,q}(\Omega)$

$$(10.1) \qquad \langle F(v),w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = \int_\Omega k(v)\mathbf{grad}v\,\mathbf{grad}w\,dx + \int_\Omega \mathbf{c}\,\mathbf{grad}v\,w\,dx - \int_\Omega fw\,dx.$$

When the domain Ω is regular, the hypotheses (5.2), (5.3), and (5.4) satisfied, and f in $L^\infty(\Omega)$, then the problem $F(v) = 0$ has a unique solution $u \in W_0^{1,p}(\Omega)$, which is in $W^{2,r}(\Omega)$ with $1 \le r < \infty$ and $DF(u) \in \mathscr{L}(W_0^{1,p}(\Omega);W^{-1,p}(\Omega))$ is an isomorphism, see Theorem 5.2.

For the sake of simplicity, we assume that Ω is a polygonal domain such that the Laplacian operator $\Delta : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ is an isomorphism. Given a family $\{\mathscr{T}_h\}_{0<h\le 1}$ of triangulations, h is the maximum of the diameter of $T \in \mathscr{T}_h$, the approximation spaces and the test spaces are the spaces V_h of continuous piecewise polynomials of degree 1. The family is assumed to be regular and quasi-uniform, see CIARLET [1978] p.132 and p.140. We study the Galerkin approximation: find $u_h \in V_h$ satisfying

$$(10.2) \qquad \text{for all } v_h \in V_h \quad \langle F(u_h),v_h\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = 0.$$

Remark 10.1. In Theorem 5.2, we have assumed that Ω is regular, so that the Laplacian operator with homogeneous boundary conditions is an isomorphism from $W_0^{1,p}(\Omega)$ onto $W^{-1,p}(\Omega)$, see SIMADER [1972]. Here we assume Ω polygonal so that $\overline{\Omega} = \cup_{T\in\mathscr{T}_h} T$ and the Laplacian operator is an isomorphism, see DAUGE [1992]. Actually when Ω is regular, we should use a family of isoparametric triangulations of Ω. $\square$

To apply Theorem 7.1, we need to prove a discrete inf-sup condition for the bilinear form $b : W_0^{1,p}(\Omega) \times W_0^{1,q}(\Omega) \to \mathbb{R}$ given by: for $v \in W_0^{1,p}(\Omega)$, $w \in W_0^{1,q}(\Omega)$

$$b(v,w) = \langle DF(u)v, w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)}.$$

Theorem 10.1. *Under the hypotheses of Theorem 5.1 let $u \in W_0^{1,p}(\Omega)$ be the unique solution to $F(v) = 0$ and $DF(u) \in \mathscr{L}(W_0^{1,p}(\Omega); W^{-1,p}(\Omega))$ be an isomorphism. We assume that the family of triangulations is regular and quasi-uniform. Then the exists $\zeta > 0$ independent of h such that*

$$\inf_{\substack{v_h\in V_h\\ |v_h|_{1,p,\Omega}=1}} \sup_{\substack{w_h\in V_h\\ |w_h|_{1,q,\Omega}=1}} b(v_h, w_h) \geq \zeta. \tag{10.3}$$

Proof. Going back to the definition of $DF(u)$ in (5.12), we have for $w \in W_0^{1,p}(\Omega)$, $\psi \in W_0^{1,q}(\Omega)$

$$\langle DF(u)w, \psi\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = \int_\Omega \mathbf{grad}(k(u)w)\,\mathbf{grad}\psi\,dx + \int_\Omega \mathbf{c}\,\mathbf{grad}w\,\psi\,dx. \tag{10.4}$$

Let $-T \in \mathscr{L}(W^{-1,p}(\Omega); W_0^{1,p}(\Omega))$ denote the inverse of the Laplacian operator with homogeneous boundary condition, which is an isomorphism. Then (10.4) reads for $w \in W_0^{1,p}(\Omega)$

$$TDF(u)w = k(u)w + T(\mathbf{c}\,\mathbf{grad}w). \tag{10.5}$$

If we introduce the mapping $R : \varphi \in W_0^{1,p}(\Omega) \to R\varphi = \varphi/k(u) \in W_0^{1,p}(\Omega)$ wich is an isomorphism, then (10.5) reads with $w = R\varphi$

$$TDF(u)R\varphi = \varphi + T(\mathbf{c}\,\mathbf{grad}(R\varphi)). \tag{10.6}$$

We have proved in Theorem 5.2 that $TDF(u)R \in \mathscr{L}(W_0^{1,p}(\Omega); W_0^{1,p}(\Omega))$ is an isomorphism.

Let $T_h \in \mathscr{L}(W^{-1,p}(\Omega); V_h)$ be the discrete analogue of T defined for $g \in W^{-1,p}(\Omega)$ by

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad}(T_h g)\,\mathbf{grad}v_h\,dx = \int_\Omega g v_h\,dx.$$

We easily check that for $\varphi_h \in V_h$

$$T_h DF(u)R\varphi_h = \varphi_h + T_h(\mathbf{c}\,\mathbf{grad}(R\varphi_h)).$$

The family of triangulations is regular and quasi-uniform, so there exists a constant $C > 0$ such that for all $w \in W^{2,p}(\Omega)$

$$\|w - \pi_h w\|_{1,p,\Omega} \leq Ch\|w\|_{2,p,\Omega},$$

where $\pi_h \in \mathscr{L}(W^{1,p}(\Omega); V_h)$ is the elliptic projector, see RANNACHER and SCOTT [1982].

Since we have $T_h = \pi_h T$, we get for $\varphi_h \in V_h$

$$T_h DF(u)R\varphi_h = TDF(u)R\varphi_h + (T_h - T)(\mathbf{c}\,\mathbf{grad}(R\varphi_h))$$

and so there exists a constant γ_1 independent of h such that

$$|T_h DF(u)R\varphi_h|_{1,p,\Omega} \geq \gamma_1|\varphi_h|_{1,p,\Omega} - \epsilon_h\|\mathbf{c}\,\mathbf{grad}(R\varphi_h)\|_{0,p,\Omega}$$

with ϵ_h tending to zero for h tending to zero, independent of φ_h. So for h small enough

$$|T_h DF(u)R\varphi_h|_{1,p,\Omega} \geq \frac{\gamma_1}{2}|\varphi_h|_{1,p,\Omega}. \tag{10.7}$$

We are working under the assumption that the Laplacian operator $\Delta : W_0^{1,p}(\Omega) \to W^{-1,p}(\Omega)$ is an isomorphism, where for all $v \in W_0^{1,p}(\Omega)$, for all $w \in W_0^{1,q}(\Omega)$

$$\langle \Delta v, w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} = \int_\Omega \mathbf{grad}v\,\mathbf{grad}w\,dx.$$

So we can check there exists a $\gamma_2 > 0$ with

$$\inf_{\substack{v\in W_0^{1,p}(\Omega)\\ |v|_{1,p,\Omega}=1}} \sup_{\substack{w\in W_0^{1,q}(\Omega)\\ |w|_{1,q,\Omega}=1}} \int_\Omega \mathbf{grad}v\,\mathbf{grad}w\,dx \geq \gamma_2 > 0.$$

The elliptic projector $\pi_h \in \mathscr{L}(W_0^{1,q}(\Omega); V_h)$ is stable in $W_0^{1,q}(\Omega)$, which means there exists a constant $C > 0$ such that for all $v \in W_0^{1,q}(\Omega)$

$$|\pi_h v|_{1,q,\Omega} \leq C|v|_{1,q,\Omega},$$

see RANNACHER and SCOTT [1982]. Let v_h be in V_h with $|v_h|_{1,p,\Omega} = 1$. Then there exists a function $w \in W_0^{1,q}(\Omega)$ with $|w|_{1,q,\Omega} = 1$ such that

$$\frac{\gamma_2}{2} \leq \int_\Omega \mathbf{grad}v_h\mathbf{grad}w\,dx = \int_\Omega \mathbf{grad}v_h\,\mathbf{grad}\pi_h w\,dx.$$

So

$$\int_\Omega \mathbf{grad}v_h\,\mathbf{grad}\frac{\pi_h w}{|\pi_h w|_{1,q,\Omega}}\,dx \geq \frac{\gamma_2}{2|\pi_h w|_{1,q,\Omega}}.$$

With the stability result of π_h, we conclude there exists a constant $\gamma_3 > 0$ independent of h such that

$$\inf_{\substack{v_h \in V_h \\ |v_h|_{1,p,\Omega}=1}} \sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} \int_\Omega \mathbf{grad} v_h \, \mathbf{grad} w_h \, dx \geq \gamma_3 . \tag{10.8}$$

Let the function φ_h be in V_h, then with the estimates (10.7) and (10.8) we get

$$\begin{aligned} \sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} b(R\varphi_h, w_h) &= \sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} \langle DF(u) R\varphi_h, w_h \rangle_{W^{-1,p}(\Omega) W_0^{1,q}(\Omega)} \\ &= \sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} \int_\Omega \mathbf{grad}(T_h DF(u) R\varphi_h) \, \mathbf{grad} w_h \, dx \geq \gamma |\varphi_h|_{1,p,\Omega}, \end{aligned}$$

and so for all $\varphi_h \in V_h$

$$\sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} b(R\varphi_h, w_h) \geq \gamma |\varphi_h|_{1,p,\Omega}. \tag{10.9}$$

Using the hypotheses (5.2) and (5.3) on k, the regularity of the triangulations and the regularity of the solution $u \in W^{2,r}(\Omega)$, $1 \leq r < \infty$, we get by standard calculations on the reference element

$$\lim_{h \to 0} \max_{\substack{\chi_h \in V_h \\ |\chi_h|_{1,p,\Omega}=1}} \| R^{-1}\chi_h - r_h(R^{-1}\chi_h) \|_{1,p,\Omega} = 0, \tag{10.10}$$

where r_h is the Lagrange interpolation operator.

Finally the estimates (10.9) and (10.10) imply there exists $\zeta > 0$ such that

$$\inf_{\substack{v_h \in V_h \\ |v_h|_{1,p,\Omega}=1}} \sup_{\substack{w_h \in V_h \\ |w_h|_{1,q,\Omega}=1}} b(v_h, w_h) \geq \zeta$$

and the proof is complete. □

We are in a position to prove the main result of the section.

Theorem 10.2. *We assume the hypotheses of Theorem 5.1 and denote by $u \in W_0^{1,p}(\Omega)$, $2 < p < \infty$, the unique solution to the problem $F(v) = 0$. Furthermore we assume that the family $\{\mathscr{T}_h\}_{0<h\leq 1}$ is regular and quasi-uniform. Then there exist positive constants h_0, δ_0, C, and for $h \leq h_0$ a unique solution u_h to problem (10.2) in the closed ball $\overline{B}(u, \delta_0)$. Moreover the following estimates hold*

$$|u - u_h|_{1,p,\Omega} \leq Ch \tag{10.11}$$

and

$$|u-u_h|_{1,p,\Omega} \le C\left(\sum_{T\in\mathscr{T}_h}\eta(T)^p\right)^{1/p}, \tag{10.12}$$

where for any $T\in\mathscr{T}_h$ *the estimator* $\eta(T)$ *is expressed by*

$$\eta(T) = h_T\|-\operatorname{div}\big(k(u_h)\mathbf{grad}u_h\big)+\mathbf{c}\,\mathbf{grad}u_h - f\|_{0,p,T} + h_T^{1-2/q}\sum_{i=1}^{3} h_{t_i}^{1/q}\Big\|\Big[k(u_h)\frac{\partial u_h}{\partial n}\Big]\Big\|_{0,p,t_i},$$

where h_T *is the diameter of the triangle* $T\in\mathscr{T}_h$, h_{t_i} *is the length of the edge* t_i *of the triangle* T, $i=1,2,3$, *and* $[.]$ *the jump across the considered edge, (we adopt the convention of a zero function outside* $\overline{\Omega}$*).*

Proof. Let us prove that we are in a position to apply Theorem 7.1. The hypothesis (7.1) is verified in Theorems 5.1 and 5.2. The inf-sup condition (7.4) is verified in Theorem 10.1 with $\beta_h=\beta$ independent of h while the hypothesis (7.5) is satisfied. With the classical interpolation result, we can verify (7.6). With the assumptions (5.2) and (5.3) on k, we easily check that DF is Lipschitzian at u.

The hypotheses of Theorem 7.1 are verified. The proof is complete once proved the two following error estimates

$$\inf_{v_h\in V_h}|u-v_h|_{1,p,\Omega}\le Ch \tag{10.13}$$

and

$$\|F(u_h)\|_{-1,p,\Omega}\le C\left(\sum_{T\in\mathscr{T}_h}\eta(T)^p\right)^{1/p}. \tag{10.14}$$

The estimate (10.13) is a consequence of the interpolation result, for $v\in W^{2,p}(\Omega)$,

$$|v-r_hv|_{1,p,\Omega}\le Ch\|v\|_{2,p,\Omega}.$$

Let us estimate now the term $\|F(u_h)\|_{-1,p,\Omega}$ with the same method as the one used in the proof of Theorem 8.1. For $w\in W_0^{1,q}(\Omega)$ and $w_h\in V_h$ we have

$$\begin{aligned}\langle F(u_h),w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} &= \langle F(u_h),w-w_h\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)}\\ &= \sum_{T\in\mathscr{T}_h}\Big\{\int_T k(u_h)\mathbf{grad}u_h\,\mathbf{grad}(w-w_h)\,dx+\int_T\mathbf{c}\,\mathbf{grad}u_h\,(w-w_h)\,dx\\ &\qquad -\int_T f(w-w_h)\,dx\Big\}\\ &= \sum_{T\in\mathscr{T}_h}\Big\{\int_T\big[-\operatorname{div}\big(k(u_h)\mathbf{grad}u_h\big)+\mathbf{c}\,\mathbf{grad}u_h-f\big](w-w_h)\,dx\\ &\qquad +\int_{\partial T}k(u_h)\frac{\partial u_h}{\partial n_T}(w-w_h)\,ds\Big\},\end{aligned}$$

where $\frac{\partial u_h}{\partial n_T}$ is the exterior normal derivative of u_h on the boundary ∂T of T. Using the Hölder inequality we get

$$\langle F(u_h), w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} \leq \tag{10.15}$$
$$\sum_{T\in\mathscr{T}_h}\left\{\| -\operatorname{div}\big(k(u_h)\mathbf{grad}u_h\big) + \mathbf{c}\,\mathbf{grad}u_h - f\|_{0,p,T}\|w-w_h\|_{0,q,T} + \sum_{i=1}^{3}\| \left[k(u_h)\frac{\partial u_h}{\partial n}\right]\|_{0,p,t_i}\|w-w_h\|_{0,q,t_i}\right\}$$

with $\left[\frac{\partial u_h}{\partial n}\right]$ the jump of a normal derivative of u_h on the side t_i of T with the convention $u_h = 0$ outside $\overline{\Omega}$.

We choose in the inequality (10.15) $w_h = \tilde{r}_h w$ where $\tilde{r}_h$ is the Clément interpolation operator, see CLÉMENT [1975]. Let $S_T = \{T' \in \mathscr{T}_h; T'\cap T \neq \emptyset\}$. Then for all functions $v \in W^{1,q}(\cup_{T'\in S_T} T')$ the following interpolation results hold

$$\|v - \tilde{r}_h v\|_{0,q,T} \leq C h_T \sum_{T'\in S_T}\|v\|_{1,q,T'}$$

and

$$\|v - \tilde{r}_h v\|_{1,q,T} \leq C \sum_{T'\in S_T}\|v\|_{1,q,T'}.$$

Using the technique of the reference element and the above interpolation result, it is a simple matter to prove there exists a constant C such that for all $v \in W^{1,q}(\cup_{T'\in S_T} T')$

$$\|v - \tilde{r}_h v\|_{0,q,t_i} \leq C h_{t_i}^{1/q} h_T^{1-2/q}\left(\sum_{T'\in S_T}\|v\|_{1,q,T'}\right).$$

It follows that

$$\langle F(u_h), w\rangle_{W^{-1,p}(\Omega)W_0^{1,q}(\Omega)} \leq C\sum_{T\in\mathscr{T}_h}\left\{h_T\| -\operatorname{div}\big(k(u_h)\mathbf{grad}u_h\big) + \mathbf{c}\,\mathbf{grad}u_h - f\|_{0,p,T} + \sum_{i=1}^{3} h_{t_i}^{1/q}h_T^{1-2/q}\| \left[k(u_h)\frac{\partial u_h}{\partial n}\right]\|_{0,p,t_i}\right\}^p |w|_{1,q,\Omega},$$

which completes the proof. □

Remark 10.2. Actually problem (10.2) is not solved as it is. The integrals are not computed exactly but with numerical quadrature rules. For the sake of simplicity we have analyzed an approximate problem without numerical integration, so that we could apply Theorem 7.1. □

Remark 10.3. We can obtain a priori and a posteriori error estimates in the norm $|.|_{1,\Omega}$ similar to the estimates (10.11) and (10.12) with the method developed in Theorem 6.4. Here we have $X = W_0^{1,p}(\Omega)$ with $p > 2$, $Z = W^{-1,p}(\Omega)$. Then let $\mathscr{X} = H_0^1(\Omega)$ and $\mathscr{Z} = H^{-1}(\Omega)$. If q denotes the conjugate of p, the following inclusions hold

$$X \subset \mathscr{X} \subset W_0^{1,q}(\Omega) \subset L^2(\Omega) \subset Z \subset \mathscr{Z}.$$

Considering the expression (5.12) for $DF(w)$ with $w \in W_0^{1,p}(\Omega)$, $2 < p < \infty$, we easily see that $DF(w) \in \mathscr{L}(W_0^{1,p}(\Omega); W^{-1,p}(\Omega))$ admits a continuous extension $DF(w) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ and the mapping $w \in W_0^{1,p}(\Omega) \to DF(w) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is continuous.

Let us check now that the mapping $DF(u) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is an isomorphism. We can check that the operator $TDF(u)R$ defined in the proof of Theorem 10.1, see (10.6), is a Fredholm operator with index zero from $H_0^1(\Omega)$ into itself. The density of $W_0^{1,p}(\Omega)$ into $H_0^1(\Omega)$ implies that the range of $TDF(u)R$ considered in $H_0^1(\Omega)$ is $H_0^1(\Omega)$ itself. It follows that $TDF(u)R$ is an isomorphism on $H_0^1(\Omega)$, and since $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$ is an isomorphism, $DF(u) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is also an isomorphism.

As a consequence of (10.11), we have

$$\lim_{h \to 0} |u - u_h|_{1,p,\Omega} = 0.$$

We proceed as in the proof of Theorem 6.4 to get if h is small enough, say $h \leq h_0$,

$$|u - u_h|_{1,\Omega} \leq 2 \|DF(u)^{-1}\|_{H^{-1}(\Omega); H_0^1(\Omega)} \|F(u_h)\|_{-1,\Omega}.$$

Such a posteriori estimates have been used in adaptive mesh refinement techniques applied to a two-dimensional regularized Stefan problem, see PICASSO [1992], and yield excellent results.

To derive a priori error estimates in the norm $|.|_{1,\Omega}$, we proceed in the same way as above but with the mapping F_h defined in (7.11). □

CHAPTER IV

PARAMETRIZED NONLINEAR PROBLEMS. APPROXIMATION OF REGULAR SOLUTIONS

Generally computational applications contain parameters. In Section 3 we have presented a parameter dependent problem, see problem (3.18), and we have proved the existence of a solution path. Our goal now is to study approximations of the problem

$$F(\lambda, u) = 0,$$

where X and Z are Banach spaces and F is a mapping defined on $\mathbb{R}^m \times X$ with values in Z. For a given value of the parameter λ the study can be reduced to the one of Sections 6 and 7, where no parameter does occur. But then we can study how the approximation varies with respect to the parameter, which is precisely the object of this chapter, when the solutions are regular.

First we introduce the notions of coordinate spaces and state spaces. Then the case where the parameter space $\mathbb{R}^m$ can be used as the coordinate space is analyzed both for general and Galerkin approximations. The results are applied to the example of Section 3. When the parameter space can no more be used as the coordinate space for the regular solution path, we transform the problem by adding equations. That situation is also analyzed and our results are illustrated with the example developed in Section 3.

11. Orientation

To present parameter dependent problems, it is adequate to work in the following general framework. Let X and Z be two real Banach spaces and $F : X \to Z$ be a C^p mapping with $p \geq 1$. A point $x \in X$ is called regular if

$$\begin{aligned} &\text{(i)} \quad DF(x) \in \mathscr{L}(X; Z) \quad \text{has a finite-dimensional kernel,} \\ &\text{(ii)} \quad DF(x) \in \mathscr{L}(X; Z) \quad \text{is surjective.} \end{aligned} \tag{11.1}$$

In this chapter we are interested to analyze and compute the solution set S of the problem

$$F(x) = 0 \tag{11.2}$$

in a neighborhood of the regular point $x_0 \in X$ satisfying $F(x_0) = 0$. Such a point x_0 is called a regular solution. Here we assume that $\dim \mathrm{Ker}(DF(x_0)) \geq 1$. The case $\dim \mathrm{Ker}(DF(x_0)) = 0$ has been studied in the previous chapter.

Since the dimension of $\mathrm{Ker}(DF(x_0))$ is finite this subspace is direct, that is there exists a closed subspace $X_1 \subset X$ such that $X = \mathrm{Ker}(DF(x_0)) \oplus X_1$. The point x_0 is regular, so the mapping $DF(x_0)_{|X_1} : X_1 \to Z$ is an isomorphism, the inverse of which is called $A \in \mathscr{L}(Z; X_1)$. A precise description of the solution set S in a neighborhood of x_0 is the following.

Theorem 11.1. *There exist a positive number ϵ, a neighborhood U of $x_0 \in X$, and a unique function $g : B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to Z$ with*

$$g(0) = 0,$$
$$S \cap U = \{x(\lambda) = x_0 + \lambda + Ag(\lambda) \quad \textit{for all}\, \lambda \in B(0,\epsilon)\}.$$

Moreover the mapping $x : \lambda \in B(0,\epsilon) \to x(\lambda) = x_0 + \lambda + Ag(\lambda) \in S \cap U$ is a C^p diffeomorphism.

Proof. The affine mapping $H : \mathrm{Ker}(DF(x_0)) \times Z \to X$ is given by

$$H(\lambda, z) = x_0 + \lambda + Az.$$

Then we define the mapping $G : \mathrm{Ker}(DF(x_0)) \times Z \to Z$ by

$$G(\lambda, z) = F(H(\lambda, z)),$$

which is of class C^p. We notice that $G(0,0) = 0$ and that the derivative $D_z G(0,0) = DF(x_0)A \in \mathscr{L}(Z; Z)$ is an isomorphism. We apply the implicit function theorem 2.3. There exist positive constants ϵ_1, η_1, and a unique mapping $g : B(0,\epsilon_1) \subset \mathrm{Ker}(DF(x_0)) \to B(0,\eta_1) \subset Z$, which is of class C^p and such that

$$g(0) = 0,$$
$$\text{for all}\, \lambda \in B(0,\epsilon_1) \quad F(x_0 + \lambda + Ag(\lambda)) = 0.$$

We set $U_1 = H(B(0,\epsilon_1) \times B(0,\eta_1))$ and the uniqueness implies

$$S \cap U_1 = \{x = x_0 + \lambda + Ag(\lambda), \quad \text{for all}\, \lambda \in B(0,\epsilon_1)\}.$$

To prove that the mapping $x(\lambda) = x_0 + \lambda + Ag(\lambda)$ is a C^p diffeomorphism, we remark that for all $\lambda \in B(0,\epsilon_1)$, $F(x(\lambda)) = 0$ and $DF(x(0))Dx(0) = 0$. So $Dx(0) \in \mathscr{L}(\mathrm{Ker}(DF(x_0)); X)$ is an isomorphism onto $\mathrm{Ker}(DF(x_0))$ and we define the mapping $K : B(0,\epsilon_1) \times X_1 \to X$ by

$$K(\lambda, x_1) = x(\lambda) + x_1.$$

Then K is of class C^p and $DK(0,0) \in \mathscr{L}(\mathrm{Ker}(DF(x_0)) \times X_1, X)$ is an isomorphism. From the inverse function theorem 2.2, we deduce that K is a C^p diffeomorphism from $B((0,0),\epsilon) \subset \mathrm{Ker}(DF(x_0)) \times X_1$ onto a neighborhood U of x_0 with $\epsilon \leq \epsilon_1$ and $U \subset U_1$. We can conclude since $K(\lambda, 0) = x(\lambda)$. $\square$

With the decomposition $X = \mathrm{Ker}(DF(x_0)) \oplus X_1$, we get a representation of the solution set close to x_0 in terms of $\lambda \in B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0))$. So we say that $\mathrm{Ker}(DF(x_0))$ is a coordinate space while X_1 is said to be a state space. Then Theorem 11.1 means that if $m = \dim \mathrm{Ker}(DF(x_0))$, the mapping $x(\lambda)$ defines a local coordinate system of the m-dimensional C^p manifold S in a neighborhood of x_0. Clearly the choice of a coordinate space is not unique. A suitable decomposition of X could be any pair of closed subspaces T and X_1 such that

$$X = T \oplus X_1, \quad \dim T = m, \quad X_1 \cap \mathrm{Ker}(DF(x_0)) = \{0\}. \tag{11.3}$$

Further developments of this analysis based on modern differential geometry can be found in Fink and Rheinboldt [1986], [1987], and in Rheinboldt [1986].

We have already noticed in the examples of Section 3 that some quantities are naturally identified as parameters. For example the space X could have the form $X = \mathbb{R}^m \times X_1$ with X_1 a Banach space and the identified parameters in $\mathbb{R}^m$.

In this chapter we shall focus on the following situation. Given a mapping $F : \mathbb{R}^m \times X \to Z$ of class C^p and a regular solution $(\lambda_0, u_0) \in \mathbb{R}^m \times X$ in the solution set $S = \{(\lambda, u); F(\lambda, u) = 0\}$, we study approximations of S in a neighborhood of (λ_0, u_0). First we shall analyze the case when the parameter space $\mathbb{R}^m$, (in fact $\mathbb{R}^m \times \{0\}$), can be used as the coordinate space, see (11.3). In that situation the solution set can be parametrized by $\lambda \in \mathbb{R}^m$ and since (λ_0, u_0) is regular the operator $D_u F(\lambda_0, u_0) \in \mathscr{L}(X; Z)$ is an isomorphism. The second case we shall analyze is when $D_u F(\lambda_0, u_0)$ is no more an isomorphism from X onto Z. Then it is no more possible to describe the solution set of $F(\lambda, u) = 0$ in a neighborhood of (λ_0, u_0) as a regular function of the parameter $\lambda \in \mathbb{R}^m$. Then one needs either to choose a coordinate space (different from $\mathbb{R}^m$) satisfying (11.3) or transform the problem. We decide for the second method. By adding equations we transform the problem to apply the theory developed in the first case. Some further developments on local coordinate systems of S can be found in Rheinboldt [1986], Chapter 4.

12. Approximation of a parametrized solution set

Let m be a positive integer, X and Z be two real Banach spaces. Given a C^1 mapping $F : \mathbb{R}^m \times X \to Z$, we consider the problem to find $(\lambda, x) \in \mathbb{R}^m \times X$ such that

$$F(\lambda, x) = 0. \tag{12.1}$$

We assume that

$$\begin{aligned}&(\lambda_0, x_0) \text{ is a solution to (12.1) and } D_x F(\lambda_0, x_0) \in \mathscr{L}(X; Z) \text{ is an}\\ &\text{isomorphism;}\end{aligned} \tag{12.2}$$

so (λ_0, x_0) is a regular solution to (12.1). We can apply the implicit function theorem and get the following result. There exist a neighborhood I of λ_0 in $\mathbb{R}^m$, a neighborhood U of x_0 in X, and a C^1 mapping $u : I \to U$ such that $\{(\lambda, u(\lambda)); \lambda \in I\}$ is a regular

solution set containing (λ_0, x_0). Moreover if $\{(\lambda, u(\lambda)); \lambda \in I\}$ and $\{(\lambda, \tilde{u}(\lambda)); \lambda \in \tilde{I}\}$ are two regular solution sets containing (λ_0, x_0), then $u(\lambda)$ and $\tilde{u}(\lambda)$ coincide on the connected component of $I \cap \tilde{I}$ containing λ_0.

Note that in the case $m = 1$, the terminology solution branch is often used.

We introduce now a family $\{F_h\}_{0<h\leq 1}$ of C^1 mappings

$$F_h : (\lambda, x) \in \mathbb{R}^m \times X \to F_h(\lambda, x) \in Z,$$

which are approximations of F. Our goal is to study the existence and the convergence of solutions of the problem: find $(\lambda, x_h) \in \mathbb{R}^m \times X$ such that

$$F_h(\lambda, x_h) = 0 \tag{12.3}$$

in a neighborhood of the regular solution set $\{(\lambda, u(\lambda)); \lambda \in I\}$.

For a fixed λ, we consider a solution $(\lambda, u) \in \mathbb{R} \times X$ of (12.1) and associate an element $\tilde{u}_h \in X$, $h \in (0, 1]$. In some cases it will be sufficient to choose $\tilde{u}_h = u$ but in other ones an appropriate choice of $\tilde{u}_h$ will lead to different error estimates.

Now for that fixed λ, the theory developed in Section 6 can be applied when $\tilde{u}_h = u$. Here we will work under analogous hypotheses of equicontinuity, consistency and stability. We assume that the mapping $D_x F_h(\lambda, .)$ is Lipschitzian at $\tilde{u}_h$ with constant $L(= L(\lambda))$, that is there exist $\eta > 0$ and L such that for all $v \in \overline{B}(\tilde{u}_h, \eta)$, for all $h \in (0, 1]$,

$$\|D_x F_h(\lambda, \tilde{u}_h) - D_x F_h(\lambda, v)\|_{X;Z} \leq L\|\tilde{u}_h - v\|_X. \tag{12.4}$$

Moreover we assume that

$$\lim_{h \to 0} \|F_h(\lambda, \tilde{u}_h)\|_Z = 0 \tag{12.5}$$

and for all $h \in (0, 1]$, $D_x F_h(\lambda, \tilde{u}_h)$ is an isomorphism from X onto Z with an inverse uniformly bounded in h, that is there exists $C_1(= C_1(\lambda))$ such that for all $h \in (0, 1]$

$$\|D_x F_h(\lambda, \tilde{u}_h)^{-1}\|_{Z;X} \leq C_1. \tag{12.6}$$

Theorem 12.1. *Let $\{F_h\}_{0<h\leq 1}$ be a family of C^1 mappings, $F_h : \mathbb{R}^m \times X \to Z$. For a given $\lambda \in \mathbb{R}^m$, we associate an element $\tilde{u}_h \in X$, $h \in (0, 1]$. We assume that the hypotheses (12.4), (12.5), and (12.6) hold. Then there exist $h_0 > 0$, $\delta_0 > 0$, and for all $0 < h \leq h_0$ a unique $u_h \in X$ satisfying*

$$F_h(\lambda, u_h) = 0 \qquad \text{and} \qquad \|u_h - \tilde{u}_h\|_X \leq \delta_0.$$

Moreover the mapping $D_x F_h(\lambda, u_h) \in \mathscr{L}(X; Z)$ is invertible with the bound

$$\|D_x F_h(\lambda, u_h)^{-1}\|_{Z;X} \leq 2\|D_x F_h(\lambda, \tilde{u}_h)^{-1}\|_{Z;X},$$

and

$$\|\tilde{u}_h - u_h\|_X \leq 2\|D_x F_h(\lambda, \tilde{u}_h)^{-1}\|_{Z;X}\|F_h(\lambda, \tilde{u}_h)\|_Z. \tag{12.7}$$

Proof. The proof goes along the same lines as the one of Theorem 6.1. For the given $\lambda \in I$, we set

$$\begin{aligned}
\epsilon_h &= \|F_h(\lambda, \tilde{u}_h)\|_Z, \\
\gamma_h &= \|D_x F_h(\lambda, \tilde{u}_h)^{-1}\|_{Z;X}, \\
\tilde{L}_h(\alpha) &= \sup_{v \in \overline{B}(\tilde{u}_h, \alpha)} \|D_x F_h(\lambda, \tilde{u}_h) - D_x F_h(\lambda, v)\|_{X;Z}.
\end{aligned}$$

From the stability hypothesis (12.6), we get $\gamma_h \leq C_1$. We set $\delta_0 = \min(\eta, (2C_1 L)^{-1})$. Then from the Lipschitzian condition (12.4) we have

$$\tilde{L}_h(\delta_0) \leq (2C_1)^{-1}.$$

Now from the consistency (12.5), there exists $h_0 \leq 1$ such that for all $h \in (0, h_0]$

$$2C_1\epsilon_h \leq \delta_0.$$

Thus for $h \leq h_0$

$$2\gamma_h \tilde{L}_h(2\gamma_h\epsilon_h) \leq 2C_1\tilde{L}_h(\delta_0) \leq 1.$$

We apply Theorem 2.1 with $G = F_h(\lambda, .)$, $v = \tilde{u}_h \in X$. For all $h \leq h_0$, there exists a unique $u_h \in \overline{B}(\tilde{u}_h, 2\gamma_h\epsilon_h)$ satisfying $F_h(\lambda, u_h) = 0$, and $D_x F_h(\lambda, u_h) \in \mathscr{L}(X; Z)$ is invertible with the bound

$$\|D_x F_h(\lambda, u_h)^{-1}\|_{Z;X} \leq 2\gamma_h.$$

The following estimate holds

$$\|\tilde{u}_h - u_h\|_X \leq 2\gamma_h\|F_h(\lambda, \tilde{u}_h)\|_Z.$$

To get the uniqueness of u_h in $\overline{B}(\tilde{u}_h, \delta_0)$ for all $h \in (0, h_0]$, we use the argument developed in Remark 2.1. □

If we choose in Theorem 12.1, $\tilde{u}_h = u$ where (λ, u) is a solution of (12.1), then the inequality (12.7) will lead to a priori error estimates for $\|u - u_h\|_X$. To get a posteriori error estimates, we use the following result.

Theorem 12.2. *Let (λ, u) be a solution of problem (12.1). We assume that the family $\{F_h\}_{0<h\leq 1}$ of C^1 mappings satisfies the hypotheses (12.4), (12.5), and (12.6) with $\tilde{u}_h = u$. Then there exists $0 < \overline{h}_0 \leq h_0$ such that for all $h \leq \overline{h}_0$ the mapping $D_x F(\lambda, u_h) \in \mathscr{L}(X; Z)$ is invertible with a uniformly bounded inverse and*

$$\|u - u_h\|_X \leq 2\|D_x F(\lambda, u_h)^{-1}\|_{Z;X}\|F(\lambda, u_h)\|_Z, \tag{12.8}$$

where h_0 and u_h are given in Theorem 12.1.

Proof. The proof is identical to the one of Theorem 6.2. It is an application of Theorem 2.1 with $G = F(\lambda, .)$ and $v = u$. □

Actually in the theorems 12.1 and 12.2 we have worked with the parameter $\lambda \in I$ fixed. We study now the dependence in λ of u_h.

The mappings $D_x F_h(\lambda, .)$ are assumed to be Lipschitzian at $\tilde{u}_h(\lambda)$, that is there exist $\eta > 0$, $L > 0$ such that for all $0 < h \leq 1$, $\lambda \in I$, and $v \in \overline{B}(\tilde{u}_h(\lambda), \eta)$

$$\|D_x F_h(\lambda, \tilde{u}_h(\lambda)) - D_x F_h(\lambda, v)\|_{X;Z} \leq L\|\tilde{u}_h(\lambda) - v\|_X; \tag{12.9}$$

moreover we assume the consistency condition

$$\lim_{h\to 0} \sup_{\lambda\in I} \|F_h(\lambda, \tilde{u}_h(\lambda))\|_Z = 0 \tag{12.10}$$

and the stability condition: there exists a constant C such that for all $0 < h \leq 1$, $\lambda \in I$,

$$\|D_x F_h(\lambda, \tilde{u}_h(\lambda))^{-1}\|_{Z;X} \leq C. \tag{12.11}$$

Theorem 12.3. *Let $\{(\lambda, u(\lambda)); \lambda \in I\}$ be a set of regular solutions of the problem (12.1). We assume that I is compact, the mappings $\tilde{u}_h : \lambda \in I \to \tilde{u}_h(\lambda) \in X$ are continuous, and the hypotheses (12.9), (12.10), (12.11) hold. Then there exist $h_0 > 0$, $\delta_0 > 0$, and for all $h \leq h_0$ a C^1 mapping $u_h : \lambda \in I \to u_h(\lambda) \in X$ such that*

$$\big(F_h(\lambda, v) = 0 \quad \text{and} \quad v \in \overline{B}(\tilde{u}_h(\lambda), \delta_0)\big) \Leftrightarrow v = u_h(\lambda).$$

Moreover the mappings $D_x F_h(\lambda, \tilde{u}_h(\lambda))$ and $D_x F(\lambda, u_h(\lambda)) \in \mathscr{L}(X; Z)$ are isomorphisms with uniformly bounded inverse and for all $\lambda \in I$

$$\|u_h(\lambda) - \tilde{u}_h(\lambda)\|_X \leq 2\|D_x F_h(\lambda, \tilde{u}_h(\lambda))^{-1}\|_{Z;X}\|F_h(\lambda, \tilde{u}_h(\lambda))\|_Z, \tag{12.12}$$

$$\|u_h(\lambda) - u(\lambda)\|_X \leq 2\|D_x F(\lambda, u_h(\lambda))^{-1}\|_{Z;X}\|F(\lambda, u_h(\lambda))\|_Z. \tag{12.13}$$

Proof. For each $\lambda \in I$ we can apply Theorem 12.1. So for each $\lambda \in I$, there exist $\delta_{0,\lambda}$, $h_{0,\lambda}$, and $u_{h,\lambda} \in X$ for $h \in (0, h_{0,\lambda}]$ satisfying to

$$F_h(\lambda, u_{h,\lambda}) = 0 \quad \text{and} \quad \|u_{h,\lambda} - \tilde{u}_h(\lambda)\|_X \leq \delta_{0,\lambda}$$

and

$$\|\tilde{u}_h(\lambda) - u_{h,\lambda}\|_X \leq 2\|D_x F_h(\lambda, \tilde{u}_h(\lambda))^{-1}\|_{Z;X}\|F_h(\lambda, \tilde{u}_h(\lambda))\|_Z.$$

Furthermore thanks to the uniformity in λ of the assumptions (12.9), (12.10), and (12.11), it is immediate to check that $\delta_{0,\lambda}$ and $h_{0,\lambda}$ can be chosen independently of $\lambda \in I$. So for $h \leq h_0$, we can define the function $u_h : \lambda \in I \to u_h(\lambda) = u_{h,\lambda} \in X$.

To prove that the function $u_h(.)$ is C^1 in I, we apply the implicit function theorem 2.3 taking into account the uniqueness of $u_h(\lambda)$ in $\overline{B}(\tilde{u}_h(\lambda), \delta_0)$.

The estimate (12.13) is a consequence of Theorem 12.2 only to restrict h_0 if necessary. □

Remark 12.1. The assumption (12.11) can be replaced by

$$\lim_{h\to 0} \sup_{\lambda\in I} \|D_x F_h(\lambda, \tilde{u}_h(\lambda)) - D_x F(\lambda, u(\lambda))\|_{X;Z} = 0. \tag{12.11'}$$

From the equality

$$D_x F_h(\lambda, \tilde{u}_h(\lambda)) = \big[D_x F_h(\lambda, \tilde{u}_h(\lambda)) - D_x F(\lambda, u(\lambda))\big] + D_x F(\lambda, u(\lambda))$$

and the estimate (12.11'), we deduce from (1.1) that for h small enough the mapping $D_x F_h(\lambda, \tilde{u}_h(\lambda)) \in \mathscr{L}(X; Z)$ is an isomorphism with an inverse uniformly bounded in h and λ. □

If we assume that the mappings F_h are of class C^p with $p \geq 1$, then $u_h(.)$ is also of class C^p. Then error estimates for the derivatives can also be deduced.

Theorem 12.4. *If the assumptions of Theorem 12.3 hold and the mappings F_h are of class C^p with $p \geq 1$, then $u_h(.)$ is of class C^p in I. Moreover if the functions $\tilde{u}_h(.)$ are of class C^p then there exist constants C_j, $j = 0, \ldots, p$, depending only on the quantities*

$$\max_{\lambda\in I} \max_{v\in\overline{B}(\tilde{u}_h(\lambda),\delta_0)} \|D^k F_h(\lambda, v)\| \quad \text{and} \quad \max_{\lambda\in I} \|D^k \tilde{u}_h(\lambda)\|, \quad 0 \leq k \leq j$$

such that for all $\lambda \in I$

$$\|D^j u_h(\lambda) - D^j \tilde{u}_h(\lambda)\| \leq C_j \sum_{k=0}^{j} \|\frac{d^k}{d\lambda^k} F_h(\lambda, \tilde{u}_h(\lambda))\| \tag{12.14}$$

for $j = 0, \ldots, p-1$ and

$$\|D^p u_h(\lambda)\| \leq C_p. \tag{12.15}$$

For simplicity $\|.\|$ denotes the norm in the different spaces.

Proof. From the implicit function theorem we deduce that u_h is of class C^p in I. We have to prove yet the estimates (12.14) and (12.15). The differentiation of the relation $F_h(\lambda, u_h(\lambda)) = 0$ for $\lambda \in I$ gives

$$Du_h(\lambda) = -D_x F_h(\lambda, u_h(\lambda))^{-1} D_\lambda F_h(\lambda, u_h(\lambda)),$$

which leads to the estimate

$$\|Du_h(\lambda)\| \le 2\|D_xF_h(\lambda,\tilde u_h(\lambda))^{-1}\|\|D_\lambda F_h(\lambda,u_h(\lambda))\|;$$

we notice that the estimate for $\|D_xF_h(\lambda,u_h(\lambda))^{-1}\|$ is given in Theorem 12.1. This last estimate and the (12.12) one give the estimates (12.14) and (12.15) with $p=1$.

Assuming the result holds for ℓ, $1\le\ell<p$, that is

$$\|D^j u_h(\lambda) - D^j\tilde u_h(\lambda)\| \le C_j\sum_{k=0}^{j}\|\frac{d^k}{d\lambda^k}F_h(\lambda,\tilde u_h(\lambda))\| \tag{12.16}$$

for $j=0,\dots,\ell-1$ and

$$\|D^\ell u_h(\lambda)\| \le C'_\ell, \tag{12.17}$$

we prove it for the index value $\ell+1$. We differentiate ℓ times with respect to λ the function $F_h(\lambda,\tilde u_h(\lambda))$, so

$$\frac{d^\ell}{d\lambda^\ell}F_h(\lambda,\tilde u_h(\lambda)) = D_xF_h(\lambda,\tilde u_h(\lambda))D^\ell\tilde u_h(\lambda) + R_\ell(\lambda,\tilde u_h(\lambda),\dots,D^{\ell-1}\tilde u_h(\lambda)) \tag{12.18}$$

where R_ℓ has the form

$$R_1(\lambda,v) = D_\lambda F_h(\lambda,v),$$

$$R_\ell(\lambda,v,\dots,D^{\ell-1}v) = \sum_{i=2}^{\ell}\sum_{\substack{k_1+k_2+\cdots+k_i=\ell\\ k_1,\dots,k_i\ge1}}\alpha_{ik}D^iF_h(\lambda,v)\left((\lambda,v)^{(k_1)},\dots,(\lambda,v)^{(k_i)}\right)$$

with

$$\alpha_{ik}\in\mathbb{R}\qquad\text{and}\qquad(\lambda,v)^{(k_j)} = (D^{k_j}\lambda, D^{k_j}v).$$

We derive now ℓ times with respect to λ the relation $F_h(\lambda,u_h(\lambda))=0$ so

$$0 = D_xF_h(\lambda,u_h(\lambda))D^\ell u_h(\lambda) + R_\ell(\lambda,u_h(\lambda),\dots,D^{\ell-1}u_h(\lambda)). \tag{12.19}$$

From both expressions (12.18) and (12.19), we get

$$\begin{aligned} D^\ell\tilde u_h(\lambda) - D^\ell u_h(\lambda) = D_xF_h(\lambda,u_h(\lambda))^{-1}\Big\{&\frac{d^\ell}{d\lambda^\ell}F_h(\lambda,\tilde u_h(\lambda))\\ &+\big[D_xF_h(\lambda,u_h(\lambda)) - D_xF_h(\lambda,\tilde u_h(\lambda))\big]D^\ell\tilde u_h(\lambda) - R_\ell(\lambda,\tilde u_h(\lambda),\dots,D^{\ell-1}\tilde u_h(\lambda))\\ &+R_\ell(\lambda,u_h(\lambda),\dots,D^{\ell-1}u_h(\lambda))\Big\}. \end{aligned} \tag{12.20}$$

The mapping R_ℓ is Lipschitzian, so we deduce from (12.20) and (12.16) the estimate (12.14) for ℓ.

The estimate (12.15) for the index value $\ell+1$ is immediately deduced from the formula

$$D^{\ell+1}u_h(\lambda) = -D_xF_h(\lambda,u_h(\lambda))^{-1}R_{\ell+1}(\lambda,u_h(\lambda),\dots,D^\ell u_h(\lambda)). \quad\square$$

Remark 12.2. The method presented in Theorem 6.4 to derive error estimates in some other norms than the norm of X can be immediately generalized to parameter dependent problems. $\square$

13. Galerkin approximation of a parametrized solution set

The goal of the present section is to transcribe the results of the previous one in the particular case of Galerkin approximations. For simplicity we will consider problem (12.1) with $m = 1$.

Let X, Y be two real Banach spaces, and let Y' denote the dual space of Y. Given a C^1 mapping $F : \mathbb{R} \times X \to Y'$, we shall analyze Galerkin approximations of the problem

$$F(\lambda, x) = 0 \tag{13.1}$$

in a neighborhood of the regular solution $(\lambda_0, x_0) \in \mathbb{R} \times X$, such that (λ_0, x_0) satisfies

$$F(\lambda_0, x_0) = 0 \text{ and } D_x F(\lambda_0, x_0) \in \mathscr{L}(X; Y') \text{ is an isomorphism.} \tag{13.2}$$

As a consequence of the hypothesis (13.2), we get the existence of a regular solution branch $\{(\lambda, u(\lambda)); \lambda \in I\}$ containing (λ_0, x_0), where the mapping $u : I \to X$ is of class C^1.

Let now $\{X_h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces of X and $\{Y_h\}_{0<h\leq 1}$ be a family of finite-dimensional subspaces of Y. A Galerkin approximation of the problem (13.1) consists in finding $\lambda \in \mathbb{R}$ and $x_h \in X_h$ such that

$$\text{for all } y_h \in Y_h \quad \langle F(\lambda, x_h), y_h \rangle_{Y'Y} = 0. \tag{13.3}$$

For each $\lambda \in I$, we define the bilinear form $b : X \times Y \to \mathbb{R}$ by

$$\text{for all } x \in X, y \in Y \quad b(x, y) = \langle D_x F(\lambda, u(\lambda))x, y \rangle_{Y'Y}. \tag{13.4}$$

Although we do not write it explicitly for notation simplicity, b is depending on λ. We assume that there exists a constant $\beta > 0$ such that for all $h \in (0, 1]$ and $\lambda \in I$

$$\inf_{\substack{x \in X_h \\ \|x\|_X = 1}} \sup_{\substack{y \in Y_h \\ \|y\|_Y = 1}} b(x, y) \geq \beta > 0 \tag{13.5}$$

and

$$\dim X_h = \dim Y_h. \tag{13.6}$$

Moreover we assume that for all $\lambda \in I$

$$\lim_{h \to 0} \inf_{x_h \in X_h} \|u(\lambda) - x_h\|_X = 0. \tag{13.7}$$

Note that the hypotheses (13.5) and (13.6) are standard stability conditions in the linear case while the (13.7) one is a standard approximation property of X by X_h.

The following theorem presents the convergence results in the case of Galerkin approximation.

Theorem 13.1. *Let $F : \mathbb{R} \times X \to Y'$ be a C^1 mapping and $\{(\lambda, u(\lambda)); \lambda \in I \subset \mathbb{R}, I$ compact interval$\}$ be a regular solution branch. We assume that there exist $L > 0$ and $\eta > 0$ such that for all $\lambda \in I$ and $v \in \overline{B}(u(\lambda), \eta)$*

$$\|D_x F(\lambda, u(\lambda)) - D_x F(\lambda, v)\|_{X;Y'} \leq L\|u(\lambda) - v\|_X.$$

Then under the assumptions (13.5), (13.6), and (13.7), there exist positive constants h_0, δ_0, and for all $\lambda \in I$ and $h \in (0, h_0]$ a unique solution $(\lambda, u_h(\lambda)) \in \mathbb{R} \times X_h$ to problem (13.3) in $I \times \overline{B}(u(\lambda), \delta_0)$. The mapping $u_h : \lambda \in I \to X$ is of class C^1 and for $\lambda \in I$, $D_x F(\lambda, u_h(\lambda)) \in \mathscr{L}(X, Y')$ is an isomorphism. Furthermore there exists a constant $C > 0$ (independent of h and λ) such that for all $h \in (0, h_0]$ and $\lambda \in I$

$$\|u(\lambda) - u_h(\lambda)\|_X \leq C \inf_{x_h \in X_h} \|u(\lambda) - x_h\|_X \tag{13.8}$$

and

$$\|u(\lambda) - u_h(\lambda)\|_X \leq 2\|D_x F(\lambda, u_h(\lambda))^{-1}\|_{Y';X}\|F(\lambda, u_h(\lambda))\|_{Y'}. \tag{13.9}$$

Proof. The proof is analogous to the one of Theorem 7.1 and will be deduced from Theorems 12.1 and 12.2. To define the mapping $F_h : I \times X \to Y'$, it is useful to introduce for each $\lambda \in I$ the two projection operators $\Pi_{X_h} \in \mathscr{L}(X; X_h)$ and $\Pi_{Y_h} \in \mathscr{L}(Y; Y_h)$ (depending on λ) given by

$$\text{for all } y_h \in Y_h \quad b(x - \Pi_{X_h} x, y_h) = 0 \tag{13.10}$$

and

$$\text{for all } x_h \in X_h \quad b(x_h, y - \Pi_{Y_h} y) = 0. \tag{13.11}$$

Under the assumptions (13.5) and (13.6) both operators Π_{X_h} and Π_{Y_h} are well defined. Then we construct the family $\{F_h\}_{0<h\leq 1}$ of mappings $F_h : I \times X \to Y'$ by: for $\lambda \in I$ and for $x \in X$, $y \in Y$,

$$\langle F_h(\lambda, x), y\rangle_{Y'Y} = \langle F(\lambda, x), \Pi_{Y_h} y\rangle_{Y'Y} + b(x, y - \Pi_{Y_h} y).$$

Then with the arguments developed in the proof of Theorem 7.1 it is a simple matter to check that problem (13.3) with $\lambda \in I$ is equivalent to find $(\lambda, x_h) \in I \times X_h$ such that

$$F_h(\lambda, x_h) = 0. \tag{13.12}$$

In the same way as in the proof of Theorem 7.1, there exist constants h_0, δ_0 and for $h \in (0, h_0]$ a unique solution $u_h \in \overline{B}(u(\lambda), \delta_0)$ to $F_h(\lambda, x_h) = 0$. Moreover the estimates (13.8) and (13.9) are got from (12.12) and (12.13).

Only to restrict h_0 if necessary, we can apply the implicit function theorem 2.3 and the above uniqueness result to prove that $u_h : \lambda \in I \to u_h(\lambda) \in X$ is a C^1 mapping. □

Remark 13.1. The duality argument used in Remark 7.3 is also valid here to get error estimates in some other norms than the norm of X. The generalization to the parameter dependent problem is immediate. □

Remark 13.2. If the mapping F is of class C^p with $p \geq 1$, then under the assumptions of Theorem 13.1, the conclusions of Theorem 12.4 are true with $\tilde{u}_h(\lambda) = u(\lambda)$ of class C^p. More precisely the function $u_h(.)$ is of class C^p in I and there exist constants C_j, $j = 0, \ldots, p$, depending only on the quantities

$$\max_{\lambda \in I} \max_{v \in \overline{B}(u(\lambda), \delta_0)} \|D^k F_h(\lambda, v)\| \quad \text{and} \quad \max_{\lambda \in I} \|D^k u(\lambda)\|, \quad 0 \leq k \leq j$$

such that for all $\lambda \in I$

$$\|D^j u_h(\lambda) - D^j u(\lambda)\| \leq C_j \sum_{k=0}^{j} \|\frac{d^k}{d\lambda^k} F_h(\lambda, u(\lambda))\| \tag{13.13}$$

for $j = 0, \ldots, p-1$ and

$$\|D^p u_h(\lambda)\| \leq C_p. \tag{13.14}$$

For simplicity $\|.\|$ denotes the norm in the different spaces. □

14. Error estimates for a parameter dependent semilinear problem

An application of the abstract results obtained in the two previous sections to the parameter dependent problems developed in Section 3 is presented here. First we shall analyze a simple case of the general problem (3.18).

Let Ω be a polygonal convex open bounded set in $\mathbb{R}^2$ and let $f : H_0^1(\Omega) \to L^2(\Omega)$ be given by: for $v \in H_0^1(\Omega)$

$$\text{for all} \quad x \in \Omega \qquad f(v)(x) = v^2(x) + 1. \tag{14.1}$$

The mapping $F : \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega)$ is defined by: for all $(\lambda, v) \in \mathbb{R} \times H_0^1(\Omega)$, $w \in H_0^1(\Omega)$

$$\langle F(\lambda, v), w \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = \int_\Omega \mathbf{grad} v \, \mathbf{grad} w \, dx - \lambda \int_\Omega f(v) w \, dx. \tag{14.2}$$

Since the mapping f is C^∞, so is F.

There exist a positive number λ^* and a C^0 mapping $u : [0, \lambda^*] \to u(\lambda) \in H_0^1(\Omega)$, which is C^∞ on $[0, \lambda^*)$, of minimal positive solutions to the problem

$$F(\lambda, v) = 0, \tag{14.3}$$

see Remark 3.2. Moreover for each $\lambda \in [0, \lambda^*)$, the solution $(\lambda, u(\lambda))$ to (14.3) is regular, in fact $D_v F(\lambda, u(\lambda)) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ given by: for all $w_1, w_2 \in H_0^1(\Omega)$

$$\langle D_v F(\lambda, u(\lambda)) w_1, w_2 \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = \int_\Omega \mathbf{grad} w_1 \, \mathbf{grad} w_2 \, dx - 2\lambda \int_\Omega u(\lambda) w_1 w_2 \, dx,$$

is an isomorphism. This means that the assumption (13.2) is satisfied for $\lambda \in [0, \lambda^*)$.

Given a regular family $\{\mathscr{T}_h\}_{0<h\leq 1}$ of triangulations of $\overline{\Omega}$, we consider the corresponding family $\{V_h\}_{0<h\leq 1}$ of finite element subspaces of degree 1,

$$V_h = \{v \in C^0(\overline{\Omega}); v_{|T} \in \mathscr{P}_1 \text{ for all } T \in \mathscr{T}_h\} \cap H_0^1(\Omega). \tag{14.4}$$

A Galerkin approximation to the problem (14.3) consists in finding $\lambda \in \mathbb{R}$ and $v_h \in V_h$ such that

$$\text{for all} \quad w_h \in V_h \qquad \langle F(\lambda, v_h), w_h \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = 0. \tag{14.5}$$

To study the discrete problem (14.5), we shall apply the results of Section 13, with $X = Y = H_0^1(\Omega)$, F given in (14.2), $X_h = Y_h = V_h$ and for $\lambda \in I = [0, \overline{\lambda}]$ with $0 < \overline{\lambda} < \lambda^*$, the bilinear form $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ given by

$$\text{for all} \quad w_1, w_2 \in H_0^1(\Omega) \qquad b(w_1, w_2) = \int_\Omega \mathbf{grad} w_1 \, \mathbf{grad} w_2 \, dx - 2\lambda \int_\Omega u(\lambda) w_1 w_2 \, dx.$$

The set $\{(\lambda, u(\lambda)); \lambda \in I\}$ is a regular solution branch. The mapping F being of class C^k for any k, the derivative $D_v F(\lambda, .)$ is clearly Lipschitzian at $u(\lambda)$. To check the inf-sup condition, we notice that for $w_h \in V_h$

$$\begin{aligned} &\int_\Omega \mathbf{grad} w_h \, \mathbf{grad} w_h \, dx - 2\lambda \int_\Omega u(\lambda) w_h w_h \, dx \\ &\quad \geq (1 - \frac{\lambda}{\lambda^*}) \int_\Omega |\mathbf{grad} w_h|^2 \, dx + \frac{\lambda}{\lambda^*} \left[\int_\Omega |\mathbf{grad} w_h|^2 \, dx - 2\lambda^* \int_\Omega u(\lambda) w_h^2 \, dx \right] \\ &\quad \geq (1 - \frac{\lambda}{\lambda^*}) \int_\Omega |\mathbf{grad} w_h|^2 \, dx \end{aligned}$$

since the expression inside the brackets is positive, see the proof of Theorem 3.4. This implies the condition (13.5) with $\beta = (\lambda^* - \overline{\lambda})/\lambda^*$, while the (13.6) one is immediate with our choice of subspaces $X_h = Y_h = V_h$. Finally the assumption (13.7) is a standard polynomial interpolation result. We can apply Theorem 13.1 and get the following result.

Theorem 14.1. *Let $F : \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega)$ be the mapping defined in (14.2). The set $\{(\lambda, u(\lambda)); \lambda \in I = [0, \overline{\lambda}]\}$ with $\overline{\lambda} < \lambda^*$ denotes a set of minimal positive regular solutions to the problem (14.3). Let the finite element subspaces V_h be given in (14.4).*

Then there exist positive constants h_0, δ_0, C, and for all $\lambda \in I$, $h \in (0, h_0]$, a solution $(\lambda, u_h(\lambda)) \in \mathbb{R} \times V_h$ to the Galerkin approximation (14.5); for $\lambda \in I$, $u_h(\lambda)$ is unique in the closed ball $\overline{B}(u(\lambda), \delta_0)$. Moreover the mapping $u_h : \lambda \in I \to H_0^1(\Omega)$ is of class C^1 and the following error estimates hold: for all $h \in (0, h_0]$ and $\lambda \in I$

$$|u(\lambda) - u_h(\lambda)|_{1,\Omega} \leq Ch \tag{14.6}$$

and

$$|u(\lambda) - u_h(\lambda)|_{1,\Omega} \leq C\Big(\sum_{T \in \mathcal{T}_h} \eta(T)^2\Big)^{1/2}, \tag{14.7}$$

where the local estimator $\eta(T)$ is

$$\eta^2(T) = h_T^2 \|\Delta u_h(\lambda) + \lambda\left(u_h^2(\lambda) + 1\right)\|_{0,T}^2 + \sum_{i=1}^{3} h_{t_i} \left\| \left[\frac{\partial u_h(\lambda)}{\partial n}\right] \right\|_{0,t_i}^2;$$

here h_T is the diameter of T, t_i $i = 1, 2, 3$ are the edges of T of length h_{t_i}, and $[.]$ denotes the jump through an edge.

Proof. The hypotheses of Theorem 13.1 were checked before with $\beta = 1 - \overline{\lambda}/\lambda^*$. From the estimates (13.8) and (13.9), we can deduce the (14.6) and (14.7) ones in the same way as in the proof of Theorem 8.1. Notice that the constant C is depending on β. □

The remainder of the section is devoted to study finite element approximations of problem (3.18) including numerical quadrature rules. In the previous example we have applied the theory of Galerkin approximations, now we will apply the general results of Section 12.

Let Ω be a regular open bounded set in $\mathbb{R}^2$ and f be the Nemytskii operator of some function $\phi : \Omega \times \mathbb{R} \to \mathbb{R}$, that is for a function $u : \Omega \to \mathbb{R}$ we have

$$\text{for all} \quad x \in \Omega \qquad f(u)(x) = \phi(x, u(x)).$$

We assume that the function ϕ satisfies

$$\begin{cases} \text{the function } \phi \text{ is twice continuously differentiable with respect to the} \\ \text{second argument uniformly in } \Omega \text{ and for all } z \in \mathbb{R},\ \phi(.,z), \dfrac{\partial \phi}{\partial z}(.,z), \\ \text{and } \dfrac{\partial^2 \phi}{\partial z^2}(.,z) \text{ are in } L^\infty(\Omega) \end{cases} \tag{14.8}$$

and the assumptions (3.16), (3.17), (3.23). Remark that the hypothesis (14.8) is stronger than the (3.15) one, so that $f : H_0^1(\Omega) \cap C^0(\overline{\Omega}) \to L^2(\Omega)$ is a C^2 mapping.

In Section 3, we have studied the positive solutions of

$$u - \lambda T f(u) = 0, \tag{14.9}$$

with $-T$ the inverse of the Laplacian operator with homogeneous Dirichlet boundary conditions, see Theorem 3.4. There exist a $\lambda^* > 0$ and a C^2 mapping $\lambda \in [0, \lambda^*) \to u(\lambda) \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$, where $u(\lambda)$ is the minimal positive solution of (14.9) with the parameter value λ. The derivative $I - \lambda T Df(u(\lambda)) \in \mathscr{L}(H_0^1(\Omega) \cap C^0(\overline{\Omega}); H_0^1(\Omega) \cap C^0(\overline{\Omega}))$ is an isomorphism.

Our goal is to study approximations of the regular solution set $\{(\lambda, u(\lambda)); \lambda \in [0, \lambda^*)\}$ when we use the schemes (3.29) and (3.30) to numerically solve (14.9). Since we have introduced numerical quadrature rules to compute the integrals, we can no longer use the Galerkin approximation approach detailed in Section 13. We need to go back to the general results of Section 12. We are within the framework of Section 12 with $m = 1$, $X = Z = H_0^1(\Omega) \cap C^0(\overline{\Omega})$, and $F : \mathbb{R} \times X \to Z$ given by: for all $\lambda \in \mathbb{R}$, $v \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$,

$$F(\lambda, v) = v - \lambda T f(v). \tag{14.10}$$

Both approximation schemes (3.29) and (3.30) can be written in the form

$$F_h(\lambda, u_h) = 0 \tag{14.11}$$

where $F_h : \mathbb{R} \times (H_0^1(\Omega) \cap C^0(\overline{\Omega})) \to H_0^1(\Omega) \cap C^0(\overline{\Omega})$ is given by: for all $\lambda \in \mathbb{R}$, $v \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$

$$F_h(\lambda, v) = v - \lambda T_h f(v). \tag{14.12}$$

Here the operator $T_h \in \mathscr{L}(C^0(\overline{\Omega}); V_h)$ is defined by: for $g \in C^0(\overline{\Omega})$

$$\text{for all} \quad v_h \in V_h \qquad a_h(T_h g, v_h) = (g, v_h)_h, \tag{14.13}$$

where

$$a_h(u_h, v_h) = \int_{\Omega_h} \mathbf{grad}(u_h) \mathbf{grad} v_h \, dx, \tag{14.14}$$

$$(g, v_h)_h = \sum_{T \in \mathscr{T}_h} \left[\frac{S_T}{3} \sum_{i=1}^{3} (g v_h)(a_{i,T}) \right], \tag{14.15}$$

in the piecewise $\mathscr{P}_1$ approximation (3.29) and

$$a_h(u_h, v_h) = \sum_{T \in \mathscr{T}_h} \left[\sum_{i=1}^{3} \frac{1}{6} \det\big(DF_T(\hat{m}_i)\big) (\mathbf{grad} u_h \, \mathbf{grad} v_h)(m_{i,T}) \right], \tag{14.16}$$

$$(g, v_h)_h = \sum_{T \in \mathscr{T}_h} \sum_{i=1}^{3} \left[\frac{1}{6} \det\big(DF_T(\hat{m}_i)\big) (g v_h)(m_{i,T}) \right], \tag{14.17}$$

in the piecewise $\mathscr{P}_2$ approximation (3.30).

The main approximation properties of the operator T_h are summarized in the next theorem.

Theorem 14.2. *Let T_h be the operator defined in (14.13) with (14.14) and (14.15) in the case $k = 1$ and with (14.16) and (14.17) in the case $k = 2$. Then there exists a constant C independent of h such that*

i) for all $f \in W^{1,q}(\Omega)$ with $q > 2$, the following estimate holds

$$|Tf - T_h f|_{1,\Omega} + \|Tf - T_h f\|_{0,\infty,\Omega} \le Ch\|f\|_{1,q,\Omega}; \tag{14.18}$$

ii) for all $f \in W^{k+1,q}(\Omega)$ with $q > 1$, the following estimate holds

$$\|Tf - T_h f\|_{0,\Omega} \le Ch^{k+1}\|f\|_{k+1,q,\Omega}; \tag{14.19}$$

iii) for $k = 2$ and $f \in W^{2,q}(\Omega)$ with $q > 1$, the following estimate holds

$$\|Tf - T_h f\|_{1,\Omega\cap\Omega_h} \le Ch^2\|f\|_{2,q,\Omega}. \tag{14.20}$$

Moreover for all $\epsilon > 0$ there exists a constant $C(\epsilon)$ such that for all $f \in W^{k+1,1}(\Omega)$,

$$\|Tf - T_h f\|_{0,\infty,\Omega} \le C(\epsilon)h^{k+1-\epsilon}\|f\|_{k+1,1,\Omega}. \tag{14.21}$$

If $r_h : C^0(\overline{\Omega}) \to V_h$ denotes the corresponding Lagrange interpolation operator, then there exists a constant C such that for all $f \in C^0(\overline{\Omega})$ and $1 \le p < \infty$

$$\|T_h f\|_{1,p,\Omega} \le C\|r_h f\|_{0,\Omega}. \quad \square \tag{14.22}$$

These estimates are standard finite element estimates. For the proofs, see for instance in CIARLET [1991], in CROUZEIX and RAPPAZ [1989], and in the paper of WAHLBIN [1978].

To each solution $(\lambda, u(\lambda))$ we associate the function $\tilde{u}_h(\lambda) = \lambda T_h f(u(\lambda)) \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$. Since $(\lambda, u(\lambda))$ is a solution to (14.9), the following relation holds

$$\tilde{u}_h(\lambda) = u(\lambda) - \lambda Tf(u(\lambda)) + \lambda T_h f(u(\lambda)) = u(\lambda) - \lambda(T - T_h)f(u(\lambda)),$$

which implies that $\tilde{u}_h(\lambda)$ is close to $u(\lambda)$ in the norm of $H_0^1(\Omega) \cap C^0(\overline{\Omega})$. This choice is well adapted to get error estimates.

From the assumption (14.8), we deduce that F is of class C^2, so the hypothesis (12.4) is clearly satisfied.

The estimate of $\|F_h(\lambda, \tilde{u}_h(\lambda))\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})}$ goes as follows. We set

$$\begin{aligned}\epsilon_h &= \|F_h(\lambda, \tilde{u}_h(\lambda))\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} = \|\tilde{u}_h(\lambda) - \lambda T_h f(\tilde{u}_h(\lambda))\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \\ &= \lambda\|T_h\big(f(u(\lambda)) - f(\tilde{u}_h(\lambda))\big)\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})}.\end{aligned}$$

As a consequence of (14.22), we get

$$\epsilon_h \le C\|r_h\big(f(u(\lambda)) - f(\tilde{u}_h(\lambda))\big)\|_{0,\Omega}.$$

Since for all $a, v \in C^0(\overline{\Omega})$

$$\|r_h(av)\|_{0,\Omega} \leq C\|a\|_{0,\infty,\Omega}\|r_h v\|_{0,\Omega},$$

we have

$$\epsilon_h \leq C\|r_h(u(\lambda) - \tilde{u}_h(\lambda))\|_{0,\Omega} = C\lambda\|\big(r_h T - T_h\big)f(u(\lambda))\|_{0,\Omega}.$$

From the classical properties of the Lagrange interpolation operator r_h and from the estimate (14.19), we get the estimate

$$\epsilon_h \leq Ch^{k+1}, \tag{14.23}$$

with $k = 1$ or 2 corresponding to both approximations (3.29) and (3.30) respectively. So the assumption (12.5) is verified.

To check the stability (12.6), let us first notice that $D_v F(\lambda, u(\lambda)) \in \mathscr{L}(H_0^1(\Omega) \cap C^0(\overline{\Omega}); H_0^1(\Omega) \cap C^0(\overline{\Omega}))$ is an isomorphism. Using a compactness argument on T, it is a simple matter to check that $D_v F(\lambda, u(\lambda)) \in \mathscr{L}(W_0^{1,p}(\Omega); W_0^{1,p}(\Omega))$, with $p > 2$, is an isomorphism. So there exists a constant C independent of h such that

$$\|D_v F(\lambda, u(\lambda))^{-1}\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega}); H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq C, \tag{14.24}$$

$$\|D_v F(\lambda, u(\lambda))^{-1}\|_{W_0^{1,p}(\Omega); W_0^{1,p}(\Omega)} \leq C. \tag{14.25}$$

Let us notice that

$$D_v F_h(\lambda, \tilde{u}_h(\lambda)) = D_v F(\lambda, u(\lambda)) + D_v F_h(\lambda, \tilde{u}_h(\lambda)) - D_v F(\lambda, u(\lambda)).$$

So to prove the invertibility of $D_v F_h(\lambda, \tilde{u}_h(\lambda)) \in \mathscr{L}(H_0^1(\Omega) \cap C^0(\overline{\Omega}); H_0^1(\Omega) \cap C^0(\overline{\Omega}))$, we will bound the operator

$$A_h(\lambda) = D_v F(\lambda, u(\lambda))^{-1}\left[D_v F_h(\lambda, \tilde{u}_h(\lambda)) - D_v F(\lambda, u(\lambda))\right].$$

Let v be in $H_0^1(\Omega) \cap C^0(\overline{\Omega})$, then

$$\begin{aligned} A_h(\lambda)v =& D_v F(\lambda, u(\lambda))^{-1}\left[\lambda T Df(u(\lambda))v - \lambda T_h Df(\tilde{u}_h(\lambda))v\right] \\ =& D_v F(\lambda, u(\lambda))^{-1}\Big[\lambda(T - T_h)Df(u(\lambda))v + \\ & \lambda T_h\big(Df(u(\lambda)) - Df(\tilde{u}_h(\lambda))\big)v\Big]. \end{aligned} \tag{14.26}$$

Since the mapping $T_h \in \mathscr{L}(C^0(\overline{\Omega}); W^{1,p}(\Omega))$ is uniformly bounded, see (14.22), we get from (14.26) and (14.25)

$$\|A_h(\lambda)v\|_{1,p,\Omega} \leq C\|v\|_{0,\infty,\Omega} \leq C\|v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})}. \tag{14.27}$$

On the other hand

$$\|A_h(\lambda)v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq C\Big[\|(T-T_h)Df(u(\lambda))v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} + \|\big[Df(u(\lambda))-Df(\tilde{u}_h(\lambda))\big]v\|_{0,\infty,\Omega}\Big],$$

so with the regularity of f,

$$\|A_h(\lambda)v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq C\Big[\|(T-T_h)Df(u(\lambda))v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} + \|(T-T_h)f(u(\lambda))\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})}\|v\|_{1,p,\Omega}\Big].$$

Using the convergence property (14.18), we get

$$\|A_h(\lambda)v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq Ch\|v\|_{1,p,\Omega}. \tag{14.28}$$

From the estimates (14.27) and (14.28) we deduce

$$\|A_h^2(\lambda)v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq Ch\|v\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})},$$

which implies

$$\|A_h^2(\lambda)\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega});H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq Ch.$$

It is not difficult to check that the condition (1.1) can be relaxed to $\|(A^{-1}B)^2\|_{X;X} \leq \ell^2$, with $\ell \in (0,1)$, see for instance Crouzeix and Rappaz [1989], to get that $A+B$ is an isomorphism. If h is chosen sufficiently small, than $D_vF_h(\lambda,\tilde{u}_h(\lambda))$ is an isomorphism with a uniformly bounded inverse. So we have checked the stability assumption (12.6).

The mapping $D_vF_h(\lambda,.)$ is Lipschitzian since $T_h \in \mathscr{L}(C^0(\overline{\Omega});H_0^1(\Omega)\cap C^0(\overline{\Omega}))$ is uniformly bounded and f is C^2, that is (12.4) is satisfied.

Theorem 14.3. *Let $\{(\lambda,u(\lambda));\lambda\in[0,\overline{\lambda}]\}$ with $\overline{\lambda}<\lambda^*$ be a regular solution set to the problem $F(\lambda,v)=0$ with F given in (14.10) and let F_h be the mapping defined in (14.12). There exist two constants $h_0>0$, $\delta_0>0$ and for $h\leq h_0$ a C^2 mapping $u_h:\lambda\in[0,\overline{\lambda}]\to u_h(\lambda)\in V_h$ such that for all $\lambda\in[0,\overline{\lambda}]$*

$$\big(F_h(\lambda,v_h)=0 \text{ and } \|v_h-\tilde{u}_h(\lambda)\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})}\leq\delta_0\big) \Longleftrightarrow v_h=u_h(\lambda),$$

and the set $\{(\lambda,u_h(\lambda));\lambda\in[0,\overline{\lambda}]\}$ is a regular solution set of problem (14.11). Moreover the following a priori estimates hold

$$\|u(\lambda)-u_h(\lambda)\|_{0,\Omega}+h\|u(\lambda)-u_h(\lambda)\|_{1,\Omega}\leq Ch^{k+1} \tag{14.29}$$

and

$$\|u(\lambda)-u_h(\lambda)\|_{0,\infty,\Omega}\leq C(\epsilon)h^{k+1-\epsilon}. \tag{14.30}$$

Proof. The hypotheses of Theorem 12.3 are a consequence of the above developments together with the compactness of the interval $[0, \overline{\lambda}]$. There exist $h_0 > 0$, $\delta_0 > 0$, and for all $h \leq h_0$ a C^2 mapping $u_h : \lambda \in [0, \overline{\lambda}] \to u_h(\lambda) \in H_0^1(\Omega) \cap C^0(\overline{\Omega})$ such that

$$\big(F_h(\lambda, v_h) = 0 \text{ and } \|v_h - \tilde{u}_h(\lambda)\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq \delta_0\big) \Longleftrightarrow v_h = u_h(\lambda)$$

and the mapping $D_v F_h(\lambda, u_h(\lambda)) \in \mathscr{L}(H_0^1(\Omega) \cap C^0(\overline{\Omega}); H_0^1(\Omega) \cap C^0(\overline{\Omega}))$ is invertible, with a uniformly bounded inverse.

The estimates (12.12) and (14.23) give

$$\|\tilde{u}_h(\lambda) - u_h(\lambda)\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq C h^{k+1}. \tag{14.31}$$

Since

$$u(\lambda) - \tilde{u}_h(\lambda) = \lambda(T - T_h) f(u(\lambda)),$$

we deduce the estimates (14.29) and (14.30) from the (14.31) one and from Theorem 14.2. □

Remark 14.1. In the same way we proved the estimate (14.23), we can prove that

$$\|\frac{d}{d\lambda} F(\lambda, \tilde{u}_h(\lambda))\|_{H_0^1(\Omega)\cap C^0(\overline{\Omega})} \leq C h^{k+1}.$$

As a consequence of Theorem 12.4, there exist constants C and for $\epsilon > 0$, $C(\epsilon)$ such that for $h \leq h_0$

$$\|Du(\lambda) - Du_h(\lambda)\|_{0,\Omega} + h\|Du(\lambda) - Du_h(\lambda)\|_{1,\Omega} \leq C h^{k+1}$$

and

$$\|Du(\lambda) - Du_h(\lambda)\|_{0,\infty,\Omega} \leq C(\epsilon) h^{k+1-\epsilon}. \quad \square$$

Remark 14.2. In Theorem 14.3, we have got a priori estimates by using the relation (12.12). The inequality (12.13) will give a posteriori estimates. We do not want to emphasize this point here. □

15. Solution set containing a simple limit point

Let X and Z be two real Banach spaces and $F : \mathbb{R} \times X \to Z$ be a nonlinear mapping of class C^1. Here for simplicity we have considered the parameter space $\mathbb{R}^m$ with $m = 1$; our approach can be immediately generalized for any $m \in \mathbb{N}$.

We assume that $(\lambda_0, x_0) \in \mathbb{R} \times X$ is a regular solution to

$$F(\lambda, x) = 0, \tag{15.1}$$

but $D_x F(\lambda_0, x_0) \in \mathscr{L}(X; Z)$ is no more an isomorphism. More precisely the following assumption holds:

$$\begin{aligned} &(\lambda_0, x_0) \text{ is a solution to (15.1), } DF(\lambda_0, x_0) \in \mathscr{L}(\mathbb{R} \times X; Z) \text{ is} \\ &\text{a Fredholm operator of index 1, } Z = \text{Range}(DF(\lambda_0, x_0)), \text{ and} \\ &D_x F(\lambda_0, x_0) \in \mathscr{L}(X; Z) \text{ is not an isomorphism.} \end{aligned} \tag{15.2}$$

The assumption (15.2) means that the solution $(\lambda_0, x_0) \in \mathbb{R} \times X$ to (15.1) is such that $D_x F(\lambda_0, x_0) \in \mathscr{L}(X; Z)$ is a Fredholm operator of index 0, $\dim \mathrm{Ker}(D_x F(\lambda_0, x_0)) = 1$, and $D_\lambda F(\lambda_0, x_0)$ does not belong to the range of $D_x F(\lambda_0, x_0)$.

Let $\varphi \in X$, $\varphi \not\equiv 0$, satisfy $D_x F(\lambda_0, x_0)\varphi = 0$; so $\mathrm{Ker}(D_x F(\lambda_0, x_0)) = \mathrm{span}\{\varphi\}$.

Since $D_x F(\lambda_0, x_0)$ is no more an isomorphism, it is not possible to describe the solution set of (15.1) in a neighborhood of (λ_0, x_0) as a solution branch parametrized by λ. To study it we introduce the mapping $\Phi : \mathbb{R} \times \mathbb{R} \times X \to \mathbb{R} \times Z$ given by

$$\Phi(s, \lambda, x) = (B(x - x_0) - s, F(\lambda, x)), \tag{15.3}$$

where B is a functional in X' with $B(\varphi) = 1$. We check that

$$\Phi(0, \lambda_0, x_0) = 0$$

and for $\delta \in \mathbb{R}$, $v \in X$,

$$D^2_{\lambda x}\Phi(0, \lambda_0, x_0)(\delta, v) = (Bv, D_\lambda F(\lambda_0, x_0)\delta + D_x F(\lambda_0, x_0)v).$$

The derivative $D^2_{\lambda x}\Phi(0, \lambda_0, x_0) \in \mathscr{L}(\mathbb{R} \times X; \mathbb{R} \times Z)$ is a Fredholm operator of index 0. Since $\dim \mathrm{Ker}(D_x F(\lambda_0, x_0))$ is finite, there exists a closed subspace $X_1 \subset X$ such that $X = \mathrm{span}\{\varphi\} \oplus X_1$. If $(\delta, v \equiv \alpha\varphi + v_1) \in \mathbb{R} \times X$ with $\alpha \in \mathbb{R}$ and $v_1 \in X_1$ satisfies

$$D^2_{\lambda x}\Phi(0, \lambda_0, x_0)(\delta, v) = 0,$$

then $v_1 = 0$ and $\delta = 0$ since $DF(\lambda_0, x_0) \in \mathscr{L}(X_1; Z)$ is an isomorphism. Since $B(\varphi) = 1$ so $\alpha = 0$. Finally $D^2_{\lambda x}\Phi(0, \lambda_0, x_0)$ is injective and therefore $D^2_{\lambda x}\Phi(0, \lambda_0, x_0) \in \mathscr{L}(\mathbb{R} \times X; \mathbb{R} \times Z)$ is an isomorphism.

To study the solutions of

$$\Phi(s, \lambda, x) = 0 \tag{15.4}$$

in a neighborhood of $(0, \lambda_0, x_0)$, we can use the implicit function theorem to prove the following result.

Theorem 15.1. *Under the assumption (15.2), there exist a positive number ϵ, a neighborhood W of $(\lambda_0, x_0) \in \mathbb{R} \times X$, and a unique mapping $(\lambda, u) : s \in [-\epsilon, \epsilon] \to (\lambda(s), u(s)) \in \mathbb{R} \times X$ of class C^1 such that for all $s \in [-\epsilon, \epsilon]$*

$$F(\lambda(s), u(s)) = 0 \qquad \text{and} \qquad s = B(u(s) - x_0), \tag{15.5}$$

$$\lambda(0) = \lambda_0 \qquad \text{and} \qquad u(0) = x_0, \tag{15.6}$$

and if $(\lambda, u) \in W$ is a solution to (15.1), then necessarily $\lambda = \lambda(s)$, $u = u(s)$ with $s = B(u - x_0)$. Moreover for $s \in [-\epsilon, \epsilon]$, the mapping $D^2_{\lambda x}\Phi(s, \lambda(s), u(s)) \in \mathscr{L}(\mathbb{R} \times X; \mathbb{R} \times Z)$ is an isomorphism.

Clearly $(\lambda(s), u(s)) \neq (\lambda_0, x_0)$ for all $s \neq 0$ since $B(u(s) - x_0) = s$. So the mapping $s \in [-\epsilon, \epsilon] \to (\lambda(s), u(s)) \in \mathbb{R} \times X$ cannot be the constant function equal to (λ_0, x_0)

and the set of solutions of (15.1) is not isolated. We can get more information on the variation of λ and u as a function of s. When differentiating with respect to s the identity $\Phi(s, \lambda(s), u(s)) = 0$, we get

$$B(\frac{d}{ds}u(0)) = 1,$$
$$D_\lambda F(\lambda_0, x_0)\frac{d}{ds}\lambda(0) + D_x F(\lambda_0, x_0)\frac{d}{ds}u(0) = 0,$$

and so $\frac{d}{ds}\lambda(0) = 0$, $\frac{d}{ds}u(0) = \varphi$.

A pair $(\lambda_0, x_0) \in \mathbb{R} \times X$ satisfying to the condition (15.2) is called a simple limit point of (15.1).

If we assume that F is a mapping of class C^p with $p \geq 2$, then we can get a more precise description of the solution set $\{(\lambda(s), u(s)); s \in [-\epsilon, \epsilon]\}$. We differentiate twice with respect to s the identity $\Phi(s, \lambda(s), u(s)) = 0$ to get

$$D_\lambda F(\lambda_0, x_0)\frac{d^2}{ds^2}\lambda(0) + D^2_{xx}F(\lambda_0, x_0)(\varphi, \varphi) + D_x F(\lambda_0, x_0)\frac{d^2}{ds^2}u(0) = 0,$$

since $\frac{d}{ds}\lambda(0) = 0$ and $\frac{d}{ds}u(0) = \varphi$, and

$$B(\frac{d^2}{ds^2}u(0)) = 0.$$

Consequently $\frac{d^2}{ds^2}\lambda(0)$ is different from 0 if and only if

$$D^2_{xx}F(\lambda_0, x_0)(\varphi, \varphi) \notin \text{Range}(D_x F(\lambda_0, x_0)). \tag{15.7}$$

If (λ_0, x_0) is a simple limit point of (15.1) satisfying to (15.7), then for s in a neighborhood of 0 the following development holds

$$\lambda(s) = \lambda_0 + \frac{\frac{d^2}{ds^2}\lambda(0)}{2}s^2 + o(s^2).$$

In the figure 15.1, we have represented the branch $\{(\lambda(s), B(u(s))\}$ in a neighborhood of 0.

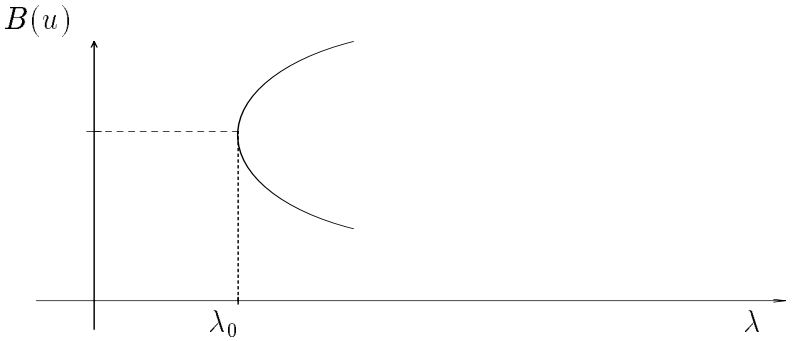

FIGURE 15.1: $D^2\lambda(0) > 0$.

A pair $(\lambda_0, x_0) \in \mathbb{R} \times X$ satisfying to the conditions (15.2) and (15.7) is a nondegenerate simple limit point.

In the degenerate case $\frac{d^2}{ds^2}\lambda(0) = 0$, different cases can occur, see for instance the figure 15.2.

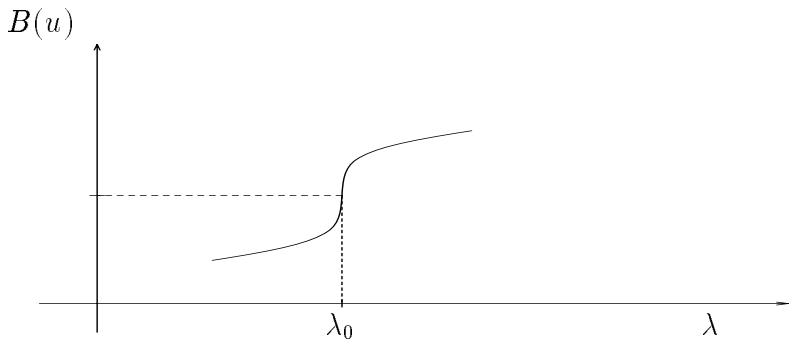

FIGURE 15.2: $D^2\lambda(0) = 0$ AND $D^3\lambda(0) \neq 0$.

Remark 15.1. A nondegenerate simple limit point is in fact a junction point of two regular solution branches parametrized by λ. Let (λ_0, x_0) be a nondegenerate simple limit point and let $s \in [-\epsilon, \epsilon] \to (\lambda(s), u(s)) \in \mathbb{R} \times X$ be the solution set given in Theorem 15.1. We assume here that F is of class C^p with $p \geq 2$. Then we will prove that there exist positive numbers $0 < \epsilon_1 \leq \epsilon$ and C such that for all $s \in [-\epsilon_1, \epsilon_1]$, $s \neq 0$,

$$D_x F(\lambda(s), u(s)) \in \mathscr{L}(X; Z) \quad \text{is an isomorphism} \tag{15.8}$$

and

$$\|D_x F(\lambda(s), u(s))^{-1}\|_{Z;X} \leq \frac{C}{|s|}. \tag{15.9}$$

Since $\frac{d^2}{ds^2}\lambda(0) \neq 0$, there exist ϵ_1, $0 < \epsilon_1 \leq \epsilon$ and a constant C_1 such that

$$\text{for all } s \in [-\epsilon_1, \epsilon_1] \quad |\frac{d}{ds}\lambda(s)| \geq C_1|s|.$$

The derivative $D^2_{\lambda x}\Phi(s, \lambda(s), u(s)) \in \mathscr{L}(\mathbb{R} \times X; \mathbb{R} \times Z)$ is an isomorphism with an inverse uniformly bounded with respect to $s \in [-\epsilon_1, \epsilon_1]$. For all $s \in [-\epsilon_1, \epsilon_1]$ and $f \in Z$, there exists a unique $(\delta, v) \in \mathbb{R} \times X$ with

$$D^2_{\lambda x}\Phi(s, \lambda(s), u(s))(\delta, v) = (0, f)$$

and

$$|\delta| + \|v\|_X \leq C_2\|f\|_Z,$$

where C_2 is independent of s. Given $f \in Z$ and $s \neq 0$, we use both relations

$$D_\lambda F(\lambda(s), u(s))\delta + D_x F(\lambda(s), u(s))v = f,$$
$$D_\lambda F(\lambda(s), u(s))\frac{d}{ds}\lambda(s) + D_x F(\lambda(s), u(s))\frac{d}{ds}u(s) = 0,$$

and set $w = v - \frac{\delta}{\frac{d}{ds}\lambda(s)}\frac{d}{ds}u(s)$, to obtain

$$D_x F(\lambda(s), u(s))w = f$$

and

$$\|w\|_X \leq C_2\left(1 + \frac{1}{C_1|s|}\|\frac{d}{ds}u(s)\|_X\right)\|f\|_Z.$$

We have to check now that $D_x F(\lambda(s), u(s))$ is injective with $s \in [-\epsilon_1, \epsilon_1]$, $s \neq 0$. Let $\psi \in X$ be such that

$$D_x F(\lambda(s), u(s))\psi = 0.$$

Then since $B(\frac{d}{ds}u(s)) = 1$, we have

$$D^2_{\lambda x}\Phi(s, \lambda(s), u(s))(-B(\psi)\frac{d}{ds}\lambda(s), \psi - B(\psi)\frac{d}{ds}u(s)) = 0,$$

and consequently $\psi = B(\psi)\frac{d}{ds}u(s)$ and $B(\psi)\frac{d}{ds}\lambda(s) = 0$. Therefore $\psi = 0$. $\square$

16. Galerkin approximation of simple limit points

Let X,Y be two real Banach spaces, and let Y' denote the dual space of Y. Given a C^p mapping $F : \mathbb{R} \times X \to Y'$, with $p \geq 2$, we shall analyze Galerkin approximations of the problem

$$F(\lambda, x) = 0 \tag{16.1}$$

in a neighborhood of a regular solution $(\lambda_0, x_0) \in \mathbb{R} \times X$ satisfying to (15.2). The study of the exact problem has been done in Section 15.

To analyze the approximation of the simple limit point (λ_0, x_0), we consider the case where the mapping F has the following structure

$$F(\lambda, x) = Lx + G(\lambda, x) \tag{16.2}$$

where

$$L \in \mathscr{L}(X; Y') \quad \text{is an isomorphism} \tag{16.3}$$

and the C^p mapping $G : \mathbb{R} \times X \to Y'$ satisfies:

$$\text{for all}\,(\lambda, x) \in \mathbb{R} \times X \quad D_x G(\lambda, x) \in \mathscr{L}(X; Y') \ \text{ is compact.} \tag{16.4}$$

Then problem (16.1) reads: find $\lambda \in \mathbb{R}$, $x \in X$ such that

$$\text{for all } y \in Y \quad \langle Lx, y\rangle_{Y'Y} + \langle G(\lambda, x), y\rangle_{Y'Y} = 0. \tag{16.5}$$

Given a family $\{X_h\}_{0<h\leq 1}$ of finite-dimensional subspaces of X and a family $\{Y_h\}_{0<h\leq 1}$ of finite-dimensional subspaces of Y, a Galerkin approximation of (16.5) consists in finding $\lambda \in \mathbb{R}$ and $x_h \in X_h$ such that

$$\text{for all } y_h \in Y_h \quad \langle Lx_h, y_h\rangle_{Y'Y} + \langle G(\lambda, x_h), y_h\rangle_{Y'Y} = 0. \tag{16.6}$$

Let $b : X \times Y \to \mathbb{R}$ be the bilinear form defined by

$$\text{for all } (x, y) \in X \times Y \quad b(x, y) = \langle Lx, y\rangle_{Y'Y}. \tag{16.7}$$

Then the stability assumption reads: for all $h \in (0, 1]$,

$$\inf_{\substack{x\in X_h \\ \|x\|_X=1}} \sup_{\substack{y\in Y_h \\ \|y\|_Y=1}} b(x, y) \geq \beta > 0 \tag{16.8}$$

and

$$\dim X_h = \dim Y_h. \tag{16.9}$$

The consistency assumption reads

$$\text{for all } x \in X \quad \lim_{h\to 0} \inf_{x_h\in X_h} \|x - x_h\|_X = 0. \tag{16.10}$$

To study the approximation (16.6) we proceed in a way similar to the one in Theorem 7.1. Let $\Pi_{X_h} \in \mathscr{L}(X; X_h)$ and $\Pi_{Y_h} \in \mathscr{L}(Y; Y_h)$ be the two projectors defined by

$$\text{for all } y_h \in Y_h \quad b(x - \Pi_{X_h} x, y_h) = 0 \tag{16.11}$$

and

$$\text{for all } x_h \in X_h \quad b(x_h, y - \Pi_{Y_h} y) = 0. \tag{16.12}$$

Then we construct the family $\{F_h\}_{0<h\leq 1}$ of mappings $F_h : \mathbb{R} \times X \to Y'$ by setting: for $\lambda \in \mathbb{R}$ and for $x \in X$, $y \in Y$,

$$\langle F_h(\lambda, x), y\rangle_{Y'Y} = \langle F(\lambda, x), \Pi_{Y_h} y\rangle_{Y'Y} + b(x, y - \Pi_{Y_h} y).$$

It is a simple matter (see the proof of Theorem 7.1) to check that problem (16.6) is equivalent to find $\lambda \in \mathbb{R}$ and $x_h \in X$ such that

$$F_h(\lambda, x_h) = 0. \tag{16.13}$$

To study problem (16.13) in some neighborhood of the simple limit point $(\lambda_0, x_0) \in \mathbb{R} \times X$, we define the mapping $\Phi_h : \mathbb{R} \times \mathbb{R} \times X \to \mathbb{R} \times Y'$ by

$$\Phi_h(s, \lambda, x) = (B(x - u(s)), F_h(\lambda, x)),$$

the discrete analogue of Φ given in (15.3). Here $(\lambda, u) : [-\epsilon, \epsilon] \to \mathbb{R} \times X$ is the exact solution set in a neighborhood of (λ_0, x_0). Clearly if $(s, \lambda_h, x_h) \in \mathbb{R} \times \mathbb{R} \times X$ satisfies

$$\Phi_h(s, \lambda_h, x_h) = 0 \tag{16.14}$$

then (λ_h, x_h) is a solution to (16.13) and consequently to (16.6). Conversely if $(\lambda_h, x_h) \in \mathbb{R} \times X$ be a solution to (16.6) or (16.13), then $(s \equiv B(x_h) - B(x_0), \lambda_h, x_h)$ is a solution to (16.14).

The study of problem (16.14) can be done as an application of the results of Section 12 with the mapping Φ_h and the solution set $\{(s, \lambda(s), u(s)); s \in [-\epsilon, \epsilon]\}$. Let us check that Φ_h satisfies successively the consistency (12.10), the stability (12.11), and the Lipschitzian condition (12.9) at the solutions $(\lambda(s), u(s))$.

For a given $s \in [-\epsilon, \epsilon]$, the following estimate holds

$$\begin{aligned} &\|\Phi_h(s, \lambda(s), u(s))\|_{\mathbb{R} \times Y'} \\ &\quad = \sup_{\substack{y \in Y \\ \|y\|_Y = 1}} \langle F_h(\lambda(s), u(s)), y \rangle_{Y'Y} = \sup_{\substack{y \in Y \\ \|y\|_Y = 1}} b(u(s), y - \Pi_{Y_h} y) \\ &\quad = \sup_{\substack{y \in Y \\ \|y\|_Y = 1}} b(u(s) - \Pi_{X_h} u(s), y) \le \|L\|_{X;Y'} \|u(s) - \Pi_{X_h} u(s)\|_X . \end{aligned} \tag{16.15}$$

For $x \in X$ and $x_h \in X_h$, the inf-sup condition (16.8) implies

$$\begin{aligned} \|x_h - \Pi_{X_h} x\|_X &\le \frac{1}{\beta} \sup_{\substack{y_h \in Y_h \\ \|y_h\|_Y = 1}} b(x_h - \Pi_{X_h} x, y_h) \\ &= \frac{1}{\beta} \sup_{\substack{y_h \in Y_h \\ \|y_h\|_Y = 1}} b(x_h - x, y_h) \le \|L\|_{X;Y'} \frac{1}{\beta} \|x_h - x\|_X . \end{aligned}$$

This last inequality and the triangular inequality give

$$\|x - \Pi_{X_h} x\|_X \le (1 + \frac{\|L\|_{X;Y'}}{\beta}) \inf_{x_h \in X_h} \|x - x_h\|_X . \tag{16.16}$$

Then we use the estimate (16.15) with the (16.16) one and the consistency assumption (16.10) to check the consistency (12.10).

Noticing that for $\delta \in \mathbb{R}$ and $v \in X$

$$\begin{aligned} D^2_{\lambda x}\Phi_h(s,\lambda(s),u(s))(\delta,v) \\ &= \big(B(v), D_\lambda F_h(\lambda(s),u(s))\delta + D_x F_h(\lambda(s),u(s))v\big) \\ &= D^2_{\lambda x}\Phi(s,\lambda(s),u(s))(\delta,v) + \Big(0, \big(D_\lambda F_h(\lambda(s),u(s)) - D_\lambda F(\lambda(s),u(s))\big)\delta \\ &\qquad + \big(D_x F_h(\lambda(s),u(s)) - D_x F(\lambda(s),u(s))\big)v\Big) \end{aligned}$$

and since $D^2_{\lambda x}\Phi(s,\lambda(s),u(s)) \in \mathscr{L}(\mathbb{R}\times X, Y')$ is an isomorphism, the stability assumption (12.11) is checked, see (1.1), if we prove

$$\lim_{h\to 0} \max_{s\in[-\epsilon,\epsilon]} \|D_\lambda F_h(\lambda(s),u(s)) - D_\lambda F(\lambda(s),u(s))\|_{Y'} = 0 \tag{16.17}$$

and

$$\lim_{h\to 0} \max_{s\in[-\epsilon,\epsilon]} \|D_x F_h(\lambda(s),u(s)) - D_x F(\lambda(s),u(s))\|_{X;Y'} = 0. \tag{16.18}$$

We have successively

$$\begin{aligned} \|D_\lambda F_h(\lambda(s),u(s)) &- D_\lambda F(\lambda(s),u(s))\|_{Y'} \\ &= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} \langle D_\lambda F_h(\lambda(s),u(s)) - D_\lambda F(\lambda(s),u(s)), y\rangle_{Y'Y} \\ &= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} [\langle D_\lambda F(\lambda(s),u(s)), \Pi_{Y_h} y - y\rangle_{Y'Y}] \\ &= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} b(L^{-1}D_\lambda F(\lambda(s),u(s)), \Pi_{Y_h} y - y) \\ &= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} b((I-\Pi_{X_h})L^{-1}D_\lambda F(\lambda(s),u(s)), y) \\ &\le \|L\|_{X;Y'} \|(I-\Pi_{X_h})L^{-1}D_\lambda F(\lambda(s),u(s))\|_X . \end{aligned}$$

Then with the estimates (16.16) and (16.10), we conclude that

$$\lim_{h\to 0} \max_{s\in[-\epsilon,\epsilon]} \|D_\lambda F_h(\lambda(s),u(s)) - D_\lambda F(\lambda(s),u(s))\|_{Y'} = 0.$$

Let us prove now the relation (16.18). For $w \in X$ and $y \in Y$ given we have

$$\begin{aligned} \langle \big(D_x F_h(\lambda(s),u(s)) &- D_x F(\lambda(s),u(s))\big)w, y\rangle_{Y'Y} \\ &= \langle D_x F(\lambda(s),u(s))w, \Pi_{Y_h} y - y\rangle_{Y'Y} + b(w, y - \Pi_{Y_h} y) \\ &= \langle \big(L + D_x G(\lambda(s),u(s))\big)w, \Pi_{Y_h} y - y\rangle_{Y'Y} + \langle Lw, y - \Pi_{Y_h} y\rangle_{Y'Y} \\ &= b(L^{-1}D_x G(\lambda(s),u(s))w, \Pi_{Y_h} y - y) \\ &= -b((I-\Pi_{X_h})L^{-1}D_x G(\lambda(s),u(s))w, y). \end{aligned}$$

Let us prove now that

$$\lim_{h\to 0} \|(I-\Pi_{X_h})L^{-1}D_xG(\lambda(s),u(s))\|_{X;X} = 0.$$

From the assumptions (16.3) and (16.4), we get that the operator $L^{-1}D_xG(\lambda(s),u(s)) \in \mathscr{L}(X;X)$ is compact. Then from the consistency (16.10), the estimate (16.16), and the result (1.3), we get that for $s \in [-\epsilon,\epsilon]$ the above limit is zero. With the compactness of the interval $[-\epsilon,\epsilon]$ and the continuity of $s \in [-\epsilon,\epsilon] \to (\lambda(s),u(s)) \in \mathbb{R}\times X$ we conclude that the relation (16.18) is true.

Finally we verify the Lipschitzian condition (12.9) for $D^2_{\lambda x}\Phi_h$. For $(\lambda,x) \in \mathbb{R}\times X$, $s \in [-\epsilon,\epsilon]$, and $(\delta,v) \in \mathbb{R}\times X$

$$\begin{aligned}(16.19)\quad &\left[D^2_{\lambda x}\Phi_h(s,\lambda(s),u(s)) - D^2_{\lambda x}\Phi_h(s,\lambda,x)\right](\delta,v)\\ &\qquad = \big(0,\left[DF_h(\lambda(s),u(s)) - DF_h(\lambda,x)\right](\delta,v)\big).\end{aligned}$$

Moreover

$$\begin{aligned}&\|\big(DF_h(\lambda(s),u(s)) - DF_h(\lambda,x)\big)(\delta,v)\|_{Y'}\\ &\quad = \sup_{\substack{y\in Y\\ \|y\|_Y=1}} \langle \left(DF_h(\lambda(s),u(s)) - DF_h(\lambda,x)\right)(\delta,v), y\rangle_{Y'Y}\\ &\quad = \sup_{\substack{y\in Y\\ \|y\|_Y=1}} \langle \left(DF(\lambda(s),u(s)) - DF(\lambda,x)\right)(\delta,v), \Pi_{Y_h}y\rangle_{Y'Y}\\ &\quad \le \|\Pi_{Y_h}\|_{Y;Y}\|DF(\lambda(s),u(s)) - DF(\lambda,x)\|_{\mathbb{R}\times X;Y'}(|\delta| + \|v\|_X). \qquad (16.20)\end{aligned}$$

With the relation (16.19), the estimate (16.20) and the fact that DF is Lipschitzian at $(\lambda(s),u(s))$, (F is assumed to be C^2), we conclude that (12.9) is verified for $D^2_{\lambda x}\Phi_h$ at the points $(s,\lambda(s),u(s))$, $s \in [-\epsilon,\epsilon]$.

We can apply now Theorem 12.3.

Theorem 16.1. *Let $F : \mathbb{R}\times X \to Y'$ satisfy the assumptions (16.2), (16.3), and (16.4) with $p \ge 2$. Let $(\lambda_0,x_0) \in \mathbb{R}\times X$ be a simple limit point of problem (16.1). Under the hypotheses (16.8), (16.9), and (16.10), there exist $h_0 > 0$, $\delta_0 > 0$, and for all $0 < h \le h_0$ a C^p mapping $(\lambda_h,u_h) : s \in [-\epsilon,\epsilon] \to (\lambda_h(s),u_h(s)) \in \mathbb{R}\times X$ such that*

$$\big(\Phi_h(s,\lambda,x_h) = 0 \quad \text{and} \quad (\lambda,x_h) \in \overline{B}((\lambda(s),u(s)),\delta_0)\big) \Longleftrightarrow \big(\lambda = \lambda_h(s), x_h = u_h(s)\big)$$

Moreover there exists a constant $C > 0$ such that for all $s \in [-\epsilon,\epsilon]$

$$(16.21)\qquad |\lambda(s)-\lambda_h(s)| + \|u(s)-u_h(s)\|_X \le C \inf_{x_h\in X_h} \|u(s)-x_h\|_X.$$

Proof. We have checked before that we can apply Theorem 12.3. The estimate (16.21) is a consequence of the (12.12) one and of the relations (16.15) and (16.16). □

We assume now that the mapping F is of class C^p, $p \geq 3$, and that (λ_0, x_0) is a nondegenerate simple limit point, which means, see (15.7),

$$D^2_{xx} F(\lambda_0, x_0)(\varphi, \varphi) \notin \text{Range}\,(D_x F(\lambda_0, x_0)).$$

We have for $j = 0, 1, 2$

$$\begin{aligned}
\|\frac{d^j}{ds^j}\Phi_h(s, \lambda(s), u(s))\|_{\mathbb{R}\times Y'} &= \|\frac{d^j}{ds^j}F_h(\lambda(s), u(s))\|_{Y'} \\
&= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} \langle \frac{d^j}{ds^j}F_h(\lambda(s), u(s)), y\rangle_{Y'Y} \\
&= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} b(\frac{d^j}{ds^j}u(s), y - \Pi_{Y_h} y) \\
&= \sup_{\substack{y\in Y \\ \|y\|_Y=1}} b(\frac{d^j}{ds^j}u(s) - \Pi_{X_h}\frac{d^j}{ds^j}u(s), y).
\end{aligned}$$

So with this last relation and the (16.16) one we get for $j = 0, 1, 2$

$$\lim_{h\to 0} \max_{s\in[-\epsilon,\epsilon]} \|\frac{d^j}{ds^j}\Phi_h(s, \lambda(s), u(s))\|_{\mathbb{R}\times Y'} = 0. \tag{16.22}$$

Then with the relation (16.22), Theorem 12.4 implies for $j = 0, 1, 2$

$$\lim_{h\to 0} \max_{s\in[-\epsilon,\epsilon]} \left\{ |\frac{d^j}{ds^j}(\lambda(s) - \lambda_h(s))| + \|\frac{d^j}{ds^j}(u(s) - u_h(s))\|_X \right\} = 0. \tag{16.23}$$

We have assumed that (λ_0, x_0) is a nondegenerate simple limit point, so

$$\lambda(0) = \lambda_0, \qquad \frac{d}{ds}\lambda(0) = 0, \qquad \frac{d^2}{ds^2}\lambda(0) \neq 0.$$

There exist ϵ_1 with $0 < \epsilon_1 \leq \epsilon$, h_1 with $0 < h_1 \leq h_0$, $C > 0$, and for $0 < h \leq h_1$ a unique $s_h \in [-\epsilon_1, \epsilon_1]$ such that

$$\frac{d}{ds}\lambda_h(s_h) = 0, \tag{16.24}$$

$$|\frac{d^2}{ds^2}\lambda_h(s)| \geq C \quad \text{for all } s \in [-\epsilon_1, \epsilon_1]. \tag{16.25}$$

So the point $(\lambda_h(s_h), u_h(s_h))$ is a nondegenerate simple limit point of the problem (16.13). We know then that the solutions $(\lambda_h(s), u_h(s))$ form two regular solution branches parametrized by λ which meet at the point $(\lambda_h(s_h), u_h(s_h))$, see Remark 15.1.

Notice that the estimate (16.21) does not give in general an error between (λ_0, x_0) and $(\lambda_h(s_h), u_h(s_h))$ where $(\lambda_h(s_h), u_h(s_h))$ is the nondegenerate simple limit point of the approximate problem. Let us try to compare $\lambda_0 \equiv \lambda(0)$ and $\lambda_{0h} \equiv \lambda_h(s_h)$. We write

$$\lambda_0 - \lambda_{0h} = \lambda_0 - \lambda_h(0) - \int_0^{s_h} \frac{d}{ds}\lambda_h(s)\, ds.$$

The estimate (16.25) implies that $\frac{d^2}{ds^2}\lambda_h(s)$ does not change sign in $[0, s_h]$ and with (16.24)

$$|\lambda_0 - \lambda_{0h}| \le |\lambda_0 - \lambda_h(0)| + |s_h||\frac{d}{ds}\lambda_h(0)|.$$

With both relations (16.24) and (16.25), we deduce that

$$|s_h| \le \frac{1}{C}|\frac{d}{ds}\lambda_h(0)| \tag{16.26}$$

and then from the above inequality we get

$$|\lambda_0 - \lambda_{0h}| \le |\lambda_0 - \lambda_h(0)| + \frac{1}{C}|\frac{d}{ds}\lambda(0) - \frac{d}{ds}\lambda_h(0)|^2. \tag{16.27}$$

From the estimate (16.26) and the boundedness of $\frac{d}{ds}u_h(s)$, $s \in [-\epsilon, \epsilon]$, we get

$$\|u_h(0) - u_h(s_h)\|_X \le \tilde{C}|\frac{d}{ds}\lambda_h(0)|, \tag{16.28}$$

and then

$$\|u(0) - u_h(s_h)\|_X \le \|u(0) - u_h(0)\|_X + \tilde{C}|\frac{d}{ds}\lambda_h(0) - \frac{d}{ds}\lambda(0)|. \tag{16.29}$$

The estimates (16.27) and (16.29) measure the error between the nondegenerate simple limit point (λ_0, x_0) and its approximation $(\lambda_h(s_h), u_h(s_h))$.

It is often possible to improve the estimate for $|\lambda(0) - \lambda_h(0)|$ along the following lines. The Taylor expansion gives

$$\begin{aligned} F(\lambda_h(0), u_h(0)) = F(\lambda_0, x_0) &+ D_x F(\lambda_0, x_0)(u_h(0) - x_0) \\ &+ D_\lambda F(\lambda_0, x_0)(\lambda_h(0) - \lambda_0) \\ &+ \bigcirc\big(|\lambda_h(0) - \lambda_0|^2 + \|u_h(0) - x_0\|_X^2\big). \end{aligned} \tag{16.30}$$

On the other hand since $F_h(\lambda_h(0), u_h(0)) = 0$, $u_h(0)$ is a solution of (16.6) and

$$\text{for all } y \in Y \quad \langle F(\lambda_h(0), u_h(0)), \Pi_{Y_h} y\rangle_{Y'Y} = 0. \tag{16.31}$$

We choose now $y_0 \in Y$, $\|y_0\| = 1$, such that

$$\text{for all } \psi \in X \quad \langle D_x F(\lambda_0, x_0)\psi, y_0\rangle_{Y'Y} = 0. \tag{16.32}$$

From the relations (16.30), (16.31), and (16.32) we deduce

$$\langle D_\lambda F(\lambda_0, x_0), \Pi_{Y_h} y_0\rangle_{Y'Y}(\lambda_h(0) - \lambda_0) \tag{16.33}$$
$$+ \langle D_x F(\lambda_0, x_0)(u_h(0) - x_0), \Pi_{Y_h} y_0 - y_0\rangle_{Y'Y} = \bigcirc\big(|\lambda_h(0) - \lambda_0|^2 + \|u_h(0) - x_0\|_X^2\big).$$

The point (λ_0, x_0) is a simple limit point, so $D_\lambda F(\lambda_0, x_0) \notin \text{Range}(D_x F(\lambda_0, x_0))$ and then

$$\langle D_\lambda F(\lambda_0, x_0), y_0\rangle_{Y'Y} \neq 0. \tag{16.34}$$

The consistency assumption (16.10) and the estimate (16.16) imply

$$\lim_{h\to 0} \langle D_\lambda F(\lambda_0, x_0), y_0 - \Pi_{Y_h} y_0\rangle_{Y'Y} = 0. \tag{16.35}$$

From the relations (16.33), (16.34), and (16.35) we deduce the existence of a constant $C > 0$ independent of h such that

$$|\lambda_0 - \lambda_h(0)| \leq C\big[\|u_h(0) - x_0\|_X + \|y_0 - \Pi_{Y_h} y_0\|_Y\big]\|u_h(0) - x_0\|_X. \tag{16.36}$$

With (16.36) the estimate (16.27) can be improved in the following way

$$|\lambda_0 - \lambda_{0h}| \leq C\big(\|u_h(0) - x_0\|_X^2 + \tag{16.37}$$
$$|\frac{d}{ds}\lambda(0) - \frac{d}{ds}\lambda_h(0)|^2 + \|u_h(0) - x_0\|_X \|y_0 - \Pi_{Y_h} y_0\|_Y\big).$$

Remark 16.1. So far we do not have presented a posteriori error estimates in the case of simple limit points. In fact under the assumptions of Theorem 16.1, we can prove there exists a constant $C > 0$ such that for all $s \in [-\epsilon, \epsilon]$

$$|\lambda_h(s) - \lambda(s)| + \|u_h(s) - u(s)\|_X \leq C\|F(\lambda_h(s), u_h(s))\|_{Y'}. \tag{16.38}$$

Indeed the estimate (12.13) of Theorem 12.3 with the mapping Φ gives

$$|\lambda_h(s) - \lambda(s)| + \|u_h(s) - u(s)\|_X \leq C\|\Phi(s, \lambda_h(s), u_h(s))\|_{\mathbb{R}\times Y'}.$$

Going back to the definition of Φ_h and Φ we deduce that

$$B(u_h(s) - x_0) - s = B(u_h(s) - u(s)) + B(u(s) - x_0) = 0$$

and then

$$\Phi(s, \lambda_h(s), u_h(s)) = (0, F(\lambda_h(s), u_h(s))). \quad \square$$

Remark 16.2. The goal of this section was to present a general but simple method to study approximations of simple limit points. We have not emphasized on what are the weakest assumptions on F.

The assumption (16.2) on the shape of F can be avoided, but then the study is different, see Caloz [1994].

The function F does not need to be a C^2 function but DF needs to be Lipschitzian at $(\lambda(s), u(s))$ for $s \in [-\epsilon, \epsilon]$.

When we have applied the results of Section 12, we have decided to associate to each solution $(s, \lambda(s), u(s))$, $s \in [-\epsilon, \epsilon]$, the solution $(\lambda(s), u(s))$ itself instead of some approximations of it, say $(\tilde{\lambda}_h(s), \tilde{u}_h(s))$. Adding the appropriate assumptions on these approximations, see Sections 12 and 14, we can reproduce the approach of the present section to get estimates better suited to study the error in different norms. An other method would be to generalize Theorem 6.4. $\square$

17. Approximation of simple limit points. An example

The goal of the section is to apply to a simple and concrete example the general results on the approximation of simple limit points in Section 16.

For simplicity we consider again problem (14.3). Let $\Omega \subset \mathbb{R}^2$ be a regular convex open bounded set and let $f : H_0^1(\Omega) \to L^2(\Omega)$ be given by: for all $v \in H_0^1(\Omega)$

$$\text{for all } x \in \Omega \quad f(v)(x) = v^2(x) + 1. \tag{17.1}$$

We define the mapping $F : \mathbb{R} \times H_0^1(\Omega) \to H^{-1}(\Omega)$ by: for all $(\lambda, v) \in \mathbb{R} \times H_0^1(\Omega)$, $w \in H_0^1(\Omega)$

$$\langle F(\lambda, v), w \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx - \lambda \int_\Omega f(v) w\, dx. \tag{17.2}$$

There exists a solution $(\lambda^*, u^*) \in \mathbb{R} \times H_0^1(\Omega)$, u^* positive in Ω, to the problem

$$F(\lambda, v) = 0, \tag{17.3}$$

such that $D_v F(\lambda^*, u^*) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is not an isomorphism, see Remark 3.2. We can study more precisely the mapping $D_v F(\lambda^*, u^*)$. If $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$ is the inverse of the $(-\Delta)$ with homogeneous Dirichlet boundary conditions, then for $w \in H_0^1(\Omega)$

$$T D_v F(\lambda^*, u^*) w = w - \lambda^* T Df(u^*) w$$

and the mapping $T D_v F(\lambda^*, u^*) \in \mathscr{L}(H_0^1(\Omega); H_0^1(\Omega))$ has a kernel of dimension 1, namely $\operatorname{span}\{\varphi\}$, where the function $\varphi \in H_0^1(\Omega)$, $\varphi > 0$ in Ω, is an eigenvector corresponding to the least eigenvalue λ^* of the problem: find $\delta \in \mathbb{R}$ and $w \in H_0^1(\Omega)$, $w \neq 0$, such that

$$-\Delta w = \delta Df(u^*) w \quad \text{in } \Omega,$$

(see Remark 3.4, the proof of Theorem 3.4, and Theorem 3.2). Therefore the mapping $T D_v F(\lambda^*, u^*)$ is a Fredholm operator of index 0. The range of $T D_v F(\lambda^*, u^*)$ is orthogonal to $\varphi \in H_0^1(\Omega)$, with respect to the scalar product $(u, v)_{1,\Omega} = \int_\Omega \mathbf{grad} u\, \mathbf{grad} v\, dx$. Since both u^* and φ are positive in Ω, we easily check that $TDF(\lambda^*, u^*) \in \mathscr{L}(\mathbb{R} \times H_0^1(\Omega); H_0^1(\Omega))$ is a Fredholm operator of index 1 and $H_0^1(\Omega) = \operatorname{Range}(TDF(\lambda^*, u^*))$. The operator $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$ is an isomorphism, so $DF(\lambda^*, u^*) \in \mathscr{L}(\mathbb{R} \times H_0^1(\Omega); H^{-1}(\Omega))$ is a Fredholm operator of index 1.

The pair $(\lambda^*, u^*) \in \mathbb{R} \times H_0^1(\Omega)$ satisfies to the condition (15.2). It is a simple limit point of the problem (17.3). Notice that this result has been already checked in Remark 3.4. It is a simple matter to prove that $D_{vv}^2 F(\lambda^*, u^*)(\varphi, \varphi)$ does not belong to the range of $D_v F(\lambda^*, u^*)$, so the relation (15.7) is satisfied and (λ^*, u^*) is a nondegenerate simple limit point.

Let $B \in H^{-1}(\Omega)$ with $B(\varphi) = 1$ denote a functional appearing in the definition of the mapping Φ in (15.3). For instance we could consider

$$B(v) = \frac{1}{\|\varphi\|_{0,\Omega}^2} \int_\Omega \varphi v\, dx.$$

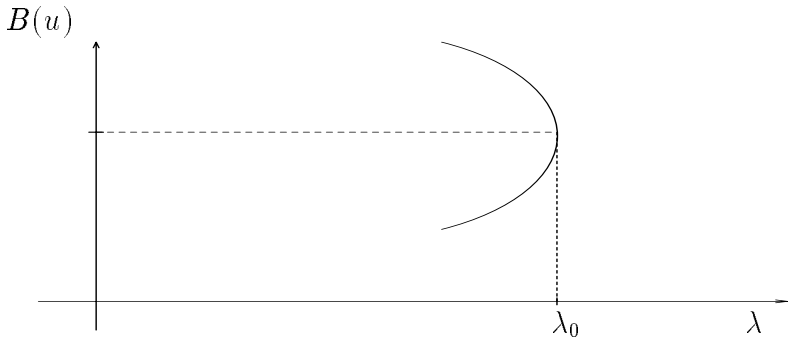

FIGURE 17.1: (λ^*, u^*) NONDEGENERATE SIMPLE LIMIT POINT.

Then graphically we have the situation of Figure 17.1.
Here $(\lambda, u) : s \in [-\epsilon, \epsilon] \to (\lambda(s), u(s)) \in \mathbb{R} \times H_0^1(\Omega)$ denotes the C^∞ mapping such that for all $s \in [-\epsilon, \epsilon]$

$$\begin{aligned} F(\lambda(s), u(s)) = 0 \quad &\text{and} \quad B(u(s) - u^*) = s, \\ \lambda(0) = \lambda^* \quad &\text{and} \quad u(0) = u^*. \end{aligned}$$

Given a family $\{\mathscr{T}_h\}_{0<h\leq 1}$ of isoparametric triangulations of Ω, we consider the corresponding family $\{V_h\}_{0<h\leq 1}$ of finite element subspaces of degree $k = 1$ or 2

$$V_h = \{v \in C^0(\mathbb{R}^2); v(x) = 0 \text{ for all } x \notin \Omega_h, v_{|T} \in P_T \text{ for all } T \in \mathscr{T}_h\}, \tag{17.4}$$

see (3.27). A Galerkin approximation to the problem (17.3) consists in finding $\lambda \in \mathbb{R}$ and $u_h \in V_h$ such that

$$\text{for all } v_h \in V_h \quad \langle F(\lambda, u_h), v_h \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} = 0. \tag{17.5}$$

We will study problem (17.5) in a neighborhood of the nondegenerate simple limit point (λ^*, u^*) by applying the results of Section 16 with $X = Y = H_0^1(\Omega)$.

It is not difficult to check that the mapping F in (17.2) has the structure (16.2) with the isomorphism $L = (-\Delta) : H_0^1(\Omega) \to H^{-1}(\Omega)$ and the C^∞ mapping $G(\lambda, .) = -\lambda f(.)$ such that $\lambda Df(v) \in \mathscr{L}(H_0^1(\Omega); H^{-1}(\Omega))$ is compact for all $v \in H_0^1(\Omega)$.

The bilinear form $b : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ in our example is

$$\text{for all } v, w \in H_0^1(\Omega) \quad b(v, w) = \int_\Omega \mathbf{grad} v\, \mathbf{grad} w\, dx.$$

Then problem (17.5) can be written equivalently in the form: find $\lambda \in \mathbb{R}$ and $u_h \in H_0^1(\Omega)$ such that for all $v \in H_0^1(\Omega)$

$$\langle F_h(\lambda, u_h), v \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} \equiv \langle F(\lambda, u_h), \Pi_{V_h} v \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)} + b(u_h, v - \Pi_{V_h} v) = 0 \tag{17.6}$$

or equivalently

$$\int_\Omega \mathbf{grad} u_h \, \mathbf{grad} v \, dx - \lambda \int_\Omega f(u_h) \Pi_{V_h} v \, dx = 0,$$

with the elliptic projector $\Pi_{V_h} \in \mathscr{L}(H_0^1(\Omega); V_h)$. Then the stability assumptions (16.8) and (16.9) are immediate. Furthermore the approximability hypothesis (16.10) is a consequence of the classical polynomial interpolation results.

We can apply Theorem 16.1.

Theorem 17.1. *There exist $h_0 > 0$, $\delta_0 > 0$, and for all $0 < h \leq h_0$ a C^∞ mapping $(\lambda_h, u_h) : s \in [-\epsilon, \epsilon] \to (\lambda_h(s), u_h(s)) \in \mathbb{R} \times H_0^1(\Omega)$ such that*

$$\Big((\lambda, u_h) \in \mathbb{R} \times V_h \text{ is a solution to (17.5)}, B(u_h - u(s)) = 0, \text{and} \quad (\lambda, u_h) \in \overline{B}\big((\lambda(s), u(s)), \delta_0\big) \Big) \iff \big(\lambda = \lambda_h(s), u_h = u_h(s)\big). \tag{17.7}$$

Moreover there exists a constant $C > 0$ such that for all $s \in [-\epsilon, \epsilon]$ and $0 < h \leq h_0$

$$|\lambda(s) - \lambda_h(s)| + |u(s) - u_h(s)|_{1,\Omega} \leq C h^k, \tag{17.8}$$

$k = 1$ or 2 depending on V_h.

There exists a $s_h \in [-\epsilon, \epsilon]$ such that the solution $(\lambda_h(s_h), u_h(s_h))$ is a nondegenerate simple limit point of the problem (17.6) and the following estimates hold

$$|\lambda^* - \lambda_h(0)| \leq C h^{2k} \tag{17.9}$$

and

$$|\lambda^* - \lambda_h(s_h)| \leq C h^{2k}. \tag{17.10}$$

Proof. We have checked before that we can apply the results of the previous section. The estimate (17.8) is a consequence of the (16.21) one. The estimates (17.9), (17.10) are got from the (16.36), (16.37) ones. Notice that an estimate for the term

$$|\frac{d}{ds}\lambda(0) - \frac{d}{ds}\lambda_h(0)|$$

is a consequence of Theorem 12.4 via the estimate (12.14). □

CHAPTER V

APPROXIMATION OF SIMPLE BIFURCATION POINTS

Let X, Z be two real Banach spaces and $F : X \to Z$ be a C^p mapping, $p \geq 1$. In this chapter we shall analyze approximations of the problem

$$F(x) = 0$$

in a neighborhood of a solution $x_0 \in X$ when $DF(x_0)$ is a not surjective Fredholm operator of index greater than 1, that is to say

(i) $DF(x_0) \in \mathscr{L}(X; Z)$ has a finite-dimensional kernel,

(ii) $\mathrm{Range}(DF(x_0))$ is closed, different from Z,

(iii) $\dim \mathrm{Ker}(DF(x_0)) - \dim\,(Z/\mathrm{Range}(DF(x_0))) \geq 1$,

where $Z/\mathrm{Range}(DF(x_0))$ denotes the quotient space of Z by $\mathrm{Range}(DF(x_0))$. Such a solution x_0 is called singular. Compared to the case treated in the previous chapter, where the solution x_0 was supposed to be regular, that is satisfying the assumption (11.1), the study is much more involved. Our goal is to present the method in a simple way and to complete it by bibliographical comments.

In Section 18 we develop a procedure to study the exact problem. This method is the standard Lyapunov-Schmidt decomposition, leading to the bifurcation equation. The common case when $X = \mathbb{R} \times \mathscr{X}$, $\mathbb{R}$ representing the parameter space and the Banach space $\mathscr{X}$ being the state space, is detailed in Section 19. Then approximations are analyzed in Section 20 where we get estimates between the bifurcation equation and the approximation bifurcation equations. In Section 21, we compare thoroughly the solutions of the bifurcation equation and the ones of the approximation bifurcation equations when these equations are defined in $\mathbb{R}^2$. A simple model example is presented in Section 22 to illustrate the general theory. Finally in Section 23 we address some further comments and bibliographical complements.

18. Lyapunov-Schmidt procedure. Bifurcation equations

Let X, Z be two real Banach spaces and $F : X \to Z$ be a C^p mapping, $p \geq 1$. A point $x \in X$ is called singular if

$$\begin{gathered} DF(x) \in \mathscr{L}(X; Z) \text{ is a Fredholm operator of index } m, m \in \mathbb{N}, m \geq 1, \\ \text{and } \mathrm{Range}(DF(x_0)) \neq Z; \end{gathered} \tag{18.1}$$

the index of the operator $DF(x)$ is

$$\text{index}\,(DF(x)) = \dim \text{Ker}(DF(x)) - \dim \text{Coker}(DF(x)) = m,$$

where $\text{Coker}(DF(x))$ is the quotient space $Z/\text{Range}(DF(x))$, and the codimension of $\text{Range}(DF(x))$ (the dimension of the cokernel) is equal to $k \geq 1$.

In this section, we will present a method to analyze the solution set S of the problem

$$F(x) = 0, \tag{18.2}$$

in a neighborhood of the singular point $x_0 \in X$ satisfying $F(x_0) = 0$. Such a point x_0 is called a singular solution.

Let m be the index of $DF(x_0)$ and k be the dimension of $\text{Coker}(DF(x_0))$.

Since the dimension of $\text{Ker}(DF(x_0))$ is finite this subspace is direct, that is there exists a closed subspace $X_1 \subset X$ such that

$$X = X_1 \oplus \text{Ker}(DF(x_0)).$$

The range of $DF(x_0) \subset Z$ is of finite codimension k, so this subspace is direct. There exists a subspace $Z_1 \subset Z$, of dimension k, such that

$$Z = Z_1 \oplus \text{Range}(DF(x_0)).$$

Subsequently in this chapter we shall use the following notations. A point $x \in X$ has a unique representation of the form

$$x = x_1 + x_2 \quad \text{with } x_1 \in X_1 \text{ and } x_2 \in \text{Ker}(DF(x_0))$$

and $P_1 \in \mathscr{L}(X; X_1)$, $P_2 \in \mathscr{L}(X; \text{Ker}(DF(x_0)))$ denote the projectors associated to the decomposition of X, given by

$$P_1(x) = x_1, \qquad P_2(x) = x_2.$$

In a similar way a point $z \in Z$ has a unique representation of the form

$$z = z_1 + z_2 \quad \text{with } z_1 \in Z_1 \text{ and } z_2 \in \text{Range}(DF(x_0))$$

and $Q_1 \in \mathscr{L}(Z; Z_1)$, $Q_2 \in \mathscr{L}(Z; \text{Range}(DF(x_0)))$ denote the projectors associated to the decomposition of Z, given by

$$Q_1(z) = z_1, \qquad Q_2(z) = z_2.$$

With these notations problem (18.2) can be written in the equivalent form

$$Q_1 F(x) = 0 \qquad \text{and} \qquad Q_2 F(x) = 0.$$

Since we study problem (18.2) in a neighborhood of the singular solution x_0, we write it in the following equivalent way: find $x_1 \in X_1$, $x_2 \in \mathrm{Ker}(DF(x_0))$ such that

$$Q_1 F(x_0 + x_1 + x_2) = 0 \tag{18.3}$$

and

$$Q_2 F(x_0 + x_1 + x_2) = 0. \tag{18.4}$$

Let us look first at the equation (18.4). The mapping $\mathscr{F} : X_1 \times \mathrm{Ker}(DF(x_0)) \to \mathrm{Range}(DF(x_0))$ is defined for $(x_1, x_2) \in X_1 \times \mathrm{Ker}(DF(x_0))$ by

$$\mathscr{F}(x_1, x_2) = Q_2 F(x_0 + x_1 + x_2). \tag{18.5}$$

We have $\mathscr{F}(0,0) = 0$ and the derivative $D_{x_1}\mathscr{F}(0,0) \in \mathscr{L}(X_1; \mathrm{Range}(DF(x_0))$ is equal to $Q_2 DF(x_0)$, which is an isomorphism from X_1 onto $\mathrm{Range}(DF(x_0))$.

We apply the implicit function theorem 2.3. There exist positive constants ϵ, η, and a mapping $g : B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to X_1$ which is of class C^p, such that for all $x_2 \in B(0,\epsilon)$

$$\big(\mathscr{F}(x_1, x_2) = 0 \quad \text{and} \quad x_1 \in B(0,\eta) \subset X_1\big) \Longleftrightarrow x_1 = g(x_2).$$

In a neighborhood of x_0, the solution set to $Q_2 F(x) = 0$ is the set

$$S_{\mathscr{F}} = \{x \in X; x = x_0 + g(x_2) + x_2,\ x_2 \in B(0,\epsilon)\},$$

so

$$Q_2 F(x_0 + g(x_2) + x_2) = 0 \quad \text{for all } x_2 \in B(0,\epsilon). \tag{18.6}$$

To complete our study, we need to analyze the equation (18.3) on the solution set $S_{\mathscr{F}}$. The mapping $\mathscr{G} : B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to Z_1$ is defined for $x_2 \in B(0,\epsilon)$ by

$$\mathscr{G}(x_2) = Q_1 F(x_0 + g(x_2) + x_2). \tag{18.7}$$

Notice that the mapping $\mathscr{G}$ can be identified to a mapping defined in $\mathbb{R}^{k+m}$ with values in $\mathbb{R}^k$. Indeed let $\zeta_1, \ldots, \zeta_k$ be a basis of Z_1 and $\varphi_1, \ldots, \varphi_{k+m}$ be a basis of $\mathrm{Ker}(DF(x_0))$. Then we can consider for a convenient $\delta > 0$ the function $\tilde{\mathscr{G}} : B(0,\delta) \subset \mathbb{R}^{k+m} \to \mathbb{R}^k$ given by: for $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_{k+m}) \in B(0,\delta)$

$$\tilde{\mathscr{G}}(\boldsymbol{\alpha}) = \left(\mathscr{G}_1\big(\sum_{j=1}^{k+m} \alpha_j \varphi_j\big), \ldots, \mathscr{G}_k\big(\sum_{j=1}^{k+m} \alpha_j \varphi_j\big)\right)$$

where

$$\mathscr{G}(x_2) = \sum_{\ell=1}^{k} \mathscr{G}_\ell(x_2)\zeta_\ell.$$

Therefore solving problem (18.2) in a neighborhood of the singular solution x_0 has been reduced to solving in a neighborhood of the origin the equation

$$\mathscr{G}(x_2) = 0 \tag{18.8}$$

or the finite nonlinear system of k equations with $k+m$ unknowns

$$\tilde{\mathscr{G}}(\boldsymbol{\alpha}) = 0. \tag{18.8'}$$

The equation (18.8) (or (18.8')) is called the bifurcation equation of problem (18.2) at x_0. The method we have used to get the bifurcation equation is called the Lyapunov-Schmidt procedure.

Remark 18.1. There is a different way to derive the bifurcation equation, as proposed in CROUZEIX and RAPPAZ [1989]. The idea is similar to the one we have used to study a solution set containing a simple limit point. It consists in transforming the problem to reduce it to a simpler one.

Let the mapping $\Phi : \mathbb{R}^k \times X \to Z$ be given by: for $(\boldsymbol{\mu}, x) \in \mathbb{R}^k \times X$

$$\Phi(\boldsymbol{\mu}, x) = F(x) + \sum_{j=1}^{k} \mu_j \zeta_j.$$

Recall that the elements $\zeta_1, \ldots, \zeta_k$ form a basis of Z_1. Then finding the solutions of problem (18.2) in a neighborhood of x_0 is equivalent to find the solutions $(\boldsymbol{\mu}, x) \in \mathbb{R}^k \times X$ of

$$\Phi(\boldsymbol{\mu}, x) = 0 \tag{18.9}$$

in a neighborhood of $(0, x_0)$ satisfying $\boldsymbol{\mu} = 0$. Then problem (18.9) is solved in two steps. First we look for the solutions of (18.9) in a neighborhood of $(0, x_0)$. Secondly we determine all the solutions $(\boldsymbol{\mu}, x)$ with $\boldsymbol{\mu} = 0$.

Remark that $D\Phi(0, x_0) \in \mathscr{L}(\mathbb{R}^k \times X; Z)$ is a Fredholm operator with index $k+m$. Indeed $\mathrm{Ker}(D\Phi(0, x_0)) = \{(0, x) \in \mathbb{R}^k \times X; x \in \mathrm{Ker}(DF(x_0))\}$. Moreover the range of $D\Phi(0, x_0)$ is equal to Z. We have a problem studied in Section 15 when $k = 1$. The case $k \geq 2$ can be treated in a similar way.

In the second step we look at the solutions of (18.9) with $\boldsymbol{\mu} = 0$, which is precisely the bifurcation equation. □

Remark 18.2. Since $F(x_0) = 0$, we easily check that $g(0) = 0$ and then $\mathscr{G}(0) = 0$. Moreover when differentiating (18.7) with respect to x_2, we get $D\mathscr{G}(0) = 0$. Hence problem (18.8) is singular. In the next section we will study the bifurcation equation more carefully. □

19. The case with parameter and state spaces

In this section, we analyze more carefully the bifurcation equation deduced from the Lyapunov-Schmidt procedure when the C^p mapping F, $p \geq 2$, is

$$F : (\lambda, \kappa) \in \mathbb{R} \times \mathscr{X} \to F(\lambda, \kappa) \in Z$$

where $\mathbb{R}$ represents the parameter space while $\mathscr{X}$ is the state space.

Let $(\lambda_0, \kappa_0) \in \mathbb{R} \times \mathscr{X}$ be a singular solution of

$$F(\lambda, \kappa) = 0 \tag{19.1}$$

such that

$$\begin{cases} \text{(i)} \quad \dim \operatorname{Ker}(D_\kappa F(\lambda_0, \kappa_0)) = 1, \\ \text{(ii)} \quad \operatorname{codim}(D_\kappa F(\lambda_0, \kappa_0)) = 1, \\ \text{(iii)} \quad D_\lambda F(\lambda_0, \kappa_0) = 0. \end{cases} \tag{19.2}$$

The last assumption in (19.2) is satisfied when we study bifurcation from the trivial branch of solutions $\{(\lambda, 0); \lambda \in \mathbb{R}\}$, when F satisfies $F(\lambda, 0) = 0$ for all $\lambda \in \mathbb{R}$. At the end of the section, we shall extend our results with weaker assumptions than (19.2).

Under the assumptions (19.2), we can write the decomposition of both spaces X and Z in a more detailed way than in the previous section. Notice that the derivative $DF(\lambda_0, \kappa_0) \in \mathscr{L}(\mathbb{R} \times \mathscr{X}; Z)$ satisfies: for all $(\delta, w) \in \mathbb{R} \times \mathscr{X}$

$$DF(\lambda_0, \kappa_0)(\delta, w) = D_\kappa F(\lambda_0, \kappa_0) w.$$

Since the kernel of $D_\kappa F(\lambda_0, \kappa_0)$ is of dimension 1, there is a closed subspace $\mathscr{X}_1 \subset \mathscr{X}$ and $\xi \in \operatorname{Ker}(D_\kappa F(\lambda_0, \kappa_0))$, $\xi \neq 0$, such that

$$\begin{aligned} \mathscr{X} &= \mathscr{X}_1 \oplus \operatorname{Ker}(D_\kappa F(\lambda_0, \kappa_0)) \\ &\equiv \mathscr{X}_1 \oplus \operatorname{span}\{\xi\}. \end{aligned} \tag{19.3}$$

From the assumption (19.2), we deduce that the kernel of $DF(\lambda_0, \kappa_0)$ is of dimension 2 and that

$$\operatorname{Ker}(DF(\lambda_0, \kappa_0)) = \mathbb{R} \times \operatorname{Ker}(D_\kappa F(\lambda_0, \kappa_0)).$$

Hence we can write a decomposition of X

$$\begin{aligned} X &= X_1 \oplus \operatorname{Ker}(DF(\lambda_0, \kappa_0)) \\ &\equiv X_1 \oplus \operatorname{span}\{\varphi_1, \varphi_2\}, \end{aligned} \tag{19.4}$$

where $X_1 = \{0\} \times \mathscr{X}_1$, $\varphi_1 = (1, 0) \in \mathbb{R} \times \mathscr{X}$, and $\varphi_2 = (0, \xi) \in \mathbb{R} \times \mathscr{X}$.

Since the range of $DF(\lambda_0, \kappa_0)$ is equal to the range of $D_\kappa F(\lambda_0, \kappa_0)$ and the codimension of $D_\kappa F(\lambda_0, \kappa_0)$ is 1, the following decomposition of Z holds

$$\begin{aligned} Z &= Z_1 \oplus \text{Range}(D_\kappa F(\lambda_0, \kappa_0)) \\ &\equiv \text{span}\{\zeta_1\} \oplus \text{Range}(D_\kappa F(\lambda_0, \kappa_0)), \end{aligned} \tag{19.5}$$

where ζ_1 does not belong to the range of $D_\kappa F(\lambda_0, \kappa_0)$. Let P_1, P_2 denote the two projectors associated to the decomposition (19.4) of X and Q_1, Q_2 associated to the decomposition (19.5) of Z.

We can apply the Lyapunov-Schmidt procedure given in Section 18. Here we have $k = m = 1$. There exist positive constants ϵ, η, and a mapping

$$g : (\alpha_1, w) \in B(0, \epsilon) \subset \mathbb{R} \times \text{Ker}(D_\kappa F(\lambda_0, \kappa_0)) \to g(\alpha_1, w) \in \mathscr{Z}_1$$

of class C^p such that for all $(\alpha_1, w) \in B(0, \epsilon)$

$$\big(Q_2 F(\lambda_0 + \alpha_1, \kappa_0 + \kappa_1 + w) = 0 \text{ and } \kappa_1 \in B(0, \eta) \subset \mathscr{Z}_1\big) \Longleftrightarrow \kappa_1 = g(\alpha_1, w).$$

In fact $w = \alpha_2 \xi$ and g can be identified to a function defined on $B(0, \delta) \subset \mathbb{R}^2$, δ small enough, by setting

$$g(\alpha_1, \alpha_2) = g(\alpha_1, \alpha_2 \xi).$$

Then g satisfies for all $(\alpha_1, \alpha_2) \in B(0, \delta)$

$$Q_2 F(\lambda_0 + \alpha_1, \kappa_0 + g(\alpha_1, \alpha_2) + \alpha_2 \xi) = 0. \tag{19.6}$$

In particuliar $g(0, 0) = 0$. Let the functional B in Z' satisfy

$$B(\zeta_1) = 1, \quad B(z) = 0 \quad \text{for all } z \in \text{Range}(DF(\lambda_0, \kappa_0)).$$

We can write then the bifurcation equation in the following way

$$f(\alpha_1, \alpha_2) \equiv B\big(F(\lambda_0 + \alpha_1, \kappa_0 + g(\alpha_1, \alpha_2) + \alpha_2 \xi)\big) = 0, \tag{19.7}$$

where $f : B(0, \delta) \subset \mathbb{R}^2 \to \mathbb{R}$ is of class C^p. We easily check that

$$f(0, 0) = 0, \quad \frac{\partial f}{\partial \alpha_1}(0, 0) = 0, \quad \text{and} \quad \frac{\partial f}{\partial \alpha_2}(0, 0) = 0,$$

which are relations we expect for a bifurcation equation, see Remark 18.2. To study the solutions of the bifurcation equation in a neighborhood of the origin, we need to compute the second derivative of f with respect to α_1 and α_2, the Hessian matrix $D^2 f(0, 0)$. When deriving the equation (19.6) with respect to α_1 and α_2, at (0,0), we get with (19.2) (iii)

$$D_\kappa F(\lambda_0, \kappa_0) \frac{\partial g}{\partial \alpha_1}(0, 0) = 0, \qquad D_\kappa F(\lambda_0, \kappa_0)\big(\frac{\partial g}{\partial \alpha_2}(0, 0) + \xi\big) = 0,$$

which implies

$$\frac{\partial g}{\partial \alpha_1}(0,0) = 0 = \frac{\partial g}{\partial \alpha_2}(0,0). \tag{19.8}$$

Then a simple calculation gives

$$\begin{aligned}
\frac{\partial^2 f}{\partial \alpha_1^2}(0,0) &= B(D^2_{\lambda\lambda} F(\lambda_0, \kappa_0)), \\
\frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2}(0,0) &= B(D^2_{\lambda\kappa} F(\lambda_0, \kappa_0)\xi), \\
\frac{\partial^2 f}{\partial \alpha_2^2}(0,0) &= B(D^2_{\kappa\kappa} F(\lambda_0, \kappa_0)(\xi,\xi)).
\end{aligned}$$

The determinant of the Hessian matrix $D^2 f(0,0)$ is

$$\begin{aligned}
\det(D^2 f(0,0)) = {} & B(D^2_{\lambda\lambda} F(\lambda_0, \kappa_0)) B(D^2_{\kappa\kappa} F(\lambda_0, \kappa_0)(\xi,\xi)) \\
& - \big(B(D^2_{\lambda\kappa} F(\lambda_0, \kappa_0)\xi)\big)^2.
\end{aligned} \tag{19.9}$$

The bifurcation equation is commonly studied with the Morse lemma. In fact the results are strongly depending on the sign of the determinant of $D^2 f(0,0)$.

Theorem 19.1. *(Morse Lemma). Let H be a Hilbert space with the scalar product $(.,.)_H$ and the norm $\|.\|_H$. Let $f : V \subset H \to \mathbb{R}$ be a mapping of class C^1 defined on a neighborhood V of the origin. We assume that $D^2 f(0)$ exists. Let the continuous symmetric linear operator $S : H \to H$ be defined by: for all u, v in H*

$$(Su, v)_H = D^2 f(0)(u, v).$$

We assume that $Df(0) = 0$ and that S is an isomorphism onto H.

Then there exist two neighborhoods $\mathscr{V}$, $\mathscr{W}$ of the origin in H, and a C^1 diffeomorphism $\Phi : \mathscr{V} \subset V \to \mathscr{W}$ satisfying

$$\Phi(0) = 0, \qquad D\Phi(0) = I,$$

and for all $v \in \mathscr{V}$

$$f(v) = f(0) + \frac{1}{2} D^2 f(0)(\Phi(v), \Phi(v)).$$

Moreover if f is in $C^m(V;\mathbb{R})$ with $m \geq 1$ and if $D^{m+1} f(0)$ exists, then Φ is of class C^m. □

We can find proofs of the Morse lemma for instance in Nirenberg [1974] or in Crouzeix and Rappaz [1989]. We do not repeat the proof here.

We assume from now on that the determinant of the Hessian matrix $D^2 f(0,0)$ is different from 0 and apply the Morse lemma to the bifurcation equation (19.7). There exist two neighborhoods $\mathscr{V}$, $\mathscr{W}$ of the origin in $\mathbb{R}^2$ and a C^1 diffeomorphism $\mathbf{\Phi} : \mathscr{V} \subset B(0,\delta) \subset \mathbb{R}^2 \to \mathscr{W} \subset \mathbb{R}^2$ satisfying

$$\mathbf{\Phi}(0,0) = 0, \qquad D\mathbf{\Phi}(0,0) = I,$$

and for all $(\alpha_1, \alpha_2) \in \mathscr{V}$

$$f(\alpha_1, \alpha_2) = \frac{1}{2}\big(D^2 f(0,0)\mathbf{\Phi}(\alpha_1, \alpha_2), \mathbf{\Phi}(\alpha_1, \alpha_2)\big),$$

here $(.,.)$ is the scalar product in $\mathbb{R}^2$.

We analyze now the two different cases which can occur depending on the sign of the determinant of $D^2 f(0,0)$. First let us assume that

$$\det(D^2 f(0,0)) > 0, \tag{19.10}$$

which is the elliptic case. Since the matrix $D^2 f(0,0)$ is symmetric, there exists a orthogonal matrix Q such that

$$Q^t D Q = D^2 f(0,0),$$

where the diagonal matrix D has the diagonal elements d_1 and d_2, d_i eigenvalue of $D^2 f(0,0)$. Let the vector $\mathbf{\Psi}$ of components Ψ_1, Ψ_2, be

$$\mathbf{\Psi}(\alpha_1, \alpha_2) = Q\mathbf{\Phi}(\alpha_1, \alpha_2),$$

then for $(\alpha_1, \alpha_2) \in \mathscr{V}$

$$f(\alpha_1, \alpha_2) = \frac{d_1}{2}\Psi_1^2(\alpha_1, \alpha_2) + \frac{d_2}{2}\Psi_2^2(\alpha_1, \alpha_2).$$

Then the bifurcation equation reads

$$d_1 \Psi_1^2(\alpha_1, \alpha_2) + d_2 \Psi_2^2(\alpha_1, \alpha_2) = 0,$$

where d_1 and d_2 have the same sign. Since we have

$$\mathbf{\Psi}(0,0) = 0 \quad \text{and} \quad D\mathbf{\Psi}(0,0) \text{ regular},$$

so in a neighborhood of (0,0), the bifurcation equation has the only solution (0,0) and the solution set

$$S = \{(\lambda, \kappa) \in \mathbb{R} \times \mathscr{X}; F(\lambda, \kappa) = 0\}$$

is reduced to $\{(\lambda_0, \kappa_0)\}$ in a neighborhood of (λ_0, κ_0). Such a solution (λ_0, κ_0) to problem (19.1) is called an isolated solution.

We assume now that

$$\det(D^2 f(0,0)) < 0, \tag{19.11}$$

which is the hyperbolic case. Since the matrix $Df(0,0)$ is regular and symmetric, there exists a regular matrix R such that

$$R^t \begin{pmatrix} 0 & \frac{1}{2} \\ \frac{1}{2} & 0 \end{pmatrix} R = D^2 f(0,0).$$

Let the vector $\mathbf{\Psi}$ of components Ψ_1, Ψ_2, be

$$\mathbf{\Psi}(\alpha_1, \alpha_2) = R\mathbf{\Phi}(\alpha_1, \alpha_2),$$

then the bifurcation equation reads: for all $(\alpha_1, \alpha_2) \in \mathscr{V}$

$$f(\alpha_1, \alpha_2) \equiv \Psi_1(\alpha_1, \alpha_2)\Psi_2(\alpha_1, \alpha_2) = 0. \tag{19.12}$$

Notice that

$$\mathbf{\Psi}(0,0) = 0 \quad \text{and} \quad D\mathbf{\Psi}(0,0) \text{ is regular.} \tag{19.13}$$

From (19.12), (19.13), and the Morse lemma, we deduce that the solution set of f

$$\mathscr{S}_f (\equiv \{(\alpha_1, \alpha_2) \in B(0,\delta); f(\alpha_1, \alpha_2) = 0\}) \cap \mathscr{V}$$

is C^{p-1} diffeomorphic to two parts of straight lines (in the coordinates Ψ_1, Ψ_2) which intersect at $(0,0) \in \mathbb{R}^2$. Consequently in a neighborhood of (λ_0, κ_0) the solution set S of (19.1) consists in two curves which intersect at the point (λ_0, κ_0).

More precisely let $\boldsymbol{\alpha}(\mathbf{\Psi})$ denote the inverse of the C^{p-1} diffeomorphism $\mathbf{\Psi}(\boldsymbol{\alpha})$ defined for $\mathbf{\Psi}$ in $R\mathscr{W}$, say for $|\Psi_1| + |\Psi_2| \leq \epsilon$ for simplicity. Then

$$\mathscr{S}_f = \{(\alpha_1(s,0), \alpha_2(s,0)); |s| \leq \epsilon\} \cup \{(\alpha_1(0,s), \alpha_2(0,s)); |s| \leq \epsilon\}.$$

For the parameter value s, with $|s| \leq \epsilon$, we set

$$\begin{aligned} \lambda_1(s) &= \lambda_0 + \alpha_1(s,0), & \kappa_1(s) &= \kappa_0 + g(\alpha_1(s,0), \alpha_2(s,0)) + \alpha_2(s,0)\xi, \\ \lambda_2(s) &= \lambda_0 + \alpha_1(0,s), & \kappa_2(s) &= \kappa_0 + g(\alpha_1(0,s), \alpha_2(0,s)) + \alpha_2(0,s)\xi. \end{aligned}$$

Clearly for $i = 1$ or 2 and $|s| \leq \epsilon$

$$F(\lambda_i(s), \kappa_i(s)) = 0,$$
$$\lambda_i(0) = \lambda_0, \kappa_i(0) = \kappa_0 \quad \text{and if } s \neq 0 \quad (\lambda_i(s), \kappa_i(s)) \neq (\lambda_0, \kappa_0).$$

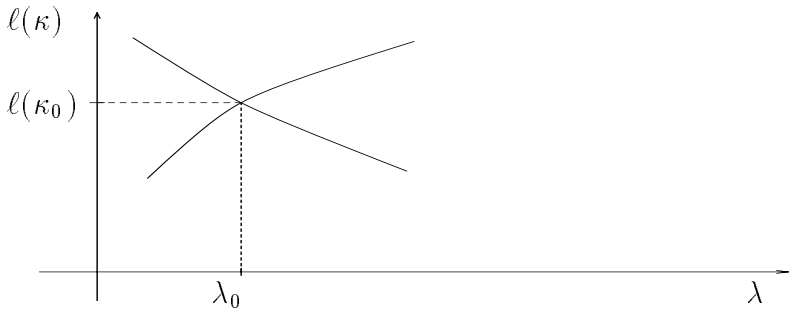

FIGURE 19.1: SIMPLE BIFURCATION POINT.

Moreover the two branches of solutions intersect at (λ_0, κ_0). If ℓ is a linear form on $\mathscr{X}$, then in a neighborhood of (λ_0, κ_0) the solution set of (19.1) can be represented as in the figure 19.1.

Such a solution (λ_0, κ_0) of (19.1) is called a simple bifurcation point.

Before studying bifurcation approximations of simple bifurcation points, we end up the section with some remarks.

Remark 19.1. If the mapping $F : \mathbb{R} \times \mathscr{X} \to Z$ satisfies

$$F(\lambda, 0) = 0 \quad \text{for all } \lambda \in \mathbb{R}$$

and if $(\lambda_0, 0)$ satisfies (19.2) (i) and (ii), then we have

$$D_\lambda F(\lambda_0, 0) = 0 \qquad \text{and} \qquad D^2_{\lambda\lambda} F(\lambda_0, 0) = 0.$$

Clearly the assumption (19.2) (iii) is satisfied. The determinant of the Hessian matrix $D^2 f(0,0)$ given in (19.9) reads

$$\det\big(D^2 f(0,0)\big) = -\big(B(D^2_{\lambda\kappa} F(\lambda_0, 0)\xi\big)^2.$$

So $(\lambda_0, 0)$ is a simple bifurcation point if and only if $B(D^2_{\lambda\kappa} F(\lambda_0, 0)\xi) \neq 0$ or in other words

$$D^2_{\lambda,\kappa} F(\lambda_0, 0)\xi \notin \text{Range}(D_\kappa F(\lambda_0, 0)),$$

which is the Crandall-Rabinowitz condition, see CRANDALL and RABINOWITZ [1973]. Notice that approximations of simple bifurcation points on the trivial branch $\{(\lambda, 0); \lambda \in \mathbb{R}\}$ is studied in CROUZEIX and RAPPAZ [1989], Chapter 5, where they use a simpler approach than the Lyapunov-Schmidt procedure. □

Remark 19.2. If we replace the assumption (19.2) (iii), namely $D_\lambda F(\lambda_0, \kappa_0) = 0$, by the weaker one

$$D_\lambda F(\lambda_0, \kappa_0) \in \text{Range}(D_\kappa F(\lambda_0, \kappa_0)), \tag{19.14}$$

then we can use the above development with minor changes and get a similar conclusion.

Indeed the kernel of $DF(\lambda_0, \kappa_0)$ is of dimension 2 and

$$\mathrm{Ker}(DF(\lambda_0, \kappa_0)) = \mathrm{span}\{\varphi_1, \varphi_2\}$$

with $\varphi_1 = (1, \eta) \in \mathbb{R} \times \mathscr{X}_1$, η the unique element satisfying

$$D_\lambda F(\lambda_0, \kappa_0) + D_\kappa F(\lambda_0, \kappa_0)\eta = 0,$$

and with $\varphi_2 = (0, \xi) \in \mathbb{R} \times \mathscr{X}$, $\xi \neq 0$ and

$$D_\kappa F(\lambda_0, \kappa_0)\xi = 0.$$

With that choice of φ_1 and φ_2, the decomposition (19.4) for X still remains valid, that is

$$X = X_1 \oplus \mathrm{span}\{\varphi_1, \varphi_2\} \tag{19.15}$$

where $X_1 = \{0\} \times \mathscr{X}_1$. Then the equation (19.6) is modified in

$$Q_2 F(\lambda_0 + \alpha_1, \kappa_0 + g(\alpha_1, \alpha_2) + \alpha_1\eta + \alpha_2\xi) = 0 \tag{19.16}$$

and the bifurcation equation (19.7) in

$$f(\alpha_1, \alpha_2) \equiv B(F(\lambda_0 + \alpha_1, \kappa_0 + g(\alpha_1, \alpha_2) + \alpha_1\eta + \alpha_2\xi)) = 0. \tag{19.17}$$

When deriving the equation (19.16) with respect to α_1 and α_2, at $(0, 0)$, we get

$$D_\lambda F(\lambda_0, \kappa_0) + D_\kappa F(\lambda_0, \kappa_0)\Big(\frac{\partial g}{\partial \alpha_1}(0, 0) + \eta\Big) = 0,$$
$$D_\kappa F(\lambda_0, \kappa_0)\Big(\frac{\partial g}{\partial \alpha_2}(0, 0) + \xi\Big) = 0,$$

which implies

$$\frac{\partial g}{\partial \alpha_1}(0, 0) = 0 = \frac{\partial g}{\partial \alpha_2}(0, 0).$$

Then a simple calculation gives the following expressions for the elements of the Hessian matrix $D^2 f(0, 0)$

$$\frac{\partial^2 f}{\partial \alpha_1^2}(0, 0) = B(D^2_{\lambda\lambda} F(\lambda_0, \kappa_0) + 2D^2_{\lambda\kappa} F(\lambda_0, \kappa_0)\eta + D^2_{\kappa\kappa} F(\lambda_0, \kappa_0)(\eta, \eta)),$$
$$\frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2}(0, 0) = B(D^2_{\lambda\kappa} F(\lambda_0, \kappa_0)\xi + D^2_{\kappa\kappa} F(\lambda_0, \kappa_0)(\eta, \xi)),$$
$$\frac{\partial^2 f}{\partial \alpha_2^2}(0, 0) = B(D^2_{\kappa\kappa} F(\lambda_0, \kappa_0)(\xi, \xi)).$$

Then the study of the bifurcation equation goes along the same line as before, depending on the sign of the determinant of $D^2 f(0,0)$. $\square$

Remark 19.3. A similar development can be carried out if we replace the assumption (19.2) by

$$\begin{cases} \text{(i)} \quad \dim \mathrm{Ker}(D_\kappa F(\lambda_0, \kappa_0)) = 2, \\ \text{(ii)} \quad \mathrm{codim}(D_\kappa F(\lambda_0, \kappa_0)) = 2, \\ \text{(iii)} \quad D_\lambda F(\lambda_0, \kappa_0) \notin \mathrm{Range}(D_\kappa F(\lambda_0, \kappa_0)). \end{cases} \tag{19.18}$$

Indeed the kernel of $DF(\lambda_0, \kappa_0)$ is of dimension 2 and

$$\mathrm{Ker}(DF(\lambda_0, \kappa_0)) = \mathrm{span}\{\varphi_1, \varphi_2\}$$

with $\varphi_1 = (0, \xi_1) \in \mathbb{R} \times \mathscr{X}$ and $\varphi_2 = (0, \xi_2) \in \mathbb{R} \times \mathscr{X}$ such that

$$\mathrm{Ker}(D_\kappa F(\lambda_0, \kappa_0)) = \mathrm{span}\{\xi_1, \xi_2\}.$$

There exists a closed subspace $\mathscr{X}_1 \subset \mathscr{X}$ such that

$$\mathscr{X} = \mathscr{X}_1 \oplus \mathrm{span}\{\xi_1, \xi_2\}$$

and we can write a decomposition of X

$$X = X_1 \oplus \mathrm{span}\{\varphi_1, \varphi_2\} \tag{19.19}$$

with $X_1 = \mathbb{R} \times \mathscr{X}_1$. The codimension of $DF(\lambda_0, \kappa_0)$ is 1, so the following decomposition of Z holds

$$Z = \mathrm{span}\{\zeta_1\} \oplus \mathrm{Range}(DF(\lambda_0, \kappa_0)). \tag{19.20}$$

The mapping g can be identified to a function defined on $B(0,\delta) \subset \mathbb{R}^2$ with values in $\mathbb{R} \times \mathscr{X}_1$, that is

$$g : (\alpha_1, \alpha_2) \in B(0,\delta) \subset \mathbb{R}^2 \to (g_1(\alpha_1, \alpha_2), g_2(\alpha_1, \alpha_2)) \in \mathbb{R} \times \mathscr{X}_1,$$

and satisfies for all $(\alpha_1, \alpha_2) \in B(0,\delta)$

$$Q_2 F(\lambda_0 + g_1(\alpha_1, \alpha_2), \kappa_0 + g_2(\alpha_1, \alpha_2) + \alpha_1 \xi_1 + \alpha_2 \xi_2) = 0. \tag{19.21}$$

The bifurcation equation reads

$$f(\alpha_1, \alpha_2) \equiv B(F(\lambda_0 + g_1(\alpha_1, \alpha_2), \kappa_0 + g_2(\alpha_1, \alpha_2) + \alpha_1 \xi_1 + \alpha_2 \xi_2)) = 0 \tag{19.22}$$

where $f : B(0,\delta) \subset \mathbb{R}^2 \to \mathbb{R}$ is of class C^p.

When deriving the equation (19.21) with respect to α_1 and α_2, at $(0,0)$, we get

$$D_\lambda F(\lambda_0,\kappa_0)\frac{\partial g_1}{\partial \alpha_1}(0,0) + D_\kappa F(\lambda_0,\kappa_0)(\frac{\partial g_2}{\partial \alpha_1}(0,0)+\xi_1) = 0,$$
$$D_\lambda F(\lambda_0,\kappa_0)\frac{\partial g_1}{\partial \alpha_2}(0,0) + D_\kappa F(\lambda_0,\kappa_0)(\frac{\partial g_2}{\partial \alpha_2}(0,0)+\xi_2) = 0.$$

Since

$$\mathrm{Range}(DF(\lambda_0,\kappa_0)) = \mathrm{span}\{D_\lambda F(\lambda_0,\kappa_0)\} \oplus \mathrm{Range}(D_\kappa F(\lambda_0,\kappa_0))$$

and since ξ_1 and ξ_2 are in the kernel of $D_\kappa F(\lambda_0,\kappa_0)$ we deduce that

$$\frac{\partial g_i}{\partial \alpha_j}(0,0) = 0, \quad 1 \le i,j \le 2.$$

Then a simple calculation gives the following expressions

$$\frac{\partial^2 f}{\partial \alpha_1^2}(0,0) = B(D^2_{\kappa\kappa}F(\lambda_0,\kappa_0)(\xi_1,\xi_1)),$$
$$\frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2}(0,0) = B(D^2_{\kappa\kappa}F(\lambda_0,\kappa_0)(\xi_1,\xi_2)),$$
$$\frac{\partial^2 f}{\partial \alpha_2^2}(0,0) = B(D^2_{\kappa\kappa}F(\lambda_0,\kappa_0)(\xi_2,\xi_2)).$$

The study of the bifurcation equation goes along the same line as before, depending on the sign of the determinant of $D^2 f(0,0)$.

In the hyperbolic case, the solution set of $F(\lambda,x)=0$ in a neighborhood of (λ_0,κ_0) is represented in the figure 19.2, where ℓ is a linear form on $\mathscr{X}$.

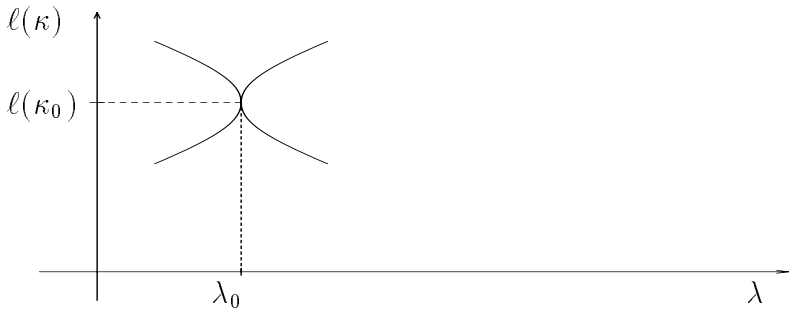

FIGURE 19.2: DOUBLE NONDEGENERATE LIMIT POINT.

Such a solution (λ_0,κ_0) of $F(\lambda,\kappa)=0$ is called a double nondegenerate limit point. □

20. Approximate bifurcation equations

Let X, Z be two real Banach spaces, $F : X \to Z$ be a C^p mapping, $p \geq 1$, and $x_0 \in X$ be a singular solution of the problem: find $x \in X$ such that

$$F(x) = 0. \tag{20.1}$$

We assume that the solution x_0 satisfies

(20.2) $DF(x_0) \in \mathscr{L}(X;Z)$ is a not surjective Fredholm operator of index $m, m \geq 1$.

The study of the solution set S of (20.1) in a neighborhood of x_0 has been done in Sections 18 and 19. We will keep the notations introduced there. Let P_1, P_2 and Q_1, Q_2 be the projectors associated to the decomposition of X and Z,

$$X = X_1 \oplus \mathrm{Ker}(DF(x_0)), \qquad Z = Z_1 \oplus \mathrm{Range}(DF(x_0)).$$

We introduce now a family $\{F_h\}_{0<h\leq 1}$ of C^p mappings, $p \geq 1$,

$$F_h : x \in X \to F_h(x) \in Z$$

which are approximations of F. Our goal is to study the existence and the convergence of solutions of the problem: find $x \in X$ such that

$$F_h(x) = 0 \tag{20.3}$$

in a neighborhood of the singular solution x_0. We assume that the family $\{F_h\}_{0<h\leq 1}$ satisfies: there exists a real number $r > 0$ such that

$$\lim_{h\to 0} \sup_{x\in B(x_0,r)} \|D^\ell F(x) - D^\ell F_h(x)\| = 0, \tag{20.4}$$

for $\ell = 0, 1$.

To study problem (20.3) under the assumption (20.4), we will apply the Lyapunov-Schmidt procedure to the mapping F_h with respect to the decomposition of X and Z given in the analysis of problem (20.1). Since we study problem (20.3) in a neighborhood of x_0, we write it in the following way: find $x_1 \in X_1$, $x_2 \in \mathrm{Ker}(DF(x_0))$ such that

$$Q_1 F_h(x_0 + x_1 + x_2) = 0 \tag{20.5}$$

and

$$Q_2 F_h(x_0 + x_1 + x_2) = 0. \tag{20.6}$$

Corresponding to the mapping $\mathscr{F}$ in (18.5) for the exact problem, we define the mapping $\mathscr{F}_h : X_1 \times \mathrm{Ker}(DF(x_0)) \to \mathrm{Range}(DF(x_0))$ by: for $x_1 \in X_1$, $x_2 \in \mathrm{Ker}(DF(x_0))$,

$$\mathscr{F}_h(x_1, x_2) = Q_2 F_h(x_0 + x_1 + x_2).$$

To study the solutions of (20.6) or of $\mathscr{F}_h(x_1, x_2) = 0$, we use Theorem 12.3, Remark 12.1, and Theorem 12.4, whose assumptions are not difficult to check with the hypothesis (20.4), since $D_{x_1}\mathscr{F}(0,0) = Q_2 DF(x_0) \in \mathscr{L}(X_1; \mathrm{Range}(DF(x_0)))$ is an isomorphism.

Let $g : B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to X_1$ be the C^p mapping corresponding to the Lyapunov-Schmidt decomposition for F, satisfying to (18.6). By Theorem 12.3, there exist positive constants $h_0 \leq 1$, η, and a C^p mapping $g_h : B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to X_1$ such that for all $x_2 \in B(0,\epsilon)$, $0 < h \leq h_0$,

$$\big(\mathscr{F}_h(x_1, x_2) = 0 \quad \text{and} \quad x_1 \in B(0,\eta) \subset X_1\big) \Longleftrightarrow x_1 = g_h(x_2).$$

Moreover the following estimates hold: for all $x_2 \in B(0,\epsilon)$

$$\|g(x_2) - g_h(x_2)\|_X \leq C\|F_h(x_0 + g(x_2) + x_2)\|_Z, \tag{20.7}$$

$$\|Dg_h(x_2)\|_{X;X} \leq C. \tag{20.8}$$

Notice that we also have estimates for higher order derivatives, see (12.14) and (12.15), which are not needed here for our purpose.

To complete our analysis we need to study the equation (20.5) on the solution set of $\mathscr{F}_h$ in $\big(B(0,\epsilon) \times B(0,\eta)\big)$. The approximate analogue of the mapping $\mathscr{G}$, is $\mathscr{G}_h$: $B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0)) \to Z_1$ defined by: for $x_2 \in B(0,\epsilon)$

$$\mathscr{G}_h(x_2) = Q_1 F_h(x_0 + g_h(x_2) + x_2). \tag{20.9}$$

Then the approximate bifurcation equation reads: find $x_2 \in B(0,\epsilon)$ such that

$$\mathscr{G}_h(x_2) = 0. \tag{20.10}$$

Before undertaking with more details the study of (20.10), which is the object of Section 21 in the case developed in Section 19, we still need a comparison result between the solutions of the exact problem (20.1) and the ones of the approximate problems (20.3). For that purpose we define the sets

$$\mathscr{S}_{\mathscr{G}} = \{x_2 \in \mathscr{V}; \mathscr{G}(x_2) = 0\},$$
$$\mathscr{S}_{\mathscr{G}_h} = \{x_2 \in \mathscr{V}; \mathscr{G}_h(x_2) = 0\},$$

where $\mathscr{V} \subset B(0,\epsilon) \subset \mathrm{Ker}(DF(x_0))$ is a compact convex neighborhood in $\mathrm{Ker}(DF(x_0))$ of 0; we also introduce

$$\mathscr{S} = \{x = x_0 + g(x_2) + x_2 \in X; x_2 \in \mathscr{S}_{\mathscr{G}}\},$$
$$\mathscr{S}_h = \{x = x_0 + g_h(x_2) + x_2 \in X; x_2 \in \mathscr{S}_{\mathscr{G}_h}\}.$$

Notice that for $x \in \mathscr{S}$, we have $F(x) = 0$ and for $x_h \in \mathscr{S}_h$, we have $F_h(x_h) = 0$. The distance we will use to compare the sets $\mathscr{S}_{\mathscr{G}}$, $\mathscr{S}_{\mathscr{G}_h}$ and $\mathscr{S}$, $\mathscr{S}_h$ is the Hausdorff metric.

Let Y be a vector space endowed with the metric d and $\mathscr{Y}$ denote the set of all nonempty bounded closed sets in Y. For $(A,B) \in \mathscr{Y} \times \mathscr{Y}$, we define

$$\rho(A,B) = \sup_{y \in A} d(y,B) \equiv \sup_{y \in A}(\inf_{z \in B} d(y,z)),$$
$$\delta(A,B) = \max(\rho(A,B), \rho(B,A)).$$

δ is a metric in $\mathscr{Y}$ called the Hausdorff metric.

Theorem 20.1. *In the framework developed above, there exist positive constants C and h_0 such that for all $0 < h \leq h_0$*

$$\delta(\mathscr{S}, \mathscr{S}_h) \leq C \left[\delta(\mathscr{S}_{\mathscr{G}}, \mathscr{S}_{\mathscr{G}_h}) + \sup_{x \in \mathscr{S}} \|F_h(x)\|_Z \right]. \tag{20.11}$$

Proof. Let $x_2 \in \mathscr{S}_{\mathscr{G}}$, $x_{2h} \in \mathscr{S}_{\mathscr{G}_h}$, and

$$x = x_0 + g(x_2) + x_2, \qquad x_h = x_0 + g_h(x_{2h}) + x_{2h}.$$

By definition we have $x \in \mathscr{S}$ and $x_h \in \mathscr{S}_h$. Furthermore

$$\|x - x_h\|_X \leq \|x_2 - x_{2h}\|_X + \|g(x_2) - g_h(x_2)\|_X + \|g_h(x_2) - g_h(x_{2h})\|_X. \tag{20.12}$$

We introduce in (20.12) both estimates (20.7) and (20.8) and deduce

$$\|x - x_h\|_X \leq C\big(\|x_2 - x_{2h}\|_X + \|F_h(x_0 + g(x_2) + x_2)\|_Z\big)$$

and

$$\|x - x_h\|_X \leq C\big(\|x_2 - x_{2h}\|_X + \sup_{x \in \mathscr{S}} \|F_h(x)\|_Z\big). \tag{20.13}$$

The infimum over $x_h \in \mathscr{S}_h$ in (20.13) leads to

$$\inf_{x_h \in \mathscr{S}_h} \|x - x_h\|_X \leq C\big(\rho(\mathscr{S}_{\mathscr{G}}, \mathscr{S}_{\mathscr{G}_h}) + \sup_{x \in \mathscr{S}} \|F_h(x)\|_Z\big)$$

and

$$\rho(\mathscr{S}, \mathscr{S}_h) \leq C\big(\rho(\mathscr{S}_{\mathscr{G}}, \mathscr{S}_{\mathscr{G}_h}) + \sup_{x \in \mathscr{S}} \|F_h(x)\|_Z\big). \tag{20.14}$$

In the same way we can prove

$$\rho(\mathscr{S}_h, \mathscr{S}) \leq C\big(\rho(\mathscr{S}_{\mathscr{G}_h}, \mathscr{S}_{\mathscr{G}}) + \sup_{x \in \mathscr{S}} \|F_h(x)\|_Z\big). \tag{20.15}$$

The two inequalites (20.14) and (20.15), and the definition of the Hausdorff metric imply the estimate (20.11). □

To obtain some estimates with respect to h for $\delta(\mathscr{S}, \mathscr{S}_h)$, it remains with (20.11) to bound the quantity $\delta(\mathscr{S}_{\mathscr{G}}, \mathscr{S}_{\mathscr{G}_h})$ to some power of h. In the next section we will study this quantity when the functions $\mathscr{G}$ and $\mathscr{G}_h$ are defined on a neighborhood of 0 in $\mathbb{R}^2$ with values in $\mathbb{R}$.

21. Approximation of simple bifurcation points

To complete our study we still need to present a precise comparison between the solution sets of the bifurcation equation $\mathscr{G}(x_2) = 0$ and of the approximate bifurcation equation $\mathscr{G}_h(x_2) = 0$. Then from the estimate (20.11), we could derive error estimates.

For simplicity we analyze the case of a simple bifurcation point which is developed in Section 19. Let $F : \mathbb{R} \times \mathscr{X} \to Z$ be a C^p mapping, $p \geq 2$, and let $(\lambda_0, \kappa_0) \in \mathbb{R} \times \mathscr{X}$ be a singular solution to (19.1) satisfying to (19.2). Then the bifurcation equation reads

$$f(\alpha_1, \alpha_2) = 0, \tag{21.1}$$

where $f : B(0, \delta) \subset \mathbb{R}^2 \to \mathbb{R}$ is of class C^p, $p \geq 2$, and satisfies

$$f(0,0) = 0, \quad \frac{\partial f}{\partial \alpha_1}(0,0) = 0, \quad \text{and} \quad \frac{\partial f}{\partial \alpha_2}(0,0) = 0.$$

The simple bifurcation case occurs under the assumption

$$\det(D^2 f(0,0)) < 0. \tag{21.2}$$

In the following we will compare the solutions of approximate bifurcation equations with respect to the solution of the given bifurcation equation (21.1).

We can apply the theory of Section 20 to a family $\{F_h\}_{0<h\leq 1}$ of C^p mappings, $p \geq 2$,

$$F_h : (\lambda, \kappa) \in \mathbb{R} \times \mathscr{X} \to F_h(\lambda, \kappa) \in Z,$$

satisfying

$$\lim_{h \to 0} \sup_{(\lambda,\kappa) \in B((\lambda_0,\kappa_0),r)} \|D^\ell F(\lambda, \kappa) - D^\ell F_h(\lambda, \kappa)\| = 0, \tag{21.3}$$

for some $r > 0$, $\ell = 0, 1$. Then we get a family of bifurcation equations, say for $h \in (0, h_0]$

$$f_h(\alpha_1, \alpha_2) = 0, \tag{21.4}$$

where $f_h : B(0, \delta) \subset \mathbb{R}^2 \to \mathbb{R}$ is of class C^p, only to restrict δ we keep the same as the one in the definition of f. With the assumption (21.3) we deduce that

$$\lim_{h \to 0} \sup_{(\alpha_1,\alpha_2) \in B(0,\delta)} |f(\alpha_1, \alpha_2) - f_h(\alpha_1, \alpha_2)| = 0. \tag{21.5}$$

We define the two solution sets

$$\mathscr{S}_f = \{(\alpha_1, \alpha_2) \in B(0, \delta); f(\alpha_1, \alpha_2) = 0\},$$
$$\mathscr{S}_{f_h} = \{(\alpha_1, \alpha_2) \in B(0, \delta); f_h(\alpha_1, \alpha_2) = 0\}.$$

We compare now the behavior of $\mathscr{S}_{f_h}$ and of $\mathscr{S}_f$. Our goal is to get error estimates for the Hausdorff distance $\delta(\mathscr{S}_f, \mathscr{S}_{f_h})$ and to give qualitative results for $\mathscr{S}_{f_h}$.

Theorem 21.1. *Under the assumptions (21.2) and (21.5), there exists a constant $C > 0$ independent of h such that*

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C\Big(\sup_{(\alpha_1,\alpha_2)\in B(0,\delta)} |f(\alpha_1,\alpha_2) - f_h(\alpha_1,\alpha_2)|\Big)^{1/2}. \tag{21.6}$$

Proof. For notation simplicity we set

$$\epsilon_h = \sup_{(\alpha_1,\alpha_2)\in B(0,\delta)} |f(\alpha_1,\alpha_2) - f_h(\alpha_1,\alpha_2)|.$$

Applying the Morse lemma, we deduce that there exists a C^{p-1} diffeomorphism $\mathbf{\Psi}$ from $B(0,\delta)$, only to restrict δ, into $\mathbb{R}^2$ such that the bifurcation equation reads

$$f(\alpha_1,\alpha_2) \equiv \Psi_1(\alpha_1,\alpha_2)\Psi_2(\alpha_1,\alpha_2) = 0,$$

see (19.12). Consequently without loss of generality, we can assume that the function f is, for $(\alpha_1,\alpha_2) \in B(0,\delta)$

$$f(\alpha_1,\alpha_2) = \alpha_1\alpha_2.$$

Then $\rho(\mathscr{S}_{f_h}, \mathscr{S}_f)$ given by

$$\rho(\mathscr{S}_{f_h}, \mathscr{S}_f) = \sup_{(\alpha_1,\alpha_2)\in\mathscr{S}_{f_h}} \Big(\inf_{(\beta_1,\beta_2)\in\mathscr{S}_f} \|(\alpha_1,\alpha_2) - (\beta_1,\beta_2)\|\Big),$$

where $\|.\|$ denotes the Euclidian norm in $\mathbb{R}^2$, can be bounded by

$$\rho(\mathscr{S}_{f_h}, \mathscr{S}_f) \leq \epsilon_h. \tag{21.7}$$

Let us estimate now $\rho(\mathscr{S}_f, \mathscr{S}_{f_h})$. We define the sets

$$C_h^+ = \{(\alpha_1,\alpha_2) \in B(0,\delta); f(\alpha_1,\alpha_2) \geq \epsilon_h\},$$
$$C_h^- = \{(\alpha_1,\alpha_2) \in B(0,\delta); f(\alpha_1,\alpha_2) \leq -\epsilon_h\}.$$

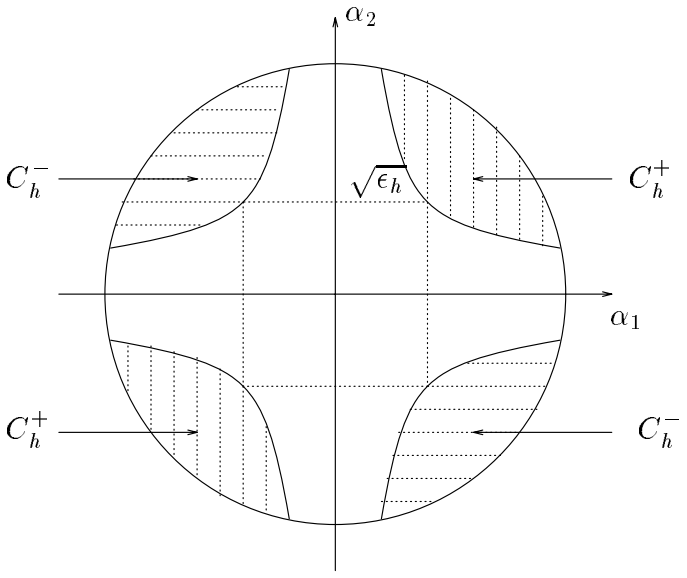

FIGURE 21.1: SETS C_h^+ AND C_h^-.

From the assumption (21.5) we know that ϵ_h tends to 0 when h tends to 0, so for h small enough both sets C_h^+ and C_h^- are not empty. Let (α_1, α_2) be in $\mathscr{S}_f$. There exist $(\alpha_{1h}^+, \alpha_{2h}^+)$ in C_h^+ and $(\alpha_{1h}^-, \alpha_{2h}^-)$ in C_h^- satisfying

$$\|(\alpha_1, \alpha_2) - (\alpha_{1h}^+, \alpha_{2h}^+)\| \leq \sqrt{2\epsilon_h}, \qquad \|(\alpha_1, \alpha_2) - (\alpha_{1h}^-, \alpha_{2h}^-)\| \leq \sqrt{2\epsilon_h}.$$

By the definition of ϵ_h we have

$$f_h(\alpha_{1h}^+, \alpha_{2h}^+) \geq f(\alpha_{1h}^+, \alpha_{2h}^+) - \epsilon_h \geq 0,$$
$$f_h(\alpha_{1h}^-, \alpha_{2h}^-) \leq f(\alpha_{1h}^-, \alpha_{2h}^-) + \epsilon_h \leq 0.$$

So there exists $\theta \in [0, 1]$ such that

$$f_h(\theta(\alpha_{1h}^+, \alpha_{2h}^+) + (1-\theta)(\alpha_{1h}^-, \alpha_{2h}^-)) = 0$$

and if we denote by $(\alpha_{1h}^0, \alpha_{2h}^0)$ that solution of f_h, we have

$$\|(\alpha_1, \alpha_2) - (\alpha_{1h}^0, \alpha_{2h}^0)\| \leq \sqrt{2\epsilon_h}.$$

Then we deduce

$$\rho(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq \sqrt{2\epsilon_h}. \tag{21.8}$$

Finally from both estimates (21.7) and (21.8) we can conclude. □

Remark 21.1. If we do not add any assumptions to (21.2) and (21.5), then the estimate (21.6) is optimal. Consider $f(\alpha_1, \alpha_2) = \alpha_1\alpha_2$ and $f_h(\alpha_1, \alpha_2) = \alpha_1\alpha_2 + \epsilon_h$, $\epsilon_h > 0$ for $h \in (0, 1]$, then we have precisely

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) = \sqrt{2\epsilon_h}. \qquad \square$$

Remark 21.2. In the proof of Theorem 21.1, we have only used the continuity of f_h and both relations (21.2) and (21.5). In fact with the assumption (21.3), we also have

$$\lim_{h\to 0} \sup_{(\alpha_1,\alpha_2)\in B(0,\delta)} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\| = 0. \tag{21.9}$$

Then we will derive in the following the estimate

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C\big(\sqrt{f_h(0,0)} + \sup_{(\alpha_1,\alpha_2)\in B(0,\delta)} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\|\big). \tag{21.10}$$

It is sometimes possible to improve the error analysis by using the estimate (21.10) rather than the (21.6) one.

To check the estimate (21.10) we proceed like in the proof of Theorem 21.1. Notice first that with the Taylor expansion we get for $(\alpha_1, \alpha_2) \in B(0, \delta)$

$$f(\alpha_1, \alpha_2) - f_h(\alpha_1, \alpha_2) = f(0,0) - f_h(0,0) + \int_0^1 \big(Df(t(\alpha_1, \alpha_2)) - Df_h(t(\alpha_1, \alpha_2))\big)(\alpha_1, \alpha_2)\, dt.$$

Since $f(0,0) = 0$, we deduce that

$$|f(\alpha_1, \alpha_2) - f_h(\alpha_1, \alpha_2)| \leq |f_h(0,0)| + \epsilon_h^1 \|(\alpha_1, \alpha_2)\|,$$

where ϵ_h^1 stands for

$$\epsilon_h^1 = \sup_{(\alpha_1, \alpha_2) \in B(0,\delta)} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\|.$$

As in the proof of Theorem 21.1, we can assume that f is given by $f(\alpha_1, \alpha_2) = \alpha_1 \alpha_2$ and satisfies

$$|f(\alpha_1, \alpha_2) - f_h(\alpha_1, \alpha_2)| \leq |f_h(0,0)| + C\epsilon_h^1(|\alpha_1| + |\alpha_2|).$$

Then we have

$$\rho(\mathscr{S}_{f_h}, \mathscr{S}_f) \leq C(|f_h(0,0)| + \epsilon_h^1). \tag{21.11}$$

To get an estimate for the distance $\rho(\mathscr{S}_f, \mathscr{S}_{f_h})$, we define the two sets

$$C_h^+ = \{(\alpha_1, \alpha_2) \in B(0,\delta); f(\alpha_1, \alpha_2) \geq |f_h(0,0)| + C\epsilon_h^1(|\alpha_1| + |\alpha_2|)\},$$
$$C_h^- = \{(\alpha_1, \alpha_2) \in B(0,\delta); f(\alpha_1, \alpha_2) \leq -\big[|f_h(0,0)| + C\epsilon_h^1(|\alpha_1| + |\alpha_2|)\big]\}.$$

From both assumptions (21.5) and (21.9), we know that the sets C_h^+ and C_h^- are not empty. Using the same method as in the proof of Theorem 21.1, we get

$$\rho(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C\big(\sqrt{|f_h(0,0)|} + \epsilon_h^1\big). \tag{21.12}$$

The two estimates (21.11) and (21.12) imply the (21.10) one. □

Remark 21.3. We still can modify the estimate (21.10) in the following way. We assume (21.2), (21.5), (21.9), and

$$\lim_{h \to 0} \sup_{(\alpha_1, \alpha_2) \in B(0,\delta)} \|D^2 f(\alpha_1, \alpha_2) - D^2 f_h(\alpha_1, \alpha_2)\| = 0. \tag{21.13}$$

Notice that the assumption (21.13) is true if we assume (21.3) with $\ell = 2$. Then the following estimate holds

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C\big(\sqrt{f_h(0,0)} + \sup_{(\alpha_1, \alpha_2) \in \mathscr{S}_f} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\|\big). \tag{21.14}$$

Indeed let (α_1, α_2) be in $B(0,\delta)$ and (α_1^0, α_2^0) be in $\mathscr{S}_f$. From the Taylor expansion we deduce that

$$\begin{aligned} f_h(\alpha_1,\alpha_2) - f(\alpha_1,\alpha_2) = \\ f_h(\alpha_1^0,\alpha_2^0) - f(\alpha_1^0,\alpha_2^0) + \big(Df_h(\alpha_1^0,\alpha_2^0) - Df(\alpha_1^0,\alpha_2^0)\big)(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0) \\ + \frac{1}{2}\big(D^2 f_h(\tilde\alpha_1,\tilde\alpha_2) - D^2 f(\tilde\alpha_1,\tilde\alpha_2)\big)\big((\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0),(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0)\big) \end{aligned}$$

where $(\tilde\alpha_1, \tilde\alpha_2)$ belongs to the segment passing by (α_1, α_2) and (α_1^0, α_2^0). So

$$\begin{aligned} |f_h(\alpha_1,\alpha_2) - f(\alpha_1,\alpha_2)| \le |f_h(\alpha_1^0,\alpha_2^0) - f(\alpha_1^0,\alpha_2^0)| \\ + \epsilon_h^1(\mathscr{S}_f)\|(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0)\| + \frac{1}{2}\epsilon_h^2\|(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0)\|^2, \end{aligned}$$

where

$$\begin{aligned} \epsilon_h^1(\mathscr{S}_f) &= \sup_{(\alpha_1,\alpha_2)\in\mathscr{S}_f} \|Df(\alpha_1,\alpha_2) - Df_h(\alpha_1,\alpha_2)\| \\ \epsilon_h^2 &= \sup_{(\alpha_1,\alpha_2)\in B(0,\delta)} \|D^2 f(\alpha_1,\alpha_2) - D^2 f_h(\alpha_1,\alpha_2)\|. \end{aligned}$$

Noticing that

$$\begin{aligned} |f_h(\alpha_1^0,\alpha_2^0) - f(\alpha_1^0,\alpha_2^0)| &\le |f_h(0,0) - f(0,0)| + C\epsilon_h^1(\mathscr{S}_f)\|(\alpha_1^0,\alpha_2^0)\| \\ &\le |f_h(0,0)| + C\epsilon_h^1(\mathscr{S}_f)\|(\alpha_1^0,\alpha_2^0)\|, \end{aligned}$$

we get

$$\begin{aligned} |f_h(\alpha_1,\alpha_2) - f(\alpha_1,\alpha_2)| \le |f_h(0,0)| \\ + C\epsilon_h^1(\mathscr{S}_f)\big(\|(\alpha_1^0,\alpha_2^0)\| + \|(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0)\|\big) \\ + \frac{1}{2}\epsilon_h^2\|(\alpha_1-\alpha_1^0,\alpha_2-\alpha_2^0)\|^2 \end{aligned}$$

We follow then the arguments in Remark 21.2 by reducing $f(\alpha_1, \alpha_2)$ to $\alpha_1\alpha_2$, to check (21.14). □

The next result describes precisely the nature of the solution set $\mathscr{S}_{f_h}$.

Theorem 21.2. *Let us assume that (21.2) holds and in addition that*

$$\lim_{h\to 0} Df_h(0,0) = 0, \tag{21.15}$$

$$\lim_{h\to 0} D^2 f_h(0,0) = D^2 f(0,0), \tag{21.16}$$

and there exists a constant L such that for all $h \in (0, h_0]$ and $(\alpha_1, \alpha_2) \in B(0,\delta)$

$$\|D^2 f_h(\alpha_1,\alpha_2) - D^2 f_h(0,0)\| \le L\|(\alpha_1,\alpha_2)\|. \tag{21.17}$$

Then there exist two positive constants C and h_0, a neighborhood $\mathscr{V}$ of $(0,0)$, and for $h \le h_0$ a point $(\alpha_{1h},\alpha_{2h}) \in \mathbb{R}^2$ and a C^1 diffeomorphism $\mathbf{\Phi}_h$ from $\mathscr{V}$ into $\mathbb{R}^2$ such that

$$\|(\alpha_{1h},\alpha_{2h})\| \le C\|Df_h(0,0)\|, \tag{21.18}$$

$$Df_h(\alpha_{1h},\alpha_{2h}) = 0, \tag{21.19}$$

$$\mathbf{\Phi}_h(\alpha_{1h},\alpha_{2h}) = (\alpha_{1h},\alpha_{2h}), \qquad D\mathbf{\Phi}_h(\alpha_{1h},\alpha_{2h}) = I,$$

$$\mathbf{\Phi}_h(\mathscr{S}_{f_h} \cap \mathscr{V}) \text{ is a part of a hyperbola.}$$

Proof. Under the assumptions (21.15), (21.16), and (21.17), we can apply Theorem 6.1 with $F_h = Df_h$ and $L_h = L$ to prove the existence of $(\alpha_{1h},\alpha_{2h}) \in B(0,\delta)$, for $h \in (0,h_0]$, only to restrict h_0 if necessary, such that both relations (21.18) and (21.19) hold. Moreover for $h \le h_0$ we have

$$\det\big(D^2 f_h(\alpha_{1h},\alpha_{2h})\big) < 0. \tag{21.20}$$

We are in a position to apply Theorem 19.1 to the mapping f_h at the point $(\alpha_{1h},\alpha_{2h})$. There exists a C^{p-1} diffeomorphism $(\alpha_1,\alpha_2) \in \mathscr{V} \to \mathbf{\Phi}_h(\alpha_1,\alpha_2) \in \mathbb{R}^2$ such that

$$f_h(\alpha_1,\alpha_2) = f_h(\alpha_{1h},\alpha_{2h}) + \frac{1}{2}\Big(D^2 f_h(\alpha_{1h},\alpha_{2h})\big(\mathbf{\Phi}_h(\alpha_1,\alpha_2) - (\alpha_{1h},\alpha_{2h}), \mathbf{\Phi}_h(\alpha_1,\alpha_2) - (\alpha_{1h},\alpha_{2h})\big)\Big) \tag{21.21}$$

where $\mathscr{V}$ is a neighborhood of $(\alpha_{1h},\alpha_{2h})$, which a priori is depending on h. Moreover $\mathbf{\Phi}_h(\alpha_{1h},\alpha_{2h}) = (\alpha_{1h},\alpha_{2h})$ and $D\mathbf{\Phi}_h(\alpha_{1h},\alpha_{2h}) = I$. Consequently the set $\mathscr{S}_{f_h}$ is diffeomorphic to a part of the hyperbola given in the $\mathbf{\Phi}$-plane by

$$f_h(\alpha_{1h},\alpha_{2h}) + \frac{1}{2}\Big(D^2 f_h(\alpha_{1h},\alpha_{2h})\big(\mathbf{\Phi} - (\alpha_{1h},\alpha_{2h}), \mathbf{\Phi} - (\alpha_{1h},\alpha_{2h})\big)\Big) = 0. \tag{21.22}$$

Going back to the proof of Morse Lemma, see the appendix in Crouzeix and Rappaz [1989], it is not difficult to check that under the assumptions (21.15), (21.16), and (21.17) and with $p = 2$, we can choose $\mathscr{V}$ independently of h and containing the point $(0,0)$, only to restrict h_0 if necessary. □

Remark 21.4. With the relation (21.21), we deduce that solving the bifurcation equation (21.4) is equivalent to solving the equation (21.22). In the $\mathbf{\Phi}$-plane, the equation

$$f_h(\alpha_{1h},\alpha_{2h}) + \frac{1}{2}\big(D^2 f_h(\alpha_{1h},\alpha_{2h})(\mathbf{\Phi} - (\alpha_{1h},\alpha_{2h}), \mathbf{\Phi} - (\alpha_{1h},\alpha_{2h}))\big) = 0$$

is the equation of a hyperbola centered at $(\alpha_{1h},\alpha_{2h})$, since $\det(D^2 f_h(\alpha_{1h},\alpha_{2h})) < 0$. The hyperbola represents the solution set to (21.4) since $(\alpha_{1h},\alpha_{2h})$ is invariant through

$\mathbf{\Phi}_h$ and $\mathbf{\Phi}_h$ is tangent to the identity at $(\alpha_{1h}, \alpha_{2h})$. The point $(\alpha_{1h}, \alpha_{2h})$ is the center of the hyperbola and with both relations (21.18) and (21.15) we know that $(\alpha_{1h}, \alpha_{2h})$ tends to 0. □

The solution set $\mathscr{S}_{f_h}$ is described by the equation (21.22) and we can distinguish two cases.

If the value of $f_h(\alpha_{1h}, \alpha_{2h})$ is not zero which is the common case,

$$f_h(\alpha_{1h}, \alpha_{2h}) \neq 0, \tag{21.23}$$

then $\mathscr{S}_{f_h}$ is diffeomorphic to a part of a nondegenerate hyperbola.

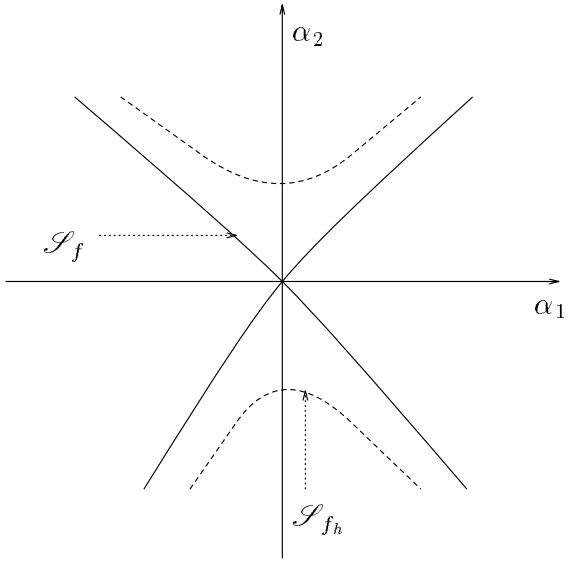

FIGURE 21.2: IMPERFECT NUMERICAL BIFURCATION.

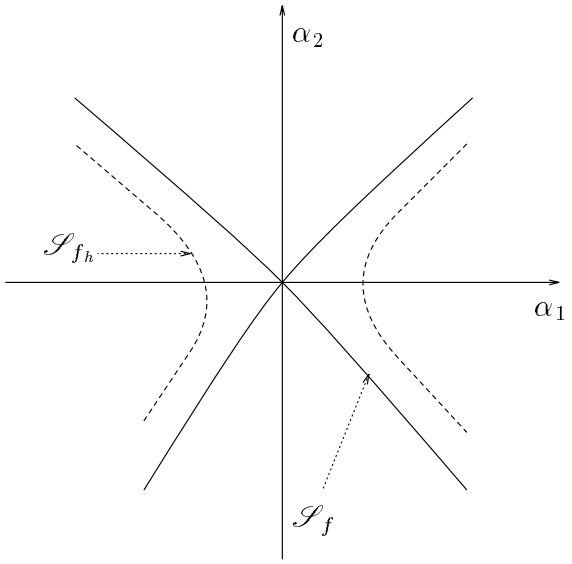

FIGURE 21.3: IMPERFECT NUMERICAL BIFURCATION.

That situation is represented on the figures 21.2 and 21.3, depending on the sign of $f_h(\alpha_{1h}, \alpha_{2h})$. We say that we have an imperfect numerical bifurcation.

The second case which can occur is

$$f_h(\alpha_{1h}, \alpha_{2h}) = 0. \tag{21.24}$$

Then $\mathscr{S}_{f_h}$ is diffeomorphic to a part of a degenerate hyperbola.

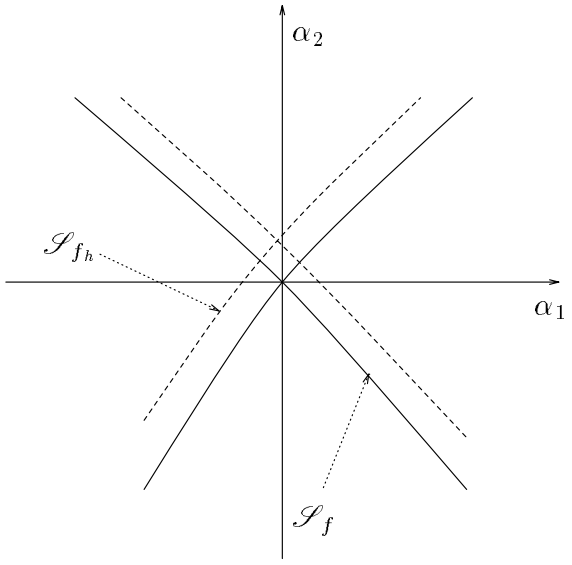

FIGURE 21.4: NUMERICAL BIFURCATION.

That situation is represented on the figure 21.4. We say that we have a numerical bifurcation. The condition (21.24) is satisfied for instance in the case of bifurcation from the trivial branch, when $F(\lambda, 0) = 0$ for all $\lambda \in \mathbb{R}$, see the example developed in the next section. It can be also satisfied when we have symmetries, see SATTINGER [1979].

Remark 21.5. In the framework of Theorem 21.2, we assume furthermore that (21.5), (21.9) hold, and that $f_h(\alpha_{1h}, \alpha_{2h}) = 0$. Here $(\alpha_{1h}, \alpha_{2h})$ is the center of the hyperbola given in Theorem 21.2. Then from the estimate (21.10) we get

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C \sup_{(\alpha_1, \alpha_2) \in B(0,\delta)} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\|. \tag{21.25}$$

Indeed we notice that

$$|f_h(0,0)| \leq C \|(\alpha_{1h}, \alpha_{2h})\|^2$$

since $f_h(\alpha_{1h}, \alpha_{2h}) = 0$ and $Df_h(\alpha_{1h}, \alpha_{2h}) = 0$, see (21.19). From the estimate (21.18) we show

$$\begin{aligned} |f_h(0,0)| &\leq C \|Df(0,0) - Df_h(0,0)\|^2 \\ &\leq C \Big(\sup_{(\alpha_1, \alpha_2) \in B(0,\delta)} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\| \Big)^2. \end{aligned}$$

The estimate (21.25) is then a direct consequence of (21.10). □

Remark 21.6. In the framework of Theorem 21.2, we assume furthermore that (21.5), (21.9), (21.13) hold, and that $f_h(\alpha_{1h}, \alpha_{2h}) = 0$. Like in Remark 21.5, we can derive from the estimate (21.14) the following one

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq C \sup_{(\alpha_1,\alpha_2)\in\mathscr{S}_f} \|Df(\alpha_1, \alpha_2) - Df_h(\alpha_1, \alpha_2)\|. \quad \square \tag{21.26}$$

Remark 21.7. It is possible to generalize the results of Section 21 to the case where the mappings f, f_h are defined in $\mathbb{R}^3$, that is

$$f : B(0, \delta) \subset \mathbb{R}^3 \to \mathbb{R}, \qquad f_h : B(0, \delta) \subset \mathbb{R}^3 \to \mathbb{R}.$$

Then under analogous assumptions, we can prove that f is C^{p-1} diffeomorphic to a part of a cone and that f_h is C^{p-1} diffeomorphic to a part of a hyperboloid with one or two sheets. □

22. A model example

To illustrate the general results in Sections 18 through 21, we will develop a simple model example putting emphasis on the methodology.

Let $\Omega \subset \mathbb{R}^2$ be a convex polygonal domain with the boundary $\partial\Omega$. We shall study the problem to find $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ satisfying

$$-\Delta u - \lambda u + u^2 = 0 \quad \text{in} \quad \Omega \tag{22.1}$$

or in variational form, for all $v \in H_0^1(\Omega)$

$$\int_\Omega \mathbf{grad} u \, \mathbf{grad} v \, dx - \lambda \int_\Omega uv \, dx + \int_\Omega u^2 v \, dx = 0. \tag{22.2}$$

If the mapping $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$ denotes the inverse of the $(-\Delta)$ operator with homogeneous boundary conditions, then problem (22.2) reads: find $\lambda \in \mathbb{R}$ and $u \in H_0^1(\Omega)$ such that

$$u + T(u^2 - \lambda u) = 0, \tag{22.3}$$

see (3.6). Clearly the trivial branch $\{(\lambda, 0); \lambda \in \mathbb{R}\}$ belongs to the solution set of (22.3). Subsequently we look for bifurcation from the trivial branch and study a finite element approximation of the problem.

We are in the framework of Section 19 with the following notations: $\mathscr{X} = H_0^1(\Omega)$, $X = \mathbb{R} \times H_0^1(\Omega)$, $Z = H_0^1(\Omega)$, and the mapping $F : \mathbb{R} \times H_0^1(\Omega) \to H_0^1(\Omega)$ given by, for $(\lambda, v) \in \mathbb{R} \times H_0^1(\Omega)$

$$F(\lambda, v) = v + T(v^2 - \lambda v). \tag{22.4}$$

Clearly the mapping F is of class C^∞. For $\lambda \in \mathbb{R}$, the derivative $DF(\lambda, 0)$ is, for $(\delta, w) \in \mathbb{R} \times H_0^1(\Omega)$

$$DF(\lambda, 0)(\delta, w) = w - \lambda T w. \tag{22.5}$$

We immediately deduce that

$$\mathrm{Ker}(DF(\lambda, 0)) = \mathbb{R} \times \mathrm{Ker}(I - \lambda T) \tag{22.6}$$

with I the identity operator in $H_0^1(\Omega)$ and T the restriction to $H_0^1(\Omega)$ of the above operator $T \in \mathscr{L}(H^{-1}(\Omega); H_0^1(\Omega))$. Let now λ_0 be the first eigenvalue of the eigenproblem: find $\lambda \in \mathbb{R}$, $\xi \in H_0^1(\Omega)$, $\xi \neq 0$ such that

$$-\Delta \xi = \lambda \xi \quad \text{in} \quad \Omega.$$

It is well known that λ_0 is a simple positive eigenvalue and that a corresponding eigenvector ξ can be chosen positive and with a norm equal to 1, that is

$$\xi > 0 \quad \text{in} \quad \Omega \qquad \text{and} \qquad \int_\Omega |\mathbf{grad}\xi|^2 \, dx = 1.$$

Then

$$\mathrm{Ker}(DF(\lambda_0, 0)) = \mathbb{R} \times \mathrm{span}\{\xi\}, \tag{22.7}$$

which is of dimension 2. With the expression (22.5), we get

$$\mathrm{Range}(DF(\lambda_0, 0)) = \mathrm{Range}(I - \lambda_0 T). \tag{22.8}$$

Since the operator $T \in \mathscr{L}(H_0^1(\Omega); H_0^1(\Omega))$ is self adjoint when $H_0^1(\Omega)$ is endowed with the scalar product, for $v_1, v_2 \in H_0^1(\Omega)$

$$(v_1, v_2)_{1,\Omega} = \int_\Omega \mathbf{grad} v_1 \, \mathbf{grad} v_2 \, dx,$$

the range of $DF(\lambda_0, 0)$ is the orthogonal complement of ξ, that is

$$\mathrm{Range}(DF(\lambda_0, 0)) = \{v \in H_0^1(\Omega); (v, \xi)_{1,\Omega} = 0\}, \tag{22.9}$$

which is of codimension 1.

The solution $(\lambda_0, 0)$ satisfies the following hypotheses

$$\begin{cases} \text{(i)} \quad \mathrm{Ker}(D_v F(\lambda_0, 0)) = \mathrm{span}\{\xi\}, \\ \text{(ii)} \quad \text{the range of } D_v F(\lambda_0, 0) \text{ is closed, of codimension 1,} \\ \text{(iii)} \quad D_\lambda F(\lambda_0, 0) = 0. \end{cases} \tag{22.10}$$

Thus $(\lambda_0, 0)$ is a singular solution of (22.3) and the assumptions (19.2) are verified. In our case a decomposition of both spaces $\mathbb{R} \times H_0^1(\Omega)$ and $H_0^1(\Omega)$ can be written explicitly. In (19.4) and (19.5), we choose

$$X_1 = \{0\} \times \mathscr{X}_1 \equiv \{0\} \times \text{Range}(I - \lambda_0 T),$$
$$Z_1 = \text{Ker}(I - \lambda_0 T) = \text{span}\{\xi\}.$$

Then the two projectors $Q_1 \in \mathscr{L}(H_0^1(\Omega); Z_1)$ and $Q_2 \in \mathscr{L}(H_0^1(\Omega); \text{Range}(I - \lambda_0 T))$ are the orthogonal projectors with respect to the scalar product $(.,.)_{1,\Omega}$ in $H_0^1(\Omega)$. The two projectors $P_1 \in \mathscr{L}(\mathbb{R} \times H_0^1(\Omega); X_1)$ and $P_2 \in \mathscr{L}(\mathbb{R} \times H_0^1(\Omega); \text{Ker}(DF(\lambda_0, 0)))$ can be written in the following way

$$P_1(\delta, v) = (0, Q_2 v), \qquad P_2(\delta, v) = (\delta, Q_1 v).$$

There exist positive constants γ, η, and a mapping $g : B(0, \gamma) \subset \mathbb{R}^2 \to \mathscr{X}_1$ of class C^∞ such that for all $(\alpha_1, \alpha_2) \in B(0, \gamma)$

$$\big(Q_2 F(\lambda_0 + \alpha_1, \kappa_1 + \alpha_2 \xi) = 0 \text{ and } \kappa_1 \in B(0, \eta) \subset \mathscr{X}_1\big) \Longleftrightarrow \kappa_1 = g(\alpha_1, \alpha_2).$$

In particuliar for $(\alpha_1, \alpha_2) \in B(0, \gamma)$

$$Q_2 F(\lambda_0 + \alpha_1, g(\alpha_1, \alpha_2) + \alpha_2 \xi) = 0, \tag{22.11}$$

and $g(\alpha_1, 0) = 0$ for $(\alpha_1, 0) \in B(0, \gamma)$.

If the functional $B \in H^{-1}(\Omega)$ satisfies

$$B(\xi) = 1 \quad \text{and} \quad B(z) = 0 \quad \text{for all } z \in \text{Range}(I - \lambda_0 T),$$

then clearly for all $z \in H_0^1(\Omega)$

$$B(z) = (z, \xi)_{1,\Omega}. \tag{22.12}$$

The bifurcation equation reads, for $(\alpha_1, \alpha_2) \in B(0, \gamma)$

$$f(\alpha_1, \alpha_2) \equiv \big(F(\lambda_0 + \alpha_1, g(\alpha_1, \alpha_2) + \alpha_2 \xi), \xi\big)_{1,\Omega} = 0. \tag{22.13}$$

The function f can be written in a simpler form. With the definition (22.4) of F and the relations

$$(g(\alpha_1, \alpha_2), \xi)_{1,\Omega} = 0 = (Tg(\alpha_1, \alpha_2), \xi)_{1,\Omega}$$

we get

$$\begin{aligned} f(\alpha_1, \alpha_2) &= \Big(\xi, \alpha_2 \xi + T\big((\alpha_2 \xi + g(\alpha_1, \alpha_2))^2 - (\lambda_0 + \alpha_1)\alpha_2 \xi\big)\Big)_{1,\Omega} \\ &= \Big(\xi, T\big((\alpha_2 \xi + g(\alpha_1, \alpha_2))^2 - \alpha_1 \alpha_2 \xi\big)\Big)_{1,\Omega} \qquad (22.14) \\ &= \int_\Omega \xi(\alpha_2 \xi + g(\alpha_1, \alpha_2))^2 \, dx - \frac{\alpha_1 \alpha_2}{\lambda_0}. \end{aligned}$$

From the relations (22.11) and (22.14), we check with some simple computations

$$g(0,0) = 0 \qquad \text{and} \qquad f(0,0) = 0, \tag{22.15}$$

$$\frac{\partial g}{\partial \alpha_1}(0,0) = 0, \quad \frac{\partial g}{\partial \alpha_2}(0,0) = 0 \quad \text{and} \quad \frac{\partial f}{\partial \alpha_1}(0,0) = 0, \quad \frac{\partial f}{\partial \alpha_2}(0,0) = 0, \tag{22.16}$$

$$\frac{\partial^2 g}{\partial \alpha_1^2}(0,0) = 0, \quad \frac{\partial^2 g}{\partial \alpha_1 \partial \alpha_2}(0,0) = 0, \quad \frac{\partial^2 g}{\partial \alpha_2^2}(0,0) = -2(I - \lambda_0 T)^{-1} Q_2 T \xi^2, \tag{22.17}$$

$$\frac{\partial^2 f}{\partial \alpha_1^2}(0,0) = 0, \quad \frac{\partial^2 f}{\partial \alpha_1 \partial \alpha_2}(0,0) = -1/\lambda_0, \quad \frac{\partial^2 f}{\partial \alpha_2^2}(0,0) = 2 \int_\Omega \xi^3 \, dx; \tag{22.18}$$

notice that here the operator $I - \lambda_0 T$ is an isomorphism onto $\mathscr{X}_1$. Since

$$\det\big(D^2 f(0,0)\big) = -(\frac{1}{\lambda_0})^2,$$

we are precisely in the case where $(\lambda_0, 0)$ is a simple bifurcation point. Moreover we notice that for all $(\alpha_1, 0) \in B(0, \gamma)$

$$f(\alpha_1, 0) = 0, \tag{22.19}$$

since $g(\alpha_1, 0) = 0$.

Once studied the solution set of (22.3) in a neighborhood of the singular solution $(\lambda_0, 0)$, we turn our attention to finite element approximations of (22.3). For ease of exposition we have assumed that Ω is a convex polygonal domain in $\mathbb{R}^2$, so we can consider a regular family $\{\mathscr{T}_h\}_{0<h\leq 1}$ of triangulations of $\overline{\Omega}$. To each triangulation we associate the finite element subspace

$$V_h = \{v \in C^0(\overline{\Omega}); v_{|T} \in \mathscr{P}_1(T), \text{ for all } T \in \mathscr{T}_h\} \cap H_0^1(\Omega), \tag{22.20}$$

see (3.12). A finite element approximation of (22.2) consists in finding $\lambda \in \mathbb{R}$ and $u_h \in V_h$ such that for all $v_h \in V_h$

$$\int_\Omega \mathbf{grad} u_h \, \mathbf{grad} v_h \, dx - \lambda \int_\Omega u_h v_h \, dx + \int_\Omega u_h^2 v_h \, dx = 0. \tag{22.21}$$

We define the mapping $T_h \in \mathscr{L}(H^{-1}(\Omega); V)$, the discrete analogue of T; given $f \in H^{-1}(\Omega)$, $T_h f \equiv w_h \in V_h$ is the unique solution of

$$\text{for all } v_h \in V_h \quad \int_\Omega \mathbf{grad} w_h \, \mathbf{grad} v_h \, dx = \langle f, v_h \rangle_{H^{-1}(\Omega)\, H_0^1(\Omega)}. \tag{22.22}$$

Then problem (22.21) can be written in the equivalent way: find $\lambda \in \mathbb{R}$ and $u_h \in H_0^1(\Omega)$ such that

$$u_h + T_h(u_h^2 - \lambda u_h) = 0. \tag{22.23}$$

We notice that if $u_h \in H_0^1(\Omega)$ satisfies (22.23), then u_h belongs to the range of T_h, which is precisely V_h, so u_h is a solution to (22.21). The discrete analogue of the mapping F in (22.4) is $F_h : \mathbb{R} \times H_0^1(\Omega) \to H_0^1(\Omega)$ defined by, for $(\lambda, v) \in \mathbb{R} \times H_0^1(\Omega)$

$$F_h(\lambda, v) = v + T_h(v^2 - \lambda v). \tag{22.24}$$

To study the approximate problem (22.23), we shall apply the results of Sections 20 and 21, with F defined in (22.4) and F_h in (22.24). We check first the assumption (20.4). For $(\lambda, v) \in \mathbb{R} \times H_0^1(\Omega)$

$$F_h(\lambda, v) - F(\lambda, v) = (T_h - T)(v^2 - \lambda v), \tag{22.25}$$

for $(\delta, w) \in \mathbb{R} \times H_0^1(\Omega)$

$$\big(DF_h(\lambda, v) - DF(\lambda, v)\big)(\delta, w) = -\delta(T_h - T)v + (T_h - T)(2vw - \lambda w), \tag{22.26}$$

and for $(\delta_1, w_1), (\delta_2, w_2) \in \mathbb{R} \times H_0^1(\Omega)$

$$\begin{aligned} &\big(D^2F_h(\lambda, v) - D^2F(\lambda, v)\big)\big((\delta_1, w_1), (\delta_2, w_2)\big) \\ &\qquad = -(T_h - T)(\delta_1 w_2 + \delta_2 w_1) + 2(T_h - T)(w_1 w_2). \end{aligned} \tag{22.27}$$

Moreover for $\ell > 2$

$$D^\ell F(\lambda, v) = D^\ell F_h(\lambda, v) = 0. \tag{22.28}$$

From the classical finite element estimates, for all $f \in L^2(\Omega)$ we have

$$|(T - T_h)f|_{1,\Omega} \le Ch\|f\|_{0,\Omega}, \tag{22.29}$$

where T and T_h are given in (3.4) and (22.22) and so

$$\lim_{h \to 0} \|T - T_h\|_{L^2(\Omega); H_0^1(\Omega)} = 0. \tag{22.30}$$

Then from the formulae (22.25)-(22.28) and the continuity of the embedding $H_0^1(\Omega) \subset L^p(\Omega)$, we get

$$\lim_{h \to 0} \sup_{(\lambda, w) \in B} \|D^\ell F_h(\lambda, w) - D^\ell F(\lambda, w)\| = 0 \tag{22.31}$$

for all $\ell = 0, 1, 2, \ldots$ and $B \subset H_0^1(\Omega)$ bounded. The assumption (20.4) is verified and we can apply the theory developed in Section 20 at the singular solution $(\lambda_0, 0)$.

There exist positive constants $h_0 \le 1$, η, and a C^∞ mapping $g_h : B(0, \gamma) \subset \mathbb{R}^2 \to \mathscr{X}_1$ such that for all $(\alpha_1, \alpha_2) \in B(0, \gamma)$, $0 < h \le h_0$,

$$\big(Q_2 F_h(\lambda_0 + \alpha_1, \kappa_1 + \alpha_2 \xi) = 0 \text{ and } \kappa_1 \in B(0, \eta) \subset \mathscr{X}_1\big) \Longleftrightarrow \kappa_1 = g_h(\alpha_1, \alpha_2).$$

In particuliar for $(\alpha_1, \alpha_2) \in B(0,\gamma)$

$$Q_2 F_h(\lambda_0 + \alpha_1, g_h(\alpha_1,\alpha_2) + \alpha_2\xi) = 0, \tag{22.32}$$

and $g_h(\alpha_1, 0) = 0$ for $(\alpha_1, 0) \in B(0,\gamma)$. Moreover we can derive the following error estimates, see (20.7), (20.8), (12.14), (12.15), (22.25)–(22.29),

$$\sup_{(\alpha_1,\alpha_2)\in B(0,\gamma)} |D^\ell g(\alpha_1,\alpha_2) - D^\ell g_h(\alpha_1,\alpha_2)|_{1,\Omega} \le C_\ell h, \tag{22.33}$$

where the constants C_ℓ, $0 \le \ell < \infty$, are independent of h.

The approximate bifurcation equation reads, for $(\alpha_1, \alpha_2) \in B(0,\gamma)$

$$f_h(\alpha_1,\alpha_2) \equiv \big(F_h(\lambda_0 + \alpha_1, g_h(\alpha_1,\alpha_2) + \alpha_2\xi), \xi\big)_{1,\Omega} = 0. \tag{22.34}$$

We can explicit with the definitions (22.13) of f and (22.34) of f_h, the difference $f - f_h$,

$$\begin{aligned} f(\alpha_1,\alpha_2) - f_h(\alpha_1,\alpha_2) &= \big(T\big(g^2(\alpha_1,\alpha_2) - g_h^2(\alpha_1,\alpha_2) + (2\alpha_2\xi - \alpha_1)(g(\alpha_1,\alpha_2) - g_h(\alpha_1,\alpha_2))\big), \xi\big)_{1,\Omega} \\ &\quad + \big((T - T_h)\big((g_h(\alpha_1,\alpha_2) + \alpha_2\xi)^2 - (\lambda_0 + \alpha_1)(g_h(\alpha_1,\alpha_2) + \alpha_2\xi)\big), \xi\big)_{1,\Omega}. \end{aligned}$$

With the estimates (22.29) and (22.33) we deduce the following

$$\sup_{(\alpha_1,\alpha_2)\in B(0,\gamma)} |D^\ell f(\alpha_1,\alpha_2) - D^\ell f_h(\alpha_1,\alpha_2)| \le C'_\ell h, \tag{22.35}$$

where the constants C'_ℓ, $0 \le \ell < \infty$, are independent of h. Since for all $(\alpha_1, 0) \in B(0,\gamma)$ we have $g_h(\alpha_1, 0) = 0$, it is not difficult to check with the definition of f_h in (22.34) that for all $(\alpha_1, 0) \in B(0,\gamma)$

$$f_h(\alpha_1, 0) = 0, \tag{22.36}$$

which is the analogue of (22.19).

To study the approximate bifurcation equation, we are exactly in the framework presented in Section 21. The functions f, f_h are of class C^∞ defined on $B(0,\gamma) \subset \mathbb{R}^2$ with values in $\mathbb{R}$, see (22.13) and (22.34). We have checked that

$$\det(D^2 f(0,0)) = -(1/\lambda_0)^2 < 0,$$

so $(\lambda_0, 0)$ is a simple bifurcation point. Moreover the difference $D^\ell f - D^\ell f_h$ is estimated in (22.35).

We can apply Theorem 21.2. For h small enough and only to restrict γ, there exists a C^∞ diffeomorphism from $\mathscr{S}_{f_h} = \{(\alpha_1,\alpha_2) \in B(0,\gamma) \subset \mathbb{R}^2; f_h(\alpha_1,\alpha_2) = 0\}$ onto a

part of a hyperbola centred at $(\alpha_{1h}, \alpha_{2h})$. Going back to definition of f_h, we can check that the center of the hyperbola is in fact $(\alpha_{1h}, 0)$ with

$$f_h(\alpha_{1h}, 0) = 0 \quad \text{and} \quad Df_h(\alpha_{1h}, 0) = 0,$$

which means that the hyperbola is degenerate and we are in the case of Figure 21.4.

From the estimates (21.18) and (22.35), we have

$$|\alpha_{1h}| \leq Ch. \tag{22.37}$$

The estimates (21.25) and (22.35) lead to

$$\delta(\mathscr{S}_f, \mathscr{S}_{f_h}) \leq Ch, \tag{22.38}$$

and the estimates (20.11), (22.29), and (22.38) imply

$$\delta(\mathscr{S}, \mathscr{S}_h) \leq Ch, \tag{22.39}$$

where $\mathscr{S}$ and $\mathscr{S}_h$ are the solution sets for F and F_h introduced in Section 20.

Remark 22.1. In fact we can check that there exists a $\xi_h \in V_h$, $\xi_h \neq 0$ such that

$$\text{for all } v_h \in V_h \qquad \int_\Omega \mathbf{grad}\xi_h \, \mathbf{grad}v_h \, dx = (\lambda_0 + \alpha_{1h}) \int_\Omega \xi_h v_h \, dx.$$

It is known from the theory of approximation of eigenproblems, see Babuška and Osborn [1991], that the following estimate holds

$$|\alpha_{1h}| \leq Ch^2.$$

The estimate (22.37) is not optimal. Optimal estimates are derived with a different method in Crouzeix and Rappaz [1989]. □

Remark 22.2. If we take polynomials of degree $k > 1$ on each triangle in the definition (22.20) of V_h, then we cannot improve the estimate (22.33) but the estimate (22.39) is modified into

$$\delta(\mathscr{S}, \mathscr{S}_h) \leq Ch^k. \quad \square$$

23. Bibliographical comments

The brief comments below are intended only to outline the wealthy literature which is related to our work. A lot of additional results could have been included or quoted.

In Sections 18 through 22 we describe the different steps to study approximations of singular solutions. Only the main features of the method are developed and the example of Section 22 is elementary but it illustrates the general theory. An example of a simple bifurcation point not on the trivial branch is studied into details in Caussignac,

DESCLOUX, and RAPPAZ [1987] and an example of bifurcation for a problem on an unbounded domain is presented in DESCLOUX and RAPPAZ [1982], see also CROUZEIX and RAPPAZ [1989].

Bifurcation phenomena often occur in parameter dependent nonlinear problems. In KELLER and ANTMAN [1969] different examples are presented. General results on bifurcation problems can be found in the books of CHOW and HALE [1982], GOLUBITSKY and SCHAEFFER [1985], GOLUBITSKY, STEWART, and SCHAEFFER [1988], IOSS and JOSEPH [1981], NIRENBERG [1974], and SATTINGER [1973] [1979], and in the bibliographies therein.

The approximation of parameter dependent problems has been widely studied in the literature. For a general approach we can mention the works of BREZZI, RAPPAZ, and RAVIART [1980] [1981a] [1981b], CROUZEIX and RAPPAZ [1989], FUJI and YAMAGUTI [1980], KELLER [1977], KIKUCHI [1976], RABIER [1985], RHEINBOLDT [1986] and SIMPSON [1972].

Further developments concerning turning point can be found for instance in GRIEWANK and REDDIEN [1984], KIKUCHI [1979], and PAUMIER [1981], and concerning bifurcation can be found in RAPPAZ and RAUGEL [1981] for bifurcation from multiple eigenvalues, in BERNARDI [1982] and DESCLOUX [1984] for Hopf bifurcation. A generalization in the direction of non differentiable problems was mentionned in Remark 2.4.

When the exact problem and its approximation are covariant with respect to representations of a finite group, see SATTINGER [1979], GOLUBITSKY and SCHAEFFER [1985], GOLUBITSKY, STEWART, and SCHAEFFER [1988], then the study can be sharpened, see CROUZEIX and RAUGEL [1988], DELLNITZ and WERNER [1989], RAUGEL [1984] for instance.

The applications of the general approximation theory are numerous and quite diverse. For instance we can find an account on the analysis of discretization of the Navier-Stokes equations in GIRAULT and RAVIART [1986]. Approximations of the von Kármán equations, see CIARLET and RABIER [1980], can be found in BREZZI, RAPPAZ, and RAVIART [1980] [1981b], KESAVAN [1979] [1980], REINHARDT [1982].

In our work we have developed a general method based on the inverse and the implicit function theorems to study approximations of regular solutions of nonlinear problems and approximations of regular and singular solutions of parametrized nonlinear problems. There are other methods to study nonlinear problems, which are not even mentionned here. For instance the ones based on the maximum principle or on monotonicity are well-suited to variational inequalities, see for instance BARRETT and ELLIOTT [1991], GLOWINSKI [1984], CONRAD and CORTEY-DUMONT [1987], ŽENÍŠEK [1990].

There is a wide variety of algorithms to solve nonlinear approximate problems. Some fundamental results are due to Keller, see KELLER [1970], KEENER and KELLER [1974], and KELLER [1987]. For an overview we refer to the conference proceedings edited by KUEPPER, MITTELMANN, and WEBER [1984], KUEPPER, SEYDEL, and TROGER [1987], MITTELMANN and WEBER [1980]. There is an extensive bibliography on the topic of numerical continuation methods in the book of ALLGOWER and GEORG [1990].

REFERENCES

Adams, R. (1975), *Sobolev spaces*, Academic Press, New York.

Allgower, E.L. and Georg K. (1990), *Numerical continuation methods. An introduction*, Springer Series in Computational Mathematics, Springer-Verlag, Berlin.

Amann, H. (1976), *Fixed point equations and nonlinear eigenvalue problems in ordered Banach space*, SIAM Review **18**, 620–709.

Babuška, I. and Aziz, A.K. (1972), *Survey lectures on the mathematical foundations of the finite element method*, in The mathematical foundations of the finite element method with application to partial differential equations (Aziz, A.K., eds.), Academic Press, New York, pp. 3–359.

Babuška, I. and Osborn, J. (1991), *Eigenvalue problems*, in Handbook of Numerical Analysis, vol. II, North-Holland, Amsterdam, pp. 641–787.

Babuška, I. and Rheinboldt, W.C. (1982), *Computational error estimates and adaptive processes for some nonlinear structural problems*, Comp. Methods Appl. Mech. Eng. **34**, 895–937.

Baranger, J. and El Amri H. (1991), *Estimateur a posteriori d'erreur pour le calcul adaptatif d'écoulements quasi-newtoniens*, M.^{2}A.N. **25**, 31–48.

Barrett, J.W. and Elliott, C.M. (1989), *Finite element approximation of a plasma equilibrium problem*, IMA J. Num. Anal. **9**, 443–464.

Barrett, J.W. and Elliott, C.M. (1991), *Finite element approximation of a free boundary problem arising in the theory of liquid drops and plasma physics*, M.^{2}A.N. **25**, 213–252.

Bernardi, C. (1982), *Approximation of Hopf bifurcation*, Numer. Math. **39**, 14–37.

Bonic, R.A. (1969), *Linear functional analysis*, Notes on mathematics and its applications, Gordon and Breach, New York.

Brezis, H. (1983), *Analyse fonctionnelle. Théorie et applications*, Collection mathématiques appliquées pour la maîtrise, Masson, Paris.

Brezzi, F., Rappaz, J., and Raviart, P.A. (1980), *Finite dimensional approximation of nonlinear problems. Branches of nonsingular solutions*, Numer. Math. **36**, 1–36.

Brezzi, F., Rappaz, J., and Raviart, P.A. (1981a), *Finite dimensional approximation of nonlinear problems. Limit points*, Numer. Math. **37**, 1–28.

Brezzi, F., Rappaz, J., and Raviart, P.A. (1981b), *Finite dimensional approximation of nonlinear problems. Simple bifurcation points*, Numer. Math. **38**, 1–30.

Caloz, G. (1987), *Simulation numérique des équilibres d'un plasma dans un tokamak: modélisation et études mathématiques*, Thèse **650**, École Polytechnique Fédérale de Lausanne.

Caloz, G. (1991), *Approximation by finite element method of the model plasma problem*, M.^{2}A.N. **25**, 213–252.

Caloz, G. (1994), *Stability on a regular branch of solutions*, Tech. Rpt. 94, I.R.M.A.R., Université de Rennes I.

Cartan, H. (1967), *Cours de calcul différentiel*, Collection méthodes, Hermann, Paris.

Caussignac, Ph., Descloux, J., and Rappaz, J. (1987), *Study of an elliptic problem with nonlinear boundary conditions*, Math. Met. in the Appl. Sci. **9**, 261–275.

Chow, S.N. and Hale, J. (1982), *Methods of bifurcation theory*, Grundlehren 251, Springer-Verlag, Berlin.

Ciarlet, P.G. (1978), *The finite element method for elliptic problems*, Studies in mathematics and its applications, North-Holland, Amsterdam.

Ciarlet, P.G. (1991), *Basic error estimates for elliptic problems*, in Handbook of Numerical Analysis, vol. II, North-Holland, Amsterdam, pp. 17–351.

Ciarlet, P.G. and Rabier, P. (1980), *Les équations de von Kármán*, Lecture notes in mathematics, 826, Springer-Verlag, Berlin.

Clément, P. (1975), *Approximation by finite element functions using local regularization*, R.A.I.R.O. Analyse Numérique **2**, 77–84.

Conrad, F. and Cortey-Dumont, P. (1987), *Nonlinear eigenvalue problems in elliptic variational inequalities: some results for the maximal branch*, Numer. Funct. Anal. and Optimiz. **9**, 1059–1114.

Crandall, M.G. and Rabinowitz, P.H. (1973), *Bifurcation, perturbation of simple eigenvalues and linearized stability*, Arch. Rational Mech. Anal. **52**, 161–180.

Crandall, M.G. and Rabinowitz, P.H. (1975), *Some continuation and variational methods for positive solutions of nonlinear eigenvalue problems*, Arch. Rational Mech. Anal. **58**, 241–269.

Crouzeix, M. and Rappaz, J. (1989), *On numerical approximation in bifurcation theory*, RMA 13, Masson, Paris.

Crouzeix, M. and Raugel, G. (1988), *Invariance under the dihedral group and application to bifurcation problems*, Nonlinear Analysis **12**, 75–99.

Dauge, M. (1992), *Neumann and mixed problems on curvilinear polyhedra*, Intregr. Equat. Op. Th. **15**, 227–261.

Dautray, R. and Lions, J.L. (1987), *Analyse mathématique et calcul numérique pour les sciences et les techniques*, Masson, Paris.

Dellnitz, M. and Werner, B. (1989), *Computational methods for bifurcation problems with symmetries*, J. Comput. Appl. Math. **26**, 97–123.

Descloux, J. (1984), *On Hopf and subharmonic bifurcations*, in Numerical methods for bifurcation problems, Birkhauser-Verlag, Basel, pp. 145–161.

Descloux, J. and Rappaz, J. (1982), *Approximation of solution branches of nonlinear equations*, R.A.I.R.O. Analyse Numérique **16**, 319–349.

Dieudonné, J. (1968), *Éléments d'analyse*, Tome I, Gauthier-Villars, Paris.

Dunford, N. and Schwartz, J.T. (1958), *Linear operators*, Interscience, New York.

Fink, J.P. and Rheinboldt, W.C. (1983a), *On the discretization error of parametrized nonlinear equations*, SIAM J. Numer. Anal. **20**, 732–746.

Fink, J.P. and Rheinboldt, W.C. (1983b), *Solution manifolds and submanifolds of parametrized equations and their discretization errors*, Tech. Rpt. ICMA-83-59, Institute for computational mathematics and applications, University of Pittsburgh (1984), Numer. Math. **45**, 323–343.

Fink, J.P. and Rheinboldt, W.C. (1986), *Folds on the solution manifold of a parametrized equation*, SIAM J. Numer. Anal. **23**, 693–706.

Fink, J.P. and Rheinboldt, W.C. (1987), *A geometric framework for the numerical study of singular points*, SIAM J. Numer. Anal. **24**, 618–633.

Fuji, H. and Yamaguti, M. (1980), *Structure of singularities and its numerical realization in nonlinear elasticity*, J. Math. Kyoto Univ. **20**, 489–590.

Gilbarg, D. and Trudinger, N.S. (1977), *Elliptic partial differential equations of second order*, Springer-Verlag, Berlin.

Girault, V. and Raviart, P.A. (1982), *An analysis of upwind schemes for the Navier-Stokes equations*, SIAM J. Numer. Anal. **19**, 312–333.

Girault, V. and Raviart, P.A. (1986), *Finite element methods for Navier-Stokes equations*, Springer-Verlag, Berlin.

Golubitsky, M. and Schaeffer, D.G. (1985), *Singularities and groups in bifurcation theory*, Volume 1, Springer-Verlag, Berlin.

Golubitsky, M., Stewart, I., and Schaeffer, D.G. (1988), *Singularities and groups in bifurcation theory*, Volume 2, Springer-Verlag, Berlin.

Griewank, A. and Reddien, G.W. (1984), *Characterization and computation of generalized turning points*, SIAM J. Numer. Anal. **21**, 176–185.

Grisvard, P. (1985), *Elliptic problems in nonsmooth domains*, Pitman, Boston.

Ioss, G. and Joseph, D.D. (1981), *Elementary stability and bifurcation theory*, Springer-Verlag, Berlin.

Joseph, D.D. (1965), *Non-linear heat generation and stability of the temperature distribution in conduction solids*, Int. J. Heat Mass Transfer **8**, 281–288.

Keener, J.P. and Keller, H.B. (1974), *Perturbed bifurcation theory*, Arch. Rational Mech. Anal. **50**, 159–175.

Keller, H.B. (1970), *Nonlinear bifurcation*, J. Diff. Equ. **7**, 417–434.

Keller, H.B. (1977), *Numerical solution of bifurcation and nonlinear eigenvalue problems*, in Application of bifurcation theory (P.H. Rabinowitz, eds.), Academic Press, pp. 359–384.

Keller, H.B. (1987), *Lectures on numerical methods in bifurcation problems*, Springer-Verlag, Berlin.

Keller, H.B. and Antman, S., Eds (1969), *Bifurcation theory and nonlinear eigenvalue problems*, W.A. Benjamin, Inc., New York.

Keller, H.B. and Cohen, D.S. (1967), *Some positone problems suggested by nonlinear heat generation*, J. Math. Mech. **16**, 1361–1376.

Kesavan, S (1979), *La méthode de Kikuchi appliquée aux équations de von Kármán*, Numer. Math. **32**, 209–232.

Kesavan, S. (1980), *Une méthode d'éléments finis mixtes pour les équations de von Kármán*, R.A.I.R.O. Analyse Numérique **14**, 149–173.

Kikuchi, F. (1976), *An iterative finite element scheme for bifurcation analysis of semilinear elliptic equations*, Report **542**, Inst. Space Aero. Sc., Tokyo University.

Kikuchi, F. (1979), *Finite element approximations to bifurcation problems of turning point type*, Theoretical and Applied Mechanics **27**, 99–144.

Kuepper, T., Mittelmann, H.D., and Weber, H., Eds (1984), *Numerical methods for bifurcation problems*, ISNM, 70, Birkhauser-Verlag, Basel.

Kuepper, T., Seydel, R., and Troger, H., Eds (1987), *Bifurcation: analysis, algorithms, applications*, ISNM, 79, Birkhauser-Verlag, Basel.

Leray, J. and Schauder, J. (1934), *Topologie et équations fonctionnelles*, Ann. Sci. École Norm. Sup. **51**, 45–78.

Lions, J.L. and Magenes, E. (1968), *Problèmes aux limites non homogènes et applications*, Travaux et recherches mathématiques, Dunod, Paris.

Lions, P.L. (1982), *On the existence of positive solutions of semilinear elliptic equations*, SIAM Review **24**, 441–467.

Maz'ya, V.G. (1985), *Sobolev spaces*, Springer, Berlin.

Mignot, F. and Puel, J.P. (1980), *Sur une classe de problèmes nonlinéaires avec non linéarité positive, croissante, convexe*, Comm. P.D.E. **5**, 791–836.

Mittelmann, H.D. and Weber, H., Eds (1980), *Bifurcation problems and their numerical solution*, ISNM, 54, Birkhauser-Verlag, Basel.

Nečas, J. (1967), *Les méthodes directes en théorie des équations elliptiques*, Masson, Paris.

Nijenhuis, A. (1974), *Strong derivatives and inverse mappings*, Amer. Math. Monthly **81**, 969–980.

Nirenberg, L. (1974), *Topics in nonlinear functional analysis*, Courant Institute of Mathematical Science, New York University, New York.

Paumier, J.C. (1981), *Une méthode numérique pour le calcul de points de retournement. Application à un problème aux limites non-linéaires*, Numer. Math. **37**, 433–452.

Picasso, M. (1992), *Simulation numérique des traitements de surface par laser*, Thèse **1011**, École Polytechnique Fédérale de Lausanne.

Pousin, J. and Rappaz, J. (1991), *Consistance, stabilité, erreurs à priori et à posteriori pour des problèmes non linéaires*, C.R. Acad. Sc. Paris, Série I **312**, 699–703.

Pousin, J. and Rappaz, J. (1992), *Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems*, Tech. Rpt. 04-92, D.M.A., École Polytechnique Fédérale de Lausanne; to appear in Numer. Math..

Rabier, P. (1985), *Topics in one-parameter nonlinear problems*, Tata Institute Lecture Notes, Springer-Verlag, Berlin.

Rannacher, R. and Scott, R. (1982), *Some optimal error estimates for piecewise linear finite element approximations*, Math. Comp. **38**, 437–445.

Rappaz, J. (1983), *Estimations d'erreur dans différentes normes pour l'approximation de problèmes de bifurcation*, C.R. Acad. Sc. Paris, Série I **296**, 179–182.

Rappaz, J. (1984), *Approximation of a nondifferentiable nonlinear problem related to MHD equilibria*, Numer. Math. **45**, 117–133.

Rappaz, J. and Raugel, G. (1982), *Approximation of double bifurcation points for nonlinear eigenvalue problems*, MAFELAP 1981, J. Whiteman editor, Academic Press, New York.

Raugel, G. (1984), *Approximation numérique de problèmes non linéaires*, Thèse, Université de Rennes I.

Reinhart, L (1982), *On the numerical analysis of the von Kármán equations*, Numer. Math. **39**, 371–404.

Rheinboldt, W.C. (1981), *Numerical analysis of continuation methods for nonlinear structural problems*, Comput. and Structures **13**, 103–116.

Rheinboldt, W.C. (1985), *Error estimates for nonlinear finite element computations*, Comput. and Structures **20**, 91–98.

Rheinboldt, W.C. (1986), *Numerical analysis of parametrized nonlinear equations*, Wiley-Interscience, John Wiley and Sons.

Sattinger, D.H. (1973), *Topics in stability and bifurcation theory*, Lecture notes in mathematics, 309, Springer-Verlag, Berlin.

Sattinger, D.H. (1979), *Group theoretic methods in bifurcation theory*, Lecture notes in mathematics, 762, Springer-Verlag, Berlin.

Simader, C.G. (1972), *On Dirichlet's boundary value problem*, Lecture notes in mathematics, 268, Springer-Verlag, Berlin.

Simpson, R.B. (1972), *Existence and error estimates for solutions of a discrete analog of nonlinear eigenvalue problems*, Math. Comp. **26**, 359–375.

Temam, R. (1977), *Theory and numerical analysis of the Navier-Stokes equations*, North-Holland, Amsterdam.

Verfürth, R. (1993), *A review of a posteriori error estimation and adaptive mesh-refinement techniques*, Tech. Rpt., Institut für Angewandte Mathematik, Universität Zürich.

Wahlbin, L.R. (1978), *Maximum norm error estimates in the finite element method with isoparametric quadratic elements and numerical integration*, R.A.I.R.O. Analyse Numérique **12**, 173–202.

Ženíšek, A. (1990), *Nonlinear elliptic and evolution problems and their finite element approximations*, Computational mathematics and applications, Academic Press Limited, London.