THE UNIVERSITY OF READING

NUMERICAL METHODS FOR THE EMDEN-FOWLER EQUATIONS

by

A.C. Lemos, M.J. Baines & N.K. Nichols

Numerical Analysis Report 9/98

DEPARTMENT OF MATHEMATICS

Numerical Methods for the Emden-Fowler Equations

A. C. Lemos[1] **, M. J. Baines and N. K. Nichols**

Numerical Analysis Report 9/98

Department of Mathematics
The University of Reading
PO Box 220 Whiteknights
Reading Berkshire RG6 6AX
United Kingdom

[1]Supported by the grant PRAXIS XXI/BD/15905/98 from the *Fundação para a Ciência e a Tecnologia* and by the *Instituto Politécnico de Leiria*

Abstract

The general Emden-Fowler equation is a nonlinear, second-order, ordinary differential equation with a singularity at the boundary. Here we study a certain class of Emden-Fowler equations with inhomogeneous boundary conditions. A particular case is the Thomas-Fermi problem for the ionized atom.

Using an iterative procedure, the solution of the original nonlinear problem is reduced to the solution of a sequence of linear boundary value problems converging monotonically to the solution of the original problem. Each iteration is then solved by a finite element method with a linear basis, thus yielding a tridiagonal linear system. Therefore this approach is more efficient than methods previously proposed. A nonuniform grid is chosen in such a way that there are more grid points near the singularity. Numerical results are obtained and compared with known results.

Contents

Notation

H	Hilbert space
H_N	finite dimensional subspace of H
u	real function defined on $\bar{\Omega}$
$u^{(\alpha)}$	multi-index notation for the partial derivative $\frac{\partial^{\lvert\alpha\rvert}u}{\partial x_1^{\alpha_1}\partial x_2^{\alpha_2}\ldots\partial x_n^{\alpha_n}}$, where $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)$ and $\lvert\alpha\rvert = \sum_{i=1}^{n} \alpha_i$
Ω	open set; typically, the interval (0,1)
$\partial\Omega$	boundary of Ω
$\bar{\Omega}$	$\Omega \cup \partial\Omega$
$L_2(\Omega)$	space of real square-integrable functions on Ω
$C^k(\Omega)$	space of real, k times differentiable functions on Ω; $C(\Omega) = C^0(\Omega)$
$(u, v) = \int_\Omega uvd\Omega$	$L_2(\Omega)$ inner product of u and v
$\lVert u\rVert = (u, u)^{1/2}$	$L_2(\Omega)$ norm of u
$a(u, v)$	typical bilinear form
$(u, v)_E = a(u, v)$	energy inner product ($a(u, v)$ symmetric)
$\lVert u\rVert_E = (u, u)_E^{1/2}$	energy norm
$H^k(\Omega)$	Sobolev space of real functions defined on Ω with square-integrable generalized derivatives of order $\leq k$
$(u, v)_k = \sum_{\lvert\alpha\rvert\leq k} \int_\Omega u^{(\alpha)}v^{(\alpha)}d\Omega$	$H^k(\Omega)$ (Sobolev) inner product of u and v
$\lVert u\rVert_k = (u, u)_k^{1/2}$	$H^k(\Omega)$ (Sobolev) norm of u
$\lvert u\rvert_k = \left(\sum_{\lvert\alpha\rvert=k} \int_\Omega (u^{(\alpha)})^2 d\Omega\right)^{1/2}$	$H^k(\Omega)$ (Sobolev) seminorm of u
$H_0^k(\Omega)$	the completion with respect to norm $\lVert.\rVert_k$ of the subset of $C^k(\Omega)$ with compact support on Ω
$H_B^k(\Omega)$	subset of $H^k(\Omega)$ defined by $H_B^k(\Omega) = \left\{v \in H^k(\Omega) : \frac{\partial^i v}{\partial x^i} = \Gamma_i \quad \text{on} \quad \partial\Omega,\ i = 0, 1, \ldots, k-1\right\}$

K	stiffness matrix
$\mathbf{G}$	vector
$\mathbf{G}^*$	vector modified by a Dirichlet boundary condition

Chapter 1

Introduction

In many physical phenomena described by partial differential equations involving the Laplacian and Dirichlet boundary conditions, it is possible to reduce the problem to a differential equation with one independent space variable (ODE) or a partial differential equation (PDE) in exactly two variables (one space, one time) using the property of radial symmetry. As pointed out by Mooney [21], in this case the dependency of the solution u on the 'radial' space variable r appears in terms of the form $u'' + \frac{b}{r}u'$. This occurs, for example, in the case of nonlinear reaction-diffusion equations (see [20, 21] and [15] for references to this application) and in other physical applications. However, if we have a boundary condition at $r = 0$, the differential form $u'' + \frac{b}{r}u'$ will have a singularity when $b \neq 0$. In order to analyse this case, we consider the nonlinear ODE

$$u'' + \frac{b}{r}u' + f(r)g(u) = 0$$

where f and g are continuous functions, or the equivalent equation

$$(r^b u')' + r^b f(r)g(u) = 0, \qquad r \geq 0. \tag{1.1}$$

By a Liouville transformation, equation (1.1) can be reduced to the form

$$y''(x) + h(x, y(x)) = 0. \tag{1.2}$$

If f, g are power functions, then h is a product of powers of x and y. The equation (1.1) is then called the *Emden-Fowler equation* and can be written in the form

$$y''(x) = cx^p y^q(x) \qquad (c, \quad \text{constant}). \tag{1.3}$$

More precisely, Emden and Fowler studied the particular form of (1.2),

$$y''(x) + x^p |y|^q sign\ (y) = 0 \tag{1.4}$$

for some constant p. Equation (1.4) is said to be *superlinear* if $q > 1$ and *sublinear* if $q < 1$. Properties of the equation in these two cases differ. Results concerning the existence and uniqueness of the solution of some boundary value problems (BVPs) for the equation (1.3) have been the object of study by several authors. Interesting surveys concerning the historical development of this problem can be found in [26] and in [19]. The former has an excellent bibliography on applications in mechanics, physics and on the study of chemically reacting systems.

We are particularly interested in the problem

$$\frac{d^2y}{dx^2} = x^p y^q, \ x \in (0,1), \tag{1.5}$$

where p, q are real numbers with $-2 < p < 0$ and $q > 1$, with boundary conditions

$$y(0) = 1, \ y(1) = 0. \tag{1.6}$$

Clearly equation (1.5) has a singularity at the endpoint $x = 0$. When $p = -1/2$ and $q = 3/2$, the equation becomes

$$y''(x) = x^{-1/2} y^{3/2}, \tag{1.7}$$

and is called the *Thomas-Fermi equation.* Equation (1.7) was studied independently by Thomas [25] and Fermi [11] and describes, as a statistical model, the completely degenerate state of the electrons in an atom. The related *Thomas-Fermi Theory* is an active subject of research. The three sets of boundary values of physical interest are (cf. [19]):

a) neutral atom

$$y(0) = 1, \ ay'(a) = y(a) \tag{1.8}$$

b) isolated neutral atom

$$y(0) = 1, \lim_{x \to \infty} y(x) = 0 \tag{1.9}$$

c) ionized atom

$$y(0) = 1, \ y(a) = 0. \tag{1.10}$$

In the following chapters we study in more detail the Emden-Fowler equation (1.5) with inhomogeneous boundary conditions (1.6). Our aim is to use the Finite Element Method with a linear basis and a nonuniform grid to solve the problem (1.5),(1.6) and discuss the results obtained. Firstly, in Chapter 2, a review of the literature is presented. In the two following chapters, Chapters 3 and 4, the methods and theory are developed. In particular, in Chapter 3 we present the work of Mooney (see [22], [20] and [21]) on the use of a Picard method and a Newton method applied to (1.5),(1.6). In Chapter 4 we apply the Finite Element Method with a nonuniform grid combined with a Picard iterative method. In Chapter 5 the sources of error are discussed and some known results concerning the error are presented. In the last chapter, Chapter 6, some numerical results are presented as well as some conclusions.

Chapter 2

A review of the literature

Here we summarise some of the studies related to the Thomas-Fermi equation (1.7) and the more general Emden-Fowler equations (1.5). Our aim is to show how these equations arose and present some essential results, mainly the more recent ones. An interesting survey on these equations is given in [26] and some history is also given in [19]. In the following we give a summary of the history of the subject, which can be found in more detail, for example, in [26] and [19].

The study of the *Emden-Fowler Equation*

$$\frac{d}{dx}\left(x^{\rho}\frac{du}{dx}\right) + x^{\sigma}u^{q} = 0,\ x \geq 0, \tag{2.1}$$

where ρ, σ, q are real numbers with $q > 0$ (see [26]), arose in relation to theories concerning gaseous dynamics in astrophysics, around the turn of the century. When studying stellar structure, researchers were concerned at that time with the equilibrium configuration of the mass of spherical clouds of gas. Assuming that the gaseous cloud is under convective equilibrium (as first proposed by Lord Kelvin in 1862), Lane (1869-70) studied the governing equation

$$\frac{1}{x^2}\frac{d}{dx}\left(x^2\frac{du}{dx}\right) + u^{q} = 0,\ x \geq 0 \tag{2.2}$$

in the cases $q = 1.5$ and $q = 2.5$. This equation is known as the *Lane-Emden equation* (cf. [26]). The study of stellar configurations governed by (2.2) culminated in a treatise by Emden in 1907. The mathematical foundations for the study of this equation and also for the more general equation (2.1) was carried out by Fowler in a series of four papers during 1914-1931. The first serious investigation concerning the *generalized Emden-Fowler equation*

$$\frac{d}{dx}\left(a_0(x)\frac{du}{dx}\right) + a_1(x)u^{q} = 0,\ x \geq 0 \tag{2.3}$$

where $a_0(x)$ is positive and absolutely continuous and $a_1(x)$ is nonnegative for $x \geq 0$, was made by Atkinson (see [26] for references).

The Thomas-Fermi equation, arose with the work of Thomas [25] and Fermi [11]. In 1926 Thomas used Adam's method of numerical integration of the differential equation to obtain approximate solutions to problem (1.7),(1.9), while Fermi, in 1927, used graphical methods. Fermi obtained the approximation for small x of

$$y(x) = 1 - 1.58x + \frac{4}{3}x^{3/2} + \cdots$$

(see [19]). In 1930 Baker [4] improved this result to

$$y(x) = 1 + b_2x + b_3x^{3/2} + + \cdots + b_kx^{k/2} + \cdots \tag{2.4}$$

with $b_2 = -1.588558$. At about the same time, Sommerfeld developed the approximate solution to (1.7),(1.9) given by

$$y(x) = y_1(x)\left(1 + [y_1(x)]^{\lambda_1/3}\right)^{\lambda_2/3}$$

where λ_1,λ_2 are zeros of the polynomial $\lambda^2 + 7\lambda - 6, \lambda_1 > 0 > \lambda_2$ and $y_1(x) = \frac{144}{x^3}$ (see [19]). This approximation is quite accurate for large x but underestimates the solution near the origin.

Since all three boundary value problems, (1.7) with boundary conditions (1.8)-(1.10), have the same boundary condition at zero, much computational use has been made of the series expansion (2.4) where the value of b_2, the slope of y at the origin, falls into three classes:

- $b_2 > -1.588\ldots$ corresponding to (1.8);
- $b_2 = -1.588\ldots$ corresponding to (1.9);
- $b_2 < -1.588\ldots$ corresponding to (1.10).

The numerical value -1.5880710 given by Bender and Orszag in 1978 is correct to seven decimal places (see [23] for references).

Hille (1970) answered questions concerning the convergence of the series (2.4). Ramnath (1970) used a technique known as multiple scales to obtain an approximate solution for (1.7),(1.9)(see [19]). The studies of Reid (1972) and Reid and Depuy (1973) apply to more generalized Emden equations (see references in [26] and [19]).

More recently we have the following studies:

- Csavinszky [9], who used an approximate analytical solution based on a variational principle to study the equation (1.7) with boundary conditions (1.9) and (1.10).
- Wong [26], who presented a survey on the generalized Emden-Fowler equation which includes an excellent bibliography on its applications.
- Luning and Perry [19], who transformed (1.7) with boundary conditions (1.10) into an eigenvalue problem and then derived an iterative scheme based on eigenpairs of linear self-adjoint integral operators of Hilbert-Schmidt type, which is shown to converge to a solution. This iteration can be used to obtain a uniform approximation to the solution of problem (1.7) with boundary conditions (1.9).
- Mooney [22, 20, 21], who studied problem (1.5),(1.6) using two iterative schemes based on the Picard and Newton algorithms previously used by him to study problem (1.7),(1.10). These iterative schemes, which were shown to converge monotonically, were solved using a central finite difference method and an extrapolation algorithm. The Newton scheme was shown to converge faster than the Picard scheme which has only first-order convergence.
- Anderson and Arthurs [2] and Burrows and Core [6], who presented a variational approach, based on the theory of variational principles, to solve problem (1.7),(1.9) using different choices of trial functions.
- Chan and Du [7] and Chan and Hon [8], who derived analytical solutions to each iteration of the Picard and Newton schemes applied by Mooney [20] to solve problem (1.7),(1.10), using modified Bessel funtions of the first and second kind, respectively.

- Kwong [15], who studied the uniqueness of a more general boundary value problem that includes the class of Emden-Fowler equations, which is of the form

$$y''(t) + a_1(t)f(y(t)) = 0, \ -\infty < a < t < b < \infty \tag{2.5}$$

with some boundary conditions, where $a_1 : (a,b) \longrightarrow \mathcal{R}$ and $f : \mathcal{R} \longrightarrow \mathcal{R}$ are continuous functions.

- Lima [17], who derived an asymptotic expansion for the Picard iterations (the iterative scheme proposed by Mooney) near the origin for the problem (1.5),(1.6) for certain values of p and q. These results were used to obtain expansions for the error of the approximate solution obtained combining the Picard method and a finite difference scheme. Acceleration of convergence with the E-algorithm of Brezinsky (see [17] for references) was used.

- Lemos and Lima [16, 18], who used the iterative schemes proposed by Mooney [21], to solve problem (1.5),(1.6) but introduced a transformation of the independent variable leading to a new equation with a solution regular in the new variable. The equation obtained at each iteration was solved numerically using both a finite difference method and a finite element method with a cubic B-spline basis.

- Al-Zanaidi, Grossman and Voller [1] who applied a monotone discretization technique based on piecewise simplifications of the nonlinearity to generate enclosures for the problem (1.7),(1.9). This technique had been applied before by Grossmann [13] to the problem (1.7),(1.10).

- Hon [14], who used a decomposition method based on Green's function and Adomian's algorithm to construct a sequence of functions approximating problem (1.7),(1.10).

In the following chapters we will present methods to study problems (1.5),(1.6) and (1.7),(1.10), which are based on [20, 21, 18, 16] and which consist in applying the finite element method to solve each linear boundary value problem corresponding to each iteration of a Picard scheme.

Chapter 3

Using Picard and Newton schemes to solve the Emden-Fowler equations

In this chapter we show how to apply two iterative procedures, the Picard method and the Newton method, to solve problem (1.5),(1.6). Thus, the solution of the nonlinear boundary value problem (1.5),(1.6) is reduced to the solution of a sequence of linear boundary value problems.

Firstly, in Section 3.1, the work of Mooney [20, 21] is summarised, defining Picard iterates and Newton iterates for problem (1.5),(1.6). Secondly, in Section 3.2, a transformation of variables is applied to problem (1.5),(1.6) and the resulting transformed Picard and Newton iterative methods are presented.

3.1 Two iterative procedures: Picard and Newton schemes

Our aim is to present two iterative schemes developed by Mooney [20, 21] to solve problem (1.5),(1.6), which correspond to modifications of the Picard and Newton iterative methods, as well as some existence and uniqueness results. Under certain assumptions on the nonlinear term, Mooney has shown that the Picard and the Newton iterates converge monotonically to the unique solution of problem (1.5),(1.6). The case of homogeneous boundary conditions will be analysed firstly and then we show how to obtain similar iterative schemes for the case of inhomogeneous boundary conditions.

In [20] Mooney applies to the Thomas-Fermi problem (1.7),(1.10) iterative procedures for general boundary value problems developed previously in [22]. Similar iterative procedures were applied later by Mooney [21] to the Emden-Fowler problem (1.5),(1.6).

Firstly, Mooney showed that problem (1.5),(1.6) can be transformed into the form

$$\begin{aligned} \mathcal{L}u(x) &= f(x, u(x)), \qquad x \in (0,1) \\ u(0) &= u(1) = 0 \end{aligned} \tag{3.1}$$

where $\mathcal{L}$ is a second-order self-adjoint operator defined by

$$\mathcal{L}u(x) = -\frac{d}{dx}\left(a_1(x)\frac{du}{dx}\right) + a_0(x)u(x), \tag{3.2}$$

with $a_1(x) \in C^1(0,1)$ and $a_0(x) \in C^0(0,1)$, $a_0(x) \geq 0$, for all $x \in (0,1)$. It is assumed that the nonlinear term f satisfies the following conditions:

(i)

$$f(x,\phi) \in C^1(D) \quad \text{where} \qquad (3.3)$$
$$D = \{(x,\phi(x)) : x \in (0,1), \phi(x) \in C^2(0,1) \quad \text{and} \quad \phi(x) \geq 0, \forall x \in (0,1)\}$$

(ii)

$$f(x,0) > 0, \forall x \in (0,1) \qquad (3.4)$$

(iii)

$$\frac{\partial f}{\partial u}(x,u)\Big|_{u=\phi} > 0, \forall x \in (0,1) \quad \text{and} \quad \phi(x) \geq 0 \quad (\textit{monotonocity condition}) \qquad (3.5)$$

(iv)

$$\frac{\partial f}{\partial u}(x,u)\Big|_{u=\phi} > \frac{\partial f}{\partial u}(x,u)\Big|_{u=\psi}, \forall x \in (0,1) \quad \text{and} \quad \phi > \psi \geq 0$$
$$(\textit{convexity condition}). \qquad (3.6)$$

The Picard method corresponds to obtaining a sequence of iterations $\left\{u^{(\nu)}(x)\right\}$, $\nu \geq 0$, defined by

$$\begin{aligned} \mathcal{L}u^{(\nu+1)}(x) &= f(x,u^{(\nu)}(x)), \qquad x \in (0,1) \\ u^{(\nu+1)}(0) &= u^{(\nu+1)}(1) = 0, \end{aligned} \qquad (3.7)$$

$\nu = 0,1,\ldots$, where $u^{(0)}(x)$ is a given function defined on $[0,1]$. A sequence of Newton iterates $\left\{v^{(\nu)}(x)\right\}$, $\nu \geq 0$, is defined by

$$\begin{aligned} \mathcal{L}v^{(\nu+1)}(x) &= f(x,v^{(\nu)}(x)) + (v^{(\nu+1)}(x) - v^{(\nu)}(x)) \left.\frac{\partial f}{\partial u}\right|_{(x,u=v^{(\nu)})}, \qquad x \in (0,1) \\ v^{(\nu+1)}(0) &= v^{(\nu+1)}(1) = 0, \end{aligned} \qquad (3.8)$$

$\nu = 0,1,\ldots$, where $v^{(0)}(x)$ is a given function defined on $[0,1]$. Note that the Newton iterative scheme involves greater differentiability requirements for f than the Picard scheme, but because both schemes were used the conditions (3.3)-(3.6) have been assumed.

In order to establish algorithms for the sequence of linear boundary value problems that can be used to solve the problem (1.5),(1.6), we first transform the problem (1.5),(1.6) to one with homogeneous boundary conditions on the interval $[0,1]$, i.e.

$$\begin{aligned} -u''(x) &= x^p\left[(1-x) - u(x)\right]^q \\ u(0) &= u(1) = 0 \end{aligned} \qquad (3.9)$$

with a solution $u(x)$ given by

$$u(x) = (1-x) - y(x), \; x \in [0,1]. \qquad (3.10)$$

The right-hand side of the equation in this problem does not satisfy the condition (3.5). Hence to permit the application of existence and uniqueness results (see [20, 21]) a term $\lambda x^p u(x)$, $\lambda \geq q > 0$, is added to both sides of (3.9) yielding

$$\begin{aligned} -u''(x) + \lambda x^p u(x) &= x^p\left[(1-x) - u(x)\right]^q + \lambda x^p u(x) \\ u(0) &= u(1) = 0 \end{aligned} \qquad (3.11)$$

with a solution $u(x)$ given by (3.10).

In this way we obtain two iterative procedures for solving the modified equation (3.11):

a) Picard

$$-\frac{d^2u^{(\nu+1)}}{dx^2}+\lambda x^p u^{(\nu+1)}(x) = x^p\left[(1-x)-u^{(\nu)}(x)\right]^q+\lambda u^{(\nu)}(x)$$
$$u^{(\nu+1)}(0) = u^{(\nu+1)}(1) = 0 \qquad (3.12)$$

with $\nu = 0, 1, \ldots$, which converges monotonically upwards from $u^{(0)}(x) = 0$ and downwards from $u^{(0)}(x) = 1 - x$ to the solution of the transformed general Emden problem (3.11);

b) Newton

$$-\frac{d^2v^{(\nu+1)}}{dx^2}+qx^p\left[(1-x)-v^{(\nu)}(x)\right]^{q-1}v^{(\nu+1)}(x) =$$
$$= x^p\left[(1-x)-v^{(\nu)}(x)\right]^q+qx^p\left[(1-x)-v^{(\nu)}(x)\right]^{q-1}v^{(\nu)}(x)$$
$$v^{(\nu+1)}(0) = v^{(\nu+1)}(1) = 0 \qquad (3.13)$$

with $\nu = 0, 1, \ldots$, which converges monotonically upwards from $v^{(0)}(x) = 0$ to the solution of the transformed general Emden problem (3.11).

Mooney [21] showed that the iterative scheme (3.12) has fastest convergence with $\lambda = q$.

Putting $u^{(i)}(x) = (1-x) - y^{(i)}(x)$ in (3.12) and $v^{(i)}(x) = (1-x) - y^{(i)}(x)$ in (3.13) we obtain the corresponding schemes with inhomogeneous boundary conditions converging to the solution to (1.5),(1.6):

a) Picard

$$\frac{d^2y^{(\nu+1)}}{dx^2}-\lambda x^p y^{(\nu+1)}(x) = x^p\left\{\left[y^{(\nu)}(x)\right]^q-\lambda y^{(\nu)}(x)\right\}$$
$$y^{(\nu+1)}(0) = 1,\ y^{(\nu+1)}(1) = 0 \qquad (3.14)$$

for $\nu = 0, \ldots$;

b) Newton

$$\frac{d^2y^{(\nu+1)}}{dx^2}-qx^p[y^{(\nu)}(x)]^{q-1}y^{(\nu+1)}(x) = (1-q)x^p\left[y^{(\nu)}(x)\right]^q$$
$$y^{(\nu+1)}(0) = 1,\ y^{(\nu+1)}(1) = 0 \qquad (3.15)$$

for $\nu = 0, \ldots$;

where $y^{(0)}(x)$ is a given function defined on $[0, 1]$.

In both schemes, Picard and Newton, $y^{(\nu)}(x)$ converges downwards from $y^{(0)}(x) = 1 - x$ to the solution of the problem (1.5),(1.6), but in the case of the Picard scheme, $y^{(\nu)}(x)$ converges also upwards from $y^{(0)}(x) = 0$ to the solution of problem (1.5),(1.6).

We are going to study in more detail problem (3.14).

3.2 Picard and Newton schemes with variable substitution

In this section the problem (1.5),(1.6) is transformed into an equivalent problem by using a transformation of variables. The Picard method and Newton method corresponding to (3.14) and (3.15) are then presented.

In order to overcome the problem of the singularity at $x = 0$, Lima and Lemos [18, 16] transformed problem (1.5),(1.6) into a problem with a solution regular in the new variable.

Suppose that p is a rational number, i.e., $p = -m/r$, where m, r are natural numbers. Since $p > -2$ then $m < 2r$. Introducing in (1.5) the variable substitution

$$x = t^r$$

and multiplying both members of the equation obtained by t^{2r}, we have

$$\begin{aligned} \tfrac{1}{r^2}\left[t^2\tfrac{d^2y}{dt^2} + (1-r)t\tfrac{dy}{dt}\right] &= t^{2r-m}y^q, \quad x \in (0,1) \\ y(0) = 1,\ y(1) &= 0. \end{aligned} \tag{3.16}$$

Hence, the Picard scheme (3.14) becomes

$$\begin{aligned} \tfrac{1}{r^2}\left[t^2\tfrac{d^2y^{(\nu+1)}}{dt^2} + (1-r)t\tfrac{dy^{(\nu+1)}}{dt}\right] - \lambda t^{2r-m}y^{(\nu+1)} &= t^{2r-m}[(y^{(\nu)})^q - \lambda y^{(\nu)}],\ x \in (0,1) \\ y^{(\nu+1)}(0) = 1,\ y^{(\nu+1)}(1) &= 0, \end{aligned} \tag{3.17}$$

with $y^{(0)}(t) \equiv 0$ or $y^{(0)}(t) \equiv 1 - t$ and $\nu = 0, 1, \ldots$. Similarly, the Newton scheme (3.15) becomes

$$\begin{aligned} \tfrac{1}{r^2}\left[t^2\tfrac{d^2y^{(\nu+1)}}{dt^2} + (1-r)t\tfrac{dy^{(\nu+1)}}{dt}\right] - qt^{2r-m}[y^{(\nu)}]^{q-1}y^{(\nu+1)} &= (1-q)t^{2r-m}[y^{(\nu)}]^q,\ x \in (0,1) \\ y^{(\nu+1)}(0) = 1,\ y^{(\nu+1)}(1) &= 0, \end{aligned} \tag{3.18}$$

with $y^{(0)}(t) \equiv 1 - t$ and $\nu = 0, 1, \ldots$.

Note that by applying a general transformation of the form $x = t^\beta$, $\beta > 0$ to (1.5) and multiplying both members of the resulting equation by $t^{2\beta}$, we obtain the following transformed problem for general p:

$$\begin{aligned} \tfrac{1}{\beta^2}\left[t^2\tfrac{d^2y}{dt^2} + (1-\beta)t\tfrac{dy}{dt}\right] - t^{(p+2)\beta}y^q &= 0, \quad x \in (0,1) \\ y(0) = 1,\ y(1) &= 0. \end{aligned} \tag{3.19}$$

Chapter 4

A finite element method

Our aim here is to apply a finite element method with a piecewise linear basis function and a nonuniform grid to solve problem (1.5), (1.6).

We can deal with our continuous problem in two different ways. One way is to apply firstly the Picard method, transforming the problem of solving a nonlinear boundary value problem into the problem of solving a sequence of continuous linear boundary value problems, and then to solve each problem by a finite element method. The other way is the reverse, that is, firstly apply the finite element method to our continuous problem, obtaining a discrete nonlinear problem approximating the original one, and then apply a Picard method generating a sequence of discrete linear boundary value problems (this approach was taken in [10], for a similar problem but without the singularity at the boundary). Concerning the former, Mooney proved some results on the convergence of the Picard method related to problem (1.5), (1.6) (see [20, 21]) and Lemos and Lima [16, 18] studied the existence and uniqueness of a weak solution of the problem after a transformation. For the latter, some results have been proved concerning convergence of both methods combined (e.g., see [10]) when the partial derivative with respect to the dependent variable in the nonlinear term is bounded. But this is not the case in our problem. Therefore both ways of combining the Picard method and the finite element method need further analysis.

In the following we concentrate on the first way; that is, we apply the finite element method to solve each iteration of the Picard method (see Chapter 3). In order to solve each iteration of the Picard (or Newton) scheme, that is, each linear boundary value problem, we will use the finite element method with piecewise polynomial basis functions. Lima and Lemos [18], dealing with a transformed equation, used the Picard method and solved each iteration by a finite element method using a cubic B-spline basis, since, they argued, the solution of the transformed equation is regular in the new variable. Instead, here we are going to use the finite element method with a *linear* B-spline basis to solve each iteration of the Picard scheme (3.14).

The Lax-Milgram Lemma [3] will be used in Section 4.1 to prove the existence of a generalised solution in the case where the parameter p satisfies $-1 < p < 0$. Firstly, we derive the weak form of the corresponding problem with homogeneous boundary conditions (Section 4.1.1) and show how to apply the Lax-Milgram Lemma in this case (Section 4.1.2) and in the case of inhomogeneous boundary conditions (Section 4.1.3). Then, in Section 4.2, we discuss the application of the finite element method.

4.1 Existence of solution

The problem we are interested in corresponds to solving an iteration of the form

$$\mathcal{L}u^{(\nu+1)}(x) = g^{(\nu)}(x) \tag{4.1}$$

$$u^{(\nu+1)}(0) = 1,\ u^{(\nu+1)}(1) = 0 \tag{4.2}$$

for $\nu = 0, 1, \ldots$, where $\mathcal{L}$ is a second-order linear differential operator, g is a given function and $u^{(\nu+1)}$ is the desired solution of the boundary value problem. (Note that we introduced a slight change in the notation of the solution of the iterative schemes given by (3.14) and (3.15), using $u^{(\nu+1)}$ instead of $y^{(\nu+1)}$). More specifically, the operator $\mathcal{L}$ is given by

$$\mathcal{L}u^{(\nu+1)}(x) = -\frac{d^2u^{(\nu+1)}}{dx^2} + Q(x)u^{(\nu+1)}(x) \tag{4.3}$$

where $Q(x) \in L_1(0,1)$ if $-1 < p < 0$.

The functions Q and g depend on the iterative method chosen (see Chapter 3). Hence, if we use the Picard scheme (3.14), then

$$Q(x) = \lambda x^p \tag{4.4}$$

$$g^{(\nu)}(x) = -x^p[u^{(\nu)}]^q + Q(x)u^{(\nu)}. \tag{4.5}$$

Similarly, if we use the Newton scheme (3.15), then we have

$$Q(x) = qx^p(u^{(\nu)})^{q-1} \tag{4.6}$$

$$g^{(\nu)}(x) = (q-1)x^p(u^{(\nu)})^q. \tag{4.7}$$

Note that the function $g^{(\nu)}(x)$ given by either (4.5) or (4.7) is a function of x and of $u^{(\nu)}(x)$ ($u^{(\nu)}(x)$ is the solution of the linear inhomogeneous BVP corresponding to the previous iteration of the Picard or Newton scheme).

In the following we will consider only the case of the Picard method.

In order to present results concerning the existence of the solution to problem (4.1),(4.2), the Lax-Milgram Lemma is used (Section 4.1.2). Although problem (4.1),(4.2) cannot be formulated directly in terms of the Lax-Milgram Lemma because it does not have homogeneous boundary conditions we will show how to apply the Lemma in this case (Section 4.1.3) by studying firstly the homogeneous boundary conditions case (Section 4.1.1). The assumptions of the Lax-Milgram Lemmma will be verified in Section 4.1.4 but only in the case $-1 < p < 0$.

4.1.1 Homogeneous boundary conditions

We start by studying a problem similar to (4.1),(4.2) but with homogeneous boundary conditions. For simplicity, we will drop the superscript notation of the iterative method returning back to it later (Section 4.1.3).

Consider the problem

$$\mathcal{L}v = g(x) \tag{4.8}$$

$$v(0) = v(1) = 0 \tag{4.9}$$

where $\mathcal{L}$ and g are, respectively, the second-order differential operator and the given function defined in the previous section (Picard method).

In order to obtain the weak form of the problem (4.8)-(4.9) we multiply both sides of the equation (4.8) by a test function $\phi \in H_0^1(0,1)$ and integrate over $I = (0,1)$, obtaining

$$\int_0^1 (-v''\phi + Q(x)v\phi)\, dx = \int_0^1 g(x)\phi dx, \qquad \phi \in H_0^1(0,1). \tag{4.10}$$

Integrating by parts, we have

$$\int_0^1 (v'\phi' + Q(x)v\phi)\, dx = \int_0^1 g(x)\phi dx, \qquad \phi \in H_0^1(0,1), \tag{4.11}$$

where we have used $\phi(0) = \phi(1) = 0$. So, the weak formulation of problem (4.8)-(4.9) is: *find a solution* $v \in H_0^1(0,1)$ *such that*

$$\int_0^1 (v'\phi' + Q(x)v\phi)\, dx = \int_0^1 g(x)\phi dx, \qquad \phi \in H_0^1(0,1). \tag{4.12}$$

In order to see how to apply the Lax-Milgram Lemma (see Section 4.1.2) we introduce a bilinear form $a(.,.)$ and a linear functional $G(.)$, given by, respectively,

$$a(v,\phi) = (\mathcal{L}v, \phi) = \int_0^1 (v'\phi' + Q(x)v\phi)\, dx \tag{4.13}$$

and

$$G(\phi) = (g, \phi) = \int_0^1 g(x)\phi dx. \tag{4.14}$$

Note that the bilinear form (4.13) is symmetric ($a(v,\phi) = a(\phi, v)$, $\forall v, \phi \in H_0^1$).

We seek a solution v satisfying

$$a(v,\phi) = G(\phi), \quad \forall \phi \in H_0^1(0,1), \tag{4.15}$$

where $H_0^1(0,1)$ is the subset of functions in $H^1(0,1)$ satisfying the homogeneous boundary conditions.

In the following sections, we present the Lax-Milgram Lemma and show how to apply it to this case (Section 4.1.2) and to the inhomogeneous boundary conditions case (Section 4.1.3) leaving the proof that the hypotheses of the Lax-Milgram Lemma are satisfied until Section 4.1.4.

4.1.2 Lax-Milgram Lemma

Firstly we recall the Lax-Milgram Lemma (for references see [3]).

Lemma 4.1 (Lax-Milgram) *Let H be a Hilbert space with inner product $(.,.)_H$ and norm*

$$\|u\|_H = (u,u)_H^{1/2}, \qquad u \in H.$$

Suppose that $a : H \times H \to \Re$ is a bilinear form such that:

(i) there exists a constant β such that

$$|a(u,v)| \leq \beta \|u\|_H \, \|v\|_H, \ \forall u, v \in H \qquad (a \text{ is bounded}); \tag{4.16}$$

(ii) there exists a constant ρ such that

$$a(u,u) \geq \rho \|u\|_H^2, \quad \forall u \in H \qquad (a \text{ is coercive}). \tag{4.17}$$

Suppose also that $G : H \to \Re$ is a linear functional such that

(iii) there exists a constant δ such that

$$|G(u)| \leq \delta \|u\|_H, \ \forall u \in H \qquad (G \text{ is bounded}). \tag{4.18}$$

Then, there exists a unique $\hat{v} \in H$ such that

$$a(\hat{v}, v) = G(v), \quad \forall v \in H. \tag{4.19}$$

Furthermore, if $a(.,.)$ is symmetric then the functional

$$f(v) = \frac{1}{2} a(v, v) - G(v), \quad \forall v \in H, \tag{4.20}$$

has a minimum at $\hat{v}$.

Note that when $a(.,.)$ is symmetric, $a(.,.)$ is an inner product on H. It is the so-called *energy inner product.* The corresponding norm is called the *energy norm.* So, we have

$$(u, v)_E \equiv a(u, v), \quad \forall u, v \in H \tag{4.21}$$

$$\|u\|_E \equiv a(u, v)^{1/2}, \quad \forall u \in H. \tag{4.22}$$

Thus, (4.16) and (4.17) imply that the energy norm and the H-norm are equivalent. Moreover, the Lemma asserts the existence of a generalized solution, $\hat{v} \in H$, of the problem (4.19). If the problem has a classical solution, then that solution is $\hat{v}$.

In the homogeneous boundary conditions problem (4.15)) we can apply the Lax-Milgram Lemma immediately (if the hypotheses (4.16)-(4.18) are satisfied), taking

$$H = H_0^1(0, 1) \tag{4.23}$$

and consequently

$$(u, v)_H = (u, v)_1 = \int_0^1 (uv + u'v')dx, \tag{4.24}$$

$$\|u\|_H = \|u\|_1 = (u, u)_1^{1/2}. \tag{4.25}$$

The bilinear form $a(.,.)$ and the functional $G(.)$ are given, respectively, by (4.13) and (4.14).

In the following section we show that problem (4.1),(4.2) can be formulated so that the Lax-Milgram Lemma is applicable.

4.1.3 Inhomogeneous boundary conditions and weak form

In the inhomogeneous boundary condition case we study a problem of the form

$$\mathcal{L}u = g(x) \tag{4.26}$$

$$u(0) = 1, \ u(1) = 0 \tag{4.27}$$

where $\mathcal{L}$ and g are, respectively, the second-order differential operator and the given function defined by (4.3) and (4.7).

Here we seek a solution u belonging to the subset

$$H_B^1(0, 1) = \left\{ u \in H^1(0, 1) : u(0) = 1, u(1) = 0 \right\} \tag{4.28}$$

which is not a linear space (see [3] for more details). Therefore the Lax-Milgram Lemma is not immediately applicable. However, for any fixed $w \in H_B^1(0,1)$ we can write

$$H_B^1(0,1) = \left\{ u \in H^1(0,1) : u = v + w,\ v \in H_0^1(0,1) \right\} \tag{4.29}$$

and define the functional

$$\begin{aligned} f^*(v) &= f(v+w) - f(w) \\ &= \frac{1}{2}a(v+w, v+w) - G(v+w) - \frac{1}{2}a(w,w) + G(w) \\ &= \frac{1}{2}a(v,v) - G^*(v), \quad v \in H_0^1(0,1) \end{aligned} \tag{4.30}$$

where

$$G^*(v) = G(v) - a(w,v) = (g,v) - a(w,v). \tag{4.31}$$

We can now apply the Lax-Milgram Lemma, taking $H_0^1(0,1)$ for the Hilbert space H, $a(.,.)$ for the bilinear form and $G^*(.)$ for the linear functional. Assuming that the Lax-Milgram Lemma hypotheses are satisfied (see Section 4.1.2), there exists a unique $\hat{v} \in H_0^1(0,1)$ such that

$$a(\hat{v}, v) = G^*(v), \quad \forall v \in H_0^1(0,1), \tag{4.32}$$

and hence

$$a(\hat{v} + w, v) = G(v), \quad \forall v \in H_0^1(0,1). \tag{4.33}$$

By the definition of $H_B^1(0,1)$ given in (4.28), it follows that there exists a unique $\hat{u} \in H_B^1(0,1)$ such that $\hat{u} = \hat{v} + w$ and

$$a(\hat{u}, v) = G(v), \quad \forall v \in H_0^1(0,1). \tag{4.34}$$

The function $\hat{v}$ minimizes the functional f^* over $H_0^1(0,1)$, so

$$\min_{u \in H_B^1(0,1)} f(u) = f(\hat{u}). \tag{4.35}$$

As pointed out in [3], $\hat{u}$ is the generalised solution of the boundary value problem with inhomogeneous boundary conditions.

Summing up, the weak form (4.12) can be reformulated for the inhomogeneous case as: *find a solution* $v \in H_0^1(0,1)$ *such that*

$$\int_0^1 (v'\phi' + Q(x)v\phi)\, dx = \int_0^1 \left[g(x)\phi - (w'\phi' + Q(x)w\phi) \right]\, dx \quad \forall \phi \in H_0^1(0,1), \tag{4.36}$$

or equivalently,
find a solution $u = v + w \in H_B^1(0,1)$ *such that*

$$\int_0^1 (u'\phi' + Q(x)u\phi)\, dx = \int_0^1 g(x)\phi\, dx \quad \forall \phi \in H_0^1(0,1). \tag{4.37}$$

Returning to the superscript notation corresponding to the Picard iteration method, the weak formulation of problem (4.1)-(4.2) can be written as:
find a solution $v^{(\nu+1)} \in H_0^1(0,1)$ *such that*

$$\int_0^1 \left(\frac{dv^{(\nu+1)}}{dx}\frac{d\phi}{dx} + Q(x)v^{(\nu+1)}\phi \right) dx = \int_0^1 \left[g^{(\nu)}(x)\phi - (w'\phi' + Q(x)w\phi) \right]\, dx \quad \forall \phi \in H_0^1(0,1), \tag{4.38}$$

where $Q(x)$ and $g^{(\nu)}(x)$ are given, respectively, by (4.4) and (4.5), $Q(x) \in L_1(0,1)$ and w is a particular solution of our problem satisfying inhomogeneous boundary conditions. In the integrand of the right-hand side of (4.38), the term $-(w'\phi' + Q(x)w\phi)$ appears precisely because of the inhomogeneous boundary conditions.

To guarantee the existence of a solution it remains to ensure that the hypotheses of the Lax-Milgram Lemma are satisfied . This is the aim of the next section.

4.1.4 Lax-Milgram assumptions

It can be proved that, in our case, the conditions (i)-(iii) of the Lax-Milgram Lemma are satisfied if $-1 < p < 0$ (see [16]). In this section, for simplicity, we will drop the superscript notation corresponding to the Picard method.

Proof (i). For $u,\ v \in H_0^1(0,1)$ we have

$$\begin{aligned} |a(u,v)| &= \left| \int_0^1 (u'v' + Q(x)uv)\ dx \right| \\ &= \left| \int_0^1 (u'v' + uv + (Q(x)-1)uv)\ dx \right| \\ &\leq \left| \int_0^1 (u'v' + uv)\ dx \right| + \left| \int_0^1 (Q(x)-1)uv\ dx \right| \end{aligned} \tag{4.39}$$

By the Cauchy-Schwartz inequality and if the second integral converges, we have

$$|a(u,v)| \leq \|u\|_1\ \|v\|_1 + \int_0^1 |Q(x)-1|\ |uv|\ dx. \tag{4.40}$$

The function $|Q(x)-1|$ does not change sign in $[0,1]$ so, using a mean value theorem for integrals and

$$\max_{x\in[0,1]} |u(x)|\ \max_{x\in[0,1]} |v(x)| \leq \|u\|_1\ \|v\|_1, \tag{4.41}$$

(see [3]) we have

$$\begin{aligned} |a(u,v)| &\leq \|u\|_1\ \|v\|_1 + \max_{x\in[0,1]} |u(x)|\ \max_{x\in[0,1]} |v(x)| \int_0^1 |Q(x)-1|\ dx \\ &\leq \left(1 + \int_0^1 |Q(x)-1|\ dx\right) \|u\|_1\ \|v\|_1, \end{aligned} \tag{4.42}$$

where the integral

$$\int_0^1 |Q(x)-1|\ dx \tag{4.43}$$

converges if $p > -1$. So there exists $\beta > 0$ such that

$$|a(u,v)| \leq \beta \|u\|_H\ \|v\|_H,\ \forall u, v \in H_0^1(0,1).$$

Proof (ii). We will use the inequality (see [3])

$$\int_0^1 u'^2\ dx\ \geq\ 2\ \int_0^1 u^2\ dx. \tag{4.44}$$

Then, because $Q(x) > 0$ and $Q(x) \in L_1(0,1)$ we have

$$a(u,u)\ =\ \int_0^1 \left(u'^2 + Q(x)u^2\right)\ dx \geq \int_0^1 u'^2 dx$$

$$
\begin{aligned}
&= \frac{1}{2}\int_0^1 u'^2\,dx + \frac{1}{2}\int_0^1 u'^2 dx \\
&\geq \int_0^1 u^2\,dx + \frac{1}{2}\int_0^1 u'^2 dx \\
&\geq \frac{1}{2}\left(\int_0^1 \left(u^2+u'^2\right)\,dx\right) \geq \frac{1}{2}\,\|u\|_1^2. \qquad (4.45)
\end{aligned}
$$

Therefore, there exists a constant ρ such that

$$a(u,u) \geq \rho\|u\|_H^2, \quad \forall\, u \in H_0^1(0,1).$$

Proof (iii). For $v \in H_0^1(0,1)$

$$
\begin{aligned}
|G^*(v)| &= \left|\int_0^1 g(x)v - w'v' - Q(x)wv\,dx\right| \\
&\leq \left|\int_0^1 g(x)v\,dx\right| + \left|\int_0^1 w'v'\,dx\right| + \left|\int_0^1 Q(x)wv\,dx\right| \qquad (4.46)
\end{aligned}
$$

provided that these last integrals converge. The first integral of (4.46) converges if $p > -1$ and because $|g(x)|$ does not change sign in $[0,1]$,

$$\left|\int_0^1 g(x)v\,dx\right| \leq \max_{x\in[0,1]}|v(x)| \int_0^1 |g(x)|\,dx. \qquad (4.47)$$

Choosing the particular solution w such that $w \in C^1(0,1)$ and using the Cauchy-Schwartz inequality, the second integral of (4.46) can be bounded in the form

$$
\begin{aligned}
\left|\int_0^1 w'v'\,dx\right| &\leq \left(\int_0^1 w'^2\,dx\right)^{\frac{1}{2}} \left(\int_0^1 v'^2\,dx\right)^{\frac{1}{2}} \\
&\leq \left(\int_0^1 w'^2\,dx\right)^{\frac{1}{2}} \|v\|_1. \qquad (4.48)
\end{aligned}
$$

Finally, the third integral of (4.46) can be bounded by

$$
\begin{aligned}
\left|\int_0^1 Q(x)wv\,dx\right| &\leq \max_{x\in[0,1]}|w(x)| \max_{x\in[0,1]}|v(x)| \int_0^1 |Q(x)|\,dx \\
&\leq \max_{x\in[0,1]}|w(x)|\,\|v\|_1 \int_0^1 |Q(x)|\,dx, \qquad (4.49)
\end{aligned}
$$

since the integral converges when $p > -1$ and the particular solution w can be chosen such that $w \in C^1(0,1)$. Using (4.47), (4.48) and (4.49) in (4.46) we see that there exists a constant δ such that

$$|G^*(v)| \leq \delta\|v\|_H, \ \forall v \in H_0^1(0,1). \qquad (4.50)$$

So, in the case $-1 < p < 0$, there is a coercive and bounded bilinear form $a(.,.)$ given by (4.13) and a bounded linear functional $G^*(.)$ given by (4.31), both defined on a Hilbert space, such that the solution of the boundary value problem is the same as the element $\hat{v} \in H$ whose existence is asserted by the Lax-Milgram Lemma. That is, $\hat{v}$ satisfies the condition

$$a(\hat{v},v) = G(v), \qquad \forall v \in H.$$

Because in our case $a(.,.)$ is symmetric, it follows that $\hat{v} \in H$ minimizes the functional

$$f(v) = \frac{1}{2}a(v,v) - G(v), \qquad \forall v \in H.$$

It was not proved that the conditions (4.16) and (4.18) of the Lax-Milgram Lemma hold in the case $-2 < p \leq -1$ because a integral of the form

$$\int_0^1 |Q(x) - c| \, dx, \tag{4.51}$$

where c is a constant and $Q(x)$ is given by (4.4), does not converge. Nevertheless, the integral we need to compute, i.e.,

$$\int_0^1 (u'v' + Q(x)uv) \, dx \tag{4.52}$$

where $Q(x) = \lambda x^p$ and u, v are the basis functions used (linear B-splines), might converge for that choice of the parameter p. Hence, we present in Chapter 6 some numerical results concerning case $-2 < p \leq -1$.

Next we show how to apply the finite element method.

4.2 Finite Element Method

Under the assumptions of the Lax-Milgram Lemma, we can obtain an approximate solution of the problem (4.15) if we introduce a finite dimensional subspace of H, say H_N (N-dimensional), and restrict our problem to this space (which is still a Hilbert space). Then we can transform the original problem into one that corresponds to solving an algebraic system of N linear equations. Since $H_N \subset H$ is a Hilbert space, in H_N the hypotheses (i)-(iii) of Lax-Milgram Lemma are satisfied and we have the following theorem.

Theorem 4.1 *Suppose that hypotheses (i)-(iii) of the Lax-Milgram Lemma are satisfied and that H_N is an N-dimensional subspace of H. Then there exists a unique $\hat{v}_N \in H_N$ such that*

$$a(\hat{v}_N, v) = G(v), \quad \forall v \in H_N. \tag{4.53}$$

Furthermore, if $a(.,.)$ is symmetric then

$$\min_{v \in H_N} f(v) = f(\hat{v}_N), \tag{4.54}$$

where f is given by (4.20).

This theorem is not constructive in the sense that it does not specify how to obtain u_N. Nevertheless, the finite dimensionality of H_N can be exploited. Since, by the Lax-Milgram Lemma, the existence of a solution in $H = H_0^1(0,1)$ is guaranteed, we can use an N-dimensional subspace of $H_0^1(0,1)$ as the trial and test spaces; in our case we use the space of continuous piecewise polynomials of degree less or equal than 1, i.e., a linear B-spline basis. We use this type of basis function because it leads us to simple banded matrices. In the particular case of the linear B-splines, we obtain tridiagonal matrices. (Note that in [18], where a transformation was applied, the resulting equation is regular in the new variable and hence the basis functions used could be cubic B-splines.)

Let $X_h : 0 \equiv x_0 < x_1 < \cdots < x_n \equiv 1$ be a partition of $I = (0,1)$ into intervals or elements $I_k = (x_{k-1}, x_k)$ of size $h_k = x_k - x_{k-1}$ $(k = 0, \ldots, n)$ and let H_N be the vector space of continuous piecewise linear functions on X_h that vanish at $x = 0$ and $x = 1$. H_N is a space of dimension $n - 1$ $(N = n - 1)$.

Let $\{\psi_k(x)\}_{k=1}^{n-1}$ be a basis of this space H_N, where

$$\psi_k(x) = \begin{cases} \frac{x-x_{k-1}}{x_k-x_{k-1}} & \text{if } x_{k-1} < x < x_k \\ \frac{x-x_{k+1}}{x_k-x_{k+1}} & \text{if } x_k < x < x_{k+1} \\ 0 & \text{otherwise} \end{cases} \tag{4.55}$$

for $k = 1, 2, \ldots, n-1$.

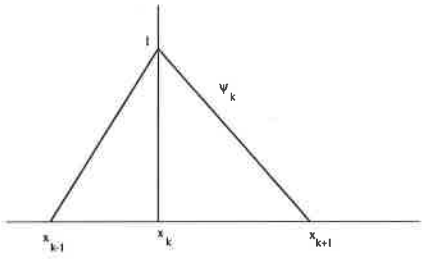

Figure 4.1: The linear B-spline $\psi_k(x)$

It was shown in Section 4.1.3 that the solution of the inhomogeneous boundary value problem is

$$\hat{u} = \hat{v} + w$$

where w is any chosen function of the set $H_B^1(0,1)$ and $\hat{v}$ the minimizer of the functional f^* given by (4.30). Hence we are looking for an approximate solution, $\hat{u}_N$, of the inhomogeneous BVP in the form

$$\hat{u}_N = \hat{v}_N + w$$

where $\hat{v}_N$ belongs to H_N (a N-dimensional subspace of $H_0^1(0,1)$) and satisfies

$$\min_{v_N \in H_N} f^*(v_N) = f^*(\hat{v}_N).$$

More specifically, if $\bar{H}_N$ is a N-dimensional subset of $H_B^1(0,1)$ given by

$$\bar{H}_N = \left\{ \bar{u} \in H_B^1(0,1) : \bar{u} = \hat{v} + w, \hat{v} \in H_N \right\} \tag{4.56}$$

we are looking for an approximate solution, $\hat{u}_N \in \bar{H}_N$ of the form

$$\hat{u}_N^{(\nu+1)} = \hat{v}_N^{(\nu+1)} + \psi_0(x) = \sum_{j=1}^{n-1} \alpha_j^{(\nu+1)} \psi_j(x) + \psi_0(x), \qquad \nu = 0, 1, \ldots \tag{4.57}$$

where $\psi_0(x)$ is a particular solution satisfying inhomogeneous boundary conditions. We let

$$\psi_0(x) = \begin{cases} \frac{x-x_1}{x_0-x_1} & \text{if } x_0 < x < x_1 \\ 0 & \text{otherwise .} \end{cases}$$

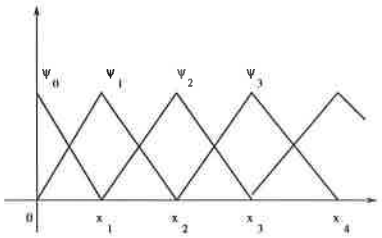

Figure 4.2: Some linear B-splines

Then we choose $v = \psi_i$ as test function and the finite dimensional formulation of (4.11) corresponding to each iteration of the Picard scheme becomes:
find $\hat{u}_N^{(\nu+1)} \in H_B^1(0,1)$ *of the form (4.57) such that*

$$\int_0^1 \left(\hat{u}_N^{(\nu+1)'} \psi_i' + Q(x) \hat{u}_N^{(\nu+1)} \psi_i \right) \, dx - \int_0^1 g(x, \hat{u}_N^{(\nu)}(x)) \psi_i dx \; = \; 0. \tag{4.58}$$

Using (4.57) we may write equation (4.58) in the form

$$\sum_{j=1}^{n-1} \left(\int_{x_{i-1}}^{x_{i+1}} (\psi_i' \psi_j' + Q(x) \psi_i \psi_j) \, dx \right) \alpha_j^{(\nu+1)} = \int_{x_{i-1}}^{x_{i+1}} g(x, \hat{u}_N^{(\nu)}) \psi_i \, dx$$
$$- \int_{x_{i-1}}^{x_{i+1}} (\psi_i' \psi_0' dx \; + \; Q(x) \psi_i \psi_0) \, dx, \tag{4.59}$$

and using the expressions for Q and g corresponding to the Picard scheme given, respectively, by (4.4) and (4.5), we obtain

$$\sum_{j=1}^{n-1} \left(\int_{x_{i-1}}^{x_{i+1}} (\psi_i' \psi_j' + \lambda x^p \psi_i \psi_j) \, dx \right) \alpha_j^{(\nu+1)} = \int_{x_{i-1}}^{x_{i+1}} \lambda x^p \psi_i \left(\sum_{j=1}^{n-1} \alpha_j^{(\nu)} \psi_j \right) dx \; -$$
$$- \int_{x_{i-1}}^{x_{i+1}} \psi_i' \psi_0' dx \; - \int_{x_{i-1}}^{x_{i+1}} \psi_i \; x^p \left(\psi_0 \; + \; \sum_{j=1}^{n-1} \alpha_j^{(\nu)} \psi_j \right)^q dx. \tag{4.60}$$

Hence, computing each iteration of the Picard scheme corresponds to solving the linear system

$$\mathbf{K} \alpha^{(\nu+1)} \; = \; \mathbf{G}^{*(\nu)} \tag{4.61}$$

where

$$K_{ij} = a(\psi_i, \psi_j), \qquad i,j = 1,2,\ldots,n-1, \tag{4.62}$$
$$G_i^{*(\nu)} = G(\psi_i) - a(\psi_i, \psi_0), \qquad i = 1,2,\ldots,n-1 \tag{4.63}$$

and $\alpha^{(\nu+1)}$ is the vector of unknowns. $\mathbf{K}$ is called the *stiffness matrix*. We have

$$K_{ij} = \int_{x_{i-1}}^{x_{i+1}} (\psi_i' \psi_j') dx \; + \; \int_{x_{i-1}}^{x_{i+1}} (\lambda x^p \psi_i \psi_j) \, dx \tag{4.64}$$

for $i,j = 1,2,\ldots,n-1$ and

$$G_i^{*(\nu)} = \int_{x_{i-1}}^{x_{i+1}} \lambda x^p \psi_i \left(\sum_{j=1}^{n-1} \alpha_j^{(\nu)} \psi_j \right) dx \; - \; \int_{x_{i-1}}^{x_{i+1}} \psi_i \; x^p \left(\psi_0 \; + \; \sum_{j=1}^{n-1} \alpha_j^{(\nu)} \psi_j \right)^q dx$$
$$- \int_{x_{i-1}}^{x_{i+1}} \psi_i' \psi_0' dx \; . \tag{4.65}$$

for $i = 1, 2, \ldots, n-1$.

Almost all the integrals in (4.64) and (4.65) are computed directly. The exception is the second integral in (4.65) which is, in general, computed numerically using a sufficiently fine composite trapezoidal rule. This integral is computed directly in the case $p = -1$, $q = 2$ and in the Thomas-Fermi case $p = -\frac{1}{2}$ and $q = \frac{3}{2}$ (with a formula given in [12]). According to the definition of the basis functions in subintervals, most of the integrals have to be split in two. The integrals involving x^p are calculated according to different ranges of the parameter p.

A tridiagonal solver is used to solve the system (4.61). In fact, since we use linear B-splines,

$$K_{ij} = 0 \qquad \text{if } |i-j| > 1, \qquad \text{for } i, j = 1, 2, \ldots, n-1, \tag{4.66}$$

and the matrix $\mathbf{K}$ is tridiagonal. The matrix $\mathbf{K}$ is also symmetric and positive definite. In fact, the symmetry of $a(.,.)$ implies that the matrix $\mathbf{K}$ is symmetric. Moreover, both the symmetry and coercivity properties of the bilinear form $a(.,.)$ imply that the matrix $\mathbf{K}$ in (4.61) is positive definite (see [3]).

Substituting

$$v_N = \sum_{j=1}^{n-1} \alpha_j \psi_j(x) \tag{4.67}$$

into (4.20) and using the bilinearity of $a(.,.)$ and the linearity of $\mathbf{G}^*(.)$, we find that

$$\begin{aligned} f^*(v_N) &= \frac{1}{2} a\left(\sum_{j=1}^{n-1} \alpha_j \psi_j(x), \sum_{j=1}^{n-1} \alpha_j \psi_j(x)\right) - G^*\left(\sum_{j=1}^{n-1} \alpha_j \psi_j(x)\right) \\ &= \frac{1}{2} \sum_{i=1}^{n-1} \sum_{j=1}^{n-1} \alpha_i \alpha_j \, a\left(\psi_i(x), \psi_j(x)\right) - \sum_{j=1}^{n-1} \alpha_j G^*\left(\psi_j(x)\right) \\ &= \frac{1}{2} \alpha^T \mathbf{K} \alpha - \alpha^T \mathbf{G}^* = \mathbf{F}^*(\alpha). \end{aligned} \tag{4.68}$$

Hence

$$\min_{v_N \in H_N} f^*(v_N) = \min_{\alpha \in \Re^N} \mathbf{F}^*(\alpha). \tag{4.69}$$

$\mathbf{F}^*$ is a quadratic functional over $\Re^N$ with a positive definite Hessian matrix and thus $\mathbf{F}^*$ is uniquely minimized by the α that satisfies

$$\mathbf{K}\alpha = \mathbf{G}^* \tag{4.70}$$

(see [3]).

In this chapter we have derived the weak form of the problem (4.1),(4.2) and we have shown the existence and uniqueness of a generalised solution in the case where the parameter p satisfies $-1 < p < 0$. Furthermore, we applied the finite element method to each iteration of a Picard scheme of the form (4.1),(4.2), yielding a linear system which is solved by a direct method.

In the next chapter different types of error, originating from the use of numerical methods, are studied. Also presented is the type of nonuniform grid chosen.

Chapter 5

Error

5.1 Sources of error

We now point out some sources of computational error in the approach, namely:

- the use of the Picard method combined with a finite element method;
- the use of a numerical integration rule;
- the solution of the linear system (tridiagonal solver or band matrix solver).

Using the Picard method combined with a finite element method, the original continuous nonlinear boundary value problem is replaced by a sequence of linear discrete BVPs and a discretization error arises in each iteration of the Picard method (linear BVP) because the solution is approximated by a piecewise polynomial (assuming that the integrals arising in the finite element method are evaluated exactly and that the resulting system of discrete equations is solved exactly).

There is also a numerical integration error that comes from evaluating the integrals arising from the finite element method using a trapezoidal rule and the error resulting from solving the discrete linear system of equations.

Our goal is to obtain a method where the discretization error is controlled within a given tolerance level (*reliability*) and that the computational work to compute a solution within the given tolerance is as small as possible (*efficiency*).

In Section 5.2 the type of grid used is presented and in Section 5.3, some of the errors are studied in more detail, namely, the discretisation error that originates from applying the finite element method (Section 5.3.1), the error that originates from solving the linear system by using a direct method (Section 5.3.2) and the numerical integration error (Section 5.3.3).

5.2 Grid selection

A nonuniform grid X_h (see Section 4.2) is chosen in accordance with the transformation used in previous work (see Section 3.2 and [16, 18] for more details), that is,

$$x_i = \left(\frac{i}{n}\right)^{\gamma}, \ \gamma > 0, \ i = 0, 1, \ldots, n. \tag{5.1}$$

This choice of grid corresponds to a nonuniform grid having more points near the singularity. In this case $h_{max} = \max_{1 \leq i \leq n} h_i = h_n = 1 - x_{n-1}$ and $h_{min} = \min_{1 \leq i \leq n} h_i = h_1 = x_1$.

In the next section different types of error arising when applying the finite element method to solve each iteration of problem (3.14) are studied.

5.3 Error Analysis

By constructing asymptotic expansions of the Picard iterates (3.14) near the singularity, Lima [17] (and also Mooney [21] for certain values of the parameters p and q), developed the corresponding asymptotic expansion for the discretisation error when applying a finite difference method. Lima [17] showed that the error expansion is $\mathcal{O}(h^{2+p})$ if $-2 < p < -1$ or $-1 < p < 0$ and $\mathcal{O}(h(1 + ln\ h))$ if $p = -1$. Although we have applied a different discretisation method (finite element method) we may expect similar results to hold.

5.3.1 Discretization error

Our aim is to show some estimates of the error $e_N = \hat{u} - \hat{u}_N$ corresponding to the case of inhomogeneous boundary conditions. Here, $\hat{u}_N$ is the finite element approximation to the exact solution $\hat{u} \in H$. Firstly we obtain some error estimates for the case of homogeneous boundary conditions.

Now we want to estimate the error $(\hat{v} - \hat{v}_N)$ where $\hat{v}_N \in H_N$ (H_N vector space of continuous piecewise linear functions) is the finite element approximation of the exact solution $\hat{v} \in H$ whose existence and uniqueness is guaranteed by Theorem 4.1.

The functional

$$f(v) = \frac{1}{2}a(v, v) - G(v), \quad v \in H \tag{5.2}$$

can be expressed as

$$f(v) = f(\hat{v}) + \frac{1}{2}a(\hat{v} - v, \hat{v} - v), \quad v \in H \tag{5.3}$$

and because

$$\min_{v_N \in H_N} f(v_N) = f(\hat{v}_N) \tag{5.4}$$

we have

$$[a(\hat{v} - \hat{v}_N, \hat{v} - \hat{v}_N)]^{1/2} = \min_{v_N \in H_N} [a(\hat{v} - v_N, \hat{v} - v_N)]^{1/2} \tag{5.5}$$

or, in the energy norm notation,

$$\|\hat{v} - \hat{v}_N\|_E = \min_{v_N \in H_N} \|\hat{v} - v_N\|_E. \tag{5.6}$$

Therefore, the finite element method minimizes the error $(\hat{v} - \hat{v}_N)$ in the energy norm over the subspace H_N. Moreover,

$$a(\hat{v} - \hat{v}_N, v_N) = 0, \quad \forall v_N \in H_N, \tag{5.7}$$

that is, $\hat{v}_N$ is the orthogonal projection with respect to the energy inner product of $\hat{v}$ onto H_N. Thus

$$a(\hat{v}, \hat{v}_N) = a(\hat{v}_N, \hat{v}_N) \tag{5.8}$$

and then

$$a(\hat{v} - \hat{v}_N, \hat{v} - \hat{v}_N) = a(\hat{v}, \hat{v}) - a(\hat{v}_N, \hat{v}_N), \tag{5.9}$$

that is, the energy in the error equals the error in the energy. Furthermore, since the left side of (5.9) is necessarily nonnegative, we have

$$a(\hat{v}_N, \hat{v}_N) \leq a(\hat{v}, \hat{v}) \tag{5.10}$$

(see [24]).

We now present an estimate of the error in the space $H = H_0^1(0,1)$ (see [3, 24] for more details). Since $a(.,.)$ satisfies (4.16) and (4.17), there exist positive constants β and ρ such that

$$\sqrt{\rho}\,\|v\|_H \leq \sqrt{a(v,v)} \leq \sqrt{\beta}\,\|v\|_H, \quad \forall v \in H, \tag{5.11}$$

that is, $\|.\|_E$ and $\|.\|_H$ are equivalent. Thus

$$\|\hat{v} - \hat{v}_N\|_H \leq C^* \sqrt{a(\hat{v} - \hat{v}_N, \hat{v} - \hat{v}_N)} \leq C \min_{v_N \in H_N} \|\hat{v} - v_N\|_H, \tag{5.12}$$

where C^* and C are constants.

In order to study the error in the inhomogeneous case, we apply (5.12) to the functional f^* given by (4.30), obtaining

$$\|\hat{v} - \hat{v}_N\|_1 \leq C \min_{v_N \in H_N} \|\hat{v} - v_N\|_1. \tag{5.13}$$

Since $\hat{v} = \hat{u} - w$ and $\hat{v}_N = \hat{u}_N - w$, we obtain

$$\|\hat{u} - \hat{u}_N\|_1 \leq C \min_{\bar{u} \in \bar{H}_N} \|\hat{u} - \bar{u}\|_1, \tag{5.14}$$

where

$$\bar{H}_N = \left\{ \bar{u} \in H_B^1(0,1) : \bar{u} = \hat{v} + w, \hat{v} \in H_N \right\}.$$

It is possible to derive other estimates of the error. A useful approximation to $\hat{u}_N$ is the linear interpolant of $\hat{u}$, $\hat{u}_I$, which can be expanded in terms of the linear B-splines $\{\psi_k(x)\}_{k=0}^n$ (see (4.55)) as

$$\hat{u}_I(x) = \sum_{k=0}^{n} \hat{u}(x_k)\psi_k(x), \tag{5.15}$$

where x_k, $k = 0, 1, \ldots, n$, are the grid points. The two approximations, $\hat{u}_N$ and $\hat{u}_I$ agree at every grid point x_k ($k = 0, 1, \ldots, n$) and $\hat{u}_I$ is linear in between. In spite of the interpolate $\hat{u}_I$ and the finite element approximation $\hat{u}_N$ being both piecewise linear, $\hat{u}_N$ is determined variationally while $\hat{u}_I$ is chosen only to be close to u. Hence, instead of studying the discretization error given by $e_N = \hat{u} - \hat{u}_N$ where $\hat{u}_N \in \bar{H}_N$ directly, we study its relation to the interpolation error $e_I = \hat{u} - \hat{u}_I$ because the latter is easier to analyse and gives a bound on the former. We have

$$\|\hat{u} - \hat{u}_N\|_1 \leq C\|\hat{u} - \hat{u}_I\|_1. \tag{5.16}$$

Using a nonuniform grid, as described before (see Section 4.2), and a linear B-spline basis of the space of continuous piecewise linear functions, $\hat{u}_I$ is the interpolant of a function $\hat{u} \in H_B^1$ between every grid point x_k, satisfying (5.15). Because of (5.16), we are interested in obtaining estimates using Sobolev norms or seminorms. It can be shown that for any $\hat{u} \in H^2(0,1)$

$$|\hat{u} - \hat{u}_I|_1 \leq \pi^{-1} h |\hat{u}|_2 \tag{5.17}$$

$$|\hat{u} - \hat{u}_I|_0 \leq \pi^{-2} h^2 |\hat{u}|_2 \tag{5.18}$$

where $h = h_{max}$ (see [3, 24]). If u has less degree of regularity, e.g. if $u \in H^1(0,1)$, the bounds (5.17),(5.17) on $|\hat{u} - \hat{u}_I|_s$ ($s = 0, 1$) deteriorate. In this case, we have

$$|\hat{u} - \hat{u}_I|_1 \leq C_1^* |\hat{u}|_1 \tag{5.19}$$

$$|\hat{u} - \hat{u}_I|_0 \leq C_0^* h |\hat{u}|_1 \tag{5.20}$$

Since

$$\|\hat{u} - \hat{u}_I\|_1^2 = |\hat{u} - \hat{u}_I|_0^2 + |\hat{u} - \hat{u}_I|_1^2, \tag{5.21}$$

using inequality (5.16) we see that if $u \in H^2(0,1)$ then

$$\|\hat{u} - \hat{u}_N\|_1 \leq \tilde{C}_1 h |\hat{u}|_2 \tag{5.22}$$

by inequalities (5.17),(5.18), for some constant $\tilde{C}$, that is,

$$\|\hat{u} - \hat{u}_N\|_1 = \mathcal{O}(h). \tag{5.23}$$

If $u \in H^1(0,1)$ it follows from (5.19) and (5.20) that (5.22) must be replaced by a weaker bound

$$\|\hat{u} - \hat{u}_N\|_1 \leq \tilde{C}_1^* |\hat{u}|_1. \tag{5.24}$$

So we see that high-accuracy basis functions tend to be wasted when the solution of the boundary value problem has low regularity.

The inequality (5.22) leads to the seminorm bounds

$$|\hat{u} - \hat{u}_N|_0 \leq \tilde{C}_1 h |\hat{u}|_2 \tag{5.25}$$

$$|\hat{u} - \hat{u}_N|_1 \leq \tilde{C}_1 h |\hat{u}|_2 \tag{5.26}$$

(see [3]). From (5.26) we can see that $|\hat{u} - \hat{u}_N|_1$ has a bound that agrees with the bound of (5.17) in the sense that the two bounds contain the same power of h. On the other hand, the order of h in (5.25) is lower than the one corresponding to (5.18). However, using the *Aubin-Nitsche method* one can derive a bound on $|\hat{u} - \hat{u}_N|_0$ that is $\mathcal{O}(h^2)$ (see [3]). Applying that method we can obtain the following bound:

$$|\hat{u} - \hat{u}_N|_0 \leq \beta h^2 |\hat{u}|_2, \quad if\ \hat{u} \in H^2(0,1) \tag{5.27}$$

for some constant β.

However, the bounds presented here do not take into account the fact that our boundary value problem has a singularity at the boundary. Therefore, near the singularity, we should expect a weaker bound for the error. The choice of a nonuniform grid can overcome that problem (see [24]). This requires further analysis.

In the next section we study the error occurring when solving each linear system arising from the use of the finite element method.

5.3.2 Linear system

In Section 4.2 we have shown that when applying the finite element method to each iteration ν of the Picard scheme (3.14) we obtain a sequence of linear systems (4.61) of the form

$$K\alpha^{(\nu+1)} = \mathbf{G}^{*(\nu)}, \qquad \nu = 0, 1, \ldots \tag{5.28}$$

where K is an $(N \times N)$, symmetric, tridiagonal and positive definite matrix. These systems are solved using a direct method.

In this case, the sources of error are

- rounding errors and numerical integration errors in the computation of the matrix and the vector;
- rounding errors inherent to the process of solving.

We recall a result concerning the spectral condition number of the stiffness matrix K, $cond(K)$. As proved in [3], if the bilinear form $a(.,.)$ is symmetric and satisfies the hypotheses of coerciveness and boundedness of the Lax-Milgram Lemma, then

$$cond(K) \leq C\frac{\lambda_N}{\lambda_1} \leq C_1 N^2, \tag{5.29}$$

where C and C_1 are constants and λ_N, λ_1 are, respectively, the maximum and minimum eigenvalues of the stiffness matrix K. Another way to write the inequality (5.29) is

$$cond(K) \leq \bar{C} h_{min}^{-2}.$$

This bound on $cond(K)$ is independent of the accuracy of the finite element basis functions, i.e., independent of the degree of the polynomial chosen.

The perturbation properties of the system (4.61) are studied in [3], namely the effect of a small change in $\mathbf{G}^*$ or K on α. It is shown that the influence of perturbations in data on the solution of the system can be significant if and only if the spectral condition number, $cond(K)$, is large.

The elements in the system (4.61) can be computed directly in most of the cases studied here but a numerical integration rule has to be employed, in general, to calculate one of the integrals of the vector $\mathbf{G}^*$ (see Section 4.2). The choice of numerical integration rule is discussed in the next section.

5.3.3 Numerical integration error

In Section 4.2 we pointed out the need for using a numerical integration rule to compute one of the integrals (nonlinear term) on the right-hand side of the linear system (4.61). The composite trapezoidal rule (with four subintervals in each element) chosen allows us to control the error to the same level as the finite element error.

A bound for the absolute error when applying this rule is given by

$$error \leq \frac{C_{max}}{12}\left(\frac{h_{max}}{4}\right)^2 \tag{5.30}$$

where C_{max} is a bound on the modulus of the second derivative of the integrand at a point $\eta \in (0,1)$ and $h_{max} = h_n$ where $h_i = x_i - x_{i-1}$, for $i = 1, \ldots, n$, and x_i are the grid points chosen (see (5.1)). Since the computational effort to compute the numerical integration is less than that of the finite element computational work (which requires the solution of a linear system), the precision of the numerical integration rule can be increased if necessary, so that it has an error at most of the same order as the bounds on the error obtained using the finite element method with linear elements (see Section 5.3.1).

In the next chapter we present and discuss some numerical results obtained for different cases of the parameters p and q of problem (1.5),(1.6).

Chapter 6

Numerical results and conclusions

We recall that to solve the nonlinear problem (1.5),(1.6) we first transformed it into a sequence of linear boundary value problems by using a Picard method (3.14) and then applied a finite element method with a linear B-spline basis and a nonuniform grid to each iteration (Section 4.2). In this way we obtain a sequence of tridiagonal linear systems (4.61) having the same matrix $\mathbf{K}$. Each system can be solved by a direct method yielding the vector $\alpha^{(\nu+1)}$. This vector allows us to compute the finite element solution $\hat{u}_N^{(\nu+1)}$ (see (4.57)) of the boundary value problem corresponding to the iteration $\nu+1$ of the Picard scheme.

In Section 6.1 we present some results as well as some details on the algorithm used. Those results are discussed in Section 6.2.

6.1 Numerical results

We present results concerning the following cases:

i) $p = -\frac{1}{2}$ and $q = \frac{3}{2}$ (*Thomas-Fermi*) ;

ii) $p = -1$ and $q = 2$;

iii) $p = -\frac{5}{4}$ and $q = \frac{9}{4}$ (which required the use of the trapezoidal rule).

The case iii) corresponds to a choice of p between -2 and -1. Note that we did not prove that the Lax-Milgram Lemma assumptions were satisfied by our problem in the cases ii) and iii). Nevertheless, we shall present the numerical results obtained by applying the algorithm to these cases. The results were obtained using double precision arithmetic and an algorithm coded in Fortran. A subroutine from LINPACK was used to solve the tridiagonal linear system.

For the stopping criterion we used the discrete Euclidean norm and both the absolute and relative errors. For example, the absolute error in this norm is computed by

$$\|\alpha^{(\nu+1)} - \alpha^{(\nu)}\| = \left(\sum_{i=1}^{n-1} \left| \alpha^{(\nu+1)}(x_i) - \alpha^{(\nu)}(x_i) \right|^2 \Delta x_i \right)^{1/2}, \qquad \nu = 0, 1, \ldots \tag{6.1}$$

where $\Delta x_i = x_i - x_{i-1}$, $i = 1, \cdots, n-1$, and $\alpha^{(\nu+1)}, \alpha^{(\nu)}$ represent two sucessive solutions of the sequence of linear systems (4.61). The tolerance used is $\epsilon = 10^{-13}$.

In order to compare our numerical results in the cases i)-iii) to others obtained previously ([21, 17, 18]), we will use different grids (i.e. different choices of γ).

Case i): $p = -\frac{1}{2}$ and $q = \frac{3}{2}$

In Table 6.1 we present the numerical results obtained using the modified Picard scheme (3.14) with $\lambda = q$ and a uniform grid ($\gamma = 1$). Note that n stands for the number of grid subintervals (finite elements). In the two first columns we recall the numerical results obtained in [21], corresponding to a finite difference method with n=400 (first column) and after extrapolation (column2). As we can see from Table 6.1, increasing the number of grid points and consequently decreasing the maximum stepsize, we obtain further common digits (comparing with [21]). The stopping criterion (6.1) is satisfied with $\nu = 12$ in the case $y^{(0)} = 0$ or $\nu = 11$ in the case $y^{(0)} = 1 - x$.

Comparing the numerical results obtained when $n = 400$ with those in [21], we have 6 or 7 common digits. For a choice of n agreeing with [17], our results have for the same n, more common digits than the ones given in [17].

Choosing a nonuniform grid with $\gamma = 2$ we could compare our results to the ones given in [18]. When $n = 400$ we obtained 5-6 digits in common against 3-4 digits in [18].

x_i	[21] n=400	[21] extrap.	$n = 50$	n =100	n =200	n =400
0.1	0.849475313	0.849474382	0.849468033	0.849472801	0.849473987	0.849474283
0.2	0.727232837	0.727231852	0.727224475	0.727230014	0.727231394	0.727231738
0.3	0.619295448	0.619294515	0.619287331	0.619292725	0.619294068	0.619294404
0.4	0.520415347	0.520414506	0.520408029	0.520412891	0.520414103	0.520414405
0.5	0.427550745	0.427550017	0.427544486	0.427548638	0.427549673	0.427549931
0.6	0.338686752	0.338686150	0.338681676	0.338685034	0.338685871	0.338686080
0.7	0.252398660	0.252398194	0.252394824	0.252397353	0.252397984	0.252398141
0.8	0.167649343	0.167649022	0.167646774	0.167648461	0.167648882	0.167648987
0.9	0.083686935	0.083686767	0.083685644	0.083686487	0.083686698	0.083686750

Table 6.1: Numerical results obtained in the case i) - Thomas-Fermi

If we compare our results with those obtained in [17] or [18] we conclude that now we have more accurate results with a more efficient algorithm. A visualization of the numerical solution in the form of a graph with 100 equally spaced points in the interval $[0, 1]$ is displayed in Figure 6.1.

We omit the corresponding graphs for cases ii) and iii) since they are similar in shape.

Case ii): $p = -1$ and $q = 2$

In Table 6.2 we present the numerical results obtained using the modified Picard scheme (3.14) with $\lambda = q$ and a uniform grid. In the first column of Table 6.2 are given the numerical results obtained in [17], where a finite difference method and extrapolation were used.

If we start the Picard iteration process with $y^{(0)} = 0$, the approximation of the solution satisfying (6.1) is obtained with $\nu = 18$ whereas if we start with $y^{(0)} = 1 - x$, the approximation of the solution satisfying (6.1) is obtained with $\nu = 17$.

The best approximation obtained ($n = 400$) has apparently from 4 to 6 digits in common with the extrapolated results in [17] depending on the grid points considered.

If we choose a nonuniform grid, the results have more common digits. For example, if $\gamma = 2$ and $n = 480$ we have 4-8 digits in common.

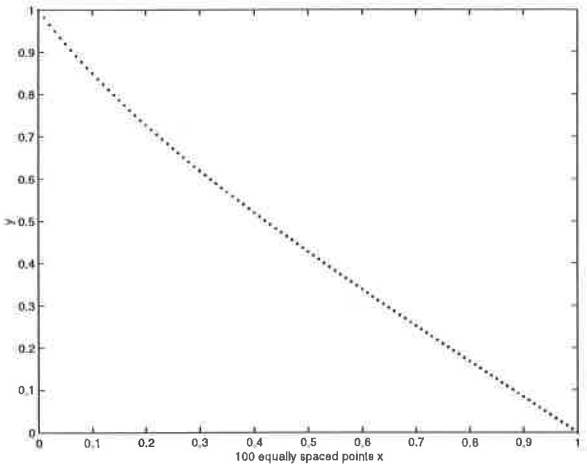

Figure 6.1: Numerical solution at 100 equally spaced points in the Thomas-Fermi case

x_i	[17] extrap.	$n = 60$	$n = 120$	$n = 240$	$n = 480$
0.1	0.780152590	0.779988960	0.780085783	0.780114027	0.780122110
0.2	0.657468748	0.657355053	0.657435994	0.657459466	0.657466153
0.3	0.558348580	0.558253581	0.558321257	0.558340847	0.558346420
0.4	0.470108552	0.470029565	0.470085848	0.470102129	0.470106758
0.5	0.387580436	0.387515870	0.387561883	0.387575189	0.387578972
0.6	0.308145660	0.308094600	0.308130989	0.308141511	0.308144501
0.7	0.230342234	0.230304182	0.230331301	0.230339142	0.230341371
0.8	0.153326388	0.153301096	0.153319122	0.153324334	0.153325815
0.9	0.076623822	0.076611189	0.076620193	0.076622796	0.076623536

Table 6.2: Numerical results obtained in the case ii)

Case iii): $p = -\frac{5}{4}$ and $q = \frac{9}{4}$

In Table 6.3 we present the numerical results obtained using the modified Picard scheme (3.14) with $\lambda = q$ (columns 3-6) and a nonuniform grid corresponding to $\gamma = 4$. The extrapolated numerical results obtained in [18] are presented in column 1. If we start the Picard iteration process with $y^{(0)} = 0$ we need $\nu = 24$ iterations to obtain a numerical solution satisfying (6.1) whereas if we start with $y^{(0)} = 1 - x$ we only need $\nu = 22$ iterations.

The best approximation obtained ($n = 400$) has apparently from 5 to 6 digits in common with the extrapolated results depending on the grid points considered, i.e., better than the one obtained in [17] or [18] without extrapolation.

The choice of a uniform grid ($\gamma = 1$) does not improve the results given in the referred papers.

As was expected the method has slower convergence in this case since p is such that $-2 < p < -1$.

In Figures 6.2-6.4 we compare the speed of convergence of the Picard scheme with and without added terms (respectively Mod Pic and Pic in the legend). The results on the Picard scheme with added terms correspond to the case where $\lambda = q$. The vertical scale is logarithmic and corresponds to the absolute value of the logarithm of the error given

x_i	[18] extrap.	$n = 50$	$n = 100$	$n = 200$	$n = 400$
0.100	0.7043964	0.704427371	0.704416130	0.704404504	0.704395963
0.200	0.5901638	0.590270710	0.590165858	0.590167169	0.590165262
0.300	0.5016889	0.501642767	0.501675452	0.501685464	0.501688100
0.400	0.4233797	0.423406970	0.423390393	0.423377564	0.423379414
0.500	0.3498278	0.349803983	0.349822741	0.349827421	0.349828052
0.600	0.2786058	0.278579929	0.278598305	0.278603915	0.278605403
0.700	0.2084974	0.208495781	0.208496859	0.208496142	0.208497144
0.800	0.1388728	0.138866435	0.138871129	0.138871991	0.138872610
0.900	0.0694182	0.069413878	0.069416707	0.069417737	0.069418065

Table 6.3: Numerical results obtained in the case iii)

by (6.1). In case i), the approximate solution satisfying (6.1) is attained with $\nu = 11$ iterations of the modified Picard scheme against $\nu = 17$ iterations of the Picard scheme without added terms ($\lambda = 0$) (see Figure 6.2). In case ii), as we can see from Figure 6.3, the approximate solution of the modified Picard scheme which satisfies (6.1) is attained with $\nu = 17$ iterations whereas for the Picard scheme without added terms the corresponding solution is attained with $\nu = 25$ iterations. As we can see from Figure 6.4, in the case iii) the speed of convergence of both schemes (modified Picard scheme and Picard scheme) is slower than in the cases i) and ii). In this case, the modified Picard solution satisfying (6.1) is attained with $\nu = 22$ iterations against $\nu = 37$ iterations for the Picard scheme without added terms.

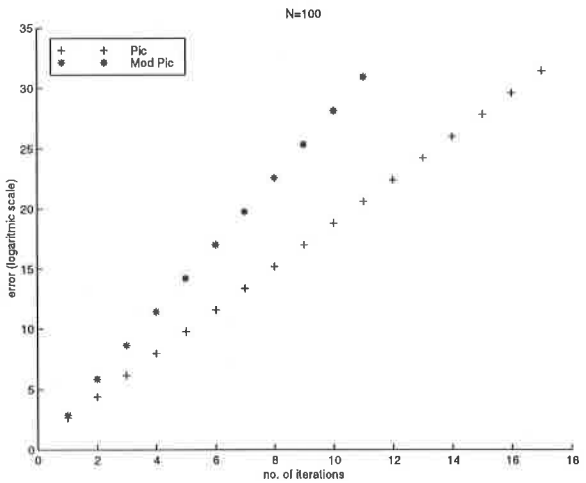

Figure 6.2: Error vs. number of iterations in the Thomas-Fermi case with $y^{(0)} = 1 - x$

As we expected (see Section 3.1), the modified Picard scheme has faster convergence than the unmodified Picard scheme ($\lambda = 0$). The speed of convergence decreases for our choice of decreasing values of p, the slowest speed of convergence corresponding to the case iii).

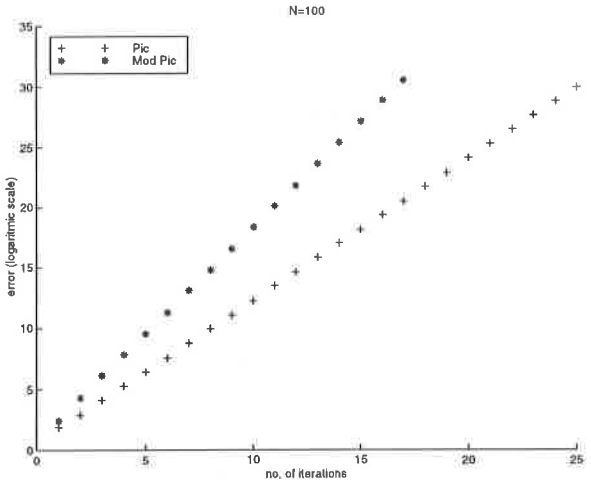

Figure 6.3: Error vs. number of iterations in the case ii) with $y^{(0)} = 1 - x$

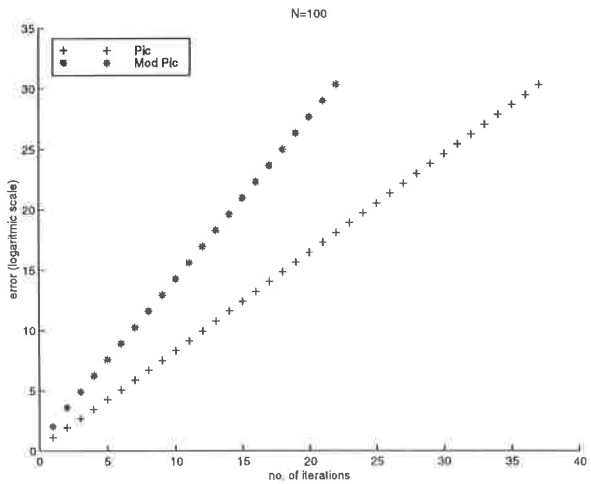

Figure 6.4: Error vs. number of iterations in the case iii) with $y^{(0)} = 1 - x$

6.2 Conclusions

Although it is possible to analyse problem (1.7),(1.10) using a variational principle (e.g. see [9]), here we adopted a different approach which is related to previous work [16, 18].

Our aim was to use the finite element method to solve problem (1.5), (1.6), using an adequate choice of the grid. The choice of a nonuniform grid of the form (5.1) corresponds to solving an equivalent transformed problem by using the finite element method with a uniform grid, which was the approach chosen in previous work (see [16, 18]).

Comparing both approaches, we immediately conclude that the new approach is more efficient in several ways. Firstly, in order to build the linear system we had to calculate fewer integrals; secondly, we used linear B-splines instead of cubic B-splines, resulting in a tridiagonal matrix (in contrast with a band matrix of width seven); thirdly, the matrix is symmetric and positive definite (in [18, 16] the matrix was not symmetric).

The numerical results we obtained have also more digits than the ones given in [17, 18, 16]. However, in the Thomas-Fermi case, the results are as good as the ones obtained in [21] with a choice of a nonuniform grid ($\gamma = 2$), but slighty better if we use a uniform grid ($\gamma = 1$). In the cases corresponding to $-2 < p \leq -1$ (cases ii) and iii) of Section 6.1), our choice of a nonuniform grid (see Section 5.2) yields more accurate results than a uniform grid.

We confirmed in practice that the modified Picard method (3.14) with $\lambda = q$ results in faster convergence (see Figures 6.2-6.4), as was proved theoretically by Mooney under certain assumptions (see Section 3.1).

In spite of not being able to prove theoretically the convergence of the finite element method for the case $-2 < p \leq -1$ (see Section 4.1) we nevertheless presented numerical results showing the convergence of the method in this case.

It remains to study theoretically the convergence of the opposite strategy; that is, the convergence of a modified Picard method applied to the discretised version of our problem obtained after using the finite element method.

Bibliography

[1] M. Al-Zanaidi, C. Grossmann, and R. L. Voller. Monotonous enclosures for the Thomas-Fermi equation in the isolated neutral atom case. *IMA J. Num. Anal.*, 16:413–434, 1996.

[2] N. Anderson and A. M. Arthurs. Variational solutions of the Thomas-Fermi equation. *Quart. Applied Mathematics*, 39:127–129, 1981.

[3] O. Axelsson and V. A. Barker. *Finite Element Solution of Boundary Value Problems - Theory and Computation.* Academic Press, 1984.

[4] E. B. Baker. The applications of the Fermi-Thomas statistical model to the calculation of potential distribution in positive ions. *Physical Review*, 36:630–647, 1930.

[5] R. E. Bellman and R. Kalaba. *Quasilinearisation and Nonlinear Boundary Value Problems.* Elsevier, 1965.

[6] B. L. Burrows and P. W. Core. A variational-iterative approximate solution of the Thomas-Fermi equation. *Quart. Applied Mathematics*, 42:73–76, 1984.

[7] C. Y. Chan and S. W. Du. A constructive method for the Thomas-Fermi equation. *Quart. Applied Mathematics*, 44:303–307, 1986.

[8] C. Y. Chan and Y. C. Hon. A constructive solution for a generalized Thomas-Fermi theory of ionized atoms. *Quart. Applied Mathematics*, 45:591–599, 1987.

[9] P. Csavinszky. Universal approximate analytical solution of the Thomas-Fermi equation for ions. *Physical Review A*, 8:1688–1701, 1973.

[10] G. Fairweather. *Finite Element Galerkin Methods for Differential Equations.* Marcel Dekker,Inc., 1978.

[11] E. Fermi. Un metodo statistico par la determinazione di alcune propietá dell' atome. *Rend. Accad. Naz. del Lincei. Cl. sci. fis., mat. e nat.*, 6:602–607, 1927.

[12] I. S. Gradshtein and I. M. Ryzhik. *Table of Integrals, Series and Products.* Academic Press, 1980.

[13] C. Grossmann. Enclosures of the solution of the Thomas-Fermi equation by monotone discretization. *J. Comp. Phys.*, 98:26–32, 1992.

[14] Y. C. Hon. A decomposition method for the Thomas-Fermi equation. *SEA Bull. Math.*, 20:55–58, 1996.

[15] M. K. Kwong. On the Kolodner-Coffman method for the uniqueness problem of emden-fowler bvp. *J. App. Math. Phys. (ZAMP)*, 41:79–104, 1990.

[16] A. C. Lemos. Solution of Emden-Fowler equations by the finite element method. Msc. thesis, IST, Lisboa, 1995.

[17] P. M. Lima. Numerical methods and asymptotic expansions for the Emden-Fowler equations. *J. Comp. App. Math.*, 70:245–266, 1996.

[18] P. M. Lima and A. C. Lemos. Finite element solution of degenerate boundary-value problems for ordinary differential equations. In A. Alekseev and N. S. Bakhvalov, editors, *Advanced Mathematics: Computations and Applications*, 1995.

[19] C. D. Luning and W. L. Perry. An iterative technique for solution of the Thomas-Fermi equation utilizing a nonlinear eigenvalue problem. *Quart. Applied Mathematics*, 35:257–268, 1977.

[20] J. W. Mooney. Monotone methods for the Thomas-Fermi equation. *Quart. Applied Mathematics*, 36:305–314, 1978.

[21] J. W. Mooney. Numerical schemes for degenerate boundary-value problems. *J. Phys. A*, 26:L413–L421, 1993.

[22] J. W. Mooney and G. F. Roach. Iterative bounds for the stable solutions of convex non-linear boundary value problems. In *Proc. Roy. Soc. Edin.(A)*, 1976.

[23] P. L. Sachdev. *Nonlinear Ordinary Differential Equations and Their Applications.* Marcel Dekker, Inc, 1991.

[24] G. Strang and G. J. Fix. *An Analysis of the Finite Element Method.* Prentice-Hall, 1973.

[25] L. H. Thomas. The calculation of atomic fields. In *Proc. Camb. Phil. Soc.*, 1927.

[26] J. W. Wong. On the generalized Emden-Fowler equations. *SIAM Review*, 17(2):339–360, 1975.