Algorithms for the accurate and efficient solution of fourth order boundary-layer problems

PhD thesis

by

Faiza Alssaedi

Supervisor: Dr Niall Madden

School of Mathematics, Statistics and Applied Mathematics

National University of Ireland, Galway

August 2021

Abstract

In this thesis, we study the analysis and numerical solution of second-order complex-valued reaction-diffusion equations, and two families of fourth-order singularly perturbed problems. The problems are all *singularly perturbed*, meaning that each has a parameter, ε, multiplying the highest derivative. This parameter is positive but maybe arbitrarily small. However, as $\varepsilon \to 0$, the differential equations become ill-posed, hence the singular nature of the perturbation.

The first problem we address is the numerical solution, by finite difference methods of a second-order, complex-valued problem. We employ specialised fitted meshes: the well-known piecewise uniform *Shishkin* mesh, and the graded *Bakhvalov* mesh. The numerical analysis of such methods usually rely on maximum principles, but these do not hold, in a direct way, for complex-valued problems. So we present an approach for rewriting the equation as a coupled system of real-valued problems, and establish that the coefficient matrix for this system is positive definite. Then we show how to adapt the analysis of Bakhvalov [2], in the style of Kellogg et al. [20], to prove convergence.

The second problem we address is the numerical solution of a fourth-order, real-valued reaction-diffusion problem. The ODE is "simply supported" (see Section 1.5.3), and so has boundary conditions that allow it to be transformed into a (weakly) coupled system of second-order reaction-diffusion equations, involving unknowns related to the solution to the fourth-order problem, and its second derivative.

When analysing a finite element method for solving this system, it is usually assumed that the coupling matrix of the coupled system is pointwise coercive. However, we show that the standard transformation (see, e.g., [47]) cannot satisfy this condition. This motivates us to propose a new transformation which resolves this issue.

Moving on to finite difference methods for this problem, we show how to adapt the transformation in a way that leads to a maximum principle-type result. Moreover, we present an iterative scheme for solving the continuous problem in order to derive a stability result for the differential operator. The convergence of the finite difference scheme on a Shishkin mesh, then follows from standard arguments.

Finally, we address the numerical solution using a fourth-order, complex-valued reaction-diffusion problem. We extend the transformation from earlier sections to deal with this case, again focusing on how to ensure coercivity (for a finite element method) and monotonicity (for analysis of a finite difference scheme).

Through all these sections, numerical results are presented that verify the convergence of the schemes, and test if the theoretical orders of convergence are observed in practice.

Contents

List of Figures

List of Tables

Declaration

This thesis is presented in fulfillment of the degree of Doctor of Philosophy requirements. It is entirely my work and has not been submitted to any other university or institute of higher education or for any other academic award in this university. Where use has been made of other people's work, it has been fully acknowledged and referenced.

Faiza Alssaedi

Acknowledgement

First of all, I would like to thank ALLAH who supported me and supplied me with power, patience, and faith to accomplish this thesis. I wish to express my deep gratitude and appreciation to Dr Niall Madden for his guidance, advice, and support throughout the entire thesis, and for his teaching of the methodologies for good scientific research.

Finally, I would like to thank the Ministry of High Studies in Libya and College of Science Scholarship in National University of Ireland, Galway for their funding.

Dedication

To Muhammed and Nidal

Chapter 1

Introduction

1.1 Aims of the thesis

The aim of this thesis to devise new numerical algorithms that efficiently and accurately compute solutions to differential equations whose solutions feature boundary layers. Moreover, we devise new transformations of challenging problems which facilitate numerical analysis of finite difference methods (FDMs) and finite element methods (FEMs).

The choice of problems we study are motivated by models based on the Rayleigh equation (see, e.g., [11]) and the Orr-Sommerfeld equation for hydrodynamic stability, (see, e.g., [11, 25]). The physical meaning of these models is not important to this thesis. What is important is that they are challenging to solve using standard numerical schemes. Therefore, novel methods are required.

The methods we consider are discretization based on finite difference and finite element methods, applied on Shishkin and Bakhvalov fitted meshes. Numerically, we see that these methods are quite successful for the problems we consider. However, the true challenge that these problems present is that their numerical analysis cannot be approached using standard techniques directly. So we have devised novel transforms of the problems into ones for which standard methods and results can be applied.

Here we give a short, chapter-by-chapter summary of the thesis; a detailed overview of the thesis' organisation is postponed until Section 1.4. Chapter 2 serves as an introduction to the area of numerical methods for singularly perturbed problems, by analysing the solution of a second-order complex-valued problem using a finite difference method. As we shall see, this can be re-cast as a coupled system of second-order real-valued problems. Variations on this idea are then developed in the rest of the thesis.

The rest of the thesis is concerned with fourth-order problems. Chapter 3 and 4 address real-valued problems where the boundary conditions allow us to transform the problem into a system of two differential equations. These problems are well-studied in the literature. However, we present a novel approach for constructing the transformation so that it yields a system which is amenable to finite element methods (Chapter 3) and finite difference methods (Chapter 4).

Chapter 5 deals with the mathematical properties of a general fourth-order complex-valued singularly perturbed problems with simple boundary conditions and these boundary conditions let us re-cast the problem into coupled systems of real-valued second-order equations.

Finally, Chapter 6 and Chapter 7, we show how to analyse and solve, numerically, particular fourth-order complex-valued singularly perturbed problems.

We conclude with some observations and suggestions for future work in Chapter 8.

1.2 Notation

We set

$$|\vec{v}| := \sqrt{\vec{v}^T\vec{v}} \quad \text{for } \vec{v} \in \mathbb{R}^2.$$

If $\vec{v} \in \mathbb{R}^{N+1}$, then

$$\|\vec{v}\|_\infty := \max_{j=0,\ldots,N} |v_j|,$$

and

$$\|\vec{v}\|_2 := \sqrt{\sum_j^N v_j^2},$$

If v and u are continuous functions on an interval D, then,

$$(u,v) = \int_D u(x)v(x)dx,$$

$$\|v\|_D := \sup_{x\in D} |v(x)|,$$

$$\|v\|_{2,D} := \sqrt{(v,v)},$$

and

$$|v|_{1,D} := \|v'\|_{2,D},$$

where $D \subset \mathbb{R}$, and most typically is a domain (or its closure), Ω, on which a differential equation is posed, or a subdomain of that. Very often, $D = [0,1]$, and in those cases we omit the D subscript.

Given an arbitrary mesh $\bar{\Omega}^N = \{0 = x_0 < x_1 < \cdots < x_N = 1\}$, and a mesh function V on $\bar{\Omega}^N$, the discrete maximum norm is

$$\|V\|_{\bar{\Omega}^N} := \max_{0\leqslant j\leqslant N} |V_j|.$$

By Ω^N we denote $\bar{\Omega}^N \cap \Omega$, i.e., $\Omega^N = \{x_1, x_2, \ldots, x_{N-1}\}$. Then,

$$\|V\|_{\Omega^N} := \max_{1\leqslant j\leqslant N-1} |V_j|.$$

Finally, C denotes a generic constant that is independent of ε and the mesh. It can take different values at different places.

1.3 Background

1.3.1 Singularly perturbed problems

Singularly perturbed second- and fourth-order differential equations are considered in this thesis. These problems have a small positive parameter ε multiplying the highest derivative. The justification for the name "singular perturbation" is that "the nature of the differential equations changes completely in the limit case when the singular perturbation parameter is equal to zero", to quote directly from [22], which is one of the seminal works in this field. To explain the key ideas, we present two types for singularly perturbed problem: algebraic equations and differential equations. Each type has two examples of perturbation problem, one regular and one singular, to show how their solutions differ as the perturbation parameter approaches zero. The examples are based on a presentation in [30], see also [36, **Section 2.2**].

First, we consider the case for an algebraic equation where the problem is regularly perturbed:

$$f(y,\varepsilon) = (4-\varepsilon)y^2 + 2\varepsilon y - 4 = 0. \tag{1.1a}$$

Its solutions are

$$y(\varepsilon) = \frac{1}{-4+\varepsilon}(\varepsilon \pm \sqrt{\varepsilon^2 - 4\varepsilon + 16}). \tag{1.1b}$$

If we set $\varepsilon = 0$ in (1.1a), and solve for y, we get $y \pm 1$. Alternatively, we can set $\varepsilon = 0$ in (1.1b), and again get $y(0) = \pm 1$. So the perturbation problem is *regular* near $\varepsilon = 0$.

Now, suppose we have a similar problem (1.1a), but with the perturbation parameter multiplying the second-order term

$$f(y,\varepsilon) = 2\varepsilon y^2 + (4-\varepsilon)y - 4 = 0. \tag{1.2a}$$

Its solutions, for $\varepsilon \neq 0$, are

$$y(\varepsilon) = \frac{1}{4\varepsilon}(-4 + \varepsilon \pm \sqrt{\varepsilon^2 + 24\varepsilon + 16}). \tag{1.2b}$$

When $\varepsilon = 0$ in (1.2a), the only solution is $y = 1$. But when $\varepsilon \to 0$ in (1.2b) the solutions tend to 1 and $\pm\infty$. So this perturbation is *singular*. Now, we consider the case of a differential equation that features a regular perturbation:

$$-y'' + 4\varepsilon y - 4 = 0 \quad \text{on} \quad (0,1) \quad \text{with} \quad y(0) = 0 \text{ and } y(1) = 0. \tag{1.3a}$$

Its solution is

$$y(x,\varepsilon) = C_1 e^{2x\sqrt{\varepsilon}} + C_2 e^{-2x\sqrt{\varepsilon}} + \frac{1}{\varepsilon}. \tag{1.3b}$$

where

$$C_1 = \frac{e^{-2\sqrt{\varepsilon}} - 1}{\varepsilon(e^{2\sqrt{\varepsilon}} - e^{-2\sqrt{\varepsilon}})} \quad \text{and} \quad C_2 = -\frac{e^{2\sqrt{\varepsilon}} - 1}{\varepsilon(e^{2\sqrt{\varepsilon}} - e^{-2\sqrt{\varepsilon}})}.$$

When $\varepsilon = 0$ in (1.3a), the solution is $y(x) = -2x^2 + 2x$. And when $\varepsilon \to 0$ in (1.3b) then $\lim_{\varepsilon \to 0} y(x,\varepsilon) = -2x^2 + 2x$. So this a regular perturbation.

In Figure 1.1 we show y with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right); notice that there are essentially identical. In addition, neither solution features a layer.

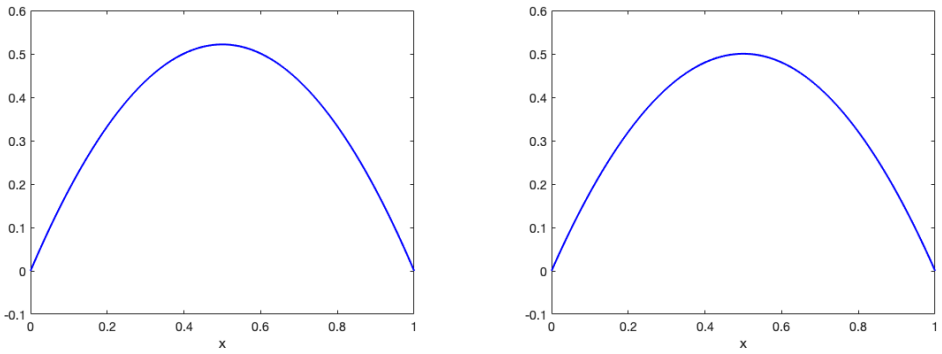

Figure 1.1: The solutions y to (1.3) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

Now, suppose we have a problem that is similar to (1.3a), but with the perturbation parameter multiplying the second derivative term:

$$-\varepsilon y'' + 4y - 4 = 0 \quad \text{on} \quad (0,1) \quad \text{with} \quad y(0) = 0 \text{ and } y(1) = 0. \tag{1.4a}$$

Its solution is

$$y(x,\varepsilon) = C_3 e^{-2x/\sqrt{\varepsilon}} + C_4 e^{2x/\sqrt{\varepsilon}} + 1, \tag{1.4b}$$

where

$$C_3 = \frac{e^{2/\sqrt{\varepsilon}} - 1}{-e^{2/\sqrt{\varepsilon}} + e^{-2/\sqrt{\varepsilon}}} \quad \text{and} \quad C_4 = -\frac{e^{-2/\sqrt{\varepsilon}} - 1}{-e^{2/\sqrt{\varepsilon}} + e^{-2/\sqrt{\varepsilon}}}.$$

When $\varepsilon = 0$ in (1.4a), and solution would have $y(x) \equiv 1$ on $(0,1)$, but $y(0) = y(1) = 0$, which is not possible. That is, (1.4a) is ill-posed if $\varepsilon = 0$. However, considering (1.4b), for example, the left boundary, we note that

$$\lim_{x\to 0}\left(\lim_{\varepsilon\to 0^+} y(x,\varepsilon)\right) = 0,$$

but

$$\lim_{\varepsilon\to 0^+}\left(\lim_{x\to 0^+} y(x,\varepsilon)\right) = 1.$$

In Figure 1.2 we show y with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right); notice that they are completely

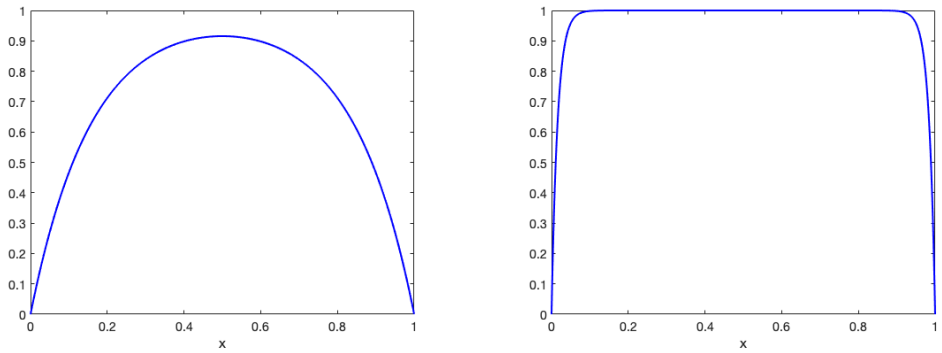

Figure 1.2: The solutions y to (1.4) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

different: for the smaller ε, we see that a layer has formed.

Following these examples, we give a formal definition of a singularly perturbed problem.

Definition 1.3.1. [22] Let P_ε be a problem that depends on a parameter ε, and let y_ε be its solution for a fixed ε. Let y_0 be the solution to P_0. Then we say P_ε is **singularly perturbed**, with respect to the norm $\|\cdot\|_*$, if

$$\lim_{\varepsilon\to 0}\|y_\varepsilon\|_* \neq \|y_0\|_*.$$

1.3.2 Uniform convergence

As seen in Figure 1.2, solutions to singularly perturbed differential equations may change abruptly, and usually exhibits layers at the boundaries and, possibly, also interior regions. Classical methods are not suitable for solving these problems for two reasons. Firstly, the analysis of such methods relies on having bounds on derivatives which are independent of the problem data. Clearly, this is not possible in layer regions. Secondly, such methods may fail to resolve these layers.

Focusing on the first issue, considering classical methods that are not uniformly convergent for (1.4), a careful examination of numerical results shows that for fixed ε, the error may initially decrease as the local (uniform) mesh width decreases, but then usually increases when the mesh is further refined, because of the boundary layer, see [35, **Section 2.1.3**].

The classical bound for a standard finite difference method is

$$\|y - y^N\| \leqslant CN^{-2}\|y''\|, \tag{1.5}$$

where y is the exact solution to a second-order linear singularly perturbed ordinary differential equation such as (1.4a), and y^N is its numerical approximation (see [21, **Chap. 1**]). However, from (1.2b), for example, we can see that

$$\|y''(x)\| = \varepsilon^{-1}|4 - 4y(x)| \leqslant C\varepsilon^{-1}.$$

This does not (necessarily) mean that the error will blow-up as $\varepsilon \to 0$, but, rather that the bound (1.5) is meaningless in the singularly perturbed case. Therefore, (1.5) does not imply convergence of the method unless $N \gg \varepsilon^{-1}$. Since we wish to solve problems for arbitrarily small ε, and the range of N is bounded by the limits of computing capacity, it is not possible to ensure $N \gg \varepsilon^{-1}$ for all ε.

It is very well-known that specialised methods, which are robust with respect to ε, are necessary for the accurate solution of such problems [14, 22, 35, 27].

Definition 1.3.2. [22] Let u_ε be the solution of a singularly perturbed problem, and let u_ε^N be a numerical approximation of u_ε obtained by a numerical method with N degrees of freedom. The numerical method is said to be "**uniformly convergent**" or "**robust** " with respect to the perturbation parameter ε in the norm $\|\cdot\|$ if

$$\|u_\varepsilon - u_\varepsilon^N\| \leqslant \vartheta(N) \text{ for } N \geqslant N_0,$$

with a function ϑ satisfying

$$\lim_{N\to\infty} \vartheta(N) = 0 \text{ and } \partial_\varepsilon \vartheta \equiv 0,$$

and with some threshold value that is independent of ε.

So, the goal of this thesis is to analyse methods which are uniformly convergent for singularly perturbed problems and, moreover, to put in place the necessary analytical tools underpinning such analyses.

1.3.3 A complex-valued example

Later chapters in this thesis are concerned with the solution of fourth-order complex-valued ordinary differential equations. These are somewhat neglected in the literature, so we introduce a classic example here from hydrodynamic stability, the Orr-Sommerfeld equation, which can be used in modelling wave-current interactions [25],

$$\varepsilon \left(\frac{d^2}{dx^2} - k^2 \right)^2 u - i \frac{d^2 u}{dx^2} + i \left(k^2 - a(k,x) \right) u = 0 \quad \text{for} \quad 0 < x < 1, \tag{1.6a}$$

with boundary conditions

$$u(0) = 0, \quad u'(0) = 0, \quad u(1) = u_1, \quad u''(1) = v_1(k), \tag{1.6b}$$

where u_1 is some specified value, and v_1 is a function of k. Notice that if one formally sets $\varepsilon = 0$, then it reduces to second-order, so it is singularly perturbed.

1.4 Organisation

The structure of this thesis is as follows.

In the remaining, final section of Chapter 1, we summarise some of the significant works on the numerical solution of singularly perturbed fourth-order ordinary differential equations in the literature, classified according to their boundary conditions. In Section 1.5.2, we discuss fourth-order problems that can not be re-cast as a system of second-order problems, and in Section 1.5.3 we focus on works on fourth-order problems that can be re-cast as a system of second-order problems; the latter class are particularly of interest for Chapters 3–7.

Although the main contributions of this thesis are on fourth-order problems, we start, in Chapter 2, with studying the robust numerical solution of second-order reaction-diffusion equations. This serves two purposes:

(i) Since our approach for fourth-order problems is to reduce them to second-order systems, we need a clear understanding of the analysis of these problems;

(ii) Since our eventual goal is to study complex-valued differential equations, we start with complex-valued second-order problems.

In Section 2.2, we present a model problem and explain the main properties of its solution by showing the exact solution to a constant coefficient problem. Also, we compare real-valued and complex-valued problems and see how the complex-valued problem differs from the real-valued one. In Section 2.3, we introduce a suitable finite difference discretization. Furthermore, we show how to construct standard Shishkin and Bakhvalov meshes. In Section 2.4, we transform the problem into a system of real-valued equations. Then, we examine this system and prove an error estimate for the computed solution. Finally, numerical results are presented in Section 2.5 as support of the theoretical analysis.

In Chapter 3 we begin our investigation of fourth-order singularly perturbed problems, with a note on the real-valued case, and its solution by a finite element method. In Section 3.2, we present a problem studied by Xenophontos et al. [47]. However, as we show, the analysis of that paper is somewhat faulty: it assumes the differential equation's coefficients are such that a certain coefficient matrix is, in their terminology, (pointwise) positive definite, but we observe that it cannot be for any choice coefficients. We then propose a new transformation for the problem that can resolve this problem. This transformation feature is a parameter that can be tuned, as needed. We determine how this is done to ensures that the mentioned coefficient matrix is, indeed, positive definite (although we prefer the term "coercive"). In Section 3.3, we describe a finite element method for this problem, applied, initially, on an arbitrary mesh. We then present a suitable Shishkin mesh for this problem. The section concludes with a report on our numerical results, for problems with both constant and variable coefficients. We also present results in a selection of norms, including the natural energy norm associated with system's bilinear form, and the discrete maximum norm, which is very commonly used for singularly perturbed problems.

In Chapter 4, we continue our investigation of real-valued fourth-order singularly perturbed problems and their solution using finite difference methods. In Section 4.2, we introduce a family of fourth-order ordinary differential equations, focusing on a problem studied by Shanthi and Ramanujam [38]. As in Chapter 3, we transform the problem into into a system of two second-order differential equations; the actual transformation is a slight simplification of that used in Chapter 3. In Section 4.3 we present a stability result for the continuous problem, which is analysed through a novel iterative method that we have proposed. In Section 4.4, we describe a finite difference method for this problem, with a suitable layer-adapted mesh, and we show the numerical analysis for this method. Finally, numerical results are presented in Section 4.5, in support of the theoretical analysis.

In Chapter 5, we extend our work from Chapters 3 and 4, and apply it to a general fourth-order, but now *complex-valued*, singularly perturbed problem. Specifically, we study a problem of the form

$$-\varepsilon u^{(4)}(x) + (1+i)a(x)u''(x) - (1+i)b(x)u(x) = f(x) \quad \text{on} \quad \Omega := (0,1), \tag{1.7}$$

with the homogeneous boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0.$$

In Section 5.2, we present a general fourth-order complex-valued problem, discuss how it can be solved using the MATLAB Chebfun toolbox [12], and present a motivating example. In Section 5.3 we present ways of rewriting the problem in terms of real-valued systems. First, in Section 5.3.1,

we transform our model problem into a system of two fourth-order real-valued problems. Then, in Section 5.3.2 (following the exposition in Chapter 3) we present a further transformation into a real-valued second-order system. Again we show how to solve this problem with Chebfun, and verify that all three formulations are essentially equivalent. In Section 5.4 we show how to determine the value of the parameters in the transformation (subject to reasonable assumptions) that ensure that the resulting coefficient matrix of the system's zero-order term is coercive. Such a result is very important, especially in the context of finite element analysis; see Chapter 6. In Sections 5.4.2 we present a general framework for applying an eigenvalue analysis to verify coercivity, and then in Section 5.4.3 we show how to apply this to specific cases of interest. In Section 5.5 we tackle a different form of analysis of the differential operator. That is, we establish the stability result of differential operator for the system of four equations solved using a Gauss-Seidel method. Such stability results are key to proving the convergence of finite difference methods: see Chapter 7. The approach is two-fold: first we present a new block-iterative version of the Gauss-Seidel, which we prove to converge. Then we extend this to a fully iterative method, in order to give precise bounds on each of the solution's components.

Finite element methods are the main theme of Chapter 6. We, again, show that a special transformation is required in order for a finite element analysis to work. In Section 6.2, we show how to build on the work of Chapter 5 to ensure that the coupling matrix is coercive. This follows some of the methodologies as Section 3.2.2, but the details are entirely different. In Section 6.3, we describe a finite element method for this problem, applied, initially, on an arbitrary mesh. We then present a suitable layer-adapted mesh, and we present the numerical analysis for this method. Finally, in Section 6.4, numerical results are presented that investigate the convergence and robustness of the method.

The final substantial chapter of the thesis is Chapter 7, where we consider a subclass of the problems introduced in Chapter 5, and their solution using *finite difference methods*. Specifically, we study

$$-\varepsilon u^{(4)}(x) + a(1+\zeta i)u''(x) - b(1+i)u(x) = (f_r + if_i)(x) \quad \text{on} \quad \Omega := (0,1),$$

with the same boundary conditions as applied to (1.7). In essence, a finite difference method is more demanding than a finite element method on the properties of the coupling matrix, so the transformation used is more general than that of Chapter 6; this is discussed in Section 7.2. In Section 7.3, we establish the stability result of the differential operator for the system of four equations solved using a Gauss-Seidel method, based on ideas in Section 5.6, leading to bounds on the coefficients which ensure convergence of the iterative method, bounds on the solutions to the iterates, and a maximum/minimum principle. In Section 7.4, we describe a finite difference method for this problem, with a suitable layer-adapted mesh, and we discuss the numerical analysis for this method. We conclude, in Section 7.5, with two examples to verify the sharpness of the analysis outlined in Section 7.4.3.

In Chapter 8, we conclude with a review of the main results of this thesis. We also outline some related open problems, potential topics for future research.

1.5 Literature Review for singularly perturbed fourth-order ordinary differential equations

1.5.1 Introduction

The majority of the published works on the numerical solution of singularly perturbed boundary value problems are focused on second-order problems. There have been some papers on first-order problems (usually systems), and a relatively small number of higher-order equations (i.e., order three or more). In this short survey, we list some of the important papers on the numerical solution of fourth-order reaction-diffusion boundary-value problems, and discuss, where appropriate, how they relate to this study.

Broadly speaking, these fourth-order problems may be classified into two types: those which can easily be rewritten as systems of two second-order problems, and those for which this is not possible. More specifically, they can be classified according to their boundary conditions: where u is the dependent variable in the boundary value problem, and we can have

Case 1: Fourth-order problems that can not be re-cast as a system of second-order problems; most typically u and u' are specified at the boundary. These are called a "clamped" problems in the seminal paper of Semper [37], and are discussed in Section 1.5.2.

Case 2: Fourth-order problems that can be re-cast as a system of second-order problems, because the boundary conditions specify u and u''. These are called "simply supported" problems in [37], and are discussed in Section 1.5.3.

We will briefly survey both of these. However, our main interest in Case 2. So, we give just a brief overview of the first case.

1.5.2 Case 1

One of the earliest papers on the numerical analysis of fourth-order reaction-diffusion ordinary differential equation was by Roos and Stynes [34]. They considered the problem

$$-\varepsilon u^{(4)} + (a(x)u')' - b(x)u' - c(x)u = f(x), \quad 0 < x < 1, \tag{1.8a}$$

with zero- and first-order boundary conditions

$$u(0) = u'(0) = u(1) = u'(1) = 0. \tag{1.8b}$$

The functions a, b, c, and f are assumed to be sufficiently smooth with

$$a(x) \geqslant \alpha > 0, \tag{1.9a}$$

$$c(x) - \frac{1}{2}b'(x) \geqslant \beta > -\alpha. \tag{1.9b}$$

Based on the asymptotic analysis of O'Malley [29], they give that

$$|u^{(k)}(x)| \leqslant C(1+\varepsilon^{(1-k)/2}e^{-x/\sqrt{\varepsilon}}). \tag{1.10}$$

The numerical approach is quite novel: the authors consider the problem in a variational setting, and replace the coefficients with piecewise polynomial approximations. Specifically, on a particular mesh, they use piecewise constant approximations, which gives an approximation to the differential equation that is first-order in the maximum norm. However, no numerical results are presented.

Sometime later, Sun and Stynes [45] began the study of arbitrarily high-order singularly perturbed reaction-diffusion problems. They studied the topic of high order elliptic two-point boundary value problems of reaction-diffusion type of the form

$$L_\varepsilon u = (-1)^{m+1}\varepsilon u^{(2m)} + (-1)^m(a_{2(m-1)}u^{(m-1)})^{(m-1)} + L_1 u = f(x),$$

$$u^{(j)}(0) = u^{(j)}(1) = 0, \quad \text{for} \quad j = 0,\ldots,m-1,$$

where $m \geqslant 2$ is an integer, and

$$L_1 u = \sum_{k=2}^{m}(-1)^{m-k+1}(a_{2(m-k)+1}u^{(m-k+1)} + a_{2(m-k)}u^{(m-k)})^{(m-k)}.$$

The functions a_r for $r = 0,\ldots,2(m-1)$ and f are assumed to be sufficiently smooth on $[0,1]$, with

$$a_{2(m-1)}(x) > \alpha > 0 \quad \text{on} \quad [0,1],$$

and

$$a_{2(m-k)}(x) - \frac{1}{2}a'_{2(m-k)}(x) > \alpha_{m-k} \quad \text{for} \quad k = 2,\ldots,m,$$

for all $x \in [0,1]$ and some constants $\alpha_{m-1} = \alpha$ and $\alpha_{m-k}(k = 2,\ldots,m)$ satisfying

$$\sum_{i=1}^{k}\alpha_{m-i} > 0, \quad \text{for} \quad k = 2,\ldots,m.$$

In the case where $m = 2$, we get (1.8) by setting $a_2 = a$, $a_1 = b$, and $a_0 = c$.

The numerical scheme analysed in [45] is a Galerkin finite-element method with piecewise polynomial basis functions, applied on a fitted Shishkin mesh. They present a uniform convergence result in a weighted energy norm. Numerical results are presented for a fourth-order problem (i.e., $m = 2$) with

$$a(x) = 1 + x(1-x), \quad b(x) = c(x) = 0, \tag{1.11a}$$

and f is chosen so that the solution to (1.8) is

$$u(x) = \sqrt{\varepsilon}\{\frac{e^{(-x/\sqrt{\varepsilon})} + e^{(-(1-x)/\sqrt{\varepsilon})}}{1+e^{(-1/\sqrt{\varepsilon})}} - 1\} + \frac{1-e^{(-1/\sqrt{\varepsilon})}}{1+e^{(-1/\sqrt{\varepsilon})}}x(1-x) + x^2(1-x)^2. \tag{1.11b}$$

Guo, Huang and Zhang [15] considered a version of (1.8), but the assumptions on the problem

data are different to (1.9). The coefficient functions a, b and f are assumed to satisfy

$$a(x) \geqslant \alpha > 0, \quad \text{and} \quad c(x) - b'(x)/2 \geqslant \beta \geqslant 0,$$

for all $x \in [0,1]$. The use a conforming finite element method (FEM) of (fixed) degree p applied on a Shishkin mesh, they were able to prove a superconvergence error bound of $(N^{-1}\ln(N+1))^p$ in a discrete energy norm. They present results for the example in (1.11), and also one for which

$$a(x) = 1, \quad b(x) = c(x) = 0, \quad \text{and} \quad f(x) = -1,$$

so, the exact solution is

$$u(x) = \sqrt{\varepsilon}\frac{e^{-(1-x/\sqrt{\varepsilon})} + e^{(-x/\sqrt{\varepsilon})} - 1 - e^{(-1/\sqrt{\varepsilon})}}{2 - 2e^{(-1/\sqrt{\varepsilon})}} + \frac{1}{2}x(1-x).$$

Panaseti et al. [31] considered a version of (1.8) with $b = 0$, and different assumptions on the coefficient functions.

$$a(x) > 0, \text{ and } \quad c(x) \geqslant 0. \tag{1.12}$$

They analysed a *hp*-FEM, which means both the local mesh width (h) and local polynomial degree (p) are allowed to vary (by contrast, in this thesis I focus on the standard h-FEM where one fixes $p = 1$ and allow only h to vary). They showed that this method applied on the Spectral Boundary Layer Mesh gives a robust approximation that converges exponentially in an energy norm. They present two examples: one with $a = c = f = 1$, and the exact solution is available, and one with variable coefficients $a(x) = e^{-x}$, $c(x) = 0$, $f(x) = e^{-x^2+1}$, for which the exact solution is not available.

Constantinou et al. [10] considered a version of (1.8) with $b = 0$. The coefficient functions a, c and f, which are assumed to be analytic on $x \in [0,1]$, but with assumptions on the problem data that are very slightly different to (1.12), specifically

$$a(x) > 0, \text{ and } \quad c(x) > 0.$$

Again, a *hp*-FEM on the Spectral Boundary Layer Mesh gives robust exponential convergence in a stronger, more balanced norm. Also, they get robust exponential convergence in maximum norm. They present results for the same examples in [31], and also the case where

$$a(x) = c(x) = 1, \quad f(x) = (x + 1/2)^{-1}.$$

In recent work, Xenophontos [46] considered a version of (1.8) with $b = 0$. He solved it using *a standard FEM* with piecewise Hermite polynomials of (fixed) degree $p \geqslant 3$ defined on an exponentially graded mesh. He showed that the method converges uniformly with respect to ε, and presented results for the same problems considered in [31].

1.5.3 Case 2

Recall that a "simply supported" fourth-order equation can be re-cast as a system of second-order problems because the boundary conditions specify u and u''. All of the papers reviewed here employ this strategy. They are relevant to the work presented in Chapters 3 and 4.

Shanthi and Ramanujan [38] studied fourth-order singularly perturbed reaction-diffusion two-point boundary value problem on the form

$$-\varepsilon u^{(4)}(x) + a(x)u''(x) - b(x)u(x) = f(x) \qquad \text{on} \qquad x \in (0,1), \tag{1.13a}$$

subject to the boundary conditions

$$u(0) = p,\ u''(0) = -r, \qquad u(1) = q,\ u''(1) = -s. \tag{1.13b}$$

The coefficient functions satisfy the following conditions

$$a(x) \geqslant \beta > 0, \tag{1.14a}$$

$$0 \geqslant b(x) \geqslant \gamma, \quad \gamma > 0, \tag{1.14b}$$

$$\beta - 2\gamma \geqslant k > 0 \quad \text{for some } k. \tag{1.14c}$$

By using the boundary conditions they transformed the problem into a system of two differential equations. They propose the transformation

$$w := -u''. \tag{1.15}$$

With this, (1.13) can be transformed into a system of two equations of the form

$$-E\vec{z}'' + A\vec{z} = \vec{f}, \tag{1.16a}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix},\ E = \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix},\ A = \begin{pmatrix} a & b \\ -1 & 0 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} -f \\ 0 \end{pmatrix}. \tag{1.16b}$$

The assumptions in (1.14) ensure the system satisfies a maximum principle, which is the key to analysis. The authors solve the system using a combination of asymptotic and numerical techniques, which combines a classical finite difference scheme and an exponentially fitted finite difference scheme.

The components of the solution to (1.16a), and their derivatives, may be bounded as

$$|u^{(k)}(x)| \leqslant C[1 + \varepsilon^{1-k/2}e(x,\beta)], \tag{1.17a}$$

and

$$|w^{(k)}(x)| \leqslant C[1 + \varepsilon^{-k/2}e(x,\beta)], \tag{1.17b}$$

where

$$e(x,\beta) = e^{-x\sqrt{\beta/\varepsilon}} + e^{-(1-x)\sqrt{\beta/\varepsilon}}.$$

Numerical results are presented for three linear fourth-order problems. The first has a constant right-hand side, and no reaction term:

$$a(x) = 4, \quad b(x) = 0, \quad \text{and} \quad f(x) = -1,$$

with the boundary conditions

$$u(0) = 1, u''(0) = -1, \qquad u(1) = 1,\ u''(1) = -1.$$

For the second example,

$$a(x) = 4,\ b(x) = -1,$$

and

$$f(x) = -\frac{1}{16}\big(2x(1-x) - 5\varepsilon + \frac{5\varepsilon(e^{-2x/\sqrt{\varepsilon}} - e^{-2(x+1)/\sqrt{\varepsilon}} + e^{-2(1-x)/\sqrt{\varepsilon}} - e^{-2(2-x)/\sqrt{\varepsilon}})}{1 - e^{-4/\sqrt{\varepsilon}}}\big),$$

with the boundary conditions

$$u(0) = 1, u''(0) = -1, \qquad u(1) = 1,\ u''(1) = -1.$$

The third example has

$$a(x) = 4, \quad b(x) = 1, \tag{1.18}$$

and

$$f(x) = -2 - (x(1-x)/8) - (5\varepsilon/16) - (5\varepsilon/16)\frac{e^{-2x/\sqrt{\varepsilon}} - e^{-2(x+1)/\sqrt{\varepsilon}} + e^{-2(1-x)/\sqrt{\varepsilon}} - e^{-2(2-x)/\sqrt{\varepsilon}}}{1 - e^{-4/\sqrt{\varepsilon}}},$$

with the boundary conditions

$$u(0) = 1, u''(0) = -1, \qquad u(1) = 1,\ u''(1) = -1.$$

They also consider the semilinear problem

$$-\varepsilon u^{(4)}(x) + 4u''(x) + u^2(x) = -f(x),$$

with the boundary conditions

$$u(0) = 1, u''(0) = -1, \qquad u(1) = 1,\ u''(1) = -1,$$

and $f(x)$ the same as in the second example.

Notice that the example (1.18) does not satisfy (1.14b), since c is positive. To deal with this, they propose a so-called adjoint system for (1.16a) as

$$-K\hat{z}'' + D\hat{z} = \vec{F}, \tag{1.19a}$$

where

$$\hat{z} = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{pmatrix}, \; K = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & \varepsilon \end{pmatrix}, \; D = \begin{pmatrix} 0 & -1 & 0 & 0 \\ c^- & a & -c^+ & 0 \\ 0 & 0 & 0 & -1 \\ -c^+ & 0 & c^- & a \end{pmatrix}, \; \vec{F} = \begin{pmatrix} 0 \\ -f \\ 0 \\ f \end{pmatrix}, \tag{1.19b}$$

and

$$c^- := c - c^+, \qquad c^+ = \begin{cases} c & \text{if } c \geqslant 0 \\ 0 & \text{if otherwise} \end{cases}$$

Shanthi and Ramanujan [39], describe the boundary value technique method to solve singularly perturbed boundary value problems for fourth-order equations of the type

$$-\varepsilon u^{(4)}(x) + a(x)u'''(x) + b(x)u''(x) - c(x)u(x) = -f(x) \qquad \text{on} \qquad x \in (0,1), \tag{1.20}$$

with a version of boundary conditions on (1.13b). The coefficient functions $a(x)$, $b(x)$, $c(x)$, and $f(x)$ are sufficiently smooth and satisfying the following conditions:

$$a(x) \geqslant \alpha > 0,$$

$$b(x) \geqslant \beta \geqslant 0,$$

$$0 \geqslant c(x) \geqslant \gamma, \quad \gamma > 0,$$

$$\alpha - 2\gamma(1 + \Delta) \geqslant k > 0 \quad \text{for some } k \text{ and } \Delta > 0.$$

The differential equation's domain is divided into two non-overlapping subintervals. The differential equation is solved in these intervals separately. The solutions obtained in these regions are combined to give a solution in the entire interval. Zero-order asymptotic expansions are used to get boundary values inside this interval. The method is applied to both linear and nonlinear equations, the latter resolved using Newton's method.

Numerical results are presented for a fourth-order problem with

$$a(x) = 4, \quad b(x) = c(x) = 0 \quad \text{and} \quad f(x) = 1,$$

Its equivalent system is decoupled. They also consider an example with

$$a(x) = 0, \quad b(x) = 4, \quad c(x) = 1,$$

Its equivalent system is a weakly coupled system. Their final example is the semilinear problem

$$-\varepsilon u^{(4)}(x) - 4u'''(x) + 4u''(x) + u^2(x) = -f(x).$$

Shanthi and Ramanujan [40], considered a version of (3.1), but with different assumptions on the data

$$a(x) \leqslant -\alpha < 0,$$

$$b(x) \geqslant \beta \geqslant 0,$$

$$0 \geqslant c(x) \geqslant \gamma, \quad \gamma > 0,$$

$$\alpha - \theta\gamma \geqslant k > 0 \quad \text{for some } \theta \text{ arbitrary close to 1.}$$

The proposed method involves a zero-order asymptotic approximation of the solution to the weakly coupled system, to construct a type of decoupling of the first equation. Then the second equation is solved by a fitted the numerical method involving a Shishkin mesh. An example is presented as follows.

$$-\varepsilon u^{(4)}(x) - 4u'''(x) + 4u''(x) = -f(x),$$

where

$$f(x) = \begin{cases} 0.7 & \text{for } 0 \leqslant x \leqslant 0.5, \\ -0.6 & \text{for } 0.5 \leqslant x \leqslant 1. \end{cases}$$

Chandru and Shanthi [8] considered a version of (2.2), but with different assumptions on the data

$$b(x) \geqslant \beta > 0,$$

$$0 \geqslant c(x) \geqslant \gamma, \quad \gamma > 0,$$

$$\beta - \theta\gamma \geqslant k > 0 \quad \text{for some } \theta \text{ arbitrarily close to 2.}$$

Using a computational method for solving this system and involving non-overlapping Schwarz method applied on a Shishkin mesh. An example presented has

$$b(x) = \begin{cases} 2x + 1, & \text{for } x \leqslant 0.5, \\ 2(1 - x) + 1 & \text{for } x > 0.5, \end{cases} \quad c(x) = 0.1, \quad f(x) = \begin{cases} -0.5, & \text{for } x \leqslant 0.5, \\ 0.5 & \text{for } x > 0.5. \end{cases}$$

Xenophontos et al. [47] is one of the few studies to consider both the clamped and simply supported cases. For the latter, the problem is transformed into a second-order system and then solved using a *hp*-FEM on the Spectral Boundary Layer Mesh resulting in the exponential convergence in the energy norm. Also, they studied the clamped case, which cannot be transformed into a system, which is again solved with a *hp*-FEM. The example presented is the same as in [31].

Chapter 2

Second-order complex-valued reaction-diffusion equations

2.1 Introduction

2.1.1 A model problem

In this chapter, we are interested in the numerical solution of a singularly perturbed, second-order, complex-valued reaction-diffusion equation. Our model differential equation is: find $u \in C^2[0,1]$ such that

$$Lu := -\varepsilon^2 u'' + bu = f \qquad \text{on} \qquad \Omega := (0,1), \tag{2.1a}$$

subject to the boundary conditions

$$u(0) = u_0, \qquad u(1) = u_1. \tag{2.1b}$$

Here ε is a positive, real-valued parameter. We assume $0 < \varepsilon \leqslant 1$, but typically have that $\varepsilon \ll 1$. The coefficient function b and right-hand side function f are complex valued functions on the real interval Ω. That is, $b : \Omega \to \mathcal{C}$, and $f : \Omega \to \mathcal{C}$. Furthermore, we assume that $b, f \in C^4(0,1) \cup C[0,1]$ (see, e.g., [33, **Remark 5.1.2**]).

2.1.2 A motivating example

We consider the following example: find $u \in C^2(\bar{\Omega})$ such that

$$-\varepsilon^2 u'' + (i+4)^2 u = (4+4i)e^x \quad \text{on } \Omega = (0,1), \qquad u(0) = 0, \qquad u(1) = 0. \tag{2.2}$$

The exact solution can be expressed as

$$u(x) = C_1 e^{\frac{-(4+i)x}{\varepsilon}} + C_2 e^{\frac{(4+i)(x-1)}{\varepsilon}} + \frac{(4+4i)e^x}{-\varepsilon^2 + 15 + 8i}, \tag{2.3}$$

where

$$C_1 = -\frac{4(i - ie^{\frac{(\varepsilon-4-i)}{\varepsilon}} + 1 - e^{\frac{(\varepsilon-4-i)}{\varepsilon}})}{(-\varepsilon^2 + 15 + 8i)(1 - e^{\frac{(-8-2i)}{\varepsilon}})} \quad \text{and } C_2 = \frac{4(ie^{\frac{(-4-i)}{\varepsilon}} - ie + e^{\frac{(-4-i)}{\varepsilon}} - e)}{(-\varepsilon^2 + 15 + 8i)(1 - e^{\frac{(-8-2i)}{\varepsilon}})}.$$

The first two terms on the right-hand side of (2.3) correspond to the left and right layer, respectively.

In Figure 2.1 we show u with $\varepsilon = 1$ (left), which does not features layers. In contrast, as shown in the graph on the right for smaller ε (in this case $\varepsilon = 0.1$), the solution possesses boundary layers near $x = 0$ and $x = 1$, in both the real and the imaginary parts.

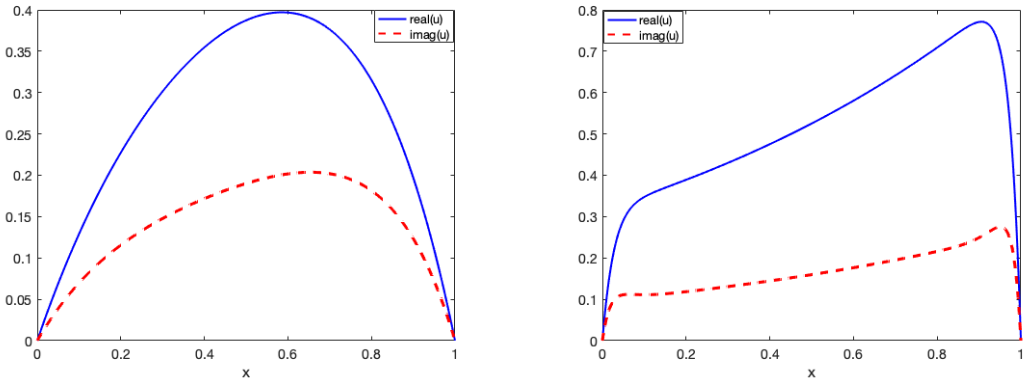

Figure 2.1: Real and imaginary parts of the solutions to (2.2) with $\varepsilon = 1$ (left) and $\varepsilon = 0.1$ (right).

2.1.3 Comparing Real-Valued and Complex-Valued Problems

This section outlines how complex-valued problems have properties that are different from their real-valued analogues. In particular, differential operators associated with many real-valued problems satisfy a maximum principle, a valuable tool for their analysis and the analysis of associated numerical schemes. However, these principles do not usually directly apply to complex-valued problems.

We begin by defining a maximum principle. Using examples of two simple cases, one for a real-valued problem and another for a complex-valued problem, we will then illustrate how the real-valued differential operator satisfies a maximum principle, but the complex-valued one does not.

Definition 2.1.1. The differential operator L satisfies a **maximum principle**, if $\psi(0) \geqslant 0$ and $\psi(1) \geqslant 0$, and $L\psi(x) \geqslant 0$, for all $x \in \Omega$, imply that $\psi(x) \geqslant 0$, for all $x \in \bar{\Omega}$ [27].

When an operator satisfies a maximum principle, solutions to associated differential equations may be analysed using barrier function techniques. For example, let us consider an analysis that gives upper and lower bounds for the solution to a real-valued differential equation. Having done so, we can then progress to establishing an upper bound for the solution by using the barrier function techniques. Establishing upper and lower bound gives the stability of the operator and,

in particular, excludes oscillations from the solution.

If $Lv \geqslant 0$, $v(0) \geqslant 0$ and $v(1) \geqslant 0$, then, we know that $v \geqslant 0$. Moreover, suppose the constant k is such that

$$L(k \pm v) \geqslant 0, \quad k \pm v(0) \geqslant 0 \quad \text{and } k \pm v(1) \geqslant 0.$$

Then,

$$k \pm v \geqslant 0.$$

So, it must be that $\|v\|_{\bar{\Omega}} \leqslant k$.

To see how a problem such as (2.1) differs from the real-valued case, consider the example

$$-\varepsilon^2 u'' + u = 0 \qquad u(0) = 1, \qquad u(1) = 0. \tag{2.4}$$

The solution is $u(x) \cong e^{-x/\varepsilon}$, which is positive and monotonic; see Figure 2.2. The associated differential operator satisfies a maximum principle; for a proof, see [27, **Chap. 6**].

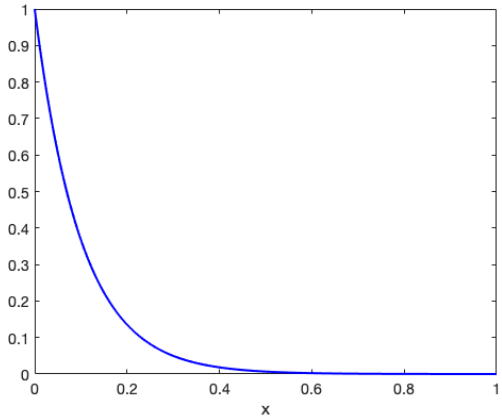

Figure 2.2: The solution to (2.4) with $\varepsilon = 0.1$.

Now consider the complex-valued problem:

$$-\varepsilon^2 u'' + (1 + \frac{i}{2})^2 u = 0 \qquad u(0) = 1 + i, \qquad u(1) = 0. \tag{2.5}$$

The solution shown in Figure 2.3 is

$$u(x) = e^{-x/\varepsilon}(\cos(x/2\varepsilon) + \sin(x/2\varepsilon)) + ie^{-x/\varepsilon}(\cos(x/2\varepsilon) - \sin(x/2\varepsilon)) + \mathcal{O}(e^{-1/\varepsilon}).$$

Notice that the imaginary part of u is neither positive nor monotonic. The associated differential operator does not satisfy a maximum principle, in the conventional sense, and the solution can oscillate.

Nonetheless, we can apply ideas based on maximum principles. Specifically, in Section 4, we show how to rewrite (2.1) as a system of real-valued problems. Of course, this system does not satisfy a maximum principle itself, but we follow the example of Kellogg et al. [20] and use an ingenious idea due to Bakhvalov to construct a related problem that does satisfy a maximum principle [2].

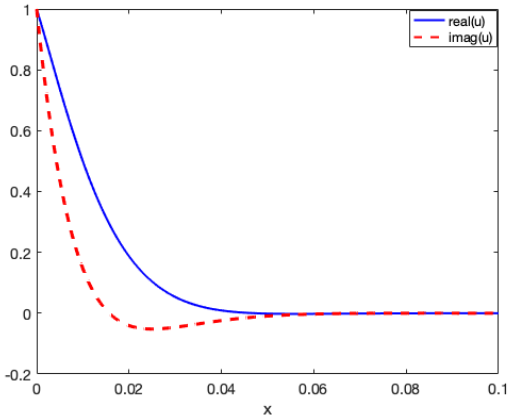

Figure 2.3: The solution to (2.5) with $\varepsilon = 0.01$.

2.1.4 Outline

The structure of this Chapter is as follows. In Section 2.2, we give bounds on the solution to (2.2) and its derivatives. In Section 2.3, we describe and analyse the finite difference method. Furthermore, we show how to construct standard Shishkin and Bakhvalov meshes. In Section 2.4, we transform the problem into a system of reaction-diffusion equations. We then examine the continuous and discrete versions of this system and prove an error estimate for the computed solution. Finally, numerical results are presented in Section 2.5 in support of the theoretical analysis. Some standard technical results used are derived in full detail in Appendix 2.A.

2.2 Analysis of the continuous problem

Lemma 2.2.1. *Let u be the solution of (2.1). Then, for $0 \leqslant k \leqslant 4$,*

$$||u^{(k)}||_{\bar{\Omega}} \leqslant C(1 + \varepsilon^{-k}). \tag{2.6}$$

Proof. Rearrange (2.1) as

$$u'' = \varepsilon^{-2}(bu - f).$$

From this, we have the bound

$$||u''||_{\bar{\Omega}} \leqslant C\varepsilon^{-2}.$$

Next, we differentiate equation (2.1) with respect to x to obtain

$$-\varepsilon^2 u''' + bu' + b'u = f'.$$

To derive a bound for u', we use the following construction, which is based on [27, **Lemma 6.1**]. Let $x \in \Omega$ and construct an associated neighbourhood $N_x = (a, a + \sigma) \subseteq \Omega$, that contains x. From the Mean Value Theorem, there is a $y \in \bar{N}_x$ such that

$$u'(y) = \frac{u(a + \sigma) - u(a)}{\sigma}.$$

So, since $\|u\|_{\bar{\Omega}} \leqslant C$,

$$|u'(y)| \leqslant C\sigma^{-1}.$$

Next, we use the Fundamental Theorem of Calculus:

$$\int_y^x g'(s)ds = g(x) - g(y).$$

Therefore

$$u'(x) = u'(y) + \int_y^x u''(s)ds.$$

Also, by the Mean Value Theorem for Integrals

$$\int_y^x u''(s)ds = (x-y)u''(q) \quad \text{for some} \quad q \in N_x.$$

Therefore

$$\begin{aligned} |u'(x)| &\leqslant |u'(y)| + |x-y||u''(q)| \\ &\leqslant C\sigma^{-1} + C\sigma\varepsilon^{-2}, \end{aligned}$$

holds for any $\sigma \in (0,1)$. But the bound is sharpest if we take $\sigma = \varepsilon$, giving

$$|u'(x)| \leqslant C\varepsilon^{-1}.$$

The bounds on the higher derivatives are obtained by using the differential equation and bounds on u and u'. □

2.3 The numerical method

2.3.1 The finite difference method

Consider an arbitrary mesh, $\Omega^N := \{0 = x_0 < x_1 < \cdots < x_N = 1\}$, where $h_j = x_j - x_{j-1}$. Suppose we want to approximate $u'(x_j)$ by a finite difference approximation based only on values of u at a finite number of points near x_j. One obvious choice is to use the forward difference approximation

$$D^+U_j := \frac{U_{j+1} - U_j}{h_{j+1}}.$$

Note that D^+U_j is the slope of the line joining (x_j, U_j) and (x_j, U_{j+1}). Another one-sided approximation is the backward difference approximation

$$D^-U_j := \frac{U_j - U_{j-1}}{h_j}.$$

Finally, we have the centred approximation, which averages D^-U_i and D^+U_i,

$$D^0U_j := \frac{1}{2}(D^+U_j + D^-U_j) = \frac{1}{2}\left(\frac{U_{j+1}}{h_{j+1}} + U_j(\frac{1}{h_j} - \frac{1}{h_{j+1}}) - \frac{U_{j-1}}{h_j}\right).$$

A standard second-order approximation of the second derivative is

$$\delta^2 U_j := D^+ D^- U_j = \frac{1}{\hbar_j}\left(\frac{U_{j-1}}{h_j} - U_j(\frac{1}{h_j} + \frac{1}{h_{j+1}}) + \frac{U_{j+1}}{h_{j+1}}\right), \tag{2.7}$$

where $\hbar_j = (x_{j+1} - x_{j-1})/2$.

We are particularly interested in the operator δ^2. Using Taylor series, it can be shown that

$$|u''(x_j) - \delta^2(x_j)| = \begin{cases} C|u''(\alpha_j)|, & \text{(2.8a)} \\ C(h_{j+1} - h_j)|u'''(\alpha_j)|, & \text{(2.8b)} \\ C\big((h_{j+1} - h_j)|u'''(x_j)| - (h_j^2 - h_j h_{j+1} + h_{j+1}^2)|u^{(4)}(\alpha_j)|\big), & \text{(2.8c)} \end{cases}$$

for some $\alpha_j \in [x_{j-1}, x_{j+1}]$. See the Appendix for details.

The finite difference operator is defined as

$$L^N \psi_j := -\varepsilon\delta^2\psi_j + b(x_j)\psi_j \quad \text{for} \quad j = 1, \ldots, N-1.$$

The finite difference method is

$$\begin{gathered} U_0 = u_0, \\ -\varepsilon^2\delta^2 U_j + b(x_j)U_j = f(x_j), \quad \text{for all} \quad x_j \in \Omega^N, \\ U_N = u(1). \end{gathered} \tag{2.9}$$

Notice that we can treat u as a mesh function. In particular u_j and $u(x_j)$ represent the same quantity, and we use whichever is most convenient.

2.3.2 Shishkin mesh

We construct a standard *Shishkin* mesh with the mesh parameter

$$\tau_S = \min\{\frac{1}{4}, 2\frac{\varepsilon}{\varrho}\ln N\},$$

where $0 < \varrho^2 \leqslant \min \Re(b(x))$. We now define two mesh transition points at $x = \tau_S$ and $x = 1 - \tau_S$. That is, we form a piecewise uniform mesh with $N/4$ equally-sized mesh intervals on each of $[0, \tau_S]$ and $[1 - \tau_S, 1]$, and $N/2$ equally-sized mesh intervals on $[\tau_S, 1 - \tau_S]$. Typically, when ε is small, $\tau_S \ll 1/4$, the mesh is very fine near the boundaries, and coarse in the interior; see Figure 2.4.

Remark 2.3.1. In this chapter, we use τ_S to represent the Shishkin transition point, whereas in later chapters we simply refer to it as τ. This is because, uniquely, this chapter considers fitted meshes of both Shishkin and Bakhvalov meshes. Following standard notation, for Bakhvalov mesh, we use τ to represent the point in the co-domain of the mesh generating function where the transition to a uniform mesh takes place.

The mesh may also be specified in terms of a mesh generating function, which we now define.

Definition 2.3.1. [22, p5] A strictly monotone function $\varphi : [0, 1] \to [0, 1]$ that maps a uniform

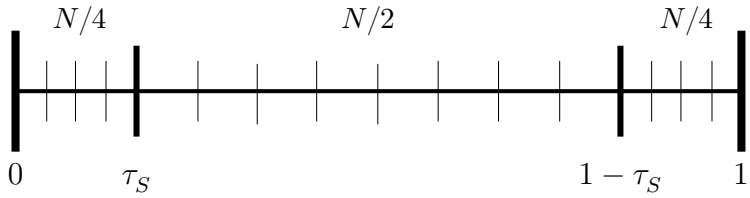

Figure 2.4: The piecewise uniform Shishkin mesh Ω^N.

mesh $t_j = j/N, i = 0, \ldots, N$, onto a layer-adapted mesh by $x_j = \varphi(t_j)$, $j = 0, \ldots, N$, is called a **mesh generating function**.

The mesh generating function φ, for Shishkin mesh described above, is

$$\varphi(t) = \begin{cases} 4t\tau_S & t \leqslant \frac{1}{4}, \\ 2(1-\tau_S)(t-\frac{1}{4}) + 2\tau_S(\frac{3}{4}-t) & \frac{1}{4} < t < \frac{3}{4}, \\ 4(1-\tau_S)(1-t) + 4(t-\frac{3}{4}) & t \geqslant \frac{3}{4}. \end{cases}$$

Notice that this is a piecewise linear function.

2.3.3 Bakhvalov mesh

We construct a standard *Bakhvalov* mesh with mesh parameters at $\sigma > 0$, $q \in (0, 1/2)$, typical values of the mesh parameters are $\sigma = 2$, $q = 1/4$, where the method has order σ, and q is the proportion of mesh points in the layer. Mesh points are $x_j = j/N$ if $\sigma\varepsilon \geqslant \varrho q$. However, when $\sigma\varepsilon < \varrho q$ one sets

$$x_j = \begin{cases} \varphi(j/N) & \text{for } j \leqslant N/2, \\ 1 - \varphi((N-j)/N) & \text{for } j > N/2, \end{cases} \tag{2.10}$$

with a mesh generating function φ defined by

$$\varphi(t) = \begin{cases} \chi(t) := -\frac{\sigma\varepsilon}{\varrho}\ln(1-\frac{t}{q}) & \text{for } t \in [0, \tau], \\ \pi(t) := \chi(\tau) + \chi'(\tau)(t-\tau) & \text{for } t \in [\tau, 1/2]. \end{cases} \tag{2.11}$$

From (2.11) we can say that $\varphi \in C^1[0,1]$. Moreover, for (2.11) to generate a mesh on [0,1/2] we want $\varphi(1/2) = 1/2$. That will happen if τ solves the nonlinear equation

$$\chi'(\tau) = \frac{1-2\chi(\tau)}{1-2\tau}. \tag{2.12}$$

The generating function given in (2.11), defines the mesh on $[0, 1/2]$ and it is extended to $[0, 1]$ by reflection about $x = 1/2$, as shown in (2.10).

The mesh is very fine and graded near the boundary, but coarse in the sense that $h_j = \mathcal{O}(N^{-1})$ in the interior. Also, for later analysis, it is useful to note that $\tau < q$, because $\ln(q - \tau/q)$ is undefined when $q - \tau \leqslant 0$.

Remark 2.3.2. Our numerical analysis for the finite difference method on this mesh requires that

$\varphi'(\tau) > 1$. This is stated without proof in [2] and [20]. To see why it must be true note that

$$\int_0^1 \varphi'(t)dt = \varphi(1) - \varphi(0) = 1.$$

But $\varphi'(t) \leqslant \varphi'(\tau)$ for all t. Clearly

$$\int_0^\tau \varphi'(t)dt + \int_{1-\tau}^1 \varphi'(t)dt < 2\tau\varphi'(\tau), \quad \text{and} \quad \int_\tau^{1-\tau} \varphi(t)dt = (1-2\tau)\varphi'(\tau).$$

So, if $\varphi'(\tau) \leqslant 1$, then

$$\int_0^1 \varphi(t)dt < 0,$$

which is not possible.

The mesh generating functions for both the Shishkin and Bakhvalov meshes are shown in Figure 2.5.

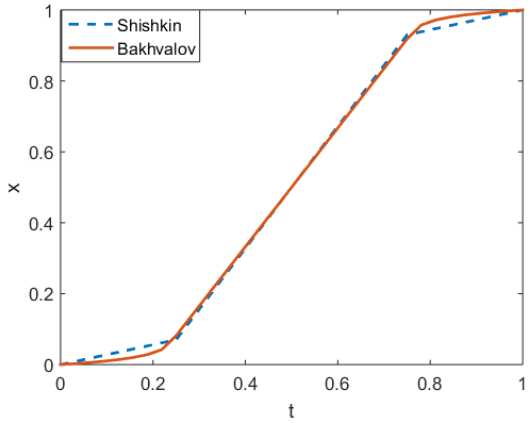

Figure 2.5: Mesh generating functions for Shishkin and Bakhvalov meshes.

2.4 A system of reaction-diffusion equations

2.4.1 The continuous problem

We now consider the following system by rewriting (2.1) as

$$-\varepsilon^2(u_r + iu_i)'' + (b_r + ib_i)(u_r + iu_i) = f_r + if_i, \tag{2.13}$$

where $u = u_r + iu_i$, $b = b_r + ib_i$ and $f = f_r + if_i$, and all of u_r, u_i, b_r, f_r and f_i are real-valued. It is assumed that

$$\alpha := \min_{0 \leqslant x \leqslant 1} (\sqrt{b_r}) > 0.$$

From (2.13), when we equate real terms and imaginary terms separately, we get

$$-\varepsilon^2 u_r'' + b_r u_r - b_i u_i = f_r, \tag{2.14a}$$

$$-\varepsilon^2 u_i'' + b_i u_r + b_r u_i = f_i. \tag{2.14b}$$

We can write this system as

$$\vec{L}\vec{u} := -\varepsilon^2 \vec{u}'' + B\vec{u} = \vec{f}, \tag{2.15}$$

where

$$\vec{u} = \begin{pmatrix} u_r \\ u_i \end{pmatrix}, \; B = \begin{pmatrix} b_r & -b_i \\ b_i & b_r \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f_r \\ f_i \end{pmatrix}.$$

Since $u = u_r + iu_i$, Lemma 2.2.1 gives that

$$\|\vec{u}^{(k)}\|_{\bar{\Omega}} \leqslant C(1 + \varepsilon^{-k}).$$

Lemma 2.4.1. *Assume that $b_r > 0$. Then the matrix B is coercive, meaning that there exists a constant α such that $\sqrt{b_r} \geqslant \alpha > 0$ and*

$$\vec{v}^T B\vec{v} \geqslant \alpha^2 \vec{v}^T\vec{v} \quad \textit{for all } \vec{v} \in \mathbb{R}^2. \tag{2.16}$$

Proof. Assume $\vec{v} = (v_1 \; v_2)$, Then

$$\vec{v}^T B\vec{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} b_r & -b_i \\ b_i & b_r \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix}.$$

So

$$\vec{v}^T B\vec{v} = b_r v_1^2 - b_i v_1 v_2 + b_i v_1 v_2 + b_r v_2^2 = b_r(v_1^2 + v_2^2) = b_r \vec{v}^T\vec{v}.$$

This implies that the matrix B is coercive, with $\alpha \leqslant \sqrt{b_r}$. □

Lemma 2.4.2. *Let $\vec{w} \in C^2(\Omega)^2 \cap C(\bar{\Omega})^2$, Then*

$$\|\vec{w}\|_{\bar{\Omega}} \leqslant \alpha^{-2}\|\vec{L}\vec{w}\|_{\Omega} + \|\vec{w}(0)\| + \|\vec{w}(1)\|. \tag{2.17}$$

Proof. Set $v = \frac{1}{2}\vec{w}^T\vec{w}$, and note that $(\vec{w}^T\vec{w})'' = 2(\vec{w}^T\vec{w}'') + 2(\vec{w}')^T\vec{w}'$. Thus,

$$-2\vec{w}^T\vec{w}'' = -(\vec{w}^T\vec{w})'' + 2|\vec{w}'|^2 \geqslant -(\vec{w}^T\vec{w})'' = -2v''. \tag{2.18}$$

Taking the scalar product of $\vec{w}^T$ with $-\varepsilon^2\vec{w}'' + B\vec{w} = \vec{L}\vec{w}$, we get

$$-\varepsilon^2\vec{w}^T\vec{w}'' + \vec{w}^T B\vec{w} = \vec{w}^T\vec{L}\vec{w}.$$

Then invoking (2.16) and (2.18), we get

$$\begin{aligned}\vec{w}^T\vec{L}\vec{w} &= -\varepsilon^2\vec{w}^T\vec{w}'' + \vec{w}^T B\vec{w}\\ &\geqslant -\varepsilon^2 v'' + \alpha^2\vec{w}^T\vec{w}\\ &\geqslant -\varepsilon^2 v'' + 2\alpha^2 v.\end{aligned}$$

Let us denote the differential operator on the right-hand side by

$$L_\alpha v := -\varepsilon^2 v'' + 2\alpha^2 v.$$

Since $\alpha > 0$, this operator satisfies the maximum principle of Definition 2.1.1.

Also, clearly $|v|_{\{0,1\}} \leqslant (\|\vec{w}(0)\|^2 + \|\vec{w}(1)\|^2)/2$. Now the standard maximum principle of Definition 2.1.1 for the scalar problem, with a constant barrier function

$$k = \frac{1}{2\alpha^2}\|\vec{w}\|_{\bar\Omega}\|\vec{L}\vec{w}\|_\Omega + \frac{1}{2}(\|\vec{w}(0)\|^2 + \|\vec{w}(1)\|^2),$$

gives

$$\|v\|_{\bar\Omega} \leqslant \frac{1}{2\alpha^2}\|\vec{w}\|_{\bar\Omega}\|\vec{L}\vec{w}\|_\Omega + \frac{1}{2}(\|\vec{w}(0)\|^2 + \|\vec{w}(1)\|^2).$$

This implies that

$$\|\vec{w}\|_{\bar\Omega}^2 = 2\|v\|_{\bar\Omega} \leqslant \|\vec{w}\|_{\bar\Omega}(\alpha^{-2}\|\vec{L}\vec{w}\|_\Omega + \|\vec{w}(0)\| + \|\vec{w}(1)\|),$$

because $\|\vec{w}(0)\| + \|\vec{w}(1)\| \leqslant \|\vec{w}\|_{\bar\Omega}$. Dividing by $\|\vec{w}\|_{\bar\Omega}$ gives

$$\|\vec{w}\|_{\bar\Omega} \leqslant \alpha^{-2}\|\vec{L}\vec{w}\|_\Omega + \|\vec{w}(0)\| + \|\vec{w}(1)\|.$$

□

We now give sharp pointwise bounds on u and its derivatives. The argument is essentially the same as in [**20, Lemma 2.3**], but simplified to the one-dimensional case.

Lemma 2.4.3. *Let $\vec{u}$ be the solution to (2.15). Let $\varrho \in (0, \sqrt{b_r})$ be arbitrary but fixed. Then there exists a constant C, which is independent of ε, such that*

$$|\vec{u}^{(k)}(x)| \leqslant C[1 + \varepsilon^{-k}(e^{-\varrho x/\varepsilon} + e^{-\varrho(1-x)/\varepsilon})], \quad \textit{for all} \quad x \in \bar\Omega, \tag{2.19}$$

and $k = 0, 1, \ldots, 4$.

Proof. For $k = 0$, this is just Lemma 4.2. Otherwise, we proceed by induction. Fix $\varrho \in (0, \sqrt{b_r})$ and set $B_k(x) = 1 + \varepsilon^{-k}(e^{-\varrho x/\varepsilon} + e^{-\varrho(1-x)/\varepsilon})$. For $k = 1, 2, 3, 4$, differentiating (2.15) k times with respect to x gives

$$-\varepsilon^2\vec{u}^{(k+2)} + B\vec{u}^{(k)} = \vec{f}^{(k)} - \sum_{l=0}^{k-1}\binom{k}{l}B^{(k-l)}\vec{u}^{(l)} =: \vec{\varphi}_k(x),$$

with $|\vec{\varphi}_k(x)| \leqslant CB_{k-1}(x)$ where the bound on φ_k is a consequence of the inductive hypothesis.

Define $\hat{u}$ by $\vec{u}^{(k)} = B_k\hat{u}$. Consider (2.15)

$$-\varepsilon^2\vec{u}'' + B\vec{u} = \vec{f}.$$

Then

$$-\varepsilon^2(B_k\hat{u})'' + B(B_k\hat{u}) = \vec{\varphi}_k,$$

and

$$-\varepsilon^2 B_k\hat{u}'' - 2\varepsilon^2 B_k'\hat{u}' - \varepsilon^2 B_k''\hat{u} + BB_k\hat{u} = \vec{\varphi}_k.$$

Dividing by B_k gives

$$-\varepsilon^2\hat{u}'' - 2\varepsilon^2\frac{B_k'}{B_k}\hat{u}' + \big(B - \varepsilon^2 I\frac{B_k''}{B_k}\big)\hat{u} = \frac{\vec{\varphi}_k}{B_k}. \tag{2.20}$$

Take the scalar product of $\hat{u}^T$ with (2.20), we get

$$-\varepsilon^2\hat{u}^T\hat{u}'' - 2\varepsilon^2\hat{u}^T\frac{B_k'}{B_k}\hat{u}' + \hat{u}^T\big(B - \varepsilon^2 I\frac{B_k''}{B_k}\big)\hat{u} = \hat{u}^T\frac{\vec{\varphi}_k}{B_k}.$$

Set $v_k = (\hat{u}^T\hat{u})/2 = |\hat{u}|^2/2$, while noting that $(\hat{u}^T\hat{u})'' = 2(\hat{u}^T\hat{u}'') + 2(\hat{u}')^T\hat{u}'$. Thus,

$$-2\hat{u}^T\hat{u}'' = -(\hat{u}^T\hat{u})'' + 2|\hat{u}'|^2 \geqslant -(\hat{u}^T\hat{u})'' = -2v_k''.$$

Then invoking (2.16) and (2.20), we get

$$-\varepsilon^2 v_k'' - 2\varepsilon^2\frac{B_k'}{B_k}v_k' + 2(b_r - \varrho^2)v_k \leqslant C\|\hat{u}\|_{\bar{\Omega}},$$

because $B_K''(x) \leqslant \varepsilon^2\varrho^2 B_k(x)$ and $\vec{\varphi}_k(x) \leqslant CB_k(x)$. Boundary conditions for v_k follow from Lemma 2.2.1, so now the standard maximum principle for scalar problem gives

$$\frac{1}{2}\|\hat{u}\|_{\bar{\Omega}}^2 = \|v_k\|_{\bar{\Omega}} \leqslant C\|\hat{u}\|_{\bar{\Omega}}.$$

Dividing by $\|\hat{u}\|_{\bar{\Omega}}$ gives $\|\hat{u}\|_{\bar{\Omega}} \leqslant C$. □

2.4.2 The discrete problem

The finite difference method for equation (2.15) is: find $\vec{U}$ such that

$$\vec{L}^N\vec{U}_j := -\varepsilon^2\delta^2\vec{U}_j + B\vec{U}_j = \vec{f}_j \quad \text{for} \quad j = 1, ..., N-1, \tag{2.21}$$
$$\vec{U}_0 = \vec{u}(0), \quad \vec{U}_N = \vec{u}(1),$$

where δ^2 is as defined in (2.7), $\vec{U}_j$ is the approximation for $\vec{u}(x_j)$, and the mesh, for now, is arbitrary.

Lemma 2.4.4. *The discrete operator $\vec{L}^N$ satisfies the stability inequality*

$$\|\vec{W}\|_{\bar{\Omega}} \leqslant \alpha^{-2}\|\vec{L}^N\vec{W}\|_{\Omega} + \|\vec{W}(0)\| + \|\vec{W}(1)\|,$$

for arbitrary vector-valued functions $\vec{W}$ defined on $\bar{\Omega}$.

Proof. Let $V = \frac{1}{2}\vec{W}^T\vec{W}$. Note that

$$\delta^2\vec{W}_j = \frac{1}{\hbar_j}\left(\frac{\vec{W}_{j-1}}{h_j} - \vec{W}_j(\frac{1}{h_j} + \frac{1}{h_{j+1}}) + \frac{\vec{W}_{j+1}}{h_{j+1}}\right).$$

So

$$\vec{W}_j^T\delta^2\vec{W}_j = \frac{1}{\hbar_j}\left(\frac{\vec{W}_j^T\vec{W}_{j-1}}{h_j} - \vec{W}_j^T\vec{W}_j(\frac{1}{h_j} + \frac{1}{h_{j+1}}) + \frac{\vec{W}_j^T\vec{W}_{j+1}}{h_{j+1}}\right).$$

One can show that, in fact,

$$\vec{W}_j^T\delta^2\vec{W}_j = \frac{1}{2}\delta^2(\vec{W}^T\vec{W})_j - \frac{1}{2\hbar_j}\left(\frac{|\vec{W}_{j+1} - \vec{W}_j|^2}{h_{j+1}} + \frac{|\vec{W}_j - \vec{W}_{j-1}|^2}{h_j}\right). \tag{2.22}$$

To see this starting from (2.22),

$$\begin{aligned}
&\frac{1}{2\hbar_j}\left(\frac{(\vec{W}^T\vec{W})_{j-1}}{h_j} - (\vec{W}^T\vec{W})_j(\frac{1}{h_j} + \frac{1}{h_{j+1}}) + \frac{(\vec{W}^T\vec{W})_{j+1}}{h_{j+1}}\right)\\
&\qquad - \frac{1}{2\hbar_j}\left(\frac{(\vec{W}_{j+1} - \vec{W}_j)^T(\vec{W}_{j+1} - \vec{W}_j)}{h_{j+1}} + \frac{(\vec{W}_j - \vec{W_{j-1}})^T(\vec{W}_j - \vec{W_{-1}})}{h_{j+1}}\right)\\
&\qquad = \frac{1}{2\hbar_j}\left(\frac{(\vec{W}^T\vec{W})_{j-1}}{h_j} - \frac{(\vec{W}^T\vec{W})_j}{h_j} - \frac{(\vec{W}^T\vec{W})_j}{h_{j+1}} + \frac{(\vec{W}^T\vec{W})_{j+1}}{h_{j+1}}\right)\\
&\qquad\quad - \frac{1}{2\hbar_j}\left(\frac{\vec{W}_{j+1}^T\vec{W_{j+1}}}{h_{j+1}} + \frac{\vec{W}_j^T\vec{W}_j}{h_{j+1}} - \frac{\vec{W}_j^T\vec{W_{j+1}}}{h_{j+1}} - \frac{\vec{W}_{j+1}^T\vec{W}_j}{h_{j+1}}\right.\\
&\qquad\qquad \left. + \frac{\vec{W}_j^T\vec{W}_j}{h_j} + \frac{\vec{W}_{j-1}^T\vec{W_{j-1}}}{h_j} - \frac{\vec{W}_j^T\vec{W_{j-1}}}{h_j} - \frac{\vec{W}_{j-1}^T\vec{W}_j}{h_j}\right)\\
&\qquad = \frac{1}{\hbar_j}\left(\frac{(\vec{W}^T\vec{W})_j}{2h_j} - \frac{\vec{W}_j^T\vec{W}_j}{2h_j} - \frac{(\vec{W}^T\vec{W})_j}{2h_{j+1}} - \frac{\vec{W}_j^T\vec{W}_j}{2h_{j+1}} + \frac{\vec{W}_j^T\vec{W}_{j+1}}{h_{j+1}} + \frac{\vec{W}_j^T\vec{W}_{j-1}}{2h_j}\right)\\
&\qquad\quad = \frac{1}{\hbar_j}\left(\frac{\vec{W}_j^T\vec{W}_{j-1}}{h_j} - \vec{W}_j^T\vec{W}_j(\frac{1}{h_j} + \frac{1}{h_{j+1}}) + \frac{\vec{W}_j^T\vec{W}_{j+1}}{h_{j+1}}\right)\\
&\qquad\qquad = \vec{W}_j^T\delta^2\vec{W}_j.
\end{aligned}$$

It follows immediately from (2.22) that

$$\vec{W}_j^T\delta^2\vec{W}_j \geqslant \delta^2 V.$$

Using (2.16), we get

$$-\varepsilon^2\delta^2 V + 2\alpha^2 V \leqslant \vec{W}^T\vec{L}^N\vec{W} \quad \textbf{on} \quad \Omega^N,$$

and

$$|V(0)| + |V(1)| \leqslant \frac{1}{2}\big(\|\vec{W}(0)\| + \|\vec{W}(1)\|\big)^2.$$

A standard discrete maximum principle for scalar problems, with a constant barrier function

$$K = \frac{1}{2\alpha^2}\|\vec{W}\|_{\bar{\Omega}^N}\|\vec{L}^N\vec{W}\|_{\Omega^N} + \frac{1}{2}\big(\|\vec{W}(0)\| + \|\vec{W}(1)\|\big)^2,$$

yields

$$\|V\|_{\bar{\Omega}^N} \leqslant \frac{1}{2\alpha^2}\|\vec{W}\|_{\bar{\Omega}^N}\|\vec{L}^N\vec{W}\|_{\Omega^N} + \frac{1}{2}\big(\|\vec{W}(0)\| + \|\vec{W}(1)\|\big)^2.$$

Hence

$$\|\vec{W}\|_{\bar{\Omega}^N}^2 = 2\|V\|_{\bar{\Omega}^N} \leqslant \|\vec{W}\|_{\bar{\Omega}^N}\left(\alpha^{-2}\|\vec{L}^N\vec{W}\|_{\Omega^N} + \|\vec{W}(0)\| + \|\vec{W}(1)\|\right).$$

Dividing by $\|\vec{W}\|_{\bar{\Omega}}$ gives

$$\|\vec{W}\|_{\bar{\Omega}^N} \leqslant \alpha^{-2}\|\vec{L}^N\vec{W}\|_{\Omega^N} + \|\vec{W}(0)\| + \|\vec{W}(1)\|.$$

□

The first of our two main results of this chapter is the following theorem, which shows that applying the finite difference method on a Bakhvalov mesh gives fully second-order ε-uniform convergence.

Theorem 2.4.5. *Let Ω^N be the Bakhvalov mesh defined in Section 2.3.3, and let $\vec{U}$ be the solution to (2.21) on this mesh. Then, if $\vec{u}$ solves (2.15),*

$$\|\vec{u} - \vec{U}\|_{\bar{\Omega}^N} \leqslant CN^{-2}. \tag{2.23}$$

Proof. Let $\vec{\eta} = \vec{u} - \vec{U}$ denote the error. Lemma 2.4.4 yields

$$\begin{aligned}\|\vec{u} - \vec{U}\|_{\bar{\Omega}^N} \leqslant \alpha^{-2}\|\vec{L}^N(\vec{u} - \vec{U})\|_{\Omega^N} &= \alpha^{-2}\|\vec{L}^N\vec{u} - \vec{L}^N\vec{U}\|_{\Omega^N} \\ &= \alpha^{-2}\|\vec{L}^N\vec{u} - \vec{f}\|_{\Omega^N} \leqslant \alpha^{-2}\|\vec{L}^N\vec{u} - \vec{L}\vec{u}\|_{\Omega^N},\end{aligned} \tag{2.24}$$

because $\vec{u}(0) = \vec{U}(0)$ and $\vec{u}(1) = \vec{U}(1)$. From the definitions of $\vec{L}$ and $\vec{L}^N$, we get

$$\vec{L}^N\vec{u} - \vec{L}\vec{u} = -\varepsilon^2\delta^2\vec{u} + \varepsilon^2\vec{u}''.$$

Then, from (2.8b),

$$(\vec{L}^N\vec{u} - \vec{L}\vec{u})(x_j) = \varepsilon^2\left(-\frac{(h_{j+1} - h_j)}{3}\vec{u}'''(x_j) - \frac{(h_j^2 - h_jh_{j+1} + h_{j+1}^2)}{12}\vec{u}^{(4)}(\gamma_j)\right).$$

Consequently,

$$\|\vec{L}^N\vec{u} - \vec{L}\vec{u}\|_{\bar{\Omega}^N} \leqslant \varepsilon^2\left(C(h_{j+1} - h_j)\|\vec{u}'''\|_{\bar{\Omega}} + C(h_j^2 - h_jh_{j+1} + h_{j+1}^2)\|\vec{u}^{(4)}\|_{\bar{\Omega}}\right). \tag{2.25}$$

We construct a standard Bakhvalov mesh with mesh parameters at $\sigma > 0$, $q \in (0, 1/2)$. If $\sigma\varepsilon \geqslant \varrho q$ the mesh is uniform with mesh size N^{-1}. Furthermore $\varepsilon^{-1} \leqslant C$. Thus

$$\varepsilon^2\|\delta^2\vec{u} - \vec{u}''\| \leqslant CN^{-2},$$

by Lemma 2.2.1 and (2.25).

We next examine the case $\sigma\varepsilon/\varrho < q$. To do so, we shall only consider the analysis where j is such that $x_j = \varphi(t_j) \leqslant 1/2$, which includes only the layer at x. The argument for $x_j > 1/2$ are essentially the same.

From the construction of φ, one must have $\tau < q$. We start by showing that

$$1 < \chi'(\tau) < \frac{1}{1 - 2q}. \tag{2.26}$$

From Remark 2.3.2, $\chi'(\tau) > 1$. Next, recall that

$$\chi(t) = -\frac{\sigma\varepsilon}{\varrho}\ln(1-\frac{t}{q}),$$

and $\chi(t)$ at the mesh parameter τ is

$$\chi(\tau) = -\frac{\sigma\varepsilon}{\varrho}\ln(\frac{q-\tau}{q}).$$

From (2.12), we have

$$\chi'(\tau) < \frac{1+2q\ln(\frac{q-\tau}{q})}{1-2\tau} < \frac{1}{1-2\tau} < \frac{1}{1-2q},$$

since $2q\ln((q-\tau)/q) > 0$ and $\tau < q$. This implies that

$$\chi'(\tau) < \frac{1}{1-2q} =: \hat{q},$$

which establishes (2.26).

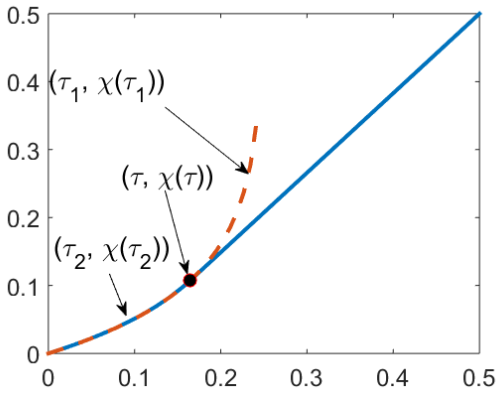

Figure 2.6: A plot of the function φ and the location of τ, τ_1 and τ_2.

Let us introduce two points, τ_1, τ_2, both in $(0, q)$ and defined so that $\chi'(\tau_1) = \hat{q}$ and $\chi'(\tau_2) = 1$; as we shall see, these bracket τ, as shown in Figure 2.6.

Recall that, for $0 \leqslant t \leqslant \tau$,

$$\chi(t) = -\frac{\sigma\varepsilon}{\varrho}\ln(1-\frac{t}{q}),$$

so that

$$\chi'(t) = \frac{\sigma\varepsilon}{\varrho(q-t)}.$$

Because $\chi'(\tau_1) = 1/(1-2q)$, then

$$\chi'(\tau_1) = \frac{\sigma\varepsilon}{\varrho(q-\tau_1)} = \frac{1}{1-2q},$$

so,

$$\tau_1 = q - \frac{\sigma\varepsilon(1-2q)}{\varrho}.$$

A similar calculation shows that

$$\tau_2 = q - \frac{\sigma\varepsilon}{\varrho}.$$

We know that $\chi'(\tau) > 1$ and $\chi'(\tau_2) = 1$. Since $\chi''(\tau) > 1$, then χ' is a strictly increasing function. Therefore $\chi'(\tau) > \chi'(\tau_2)$ for $\tau > \tau_2$. Similarly $\tau_1 > \tau$. This shows that

$$\tau_2 = q - \sigma\varepsilon/\varrho < \tau < \tau_1 = q - \sigma\varepsilon(1-2q)/\varrho.$$

To bound all terms in (2.25), we separately analyse the layer regions, the interior, and transitional regions between these.

Since $\varphi(t) = \chi(\tau)$ for $t \in [0,\tau]$ and $\varphi'(t) = \chi'(\tau)$ for $t \in [\tau, 1/2]$, so $\varphi'(t) \leqslant \chi'(\tau) \leqslant \hat{q}$ for $t \in [0,1]$. Thus,

$$h_j = x_j - x_{j-1} = \varphi(t_j) - \varphi(t_{j-1}) = \int_{t_{j-1}}^{t_j} \varphi'(t)dt \leqslant \hat{q}(t_j - t_{j-1}) = \hat{q}N^{-1}, \tag{2.27}$$

for $j = 1, \ldots, N$.

Of course, in the region closest to the layer, the mesh width is very fine. Specifically, if j is such that $t_j < q$, then, for $t < t_j$, we have that $\varphi'(t) \leqslant \chi'(\tau) = \sigma\varepsilon/\varrho(q-t) \leqslant \sigma\varepsilon/\varrho(q-t_j)$. Hence,

$$h_j = \int_{t_{j-1}}^{t_j} \varphi'(t)dt \leqslant N^{-1}\varphi'(t_j) \leqslant N^{-1}\frac{\sigma\varepsilon}{\varrho(q-t_j)} \leqslant N^{-1}\frac{2\sigma\varepsilon}{\varrho(q-t_{j-1})} \quad \text{for } t_j \leqslant q - N^{-1}. \tag{2.28}$$

The difference between two adjacent mesh sizes on $[\tau, 1-\tau]$ is $h_{j+1} - h_j = x_{j+1} - 2x_j + x_{j-1} = \varphi''(t_j^*)N^{-2}$ for some $t_j^* \in [t_{j-1}, t_{j+1}]$. Now

$$\varphi''(t) \leqslant \chi''(\tau) = \frac{\sigma\varepsilon}{\varrho(q-\tau)^2} \quad \text{and} \quad \frac{1}{q-\tau} \leqslant \frac{1}{q-\tau_1} = \frac{\varrho\hat{q}}{\sigma\varepsilon},$$

which gives

$$|h_{j+1} - h_j| \leqslant \frac{\varrho\hat{q}}{\sigma\varepsilon}N^{-2}. \tag{2.29}$$

Also, we have to bound the difference between adjacent mesh sizes on $[0,\tau]$. In this region

$$\varphi''(t_j^*) \leqslant \frac{\sigma\varepsilon}{\varrho(q-t_{j-1})^2} \leqslant \frac{4\sigma\varepsilon}{\varrho(q-t_j)^2} \quad \text{for } t_j \leqslant q - 2N^{-1}, \tag{2.30}$$

which yields

$$|h_{j+1} - h_j| \leqslant \frac{4\sigma\varepsilon}{N^2\varrho(q-t_j)^2} \quad \text{for } t_j \leqslant q - 2N^{-1}. \tag{2.31}$$

From the mesh generating function, we have

$$e^{-\varrho x_j/\varepsilon} = \left(\frac{q-t_j}{q}\right)^\sigma \quad \text{for } t_j \leqslant \tau, \tag{2.32}$$

and

$$e^{-\varrho x_j/\varepsilon} \leqslant \left(\frac{\sigma\varepsilon}{\varrho q}\right)^\sigma \quad \text{for } t_j \geqslant \tau_2. \tag{2.33}$$

Recalling the assumption that $\sigma \geqslant 2$, and using (2.25),(2.19), (2.27) and (2.33), we get

$$\varepsilon^2|(\delta^2\vec{u} - \vec{u}'')(x_j)| \leqslant CN^{-2} \quad \text{for } \tau_2 \leqslant t_{j-1},$$

which is the region outside the layer.

If j is such that $t_j \leqslant q - 2N^{-1}$, and consequently, the corresponding x_j is the layer region, using (2.25) and (2.19), we get

$$\varepsilon^2\|[\delta^2\vec{u} - \vec{u}''](x_j)\| \leqslant C\varepsilon^2|h_{j+1} - h_j| + C\varepsilon^{-1}|h_{j+1} - h_j|e^{-\varrho x_j/\varepsilon} \\ + C\varepsilon^2|h_j^2 - h_jh_{j+1} + h_{j+1}^2| + C\varepsilon^{-2}|h_j^2 - h_jh_{j+1} + h_{j+1}^2|e^{-\varrho x_{j-1}/\varepsilon}.$$

To bound the first term, use (2.29); for the second term, use (2.31) and (2.32); for the third term, use (2.27); and for the fourth, use (2.28), (2.32) and $q - t_{j-1} \leqslant 3(q - t_j)/2$. This yields

$$\varepsilon^2\|[\delta^2\vec{u} - \vec{u}''](x_j)\| \leqslant CN^{-2} \quad \text{for } t_j \leqslant q - 2N^{-1}.$$

If j is such that $t_j > q - 2N^{-1}$ and $\tau_2 > t_{j-1}$, the corresponding x_j is the transition region. Thus,

$$q - \frac{2}{N} < t_j < \tau_2 + \frac{1}{N} = q - \frac{\sigma\varepsilon}{\varrho} + \frac{1}{N} < q + \frac{1}{N}.$$

It is clear that the first two inequalities here imply that $\varepsilon < 3\varrho/(\sigma N)$. Use (2.8a):

$$\varepsilon^2\|[\delta^2\vec{u} - \vec{u}''](x_j)\| \leqslant C(\varepsilon^2 + e^{-\varrho x_{j-1}/\varepsilon}) \leqslant CN^{-2},$$

by (2.33) and $\varepsilon \leqslant CN^{-1}$.

Consequently, we get the bound for the truncation error in the maximum norm on the Bakhvalov mesh. Then, from (2.24), the bound in (2.23) follows immediately. □

Moving on from Theorem 2.4.5, we will prove an error estimate for the solution on the Shishkin mesh.

Theorem 2.4.6. *Let Ω^N be the Shishkin mesh defined in Section 2.3.2, and let $\vec{U}$ be the solution to (2.21) on this mesh. If $\vec{u}$ solves (2.15), then*

$$\|\vec{u} - \vec{U}\|_{\bar{\Omega}^N} \leqslant CN^{-2}\ln^2 N. \tag{2.34}$$

Proof. Let us first consider the case where ε is so large that $\tau_S = 1/4$, and, so, the mesh is uniform with mesh size $h_j = N^{-1}$ for all j. Then

$$\frac{1}{4} \leqslant \sigma\varepsilon\varrho^{-1}\ln N,$$

and, so, $\varepsilon^{-1} \leqslant 4\sigma\varrho^{-1}\ln N \leqslant C\ln N$. Hence (4.21) and (2.25) give

$$\|\vec{L}^N\vec{u} - \vec{L}\vec{u}\|_{\Omega^N} \leqslant CN^{-2}\ln^2 N.$$

Now consider the case where ε is small enough (relative to N) so that $\tau_S = \sigma\varepsilon\varrho^{-1}\ln N \leqslant 1/4$.

As a consequence, there exist a point $x^* = 2\varepsilon\varrho^{-1}\ln N \in [0, 1/2]$, for which we may define

$$v(x) = \begin{cases} \sum_{i=0}^{4} \frac{(x-x^*)}{i!} u^{(i)}(x^*) & \text{for } 0 \leqslant x \leqslant x^*, \\ u(x) & \text{for } x^* \leqslant x \leqslant 1-x^*, \\ \sum_{i=0}^{4} \frac{(x-x^*)}{i!} u^{(i)}(1-x^*) & \text{for } 1-x^* \leqslant x \leqslant 1, \end{cases}$$

and $w(x) = u(x) - v(x)$. Then Lemma 2.2.1, and the choice of x^* demonstrate that

$$\|v^{(k)}(x)\|_{\bar{\Omega}} \leqslant C(1 + \varepsilon^{2-k}), \tag{2.35}$$

and

$$\|w^{(k)}(x)\|_{\bar{\Omega}} \leqslant C\varepsilon^{-k}(e^{-\varrho x/\varepsilon} + e^{-\varrho(1-x)/\varepsilon}) \text{ for } 0 \leqslant k \leqslant 4. \tag{2.36}$$

Thus, the solution $\vec{u}$ to (2.15) has a decomposition

$$\vec{u} = \vec{w} + \vec{v}. \tag{2.37}$$

Here $\vec{w}$ is the *boundary layer* component and $\vec{v}$ is the regular component. Notice that $\vec{w}$ is not explicitly constructed as the solution to a differential equation.

The error in the regular and boundary components can be written as

$$\|\delta^2\vec{u} - \vec{u}''\|_{\bar{\Omega}^N} = \|\delta^2\vec{v} - \vec{v}''\|_{\bar{\Omega}^N} + \|\delta^2\vec{w} - \vec{w}''\|_{\bar{\Omega}^N}.$$

To bound the error in the $\vec{v}$ term, use (2.8b) at the mesh transition points, and at other points use (2.8c). For the layer term $\vec{w}$, use (2.8a) at the mesh points and (2.8c) for other points. This gives

$$\varepsilon^2\|\delta^2\vec{u} - \vec{u}''\|_{\bar{\Omega}^N} \leqslant CN^{-2}\ln^2 N + \begin{cases} C\varepsilon N^{-1} & \text{for } j \in \{\frac{N}{4}, \frac{3}{4}N\}, \\ 0 & \text{otherwise.} \end{cases}.$$

Next, suppose that $\vec{\eta}$ is the solution to

$$\vec{L}^N\vec{\eta} = \varepsilon^2(\delta^2\vec{u} - \vec{u}'').$$

To get the bound to η we use the technique of Lemma 2.2.1. Set $\vec{V} = \frac{1}{2}\vec{\eta}^T\vec{\eta}$. Then

$$-\varepsilon^2\delta^2\vec{V} + 2\alpha^2\vec{V} \leqslant C\|\vec{\eta}\|_{\bar{\Omega}}|\vec{L}^N\vec{\eta}|.$$

We can apply a discrete maximum principle for scalar equation by using a barrier function

$$\vec{V}(x_j) = C\|\vec{\eta}\|N^{-2}(\ln^2 N + \tau_S\varepsilon^{-1}\varphi_j)$$

where

$$\varphi_j = \begin{cases} x_j\tau_S^{-1} & \text{for } j = 0, \ldots, \frac{N}{4}, \\ 1 & \text{for } j = \frac{N}{4}, \ldots, \frac{3}{4}N, \\ (1-x_j)\tau_S^{-1} & \text{for } j = \frac{3}{4}N, \ldots, N. \end{cases}$$

Then, we can obtain the bound for $\|\vec{V}\|$ and hence $\|\vec{\eta}\| \leqslant CN^{-1}\ln^2 N$, independent on C and ε.

□

2.5 Numerical results

We now present numerical results in support of Theorem 2.4.5 and Theorem 2.4.6. For a specific example, we tabulate global and pointwise errors and convergence rates. We do this for uniform, Bakhvalov and Shishkin meshes. The theoretical results of Theorems 2.4.5 and 2.4.6 show that the pointwise solutions are computed robustly on the fitted Bakhvalov and Shishkin meshes. To verify this numerically, we compute the pointwise errors, which are defined as

$$E_\varepsilon^N := \max_{i=0,\ldots,N} |u(x_i) - U_i|.$$

The associated rate of convergence is

$$\rho_\varepsilon^N := \log_2 \left(\frac{E_\varepsilon^N}{E_\varepsilon^{N/2}} \right). \tag{2.38}$$

Recall our example problem (2.2)

$$-\varepsilon^2 u'' + (i+4)^2 u = (4+4i)e^x, \qquad u(0) = 0, \qquad u(1) = 0.$$

We begin our numerical investigations by presenting, in Table 2.1, the pointwise errors computed when this equation is solved numerically by our finite difference method on a uniform mesh. Table 2.1 gives the pointwise errors. For $\varepsilon = 1$, it is clear that the error is approximately CN^{-2} where $C \approx 3.05 \times 10^{-2}$. But for $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-6}$, the error actually grows proportionally to N^2. From this, we can see that the error depends strongly and adversely on ε.

Table 2.1: Errors, E_ε^N, for problem (2.2), solved on a uniform mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
1	1.105e-03	2.781e-04	6.961e-05	1.741e-05	4.353e-06	1.088e-06	2.720e-07
ρ_ε^N		1.991	1.999	1.999	2.000	2.000	2.000
1e-01	3.988e-02	2.050e-02	5.595e-03	1.468e-03	3.694e-04	9.267e-05	2.318e-05
ρ_ε^N		0.960	1.874	1.930	1.991	1.995	1.999
1e-02	1.359e-03	5.394e-03	2.016e-02	4.005e-02	2.794e-02	8.536e-03	2.257e-03
ρ_ε^N		-1.989	-1.902	-0.990	0.520	1.711	1.919
1e-03	1.362e-05	5.448e-05	2.178e-04	8.703e-04	3.464e-03	1.355e-02	3.593e-02
ρ_ε^N		-2.000	-2.000	-1.998	-1.993	-1.968	-1.407
1e-04	1.362e-07	5.448e-07	2.179e-06	8.717e-06	3.487e-05	1.394e-04	5.573e-04
ρ_ε^N		-2.000	-2.000	-2.000	-2.000	-2.000	-1.999
1e-05	1.362e-09	5.448e-09	2.179e-08	8.717e-08	3.487e-07	1.395e-06	5.579e-06
ρ_ε^N		-2.000	-2.000	-2.000	-2.000	-2.000	-2.000
1e-06	1.362e-11	5.448e-11	2.179e-10	8.717e-10	3.487e-09	1.395e-08	5.579e-08
ρ_ε^N		-2.000	-2.000	-2.000	-2.000	-2.000	-2.000

In essence, the problem with the results shown in Table 2.1 is that, since a uniform mesh is used, and only pointwise errors reported, the error is not computed at any point inside the layers. To address this, and to determine if layers are resolved, we compute the maximum global error,

which is defined as

$$G_\varepsilon^N := \max_{0\leqslant x\leqslant 1} |u(x) - U(x)|,$$

where $U(x)$ represents the piecewise linear interpolant to the mesh function U, evaluated at x. Note that, since this interpolant is globally defined, we are estimating the maximum error at all points, including within the layers (even when the numerical solution does not resolve the layers). As we shall see, in practice, for layer resolving meshes G_ε^N satisfies the same bounds as E_ε^N; this is explained for Shishkin meshes in [**27, Thm. 8.3**]. It is known that an analogous statement does not hold for uniform meshes [**14, Thm. 3.12**].

In Table 2.2 we give the computed global errors, G_ε^N, for when (2.2) is solved on a uniform mesh (the rates of convergence are defined in the obvious way based on (2.38)). For $\varepsilon = 1$, it is again clear that the error is proportional to CN^{-2} where $C \approx 0.33$. But for $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-6}$, no convergence is observed. We conclude from this that uniform meshes are unsuitable for this problem and a layer-adapted mesh is needed.

Table 2.2: Error, G_ε^N, for problem (2.2), solved on a uniform mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
1	6.874e-03	1.798e-03	4.594e-04	1.161e-04	2.918e-05	7.313e-06	1.831e-06
Rate		1.935	1.969	1.985	1.992	1.996	1.998
1e-01	2.683e-01	1.139e-01	3.716e-02	1.050e-02	2.779e-03	7.141e-04	1.809e-04
ρ_ε^N		1.236	1.615	1.823	1.918	1.961	1.981
1e-02	7.564e-01	6.657e-01	5.185e-01	3.304e-01	1.556e-01	5.445e-02	1.593e-02
ρ_ε^N		0.184	0.361	0.650	1.086	1.515	1.773
1e-03	8.592e-01	8.593e-01	8.316e-01	7.864e-01	6.986e-01	5.711e-01	3.930e-01
ρ_ε^N		-0.000	0.047	0.081	0.171	0.291	0.539
1e-04	8.592e-01	8.593e-01	8.593e-01	8.593e-01	8.594e-01	8.493e-01	8.041e-01
ρ_ε^N		-0.000	-0.000	-0.000	-0.000	0.017	0.079
1e-05	8.592e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01
ρ_ε^N		-0.000	-0.000	-0.000	-0.000	-0.000	-0.000
1e-06	8.592e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01	8.593e-01
ρ_ε^N		-0.000	-0.000	-0.000	-0.000	-0.000	-0.000

Tables 2.3 and 2.4 present numerical results for problem (2.2) on the Shishkin mesh. Table 2.3 shows pointwise errors. Since the Shishkin mesh is uniform for large values of ε, we observe that the first two rows of Tables 2.1 and 2.3 are the same. However, in Table 2.3, for $\varepsilon = 10^{-2}$ to $\varepsilon = 10^{-6}$ the error is independent of ε and it is proportional to N^{-2}, verifying that Theorem 2.4.6 is sharp. To confirm that the layers are resolved, in Table 2.4 we show the global errors. We see that, for $\varepsilon \leqslant 10^{-2}$, the method is robust with respect to ε, and, just like the pointwise errors, almost second-order convergent in N.

Table 2.3: Errors, E_ε^N, for problem (2.2), solved on a Shishkin mesh

ε	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
1	1.105e-03	2.781e-04	6.961e-05	1.741e-05	4.353e-06	1.088e-06	2.720e-07
ρ_ε^N		1.991	1.999	1.999	2.000	2.000	2.000
1e-01	3.988e-02	2.050e-02	5.595e-03	1.468e-03	3.694e-04	9.267e-05	2.318e-05
ρ_ε^N		0.960	1.874	1.930	1.991	1.995	1.999
1e-02	2.395e-02	3.844e-02	3.680e-02	1.986e-02	6.728e-03	2.291e-03	7.209e-04
ρ_ε^N		-0.683	0.063	0.890	1.561	1.554	1.668
1e-03	2.395e-02	3.844e-02	3.680e-02	1.986e-02	6.728e-03	2.291e-03	7.209e-04
ρ_ε^N		-0.683	0.063	0.890	1.561	1.554	1.668
1e-04	2.395e-02	3.844e-02	3.680e-02	1.986e-02	6.728e-03	2.291e-03	7.209e-04
ρ_ε^N		-0.683	0.063	0.890	1.561	1.554	1.668
1e-05	2.395e-02	3.844e-02	3.680e-02	1.986e-02	6.728e-03	2.291e-03	7.209e-04
ρ_ε^N		-0.683	0.063	0.890	1.561	1.554	1.668
1e-06	2.395e-02	3.844e-02	3.680e-02	1.986e-02	6.728e-03	2.291e-03	7.209e-04
ρ_ε^N		-0.683	0.063	0.890	1.561	1.554	1.668

Table 2.4: Error, G_ε^N, for problem (2.2), solved on a Shishkin mesh.

ε	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
1	6.874e-03	1.798e-03	4.594e-04	1.161e-04	2.918e-05	7.313e-06	1.831e-06
ρ_ε^N		1.935	1.969	1.985	1.992	1.996	1.998
1e-01	2.683e-01	1.139e-01	3.716e-02	1.050e-02	2.779e-03	7.141e-04	1.809e-04
ρ_ε^N		1.236	1.615	1.823	1.918	1.961	1.981
1e-02	4.908e-01	3.624e-01	2.235e-01	1.108e-01	4.511e-02	1.615e-02	5.329e-03
ρ_ε^N		0.438	0.697	1.013	1.296	1.482	1.599
1e-03	4.908e-01	3.624e-01	2.235e-01	1.108e-01	4.511e-02	1.615e-02	5.329e-03
ρ_ε^N		0.438	0.697	1.013	1.296	1.482	1.599
1e-04	4.908e-01	3.624e-01	2.235e-01	1.108e-01	4.511e-02	1.615e-02	5.329e-03
ρ_ε^N		0.438	0.697	1.013	1.296	1.482	1.599
1e-05	4.908e-01	3.624e-01	2.235e-01	1.108e-01	4.511e-02	1.615e-02	5.329e-03
ρ_ε^N		0.438	0.697	1.013	1.296	1.482	1.599
1e-06	4.908e-01	3.624e-01	2.235e-01	1.108e-01	4.511e-02	1.615e-02	5.329e-03
ρ_ε^N		0.438	0.697	1.013	1.296	1.482	1.599

Finally, Tables 2.5 and 2.6 present the pointwise and global errors computed when (2.2) is solved by the finite difference scheme on a Bakhvalov mesh. These numerical results are in agreement with the theoretical result of Theorem 2.4.5: now, rather than seeing only almost second-order convergence (i.e., with the spoiling $\ln^2 N$ term), we have full ε-uniform second-order convergence.

Table 2.5: Errors, E_ε^N, for problem (2.2), solved on a Bakhvalov mesh.

ε	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
1	1.105e-03	2.781e-04	6.961e-05	1.741e-05	4.353e-06	1.088e-06	2.720e-07
ρ_ε^N		1.991	1.999	1.999	2.000	2.000	2.000
1e-01	1.258e-02	3.321e-03	8.419e-04	2.112e-04	5.291e-05	1.323e-05	3.308e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000
1e-02	1.257e-02	3.319e-03	8.415e-04	2.111e-04	5.288e-05	1.322e-05	3.307e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000
1e-03	1.257e-02	3.319e-03	8.415e-04	2.111e-04	5.288e-05	1.322e-05	3.307e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000
1e-04	1.257e-02	3.319e-03	8.415e-04	2.111e-04	5.288e-05	1.322e-05	3.307e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000
1e-05	1.257e-02	3.319e-03	8.415e-04	2.111e-04	5.288e-05	1.322e-05	3.307e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000
1e-06	1.257e-02	3.319e-03	8.415e-04	2.111e-04	5.288e-05	1.322e-05	3.307e-06
ρ_ε^N		1.921	1.980	1.995	1.997	2.000	2.000

Table 2.6: Error, G_ε^N, for problem (2.2), solved on a Bakhvalov mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
1	6.874e-03	1.798e-03	4.594e-04	1.161e-04	2.918e-05	7.313e-06	1.831e-06
ρ_ε^N		1.935	1.969	1.985	1.992	1.996	1.998
1e-01	1.051e-01	3.190e-02	8.714e-03	2.271e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993
1e-02	1.051e-01	3.191e-02	8.714e-03	2.272e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993
1e-03	1.051e-01	3.191e-02	8.714e-03	2.272e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993
1e-04	1.051e-01	3.191e-02	8.714e-03	2.272e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993
1e-05	1.051e-01	3.191e-02	8.714e-03	2.272e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993
1e-06	1.051e-01	3.191e-02	8.714e-03	2.272e-03	5.795e-04	1.463e-04	3.676e-05
ρ_ε^N		1.720	1.872	1.940	1.971	1.986	1.993

2.A Derivation of truncation errors

In this appendix, we prove the following bound for $u''(x_j) - \delta^2 u(x_j)$ on $\bar{\Omega}$ which have been stated in (2.8) by using Taylor series techniques

$$|u''(x_j) - \delta^2(x_j)| = \begin{cases} C|u''(\alpha_j)|, \\ C(h_{j+1} - h_j)|u'''(\alpha_j)|, \\ C\big((h_{j+1} - h_j)|u'''(x_j)| - (h_j^2 - h_j h_{j+1} + h_{j+1}^2)|u^{(4)}(\alpha_j)|\big), \end{cases}$$

for some $\alpha_j \in [x_{j-1}, x_{j+1}]$.

First, to prove (2.8a), by the triangle inequality

$$|u''(x_j) - \delta^2 u(x_j)| \leqslant |u''(x_j)| + |\delta^2 u(x_j)| \leqslant \|u''(x)\|_{[x_{j-1}, x_{j+1}]} + |\delta^2 u(x_j)|.$$

So we need to show that

$$|\delta^2 u(x_j)| \leqslant C\|u''(x_j)\|_{\bar{\Omega}}.$$

By using the second order Taylor series expansion of $u(x_j)$ about the points x_{j-1} and x_{j+1}:

$$u(x_{j-1}) = u(x_j) + (x_{j-1} - x_j)u'(x_j) + \frac{1}{2}(x_{j-1} - x_j)^2 u''(\alpha_j), \tag{2.40}$$

and

$$u(x_{j+1}) = u(x_j) + (x_{j+1} - x_j)u'(x_j) + \frac{1}{2}(x_{j+1} - x_j)^2 u''(\beta_j). \tag{2.41}$$

Multiplying h_{j+1} in (2.40) and h_j in (2.41)

$$h_{j+1}u(x_{j-1}) = h_{j+1}u(x_j) - h_{j+1}h_j u'(x_j) + \frac{1}{2}h_{j+1}h_j^2 u''(\alpha_j), \tag{2.42}$$

$$h_j u(x_{j+1}) = h_j u(x_j) + h_j h_{j+1} u'(x_j) + \frac{1}{2}h_j h_{j+1}^2 u''(\beta_j). \tag{2.43}$$

since $h_j = x_j - x_{j-1}$. Adding (2.42) and (2.43), gives

$$h_{j+1}u(x_{j-1}) + h_j u(x_{j+1}) - (h_j + h_{j+1})u(x_j) = \big(\frac{h_j + h_{j+1}}{2}\big)h_j h_{j+1} u''(\gamma_j),$$

for some $\gamma_j \in [x_{j-1}, x_{j+1}]$. Dividing by $h_j h_{j+1}$ gives

$$\frac{u(x_{j-1})}{h_j} + \frac{u(x_{j+1})}{h_{j+1}} - \frac{(h_j + h_{j+1})u(x_j)}{h_j h_{j+1}} = \hbar_j u''(\gamma_j). \tag{2.44}$$

where $\hbar_j = (x_{j+1} - x_{j-1})/2$. Dividing (2.44) by $\hbar_j$, we get

$$\delta^2 u(x_j) = \frac{1}{\hbar_j}\left[\frac{u(x_{j-1})}{h_j} + \frac{u(x_{j+1})}{h_{j+1}} - \frac{(h_j + h_{j+1})u(x_j)}{h_j h_{j+1}}\right] = u''(\gamma_j). \tag{2.45}$$

This implies that

$$|u''(x_j) - \delta^2 u(x_j)| \leqslant \|u''(x)\|_{[x_{j-1}, x_{j+1}]} + |u''(\gamma_j)|.$$

Second, to prove (2.8b), using the third order Taylor series expansion of $u(x_j)$ about the points x_{j-1} and x_{j+1}, we get

$$u(x_{j-1}) = u(x_j) + (x_{j-1} - x_j)u'(x_j) + \frac{1}{2}(x_{j-1} - x_j)^2 u''(x_j) + \frac{1}{6}(x_{j-1} - x_j)^3 u'''(\alpha_j), \tag{2.46}$$

and

$$u(x_{j+1}) = u(x_j) + (x_{j+1} - x_j)u'(x_j) + \frac{1}{2}(x_{j+1} - x_j)^2 u''(x_j) + \frac{1}{6}(x_{j+1} - x_j)^3 u'''(\beta_j). \tag{2.47}$$

Multiplying by h_{j+1} in (2.46) and by h_j in (2.47)

$$h_{j+1}u(x_{j-1}) = h_{j+1}u(x_j) - h_{j+1}h_j u'(x_j) + \frac{1}{2}h_{j+1}h_j^2 u''(x_j) - \frac{1}{6}h_{j+1}h_j^3 u'''(\alpha_j), \tag{2.48}$$

$$h_j u(x_{j+1}) = h_j u(x_j) + h_j h_{j+1} u'(x_j) + \frac{1}{2}h_j h_{j+1}^2 u''(x_j) + \frac{1}{6}h_j h_{j+1}^3 u'''(\beta_j). \tag{2.49}$$

since $h_j = x_j - x_{j-1}$. Adding (2.48) and (2.49), gives

$$\begin{aligned} h_{j+1}u(x_{j-1}) + h_j u(x_{j+1}) - (h_j + h_{j+1})u(x_j) = &\left(\frac{h_j + h_{j+1}}{2}\right)h_j h_{j+1} u''(x_j) \\ &+ \left(\frac{h_j h_{j+1}}{3}\right)\left(\frac{h_{j+1}^2 - h_j^2}{2}\right)u'''(\gamma_j) \quad \text{for some } \gamma_j \in [x_{j-1}, x_{j+1}]. \end{aligned}$$

Dividing by $\hbar_j h_j h_{j+1}$ gives

$$\frac{1}{\hbar_j}\left[\frac{u(x_{j-1})}{h_j} + \frac{u(x_{j+1})}{h_{j+1}} - \frac{(h_j + h_{j+1})u(x_j)}{h_j h_{j+1}}\right] = u''(x_j) + \frac{1}{3}(h_{j+1} - h_j)u'''(\gamma_j), \tag{2.50}$$

where $\hbar_j = (x_{j+1} - x_{j-1})/2$. This implies that

$$u''(x_j) = \delta^2 u(x_j) - \frac{1}{3}(h_{j+1} - h_j)u'''(\gamma_j),$$

which proves (2.8b).

Finally, to prove (2.8c) use a fourth order Taylor series expansion of $u(x_j)$ about the points x_{j-1} and x_{j+1}

$$\begin{aligned} u(x_{j-1}) = u(x_j) + (x_{j-1} - x_j)u'(x_j) + \frac{1}{2}(x_{j-1} - x_j)^2 u''(x_j) + \frac{1}{6}(x_{j-1} - x_j)^3 u'''(x_j) \\ + \frac{1}{24}(x_{j-1} - x_j)^4 u''''(\alpha_j), \end{aligned} \tag{2.51}$$

and

$$\begin{aligned} u(x_{j+1}) = u(x_j) + (x_{j+1} - x_j)u'(x_j) + \frac{1}{2}(x_{j+1} - x_j)^2 u''(x_j) + \frac{1}{6}(x_{j+1} - x_j)^3 u'''(x_j) \\ + \frac{1}{24}(x_{j+1} - x_j)^4 u''''(\beta_j). \end{aligned} \tag{2.52}$$

Multiplying h_{j+1} in (2.52) and h_j in (2.51)

$$h_{j+1}u(x_{j-1}) = h_{j+1}u(x_j) - h_{j+1}h_j u'(x_j) + \frac{1}{2}h_{j+1}h_j^2 u''(x_j) - \frac{1}{6}h_{j+1}h_j^3 u'''(x_j) + \frac{1}{24}h_{j+1}h_j^4 u^{(4)}(\alpha_j), \tag{2.53}$$

$$h_j u(x_{j+1}) = h_j u(x_j) + h_j h_{j+1} u'(x_j) + \frac{1}{2} h_j h_{j+1}^2 u''(x_j) + \frac{1}{6} h_j h_{j+1}^3 u'''(\beta_j) + \frac{1}{24} h_j h_{j+1}^4 u^{(4)}(\beta_j). \tag{2.54}$$

since $h_j = x_j - x_{j-1}$. Adding (2.54) and (2.53), gives

$$\begin{aligned} h_{j+1} u(x_{j-1}) + h_j u(x_{j+1}) - (h_j + h_{j+1}) u(x_j) &= \big(\frac{h_j + h_{j+1}}{2}\big) h_j h_{j+1} u''(x_j) \\ &+ \big(\frac{h_j h_{j+1}}{3}\big)\big(\frac{h_{j+1}^2 - h_j^2}{2}\big) u'''(x_j) + \big(\frac{h_j h_{j+1}}{24}\big)\big(h_{j+1}^3 + h_j^3\big) u^{(4)}(\gamma_j) \quad \text{for some } \gamma_j \in \Omega. \end{aligned} \tag{2.55}$$

Dividing by $\hbar_j h_j h_{j+1}$ gives

$$\begin{aligned} \frac{1}{\hbar_j}\left[\frac{u(x_{j-1})}{h_j} + \frac{u(x_{j+1})}{h_{j+1}} - \frac{(h_j + h_{j+1}) u(x_j)}{h_j h_{j+1}}\right] &= u''(x_j) + \frac{1}{3}(h_{j+1} - h_j) u'''(x_j) \\ &+ \frac{1}{12}(h_j^2 - h_j h_{j+1} + h_{j+1}^2) u^{(4)}(\gamma_j), \end{aligned} \tag{2.56}$$

where $\hbar_j = (x_{j+1} - x_{j-1})/2$. Since

$$\frac{1}{24}(h_j^3 + h_{j+1}^3) = \frac{1}{12}\big(\frac{h_j + h_{j+1}}{2}\big)(h_j^2 - h_j h_{j+1} + h_{j+1}^2) = \frac{1}{12}\hbar_j (h_j^2 - h_j h_{j+1} + h_{j+1}^2).$$

Then,

$$u''(x_j) = \delta^2 u(x_j) - \frac{1}{3}(h_{j+1} - h_j) u'''(x_j) - \frac{1}{12}(h_j^2 - h_j h_{j+1} + h_{j+1}^2) u^{(4)}(\gamma_j).$$

which proves (2.8c).

Chapter 3

A note on a finite element analysis of a fourth-order real-valued singularly perturbed problem

3.1 Introduction

In this chapter, we are interested in the numerical solution of a singularly perturbed, fourth-order, real-valued reaction diffusion equation. Our model differential equation is

$$-\varepsilon u^{(4)}(x) + au''(x) - bu(x) = f(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{3.1a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{3.1b}$$

This problem is of the "simply supported" type reviewed in Section 1.5.3. As usual, ε is a positive, real-valued parameter, and we assume $0 < \varepsilon \leqslant 1$, but typically have that $\varepsilon \ll 1$. The coefficient functions, a and b, and right-hand side function, f, are real-valued functions on the interval Ω. For these problems, it is typical to transform the problem into a (weakly) coupled system of second-order reaction-diffusion problems. Then, when analysing a finite element method for solving this system, it is usually assumed that the coupling matrix is pointwise coercive, which is sometimes referred to as "positive definite but not necessarily symmetric", in the literature. This is the approach taken, for example, in [47]. However, in Section 3.2.1, we show that standard ideas for transforming the problem into a system of two second-order differential equations do not yield a coupling matrix that satisfies the coercivity condition. We then propose a new transformation, in Section 3.2.2, which involves a coefficient-dependent parameter. We shall show how to, under reasonable assumptions, determine the value of the parameter in the transformation that ensures that the coupling matrix is coercive. In Section 3.3, we describe a variational formulation of (3.1) and an associated energy norm. This leads to a finite element method for the problem, for which

we present a suitable layer-adapted mesh. Finally, the section concludes with numerical validating the expected error bounds.

We emphasise that the primary contribution of this chapter is the transformation that gives a coercive system in Section 3.2.2. The remaining section validates the usefulness of this. Although we briefly discuss the analysis of the finite element method and give numerical results, it is not our main concern.

3.2 From a fourth-order problem to a coupled system

In this section we investigate how to transform (3.1) into a coupled system, in such a way that the coupling matrix is coercive.

Definition 3.2.1. A matrix M is *coercive*, if there exists a constant $\gamma > 0$ such that

$$\frac{\vec{v}^T M \vec{v}}{\vec{v}^T \vec{v}} \geqslant \gamma \quad \text{for all } \vec{v} \in \mathbb{R}^2/\{(0,0)^T\}. \tag{3.2}$$

We have already introduced this in Lemma 2.4.1, but we have repeated it here to make the presentation self-contained. We also note that some papers in the literature, such as [47], refer to such as matrix as being "positive definite". However, most standard references define positive definiteness as a property that applies only to symmetric matrices. So, we will avoid this terminology.

It will be important to determine when a given matrix is coercive. The following classical results are very useful.

Theorem 3.2.1. [18] *A real-valued $n \times n$ matrix B satisfies*

$$\vec{v}^T B \vec{v} > 0 \quad \textit{for all } \vec{v} \in \mathbb{R}^N/\{\vec{0}\},$$

if and only if $M = (B + B^T)/2$ is symmetric positive definite.

Theorem 3.2.2. [16, p. 402] *A real-valued $n \times n$ symmetric matrix M is positive definite if and only if all of its eigenvalues are positive.*

3.2.1 A simple, and inadequate, transformation

Recall the problem stated in (3.1). In [47], it is assumed that the functions a, b and f are given sufficiently smooth and that

$$a(x) \geqslant 0 \quad \text{and} \quad b(x) \geqslant 0 \quad \text{for} \quad x \in \bar{\Omega}. \tag{3.3}$$

By using the boundary conditions, they transform the problem into a system of two differential equations. They set $\vec{z} = (u, w)^T$, where

$$w := -u''. \tag{3.4}$$

With this, (3.1) can be transformed into a system of two equations of the form

$$-E\vec{z}'' + A\vec{z} = \vec{f}, \tag{3.5}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; E = \begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix}, \; A = \begin{pmatrix} a & -b \\ 1 & 0 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} -f \\ 0 \end{pmatrix}. \tag{3.6}$$

In [47] it is written:

> *We assume that the matrix A is point-wise positive definite (but not necessarily symmetric), i.e., for some fixed $\gamma > 0$*
>
> $$\vec{v}^T A\vec{v} \geqslant \gamma^2 \vec{v}^T\vec{v} \;\text{ for all } \vec{v} \in \mathbb{R}^2/\{(0,0)^T\}. \tag{3.7}$$

That is, in our terminology, for all $x \in [0,1]$, A is coercive. However, for any $\vec{v} \in \mathbb{R}^2$,

$$\vec{v}^T A\vec{v} = \begin{pmatrix} v_1 & v_2 \end{pmatrix} \begin{pmatrix} a & -b \\ 1 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = av_1^2 + v_1v_2(1-b).$$

So, for example, if $a = 1$ and $b = 2$, then $\vec{v}^T A\vec{v} = -1$ when $\vec{v} = (1,2)^T$. More generally, if $b = 1$ then for any $\vec{v}$ with $v_1 = 0$, one gets $\vec{v}^T A\vec{v} = 0$. Moreover, if $b \neq 1$, then for any other $b \geqslant 0$ and $a \geqslant 0$, we can find $\vec{v}$ such that $\vec{v}^T A\vec{v} < 0$. So there is no sense in which both (3.3) and (3.7) are satisfied.

3.2.2 Coercive system

In this section we propose a new transformation for (3.1), which transforms it into a system of two differential equations. This transformation features a parameter that depends on the problem data. We determine the parameter's value that ensures that the coefficient matrix for the zero-order term in the system is coercive.

In our case, we assume that a and b satisfy the following conditions:

$$a \geqslant \varepsilon + r^* > 0, \tag{3.8a}$$

$$b \geqslant 1, \tag{3.8b}$$

for some positive constant r^*.

Remark 3.2.1. Since we place no other assumptions on a and ε, other than their sign, the problem (3.1) can be rescaled so that (3.8b) holds, providing, of course, that $b > 0$.

We propose the transformation

$$w := \frac{u'' - u}{\alpha}. \tag{3.9}$$

That is,

$$u'' = \alpha w + u, \tag{3.10}$$

where α is a non-zero constant chosen depending on the problem data. When (3.10) is used repeatedly, it gives

$$u^{(4)} = \alpha w'' + \alpha w + u.$$

With this, (3.1) can be transformed into a system of two equations of the form

$$-\varepsilon\alpha w'' + \alpha(a-\varepsilon)w + (a-\varepsilon-b)u = f, \tag{3.11a}$$

$$-u'' + \alpha w + u = 0, \tag{3.11b}$$

subject to the boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{3.11c}$$

We can write this system as

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon\alpha & 0 \\ 0 & 1 \end{pmatrix}\vec{z}'' + B\vec{z} = \vec{f}, \tag{3.12a}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; B = \begin{pmatrix} \alpha(a-\varepsilon) & a-\varepsilon-b \\ \alpha & 1 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{3.12b}$$

Next, we will use the conditions (3.8a) and (3.8b) to determine the value of the parameter in the transformation that ensure that the coefficient matrix for the zero-order term in (3.12) is coercive. Recall from Theorems 3.2.2 and 3.2.1 that the matrix B satisfies $\vec{v}^T B\vec{v} > 0$ for all $\vec{v}$ if, and only if, $M = (B^T + B)/2$ is symmetric positive definite. Here

$$M = \begin{pmatrix} \alpha(a-\varepsilon) & \frac{1}{2}(a-\varepsilon-b+\alpha) \\ \frac{1}{2}(a-\varepsilon-b+\alpha) & 1 \end{pmatrix}. \tag{3.13}$$

Clearly, M is symmetric. In addition M is positive definite if and only if all of its eigenvalues are positive. We now will show that it is possible to select α in (3.11) so that its eigenvalues are positive. The eigenvalues of M are

$$\lambda_1 = \frac{1}{2}\Big[-\alpha\varepsilon + \alpha a + 1 + \big(\alpha^2\varepsilon^2 + \varepsilon^2 - 2a\alpha^2\varepsilon - 2a\varepsilon + a^2\alpha^2 + a^2 + 2b\varepsilon - 2ab + \alpha^2 - 2\alpha b + b^2 + 1\big)^{(1/2)}\Big], \tag{3.14a}$$

$$\lambda_2 = \frac{1}{2}\Big[-\alpha\varepsilon + \alpha a + 1 - \big(\alpha^2\varepsilon^2 + \varepsilon^2 - 2a\alpha^2\varepsilon - 2a\varepsilon + a^2\alpha^2 + a^2 + 2b\varepsilon - 2ab + \alpha^2 - 2\alpha b + b^2 + 1\big)^{(1/2)}\Big]. \tag{3.14b}$$

Since M is symmetric, λ_1 and λ_2 are real numbers. Clearly $\lambda_1 \geqslant \lambda_2$ for any a, b and α. We need to find the set of values of α for which both $\lambda_i > 0$, so we will find the range of α for which $\lambda_2 > 0$. By inspection we can see that this is

$$-\varepsilon + a + b - 2\sqrt{ab - \varepsilon b} \leqslant \alpha \leqslant -\varepsilon + a + b + 2\sqrt{ab - \varepsilon b}. \tag{3.15}$$

For any choice of α that satisfies (3.15), M is positive definite for all a and b satisfying (3.8).

For the simple case where a and b are constants, we propose taking

$$\alpha = a + b - \varepsilon, \tag{3.16}$$

but we emphasise that any choice of α between $-\varepsilon + a + b - 2\sqrt{ab - \varepsilon b}$ and $-\varepsilon + a + b + 2\sqrt{ab - \varepsilon b}$ will suffice.

Suppose we use the same example as presented in [47], where $a = 1$, $b = 1$ and $\varepsilon = 10^{-4}$. Then, from (3.6)

$$A = \begin{pmatrix} 1 & -1 \\ 1 & 0 \end{pmatrix},$$

which is not coercive, since

$$(A + A^T)/2 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix},$$

has zero as an eigenvalue. On another hand, from (3.12b) and (3.13), we have

$$B = \begin{pmatrix} 1.9997 & -0.00010 \\ 1.9999 & 1 \end{pmatrix} \quad \text{and} \quad M = \begin{pmatrix} 1.9997 & 0.9999 \\ 0.9999 & 1 \end{pmatrix}.$$

The eigenvalues of M are $\lambda_1 = 2.617727$ and $\lambda_2 = 0.381972$. So, M is symmetric positive matrix, and, consequently, the matrix B is coercive.

For the analysis, it is helpful not only to show that the eigenvalues of M are positive but also to give an ε-independent lower bound for them, which, in turn, gives an ε-independent bound for γ in (3.7). This, we now do.

Lemma 3.2.3. *Let M be the matrix defined in (3.13). If we set $\alpha = a + b - \varepsilon$ as in (3.16), then M is positive definite, and its smaller eigenvalue λ_2, is bounded below as*

$$\lambda_2 > \frac{b(a - \varepsilon)}{a^2 + ab + 1} > \frac{br^*}{a^2 + ab + 1}, \tag{3.17}$$

independently of ε, where r^* is as given in (3.8).

Proof. First, we will show that both of eigenvalues of M are positive, and, consequently, M is symmetric positive definite. When $\alpha = a + b - \varepsilon$, we have

$$M = \begin{pmatrix} (a + b - \varepsilon)(a - \varepsilon) & a - \varepsilon \\ a - \varepsilon & 1 \end{pmatrix}.$$

The smaller eigenvalue of M is λ_2, as given in (3.14b), and with $\alpha = a + b - \varepsilon$, it is

$$\begin{aligned} \lambda_2 = \frac{1}{2}\Big[& (a^2 + ab - 2a\varepsilon - b\varepsilon + \varepsilon^2 + 1) \\ & - \Big(a^4 + (2b - 4\varepsilon)a^3 + (b^2 - 6b\varepsilon + 6\varepsilon^2 + 2)a^2 + (-2b^2\varepsilon + 6b\varepsilon^2 - 4\varepsilon^3 - 2b - 4\varepsilon)a \\ & \qquad + b^2\varepsilon^2 + (-2\varepsilon^3 + 2\varepsilon)b + (\varepsilon^2 + 1)^2\Big)^{(1/2)}\Big]. \end{aligned} \tag{3.18}$$

We can write λ_2 as the difference of two functions in a, b and ε:

$$\lambda_2 = Q(a,b,\varepsilon) - N(a,b,\varepsilon),$$

where

$$Q(a,b,\varepsilon) = (1/2)(a^2 + ab - 2a\varepsilon - b\varepsilon + \varepsilon^2 + 1),$$

and

$$N(a,b,\varepsilon) = (1/2)\big(a^4 + (2b-4\varepsilon)a^3 + (b^2 - 6b\varepsilon + 6\varepsilon^2 + 2)a^2 \\ + (-2b^2\varepsilon + 6b\varepsilon^2 - 4\varepsilon^3 - 2b - 4\varepsilon)a + b^2\varepsilon^2 + (-2\epsilon^3 + 2\varepsilon)b + (\varepsilon^2+1)^2\big)^{(1/2)}.$$

Note that $Q(a,b,\varepsilon) > 0$, $N(a,b,\varepsilon) \geqslant 0$, and

$$Q(a,b,\varepsilon)^2 - N(a,b,\varepsilon)^2 = b(a-\varepsilon) > 0, \tag{3.19}$$

for any a, b and ε satisfying (3.8). From this we can see that $Q(a,b,\varepsilon) > N(a,b,\varepsilon)$ and, so, $\lambda_2 > 0$. To establish a lower bound for λ_2, we write

$$Q(a,b,\varepsilon)^2 - N(a,b,\varepsilon)^2 = (Q(a,b,\varepsilon) + N(a,b,\varepsilon))(Q(a,b,\varepsilon) - N(a,b,\varepsilon)).$$

Therefore,

$$\lambda_2 = Q(a,b,\varepsilon) - N(a,b,\varepsilon) = \frac{Q(a,b,\varepsilon)^2 - N(a,b,\varepsilon)^2}{Q(a,b,\varepsilon) + N(a,b,\varepsilon)} = \frac{b(a-\varepsilon)}{Q(a,b,\varepsilon) + N(a,b,\varepsilon)}.$$

We know that

$$Q(a,b,\varepsilon) + N(a,b,\varepsilon) < 2Q(a,b,\varepsilon)$$

because $Q(a,b,\varepsilon) > N(a,b,\varepsilon)$. Furthermore,

$$2Q(a,b,\varepsilon) = a^2 + ab - 2a\varepsilon - b\varepsilon + \varepsilon^2 + 1 = (a-\varepsilon)^2 + b(a-\varepsilon) + 1 < a^2 + ab + 1.$$

So this, combined with (3.19) yields

$$\lambda_2 > \frac{b(a-\varepsilon)}{2Q(a,b,\varepsilon)} > \frac{br^*}{a^2+ab+1},$$

for all a, b and ε satisfying (3.8a) and (3.8b).

□

Example 3.2.1. Suppose we take $\varepsilon = 10^{-4}$, $a = 2$ and $b = 4$ in (3.1). Then

$$B = \begin{pmatrix} 11.9992 & -2.0001 \\ 5.9999 & 1 \end{pmatrix}, \tag{3.20}$$

and

$$M = \begin{pmatrix} 11.9992 & 1.9999 \\ 1.9999 & 1 \end{pmatrix}. \tag{3.21}$$

The eigenvalues of M are

$$\lambda_1 = 12.3515 \quad \text{and} \quad \lambda_2 = 0.6476.$$

Clearly, M is symmetric. In addition M is positive definite since the eigenvalues are positive. Using the same data in (3.17), we get that

$$\lambda_2 > 0.6153,$$

which shows that the bound in (3.17) is sharp.

Theorem 3.2.4. *Let B be the matrix in (3.12). If $\alpha = a + b - \varepsilon$, then B is coercive (see Definition 3.2.1), with*

$$\gamma \geqslant \frac{br^*}{a^2 + ab + 1}. \tag{3.22}$$

Proof. Let $M = (B^T + B)/2$. The quantity

$$R_{\vec{v}}(M) = \frac{\vec{v}^T M \vec{v}}{\vec{v}^T \vec{v}},$$

is called the Rayleigh quotient of M, for the vector $\vec{v}$. From, e.g., [44, **Thm 5.12**],

$$R_{\vec{v}}(M) \geqslant \lambda_2 \quad \text{for all } \vec{v} \in \mathbb{R}^2/\{(0,0)^T\},$$

where λ_2 is the smaller of the eigenvalues of M. But, for any $\vec{v}$,

$$\vec{v}^T M \vec{v} = 1/2\Big(\vec{v}^T(B^T\vec{v} + B\vec{v})\Big) = 1/2\Big(\vec{v}^T B^T \vec{v} + \vec{v}^T B\vec{v}\Big) = \vec{v}^T B\vec{v},$$

which completes the proof. □

3.3 The numerical method

3.3.1 Variational formulation

First, we denote the usual L^2-inner product on the unit interval as

$$(q, p) := \int_0^1 q(x)p(x)dx.$$

Then, the variational formulation of (3.11) is: find $\vec{z} \in (H_0^1(0,1))^2$ such that

$$\mathcal{B}(\vec{z}, \vec{v}) = \mathbf{F}(\vec{v}) \quad \text{for all} \quad \vec{v} \in (H_0^1(0,1))^2, \tag{3.23}$$

where

$$\mathcal{B}(\vec{z}, \vec{v}) := \varepsilon\alpha(z_1', v_1') + (z_2', v_2') + (b_{11}z_1, v_1) + (b_{12}z_2, v_1) + (b_{21}z_1, v_2) + (b_{22}z_2, v_2), \tag{3.24}$$

and

$$\mathbf{F}(\vec{v}) := (f, v_1)$$

where $b_{11} = \alpha(a-\varepsilon)$, $b_{12} = a-\varepsilon-b$, $b_{21} = \alpha$, and $b_{22} = 1$.

The energy norm on $(H_0^1(0,1))^2$ associated with the bilinear form $\mathcal{B}(\cdot,\cdot)$ on $(H_0^1(0,1))^2$ is $\|\cdot\|_{\mathcal{B}}$, defined by

$$\|\vec{z}\|_{\mathcal{B}}^2 := \varepsilon\alpha \left\|z_1'\right\|_2^2 + \left\|z_2'\right\|_2^2 + \gamma(\|z_1\|_2^2 + \|z_2\|_2^2), \tag{3.25}$$

where, as usual, γ is the (coercivity) constant for the matrix B as in Definition 3.2.1.

Lemma 3.3.1. *Let $\mathcal{B}$ be a bilinear form defined in (3.24). If we set $\alpha = a+b-\varepsilon$ as in (3.16) and γ defined in (3.22), then $\mathcal{B}$ is is coercive, with respect to $\|\cdot\|_{\mathcal{B}}$; that is*

$$\mathcal{B}(\vec{v},\vec{v}) \geqslant \gamma \|\vec{v}\|_{\mathcal{B}}^2 \quad \text{for all} \quad \vec{v} = (v_1, v_2) \in (H_0^1(0,1))^2. \tag{3.26}$$

Also, $\mathcal{B}$ is continuous (bounded), i.e., there exists on constant C such that

$$|\mathcal{B}(\vec{z},\vec{v})| \leqslant C \|\vec{z}\|_{\mathcal{B}} \|\vec{v}\|_{\mathcal{B}} \quad \text{for all} \quad \vec{z},\vec{v} \in (H_0^1(0,1))^2. \tag{3.27}$$

Proof. From (3.24), we get

$$\mathcal{B}(\vec{v_1},\vec{v_2}) := \varepsilon\alpha(v_1', v_1') + (v_2', v_2') + (b_{11}v_1, v_1) + (b_{12}v_1, v_2) + (b_{21}v_2, v_1) + (b_{22}v_2, v_2).$$

Note that $(v_1', v_1') = \|v_1'\|_2^2$ and $(v_2', v_2') = \|v_2'\|_2^2$. Also, from Theorem 3.2.4,

$$\begin{aligned}\mathcal{B}(\vec{v},\vec{v}) &\geqslant (b_{11}v_1, v_1) + (b_{12}v_2, v_1) + (b_{21}v_1, v_2) + (b_{22}v_2, v_2)\\ &= \int_0^1 \vec{v}^T(x)B(x)\vec{v}(x)dx\\ &\geqslant \int_0^1 \gamma\vec{v}^T(x)\vec{v}(x)dx = \gamma\big(\|v_1\|_2^2 + \|v_2\|_2^2\big).\end{aligned}$$

Then (3.26) follows.

For the continuity result, using the Cauchy-Schwarz inequality, there is a constant C such that

$$\begin{aligned}|\mathcal{B}(\vec{z},\vec{v})| &\leqslant \varepsilon\alpha|(z_1', v_1')| + |(z_2', v_2')| + |(b_{11}z_1, v_1) + (b_{12}z_2, v_1) + (b_{21}z_1, v_2) + (b_{22}z_2, v_2)|\\ &\leqslant \varepsilon\alpha\|z_1'\|\|v_1'\| + \|z_2'\|\|v_2'\| + |(\vec{z}^T B, \vec{v})|\\ &\leqslant \varepsilon\alpha\|z_1'\|\|v_1'\| + \|z_2'\|\|v_2'\| + \|\vec{z}^T B\|\|\vec{v}\|\\ &\leqslant \varepsilon\alpha\|z_1'\|\|v_1'\| + \|z_2'\|\|v_2'\| + \|B\|_{\max}\|\vec{z}\|\|\vec{v}\|\\ &\leqslant C(\varepsilon\alpha\|z_1'\|\|v_1'\| + \|z_2'\|\|v_2'\| + \gamma\|\vec{z}\|\|\vec{v}\|) \leqslant C\|\vec{z}\|_{\mathcal{B}}\|\vec{v}\|_{\beta},\end{aligned}$$

where here $\|B\|_{\max}$ denotes $\max_{i,j}\max_{0\leqslant x\leqslant 1}|b_{ij}(x)|$, which is clearly bounded above. Thus (3.27) holds. □

Note that the demonstration of Theorem 3.3.1 is largely elementary, once the coercive property of B are established. In Chapter 6 identical reasoning will be applied for a system of four equations.

The results of Lemma 3.3.1 apply directly to show that (3.23) has a unique solution. This result is called the Lax-Milgram Lemma, and is standard in any text addressing the mathematics of finite element methods, e.g., [**6, Theorem 2.7.7**] or [**35, Theorem 2.43**].

3.3.2 Shishkin mesh

We construct a standard *Shishkin* mesh with the mesh parameter

$$\tau = \min\{\frac{1}{4}, 2\sqrt{\varepsilon}\ln N\}. \tag{3.28}$$

We now define two mesh transition points at $x = \tau$ and $x = 1 - \tau$. That is, we form a piecewise uniform mesh with $N/4$ equally-sized mesh intervals on each of $[0, \tau]$ and $[1 - \tau, 1]$, and $N/2$ equally-sized mesh intervals on $[\tau, 1-\tau]$. Typically, when ε is small, $\tau \ll 1/4$, the mesh is very fine near the boundaries, and coarse in the interior. We refer to Section 2.3 for more details.

3.3.3 Finite element method

We define S to be the subspace of $(H_0^1(0,1))^2$ made up of piecewise linear functions on the mesh of Section 3.3.2. Then the discrete version of (3.23) is: find $\vec{Z} \in S$ such that

$$\mathcal{B}(\vec{Z}, \vec{V}) = \mathbf{F}(\vec{V}) \quad \text{for all} \quad \vec{V} \in S. \tag{3.29}$$

Then, noting Lemma 3.3.1, standard finite element numerical analysis can proceed based on quasi-optimal approximation properties of the finite element space, and interpolation error estimates. This is due to Céa's Lemma (e.g., [**6, Theorem 2.8.1**] or [**35, Theorem 2.44**]). That is, Céa's Lemma gives that, for *any* function $\hat{Z}$ in S.

$$\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}^2 \leqslant \frac{C}{\gamma}\left\|\vec{z} - \hat{Z}\right\|_{\mathcal{B}}^2,$$

where C and γ are the constants in Lemma 3.3.1. This means, to conclude the analysis, we can consider any $\hat{Z} \in S$ for which one has a bound on $\left\|\vec{z} - \hat{Z}\right\|_{\mathcal{B}}^2$. The most popular choice is to take $\hat{Z}$ to be the piecewise linear interpolant to $\vec{z}$, see, e.g., [**44, Chap. 11**]. The details are standard, so we do not give that here (see, e.g., Liu et al. [**24**]), but we can conclude that

$$\begin{aligned}\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}^2 = \left\|u' - U'\right\|_2^2 + \varepsilon\alpha\left\|w' - W'\right\|_2^2 + \gamma(\|u - U\|_2^2 + \|w - W\|_2^2)\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}} \\ \leqslant C_1\varepsilon^{1/2}N^{-1}\ln N + C_2N^{-1} + C_3N^{-2},\end{aligned} \tag{3.30}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix} \text{ and } \vec{Z} = \begin{pmatrix} W \\ U \end{pmatrix}. \tag{3.31}$$

We are specifically interested in the singularly perturbed case, where $\varepsilon \ll N^{-1}$. Thus, for sufficiently small ε, and large enough N, on expects the bound in (3.30) to simplify to

$$\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}^2 \leqslant CN^{-1}. \tag{3.32}$$

This is investigated in the following section.

3.3.4 Numerical results

In this section, we present two examples. The first equation features constant coefficients and can be solved exactly. The second equation has non-constant coefficients, and so we estimate the errors in the numerical solutions based on a computed benchmark solution. Even though the coefficients are non-constant, we show it is possible to find a constant α which satisfies (3.15).

We are primarily interested in the convergence of the finite element solution in the energy norm. In particular, we would like to verify (7.32) and examine the contribution of each component in (3.30). Although not covered by theory, we will also verify pointwise convergence.

We denote the error for given N and ε as

$$E_{\mathcal{B}}^N := \left\| \vec{z} - \vec{Z} \right\|_{\mathcal{B}},$$

where $\vec{Z}$ is the finite element solution, and $\vec{z}$ is either the true or benchmark solution, as appropriate. In addition, $\rho_{\mathcal{B}}^N$ denotes the convergence rate of the error in the energy norm. It is computed as

$$\rho_{\mathcal{B}}^N := \log_2 \left(\frac{E_{\mathcal{B}}^N}{E_{\mathcal{B}}^{N/2}} \right). \tag{3.33}$$

By $E_\infty^N(u)$ and $E_\infty^N(w)$ we denote the true or estimated maximum pointwise error in u and w, respectively, and by $\rho_\infty^N(u)$ and $\rho_\infty^N(w)$ the corresponding rates of convergence of u and w, i.e.,

$$\rho_\infty^N(u) := \log_2 \left(\frac{E_\infty^N(u)}{E_\infty^{N/2}(u)} \right) \quad \text{and} \quad \rho_\infty^N(w) := \log_2 \left(\frac{E_\infty^N(w)}{E_\infty^{N/2}(w)} \right). \tag{3.34}$$

Example 3.3.1. We consider the following specific example of (3.1)

$$-\varepsilon u^{(4)}(x) + (4 + \frac{\varepsilon}{4}) u''(x) - u(x) = 1 + x \qquad \text{on} \qquad \Omega := (0,1), \tag{3.35}$$

with boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0.$$

The solution is

$$u(x) = -(1+x) + \frac{\varepsilon e^{2x/\sqrt{\varepsilon}}(e^{-2/\sqrt{\varepsilon}} - 2) - \varepsilon e^{-2x/\sqrt{\varepsilon}}(e^{2/\sqrt{\varepsilon}} - 2)}{(\varepsilon - 16)(e^{-2/\sqrt{\varepsilon}} - e^{2/\sqrt{\varepsilon}})} + \frac{16e^{-x/2}(e^{1/2} - 2) - 16e^{x/2}(e^{-1/2} - 2)}{(\varepsilon - 16)(e^{-1/2} - e^{1/2})}.$$

As per (3.15), we take $\alpha = 5 - 3\varepsilon/4$, and then the system we solve is

$$-\begin{pmatrix} \varepsilon\alpha & 0 \\ 0 & 1 \end{pmatrix} \vec{z}'' + B\vec{z} = \vec{f}, \tag{3.36a}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; B = \begin{pmatrix} (5 - 3\varepsilon/4)(4 - 3\varepsilon/4) & 3 - 3\varepsilon/4 \\ 5 - 3\varepsilon/4 & 1 \end{pmatrix} \quad \text{and} \quad \vec{f} = \begin{pmatrix} 1 + x \\ 0 \end{pmatrix}, \tag{3.36b}$$

with boundary conditions

$$u(0) = w(0) = u(1) = w(1) = 0. \tag{3.36c}$$

For this system, γ is bounded by using Corollary 3.2.4, giving $\gamma \approx 0.1394$. It can be verified numerically that $\gamma \approx 0.1922$ for all ε, so this is quite sharp.

In Figure 3.1 we show u with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right), which does not features layers. In Figure 3.2 we show w with $\varepsilon = 10^{-1}$ (left), which does not features layers. In contrast, as shown in the graph on the right for smaller ε (in this case $\varepsilon = 10^{-3}$), the solution possesses boundary layers near $x = 0$ and $x = 1$.

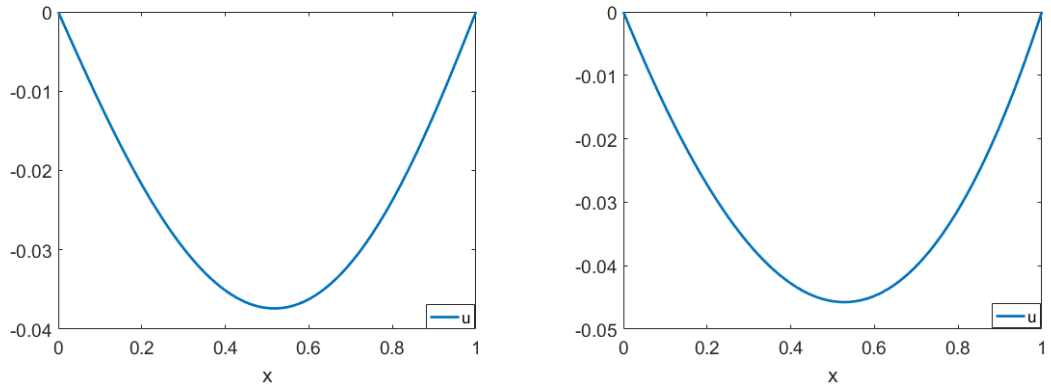

Figure 3.1: The solution u to (3.35) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

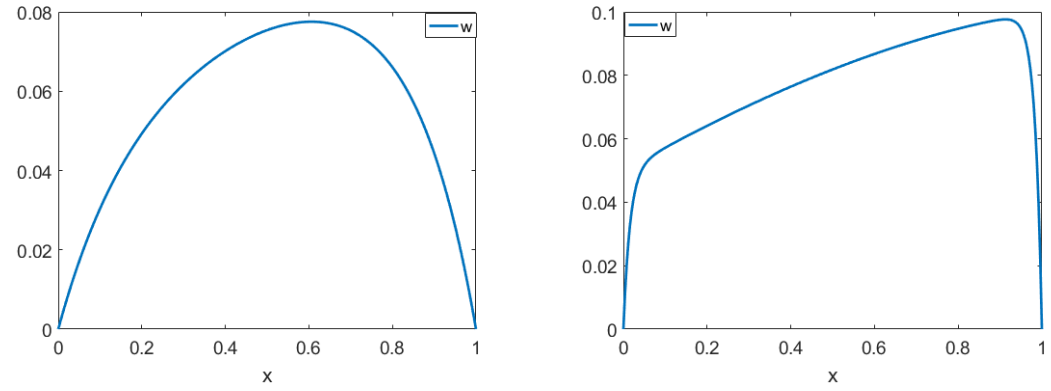

Figure 3.2: The solution w in (3.35) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

In Tables 3.1 and 3.2, we present the error in the energy norm and the associated rates of convergence computed when (3.35) is solved by the finite element method on the Shishkin mesh of Section 3.3.2, but with a minor change to the transition point in (3.28), since the effective perturbation parameter in (3.36) is $\alpha\varepsilon$. The numerical solution converges at a rate that is fully first-order, independently of ε. For large ε, it is clear that the error increases as ε initially decreases. But for $\varepsilon = 10^{-8}$ to $\varepsilon = 10^{-12}$, the method is clearly robust.

From (3.30), we see that E_B^N is comprised of four components. We now investigate each of these. First in Table 3.3, we present $\|u' - U'\|_2$; it clearly shows that u' is robustly estimated. Also, the associated rates of convergence are fully first-order. Note that, particularly for smaller ε, the

Table 3.1: $E_{\mathcal{B}}^N$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.103e-02	5.515e-03	2.758e-03	1.379e-03	6.895e-04
1.0e-02	2.750e-02	1.439e-02	7.285e-03	3.654e-03	1.828e-03
1.0e-04	2.120e-02	1.825e-02	1.384e-02	9.107e-03	5.465e-03
1.0e-06	1.434e-02	8.548e-03	5.381e-03	3.272e-03	1.891e-03
1.0e-08	1.359e-02	6.953e-03	3.627e-03	1.907e-03	9.994e-04
1.0e-10	1.351e-02	6.774e-03	3.403e-03	1.712e-03	8.610e-04
1.0e-12	1.351e-02	6.756e-03	3.380e-03	1.691e-03	8.460e-04

Table 3.2: $\rho_{\mathcal{B}}^N$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.000	1.000	1.000	1.000
1.0e-02	0.934	0.982	0.996	0.999
1.0e-04	0.216	0.399	0.604	0.737
1.0e-06	0.747	0.668	0.718	0.791
1.0e-08	0.967	0.939	0.928	0.932
1.0e-10	0.996	0.993	0.991	0.991
1.0e-12	0.999	0.999	0.999	0.999

quantities in Table 3.1 agree with those in Table 3.3, up to 3 or 4 digits, showing that $\|u' - U'\|_2$ is the dominating term in $\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}$. That is,

$$\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}} \approx \|u' - U'\|_2.$$

Table 3.3: $\|u' - U'\|_2$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.719e-03	8.592e-04	4.296e-04	2.148e-04	1.074e-04
1.0e-02	6.176e-03	3.089e-03	1.545e-03	7.725e-04	3.862e-04
1.0e-04	1.126e-02	5.370e-03	2.558e-03	1.218e-03	5.798e-04
1.0e-06	1.327e-02	6.609e-03	3.290e-03	1.638e-03	8.153e-04
1.0e-08	1.348e-02	6.739e-03	3.368e-03	1.683e-03	8.414e-04
1.0e-10	1.350e-02	6.752e-03	3.376e-03	1.688e-03	8.440e-04
1.0e-12	1.350e-02	6.754e-03	3.377e-03	1.688e-03	8.442e-04

Next, in Tables 3.4, 3.5 and 3.6 we present the three other components in (3.30). From Table 3.4, we can see that, for small ε, the error term $\sqrt{\alpha\varepsilon}\|w' - W'\|_2$ (for fixed N) scales like $\varepsilon^{1/2}$. Since typically $\varepsilon \leqslant \ln^2 N$, this term does not dominate in Table 3.1. Furthermore, we can see that the associated rates of convergence is only $\vartheta(N^{-1} \ln N)$. Table 3.5 verifies that $\|u - U\|_2$ is $\vartheta(N^{-2})$, uniformly in ε, while in Table 3.6 we see that $\|w - W\|_2$ is $\vartheta(\varepsilon^{1/2} N^{-1} \ln N)$. But in all three cases, as expected, the quantities are dominated by those in Table 3.3.

Finally, for curiosity, we verify the pointwise convergence of the method. Tables 3.7 and 3.9 show that, for sufficiently small ε, the pointwise convergence is parameter uniform. Tables 3.8 and 3.10 demonstrate that this convergence is fully second-order for u, and almost second-order for w. This is not surprising, since, as shown in Figure 3.1 and Figure 3.2 there is no (strong) boundary layer in u, but there is in w. This will be investigated more deeply in Chapter 4.

Table 3.4: $\sqrt{\alpha\varepsilon}\|w' - W'\|_2$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.089e-02	5.448e-03	2.724e-03	1.362e-03	6.811e-04
1.0e-02	2.678e-02	1.406e-02	7.119e-03	3.571e-03	1.787e-03
1.0e-04	1.911e-02	1.775e-02	1.370e-02	9.062e-03	5.448e-03
1.0e-06	6.041e-03	5.614e-03	4.332e-03	2.866e-03	1.723e-03
1.0e-08	1.910e-03	1.775e-03	1.370e-03	9.062e-04	5.448e-04
1.0e-10	6.041e-04	5.614e-04	4.332e-04	2.866e-04	1.723e-04
1.0e-12	1.910e-04	1.775e-04	1.370e-04	9.062e-05	5.448e-05

Table 3.5: $\|u - U\|_2$ for (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	5.407e-05	1.353e-05	3.384e-06	8.460e-07	2.115e-07
1.0e-02	1.250e-04	3.126e-05	7.816e-06	1.954e-06	4.886e-07
1.0e-04	2.542e-04	5.150e-05	1.044e-05	2.192e-06	5.103e-07
1.0e-06	4.924e-04	1.211e-04	2.978e-05	7.323e-06	1.800e-06
1.0e-08	5.219e-04	1.303e-04	3.252e-05	8.116e-06	2.026e-06
1.0e-10	5.249e-04	1.312e-04	3.280e-05	8.199e-06	2.049e-06
1.0e-12	5.252e-04	1.313e-04	3.283e-05	8.207e-06	2.052e-06

Table 3.7: $E_\infty^N(u)$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	2.863e-05	7.147e-06	1.786e-06	4.465e-07	1.116e-07
1.0e-02	3.280e-05	7.310e-06	1.841e-06	4.569e-07	1.140e-07
1.0e-04	7.744e-06	1.741e-06	9.876e-07	3.947e-07	1.183e-07
1.0e-06	1.378e-05	3.399e-06	8.413e-07	2.075e-07	5.118e-08
1.0e-08	1.443e-05	3.601e-06	9.021e-07	2.252e-07	5.624e-08
1.0e-10	1.449e-05	3.622e-06	9.082e-07	2.270e-07	5.675e-08
1.0e-12	1.450e-05	3.624e-06	9.088e-07	2.272e-07	5.680e-08

Table 3.8: $\rho_\infty^N(u)$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.002	2.001	2.000	2.000
1.0e-02	2.166	1.990	2.010	2.002
1.0e-04	2.153	0.818	1.323	1.738
1.0e-06	2.019	2.015	2.019	2.020
1.0e-08	2.002	1.997	2.002	2.002
1.0e-10	2.001	1.996	2.000	2.000
1.0e-12	2.001	1.995	2.000	2.000

Table 3.6: $\|w - W\|_2$ for (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	9.123e-05	2.281e-05	5.702e-06	1.426e-06	3.564e-07
1.0e-02	1.939e-03	5.136e-04	1.304e-04	3.272e-05	8.187e-06
1.0e-04	6.424e-03	3.763e-03	1.801e-03	7.103e-04	2.463e-04
1.0e-06	2.033e-03	1.190e-03	5.695e-04	2.246e-04	7.788e-05
1.0e-08	6.471e-04	3.768e-04	1.802e-04	7.103e-05	2.463e-05
1.0e-10	2.179e-04	1.206e-04	5.716e-05	2.249e-05	7.794e-06
1.0e-12	1.017e-04	4.247e-05	1.867e-05	7.208e-06	2.482e-06

Table 3.9: $E_\infty^N(w)$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.290e-04	3.221e-05	8.050e-06	2.013e-06	5.032e-07
1.0e-02	1.350e-02	3.018e-03	7.542e-04	1.873e-04	4.683e-05
1.0e-04	1.029e-01	7.394e-02	4.005e-02	1.586e-02	4.742e-03
1.0e-06	1.029e-01	7.394e-02	4.005e-02	1.586e-02	4.742e-03
1.0e-08	1.029e-01	7.394e-02	4.005e-02	1.586e-02	4.742e-03
1.0e-10	1.029e-01	7.394e-02	4.005e-02	1.586e-02	4.742e-03
1.0e-12	1.029e-01	7.394e-02	4.005e-02	1.586e-02	4.742e-03

Table 3.10: $\rho_\infty^N(w)$ for problem (3.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.002	2.000	2.000	2.000
1.0e-02	2.161	2.001	2.010	2.000
1.0e-04	0.477	0.885	1.336	1.742
1.0e-06	0.477	0.885	1.336	1.742
1.0e-08	0.477	0.885	1.336	1.742
1.0e-10	0.477	0.885	1.336	1.742
1.0e-12	0.477	0.885	1.336	1.742

Example 3.3.2. We consider another example, but with variable coefficients. We will take

$$a = 2x + 1, \quad b = 4x + 1 \quad \text{and} \quad f = 1 + x,$$

in (3.1). That is, we solve

$$-\varepsilon u^{(4)}(x) + (2x+1)u''(x) - (4x+1)u(x) = 1 + x \qquad \text{on} \qquad \Omega := (0,1), \tag{3.37a}$$

with boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0. \tag{3.37b}$$

We can not take α as in (3.16), since then it will be variable so (3.10) could not directly apply. However, we can still choose an α that is constant (in x), and that satisfies (3.15). Specifically, we choose $\alpha = 2 - \varepsilon$.

In Figure 3.3, we show that α satisfies (3.15).

The system we solve is

$$-\begin{pmatrix} \varepsilon(2-\varepsilon) & 0 \\ 0 & 1 \end{pmatrix} \vec{z}'' + \begin{pmatrix} (2-\varepsilon)(2x+1-\varepsilon) & -2x-\varepsilon \\ (2-\varepsilon) & 1 \end{pmatrix} \vec{z} = \begin{pmatrix} 1+x \\ 0 \end{pmatrix}, \tag{3.38a}$$

with boundary conditions

$$u(0) = w(0) = u(1) = w(1) = 0. \tag{3.38b}$$

For this system, γ is bounded numerically, giving $\gamma \approx 0.38197$.

In Figure 3.4 we show u (left) and w (right) both for $\varepsilon = 10^{-3}$; as with u from Example 3.3.1, the former does not features layers, but the latter does.

We will report the error for given N and ε. As before, $\vec{Z}$ is the finite element solution, but since the exact solution to (3.37) is not available to us, $\vec{z}$ will be taken from a benchmark solution.

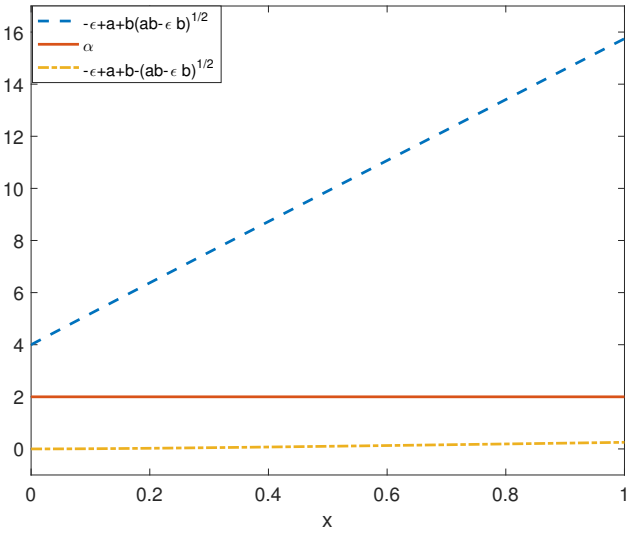

Figure 3.3: Our chosen α, and its upper and lower bounds from (3.15).

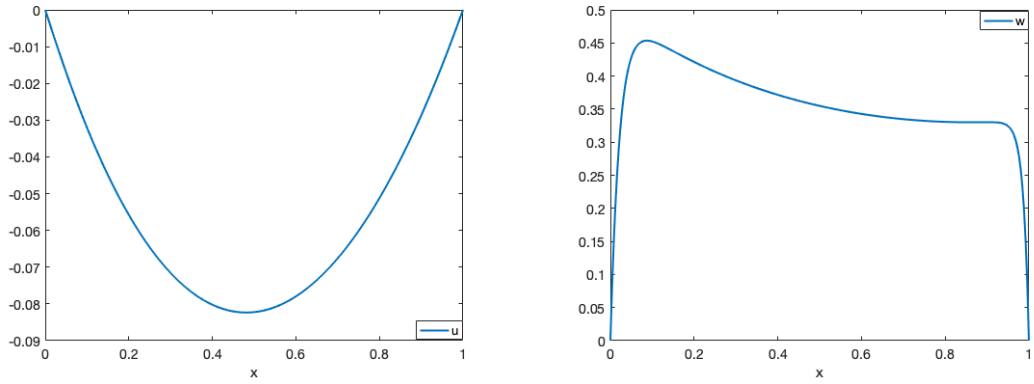

Figure 3.4: The solutions $z_1 = u$ and $z_2 = w$ to (3.38) with $\varepsilon = 10^{-3}$.

Specifically, suppose $\vec{z}$ is computed on a mesh with N intervals. In that case, $\vec{z}$ is calculated on a mesh with $8N$ intervals, but the same transition points (that is, the computational mesh is uniformly refined three times to obtain the mesh on winch the benchmark solution is computed).

In Tables 3.11 and 3.12, we present the error in the energy norm, and the associated rates of convergence computed when (3.37) is solved by the finite element method on the Shishkin mesh. The numerical solution converges at a rate that is fully first-order, independently of ε: although the error increases as ε initially decreases, for $\varepsilon = 10^{-8}$ to $\varepsilon = 10^{-12}$, the method is clearly robust.

From (3.30), we see that $E_{\mathcal{B}}^N$ is comprised of four components. But we will present the first component $\|u' - U'\|_2$ in Table 3.13, which the results of this component are smaller to 3.11, up to 3 or 4 digits, showing that $\|u' - U'\|_2$ is the dominating term in $\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}$.

As discussed for Example 3.3.1, although $E_{\mathcal{B}}^N$ is comprised of four other error terms, it is dominated by $\|u' - U'\|_2$. This is verified in Table 3.13: The entries shown in it agree with those in Table 3.11, up to three or four digits.

Table 3.14 shows pointwise errors for the solution u. The results are qualitatively similar to

Table 3.11: $E_{\mathcal{B}}^N$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	2.812e-03	1.405e-03	7.025e-04	3.513e-04	1.756e-04
1.0e-02	1.126e-02	5.631e-03	2.816e-03	1.408e-03	7.040e-04
1.0e-04	1.889e-02	8.774e-03	4.079e-03	1.902e-03	8.913e-04
1.0e-06	2.445e-02	1.214e-02	6.026e-03	2.992e-03	1.486e-03
1.0e-08	2.506e-02	1.252e-02	6.254e-03	3.125e-03	1.561e-03
1.0e-10	2.512e-02	1.256e-02	6.277e-03	3.138e-03	1.569e-03
1.0e-12	2.512e-02	1.256e-02	6.279e-03	3.139e-03	1.570e-03

Table 3.12: $\rho_{\mathcal{B}}^N$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.001	1.000	1.000	1.000
1.0e-02	1.000	1.000	1.000	1.000
1.0e-04	1.107	1.105	1.101	1.093
1.0e-06	1.010	1.010	1.010	1.010
1.0e-08	1.001	1.001	1.001	1.001
1.0e-10	1.000	1.000	1.000	1.000
1.0e-12	1.000	1.000	1.000	1.000

Table 3.7: for sufficiently small ε, the method is ε-uniformly pointwise convergent. Table 3.15 shows, however, that the rate of convergence is fully second-order again. Table 3.16 show that the numerical approximation of w converges pointwise, uniformly in ε, at an almost second-order rate. We speculate that. Therefore, there is a constant C independent of ε and N, such that

$$\|u - U\|_{\infty,\Omega^N} \leqslant CN^{-2} \quad \text{and} \quad \|w - W\|_{\infty,\Omega^N} \leqslant CN^{-2}\ln^2 N.$$

Table 3.14: $E_{\infty}^N(u)$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	3.746e-05	9.383e-06	2.345e-06	5.863e-07	1.466e-07
1.0e-02	5.862e-05	1.368e-05	3.362e-06	8.370e-07	2.090e-07
1.0e-04	9.283e-05	2.076e-05	4.616e-06	1.020e-06	2.241e-07
1.0e-06	1.302e-04	3.219e-05	7.975e-06	1.977e-06	4.901e-07
1.0e-08	1.340e-04	3.337e-05	8.329e-06	2.080e-06	5.197e-07
1.0e-10	1.344e-04	3.349e-05	8.365e-06	2.091e-06	5.226e-07
1.0e-12	1.344e-04	3.350e-05	8.368e-06	2.092e-06	5.229e-07

Table 3.13: $\|u' - U'\|_2$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.988e-03	9.936e-04	4.968e-04	2.484e-04	1.242e-04
1.0e-02	1.115e-02	5.576e-03	2.788e-03	1.394e-03	6.971e-04
1.0e-04	1.887e-02	8.771e-03	4.079e-03	1.902e-03	8.913e-04
1.0e-06	2.442e-02	1.213e-02	6.026e-03	2.992e-03	1.486e-03
1.0e-08	2.502e-02	1.251e-02	6.253e-03	3.125e-03	1.561e-03
1.0e-10	2.509e-02	1.255e-02	6.276e-03	3.138e-03	1.569e-03
1.0e-12	2.509e-02	1.255e-02	6.279e-03	3.139e-03	1.570e-03

Table 3.15: $\rho_\infty^N(u)$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.997	2.000	2.000	2.000
1.0e-02	2.100	2.024	2.006	2.002
1.0e-04	2.161	2.169	2.178	2.186
1.0e-06	2.016	2.013	2.012	2.012
1.0e-08	2.006	2.002	2.001	2.001
1.0e-10	2.005	2.001	2.000	2.000
1.0e-12	2.005	2.001	2.000	2.000

Table 3.16: $E_\infty^N(w)$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.061e-04	2.651e-05	6.628e-06	1.657e-06	4.142e-07
1.0e-02	2.575e-02	5.903e-03	1.442e-03	3.600e-04	8.977e-05
1.0e-04	1.176e-01	7.448e-02	3.511e-02	1.232e-02	3.671e-03
1.0e-06	1.173e-01	7.448e-02	3.516e-02	1.234e-02	3.678e-03
1.0e-08	1.173e-01	7.448e-02	3.517e-02	1.234e-02	3.679e-03
1.0e-10	1.173e-01	7.448e-02	3.517e-02	1.234e-02	3.679e-03
1.0e-12	1.173e-01	7.448e-02	3.517e-02	1.234e-02	3.679e-03

Table 3.17: $\rho_\infty^N(w)$ for problem (3.37) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.001	2.000	2.000	2.000
1.0e-02	2.125	2.033	2.002	2.004
1.0e-04	0.659	1.085	1.511	1.746
1.0e-06	0.655	1.083	1.511	1.746
1.0e-08	0.655	1.083	1.511	1.746
1.0e-10	0.655	1.083	1.511	1.746
1.0e-12	0.655	1.083	1.511	1.746

Chapter 4

Finite differences for fourth-order real-valued singularly perturbed problems

4.1 Introduction

In this chapter we study the numerical solution, by finite difference methods, of a singularly perturbed, fourth-order, real-valued reaction diffusion equations. Our model differential equation is the same as Chapter 3:

$$-\varepsilon u^{(4)}(x) + au''(x) - bu(x) = f(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{4.1a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{4.1b}$$

As in Chapter 3, ε is a positive, real-valued parameter that is always at most 1, but the cases of particular interest are when $\varepsilon \ll 1$. The coefficient functions, a and b, and right-hand side function, f, are real-valued functions on the interval Ω.

4.1.1 Outline

In Section 4.2, we introduce a family of fourth-order ordinary differential equations, focusing on a problem studied by Shanthi and Ramanujam [38]. As in Chapter 3, we propose a new transformation for the problem, which transforms it into a system of two differential equations. However, whereas in Chapter 3, this was to ensure the resulting system matrix is positive definite, here we wish to use maximum/minimum principle techniques. In Section 4.3, in Lemma 4.3.1, we establish a stability result and maximum principle for the differential operator of the system. Also, we

state and prove the bounds for the solution and its derivatives in Lemma 4.3.5. Section 4.4, we describe a finite difference method for this problem, applied, initially, on an arbitrary mesh. We then present a suitable layer adapted mesh, and the numerical analysis of method on this mesh. Finally, in Section 4.5, numerical results are shown in support of the theoretical analysis.

4.2 Analysis of the continuous problem

We consider the real-valued fourth-order ordinary differential equation (4.1), and begin by discussing the assumptions made by Shanthi and Ramanujam [38]. They assumed that there exist positive constants a^*, b^* and k that satisfy the following conditions:

$$a(x) \geqslant a^* > 0, \tag{4.2a}$$

$$0 \geqslant b(x) \geqslant -b^*, \quad b^* > 0, \tag{4.2b}$$

$$a^* - 2b^* \geqslant k > 0. \tag{4.2c}$$

In our case, we assume that a and b satisfy the following conditions:

$$a \geqslant \varepsilon + r^* > 0, \tag{4.3a}$$

$$b \geqslant s^* > 0, \tag{4.3b}$$

$$b \leqslant \varepsilon + r^*, \tag{4.3c}$$

for some positive constants r^* and s^*.

It appears that (4.3a) is a stronger assumption than (4.2a), since the latter gives no specific lower bound on a. Also, (4.2b) is quite different to (4.3b), in our case the function b is positive and having the upper and lower bound, but in (4.2b) the function is negative. Finally, note that (4.2c) is quite restrictive. For example, (4.3b) is not satisfied for $-\varepsilon u^{(4)} + u'' - u = f$. For that Shanthi and Ramanujam present another technique in [38, **Section 6**]. This so-called "adjoint" approach involves a system of four equations. For details, see Section 1.5.3.

4.2.1 A second-order system

To analyse and solve (4.1), we follow the approach presented in Chapter 3, and transform it into a system of two differential equations. We propose the transformation

$$w := u'' - u. \tag{4.4}$$

That is,

$$u'' = w + u, \tag{4.5}$$

which, when used repeatedly, gives

$$u^{(4)} = w'' + w + u.$$

With this, (4.1) can be transformed into a system of two equations of the form

$$-\varepsilon w'' + (a - \varepsilon)w + (a - \varepsilon - b)u = f, \tag{4.6a}$$

$$-u'' + w + u = 0, \tag{4.6b}$$

subject to the boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{4.6c}$$

We can write this system as

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix} \vec{z}'' + B\vec{z} = \vec{f}, \tag{4.7a}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; B = \begin{pmatrix} a - \varepsilon & a - \varepsilon - b \\ 1 & 1 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{4.7b}$$

The boundary conditions are

$$\vec{z}(0) = \vec{z}(1) = \begin{pmatrix} 0 \\ 0 \end{pmatrix}. \tag{4.7c}$$

4.3 Stability result and maximum principle

This section considers how a maximum principle analysis may be applied to the system. We will use a Jacobi-type analysis to prove stability, which follows ideas from [19]. In doing so, we will assume that all of the conditions in (4.3) hold.

From (4.3a) and (4.3b) note that the diagonal entries of B are positive. We can now define decoupled operators associated with the diagonal entries of B,

$$L_1 w := -\varepsilon w'' + b_{11} w, \tag{4.8a}$$

$$L_2 u := -u'' + u, \tag{4.8b}$$

where $b_{11} = a - \varepsilon$.

Let us recall the standard concept of a *maximum principle*.

Definition 4.3.1. A differential operator L, satisfies a **maximum principle** on the domain Ω, if, for any ψ for which $L\psi(x)$ is defined and non-negative, and $\psi|_\Gamma \geqslant 0$, then $\psi(x) \geqslant 0$, for all $x \in \bar{\Omega}$.

Note that, if L satisfies a maximum principle and, furthermore, $L\psi(x) \leqslant 0$ and $\psi|_\Gamma \leqslant 0$, then $\psi(x) \leqslant 0$, for all $x \in \bar{\Omega}$.

Lemma 4.3.1. *Let* $\vec{v} = (v_1, v_2)^T \in C^2(\Omega)^2 \cap C(\bar{\Omega})^2$, *Then*

$$\|v_1\|_{\bar{\Omega}} \leqslant \|L_1 v_1 / b_{11}\|_{\Omega}, \tag{4.9a}$$

and

$$\|v_2\|_{\bar{\Omega}} \leqslant \|L_2 v_2\|_{\Omega}, \tag{4.9b}$$

where $b_{11} = a - \varepsilon$.

Proof. Since the arguments are analogous, only the details for L_1 are given. Define the barrier function $\xi_1(x) = \|(Lv_1)/b_{11}\|_{\Omega}$ for all $x \in \Omega$. We have

$$L_1\xi_1(x) = L_1\big(\|(L_1v_1)/b_{11}\|_{\Omega}\big)(x) = b_{11}(x)\big(\|(L_1v_1)/b_{11}\|_{\Omega}\big) \geqslant b_{11}(x)\big(|(L_1v_1)/b_{11}(x)|\big) = |L_1v_1(x)|.$$

Applying the maximum principle of Definition 4.3.1 with this barrier function gives

$$L_1(\xi_1(x) \pm v_1(x)) = L_1\xi_1(x) \pm L_1v_1(x) \geqslant 0,$$

since $L_1\xi_1(x) \geqslant |L_1v_1(x)|$. Thus $\xi_1(x) \pm v_1(x) \geqslant 0$, and, therefore, $|v_1(x)| \leqslant \|(L_1v_1)/b_{11}\|_{\Omega}$. □

Next, let us define $\rho_1 := b_{12}/b_{11} = (a - \varepsilon - b)/(a - \varepsilon)$ and $\rho_2 := b_{21} = 1$. Then, from (4.3)

$$0 < (\rho_1\rho_2)(x) = \frac{(a - \varepsilon - b)}{(a - \varepsilon)}(x) < 1, \tag{4.10}$$

for all $x \in \bar{\Omega}$.

Lemma 4.3.2. *Let* $\vec{z} = (w, u)^T$ *be the solution to* (4.7). *For* $k = 0, 1, 2, \ldots$, *we define the sequence of vector-valued functions*

$$\vec{z}^{[k]} = (w^{[k]}, u^{[k]}),$$

where $\vec{z}^{[0]} = \vec{0}$, *and* $\vec{z}^{[k]}$ *solves*

$$L_1w^{[k]} = f - b_{12}u^{[k-1]} \text{ on } \Omega, \qquad \text{and } w^{[k]}(0) = w^{[k]}(1) = 0, \tag{4.11a}$$

$$L_2u^{[k]} = -w^{[k]} \text{ on } \Omega, \qquad \text{and } u^{[k]}(0) = u^{[k]}(1) = 0, \tag{4.11b}$$

for $k = 1, 2, \ldots$. *Then* $\lim_{k\to\infty} \vec{z}^{[k]} = \vec{z}$. *Moreover*

$$\|\vec{z}\|_{\Omega} \leqslant \frac{1}{1 - \rho_1\rho_2}\big(\|f\|/\|b_{11}\|\big).$$

Proof. For $k = 0, 1, 2, \ldots$, we set $\vec{n}^{[k]} = (n_1^{[k]}, n_2^{[k]})^T$ and $\vec{n}^{[k]} = \vec{z} - \vec{z}^{[k]}$. For $k \geqslant 1$, from (4.8) and (4.11), we have

$$L_1n_1^{[k]} = -b_{12}n_2^{[k-1]} \quad \text{on } \Omega, \tag{4.12a}$$

and

$$L_2n_2^{[k]} = -n_1^{[k]} \quad \text{on } \Omega. \tag{4.12b}$$

From Lemma (4.3.1), we get

$$\|n_1^{[k]}\|_{\bar{\Omega}} \leqslant \|b_{12}n_2^{[k-1]}/b_{11}\|_{\Omega} \leqslant \rho_1\|n_2^{[k-1]}\|_{\Omega}, \tag{4.13}$$

and

$$\|n_2^{[k]}\|_{\bar{\Omega}} \leqslant \|n_1^{[k]}\|_{\Omega} \leqslant \rho_2\|n_1^{[k]}\|_{\Omega}. \tag{4.14}$$

So, $\|\vec{n}^{[k]}\|_{\bar{\Omega}} \leqslant \rho_1\rho_2\|\vec{n}^{[k-1]}\| \leqslant (\rho_1\rho_2)^k\|\vec{n}^{[0]}\|$. It follows from (4.10) that $\|\vec{n}^{[k]}\|_{\bar{\Omega}} \to 0$, and so $\lim_{k\to\infty}\vec{z}^{[k]} = \vec{z}$.

Setting $\vec{R}^{[k]} = \vec{z}^{[k]} - \vec{z}^{[k-1]}$ for $k = 1, 2, \ldots$, (4.11) implies that

$$L_1R_1^{[k]} = b_{12}R_2^{[k-1]} \quad \text{on } \Omega, \tag{4.15a}$$

$$L_2R_2^{[k]} = b_{21}R_1^{[k]} \quad \text{on } \Omega. \tag{4.15b}$$

From Lemma (4.3.1), we have

$$\|R_1^{[k]}\|_{\bar{\Omega}} \leqslant \|b_{12}R_2^{[k-1]}/b_{11}\|_{\Omega}, \tag{4.16}$$

and

$$\|R_2^{[k]}\|_{\bar{\Omega}} \leqslant \|R_1^{[k]}/b_{22}\|_{\Omega}. \tag{4.17}$$

Thus $\|R^{[k]}\|_{\Omega} \leqslant \rho_1\rho_2\|R^{[k-1]}\|_{\Omega}$. Therefore $\|R^{[k]}\|_{\Omega} \leqslant (\rho_1\rho_2)^{k-1}\|R^{[1]}\|_{\Omega}$ for $k = 1, 2, 3, \ldots$. Also, since $\vec{z}^{[0]} = \vec{0}$,

$$\|R^{[1]}\|_{\Omega} = \rho_1\rho_2\|\vec{z}^{[1]}\|_{\Omega} \leqslant \|f\|/\|b_{11}\|,$$

and, consequently,

$$\|R^{[k]}\|_{\Omega} \leqslant (\rho_1\rho_2)^{k-1}\big(\|f\|/\|b_{11}\|\big).$$

Finally, since $\vec{z}^{[0]} = 0$, and

$$\vec{z}^{[k]} = R^{[k]} + R^{[k-1]} + \cdots + R^{[1]},$$

$$\|\vec{z}\| = \lim_{j\to\infty}\|\sum_{k=1}^{j}\vec{R}^{[k]}\| \leqslant \lim_{j\to\infty}\sum_{k=1}^{j}(\rho_1\rho_2)^{k-1}\big(\|f\|/\|b_{11}\|\big) = \frac{1}{1-\rho_1\rho_2}\big(\|f\|/\|b_{11}\|\big).$$

□

Lemma 4.3.3. *Let $\vec{z} = (w, u)^T$ be the solution to (4.7) with $f(0) > 0$ for all $x \in [0, 1]$, then $w(x) \geqslant 0$ and $u(x) \leqslant 0$.*

Proof. Consider the sequence $\vec{z}^{[1]}, \vec{z}^{[2]}, \ldots$ of solutions to (4.11). We use induction on k. For $k = 0$, we have $\vec{z}^{[0]} = 0$ and

$$L_1w^{[1]} = f - b_{12}u^{[0]}.$$

Since $u^{[0]} = 0$, and $f \geqslant 0$, so $L_1w^{[1]} \geqslant 0$. Thus, from Definition 4.3.1,

$$w^{[1]} \geqslant 0.$$

Next, we have

$$L_2u^{[1]} = -w^{[1]}.$$

Since $w^{[1]} \geqslant 0$, we get $L_2 u^{[1]} \leqslant 0$, and thus

$$u^{[1]} \leqslant 0.$$

For $k = n$, suppose $w^{[n]} \geqslant 0$ and $u^{[n]} \leqslant 0$. Then

$$L_1 w^{[n+1]} = f - b_{12} u^{[n]} \geqslant 0,$$

since $u^{[n]} \leqslant 0$. Thus $w^{[n+1]} \leqslant 0$. Next

$$L_2 u^{[n+1]} = -w^{[n+1]} \geqslant 0,$$

since $w^{[n+1]} \geqslant 0$. So

$$u^{[n+1]} \leqslant 0.$$

Thus, by induction, $w^{[k]} \geqslant 0$ and $u^{[k]} \leqslant 0$ for all k. Combining with Lemma 4.3.2, $w(x) \geqslant 0$ and $u(x) \leqslant 0$ for all x. □

Now we give an expression for the solution and its derivative for the non-singularly perturbed problem, which we will use on the proof for the following lemma. The proof is elementary: substitute (4.19) into (4.18) to verify it, and differentiate it to verify (4.20).

Lemma 4.3.4. *If $u(x)$ solves*

$$-u''(x) + u(x) = g(x) \text{ on } (0,1), \text{ subject to } u(0) = u(1) = 0, \tag{4.18}$$

then

$$u(x) = \frac{1}{2}\left(- e^{-x} \int_0^x e^s g(s) ds + e^x \int_0^x e^{-s} g(s) ds \right) + c(e^{-x} - e^x), \tag{4.19}$$

where c is a constant that ensures $u(1) = 0$, and depends on (integrals of) g, but not its derivatives. Furthermore,

$$u^{(k)}(x) = \begin{cases} \frac{1}{2}\left(- e^{-x} \int_0^x e^s g(s) ds + e^x \int_0^x e^{-s} g(s) ds \right) + \sum_{j=0}^{k/2-1} g^{(2j)}(x) \\ \qquad +c(e^{-x} - e^x) & k \text{ is even}, \\ \frac{1}{2}\left(e^{-x} \int_0^x e^s g(s) ds + e^x \int_0^x e^{-s} g(s) ds \right) + \sum_{j=0}^{(k-1)/2-1} g^{(2j+1)}(x) \\ \qquad -c(e^{-x} + e^x) & k \text{ is odd}. \end{cases} \tag{4.20}$$

where $k = 1, 2, \ldots$.

We now give sharp pointwise bounds on u, w and their derivatives.

Lemma 4.3.5. *Let $\vec{z} = (w, u)^T$ be the solution to (4.7). Then there exists a constant C, which is independent of ε, such that*

$$|u^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)} \beta_\varepsilon(x)], \tag{4.21}$$

and

$$|w^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)} \beta_\varepsilon(x)], \tag{4.22}$$

where $\beta_\varepsilon(x) := e^{-x/\sqrt{\varepsilon}} + e^{-(1-x)/\sqrt{\varepsilon}}$ *and* $l = 0, 1, \ldots, 4$.

Proof. From Lemma 4.3.2, we have $(w^{[k]}, u^{[k]}) \longrightarrow (w, u)$ as $k \longrightarrow \infty$. We use induction on k. For $k = 0$, we have

$$\|(w^{[0]})^{(l)}\| = 0 \quad \text{and} \quad \|(u^{[0]})^{(l)}\| = 0,$$

for all l. Next, for $k = 1$, note that $w^{[1]}$ solves

$$L_1 w^{[1]} = f,$$

since $u^{[0]} = 0$. So we can apply the result in e.g., [27, **Lemma 6.2**] to get that

$$|(w^{[1]})^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\beta_\varepsilon(x)], \tag{4.23}$$

for $l = 0, 1, \ldots, 4$.

Next, $u^{[1]}$ solves

$$L_2 u^{[1]} = -w^{[1]}. \tag{4.24}$$

From (4.23), it clear that

$$|u^{[1]}(x)| \leqslant C. \tag{4.25}$$

Also, $|(u^{[1]})''(x)| \leqslant C$, because from (4.11), we have

$$L_2 u^{[1]} = -(u^{[1]})'' + u^{[1]} = -w^{[1]}.$$

Therefore

$$(u^{[1]})''(x) = w^{[1]}(x) + u^{[1]}(x).$$

Now, we need to check $|(u^{[1]})'(x)|$, $|(u^{[1]})'''(x)|$, and $|(u^{[1]})^{(4)}(x)|$. From Lemma 4.3.4, we can see that

$$(u^{[1]})'(x) = \frac{1}{2}\left(e^{-x}\int_0^x e^s w^{[1]}(s)ds - e^x \int_0^x e^{-s} w^{[1]}(s)ds\right) - w^{[1]}(x).$$

Combining this with (4.23), we get

$$|(u^{[1]})'(x)| \leqslant C. \tag{4.26}$$

Differentiating (4.11), we get

$$(u^{[1]})'''(x) = (w^{[1]})'(x) + (u^{[1]})'(x), \tag{4.27}$$

and so

$$|(u^{[1]})'''(x)| \leqslant C[1 + \varepsilon^{-1/2}\beta_\varepsilon(x)]. \tag{4.28}$$

Differentiating (4.27), we get

$$(u^{[1]})^{(4)}(x) = (w^{[1]})''(x) + (u^{[1]})''(x). \tag{4.29}$$

From (4.23) and $|(u^{[1]})''(x)| \leqslant C$, we can see that

$$|(u^{[1]})^{(4)}(x)| \leqslant C[1 + \varepsilon^{(-1)}\beta_\varepsilon(x)]. \tag{4.30}$$

For $k = n$, suppose

$$|(w^{[n]})^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\beta_\varepsilon(x)], \tag{4.31}$$

and

$$|(u^{[n]})^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)}\beta_\varepsilon(x)], \tag{4.32}$$

hold for all $l = 0, 1, \ldots, 4$.

For $k = n + 1$, note that $w^{[n+1]}$ solves

$$L_1 w^{[n+1]} = f - b_{12}u^{[n]}. \tag{4.33}$$

From (4.25), we can see that

$$|w^{[n+1]}(x)| \leqslant C. \tag{4.34}$$

We need to check $|(w^{[n+1]})'(x)|$, $|(w^{[n+1]})''(x)|$, $|(w^{[n+1]})'''(x)|$, and $|(w^{[n+1]})^{(4)}(x)|$. By differentiating (4.33) we get

$$(L_1 w^{[n+1]})' = (f - b_{12}u^{[n]})'(x) = f'(x) - b_{12}(u^{[n]})'(x) - b'_{12}u^{[n]}(x). \tag{4.35}$$

From (4.32), we can see that

$$|(w^{[n+1]})'(x)| \leqslant C[1 + \varepsilon^{(-1/2)}\beta_\varepsilon(x)]. \tag{4.36}$$

Repeating the process, and using the bounds established for the lower-order derivatives, gives

$$|(w^{[n+1]})''(x)| \leqslant C[1 + \varepsilon^{(-1)}\beta_\varepsilon(x)], \tag{4.37}$$

$$|(w^{[n+1]})'''(x)| \leqslant C[1 + \varepsilon^{(-3/2)}\beta_\varepsilon(x)], \tag{4.38}$$

and

$$|(w^{[n+1]})^{(4)}(x)| \leqslant C[1 + \varepsilon^{(-2)}\beta_\varepsilon(x)]. \tag{4.39}$$

Next, $u^{[n+1]}$ solves

$$L_2 u^{[n+1]} = -b_{21}w^{[n+1]}. \tag{4.40}$$

So, using the above bounds, along with Lemma (4.3.4), gives

$$|u^{[n+1]}(x)| \leqslant C, \tag{4.41}$$

$$|(u^{[n+1]})'(x)| \leqslant C, \tag{4.42}$$

$$|(u^{[n+1]})''(x)| \leqslant C,$$

$$|(u^{[n+1]})'''(x)| \leqslant C[1+\varepsilon^{-1/2}\beta_\varepsilon(x)], \tag{4.43}$$

and

$$|(u^{[n+1]})^{(4)}(x)| \leqslant C[1+\varepsilon^{(-1)}\beta_\varepsilon(x)]. \tag{4.44}$$

Thus, by induction,

$$|(w^{[k]})^{(l)}(x)| \leqslant C[1+\varepsilon^{(-l/2)}\beta_\varepsilon(x)], \tag{4.45}$$

and

$$|(u^{[k]})^{(l)}(x)| \leqslant C[1+\varepsilon^{(1-l/2)}\beta_\varepsilon(x)], \tag{4.46}$$

for all k. □

4.4 The numerical method

4.4.1 The finite difference method for system

Consider an arbitrary mesh, $\Omega^N := \{0 = x_0 < x_1 < \cdots < x_N = 1\}$, where $h_i = x_i - x_{i-1}$ and $\hbar_i = (x_{i+1} - x_{i-1})/2$. The standard second-order approximation of a second derivative is

$$u''(x_i) \approx D^2 u_i := \frac{1}{h_i\hbar_i}u_{i-1} - (\frac{1}{h_i\hbar_i} + \frac{1}{h_{i+1}\hbar_i})u_i + \frac{1}{\hbar_i h_{i+1}}u_{i+1}. \tag{4.47}$$

The finite difference method for equation (4.7) is: find $\vec{Z}(x_i) = (W_i, U_i)^T$ for $i = 1, \ldots, N-1$. such that

$$\vec{L}^N\vec{Z}(x_i) := -\begin{pmatrix}\varepsilon & 0\\ 0 & 1\end{pmatrix}D^2\vec{Z}(x_i) + B(x_i)\vec{Z}(x_i) = \vec{f}(x_i), \tag{4.48}$$

with the boundary conditions

$$\vec{Z}(x_0) = \vec{Z}(x_N) = 0. \tag{4.49}$$

Writing this as a system of two equations gives

$$\vec{L}^N\vec{Z}(x_i) := \begin{cases} -\varepsilon D^2 W_i + (a(x_i) - \varepsilon)W_i + (a(x_i) - b(x_i) - \varepsilon)U_i = f(x_i), & \text{(4.50a)}\\ -D^2U_i + W_i + U_i = 0, & \text{(4.50b)}\end{cases}$$

with the boundary conditions

$$U_0 = W_0 = U_N = W_N = 0. \tag{4.50c}$$

4.4.2 Shishkin mesh

We construct a standard *Shishkin* mesh with the mesh parameter

$$\tau = \min\{\frac{1}{4}, 2\sqrt{\varepsilon}\ln N\}. \tag{4.51}$$

We now define two mesh transition points at $x = \tau$ and $x = 1 - \tau$. That is, we form a piecewise uniform mesh with $N/4$ equally-sized mesh intervals on each of $[0, \tau]$ and $[1 - \tau, 1]$, and $N/2$ equally-sized mesh intervals on $[\tau, 1-\tau]$. Typically, when ε is small, $\tau \ll 1/4$, the mesh is very fine near the boundaries, and coarse in the interior.

4.4.3 Numerical analysis

Define the decoupled operators

$$L_1^N W := -\varepsilon D^2 W + b_{11} W, \tag{4.52a}$$

$$L_2^N U := -D^2 U + U. \tag{4.52b}$$

We also mention that these scalar operators satisfy a *discrete maximum principle* for finite difference operators.

Definition 4.4.1 (Discrete maximum principle)**.** A finite difference operator L^N, satisfies a **discrete maximum principle** on the domain Ω^N, if, for any ψ for which $L^N\psi(x_i)$ is defined and non-negative, and $\psi|_\Gamma \geqslant 0$, then $\psi(x_i) \geqslant 0$, for all $x_i \in \bar{\Omega}^N$ for $i = 1, \ldots, N-1$.

Note that, if L^N satisfies a maximum principle, then, $L^N\psi(x_i) \leqslant 0$ and $\psi|_\Gamma \leqslant 0$, furthermore, $\psi(x_i) \leqslant 0$, for all $x \in \bar{\Omega}^N$.

Lemma 4.4.1. *Let* $\vec{V} = (V_1, V_2)^T \in C^2(\Omega)^2 \cap C(\bar{\Omega})^2$*, Then*

$$\|V_1\|_{\bar{\Omega}} \leqslant \|(L_1^N V_1)/b_{11}\|_\Omega, \tag{4.53}$$

and

$$\|V_2\|_{\bar{\Omega}} \leqslant \|(L_2^N V_2)\|_\Omega. \tag{4.54}$$

Proof. Since the arguments are analogous, only the details for L_1^N are given. Define the barrier function $\eta_1(x) = \|(L^N V_2)/b_{11}\|_\Omega$ for all $x \in \Omega$. We have

$$L_1^N \eta_1(x) = L_1^N\big(\|(L_1^N V_1)/b_{11}\|\big)(x) = b_{11}(x)\big(\|(L_1^N V_1)/b_{11}\|\big) \geqslant b_{11}(x)\big(|(L_1^N V_1)/b_{11}|\big) = |L_1^N V_1(x)|.$$

Applying the maximum principle of Definition 4.4.1 with this barrier function gives

$$L_1^N(\eta_1(x) \pm V_1(x)) = L_1^N \eta_1(x) \pm L_1^N V_1(x) \geqslant 0,$$

since $L_1^N \eta_1(x) \geqslant |L_1^N V_1(x)|$. Thus $\eta_1(x) \pm V_1(x) \geqslant 0$, and, therefore, $|V_1(x)| \leqslant \|(L_1^N V_1)/b_{11}\|_\Omega$. □

4.4.4 Theorem

We can now state the error result for the method. The proof result follows from the standard arguments (e.g., [27, **Chapter 6**]) thanks to the ε-uniform stability that we have established, and so is not included.

Theorem 4.4.2. *Let Ω^N be the Shishkin mesh defined in Section 4.4.2, and let $\vec{Z} = (W, U)^T$ be the solution to (4.48) on this mesh. If $\vec{z} = (w, u)^T$ solves (4.7), then*

$$\|\vec{z} - \vec{Z}\|_{\bar{\Omega}^N} := \|\hat{u} - \hat{U}\|_{\bar{\Omega}^N} + \|\hat{w} - \hat{W}\|_{\bar{\Omega}^N} \leqslant CN^{-2} \ln^2 N, \tag{4.55}$$

for some constant C.

Remark 4.4.1. From the result of (4.55), we can see that

$$\|\hat{w} - \hat{W}\|_{\bar{\Omega}^N} \leqslant C_1 N^{-2} \ln^2 N, \tag{4.56}$$

and

$$\|\hat{u} - \hat{U}\|_{\bar{\Omega}^N} \leqslant C_2 N^{-2} \ln^2 N, \tag{4.57}$$

for some constants C_1 and C_2. From the numerical results in Section 4.5, we can see that the error bound in (4.56) appears sharp, but, in the bound in (4.57) appears not to be, and, in practice on observes that

$$\|\hat{\mathbf{u}} - \hat{\mathbf{U}}\|_{\bar{\Omega}^N} \leqslant C_2 N^{-2}, \tag{4.58}$$

4.5 Numerical results

In this section, we will present two examples, The first one has constant coefficients, and so the exact solution is available to us when computing errors. The second one has variable coefficients, we do not have its exact solution, so we estimate the error using the benchmark solution computed numerically. We report the error for given N and ε as

$$\tilde{E}_\infty^N := \max_{i=0,\ldots,N} |u(x_i) - U_i|.$$

The associated rate of convergence is

$$\tilde{\rho}^N := \log_2 \left(\frac{E_\infty^N}{E_\infty^{N/2}} \right). \tag{4.59}$$

Example 4.5.1. Consider the following specific example of (4.1). We will take

$$a = 4, \quad b = 2 \quad \text{and} \quad f = 1,$$

so the problem is

$$-\varepsilon u^{(4)}(x) + 4u''(x) - 2u(x) = 1.$$

Written as a system this is

$$-\varepsilon w'' + (4 - \varepsilon)w + (2 - \varepsilon)u = 1, \tag{4.60a}$$

$$-u'' + w + u = 0, \tag{4.60b}$$

subject to the boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{4.60c}$$

The coupling matrix is

$$B = \begin{pmatrix} (4-\varepsilon) & 2-\varepsilon \\ 1 & 1 \end{pmatrix}.$$

We can see that $\rho_1 := b_{12}/b_{11} = (2-\varepsilon)/(4-\varepsilon)$ and $\rho_2 := b_{21} = 1$. Then, from (4.3)

$$|\rho_1\rho_2| = \frac{2-\varepsilon}{4-\varepsilon} < 1. \tag{4.61}$$

In Figure 4.1 we show u with $\varepsilon = 10^{-1}$ (left), and $\varepsilon = 10^{-3}$ (right), neither of which features layers. In Figure 4.2 we show w with $\varepsilon = 10^{-1}$ (left), which does not features layers. In contrast, as shown in the graph on the right for smaller ε (in this case $\varepsilon = 10^{-3}$), the solution possesses boundary layers near $x = 0$ and $x = 1$.

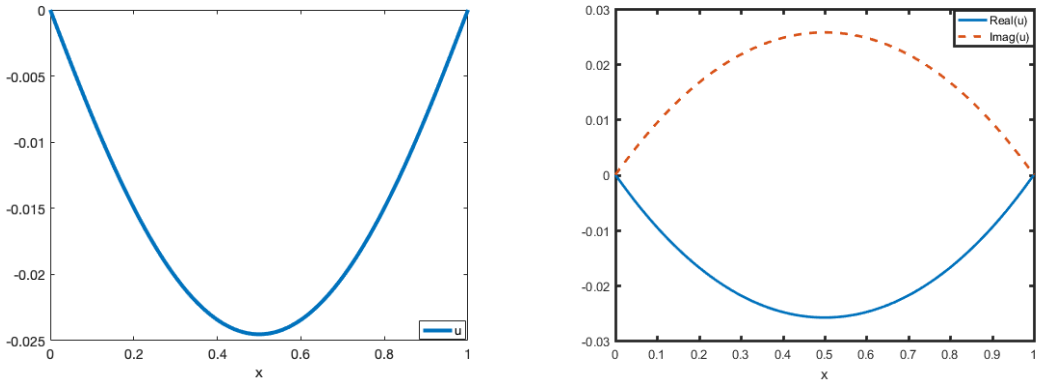

Figure 4.1: The solutions u to (4.60) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

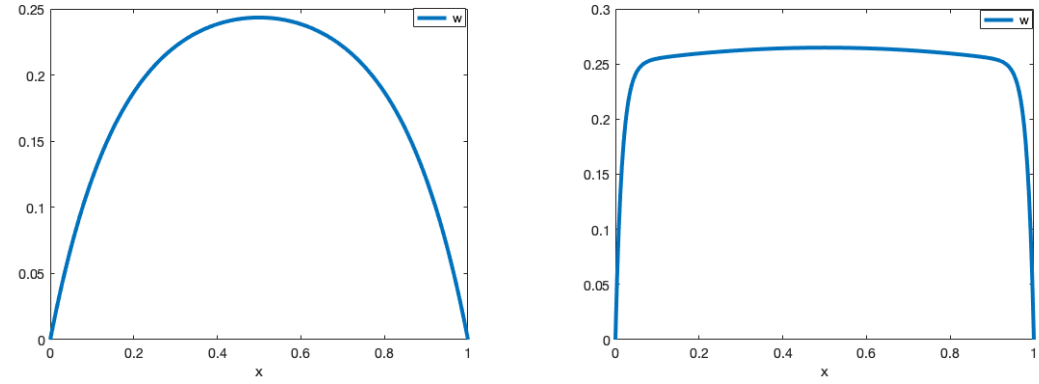

Figure 4.2: The transformation w to (4.60) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

Tables 4.1 and 4.3 present the pointwise computed when (4.60) is solved by the finite difference scheme on Shishkin mesh. So, for small ε, we can see that the pointwise convergence is parameter uniform. Tables 4.2 and 4.4 demonstrate that this convergence is fully second-order for u which, as previously noted in Remark 4.4.1, suggests that the error bound is suboptimal for this component. For the w component, we observe almost second-order convergence. That is, we observe the logarithmic factor in practice, suggesting the bound for w is sharp.

Table 4.1: $\tilde{E}_{\infty}^{N}(u)$ for problem (4.60) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.8824e-05	4.7130e-06	1.1787e-06	2.9470e-07	7.3676e-08
1.00e-02	1.1731e-05	3.2272e-06	8.6635e-07	2.1815e-07	5.4638e-08
1.00e-04	8.6333e-06	1.7469e-06	3.5013e-07	6.9793e-08	1.3938e-08
1.00e-06	1.6993e-05	4.1795e-06	1.0273e-06	2.5245e-07	6.2028e-08
1.00e-08	1.8054e-05	4.5095e-06	1.1257e-06	2.8097e-07	7.0127e-08
1.00e-10	1.8163e-05	4.5434e-06	1.1359e-06	2.8394e-07	7.0972e-08
1.00e-12	1.8174e-05	4.5468e-06	1.1369e-06	2.8423e-07	7.1059e-08

Table 4.2: $\tilde{\rho}^{N}(u)$ for problem (4.60) computed on a Shishkin mesh.

ε	$N = 16 - 32$	$N = 32 - 64$	$N = 64 - 128$	$N = 128 - 256$
1	1.998	1.999	2.000	2.000
1e-02	1.862	1.897	1.990	1.997
1e-04	2.305	2.319	2.327	2.324
1e-06	2.024	2.024	2.025	2.025
1e-08	2.001	2.002	2.002	2.002
1e-10	1.999	2.000	2.000	2.000
1e-12	1.999	2.000	2.000	2.000

Table 4.3: $\tilde{E}_{\infty}^{N}(w)$ for problem (4.60) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	6.6395e-05	1.6617e-05	4.1553e-06	1.0389e-06	2.5973e-07
1.00e-02	5.1087e-03	1.3945e-03	3.6594e-04	9.2109e-05	2.3112e-05
1.00e-04	1.0325e-02	7.8559e-03	3.7418e-03	1.3257e-03	4.5039e-04
1.00e-06	1.0325e-02	7.8560e-03	3.7419e-03	1.3257e-03	4.5039e-04
1.00e-08	1.0325e-02	7.8560e-03	3.7419e-03	1.3257e-03	4.5039e-04
1.00e-10	1.0325e-02	7.8560e-03	3.7419e-03	1.3257e-03	4.5039e-04
1.00e-12	1.0325e-02	7.8560e-03	3.7419e-03	1.3257e-03	4.5039e-04

Table 4.4: $\tilde{\rho}^{N}(w)$ for problem (4.60) computed on a Shishkin mesh.

ε	$N = 16 - 32$	$N = 32 - 64$	$N = 64 - 128$	$N = 128 - 256$
1	1.998	2.000	2.000	2.000
1e-02	1.873	1.930	1.990	1.995
1e-04	0.394	1.070	1.497	1.558
1e-06	0.394	1.070	1.497	1.558
1e-08	0.394	1.070	1.497	1.558
1e-10	0.394	1.070	1.497	1.558
1e-12	0.394	1.070	1.497	1.558

Example 4.5.2. We consider another example for (4.1), but with variable coefficients. We will take

$$a(x) = 2x + 1, \quad b(x) = 4x + 1, \quad \text{and} \quad f(x) = 1 + x.$$

We have

$$-\varepsilon w'' + (2x + 1 - \varepsilon)w + (-2x - \varepsilon)u = 1 + x, \tag{4.62a}$$

$$-u'' + w + u = 0, \tag{4.62b}$$

subject to the boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{4.62c}$$

We can write this system as

$$-\begin{pmatrix} \varepsilon & 0 \\ 0 & 1 \end{pmatrix} \vec{z}'' + B\vec{z} = \vec{f},$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; B = \begin{pmatrix} (2x+1-\varepsilon) & -2x-\varepsilon \\ 1 & 1 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} 1+x \\ 0 \end{pmatrix}.$$

We can see that $\rho_1 := b_{12}/b_{11} = (-2x-\varepsilon)/(2x+1-\varepsilon)$ and $\rho_2 := b_{21} = 1$. Then, from (4.3)

$$|\rho_1\rho_2| = |\frac{-2x-\varepsilon}{2x+1-\varepsilon}| < 1, \tag{4.64}$$

providing $\varepsilon < 1/2$.

For this system, we will report the maximum pointwise error for given N and ε as $\widetilde{E}_\infty^N(u)$ and $\widetilde{E}_\infty^N(w)$ in u and w, respectively, and by $\widetilde{\rho}_\infty^N(u)$ and $\widetilde{\rho}_\infty^N(w)$ the corresponding rates of convergence of u and w.

In Figure 4.3, we plot solutions to u (left) and w (right) with $\varepsilon = 10^{-3}$. As in the previous example, w exhibits a layer, but u does not.

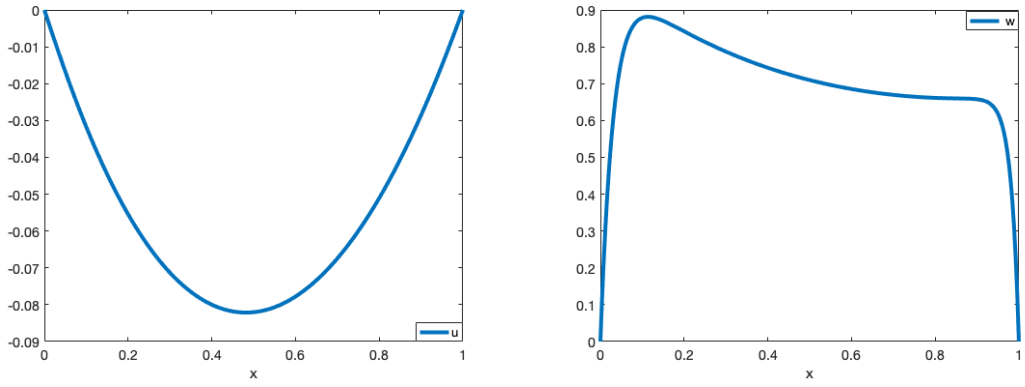

Figure 4.3: The solutions to u (left) and w (right) with $\varepsilon = 10^{-3}$ to (4.62).

Tables 4.5 and 4.7 present the pointwise errors computed when (4.62) is solved by the finite difference scheme on Shishkin mesh. The error initially increases as ε decreases, but for the smallest values of ε, the method is robust. Tables 4.6 and 4.8 we can see that the numerical solution converges at a rate that is a full second-order for u, and an almost second-order for w, independently of ε.

Table 4.5: $\tilde{E}_\infty^N(u)$ for problem (4.62) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	3.7820e-05	9.4743e-06	2.3698e-06	5.9253e-07	1.4814e-07
1.00e-02	6.5015e-05	1.7221e-05	4.3392e-06	1.0904e-06	2.7275e-07
1.00e-04	1.0654e-04	2.2285e-05	5.8369e-06	1.4701e-06	3.5802e-07
1.00e-06	2.1411e-04	5.2935e-05	1.3039e-05	3.2015e-06	7.8587e-07
1.00e-08	2.2798e-04	5.7489e-05	1.4366e-05	3.5885e-06	8.9557e-07
1.00e-10	2.2940e-04	5.7959e-05	1.4503e-05	3.6291e-06	9.0718e-07
1.00e-12	2.2954e-04	5.8006e-05	1.4517e-05	3.6332e-06	9.0835e-07

Table 4.6: $\tilde{\rho}^N(u)$ for problem (4.62) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.997	1.999	2.000	2.000
1e-02	1.917	1.989	1.993	1.999
1e-04	2.257	1.933	1.989	2.038
1e-06	2.016	2.021	2.026	2.026
1e-08	1.988	2.001	2.001	2.002
1e-10	1.985	1.999	1.999	2.000
1e-12	1.984	1.998	1.998	2.000

Table 4.7: $\tilde{E}^N_\infty(w)$ for problem (4.62) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	7.5637e-05	1.8954e-05	4.7427e-06	1.1857e-06	2.9645e-07
1.00e-02	1.1253e-02	3.0288e-03	7.7268e-04	1.9439e-04	4.8742e-05
1.00e-04	2.6725e-02	1.7795e-02	7.6278e-03	2.7326e-03	9.0323e-04
1.00e-06	2.6581e-02	1.7723e-02	7.5971e-03	2.7210e-03	8.9932e-04
1.00e-08	2.6567e-02	1.7716e-02	7.5940e-03	2.7198e-03	8.9894e-04
1.00e-10	2.6566e-02	1.7715e-02	7.5937e-03	2.7197e-03	8.9890e-04
1.00e-12	2.6566e-02	1.7715e-02	7.5937e-03	2.7197e-03	8.9889e-04

Table 4.8: $\tilde{\rho}^N(w)$ for problem (4.62) computed on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.997	1.999	2.000	2.000
1e-02	1.893	1.971	1.991	1.996
1e-04	0.587	1.222	1.481	1.597
1e-06	0.585	1.222	1.481	1.597
1e-08	0.585	1.222	1.481	1.597
1e-10	0.585	1.222	1.481	1.597
1e-12	0.585	1.222	1.481	1.597

Chapter 5

A general fourth-order complex-valued singularly perturbed problem

5.1 Introduction

In this chapter, we are interested in the properties of certain fourth-order, complex-valued reaction-diffusion differential equations. Specifically, we discuss how these equations can be reformulated as coupled systems in various ways and what these formulations reveal about the equations and their solutions. For example, they can be used to study the stability of the differential operators or the positivity/negativity of their solutions.

Specifically, this chapter aims to present a new technique for transforming general fourth-order complex-valued singularly perturbed problems into coupled systems of real-valued second-order equations. This then allows us to prove the coercivity of the coupling matrix of this system, subject to certain conditions on the coefficients. Such results can be used, in turn, in finite element analyses; see Section 3.3.3.

We also propose an iterative approach for solving (exactly) this system, and we show that it converges uniformly with respect to the singular perturbation parameter. This iterative process is useful for several theoretical reasons: it allows us to establish bounds on the components of the solution, determine their sign, and derive bounds on their derivatives. All these are useful in the analysis of finite difference methods for these problems.

Since this chapter features a study of the mathematical properties of the equations, and not their numerical solution, we do not discuss the approximation of solutions in this chapter. That is the subject of Chapters 6 and 7. However, to help with the exposition, we will present the results of some computations based on the MATLAB "Chebfun" package [12].

This chapter is organized as follows. In Section 5.2, we present a general fourth-order complex-

valued problem, discuss how it can be solved using Chebfun, and present a motivating example. In Section 5.3 we show ways to rewrite such a problem in terms of real-valued systems. First, in Section 5.3.1, we transform our model fourth-order complex-valued problem into a system of two fourth-order real-valued problems. Then, in Section 5.3.2 (following the exposition in Chapter 3) we present a further transformation into a real-valued second-order system. Again we show how to solve this problem with "Chebfun", and verify that all three formulations are essentially equivalent. In Section 5.4 we show how to determine the value of the parameters in the transformation (subject to reasonable assumptions) that ensure that the resulting coefficient matrix of the system's zero-order term is coercive. Such a result is very important, especially in the context of finite element analysis; see Chapter 6. In Sections 5.4.2 and 5.4.3 we demonstrate how to use an eigenvalue test to apply this analysis to different cases.

In Section 5.5 we tackle a different form of analysis of the differential operator. That is, we establish the stability result of the differential operator for the system of four equations solved using a Gauss-Seidel method. Such stability results are key to proving the convergence of finite difference methods: see Chapter 7. We present a new block-iterative method that extends the analysis to this system. Also, we present a fully iterative method by using the ideas from Lemma 4.3.2 to show that the method converges. In Section 5.6 we establish the Maximum Principle of differential operator for the system of four equations solve by the fully iterative method. Also, we state the bounds for the solution and its derivative in Lemma 5.6.2.

5.2 A general fourth-order problem

5.2.1 The equation

The fourth-order complex-valued reaction-diffusion differential equation we will study is, in its most general form,

$$-\varepsilon u^{(4)}(x) + au''(x) - bu(x) = f(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{5.1a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{5.1b}$$

Here, as usual in this thesis, $\varepsilon \in (0,1]$. However, unlike Chapters 3 and 4, we assume that the coefficient functions, a and b, and the right-hand side function, f, are complex-valued functions on the interval Ω, and may have non-zero imaginary parts. However, it is often convenient to express (5.1a) in terms of real-valued coefficients:

$$-\varepsilon u^{(4)}(x) + (a_r + ia_i)u''(x) - (b_r + ib_i)u(x) = (f_r + if_i)(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{5.2}$$

where

$$a = a_r + ia_i, \quad b = b_r + ib_i, \quad \text{and } f = f_r + if_i, \tag{5.3}$$

and a_r, a_i, b_r, b_i, f_r and f_i are all real-valued functions; of course, $i = \sqrt{-1}$ is the imaginary unit.

As alluded to earlier, we are interested in the properties of the differential operator when recast as a coupled system. Throughout this chapter, we shall endeavour to make minimal assumptions on the problem data, and will distinguish between those needed to ensure that, for example, the operator is coercive (see Section 5.4) and those that ensure an iterative algorithm for the system converges (see Section 5.5). However, in all instances, we will assume that there exists a positive ϱ such that

$$a_r(x) \geqslant \varrho > 0 \qquad \text{for all } x \in \bar{\Omega}. \tag{5.4}$$

5.2.2 Solving with Chebfun

As mentioned in the introduction, in this chapter, we are not interested in the numerical solution of this equation. However, we do wish to compute (extremely accurate) solutions as part of the exposition.

Chebfun [12] is a freely available MATLAB [26] toolbox for working with functions of real variables. It aims to combine the efficiency of numerical computing systems (like MATLAB or Octave [13]) and the ease of use of symbolic computing systems (like Maple [28], SageMath [41] or Mathematica [17]). It does this by representing functions that a user defines by extremely accurate polynomial approximations, and then performing computations on these polynomials. Chebyshev methods are used to construct these polynomial approximations (called "*chebfuns*"). Then the toolbox allows the user to perform standard tasks, such as integration or differentiation of the functions, or even solving differential equations with these Chebfuns as coefficients. The software makes use of MATLAB's object-oriented programming, so that the technical details are hidden from the user while making the package very easy to use.

The original version of Chebfun dates from 2004 [4]. However, it is under continuous development, particularly to extend its functionality to higher-dimensional problems [12].

There has been relatively little work on applying Chebfun to solving singularly perturbed problems, particularly because representing functions with layer-type behaviour can be difficult. However, it is known to be feasible, if the correct splitting is used [1].

In this chapter, we wish to present methods of solving fourth-order problems iterative, where each iterate is a second-order equation. If each of these equations can be solved exactly (rather than numerically), we will see that the sequence converges. This is why we use Chebfun.

First, however, we will solve (5.2) non-iteratively. In Figure 5.1 to do this, where ε, a_r, a_i, b_r, b_i, f_r and f_i are defined terms.

Figure 5.1: MATLAB/Chebfun code to solve (5.2)

```
L = chebop(@(x,u) -epsilon*diff(u,4) + (ar+i*ai)*diff(u,2) - (br+i*bi)*u -
      f, domain);
rhs=0;
L.bc = @(x,u) [u(0); feval(diff(u,2),0); u(1); feval(diff(u,2),1)];
u = solvebvp(L, rhs, options);
w = diff(u,2);
```

5.2.3 A motivating example

We consider an example of (5.2) with $a_r = 4$, $a_i = 2$, $b_r = 6$, $b_i = 2$, $f_r = 1$ and $f_i = 0$, i.e.,

$$-\varepsilon u^{(4)}(x) + (4 + 2i)u''(x) - (6 + 2i)u(x) = 1 \qquad \text{on} \qquad \Omega := (0, 1). \tag{5.5a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{5.5b}$$

This is easily solved using the code presented in Figure 5.1, setting the appropriate values for the coefficients. In Figure 5.2, we show the real and imaginary parts of u computed with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right). Note that neither feature obvious layers.

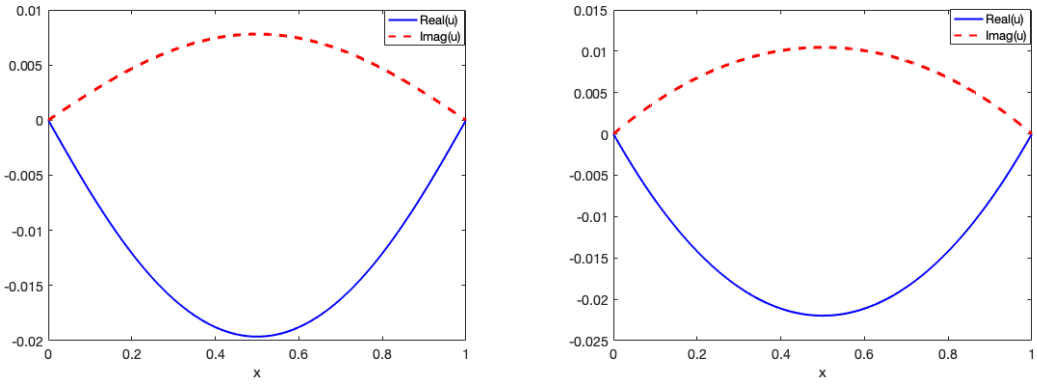

Figure 5.2: Solutions to (5.5) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

However, there are weak layers present in the solutions to (5.5). These can be seen in Figure 5.3, where we plot the second derivative of the solution, u''. Again, on the left, we have $\varepsilon = 10^{-1}$, which does not feature layers, while, in contrast, on the right, for smaller ε (in this case $\varepsilon = 10^{-3}$), there are layers near $x = 0$ and $x = 1$, in both the real and the imaginary parts.

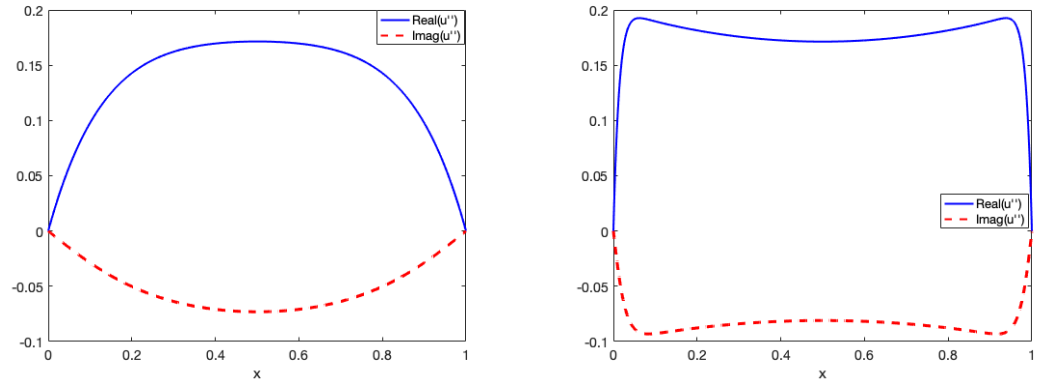

Figure 5.3: The second derivative of solutions to (5.5) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

5.3 From a 4th-order complex-valued problem to a real-valued system of 2nd-order problems

In this section, we investigate how to transform (5.2) into a coupled system of four second-order real-valued problems in such a way that the coupling matrix is coercive. The first step is to transform (5.2) into a system of two real-valued fourth-order problems. Then we transform this into a system of four real-valued, second-order differential equations.

5.3.1 Writing (5.2) as a system of real-valued, fourth-order equations

We will show how to transform a fourth-order complex-valued problem into a coupled system of two fourth-order problems, following the ideas in Section 2.4.1 for a second-order complex-valued problem. That is, we equate the real and imaginary parts in (5.2) separately. Specifically, as with (5.3), we let u_r and u_i be real-valued functions such that

$$u = u_r + iu_i,$$

and which satisfy

$$-\varepsilon u_r^{(4)} + a_r u_r'' - a_i u_i'' - b_r u_r + b_i u_i = f_r, \tag{5.6a}$$

$$-\varepsilon u_i^{(4)} + a_i u_r'' + a_r u_i'' - b_i u_r - b_r u_i = f_i. \tag{5.6b}$$

It is clear that the solution to (5.6) satisfies (5.2). Although it is not crucial to our approach, we note that each of these problems could be solved using the real-valued finite difference method of Section 4.4.

The system in (5.6) can also be solved using Chebfun, with, for example, the code in Figure 5.4. We don't present graphs of the results since they are indistinguishable from those shown in Figure 5.3: they differ by approximately 10^{-16}.

Figure 5.4: MATLAB/Chebfun code for solving (5.6)

```
domain = [0,1];
L_system = chebop(@(x,ur,ui) ...
  [-epsilon*diff(ur,4) + ar*diff(ur,2) - ai*diff(ui,2) - br*ur + bi*ui - fr;
   -epsilon*diff(ui,4) + ai*diff(ur,2) + ar*diff(ui,2) - bi*ur - br*ui - fi], ...
      domain);
options = chebopppref();
rhs_system = [0;0];
L_system.bc = @(x,ur,ui) ...
   [ur(0); feval(diff(ur,2),0); ur(1); feval(diff(ur,2),1);
    ui(0); feval(diff(ui,2),0); ui(1); feval(diff(ui,2),1)];
[ur, ui] = solvebvp(L_system, rhs_sys2, options);
wr = diff(ur,2);
wi = diff(ui,2);
```

However, (5.6) is just an interim transformation: we actually wish to reduce (5.6) to a system of four second-order real-valued equations. But there is more than one way to do this, and different approaches may have different properties (desirable or otherwise).

5.3.2 Writing (5.2) as a system of real-valued, second-order equations

In this section, we will propose a transformation for converting the pair of real-valued fourth-order equations (5.6) into a system of four second-order real-valued equations. The transformation features two parameters that can be chosen to ensure that the resulting system has certain properties. We'll discuss a specific example that ensures that the coupling matrix is coercive.

We propose a transformation defined by introducing the term

$$w := \frac{u'' - \beta u}{\alpha}, \tag{5.7}$$

where α and β are non-zero, real-valued constants chosen depending on the problem data. Clearly (5.7) makes sense only when $\alpha \neq 0$. We discuss why $\beta \neq 0$ in Remark 5.4.1.

The expression in (5.7) can be written as

$$u'' = \alpha w + \beta u. \tag{5.8}$$

Differentiating twice, we get

$$u^{(4)} = \alpha w'' + \alpha\beta w + \beta^2 u.$$

With this, (5.6) can be transformed to a system of four equations of the form

$$-\varepsilon\alpha w_r'' + \alpha(a_r - \varepsilon\beta)w_r - \alpha a_i w_i + (a_r\beta - \varepsilon\beta^2 - b_r)u_r + (b_i - a_i\beta)u_i = f_r, \tag{5.9a}$$
$$-\varepsilon\alpha w_i'' + \alpha a_i w_r + \alpha(a_r - \varepsilon\beta)w_i + (a_i\beta - b_i)u_r + (a_r\beta - \varepsilon\beta^2 - b_r)u_i = f_i, \tag{5.9b}$$
$$-u_r'' + \alpha w_r + \beta u_r = 0, \tag{5.9c}$$
$$-u_i'' + \alpha w_i + \beta u_i = 0, \tag{5.9d}$$

with boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{5.9e}$$

Setting $\vec{z} = (w_r, w_i, u_r, u_i)^T$, we can write (5.9) in matrix form:

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon\alpha & 0 & 0 & 0 \\ 0 & \varepsilon\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\vec{z}'' + B\vec{z} = \vec{f}, \tag{5.10a}$$

where

$$B = \begin{pmatrix} \alpha(a_r - \varepsilon\beta) & -\alpha a_i & a_r\beta - \varepsilon\beta^2 - b_r & b_i - a_i\beta \\ \alpha a_i & \alpha(a_r - \varepsilon\beta) & a_i\beta - b_i & a_r\beta - \varepsilon\beta^2 - b_r \\ \alpha & 0 & \beta & 0 \\ 0 & \alpha & 0 & \beta \end{pmatrix} \text{ and } \vec{f} = \begin{pmatrix} f_r \\ f_i \\ 0 \\ 0 \end{pmatrix}. \tag{5.10b}$$

It is easy to verify that any $\vec{z}$ satisfying (5.10) will correspond to a solution (5.1). However, the form of the problem in (5.10) has numerous advantages, in theory and practice, over (5.1).

From a theoretical view point, the analysis of second-order problems is more developed than that for fourth-order problems, so these analyses can be applied to (5.10) (most books on boundary value problems are primarily focused on second-order equations; few even mention fourth-order problems).

From a practical view-point, there are far more numerical solvers available for second-order problems compared to fourth-order ones. If these solvers are for coupled systems, they can be applied directly to (5.10); this is done in Figure 5.5 below where the `Chebfun` solver in MATLAB is used. (Later, in Section 5.5 we discuss how to apply solvers for scalar equations to this problem). We have verified that the solution given by the code in Figure 5.5 agrees with that given in Figure 5.4 up to machine precision.

Figure 5.5: MATLAB/Chebfun code that solves (5.10)

```
B = [alpha*(ar-epsilon*beta),-alpha*ai, ar*beta-br-epsilon*beta^2, bi-ai*beta;
     alpha*ai, alpha*(ar-epsilon*beta), ai*beta-bi, ar*beta-br-epsilon*beta^2;
     alpha,    0,                       beta,       0;
     0,        alpha,                   0,          beta];
L_system = chebop(@(x,wr,wi,ur,ui)...
[-epsilon*alpha*diff(wr,2)+B(1,1)*wr + B(1,2)*wi + B(1,3)*ur + B(1,4)*ui - real(f);
 -epsilon*alpha*diff(wi,2)+B(2,1)*wr + B(2,2)*wi + B(2,3)*ur + B(2,4)*ui - imag(f);
               -diff(ur,2)+B(3,1)*wr + B(3,2)*wi + B(3,3)*ur + B(3,4)*ui - 0;
               -diff(ui,2)+B(4,1)*wr + B(4,2)*wi + B(4,3)*ur + B(4,4)*ui - 0],
    domain);
options = chebopppref();
rhs_system = [0;0;0;0];
L_system.bc = @(x, wr, wi, ur, ui) [wr(0); wi(0); ur(0); ui(0);
                                    wr(1); wi(1); ur(1); ui(1)];
[Wr, Wi, Ur, Ui] = solvebvp(L_system, rhs_system, options);
```

For the remainder of this chapter, we will consider the problem (5.1) in the form given in (5.10), for a coercivity analysis (Section 5.4.2) and convergence of an iterative method (Section 5.4.3).

5.4 Ensuring a coercive system matrix

5.4.1 Coercivity

Recall from Definition 3.2.1, that a matrix A is *coercive*, if there exists a constant $\gamma > 0$ such that

$$\frac{\vec{v}^T A \vec{v}}{\vec{v}^T \vec{v}} \geqslant \gamma \quad \text{for all } \vec{v} \in \mathbb{R}^2/\{(0,0)^T\}. \tag{5.11}$$

Furthermore, it is coercive if, and only if, the symmetric part of A, $M = (A + A^T)/2$, has positive eigenvalues.

Remark 5.4.1. It is important to note that, although for any particular problem's data, there may a large range of possible α and β in (5.7), that can be chosen to ensure the coercivity of the coupling matrix, it is typically required that $\beta \neq 0$. This is significant, because it excludes the most common form of transformations from fourth-order problems to second-order systems: setting $w := u''$.

To see this, suppose that we did choose $\beta = 0$. Now let $\vec{v} = (1, 0, 0, v_4)^T$. Then we have

$$\vec{v}^T B\vec{v} = \begin{pmatrix} 1 & 0 & 0 & v_4 \end{pmatrix} \begin{pmatrix} \alpha a_r & -\alpha a_i & -b_r & b_i \\ \alpha a_i & \alpha a_r & -b_i & -b_r \\ \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \end{pmatrix} \begin{pmatrix} 1 \\ 0 \\ 0 \\ v_4 \end{pmatrix},$$

giving

$$\vec{v}^T B\vec{v} = \alpha a_r + b_i v_4.$$

So, for any α, a_r and $b_i \neq 0$, we can choose v_4, such that $\vec{v}^T A\vec{v} = -1 < 0$. For example, for the problem in (5.5) where $a_r = 4$ and $b_i = 2$, if, for example, $\alpha = 1$, then $\vec{v}^T A\vec{v} = -1$ when $v_4 = -5/2$.

In the case where $b_i = 0$, a similar calculation can be preformed with $\vec{v} = (0, 1, 0, v_4)^T$ to show that, if $\beta = 0$ then for any α, one can chose a v_4 such that $\vec{v}^T B\vec{v} < 0$.

A further justification that one cannot take $\beta = 0$, but based on a spectral analysis of B, is given in Remark 5.4.2.

5.4.2 The eigenvalue test for coercivity

In this section, we apply the eigenvalue test of Section 3.2.2 to the matrix B in (5.10b), in order to determine values of α and β that ensure it is coercive.

Recall from Theorems 3.2.2 and 3.2.1 that the matrix B satisfies $\vec{v}^T B\vec{v} > 0$ for all $\vec{v}$ if, and only if, $M = (B^T + B)/2$ is symmetric positive definite. From (5.10b), we have

$$B = \begin{pmatrix} \alpha(a_r - \varepsilon\beta) & -\alpha a_i & a_r\beta - \varepsilon\beta^2 - b_r & b_i - a_i\beta \\ \alpha a_i & \alpha(a_r - \varepsilon\beta) & a_i\beta - b_i & a_r\beta - \varepsilon\beta^2 - b_r \\ \alpha & 0 & \beta & 0 \\ 0 & \alpha & 0 & \beta \end{pmatrix}, \tag{5.12}$$

and

$$M = (B^T + B)/2 = \tag{5.13}$$

$$\begin{pmatrix} \alpha(\alpha a_r - \varepsilon\beta) & 0 & \dfrac{\beta a_r - \varepsilon\beta^2 - b_r + \alpha}{2} & \dfrac{-\beta a_i + b_i}{2} \\ 0 & \alpha(a_r - \beta\varepsilon) & \dfrac{\beta a_i - b_i}{2} & \dfrac{-\varepsilon\beta^2 + \beta a_r - b_r + \alpha}{2} \\ \dfrac{\beta a_r - \varepsilon\beta^2 - b_r + \alpha}{2} & \dfrac{\beta a_i - b_i}{2} & \beta & 0 \\ \dfrac{-\beta a_i + b_i}{2} & \dfrac{-\varepsilon\beta^2 + \beta a_r - b_r + \alpha}{2} & 0 & \beta \end{pmatrix}.$$

Obviously, the matrix M is symmetric. Furthermore, M is positive definite if and only if all of its eigenvalues are positive. By direct calculation, one can observe that M has two distinct

eigenvalues, both with geometric multiplicity two:

$$\lambda_1 = \frac{1}{2}\Big[(-\beta\varepsilon + a_r)\alpha + \beta + \Big(\varepsilon^2\beta^4 - 2\varepsilon a_r\beta^3 + (\alpha^2\varepsilon^2 + a_i^2 + a_r^2 + 2b_r\varepsilon + 1)\beta^2 + (-2\alpha^2 a_r\varepsilon - 2a_i b_i - 2a_r b_r)\beta + (a_r^2 + 1)\alpha^2 - 2b_r\alpha + b_i^2 + b_r^2\Big)^{1/2}\Big], \quad (5.14a)$$

and

$$\lambda_2 = \frac{1}{2}\Big[(-\beta\varepsilon + a_r)\alpha + \beta - \Big(\varepsilon^2\beta^4 - 2\varepsilon a_r\beta^3 + (\alpha^2\varepsilon^2 + a_i^2 + a_r^2 + 2b_r\varepsilon + 1)\beta^2 + (-2\alpha^2 a_r\varepsilon - 2a_i b_i - 2a_r b_r)\beta + (a_r^2 + 1)\alpha^2 - 2b_r\alpha + b_i^2 + b_r^2\Big)^{1/2}\Big]. \quad (5.14b)$$

Since M is symmetric, λ_1 and λ_2 are real numbers. Consequently, we can see that $\lambda_1 \geqslant \lambda_2$ for any a_r, a_i, b_r, b_i, α and β.

Example 5.4.1. Recall the example in (5.5) where $a_r = 4$, $a_i = 2$, $b_r = 6$, $b_i = 2$ and $\varepsilon = 10^{-4}$. Suppose we choose that $\alpha = 14$ and $\beta = 2$. Then, from (5.10b), we have

$$B = \begin{pmatrix} 55.9972 & -28 & 1.9996 & -2 \\ 28 & 55.9972 & 2 & 1.9996 \\ 14 & 0 & 2 & 0 \\ 0 & 14 & 0 & 2 \end{pmatrix},$$

and

$$M = (B + B^T)/2 = \begin{pmatrix} 55.9972 & 0 & 7.9998 & -1 \\ 0 & 55.9972 & 1 & 7.9998 \\ 7.9998 & 1 & 2 & 0 \\ -1 & 7.9998 & 0 & 2 \end{pmatrix}.$$

The eigenvalues of M are $\lambda_1 = 57.17521$ and $\lambda_2 = 0.821993$. So, M is symmetric positive matrix, and, consequently, the matrix B is coercive, and $\gamma \approx 0.821993$.

Remark 5.4.2. As discussed in Remark 5.4.1, it is important to note that, although typically there maybe some freedom to chose α and β in (5.7), this freedom is not absolute. Specifically, one may not take $\beta = 0$. This can be observed from the calculation given in (5.14), as the following example illustrates.

Suppose $\beta = 0$ in (5.8). Then, we have

$$w := \frac{u''}{\alpha}. \quad (5.15)$$

With this, we get

$$B = \begin{pmatrix} a_r\alpha & -a_i\alpha & -b_r & b_i \\ a_i\alpha & a_r\alpha & -b_i & -b_r \\ \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \end{pmatrix}. \quad (5.16)$$

Suppose we use the same example as presented in (5.4.1) with $a_r = 4$, $a_i = 2$, $b_r = 6$ and $b_i = 2$.

Then, from (5.16)

$$B = \begin{pmatrix} 4\alpha & -2\alpha & -6 & 2 \\ 2\alpha & 4\alpha & -2 & -6 \\ \alpha & 0 & 0 & 0 \\ 0 & \alpha & 0 & 0 \end{pmatrix},$$

which is not coercive, because from Theorems 3.2.1 and 3.2.2, since

$$(B + B^T)/2 = \begin{pmatrix} 4\alpha & 0 & -3 + \alpha/2 & 1 \\ 0 & 4\alpha & -1 & -3 + \alpha/2 \\ -3 + \alpha/2 & -1 & 0 & 0 \\ 1 & -3 + \alpha/2 & 0 & 0 \end{pmatrix},$$

has the eigenvalues $\lambda_1 = (4\alpha + \sqrt{17\alpha^2 - 12\alpha + 40})/2$ and $\lambda_2 = (4\alpha - \sqrt{17\alpha^2 - 12\alpha + 40})/2$.

Suppose (for argument) that $\lambda_2 > 0$ for *some* α. In that case one would have $4\alpha - \sqrt{17\alpha^2 - 12\alpha + 40} > 0$, which would imply $16\alpha^2 > 17\alpha^2 - 12\alpha + 40$. But it is easy to verify that $-\alpha^2 + 12\alpha - 40$ is negative for all α (with a maximum value of -4). Thus there is no real α for which $\lambda_2 > 0$.

Of course, one can also directly verify that the quantity $\lambda_2 = 2\alpha - \sqrt{17\alpha^2 - 12\alpha + 40})2$ is maximized when $\alpha = (6 + 8\sqrt{171})/17 \approx 6.324$, which would give $\lambda_2 \approx -0.00405$.

5.4.3 Using (5.14) to determine α and β

Equipped with the results of the calculation in (5.14), for specific problem data, one can determine if there are values of α and β for which (5.11) is satisfied and, if so, what are suitable choices. In certain sub-classes of (5.2) one may also derive general bounds for α and β. As a demonstration, in this section, we show how to do that for two cases. To simplify the presentation, we will always take $\beta = 1$; it is our experience that for any reasonable problem data, unless β is exceptionally small, one can still find a suitable α.

Case 1: Suppose that $a_i = 0$ in the problem data. Then, from (5.14b), $\lambda_2 > 0$ providing that

$$a_r + b_r - \varepsilon - \sqrt{4a_r b_r - 4b_r\varepsilon - b_i^2} < \alpha < a_r + b_r - \varepsilon + \sqrt{4a_r b_r - 4b_r\varepsilon - b_i^2}. \tag{5.17}$$

Suppose we use the same example as presented in (5.5) but with $a_i = 0$, $a_r = 4$, $b_r = 6$, $b_i = 2$ and $\varepsilon = 10^{-4}$, so from (5.17), then one can choose any α such that

$$0.4083620628 < \alpha < 19.59143794.$$

If one wants to maximise γ, then (to 8 digits) one should take $\alpha = 2.89811046$, which would yield $\gamma = 0.88774662$.

Case 2. Suppose that $b_i = 0$ in the problem data. Then, from (5.14b), $\lambda_2 > 0$ providing that

$$a_r + b_r - \varepsilon - \sqrt{4a_r b_r - a_i^2 - 4b_r\varepsilon} < \alpha < a_r + b_r - \varepsilon + \sqrt{4a_r b_r - a_i^2 - 4b_r\varepsilon}. \tag{5.18}$$

Suppose we have another example with $a_r = 7$, $a_i = 5$, $b_r = 3$ and $\varepsilon = 10^{-4}$, so from (5.18),

we have (to 8 digits)

$$2.318832366 < \alpha < 17.68096763.$$

Again, if one wants to maximise γ, then one should take $\alpha = 6.4617817$, which would yield $\gamma = 0.251261318$.

5.5 Iterative solution of the second-order system

In this section, we establish a stability result for the differential operator in (5.9). To do this, we show that (5.9) may be solved using a Gauss-Seidel method. This is for theoretical (rather than practical purposes) since it allows us to invoke the stability theory developed for uncoupled problems. Our general approach is quite different from that presented in [19, §II], and which we used in Section 4.3, since that was only for two equations. Especially, the analysis in [19, §II], for a general system of $\ell \geqslant 2$ equations require that the coupling matrix be strictly diagonally dominant, a property which our system does not enjoy. In Section 4.3 we exploited the observation in [19, **Remark 2.4**] to gave much less stringent conditions, but only in the case of a system of two equations. However, we now present a new block-iterative method that extends the analysis to a system of four equations.

To ensure convergence, we require that the problem data, and our choice of the positive transformation parameters, α and β, satisfy the following conditions. Recalling from (5.4) that $a_r(x) \geqslant \varrho > 0$, we shall, in addition, assume

$$0 \leqslant |a_i(x)| < a_r(x) - \varepsilon\beta \qquad \text{for all } x \in [0,1]. \tag{5.19}$$

Note that this implies $0 < \varepsilon\beta < a_r(x)$, which we also use.

5.5.1 A block iterative method

To prove the convergence of the method, we'll introduce a new "block" iterative scheme. This is similar in style to the version for two systems, but used recursively. The key idea is that we split the system of four equations into two systems of two equations. The semi-decoupled operators are

$$\hat{L}_1\hat{\mathbf{w}} := -\varepsilon\alpha\hat{\mathbf{w}}'' + \hat{B}_{11}\hat{\mathbf{w}}, \tag{5.20a}$$

$$\hat{L}_2\hat{\mathbf{u}} := -\hat{\mathbf{u}}'' + \hat{B}_{22}\hat{\mathbf{u}} \tag{5.20b}$$

where

$$\hat{B}_{11} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} \alpha(a_r - \varepsilon\beta) & -\alpha a_i \\ \alpha a_i & \alpha(a_r - \varepsilon\beta) \end{pmatrix},$$

$$\text{and} \quad \hat{B}_{22} = \begin{pmatrix} b_{33} & b_{34} \\ b_{43} & b_{44} \end{pmatrix} = \begin{pmatrix} \beta & 0 \\ 0 & \beta \end{pmatrix}. \tag{5.20c}$$

For later use, we also define

$$\hat{B}_{12} = \begin{pmatrix} b_{13} & b_{14} \\ b_{23} & b_{24} \end{pmatrix} = \begin{pmatrix} a_r\beta - \varepsilon\beta^2 - b_r & b_i - a_i\beta \\ a_i\beta - b_i & a_r\beta - \varepsilon\beta^2 - b_r \end{pmatrix},$$

$$\text{and} \quad \hat{B}_{21} = \begin{pmatrix} b_{31} & b_{32} \\ b_{41} & b_{42} \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}. \tag{5.20d}$$

We'll also denote

$$\hat{\mathbf{f}}_1 = \begin{pmatrix} f_r \\ f_i \end{pmatrix}.$$

We define

$$\rho := \left|\frac{b_{12}}{b_{11}}\right| = \left|\frac{b_{21}}{b_{22}}\right| = \frac{|a_i|}{a_r - \varepsilon\beta},$$

and

$$\theta = \min_{j=1,2} \min_{x\in\bar{\Omega}} |b_{jj}(x)| = \min_{x\in\bar{\Omega}} \big(\alpha(a_r - \varepsilon\beta)\big).$$

Note that, from (5.19), we have

$$0 < \rho(x) < 1 \quad \text{for all } x \in \bar{\Omega}.$$

Now we can state that, if a $\hat{\mathbf{w}}$ solves

$$\hat{L}_1\hat{\mathbf{w}} = \hat{\mathbf{g}}, \qquad \hat{\mathbf{w}}(0) = \hat{\mathbf{w}}(0) = \vec{0},$$

for arbitrary $\hat{\mathbf{g}}$, then it follows from Lemma 4.3.2, and the fact that $0 < \rho^2 < 1$, that

$$\|\hat{\mathbf{w}}\| \leqslant \frac{1}{1-\rho^2}\frac{\|\hat{\mathbf{g}}\|}{\theta}. \tag{5.21a}$$

Also, if $\hat{\mathbf{u}}$ solves

$$\hat{L}_2\hat{\mathbf{u}} = \hat{\mathbf{g}}, \qquad \hat{\mathbf{u}}(0) = \hat{\mathbf{u}}(0) = \vec{0},$$

then,

$$\|\hat{\mathbf{u}}\| \leqslant \frac{1}{\beta}\|\hat{\mathbf{g}}\|, \tag{5.21b}$$

which comes from the standard maximum principle, since $\hat{B}_{22}$ is a diagonal matrix.

Theorem 5.5.1. *Let $\hat{\mathbf{w}} = (w_r, w_i)^T$ and $\hat{\mathbf{u}} = (u_r, u_i)^T$ be solution to (5.20). For $k = 0, 1, 2, \ldots,$ let $\hat{\mathbf{w}}^{[k]}$ and $\hat{\mathbf{u}}^{[k]}$ be defined as follows: set $\hat{\mathbf{w}}^{[0]}$ and $\hat{\mathbf{u}}^{[0]}$ to be the vector-valued zero functions, and, for $k = 1, 2, 3, \ldots,$ $\hat{\mathbf{w}}^{[k]}$ and $\hat{\mathbf{u}}^{[k]}$ to be the solutions to*

$$\hat{L}_1\hat{\mathbf{w}}^{[k]} = \hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]} \qquad \text{subject to } \hat{\mathbf{w}}^{[k]}(0) = \hat{\mathbf{w}}^{[k]}(1) = \vec{0} \quad \text{on } (0,1), \tag{5.22a}$$

$$\hat{L}_2\hat{\mathbf{u}}^{[k]} = -\hat{B}_{21}\hat{\mathbf{w}}^{[k]} \qquad \text{subject to } \hat{\mathbf{u}}^{[k]}(0) = \hat{\mathbf{u}}^{[k]}(1) = \vec{0} \quad \text{on } (0,1), \tag{5.22b}$$

If

$$0 < \frac{\alpha}{\beta}\frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} < 1, \tag{5.23}$$

then $\lim_{k\to\infty} \hat{\mathbf{w}}^{[k]} = \hat{\mathbf{w}}$, and $\lim_{k\to\infty} \hat{\mathbf{u}}^{[k]} = \hat{\mathbf{u}}$.

Proof. For $k = 0, 1, 2, \ldots$, we set $\hat{\mathbf{m}}^{[k]} = \hat{\mathbf{u}} - \hat{\mathbf{u}}^{[k]}$, and $\hat{\mathbf{n}}^{[k]} = \hat{\mathbf{w}} - \hat{\mathbf{w}}^{[k]}$. For $k \geqslant 1$, from (5.20a) and (5.22a) we have

$$\hat{L}_1 \hat{\mathbf{n}}^{[k]} = -\hat{B}_{12} \hat{\mathbf{m}}^{[k-1]} \text{ on } \Omega. \tag{5.24}$$

Then, from (5.21a),

$$\|\hat{\mathbf{n}}^{[k]}\|_{\bar{\Omega}} \leqslant \frac{1}{(1-\rho^2)\theta} \|\hat{B}_{12} \hat{\mathbf{m}}^{[k-1]}\| \leqslant \frac{1}{(1-\rho^2)\theta} \|\hat{B}_{12}\| \|\hat{\mathbf{m}}^{[k-1]}\|, \tag{5.25}$$

where here $\|\hat{B}_{12}\|$ is the usual (sub-multiplicative) matrix norm induced by the maximum vector norm $\|\cdot\|$. Furthermore, since

$$\hat{L}_2 \hat{\mathbf{m}}^{[k]} = -\hat{B}_{21} \hat{\mathbf{n}}^{[k]} \text{ on } \Omega. \tag{5.26}$$

we get from (5.21b),

$$\|\hat{\mathbf{m}}^{[k]}\|_{\bar{\Omega}} \leqslant \frac{\alpha}{\beta} \|\hat{\mathbf{n}}^{[k]}\|_{\Omega}. \tag{5.27}$$

Combining these we get that

$$\|\hat{\mathbf{m}}^{[k]}\| \leqslant \frac{\alpha}{\beta} \frac{1}{(1-\rho^2)\theta} \|\hat{B}_{12}\| \|\hat{\mathbf{m}}^{[k-1]}\|. \tag{5.28}$$

It follows from (5.23) that $\|\hat{\mathbf{m}}^{[k]}\|_{\bar{\Omega}} \to 0$, and so $\lim_{k\to\infty} \hat{\mathbf{u}}^{[k]} = \hat{\mathbf{u}}$. This in turn implies that $\lim_{k\to\infty} \hat{\mathbf{w}}^{[k]} = \hat{\mathbf{w}}$. □

In Figure 5.6, we show the way to solve the system of two equations by a block-iterative method that extends the analysis to a system of four equations using Chebfun.

We now turn to investigate how one can ensure that (5.23) holds for a specific example. First note, that all the terms in inequality are positive, so it is enough to ensure that the upper bound holds. We start by writing out the full expression:

$$\begin{aligned}
\frac{\alpha}{\beta} \frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} &= \frac{\alpha}{\beta} \frac{|a_r\beta - \varepsilon\beta^2 - b_r| + |b_i - a_i\beta|}{1-\rho^2} \frac{1}{\min_{x\in\bar{\Omega}} \alpha(a_r - \varepsilon\beta)} \\
&= \frac{1}{\beta} \frac{|a_r\beta - \varepsilon\beta^2 - b_r| + |b_i - a_i\beta|}{1-\rho^2} \frac{1}{\min_{x\in\bar{\Omega}} a_r - \varepsilon\beta} \\
&\leqslant \frac{1}{\beta} \frac{|a_r\beta - \varepsilon\beta^2 - b_r| + |b_i - a_i\beta|}{1-\rho^2} \frac{1}{a_r - \varepsilon\beta}.
\end{aligned}$$

Notice that this expression is independent of the choice of α. Furthermore, we are interested in the case where ε is small; specifically, where $a_r - \varepsilon\beta \approx a_r$. In this case we get that $\rho \approx a_i/a_r$, and so $1/(1-\rho^2) \approx a_i^2/(a_r^2 - a_i^2)$. In addition, $1/(a_r - \varepsilon\beta) \approx 1/a_r$. Thus

$$\frac{\alpha}{\beta} \frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} \lessapprox \frac{1}{\beta} \big(|a_r\beta - b_r| + |b_i - a_i\beta|\big) \frac{a_i^2}{a_r^2 - a_i^2} \frac{1}{a_r}.$$

This quantity can be (approximately, up to terms of $\mathcal{O}(\varepsilon)$) minimised by taking

$$\beta = b_r/a_r \qquad \text{or} \qquad \beta = b_i/a_i, \tag{5.29}$$

as appropriate to the specific case.

Remark 5.5.1. We stress that the calculation leading to (5.29) is only with regard to the optimal

Figure 5.6: MATLAB/Chebfun code that implements the block-iterative algorithm in (5.20)

```
domain = [0, 1];
B = [alpha*(ar-epsilon*beta), -alpha*ai, ar*beta-br-epsilon*beta^2, bi-ai*beta;
     alpha*ai,  alpha*(ar-epsilon*beta), ai*beta-bi, ar*beta-br-epsilon*beta^2;
     alpha,                          0,       beta,            0;
         0,                      alpha,          0,         beta];
L1 = chebop(@(x,wr,wi) [-epsilon*alpha*diff(wr,2)+B(1,1)*wr + B(1,2)*wi;
                        -epsilon*alpha*diff(wi,2)+B(2,1)*wr + B(2,2)*wi], domain);
F1 = f(1:2);
L1.bc = @(x, wr, wi) [wr(0); wi(0); wr(1); wi(1)];
L2 = chebop(@(x,ur,ui) [-diff(ur,2) + B(3,3)*ur;
                        -diff(ui,2) + B(4,4)*ui], domain);
F2 = f(3:4);
L2.bc = @(x, ur, ui) [ur(0); ui(0); ur(1); ui(1)];
Wr = chebfun(@(x)0, domain); Wi = chebfun(@(x)0, domain);
Ur = chebfun(@(x)0, domain); Ui = chebfun(@(x)0, domain);

MaxIterations = 12;
BlockTOL = 1.0e-12;
Diff_k = BlockTOL+1;
wr=Wr; wi=Wi; ur=Ur; ui=Ui;
k=0;
while ( (k<MaxIterations) && (Diff_k > BlockTOL) )
   k=k+1;
   RHS1 = F1 - B(1:2,3:4)*[Ur;Ui];
   [Wr,Wi] = solvebvp(L1, RHS1);
   RHS2 = F2 - B(3:4,1:2)*[Wr;Wi];
   [Ur,Ui] = solvebvp(L2, RHS2);
   Diffs(k, :) = [norm(wr-Wr), norm(wi-Wi), norm(ur-Ur), norm(ui-Ui)];
   fprintf('Iteration: (%3d) ||wr-Wr||=%9.3e, ||wi-Wi||=%9.3e ||ur-Ur||=%9.3e, ||ui
    -Ui||=%9.3e \n', k, Diffs(k,:));
   Diff_k = max(Diffs(k,:));
   wr=Wr; wi=Wi; ur=Ur; ui=Ui;
end
```

choice of β, with respect to optimising the rate of the convergence of the method. Indeed, in certain cases of constant coefficients, it can ensure convergence to machine precision in two or three iterations. But we emphasise that *any* choice of β for which (5.23) holds is acceptable in theory.

Example 5.5.1. Suppose we use the same example as presented in (5.5) with $a_r = 4$, $a_i = 2$, $b_r = 6$ and $b_i = 2$, also, we suppose we choose $\alpha = 14$ and $\beta = b_r/a_r = 1.5$. Then, from (5.23) we have

$$\frac{\alpha}{\beta}\frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} = 0.2223 < 1.$$

In Figure 5.7, we show the convergence of the block iterative method, implemented in Chebfun, for a solving system of four equations of the example (5.5.1). Notice that it converges rapidly: after 8 iterations, the errors are essential "machine epsilon" (roughly 2.2×10^{-16}).

Theorem 5.5.2. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to* (5.20). *Define*

$$\rho_1 := \frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} \quad \textit{and} \quad \rho_2 = \frac{\alpha}{\beta}.$$

Then

$$\|\hat{\mathbf{w}}\| \leqslant \left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta}, \tag{5.30}$$

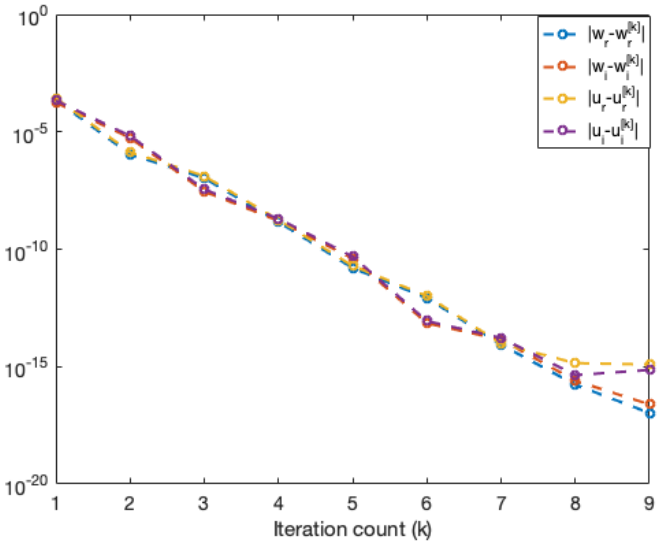

Figure 5.7: Convergence of a block iterative method for solving Example 5.5.1

and

$$\|\hat{\mathbf{u}}\| \leqslant \frac{\alpha}{\beta}\|\hat{\mathbf{w}}\| \leqslant \frac{\alpha}{\beta}\left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta}. \tag{5.31}$$

Proof. Set $\hat{\mathbf{R}}_1^{[k]} = \hat{\mathbf{w}}^{[k]} - \hat{\mathbf{w}}^{[k-1]}$ and $\hat{\mathbf{R}}_2^{[k]} = \hat{\mathbf{u}}^{[k]} - \hat{\mathbf{u}}^{[k-1]}$. Since $\hat{\mathbf{w}}^{[0]} \equiv \vec{0}$ and $\hat{\mathbf{u}}^{[0]} \equiv \vec{0}$, we have

$$\hat{\mathbf{w}}^{[k]} = \sum_{j=1}^{k} \hat{\mathbf{R}}_1^{[k]} \quad \text{and} \quad \hat{\mathbf{u}}^{[k]} = \sum_{j=1}^{k} \hat{\mathbf{R}}_2^{[k]},$$

for $k = 1, 2, \ldots$. From (5.21) we get

$$\|\hat{\mathbf{R}}_1^{[k]}\| \leqslant \frac{1}{(1-\rho^2)\theta}\|\hat{B}_{12}\|\|\hat{\mathbf{R}}_2^{[k-1]}\|, \tag{5.32}$$

and

$$\|\hat{\mathbf{R}}_2^{[k]}\| \leqslant \frac{\alpha}{\beta}\|\hat{\mathbf{R}}_1^{[k-1]}\|. \tag{5.33}$$

So, from (5.32), we have

$$\|\hat{\mathbf{R}}_1^{[k]}\| \leqslant \rho_1\rho_2\|\hat{\mathbf{R}}_1^{[k-1]}\| \leqslant (\rho_1\rho_2)^{k-1}\|\hat{\mathbf{R}}_1^{[1]}\|.$$

To get a bound for $\hat{\mathbf{R}}_1^{[1]}$, use that $\hat{\mathbf{R}}_1^{[1]} = \hat{\mathbf{w}}_1^{[1]}$. Since $\hat{\mathbf{w}}_1^{[1]}$ satisfies

$$\hat{L}_1\hat{\mathbf{w}}^{[1]} = \hat{\mathbf{f}}_1, \quad \text{subject to} \quad \hat{\mathbf{w}}^{[1]}(0) = \hat{\mathbf{w}}^{[1]}(1) = \vec{0},$$

we can deduce that

$$\|\hat{\mathbf{w}}^{[1]}\| \leqslant \frac{1}{(1-\rho^2)\theta}\|\hat{\mathbf{f}}_1\|.$$

Therefore,

$$\|\hat{\mathbf{R}}_1^{[k]}\| \leqslant (\rho_1\rho_2)^{k-1}\frac{1}{(1-\rho^2)\theta}\|\hat{\mathbf{f}}_1\|.$$

Since $\hat{\mathbf{w}}^{[k]} = \sum_{j=1}^{k} \hat{\mathbf{R}}_1^{[k]}$ and $\hat{\mathbf{w}} = \lim_{k\to\infty} \hat{\mathbf{w}}^{[k]}$, we get that

$$\|\hat{\mathbf{w}}\| = \lim_{k\to\infty} \|\sum_{j=1}^{k} \hat{\mathbf{R}}_1^{[k]}\| \leqslant \lim_{k\to\infty} \sum_{j=1}^{k} \|\hat{\mathbf{R}}_1^{[k]}\|$$
$$\leqslant \lim_{k\to\infty} \sum_{j=1}^{k} (\rho_1\rho_2)^{k-1} \frac{1}{(1-\rho^2)\theta} \|\hat{\mathbf{f}}_1\| = \left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta}.$$

Since the operator $\hat{L}_2$ is uncoupled, it easily follows that

$$\|\hat{\mathbf{u}}\| \leqslant \frac{\alpha}{\beta}\|\hat{\mathbf{w}}\|,$$

which concludes the proof. □

The bounds given in Theorem 5.5.2 shows that $\|\hat{\mathbf{u}}\|$ and $\|\hat{\mathbf{w}}\|$ are bounded independently of ε (ρ and ρ_1 do depend on ε, but upper and lower bounds on them to not). These bounds can be verified numerically. The following example does that and investigates their sharpness.

Example 5.5.2. Suppose we use the same example as presented in (5.5) with $a_r = 4$, $a_i = 2$, $b_r = 6$ and $b_i = 2$, also, we suppose we choose $\alpha = 14$ and $\beta = b_r/a_r = 1.5$. The solutions are shown in Figure 5.8. One can observe that

$$\|\hat{\mathbf{w}}\| \approx 1.46 \times 10^{-2} \qquad \text{and} \qquad \|\hat{\mathbf{u}}\| \approx 2.20 \times 10^{-2}.$$

Applying from (5.30) and (5.31) we get

$$\|\hat{\mathbf{w}}\| \leqslant \left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta} = 0.0306,$$

and

$$\|\hat{\mathbf{u}}\| \leqslant \frac{\alpha}{\beta}\|\hat{\mathbf{w}}\| = 0.2858.$$

These bounds are also shown in Figure 5.8: Notice that the bounds are correct, but especially in the case of $\|\hat{\mathbf{u}}\|$, not particularly sharp. This is largely because the bound we use in (5.21a) is quite sharp for small ε, the same is not true of (5.21b), where the diffusion coefficient is 1.

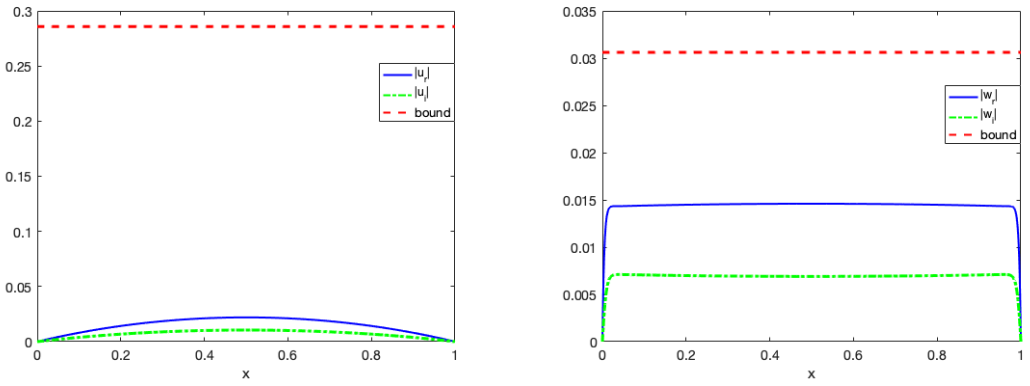

Figure 5.8: The bound of $\hat{u}$ (left) and the bound of $\hat{w}$ (right) of the example (5.5.2).

5.5.2 A fully iterative method

In Section 5.5.1 we presented the following block-iterative method for solving (5.9): set $\hat{\mathbf{w}}^{[0]}$ and $\hat{\mathbf{u}}^{[0]}$ to be the vector-valued zero functions, and, for $k = 1, 2, 3, \ldots$, $\hat{\mathbf{w}}^{[k]}$ and $\hat{\mathbf{u}}^{[k]}$ to be the solutions to

$$\hat{L}_1 \hat{\mathbf{w}}^{[k]} = \hat{\mathbf{f}}_1 - \hat{B}_{12} \hat{\mathbf{u}}^{[k-1]}, \tag{5.34a}$$

$$\hat{L}_2 \hat{\mathbf{u}}^{[k]} = -\hat{B}_{21} \hat{\mathbf{w}}^{[k]}, \tag{5.34b}$$

where

$$\hat{L}_1 \hat{\mathbf{w}}^{[k]} := -\begin{pmatrix} \varepsilon\alpha & 0 \\ -0 & \varepsilon\alpha \end{pmatrix} \begin{pmatrix} \hat{w}_r^{[k]} \\ \hat{w}_i^{[k]} \end{pmatrix}'' + \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} \begin{pmatrix} \hat{w}_r^{[k]} \\ \hat{w}_i^{[k]} \end{pmatrix}$$

and

$$\hat{L}_2 \hat{\mathbf{u}}^{[k]} := -\begin{pmatrix} \hat{u}_r^{[k]} \\ \hat{u}_i^{[k]} \end{pmatrix}'' + \begin{pmatrix} b_{33} & 0 \\ 0 & b_{44} \end{pmatrix} \begin{pmatrix} \hat{u}_r^{[k]} \\ \hat{u}_i^{[k]} \end{pmatrix}$$

So the operator $\hat{L}_1$ is itself a coupled pair of linear differential operators, since typically $b_{12} \neq 0$ and $b_{21} \neq 0$ (on the other hand, $\hat{L}_2$ is also a pair of linear differential operators, but they are uncoupled). So, to get a fully iterative method, we can solve (5.34a) and (5.34b) in an iterative way.

We now define decoupled operators associated with the diagonal entries of B in (5.12):

$$L_{1;1}\psi := -\varepsilon\alpha\psi'' + b_{11}\psi, \tag{5.35a}$$

$$L_{1;2}\psi := -\varepsilon\alpha\psi'' + b_{22}\psi, \tag{5.35b}$$

$$L_{2;1}\psi := -\psi'' + b_{33}\psi, \tag{5.35c}$$

$$L_{2;2}\psi := -\psi'' + b_{44}\psi. \tag{5.35d}$$

Then we can solve (5.34a) and (5.34b) iteratively, for a fixed k, as follows. We set $\hat{w}_i^{[k;0]} = \hat{w}_i^{[k-1]}$, and solve

$$\begin{aligned} L_{1;1}\hat{w}_r^{[k;j]} &= (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]})_1 - b_{12}\hat{w}_i^{[k;j-1]} \qquad \text{subject to } \hat{w}_r^{[k;j]}(0) = \hat{w}_r^{[k;j]}(1) = \vec{0}, \\ L_{1;2}\hat{w}_i^{[k;j]} &= (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]})_2 - b_{21}\hat{w}_r^{[k;j]} \qquad \text{subject to } \hat{w}_i^{[k;j]}(0) = \hat{w}_i^{[k;j]}(1) = \vec{0}, \end{aligned} \tag{5.36a}$$

for $j = 1, 2, 3, \ldots$, setting $\hat{w}^{[k]}$ to be the limit of the sequence $\{\hat{w}^{[k;j]}\}_{j=0}^{\infty}$. Then solve

$$\begin{aligned} L_{2;1}\hat{u}_r^{[k]} &= (-\hat{B}_{21}\hat{\mathbf{w}}^{[k]})_1 \qquad \text{subject to } \hat{u}_r^{[k]}(0) = \hat{u}_r^{[k]}(1) = \vec{0}, \\ L_{2;2}\hat{u}_2^{[k]} &= (-\hat{B}_{21}\hat{\mathbf{w}}^{[k]})_2 \qquad \text{subject to } \hat{u}_i^{[k]}(0) = \hat{u}_i^{[k]}(1) = \vec{0}. \end{aligned} \tag{5.36b}$$

Using the ideas from Lemma 4.3.2, and the fact that $b_{11}(x)b_{22}(x) > b_{12}(x)b_{21}(x)$, the iteration in (5.36a) converges in the sense that $\hat{\mathbf{w}}^{[k;j]} \to \hat{\mathbf{w}}^{[k]}$, as $j \to \infty$, where $\hat{\mathbf{w}}^{[k]}$ is as defined in (5.22a). So this now gives a fully iterative method (i.e., we solve scalar equations at every step).

As mentioned previously, standard maximum principle results for coupled systems of two linear reaction-diffusion operators, such as $\hat{L}_1$ assume the coupling (i.e., reaction) matrix is an M-matrix: it is strictly diagonally dominant with positive diagonal entries, and non-positive off-diagonal entries. But, of course, this last assumption cannot hold for $\hat{L}_1$, since $b_{21} = -b_{12}$. However, we still have diagonal dominance, so, using the iterative method just described, one can adapt the

arguments [19, **Lemma 2.3**], to show that, if $\hat{w}^{[k]}$ solves

$$\hat{L}_1\hat{w} = \hat{f} \text{ on } (0,1), \qquad \text{and } \hat{w}(0) = \hat{w}(1) = \vec{0},$$

and $\vec{f}$ is such that $\hat{f}_1 \geqslant 0$ and $\hat{f}_2 \leqslant 0$, then $\hat{w}_1(x) \geqslant 0$ and $\hat{w}_2(x) \leqslant 0$; the arguments only require applying the appropriate maximum/minimum condition for scalar problems.

5.6 A monotonicity result for the differential operator defined in (5.9)

In this section, we consider how a maximum principle-type analysis may be applied to the system in (5.9), by using the iteration in (5.36). This is for theoretical (rather than practical purposes) since it allows us to invoke the stability theory developed for uncoupled problems. As such, we are using the idea of *Gauss-Seidel Iteration*, which was introduced as a theoretical tool in [19, **§II**], and which we applied earlier in Section 4.3. However, that was for a much simpler setting, involving just two equations. So, although the iteration presented here is similar to that of [19, **§II**], the supporting theory is very different. Specifically, in that work it is assumed that the coupling matrix is row-diagonally dominant, which is not the case here. Thus, the details are quite different.

Recall the concept of a maximum principle (Definition 4.3.1) and the associated minimum principle. We now wish to investigate how related ideas can be extended to the system we consider here.

For this analysis we require the assumption on a and b given in (5.19), which gives that

$$b_{11} = b_{22} > 0, \qquad \text{and} \qquad 0 < \frac{|b_{12}|}{b_{11}} = \frac{b_{21}}{b_{22}} < 1.$$

In addition, we assume that we can choose β sufficiently large that

$$0 < \frac{\varepsilon\beta^2 + b_r}{\beta} < \varrho. \tag{5.37a}$$

and

$$\frac{b_i}{\beta} < a_i. \tag{5.37b}$$

Respectively, these ensure that

$$b_{13} = b_{24} > 0, \quad \text{and} \quad b_{14} = -b_{23} < 0.$$

It transpires that the signs of the terms u_r, u_i, w_r, and w_i depend on the signs of f_r and f_i (of course, other assumptions on the problem data determine if u_r, u_i, w_r, and w_i change sign or not). The results of several examples are shown in Table 5.1, where we have taken

$$a_r(x) = 4 + x, \qquad a_i(x) = e^x, \qquad b_r(x) = 6 + \sqrt{x}, \qquad \text{and} \qquad b_i(x) = e^{-x}, \tag{5.38}$$

and, for simplicity, only one of f_r and f_i are non-zero.

In Figure 5.9 we show u_r and u_i (left) and w_r and w_i (right) with $\varepsilon = 10^{-4}$, a and b as given in (5.38), and

$$f_r = 1, \quad f_i = 0, \quad \alpha = 14, \quad \text{and} \quad \beta = 2. \tag{5.39}$$

In Figure 5.10 we corresponding results, but for

$$f_r = 0, \quad f_i = 1, \quad \alpha = 14, \quad \text{and} \quad \beta = 2. \tag{5.40}$$

Table 5.1: How the signs of components of u and w depend on the sign of f_r and f_i, with problem data as in (5.38)

f_r	f_i	w_r	w_i	u_r	u_i
$f_r > 0$	$f_i = 0$	$w_r > 0$	$w_i < 0$	$u_r < 0$	$u_i > 0$
$f_r = 0$	$f_i > 0$	$w_r > 0$	$w_i > 0$	$u_r < 0$	$u_i < 0$
$f_r < 0$	$f_i = 0$	$w_r < 0$	$w_i > 0$	$u_r > 0$	$u_i < 0$
$f_r = 0$	$f_i < 0$	$w_r < 0$	$w_i < 0$	$u_r > 0$	$u_i > 0$

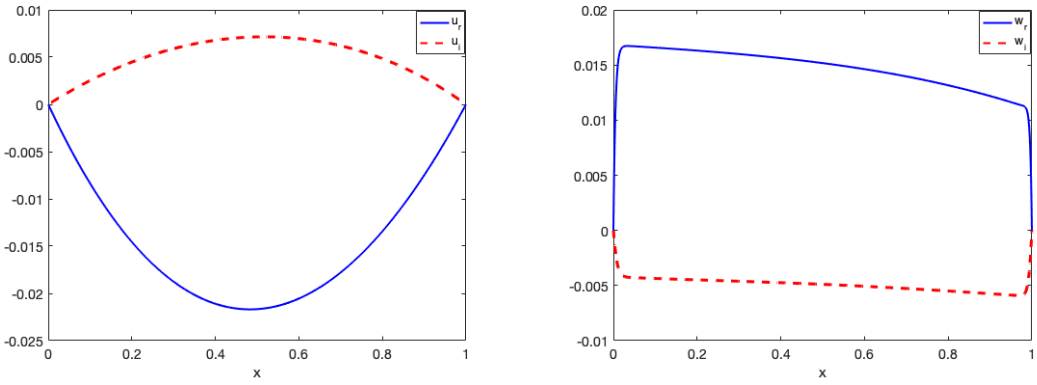

Figure 5.9: The solutions, u and w to (5.9), with data as in (5.38) and (5.39).

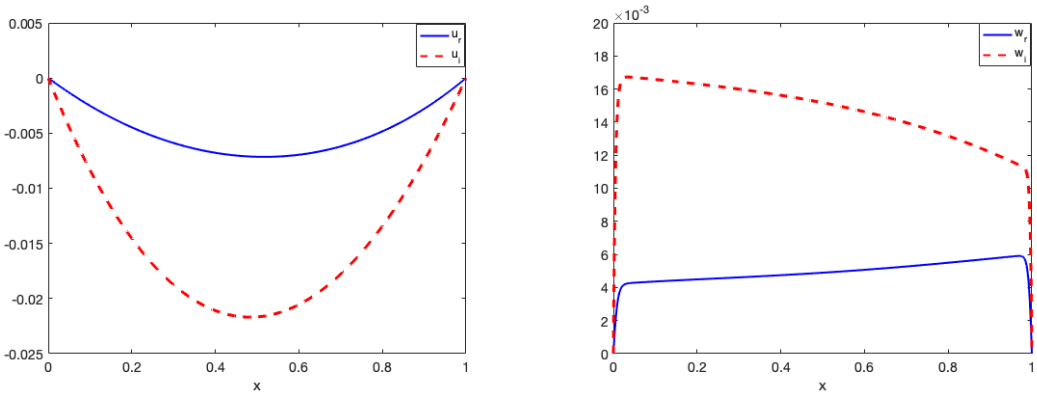

Figure 5.10: The solutions, u and w to (5.9), with data as in (5.38) and (5.40).

First note, as is easy to prove, that the sign of u_r is always the opposite of w_r, and the sign of u_i is always the opposite of w_i. In addition, if the sign of f_r (say) does not change for any x, and $f_i \equiv 0$, then sign of w_r is the same as f_r.

Remark 5.6.1. One can also investigate the signs of the components of u and w in the cases where neither f_r and f_i are zero. The deductions are somewhat more complex, since very particular assumptions on the relative magnitudes of f_r and f_i are needed to ensure that each component is of a particular sign (i.e., does not change the sign on the interval). Detailed derivation on the necessary conditions to avoid this is beyond the scope of this thesis, but would make for some interesting future work.

To establish one of these results formally, and thus establish a maximum/minimum principle for the differential operators in (5.20), we will study the case where $f_r > 0$ and $f_i = 0$. We will use all the assumptions on the coefficient functions, a and b in (5.37).

Remark 5.6.2. Recall the problem stated in [**27, Chap. 6**]

$$Lu := -\varepsilon u'' + bu = f \qquad \text{on} \qquad \Omega := (0,1), \tag{5.41a}$$

with the boundary conditions

$$u(0) = 0, \qquad u(1) = 0, \tag{5.41b}$$

where $b \geqslant \beta_0 > 0$ for all x. For follows directly from the usual Maximum Principle that $\|u\| \leqslant \|f\|/\beta_0$. In our case we have

$$L_{1;1}\hat{w}_r^{[k;j]} = f_r - b_{12}\hat{w}_i^{[k;j-1]}, \tag{5.42}$$

so

$$\|\hat{w}_r^{[k;j]}\| \leqslant \frac{1}{\min_{x\in\bar{\Omega}}(b_{11})}\|f_r - b_{12}\hat{w}_i^{[k;j-1]}\|, \tag{5.43}$$

for $k, j = 1, 2, \ldots$.

Theorem 5.6.1. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to* (5.36) *with* $f_r(x) > 0$, $f_i(x) = 0$ *for all* $x \in [0,1]$, *and* a *and* b *satisfying the assumptions in* (5.19) *and* (5.37). *Then*

$$w_r(x) \geqslant 0, \quad w_i(x) \leqslant 0, \quad u_r(x) \leqslant 0, \quad \text{and} \quad u_i(x) \geqslant 0. \tag{5.44}$$

Proof. Consider the sequence $\hat{\mathbf{w}}^{[1]}, \hat{\mathbf{w}}^{[2]}, \ldots$ of solutions to (5.36a), and $\hat{\mathbf{u}}^{[1]}, \hat{\mathbf{u}}^{[2]}, \ldots$ of solutions to (5.36b). We use induction on k, and on j for each k, to show that

$$w_r^{[k]}(x) \geqslant 0, \quad w_i^{[k]}(x) \leqslant 0, \quad u_r^{[k]}(x) \leqslant 0, \quad \text{and} \quad u_i^{[k]}(x) \geqslant 0. \tag{5.45}$$

For $k = 0$, we have $\hat{\mathbf{w}}^{[0]} \equiv 0$ and $\hat{\mathbf{u}}^{[0]} \equiv 0$. For $k = 1$ we have

$$L_{1;1}\hat{w}_r^{[1;j]} = (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[0]})_1 - b_{12}\hat{w}_i^{[1;j-1]},$$

Recall that $\hat{\mathbf{w}}^{[1;0]} = \hat{\mathbf{w}}^{[0]}$. So, when $j = 1$, we have

$$L_{1;1}\hat{w}_r^{[1;1]} = (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[0]})_1 - b_{12}\hat{w}_i^{[1;0]} = f_r > 0,$$

since $f_r > 0$, $\hat{\mathbf{u}}^{[0]} = 0$ and $\hat{w}_i^{[1;0]} = 0$. Consequently, $\hat{w}_r^{[1;1]} \geqslant 0$. Also, $\|\hat{w}_r^{[1;1]}\| \leqslant f_r/\|b_{11}\|$. Next,

$$L_{1;2}\hat{w}_i^{[1;1]} = -b_{21}\hat{w}_r^{[1;1]} < 0,$$

since $f_i = 0$, $\hat{\mathbf{u}}^{[0]} = 0$, $\hat{w}_r^{[1;1]}(x) \geqslant 0$, $b_{21} > 0$ (from the assumption in 5.19). Furthermore, this also gives that, for all x, $|\hat{w}_i^{[1;1]}(x)| < (\|b_{21}\|/\|b_{22}\|)\hat{w}_r^{[1;1]}(x)$.

Next, for $j = 2, 3, \ldots$ we have

$$L_{1;1}\hat{w}_r^{[1;j]} = f_r - b_{12}\hat{w}_i^{[1;j-1]} > 0,$$

since

$$b_{12}\hat{w}_i^{[1;j-1]} \leqslant \frac{\|b_{12}\|\|b_{21}\|}{\|b_{22}\|}\hat{w}_r^{[1;j-1]} \leqslant \frac{\|b_{12}\|\|b_{21}\|}{\|b_{11}\|\|b_{22}\|}f_r < f_r. \tag{5.46}$$

Hence, $\hat{w}_r^{[1;j]} \geqslant 0$. Also,

$$L_{1;2}\hat{w}_i^{[1;j]} = (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[0]})_2 - b_{21}\hat{w}_r^{[1;j]} = -b_{21}\hat{w}_r^{[1;j]} < 0,$$

so $\hat{w}_i^{[1;j]} \leqslant 0$. Since this holds for all j, and because $w_r^{[1,j]} \to w_r^{[1]}$, while $w_i^{[1,j]} \to w_i^{[1]}$, we conclude that

$$w_r^{[1]}(x) \geqslant 0 \text{ and } w_i^{[1]}(x) \leqslant 0 \text{ for all } x \in [0, 1].$$

For $k = 2$, we have

$$L_{1;1}\hat{w}_r^{[2;j]} = (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[1]})_1 - b_{12}\hat{w}_i^{[2;j-1]} = f_r - b_{13}u_r^{[1]} - b_{14}u_i^{[1]} - b_{12}\hat{w}_i^{[2;j-1]}, \tag{5.47}$$

and

$$L_{1;2}\hat{w}_i^{[2;j]} = (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[1]})_2 - b_{21}\hat{w}_r^{[2;j]} = -b_{23}u_r^{[1]} - b_{24}u_i^{[1]} - b_{21}\hat{w}_r^{[2;j]}. \tag{5.48}$$

Recall that

$$\hat{B}_{12} = \begin{pmatrix} b_{13} & b_{14} \\ b_{23} & b_{24} \end{pmatrix} = \begin{pmatrix} a_r\beta - \varepsilon\beta^2 - b_r & b_i - a_i\beta \\ a_i\beta - b_i & a_r\beta - \varepsilon\beta^2 - b_r \end{pmatrix},$$

When $j = 1$, we have

$$L_{1;1}w_r^{[2;1]} = f_r - \underbrace{b_{13}u_r^{[1]}}_{<0} - \underbrace{b_{14}u_i^{[1]}}_{<0} - b_{12}\hat{w}_i^{[2;0]} > f_r - b_{12}\hat{w}_i^{[1]} > 0 \tag{5.49}$$

from (5.46).

Also, we have

$$L_{1;2}\hat{w}_i^{[2;j]} = -b_{23}u_r^{[1]} \underbrace{-b_{24}u_i^{[1]}}_{<0} - b_{21}\hat{w}_r^{[2;1]} < -b_{23}u_r^{[1]} - b_{21}\hat{w}_r^{[2;1]}.$$

But

$$b_{21}\hat{w}_r^{[2;j]} = \alpha a_i\hat{w}_r^{[2;1]} \geqslant \alpha a_i\frac{\beta}{\alpha}\hat{u}_r^{[1]} = a_i\beta\hat{u}_r^{[1]} > (a_i\beta - b_i)\hat{u}_r^{[1]} = b_{23}\hat{u}_r^{[1]}.$$

So we can conclude that $L_{1;2}\hat{w}_i^{[2;j]} < 0$, and, hence, that $\hat{w}_i^{[2;j]} < 0$.

The same reasoning applies for $j = 2, 3, \ldots$, showing that, for all these j, that $\hat{w}_r^{[2;j]} > 0$ and $\hat{w}_i^{[2;j]} < 0$. Again this shows that (5.45) holds for $k = 2$. Repeating the arguments inductively gives (5.44). □

As a corollary Theorem 5.6.1, we now give pointwise bounds on w_r, w_i, u_r, u_i and their derivatives, which are useful when deriving error estimates for finite difference and finite element methods. The arguments are the same as those used to prove Lemma 4.3.5.

Lemma 5.6.2. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to (5.36), where the assumptions required by Theorem 5.6.1 hold. Then there exists a constant* C*, which is independent of* ε*, such that*

$$|{w_r}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\psi_\varepsilon(x)], \tag{5.50}$$

$$|{w_i}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\psi_\varepsilon(x)], \tag{5.51}$$

$$|{u_r}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)}\psi_\varepsilon(x)], \tag{5.52}$$

and

$$|{u_i}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)}\psi_\varepsilon(x)], \tag{5.53}$$

where $\psi_\varepsilon(x) := e^{-x/\sqrt{\varepsilon}} + e^{-(1-x)/\sqrt{\varepsilon}}$ *and* $l = 0, 1, \ldots, 4$.

Example 5.6.1. Recall the example presented in (5.5), which has $a_r = 4$, $a_i = 2$, $b_r = 6$ and $b_i = 2$. We choose $\alpha = 14$ and $\beta = b_r/a_r = 1.5$. Figure 5.11 shows plots of scaled fourth derivatives (computed using Chebfun), which suggest that

$$|{w_r}^{(4)}(x)| \leqslant C[1 + \varepsilon^{-2}\psi_\varepsilon(x)], \quad |{w_i}^{(4)}(x)| \leqslant C[1 + \varepsilon^{-2}\psi_\varepsilon(x)],$$

$$|{u_r}^{(4)}(x)| \leqslant C[1 + \varepsilon^{(-1)}\psi_\varepsilon(x)], \text{and} \quad |{u_i}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-1)}\psi_\varepsilon(x)],$$

That is, the results are shown in Figure 5.11 support the assertion that the results of Lemma 5.6.2 are sharp.

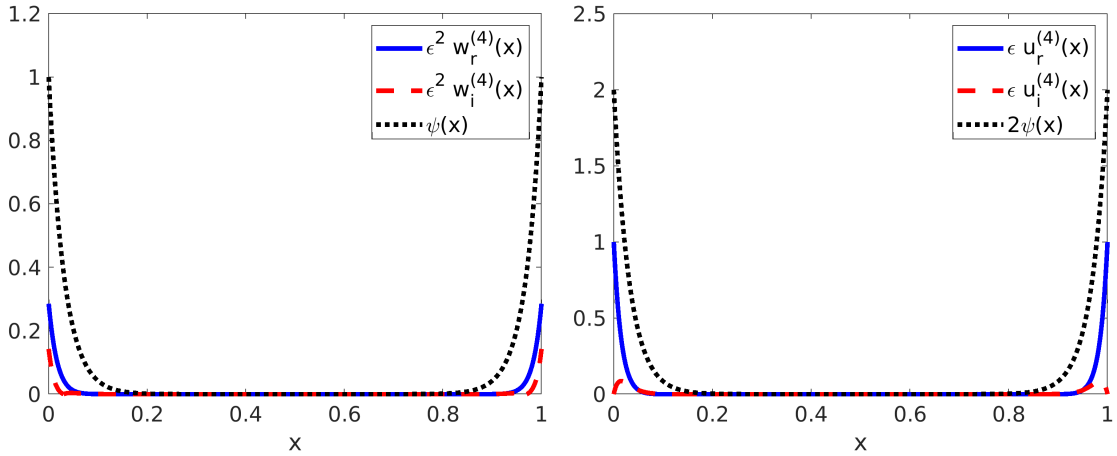

Figure 5.11: Scale fourth-order derivatives of the solution to the problem in Example 5.6.1 with $\varepsilon = 10^{-3}$, showing that $\varepsilon^2 w^{(4)}(x)$ and $\varepsilon u^{(4)}(x)$ are bounded.

Remark 5.6.3. As shown in Lemma 5.6.2, it is important to note that, if we chose $a_i = 0$ and $b_i = 0$ and $f_i = 0$ in (5.2), then we will have the case of real-valued fourth-order problem that was proved in Chapter 4 with $|u_i^{(l)}(x)| = 0$ and $|w_i^{(l)}(x)| = 0$, for $l = 0, 1, ..., 4$.

Chapter 6

A finite element analysis of a fourth-order complex-valued singularly perturbed problem

6.1 Introduction

In this chapter we are interested in the numerical solution of the singularly perturbed, fourth-order, complex-valued reaction-diffusion equation introduced in Chapter 5, by *finite element methods*, specifically focusing on transforming a special case of this problem into a coupled system of second-order reaction-diffusion problems. As in Chapter 3, a key step in the analysis involves determining if and then the coupling matrix is pointwise coercive.

In Chapter 5, the general problem, (5.2), was stated as

$$-\varepsilon u^{(4)}(x) + (a_r + ia_i)u''(x) - (b_r + ib_i)u(x) = (f_r + if_i)(x) \quad \text{on} \quad \Omega := (0,1), \tag{6.1a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{6.1b}$$

To focus on a specific sub-class of problems of the form in (6.1), we will consider the case where

$$a_r = a_i =: a, \quad \text{and} \quad b_r = b_i =: b, \tag{6.2}$$

for some real-valued functions a and b. Then our model differential equation is

$$-\varepsilon u^{(4)}(x) + a(1+i)u''(x) - b(1+i)u(x) = (f_r + if_i)(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{6.3a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{6.3b}$$

As is always the case in this thesis, the singular perturbation parameter ε lies in the interval $(0, 1]$. The right-hand side terms, f_r and f_i denote real-valued functions on the interval Ω.

As in Chapter 5 we transform the problem to a real-valued, second-order one. The simplification of (6.2), i.e., taking $a_r = a_i$ and $b_r = b_i$ is largely for exposition, since it simplifies the calculations enough so that we can provide explicit formulae for bounds on the problem's coefficients, and the parameters in the transformation, that ensure the couple matrix is coercive. This is different from what was possible Chapter 5 where we showed how to choose the parameters numerically, rather than via a formula.

6.1.1 A motivating example

We consider the following example with $a = 2$, $b = 4$, $f_r = 1$, and $f_i = 0$

$$-\varepsilon u^{(4)}(x) + 2(1+i)u''(x) - 4(1+i)u(x) = 1 \qquad \text{on} \qquad \Omega := (0, 1). \tag{6.4a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{6.4b}$$

In Figure 6.1 we show u with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right), note that the graphs are very similar, and neither feature strong layers. However, (6.4) is still singularly perturbed, in the sense that, as $\varepsilon \to 0$, it becomes ill-posed: it reduces to a second-order differential equation with four boundary conditions. Furthermore, as we shall see, derivatives of u are not bounded as $\varepsilon \to 0$, and so classical numerical methods cannot be analysed with standard techniques.

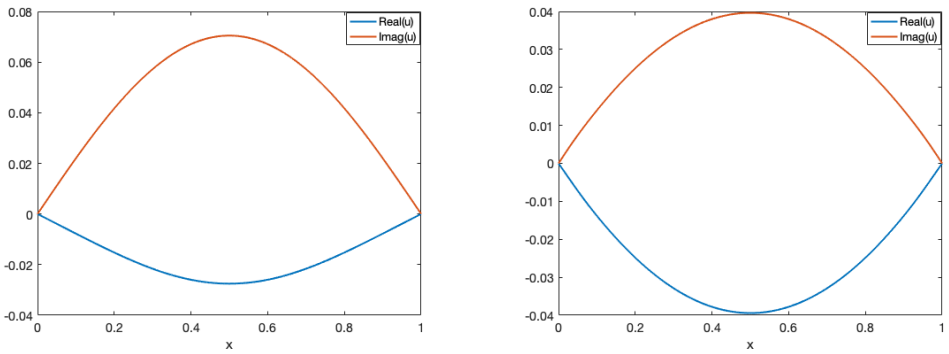

Figure 6.1: Real and imaginary parts of u to (6.4) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

6.1.2 Outline

In Section 6.2, we apply ideas based on those in Section 5.3 to rewrite (6.1), first as a system of two fourth-order real-valued problems, and then as a coupled system of four second-order ones. As in Section 3.2, we show that if this is done naively, the resulting coupling matrix can not be coercive (in the sense of Definition 3.2.1). This is done in Section 6.2.2. However, most of Section 6.2 is devoted to show how the transformation, from fourth-order to second-order, can be adapted so that the coupling matrix is coercive. This follows the methodology of Section 3.2.2, but the details are entirely different. In Section 6.3, we describe a finite element method for this problem, applied, initially, on an arbitrary mesh. We then present a suitable layer-adapted mesh, and we present the numerical analysis for this method. Finally, in Section 6.4, numerical results are presented in support of the theoretical analysis.

6.2 From a 4th-order complex-valued problem to a coupled system of real-valued 2nd-order problems

Recalling the ideas in Section 5.3, we show how to change the 4th-order complex-valued problem (6.3) first into coupled system of two fourth-order real-valued problems, and then into a coupled system of four second-order equations. We begin by writing (6.3) as

$$-\varepsilon(u_r + iu_i)^{(4)}(x) + (a + ia)(u_r + iu_i)''(x) - (b + ib)(u_r + iu_i) = (f_r + if_i), \tag{6.5}$$

where $u = u_r + iu_i$, When we equate real terms and imaginary terms separately, we get the system

$$-\varepsilon u_r^{(4)} + au_r'' - au_i'' - bu_r + bu_i = f_r, \tag{6.6a}$$
$$-\varepsilon u_i^{(4)} + au_r'' + au_i'' - bu_r - bu_i = f_i. \tag{6.6b}$$

Following the technique of Section 5.3, we will use the transformation in (5.7) to transform this system of real-valued fourth-order into real-valued, second-order system of four differential equations. The aim of this is to determine how to parametrise the transformation in order to ensure that the resulting system matrix is coercive.

6.2.1 Coercive system

We begin by assuming that a, and b satisfy the following conditions:

$$b > 0. \tag{6.7a}$$

$$a \in [3b - 2\sqrt{2b^2 - \varepsilon b}, 3b + 2\sqrt{2b^2 - \varepsilon b}]. \tag{6.7b}$$

Although this is a restriction on the range of problems that can be considered, we do not consider to to be excessively so. For example, if $b \equiv 4$ and $\varepsilon = 10^{-4}$, the analysis we will present is viable for any a with $0.6863 \leqslant a(x) \leqslant 23.3136$.

Now we use a variation on the transformation (5.7) to convert (6.6) into a system of four differential equations. As before, this transformation features a parameter that depend on the problem data.

The transformation we propose is

$$w := \frac{u'' - u}{\alpha}. \tag{6.8}$$

which is (5.7) but with $\beta = 1$. So now

$$u'' = \alpha w + u, \tag{6.9}$$

where α is a non-zero constant chosen depending on the problem data. From (6.9),

$$u^{(4)} = \alpha w'' + \alpha w + u.$$

With this, (6.6) can be transformed to a system of four equations of the form

$$-\varepsilon\alpha w_r'' + \alpha(a-\varepsilon)w_r - \alpha a w_i + (a-\varepsilon-b)u_r + (b-a)u_i = f_r, \tag{6.10a}$$
$$-\varepsilon\alpha w_i'' + \alpha a w_r + \alpha(a-\varepsilon)w_i + (a-b)u_r + (a-\varepsilon-b)u_i = f_i, \tag{6.10b}$$
$$-u_r'' + \alpha w_r + u_r = 0, \tag{6.10c}$$
$$-u_i'' + \alpha w_i + u_i = 0, \tag{6.10d}$$

with boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{6.10e}$$

We can write (6.10) in matrix form, this is

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon\alpha & 0 & 0 & 0 \\ 0 & \varepsilon\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\vec{z}'' + B\vec{z} = \vec{f}, \tag{6.11a}$$

where

$$\vec{z} = \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix}, \quad B = \begin{pmatrix} \alpha(a-\varepsilon) & -\alpha a & a-b-\varepsilon & b-a \\ \alpha a & \alpha(a-\varepsilon) & a-b & a-b-\varepsilon \\ \alpha & 0 & 1 & 0 \\ 0 & \alpha & 0 & 1 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f_r \\ f_i \\ 0 \\ 0 \end{pmatrix}. \tag{6.11b}$$

Recall from Theorems 3.2.2 and 3.2.1 that the matrix B satisfies $\vec{v}^T B\vec{v} > 0$ for all $\vec{v}$ if, and only if, $M = (B^T + B)/2$ is symmetric positive definite. Here

$$M = \begin{pmatrix} \alpha(a-\varepsilon) & 0 & (a-b-\varepsilon+\alpha)/2 & (b-a)/2 \\ 0 & \alpha(a-\varepsilon) & (a-b)/2 & (a-b-\varepsilon+\alpha)/2 \\ (a-b-\varepsilon+\alpha)/2 & (b-a)/2 & 1 & 0 \\ (b-a)/2 & (a-b-\varepsilon+\alpha)/2 & 0 & 1 \end{pmatrix}. \tag{6.12}$$

We now will show that it is possible to select α in (6.8) so that the eigenvalues of M are positive.

The eigenvalues of M are

$$\lambda_1 = \frac{1}{2}\Big[a\alpha - \varepsilon\alpha + 1 + \big(a^2\alpha^2 - 2a\alpha^2\varepsilon + \alpha^2\varepsilon^2 + 2a^2 - 4ab - 2a\varepsilon + \alpha^2 - 2\alpha b + 2b^2 + 2b\varepsilon + \varepsilon^2 + 1\big)^{(1/2)}\Big], \quad (6.13)$$

and

$$\lambda_2 = \frac{1}{2}\Big[a\alpha - \varepsilon\alpha + 1 - \big(a^2\alpha^2 - 2a\alpha^2\varepsilon + \alpha^2\varepsilon^2 + 2a^2 - 4ab - 2a\varepsilon + \alpha^2 - 2\alpha b + 2b^2 + 2b\varepsilon + \varepsilon^2 + 1\big)^{(1/2)}\Big], \quad (6.14)$$

which both have geometric multiplicity two. Since M is symmetric, λ_1 and λ_2 are real numbers. Clearly $\lambda_1 \geqslant \lambda_2$ for any a, b and α. We need to find the set of values of α for which both $\lambda_i > 0$, so we will find the range of α for which $\lambda_2 > 0$. By inspection, we can see that this is

$$a + b - \varepsilon - \sqrt{-a^2 + 6ab - b^2 - 4b\varepsilon} \leqslant \alpha \leqslant a + b - \varepsilon + \sqrt{-a^2 + 6ab - b^2 - 4b\varepsilon}. \quad (6.15)$$

For any choice of α that satisfies (6.15), M is positive definite for all a and b satisfying (6.7).

For the simple case where a and b are constants, to fix a value, we choose

$$\alpha = a + b - \varepsilon. \quad (6.16)$$

But we emphasise that any choice of α between $a + b - \varepsilon - \sqrt{-a^2 + 6ab - b^2 - 4b\varepsilon}$ and $a + b - \varepsilon + \sqrt{-a^2 + 6ab - b^2 - 4b\varepsilon}$ will suffice.

Example 6.2.1. Suppose we use the same example as presented in Section 6.1.1, where $a = 2$, $b = 4$ and $\varepsilon = 10^{-4}$. As stated in (6.15), we can choose any $\alpha \in [0.7085, 11.2912]$, but prefer $\alpha = 5.9999$, as per (6.16). Then, from (6.11b) and (6.12), we have

$$B = \begin{pmatrix} 11.9992 & -11.9998 & -2.0001 & 2 \\ 11.9998 & 11.9992 & -2 & -2.0001 \\ 5.9999 & 0 & 1 & 0 \\ 0 & 5.9999 & 0 & 1 \end{pmatrix},$$

and

$$M = \begin{pmatrix} 11.9992 & 0 & 1.9999 & 1 \\ 0 & 11.9992 & -1 & 1.9999 \\ 1.9999 & -1 & 1 & 0 \\ 1 & 1.9999 & 0 & 1 \end{pmatrix}.$$

The eigenvalues of M are $\lambda_1 = 12.436366$ and $\lambda_2 = 0.5628331$. So, M is a symmetric positive matrix, and, consequently, the matrix B is coercive.

6.2.2 A non-coercive transformation

We briefly digress to explain that it is not possible to choose $\beta = 0$ in (5.7), even though this is the most common approach in the literature.

Recall the transformation in [47], they set $\vec{z} = (u, w)^T$, where

$$w := -u'', \tag{6.17}$$

which is equivalent to (5.7) but with $\alpha = -1$ and $\beta = 0$. Note that this choice of α will not satisfy (6.15). With (6.17), the transformed (6.6) can be expressed as

$$-E\vec{z}'' + B\vec{z} = \vec{f}, \tag{6.18}$$

where

$$\vec{z} = \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix}, \; E = \begin{pmatrix} \varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}, \; B = \begin{pmatrix} a & -a & b & -b \\ a & a & b & b \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} -f_r \\ -f_i \\ 0 \\ 0 \end{pmatrix}, \tag{6.19}$$

with boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{6.20}$$

For B to be coercive, we need the symmetric part of B, denoted $M = (B + B^T)/2$ to be positive definite: equivalently, to have positive eigenvalues. It is easy to show, using either of these approaches, that M is not positive definite.

First, suppose we choose $\vec{v} = (1, 0, 0, v_4)^T$, then for any a and b, we can choose v_4, such that $\vec{v}^T B\vec{v} < 0$. For example, if $a = 1$ and $b = 2$, then $\vec{v}^T B\vec{v} = -1$ when $v_4 = 1$.

Alternatively, one can use the spectral approach. Suppose we use the same example as presented in Section 6.1.1, with $a = 2$ and $b = 4$. Then, from (6.19)

$$B = \begin{pmatrix} 2 & -2 & 4 & -4 \\ 2 & 2 & 4 & 4 \\ -1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \end{pmatrix},$$

which is not coercive, because

$$(B + B^T)/2 = \begin{pmatrix} 2 & 0 & 3/2 & -2 \\ 0 & 2 & 2 & 3/2 \\ 3/2 & 2 & 0 & 0 \\ -2 & 3/2 & 0 & 0 \end{pmatrix},$$

has the negative eigenvalue $\lambda_2 = -1.692582$. It follows then from Theorems 3.2.1, and 3.2.2 that B is not coercive.

6.2.3 Bounds on γ

Because of the theoretical development of Section 6.2.1, we now know that it is acceptable to set $\alpha = a + b - \varepsilon$ in the constant coefficient case. This allows us to deduce a bound on the coercivity parameter γ (i.e., the minimum Rayleigh quotient).

Lemma 6.2.1. *Let M be the matrix defined in (6.12). If we set $\alpha = a + b - \varepsilon$ as in (6.16), then M is positive definite, has two distinct eigenvalues, and the smaller of them λ_2, is bounded below as*

$$\lambda_2 > \frac{-b^2 + (6a - 4\varepsilon)b - a^2}{4\varepsilon^2 + (-8a - 4b)\varepsilon + 4a^2 + 4ab + 4}, \tag{6.21}$$

independently of ε.

Proof. Recall that M has two distinct eigenvalues, given in (6.13) and (6.14). First, we will show that both of eigenvalues of M are positive, and, consequently, M is symmetric positive definite. When $\alpha = a + b - \varepsilon$, we have

$$M = \begin{pmatrix} a(a+b-\varepsilon) - (a+b-\varepsilon)\varepsilon & 0 & a-\varepsilon & (b-a)/2 \\ 0 & a(a+b-\varepsilon) - (a+b-\varepsilon)\varepsilon & (a-b)/2 & a-\varepsilon \\ a-\varepsilon & (b-a)/2 & 1 & 0 \\ (b-a)/2 & a-\varepsilon & 0 & 1 \end{pmatrix}$$

The smaller eigenvalue of M is λ_2, as given in (6.14), and with $\alpha = a + b - \varepsilon$, it is

$$\begin{aligned} \lambda_2 = \frac{1}{2}\Big[&(a^2 + ab - 2a\varepsilon - b\varepsilon + \varepsilon^2 + 1) \\ &- \Big(a^4 + (2b - 4\varepsilon)a^3 + (b^2 - 6b\varepsilon + 6\varepsilon^2 + 2)a^2 + (-2b^2\varepsilon + 6b\varepsilon^2 - 4\varepsilon^3 - 2b - 4\varepsilon)a \\ &\qquad + b^2\varepsilon^2 + (-2\epsilon^3 + 2\varepsilon)b + (\varepsilon^2 + 1)^2\Big)^{(1/2)}\Big]. \end{aligned} \tag{6.22}$$

We can write λ_2 as the difference of two functions in a, b and ε:

$$\lambda_2 = Q(a, b, \varepsilon) - N(a, b, \varepsilon),$$

where

$$Q(a, b, \varepsilon) = (1/2)(a^2 + ab - \varepsilon a - \varepsilon b + \varepsilon^2 + 1),$$

and

$$\begin{aligned} N(a, b, \varepsilon) = (1/2)\big(&a^4 + (2b - 4\varepsilon)a^3 + (b^2 - 6b\varepsilon + 6\varepsilon^2 + 3)a^2 \\ &+ (-2b^2\varepsilon + 6b\varepsilon^2 - 4\varepsilon^3 - 4b - 4\varepsilon)a + (\varepsilon^2 + 1)b^2 + (-2\varepsilon^3 + 2\varepsilon)b + (\varepsilon^2 + 1)^2\big)^{(1/2)}. \end{aligned}$$

Note that $Q(a, b, \varepsilon) > 0$ for any a, b and ε satisfying (6.7), and $N(a, b, \varepsilon) > 0$, since λ_2 is real valued. Furthermore, $Q(a, b, \varepsilon)^2 - N(a, b, \varepsilon)^2 = (Q(a, b, \varepsilon) + N(a, b, \varepsilon))(Q(a, b, \varepsilon) - N(a, b, \varepsilon))$, so, if

$$Q(a, b, \varepsilon)^2 - N(a, b, \varepsilon)^2 > 0 \quad \text{and} \quad Q(a, b, \varepsilon) + N(a, b, \varepsilon) > 0,$$

then $Q(a, b, \varepsilon) - N(a, b, \varepsilon) > 0$. We already have that $Q(a, b, \varepsilon) + N(a, b, \varepsilon) > 0$, where a and b

satisfy (6.7). By squaring $Q(a,b,\varepsilon)$ and $N(a,b,\varepsilon)$ and subtracting them, we have

$$Q(a,b,\varepsilon)^2 - N(a,b,\varepsilon)^2 = -(1/4)(b^2 - 6ab + 4\varepsilon b + a^2).$$

So, $\lambda_2 > 0$, since a, b and ε satisfy (6.7a).

To get a sharper lower bound, note that

$$\lambda_2 = Q(a,b,\varepsilon) - N(a,b,\varepsilon) = \frac{Q(a,b,\varepsilon)^2 - N(a,b,\varepsilon)^2}{Q(a,b,\varepsilon) + N(a,b,\varepsilon)} = \frac{-(1/4)(b^2 - 6ab + 4\varepsilon b + a^2)}{Q(a,b,\varepsilon) + N(a,b,\varepsilon)}.$$

We know that

$$Q(a,b,\varepsilon) + N(a,b,\varepsilon) > 2Q(a,b,\varepsilon)$$

because $Q(a,b,\varepsilon) > N(a,b,\varepsilon)$. Furthermore,

$$2Q(a,b,\varepsilon) = a^2 + ab - \varepsilon a - \varepsilon b + \varepsilon^2 + 1.$$

Therefore

$$\lambda_2 > \frac{-(1/4)(b^2 - 6ab + 4\varepsilon b + a^2)}{2Q(a,b,\varepsilon)} = \frac{-b^2 + (6a - 4\varepsilon)b - a^2 + 4}{4\varepsilon^2 + (-8a - 4b)\varepsilon + 4a^2 + 4ab + 4},$$

for all a, b and ε satisfying (6.7a) and (6.7b). □

Recall Example 6.2.1, which had $a \equiv 2$, $b \equiv 4$, $\varepsilon = 10^{-4}$ and $\alpha = 5.9999$. In that case, $\lambda_2 = 0.5628331$. The same data in (6.21) gives the bound $\lambda_2 > 0.53841$, which is reasonably sharp.

6.3 The numerical method

6.3.1 Variational formulation

The variational formulation of (6.10) is: find $\vec{z} \in (H_0^1(0,1))^2$ such that

$$\mathcal{B}(\vec{z},\vec{v}) = \mathbf{F}(\vec{v}) \quad \text{for all} \quad \vec{v} \in (H_0^1(0,1))^2, \tag{6.23}$$

where

$$(q,p) = \int_0^1 q(x)p(x)dx,$$

$$\begin{aligned}\mathcal{B}(\vec{z},\vec{v}) := {} & \varepsilon\alpha(z_1',v_1') + \varepsilon\alpha(z_2',v_2') + (z_3',v_3') + (z_4',v_4') \\ & + (b_{11}z_1,v_1) + (b_{12}z_2,v_1) + (b_{13}z_3,v_1) + (b_{14}z_4,v_1) \\ & + (b_{21}z_1,v_2) + (b_{22}z_2,v_2) + (b_{23}z_3,v_2) + (b_{24}z_4,v_2) \\ & + (b_{31}z_1,v_2) + (b_{33}z_3,v_3) + (b_{42}z_2,v_2) + (b_{44}z_4,v_4),\end{aligned} \tag{6.24}$$

and

$$\mathbf{F}(\vec{v}) := (f_r, f_i, v_1, v_2),$$

where $b_{11} = \alpha(a-\varepsilon)$, $b_{12} = -\alpha a$, $b_{13} = a-b-\varepsilon$, $b_{14} = b-a$, $b_{21} = \alpha a$, $b_{22} = \alpha(a-\varepsilon)$, $b_{23} = a-b$, $b_{24} = a-b-\varepsilon$, $b_{31} = \alpha$, $b_{33} = 1$,, $b_{42} = \alpha$, and $b_{44} = 1$.

The energy norm associated with the bilinear form $\mathcal{B}(w_r, w_i, u_r, u_i)$ on $(H_0^1(0,1))^2$ is

$$\|\vec{z}\|_{\mathcal{B}}^2 = \varepsilon\alpha \left\|w_r'\right\|_2^2 + \varepsilon\alpha \left\|w_i'\right\|_2^2 + \left\|u_r'\right\|_2^2 + \left\|u_i'\right\|_2^2 + \gamma(\|w_r\|_2^2 + \|w_i\|_2^2 + \|u_r\|_2^2 + \|u_i\|_2^2). \tag{6.25}$$

Because we have shown that the coupling matrix is coercive, it is easy to show that the bilinear form is too, with respect to $\|\cdot\|_{\mathcal{B}}$, i.e.,

$$\mathcal{B}(\vec{v}, \vec{v}) \geqslant \|\vec{v}\|_{\mathcal{B}}^2 \quad \text{for all} \quad \vec{v} = (v_1, v_2) \in (H_0^1(0,1))^2. \tag{6.26}$$

The arguments are identical to those in Section 3.3, and, specifically, Lemma 3.3.1. The reasoning there can also be adapted to show that $\mathcal{B}$ is continuous, i.e.,

$$\mathcal{B}(\vec{z}, \vec{v}) \leqslant C \|\vec{z}\|_{\mathcal{B}} \|\vec{v}\|_{\mathcal{B}} \quad \text{for all} \quad \vec{z}, \vec{v} \in (H_0^1(0,1))^2. \tag{6.27}$$

It follows from the Lax-Milgram Lemma that there exists unique solution to (6.23).

6.3.2 Shishkin mesh

We construct a standard *Shishkin* mesh with the mesh parameter

$$\tau = \min\{\frac{1}{4}, \sqrt{\varepsilon/\varrho}\log(N)\}, \tag{6.28}$$

where $\varrho = \min_{\Omega}(a-\varepsilon)$. Note that this transition point does not depend of α, since it scaled the diffusion and reaction terms equally (and, thus, cancelled in the mesh parameter). We now define two mesh transition points at $x = \tau$ and $x = 1-\tau$. That is, we form a piecewise uniform mesh with $N/4$ equally-sized mesh intervals on each of $[0, \tau]$ and $[1-\tau, 1]$, and $N/2$ equally-sized mesh intervals on $[\tau, 1-\tau]$. Typically, when ε is small, $\tau \ll 1/4$, the mesh is very fine near the boundaries, and coarse in the interior. We refer to Section 2.3 for more details.

6.3.3 Finite element method

We define S to be the subspace of $(H_0^1(0,1))^2$ made up of piecewise linear functions on the mesh of Section 6.3.2. Then the discrete version of (6.23) is: find $\vec{Z} \in S$ such that

$$\mathcal{B}(\vec{Z}, \vec{V}) = \mathbf{F}(\vec{V}) \quad \text{for all} \quad \vec{V} \in S. \tag{6.29}$$

As previously noted in Section 3.3.3, the standard finite element numerical analysis can proceed based on quasi-optimal approximation properties of the finite element space, and an interpolation error estimates. That is, the error

$$\begin{aligned}\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}^2 := \left\|u_r' - U_r'\right\|_2^2 + \left\|u_i' - U_i'\right\|_2^2 + \varepsilon\alpha \left\|w_r' - W_r'\right\|_2^2 + \varepsilon\alpha \left\|w_i' - W_i'\right\|_2^2 \\ + \gamma(\|u_r - U_r\|_2^2 + \|u_i - U_i\|_2^2 + \|w_r - W_r\|_2^2 + \|w_i - W_i\|_2^2),\end{aligned} \tag{6.30}$$

is bounded as

$$\left\|\vec{z}-\vec{Z}\right\|_{\mathcal{B}} \leqslant C_1 N^{-2} + C_2 \varepsilon^{1/2} N^{-1} \ln N + C_3 N^{-1}. \tag{6.31}$$

If the constants C_2 and C_3 are of the some order, we expect the ε-robust estimate

$$\left\|\vec{z}-\vec{Z}\right\|_{\mathcal{B}} \leqslant C N^{-1}. \tag{6.32}$$

This is verified in the following section.

6.4 Numerical results

In this section, we present two examples. The first equation features constant coefficients and the second equation has variable coefficients, and in both examples, we estimate the errors in the numerical solutions based on a computed benchmark solution. Even though the coefficients are variable in the second example, we show it is possible to find a constant α which satisfies (6.15).

We denote the error for given N and ε as

$$E_{\mathcal{B}}^N := \left\|\vec{z}-\vec{Z}\right\|_{\mathcal{B}},$$

where $\vec{Z}$ is the finite element solution, and $\vec{z}$ is either the true or benchmark solution, as appropriate. In addition, $\rho_{\mathcal{B}}^N$ denotes the rates of convergence of the error in the energy norm. It is computed as

$$\rho_{\mathcal{B}}^N := \log_2\left(\frac{E_{\mathcal{B}}^N}{E_{\mathcal{B}}^{N/2}}\right). \tag{6.33}$$

By $E_\infty^N(u)$ and $E_\infty^N(w)$ we denote the true or estimated maximum pointwise error in u and w, respectively, and by $\rho_\infty^N(u)$ and $\rho_\infty^N(w)$ the corresponding rates of convergence of u and w, i.e.,

$$\rho_\infty^N(u) := \log_2\left(\frac{E_\infty^N(u)}{E_\infty^{N/2}(u)}\right) \quad \text{and} \quad \rho_\infty^N(w) := \log_2\left(\frac{E_\infty^N(w)}{E_\infty^{N/2}(w)}\right). \tag{6.34}$$

Example 6.4.1. Suppose we take $a = 2$, $b = 4$, $f_r = 1$ and $f_i = 0$ in (6.3). That is, we solve

$$-\varepsilon u^{(4)}(x) + 2(1+i)u''(x) - 4(1+i)u(x) = 1 \qquad \text{on} \qquad \Omega := (0,1), \tag{6.35}$$

with boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0.$$

The solution u with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right) was already shown in Figure 6.1. In Figure 6.2 we show w with $\varepsilon = 10^{-1}$ (left), which does not feature layers. In contrast, as shown in the graph on the right for smaller ε (in this case, $\varepsilon = 10^{-3}$), w does possess boundary layers near $x = 0$ and $x = 1$, in both the real and the imaginary parts.

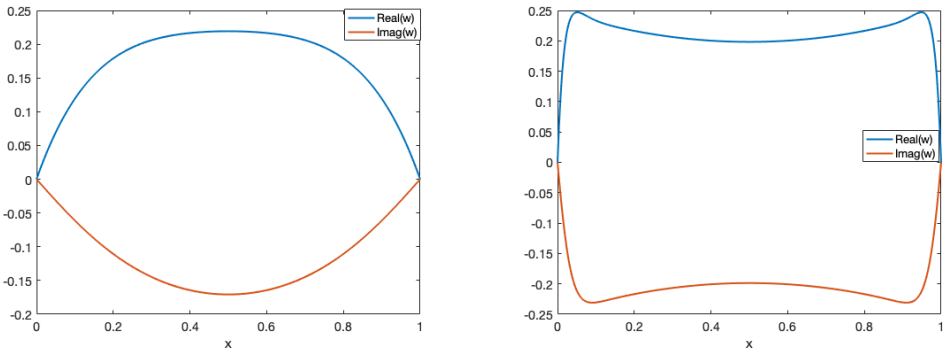

Figure 6.2: Real and imaginary parts of w to (6.4.1) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

From (6.16), we take $\alpha = 6 - \varepsilon$, and then the system we solve is

$$-\begin{pmatrix} \varepsilon(6-\varepsilon) & 0 & 0 & 0 \\ 0 & \varepsilon(6-\varepsilon) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_r'' \\ w_i'' \\ u_r'' \\ u_i'' \end{pmatrix} + B \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where

$$B = \begin{pmatrix} \varepsilon^2 - 8\varepsilon + 12 & -12 + 2\varepsilon & -2 - \varepsilon & 2 \\ 12 - 2\varepsilon & \varepsilon^2 - 8\varepsilon + 12 & -2 & -2 - \varepsilon \\ 6 - \varepsilon & 0 & 1 & 0 \\ 0 & 6 - \varepsilon & 0 & 1 \end{pmatrix}.$$

For this system, we have verified numerically that $\gamma \approx 0.5631$.

In Tables 6.1 and 6.2, we present the error in the energy norm and the associated rates of convergence computed when (6.35) is solved by using the finite element method on the Shishkin mesh. We can see that the numerical solution for this problem converges at a rate that is at an almost first-order, independently of ε. Also, the error increases as ε initially decreases, for $\varepsilon = 10^{-6}$ to $\varepsilon = 10^{-12}$, the method is clearly robust.

Table 6.1: $E_{\mathcal{B}}^N$ for problem (6.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	7.271e-03	3.636e-03	1.818e-03	9.089e-04	4.545e-04
1.0e-02	1.846e-02	9.494e-03	4.782e-03	2.395e-03	1.198e-03
1.0e-04	1.269e-02	7.198e-03	4.063e-03	2.271e-03	1.259e-03
1.0e-06	1.110e-02	5.651e-03	2.886e-03	1.477e-03	7.577e-04
1.0e-08	1.094e-02	5.482e-03	2.747e-03	1.377e-03	6.907e-04
1.0e-10	1.093e-02	5.465e-03	2.733e-03	1.367e-03	6.837e-04
1.0e-12	1.093e-02	5.463e-03	2.732e-03	1.366e-03	6.830e-04

In Table 6.3, we present $\|u_r' - U_r'\|_2 + \|u_i' - U_i'\|_2$; it clearly shows that, for sufficiently small ε, the quantities in Table 6.1 agree with those in Table 6.3, up to 3 or 4 digits, showing that $\|u_r' - U_r'\|_2 + \|u_i' - U_i'\|_2$ is the dominating term in $\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}}$. That is,

$$\left\|\vec{z} - \vec{Z}\right\|_{\mathcal{B}} \approx \|u_r' - U_r'\|_2 + \|u_i' - U_i'\|_2.$$

Table 6.2: ρ_B^N for problem (6.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.000	1.000	1.000	1.000
1.0e-02	0.959	0.989	0.997	0.999
1.0e-04	0.818	0.825	0.839	0.851
1.0e-06	0.974	0.970	0.966	0.963
1.0e-08	0.997	0.997	0.996	0.996
1.0e-10	1.000	1.000	1.000	1.000
1.0e-12	1.000	1.000	1.000	1.000

Given the close agreement between the data in Tables 6.1 and 6.3, we do not present rates of convergence for the latter; for small ε they would essentially be the same as those shown in Table 6.2.

Table 6.3: $\|u_r' - U_r'\|_2 + \|u_i' - U_i'\|_2$ for problem (6.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.290e-03	6.450e-04	3.225e-04	1.612e-04	8.062e-05
1.0e-02	4.935e-03	2.469e-03	1.234e-03	6.173e-04	3.086e-04
1.0e-04	9.543e-03	4.609e-03	2.225e-03	1.074e-03	5.179e-04
1.0e-06	1.078e-02	5.374e-03	2.678e-03	1.335e-03	6.651e-04
1.0e-08	1.091e-02	5.454e-03	2.726e-03	1.363e-03	6.811e-04
1.0e-10	1.092e-02	5.462e-03	2.731e-03	1.365e-03	6.827e-04
1.0e-12	1.092e-02	5.462e-03	2.731e-03	1.366e-03	6.829e-04

Finally, as we did for the real-valued problem in Section 3.3.4, we verify the pointwise convergence of the method. This is not covered by the theory of the method, but it is interesting, at least to compare the accuracy of the FEM and the finite difference method of Chapter 7. Tables 6.4 and 6.6 show that, for sufficiently small ε, the pointwise convergence is parameter uniform. In Table 6.5 we show that the rate of convergence is fully second-order for u. This is to be expected since, as shown in Figure 6.1, there is no (strong) boundary layer in u. That is $u''(x)$ is bounded independently of ε. In contrast, in Table 6.7 one sees only almost second-order for w, due to the presence of the layer: see Figure 6.2.

Table 6.4: $E_\infty^N(u)$ for problem (6.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	2.129e-05	5.313e-06	1.328e-06	3.319e-07	8.296e-08
1.0e-02	1.494e-05	3.564e-06	8.904e-07	2.226e-07	5.564e-08
1.0e-04	4.630e-05	1.102e-05	2.623e-06	6.235e-07	1.480e-07
1.0e-06	5.406e-05	1.342e-05	3.340e-06	8.315e-07	2.070e-07
1.0e-08	5.484e-05	1.367e-05	3.413e-06	8.527e-07	2.131e-07
1.0e-10	5.492e-05	1.369e-05	3.420e-06	8.548e-07	2.137e-07
1.0e-12	5.492e-05	1.369e-05	3.421e-06	8.550e-07	2.137e-07

Table 6.5: $\rho_\infty^N(u)$ for problem (6.35) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.002	2.001	2.000	2.000
1.0e-02	2.067	2.001	2.000	2.000
1.0e-04	2.071	2.071	2.073	2.075
1.0e-06	2.010	2.007	2.006	2.006
1.0e-08	2.005	2.002	2.001	2.001
1.0e-10	2.004	2.001	2.000	2.000
1.0e-12	2.004	2.001	2.000	2.000

Table 6.6: $E_\infty^N(w)$ for problem (6.35) solved on a Shishkin mesh.

ε	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$
1	1.406e-05	3.518e-06	8.795e-07	2.199e-07	5.497e-08
1.0e-02	1.154e-03	2.723e-04	6.706e-05	1.670e-05	4.172e-06
1.0e-04	2.841e-03	1.100e-03	3.739e-04	1.263e-04	4.095e-05
1.0e-06	2.841e-03	1.100e-03	3.738e-04	1.263e-04	4.095e-05
1.0e-08	2.841e-03	1.100e-03	3.738e-04	1.263e-04	4.095e-05
1.0e-10	2.841e-03	1.100e-03	3.738e-04	1.263e-04	4.095e-05
1.0e-12	2.841e-03	1.100e-03	3.738e-04	1.263e-04	4.095e-05

Table 6.7: $\rho_\infty^N(w)$ for problem (6.35) solved on a Shishkin mesh.

ε	$N=16$	$N=32$	$N=64$	$N=128$
1	1.999	2.000	2.000	2.000
1.0e-02	2.084	2.022	2.005	2.001
1.0e-04	1.369	1.557	1.565	1.625
1.0e-06	1.369	1.557	1.565	1.625
1.0e-08	1.369	1.557	1.565	1.625
1.0e-10	1.369	1.557	1.565	1.625
1.0e-12	1.369	1.557	1.565	1.625

Example 6.4.2. We consider another example, where a and b are variable coefficients. We will take

$$a = 2x+1, \quad b = 4x+1, \quad f_r = 1+x \quad \text{and} \quad f_i = 0,$$

in (6.3). That is, we solve

$$-\varepsilon u^{(4)}(x) + (2x+1)(1+i)u''(x) - (4x+1)(i+1)u(x) = 1+x \quad \text{on} \quad \Omega := (0,1), \tag{6.36a}$$

with boundary conditions

$$u(0) = u''(0) = u(1) = u''(1) = 0. \tag{6.36b}$$

We can not take α as in (6.16), since then it will be variable, and thus (6.9) would not apply directly. However, we can still chose an α that is constant (in x), and that satisfies (6.15). Specifically, we choose $\alpha = 2-\varepsilon$; in Figure 6.3, we show that this α satisfies (6.15).

The system we solve is

$$-\begin{pmatrix} \varepsilon(2-\varepsilon) & 0 & 0 & 0 \\ 0 & \varepsilon(2-\varepsilon) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_r'' \\ w_i'' \\ u_r'' \\ u_i'' \end{pmatrix} + B \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix} = \begin{pmatrix} 1+x \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where

$$B = \begin{pmatrix} -(2-\varepsilon)\varepsilon + (2x+1)(2-\varepsilon) & -(2x+1)(2-\varepsilon) & -\varepsilon - 2x & 2x \\ (2x+1)(2-\varepsilon) & -(2-\varepsilon)\varepsilon + (2x+1)(2-\varepsilon) & -2x & -\varepsilon - 2x \\ 2-\varepsilon & 0 & 1 & 0 \\ 0 & 2-\varepsilon & 0 & 1 \end{pmatrix}.$$

For this system, we have verified numerically that $\gamma \approx 0.382$.

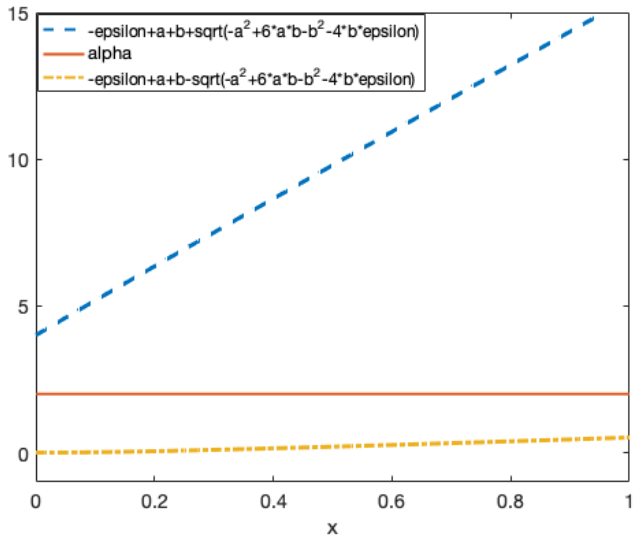

Figure 6.3: Our chosen α, and its upper and lower bounds from (6.15).

In Tables 6.8 and 6.9, we present the error in the energy norm and the associated rates of convergence computed when the coefficient functions are variable of the problem (6.36) which solved by the finite element method on the Shishkin mesh.

We can see that the numerical solution converges at a rate that is an almost first-order, independently of ε. Again, as observed with the constant-coefficient problem, the error initially increases as ε decreases, but for the smallest values of ε, the method is clearly robust.

Table 6.8: $E_{\mathcal{B}}^{N}$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	2.460e-02	1.230e-02	6.150e-03	3.075e-03	1.537e-03
1.0e-02	4.783e-02	2.472e-02	1.247e-02	6.249e-03	3.126e-03
1.0e-04	3.199e-02	2.096e-02	1.294e-02	7.618e-03	4.357e-03
1.0e-06	1.960e-02	1.069e-02	5.852e-03	3.187e-03	1.727e-03
1.0e-08	1.794e-02	9.075e-03	4.600e-03	2.335e-03	1.186e-03
1.0e-10	1.777e-02	8.898e-03	4.456e-03	2.232e-03	1.118e-03
1.0e-12	1.775e-02	8.881e-03	4.441e-03	2.221e-03	1.111e-03

Table 6.9: $\rho_{\mathcal{B}}^{N}$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.000	1.000	1.000	1.000
1.0e-02	0.952	0.987	0.997	0.999
1.0e-04	0.610	0.696	0.764	0.806
1.0e-06	0.875	0.869	0.876	0.884
1.0e-08	0.983	0.980	0.979	0.977
1.0e-10	0.998	0.998	0.998	0.997
1.0e-12	0.999	1.000	1.000	1.000

Finally, Tables 6.10 and 6.12 present the pointwise errors computed when (6.36) is solved by the finite element method on a Shishkin mesh, and these tables show for sufficiently small ε, the pointwise convergence is a parameter uniform. In Tables 6.11 and 6.13 demonstrate that this convergence is fully second-order for u, and almost second-order for w.

Table 6.10: $E_\infty^N(u)$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	3.294e-05	8.248e-06	2.061e-06	5.153e-07	1.288e-07
1.0e-02	3.330e-05	7.829e-06	1.927e-06	4.800e-07	1.199e-07
1.0e-04	5.239e-05	1.217e-05	2.822e-06	6.527e-07	1.504e-07
1.0e-06	6.573e-05	1.629e-05	4.045e-06	1.005e-06	2.499e-07
1.0e-08	6.708e-05	1.670e-05	4.170e-06	1.042e-06	2.603e-07
1.0e-10	6.721e-05	1.675e-05	4.183e-06	1.045e-06	2.614e-07
1.0e-12	6.722e-05	1.675e-05	4.184e-06	1.046e-06	2.615e-07

Table 6.11: $\rho_\infty^N(u)$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.998	2.001	2.000	2.000
1.0e-02	2.089	2.022	2.006	2.001
1.0e-04	2.106	2.108	2.112	2.117
1.0e-06	2.013	2.009	2.009	2.008
1.0e-08	2.006	2.002	2.001	2.001
1.0e-10	2.005	2.001	2.000	2.000
1.0e-12	2.005	2.001	2.000	2.000

Table 6.12: $E_\infty^N(w)$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.008e-04	2.519e-05	6.300e-06	1.575e-06	3.938e-07
1.0e-02	6.946e-03	1.583e-03	3.912e-04	9.778e-05	2.442e-05
1.0e-04	2.575e-02	1.304e-02	4.779e-03	1.521e-03	4.938e-04
1.0e-06	2.574e-02	1.306e-02	4.788e-03	1.524e-03	4.948e-04
1.0e-08	2.573e-02	1.306e-02	4.789e-03	1.525e-03	4.949e-04
1.0e-10	2.573e-02	1.306e-02	4.789e-03	1.525e-03	4.949e-04
1.0e-12	2.573e-02	1.306e-02	4.789e-03	1.525e-03	4.949e-04

Table 6.13: $\rho_\infty^N(w)$ for problem (6.36) solved on a Shishkin mesh.

ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.000	2.000	2.000	2.000
1.0e-02	2.134	2.016	2.000	2.001
1.0e-04	0.981	1.448	1.651	1.623
1.0e-06	0.979	1.448	1.651	1.623
1.0e-08	0.978	1.447	1.651	1.623
1.0e-10	0.978	1.447	1.651	1.623
1.0e-12	0.978	1.447	1.651	1.623

Chapter 7

The analysis of a finite difference method of a fourth-order complex-valued singularly perturbed problem

7.1 Introduction

In this chapter, we study the numerical solution of the singularly perturbed, fourth-order, complex-valued reaction-diffusion equation introduced in Chapter 5, by *finite difference methods*. As in Chapter 6, we transform a special case of this problem into a coupled system of second-order reaction-diffusion problems. However, whereas in Chapter 6, this was to ensure that the resulting system matrix is positive definite, and here we show how to derive a maximum/minimum-type principle principal for this system, which is typically required for the numerical analysis of finite difference methods.

Recall the general singularly perturbed, fourth-order, complex-valued reaction diffusion equations from (5.2),

$$-\varepsilon u^{(4)}(x) + (a_r + ia_i)u''(x) - (b_r + ib_i)u(x) = (f_r + if_i)(x) \qquad \text{on} \qquad \Omega := (0,1), \tag{7.1a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{7.1b}$$

We focus on a special case of problems of the form in (7.1) where

$$a := a_r, \ a_i = \zeta a, \qquad \text{and} \qquad b_r = b_i =: b,$$

where $\zeta \in (0, 1-\varepsilon)$, for some real-valued functions a and b (a slightly stronger upper bound on ζ is given in (7.8), along with other assumptions on a and b). Then our model differential equation becomes

$$-\varepsilon u^{(4)}(x) + a(1+\zeta i)u''(x) - b(1+i)u(x) = (f_r + if_i)(x) \quad \text{on} \quad \Omega := (0,1), \tag{7.2a}$$

subject to the boundary conditions

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0. \tag{7.2b}$$

The singular perturbation parameter, ε, is in the interval $(0, 1]$. Also, the right-hand side terms, f_r and f_i denote real-valued functions on the interval Ω.

7.1.1 Outline

In Section 7.2, we apply ideas similar to those in Section 5.3, to show how to first rewrite (7.2) as a system of fourth-order real-valued problems, and then as a coupled system of second-order ones. This yields a coupled system of four second-order real-valued differential equations.

In Section 7.3, we establish the stability result of differential operator for the system of four equations solved using a Gauss-Seidel method ideas based on those in Section 5.6, leading to bounds on the coefficients which ensure convergence of the Gauss-Seidel method (Corollary 7.3.1), bounds on the solutions to the iterates (Corollary 7.3.2), and a maximum/minimum principle (Corollary 7.3.3).

In Section 7.4, we describe a finite difference method for this problem, applied, initially, on an arbitrary mesh. We then present a suitable layer-adapted mesh, and we outline the numerical analysis for this method. We conclude, in Section 7.5, with two examples to verify the sharpness of the analysis outlined in Section 7.4.3.

7.2 Transformation into a system of four second-order, real-valued problems

In Section 5.3 we showed how to transform a general, fourth-order complex-valued problem into a coupled system of four real-valued second-order ones. Here, we reuse that approach but restate it since there are some simplifications in the choice of coefficients, but also because we are not interested in the coercivity of the resulting coupling matrix.

Starting as before, we rewrite (7.2) as the fourth-order real-valued system:

$$-\varepsilon u_r^{(4)} + au_r'' - \zeta a u_i'' - bu_r + bu_i = f_r, \tag{7.3a}$$
$$-\varepsilon u_i^{(4)} + \zeta a u_r'' + au_i'' - bu_r - bu_i = f_i. \tag{7.3b}$$

Using the transformation in (5.7), we set

$$w := \frac{u'' - \beta u}{\alpha} \tag{7.4}$$

where α and β are non-zero constants chosen depending on the problem data. This gives

$$u'' = \alpha w + \beta u, \qquad \text{and, thus,} \qquad u^{(4)} = \alpha w'' + \alpha\beta w + \beta^2 u. \tag{7.5}$$

With this, (7.3) is transformed to a system of four equations of the form

$$-\varepsilon\alpha w_r'' + \alpha(a - \varepsilon\beta)w_r - \alpha\zeta a w_i + (a\beta - b - \varepsilon\beta^2)u_r + (b - \zeta a\beta)u_i = f_r, \tag{7.6a}$$

$$-\varepsilon\alpha w_i'' + \alpha\zeta a w_r + \alpha(a - \varepsilon\beta)w_i + (\zeta a\beta - b)u_r + (a\beta - b - \varepsilon\beta^2)u_i = f_i, \tag{7.6b}$$

$$-u_r'' + \alpha w_r + \beta u_r = 0, \tag{7.6c}$$

$$-u_i'' + \alpha w_i + \beta u_i = 0, \tag{7.6d}$$

with boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{7.6e}$$

Writing (7.6) in matrix form gives

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon\alpha & 0 & 0 & 0 \\ 0 & \varepsilon\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}\vec{z}'' + B\vec{z} = \vec{f}, \tag{7.7a}$$

where

$$\vec{z} = \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix}, \; B = \begin{pmatrix} \alpha(a - \varepsilon\beta) & -\alpha\zeta a & a\beta - b - \varepsilon\beta^2 & b - \zeta a\beta \\ \alpha\zeta a & \alpha(a - \varepsilon\beta) & \zeta a\beta - b & a\beta - b - \varepsilon\beta^2 \\ \alpha & 0 & \beta & 0 \\ 0 & \alpha & 0 & \beta \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f_r \\ f_i \\ 0 \\ 0 \end{pmatrix}. \tag{7.7b}$$

7.3 Stability result and maximum principle

In this section, we consider how a maximum principle analysis may be applied to the system in (7.7). We will use the Gauss-Seidel analysis approach presented in Section 5.5. When doing so, we assume that a, b, and ζ satisfy the following conditions:

$$a - \varepsilon\beta > \zeta a > b > 0. \tag{7.8}$$

In addition, we assume that we can choose β such that

$$0 < \frac{\varepsilon\beta^2 + b}{\beta} < a, \tag{7.9a}$$

and

$$\frac{b}{\beta} < a\zeta. \tag{7.9b}$$

These assumptions, with $\zeta \in (0, 1-\varepsilon)$, mean that all the results of Section 5.5 and Section 5.6 now hold for this problem, including the results of Theorem 5.5.1, Theorem 5.5.2, Theorem 5.6.1, and Lemma 5.6.2.

Recall the block iterative method from Section 5.5, and we apply to this system (7.6) The semi-decoupled operators are

$$\hat{L}_1\hat{\mathbf{w}} := -\varepsilon\alpha\hat{\mathbf{w}}'' + \hat{B}_{11}\hat{\mathbf{w}}, \tag{7.10a}$$

$$\hat{L}_2\hat{\mathbf{u}} := -\hat{\mathbf{u}}'' + \hat{B}_{22}\hat{\mathbf{u}} \tag{7.10b}$$

where

$$\begin{aligned}
\hat{B}_{11} = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \end{pmatrix} = \begin{pmatrix} \alpha(a-\varepsilon\beta) & -\alpha\zeta a \\ \alpha\zeta a & \alpha(a-\varepsilon\beta) \end{pmatrix}, \\
\hat{B}_{12} = \begin{pmatrix} b_{13} & b_{14} \\ b_{23} & b_{24} \end{pmatrix} = \begin{pmatrix} a\beta - b - \varepsilon\beta^2 & b - \zeta a\beta \\ \zeta a\beta - b & a\beta - b - \varepsilon\beta^2 \end{pmatrix}, \\
\hat{B}_{21} = \begin{pmatrix} b_{31} & b_{32} \\ b_{41} & b_{42} \end{pmatrix} = \begin{pmatrix} \alpha & 0 \\ 0 & \alpha \end{pmatrix}, \quad \text{and} \quad \hat{B}_{22} = \begin{pmatrix} b_{33} & b_{34} \\ b_{43} & b_{44} \end{pmatrix} = \begin{pmatrix} \beta & 0 \\ 0 & \beta \end{pmatrix}.
\end{aligned} \tag{7.10c}$$

We'll also denote

$$\hat{\mathbf{f}}_1 = \begin{pmatrix} f_r \\ f_i \end{pmatrix}.$$

We define

$$\rho := \left|\frac{b_{12}}{b_{11}}\right| = \left|\frac{b_{21}}{b_{22}}\right| = \frac{|\zeta a|}{a - \varepsilon\beta}, \tag{7.11}$$

and

$$\theta := \min_{j=1,2} \min_{x\in\bar{\Omega}} |b_{jj}(x)| = \min_{x\in\bar{\Omega}} \big(\alpha(a-\varepsilon\beta)\big). \tag{7.12}$$

Note that, from (7.8), we have

$$0 < \rho(x) < 1 \quad \text{for all } x \in \bar{\Omega}. \tag{7.13}$$

We restate the specific version of Theorem 5.5.1 for the ODE formulated as in (7.6).

Corollary 7.3.1. *Let $\hat{\mathbf{w}} = (w_r, w_i)^T$ and $\hat{\mathbf{u}} = (u_r, u_i)^T$ be solution to (7.10). For $k = 0, 1, 2, \ldots,$ let $\hat{\mathbf{w}}^{[k]}$ and $\hat{\mathbf{u}}^{[k]}$ be defined as follows: set $\hat{\mathbf{w}}^{[0]}$ and $\hat{\mathbf{u}}^{[0]}$ to be the vector-valued zero functions, and, for $k = 1, 2, 3, \ldots,$ $\hat{\mathbf{w}}^{[k]}$ and $\hat{\mathbf{u}}^{[k]}$ to be the solutions to*

$$\hat{L}_1\hat{\mathbf{w}}^{[k]} = \hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]} \qquad \textit{subject to } \hat{\mathbf{w}}^{[k]}(0) = \hat{\mathbf{w}}^{[k]}(1) = \vec{0} \quad \textit{on } (0,1), \tag{7.14a}$$

$$\hat{L}_2\hat{\mathbf{u}}^{[k]} = -\hat{B}_{21}\hat{\mathbf{w}}^{[k]} \qquad \textit{subject to } \hat{\mathbf{u}}^{[k]}(0) = \hat{\mathbf{u}}^{[k]}(1) = \vec{0} \quad \textit{on } (0,1), \tag{7.14b}$$

Taking ρ and θ as defined in (7.11) and (7.12) if

$$0 < \frac{\alpha}{\beta}\frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} < 1, \tag{7.15}$$

then $\lim_{k\to\infty} \hat{\mathbf{w}}^{[k]} = \hat{\mathbf{w}}$, *and* $\lim_{k\to\infty} \hat{\mathbf{u}}^{[k]} = \hat{\mathbf{u}}$.

7.3.1 Using (7.15) to determine β

We now turn to investigate how to use (7.15) in order to determine β to ensure the sequence defined in Corollary 7.3.1 converges. First note that α, β, $\|\hat{B}_{12}\|$, $(1-\rho^2)$ and θ are all positive, so the lower bound in (7.15) must always hold. Also,

$$\frac{\alpha}{\beta}\frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} = \frac{|a\beta - \varepsilon\beta^2 - b| + |a\beta\zeta - b|}{\beta(1-\frac{a^2\zeta^2}{(a-\varepsilon\beta)^2})(a-\varepsilon\beta)} = \frac{(a-\varepsilon\beta)[|a\beta - \varepsilon\beta^2 - b| + |a\beta\zeta - b|]}{\beta(a-\varepsilon\beta-a\zeta)(a-\varepsilon\beta+a\zeta)}.$$

We are interested in case where ε is small; specifically, $a-\varepsilon\beta \approx a$. In this case, our goal is to chose β so that

$$\frac{a[|a\beta - b| + |a\beta\zeta - b|]}{\beta(a-a\zeta)(a+a\zeta)} < 1. \tag{7.16}$$

For any positive β, and since $(a-a\zeta)(a+a\zeta) = a^2 - a^2\zeta^2 > 0$, the inequality (7.16) is equivalent to

$$a(|a\beta - b| + |a\beta\zeta - b|) < \beta(a^2 - a^2\zeta^2).$$

If $a\beta - b > 0$ (which, we shall see presently is consistent with (7.15)), then we are attempting to ensure that

$$a^2\beta - ab + a|a\beta\zeta - b| < a^2\beta - a^2\zeta^2\beta.$$

That is,

$$-b + |a\beta\zeta - b| < -a\zeta^2\beta,$$

or, equivalently,

$$b - a\zeta^2\beta - |a\beta\zeta - b| > 0.$$

By inspection, we can see that this can be rearranged as

$$\beta(\beta - \frac{2b}{a\zeta(\zeta+1)}) > 0. \tag{7.17}$$

Since β is positive, we can see that this requires

$$\beta < \frac{2b}{a\zeta(\zeta+1)}.$$

In fact, not only can one ensure that the convergence factor in (7.15) is less that 1, one can also minimize it. From (7.16), one should choose β in the range

$$\frac{b}{a} \leqslant \beta \leqslant \frac{b}{\zeta a}, \tag{7.18}$$

for constant a and b. For variable a and b, (7.18) becomes

$$\max_{0\leqslant x\leqslant 1}\frac{b(x)}{a(x)}\leqslant\beta\leqslant\min_{0\leqslant x\leqslant 1}\frac{b(x)}{\zeta a(x)}. \tag{7.19}$$

We discuss this further in Section 7.5

Example 7.3.1. We consider particular case of (7.2) with $\varepsilon = 10^{-4}$, $a = 6$, $b = 4$, and $\zeta = 0.9$. We choose $\beta = 0.7407$ as in (7.18). If we take $\alpha = 10 - \varepsilon$, then from (7.11) we have

$$\rho = 0.900011 < 1.$$

Furthermore,

$$\frac{\alpha}{\beta}\frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} = 0.5263 < 1,$$

which demonstrates that (7.15) is satisfied. Although we don't present it here, we have verified the algorithm converges in a manner that is very similar to that shown for Example 5.5.1; see Figure 5.7.

The bounds on the solution to (7.10) are given in Theorem 5.5.2, and yield the following result.

Corollary 7.3.2. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to* (7.10). *Define*

$$\rho_1 := \frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} \quad \textit{and} \quad \rho_2 = \frac{\alpha}{\beta}.$$

Then

$$\|\hat{\mathbf{w}}\| \leqslant \left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta}, \tag{7.20}$$

and

$$\|\hat{\mathbf{u}}\| \leqslant \frac{\alpha}{\beta}\|\hat{\mathbf{w}}\| \leqslant \frac{\alpha}{\beta}\left(\frac{1}{1-\rho_1\rho_2}\right)\left(\frac{1}{1-\rho^2}\right)\frac{\|\hat{\mathbf{f}}_1\|}{\theta}. \tag{7.21}$$

Recalling (5.35), we define decoupled operators associated with the diagonal entries of B in (7.7b):

$$L_{1;1}\psi := -\varepsilon\alpha\psi'' + b_{11}\psi, \tag{7.22a}$$
$$L_{1;2}\psi := -\varepsilon\alpha\psi'' + b_{22}\psi, \tag{7.22b}$$
$$L_{2;1}\psi := -\psi'' + b_{33}\psi, \tag{7.22c}$$
$$L_{2;2}\psi := -\psi'' + b_{44}\psi. \tag{7.22d}$$

As before, one can solve (7.14a) and (7.14b) iteratively, for a fixed k, as follows. We set $\hat{w}_i^{[k;0]} = \hat{w}_i^{[k-1]}$, and solve

$$\begin{aligned} L_{1;1}\hat{w}_r^{[k;j]} &= (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]})_1 - b_{12}\hat{w}_i^{[k;j-1]} \qquad \text{subject to } \hat{w}_r^{[k;j]}(0) = \hat{w}_r^{[k;j]}(1) = \vec{0}, \\ L_{1;2}\hat{w}_i^{[k;j]} &= (\hat{\mathbf{f}}_1 - \hat{B}_{12}\hat{\mathbf{u}}^{[k-1]})_2 - b_{21}\hat{w}_r^{[k;j]} \qquad \text{subject to } \hat{w}_i^{[k;j]}(0) = \hat{w}_i^{[k;j]}(1) = \vec{0}, \end{aligned} \tag{7.23a}$$

for $j = 1, 2, 3, \ldots$, setting $\hat{w}^{[k]}$ to be the limit of the sequence $\{\hat{w}^{[k;j]}\}_{j=0}^{\infty}$. Then solve

$$\begin{aligned} L_{2;1}\hat{u}_r^{[k]} &= (-\hat{B}_{21}\hat{\mathbf{w}}^{[k]})_1 \qquad \text{subject to } \hat{u}_r^{[k]}(0) = \hat{u}_r^{[k]}(1) = \vec{0}, \\ L_{2;2}\hat{u}_2^{[k]} &= (-\hat{B}_{21}\hat{\mathbf{w}}^{[k]})_2 \qquad \text{subject to } \hat{u}_i^{[k]}(0) = \hat{u}_i^{[k]}(1) = \vec{0}, \end{aligned} \tag{7.23b}$$

The following result comes from Theorem 5.6.1.

Corollary 7.3.3. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to* (7.23) *with* $f_r(x) > 0$, $f_i(x) = 0$ *for all* $x \in [0,1]$, *and* a *and* b *satisfying the assumptions in* (7.8), (7.9a) *and* $\zeta \in (0, 1-\varepsilon)$. *Then*

$$w_r(x) \geqslant 0, \quad w_i(x) \leqslant 0, \quad u_r(x) \leqslant 0, \quad \text{and} \quad u_i(x) \geqslant 0. \tag{7.24}$$

Example 7.3.2. Suppose we use the same example as presented in (7.3.1) with $f_r = x + 1$ and $f_i = 0$. Then, from (7.6) we have

$$-\varepsilon(10-\varepsilon)w_r'' + (10-\varepsilon)(6-\beta\varepsilon)w_r - 4.4(10-\varepsilon)w_i + (0.4-0.5\varepsilon)u_r + 0.0002u_i = x+1, \tag{7.25a}$$

$$-\varepsilon(10-\varepsilon)w_i'' + 4.4(10-\varepsilon)w_r + (10-\varepsilon)(6-\beta\varepsilon)w_i - 0.0002u_r + (0.4-0.5\varepsilon)u_i = 0, \tag{7.25b}$$

$$-u_r'' + (10-\varepsilon)w_r + \beta u_r = 0, \tag{7.25c}$$

$$-u_i'' + (10-\varepsilon)w_i + \beta u_i = 0, \tag{7.25d}$$

where $\beta = 0.7407$.

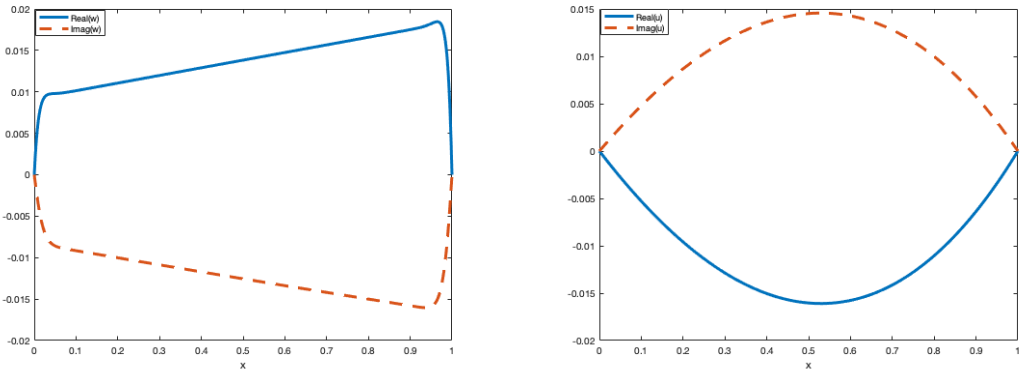

Figure 7.1: The solutions u_r, u_i, w_r and w_i to 7.3.2 with $\varepsilon = 10^{-3}$.

In Figure 7.1 we can see that the signs of components of u and w are consistent with Corollary 7.3.3. That is

$$w_r(x) \geqslant 0, \quad w_i(x) \leqslant 0, \quad u_r(x) \leqslant 0, \quad \text{and} \quad u_i(x) \geqslant 0.$$

Also, we have straight from Lemma 5.6.2 the following result.

Corollary 7.3.4. *Let* $\hat{\mathbf{w}} = (w_r, w_i)^T$ *and* $\hat{\mathbf{u}} = (u_r, u_i)^T$ *be solution to* (7.23), *where the assumptions required by Corollary 7.3.3 hold. Then there exists a constant* C, *which is independent of* ε, *such that*

$$|w_r{}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\psi_\varepsilon(x)], \tag{7.26a}$$

$$|w_i{}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(-l/2)}\psi_\varepsilon(x)], \tag{7.26b}$$

$$|u_r{}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)}\psi_\varepsilon(x)], \tag{7.26c}$$

and

$$|u_i{}^{(l)}(x)| \leqslant C[1 + \varepsilon^{(1-l/2)}\psi_\varepsilon(x)], \tag{7.26d}$$

where $\psi_\varepsilon(x) := e^{-x/\sqrt{\varepsilon}} + e^{-(1-x)/\sqrt{\varepsilon}}$ *and* $l = 0, 1, \ldots, 4$.

Example 7.3.3. Suppose we use the same example as presented in (7.3.1) with $f_r = x + 1$ and $f_i = 0$. Then, in Figure 7.2 we can see that the bound of components of $u^{(4)}$ and $w^{(4)}$ satisfied as per Corollary 7.3.4, and are, in fact quite sharp. That is that we observe

$$|w_r{}^{(4)}(x)| \approx 1.2\varepsilon^{-2}\psi_\varepsilon(x), \quad |w_i{}^{(4)}(x)| \approx 1.2\varepsilon^{-2}\psi_\varepsilon(x),$$

$$|u_r{}^{(4)}(x)| \approx 2\varepsilon^{-1}\psi_\varepsilon(x), \text{and} \quad |u_i{}^{(l)}(x)| \approx 2\varepsilon^{-1}\psi_\varepsilon(x).$$

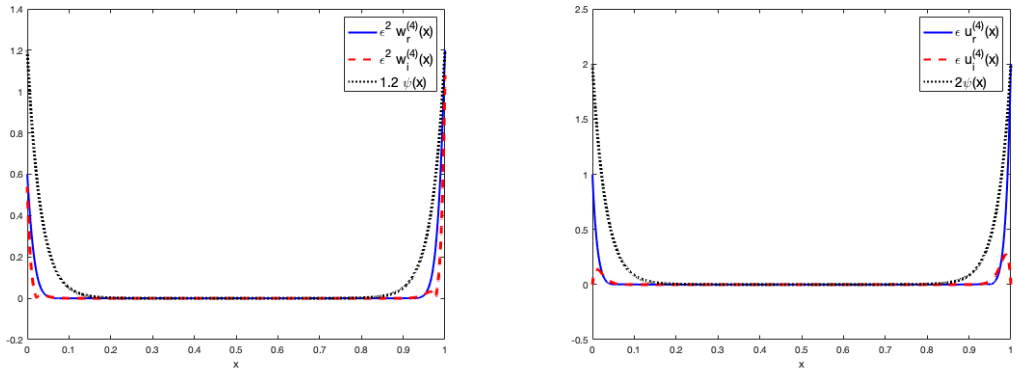

Figure 7.2: Plots of $u_r{}^{(4)}$, $u_i{}^{(4)}$, $w_r{}^{(4)}$ and $w_i{}^{(4)}$ to the problem in Example 7.3.3 with $\varepsilon = 10^{-3}$, showing that the bounds in (7.26) are quite sharp in this case.

7.4 The numerical method

7.4.1 The finite difference scheme

First, for convenience, we recall from (4.47) that the definition of the standard second-order finite difference operator on an arbitrary mesh $\{x_0 < x_1 < \cdots < x_{N-1} < x_N\}$, is

$$D^2 u_i := \frac{1}{h_i \hbar_i} u_{i-1} - \left(\frac{1}{h_i \hbar_i} + \frac{1}{h_{i+1}\hbar_i}\right) u_i + \frac{1}{\hbar_i h_{i+1}} u_{i+1},$$

where $h_i = x_i - x_{i-1}$, and $\hbar_i = (h_i + h_{i+1})/2$, and $\{u_i\}_{i=0}^N$ is any mesh function. Then the finite difference method for (7.7) is: find $\vec{Z}(x_i) = (W_r(x_i), W_i(x_i), U_r(x_i), U_i(x_i))^T$ such that

$$\vec{L}^N \vec{Z}(x_i) := -\begin{pmatrix} \varepsilon\alpha & 0 & 0 & 0 \\ 0 & \varepsilon\alpha & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} D^2 \vec{Z}(x_i) + B(x_i)\vec{Z}(x_i) = \vec{f}(x_i), \tag{7.27a}$$

for $i = 1, \ldots, N-1$, and

$$\vec{Z}(x_0) = \vec{Z}(x_N) = 0. \tag{7.27b}$$

7.4.2 Shishkin mesh

We construct a standard *Shishkin* mesh with the mesh parameter

$$\tau = \min\{\frac{1}{4}, \sqrt{\varepsilon/\varrho}\log(N)\}, \tag{7.28}$$

where $\varrho = \min_\Omega(a - \varepsilon)$. Notice that, although the perturbation parameter in (7.27) appears to be $\varepsilon\alpha$, in fact the α term cancels with that in B.

We now define two mesh transition points at $x = \tau$ and $x = 1 - \tau$. That is, we form a piecewise uniform mesh with $N/4$ equally-sized mesh intervals on each of $[0, \tau]$ and $[1 - \tau, 1]$, and $N/2$ equally-sized mesh intervals on $[\tau, 1 - \tau]$. Typically, when ε is small, $\tau \ll 1/4$, the mesh is very fine near the boundaries, and coarse in the interior. We refer to Section 2.3 for more details.

7.4.3 Numerical analysis

We have shown that the (continuous) iterative method applied to (7.7) converges, which has allowed us to deduce the stability of the continuous operator, subject to the choice of β in the transformation. That is the primary result of this chapter. But of course, we also wish to establish that we can compute a uniformly convergent solution using the finite difference method of Section 7.4.1 on the mesh described in Section 7.4.2. In some sense, such a result is standard, as there are many papers and books on this topic. However, the problem we consider is a little different from many of those in the literature, especially regarding the finite difference solution of coupled systems of second-order equations. Specifically, many papers consider systems which are "fully singularly perturbed", meaning that each equation features an arbitrarily small parameter multiplying its leading term. However, (in the terminology of Valarmathi and co-authors, e.g. [32]) our system is "partially singularly perturbed", since two of the equations do not feature a small parameter. This is different from the analysis of many papers (e.g., [23]) requires that all parameters are small, so that the solution can be decomposed into regular and layer parts; thus, their analysis does not apply directly to our scenario.

Fortunately, the arguments of Paramasivam et al. [32] can be applied, with some minor modifications.

First, we can see that in our problem, the coefficients of the terms in the first two differential

equations are singular perturbation parameters, $\varepsilon\alpha$, but the leading term in the third and fourth equations are not singular perturbation parameters (i.e., they are 1). So this fits into the notation of [32], with $n = 4$, and $m = 2$. (In [32], the presentation is also more general in that they allow for $\varepsilon_1 < \varepsilon_2$, but we have $\varepsilon_1 = \varepsilon_2$, which is still covered by the theory).

Next, in [32] paper it is assumed that the coupling matrix, which we denote as B in (7.7), has non-negative off-diagonal entries, positive diagonal entries and is strictly diagonally dominant for all x. That is, B is an M-matrix (e.g., [5]). This in turn leads to a maximum principle, for the continuous and discrete operators, which are then applied in the analysis; see [32, **Lemma 1**].

In our setting, we do not have these properties on B; indeed, as we have discussed, there is no set of problem coefficients or choices of the transformation parameters, α and β that would give this. But, nonetheless, we have shown the stability of the operator in Corollary 7.3.3, which is essentially a combined maximum/minimum principle, analogous to [32, **Lemma 2**].

The bounds on the solution and its derivatives are presented [32, **Lemma 3**]. These are for a very general problem, and so are somewhat pessimistic when translated into our simpler setting. But, importantly, the bounds presented in Corollary 7.3.4 imply those in [32, **Lemma 3**]. However, in Section 3 of [32] "improved estimates" are presented for a decomposed solution which agree exactly with those (for the undecomposed problems) in Corollary 7.3.4; see [32, **Lemma 6**], where the scenario we consider is a simplified version of Case 2 in the proof (which allows for more than one $\mathcal{O}(1)$ term in the second order coefficients).

Next, we note that the Shishkin mesh and finite difference schemes presented in Sections 6 and 7, respectively, of [32], are exactly the same as we use in Sections 7.4.2 and 7.4.1. The stability results for the discrete operator follow by standard arguments, but the proof of the robust convergence of the numerical method, culminating in [32, **Thm. 3**], follows detailed numerical analysis. However, the complications are largely due to the presence of multiple singular perturbation parameters. In our simpler setting, we can deduce from [32, **Thm. 3**] the following result.

Theorem 7.4.1. *Let Ω^N be the Shishkin mesh defined in Section 7.4.2, and let $\hat{\mathbf{w}}$ and $\hat{\mathbf{u}}$ be the solution to (7.23) on this mesh. If $\hat{\mathbf{W}}$ and $\hat{\mathbf{U}}$ solves (7.27), then*

$$\|\vec{z} - \vec{Z}\|_{\bar{\Omega}^N} := \|\hat{\mathbf{u}} - \hat{\mathbf{U}}\|_{\bar{\Omega}^N} + \|\hat{\mathbf{w}} - \hat{\mathbf{W}}\|_{\bar{\Omega}^N} \leqslant CN^{-2}\ln^2 N, \tag{7.29}$$

for some constant C.

Proof. This follows from inspecting the arguments leading to [32, **Thm. 3**]. □

Note that (7.29) implies both

$$\|\hat{\mathbf{w}} - \hat{\mathbf{W}}\|_{\bar{\Omega}^N} \leqslant C_1 N^{-2}\ln^2 N, \tag{7.30}$$

and

$$\|\hat{\mathbf{u}} - \hat{\mathbf{U}}\|_{\bar{\Omega}^N} \leqslant C_2 N^{-2}\ln^2 N, \tag{7.31}$$

for some constants C_1 and C_2. As we shall see in Section 7.5, the error bound in (7.30) appears

sharp, but, in fact the bound in (7.31) appears not to be, and, in practice on observes that

$$\|\hat{\mathbf{u}} - \hat{\mathbf{U}}\|_{\bar{\Omega}^N} \leqslant C_2 N^{-2}, \tag{7.32}$$

7.5 Numerical results

In this section, we will present numerical results for two examples. The first example has constant coefficient functions, a and b are constants, which simplifies the transformation, since the (constant) parameter β can be expressed as an expression in a and b. For the second example, the coefficient functions, a and b, and right-hand side function, f, are variable. Even though the coefficients are variable in the second example, we show it is possible to find a constant β which satisfies (7.19).

In both examples, we estimate the errors in the numerical solutions based on a computed benchmark solution: the errors in the tables below are computed by comparing a numerical solution on N intervals with one computed on a mesh with $100N$ intervals, but using the same transition point (so interpolation is not required). Then, for given N and ε, we denote by $E_\varepsilon^N(u)$ and $E_\varepsilon^N(w)$ the estimated maximum pointwise error in u and w, respectively. In addition, $\rho_\varepsilon^N(u)$ and $\rho_\varepsilon^N(w)$ represent the estimated rates of convergence for u and w, computed in the same way as in earlier chapters (see, e.g., Section 4.5).

7.5.1 A constant coefficient example

Example 7.5.1. In our first example, we use the data from Example 7.3.2; see (7.3.1). That is, we will solve

$$-\varepsilon u^{(4)}(x) + 6(1+0.9i)u''(x) - 4(1+i)u(x) = x + 1. \tag{7.33}$$

For the simple case we choose $\alpha = 10 - \varepsilon$ as in (3.16), and $\beta = 0.7407$ as in (7.18). The transformed system can be written in matrix form (7.7) with

$$-\begin{pmatrix} \varepsilon(10-\varepsilon) & 0 & 0 & 0 \\ 0 & \varepsilon(10-\varepsilon) & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_r'' \\ w_i'' \\ u_r'' \\ u_i'' \end{pmatrix} + B \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{7.34}$$

where

$$B = \begin{pmatrix} (10-\varepsilon)(6-0.7407\varepsilon) & -4.4(10-\varepsilon) & 0.4-0.5\varepsilon & 0.0002 \\ 4.4(10-\varepsilon) & (10-\varepsilon)(6-0.7407\varepsilon) & -0.0002 & 0.4-0.5\varepsilon \\ 10-\varepsilon & 0 & 0.7407 & 0 \\ 0 & 10-\varepsilon & 0 & 0.7407 \end{pmatrix}.$$

In Figure 7.3 we show the solutions u_r, u_i, w_r and w_i to (7.34) with $\varepsilon = 10^{-1}$; note that layers are not obvious in either component. This contrasts with Figure 7.1 where, for smaller ε (in this case, $\varepsilon = 10^{-3}$), w does possesses boundary layers near $x = 0$ and $x = 1$, in both the real and the imaginary parts.

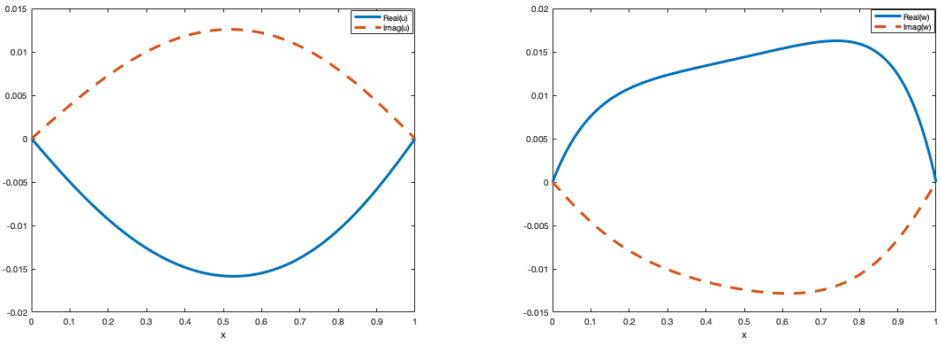

Figure 7.3: The solutions u_r, u_i, w_r and w_i to Example 7.5.1 with $\varepsilon = 10^{-1}$; compare with Figure 7.1

We now present numerical results for Example 7.5.1, where the solution is computed on the Shishkin mesh of Section 7.4.3. Tables 7.1 and Tables 7.2, we present the pointwise errors for the solutions z present it in (7.29), and the rate of convergence, respectively. We can see that the numerical solution for this problem converges at a rate that is at an almost second-order, independently of ε. Also, the error increases as ε initially decreases, for reported values of ε less than 10^{-3} the method is clearly robust.

Table 7.1: $E_\varepsilon^N(z)$ for problem (7.5.1) computed on a Shishkin mesh

Real part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	3.754e-05	9.404e-06	2.353e-06	5.883e-07	1.471e-07
1.0e-02	8.268e-04	3.014e-04	7.869e-05	2.021e-05	5.067e-06
1.0e-04	8.294e-04	3.846e-04	1.438e-04	5.006e-05	1.640e-05
1.0e-06	8.331e-04	3.857e-04	1.441e-04	5.014e-05	1.642e-05
1.0e-08	8.335e-04	3.858e-04	1.442e-04	5.015e-05	1.642e-05
1.0e-10	8.336e-04	3.858e-04	1.442e-04	5.015e-05	1.642e-05
1.0e-12	8.336e-04	3.858e-04	1.442e-04	5.015e-05	1.642e-05

Table 7.2: $\rho_\varepsilon^N(z)$ for problem (7.5.1) computed on a Shishkin mesh

Real part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.997	1.999	2.000	2.000
1.0e-02	1.456	1.938	1.961	1.996
1.0e-04	1.109	1.419	1.523	1.610
1.0e-06	1.111	1.420	1.524	1.611
1.0e-08	1.111	1.420	1.524	1.611
1.0e-10	1.111	1.420	1.524	1.611
1.0e-12	1.111	1.420	1.524	1.611

Tables 7.3 and Tables 7.4, we present the pointwise errors for the solutions u, and the rate of convergence, respectively. For the former, we see that the error in both the real and imaginary components of u is, essentially, robust for all values of ε, with minor fluctuations in the error for larger ε. From the latter, we see that the rate of convergence is fully second-order. Note that this suggests that the result in Theorem 7.4.1, which predicts the usual logarithmic factor in the rate of convergence, is not entirely sharp. However, since (as we shall see) the error is dominated by the w component, this is not consequential.

Table 7.3: $E_\varepsilon^N(u)$ for problem (7.5.1) computed on a Shishkin mesh.

Real part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.377e-05	3.462e-06	8.663e-07	2.167e-07	5.425e-08
1.0e-02	5.362e-06	2.635e-06	7.083e-07	1.804e-07	4.562e-08
1.0e-04	1.081e-05	2.584e-06	6.168e-07	1.478e-07	4.037e-08
1.0e-06	1.382e-05	3.435e-06	8.534e-07	2.120e-07	5.259e-08
1.0e-08	1.416e-05	3.543e-06	8.884e-07	2.220e-07	5.543e-08
1.0e-10	1.420e-05	3.554e-06	8.920e-07	2.230e-07	5.573e-08
1.0e-12	1.420e-05	3.555e-06	8.924e-07	2.231e-07	5.579e-08
Imaginary part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.290e-05	3.233e-06	8.085e-07	2.022e-07	5.061e-08
1.0e-02	8.828e-06	2.595e-06	6.838e-07	1.724e-07	4.326e-08
1.0e-04	8.972e-06	2.150e-06	5.153e-07	1.298e-07	3.556e-08
1.0e-06	1.135e-05	2.821e-06	7.005e-07	1.740e-07	4.317e-08
1.0e-08	1.163e-05	2.909e-06	7.293e-07	1.822e-07	4.550e-08
1.0e-10	1.166e-05	2.918e-06	7.323e-07	1.831e-07	4.575e-08
1.0e-12	1.166e-05	2.918e-06	7.326e-07	1.831e-07	4.580e-08

Table 7.4: $\rho_\varepsilon^N(u)$ for problem (7.5.1) computed on a Shishkin mesh.

Real part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.992	1.999	1.999	1.998
1.0e-02	1.025	1.895	1.973	1.984
1.0e-04	2.065	2.067	2.061	1.873
1.0e-06	2.008	2.009	2.009	2.011
1.0e-08	1.999	1.995	2.001	2.001
1.0e-10	1.999	1.994	2.000	2.000
1.0e-12	1.998	1.994	2.000	1.999
Imaginary part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.996	2.000	1.999	1.998
1.0e-02	1.767	1.924	1.988	1.995
1.0e-04	2.061	2.061	1.989	1.868
1.0e-06	2.008	2.010	2.009	2.011
1.0e-08	2.000	1.996	2.001	2.002
1.0e-10	1.999	1.994	2.000	2.000
1.0e-12	1.999	1.994	2.000	1.999

In Table 7.5 we present the errors for the w component, along with the estimate rates of convergence in Table 7.6. The errors in Table 7.5 follow those typically observed for a second-order reaction-diffusion equations: the error initially increases as ε decreases, but for the smallest values of ε, the method is robust. Further, we see almost second-order for w, independently of ε, which is entirely in agreement with Theorem 7.4.1

Table 7.5: $E_\varepsilon^N(w)$ for problem (7.5.1) computed on a Shishkin mesh.

Real part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.885e-05	4.720e-06	1.181e-06	2.951e-07	7.373e-08
1.0e-02	7.815e-04	2.979e-04	7.771e-05	1.997e-05	5.005e-06
1.0e-04	7.804e-04	3.809e-04	1.425e-04	4.986e-05	1.635e-05
1.0e-06	7.804e-04	3.808e-04	1.425e-04	4.985e-05	1.635e-05
1.0e-08	7.804e-04	3.809e-04	1.425e-04	4.985e-05	1.635e-05
1.0e-10	7.804e-04	3.809e-04	1.425e-04	4.985e-05	1.635e-05
1.0e-12	7.804e-04	3.809e-04	1.425e-04	4.985e-05	1.635e-05
Imaginary part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	3.124e-06	7.873e-07	1.972e-07	4.935e-08	1.237e-08
1.0e-02	2.742e-04	9.834e-05	2.494e-05	6.254e-06	1.565e-06
1.0e-04	2.747e-04	1.192e-04	4.534e-05	1.555e-05	5.122e-06
1.0e-06	2.744e-04	1.192e-04	4.534e-05	1.555e-05	5.122e-06
1.0e-08	2.728e-04	1.191e-04	4.534e-05	1.555e-05	5.122e-06
1.0e-10	2.725e-04	1.191e-04	4.534e-05	1.555e-05	5.122e-06
1.0e-12	2.725e-04	1.191e-04	4.534e-05	1.555e-05	5.122e-06

Table 7.6: $\rho_\varepsilon^N(w)$ for problem (7.5.1) computed on a Shishkin mesh.

Real part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.998	1.999	2.000	2.001
1.0e-02	1.391	1.939	1.961	1.996
1.0e-04	1.035	1.418	1.516	1.608
1.0e-06	1.035	1.418	1.516	1.608
1.0e-08	1.035	1.418	1.516	1.608
1.0e-10	1.035	1.418	1.516	1.608
1.0e-12	1.035	1.418	1.516	1.608
Imaginary part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.988	1.997	1.999	1.997
1.0e-02	1.479	1.979	1.996	1.999
1.0e-04	1.205	1.394	1.544	1.602
1.0e-06	1.203	1.395	1.544	1.602
1.0e-08	1.195	1.394	1.544	1.602
1.0e-10	1.194	1.394	1.544	1.602
1.0e-12	1.194	1.394	1.544	1.602

7.5.2 A variable coefficient example

Example 7.5.2. In this example, we take $a = \sqrt{x+2}$, $b = e^x/2$, $\zeta = 0.4$, and $f_r = x^2 + 2$ and $f_i = 0$ in (7.2). Then the problem we are solving is

$$-\varepsilon u^{(4)}(x) + \sqrt{x+2}(1+0.4i)u''(x) - \frac{e^x}{2}(1+i)u(x) = x^2 + 2 \qquad on \qquad (0,1), \tag{7.35}$$

with, as usual, homogeneous Dirichlet boundary conditions.

As per (7.5), α and β must be chosen to be constant when transforming to a system of four second-order equations. Recalling (7.19), we must choose β so that

$$\max_{0\leqslant x\leqslant 1} \frac{b(x)}{a(x)} \leqslant \beta \leqslant \min_{0\leqslant x\leqslant 1} \frac{b(x)}{\zeta a(x)}.$$

That is, for this data of this problem,

$$\frac{b(1)}{a(1)} \approx 0.7847 \leqslant \beta \leqslant \frac{b(0)}{\zeta a(0)} \approx 0.8839.$$

Therefore, we choose $\beta = 0.8$ for the numerical results in this example; see Figure 7.4;

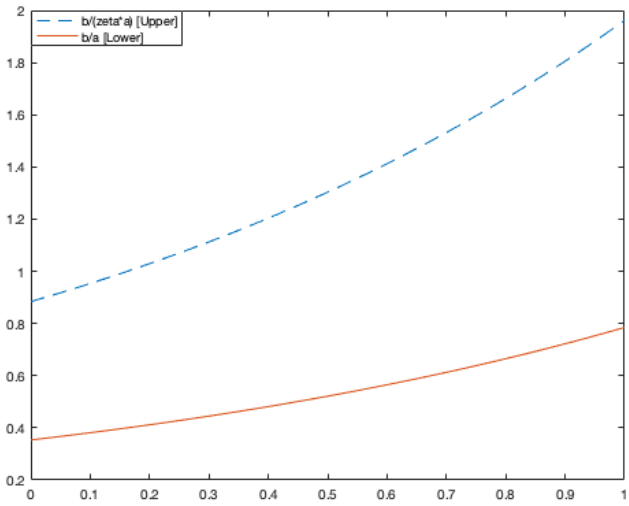

Figure 7.4: The upper and lower bounds for β from (7.19) when $\varepsilon = 10^{-4}$. Notice we can choose, e.g., $\beta = 0.8$.

We are free to choose α, since it does no impact on the convergence of the scheme; arbitrarily, we take $\alpha = 1$. From (7.11) we have

$$\max_{0 \leqslant x \leqslant 1} \rho(x) = 0.400018 < 1,$$

so (7.13) is satisfied. Furthermore, from (7.15) when $\varepsilon = 10^{-4}$, we have and

$$\frac{\alpha}{\beta} \frac{\|\hat{B}_{12}\|}{(1-\rho^2)\theta} \leqslant 0.7143 < 1,$$

for all $x \in [0, 1]$, so convergence is assured via Corollary 7.3.1.

The system we solve is

$$-\begin{pmatrix} \varepsilon & 0 & 0 & 0 \\ 0 & \varepsilon & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_r'' \\ w_i'' \\ u_r'' \\ u_i'' \end{pmatrix} + B \begin{pmatrix} w_r \\ w_i \\ u_r \\ u_i \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{7.36}$$

where

$$B = \begin{pmatrix} \sqrt{x+2} - \varepsilon & -0.4\sqrt{x+2} & \sqrt{x+2} - e^x/2 - \varepsilon & e^x/2 - 0.4\sqrt{x+2} \\ 0.4\sqrt{x+2} & \sqrt{x+2} - \varepsilon & 0.4\sqrt{x+2} - e^x/2 & \sqrt{x+2} - e^x/2 - \varepsilon \\ 1 & 0 & 0.8 & 0 \\ 0 & 1 & 0 & 0.8 \end{pmatrix}.$$

In Figure 7.5 we show u with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right), note that the graphs are very similar, and neither feature layers. In Figure 7.6 we show w with $\varepsilon = 10^{-1}$ (left), which does not features layers. In contrast, as shown in the graph on the right for smaller ε (in this case, $\varepsilon = 10^{-3}$), w does possesses boundary layers near $x = 0$ and $x = 1$, in both the real and the imaginary parts.

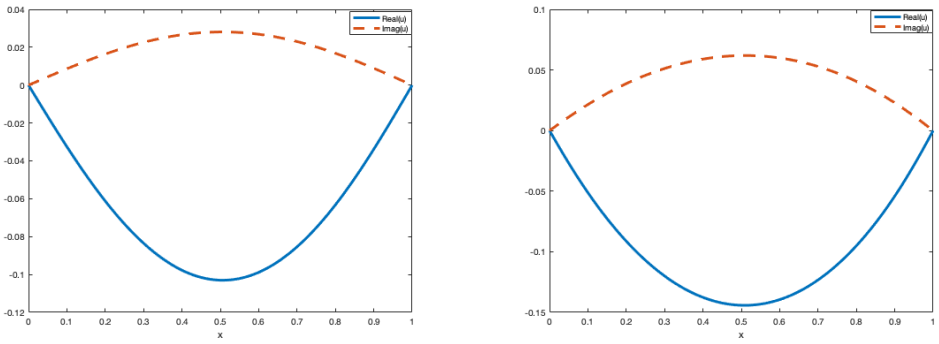

Figure 7.5: Real and imaginary parts of the solutions u to (7.36) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right)

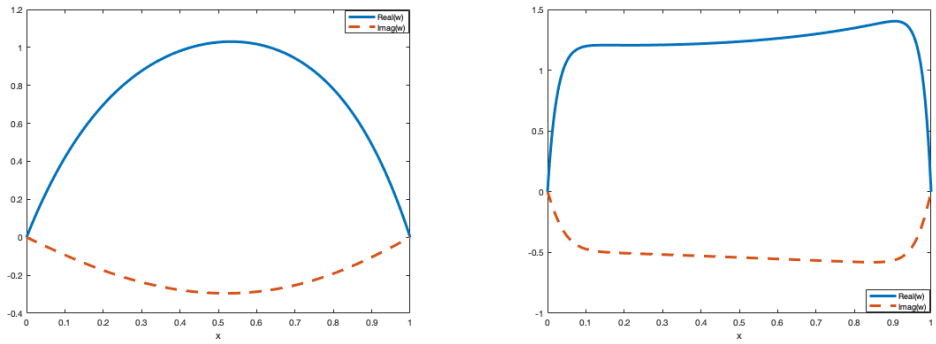

Figure 7.6: Real and imaginary parts of w to (7.36) with $\varepsilon = 10^{-1}$ (left) and $\varepsilon = 10^{-3}$ (right).

Tables 7.7 and 7.9 present the pointwise errors for the solutions u and w computed when (7.5.2) is solved by the finite difference scheme on Shishkin mesh. The results are qualitatively similar to Tables 7.3 and 7.5: the error initially increases as ε decreases, but for the smallest values of ε, the method is clearly robust.

Tables 7.8 and 7.10, we can see that the numerical solution converges at a rate that is a full second-order for u, and an almost second-order for w, independently of ε.

Table 7.7: $E_\varepsilon^N(u)$ for problem (7.5.2) computed on a Shishkin mesh.

Real part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	5.779e-05	1.450e-05	3.625e-06	9.065e-07	2.266e-07
1.0e-02	7.110e-05	1.782e-05	4.458e-06	1.115e-06	2.787e-07
1.0e-04	1.596e-04	3.420e-05	7.271e-06	1.541e-06	3.551e-07
1.0e-06	2.674e-04	6.622e-05	1.640e-05	4.044e-06	9.973e-07
1.0e-08	2.805e-04	7.043e-05	1.763e-05	4.403e-06	1.100e-06
1.0e-10	2.818e-04	7.086e-05	1.776e-05	4.441e-06	1.110e-06
1.0e-12	2.819e-04	7.090e-05	1.777e-05	4.444e-06	1.111e-06
Imaginary part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	8.618e-06	2.152e-06	5.384e-07	1.346e-07	3.365e-08
1.0e-02	3.086e-05	8.163e-06	2.055e-06	5.160e-07	1.290e-07
1.0e-04	4.010e-05	8.810e-06	1.896e-06	4.070e-07	9.481e-08
1.0e-06	6.669e-05	1.645e-05	4.069e-06	1.004e-06	2.478e-07
1.0e-08	6.987e-05	1.744e-05	4.359e-06	1.089e-06	2.719e-07
1.0e-10	7.019e-05	1.754e-05	4.389e-06	1.098e-06	2.744e-07
1.0e-12	7.022e-05	1.755e-05	4.392e-06	1.099e-06	2.747e-07

Table 7.8: $\rho_\varepsilon^N(u)$ for problem (7.36) computed on a Shishkin mesh.

Real part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.995	2.000	2.000	2.000
1.0e-02	1.996	1.999	2.000	2.000
1.0e-04	2.223	2.234	2.239	2.117
1.0e-06	2.014	2.014	2.019	2.020
1.0e-08	1.994	1.998	2.002	2.002
1.0e-10	1.992	1.996	2.000	2.000
1.0e-12	1.991	1.996	2.000	2.000
Imaginary part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	2.002	1.999	2.000	2.000
1.0e-02	1.919	1.990	1.994	2.000
1.0e-04	2.186	2.216	2.220	2.102
1.0e-06	2.019	2.016	2.019	2.019
1.0e-08	2.002	2.000	2.001	2.002
1.0e-10	2.000	1.999	1.999	2.000
1.0e-12	2.000	1.999	1.999	2.000

Table 7.9: $E_\varepsilon^N(w)$ for problem (7.36) computed on a Shishkin mesh.

Real part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.441e-04	3.613e-05	9.034e-06	2.259e-06	5.647e-07
1.0e-02	1.701e-02	4.458e-03	1.145e-03	2.871e-04	7.189e-05
1.0e-04	4.111e-02	1.854e-02	7.014e-03	2.423e-03	7.968e-04
1.0e-06	4.104e-02	1.851e-02	7.004e-03	2.420e-03	7.956e-04
1.0e-08	4.102e-02	1.851e-02	7.003e-03	2.419e-03	7.955e-04
1.0e-10	4.102e-02	1.851e-02	7.003e-03	2.419e-03	7.955e-04
1.0e-12	4.102e-02	1.851e-02	7.003e-03	2.419e-03	7.955e-04
Imaginary part					
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	4.127e-05	1.033e-05	2.583e-06	6.459e-07	1.615e-07
1.0e-02	2.948e-03	7.454e-04	1.868e-04	4.680e-05	1.170e-05
1.0e-04	7.199e-03	3.129e-03	1.146e-03	3.916e-04	1.281e-04
1.0e-06	7.269e-03	3.126e-03	1.144e-03	3.909e-04	1.279e-04
1.0e-08	7.298e-03	3.127e-03	1.144e-03	3.908e-04	1.279e-04
1.0e-10	7.302e-03	3.127e-03	1.144e-03	3.908e-04	1.279e-04
1.0e-12	7.302e-03	3.127e-03	1.144e-03	3.908e-04	1.279e-04

Table 7.10: $\rho_\varepsilon^N(w)$ for problem (7.36) computed on a Shishkin mesh.

Real part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
ε	$N = 32$	$N = 64$	$N = 128$	$N = 256$
1	1.996	2.000	2.000	2.000
1.0e-02	1.932	1.961	1.996	1.997
1.0e-04	1.149	1.402	1.533	1.605
1.0e-06	1.148	1.402	1.533	1.605
1.0e-08	1.148	1.402	1.533	1.605
1.0e-10	1.148	1.402	1.533	1.605
1.0e-12	1.148	1.402	1.533	1.605

Imaginary part				
ε	$N = 16$	$N = 32$	$N = 64$	$N = 128$
1	1.998	2.000	2.000	2.000
1.0e-02	1.984	1.996	1.997	2.000
1.0e-04	1.202	1.450	1.549	1.611
1.0e-06	1.217	1.450	1.549	1.612
1.0e-08	1.223	1.451	1.549	1.612
1.0e-10	1.223	1.451	1.549	1.612
1.0e-12	1.223	1.451	1.549	1.612

Chapter 8

Conclusion

8.1 Summary of thesis

The aim of this thesis has been to design algorithms for the accurate and efficient analysis and solution of second-and fourth-order boundary-layer problems, with a particular emphasis on complex-valued problems. Although complex-valued problems are very important in applications, they have not received much attention in the literature concerning the parameter robust solution singularly perturbed problems. Thus there are numerous fundamental questions to be addressed, and opportunities to contribute to the emerging science in this area. This has been achieved through the following contributions.

- We have introduced the numerical analysis of complex-valued singularly perturbed differential equations, beginning with the convergence of a finite difference scheme for a second-order equation on a layer-adapted mesh (Chapter 2). This is important to the rest of the thesis because it shows, in great detail, how to analyse a finite difference method for a coupled system of second-order equations.
- We have proposed a new transformation for a real-valued fourth-order problem to a coupled system which gives a simple resolution to a short-coming in some published literature. There is a parameter in the transformation that depends on the problem data, but we have shown how it can be chosen when the problem at hand is to be solved by a finite element method (Chapter 3), or a finite difference method (Chapter 4).
- Next we have approached the problem of analysing complex-valued fourth-order problem; in a general setting we have shown how this can be transformed into a system of four real-valued second-order equations. We then discussed a framework for ensuring the coupling matrix is coercive, or the operator satisfies a monotonicity result, when given very specific values of the problem data. (Chapter 5).
- We have investigated the applicability of that framework in the context of a special case of problem in Chapter 5:

$$-\varepsilon u^{(4)}(x) + (1+i)a(x)u''(x) - (1+i)b(x)u(x) = f(x) \quad \text{on} \quad \Omega := (0,1),$$

$$u(0) = u''(0) = 0 = u(1) = u''(1) = 0,$$

(Chapter 6). We have shown how formulate the problem to ensure coercivity of coupling matrix, without having to consider specific values of the problem data (other than certain assumptions). The exact choice of the special case is mainly for exposition; the approach we present can be applied to other, or more general, cases.

- Finally, we study another special case of the problem from Chapter 5, this time of form

$$-\varepsilon u^{(4)}(x) + a(1+\zeta i)u''(x) - b(1+i)u(x) = (f_r + if_i)(x) \quad \text{on} \quad \Omega := (0,1),$$

$$u(0) = u''(0) = 0, \qquad u(1) = u''(1) = 0,$$

but from the perspective of a maximum/minimum principle-type result, which is often required for the numerical analysis of finite difference methods (Chapter 7).

8.2 Future work

There are many possible extensions of the work of this thesis that can be addressed in the future. Of course, we do not aim to discuss all possibilities of further work related to this thesis. Instead, I mention some directions which I think would be particularly interesting. Of course, this is my own view and is based on my own mathematical interests.

1. In this thesis we have considered so-called "simply supported" fourth-order problems, where the boundary conditions allow for relatively direct reframing of the problem as coupled systems. The so-called "clamped" problems are also of huge interest, but are more difficult to solve in several ways. Firstly, it is not as clear to express the problems as lower-order systems. Secondly, the layers present are more strongly dependent on ε. That is, for the simply supported problems, the solution and its first two derivatives are bounded independently of ε. But for the clamped case, only the solution and its first derivative are bounded. This will complicate both the numerical methods and their analysis.

2. The complex-valued problems we have studied feature the same boundary conditions on the real and imaginary parts, and so, the layers present in the two components are very similar, In particular, a simple layer-adapted mesh with a single transition point is needed near each boundary. However, one can formulate a complex-valued problem where the boundary conditions are different for each component, requiring a more complicated mesh for their solution.

3. It would be very interesting to apply the methods presented here to the full Orr-Sommerfeld problem (mentioned briefly in (1.6)), which is a complex-valued, parametrised problem, with mixed boundary conditions (see, e.g, [11]). Moreover, the solution and its derivatives possess boundary layers. A particular version of interest, from a model of wave-current interactions [25], can be expressed as follows: *find the function, u, and (complex) parameter k,*

such that

$$\varepsilon\left(\frac{d^2}{dx^2}-k^2\right)^2 u - i\frac{d^2u}{dx^2} + i\left(k^2 - a(k,x)\right)u = 0 \quad \text{for} \quad 0<x<1, \tag{8.1a}$$

with boundary conditions

$$u(0)=0, \quad u'(0)=0, \quad u(1)=u_1, \quad u''(1)=v_1(k), \tag{8.1b}$$

where u_1 is some specified value, and v_1 is a function of k. Since the parameter k is to be determined, an extra condition is needed. This is provided by the "extra" boundary condition

$$u'''(1) = g(u(1), u'(1), k). \tag{8.1c}$$

This problem is complicated, and interesting, for a number of reasons.

(a) It is a boundary layer problem, so suitable layer-adapted meshes are required for accurate numerical solutions.

(b) It has mixed boundary conditions, and most papers on fourth-order problems in the literature consider only first or second-order boundary conditions. At first, the problem (8.1a)–(8.1b) is not suitable for conversion to a coupled system. However, it can be shown that if the boundary condition on $u'(0)$ is replaced with one on $u''(0)$, then the solution remains largely unchanged away from $x=0$. So, as a first step towards investigating this problem, one could make this simplification.

(c) This problem is *parametrised* meaning that we must also solve for the unknown (complex) number k using (8.1c). For that, a very accurate estimate of $u'(1)$ is needed, again demonstrating that layer-adapted meshes are important.

4. In [11] a version of the Orr-Sommerfeld equation is proposed which features Robin-type boundary conditions at one boundary. These have been studied in the context of coupled systems of singularly perturbed problems (see, e.g., [7]) and would be worthy of further consideration, especially for complex-valued problems.

5. It would be interesting to construct a transformation based on non-constant parameters. In various places in this thesis, we have transformed a fourth-order problem to a system by using the transformation

$$w := \frac{u''-\beta u}{\alpha}, \tag{8.2}$$

where α and β are *constants* chosen depending on the problem data. (In some cases, we have taken $\alpha=1$ or $\beta=1$, but (8.2) shows the most general version to date). This gives

$$u'' = \alpha w + \beta u. \tag{8.3}$$

To proceed with the transformation, we then differentiate to get

$$u^{(4)} = \alpha w'' + \beta u'' = \alpha w'' + \beta\alpha w + \beta^2 u. \tag{8.4}$$

For certain ranges of values of the problem data, we have shown how to determine values of

α and β so that, for example, the coupling matrix is coercive. Outside of those ranges, it may be that there there are no suitable constants α and β. To circumvent this, one could generalise (8.2) further, and allow α and β to be non-constant. That is, one would set $u''(x) = \alpha(x)w(x) + \beta(x)u(x)$. In such a scenario, instead of (8.4), we would have

$$u^{(4)} = \alpha w'' + 2\alpha' w' + \alpha'' w + \beta\alpha w + \beta^2 u + 2\beta' u' + \beta'' u. \tag{8.5}$$

With this, the singularly perturbed, fourth-order, real-valued reaction-diffusion equation (3.1), which presented in Chapter 3 can be transformed into a system of two equations of the form

$$-\varepsilon\alpha w'' - 2\varepsilon\alpha' w' - 2\varepsilon\beta' u' + (a\alpha - \varepsilon\alpha'' - \varepsilon\alpha\beta)w + (a\beta - \varepsilon\beta^2 - \varepsilon\beta'' - b)u = f, \tag{8.6a}$$

$$-u'' + \alpha w + u = 0, \tag{8.6b}$$

subject to the boundary conditions

$$u(0) = w(0) = 0, \qquad u(1) = w(1) = 0. \tag{8.6c}$$

We can write this system as

$$\vec{L}\vec{z} := -\begin{pmatrix} \varepsilon\alpha & 0 \\ 0 & 1 \end{pmatrix}\vec{z}'' - \begin{pmatrix} 2\varepsilon\alpha' & 2\varepsilon\beta' \\ 0 & 0 \end{pmatrix}\vec{z}' + B\vec{z} = \vec{f}, \tag{8.7a}$$

where

$$\vec{z} = \begin{pmatrix} w \\ u \end{pmatrix}, \; B = \begin{pmatrix} a\alpha - \varepsilon\alpha'' - \varepsilon\alpha\beta & a\beta - \varepsilon\beta^2 - \varepsilon\beta'' - b \\ \alpha & 1 \end{pmatrix} \quad \text{and } \vec{f} = \begin{pmatrix} f \\ 0 \end{pmatrix}. \tag{8.7b}$$

This is feasible, but leads to other challenges.

- A system has first-order derivatives. That is, (8.6a) is a convection-diffusion problem. This presents difficulties for the finite element and finite difference solution of the equations, since the discrete systems are not stable on arbitrary meshes: solution can have large oscillations. So the numeric methods needed would need upwinding (or similar) for stability. The literature on the numerical solution of these problems is *vast* (see surveys [**35**, **42**, **43**]) and beyond the scope of this thesis.
- The resulting system, which would be reaction-convection-diffusion in nature, would be coupled by both first derivative and second derivative terms. That is, whereas as the systems we have previously studied are considered to be *weakly coupled*, (8.6) is *strongly coupled*, at least in the first equation. Again, this makes it more difficult to solve, and to analyse.
- The task of determining the range of values of the functions α and β that ensure that the bilinear form (for example) is coercive is clearly more challenging since, instead of computing a single value, we need to determine suitable functions.

6. We have considered only one-dimensional differential equations, but problems in higher dimensions are also very interesting. Furthermore, here the coercivity results for finite element methods will be even more important. This is because, even if an operator does not satisfy

a maximum/minimum principle, one still finds that finite difference methods can be applied and may give accurate solutions: only the proofs are harder. But with finite element methods, coercivity is required for even the existence of the solutions.

Furthermore, although just about any differential equation in one dimension can be solved by one's choice of finite difference or finite element method, the same is not true in higher dimensions. This is because finite difference methods are mainly applied only on simple domain shapes (like rectangles or circles, in two-dimensions), and on tensor product grids. For arbitrary shaped domains finite difference methods may not be feasible, and so finite elements are required.

7. We have considered only problems with smooth data. If, for example, the right-hand side has a discontinuity, then layers can develop in the interior of the domain. There are some very recent papers considering this topic, see, e.g., [**3**, **9**]. So it would be interesting to consider such problems in the context of complex-valued problems.

8. Recall Section 5.6 where we investigated the signs of solution components under strong assumptions on the right-hand side function f. We restricted our attention to the case where f was strictly real or purely imaginary. As mentioned in Remark 5.6.1 more general problems would required much more detailed analysis. This is because the sign of the solution components would depend on relative size of f_r and f_i, and in a way that depends on the other problem data. Indeed, it is possible to construct cases where solutions change sign in the interior of the domain. A full investigation would interesting, but necessarily very detailed.

Bibliography

[1] Kevin W. Aiton and Tobin A. Driscoll. An adaptive partition of unity method for chebyshev polynomial interpolation. *SIAM Journal on Scientific Computing*, 40(1):A251–A265, 2018. (Cited on page 74.)

[2] Nikolai Sergeevich Bakhvalov. On the optimization of the methods for solving boundary value problems in the presence of a boundary layer. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 9(4):841–859, 1969. (Cited on pages i, 18, and 23.)

[3] Mahabub Basha Pathan and Vembu Shanthi. A parameter-uniform non-standard finite difference method for a weakly coupled system of singularly perturbed convection-diffusion equations with discontinuous source term. *Int. J. Adv. Appl. Math. Mech.*, 3(2):5–15, 2015. (Cited on page 131.)

[4] Zachary Battles and Lloyd N. Trefethen. An extension of MATLAB to continuous functions and operators. *SIAM J. Sci. Comput.*, 25(5):1743–1770, 2004. (Cited on page 74.)

[5] Abraham Berman and Robert J. Plemmons. *Nonnegative matrices in the mathematical sciences*, volume 9. Siam, 1994. (Cited on page 118.)

[6] Susanne C. Brenner and L. Ridgway Scott. *The mathematical theory of finite element methods*, volume 15 of *Texts in Applied Mathematics*. Springer, New York, third edition, 2008. (Cited on pages 47 and 48.)

[7] M. Chandru and V. Shanthi. Fitted mesh method for singularly perturbed Robin type boundary value problem with discontinuous source term. *Int. J. Appl. Comput. Math.*, 1(3):491–501, 2015. (Cited on page 129.)

[8] M Chandru and V Shanthi. A schwarz method for fourth-order singularly perturbed reaction-diffusion problem with discontinuous source term. *J. Appl. Math. & Informatics Vol*, 34(5-6):495–508, 2016. (Cited on page 15.)

[9] M. Chandru and V. Shanthi. A Schwarz method for fourth-order singularly perturbed reaction-diffusion problem with discontinuous source term. *J. Appl. Math. Inform.*, 34(5-6):495–508, 2016. (Cited on page 131.)

[10] Philippos Constantinou and Christos Xenophontos. An *hp* finite element method for a 4^{th} order singularly perturbed boundary value problem in two dimensions. *Computers & Mathematics with Applications*, 74(7):1565–1575, 2017. (Cited on page 11.)

[11] P. G. Drazin and W. H. Reid. *Hydrodynamic stability*. Cambridge University Press, 2nd edition, 2004. (Cited on pages 1, 128, and 129.)

[12] Tobin A. Driscoll, Nicholas Hale, and Lloyd N. Trefethen. *Chebfun Guide*. Pafnuty Publications, Oxford, 2014. (Cited on pages 7, 72, and 74.)

[13] John W. Eaton. *GNU Octave Manual*. Network Theory Limited, 2002. (Cited on page 74.)

[14] P. A. Farrell, A. F. Hegarty, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. *Robust computational techniques for boundary layers*, volume 16 of *Applied Mathematics (Boca Raton)*. Chapman & Hall/CRC, Boca Raton, FL, 2000. (Cited on pages 5 and 34.)

[15] Hailong Guo, Can Huang, and Zhimin Zhang. Superconvergence of conforming finite element for fourth-order singularly perturbed problems of reaction diffusion type in 1d. *Numerical Methods for Partial Differential Equations*, 30(2):550–566, 2014. (Cited on page 10.)

[16] Roger A. Horn and Charles R. Johnson. *Matrix analysis*. Cambridge university press, 1990. (Cited on page 41.)

[17] Wolfram Research, Inc. Mathematica, Version 12.0. Champaign, IL, 2019. (Cited on page 74.)

[18] C. R. Johnson. Positive definite matrices. *Amer. Math. Monthly*, 77:259–264, 1970. (Cited on page 41.)

[19] R. B. Kellogg, Niall Madden, and Martin Stynes. A parameter-robust numerical method for a system of reaction-diffusion equations in two dimensions. *Numer. Methods Partial Differential Equations*, 24(1):312–334, 2008. (Cited on pages 59, 82, and 89.)

[20] R. Bruce Kellogg, Torsten Linß, and Martin Stynes. A finite difference method on layer-adapted meshes for an elliptic reaction-diffusion system in two dimensions. *Math. Comp.*, 77(264):2085–2096, 2008. (Cited on pages i, 18, 23, and 25.)

[21] Randall J. LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*, volume 98. Siam, 2007. (Cited on page 5.)

[22] Torsten Linß. *Layer-adapted meshes for reaction-convection-diffusion problems*, volume 1985 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010. (Cited on pages 3, 5, and 21.)

[23] Torsten Linß and Niall Madden. Layer-adapted meshes for a linear system of coupled singularly perturbed reaction-diffusion problems. *IMA J. Numer. Anal.*, 29(1):109–125, 2009. (Cited on page 117.)

[24] Fang Liu, Niall Madden, Martin Stynes, and Aihui Zhou. A two-scale sparse grid method for a singularly perturbed reaction–diffusion problem in two dimensions. *IMA Journal of Numerical Analysis*, 29(4):986–1007, 2008. (Cited on page 48.)

[25] N. Madden, M. Stynes, and G.P. Thomas. On the application of robust numerical methods to a complete-flow wave-current model. In *Proc. Bail, Toulouse*, 2004. (Cited on pages 1, 6, and 128.)

[26] Matlab optimization toolbox, R2021a. The MathWorks, Natick, MA, USA. (Cited on page 74.)

[27] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. *Fitted numerical methods for singular perturbation problems*. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, revised edition, 2012. (Cited on pages 5, 17, 18, 19, 34, 63, 66, and 91.)

[28] Michael B. Monagan, Keith O. Geddes, K. Michael Heal, George Labahn, Stefan M. Vorkoetter, James McCarron, and Paul DeMarco. *Maple 10 Programming Guide*. Maplesoft, Waterloo ON, Canada, 2005. (Cited on page 74.)

[29] Robert E. O'Malley, Jr. *Singular perturbation methods for ordinary differential equations*, volume 89 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1991. (Cited on page 10.)

[30] Robert E. O'Malley, Jr. *Thinking about ordinary differential equations*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, 1997. (Cited on page 3.)

[31] P. Panaseti, A. Zouvani, N. Madden, and C. Xenophontos. A C^1-conforming hp finite element method for fourth order singularly perturbed boundary value problems. *Applied Numerical Mathematics*, 104:81 – 97, 2016. (Cited on pages 11 and 15.)

[32] Mathiyazhagan Paramasivam, John J. H. Miller, and Sigamani Valarmathi. Second order parameter-uniform numerical method for a partially singularly perturbed linear system of reaction-diffusion type. *Math. Commun.*, 18(1):271–295, 2013. (Cited on pages 117 and 118.)

[33] Alfio Quarteroni and Alberto Valli. *Numerical approximation of partial differential equations*, volume 23. Springer Science & Business Media, 2008. (Cited on page 16.)

[34] Hans-Görg Roos and Martin Stynes. A uniformly convergent discretization method for a fourth order singular perturbation problem. In *Extrapolation and defect correction (1990)*, volume 228 of *Bonner Math. Schriften*, pages 30–40. Univ. Bonn, Bonn, 1991. (Cited on page 9.)

[35] Hans-Görg Roos, Martin Stynes, and Lutz Tobiska. *Robust numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2008. (Cited on pages 5, 47, 48, and 130.)

[36] Stephen Russell. *Sparse grid methods for singularly perturbed problems*. National University of Ireland, Galway, Theses ; 12317. National University of Ireland, Galway, Galway, 2016. (Cited on page 3.)

[37] Bill Semper. Locking in finite-element approximations to long thin extensible beams. *IMA Journal of Numerical Analysis*, 14(1):97–109, 1994. (Cited on page 9.)

[38] V. Shanthi and N. Ramanujam. A numerical method for boundary value problems for singularly perturbed fourth-order ordinary differential equations. *Appl. Math. Comput.*, 129(2-3):269–294, 2002. (Cited on pages 7, 12, 57, and 58.)

[39] V. Shanthi and N. Ramanujam. A boundary value technique for boundary value problems for singularly perturbed fourth-order ordinary differential equations. *Comput. Math. Appl.*, 47(10-11):1673–1688, 2004. (Cited on page 14.)

[40] V. Shanthi and N. Ramanujam. Asymptotic numerical method for boundary value problems for singularly perturbed fourth-order ordinary differential equations with a weak interior layer. *Appl. Math. Comput.*, 172(1):252–266, 2006. (Cited on page 14.)

[41] W. A. Stein et al. *Sage Mathematics Software (Version 9.0)*. The Sage Development Team, 2020. `http://www.sagemath.org`. (Cited on page 74.)

[42] Martin Stynes. Steady-state convection-diffusion problems. *Acta Numer.*, 14:445–508, 2005. (Cited on page 130.)

[43] Martin Stynes and David Stynes. *Convection-diffusion problems*, volume 196 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, 2018. (Cited on page 130.)

[44] Endre Süli and David F Mayers. *An introduction to numerical analysis*. Cambridge university press, 2003. (Cited on pages 46 and 48.)

[45] Guang Fu Sun and Martin Stynes. Finite-element methods for singularly perturbed high-order elliptic two-point boundary value problems. I. Reaction-diffusion-type problems. *IMA J. Numer. Anal.*, 15(1):117–139, 1995. (Cited on page 10.)

[46] Christos Xenophontos. A parameter robust finite element method for fourth order singularly perturbed problems. *Comput. Methods Appl. Math.*, 17(2):337–349, 2017. (Cited on page 11.)

[47] Christos Xenophontos, Markus Melenk, Niall Madden, Lisa Oberbroeckling, Pandelitsa Panaseti, and Antri Zouvani. hp finite element methods for fourth order singularly perturbed boundary value problems. In *International Conference on Numerical Analysis and Its Applications*, pages 532–539. Springer, 2012. (Cited on pages i, 7, 15, 40, 41, 42, 44, and 99.)