DISCRETE SUB- AND SUPERSOLUTIONS METHOD

David González de la Aleja Gallego

Doctoral thesis submitted in partial fulfillment of requirements for the degree of Ph.D. in

MATHEMATICAL ENGINEERING

UNIVERSIDAD CARLOS III DE MADRID

Advisor:

Dra. Marcela Molina-Meyer

Madrid, Spain
July 2021

Este trabajo ha sido desarrollado en el Departamento de Matemáticas de la Universidad Carlos III de Madrid (UC3M) bajo la dirección de la profesora Marcela Molina Meyer. Se contó con una beca de la UC3M de ayuda al estudio de máster de dos años de duración y posteriormente con un contrato de investigador en formación de la UC3M durante tres años más. Adicionalmente se recibió una ayuda de la UC3M para realizar una estancia de investigación, financiada por el programa ”Ayudas para la movilidad de investigadores de la Universidad Carlos III de Madrid en centros de investigación nacionales o extranjeros, Convocatoria 2015. Modalidad B: investigadores en formación predoctoral”. La estancia de investigación se realizó en el Departamento de Matematicas de Heriot-Watt University (Edimburgo). Además, esta tesis ha sido respaldada parcialmente por el Ministerio de Economia y Competitividad de España con el proyecto MTM2015-65899-P y por el Ministerio de Ciencia, Innovación y Universidades de España con el proyecto PGC2018-097104-B-100.

Published and submitted content

The main results of this thesis are based on the following paper.

- D. ALEJA AND M. MOLINA-MEYER, *Nonlinear finite elements: Sub- and supersolutions for the heterogeneous logistic equation*, Journal of Differential Equations, 278 (2021), 189-219. `https://doi.org/10.1016/j.jde.2020.12.026`

It is partially included in Chapter 1, Chapter 2 and Chapter 5 of this thesis. The material from this source included in this thesis is not singled out with typographic means and references.

The previous paper were presented by the author in the next following talks.

- Elementos finitos para la ecuación logística con coeficientes variables, Seminario de Matemáticas, Universidad Rey Juan Carlos (May 2019).
- Finite elements for the heterogeneous logistic equation, Sharp Dynamics of Differential Equations and Systems (congress), Universidad Complutense de Madrid (March 2019).

Agradecimientos

"Dando una patada a la sílaba im- de la palabra imposible, cualquier persona estará segura de salir adelante." Robert Baden-Powell, fundador del movimiento Scout.

En estos años, imposible es la palabra que más se me ha pasado por la cabeza. Cada ocasión de superación ha dado un golpe positivo a mi aprendizaje, superando así todas mis expectativas. Esta tesis ha marcado un cambio en mi carrera investigadora y no por el número de publicaciones, que solamente es una, sino por todos los conocimientos adquiridos. Ha potenciado mis habilidades matemáticas en el análisis numérico, me ha ayudado a ser más autónomo y a valorar lo difícil que es publicar en una revista internacional, y lo más importante, a nunca rendirse, ya que al final, siempre es posible.

Sin ninguna duda, esta tesis ha sido posible gracias a Marcela. Sin conocerme, se aventuró a ser mi directora de un problema bastante difícil y novedoso. Recuerdo los primero años que cada término de la ecuación nos proponía un auténtico reto. Gracias por hacer posible esta tesis y por tu entrega incondicional.

También, me gustaría agradecer al departamento de Matemáticas de la Universidad Carlos III de Madrid la posibilidad de poder realizar esta tesis doctoral, además de darme la oportunidad de empezar mis primeros pasos como docente. Gracias por la formación recibida y por el gran empujón que habéis dado a mi trayectoria profesional.

Y no podría dejar de mencionar a mis padres, mi familia, mis amigos y a todo la gente que directa o indirectamente ha hecho posible este sueño. Gracias por vuestros ánimos y por apoyarme en todo momento. A todos vosotros, mil gracias.

Contents

1 Summary 1

1.1 Introduction and goal . . . 1

1.2 Finite element and notation . . . 3

2 The heterogeneous logistic equation 5

2.1 Introduction . . . 5

2.2 The discrete bilinear form . . . 7

2.3 The Discrete Maximum Principle . . . 9

2.4 Discrete Sub- and Supersolutions Method . . . 13

2.5 Errors . . . 18

2.6 Some numerical examples . . . 22

2.6.1 Numerical results for the choice $D = 1$, $\lambda = 1$, $\delta = 1$, $c(x) = x + 1$, $a \equiv 1$ and $\beta_0 = \beta_1 = 0$. . . . 23

2.6.2 Numerical results for the choice $D = 10^{-4}$, $\lambda = 1$, $\delta = 1$, $c \equiv 1$, $a(x) = 100x+1$ and $\beta_0 = \beta_1 = 0$. . . . 24

2.6.3 Numerical results for the choice $D = 10^{-3}$, $\lambda = 1$, $\delta = 1, \beta_0 = \beta_1 = 10^{-2}$, $a \equiv 1$ and $c(x) = 10 - (x - 0.5)^2$. . . . 24

3 The generalized diffusive logistic equation 29

3.1 Introduction . . . 29

3.2 The discrete bilinear form . . . 30

3.3 Discrete Sub- and Supersolutions Method . . . 31

3.4 Errors . . . 36

3.5 Simulations . . . 41

3.5.1 Numerical results for $m(x) = \cos(x)$, $a(x) = 1$ and $D = \alpha = \lambda = \delta = 1$. . . . 41

3.5.2 Numerical results for the case $m(x) = \sin(x + 1)$, $a(x) = 1$, $D = \delta = \lambda = 1$ and $\alpha = 200$. . . . 44

4 The generalized diffusive logistic equation with change of variable 47

4.1 Introduction . . . 47

4.2 Applying and improving the Chapter 2 . . . 48

4.3 Simulations . . . 54

5 Conclusions 57

Chapter 1

Summary

1.1 Introduction and goal

This thesis develops the discrete sub- and supersolutions method and applies it to prove the convergence of the nonlinear finite element method applied to the generalized diffusive logistic equation

$$\begin{cases} -(Du' - \alpha m'(x)u)' = \lambda m(x)u - \delta a(x)u^2, & x \in (0,1), \\ Du'(0) = \alpha m'(0)u(0), \\ Du'(1) = \alpha m'(1)u(1), \end{cases} \tag{1.1}$$

where $D > 0$, $\alpha \geq 0$, $\lambda > 0$, $\delta > 0$, $m \in C^2[0,1]$ such that $m(x) > 0$, $m'(0) \geq 0$ and $m'(1) \leq 0$ and $a \in C[0,1]$ with $a(x) > 0$ for all $x \in [0,1]$. This model was introduced in [1], [2], [3] and [4] as a general version of the advection-diffusion model proposed in [10]. The parameter λ permits to amplify or reduce the influence of the population growth rate on the drift term, meanwhile provides a different insight in the theoretical analysis of (1.1).

The dynamics of the parabolic problem associated with (1.1) in the case of positive initial conditions is regulated by the non-negative solutions of (1.1). Assuming the description of the mathematical model in [10], and noting with $u(x,t)$ to the solution of the parabolic problem associated with (1.1), we have that $\lambda m(x)$ is the per capita growth rate, it is not constant along the habitat and it depends on the location x. Moreover, if $m(x) > 0$ for all $x \in [0,1]$, the entire habitat is considered favourable and $\lambda m(x)$ behaves like a source in the whole habitat. There are neither unfavourable regions where the population dies, i.e. $\{x \in (0,1) \text{ such that } m(x) < 0\}$, nor regions where the population does not reproduce, i.e. $\{x \in (0,1) \text{ such that } m(x) = 0\}$. The flux of the population density $u(x,t)$ is $J = -D\frac{\partial u}{\partial x} + \alpha m'(x)u$. Hence the dispersal of $u(x,t)$ consists of two parts: the diffusion $-D\frac{\partial u}{\partial x}$ provide by the Fick's Law and the directed movement upward along $m'(x)$. This type of dispersal is called conditional dispersal. The parameter α measures the rate at which the population moves up $m'(x)$. When we consider $\alpha > 0$, the population $u(x,t)$ shifts in the direction along which m is increasing. It is also assumed that the boundary acts as a reflecting barrier to the population. Moreover, as the dispersal term $\frac{\partial J}{\partial x}$ is in divergence form, the dispersal per se does not increase or decrease the population, [10].

One of the objectives of this thesis is to obtain the discrete version of the characterization (necessary and sufficient conditions) of the Maximum Principle in [40] and [8], that will be the essential tool in the method and the construction of the nonlinear finite element solution. Some results concerning the Discrete Maximum Principle (DMP) can be found in [23], [50], [12], [33] and [34], but in none of them there is a DMP characterization.

The very well known Nonlinear Galerkin Method (NGM) has been applied in [24], [34] and [50] to equations in divergence form, i.e. $-\nabla \cdot (g(x, u)\nabla u) = f(x)$, and to equations that match in the framework of nonlinear monotone operators in [7], [19], [51] and [11]. Unfortunately, neither of these results can be applied to the logistic equation. On the other hand, as the logistic equation require a positive subsolution subject to homogeneous boundary conditions, the results in [32], [29] and [30] can not be applied because they assume positive boundary conditions. In fact, in [29] is said that the finite element solution may not preserve the properties of the solution of the original equation, however, using the DMP, we guarantee that all the properties of the positive solutions of problem (1.1) are actually preserved.

Moreover, a detailed list of the first references to the Galerkin Method can be found in [27], two key works on NGM are [16] and [26], [35] studies the stability and consistency for nonlinear problems and [47] provides a posteriori estimates of the error for nonlinear finite element approximations using the solution of linearized associated problem. More generally, sufficient conditions ensuring that the approximate problem preserves the structure of the manifold of solutions of the continuous problems depending on a parameter have been studied in [48], [36], [13], [14], [15] and [39].

To the best of our knowledge, this thesis proves the convergence of nonlinear finite elements for the logistic equation for the first time in the literature. When a nonlinear elliptic equation is discretized using finite differences, finite element or spectral methods, a nonlinear algebraic system arises. In our case, considering the space of piecewise linear finite elements V_h, $h > 0$, it follows that the approximation of u, denoted by $u_h \in V_h$, or equivalently the vector $\mathbf{u}_h$, where the components of the vector are the coefficients of u_h in the finite element basis, satisfies a nonlinear algebraic system. Now, our goal is to approximate $\mathbf{u}_h > \mathbf{0}$ (all the components of the vector are nonnegative and $\mathbf{u}_h \neq \mathbf{0}$) and compare it with the positive solution u.

In a similar way to the continuous case, we propose the discrete sub- and supersolutions method. This method is entirely new, it can be extended to many other differential problems of elliptic type and it supports routine implementation compared with the proofs of convergence developed in [45], [18], [19] or [11]. It requires an ordered pair of a couple of positive subsolution and supersolution of the finite element discretization and to check that the DMP is fulfilled. We prove that if the mesh size is less than a first critical value, that depends on the smallest constant positive supersolution, denoted by M, of (1.1), the Jacobian matrix evaluated in any positive supersolution less than $\mathbf{M} := M(1, 1, \dots, 1)^T$ satisfies the DMP, meanwhile a positive strict subsolution, denoted by $\mathbf{E}$, gives the coercivity constant.

In the end, we generate via the Newton Method a decreasing sequence of supersolutions $\bar{\mathbf{u}}_h^{(k)}$ such that

$$\mathbf{0} \ll \mathbf{E} \ll \bar{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k)} \leq \mathbf{M}, \quad \text{for all } k \geq 0,$$

where $\ll$ stands for component-wise strict inequalities. And best of all, we prove that the sequence converges to $\mathbf{u}_h$. In [42], the Jacobian is required to be invertible in $[\mathbf{E}, \mathbf{M}]$, however, in this thesis we only need to assume that $\mathbf{E} \ll \mathbf{M}$. Moreover, we obtain explicit bounds of the error in the case that u_h is bounded from below by a positive constant, because in this case the coercivity constant does not depend on h. Consequently, under this condition, we prove that u_h or $\bar{u}_h^{(k)}$ for some $k \geq 0$ approximates to u

when $h \to 0$.

The structure of this thesis is the following. In Chapter 2, we analyze the method exposed above for a heterogeneous logistic equation without the drift term, subject to Neumann or Robin condition boundary conditions. This problem covers the cases when $\alpha = 0$ or $m' \equiv 0$. Moreover, we introduce the DMP and give necessary and sufficient conditions for it to be fulfilled. In Chapter 3, the discrete sub- and supersolutions method is applied to approximate the positive solution of (1.1) when $\alpha m' \not\equiv 0$. The proof of the convergence is complex because the differential operator in (1.1) is non-selfadjoint, but thanks to a change of variable, it is possible to adapt the procedure to deduce explicit bounds of the error. In Chapter 3, we apply a change of variable to transform the problem (1.1) in a problem with a self-adjoint differential operator, consequently allowing to use the results of Chapter 2.

1.2 Finite element and notation

In this thesis, we consider a partition of $[0, 1]$ in N subintervals, $[x_{i-1}, x_i]$, where $x_i = ih$ with $h = \frac{1}{N}$, and the continuous and piecewise linear functions on this partition defined by

$$\varphi_i(x) = \begin{cases} (x - x_{i-1})/h, & x_{i-1} \le x \le x_i, \\ (x_{i+1} - x)/h, & x_i \le x \le x_{i+1}, \\ 0 & \text{otherwise,} \end{cases}$$

for $0 \le i \le N$, which generates the space of finite elements,

$$V_h := \left\{ v_h = \sum_{i=0}^{N} v_{h,i}\varphi_i \quad : \quad \mathbf{v}_h = (v_{h,0}, v_{h,1}, ..., v_{h,N})^T \in \mathbb{R}^{N+1} \right\}.$$

Given a function $f : [0, 1] \to \mathbb{R}$, we denote

$$\|f\|_2 := \left(\int_0^1 f^2\right)^{1/2}, \quad \|f\|_{H^1} := \left(\|f\|_2^2 + \|f'\|_2^2\right)^{1/2} \quad \text{and} \quad \|f\|_\infty := \text{ess sup}_{x\in[0,1]}|f(x)|. \tag{1.2}$$

If $\mathbf{v} = (v_0, v_1, ..., v_n) \in \mathbb{R}^{n+1}$, $n \ge 0$, we denote

$$|\mathbf{v}|_2 := \left(\sum_{i=0}^{n} v_i^2\right)^{1/2}. \tag{1.3}$$

Let us introduce the interpolation operator $\pi_h : C[0, 1] \to V_h$ defined as

$$\pi_h(v) = \sum_{i=0}^{N} v(x_i)\varphi_i. \tag{1.4}$$

It is well known that V_h approximates to $H^2((0, 1))$ in the sense that

$$\|v - \pi_h(v)\|_2 \le h^2\|v''\|_2, \quad \|v - \pi_h(v)\|_{H^1} \le h\sqrt{h^2 + 1}\|v''\|_2 \quad \text{and} \quad \|(v - \pi_h(v))'\|_2 \le h\|v''\|_2. \tag{1.5}$$

The above properties guarantee that V_h is a good subspace where to find the numerical solution, u_h, that approximates u. Moreover, if we consider changing the degree of the finite elements to a higher one, it does not ensure a better convergence if the function belongs to $H^2((0,1))$. The proof of (1.5) and the optimal choice of the polynomial degree can be found in [46, Th. 4.2] and [46, Table 4.1], respectively.

Also, we precise of an analogous result for the multiplication of certain functions which enunciate and prove in the following proposition.

Proposition 1.2.1. *Let $f \in C^1[0,1]$ and $v_h \in V_h$. Then*

$$\|fv_h - \pi_h(fv_h)\|_\infty \leq \|v_h\|_\infty \|f'\|_\infty h.$$

Proof. For $f \in C^1[0,1]$ and $x \in (x_i, x_{i+1})$, $i \in \{0,1,\dots,N\}$,

$$|f(x) - f(x_i)| = \left| \int_{x_i}^{x} f' \right| \leq \int_{x_i}^{x_{i+1}} |f'| \quad \text{and} \quad |f(x) - f(x_{i+1})| = \left| \int_{x}^{x_{i+1}} f' \right| \leq \int_{x_i}^{x_{i+1}} |f'|.$$

Moreover,

$$(fv_h - \pi_h(fv_h))(x) = [f(x) - f(x_i)]\, v_{h,i}\varphi_i(x) + [f(x) - f(x_{i+1})]\, v_{h,i+1}\varphi_{i+1}(x).$$

Hence, we obtain that

$$|(fv_h - \pi_h(fv_h))(x)| \leq \|v_h\|_\infty \left(\varphi_i(x) + \varphi_{i+1}(x)\right) \int_{x_i}^{x_{i+1}} |f'| \leq \|v_h\|_\infty \|f'\|_\infty h$$

because $\varphi_i(x) + \varphi_{i+1}(x) = 1$ for each $x \in (x_i, x_{i+1})$. □

On the other hand, we provide an upper bound for $|\mathbf{v}_h|_2$ in connection with $\|v_h\|_2$ for some $v_h \in V_h$. This result will be necessary for the proof of the convergence.

Proposition 1.2.2. *Let $v_h = \sum_{i=0}^{N} v_i\varphi_i \in V_h$, or equivalently, $\boldsymbol{v}_h = (v_0, v_1, ..., v_N)^T$, then*

$$h\, |\boldsymbol{v}_h|_2^2 \leq 6\, \|v_h\|_2^2.$$

Proof. Integrating the finite elements, we deduce that

$$\|v_h\|_2^2 = \frac{h}{3}\left(v_0^2 + 2\sum_{i=1}^{N-1} v_i^2 + v_N^2\right) + \frac{h}{6}\sum_{i=0}^{N-1} 2v_iv_{i+1}.$$

Then, since

$$2v_iv_{i+1} = (v_i + v_{i+1})^2 - v_i^2 - v_{i+1}^2 \geq -v_i^2 - v_{i+1}^2,$$

it follows that

$$\|v_h\|_2^2 \geq \frac{h}{3}\left(v_0^2 + 2\sum_{i=1}^{N-1} v_i^2 + v_N^2\right) - \frac{h}{6}\sum_{i=0}^{N-1}\left(v_i^2 + v_{i+1}^2\right) \geq \frac{h}{6}|\mathbf{v}_h|_2^2.$$

The proof is finished. □

Finally, we introduce the following inequalities between matrices and therefore also for vectors. Given $R = \{R_{ij}\}$ and $S = \{S_{ij}\}$ two real matrices, we denote

$$R \geq S \quad \text{if } R_{ij} \geq S_{ij} \ \forall\, i,\ j, \qquad R \gg S \ \text{ if } R_{ij} > S_{ij} \ \forall\, i,\ j,$$

and

$$R > S \quad \text{if } R \geq S \text{ and there exist at least } i_0 \text{ and } j_0, \text{ such that } R_{i_0j_0} > S_{i_0j_0}.$$

Chapter 2

The heterogeneous logistic equation

2.1 Introduction

Let us consider the heterogeneous logistic equation subject to general boundary conditions

$$\begin{cases} -Du'' = \lambda c(x)u - \delta a(x)u^2, & x \in (0,1), \\ -Du'(0) + \beta_0 u(0) = 0, & \\ Du'(1) + \beta_1 u(1) = 0, & \end{cases} \tag{2.1}$$

where $D > 0$, $\lambda \in \mathbb{R}$, $\delta > 0$, $\beta_0 \geq 0$, $\beta_1 \geq 0$, $c \in C[0,1]$ and $a \in C[0,1]$, with $a(x) > 0$ for all $x \in [0,1]$. This equation models the asymptotic behavior of a population within a spatially heterogeneous region, whose growth is subject to logistic self-regulation, with a nonnegative nontrivial initial population, where $u(x)$ represents the density of the population at location x, $D > 0$ is the random diffusion rate, $\lambda c(x)$ is the spatially varying local growth rate, and $\delta a(x)$ measures the strength of the density dependence of the logistic self-limitation. In particular, if Robin boundary conditions are imposed, i.e. $\beta_0 > 0$ and $\beta_1 > 0$, the surrounding medium is not considered inhospitable for the population. It is well known that in a second stage, to simulate competition or cooperation between different populations, it is first needed to simulate positive solutions of the logistic equation (2.1).

There has been a great progress in the study of the existence of positive solutions of the logistic equation over the last thirty years, from the pionering modelling in [41] to the existence of metasolutions in [38]. During this time different techniques have been applied to overcome the ongoing problems in solving the heterogeneous logistic equation, such as sub- and supersolutions, bifurcation theory and degree theory, see for example [28], [8], [20], [21], [2], [3] and [4]. Undoubtedly, the characterization of the Maximum Principle established in [40], and later refined in [8], is one fundamental tool in the proof of existence and uniqueness of u. Concerning sub- and supersolutions of (2.1), it is demonstrated in [6] that, if there exists a supersolution, $\bar{U} \in H^2\left((0,1)\right)$, and a subsolution $\underline{U} \in H^2((0,1))$, in the sense that

$$\begin{gathered} -D\bar{U}'' \geq \lambda c(x)\bar{U} - \delta a(x)\bar{U}^2, \quad -D\underline{U}'' \leq \lambda c(x)\underline{U} - \delta a(x)\underline{U}^2 \quad \text{almost everywhere in } (0,1), \\ D\bar{U}'(0) \leq \beta_0\bar{U}(0), \quad D\bar{U}'(1) \geq -\beta_1\bar{U}(1), \quad D\underline{U}'(0) \geq \beta_0\underline{U}(0), \quad D\underline{U}'(1) \leq -\beta_1\underline{U}(1), \end{gathered} \tag{2.2}$$

and satisfying $0 < \underline{U} \leq \bar{U}$, then there exists a positive solution $u \in H^2((0,1))$ of (2.1), $u > 0$ ($u \geq 0$ and $u \neq 0$), such that $\underline{U} \leq u \leq \bar{U}$ which is the unique positive solution by [21].

Now, consider the space of piecewise linear finite elements $V_h \subset H^1((0,1))$, $h > 0$. Hence, if u is the solution of (2.1), u satisfies

$$D\int_0^1 u'v' - \lambda\int_0^1 cuv + \int_{\{0,1\}} \beta uv + \delta\int_0^1 au^2v = 0, \quad \forall v \in H^1((0,1)), \tag{2.3}$$

where $\beta(0) = \beta_0$ and $\beta(1) = \beta_1$ and then, it follows that the approximation of u, denoted by $u_h \in V_h$, satisfies

$$D\int_0^1 u_h'v_h' - \lambda\int_0^1 \pi_h(c)u_hv_h + \int_{\{0,1\}} \beta u_hv_h + \delta\int_0^1 \pi_h(a)u_h^2v_h = 0, \ \ \forall v_h \in V_h, \tag{2.4}$$

where π_h is the interpolation operator onto V_h. Thus, (2.4) can be written as

$$\mathbf{F}_h(\mathbf{u}_h) = \mathbf{0}, \tag{2.5}$$

where the components of the vector $\mathbf{u}_h$ are the coefficients of u_h in the finite element basis. As in the continuous case, we define $\bar{\mathbf{z}}$ as a positive strict supersolution of (2.5) if

$$\mathbf{F}_h(\bar{\mathbf{z}}) > \mathbf{0} \quad \text{with} \ \ \bar{\mathbf{z}} > \mathbf{0}, \tag{2.6}$$

and $\underline{\mathbf{z}}$ as a positive strict subsolution of (2.5) if

$$\mathbf{F}_h(\underline{\mathbf{z}}) < \mathbf{0} \quad \text{with} \ \ \underline{\mathbf{z}} > \mathbf{0}, \tag{2.7}$$

and propose the method of the sub- and supersolutions exposed in the introduction.

In this chapter, we generate via the Newton Method a decreasing sequence of supersolutions $\bar{\mathbf{u}}_h^{(k)}$ and at the same time, via a Modified Newton Method, we also get an increasing sequence of subsolutions $\underline{\mathbf{u}}_h^{(k)}$ such that

$$\underline{\mathbf{u}}_h^{(k)} \ll \underline{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k)}, \quad \text{for all} \ \ k \geq 0. \tag{2.8}$$

Then, we prove that both sequences converge to $\mathbf{u}_h$. Moreover, we obtain explicit a priori bounds of the error in the case that $\mathbf{u}_h$ is bounded from below by a positive constant, because in this case the coercivity constant does not depend on h. Also, we determine the a posteriori bounds of the distance between each iteration of the Newton Method and the exact solution of (2.1). In addition, the a priori bounds depend on the positive strict subsolution and the positive strict supersolution of (2.5), which in turn depend on the maximum and minimum values of the coefficients $a(x)$ and $c(x)$ and the parameters δ and λ. Nevertheless, the a posteriori bounds depend on the stopping criterion established for the Newton Method.

The structure of the chapter is now detailed. In Section 2.2, problem (2.5) is re-written using matrices notations and its Jacobian is calculated. In Section 2.3, sufficient conditions that ensure the Jacobian matrix to be non-singular M-matrix are given. The existence of the principal eigenvalue is proven for tridiagonal matrices whose element in both principal subdiagonals are negative. Meanwhile, the characterization of the DMP is proven in Theorem 2.3.2. Thereupon, some very important properties of the principal eigenvalue are proven and then used to obtain the uniqueness of the positive solution of (2.5). In Section 2.4, the existence of the positive solution of problem (2.5) is proven, initiating

Newton Method in $\mathbf{M} = M(1,1,...,1)$ for large enough $M > 0$. Moreover, two sequences verifying (2.6) and (2.7) are built, one of positive strict supersolutions and the other of positive strict subsolutions, respectively, that also satisfy (2.8). In some situations searching for the initial subsolution $\underline{\mathbf{u}}_h^{(0)}$ is really a difficult task. In Theorem 2.4.2, $\underline{\mathbf{u}}_h^{(0)}$ is constructed in the case $\lambda c(x) > 0$ and with Neumann boundary conditions. Henceforth, Theorem 2.4.3 deals with the general case. In Section 2.5, the convergence of the nonlinear finite elements is proved, using sub- and supersolutions, the monotonicity in (2.8) and the coercivity of two different bilinear forms. Finally, in Section 2.6, three very qualitatively different examples are exposed, showing the different behaviours of the nonlinear finite element solutions. The first one, with non constant $c(x)$, large diffusion and mesh size guaranteeing the DMP, provides a small approximation error. The second case corresponds to small diffusion and non constant coefficient $a(x)$, the simulations present oscillations when the mesh size is not sufficiently small, but once the mesh size is less than a critical level, the DMP is fulfilled, the iterations of the Newton Method are ordered -they do not cross each other- and the oscillations disappear. Although the mesh size has to be very small to obtain the same error order of the first example. In the third and final example the solution is subject to Robin boundary conditions, the coefficient $c(x)$ is non constant and the diffusion is equal to 10^{-4}. In this example, even though the Newton Method converges, the discrete solution oscillates in the cases that the DMP is not satisfied. The numerical simulations make us conjecture that if the DMP is satisfied the oscillations disappear.

2.2 The discrete bilinear form

Throughout this chapter, given $f \in C[0,1]$, we define the following bilinear forms as

$$b_f : H^1((0,1)) \times H^1((0,1)) \to \mathbb{R} \quad \text{and} \quad b_f^h : V_h \times V_h \to \mathbb{R}$$

where

$$b_f(w,v) := D\int_0^1 w'v' - \lambda\int_0^1 cwv + \int_{\{0,1\}} \beta wv + \delta\int_0^1 afwv$$

and

$$b_f^h(w_h,v_h) := D\int_0^1 w_h'v_h' - \lambda\int_0^1 \pi_h(c)w_hv_h + \int_{\{0,1\}} \beta w_hv_h + \delta\int_0^1 \pi_h(a)fw_hv_h,$$

with $\beta(0) = \beta_0$ and $\beta(1) = \beta_1$. By (2.3), we observe that

$$b_u(u,v) = 0 \quad \text{for all} \ \ v \in H^1((0,1)), \tag{2.9}$$

so that, our purpose is to find an approximation of u,

$$u_h = \sum_{i=0}^{N} u_{h,i}\varphi_i, \quad \text{or equivalently,} \quad \mathbf{u}_h = (u_{h,0}, u_{h,1}, ..., u_{h,N})^T,$$

such that

$$b_{u_h}^h(u_h,v_h) = 0 \quad \text{for all} \ \ v_h \in V_h. \tag{2.10}$$

Notice that (2.10) may be considered as a system of nonlinear equations. Indeed, as $b^h_{u_h}(u_h,\cdot)$ is linear, it is sufficient to check that

$$b^h_{u_h}(u_h,\varphi_i)=0 \quad \text{for all} \;\; i\in\{0,1,...,N\}\,.$$

Analogously, using that $b^h_{u_h}$ is also linear with respect to the first entry, (2.10) is equivalent to the nonlinear mapping

$$\mathbf{F}_h:\mathbb{R}^{N+1}\to\mathbb{R}^{N+1},\quad \mathbf{F}_h(\mathbf{z}):=(A_h+Y_h(\mathbf{z}))\,\mathbf{z},\quad \mathbf{z}=(z_0,z_1,...,z_N)^T,$$

where $A_h\in\mathbb{R}^{N+1\times N+1}$ and $Y_h(\mathbf{z})\in\mathbb{R}^{N+1\times N+1}$ are defined by

$$A_h=\left\{b^h_0(\varphi_j,\varphi_i)\right\}_{i,j=0}^N \quad \text{and} \quad Y_h(\mathbf{z})=\left\{\delta\int_0^1\pi_h(a)z\varphi_j\varphi_i\right\}_{i,j=0}^N,\quad z=\sum_{i=0}^N z_i\varphi_i,$$

where i indicates the row and j indicates the column, satisfying

$$\mathbf{F}_h(\mathbf{u}_h)=\mathbf{0}.$$

Clearly, the matrices A_h and $Y_h(\mathbf{z})$ are easily obtained integrating the functions of the basis of finite elements and/or their derivatives. Hence,

$$A_h=\frac{D}{h}A_{h,2}-\frac{\lambda h}{12}A_{h,1}+A_{h,0}, \tag{2.11}$$

where the element of the symmetric and tridiagonal matrices $A_{h,2}$, $A_{h,1}$ and $A_{h,0}$ are the following

$$\begin{aligned}&(A_{h,2})_{00}=1, \qquad &&(A_{h,1})_{00}=3c_0+c_1, \qquad &&(A_{h,0})_{00}=\beta_0,\\ &(A_{h,2})_{NN}=1, &&(A_{h,1})_{NN}=c_{N-1}+3c_N, &&(A_{h,0})_{NN}=\beta_1,\end{aligned}$$

if $i\in\{1,2,...,N-1\}$,

$$(A_{h,2})_{i,i}=2,\quad (A_{h,1})_{i,i}=c_{i-1}+6c_i+c_{i+1}\quad\text{and}\quad (A_{h,0})_{i,i}=0,$$

and finally, if $i\in\{0,1,...,N-1\}$,

$$(A_{h,2})_{i,i+1}=-1,\quad (A_{h,1})_{i,i+1}=c_i+c_{i+1}\quad\text{and}\quad (A_{h,0})_{i,i+1}=0,$$

where $c_i=c(x_i)$. In the same way, we obtain that

$$Y_h(\mathbf{z})=\frac{\delta h}{60}Y_{h,0}(\mathbf{z}), \tag{2.12}$$

where the element of the symmetric and tridiagonal matrix $Y_{h,0}(\mathbf{z})$ are

$$\begin{aligned}(Y_{h,0}(\mathbf{z}))_{00}&=z_0(12a_0+3a_1)+z_1(3a_0+2a_1),\\ (Y_{h,0}(\mathbf{z}))_{NN}&=z_{N-1}(2a_{N-1}+3a_N)+z_N(3a_{N-1}+12a_N),\end{aligned}$$

if $i \in \{1, 2, ..., N - 1\}$,

$$(Y_{h,0}(\mathbf{z}))_{ii} = z_{i-1}(2a_{i-1} + 3a_i) + z_i(3a_{i-1} + 24a_i + 3a_{i+1}) + z_{i+1}(3a_i + 2a_{i+1}),$$

and if $i \in \{0, 1, ..., N - 1\}$,

$$(Y_{h,0}(\mathbf{z}))_{i,i+1} = z_i(3a_i + 2a_{i+1}) + z_{i+1}(2a_i + 3a_{i+1}),$$

where $a_i = a(x_i)$ and $\mathbf{z} = (z_0, z_1, ..., z_N)^T$.

Newton Method is a fundamental tool used in this chapter, it requires the Jacobian of $\mathbf{F}_h$ evaluated in some $\mathbf{z}$, which is expressed as

$$J_h(\mathbf{z}) := \left\{ \frac{\partial F_{h,i}(\mathbf{z})}{\partial z_j} \right\}_{i,j=0}^{N} = A_h + 2Y_h(\mathbf{z}), \tag{2.13}$$

where $\mathbf{F}_h = (F_{h,0}, F_{h,1}, ..., F_{h,N})^T$.

2.3 The Discrete Maximum Principle

In this section, we will derive an upper bound for the finite element mesh size h in order to enforce the Jacobian Matrix $J_h(\mathbf{z})$ in (2.13) to be of the form

$$E = \begin{bmatrix} e_1 & -\hat{e}_1 & 0 & 0 & ... \\ -\tilde{e}_1 & e_2 & -\hat{e}_2 & 0 & ... \\ 0 & -\tilde{e}_2 & e_3 & -\hat{e}_3 & ... \\ 0 & 0 & -\tilde{e}_3 & e_4 & ... \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}, \quad \text{where } \tilde{e}_i > 0 \text{ and } \hat{e}_i > 0. \tag{2.14}$$

Note that the matrices satisfying (2.14) can be expressed in the form

$$E = sI - B, \quad \text{where} \;\; s = \max\{|e_i|\} + 1 \;\; \text{and} \;\; B \;\; \text{is a nonnegative matrix.}$$

Moreover, the matrix B is irreducible because it is a tridiagonal matrix with $\tilde{e}_i \neq 0$ and $\hat{e}_i \neq 0$ (see [31, 6.2.24]). We will provide below, as a consequence of the Perron-Frobenius theory on non-negative matrices applied to B, some very useful properties of E. In the following theorem, we denote with $\rho(B)$ the spectral radius of B.

Theorem 2.3.1. *Let $E \in \mathbb{R}^{n \times n}$ be a matrix satisfying* (2.14). *Then,*

$$\sigma_0(E) := s - \rho(B)$$

is the unique eigenvalue of E, called the principal eigenvalue of E, that admits a real eigenvector $\mathbf{y}_E > 0$, called the principal eigenvector. Moreover, $\mathbf{y}_E \gg 0$, $\sigma_0(E)$ is algebraically and geometrically simple, and if $\tau \neq \sigma_0(E)$ is an eigenvalue of E,

$$\sigma_0(E) < Re(\tau),$$

where $Re(\tau)$ is the real part of τ.

Proof. As B is an irreducible and nonnegative matrix, $\rho(B)$ is a positive algebraically and geometrically simple eigenvalue of B associated to $\mathbf{y}_B \gg \mathbf{0}$ (see [31, Th. 8.4.4]). Thus, $\sigma_0(E)$ is an eigenvalue of E associated to $\mathbf{y}_B \gg \mathbf{0}$ and it is algebraically and geometrically simple.

Now, we suppose that there exists an eigenvalue of E, μ, associated to $\mathbf{y} > 0$. Since B is an irreducible and nonnegative matrix, then $(I+B)^{n-1} \gg 0$ (see [31, 8.4.1]), as well as $s-\mu$ is an eigenvalue of B associated to $\mathbf{y}$, we have that

$$0 \ll (I + B)^{n-1}\mathbf{y} = (1 + s - \mu)^{n-1}\mathbf{y}.$$

Consequently, $\mathbf{y} \gg 0$. Moreover, applying [31, Cor. 8.1.30] to B, we obtain that $\mu = \sigma_0(E)$ and therefore, the uniqueness of the principal eigenvalue is proved, besides $\mathbf{y}_E := \mathbf{y} \gg \mathbf{0}$.

Finally, if we suppose that τ is an eigenvalue of E, $s - \tau$ is also an eigenvalue of B and then

$$s - \mathrm{Re}(\tau) \leq |s - \tau| \leq \rho(B) \quad \Rightarrow \quad \sigma_0(E) \leq \mathrm{Re}(\tau).$$

In the case that $\sigma_0(E) = \mathrm{Re}(\tau)$, we have that $s - \mathrm{Re}(\tau) < |s - \tau|$ because $\tau \neq \sigma_0(E)$. Then, we obtain that $\sigma_0(E) < \mathrm{Re}(\tau)$ which is a contradiction. □

Note that in the special case in which the principal eigenvalue of E is positive, it is guaranteed that E is an invertible matrix. Moreover, it holds that $s > \rho(B)$ and thus, E is a non-singular M-matrix. The theory on M-matrices has been extensively developed, i.e., [17], [25], [43], [44] and [49]. In this sense, the following theorem provides us with the Discrete Maximum Principle.

Theorem 2.3.2. Discrete Maximum Principle *Let $E \in \mathbb{R}^{n\times n}$ be a matrix satisfying* (2.14). *Then, each of the following conditions are equivalent:*

i) $\sigma_0(E) > 0$.

ii) There exists $\mathbf{y} > \mathbf{0}$ *with* $E\mathbf{y} > \mathbf{0}$.

iii) $E^{-1} \gg 0$.

Proof. Condition *ii*) follows immediately from condition *i*), choosing the principal eigenvector of E given in Theorem 2.3.1. Also, due to Theorem 2.3.1 and condition *iii*), we prove *i*), multiplying $E\mathbf{y}_E = \sigma_0(E)\mathbf{y}_E$ by $E^{-1} \gg 0$, because $\mathbf{y}_E \gg 0$. Finally, suppose condition *ii*). As $s\mathbf{y} > B\mathbf{y}$ and $I + B > 0$, we have that

$$(1 + s)^{n-1}\mathbf{y} > (I + B)^{n-1}\mathbf{y},$$

as well as $(I+B)^{n-1} \gg 0$ because B is an irreducible and nonnegative matrix (see [31, 8.4.1]), this implies that $\mathbf{y} \gg 0$. Consequently, condition *iii*) is deduced by [17, Th. 6.2.7]. Note that E is a irreducible matrix with negative element in both its principal sub-diagonals. The proof is complete. □

At this regard, we obtain a principal eigenvalue comparison result for matrices satisfying (2.14).

Theorem 2.3.3. *Let P, $Q \in \mathbb{R}^{n\times n}$ be matrices, both satisfying* (2.14)*, such that $P > Q$. Then*

$$\sigma_0(P) > \sigma_0(Q).$$

Proof. From Theorem 2.3.1, it follows that there exists the principal eigenvector of Q, $\mathbf{y}_Q \gg \mathbf{0}$. Let us define $R := P - \sigma_0(Q)I$. Then,

$$R\mathbf{y}_Q = (P - \sigma_0(Q)I)\,\mathbf{y}_Q = (P - Q)\,\mathbf{y}_Q > \mathbf{0},$$

because $P > Q$ and $\mathbf{y}_Q \gg \mathbf{0}$. Moreover, R satisfies (2.14) because P satisfies it. Now, we can apply Theorem 2.3.2 to deduce that

$$\sigma_0(R) > 0. \tag{2.15}$$

Thus, applying Theorem 2.3.1 to R and using its definition, it follows that

$$P\mathbf{y}_R = (\sigma_0(Q) + \sigma_0(R))\,\mathbf{y}_R \quad \text{with} \;\; \mathbf{y}_R \gg \mathbf{0},$$

and consequently, by uniqueness of the principal eigenvalue, we conclude that

$$\sigma_0(P) = \sigma_0(Q) + \sigma_0(R).$$

Finally, from (2.15), we obtain the desired result. □

The following theorem provides the critical size h_M, such that, for all $h < h_M$, $J_h(\mathbf{z})$ fulfills (2.14) for $\mathbf{z} \le \mathbf{M} := M(1, 1, ..., 1)^T$. Note that h_M depends on the parameters present in problem (2.1).

Theorem 2.3.4. *Let $M > 0$. Assume that $h \in (0, h_M)$, where*

$$h_M := \left(\frac{6D}{\|2\delta Ma - \lambda c\|_\infty} \right)^{1/2}, \tag{2.16}$$

with $h_M = \infty$ if the denominator is zero. Then, for all $z \le \boldsymbol{M} := M(1, 1, ..., 1)^T$, the symmetric and tridiagonal matrix

$$J_h(z) := A_h + 2Y_h(z)$$

fulfills (2.14).

Proof. We need to prove that $A_h + 2Y_h(\mathbf{z})$ has negative sub-diagonal entries. It is enough to prove that $A_h + 2Y_h(\mathbf{M})$ has negative sub-diagonal entries because

$$A_h + 2Y_h(\mathbf{z}) \le A_h + 2Y_h(\mathbf{M}) \quad \text{when} \;\; \mathbf{z} \le \mathbf{M}.$$

Then, for $i \in \{0, 1, ..., N - 1\}$, using the expressions in (2.11) and (2.12), we conclude that

$$(A_h + 2Y_h(\mathbf{M}))_{i,i+1} = -\frac{D}{h} - \frac{\lambda h}{12}(c_i + c_{i+1}) + \frac{M\delta h}{6}\,(a_i + a_{i+1}) \le -\frac{D}{h} + \frac{h}{6}\|2\delta Ma - \lambda c\|_\infty < 0,$$

if $h \in (0, h_M)$. □

In [46], the linear finite element method on a uniform grid is used to solve a simpler linear equation with constant coefficients, i.e. $c(x) = -1$, $\delta = 0$, subjected to non-homogeneous Dirichlet boundary conditions. There, in order to avoid oscillations of the numerical solution, it is obtained (2.16) with $M = 0$.

In our case, when $\bar{\mathbf{z}}$ satisfies (2.6), we will need to prove that $J_h(\bar{\mathbf{z}})$ satisfies the Discrete Maximum Principle (Theorem 2.3.2) and hence it is non-singular. This result is shown in the following theorem.

Theorem 2.3.5. *Let $M > 0$ and $h \in (0, h_M)$, where h_M is given in Theorem 2.3.4. Then,*

$$\sigma_0(J_h(\bar{z})) > 0 \tag{2.17}$$

and

$$(J_h(\bar{z}))^{-1} \gg 0, \tag{2.18}$$

for each $\bar{z} \leq M(1, 1, ..., 1)^T$ satisfying (2.6).

Proof. Due to Theorem 2.3.4, both matrices $J_h(\bar{\mathbf{z}})$ and $A_h + Y_h(\bar{\mathbf{z}})$ satisfy (2.14). Moreover, as $\bar{\mathbf{z}} > 0$, we have the following inequality

$$J_h(\bar{\mathbf{z}}) > A_h + Y_h(\bar{\mathbf{z}}).$$

Then, applying Theorem 2.3.3, we deduce that

$$\sigma_0(J_h(\bar{\mathbf{z}})) > \sigma_0(A_h + Y_h(\bar{\mathbf{z}})). \tag{2.19}$$

Now, as $\bar{\mathbf{z}} \in \mathbb{R}^{N+1}$ satisfies (2.6), we find that $\sigma_0(A_h + Y_h(\bar{\mathbf{z}})) > 0$ thanks to Theorem 2.3.2, and consequently, by (2.19), we obtain (2.17). Finally, applying once again Theorem 2.3.2, we obtain (2.18). □

At this stage, it remains to study the uniqueness of the positive solutions $\mathbf{u}_h$ of problem (2.5). The uniqueness is obtained as a consequence of comparing principal eigenvalues together with Theorem 2.3.1. Moreover, we calculate the principal eigenvalues of A_h, $A_h + Y_h(\mathbf{u}_h)$ and $J_h(\mathbf{u}_h)$.

Theorem 2.3.6. *Let $M > 0$ and $h \in (0, h_M)$, where h_M is given in Theorem 2.3.4. Suppose that*

$$\boldsymbol{F}_h(\boldsymbol{u}_h) = \boldsymbol{0} \quad \textit{and} \quad \boldsymbol{0} < \boldsymbol{u}_h \leq M(1, 1, ..., 1)^T. \tag{2.20}$$

Then, $\boldsymbol{u}_h$ is the unique vector satisfying (2.20). *Moreover,*

$$\sigma_0(A_h) < 0, \quad \sigma_0(A_h + Y_h(\boldsymbol{u}_h)) = 0 \quad \textit{and} \quad \sigma_0(J_h(\boldsymbol{u}_h)) > 0. \tag{2.21}$$

Proof. First of all, from (2.20) and Theorem 2.3.4 we deduce that A_h, $A_h + Y_h(\mathbf{u}_h)$ and $J_h(\mathbf{u}_h)$ satisfy (2.14). Then, the principal eigenvalues $\sigma_0(A_h)$, $\sigma_0(A_h + Y_h(\mathbf{u}_h))$ and $\sigma_0(J_h(\mathbf{u}_h))$ are well defined. Moreover, using once again (2.20) combined with Theorem 2.3.1, it follows that

$$\sigma_0(A_h + Y_h(\mathbf{u}_h)) = 0. \tag{2.22}$$

Now, as $\mathbf{u}_h > \mathbf{0}$, we can compare the above principal eigenvalues thanks to Theorem 2.3.3 and obtain that

$$\sigma_0(A_h) < \sigma_0(A_h + Y_h(\mathbf{u}_h)) < \sigma_0(J_h(\mathbf{u}_h)). \tag{2.23}$$

Therefore, (2.21) follows from (2.22) and (2.23).

Hereafter, we proceed by contradiction to prove the uniqueness. If we suppose that there exist two vectors $\mathbf{u}_h$ and $\mathbf{v}_h$, with $\mathbf{u}_h \neq \mathbf{v}_h$ satisfying (2.20), then the following equations hold true

$$0 = \mathbf{F}_h(\mathbf{u}_h) - \mathbf{F}_h(\mathbf{v}_h) = A_h(\mathbf{u}_h - \mathbf{v}_h) + Y_h(\mathbf{u}_h)\mathbf{u}_h - Y_h(\mathbf{v}_h)\mathbf{v}_h = (A_h + Y_h(\mathbf{u}_h + \mathbf{v}_h))(\mathbf{u}_h - \mathbf{v}_h).$$

Hence, $\mathbf{u}_h - \mathbf{v}_h \neq \mathbf{0}$ is an eigenvector of the matrix $A_h + Y_h(\mathbf{u}_h + \mathbf{v}_h)$ associated to the eigenvalue 0. Note that $A_h + Y_h(\mathbf{u}_h + \mathbf{v}_h)$ satisfies property (2.14) due to Theorem 2.3.4, since $h < h_M$ and $\mathbf{u}_h + \mathbf{v}_h \leq 2M(1, 1, ..., 1)^T$. Therefore, using Theorem 2.3.1 we obtain that

$$\sigma_0(A_h + Y_h(\mathbf{u}_h + \mathbf{v}_h)) \leq 0. \tag{2.24}$$

On the other hand, as $\mathbf{v}_h > 0$, applying Theorem 2.3.3 and subsequently (2.22), it follows that

$$\sigma_0(A_h + Y_h(\mathbf{u}_h + \mathbf{v}_h)) > \sigma_0(A_h + Y_h(\mathbf{u}_h)) = 0,$$

which is a contradiction with (2.24). □

2.4 Discrete Sub- and Supersolutions Method

In this section, we prove the existence of the positive solution of the nonlinear system (2.5) by using the sequence $\bar{\mathbf{u}}_h^{(k+1)}$ provided by the Newton Method:

$$\bar{\mathbf{u}}_h^{(k+1)} = \bar{\mathbf{u}}_h^{(k)} - (J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1}\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}), \quad k \geq 0, \tag{2.25}$$

where $(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1}$ is the inverse of the matrix $J_h(\bar{\mathbf{u}}_h^{(k)}) := A_h + 2Y_h(\bar{\mathbf{u}}_h^{(k)})$. For that end, we take as initial guess $\bar{\mathbf{u}}_h^{(0)}$ a positive strict supersolution of problem (2.5) in the sense of (2.6). One of our aims is to prove that, for this choice of the initial guess $\bar{\mathbf{u}}_h^{(0)}$, each one of the iterations of Newton Method (2.25) is a positive strict supersolutions of problem (2.5). Then, thanks to Theorem 2.3.5 all $\bar{\mathbf{u}}_h^{(k+1)}$ are well defined and moreover,

$$\bar{\mathbf{u}}_h^{(k)} \gg \bar{\mathbf{u}}_h^{(k+1)}. \tag{2.26}$$

Additionally, we consider another sequence $\underline{\mathbf{u}}_h^{(k+1)}$ provided by a modified Newton Method:

$$\underline{\mathbf{u}}_h^{(k+1)} = \underline{\mathbf{u}}_h^{(k)} - (J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1}\mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}), \quad k \geq 0, \tag{2.27}$$

choosing in this case, as initial guess $\underline{\mathbf{u}}_h^{(0)}$ a positive strict subsolution of problem (2.5) in the sense of (2.7). We will prove that all the iterations in (2.27) are positive strict subsolutions of problem (2.5), and consequently

$$\underline{\mathbf{u}}_h^{(k+1)} \gg \underline{\mathbf{u}}_h^{(k)}. \tag{2.28}$$

In the following theorem, we prove the above remarks by using the convexity of each component of $\mathbf{F}_h$.

Theorem 2.4.1. *Let $k \geq 0$. Suppose that the iterations $\bar{\boldsymbol{u}}_h^{(k)}$ and $\bar{\boldsymbol{u}}_h^{(k+1)}$ in* (2.25) *are well defined and that* (2.26) *is satisfied, then*

$$\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k+1)}) \gg \boldsymbol{0}. \tag{2.29}$$

Moreover, suppose that the iterations $\underline{\boldsymbol{u}}_h^{(k)}$ and $\underline{\boldsymbol{u}}_h^{(k+1)}$ in (2.27) *are well defined,* (2.28) *is satisfied and $\bar{\boldsymbol{u}}_h^{(k)} \geq \underline{\boldsymbol{u}}_h^{(k+1)}$, then*

$$\boldsymbol{F}_h(\underline{\boldsymbol{u}}_h^{(k+1)}) \ll \boldsymbol{0}. \tag{2.30}$$

Proof. First of all, note that $\mathbf{F}_h$ is convex in each component because

$$F_{h,i}(\mathbf{z}) = b_0^h(z,\varphi_i) + \delta\int_0^1 \pi_h(a)z^2\varphi_i, \quad \text{for } \mathbf{z} = (z_0, z_1, ..., z_N)^T \in \mathbb{R}^{N+1}, \quad z = \sum_{j=0}^{N} z_j\varphi_j,$$

where the first term is linear and the second term is convex. Hence, defining

$$\mathbf{v} := \left(\bar{\mathbf{u}}_h^{(k+1)} + \bar{\mathbf{u}}_h^{(k)}\right)/2,$$

it follows that

$$\mathbf{F}_h(\mathbf{v}) - \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \geq J_h(\bar{\mathbf{u}}_h^{(k)})(\mathbf{v} - \bar{\mathbf{u}}_h^{(k)}). \tag{2.31}$$

Now, by (2.26), we have that $\bar{u}_h^{(k+1)}(s) \neq \bar{u}_h^{(k)}(s)$ for each $s \in [0,1]$, and then,

$$2\left(\bar{u}_h^{(k+1)}(s)\right)^2 + 2\left(\bar{u}_h^{(k)}(s)\right)^2 > \left(\bar{u}_h^{(k+1)}(s) + \bar{u}_h^{(k)}(s)\right)^2.$$

Therefore, as $\delta\pi_h(a)\varphi_i$ is positive for s sufficiently near to x_i, we obtain that

$$\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k+1)}) + \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \gg 2\mathbf{F}_h(\mathbf{v}). \tag{2.32}$$

Multiplying (2.31) by 2 and applying (2.32), we find that

$$\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k+1)}) \gg \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) + J_h(\bar{\mathbf{u}}_h^{(k)})(\bar{\mathbf{u}}_h^{(k+1)} - \bar{\mathbf{u}}_h^{(k)}) = \mathbf{0},$$

and (2.29) follows from (2.25).

In the same way, we deduce that

$$\mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}) \gg \mathbf{F}_h(\underline{\mathbf{u}}_h^{(k+1)}) + J_h(\underline{\mathbf{u}}_h^{(k+1)})(\underline{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}),$$

and consequently, as

$$J_h(\bar{\mathbf{u}}_h^{(k)})(\underline{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k)}) \geq J_h(\underline{\mathbf{u}}_h^{(k+1)})(\underline{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k)})$$

because $\bar{\mathbf{u}}_h^{(k)} \geq \underline{\mathbf{u}}_h^{(k+1)}$ and (2.28) is satisfied, it becomes apparent that

$$\mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}) + J_h(\bar{\mathbf{u}}_h^{(k)})(\underline{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k)}) \gg \mathbf{F}_h(\underline{\mathbf{u}}_h^{(k+1)}).$$

Finally, from (2.27) it follows (2.30). □

From now on we focus our efforts to find the initial positive strict supersolution $\bar{\mathbf{u}}_h^{(0)}$ and the positive strict subsolution $\underline{\mathbf{u}}_h^{(0)}$ satisfying (2.6) and (2.7), respectively. In this sense, note that the constant function,

$$M := \max_{x\in[0,1]} \frac{\lambda c(x)}{\delta a(x)} \tag{2.33}$$

is a supersolution of problem (2.1) if $M > 0$, and also, in the particular case that $\beta_0 = \beta_1 = 0$ and $\lambda c(x) > 0$ for all $x \in [0,1]$,

$$\varepsilon := \min_{x\in[0,1]} \frac{\lambda c(x)}{\delta a(x)} \tag{2.34}$$

is a positive subsolution of problem (2.1). In the following theorem the constants ε and M are used to prove the existence of positive sub- and supersolutions of problem (2.5).

Theorem 2.4.2. *Let* $\mathbf{M} = M(1,1,...,1)^T$ *and* $\boldsymbol{\varepsilon} = \varepsilon(1,1,...,1)^T$ *where* M *and* ε *are given in* (2.33) *and* (2.34)*, respectively. Then,*

a) $\boldsymbol{F}_h(\boldsymbol{M}) \geq \boldsymbol{0}$ *if* $M > 0$.

b) $\boldsymbol{F}_h(\boldsymbol{\varepsilon}) \leq \boldsymbol{0}$ *if* $\beta_0 = \beta_1 = 0$ *and* $\lambda c(x) > 0$ *for all* $x \in [0,1]$.

Proof. Firstly, we prove *a*) using the expressions (2.11) and (2.12), $M > 0$ and

$$M\delta a_i \geq \lambda c_i \ \text{ for each } \ i \in \{0,1,...,N\}.$$

Indeed, we have that

$$F_{h,0}(\mathbf{M}) = Mh\Big[-\frac{\lambda}{12}(4c_0 + 2c_1) + \beta_0 + \frac{M\delta}{60}(20a_0 + 10a_1)\Big] \geq 0,$$

because $\beta_0 \geq 0$. Now, if $i \in \{1,2,...,N-1\}$,

$$F_{h,i}(\mathbf{M}) = Mh\Big[-\frac{\lambda}{12}(2c_{i-1} + 8c_i + 2c_{i+1}) + \frac{M\delta}{60}(10a_{i-1} + 40a_i + 10a_{i+1})\Big] \geq 0.$$

Finally,

$$F_{h,N}(\mathbf{M}) = Mh\Big[-\frac{\lambda}{12}(2c_{N-1} + 4c_N) + \beta_1 + \frac{M\delta}{60}(10a_{N-1} + 20a_N)\Big] \geq 0,$$

because $\beta_1 \geq 0$.

In an analogous way, it can be proved *b*), assuming that $\beta_0 = \beta_1 = 0$ and $\lambda c(x) > 0$ for all $x \in [0,1]$. □

In the case that $\lambda c(x) \leq 0$ for all $x \in [0,1]$ ($M \leq 0$), it is not necessary to find a supersolution because (2.1) doesn't admit a positive solution (for more details see [28] and [37]). Therefore, if (2.1) has a positive solution, we can choose M given in (2.33) such that $\mathbf{M}$ is a good candidate to start sequence (2.25).

Now, we obtain $\underline{\mathbf{u}}_h^{(0)}$ in terms of a subsolution $\underline{u}$ of (2.1) satisfying

$$-D\underline{u}'' - \lambda c\underline{u} + \delta a\underline{u}^2 \leq -C \ \text{ in } \ (0,1), \quad D\underline{u}'(0) \geq \beta_0\underline{u}(0), \quad D\underline{u}'(1) \leq -\beta_1\underline{u}(1) \tag{2.35}$$

for some constant $C > 0$ and

$$\underline{u}(x) > 0 \quad \text{for all } \ x \in [0,1]. \tag{2.36}$$

Note that (2.35) is more restrictive than (2.2).

Theorem 2.4.3. *Suppose* $a, c \in H^2((0,1))$ *and* $\underline{u} \in H^2((0,1))$ *satisfying* (2.35) *and* (2.36)*. Then*

$$\boldsymbol{F}_h(\boldsymbol{\pi}_h(\underline{u})) \ll \boldsymbol{0} \quad \textit{with} \ \ \boldsymbol{\pi}_h(\underline{u}) = (\underline{u}(x_0), \underline{u}(x_1), ..., \underline{u}(x_N))^T \gg \boldsymbol{0}, \tag{2.37}$$

for each $h \in (0, h^*)$ *where*

$$h^* := 3^{1/3}\left(\frac{C/2}{\Big(\|\delta a\underline{u} - \lambda c\|_\infty + \delta\|a\|_\infty\|\underline{u}\|_\infty\Big)\|\underline{u}''\|_2 + \|\underline{u}\|_\infty\|\lambda c''\|_2 + \delta\|\underline{u}\|_\infty^2\|a''\|_2}\right)^{2/3}. \tag{2.38}$$

In the case in which the denominator in (2.38) *is equal to zero,* $h^* = \infty$.

Proof. In this proof, we use the notation $v_h := \pi_h(\underline{u})$. From (2.36), we know that $\mathbf{v}_h = \boldsymbol{\pi}_h(\underline{u}) \gg 0$; thus, (2.37) is satisfied if

$$F_{h,i}(\mathbf{v}_h) = b^h_{v_h}(v_h, \varphi_i) < 0 \quad \text{for each} \;\; i \in \{0, 1, ..., N\}\,. \tag{2.39}$$

We now prove (2.39). Firstly, by multiplying the first inequality in (2.35) by φ_i, we find that

$$-D \int_0^1 \underline{u}''\varphi_i - \lambda \int_0^1 c\underline{u}\varphi_i + \delta \int_0^1 a\underline{u}^2\varphi_i \leq -C \int_0^1 \varphi_i.$$

Integrating by parts and using the remaining inequalities of (2.35), it follows that

$$b_{\underline{u}}(\underline{u}, \varphi_i) \leq -Ch\gamma_i, \quad \text{where } \gamma_i = \begin{cases} 1/2, & i = 0, N, \\ 1, & i \in \{1, 2, ..., N-1\}\,, \end{cases}$$

and then,

$$b^h_{v_h}(v_h, \varphi_i) \leq b^h_{v_h}(v_h, \varphi_i) - b_{\underline{u}}(\underline{u}, \varphi_i) - Ch\gamma_i.$$

Hence, (2.39) is satisfied if

$$b^h_{v_h}(v_h, \varphi_i) - b_{\underline{u}}(\underline{u}, \varphi_i) < Ch\gamma_i. \tag{2.40}$$

We finally prove that, for each $h \in (0, h^*)$, (2.40) is fulfilled if h^* is given by (2.38). Indeed, we know both that

$$\int_0^1 v_h'\varphi_i' = \int_0^1 \underline{u}'\varphi_i' \quad \text{and} \quad \int_{\{0,1\}} \beta v_h \varphi_i = \int_{\{0,1\}} \beta \underline{u}\varphi_i,$$

and consequently,

$$b^h_{v_h}(v_h, \varphi_i) - b_{\underline{u}}(\underline{u}, \varphi_i) = -\lambda \int_0^1 \pi_h(c) v_h \varphi_i + \delta \int_0^1 \pi_h(a) v_h^2 \varphi_i + \lambda \int_0^1 c\underline{u}\varphi_i - \delta \int_0^1 a\underline{u}^2\varphi_i.$$

Now, adding and subtracting $\lambda \int_0^1 cv_h\varphi_i$ and $\delta \int_0^1 av_h^2\varphi_i$ and denoting $E_f := \pi_h(f) - f$, we deduce that

$$b^h_{v_h}(v_h, \varphi_i) - b_{\underline{u}}(\underline{u}, \varphi_i) = \int_0^1 (-\lambda c + \delta a(v_h + \underline{u}))E_{\underline{u}}\varphi_i - \lambda \int_0^1 E_c v_h \varphi_i + \delta \int_0^1 E_a v_h^2 \varphi_i.$$

Subsequently, using $\|v_h\|_\infty \leq \|\underline{u}\|_\infty$, Hölder's inequality and (1.5), we can bound the above three integrals in the following way,

$$\left| \int_0^1 (-\lambda c + \delta a(v_h + \underline{u}))E_{\underline{u}}\varphi_i \right| \leq \Big(\|\delta a\underline{u} - \lambda c\|_\infty + \delta \|a\|_\infty \|\underline{u}\|_\infty \Big) \|\underline{u}''\|_2 \|\varphi_i\|_2 h^2,$$

$$\left| \lambda \int_0^1 E_c v_h \varphi_i \right| \leq \|\underline{u}\|_\infty \|\lambda c''\|_2 \|\varphi_i\|_2 h^2 \quad \text{and} \quad \left| \delta \int_0^1 E_a v_h^2 \varphi_i \right| \leq \|\underline{u}\|_\infty^2 \|\delta a''\|_2 \|\varphi_i\|_2 h^2,$$

to obtain that

$$\begin{aligned} &\left| b^h_{v_h}(v_h, \varphi_i) - b_{\underline{u}}(\underline{u}, \varphi_i) \right| \leq \\ &\qquad \Big[\Big(\|\delta a\underline{u} - \lambda c\|_\infty + \|\delta a\|_\infty \|\underline{u}\|_\infty \Big) \|\underline{u}''\|_2 + \|\underline{u}\|_\infty \|\lambda c''\|_2 + \|\underline{u}\|_\infty^2 \|\delta a''\|_2 \Big] \|\varphi_i\|_2 h^2. \end{aligned} \tag{2.41}$$

Finally, as $\|\varphi_i\|_2 = (2\gamma_i h/3)^{1/2}$ and assuming that $h \leq h^*$, we obtain (2.40) using $\gamma_i \geq 1/2$. In the case that the denominator in (2.38) is zero, we obtain that $b^h_{v_h}(v_h, \varphi_i) = b_{\underline{u}}(\underline{u}, \varphi_i)$ in (2.41), and therefore, (2.40) is satisfied for all $h > 0$. □

The main result of this section is stated in the following theorem. We will denote as

$$M_{k,h} := \max_{x\in[0,1]} \bar{u}_h^{(k)}(x) \quad \text{and} \quad \varepsilon_{k,h} := \min_{x\in[0,1]} \underline{u}_h^{(k)}(x). \tag{2.42}$$

Also, we will choose M given in (2.33) such that $\mathbf{M} = M(1,1,...1)^T$ is a positive strict supersolution. Nevertheless, if $\mathbf{M}$ is not a positive strict supersolution, we are in the case that either (2.5) doesn't admit a positive solution or $\mathbf{M}$ is itself a positive solution of (2.5) (see Theorem 2.4.2).

Theorem 2.4.4. *Let M and h_M given in* (2.33) *and Theorem 2.3.4, respectively, such that $\boldsymbol{M}$ satisfies* (2.6)*. Suppose $h < h_M$ so that there exists $\underline{\boldsymbol{u}}_h^{(0)}$ satisfying* (2.7) *and $\underline{\boldsymbol{u}}_h^{(0)} \ll \boldsymbol{M}$. Then, if $\bar{\boldsymbol{u}}_h^{(0)} := \boldsymbol{M}$, the sequences in* (2.25) *and* (2.27) *are well defined, $\sigma_0(J_h(\bar{\boldsymbol{u}}_h^{(k)})) > 0$ and*

$$\boldsymbol{0} < \underline{\boldsymbol{u}}_h^{(k)} \ll \underline{\boldsymbol{u}}_h^{(k+1)} \ll \bar{\boldsymbol{u}}_h^{(k+1)} \ll \bar{\boldsymbol{u}}_h^{(k)} \leq \boldsymbol{M} \tag{2.43}$$

for all $k \geq 0$, and both sequences converge to $\boldsymbol{u}_h$ which is the unique solution of (2.5) *satisfying* (2.20)*. Moreover, $\boldsymbol{u}_h$ satisfies that for each $k \geq 0$,*

$$\varepsilon_{k,h}(1,1,...,1)^T \leq \boldsymbol{u}_h \leq M_{k,h}(1,1,...,1)^T. \tag{2.44}$$

Proof. First of all, note that $\bar{\mathbf{u}}_h^{(0)}$ and $\underline{\mathbf{u}}_h^{(0)}$ satisfy (2.6) and (2.7), respectively, and $\underline{\mathbf{u}}_h^{(0)} \ll \bar{\mathbf{u}}_h^{(0)} := \mathbf{M}$. Now, we proceed by induction. We suppose that $\bar{\mathbf{u}}_h^{(k)}$ and $\underline{\mathbf{u}}_h^{(k)}$ are well defined and satisfy

$$\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) > \mathbf{0}, \quad \mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}) < \mathbf{0} \quad \text{and} \quad \mathbf{0} < \underline{\mathbf{u}}_h^{(k)} \ll \bar{\mathbf{u}}_h^{(k)} \leq \mathbf{M}. \tag{2.45}$$

Then, applying Theorem 2.3.5, $\sigma_0(J_h(\bar{\mathbf{u}}_h^{(k)})) > 0$ and $(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \gg 0$. Consequently, $\bar{\mathbf{u}}_h^{(k+1)}$ and $\underline{\mathbf{u}}_h^{(k+1)}$ are well defined and satisfy (2.26) and (2.28), respectively.

On the other hand, by (2.27), we deduce that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right) = J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k)}\right) + \mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}), \tag{2.46}$$

where we have added and subtracted $J_h(\bar{\mathbf{u}}_h^{(k)})\underline{\mathbf{u}}_h^{(k)}$. Developing (2.46), it follows that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right) = Y_h\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k)}\right)\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k)}\right) + \mathbf{F}_h\left(\bar{\mathbf{u}}_h^{(k)}\right),$$

and owing to (2.45), we have that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right) \gg \mathbf{0}.$$

Then, as $(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \gg 0$, we deduce $\underline{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k)}$. Therefore, by Theorem 2.4.1, we obtain (2.29) and (2.30).

Now, we prove that

$$\underline{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k+1)}, \tag{2.47}$$

and the proof by induction will be finish. Repeating a similar process to (2.46), we affirm that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k+1)}\right) = -\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) + J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right),$$

and developing the right side, adding and subtracting $Y_h\left(\underline{\mathbf{u}}_h^{(k+1)}\right)\left(\underline{\mathbf{u}}_h^{(k+1)}\right)$, it follows that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k+1)}\right) = Y_h\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right)\left(\bar{\mathbf{u}}_h^{(k)} - \underline{\mathbf{u}}_h^{(k+1)}\right) - \mathbf{F}_h\left(\underline{\mathbf{u}}_h^{(k+1)}\right).$$

Then, as $\underline{\mathbf{u}}_h^{(k+1)} \ll \bar{\mathbf{u}}_h^{(k)}$, by (2.30) we obtain that

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k+1)}\right) \gg \mathbf{0}$$

and consequently, we have (2.47) because $(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \gg 0$.

The rest of the proof is a consequence of (2.43) and the uniqueness given in Theorem 2.3.6. Indeed, as the sequence $\left\{\bar{\mathbf{u}}_h^{(k)}\right\}_{k\geq 0}$ is strictly monotonically decreasing and bounded from below by $\underline{\mathbf{u}}_h^{(0)} > \mathbf{0}$ and the sequence $\left\{\underline{\mathbf{u}}_h^{(k)}\right\}_{k\geq 0}$ is strictly monotonically increasing and bounded from above by $\mathbf{M}$, the following limits exist

$$\bar{\mathbf{u}}_h := \lim_{k\to\infty} \bar{\mathbf{u}}_h^{(k)} \quad \text{and} \quad \underline{\mathbf{u}}_h := \lim_{k\to\infty} \underline{\mathbf{u}}_h^{(k)}, \tag{2.48}$$

satisfying $\mathbf{0} < \underline{\mathbf{u}}_h, \bar{\mathbf{u}}_h \leq \mathbf{M}$. Now, making k tend to infinity in

$$J_h(\bar{\mathbf{u}}_h^{(k)})\left(\underline{\mathbf{u}}_h^{(k+1)} - \underline{\mathbf{u}}_h^{(k)}\right) = -\mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}) \quad \text{and} \quad J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \bar{\mathbf{u}}_h^{(k)}\right) = -\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}),$$

because $\mathbf{F}_h \in C^\infty(\mathbb{R}^{N+1})$, we conclude that both $\underline{\mathbf{u}}_h$ and $\bar{\mathbf{u}}_h$ are solutions of (2.5). Moreover, as $\mathbf{0} < \underline{\mathbf{u}}_h, \bar{\mathbf{u}}_h \leq \mathbf{M}$, owing to Theorem 2.3.6, we obtain $\underline{\mathbf{u}}_h = \bar{\mathbf{u}}_h$, therefore, it is the unique solution satisfying (2.20). Finally, (2.44) is obtained from (2.43) and (2.48). The proof is completed. □

Let us remark that, as A_h is not a non-singular M-matrix, it is not possible to use the results in [42], because $\sigma_0(A_h) < 0$, as we proved in (2.21). Nevertheless, we can apply Theorem 2.3.5 in each iteration of (2.25), to prove that $(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \gg 0$.

2.5 Errors

In practice, we need reliable and accurate estimates of the error, understanding the error as the norm of the difference between the corresponding iteration in (2.25), $\bar{u}_h^{(k)}$, and the positive solution of the problem (2.1), u. We will estimate the error using the triangular inequality

$$\|u - \bar{u}_h^{(k)}\| \leq \|u - u_h\| + \|u_h - \bar{u}_h^{(k)}\|, \quad k \geq 1, \tag{2.49}$$

where u_h is given in the Theorem 2.4.4. Moreover the error estimates will provide us with the mesh size h required to fulfill a previously given bound of the error. This critical size of h will depend strongly on the parameters and coefficients appearing in (2.1) and of the a posteriori bounds, $\varepsilon_{k,h}$ and $M_{k,h}$ given in (2.42), which become tighter bounds for u_h as $k \to \infty$. Furthermore, it will also provided us with the exit criterion for Newton Method. Note that the norms used in this section have been defined in (1.2) and (1.3).

Also, it is very important to highlight that, under the same conditions of Theorem 2.4.4,

$$\mathbf{0} \ll \varepsilon_{k,h}(1, 1, ..., 1)^T \leq \mathbf{u}_h \leq M_{k,h}(1, 1, ..., 1)^T \ll \mathbf{M} \quad \text{if } k \geq 1.$$

Moreover, in the cases that $\underline{u}_h^{(0)} := \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon}$ is given in Theorem 2.4.2 or $\underline{\mathbf{u}}_h^{(0)} := \boldsymbol{\pi}_h(\underline{u})$ if the hypotheses of Theorem 2.4.3 are satisfied, $\varepsilon_{k,h}$ is lower bounded by $\varepsilon > 0$ or by $\min_{x\in[0,1]} \underline{u}(x) > 0$, respectively. Therefore, in the mentioned cases, $\varepsilon_{k,h}$ and $M_{k,h}$ are bounded by positive constants independently of h and k.

To begin, we analyse the three bilinear forms b_{2u}, b_{u+u_h} and $b^h_{2u_h}$, as described in the following theorem.

Theorem 2.5.1. *Let the assumptions of Theorem 2.4.4 be satisfied. Then, for $k \geq 1$, it holds that*

$$b_u(w,w) \geq 0, \quad b_{u+u_h}(w,w) \geq C_{1,1}\,\|w\|^2_{H^1} \quad \textit{and} \quad b_{u+u_h}(w,w) \geq C_2\,\|w\|^2_2 \quad \forall\, w \in H^1((0,1)), \tag{2.50}$$

and

$$b^h_{u_h}(w_h,w_h) \geq 0, \quad b^h_{2u_h}(w_h,w_h) \geq \; C_{1,2}\,\|w_h\|^2_{H^1} \quad \textit{and} \quad b^h_{2u_h}(w_h,w_h) \geq C_2\,\|w_h\|^2_2 \quad \forall\, w_h \in V_h, \tag{2.51}$$

where the positive constants $C_{1,1} = C_1(u)$, $C_{1,2} := C_1(\varepsilon_{k,h})$ and C_2 are

$$C_2 := \delta \min_{x\in[0,1]} a(x)\varepsilon_{k,h} \quad \textit{and} \quad C_1(\varphi) := \min\left\{D, \frac{DC_2}{D + \|(\lambda c - \delta a\varphi)^+\|_\infty}\right\}, \tag{2.52}$$

with $(\lambda c - \delta a\varphi)^+$ the part positive of $\lambda c - \delta a\varphi$.

Proof. First of all, notice that, as u is a positive solution of the problem (2.1), u is a positive eigenfunction associated with the eigenvalue 0 of the following eigenvalue problem

$$\begin{cases} -D\psi'' - \lambda c\psi + \delta a u\psi = \sigma\psi, & \text{in } (0,1), \\ -D\psi'(0) + \beta_0\psi(0) = 0, \\ D\psi'(1) + \beta_1\psi(1) = 0. \end{cases}$$

Thus, by [37, Th. 7.7 and Th. 7.8], 0 is the corresponding smallest eigenvalue. Now, we obtain the first inequality in (2.50) from the Rayleigh formula applied to the symmetric form b_u. Moreover, since $u_h \geq \varepsilon_{k,h}$, we conclude that

$$b_{u+u_h}(w,w) \geq \delta \int_0^1 a u_h w^2 \geq C_2\,\|w\|^2_2, \quad \forall\, w \in H^1((0,1)).$$

On the other hand, by Rayleigh-Ritz Theorem (see [31, Th. 4.2.2]) applied to $A_h + Y_h(\mathbf{u}_h)$, as it is a symmetric matrix, and using Theorem 2.3.6, we have that

$$\mathbf{w}_h^T(A_h + Y_h(\mathbf{u}_h))\mathbf{w}_h \geq 0, \quad \forall \;\, \mathbf{w}_h \in \mathbb{R}^{N+1},$$

or equivalently, the first inequality in (2.51) is satisfied. Consequently, as $u_h \geq \varepsilon_{k,h}$, it follows that

$$b^h_{2u_h}(w_h,w_h) \geq \delta \int_0^1 \pi_h(a) u_h w_h^2 \geq C_2\,\|w_h\|^2_2, \quad \forall \;\, w_h \in V_h.$$

Finally, we will prove the coercivity in $H^1((0,1))$. Let $\mu \in [0,1]$, we know that for all $w \in H^1((0,1))$,

$$b_{u+u_h}(w,w) \geq \mu b_u(w,w) + C_2 \int_0^1 w^2 \geq \mu D \int_0^1 (w')^2 + (C_2 - \mu\|(\lambda c - \delta a u)^+\|_\infty) \int_0^1 w^2,$$

and for all $w_h \in V_h$,

$$b^h_{2u_h}(w_h, w_h) \geq \mu b^h_{u_h}(w_h, w_h) + C_2 \int_0^1 w_h^2 \geq \mu D \int_0^1 (w_h')^2 + (C_2 - \mu \|(\lambda c - \delta a \varepsilon_{k,h})^+\|_\infty) \int_0^1 w_h^2.$$

Now, we choose μ such that both second inequalities, in (2.50) and in (2.51), hold true, considering $\varphi = u$ and $\varphi = \varepsilon_{k,h}$, respectively. In the case $C_2 - \|(\lambda c - \delta a\varphi)^+\|_\infty \geq D$, we choose $\mu := 1$ and $C_1(\varphi) := D$. However, when $C_2 - \|(\lambda c - \delta a\varphi)^+\|_\infty < D$, we have to choose μ such that

$$\mu D = C_2 - \mu \|(\lambda c - \delta a\varphi)^+\|_\infty, \quad \text{or, equivalently} \quad \mu := \frac{C_2}{D + \|(\lambda c - \delta a\varphi)^+\|_\infty} \in (0, 1),$$

and

$$C_1(\varphi) := \frac{DC_2}{D + \|(\lambda c - \delta a\varphi)^+\|_\infty}.$$

The proof is concluded. □

In the following theorem we obtain an upper bound of the error between the iteration $u_h^{(k)}$ and u_h. This upper bound depends on $\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})$, which tends to $\mathbf{0}$ when k tends to infinity.

Theorem 2.5.2. *Under the assumptions of Theorem 2.4.4, the following estimates hold*

$$C_{1,2}\|\bar{u}_h^{(k)} - u_h\|_{H^1} \leq (3/h)^{1/2}\, |\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k)})|_2, \quad \textit{and} \quad C_2\|\bar{u}_h^{(k)} - u_h\|_2 \leq (3/h)^{1/2}\, |\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k)})|_2, \quad \forall\, k \geq 1, \tag{2.53}$$

where $C_{1,2} = C_1(\varepsilon_{k,h})$ and C_2 are those described in (2.52).

Proof. As $\mathbf{F}_h$ is convex (see the proof of Theorem 2.4.1), we deduce that

$$\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \geq J_h(\mathbf{u}_h)\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h\right). \tag{2.54}$$

In particular, from Theorem 2.4.4 we know that $\bar{\mathbf{u}}_h^{(k)} \geq \mathbf{u}_h$ and thus, from (2.54) we have

$$\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h\right)^T \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \geq \left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h\right)^T J_h(\mathbf{u}_h)\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h\right) = b^h_{2u_h}(\bar{u}_h^{(k)} - u_h, \bar{u}_h^{(k)} - u_h). \tag{2.55}$$

Now, using Cauchy-Schwarz inequality, from (2.55) it follows that

$$b^h_{2u_h}(\bar{u}_h^{(k)} - u_h, \bar{u}_h^{(k)} - u_h) \leq |\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h|_2 |\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})|_2. \tag{2.56}$$

Since

$$\|\bar{u}_h^{(k)} - u_h\|_2^2 = \int_0^1 \left(\sum_{i=0}^N (\bar{u}_{h,i}^{(k)} - u_{h,i})\varphi_i\right)^2 \geq \sum_{i=0}^N (\bar{u}_{h,i}^{(k)} - u_{h,i})^2 \int_0^1 \varphi_i^2 \geq (h/3)\, |\bar{\mathbf{u}}_h^{(k)} - \mathbf{u}_h|_2^2,$$

because $\bar{\mathbf{u}}_h^{(k)} \geq \mathbf{u}_h$, we obtain from (2.56),

$$b^h_{2u_h}(\bar{u}_h^{(k)} - u_h, \bar{u}_h^{(k)} - u_h) \leq (3/h)^{1/2}\|\bar{u}_h^{(k)} - u_h\|_2 |\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})|_2. \tag{2.57}$$

Finally, (2.53) follows from (2.51) and (2.57). The proof is complete. □

The next result provides us with the bound for the error between the solutions of (2.1) and (2.5). We use a similar proof to the famous Céa's Lemma applied to b_{u+u_h}. Usually in the literature the error $\|u - u_h\|$ is estimated by Ch, where C is a large constant. We improve this estimate by developing the error in

$$\|u - u_h\| \le \|u - \pi_h(u)\| + \|\pi_h(u) - u_h\|,$$

and using the Céa's Lemma to obtain the bound of $\|\pi_h(u) - u_h\|$, which we will obtain that is $O(h^2)$ (see [22]).

Theorem 2.5.3. *Assume $a,\ c \in H^2((0,1))$. Then, under the assumptions of Theorem 2.4.4, for each $k \ge 1$ we have that*

$$\|u_h - u\|_{H^1} \le \left(\frac{C_3}{C_{1,1}} h + \sqrt{h^2+1}\|u''\|_2\right) h \quad \text{and} \quad \|u_h - u\|_2 \le \left(\frac{C_3}{C_2} + \|u''\|_2\right) h^2,$$

where $C_{1,1} = C_1(u)$ and C_2 are given by (2.52) *and*

$$C_3 := (\|\delta a u - \lambda c\|_\infty + \delta\|a\|_\infty M_{k,h})\, \|u''\|_2 + M_{k,h}\|\lambda c''\|_2 + M_{k,h}^2\|\delta a''\|_2. \tag{2.58}$$

Proof. First of all, in order to simplify the notation, we define $E_f := f - \pi_h(f)$. Now by (2.9) and (2.10), we obtain

$$b_{u+u_h}(u_h - \pi_h(u), v_h) = b_{u+u_h}(u_h - \pi_h(u), v_h) + b_u(u, v_h) - b^h_{u_h}(u_h, v_h), \qquad \forall\ v_h \in V_h,$$

and, adding and subtracting $\lambda\int_0^1 c u_h v_h$ and $\delta\int_0^1 a u_h^2 v_h$,

$$\begin{aligned} b_{u+u_h}(u_h - \pi_h(u), v_h) = {} & D\int_0^1 E_u' v_h' - \lambda\int_0^1 c E_u v_h - \lambda\int_0^1 E_c u_h v_h \\ & + \int_{\{0,1\}} \beta E_u v_h + \delta\int_0^1 E_a u_h^2 v_h + \delta\int_0^1 a(u + u_h) E_u v_h. \end{aligned} \tag{2.59}$$

Meanwhile,

$$D\int_0^1 E_u' v_h' = 0 \quad \text{and} \quad \int_{\{0,1\}} \beta E_u v_h = 0,$$

because $\pi_h(u)(x_i) = u(x_i)$ for each $i \in \{0, 1, ..., N\}$, (2.59) can be rewritten as

$$b_{u+u_h}(u_h - \pi_h(u), v_h) = \int_0^1 (\delta a(u + u_h) - \lambda c)\, E_u v_h - \lambda\int_0^1 E_c u_h v_h + \delta\int_0^1 E_a u_h^2 v_h.$$

Moreover, by Hölder's inequality and (2.44) we deduce that

$$|b_{u+u_h}(u_h - \pi_h(u), v_h)| \le \Big[(\|\delta a u - \lambda c\|_\infty + \delta\|a\|_\infty M_{k,h})\, \|E_u\|_2 + M_{k,h}\|\lambda E_c\|_2 + M_{k,h}^2\|\delta E_a\|_2\Big]\, \|v_h\|_2.$$

Then, applying (1.5) to E_u, E_c and E_a, we obtain

$$|b_{u+u_h}(u_h - \pi_h(u), v_h)| \le C_3 h^2 \|v_h\|_2, \tag{2.60}$$

where C_3 is given by (2.58). Considering $v_h := u_h - \pi_h(u)$ in (2.60), and using the corresponding inequality in $L^2((0,1))$ given by (2.50), we have

$$C_2 \|u_h - \pi_h(u)\|_2 \le C_3 h^2. \tag{2.61}$$

Hereafter, if we bound $\|v_h\|_2$ by $\|v_h\|_{H^1}$ and once again take $v_h := u_h - \pi_h(u)$ in (2.60), using the coercivity in $H^1((0,1))$ given by (2.50), we infer that

$$C_{1,1} \|u_h - \pi_h(u)\|_{H^1} \le C_3 h^2. \tag{2.62}$$

Finally, using the triangle inequality, together with (1.5), (2.61) and (2.62), we state that

$$\|u_h - u\|_2 \le \|u_h - \pi_h(u)\|_2 + \|\pi_h(u) - u\|_2 \le \left(\frac{C_3}{C_2} + \|u''\|_2 \right) h^2,$$

and

$$\|u_h - u\|_{H^1} \le \|u_h - \pi_h(u)\|_{H^1} + \|\pi_h(u) - u\|_{H^1} \le \left(\frac{C_3}{C_{1,1}} h + \sqrt{h^2+1} \|u''\|_2 \right) h.$$

The proof is complete. □

Now, combining Theorem 2.5.2 with Theorem 2.5.3 in (2.49), yields

Theorem 2.5.4. *Suppose $a, c \in H^2((0,1))$, under the assumptions of Theorem 2.4.4, for $k \ge 1$,*

$$\|\bar{u}_h^{(k)} - u\|_{H^1} \le \frac{(3/h)^{1/2} \, |\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k)})|_2}{C_{1,2}} + \left(\frac{C_3}{C_{1,1}} h + \sqrt{h^2+1} \|u''\|_2 \right) h,$$

and

$$\|\bar{u}_h^{(k)} - u\|_2 \le \frac{(3/h)^{1/2} \, |\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k)})|_2}{C_2} + \left(\frac{C_3}{C_2} + \|u''\|_2 \right) h^2, \tag{2.63}$$

where $C_{1,1} = C_1(u)$, $C_{1,2} = C_1(\varepsilon_{k,h})$, C_2 and C_3 are given in (2.52) *and* (2.58).

2.6 Some numerical examples

In this section, we study three problems which highlight different aspects of the theory developed in this chapter. First, we fix the following exit criterion for the Newton Method

$$|\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k_{exit})})|_2 \le \xi \quad \text{and} \quad |\bar{\mathbf{u}}_h^{(k_{exit})} - \underline{\mathbf{u}}_h^{(k_{exit})}|_2 \le \xi, \tag{2.64}$$

for some $\xi \in (0,1)$. As $J_h(\bar{\mathbf{u}}_h^{(k)})$ is a tridiagonal matrix such that $\sigma_0(J_h(\bar{\mathbf{u}}_h^{(k)})) > 0$, $J_h(\bar{\mathbf{u}}_h^{(k)})$ is a non-singular M-matrix. Therefore, we can solve

$$(J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \quad \text{and} \quad (J_h(\bar{\mathbf{u}}_h^{(k)}))^{-1} \mathbf{F}_h(\underline{\mathbf{u}}_h^{(k)}), \quad k \ge 0,$$

using Thomas algorithm due to its advantages (see [46, Section 3.7] on banded systems).

2.6.1 Numerical results for the choice $D = 1$, $\lambda = 1$, $\delta = 1$, $c(x) = x + 1$, $a \equiv 1$ and $\beta_0 = \beta_1 = 0$.

In this case, we rewrite (2.1) as follows

$$\begin{cases} -u'' = (x+1)\,u - u^2, & \text{in } (0,1), \\ u'(0) = u'(1) = 0. \end{cases} \tag{2.65}$$

Then, the constants $M = 2$ and $\varepsilon = 1$ defined in (2.33) and (2.34), are respectively a supersolution and a subsolution of (2.65). Consequently, (2.65) has a unique positive solution, u, satisfying

$$1 \le u(x) \le 2, \quad \text{for all } x \in [0,1].$$

The hypotheses of Theorem 2.4.4 are satisfied if $\underline{\mathbf{u}}_h^{(0)} := \boldsymbol{\varepsilon}$ and $\bar{\mathbf{u}}_h^{(0)} := \mathbf{M}$ (see Theorem 2.4.2). Moreover, both sequences, (2.25) and (2.27), converge to $\mathbf{u}_h$ and since in this case $h_M > 1$ it also holds

$$\mathbf{0} \ll \varepsilon_{k,h}(1,1,...,1)^T \le \mathbf{u}_h \le M_{k,h}(1,1,...,1)^T, \quad \text{for all } k \ge 0 \text{ and } h > 0,$$

where $1 \le \varepsilon_{k,h} \le M_{k,h} \le 2$. Moreover, thanks to Theorem 2.5.4, we have

$$\|\bar{u}_h^{(k)} - u\|_{H^1} \le \frac{(3/h)^{1/2}\,|\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})|_2}{c_{1,2}} + \frac{4\,(1+M_{k,h})}{\varepsilon_{k,h}}h^2 + 2h\sqrt{h^2+1} := E_1, \tag{2.66}$$

and

$$\|\bar{u}_h^{(k)} - u\|_2 \le \frac{(3/h)^{1/2}\,|\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})|_2}{\varepsilon_{k,h}} + \left(\frac{2\,(1+M_{k,h})}{\varepsilon_{k,h}} + 2\right)h^2 := E_2,$$

where we have used that

$$\|u''\|_2 \le \|u''\|_\infty \le 2\|x+1-u\|_\infty \le 2$$

by (2.65). Hence

$$C_2 := \varepsilon_{k,h}, \quad C_{1,1} \ge \varepsilon_{k,h}/2, \quad C_{1,2} \ge c_{1,2} := \min\left\{1, \frac{\varepsilon_{k,h}}{3-\varepsilon_{k,h}}\right\} \quad \text{and} \quad C_3 \le 2(1+M_{k,h}).$$

In the same way, if we fix the exit criterion given in (2.64), as $1 \le \varepsilon_{k,h} \le M_{k,h} \le 2$, it follows that

$$E_1 \le 2(3/h)^{1/2}\xi + 12h^2 + 2h\sqrt{h^2+1} \quad \text{and} \quad E_2 \le (3/h)^{1/2}\xi + 8h^2.$$

Finally, we check that the derivative of $v := \bar{u}_h^{(k_{exit})}$, obtained from the simulations, near the boundary point $x = 0$, satisfies the following inequality

$$h|v'(x)| = |v(h) - v(0)| \le \int_0^h |v'(t) - u'(t)|dt + |u(h) - u(0)| \le h^{1/2}E_1 + h^2, \quad \text{for all } x \in (0,h),$$

where E_1 is given in (2.66), with $k = k_{exit}$ and $|u(h) - u(0)|$ is estimated using Taylor's Theorem. In particular, Table 2.1 shows that the numerical solutions of (2.65) satisfy the above inequality for different values of h and $\xi = 10^{-7}$.

h	k_{exit}	$\lvert\bar{u}_h^{(k_{exit})}(h) - \bar{u}_h^{(k_{exit})}(0)\rvert$	E_1	$h^{1/2}E_1 + h^2$
10^{-1}	5	$3 \cdot 10^{-3}$	$2 \cdot 10^{-1}$	$9 \cdot 10^{-2}$
10^{-2}	5	$3 \cdot 10^{-5}$	$2 \cdot 10^{-2}$	$2 \cdot 10^{-3}$
10^{-3}	5	$3 \cdot 10^{-7}$	$2 \cdot 10^{-3}$	$6 \cdot 10^{-5}$
10^{-4}	5	$3 \cdot 10^{-9}$	$2 \cdot 10^{-4}$	$2 \cdot 10^{-6}$

Table 2.1: Simulations of example 2.6.1, for $\xi = 10^{-7}$.

2.6.2 Numerical results for the choice $D = 10^{-4}$, $\lambda = 1$, $\delta = 1$, $c \equiv 1$, $a(x) = 100x + 1$ and $\beta_0 = \beta_1 = 0$.

Consider now the singular perturbation problem

$$\begin{cases} -10^{-4}\, u'' = u - (100x + 1)\, u^2 & \text{in } (0,1), \\ u'(0) = u'(1) = 0. \end{cases} \tag{2.67}$$

In an analogous way to the previous example, we deduce that the positive solution of (2.67) satisfies

$$\varepsilon := \frac{1}{101} \leq u(x) \leq 1 =: M, \quad \text{for all } \; x \in [0,1].$$

Consequently, we pick again $\underline{\mathbf{u}}_h^{(0)} := \boldsymbol{\varepsilon}$ and $\bar{\mathbf{u}}_h^{(0)} := \mathbf{M}$ in Theorem 2.4.4, but for this example

$$h_M \approx 0.0017,$$

which implies that we can only ensure the convergence of (2.25) if $h < h_M$. In Figure 2.1 we present three plots, each one for a different value of h, of an iteration of the Newton Method satisfying the same fixed exit criterion. Firstly, for $h = 0.125$ the Newton Method does not converge. Nevertheless, the Newton Method converges for $h = 0.0625$ and $h = 0.0016$. In particular, for $h = 0.0625$, $h \nleq h_M$, the iterations of the Newton Method do not satisfy (2.43), they intersect and the numerical solution oscillates (see Figure 2.2 (top)). But in the case $h = 0.0016$, $h \leq h_M$, the iterations of the Newton Method are ordered and the final numerical solution does not oscillate as it can be seen in Figure 2.2 (bottom). Figure 2.1 is similar to [46, Figure 11.3].

2.6.3 Numerical results for the choice $D = 10^{-3}$, $\lambda = 1$, $\delta = 1$, $\beta_0 = \beta_1 = 10^{-2}$, $a \equiv 1$ and $c(x) = 10 - (x - 0.5)^2$.

In this last example we consider Robin boundary conditions,

$$\begin{cases} -10^{-3}\, u'' = (10 - (x - 0.5)^2)\, u - u^2, & \text{in } (0,1), \\ -10^{-3}\, u'(0) + 10^{-2}\, u(0) = 0, \\ 10^{-3}\, u'(1) + 10^{-2}\, u(1) = 0. \end{cases} \tag{2.68}$$

We emphasize that in this case we can not use Theorem 2.4.2 to obtain a positive strict subsolution of problem (2.5), however, we will use Theorem 2.4.3. In that sense, firstly we prove that

$$\underline{u}(x) = \mu e^{-10(x-0.5)^2} \;\text{ for }\; \mu \in (0, 9.83) \tag{2.69}$$

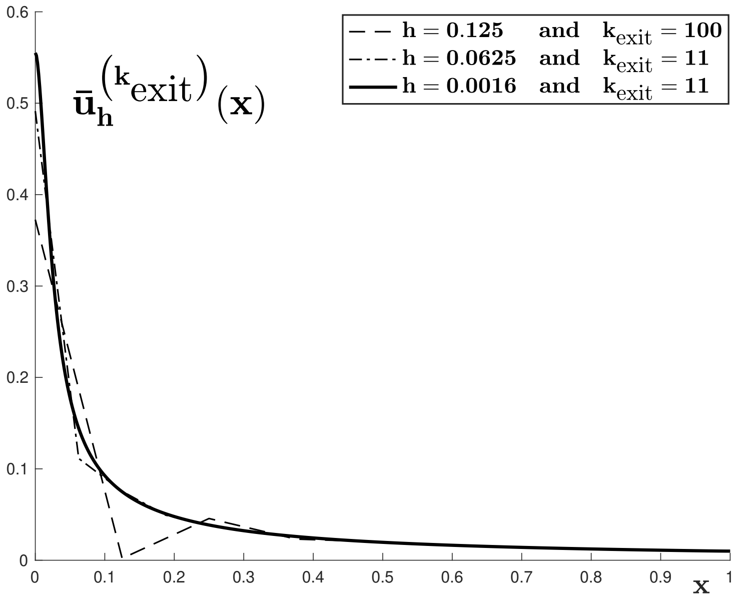

Figure 2.1: Simulations of example 2.6.2, for $\xi = 10^{-7}$ and number of iterations in Newton Method, $k_{exit} \leq 100$.

satisfies (2.35). Indeed, as $\underline{u}$ satisfies the boundary conditions in (2.68), it only remains to find $C > 0$ such that

$$-10^{-3}\underline{u}''(x) - \left(10 - (x-0.5)^2\right)\underline{u}(x) + \underline{u}^2(x) \leq -C \quad \text{for all } \; x \in (0,1). \tag{2.70}$$

Replacing $\underline{u}$ defined in (2.69) and its second derivative in the expression in the left of (2.70), we obtain that for all $x \in [0,1]$,

$$\underline{u}(x)\left(-9.98 + 0.6(x-0.5)^2 + \underline{u}(x)\right) \leq \underline{u}(x)\left(-9.83 + \underline{u}(x)\right) \leq \mu e^{-5/2}\left(-9.83 + \mu\right),$$

and thus, we obtain that

$$C = \mu e^{-5/2}\left(9.83 - \mu\right)$$

satisfies (2.70). Then, due to Theorem 2.4.3 we have that $\boldsymbol{\pi}_h(\underline{u})$ is a positive strict subsolution of problem (2.5) if $h \leq h^*$, where

$$h^* := h^*(\mu) := 3^{1/3}\left(\frac{\mu e^{-5/2}\left(9.83 - \mu\right)/2}{\left(\|\underline{u} - c\|_\infty + \|\underline{u}\|_\infty\right)\|\underline{u}''\|_2 + 2\|\underline{u}\|_\infty}\right)^{2/3}.$$

Now, using that

$$\|\underline{u}\|_\infty = \mu, \quad \|\underline{u} - c\|_\infty \leq 10 - \mu e^{-5/2} \quad \text{and} \quad \|\underline{u}''\|_\infty \leq 80\mu,$$

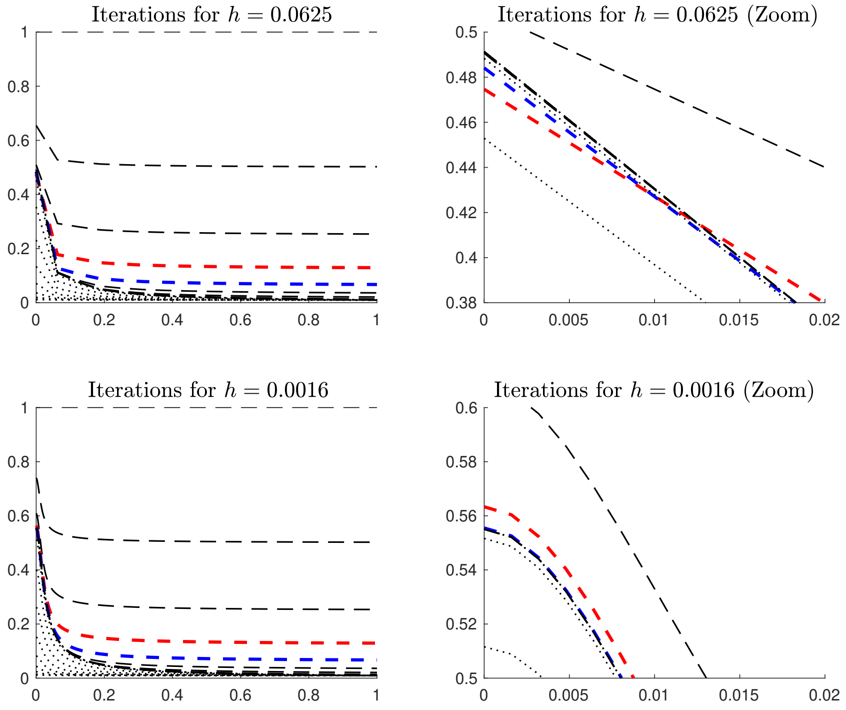

Figure 2.2: Iterations of Figure 2.1: 'dashed' - upper iterations and 'dotted' - lower iterations.

we have that

$$h^*(\mu) \geq 3^{1/3}\left(\frac{e^{-5/2}\,(9.83-\mu)\,/4}{401+40\mu\,(1-e^{-5/2})}\right)^{2/3} := h_0^*(\mu),$$

and, as $h_0^*(\mu)$ is a decreasing function for all $\mu \in [0, 9.83]$, we are interested in choosing $\mu \approx 0$ to obtain a larger range for h. In particular,

$$h \leq h_0^*(0) \approx 0.009. \tag{2.71}$$

Therefore, we start the sequence in (2.27), taking as initial guess

$$\underline{\mathbf{u}}_h^{(0)} := \left(\mu e^{-10(x_i-0.5)^2}\right)_{i=0,1,\dots,N} \quad \text{where} \;\; \mu \approx 0, \;\; \mu > 0, \tag{2.72}$$

Consequently, the hypotheses of Theorem 2.4.4 are satisfied for $\bar{\mathbf{u}}_h^{(0)} = \mathbf{M}$, for $M = 10$ (see Theorem 2.4.2). Also (2.72) is satisfied for all h verifying (2.71), because $h_M \approx 0.024$ (see (2.16)). Hence, we obtain that both sequences, (2.25) and (2.27), converge to $\mathbf{u}_h$ satisfying

$$\mathbf{0} \ll \varepsilon_{k,h}(1,1,...,1)^T \leq \mathbf{u}_h \leq M_{k,h}(1,1,...,1)^T, \quad \text{for all} \;\; k \geq 0,$$

where $0 < \mu e^{-5/2} \leq \varepsilon_{k,h} \leq M_{k,h} \leq M$.

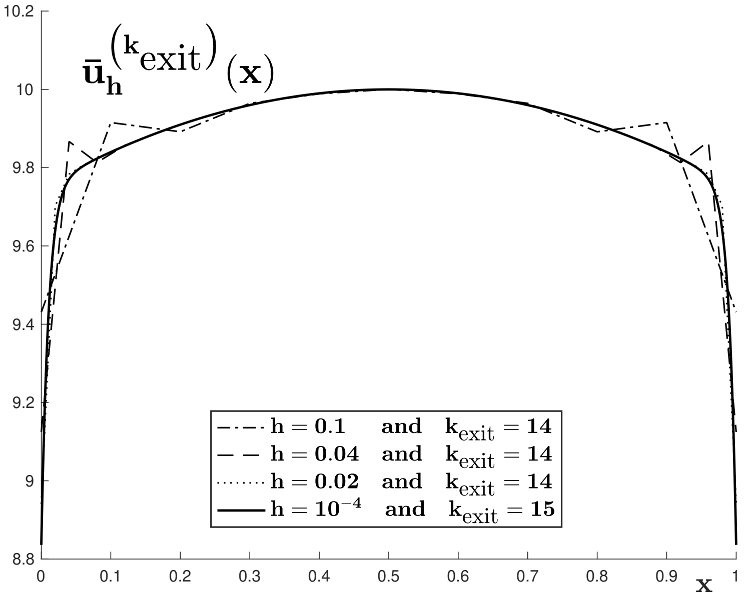

Figure 2.3: Simulations for example 2.6.3, for $\xi = 10^{-7}$ and $\mu = 0.1$.

Now, we study the error in the L^2-norm. Firstly, as (2.33) and (2.69) with $\mu = 9.83$ are a subsolution and a supersolution of (2.68) respectively, it follows that the unique positive solution of (2.68) satisfies

$$0.8 \leq u(x) \leq 10, \quad \text{for all } \ x \in [0,1].$$

Then, we have that

$$C_2 := \varepsilon_{k,h} \quad \text{and} \quad C_3 \leq (9.2 + M_{k,h})\|u''\|_2 + 2M_{k,h},$$

where the constants are defined in (2.52) and (2.58). Moreover, by proceeding as in the first example, we bound the second derivative by $\|u''\|_\infty \leq 92200$. Therefore, by (2.63) we obtain that

$$\|\bar{u}_h^{(k)} - u\|_2 \leq \frac{(3/h)^{1/2}\,|\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)})|_2}{\varepsilon_{k,h}} + \left(\frac{(9.2 + M_{k,h})92200 + 2M_{k,h}}{\varepsilon_{k,h}} + 92200\right)h^2 := E_2.$$

Table 2.2 shows E_2 for $k = k_{exit}$ and different values of h.

Note that $\varepsilon_{0,h} = \mu e^{-5/2}$ is near to zero because $\mu \approx 0$ but this is not a problem because we expect $\varepsilon_{exit,h}$ to be bigger and near to the minimum of u. In the Table 2.2 we can observe this fact. Also, the constant C_3 could have been taken smaller, if we would have had more information about u.

h	k_{exit}	$\varepsilon_{k_{exit},h}$	E_2
10^{-3}	14	8.83	0.29
10^{-4}	15	8.83	$2 \cdot 10^{-3}$
10^{-5}	15	8.83	$2 \cdot 10^{-5}$

Table 2.2: Simulations for example 2.6.3, for $\xi = 10^{-7}$ and $\mu = 0.1$.

Finally, Figure 2.3 shows some simulations of this last example, obtained for decreasing values of h. We highlight again, the effects in the simulations when the hypotheses of Theorem 2.4.4 are not satisfied. Indeed, in the two cases that $h > h_M$, Newton Method converges but the numerical solutions oscillate. In the Figure 2.4 we have done a zoom to better observe the case $h = 0.02$.

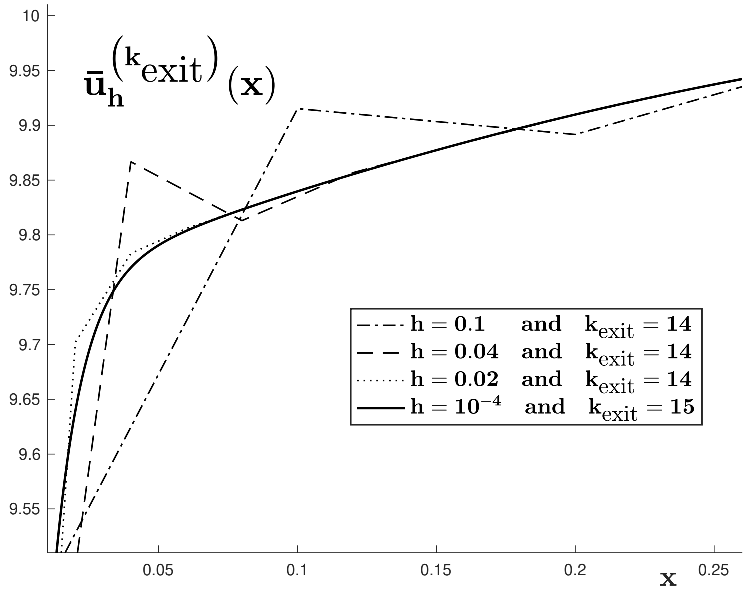

Figure 2.4: Zoom of Figure 2.3.

Chapter 3

The generalized diffusive logistic equation

3.1 Introduction

In this chapter, we apply the discrete sub- and supersolutions method to prove the convergence of the approximate solution given by the nonlinear finite element method to the positive solution of the problem (1.1) that we write here again

$$\begin{cases} -(Du' - \alpha m'(x)u)' = \lambda m(x)u - \delta a(x)u^2, & x \in (0,1), \\ Du'(0) = \alpha m'(0)u(0), & \\ Du'(1) = \alpha m'(1)u(1), & \end{cases} \tag{3.1}$$

where $D > 0$, $\alpha > 0$, $\lambda > 0$, $\delta > 0$, $m \in C^2[0,1]$ such that $m(x) > 0$ for all $x \in [0,1]$, $m' \not\equiv 0$, $m'(0) \geq 0$ and $m'(1) \leq 0$, and $a \in C[0,1]$, with $a(x) > 0$ for all $x \in [0,1]$. As we mentioned in the introduction, if there exists a supersolution, $\bar{U} \in H^2\left((0,1)\right)$, and a subsolution $\underline{U} \in H^2((0,1))$, in the sense that

$$-(D\bar{U}' - \alpha m'(x)\bar{U})' \geq \lambda m(x)\bar{U} - \delta a(x)\bar{U}^2, \quad -(D\underline{U}' - \alpha m'(x)\underline{U})' \leq \lambda m(x)\underline{U} - \delta a(x)\underline{U}^2$$

almost everywhere in $(0,1)$, and

$$\begin{aligned} D\bar{U}'(0) \leq \alpha m'(0)\bar{U}(0), \quad D\bar{U}'(1) \geq \alpha m'(1)\bar{U}(1), \\ D\underline{U}'(0) \geq \alpha m'(0)\underline{U}(0), \quad D\underline{U}'(1) \leq \alpha m'(1)\underline{U}(1), \end{aligned}$$

and satisfying $0 < \underline{U} \leq \bar{U}$, then, by [6], there exists a positive solution $u \in H^2((0,1))$ of (3.1), $u > 0$ ($u \geq 0$ and $u \neq 0$), such that $\underline{U} \leq u \leq \bar{U}$ which is the unique positive solution by [2]. Moreover, if u is the solution of (3.1), u satisfies

$$\int_0^1 (Du' - \alpha m'u)\, v' - \int_0^1 \left(\lambda mu - \delta au^2\right) v = 0, \quad \forall v \in H^1((0,1)), \tag{3.2}$$

and then, it follows that the approximation of u, in the space of finite element, denoted by $u_h \in V_h$, satisfies

$$\int_0^1 \left(Du_h' - \alpha\pi_h(m')u_h\right) v_h' - \int_0^1 \left(\lambda\pi_h(m)u_h - \delta\pi_h(a)u_h^2\right) v_h = 0, \quad \forall v_h \in V_h, \tag{3.3}$$

where π_h is defined in (1.4). Thus, the discrete sub- and supersolutions method is proposed to approximate $\mathbf{u}_h > \mathbf{0}$ satisfying

$$\mathbf{F}_h(\mathbf{u}_h) = \mathbf{0}, \tag{3.4}$$

and compare it with the positive solution u.

3.2 The discrete bilinear form

Our aim now is to calculate the nonlinear map $\mathbf{F}_h$ in (3.4). In order to do so, we introduce the approximation u_h of u:

$$u_h = \sum_{i=0}^{N} u_{h,i}\varphi_i, \quad \text{or equivalently,} \quad \mathbf{u}_h = (u_{h,0}, u_{h,1}, ..., u_{h,N})^T.$$

The weak formulations (3.2) and (3.3) can be rewritten as

$$b_u(u, v) = 0 \ \ \forall v \in H^1((0,1)) \quad \text{and} \quad b^h_{u_h}(u_h, v_h) = 0 \ \ \forall v_h \in V_h,$$

respectively, where

$$b_f(w, v) := D\int_0^1 w'v' - \alpha\int_0^1 m'wv' - \lambda\int_0^1 mwv + \delta\int_0^1 afwv$$

and

$$b^h_f(w_h, v_h) := D\int_0^1 w_h'v_h' - \alpha\int_0^1 \pi_h(m')w_hv_h' - \lambda\int_0^1 \pi_h(m)w_hv_h + \delta\int_0^1 \pi_h(a)fw_hv_h$$

for w, $v \in H^1((0,1))$, w_h, $v_h \in V_h$ and $f \in C[0,1]$. Therefore, $\mathbf{u}_h$ solves the nonlinear system

$$\mathbf{F}_h(\mathbf{z}) = \mathbf{0}, \quad \text{with } \ \mathbf{F}_h(\mathbf{z}) := (A_h + Y_h(\mathbf{z}))\,\mathbf{z}, \quad \mathbf{z} = (z_0, z_1, ..., z_N)^T,$$

where the matrices $A_h \in \mathbb{R}^{N+1\times N+1}$ and $Y_h(\mathbf{z}) \in \mathbb{R}^{N+1\times N+1}$ are defined as follows

$$A_h = \left\{b^h_0(\varphi_j, \varphi_i)\right\}_{i,j=0}^{N} \quad \text{and} \quad Y_h(\mathbf{z}) = \left\{\delta\int_0^1 \pi_h(a)z\varphi_j\varphi_i\right\}_{i,j=0}^{N}, \quad z = \sum_{i=0}^{N} z_i\varphi_i.$$

Note that i indicates the row and j indicates the column. It is easily seen that

$$A_h = \frac{D}{h}A_{h,2} - \frac{\alpha}{6}A_{h,1} - \frac{\lambda h}{12}A_{h,0}, \tag{3.5}$$

where the element of the tridiagonal matrices $A_{h,2}$, $A_{h,1}$ and $A_{h,0}$ are the following

$$\begin{array}{lll} (A_{h,2})_{00} = 1, & (A_{h,1})_{00} = -2m_0' - m_1', & (A_{h,0})_{00} = 3m_0 + m_1, \\ (A_{h,2})_{NN} = 1, & (A_{h,1})_{NN} = m_{N-1}' + 2m_N', & (A_{h,0})_{NN} = m_{N-1} + 3m_N, \end{array}$$

if $i \in \{1, 2, ..., N-1\}$,

$$(A_{h,2})_{i,i} = 2, \quad (A_{h,1})_{i,i} = m'_{i-1} - m'_{i+1} \quad \text{and} \quad (A_{h,0})_{i,i} = m_{i-1} + 6m_i + m_{i+1},$$

and finally, if $i \in \{0, 1, ..., N-1\}$,

$$\begin{aligned}(A_{h,2})_{i,i+1} = -1, \quad (A_{h,1})_{i,i+1} = -m'_i - 2m'_{i+1} \quad &\text{and} \quad (A_{h,0})_{i,i+1} = m_i + m_{i+1},\\ (A_{h,2})_{i+1,i} = -1, \quad (A_{h,1})_{i+1,i} = 2m'_i + m'_{i+1} \quad &\text{and} \quad (A_{h,0})_{i+1,i} = m_i + m_{i+1},\end{aligned}$$

where $m_i = m(x_i)$ and $m'_i = m'(x_i)$. In the same way,

$$Y_h(\mathbf{z}) = \frac{\delta h}{60} Y_{h,0}(\mathbf{z}), \tag{3.6}$$

where the elements of the symmetric and tridiagonal matrix $Y_{h,0}(\mathbf{z})$ are

$$\begin{aligned}(Y_{h,0}(\mathbf{z}))_{00} &= z_0(12a_0 + 3a_1) + z_1(3a_0 + 2a_1),\\ (Y_{h,0}(\mathbf{z}))_{NN} &= z_{N-1}(2a_{N-1} + 3a_N) + z_N(3a_{N-1} + 12a_N),\end{aligned}$$

if $i \in \{1, 2, ..., N-1\}$,

$$(Y_{h,0}(\mathbf{z}))_{ii} = z_{i-1}(2a_{i-1} + 3a_i) + z_i(3a_{i-1} + 24a_i + 3a_{i+1}) + z_{i+1}(3a_i + 2a_{i+1}),$$

and, if $i \in \{0, 1, ..., N-1\}$,

$$(Y_{h,0}(\mathbf{z}))_{i,i+1} = z_i(3a_i + 2a_{i+1}) + z_{i+1}(2a_i + 3a_{i+1}),$$

where $a_i = a(x_i)$ and $\mathbf{z} = (z_0, z_1, ..., z_N)^T$.

3.3 Discrete Sub- and Supersolutions Method

In the present section we develop the main steps of the Discrete Sub- and Supersolution Method in Chapter 2 applied to the discrete equation (3.3). The first step is to construct a suitable constant positive strict supersolution $\mathbf{M}$ satisfying

$$\mathbf{F}_h(\mathbf{M}) > 0 \quad \text{with} \quad \mathbf{M} := M(1, 1, \ldots, 1)^T > 0. \tag{3.7}$$

Let $J_h(\mathbf{z})$ denote the Jacobian of $\mathbf{F}_h$. The second step is to obtain $h_M > 0$ such that $J_h(\mathbf{z})$ evaluated at $\mathbf{z} \leq \mathbf{M}$ possesses negative principal subdiagonals for all $h < h_M$, i.e.

$$J_h(\mathbf{z}) = \begin{bmatrix} * & - & 0 & 0 & ... \\ - & * & - & 0 & ... \\ 0 & - & * & - & ... \\ 0 & 0 & - & * & ... \\ \vdots & \vdots & \vdots & \vdots & \ddots \end{bmatrix}. \tag{3.8}$$

Let $\sigma_0(J_h(\mathbf{z}))$ denote the principal eigenvalue of $J_h(\mathbf{z})$. Property (3.8) guarantees the existence of $\sigma_0(J_h(\mathbf{z}))$, which is the unique eigenvalue that admits a real positive eigenvector (see Theorem 2.3.1 for more details). Moreover, this condition enable the Discrete Maximum Principle introduced in Theorem 2.3.2.

The key point through the ongoing procedure is the choice of h_M, given $M > 0$. With this choice and the several consequences of the DMP next theorem proves the uniqueness of the positive approximate solution $\mathbf{u}_h \in [\mathbf{0}, \mathbf{M}]$ satisfying (3.4).

Theorem 3.3.1. *Let $M > 0$. If*

$$h < h_M := \frac{6D}{3\alpha\|m'\|_\infty + \|2M\delta a - \lambda m\|_\infty} \tag{3.9}$$

holds, then $J_h(z) := A_h + 2Y_h(z)$ satisfies (3.8) *for each $z \leq \boldsymbol{M} := M(1, 1, \dots, 1)$, and the positive solution $\boldsymbol{u}_h$ of* (3.4) *satisfying*

$$\boldsymbol{0} < \boldsymbol{u}_h \leq \boldsymbol{M}$$

is unique, if it exists.

Proof. Let $M > 0$ and $h < h_M$. Since the $N + 1$ components of $\mathbf{F}_h$ are defined as

$$F_{h,i}(\mathbf{z}) = b_0^h(z, \varphi_i) + \delta \int_0^1 \pi_h(a) z^2 \varphi_i, \quad \mathbf{z} = (z_0, z_1, \dots, z_N)^T, \quad z = \sum_{j=0}^N z_j \varphi_j,$$

we know that

$$\frac{\partial F_{h,i}(\mathbf{z})}{\partial z_j} = b_0^h(\varphi_j, \varphi_i) + 2\delta \int_0^1 \pi_h(a) z \varphi_i \leq b_0^h(\varphi_j, \varphi_i) + 2\delta \int_0^1 \pi_h(a) M \varphi_i = \frac{\partial F_{h,i}(\mathbf{M})}{\partial z_j}.$$

So $J_h(\mathbf{z}) = A_h + 2Y_h(\mathbf{z})$ and $J_h(\mathbf{z}) \leq J_h(\mathbf{M})$, if $\mathbf{z} \leq \mathbf{M}$. Now we will prove that $J_h(\mathbf{M})$ satisfies (3.8), and so does $J_h(\mathbf{z})$, if $h < h_M$. From (3.5) and (3.6) we obtain that

$$(J_h(\mathbf{M}))_{i,i+1} = -\frac{D}{h} + \frac{\alpha}{6}(m_i' + 2m_{i+1}') - \frac{\lambda h}{12}(m_i + m_{i+1}) + \frac{M\delta h}{6}(a_i + a_{i+1})$$

and

$$(J_h(\mathbf{M}))_{i+1,i} = -\frac{D}{h} - \frac{\alpha}{6}(2m_i' + m_{i+1}') - \frac{\lambda h}{12}(m_i + m_{i+1}) + \frac{M\delta h}{6}(a_i + a_{i+1})$$

for each $i \in \{0, 1, \dots, N-1\}$. Hence, the upper bound

$$-\frac{D}{h} + \frac{\alpha}{2}\|m'\|_\infty + \frac{h}{6}\|2M\delta a - \lambda m\|_\infty,$$

of the two above subdiagonals elements of $J_h(\mathbf{M})$ is negative if $h < h_M$. Since $\mathbf{F}_h(\mathbf{z})$ is the sum of the matrix A_h and the non-linear term $Y_h(\mathbf{z})\mathbf{z}$, the proof finishes in the same way that the proof of Theorem 2.3.6.

□

Next theorem proves that if $h < h_M$, the sequence of the consecutive iterations of the Newton Method applied to (3.4), with initial data $\mathbf{M}$, converges and it is itself a positive monotone decreasing sequence. The theorem also proves that if there exists a positive strict subsolution $\mathbf{E}$, satisfying

$$\mathbf{F}_h(\mathbf{E}) < \mathbf{0} \quad \text{and} \quad \mathbf{0} < \mathbf{E} \ll \mathbf{M} := M(1, 1, \ldots, 1), \tag{3.10}$$

the iterations of the Newton Method converge to a positive solution of (3.4). Observe that $\mathbf{M}$ can not be a solution of (3.4).

Theorem 3.3.2. *Let h_M defined in Theorem 3.3.1 and $\bar{\boldsymbol{u}}_h^{(k+1)}$ be the sequence of supersolutions of problem* (3.4) *given by the Newton Method*

$$\bar{\boldsymbol{u}}_h^{(k+1)} = \bar{\boldsymbol{u}}_h^{(k)} - (J_h(\bar{\boldsymbol{u}}_h^{(k)}))^{-1}\boldsymbol{F}_h(\bar{\boldsymbol{u}}_h^{(k)}), \quad k \geq 0, \tag{3.11}$$

where $J_h(\bar{\boldsymbol{u}}_h^{(k)}) := A_h + 2Y_h(\bar{\boldsymbol{u}}_h^{(k)})$. If $h < h_M$, $\bar{\boldsymbol{u}}_h^{(0)} := \boldsymbol{M}$,

$$\boldsymbol{M} := M(1, 1, \ldots, 1) \quad \textit{with} \quad M := \frac{\lambda\|m\|_\infty + \alpha\|m''\|_\infty}{\delta \min_{x\in[0,1]} a(x)}, \tag{3.12}$$

and $\boldsymbol{F}_h(\boldsymbol{M}) \neq \boldsymbol{0}$. Then $\bar{\boldsymbol{u}}_h^{(k+1)}$ is well defined and it converges to a solution $\boldsymbol{u}_h \in [\boldsymbol{0}, \boldsymbol{M}]$ of (3.4)*. Moreover,*

$$\boldsymbol{0} \ll \bar{\boldsymbol{u}}_h^{(k+1)} \ll \bar{\boldsymbol{u}}_h^{(k)} \ll \boldsymbol{M} \ \textit{ for all } \ k \geq 0. \tag{3.13}$$

If $\boldsymbol{E}$ satisfying (3.10) *exists. Then*

$$\boldsymbol{E} \ll \bar{\boldsymbol{u}}_h^{(k)} \ \textit{ for all } \ k \geq 0. \tag{3.14}$$

Proof. First of all, we will prove that $\mathbf{M}$ is a positive strict supersolution of (3.4), i.e. $\mathbf{M}$ satisfies (3.7). Observe that $M > 0$ because $m(x) > 0$ for each $x \in [0, 1]$ and $\mathbf{F}_h(\mathbf{M}) \neq \mathbf{0}$ by hypothesis, thus, (3.7) is satisfied if

$$F_{h,0}(\mathbf{M}) = Mh\left[\frac{\alpha}{2h}(m_0' + m_1') - \frac{\lambda}{12}(4m_0 + 2m_1) + \frac{M\delta}{60}(20a_0 + 10a_1)\right] \geq 0,$$

$$F_{h,N}(\mathbf{M}) = Mh\left[-\frac{\alpha}{2h}(m_{N-1}' + m_N') - \frac{\lambda}{12}(2m_{N-1} + 4m_N) + \frac{M\delta}{60}(10a_{N-1} + 20a_N)\right] \geq 0$$

and

$$F_{h,i}(\mathbf{M}) = Mh\left[\frac{\alpha}{2h}(m_{i+1}' - m_{i-1}') - \frac{\lambda}{12}(2m_{i-1} + 8m_i + 2m_{i+1}) + \frac{M\delta}{60}(10a_{i-1} + 40a_i + 10a_{i+1})\right] \geq 0$$

for each $i \in \{1, 2, \ldots, N-1\}$. Indeed, as

$$\delta M a_i \geq \lambda\|m\|_\infty + \alpha\|m''\|_\infty \ \text{ for each } \ i \in \{0, 1, \ldots, N\}, \quad m_0' \geq 0 \quad \text{and} \quad m_N' \leq 0,$$

it follows that

$$F_{h,0}(\mathbf{M}) \geq \frac{\alpha M h}{2}\left[\frac{m_1' - m_0'}{h} + \|m''\|_\infty\right] \geq 0, \quad F_{h,N}(\mathbf{M}) \geq \frac{\alpha M h}{2}\left[\frac{m_N' - m_{N-1}'}{h} + \|m''\|_\infty\right] \geq 0$$

and

$$F_{h,i}(\mathbf{M}) = \alpha M h\left[\frac{m'_{i+1} - m'_{i-1}}{2h} + \|m''\|_\infty\right] \geq 0 \text{ for each } i \in \{0, 1, \ldots, N\},$$

because

$$m'_{i+1} - m'_i = \int_{x_i}^{x_{i+1}} m'' \geq -h\|m''\|_\infty \text{ for each } i \in \{0, 1, \ldots, N-1\}.$$

Therefore, $\mathbf{M}$ satisfies (3.7).

Now we will prove that if $h < h_M$ and $\mathbf{F}_h(\mathbf{M}) \neq \mathbf{0}$, $\bar{\mathbf{u}}_h^{(1)}$ is well defined, it holds that

$$\mathbf{0} \ll \bar{\mathbf{u}}_h^{(1)} \ll \bar{\mathbf{u}}_h^{(0)} := \mathbf{M}.$$

According to Theorem 3.3.1, since $h < h_M$, the elements of the subdiagonals of $J_h(\mathbf{M})$ are negative. Thus, as $\mathbf{M} \gg 0$ and

$$J_h(\mathbf{M})\mathbf{M} = (A_h + 2Y_h(\mathbf{M}))\,\mathbf{M} \gg \mathbf{F}_h(\mathbf{M}) > \mathbf{0},$$

it follows from Theorem 2.3.2 that $J_h(\mathbf{M})$ is invertible and $J_h^{-1}(\mathbf{M}) \gg \mathbf{0}$. Then, $\bar{\mathbf{u}}_h^{(1)}$ is well defined and $\bar{\mathbf{u}}_h^{(1)} \ll \mathbf{M}$. Moreover, as

$$J_h(\mathbf{M})\bar{\mathbf{u}}_h^{(1)} = J_h(\mathbf{M})\left(\bar{\mathbf{u}}_h^{(1)} - \mathbf{M}\right) + J_h(\mathbf{M})\mathbf{M} \gg -\mathbf{F}_h(\mathbf{M}) + \mathbf{F}_h(\mathbf{M}) = \mathbf{0},$$

$\bar{\mathbf{u}}_h^{(1)} \gg \mathbf{0}$, and thus, (3.13) is satisfied for $k = 0$. The proof continues by induction in k. To obtain that $\mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) \gg \mathbf{0}$, since $\mathbf{F}_h$ is convex, we use a similar argument of the one used in the proof of Theorem 2.4.1. Thereafter, we repeat the process done before. The converge of $\bar{\mathbf{u}}_h^{(k)}$ to a solution $\mathbf{u}_h \in [\mathbf{0}, \mathbf{M}]$ satisfying (3.4) is a consequence of both (3.13) and $\mathbf{F}_h \in C^\infty(\mathbb{R}^{N+1})$.

Lastly, we prove (3.14) by induction. Since $\mathbf{E} \ll \mathbf{M}$, (3.14) is true for $k = 0$. As a consequence of $Y_h(\mathbf{z})\mathbf{v} = Y_h(\mathbf{v})\mathbf{z}$ for all $\mathbf{z}, \mathbf{v} \in \mathbb{R}^{N+1}$, we obtain that

$$\begin{aligned} J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) &= A_h\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) + 2Y_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) \\ &= \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) - \mathbf{F}_h(\mathbf{E}) + Y_h(\mathbf{E} - \bar{\mathbf{u}}_h^{(k)})\mathbf{E} + Y_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) \\ &= \mathbf{F}_h(\bar{\mathbf{u}}_h^{(k)}) - \mathbf{F}_h(\mathbf{E}) + Y_h(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right), \end{aligned}$$

Now by (3.11), since $\mathbf{F}_h(\mathbf{E}) < 0$ and $\mathbf{E} \ll \bar{\mathbf{u}}_h^{(k)}$ by hypothesis of induction, we deduce that

$$\begin{aligned} J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \mathbf{E}\right) &= J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k+1)} - \bar{\mathbf{u}}_h^{(k)}\right) + J_h(\bar{\mathbf{u}}_h^{(k)})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) \\ &= -\mathbf{F}_h(\mathbf{E}) + Y_h(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E})\left(\bar{\mathbf{u}}_h^{(k)} - \mathbf{E}\right) \gg 0. \end{aligned}$$

Thus, using that $\left(J_h(\bar{\mathbf{u}}_h^{(k)})\right)^{-1} \gg 0$, we get $\mathbf{E} \ll \bar{\mathbf{u}}_h^{(k+1)}$. □

The following theorem provides us with the positive strict subsolution of (3.4) verifying (3.10). The key point is to use the subsolution

$$\underline{u} := \varepsilon e^{\alpha m(x)/D}$$

of the problem (3.1), where $\varepsilon > 0$ is sufficiently small. Hence, the interpolation in the space of finite elements of $\underline{u}$ is the desire subsolution of (3.4). This idea was already used in Theorem 2.4.3. So, due to the presence of the drift term in the equation (3.1), we have to adapt the argument of the proof of Theorem 2.4.3.

Theorem 3.3.3. *Suppose* $m \in H^3((0,1))$. *If*

$$h \leq h^* := \left(\frac{\lambda \min_{x\in[0,1]} m(x)}{4\sqrt{2}\alpha \|e^{-\alpha m/D}\|_\infty \left(\|m'\|_\infty \| \left(e^{\alpha m/D}\right)'' \|_2 + \|e^{\alpha m/D}\|_\infty \|m'''\|_2 \right)} \right)^2 .$$

Then

$$\boldsymbol{E} := \varepsilon \left(e^{\alpha m(x_0)/D}, e^{\alpha m(x_1)/D}, \ldots, e^{\alpha m(x_N)/D} \right)$$

where

$$0 < \varepsilon < \min\left\{ \frac{M}{\|e^{\alpha m/D}\|_\infty}, \ \frac{\lambda \min_{x\in[0,1]} m(x)}{2\delta \|a\|_\infty \|e^{\alpha m/D}\|_\infty} \right\} \tag{3.15}$$

satisfies (3.10) *for* $M > 0$.

Proof. As

$$0 < \varepsilon e^{\alpha m(x_i)/D} < \frac{M e^{\alpha m(x_i)/D}}{\|e^{\alpha m/D}\|_\infty} \leq M \quad \forall i \in \{0, 1, \ldots, N\}$$

by (3.15), it follows that $\mathbf{0} < \mathbf{E} \ll \mathbf{M} := M(1, 1, \ldots, 1)$, for $M > 0$. Now will prove that

$$F_{h,i}(\mathbf{E}) = b_E^h(E, \varphi_i) < 0 \ \text{ for each } \ i \in \{0, 1, \ldots, N\}, \tag{3.16}$$

where $E = \pi_h(\underline{u})$ and $\underline{u} := \varepsilon e^{\alpha m(x)/D}$. Indeed, as

$$\int_0^1 \left(D\underline{u}' - \alpha m' \underline{u} \right) \varphi_i' = 0 \quad \text{and} \quad \int_0^1 \left(E - \underline{u} \right)' \varphi_i' = 0,$$

we have that

$$b_E^h(E, \varphi_i) = -\alpha \int_0^1 m'(E - \underline{u})\varphi_i' - \alpha \int_0^1 (\pi_h(m') - m') E \varphi_i' + \int_0^1 (\delta \pi_h(a) E - \lambda \pi_h(m)) E \varphi_i. \tag{3.17}$$

Moreover, by (3.15), we know that

$$g(\varepsilon) := \delta \|a\|_\infty \|e^{\alpha m/D}\|_\infty \, \varepsilon - \lambda \min_{x\in[0,1]} m(x) < 2\delta \|a\|_\infty \|e^{\alpha m/D}\|_\infty \, \varepsilon - \lambda \min_{x\in[0,1]} m(x) < 0.$$

Thus, it is possible to find an upper bound for the third term of (3.17) in the following way

$$\int_0^1 (\delta \pi_h(a) E - \lambda \pi_h(m)) E \varphi_i \leq \left(\delta \|a\|_\infty \|\underline{u}\|_\infty - \lambda \min_{x\in[0,1]} m(x) \right) \int_0^1 E \varphi_i \leq \frac{g(\varepsilon)\varepsilon h}{2\|e^{-\alpha m/D}\|_\infty}.$$

Consequently, using Hölder inequality and (1.5), we deduce that

$$b_E^h(E, \varphi_i) \leq \alpha \left(\|m'\|_\infty \|\underline{u}''\|_2 + \|\underline{u}\|_\infty \|m'''\|_2 \right) \|\varphi_i'\|_2 h^2 + \frac{g(\varepsilon)\varepsilon h}{2\|e^{-\alpha m/D}\|_\infty}.$$

Finally, as $\|\varphi_i'\|_2 \leq \sqrt{2/h}$, (3.16) is satisfied if

$$g(\varepsilon) < -2\sqrt{2}\alpha \|e^{-\alpha m/D}\|_\infty \left(\|m'\|_\infty \| \left(e^{\alpha m/D}\right)'' \|_2 + \|e^{\alpha m/D}\|_\infty \|m'''\|_2 \right) h^{1/2},$$

or equivalently, if

$$\varepsilon < \frac{\lambda \min_{x\in[0,1]} m(x) - 2\sqrt{2}\alpha ||e^{-\alpha m/D}||_\infty \left(||m'||_\infty || \left(e^{\alpha m/D}\right)'' ||_2 + ||e^{\alpha m/D}||_\infty ||m'''||_2 \right) h^{1/2}}{\delta ||a||_\infty ||e^{\alpha m/D}||_\infty},$$

which is true because $h \le h^*$ and

$$\varepsilon < \frac{\lambda \min_{x\in[0,1]} m(x)}{2\delta ||a||_\infty ||e^{\alpha m/D}||_\infty}.$$

□

3.4 Errors

In this section, we obtain the upper bounds of the $L^2((0,1))$ norm of the difference between the positive solution of (3.1) and the positive solution of (3.3) and its first derivatives, respectively. As the existence of a positive solution u_h of (3.3) is guaranteed by Theorem 3.3.2, for h is sufficiently small and Theorem 3.3.3 ensures the existence of $\mathbf{E} = (E_0, E_1, \dots, E_N)$ satisfying (3.10) for $M > 0$, if the hypothesis of Theorem 3.3.3 are satisfied and ε satisfies (3.15), we can finally bound u_h as follows

$$0 < \varepsilon e^{\alpha \min_{x\in[0,1]} m(x)/D} \le u_h(x) \le M := \frac{\lambda ||m||_\infty + \alpha ||m''||_\infty}{\delta \min_{x\in[0,1]} a(x)} \quad \forall x \in [0,1]. \tag{3.18}$$

Thanks to (3.18), the constants γ_0, γ_1 and γ_2 in the next theorem are independent of h if the hypothesis of the Theorems 3.3.2 and 3.3.3 hold true.

Theorem 3.4.1. *Assume that $a \in H^2((0,1))$ and $m \in H^3((0,1))$. If u and u_h are the positive solutions of* (3.1) *and* (3.3)*, respectively and $u_h(x) > 0$ for each $x \in [0,1]$. Then*

$$\gamma_0 ||u - u_h||_2^2 \le \gamma_2 h \quad \text{and} \quad \gamma_1 ||\,(u-u_h)'\,||_2^2 \le \gamma_2 h$$

where

$$\gamma_0 := \frac{\delta \min_{x\in[0,1]} (a(x) u_h(x))}{||e^{\alpha m/D}||_\infty}, \quad \gamma_1 = \min\left\{ \frac{\delta \min_{x\in[0,1]} (a(x)u_h(x))}{||g^+||_\infty}, 1 \right\} \frac{D}{||e^{\alpha m/D}||_\infty},$$

$g := \lambda m - \alpha m'' - \delta a u$, g^+ is the positive part of g, $\gamma_1 = D/||e^{\alpha m/D}||_\infty$ if $g^+ \equiv 0$, and $\gamma_2 := \sum_{i=1}^4 \gamma_{2,i}$ with

$$\begin{aligned}
\gamma_{2,1} &:= 2\alpha ||e^{-\alpha m/D}||_\infty \left(||u||_\infty + ||u_h||_\infty\right) ||u_h||_\infty ||m'''||_2,\\
\gamma_{2,2} &:= \alpha \left(||m''||_\infty ||u_h||_\infty + ||m'||_\infty (\tilde{\gamma}/D)^{1/2} \right) \left[|| \left(e^{-\alpha m/D} u\right)'' ||_2 h + ||u_h||_\infty || \left(e^{-\alpha m/D}\right)' ||_\infty \right],\\
&\quad \tilde{\gamma} := \alpha ||m'||_\infty ||u_h||_\infty^2 + \frac{\alpha}{2} \left(||m'''||_2 h + ||m''||_\infty \right) ||u_h||_\infty^2 + \lambda ||m||_\infty ||u_h||_\infty^2 + \delta ||a||_\infty ||u_h||_\infty^3,\\
\gamma_{2,3} &:= \left(\delta ||a''||_2 ||u_h||_\infty + \lambda ||m''||_2\right) ||e^{-\alpha m/D}||_\infty \left(||u||_\infty + ||u_h||_\infty\right) ||u_h||_\infty h,\\
\gamma_{2,4} &:= \left(\delta ||a||_\infty ||u_h||_\infty^2 + \lambda ||m||_\infty ||u_h||_\infty \right) \left[|| \left(e^{-\alpha m/D} u\right)'' ||_2 h + ||u_h||_\infty || \left(e^{-\alpha m/D}\right)' ||_\infty \right].
\end{aligned}$$

Consequently,

$$||u - u_h||_\infty \le \left(\sqrt{\frac{\gamma_2}{\gamma_0}} + \sqrt{\frac{\gamma_2}{\gamma_1}} \right) h^{1/2}.$$

Proof. Firstly, we will prove that

$$b_u(e^{\alpha m/D}v, v) \geq 0 \quad \text{for all } \; v \in H^1((0,1)). \tag{3.19}$$

Indeed, deriving and simplifying, we obtain that

$$b_u(e^{\alpha m/D}v, v) = c_u(v, v)$$

where c_u is the symmetric bilinear form defined as

$$c_u(v, v) = \int_0^1 De^{\alpha m/D}(v')^2 + \int_0^1 (\delta au - \lambda m)\, e^{\alpha m/D}v^2.$$

Now, as u is a positive solution of (3.1), we have that $e^{-\alpha m/D}u$ is a positive eigenfunction associated with the eigenvalue 0 of the eigenvalue problem

$$\begin{cases} -\left(De^{\alpha m/D}\varphi'\right)' + (au - \lambda m)\, e^{\alpha m/D}\varphi = \sigma\varphi, & \text{in } \; (0,1), \\ \partial_\nu\varphi := De^{\alpha m/D}\varphi' n = 0, & \text{on } \; \{0,1\}, \end{cases}$$

where $n(0) = -1$ and $n(1) = 1$. Moreover, it follows from [37, Th. 7.7 and Th. 7.8] that 0 is the corresponding smallest eigenvalue. Therefore, as c_u is the symmetric bilinear form associated with

$$\left(-\left(De^{\alpha m/D}\cdot'\right)' + (au - \lambda m)\, e^{\alpha m/D}, \; \partial_\nu, \; [0,1]\right),$$

we conclude (3.19), due to the Rayleigh formula applied to c_u.

Secondly, we will obtain that

$$\gamma_0\|u - u_h\|_2^2 \leq b_{u+u_h}(u - u_h, e^{-\alpha m/D}(u - u_h)) \tag{3.20}$$

and

$$\gamma_1\| \, (u - u_h)' \, \|_2^2 \leq b_{u+u_h}(u - u_h, e^{-\alpha m/D}(u - u_h)). \tag{3.21}$$

Indeed, applying (3.19) with $v := e^{-\alpha m/D}(u - u_h)$, it becomes apparent that

$$b_{u+u_h}(u - u_h, v) \geq \delta \int_0^1 au_h e^{-\alpha m/D}(u - u_h)^2 \geq \gamma_0\|u - u_h\|_2^2$$

and (3.20) is proved. On the other hand, from (3.19) it follows that

$$b_{u+u_h}(u - u_h, v) \geq \mu b_u(u - u_h, v) + \delta \int_0^1 au_h e^{-\alpha m/D}(u - u_h)^2 \tag{3.22}$$

for some $\mu \in [0,1]$. Moreover, integrating by parts, we obtain that

$$-2\alpha \int_0^1 m' e^{-\alpha m/D}(u - u_h)(u - u_h)' \geq \alpha \int_0^1 \left(m' e^{-\alpha m/D}\right)' (u - u_h)^2 \tag{3.23}$$

because $m'(0) \geq 0$ and $m'(1) \leq 0$. Thus, using (3.22) and (3.23), it follows that

$$b_{u+u_h}(u-u_h, v) \geq \frac{\mu D}{\|e^{\alpha m/D}\|_\infty}\|(u-u_h)'\|_2^2 + \int_0^1 [\mu\,(\alpha m'' - \lambda m + \delta a u) + \delta a u_h]\, e^{-\alpha m/D}(u-u_h)^2.$$

Now, (3.21) is satisfied if

$$\mu := \min\left\{\frac{\delta \min_{x\in[0,1]}(a(x)u_h(x))}{\|\,(\lambda m - \alpha m'' - \delta a u)^+\,\|_\infty}, 1\right\},$$

$(\lambda m - \alpha m'' - \delta a u)^+$ being the positive part of $\lambda m - \alpha m'' - \delta a u$, or $\mu := 1$ in the case that the denominator is equal to zero.

Thridly, we will prove that

$$b_{u+u_h}(u-u_h, e^{-\alpha m/D}(u-u_h)) \leq \gamma_2 h. \tag{3.24}$$

It suffices to show that

$$-b_u(u_h, v) + \delta \int_0^1 a u_h e^{-\alpha m/D}(u-u_h)^2 \leq \gamma_2 h, \quad \text{with} \quad v := e^{-\alpha m/D}(u-u_h),$$

because u satisfies (3.2). Moreover, by (3.3) we have that

$$b^h_{u_h}(u_h, v_h) = 0, \quad \text{with} \quad v_h := \pi_h(v).$$

Thus, to show that (3.24) is verified is equivalent to prove that

$$b^h_{u_h}(u_h, v_h) - b_u(u_h, v) + \delta \int_0^1 a u_h e^{-\alpha m/D}(u-u_h)^2 \leq \gamma_2 h. \tag{3.25}$$

Next, we will prove (3.25). Define $E_f := \pi_h(f) - f$. As $v_h(x_i) = v(x_i)$ for each $i \in \{0, 1, \ldots, N\}$, it follows that

$$\begin{aligned}
& b^h_{u_h}(u_h, v_h) - b_u(u_h, v) + \delta \int_0^1 a u_h e^{-\alpha m/D}(u-u_h)^2 \\
&= -\alpha \int_0^1 \left(\pi_h(m')v_h' - m'v'\right) u_h - \lambda \int_0^1 (\pi_h(m)v_h - mv)\, u_h + \delta \int_0^1 (\pi_h(a)v_h - av)\, u_h^2 \\
&= -\alpha \int_0^1 E_{m'} v_h' u_h - \alpha \int_0^1 m' u_h E_v' + \int_0^1 (\delta E_a u_h - \lambda E_m)\, v_h u_h + \int_0^1 (\delta a u_h - \lambda m)\, E_v u_h
\end{aligned} \tag{3.26}$$

where we have added and subtracted the terms $\alpha \int_0^1 m' v_h' u_h$, $\lambda \int_0^1 m v_h u_h$ and $\delta \int_0^1 a v_h u_h^2$. Now, by (1.5), due to

$$\|v_h'\|_\infty \leq 2\|v\|_\infty / h$$

and applying integration by parts in the second term, to find the corresponding upper bounds, we consider separately each in (3.26):

- $-\alpha \int_0^1 E_{m'} v_h' u_h \leq \alpha \|v_h'\|_\infty \|u_h\|_\infty \|E_{m'}\|_2 \leq 2\alpha \|e^{-\alpha m/D}\|_\infty (\|u\|_\infty + \|u_h\|_\infty)\|u_h\|_\infty \|m'''\|_2 h.$

- $-\alpha \int_0^1 m' u_h E_v' = \alpha \int_0^1 (m' u_h)' E_v \leq \alpha \left(\|m''\|_\infty \|u_h\|_\infty + \|m'\|_\infty \|u_h'\|_2 \right) \|E_v\|_2.$

- $\int_0^1 (\delta E_a u_h - \lambda E_m) v_h u_h \leq (\delta \|E_a\|_2 \|u_h\|_\infty + \lambda \|E_m\|_2) \|v\|_\infty \|u_h\|_\infty$

  $\leq (\delta \|a''\|_2 \|u_h\|_\infty + \lambda \|m''\|_2) \|e^{-\alpha m/D}\|_\infty (\|u\|_\infty + \|u_h\|_\infty) \|u_h\|_\infty h^2.$

- $\int_0^1 (\delta a u_h - \lambda m) E_v u_h \leq \left(\delta \|a\|_\infty \|u_h\|_\infty^2 + \lambda \|m\|_\infty \|u_h\|_\infty \right) \|E_v\|_2.$

Moreover, from Proposition 1.2.1 and (1.5), we deduce that

$$\begin{aligned} \|E_v\|_2 &\leq \|\pi_h \left(e^{-\alpha m/D} u \right) - e^{-\alpha m/D} u\|_2 + \|\pi_h \left(e^{-\alpha m/D} u_h \right) - e^{-\alpha m/D} u_h\|_\infty \\ &\leq \left[\| \left(e^{-\alpha m/D} u \right)'' \|_2 h + \|u_h\|_\infty \| \left(e^{-\alpha m/D} \right)' \|_\infty \right] h. \end{aligned}$$

Consequently, to obtain (3.25) it is enough to find an upper bound of $\|u_h'\|_2$. For this goal we use that, due to (3.3), u_h satisfies

$$D \int_0^1 \left(u_h' \right)^2 = \alpha \int_0^1 \pi_h(m') u_h u_h' + \int_0^1 \left(\lambda \pi_h(m) u_h - \delta \pi_h(a) u_h^2 \right) u_h.$$

As

$$2 \int_0^1 \pi_h(m') u_h u_h' = m' u_h^2 \Big|_0^1 - \int_0^1 (\pi_h(m'))' u_h^2 = m' u_h^2 \Big|_0^1 - \int_0^1 (\pi_h(m') - m')' u_h^2 - \int_0^1 m'' u_h^2,$$

applying (1.5), we conclude that

$$D \int_0^1 (u_h')^2 \leq \alpha \|m'\|_\infty \|u_h\|_\infty^2 + \frac{\alpha}{2} (\|m'''\|_2 h + \|m''\|_\infty) \|u_h\|_\infty^2 + \lambda \|m\|_\infty \|u_h\|_\infty^2 + \delta \|a\|_\infty \|u_h\|_\infty^3.$$

Finally, the proof finishes as a consequence of the inequality

$$\|u - u_h\|_\infty \leq \|u - u_h\|_2 + \|(u - u_h)'\|_2.$$

□

We now discuss error estimates for the iterations

$$\bar{\mathbf{u}}_h^{(k)} = \left(\bar{u}_{h,0}^{(k)}, \bar{u}_{h,1}^{(k)}, \ldots, \bar{u}_{h,N}^{(k)} \right), \qquad \bar{u}_h^{(k)} = \sum_{i=0}^{N} \bar{u}_{h,i}^{(k)} \varphi_i,$$

of the Newton Method and the positive solution u of (3.1), in terms of $\mathbf{F}_h \left(\bar{\mathbf{u}}_h^{(k)} \right)$. We will denote by $| \cdot |_1$ the 1-norm in $\mathbb{R}^{N+1}$.

Theorem 3.4.2. *Assume that $a \in H^2((0,1))$ and $m \in H^3((0,1))$. Under the assumptions of Theorem 3.3.2, the following estimates hold for each $k \geq 0$,*

$$\gamma_0 \|u - \bar{u}_h^{(k)}\|_2^2 \leq \gamma_2 h + \gamma_3 \left| \boldsymbol{F}_h \left(\bar{\boldsymbol{u}}_h^{(k)} \right) \right|_1 \quad \textit{and} \quad \gamma_1 \| \left(u - \bar{u}_h^{(k)} \right)' \|_2^2 \leq \gamma_2 h + \gamma_3 \left| \boldsymbol{F}_h \left(\bar{\boldsymbol{u}}_h^{(k)} \right) \right|_1,$$

where

$$\gamma_0 := \frac{\delta \min_{x\in[0,1]} \left(a(x) \bar{u}_h^{(k)}(x) \right)}{\|e^{\alpha m/D}\|_\infty}, \quad \gamma_1 = \min \left\{ \frac{\delta \min_{x\in[0,1]} \left(a(x) \bar{u}_h^{(k)}(x) \right)}{\|g^+\|_\infty}, 1 \right\} \frac{D}{\|e^{\alpha m/D}\|_\infty}$$

$g := \lambda m - \alpha m'' - \delta a u$, g^+ is the positive part of g , $\gamma_1 = D/\|e^{\alpha m/D}\|_\infty$ if $g^+ \equiv 0$, $\gamma_2 := \sum_{i=1}^4 \gamma_{2,i}$ with

$$\begin{aligned}
\gamma_{2,1} :=& 2\alpha \|e^{-\alpha m/D}\|_\infty \left(\|u\|_\infty + \|\bar{u}_h^{(k)}\|_\infty \right) \|\bar{u}_h^{(k)}\|_\infty \|m'''\|_2, \\
\gamma_{2,2} :=& \alpha \left(\|m''\|_\infty \|\bar{u}_h^{(k)}\|_\infty + \|m'\|_\infty (\hat{\gamma}/D)^{1/2} \right) \left[\| \left(e^{-\alpha m/D} u \right)'' \|_2 h + \|\bar{u}_h^{(k)}\|_\infty \| \left(e^{-\alpha m/D} \right)' \|_\infty \right], \\
& \hat{\gamma} := \|\bar{u}_h^{(k)}\|_\infty \left| \boldsymbol{F}_h \left(\bar{\boldsymbol{u}}^{(k)} \right) \right|_1 + \tilde{\gamma}, \\
& \tilde{\gamma} := \alpha \|m'\|_\infty \|\bar{u}_h^{(k)}\|_\infty^2 + \frac{\alpha}{2} \left(\|m'''\|_2 h + \|m''\|_\infty \right) \|\bar{u}_h^{(k)}\|_\infty^2 + \lambda \|m\|_\infty \|\bar{u}_h^{(k)}\|_\infty^2 + \delta \|a\|_\infty \|\bar{u}_h^{(k)}\|_\infty^3, \\
\gamma_{2,3} :=& \left(\delta \|a''\|_2 \|\bar{u}_h^{(k)}\|_\infty + \lambda \|m''\|_2 \right) \|e^{-\alpha m/D}\|_\infty \left(\|u\|_\infty + \|\bar{u}_h^{(k)}\|_\infty \right) \|\bar{u}_h^{(k)}\|_\infty h, \\
\gamma_{2,4} :=& \left(\delta \|a\|_\infty \|\bar{u}_h^{(k)}\|_\infty^2 + \lambda \|m\|_\infty \|\bar{u}_h^{(k)}\|_\infty \right) \left[\| \left(e^{-\alpha m/D} u \right)'' \|_2 h + \|\bar{u}_h^{(k)}\|_\infty \| \left(e^{-\alpha m/D} \right)' \|_\infty \right].
\end{aligned}$$

and

$$\gamma_3 := \left(\|u\|_\infty + \|\bar{u}_h^{(k)}\|_\infty \right) \|e^{-\alpha m/D}\|_\infty.$$

Consequently,

$$\|u - \bar{u}_h^{(k)}\|_\infty \leq \left(\frac{\gamma_2 h + \gamma_3 \left| \boldsymbol{F}_h \left(\bar{\boldsymbol{u}}_h^{(k)} \right) \right|_1}{\gamma_0} \right)^{1/2} + \left(\frac{\gamma_2 h + \gamma_3 \left| \boldsymbol{F}_h \left(\bar{\boldsymbol{u}}_h^{(k)} \right) \right|_1}{\gamma_1} \right)^{1/2}.$$

Proof. Let $k \geq 0$. In an analogous way to the proof of (3.20) and (3.21), we have that

$$\gamma_0 \|w\|_2^2 \leq b_{u+\bar{u}_h^{(k)}}(w, e^{-\alpha m/D} w) \quad \text{and} \quad \gamma_1 \|w'\|_2^2 \leq b_{u+\bar{u}_h^{(k)}}(w, e^{-\alpha m/D} w), \quad w := u - \bar{u}_h^{(k)}.$$

Now we will prove that

$$b_{u+\bar{u}_h^{(k)}}(w, v) \leq \gamma_2 h + \gamma_3 \left| \mathbf{F}_h \left(\bar{\mathbf{u}}_h^{(k)} \right) \right|_1 \quad \text{where} \quad v := e^{-\alpha m/D}(u - \bar{u}_h^{(k)}).$$

Considering $v_h := \pi_h(v)$, it suffices to show that

$$b^h_{\bar{u}_h^{(k)}}(\bar{u}_h^{(k)}, v_h) - b_u(\bar{u}_h^{(k)}, v) + \delta \int_0^1 a \bar{u}_h^{(k)} e^{-\alpha m/D} w^2 \leq \gamma_2 h, \tag{3.27}$$

because u satisfies (3.2) and

$$-b^h_{\bar{u}_h^{(k)}}(\bar{u}_h^{(k)}, v_h) \leq \left| \mathbf{v}_h^T \mathbf{F}_h (\bar{\mathbf{u}}_h^{(k)}) \right| \leq \left(\|u\|_\infty + \|\bar{u}_h^{(k)}\|_\infty \right) \|e^{-\alpha m/D}\|_\infty \left| \mathbf{F}_h \left(\bar{\mathbf{u}}_h^{(k)} \right) \right|_1.$$

Therefore, it remains to prove (3.27). As in the proof of Theorem 3.4.1, we know that

$$\begin{aligned}
&b^h_{\bar{u}_h^{(k)}}(\bar{u}_h^{(k)}, v_h) - b_u(\bar{u}_h^{(k)}, v) + \delta \int_0^1 a\bar{u}_h^{(k)} e^{-\alpha m/D} w^2 \\
&= -\alpha \int_0^1 E_{m'} v_h' \bar{u}_h^{(k)} - \alpha \int_0^1 m' \bar{u}_h^{(k)} E_v' + \int_0^1 \left(\delta E_a \bar{u}_h^{(k)} - \lambda E_m\right) v_h \bar{u}_h^{(k)} + \int_0^1 \left(\delta a \bar{u}_h^{(k)} - \lambda m\right) E_v \bar{u}_h^{(k)} \\
&\leq \gamma_{2,1} h + \alpha \left(\|m''\|_\infty \|\bar{u}_h^{(k)}\|_\infty + \|m'\|_\infty \|\bar{u}_h'^{(k)}\|_2\right) \left[\| \left(e^{-\alpha m/D} u\right)'' \|_2 h + \|\bar{u}_h^{(k)}\|_\infty \| \left(e^{-\alpha m/D}\right)' \|_\infty\right] h + \gamma_{2,3} h + \gamma_{2,4} h.
\end{aligned}$$

Then, it is necessary to find an upper bound for $\|\bar{u}_h'^{(k)}\|_2$. Indeed, denoting $z := \bar{u}_h^{(k)}$, we obtain that

$$D \int_0^1 (z')^2 = b^h_z(z, z) + \alpha \int_0^1 \pi_h(m') z z' + \lambda \int_0^1 \pi_h(m) z^2 - \delta \int_0^1 \pi_h(a) z^3. \tag{3.28}$$

Consequently, as

$$2 \int_0^1 \pi_h(m') z z' = m' z^2 \Big|_0^1 - \int_0^1 (\pi_h(m'))' z^2 = m' z^2 \Big|_0^1 - \int_0^1 (\pi_h(m') - m')' z^2 - \int_0^1 m'' z^2,$$

and

$$b^h_z(z, z) = \mathbf{z}^T \mathbf{F}_h(\mathbf{z}) \leq \|\bar{u}_h^{(k)}\|_\infty \, |\mathbf{F}_h(\mathbf{z})|_1$$

we deduce from (3.28) and (1.5) that

$$D\|z'\|_2^2 \leq \|\bar{u}_h^{(k)}\|_\infty \, |\mathbf{F}_h(\mathbf{z})|_1 + \tilde{\gamma}.$$

The proof finishes using that

$$\|u - \bar{u}_h^{(k)}\|_\infty \leq \|u - \bar{u}_h^{(k)}\|_2 + \| \left(u - \bar{u}_h^{(k)}\right)' \|_2.$$

□

3.5 Simulations

In this section we present the results given by Theorem 3.3.1, Theorem 3.4.1 and Theorem 3.4.2 applied to a first problem and Theorem 3.3.1 and Theorem 3.3.2 applied to a second different problem.

3.5.1 Numerical results for $m(x) = \cos(x)$, $a(x) = 1$ and $D = \alpha = \lambda = \delta = 1$.

We will obtain the a priori and a posteriori estimations of the error for the following problem

$$\begin{cases} -(u' + \sin(x)u)' = \cos(x)u - u^2, & x \in (0, 1), \\ u'(0) = 0, \\ u'(1) = -\sin(1)u(1). \end{cases} \tag{3.29}$$

Firstly, using the existence of the supersolution $\bar{U}$ and the subsolution $\underline{U}$ of (3.29),

$$\bar{U} := 2 \quad \text{and} \quad \underline{U} = \cos(1)e^{\cos(x)-\cos(1)},$$

satisfying $0 < \underline{U} \leq \bar{U}$, we ensure the existence of a unique positive solution u of (3.29). Moreover, u is bounded as follows

$$0.5 \leq \cos(1)e^{\cos(x)-\cos(1)} \leq u(x) \leq 2 \quad \text{for all} \quad x \in [0,1]. \tag{3.30}$$

Hence, from Theorem 3.3.3, if $h \leq 0.0012 < h^*$, it is true that

$$\mathbf{E} := 0.0993\left(e^{\cos(x_0)}, e^{\cos(x_1)}, \ldots, e^{\cos(x_N)}\right)$$

satisfies (3.10). Now, as $h_M > 1/2$, the hypothesis of the Theorem 3.3.2 are satisfied choosing

$$\underline{\mathbf{u}}_h^{(0)} := \mathbf{E} \qquad \bar{\mathbf{u}}_h^{(0)} := M(1,1,\ldots,1), \quad M := 2, \quad \text{and} \quad h \leq 0.0012.$$

Thus, the sequence in (3.11) converges to a positive solution $\mathbf{u}_h = (u_{h,0}, u_{h,1}, \ldots, u_{h,N})$ of (3.4), this solution is unique in $[\mathbf{0}, \mathbf{M}]$ and it satisfies

$$\underline{\varepsilon} := 0.17 \leq u_h(x) \leq 2, \quad \text{where} \quad u_h(x) = \sum_{i=0}^{N} u_{h,i}\varphi_i(x). \tag{3.31}$$

We plot in Figure 3.1 the a priori estimates of $\|u - u_h\|^2$, $\|(u - u_h)'\|^2$ and $\|u_h - u\|_\infty$ given by Theorem 3.4.1. To do so we proceed as follows, using the bounds given by (3.30) and (3.31), we have that

$$\gamma_0 \geq \underline{\varepsilon}e^{-1}, \quad \gamma_1 \geq \frac{\underline{\varepsilon}e^{-1}}{2-\cos(1)e^{1-\cos(1)}}, \quad \gamma_{2,1} \leq 16e^{-\cos(1)}\sin(1), \quad \tilde{\gamma} \leq 2\sin(1)(2+h) + 14,$$

$$\gamma_{2,2} \leq \left(2 + \sin(1)\sqrt{\tilde{\gamma}}\right)\left(\eta h + 2\sin(1)e^{-\cos(1)}\right), \quad \gamma_{2,3} \leq 8e^{-\cos(1)}h, \quad \gamma_{2,4} \leq 6\left(\eta h + 2\sin(1)e^{-\cos(1)}\right),$$

where $\eta := \|\left(e^{-\cos(x)}u\right)''\|_2$. Furthermore,

$$\eta \leq \|\left(e^{-\cos(x)}\right)'\|_\infty \|u' + \sin(x)u\|_\infty + \|e^{-\cos(x)}\|_\infty \|(u' + \sin(x)u)'\|_\infty.$$

To estimate $\|(u' + \sin(x)u)'\|_\infty$, we use that u is a solution of (3.29). Hence

$$\|(u' + \sin(x)u)'\|_\infty = \|\cos(x)u - u^2\|_\infty.$$

Now we integrate the corresponding equation of (3.29) to obtain that

$$\int_0^x u^2 - \cos(t)u \, dt = \int_0^x (u' + \sin(t)u)' \, dt = u'(x) + \sin(x)u(x)$$

holds true for each $x \in (0,1]$, because $u'(0) = 0$. Thus

$$\|u' + \sin(x)u\|_\infty \leq \|\cos(x)u - u^2\|_\infty,$$

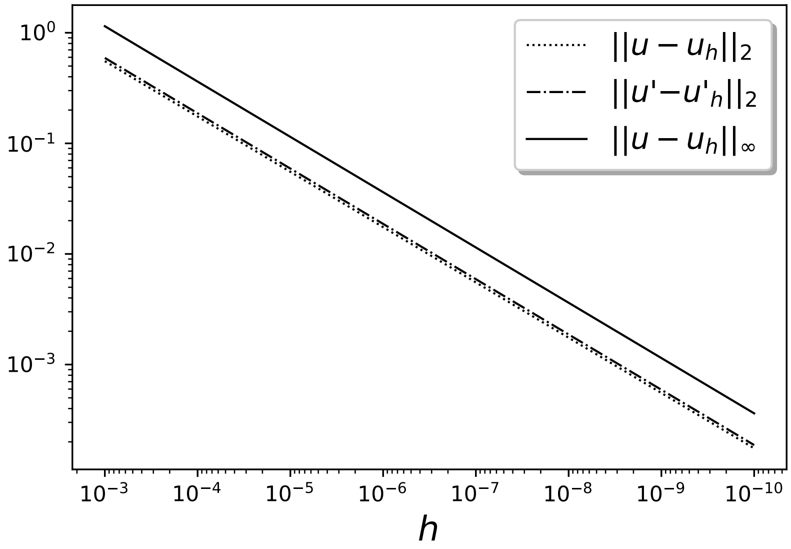

Figure 3.1: A priori errors for the problem (3.29).

which due to (3.30) leads to

$$\eta \leq \left(\sin(1)e^{-\cos(1)} + e^{-\cos(1)}\right) \|\cos(x)u - u^2\|_\infty \leq e^{-\cos(1)}\,(\sin(1)+1)\left(4-\cos^2(1)\right).$$

Finally we will obtain the estimations of $\|u-\bar{u}_h^{(k)}\|^2$, $\|\left(u-\bar{u}_h^{(k)}\right)'\|^2$ and $\|u-\bar{u}_h^{(k)}\|_\infty$, given by Theorem 3.4.2 and afterwards, we will present them in Table 3.1. It is important to highlight that, in our case, higher levels of precision (see [9] for more details) are required to obtain that

$$\left|\mathbf{F}_h\left(\bar{\mathbf{u}}_h^{(k)}\right)\right|_1 \leq (N+1)\max_i \left|F_{h,i}\left(\bar{\mathbf{u}}_h^{(k)}\right)\right| \to 0 \quad \text{as} \quad k\to\infty.$$

Using the following notation

$$\varepsilon_k := \min_{x\in[0,1]} \bar{u}_h^{(k)}(x), \quad M_k := \max_{x\in[0,1]} \bar{u}_h^{(k)}(x), \quad \text{and} \quad \xi := \left|\mathbf{F}_h\left(\bar{\mathbf{u}}_h^{(k)}\right)\right|_1,$$

we deduce that the constants in the Theorem 3.4.2 are bounded as follows

$$\gamma_0 \geq \varepsilon_k e^{-1}, \quad \gamma_1 \geq \frac{\varepsilon_k e^{-1}}{2-\cos(1)e^{1-\cos(1)}}, \quad \gamma_{2,1} \leq 2(2+M_k)M_k e^{-\cos(1)}\sin(1), \quad \hat{\gamma} \leq M_k\xi + \tilde{\gamma},$$

$$\tilde{\gamma} \leq \sin(1)M_k^2 + \frac{1}{2}\left(\sin(1)h+1\right)M_k^2 + M_k^2 + M_k^3, \quad \gamma_{2,2} \leq \left(M_k + \sin(1)\sqrt{\hat{\gamma}}\right)\left(\eta h + M_k\sin(1)e^{-\cos(1)}\right),$$

$$\gamma_{2,3} \leq (2 + M_k)M_k e^{-\cos(1)}h, \quad \gamma_{2,4} \leq (M_k^2 + M_k)\left(\eta h + M_k \sin(1)e^{-\cos(1)}\right), \quad \gamma_3 \leq (2 + M_k)e^{-\cos(1)},$$

where $\eta := \| \left(e^{-\cos(x)}u\right)'' \|_2$ is estimated as before.

h	k	$\left\|\mathbf{F}_h\left(\bar{\mathbf{u}}_h^{(k)}\right)\right\|_1$	$\|u - \bar{u}_h^{(k)}\|_2$	$\|(u - \bar{u}_h^{(k)})'\|_2$	$\|u - \bar{u}_h^{(k)}\|_\infty$
10^{-3}	6	$4.52 \cdot 10^{-16}$	$1.48 \cdot 10^{-1}$	$1.59 \cdot 10^{-1}$	$3.08 \cdot 10^{-1}$
10^{-4}	6	$4.53 \cdot 10^{-16}$	$4.7 \cdot 10^{-2}$	$5.02 \cdot 10^{-2}$	$9.73 \cdot 10^{-2}$
10^{-5}	6	$4.53 \cdot 10^{-16}$	$1.48 \cdot 10^{-2}$	$1.59 \cdot 10^{-2}$	$3.07 \cdot 10^{-2}$
10^{-6}	6	$4.53 \cdot 10^{-16}$	$4.7 \cdot 10^{-3}$	$5.02 \cdot 10^{-3}$	$9.72 \cdot 10^{-3}$

Table 3.1: A posteriori error for the problem 3.29.

3.5.2 Numerical results for the case $m(x) = \sin(x + 1)$, $a(x) = 1$, $D = \delta = \lambda = 1$ and $\alpha = 200$.

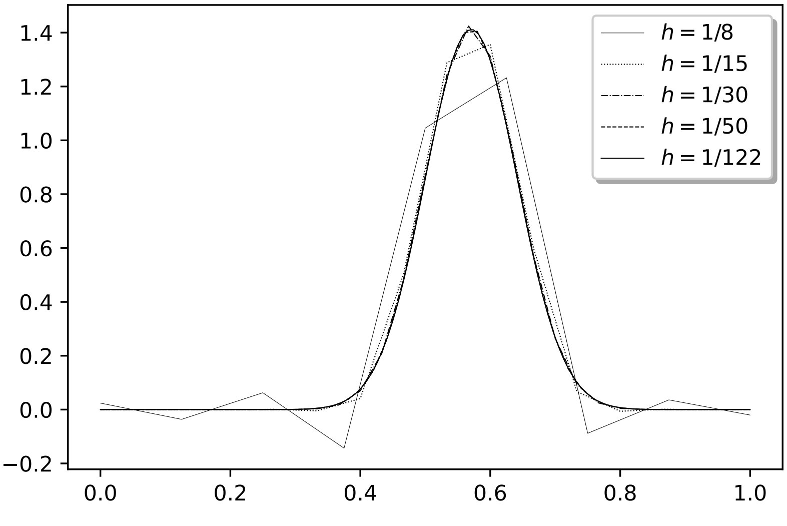

Figure 3.2: The iterations $\bar{u}_h^{(14)}$ for the problem (3.32) with different h.

Consider the following generalized diffusive logistic equation

$$\begin{cases} -(u' - 200\cos(x+1)u)' = \sin(x+1)u - u^2, & x \in (0,1), \\ u'(0) = 200\cos(1)u(0), \\ u'(1) = 200\cos(2)u(1). \end{cases} \tag{3.32}$$

We will study the behaviour of the iterations of the Newton Method for different choices of h. For this problem, Theorem 3.3.2 provides $M = 201$ in (3.12) and ensures that the Newton Method converges for an initial data $\mathbf{M} := 201(1,1,\dots,1)$ if $h < h_{201} \approx 1/121$. Recall that, by (3.9), Theorem 3.3.1 yields $h_{201} = 1/121$.

We plot in Figure 3.2 the corresponding fourteenth iteration of the Newton Method for a set of five decreasing values of h verifying

$$\left|\mathbf{F}_h\left(\bar{\mathbf{u}}_h^{(14)}\right)\right|_1 \approx 10^{-15}.$$

The effect in the fourteenth iteration when $h > h_{201}$ is appreciated in the simulations shown in Figure 3.2 and Figure 3.3. Moreover, we have done a zoom in Figure 3.3 to better observe the different cases: on the rise (left) and in the peak (right). Thus, Theorem 3.3.2 ensures the convergence and the monotonicity of the iterations of the Newton Method, as in (3.13), only for $h = 1/122$ among all h considered in Figure 3.2. It can be seen in Table 3.2 some cases in which the difference between two consecutive iterations of the Newton Method changes sign, consequently they are not ordered as in (3.13), for $h > h_{201}$. We

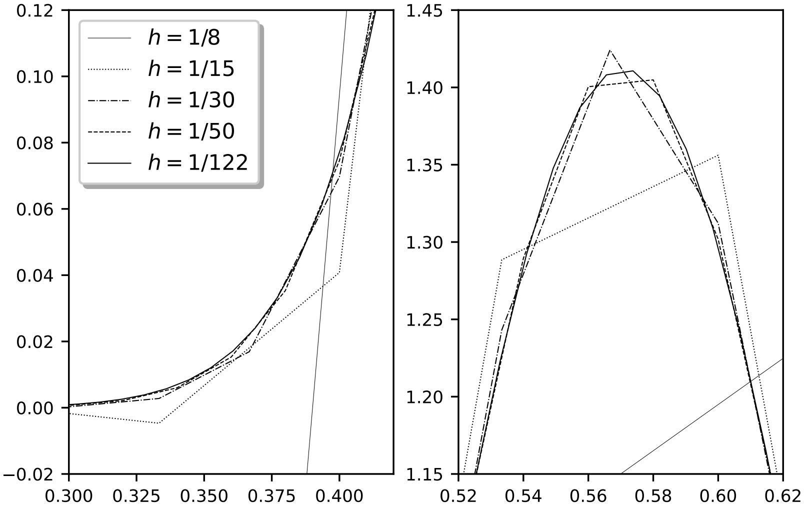

Figure 3.3: Zoom of Figure 3.2.

also provide in Table 3.2 the values of

$$\min\left\{\bar{\mathbf{u}}_h^{(k)} - \bar{\mathbf{u}}_h^{(k+1)}\right\} := \min_i \left\{\bar{u}_{h,i}^{(k)} - \bar{u}_{h,i}^{(k+1)}\right\} \tag{3.33}$$

for two different situations for each h: when (3.33) is more negative and for $k = 13$. Finally, we show in Figure 3.4 that the iterations of the Newton Method are ordered if $h < h_{201}$, however when $h > h_{201}$ they intersect.

h	k	$\min\left\{\bar{\mathbf{u}}_h^{(k)} - \bar{\mathbf{u}}_h^{(k+1)}\right\}$	h	k	$\min\left\{\bar{\mathbf{u}}_h^{(k)} - \bar{\mathbf{u}}_h^{(k+1)}\right\}$
1/8	4	-2.09	1/30	2	$-1.7 \cdot 10^{-3}$
1/8	13	$-1.42 \cdot 10^{-8}$	1/30	13	$-2.52 \cdot 10^{-14}$
1/15	4	-0.1	1/50	5	$-2.16 \cdot 10^{-18}$
1/15	13	$-8.29 \cdot 10^{-10}$	1/50	13	$-4.22 \cdot 10^{-26}$

Table 3.2: Several cases when the iterations do not satisfy (3.13).

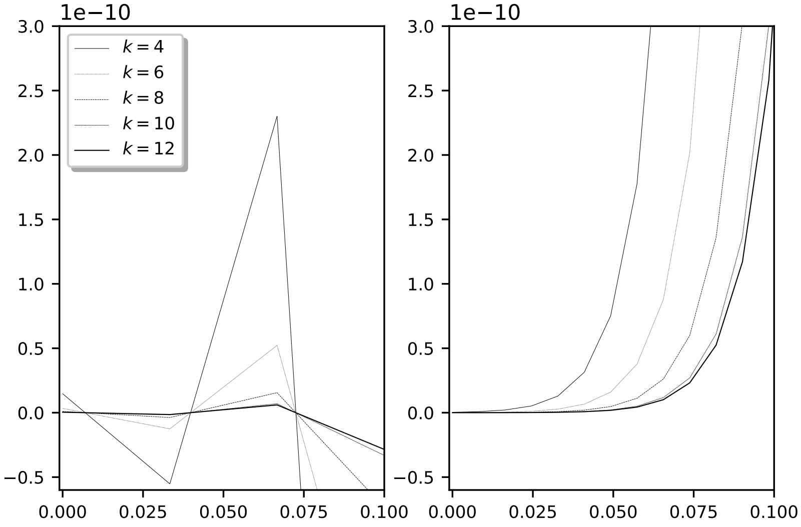

Figure 3.4: Zoom of several iterations for the case $h = 1/30$ (left) and $h = 1/122$ (right).

Chapter 4

The generalized diffusive logistic equation with change of variable

4.1 Introduction

In this chapter, we will apply the results of Chapter 2 to the problem (1.1) with $\alpha > 0$ and $m' \not\equiv 0$. For this goal, it should be noted that the change of variable

$$u = e^{\alpha m/(2D)}w, \tag{4.1}$$

transforms (1.1) in the following heterogeneous logistic equation

$$\begin{cases} -w'' = \left(\dfrac{\lambda}{D}m(x) - \dfrac{\alpha}{2D}m''(x) - \dfrac{\alpha^2}{4D^2}(m'(x))^2\right)w - \dfrac{\delta}{D}a(x)e^{\alpha m(x)/(2D)}w^2, \qquad x \in (0,1), \\ \\ -w'(0) + \dfrac{\alpha}{2D}m'(0)w(0) = 0, \quad w'(1) - \dfrac{\alpha}{2D}m'(1)w(1) = 0. \end{cases} \tag{4.2}$$

Then, (4.2) is a self-adjoint problem and it can be rewritten as

$$\begin{cases} -w'' = c(x)w - \tilde{a}(x)w^2, \qquad x \in (0,1), \\ -w'(0) + \beta_0 w(0) = 0, \\ w'(1) + \beta_1 w(1) = 0, \end{cases}$$

where

$$c(x) := \frac{\lambda}{D}m(x) - \frac{\alpha}{2D}m''(x) - \frac{\alpha^2}{4D^2}(m'(x))^2, \quad \tilde{a}(x) := \frac{\delta}{D}a(x)e^{\alpha m(x)/(2D)}$$
$$\beta_0 := \frac{\alpha}{2D}m'(0), \quad \beta_1 := -\frac{\alpha}{2D}m'(1). \tag{4.3}$$

Since we have assumed that $D > 0$, $\delta > 0$, $\alpha > 0$, $m \in C^2[0,1]$, $m'(0) \geq 0$, $m'(1) \leq 0$, $a \in C[0,1]$ and $a(x) > 0$ for all $x \in [0,1]$, it follows that $\beta_0 \geq 0$, $\beta_1 \geq 0$, $c \in C[0,1]$, $\tilde{a} \in C[0,1]$ and $\tilde{a}(x) > 0$ for all $x \in [0,1]$. Then, the hypotheses on the heterogeneous logistic equation (2.1) of Chapter 2 are verified.

On the other hand, note that (4.2) has a unique positive solution, denoted by w, in the same way that (2.1) possesses a unique positive solution u. Moreover, w satisfies

$$\int_0^1 w'v' - \int_0^1 cwv + \int_{\{0,1\}} \beta wv + \int_0^1 \tilde{a}w^2v = 0, \quad \forall v \in H^1((0,1)), \tag{4.4}$$

where $\beta(0) = \beta_0$ and $\beta(1) = \beta_1$. Thus, considering the space of piecewise linear finite elements $V_h \subset H^1((0,1))$, $h > 0$, it follows that the approximation of w, denoted by $w_h \in V_h$, satisfies

$$\int_0^1 w_h'v_h' - \int_0^1 \pi_h(c)w_hv_h + \int_{\{0,1\}} \beta w_hv_h + \int_0^1 \pi_h(\tilde{a})w_h^2v_h = 0, \ \ \forall v_h \in V_h, \tag{4.5}$$

where π_h is the interpolation operator onto V_h. Consequently, we propose

$$u_h := \pi_h(e^{\alpha m/(2D)}w_h)$$

as the candidate to approximate to u in V_h.

4.2 Applying and improving the Chapter 2

In a similar way to Chapter 2 for $D = \lambda = \delta = 1$, we define the following bilinear forms

$$b_f(w,v) := \int_0^1 w'v' - \int_0^1 cwv + \int_{\{0,1\}} \beta wv + \int_0^1 \tilde{a}fwv$$

and

$$b_f^h(w_h,v_h) := \int_0^1 w_h'v_h' - \int_0^1 \pi_h(c)w_hv_h + \int_{\{0,1\}} \beta w_hv_h + \int_0^1 \pi_h(\tilde{a})fw_hv_h,$$

with $\beta(0) = \beta_0$ and $\beta(1) = \beta_1$. Then, obtaining $w_h \in V_h$ satisfying (4.5) is equivalent to solve

$$\mathbf{F}_h(\mathbf{w}_h) = \mathbf{0} \tag{4.6}$$

where $\mathbf{F}_h : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ is defined by

$$\mathbf{F}_h(\mathbf{z}) := (A_h + Y_h(\mathbf{z}))\,\mathbf{z}, \quad A_h = \left\{b_0^h(\varphi_j,\varphi_i)\right\}_{i,j=0}^N \quad \text{and} \quad Y_h(\mathbf{z}) = \left\{\int_0^1 \pi_h(\tilde{a})z\varphi_j\varphi_i\right\}_{i,j=0}^N,$$

where i indicates the row and j indicates the column, and $\mathbf{z} = (z_0, z_1, ..., z_N)^T$ and $z = \sum_{i=0}^N z_i\varphi_i$.

Now, we look for both a positive strict supersolution and a subsolution of (4.6) (see (2.6) and (2.7)). Moreover, taking into account the hypothesis assumed on $m(x)$, we have that

$$M = \max_{x\in[0,1]} \frac{c(x)}{\tilde{a}(x)} = \max_{x\in[0,1]} \frac{4D\lambda m(x) - 2\alpha m''(x) - \alpha^2(m'(x))^2}{4D\delta a(x)e^{\alpha m/(2D)}} > 0. \tag{4.7}$$

Then, by the Theorem 2.4.2, $\mathbf{M} := M(1,1,\dots,1)^T$ satisfies

$$\mathbf{F}_h(\mathbf{M}) \geq \mathbf{0} \quad \text{and} \quad \mathbf{M} \gg \mathbf{0}.$$

Thus, $\mathbf{M}$ is a positive strict supersolution if $\mathbf{F}_h(\mathbf{M}) \neq \mathbf{0}$. On the other hand, since $m(x) > 0$ for all $x \in [0,1]$, we know that

$$\underline{w} := \varepsilon e^{\alpha m/(2D)}$$

is a positive subsolution on (4.2) satisfying (2.35) if ε is sufficiently small. Consequently, if h is sufficiently small and $m \in H^4((0,1))$, we apply the Theorem 2.4.3 to prove that

$$\mathbf{F}_h(\mathbf{E}) \ll \mathbf{0} \quad \text{with } \ \mathbf{E} = (\underline{w}(x_0), \underline{w}(x_1), \ldots, \underline{w}(x_N))^T. \tag{4.8}$$

The following theorem gives a positive strict subsolution of (4.6) in a similar way as Theorem 3.3.3 does.

Theorem 4.2.1. *Suppose $m \in H^4((0,1))$. Then,*

$$\boldsymbol{E} := \varepsilon \left(e^{\alpha m(x_0)/(2D)}, e^{\alpha m(x_1)/(2D)}, \ldots, e^{\alpha m(x_N)/(2D)}\right)$$

where

$$0 < \varepsilon < \frac{\lambda \min_{x\in[0,1]} m(x)}{2\delta \|a\|_\infty \|e^{\alpha m/D}\|_\infty} \tag{4.9}$$

satisfies (4.8) *if*

$$h \leq h^* := \left(\frac{\sqrt{3}\lambda \min_{x\in[0,1]} m(x)}{4\sqrt{2}D\|e^{-\alpha m/(2D)}\|_\infty \left(\left(\|c\|_\infty + \frac{\lambda}{D}\|m\|_\infty\right)\|\left(e^{\alpha m/(2D)}\right)''\|_2 + \|e^{\alpha m/(2D)}\|_\infty \|c''\|_2\right)}\right)^{2/3}$$

where $c(x) := \dfrac{\lambda}{D}m(x) - \dfrac{\alpha}{2D}m''(x) - \dfrac{\alpha^2}{4D^2}(m'(x))^2$.

Proof. We will prove that

$$F_{h,i}(\mathbf{E}) = b_E^h(E, \varphi_i) < 0 \ \text{ for each } \ i \in \{0, 1, \ldots, N\}, \tag{4.10}$$

if $E = \pi_h(\underline{w})$ and $\underline{w} := \varepsilon e^{\alpha m(x)/(2D)}$. Note that

$$\int_0^1 \underline{w}'\varphi_i' = -\int_0^1 \underline{w}''\varphi_i \ \text{ for each } \ i \in \{1, 2, \ldots, N-1\},$$

$$\int_0^1 \underline{w}'\varphi_0' = -\beta_0\underline{w}(0) - \int_0^1 \underline{w}''\varphi_0 \quad \text{and} \quad \int_0^1 \underline{w}'\varphi_N' = -\beta_1\underline{w}(1) - \int_0^1 \underline{w}''\varphi_N.$$

Thus, it is easy to obtain that

$$b_E^h(E, \varphi_i) = \int_0^1 (E' - \underline{w}')\varphi_i' + \int_0^1 \left(-\underline{w}'' - \pi_h(c)E + \pi_h(\tilde{a})E^2\right)\varphi_i$$

for each $i \in \{0, 1, \ldots, N\}$. Moreover, since

$$\underline{w}'' = \left(\frac{\lambda}{D}m - c\right)\underline{w} \quad \text{and} \quad \int_0^1 (E' - \underline{w}')\varphi_i' = 0 \ \text{ for each } \ i \in \{0, 1, \ldots, N\},$$

it is enough that

$$b_E^h(E,\varphi_i) = \int_0^1 \left(c\underline{w} - \pi_h(c)E\right)\varphi_i + \int_0^1 \left(\pi_h(\tilde{a})E^2 - \frac{\lambda}{D}m\underline{w}\right)\varphi_i < 0 \tag{4.11}$$

to prove (4.10). Thus, we deduce (4.11) if $h \leq h^*$. Indeed, by (1.5) and the Hölder inequality, we have that

$$\int_0^1 \left(c\underline{w} - \pi_h(c)E\right)\varphi_i = \int_0^1 (c - \pi_h(c))\,\underline{w}\varphi_i + \int_0^1 \pi_h(c)\left(\underline{w} - E\right)\varphi_i \leq \left(\|\underline{w}\|_\infty \|c''\|_2 + \|c\|_\infty \|\underline{w}''\|_2\right) h^2 \|\varphi_i\|_2$$

and

$$\int_0^1 m\left(E - \underline{w}\right)\varphi_i \leq \|m\|_\infty \|\underline{w}''\|_2 h^2 \|\varphi_i\|_2.$$

Then, we deduce that

$$b_E^h(E,\varphi_i) \leq \left(\|\underline{w}\|_\infty \|c''\|_2 + \left(\|c\|_\infty + \frac{\lambda}{D}\|m\|_\infty\right)\|\underline{w}''\|_2\right) h^2 \|\varphi_i\|_2 + \int_0^1 \left(\pi_h(\tilde{a})E - \frac{\lambda}{D}m\right)E\varphi_i. \tag{4.12}$$

Now, by (4.9), we know that

$$g(\varepsilon) := \frac{\delta}{D}\|a\|_\infty \|e^{\alpha m/D}\|_\infty\ \varepsilon - \frac{\lambda}{D}\min_{x\in[0,1]} m(x) < \frac{2\delta}{D}\|a\|_\infty \|e^{\alpha m/D}\|_\infty\ \varepsilon - \frac{\lambda}{D}\min_{x\in[0,1]} m(x) < 0.$$

Thus, it is possible to find an upper bound for the second integral of (4.12) in the following way

$$\int_0^1 \left(\pi_h(\tilde{a})E - \frac{\lambda}{D}m\right)E\varphi_i \leq g(\varepsilon)\int_0^1 E\varphi_i \leq \frac{g(\varepsilon)\varepsilon h}{2\|e^{-\alpha m/(2D)}\|_\infty}. \tag{4.13}$$

Consequently, by (4.12), (4.13) and $\|\varphi_i\|_2 \leq (2h/3)^{1/2}$, we obtain that

$$b_E^h(E,\varphi_i) \leq \left(\|\underline{w}\|_\infty \|c''\|_2 + \left(\|c\|_\infty + \frac{\lambda}{D}\|m\|_\infty\right)\|\underline{w}''\|_2\right) h^2 (2h/3)^{1/2} + \frac{g(\varepsilon)\varepsilon h}{2\|e^{-\alpha m/(2D)}\|_\infty}.$$

Finally, (4.11) is satisfied if

$$g(\varepsilon) < -2\|e^{-\alpha m/(2D)}\|_\infty \left(\|e^{\alpha m/(2D)}\|_\infty \|c''\|_2 + \left(\|c\|_\infty + \frac{\lambda}{D}\|m\|_\infty\right)\|\left(e^{\alpha m/(2D)}\right)''\|_2\right) h (2h/3)^{1/2},$$

or equivalently, if

$$\varepsilon < \frac{\lambda \min_{x\in[0,1]} m(x) - 2D\|e^{-\alpha m/(2D)}\|_\infty \left(\|e^{\alpha m/(2D)}\|_\infty \|c''\|_2 + \left(\|c\|_\infty + \frac{\lambda}{D}\|m\|_\infty\right)\|\left(e^{\alpha m/(2D)}\right)''\|_2\right) h (2h/3)^{1/2}}{\delta \|a\|_\infty \|e^{\alpha m/D}\|_\infty}$$

which holds true because $h \leq h^*$ and (4.9). □

Now, according to Theorem 2.3.4, to guarantee that the elements of the subdiagonals of

$$J_h(\mathbf{z}) := A_h + 2Y_h(\mathbf{z})$$

are negative for all $h < h_M$ and for each $\mathbf{z} \leq \mathbf{M} := M(1, 1, ..., 1)^T$, we need that

$$h_M := \left(\frac{24D^2}{\|8D\delta M a e^{\alpha m/(2D)} - 4D\lambda m + 2D\alpha m'' + \alpha^2 (m')^2\|_\infty} \right)^{1/2},$$

with $h_M = \infty$ if the denominator is equal to zero. Now we can apply Theorem 2.4.4. Note that the sequence in (2.27) is not needed in this chapter. Thus, the following theorem is analogous to Theorem 3.3.2. We will omit the its proof because it is similar to the proofs of Theorem 3.3.2 and Theorem 2.3.6. As before, the case in which $\mathbf{M}$ is itself a positive solution of 4.6 is excluded.

Theorem 4.2.2. *If $\bar{\boldsymbol{w}}_h^{(0)} := \boldsymbol{M} = M(1, 1, \ldots, 1)^T$, with M defined in* (4.7)*, $\boldsymbol{F}_h(\boldsymbol{M}) \neq \boldsymbol{0}$ and $h < h_M$, the sequence*

$$\bar{\boldsymbol{w}}_h^{(k+1)} = \bar{\boldsymbol{w}}_h^{(k)} - (J_h(\bar{\boldsymbol{w}}_h^{(k)}))^{-1} \boldsymbol{F}_h(\bar{\boldsymbol{w}}_h^{(k)}), \quad k \geq 0, \tag{4.14}$$

where $J_h(\bar{\boldsymbol{w}}_h^{(k)}) := A_h + 2Y_h(\bar{\boldsymbol{w}}_h^{(k)})$, is well defined, it converges to some solution of (4.6) *and*

$$\boldsymbol{0} \ll \bar{\boldsymbol{w}}_h^{(k+1)} \ll \bar{\boldsymbol{w}}_h^{(k)} \ll \boldsymbol{M} \ \textit{for all} \ k \geq 0.$$

Moreover, if $m \in H^4((0, 1))$, h^* and $\boldsymbol{E}$ are as defined in Theorem 4.2.1, $\boldsymbol{E} \ll \boldsymbol{M}$ and $h \leq h^*$. Then

$$\boldsymbol{0} \ll \boldsymbol{E} \ll \bar{\boldsymbol{w}}_h^{(k)} \ \textit{for all} \ k \geq 0$$

and consequently, (4.14) *converges to the unique positive solution of* (4.6) *between* $[\boldsymbol{0}, \boldsymbol{M}]$.

As we have already seen $\mathbf{w}_h = (w_{h,0}, w_{h,1}, \ldots, w_{h,N})^T$ is a solution of (4.6) if, and only if, $w_h = \sum_{i=0}^N w_{h,i}\varphi_i$ satisfies (4.5). Hence, thanks to Theorem 4.2.2, the function w_h and the iterations of the Newton Method

$$\bar{w}_h^{(k)} = \sum_{i=0}^{N} \bar{w}_{h,i}^{(k)} \varphi_i, \quad \text{where} \quad \bar{\mathbf{w}}_h^{(k)} = \left(\bar{w}_{h,0}^{(k)}, \bar{w}_{h,1}^{(k)}, \ldots, \bar{w}_{h,N}^{(k)} \right)$$

are bounded by constants that do not depend on h :

$$0 < \varepsilon e^{\alpha \min_{x \in [0,1]} m(x)/(2D)} \leq w_h(x) \leq \bar{w}_h^{(k)}(x) \leq M \quad \text{for all} \ \ x \in [0, 1] \tag{4.15}$$

and ε satisfies (4.9). Hence, to prove that w_h converges to w when h tends to 0, we may apply Theorem 2.5.3 if $a \in H^2((0, 1))$.

Now, consider $u_h = \pi_h(e^{\alpha m/(2D)} w_h)$. We prove that u_h approximates to u. Indeed, by (4.1) and Proposition 1.2.1, it is apparent that

$$\begin{aligned} \|u_h - u\|_\infty &\leq \|\pi_h(e^{\alpha m/(2D)} w_h) - e^{\alpha m/(2D)} w_h\|_\infty + \|e^{\alpha m/(2D)} w_h - e^{\alpha m/(2D)} w\|_\infty \\ &\leq \|w_h\|_\infty \| \left(e^{\alpha m/(2D)} \right)' \|_\infty h + \|e^{\alpha m/(2D)}\|_\infty \|w_h - w\|_\infty. \end{aligned} \tag{4.16}$$

Moreover, by Theorem 2.5.3, there exists a constant $C > 0$ such that

$$\|w_h - w\|_\infty \leq 2\|w_h - w\|_{H^1} \leq 2 \left(Ch + \sqrt{h^2 + 1} \|w''\|_2 \right) h. \tag{4.17}$$

Therefore, from (4.15), (4.16) and (4.17), we deduce that u_h converges to u whe h tends to 0.

The next theorem provides the estimates for the norms of the difference between u and the iteration

$$\bar{u}_h^{(k)} := \pi_h\left(e^{\alpha m/(2D)}\bar{w}_h^{(k)}\right),$$

and w and $\bar{w}_h^{(k)}$. These errors depend on $\mathbf{F}_h\left(\bar{\mathbf{w}}_h^{(k)}\right)$, that converges to $\mathbf{0}$ when $k \to \infty$.

Theorem 4.2.3. *Suppose $a \in H^2((0,1))$ and $m \in H^4((0,1))$. Under the assumptions of Theorem 4.2.2, the following estimates hold true for each $k \geq 0$,*

$$\|\bar{u}_h^{(k)} - u\|_\infty \leq \|\bar{w}_h^{(k)}\|_\infty \| \left(e^{\alpha m/(2D)}\right)' \|_\infty h + \|e^{\alpha m/(2D)}\|_\infty \|\bar{w}_h^{(k)} - w\|_\infty, \tag{4.18}$$

and

$$\|\bar{w}_h^{(k)} - w\|_\infty \leq \left(h^2 + h\sqrt{h^2+1}\right)\|w''\|_2 + \left(\frac{\gamma_2}{\gamma_0} + \frac{\gamma_2}{\gamma_1}\right)h^2 + \left(\frac{1}{\gamma_0} + \frac{1}{\gamma_1}\right)(6/h)^{1/2}\left|\boldsymbol{F}_h\left(\bar{\boldsymbol{w}}_h^{(k)}\right)\right|_2, \tag{4.19}$$

where

$$\gamma_0 := \min_{x\in[0,1]}\left(\tilde{a}(x)\bar{w}_h^{(k)}(x)\right), \quad \gamma_1 = \min\left\{\frac{\gamma_0}{1+\|g^+\|_\infty}, 1\right\}$$

and

$$\gamma_2 = \|\bar{w}_h^{(k)}\|_\infty\|c''\|_2 + \|\tilde{a}''\|_2\|\bar{w}_h^{(k)}\|_\infty^2 + \left(\|\tilde{a}w - c\|_\infty + \|\tilde{a}\|_\infty\|\bar{w}_h^{(k)}\|_\infty\right)\|w''\|_2,$$

where $\tilde{a}$ and c are defined in (4.3) *and g^+ is the positive part of $g := c - \tilde{a}w$.*

Proof. Since w is a positive solution of the problem (4.2), w is a positive eigenfunction associated with the eigenvalue 0 of the following eigenvalue problem

$$\begin{cases} -\psi'' - c\psi + \tilde{a}w\psi = \sigma\psi, & \text{in } (0,1),\\ -\psi'(0) + \beta_0\psi(0) = 0, & \\ \psi'(1) + \beta_1\psi(1) = 0, & \end{cases}$$

where c, $\tilde{a}$, β_0 and β_1 are given in (4.3). Thus, by [37, Th. 7.7 and Th. 7.8], 0 is the corresponding smallest eigenvalue. Moreover, from the Rayleigh formula applied to the symmetric form b_w, it follows that

$$b_w(v,v) \geq 0, \quad \forall v \in H^1((0,1)). \tag{4.20}$$

Now, for $k \geq 0$, we will prove that

$$b_{w+\bar{w}_h^{(k)}}(v,v) \geq \gamma_0\|v\|_2^2 \quad \text{and} \quad b_{w+\bar{w}_h^{(k)}}(v,v) \geq \gamma_1\|v\|_{H^1}^2. \tag{4.21}$$

Indeed, by (4.20), we have that

$$b_{w+\bar{w}_h^{(k)}}(v,v) \geq \int_0^1 \tilde{a}\bar{w}_h^{(k)}v^2 \geq \gamma_0\|v\|_2^2.$$

Hence, the first inequality of (4.21) is proved. On the other hand, thanks also to (4.20) we know that

$$b_{w+\bar{w}_h^{(k)}}(v,v) \geq \mu b_w(v,v) + \int_0^1 \tilde{a}\bar{w}_h^{(k)}v^2$$

for some $\mu \in [0,1]$. Then, as $\beta_0 \geq 0$ and $\beta_1 \geq 0$, we have that

$$b_{w+\bar{w}_h^{(k)}}(v,v) \geq \mu \int_0^1 (v')^2 + \int_0^1 (\gamma_0 - \mu \| (c - \tilde{a}w)^+ \|_\infty)\, v^2 \tag{4.22}$$

and, choosing $\mu = 1$ if $\gamma_0 - \| (c - \tilde{a}w)^+ \|_\infty \geq 1$, or otherwise,

$$\mu = \gamma_0 - \mu \| (c - \tilde{a}w)^+ \|_\infty, \quad \text{i.e.,} \quad \mu = \frac{\gamma_0}{1 + \| (c - \tilde{a}w)^+ \|_\infty} \in [0,1],$$

we obtain from (4.22) the second inequality in (4.21).

Next, we will prove that

$$b_{w+\bar{w}_h^{(k)}}(v_h, v_h) \leq \gamma_2 h^2 \|v_h\|_2 + (6/h)^{1/2} \|v_h\|_2 \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2 \quad \text{where} \quad v_h := \pi_h(w) - \bar{w}_h^{(k)}. \tag{4.23}$$

First, by the Cauchy-Schwarz inequality and Proposition 1.2.2, it follows that

$$-b^h_{\bar{w}_h^{(k)}}(\bar{w}_h^{(k)}, v_h) \leq \left| \mathbf{v}_h^T \mathbf{F}_h(\bar{\mathbf{w}}_h^{(k)}) \right| \leq |\mathbf{v}_h|_2 \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2 \leq (6/h)^{1/2} \|v_h\|_2 \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2.$$

Moreover, as w satisfies (4.4), it is enough to obtain that

$$b_{w+\bar{w}_h^{(k)}}(v_h, v_h) - b_w(w, v_h) + b^h_{\bar{w}_h^{(k)}}(\bar{w}_h^{(k)}, v_h) \leq \gamma_2 h^2 \|v_h\|_2. \tag{4.24}$$

Now, we will prove (4.24). Indeed, defining $E_f := \pi_h(f) - f$, we have that

$$\int_0^1 E_w' v_h' = 0 \quad \text{and} \quad \int_{\{0,1\}} \beta E_w v_h = 0.$$

Therefore, it follows that

$$\begin{aligned} &b_{w+\bar{w}_h^{(k)}}(v_h, v_h) - b_w(w, v_h) + b^h_{\bar{w}_h^{(k)}}(\bar{w}_h^{(k)}, v_h) \\ &\quad = -\int_0^1 E_c \bar{w}_h^{(k)} v_h + \int_0^1 E_{\tilde{a}} \left(\bar{w}_h^{(k)} \right)^2 v_h + \int_0^1 \left[\tilde{a} \left(w + \bar{w}_h^{(k)} \right) - c \right] E_w v_h. \end{aligned}$$

Consequently, by the Hölder's inequality and (1.5), we obtain (4.23).

Finally, if $v := \pi_h(w) - \bar{w}_h^{(k)}$ in (4.21), from (4.23) it is apparent that

$$\gamma_0 \|\pi_h(w) - \bar{w}_h^{(k)}\|_2^2 \leq \gamma_2 h^2 \|\pi_h(w) - \bar{w}_h^{(k)}\|_2 + (6/h)^{1/2} \|\pi_h(w) - \bar{w}_h^{(k)}\|_2 \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2 \tag{4.25}$$

and

$$\gamma_1 \|\pi_h(w) - \bar{w}_h^{(k)}\|_{H^1}^2 \leq \gamma_2 h^2 \|\pi_h(w) - \bar{w}_h^{(k)}\|_2 + (6/h)^{1/2} \|\pi_h(w) - \bar{w}_h^{(k)}\|_2 \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2. \tag{4.26}$$

Then, removing terms in (4.25), we deduce that

$$\gamma_0 \|\pi_h(w) - \bar{w}_h^{(k)}\| \leq \gamma_2 h^2 + (6/h)^{1/2} \left| \mathbf{F}_h \left(\bar{\mathbf{w}}_h^{(k)} \right) \right|_2, \tag{4.27}$$

and thus, by (1.5), we have that

$$\|w-\bar{w}_h^{(k)}\|_2 \le \|w-\pi_h(w)\|_2 + \|\pi_h(w)-\bar{w}_h^{(k)}\|_2 \le h^2\|w''\|_2 + \frac{\gamma_2}{\gamma_0}h^2 + \frac{(6/h)^{1/2}}{\gamma_0}\left|\mathbf{F}_h\left(\bar{\mathbf{w}}_h^{(k)}\right)\right|_2. \tag{4.28}$$

And, in the same way, since

$$\|\pi_h(w)-\bar{w}_h^{(k)}\|_2 \le \|\pi_h(w)-\bar{w}_h^{(k)}\|_{H^1},$$

we obtain from (4.26) and (1.5) that

$$\|w-\bar{w}_h^{(k)}\|_{H^1} \le \|w-\pi_h(w)\|_{H^1} + \|\pi_h(w)-\bar{w}_h^{(k)}\|_{H^1} \le h\sqrt{h^2+1}\|w''\|_2 + \frac{\gamma_2}{\gamma_1}h^2 + \frac{(6/h)^{1/2}}{\gamma_1}\left|\mathbf{F}_h\left(\bar{\mathbf{w}}_h^{(k)}\right)\right|_2.$$

Therefore, (4.19) follows from

$$\|w-\bar{w}_h^{(k)}\|_\infty \le \|w-\bar{w}_h^{(k)}\|_2 + \|w-\bar{w}_h^{(k)}\|_{H^1}.$$

Moreover, replacing w_h and u_h by $\bar{w}_h^{(k)}$ and $\bar{u}_h^{(k)}$, respectively, (4.16) implies (4.18). □

4.3 Simulations

We will apply the results of this chapter to the problem

$$\begin{cases} -(u'+\sin(x)u)' = \cos(x)u - u^2, & x\in(0,1),\\ u'(0)=0, & \\ u'(1) = -\sin(1)u(1). & \end{cases} \tag{4.29}$$

First, we propose the change of variable

$$u = e^{\cos(x)/2}w$$

that transforms (4.29) into

$$\begin{cases} -w'' = 0.25\left(6\cos(x)-\sin^2(x)\right)w - e^{\cos(x)/2}w^2, & x\in(0,1),\\ \\ w'(0)=0, \quad w'(1)+0.5\sin(1)w(1)=0. & \end{cases} \tag{4.30}$$

Note that (4.30) possesses a unique positive solution w, because (4.29) has a unique positive solution as we have proved in the Subsection 3.5.1. Moreover, we know that

$$\bar{W} := 1.5e^{-1/2} \quad \text{and} \quad \underline{W} = \cos(1)e^{0.5\cos(x)-\cos(1)}$$

are a supersolution and a subsolution of (4.30), respectively, and they satisfy $0<\underline{W}\le\bar{W}$. Moreover,

$$0.41 \le \cos(1)e^{0.5\cos(x)-\cos(1)} \le w(x) \le 1.5e^{-1/2} \le 0.91 \quad \text{for all } x\in[0,1]. \tag{4.31}$$

Second, from the Theorem 4.2.1, we deduce that

$$\mathbf{E} := 0.0993\left(e^{\cos(x_0)/2}, e^{\cos(x_1)/2}, \dots, e^{\cos(x_N)/2}\right)$$

satisfies (4.8) if $h \leq 0.141 < h^*$. Now, if $M := \bar{W}$ and $h \leq 0.141 < h^*$ the hypothesis of the Theorem 4.2.2 are satisfied because $h_M > 1/2$. Thus, the sequence in (4.14) converges to a positive solution $\mathbf{w}_h = (w_{h,0}, w_{h,1}, \ldots, w_{h,N})$ of (4.6). This solution is unique in $[\mathbf{0}, \mathbf{M}]$ and it satisfies

$$0.13 \leq w_h(x) \leq 0.91, \quad \text{where} \quad w_h(x) = \sum_{i=0}^{N} w_{h,i}\varphi_i(x). \tag{4.32}$$

Now we will obtain the estimations of $\|\bar{u}_h^{(k)} - u\|_\infty$, given by Theorem 4.2.3. Afterwards, we will present these estimations in Table 4.1. Using the notation

$$\varepsilon_k := \min_{x\in[0,1]} \bar{u}_h^{(k)}(x) \quad \text{and} \quad M_k := \max_{x\in[0,1]} \bar{u}_h^{(k)}(x),$$

we have that the constants in the Theorem 4.2.3 are bounded as follows

$$\gamma_0 \geq \varepsilon_k e^{\cos(1)/2}, \quad \gamma_1 \geq \frac{\varepsilon_k e^{\cos(1)/2}}{2.5 - \cos(1)e^{1-\cos(1)}}, \quad \gamma_2 \leq 2M_k + 0.5e^{1/2}M_k^2 + (1.5 - \cos(1)e^{1-\cos(1)} + e^{1/2}M_k)\eta,$$

where $\eta := \|w''\|_2$ is estimated from problem (4.30) and (4.31) as follows

$$\|w''\|_2 \leq \|w''\|_\infty = \|cw - \tilde{a}w^2\|_\infty \leq 2.25e^{-1/2} - \cos^2(1)e^{1.5-2\cos(1)} \approx 0.92$$

h	k	$\left\lvert \mathbf{F}_h\left(\bar{\mathbf{w}}_h^{(k)}\right)\right\rvert_2$	$\|w - \bar{w}_h^{(k)}\|_\infty$	$\|u - \bar{u}_h^{(k)}\|_\infty$
10^{-2}	4	$1.4 \cdot 10^{-8}$	$1.05 \cdot 10^{-2}$	$2.07 \cdot 10^{-2}$
10^{-3}	4	$4.44 \cdot 10^{-9}$	$9.35 \cdot 10^{-4}$	$1.87 \cdot 10^{-3}$
10^{-4}	4	$1.41 \cdot 10^{-9}$	$9.36 \cdot 10^{-5}$	$1.87 \cdot 10^{-4}$
10^{-5}	5	$1.30 \cdot 10^{-16}$	$9.2 \cdot 10^{-6}$	$1.84 \cdot 10^{-5}$

Table 4.1: A posteriori error for the problem 4.29.

Finally, it is important to comment that 128-bit floating-point arithmetic (quad precision) is required for the case $h = 10^{-5}$ because of the bad influence of the rounding errors appearing if 64-bit floating-point arithmetic (double precision) is used. In particular, the results of Theorem 4.2.2 are not satisfied for $k = 4$ and it is not possible to obtain

$$\left\lvert \mathbf{F}_h\left(\bar{\mathbf{w}}_h^{(k)}\right)\right\rvert_\infty < 10^{-11}$$

for some $k \geq 0$ if the double precision is activated. Fortunately, the problems are eliminated using quad precision.

Chapter 5

Conclusions

This thesis introduces the discrete sub- and supersolutions method to simulate the positive solutions of the generalize logistic diffusive equation. This new method provides exactly an algorithm with all the necessary details, like the initial data, the step size and the stop criterion among others, to approximate the positive solution of a current research problem. In addition the method calculates explicitly the difference between the approximated solution and the solution. Consequently, the discrete sub and supersolutions method provides the keys to optimize the finite element mesh size when applied to nonlinear equations with varying coefficients.

The necessary and sufficient conditions that guarantee that the Discrete Maximum Principle (DMP) is fulfilled are the key tools throughout the thesis. The DMP in this thesis is the discrete version of the characterization of the Maximum Principle in [40], [8] and [37] that has inspired us. To ensure the convergence of the Newton Method, it is enough that the iterations are positive, are ordered and do not intercross. For all this to be fulfilled, it is required to restrict the mesh size so that the DMP is satisfied. Also, it is required the initial data of the Newton Method to be a positive strict supersolution. In this thesis, we provide this mesh size and one positive strict supersolution for each discrete problem studied.

Once the converge of the Newton Method is guaranteed, it is necessary to find a positive constant limiting below all iterations, independently of the mesh size. This constant ensures that the coercivity constant does not depend on the mesh size which is essential to prove that the approximate solution effectively converges to the solution of the problem. In this thesis, we prove that, thanks to the DMP, a positive strict discrete subsolution limits bellow all iterations but it would be necessary to check that this subsolution is bounded below by a positive constant, independently of the mesh size. Fortunately, if the mesh size is less that a second threshold, a theoretical subsolution, maybe more restrictive, provides us this positive constant. In addition, in this thesis, the second threshold is provided for each discrete problem.

The discrete sub- and supersolutions was introduced for the first time in the recent paper [5] for a heterogeneous diffusive logistic equation, which is written up in Chapter 2. The referee of this paper exposed:

“This works considers the stationary logistic equation, set in one spatial dimension, with variable coefficients and Neumann or Robin boundary conditions. The authors establish some conditions to ensure that the discretization of the problem via the nonlinear finite elements method converges to and

has the same shape of the solution of the original continuous problem.

The main novelty is the use of discrete sub- and supersolutions to guarantee that the Newton (or modified Newton) method converges to the solution of the discrete problem. The key for that is a characterization of the validity of the discrete maximum principle, which is a counterpart of the continuous version established in [40] and later refined in [8].

The techniques developed in this work are very innovative since they build an easily implementable bridge between continuous and discrete problems, that can be extended to many other differential problems of elliptic type. Moreover, the mathematical presentation of the results is pretty clear and the simulations of Section 6 help to understand how the developed theory works. For all this reasons, I recommend this work for publication in the Journal of Differential Equations."

The main goal of this thesis is to approximate the positive solution of the generalized diffusive logistic equation. In Chapter 3, we apply the discrete sub- and supersolutions method to this model and successfully achieved our goal. Similar results to those of Chapter 2 are achieved depending on the parameters of the model. The great difficulty of this problem is that is not self-adjoint, but it is possible to apply a change of variable to obtain the coercivity constant and thus, to prove the convergence. Also, a different strategy is considered in Chapter 4. Applying a change of variable, we transform the model in a self-adjoint problem in order to use the results of Chapter 2.

The two strategies provide methods to approximate the positive solution of the generalized diffusive logistic equation. It is important to highlight that each strategy either provides different initial data, step size, stop criterion and explicit error. We provide them for the same example in Chapter 3 and Chapter 4. Clearly, the errors in infinity norm are better for "normal" parameters if we apply the change of variable because in this case the error is bounded above by a constant dependently of h against $h^{1/2}$ without the change of variable. Nevertheless, when we apply the change of variable using the exponential function, the underflow and overflows have to be checked. Also, it is necessary to approximate the second derivative of m, thus more regularity on m is needed. Consequently, we suggest taking into account both strategies and to calculate both errors.

Finally, we would like to comment that the discrete sub- and supersolutions method opens the possibility of being able to use it with other numerical methods different to the Finite Element Method and also to apply it to many others nonlinear partial differential equations. Moreover, the following step is to treat the case of several dimensions, but we have preferred to choose it in one dimension, because the method is new.

Bibliography

[1] D. Aleja, *Ecuaciones adventivas-difusivas de la dinámica de poblaciones*, Tesis Doctoral en Investigación Matemática, Universidad Complutense de Madrid (2015).

[2] D. Aleja and J. López-Gómez, *Some paradoxical effects of the advection on a class of diffusive equations in Ecology*, Disc. Cont. Dyn. Systems B **19** (2014) 10, 3031–3056.

[3] D. Aleja and J. López-Gómez, *Concentration through large advection*, J. Diff. Eqns. **257** (2014), 3135–3164.

[4] D. Aleja and J. López-Gómez, *Dynamics of a class of advective-diffusive equations in Ecology*, Advanced Nonlinear Studies **15** (2015), 557–587.

[5] D. Aleja and M. Molina-Meyer, *Nonlinear finite elements: Sub- and supersolutions for the heterogeneous logistic equation*, J. Diff. Eqns. **278** (2021), 189–219.

[6] H. Amann, *On the existence of positive solutions of nonlinear elliptic boundary value problems*, Indiana Univ. Math. **21** (1971), 125–146.

[7] H. Amann, *Zum Galerkin-Verfahren für die Hammersteinsche Gleichung*, Arch. Rational Mech. Anal. **35** (1969), 114–121.

[8] H. Amann and J. López-Gómez, *A priori bounds and multiple solutions for superlinear indefinite elliptic problems*, J. Diff. Eqns. **146** (1998), 336–374.

[9] D.H. Bailey and J.M. Borwein, *High-precision arithmetic in mathematical physics.*, Mathematics **3** (2015), 337–367.

[10] F. Belgacem and C. Cosner, *The effects of dispersal along environmental gradients on the dynamics of populations in heterogeneous environments*, Can. Appl. Math. Quart., **3** (1995), 379–397.

[11] K. Böhmer, *Numerical Methods for Nonlinear Elliptic Differential Equations: A Synopsis*, Oxford University Press, New York, 2010.

[12] J. H. Brandts, S. Korotov and M. Křížek, *The discrete maximum principle for linear simplicial finite element approximations of a reaction diffusion problem*, Linear Algebra and its Applications **429** (2008), 2344-2357.

[13] F. Brezzi, J. Rappaz and P.-A. Raviart, *Finite-dimensional approximation of nonlinear problems. Part I: Branches of nonsingular solutions*, Numer. Math. **36** (1980), 1–25.

[14] F. Brezzi, J. Rappaz and P.-A. Raviart, *Finite-dimensional approximation of nonlinear problems. Part II: Limit points*, Numer. Math. **37** (1981), 1–28.

[15] F. Brezzi, J. Rappaz and P.-A. Raviart, *Finite-dimensional approximation of nonlinear problems. Part III: Simple bifurcation points*, Numer. Math. **38** (1981), 1–30.

[16] F. E. Browder, *Nonlinear elliptic boundary value problems*, Bull. Amer. Math. Soc. **69** (1963), 862–874.

[17] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979.

[18] G. Caloz, *Stability of the approximation of a regular solution branch*, IMA Journal of Numerical Analysis **17** (1997), 2, 285–303.

[19] G. Caloz and J. Rappaz, *Numerical analysis for nonlinear and bifurcation problems*, Handbook of Numerical Analysis **5** (1997), 487–637.

[20] S. Cano-Casanova, *Existence and structure of the set of positive solutions of a general class of sublinear elliptic non-classical mixed boundary value problem*, Nonlinear Anal. **49** (2002), 361–430.

[21] S. Cano-Casanova and J. López-Gómez, *Varying boundary conditions in a general class of elliptic problems of mixed type*, Nonlinear Anal. **55** (2003), 47-62.

[22] W. Chen and M. Křížek, *What is the smallest possible constant in Céa's lemma?*, Appl. Math. **51** (2006), 2, 128–144.

[23] P. G. Ciarlet and P.-A. Raviart, *Maximum principle and uniform convergence for the finite element method*, Computer Methods in Applied Mechanics and Engineering **2** (1973), 1, 17-31.

[24] J. Douglas and T. Dupont, *A Galerkin method for a nonlinear Dirichlet problem*, Math. Comput. **29** (1975), 689–696.

[25] M. Fiedler and V. Pták, *On matrices with non-positive off-diagonal and positive principal minors*, Czech. Math. J. **12** (1962), 382–400.

[26] B. Finlayson, *The Method of Weighted Residuals and Variational Principles, with Application in Fluid Mechanics*, Heat and Mass Transfer, Volume 87, Academic Press, New York and London, 1972.

[27] B. Finlayson, L. Scriven, *The Method of Weighted Residuals – a Review*, Appl. Mech. Rev. **19** (9) (1966), pp. 735-748.

[28] J. M. Fraile, P. Koch-Medina, J. López-Gómez and S. Merino, *Elliptic eigenvalue problems and unbounded continua of positive solutions of a semilinear elliptic equation*, J. Diff. Eqns. **127** (1996), 295–319.

[29] B. Guo and Y. Wang, *An almost monotone approximation for a nonlinear two-point boundary value problem*, Advances in Computational Mathematics **8** (1998), 65–96.

[30] B. Guo and J. Miller, *Iterative and Petrov-Galerkin methods for a nonlinear two-point boundary value problems with application to semiconductor devices*, Mathematics of Computation **58** (1992), 198, 531–547.

[31] R. A. Horn and C. R. Johnson, *Matrix Analysis*, Cambridge University Press, Cambridge, 1985.

[32] K. Ishihara, *Finite Element Approximations Applied to the Nonlinear Boundary Value Problem* $\Delta u = bu^2$, RIMS, Kyoto Univ. **18** (1982), 17–34.

[33] J. Karátson and S. Korotov, *Discrete maximum principles for finite element solutions of nonlinear elliptic problems with mixed boundary conditions*, Numer. Math. **99** (2005), 669–698.

[34] J. Karátson and S. Korotov, *Some discrete maximum principles arising for nonlinear elliptic finite element problems*, Comput. Math. Appl. **70** (2015), 11, 2732–2741.

[35] H. B. Keller, *Approximation methods for nonlinear problems with applications to two-point boundary value problems*, Math. Comput. **29** (1975), 464–474.

[36] H. B. Keller, *Geometrically isolated nonisolated solutions and their approximation*, SIAM J. Numer. Anal. **18** (1981), 822–838.

[37] J. López-Gómez, *Linear Second Order Elliptic Operators*, World Scientific Publishing, Singapore, 2013.

[38] J. López-Gómez, *Metasolutions of Parabolic Problems in Population Dynamics*, CRC Press, Boca Raton, 2015.

[39] J. López-Gómez, M. Molina-Meyer and M. Villarreal, *Numerical simulation of coexistence states*, SIAM J. Numer. Anal. **29** (1992), 4, 1154–1165.

[40] J. López-Gómez and M. Molina-Meyer, *The maximum principle for cooperative weakly elliptic systems and some applications*, Diff. Int. Eqs. **7** (1994), 383–398.

[41] J. Murray, *Mathematical Biology: I. An Introduction*, Interdisciplinary Applied Mathematics, 3rd edition, Springer-Verlag, New york, 2002.

[42] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables*, Academic Press, New York and London, 1970.

[43] A. Ostrowski, *Über die Determinanten mit überwiegender Hauptdiagonal*, Comment. Math. Helv. **10** (1937), 69–96.

[44] R. Plemmons, *M-matrix characterizations. I: Nonsingular M-matrices*, Linear Algebra and Appl. **18** (1977), 175–188.

[45] J. Pousin and J. Rappaz, *Consistency, stability, a priori and a posteriori errors for Petrov-Galerkin methods applied to nonlinear problems*, Numer. Math. **69** (1994), 213-231.

[46] A. Quarteroni, *Numerical Models for Differential Problems*, Springer Series in Modeling, Simulations & Applications **2**, 2nd edition, Springer-Verlag Italia, 2009.

[47] W. C. Rheinboldt, *Error estimates for nonlinear finite element computations*, Computers & Structures **20** (1985), 91–98.

[48] W. C. Rheinboldt, *Numerical analysis of continuation methods for nonlinear structural problems*, Computers & Structures **13** (1981), 103-113.

[49] R. S. Varga, *Matrix Iterative Analysis*, Springer Series in Computational Mathematics **27**, 2nd edition, Springer-Verlag, Berlin Heidelberg New York, 2000.

[50] J. Xu and L. Zikatanov, *A monotone finite element scheme for convection-diffusion equations*, Math. Comp. **68** (1999), 1429-1446.

[51] E. Zeidler, *Nonlinear Functional Analysis and Its Applications II/A: Linear Monotone Operators*, Springer-Verlag, New York, 1990.