Preconditioning techniques for singularly perturbed differential equations

A dissertation submitted

by

Anh Thái Nhan (M.Sc.)

to

The School of Mathematics, Statistics and Applied Mathematics,
National University of Ireland, Galway

in fulfilment of the requirements for the degree of

Doctor of Philosophy

Thesis Supervisor: Dr Niall Madden
Head of School: Dr Ray Ryan

July 2015

Abstract

This dissertation is concerned with the numerical solution of linear systems arising from finite difference and finite element discretizations of singularly perturbed reaction-diffusion problems. Such linear systems present several difficulties that make computing accurate solutions efficiently a nontrivial challenge for both direct and iterative solvers.

The poor performance of direct solvers, such as Cholesky factorization, is due to the presence of *subnormal floating point numbers* in the factors. This thesis provides a careful analysis of this phenomenon by giving a concrete formula for the magnitude of the fill-in entries in the Cholesky factors in terms of the perturbation parameter, ε, and the discretization parameter, N. It shows that, away from the main diagonal, the magnitude of fill-in entries decreases exponentially. Furthermore, with our analysis, the location of corresponding fill-in entries associated with some given magnitude can also be determined. This can be used to predict the number and location of subnormals in the factors.

Since direct solvers scale badly with ε, one must use iterative solvers. However, the application of finite difference and finite element discretizations on layer-adapted meshes results in ill-conditioned linear systems. The use of suitable preconditioners is essential. In this thesis we analyze several preconditioning techniques. They include the diagonal and incomplete Cholesky preconditioners for finite difference discretized systems, and a specially designed *boundary layer preconditioner* for a finite element discretized system. The study of the diagonal and incomplete Cholesky preconditioners focuses on the simplicity and robustness of these techniques; while that of the boundary layer preconditioner is concerned with optimality.

Finally, a novel contribution of this thesis is a pointwise uniform convergence proof for one-dimensional singularly perturbed problems. The central idea of the proof is based on the the preconditioning of the discrete system.

Contents

List of Figures

List of Tables

Declaration

I declare that this dissertation is my own work, and I have not obtained a degree in this University, or elsewhere, on the basis of this work. All the sources have been quoted and acknowledged by means of complete references.

ACKNOWLEDGMENT

This dissertation would not have been completed without the endless support from people around me. Therefore, it is such a pleasure for me to take this opportunity to express my deepest gratitude.

First and foremost, I would like to thank my supervisor Dr Niall Madden for introducing me to an interesting research topic to explore and for all the good and helpful advice during my study at National University of Ireland Galway.

I am very grateful to Dr Scott MacLachlan, Memorial University of Newfoundland, Canada (formerly of Tufts University, USA). I will never forget my two week visit in Tufts in September 2013 which resulted in Chapter 5 of this dissertation. I am also grateful to Scott for the implementation of the boundary layer preconditioned Conjugate Gradient which we used to generate the two-dimensional results in that chapter.

My special thanks go to Professor Relja Vulanović. It has been a great pleasure to work with him. I am also indebted to Dr Petri Piiroinen, and Le Phuong Quan who wrote me reference letters in order to secure the Irish Research Council funding for my PhD study in Galway.

I would like to share this moment with my beloved parents. I am genuinely grateful for all their unbounded and unconditional love and support no matter what paths I have chosen to follow. Special thanks to my younger sister who always cheers me up every weekend by her hilarious stories from home.

Last but certainly not least, I find that words would not be enough to thank my beloved wife, Thuong Le, who always stays by my side. Thanks for waiting for me when I was away in the is(Ire)land of mathematics. I cannot wait to fly over the Atlantic and reunite with you for good!

Dedication

To my parents

Chapter 1

Introduction

1.1 Overview

This dissertation addresses the theory and application of solving linear systems that arise from finite difference and finite element discretizations of certain linear singularly perturbed boundary value problems in one and two dimensions. These singularly perturbed problems are ordinary or partial differential equations in which the highest order derivatives are multiplied by a small parameter, the *perturbation parameter*, which we denote by ε. A simple example of a singularly perturbed *convection-reaction-diffusion* problem in one dimension is

$$-\varepsilon^2 u'' + a(x)u' + b(x)u = f(x), \quad \text{on } (0,1), \quad u(0) = u(1) = 0. \tag{1.1}$$

The terms u'', u' and u are the diffusion, convection and reaction terms respectively, while f is the source term. Solutions of these differential equations usually exhibit sharp boundary and/or interior layers, which are narrow regions where solutions, and their derivatives, change abruptly. When $a \not\equiv 0$ the problem (1.1) is of *convection-diffusion* type, whereas if $a \equiv 0$ and $b \not\equiv 0$, it is known as a *reaction-diffusion* problem. Such problems, and their two dimensional analogues are the subject of study of this thesis.

A formal definition of a singularly perturbed problem can be found in [61, Def. 1.1]. It can be summarized for the problem (1.1) as follows. Let u_ε be the solution to (1.1) for a specific ε. We say that the problem (1.1) is *singularly perturbed* for $\varepsilon \to 0$ in certain norm, $\|\cdot\|_*$, if

$$\lim_{\varepsilon \to 0} \|u_\varepsilon - u_0\|_* \neq 0,$$

where u_0 is the solution of the reduced problem

$$a(x)u_0' + b(x)u_0 = f(x), \quad \text{on } (0,1),$$

with a suitably chosen boundary condition. It is very important to note that this definition is norm dependent, see, e.g., [61, Remark 1.2] and also [33, §1.2].

Singularly perturbed differential equations such as (1.1) arise in various practical applications and mathematical models. For example, convection-diffusion problems are found in many formulations of fluid flow problems (such as the linearization of the Navier-Stokes equations, and transport problems), and semi-conductor device simulation. More details on these two significant examples can be found in [96, pages 1–4]. Further examples are given in [78, 82]. Mathematical models involving systems of reaction-diffusion problems appear, for example, in simulation of chemical reactions, wave-current interaction, and biological applications [10, 32, 74].

Finding numerical solutions to the singularly perturbed problems is a great challenge. The difficulties in applying classical numerical schemes to (1.1) stem from the fact that, typically, derivatives of u of order p have magnitude $\mathcal{O}(\varepsilon^{-p})$. Classical techniques do not have the property of being *parameter robust* (also known as "uniformly convergent" and "ε-uniform" in the literature) because they rely on certain derivatives being bounded, which is not the case as $\varepsilon \to 0$. That is, methods that work well when $\varepsilon = \mathcal{O}(1)$, may fail to give meaningful solutions when ε is small, unless one makes unreasonable assumptions such as the discretization parameter, N, being $\mathcal{O}(\varepsilon^{-1})$. A definition of uniform convergence of a numerical method can be formulated as follows (see [61, Def. 1.4]):

> Let u_ε^N be a numerical approximation of u_ε obtained by a numerical method. A numerical method is said to be *uniformly convergent* with respect to the perturbation parameter ε in the norm $\|\cdot\|_*$ if there exists a positive integer, N_0, independent of ε, such that
>
> $$\|u_\varepsilon - u_\varepsilon^N\|_* \leq \eta(N), \quad \text{for } N \geq N_0,$$
>
> with the function η satisfying
>
> $$\lim_{N\to\infty} \eta(N) = 0, \quad \text{and} \quad \partial_\varepsilon \eta \equiv 0. \tag{1.2}$$

Although many methods which are tailored to singularly perturbed problems are uniformly convergent in this strict sense, there are methods which are referred to as "parameter robust" in the literature, but for which the condition $\partial_\varepsilon \eta \equiv 0$ is not satisfied. For example, the convergence of the standard finite element method on a boundary layer-adapted mesh (discussed in Chapter 5) can be proven to be of $\mathcal{O}(\varepsilon^{1/2} N^{-1} \ln N + N^{-2} \ln^2 N)$ in the energy norm. Similarly, the patched mesh method proposed by de Falco and O'Riordan [25] has a pointwise error estimate of $\mathcal{O}(N^{-1} \ln N + \varepsilon)$. Strictly speaking, these methods do not converge uniformly in ε in the sense of (1.2). However,

since error bounds depend on a positive power of ε, they behave well as $\varepsilon \to 0$, and so we will consider them to be parameter robust (this is in contrast to other convergence estimates which have a negative power of ε. For instance, a naïve analysis of finite difference method applied to reaction-diffusion problems on a uniform mesh yields an error bound that is $\mathcal{O}(N^{-2} + \varepsilon^{-2}N^{-2})$, which is not robust with respect to ε).

The ultimate goal of numerical methods for singularly perturbed problems is to achieve the uniform convergence described above. On the other hand, in this thesis, our primary interest is in the theoretical and practical questions raised by direct solution and robust iterative solution methods of the arising linear systems. In next section, we outline some recent developments on this topic.

1.2 Linear solvers for singularly perturbed problems

The design of numerical methods that are robust with respect to the perturbation parameter has been of significant mathematical interest over the past few decades. For evidence of this, see the monographs [33, 61, 77, 102], the state-of-the-art textbook [96], and the many references therein. We also refer the reader to the excellent surveys [36, 53, 58, 91, 95, 105] for further exposition and development of this fascinating topic. Most of these research papers concern the application of finite difference and/or finite element methods on layer-adapted meshes to achieve ε-uniform convergence. These robust schemes often lead to a large of linear systems that need to solved. Most studies proposing these parameter robust methods make the tacit assumption that the complexity of their algorithms is solely dependent on the discretization parameter, and is independent of ε. However, in a recent study [72], MacLachlan and Madden point out that this is not necessarily the case because direct solvers may not be robust for small ε. Therefore, further detailed studies of linear solvers for singularly perturbed problems are required.

Surprisingly, there has been very few studies that consider the issue of solving the linear systems with efficiency that is robust with respect to the perturbation parameter. The exceptions to this are mainly for convection-diffusion problems. In [34], a short analysis of a Gauss-Seidel method for a convection-diffusion problem discretized on a uniform mesh is given. Roos [90] shows that the condition number of matrix arising from the standard upwind scheme discretized on a Shishkin mesh for convection-diffusion problems in one and two dimensions grows unboundedly as ε tends to zero, and a diagonal preconditioner is proposed to improve this situation. In [6], empirical results for ILU-based preconditioners are reported, and it is observed that the iteration

count can be significantly reduced when this strategy is used. The comprehensive textbook by Elman et al. [30] gives a broad account of iterative solvers of systems resulting from finite element discretizations; however, uniform convergence and use of boundary layer-fitted meshes are not discussed. Basic frameworks and several different multigrid components are presented in [47], though, again, with no reference to fitted meshes. Similarly, the recent textbook [86] devotes 30 pages to the topic of linear solvers for singularly perturbed problems (convection-diffusion and reaction-diffusion). While it discusses the need for parameter robust solvers, it considers only discretizations on quasi-uniform meshes, which excludes the possibility of layer-adapted meshes where the local mesh width is ε-dependent. Studies of multigrid methods for convection-dominated problems on Shishkin meshes can be found in [38, 39], where Gaspar et al. use a scalable multigrid scheme. By contrast, there has been very few studies for reaction-diffusion problems in literature compared to those for convection-diffusion problems. An exception to this is the recent article of MacLachlan and Madden [72]. That paper raised several open questions which are subjects of this thesis, and are described in detail in next section.

1.3 Thesis outline

In this thesis, we are particularly interested in developing robust algorithms for solving systems arising from discretizations of singularly perturbed problems. More precisely, we consider the numerical solution of singularly perturbed reaction-diffusion problems in one and two dimensions. Our model problems are:

$$-\varepsilon^2 u'' + b(x)u = f(x), \quad x \in \omega_x := (0,1), \quad u(0) = u(1) = 0, \tag{1.3}$$

and

$$-\varepsilon^2 \Delta u + b(x,y)u = f(x,y), \quad \text{on } \Omega = (0,1)^2, \quad u(\partial\Omega) = g(x,y). \tag{1.4}$$

We are interested in the question of how to robustly and efficiently solve the linear systems of equations when we discretize the above equations by finite difference or finite element methods.

In Chapter 3, we analyze the performance of direct solvers for the problem (1.4). More precisely, we consider the solution of large linear systems of equations that arise when the two-dimensional singularly perturbed reaction-diffusion equation (1.4) is discretized by finite difference methods. This leads to system matrices that are positive definite. The direct solvers of choice for such systems are based on Cholesky factorization. However, as observed in [72], these solvers may exhibit poor performance for singularly perturbed problems, and so their efficiency is not robust with respect to the perturbation parameter, ε. We investigate these limitations of a standard direct solver

for such linear systems, in which the magnitudes of diagonal entries of the system matrices are dominant compared with small off-diagonal entries. As observed in [72, §4.1], the filled-in entries in the standard Cholesky factorization decay exponentially away from the main diagonal. Thus, for small ε and large N, the Cholesky factors contain many subnormal numbers which are very small (more details can be found in Section 3.3.1). In practice, this affects computational speed considerably, and so the amount of time required to solve these linear systems depends badly on the perturbation parameter. There is no mathematical justification of this phenomenon provided in [72], so we consider it in depth in Chapter 3. We provide an analysis of the distribution of entries in the factors based on their magnitude that explains this phenomenon, and give bounds on the ranges of the perturbation and discretization parameters where poor performance is to be expected.

In Chapter 4, we study the use of diagonal and incomplete Cholesky preconditioners for the Conjugate Gradient method. This is motivated, in part, by the fact that there are difficulties in solving such linear systems by direct methods, as discussed above. Therefore, iterative methods are natural choices. However, in Chapter 4, we show that the condition number of the coefficient matrix grows unboundedly when ε tends to zero, and so unpreconditioned iterative schemes, such as the Conjugate Gradient algorithm, perform poorly with respect to ε. Hence, preconditioners are required in order to robustly and efficiently solve these linear systems. We provide a careful analysis of diagonal and incomplete Cholesky preconditionings, and show that the condition numbers of the preconditioned linear systems are robust and independent of the perturbation parameter. We demonstrate numerically the surprising fact that these schemes are more efficient when ε is small, than when ε is $\mathcal{O}(1)$. The analysis of the incomplete Cholesky preconditioner in this chapter also provides an explanation of why, as shown experimentally in [6], ILU-based preconditioned iterative schemes are very efficient for singularly perturbed problems.

Our motivation for studying incomplete Cholesky preconditioning stems from the analysis of Cholesky factorization in Chapter 3, which shows the exponential decrease in magnitude of the fill-in entries in the factors. This suggests that the incomplete Cholesky factorization resembles the full Cholesky factorization of the original matrix when $\varepsilon \ll 1$, and, so, the incomplete Cholesky approximation of the system matrix should be a very good preconditioner.

A further motivation is that, although very successful boundary layer preconditioners for the problems (1.3) and (1.4) have been proposed in [72], these preconditioners are not trivial to implement in practice. Their design is based on the structure of layer-fitted meshes. More precisely, the preconditioners are constructed using *a priori* information about the location and width of layers, and also involve different strategies

in different regions. On the other hand, the diagonal and incomplete Cholesky preconditioners considered in Chapter 4 of the thesis are very easy to implement: they are widely supported on standard platforms, such as MATLAB, and their application does not require any *a priori* information.

An offshoot of this work, presented in Chapter 2, though which is not directly related to the topic of solving linear systems, is an application of preconditioning to prove ε-uniform convergence for singularly perturbed problems in one dimension. For these problems, we consider an upwind scheme on a Shishkin mesh. The resulting linear system is ill-conditioned because the condition number of its matrix grows unboundedly as ε tends to 0. This is shown in [90], where a relatively simple method for conditioning the system is also proposed. We modify this method to obtain the same result, but, at the same time, to make the preconditioned consistency error convergent uniformly in ε. Therefore, ε-uniform pointwise convergence follows from the standard stability-consistency principle. In addition, our approach is interesting because it points to a connection between preconditioning and ε-uniform convergence.

In Chapter 5, we consider the iterative solution of linear systems of equations arising from the discretization of the problems (1.3) and (1.4) by finite element methods on layer-adapted meshes. Motivated by the ideas in [72], we present an analysis for a specially designed boundary layer preconditioner for the one-dimensional reaction-diffusion problem. We prove optimality of the proposed preconditioner in the sense of *spectral equivalence*. Furthermore, appropriate stopping criteria are derived to ensure that the iterative scheme recovers discretization accuracy, but over-solving is avoided. We show how the algorithm can be extended to the two-dimensional problem, and provide numerical experiments which show the efficiency and robustness of this boundary layer preconditioner.

1.4 Notation

Throughout this thesis, C denotes a generic constant that is independent of both the perturbation parameter ε and the mesh parameter N. We write $f(\cdot) = \mathcal{O}(g(\cdot))$ if there exist positive constants C_0, and C_1, independent of the arguments of f and g, such that $C_0|g(\cdot)| \leq f(\cdot) \leq C_1|g(\cdot)|$.

We denote open unit intervals in the x- and y-directions by ω_x, and ω_y, respectively. The open unit square is denoted by $\Omega := (0,1)^2$. Notation for one-dimensional grids is

$$\omega_x^N := \{0 = x_0 < x_1 \ldots < x_N = 1\}, \quad \omega_y^N := \{0 = y_0 < y_1 \ldots < y_N = 1\}.$$

The two-dimensional Cartesian product grid is

$$\Omega^{N,N} := \omega_x^N \times \omega_y^N.$$

By $\|\cdot\|$ we denote the maximum vector norm, $\|\omega\| = \max_i |\omega_i|$, as well as its subordinate matrix norm. Otherwise, a generic norm for vectors and matrices will be denoted by $\|\cdot\|_*$. For a real-valued function $g \in C(\bar{\omega}_x)$, and a mesh function $G^N = [G_0^N, G_1^N, \ldots, G_N^N]^T$ on ω_x^N (say), we denote

$$\|g\|_{\bar{\omega}_x} = \max_{0\le x\le 1} |g(x)|, \quad \|g\|_{\omega_x^N} = \max_{0\le i\le N} |g(x_i)|, \quad \|g - G^N\|_{\omega_x^N} = \max_{0\le i\le N} |g(x_i) - G_i^N|.$$

Similar notation is extended to the two-dimensional domain Ω, and to $\Omega^{N,N}$.

We denote the energy norm, used in finite element analysis, by

$$\|u\|_\varepsilon := \sqrt{\varepsilon^2 \|u'\|_0^2 + \beta_0^2 \|u\|_0^2},$$

where

$$(u, v) := \int_0^1 u(x)v(x)dx, \quad \text{and} \quad \|u\|_0 = (u, u)^{1/2}.$$

A numerical method is said to be convergent of *order p* if the computed error is bounded by $\mathcal{O}(N^{-p})$. We say a method is referred to being convergent at a rate that is *almost* of order p, if the discretization error is greater than $\mathcal{O}(N^{-p})$, but only by some small factor—typically logarithmic. For example, the rate of convergence of the finite difference discretization for singularly perturbed problems (1.3) and (1.4) is $\mathcal{O}(N^{-2}\ln^2 N)$ on a Shishkin mesh (see 1.7.1), which is referred to as "almost second-order".

A *matrix stencil* is used to represent the connection between a discretization and the corresponding entries in the system matrix. The center value of the stencil corresponds to a diagonal entry of the matrix, with the other terms in the stencil corresponding to nonzero entries in the same row. Their position in the stencil is related to the location of the associated nodes in the grid. A simple example is the finite difference approximation of the second-order derivative in one dimension on ω_x, with the equidistant mesh width $h = N^{-1}$:

$$u''(x_i) \approx \frac{u(x_i + h) - 2u(x_i) + u(x_i - h)}{h^2}.$$

Then, the resulting system matrix of this discretization on ω_x^N is tridiagonal, and can be expressed using the matrix stencil

$$A := \begin{bmatrix} \frac{1}{h^2} & \frac{-2}{h^2} & \frac{1}{h^2} \end{bmatrix}.$$

This notation extends to two-dimensional cases. For example, the following 5-point stencil can be used to express the central finite difference approximation of the second-

order derivative in two-dimensions on a uniform grid:

$$A := \frac{1}{h^2}\begin{bmatrix} 0 & 1 & 0 \\ 1 & -4 & 1 \\ 0 & 1 & 0 \end{bmatrix}.$$

The corresponding matrix has at most five nonzero entries per row, with $-4/h^2$ on the diagonal, and the location of these entries depending on the ordering of nodes in the grid. For example, when lexicographical ordering is used on a grid with $(N+1)\times(N+1)$ points, the corresponding row of the matrix (neglecting the boundary condition) is

$$0 \quad \dots \quad 0 \quad \frac{1}{h^2} \quad \underbrace{0 \quad \dots \quad 0}_{N-2 \text{ zeros}} \quad \frac{1}{h^2} \quad \frac{-4}{h^2} \quad \frac{1}{h^2} \quad \underbrace{0 \quad \dots \quad 0}_{N-2 \text{ zeros}} \quad \frac{1}{h^2} \quad 0 \quad \dots \quad 0.$$

For two real $n \times n$ matrices $A = (a_{ij})$, and $B = (b_{ij})$, we write $A \geq B$ if $a_{ij} \geq b_{ij}$ for all $i, j = 1, \dots, n$.

1.5 Introduction to singularly perturbed differential equations

1.5.1 Problems in one and two dimensions

This thesis is primarily concerned with linear singularly perturbed reaction-diffusion problems in one and two dimensions. Therefore, we shall describe these problems in detail.

A singularly perturbed, one dimensional, reaction-diffusion differential equation can be written as

$$\mathcal{L}u := -\varepsilon^2 u'' + b(x)u = f(x), \quad x \in \omega_x, \quad u(0) = u(1) = 0. \tag{1.5}$$

Here we shall assume that the *perturbation parameter*, ε, belongs to $(0, 1]$, and b and f are $C^1(\omega_x)$-functions, where b satisfies

$$b(x) \geq \beta^2 > 0, \quad \text{with } \beta > 0, \quad \text{for } x \in \bar{\omega}_x.$$

The operator $\mathcal{L}$ satisfies a maximum principle (see, e.g., [89, 96]), and therefore, the problem has a unique solution. When $\varepsilon \ll 1$, its solution has two boundary layers: near $x = 0$, and $x = 1$.

To demonstrate the behavior of the solution to (1.5) with respect to the perturbation parameter, let us consider the following simple example

$$-\varepsilon^2 u'' + u = e^x, \quad x \in \omega_x, \quad u(0) = u(1) = 0. \tag{1.6}$$

The solution, which is plotted in Figure 1.1, is easily shown to be:

$$u(x) = \frac{1}{2(1-e^{-2})}\left(e^{x} - e^{-x}\right) - \frac{xe^{x}}{2}, \quad \text{for } \varepsilon = 1,$$

and

$$u(x) = \frac{e^{-x/\varepsilon}\left(e^{(1-1/\varepsilon)} - 1\right) + e^{-(1-x)/\varepsilon}\left(e^{1/\varepsilon - e}\right)}{(1-\varepsilon^2)\left(1-e^{-2/\varepsilon}\right)} + \frac{e^{x}}{1-\varepsilon^2}, \quad \text{for } \varepsilon \in (0,1).$$

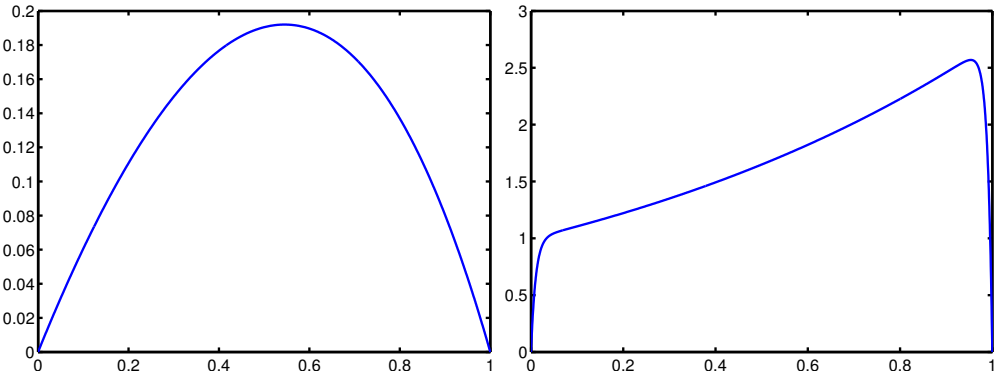

Figure 1.1: Solution to (1.6) with $\varepsilon = 1$ (left) and $\varepsilon = 10^{-2}$ (right).

In two dimensions, the singularly perturbed reaction-diffusion boundary value problem is of the form

$$\mathcal{L}u := -\varepsilon^2 \Delta u + b(x,y)u = f(x,y), \quad \text{on } \Omega = (0,1)^2, \quad u(\partial\Omega) = g(x,y), \tag{1.7}$$

where β is a positive constant such that $b(x,y) \geq \beta^2 > 0$ for all $(x,y) \in \Omega$. When ε is small, the solution to (1.7) typically has four boundary and four corner layers along the edges of the unit square. As an example, consider the following problem,

$$-\varepsilon^2 \Delta u + u = e^{x} + y, \quad (x,y) \in \Omega, \quad u(\partial\Omega) = 0. \tag{1.8}$$

The solution to this problem is plotted in Figure 1.2, where the boundary and corner layers are evident.

1.5.2 Bounds on derivatives and solution decomposition

Estimates of the derivatives of the solution to singularly perturbed problems are required in order to carry out the numerical analysis of discretization methods. As we shall see, the derivatives of u grow unboundedly as ε tends to zero, which presents a challenge when designing robust numerical methods. Here we present bounds on the derivatives of the solution to (1.5), taken from Linß [61, §3.3.1.2].

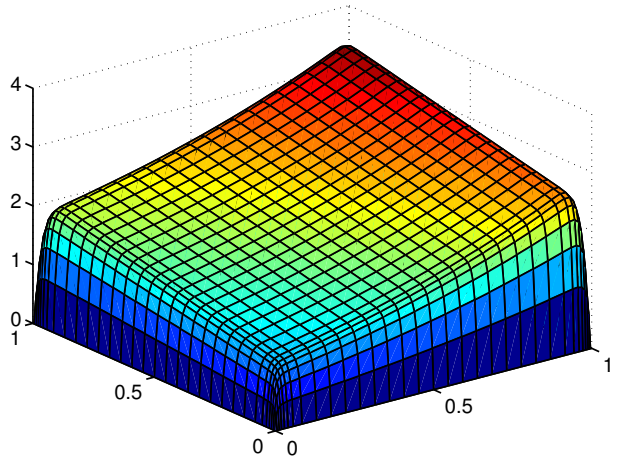

Figure 1.2: Solution to (1.8) with $\varepsilon = 10^{-2}$.

Let $b, f \in C^q[0,1]$ for some positive integer q. Then

$$\left|u^{(m)}(x)\right| \leq C\left\{1 + \varepsilon^{-m}e^{-\beta x/\varepsilon} + \varepsilon^{-m}e^{-\beta(1-x)/\varepsilon}\right\} \text{ for } x \in (0,1) \text{ and } m = 0, 1, \ldots, q.$$

Furthermore, u can be decomposed into the regular and layer components as follows

$$u = v + w_0 + w_1,$$

where

$$\mathcal{L}v = f, \quad \mathcal{L}w_0 = 0, \quad \text{and} \quad \mathcal{L}w_1 = 0 \quad \text{for} \quad x \in (0,1).$$

The derivatives of these components satisfy

$$\left|v^{(m)}(x)\right| \leq C\left(1 + \varepsilon^{q-m}\right),$$

and

$$\left|w_0^{(m)}(x)\right| \leq C\varepsilon^{-m}e^{-\beta x/\varepsilon}, \quad \text{and} \quad \left|w_1^{(m)}(x)\right| \leq C\varepsilon^{-m}e^{-\beta(1-x)/\varepsilon},$$

for $x \in [0,1]$ and $m = 0, 1, \ldots, q$.

Such a decomposition is usually called a "Shishkin decomposition", named after Grigorii I. Shishkin who introduced this concept in [100, 101]. In particular, [101] contains a decomposition for a reaction-diffusion problem in n dimensions. Here we consider the decomposition of the two dimensional problem (1.7), for which the detailed analysis is given in [20]. Let v, w_i and z_i, $i = 1, \ldots, 4$, denote the regular, boundary and corner components respectively. Then, subject to sufficient regularity and compatibility of b, f and g, the solution u can be decomposed as

$$u = v + \sum_{i=1}^{4} w_i + \sum_{i=1}^{4} z_i,$$

where

$$\mathcal{L}v = f, \quad \mathcal{L}w_i = 0, \quad \text{and} \quad \mathcal{L}z_i = 0, \quad i = 1, 2, 3, 4.$$

The regular component v satisfies

$$\left\| \frac{\partial^{p+q} v}{\partial x^p \partial y^q} \right\| \leq C(1 + \varepsilon^{2-(p+q)}), \quad 0 \leq p+q \leq 4.$$

For the boundary layer component, w_1, associated with the edge $y = 0$, we have

$$|w_1(x,y)| \leq Ce^{-y\beta/\varepsilon},$$

and

$$\left\| \frac{\partial^q w_1}{\partial y^q} \right\| \leq C\varepsilon^{-q}, \quad 1 \leq q \leq 4, \quad \left\| \frac{\partial^p w_1}{\partial x^p} \right\| \leq C(1 + \varepsilon^{2-p}), \quad 1 \leq p \leq 4.$$

The corner function z_1 associated with the corner $(0,0)$ satisfies

$$|z_1(x,y)| \leq Ce^{-(x+y)\beta/\varepsilon}, \quad \left\| \frac{\partial^{p+q} z_1}{\partial x^p \partial y^q} \right\| \leq C\varepsilon^{-(p+q)}, \quad 1 \leq p+q \leq 4.$$

Analogous bounds hold for w_2, w_3, w_4, and z_2, z_3, z_4.

These decompositions together with the derivative bounds play a key role in numerical analysis of methods for the singularly perturbed problems (1.5) and (1.7).

1.6 Numerical methods

In this section, we present basic ideas of finite difference discretizations for the problems (1.5) and (1.7), and a finite element discretization for the problem (1.5). As we shall see, for instance, the standard finite difference discretization on uniform meshes is inadequate for singularly perturbed problems. To demonstrate this, we report the numerical results for the finite difference scheme applied to the problem (1.7) on a uniform mesh.

1.6.1 The finite difference method

A finite difference method for one-dimensional problems

Let N be the mesh parameter. Given an arbitrary one-dimensional grid

$$\omega_x^N := \{0 = x_0 < x_1 < \cdots < x_N = 1\}, \text{ with } h_i = x_i - x_{i-1}, \ i = 1, \ldots, N,$$

the standard second order finite difference discretization of the problem (1.5) is given by

$$U_0 = 0,$$

$$-\frac{\varepsilon^2}{\bar{h}_i}\left(\frac{U_{i+1}-U_i}{h_{i+1}}-\frac{U_i-U_{i-1}}{h_i}\right)+b(x_i)U_i=f(x_i),\quad i=1,\ldots,N-1,$$
$$U_N=0,$$

where

$$\bar{h}_i=\frac{h_i+h_{i+1}}{2},\quad i=1,\ldots,N-1.$$

We denote mesh functions on ω_x^N by U^N, W^N, etc.

In general, $\bar{h}_i \neq \bar{h}_{i+1}$ on an arbitrary mesh, so the above discretization results in an unsymmetric linear system of equations. It is natural to consider symmetrising this operator because of the many advantages of dealing with symmetric and positive definite linear systems. When the above system is multiplied by a diagonal matrix with entries $\bar{h}_i$ (the first and last entries on the diagonal equal to 1), the symmetrised finite difference method for the problem (1.5) is given by

$$U_0=0,$$
$$-\varepsilon^2\left(\frac{U_{i+1}-U_i}{h_{i+1}}-\frac{U_i-U_{i-1}}{h_i}\right)+\bar{h}_i b(x_i)U_i=\bar{h}_i f(x_i),\quad i=1,\ldots,N-1, \tag{1.9}$$
$$U_N=0.$$

The discretization (1.9) can be written down in matrix form, with boundaries eliminated, as

$$AU^N=f^N, \tag{1.10}$$

where A is an $(N-1)\times(N-1)$ matrix.

When ε is $\mathcal{O}(1)$, one can use a Taylor's series expansion and a maximum principle technique to prove the convergence of finite difference scheme on a uniform mesh, with the following pointwise error bound

$$\|u-U^N\|_{\omega_x^N}\leq\frac{M}{12}N^{-2}, \tag{1.11}$$

where $\|u^{(iv)}(x)\|\leq M$ for all $x\in\bar{\omega}_x$.

A finite difference method for two-dimensional problems

For the two-dimensional problem (1.7), we take two arbitrary grids, ω_x^N and ω_y^N, in the x- and y-directions respectively. We then apply the natural extension of the method (1.9) to a grid that is the Cartesian product of ω_x^N and ω_y^N, which is denoted $\Omega^{N,N}$. Denote the mesh points of an arbitrary rectangular mesh as (x_i, y_j) for

$i, j \in \{0, 1, \ldots, N\}$, write the local mesh widths as $h_i = x_i - x_{i-1}$ and $k_j = y_j - y_{j-1}$, and let $\bar{h}_i = (h_i + h_{i+1})/2$, and $\bar{k}_j = (k_j + k_{j+1})/2$. Then the numerical scheme is

$$\begin{aligned} \left(-\varepsilon^2 \Delta^N + \bar{h}_i \bar{k}_j b(x_i, y_j) U_{i,j}\right) &= \bar{h}_i \bar{k}_j f(x_i, y_j), \qquad i = 1, \ldots, N-1,\ j = 1, \ldots, N-1, \\ U_{i,j} &= g(x_i, y_j), \qquad \text{otherwise}, \end{aligned} \tag{1.12}$$

where Δ^N is the symmetrised 5-point second-order central difference operator that can be expressed in stencil notation as

$$\Delta^N := \begin{bmatrix} & \dfrac{\bar{h}_i}{k_{j+1}} & \\ \dfrac{\bar{k}_j}{h_i} & -\left(\bar{k}_j\left(\dfrac{1}{h_i} + \dfrac{1}{h_{i+1}}\right) + \bar{h}_i\left(\dfrac{1}{k_j} + \dfrac{1}{k_{j+1}}\right)\right) & \dfrac{\bar{k}_j}{h_{i+1}} \\ & \dfrac{\bar{h}_i}{k_j} & \end{bmatrix}. \tag{1.13}$$

The linear system for the finite difference method can be, again, written as

$$AU^N = f^N, \quad \text{where} \quad A = \left(-\varepsilon^2 \Delta^N + \bar{h}_i \bar{k}_j b(x_i, y_j)\right), \tag{1.14}$$

and A is a symmetric positive definite $(N-1)^2 \times (N-1)^2$ matrix. When the problem (1.7) is not singularly perturbed, an analogue of the one dimensional error estimate (1.11) holds true for this scheme applied on a uniform mesh; see the first two rows of Table 1.1 below. By contrast, it is well-known that when ε is small, standard numerical methods, applied on a uniform mesh, do not yield satisfactory computed solutions. To obtain a meaningful approximation one must make the assumption that the mesh parameter N is $\mathcal{O}(\varepsilon^{-1})$, which is impractical and unrealistic. To demonstrate this phenomenon, let us consider the following example of the two-dimensional problem (1.7) which has only two boundary layers, near the edges $x = 0$ and $y = 0$, and one corner layer near $(0, 0)$. Let $b(x, y) = 1$. The functions f, g are chosen so that

$$u(x, y) = x^3(1 + y^2) + \sin(\pi x^2) + \cos(\pi y/2) + (1 + x + y)\left(e^{-2x/\varepsilon} + e^{-2y/\varepsilon}\right), \tag{1.15}$$

which is plotted in Figure 4.1. Table 1.1 below gives the maximum pointwise errors in the numerical solution to the problem (1.7), which has the solution (1.15), for a range of values of ε and N on a uniform mesh. It is easy to observe that when ε is $\mathcal{O}(1)$, the method is second-order accurate, as expected. However, when ε is small, for example $\varepsilon^2 = 10^{-12}$, the reported error increases as N increases, as seen in the last row in Table 1.1. This is because the boundary layers are not resolved whenever $N \ll \varepsilon^{-1}$. As N increases, but is still much smaller than ε^{-1}, more mesh points are close to, or within, the layers, where the problem is most difficult to solve. Thus the pointwise error increases, and so does not satisfy the definition of uniform convergence discussed in Section 1.1.

To see the approximation obtained from the uniform mesh is unsatisfactory, in Table 1.2, we report the maximum global error in the approximation to (1.15) taken

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	6.17e-03	1.55e-03	3.90e-04	9.76e-05	2.44e-05	6.10e-06
10^{-2}	6.36e-02	1.71e-02	4.36e-03	1.10e-03	2.75e-04	6.88e-05
10^{-4}	5.06e-02	1.52e-01	2.26e-01	1.13e-01	3.15e-02	8.09e-03
10^{-6}	5.44e-04	2.11e-03	8.22e-03	3.15e-02	1.13e-01	2.32e-01
10^{-8}	5.44e-06	2.11e-05	8.32e-05	3.30e-04	1.31e-03	5.21e-03
10^{-10}	5.44e-08	2.11e-07	8.32e-07	3.30e-06	1.32e-05	5.25e-05
10^{-12}	5.44e-10	2.11e-09	8.32e-09	3.30e-08	1.32e-07	5.25e-07

Table 1.1: $\|u - U^N\|_{\Omega^{N,N}}$ with u defined in (1.15) approximated by a FDM on a uniform mesh.

as the piecewise linear interpolant, $(U^N)^I$, of the finite difference solution of (1.7). Clearly, when ε is small, there is no noticeable decrease in the error as N increases: the method is not convergent.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	9.86e-03	2.49e-03	6.23e-04	1.56e-04	3.90e-05	9.75e-06
10^{-2}	2.28e-01	7.30e-02	2.05e-02	5.45e-03	1.41e-03	3.59e-04
10^{-4}	1.44e+00	1.10e+00	7.25e-01	3.45e-01	1.17e-01	3.34e-02
10^{-6}	1.52e+00	1.51e+00	1.51e+00	1.46e+00	1.24e+00	8.50e-01
10^{-8}	1.52e+00	1.51e+00	1.51e+00	1.50e+00	1.50e+00	1.50e+00
10^{-10}	1.52e+00	1.51e+00	1.51e+00	1.50e+00	1.50e+00	1.50e+00
10^{-12}	1.52e+00	1.51e+00	1.51e+00	1.50e+00	1.50e+00	1.50e+00

Table 1.2: $\|u - (U^N)^I\|_{\Omega}$ with u defined in (1.15) approximated by a FDM on a uniform mesh.

1.6.2 The finite element method

A finite element method for one-dimensional problems

Let $H_0^1(\omega_x)$ be the space of continuous functions on ω_x such that, if $w \in H_0^1(\omega_x)$, then

$$\sqrt{\int_0^1 [w'(x)]^2 dx} < \infty, \quad \text{and} \quad w(0) = w(1) = 0.$$

The variational formulation of (1.5) is: find $u \in H_0^1(\omega_x)$ such that

$$B_\varepsilon(u, v) := \varepsilon^2(u', v') + (bu, v) = (f, v), \quad \text{for all } v \in H_0^1(\omega_x),$$

where, as usual, $(u, v) := \int_0^1 u(x)v(x)dx$. The finite element (FE) formulation is arrived at by replacing $H_0^1(\omega_x)$ with a suitably chosen finite dimensional subspace. A natural choice is the space of piecewise linear functions on the arbitrary mesh ω_x^N. Since it may be that b is not easily integrated, we use a mid-point quadrature rule, equivalent to approximating b by a piecewise constant function. We will use b_i to denote $b((x_{i-1} + x_i)/2)$. Then the finite element method on an arbitrary mesh for (1.5) leads to the linear system

$$AU^N = f^N, \tag{1.16}$$

where the system matrix can be written as $A = S + M$, with

$$S = \begin{bmatrix} -\dfrac{\varepsilon^2}{h_i} & \dfrac{\varepsilon^2}{h_i} + \dfrac{\varepsilon^2}{h_{i+1}} & -\dfrac{\varepsilon^2}{h_{i+1}} \end{bmatrix},$$

and

$$M = \begin{bmatrix} \dfrac{h_i b_i}{6} & \dfrac{h_i b_i + h_{i+1} b_{i+1}}{3} & \dfrac{h_{i+1} b_{i+1}}{6} \end{bmatrix}.$$

The bilinear form (1.6.2) is continuous and coercive, so standard finite element arguments (see, e.g., the textbook by Brenner and Scott [8]), can be applied to give a quasi-optimal error estimate

$$\|u - U^N\|_\varepsilon \leq C\|u - v^N\|_\varepsilon,$$

where v^N is any piecewise linear function defined on the mesh ω_x^N. In particular,

$$\|u - U^N\|_\varepsilon \leq C\|u - u^I\|_\varepsilon,$$

where u^I is the nodal interpolant of u on ω_x^N. Since, if $\varepsilon \ll 1$ and the mesh is uniform, the interpolant does not capture the layers in u, so one would not expect to obtain an accurate solution. (Actually, it has been shown by, e.g., Schopf [99], that $\|u - U^N\|_\varepsilon \leq CN^{-1/2}$, independent of ε, but details of this are beyond the scope of this thesis. We also note that $\|u - U^N\| \simeq \mathcal{O}(1)$.)

1.7 Uniform convergence on fitted meshes

Solutions of singularly perturbed boundary value problems change abruptly in the layer regions. As observed in Section 1.6, a standard finite difference scheme on an equidistant mesh may not yield a satisfactory numerical solution unless we assume that N is $\mathcal{O}(\varepsilon^{-1})$. To resolve the layers, it is natural to have grids that condense in the layer regions. However, the construction of such grids usually requires *a priori* knowledge of the behaviour of the exact solution, see Section 1.5.2. The present section is devoted to such meshes, specifically the piecewise uniform *Shishkin mesh*, and

the graded *Bakhvalov mesh*. We outline the construction of these meshes for one and two-dimensional problems, and give numerical results for a two-dimensional problem (results for a one-dimensional problem are very similar). We postpone a detailed numerical analysis to Chapter 2, where a novel preconditioning-based proof technique is introduced.

1.7.1 Shishkin meshes

The piecewise uniform meshes of Shishkin have gained much popularity since the mid-1990s because of the simplicity of their construction and analysis. The monographs [33, 77], and the textbook [96, §2.4.2] discuss the use of these meshes at length. Miller et al. [77] provide the analysis of finite difference methods on this mesh for convection-diffusion and reaction-diffusion problems, whereas Farrell et al [33] focus on detailed numerical results on Shishkin meshes for various problems in one and two dimensions. A comprehensive survey, which is devoted to Shishkin's great contribution to singularly perturbed problems, is given in [53]. That review paper also fully explains the analysis that leads to the construction of Shishkin meshes, as well as the Shishkin solution decomposition for convection-diffusion problems in two dimensions.

We describe here the construction of a Shishkin mesh for the reaction-diffusion problem in one dimension (1.5), which condenses in the regions near boundary layers (recall Figure 1.1).

Recall that b in (1.5) is bounded below by $\beta^2 > 0$. Define the *mesh transition point*

$$\tau_x := \min\left\{\frac{1}{4},\, 2\frac{\varepsilon}{\beta}\ln N\right\}. \tag{1.18}$$

Then the interval $[0,1]$ is divided into three subintervals: $[0,\tau_x]$, $[\tau_x, 1-\tau_x]$ and $[1-\tau_x, 1]$. The mesh is then constructed by subdividing $[\tau_x, 1-\tau_x]$ into $N/2$ equidistant mesh intervals of length $H = 2(1-2\tau_x)/N$, and subdividing each of $[0,\tau_x]$ and $[1-\tau_x, 1]$ into $N/4$ equidistant mesh intervals of length $h = 4\tau_x/N$, as shown in Figure 1.3. (We are assuming that N is divisible by 4). When the problem (1.5) is discretized by the scheme (1.9) on this Shishkin mesh, one can show almost first-order ε-uniform convergence (see [77, Chapter 6]):

$$\|u - U^N\|_{\omega_x^N} \le CN^{-1}\ln N. \tag{1.19}$$

A refined analysis using a more sophisticated barrier function technique can be used to prove that (see [78, Theorem 6.1])

$$\|u - U^N\|_{\omega_x^N} \le CN^{-2}\ln^2 N. \tag{1.20}$$

Figure 1.3: A Shishkin mesh for a reaction-diffusion problem in one dimension.

See Section 2.5 below for an alternative proof.

A Shishkin mesh for the two dimensional problem (1.7), which typically have four boundary and four corner layers, (see Figure 1.2), is constructed by taking a Cartesian product of Shishkin grids on each direction. In Figure 1.4, we show a Shishkin mesh for the two dimensional reaction-diffusion problems.

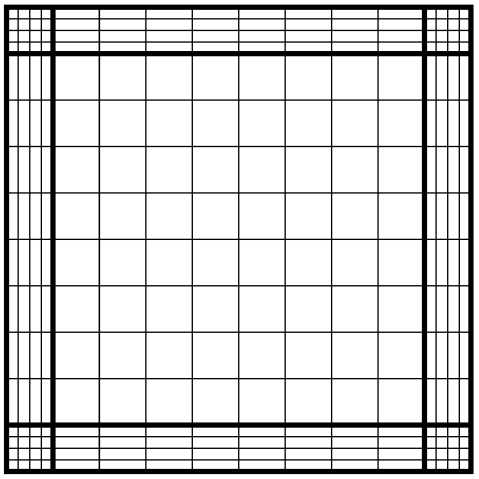

Figure 1.4: A Shishkin mesh for a reaction-diffusion problem in two dimensions.

Recall example (1.15), which has only two boundary layers and one corner layer. For this problem, the transition point is taken as $\tau_x = \min\{1/2, 2\varepsilon \ln N/\beta\}$. Then, the unit interval is divided into two subintervals $[1, \tau_x]$ and $[\tau_x, 1]$. We divide each of these intervals into $N/2$ equidistant mesh intervals. On y-direction, we take $\tau_y = \tau_x$, with a similar construction for subintervals of $[1, \tau_y]$ and $[\tau_y, 1]$.

The first full analysis for a standard finite difference scheme applied on a Shishkin mesh to solve a two-dimensional reaction-diffusion problem is due to Clavaro et al. [20] (see also [52] for an extension to coupled systems). More precisely, in [20], the following parameter robust error estimate is proved: there is a constant C independent of N and ε such that

$$\|u - U^N\|_{\Omega^{N,N}} \leq CN^{-2} \ln^2 N. \tag{1.21}$$

In Table 1.3, we report the maximum pointwise error when (1.7) is solved on the Shishkin mesh described above. When ε is $\mathcal{O}(1)$, the mesh is uniform, and so the numerical results are identical to those reported in Table 1.1 (first and second rows). When ε is small, the error decreases with a convergence rate of $\mathcal{O}(N^{-2} \ln^2 N)$ when N

increases, as seen in (1.21). More importantly, for a fixed N, although we observe that the error initially increases as ε deceases, due mainly to the mesh jumping from being uniform to being piecewise uniform, the observed error is robust with respect to ε.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	6.17e-03	1.55e-03	3.90e-04	9.76e-05	2.44e-05	6.10e-06
10^{-2}	6.36e-02	1.71e-02	4.36e-03	1.10e-03	2.75e-04	6.88e-05
10^{-4}	9.04e-02	3.76e-02	1.44e-02	5.00e-03	1.65e-03	5.23e-04
10^{-6}	9.08e-02	3.82e-02	1.47e-02	5.11e-03	1.68e-03	5.35e-04
10^{-8}	9.08e-02	3.83e-02	1.47e-02	5.12e-03	1.69e-03	5.37e-04
10^{-10}	9.08e-02	3.83e-02	1.47e-02	5.12e-03	1.69e-03	5.37e-04
10^{-12}	9.08e-02	3.83e-02	1.47e-02	5.12e-03	1.69e-03	5.37e-04

Table 1.3: $\|u - U^N\|_{\Omega^{N,N}}$ with u defined in (1.15) solved by a FDM on a Shishkin mesh.

1.7.2 Bakhvalov meshes

When the problem is not singularly perturbed, the scheme (1.12) applied on a uniform mesh yields a discrete solution which is fully second-order accurate. However, as ε becomes small, the convergence rate on the Shishkin mesh is reduced by a logarithmic factor, as shown in (1.21). However, the more sophisticated graded boundary layer-adapted mesh of Bakhvalov [7] can be used to recover second-order convergence, i.e., if U^N is computed on a suitable Bakhvalov mesh, then there exists a constant, C, independent of both N and ε such that

$$\|u - U^N\|_{\omega_x^N} \leq CN^{-2}.$$

Away from the layers, the mesh is equidistant, like the Shishkin mesh. Inside the boundary layers, the mesh is graded. It can be described by a *mesh generating* function, ψ, defined as

$$\psi(t) = \begin{cases} \chi(t) := -\dfrac{\sigma\varepsilon}{\beta} \ln\left(1 - t/q\right), & \text{for } t \in [0, \tau_B], \\ \phi(t) := \chi(\tau_B) + \chi'(\tau_B)(t - \tau_B), & \text{for } t \in [\tau_B, 1/2], \\ 1 - \psi(1 - t), & \text{for } t \in (1/2, 1], \end{cases}$$

where the Bakhvalov mesh transition point τ_B is chosen so that $\psi \in C^1[0, 1]$. The mesh parameters q and σ are user-chosen and control, respectively, the proportion of the mesh points in the layer regions, and the grading of the mesh within the layers. A diagram of such mesh is shown in Figure 1.5.

This mesh was first introduced in [7] in late 1960s for one and two-dimensional reaction-diffusion problems. A generalization, referred to as a *Bakhvalov-type* mesh is

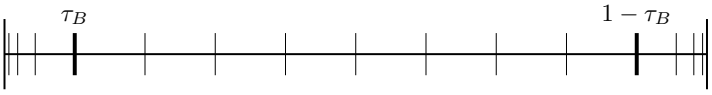

Figure 1.5: A Bakhvalov mesh for a reaction-diffusion problem in one dimension.

introduced in [112], where uniform convergence for one dimensional semilinear reaction-diffusion problems is proved. The extension to coupled systems of two-dimensional reaction-diffusion problems is presented in [49].

Table 1.4 shows the maximum pointwise error when the two dimensional reaction-diffusion problem (1.7) is discretized on the Bakhvalov mesh described above. Full second-order accuracy of the discrete solution is easily observed.

ε^2	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$
1	6.17e-03	1.55e-03	3.90e-04	9.76e-05	2.44e-05	6.10e-06
10^{-2}	3.94e-02	1.00e-02	2.56e-03	6.39e-04	1.60e-04	4.00e-05
10^{-4}	3.41e-02	9.56e-03	2.44e-03	6.13e-04	1.54e-04	3.84e-05
10^{-6}	3.42e-02	9.72e-03	2.49e-03	6.26e-04	1.57e-04	3.94e-05
10^{-8}	3.41e-02	9.72e-03	2.49e-03	6.28e-04	1.58e-04	3.95e-05
10^{-10}	3.41e-02	9.72e-03	2.49e-03	6.28e-04	1.58e-04	3.95e-05
10^{-12}	3.41e-02	9.72e-03	2.49e-03	6.28e-04	1.58e-04	3.95e-05

Table 1.4: $\|u - U^N\|_{\Omega^{N,N}}$ with u defined in (1.15) approximated by a FDM on a Bakhvalov mesh.

1.8 Preliminaries for linear solvers

Our goal in this section is to present some fundamental ideas from linear algebra, including Geršgorin's Theorem in Section 1.8.1, that are used in the rest of this thesis. In Section 1.8.2, we provide some initial analysis of classical iterative linear solvers such as the Jacobi, Gauss-Seidel and Successive Overrelaxation (SOR) methods for the problem (1.5). We will show that, for a fixed ε, these schemes are slow to converge, and scale badly in N, even for one-dimensional problems. Hence, in the rest of the thesis, we focus on iterative methods that are based on Krylov subspaces, such as the Conjugate Gradient algorithm and preconditioners for them. Therefore, we conclude, in Section 1.8.3, with a discussion of preconditioning for solving linear systems.

1.8.1 Geršgorin's theorem

A very famous, and useful, theorem to bound the spectral radius of the matrix A is due to Geršgorin [40].

Theorem 1.1 ([110, Theorem 1.11]). *Let $A = [a_{i,j}]$ be an arbitrary $n \times n$ complex matrix, and let*

$$\Lambda_i := \sum_{j \neq i}^{n} |a_{i,j}|, \quad 1 \leq i \leq n,$$

where $\Lambda_i = 0$ if $n = 1$. If λ is an eigenvalue of A, then there is a positive integer r, with $1 \leq r \leq n$, such that

$$|\lambda - a_{r,r}| \leq \Lambda_r.$$

Consequently, if $\lambda_{\max}$ and $\lambda_{\min}$ are the largest and smallest eigenvalues, respectively, of the real symmetric matrix A, from Geršgorin's Theorem, we get that

$$\lambda_{\max} \leq \max_{i=1,\ldots,n} \left\{ \sum_{j=1}^{n} |a_{i,j}| \right\}, \quad \text{and} \quad \lambda_{\min} \geq \min_{i=1,\ldots,n} \left\{ |a_{i,i}| - \sum_{j \neq i}^{n} |a_{i,j}| \right\}.$$

1.8.2 Classical iterative schemes for solving linear systems

For solving the linear system

$$Ax = b, \tag{1.22}$$

one can use direct methods (such as Gaussian Elimination, Cholesky factorization, or LU factorization). However, when the linear system is large, direct methods are not suitable since they require too much memory. For this reason, one uses iterative methods. The idea of these methods is that, starting from an initial guess $x^{(0)}$, the method generates a sequence of vectors $\{x^{(1)}, x^{(2)}, \ldots\}$ that *converges* to the true solution. Many such schemes can be written in the form:

$$x^{(m+1)} = Rx^{(m)} + c, \tag{1.23}$$

where R is some suitable matrix.

We present here some analysis of classical iterative solvers such as the Jacobi, Gauss-Seidel and Successive Overrelaxation (SOR) methods for the linear system arising when the problem (1.3) is discretized by the finite difference method (1.9) on a Shishkin mesh. These algorithms, especially the Jacobi and Gauss-Seidel methods, can be used as smoothers for multigrid methods. Detailed descriptions of these classical schemes can be found in many standard textbooks, e.g. [26, 110]. Suppose the matrix A has no zeros on its diagonal. We write $A = D - L - U$, where D is the diagonal of A, and $-L$

and $-U$ are the strictly lower triangular and upper triangular parts of A, respectively. Then Table 1.5 below summarizes these schemes in terms of (1.23), while Theorem 1.2 gives us details of their convergence properties.

Method	R	c
Jacobi's	$R_J = D^{-1}(L+U)$	$D^{-1}b$
Gauss-Seidel	$R_{GS} = (D-L)^{-1}U$	$(D-L)^{-1}$
SOR(ω)	$R_{SOR} = (D-\omega L)^{-1}((1-\omega)D+\omega U)$	$\omega(D-\omega L)^{-1}b$

Table 1.5: Classical iterative schemes.

Theorem 1.2 ([26, Theorem 6.1])**.** *The iteration* (1.23) *converges to the solution of* (1.22) *for all starting vectors* x_0 *and for all* b *if and only if* $\rho(R) < 1$*, where* $\rho(R)$ *is spectral radius of* R*.*

We now describe the convergence analysis when applying these methods to the linear system generated by finite difference scheme (1.9) on the Shishkin mesh described in Section 1.7 for the problem (1.5). The matrix A defined in (1.16) is a symmetric positive definite $(N-1)\times(N-1)$ matrix. Furthermore, since $b(x) \geq \beta^2 > 0$, the matrix A is strictly row diagonally dominant. Thus, it is easy to show that the Jacobi and Gauss-Seidel methods converge [26, Theorem 6.2]. Moreover, A is symmetric and has positive diagonal entries, so SOR(ω) converges for all $0 < \omega < 2$ [26, Theorem 6.5]. However, these results only tell us that the method will converge, and not how quickly, or how the performance depends on ε. In order to compare the speed of convergence of different solvers and discretizations, we need to define:

Definition 1.1 ([26, Def. 6.5])**.** *The* rate of convergence *of* (1.23) *is* $r(R) \equiv -\log_{10}\rho(R)$.

The value of $r(R)$ in above definition tells us the increase in the number of correct decimal digits per iteration.

From the above definition, we observe that the smaller the spectral radius, the greater the rate of convergence. We investigate use of the iterative schemes described in Table 1.5 applied to the example (1.6). The left of Figure 1.6 below shows a plot of the residuals, i.e., $\|b - Ax^{(m)}\|$, for each of the Jacobi, Gauss-Seidel and SOR methods, respectively, using the finite difference scheme (1.9) on a Shishkin mesh with $N = 2^5$ and $\varepsilon = 1$, for the first 20 iterations. The right of Figure 1.6 shows a plot of the spectral radii of the Jacobi, Gauss-Seidel and SOR methods. The initial guess is taken to be zero in the experiments. It clearly shows that these iterative schemes are linearly convergent, but the convergence is slow. In contrast, when $\varepsilon = 10^{-3}$, these methods converge rapidly as seen in Figure 1.7. This is because the off-diagonal entries of the

symmetrised matrix are very small as ε is small. Therefore, the matrix resembles the diagonal matrix, which is easy to invert.

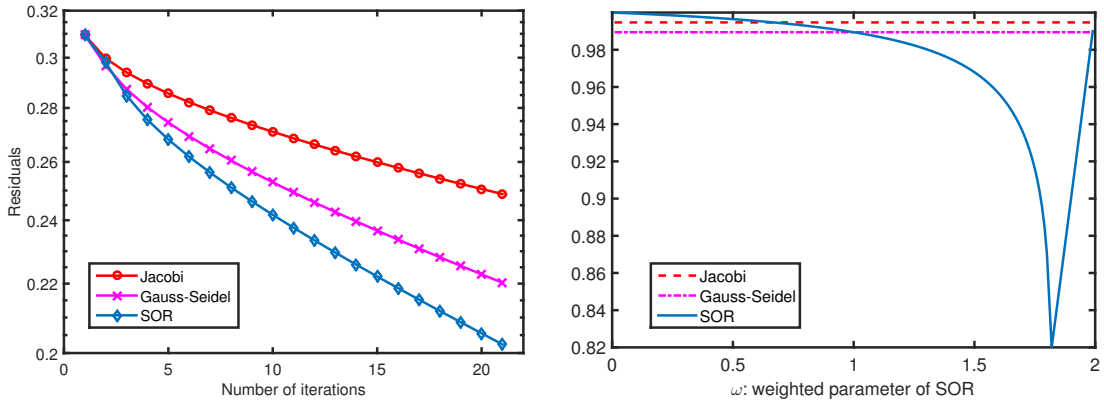

Figure 1.6: Residuals and spectral radii of classical methods for a 1D problem, with $\varepsilon = 1$ and $N = 2^5$.

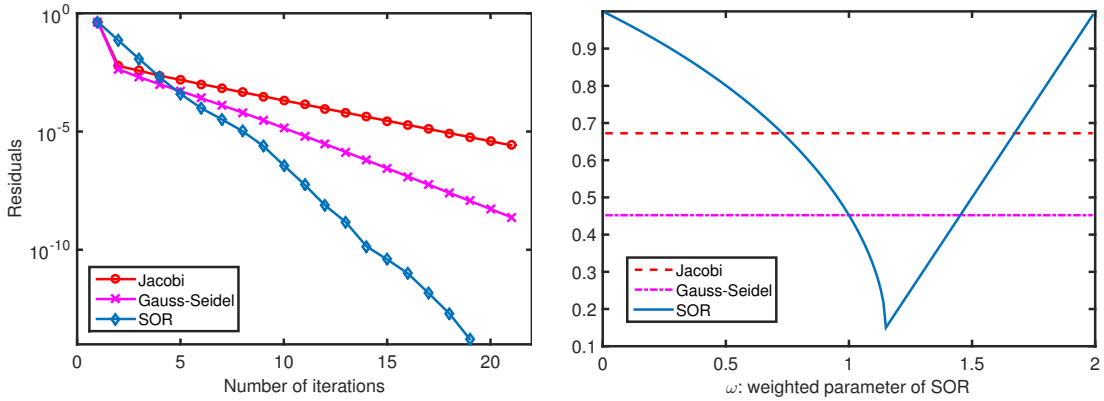

Figure 1.7: Residuals and spectral radii of classical methods for a 1D problem, with $\varepsilon = 10^{-3}$ and $N = 2^5$.

As an example of an analysis of an iterative method for the singularly perturbed problems, we consider the Jacobi method for the ordinary differential equation (1.5) on the Shishkin mesh. Applying the Geršgorin's Theorem to R_J from the table above, we can give the upper bound for the spectral radius:

$$\rho(R_J) \leq \max\{K_1, K_2, K_3\},$$

where

$$K_1 = \frac{N^2}{N^2 + 32\ln^2 N}, \quad K_2 = \frac{N^2\varepsilon}{N^2\varepsilon + \beta\ln N(\beta - 3\varepsilon\ln N)/2},$$

and

$$K_3 = \frac{N^2\varepsilon^2}{N^2\varepsilon^2 + (\beta - 4\varepsilon\ln N)^2/8}.$$

It is easy to see that when N increases, K_1 becomes dominant and tends to 1. Therefore, for a fixed ε, the rate of convergence of the schemes in Table 1.5 decreases as N

increases: compare Figure 1.7 with Figure 1.8, which shows the corresponding residuals and spectral radii when $\varepsilon = 10^{-3}$ and $N = 2^8$.

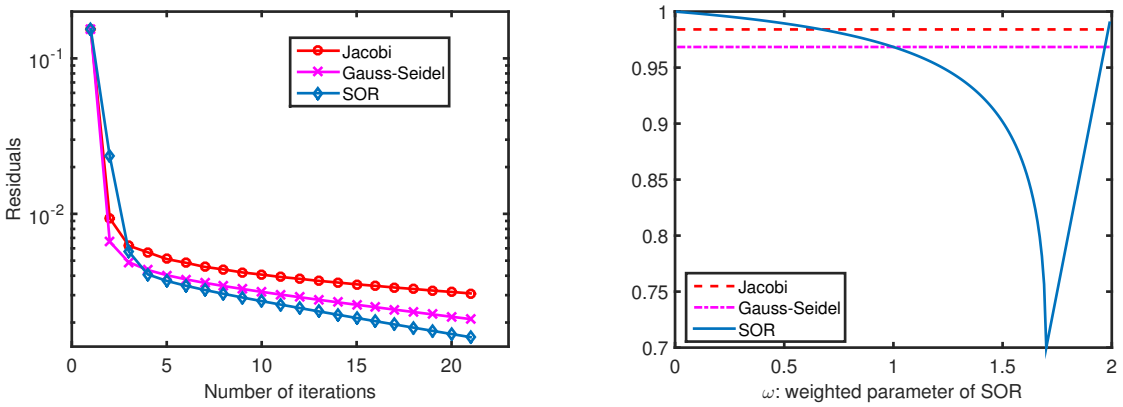

Figure 1.8: Residuals and spectral radii of classical methods for a 1D problem with $\varepsilon = 10^{-3}$ and $N = 2^8$.

For two-dimensional problems, these classical iterative schemes become inefficient due to the increase in the number of degree of freedom from $\mathcal{O}(N)$ to $\mathcal{O}(N^2)$. Figure 1.9 shows the residuals for the Jacobi, Gauss-Seidel, and SOR methods with $\varepsilon = 10^{-3}$, using the finite difference scheme (1.12) on a Shishkin mesh with $N = 2^5$ (left), and $N = 2^8$ (right) to the problem (1.7). It is easy to see that convergence slows considerably as N increases.

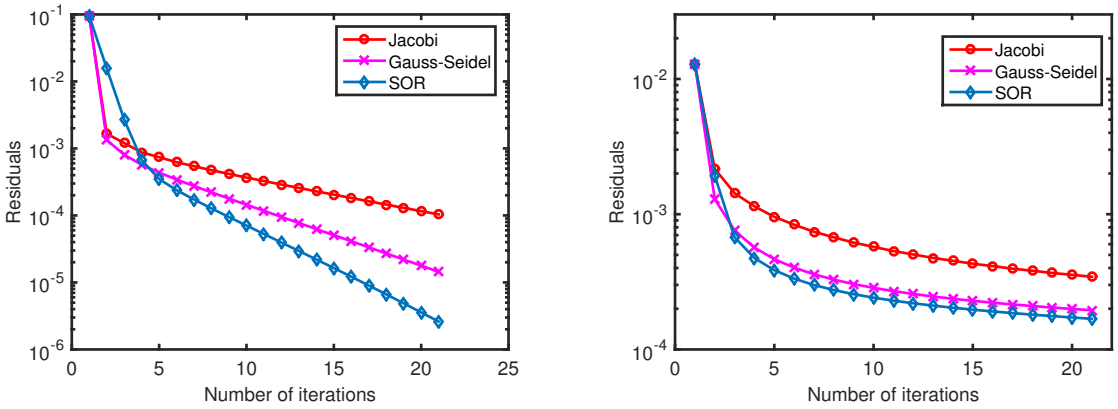

Figure 1.9: Residuals of classical methods for a 2D problem with $\varepsilon = 10^{-3}$, $N = 2^5$ (left) and $N = 2^8$ (right).

In Figure 1.10, we plot the residuals computed when the SOR method is applied to the two-dimensional problem (1.7). On the left of Figure 1.10, we fix $N = 2^6$, and vary ε. As observed for the one-dimensional problem, it is seen that the rate of convergence increases for smaller ε. However, although the convergence is rapid for the first few iterations, it then slows with the same convergence rate observed when $\varepsilon = 1$. This is

because when ε is small, the error associated with the interior region dominates and easily damped by the SOR method. After that, the error in different regions becomes balanced. However, in the layer regions, the small local mesh width means that the discretization resembles that of a diffusion-dominated problem. Therefore, after the first few iterations, we see the same rate of convergence as when $\varepsilon = 1$. On the right of Figure 1.10, for a fixed $\varepsilon = 10^{-3}$, we observe slower convergence when N increases.

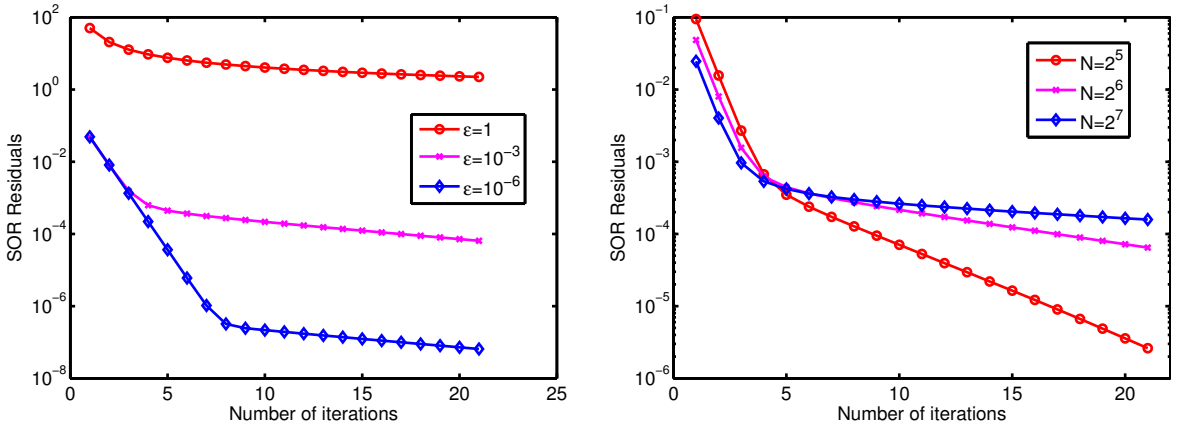

Figure 1.10: Residuals of SOR for a 2D problem, for various N and ε.

1.8.3 Preconditioners

As we have seen, the matrices defined in (1.10) and (1.14) are symmetric positive definite. The most commonly used iterative method for these linear systems is the Conjugate Gradient (CG) method. CG converges rapidly if the coefficient matrix A is close to the identity [44]. However, as we shall see in Chapter 4 and Chapter 5, the matrix A is very ill-conditioned due to the nature of discretization on a special layer-adapted mesh. Therefore, one should use a *preconditioning* technique. That is, we replace the original system by an equivalent one in which it is easier to solve by iterative methods. For example, the linear system (1.22) can be rewritten as

$$M^{-1}Ax = M^{-1}b, \tag{1.24}$$

where M is called a *preconditioner*. Ideally, M is chosen so that it satisfies the following properties [97, 44]:

P1) M is a good approximation of A,

P2) it is inexpensive to solve linear system $Mx = b$.

For example, if we choose $M = A$, then the condition (P1) is satisfied, but the system $Mx = b$ in (P2) has the same complexity as the original system. If M is chosen to

satisfy (P2), for example, by taking M to be the identity matrix, then M will probably not approximate A very well. Therefore, a suitable preconditioner must balance these properties. This can be difficult for singularly perturbed problems since the nature of the linear system can vary so much between subregions. There are many proposed preconditioners in the literature. We shall study the diagonal and incomplete Cholesky preconditioners in Chapter 4, and the specially designed boundary layer preconditioner in Chapter 5.

1.9 Other literature on singularly perturbed problems

As stated in [95, §1]:

> A search of the MathSciNet database for papers published in the years 2005–2014 with MSC Primary Classification 65 (viz., Numerical Analysis) and the phrase singular* perturb* [in MathSciNet asterisks are wildcards] yields 879 published works.

This clearly shows that the numerical analysis of singularly perturbed problems is a very active research topic. In this section we briefly review some other related developments on singularly perturbed problems. We do not aim to give a full discussion; nevertheless, we try to provide some recent progress in the field including the approaches of *fitted operator* methods and *layer-fitted mesh* methods.

Before the 1990s, the majority of research papers that focused on ε-uniform convergence where concerned with fitted operators, used on equidistant grids, see, e.g., [28], [77, Chapter 4], as well as [96, §I.2.1] for a convection-diffusion problem in one dimension. These schemes are often referred as *Il'in-Allen-Southwell* schemes [3, 48], and, typically yield uniform convergence in the maximum norm for one-dimensional singularly perturbed problems as follows

$$\|u - U^N\| \leq CN^{-1}.$$

Recently, Roos and Schopf [94] extended this scheme to the two dimensional case by using the hybrid stability approach which is based on discrete Green's functions. They show that, under some conditions, the uniform first-order convergence of the Il'in scheme is retained for the two-dimensional case. The sufficient conditions for uniform convergence of Il'in-Allen-Southwell schemes were derived by Farrell [35]. Exponentially fitted finite element methods, which also turn out to be uniformly convergent, have been derived by a fitted operator approach, see, e.g., [96, §2.2.5] and also [51, 106].

The 1990s saw the introduction of the simple piecewise uniform Shishkin mesh (see Section 1.7.1), on which one can obtain convergence, uniformly in ε, by standard discretizations. Since then much attention has turned into the approach of fitted mesh methods. For example, in [107], Stynes and Roos extended the analysis based on barrier functions, which was introduced in the well-known paper [50], to an arbitrary grid for convection-diffusion problems. More precisely, for one-dimensional convection-diffusion problems discretized by upwind schemes on the Shishkin mesh, we have

$$\|u - U^N\| \leq \begin{cases} CN^{-1} \ln N, & \text{simple upwind scheme,} \\ CN^{-2} \ln^2 N, & \text{modified upwind scheme.} \end{cases}$$

The analogous extensions to two-dimensional convection-diffusion problems are due to Linß and Stynes [68, 57]. The key ingredients in these papers are the barrier functions introduced in [107], and Shishkin decomposition for two-dimensional convection-diffusion problems analyzed in [69]. The analysis for systems of singularly perturbed convection-diffusion problems is due to Linß [59, 60]. The barrier function technique works fine for Shishkin-type meshes [92]; however, it is unknown that if it can be applied to Bakhvalov type meshes [61, Remark 4.21]. In particular, uniform convergence analysis of the upwind finite difference method applied on a Bakhvalov-type mesh to the two-dimensional convection-diffusion problem is still an open question [95, Question 6]. As for non-stationary convection-diffusion problems, the reader is referred to the monograph [96, Part II] and research papers [16, 17], as well as [43], where a time-dependent convection-diffusion problem with interior layers is analyzed.

For one-dimensional reaction-diffusion problems, as discussed in Section 1.7, the detailed convergence analysis is given in [77, Chapter 6] and [78, Chapter 6]. The extension to two-dimensional problems is provided in [20]. The study of systems of singularly perturbed reaction-diffusion problems is given in [52, 63, 62, 65, 73]. Higher-order schemes for reaction-diffusion singularly perturbed systems can be found in, e.g., [18, 19]. For time-dependent reaction-diffusion problems, we refer the reader to [11, 13, 14, 15, 64]. In particular, Miller et al. [79] proved almost second-order convergence in space by employing a piecewise uniform barrier function, and so improved on the almost first-order convergence analysis for the steady-state problem proved by the same authors in [78, Chapter 6].

Within the scope of this thesis, it is impossible to present all advances of this fascinating topic. We hope the discussion above convinces the reader that this is a wide and rapidly developing area of research. Although remarkable results have been achieved over last few decades, there are still many interesting open questions relating to the numerical solution of singularly perturbed differential equations [95].

Chapter 2

Uniform convergence via preconditioning

The concept of uniform convergence is crucial to this thesis which is primarily concerned with direct and iterative solvers of linear systems of equations that arise when specialized parameter robust methods are used to solve singularly perturbed problems. Therefore, a detailed theoretical discussion of uniform convergence is warranted. In this chapter, the main focus is convergence analysis of the singularly perturbed convection-diffusion problems in one dimension. However, rather than repeating standard theory, we give a new proof of pointwise uniform convergence when these problems are discretized by a finite difference scheme. This proof technique has been published as [115]: R. Vulanović and T. A. Nhan, *Uniform convergence via preconditioning*, Int. J. Numer. Anal. Model. Ser. B., 5(4):347-356, 2014. Furthermore, to provide a connection with later chapters, we also present a simple uniform convergence proof for singularly perturbed reaction-diffusion problems in one dimension.

2.1 Introduction

We consider the following one-dimensional singularly perturbed problem of convection-reaction-diffusion type,

$$\mathcal{L}u := -\varepsilon u'' - b(x)u' + c(x)u = f(x),\ x \in \omega_x,\ u(0) = u(1) = 0, \tag{2.1}$$

where, as usual, ε, is a small positive perturbation parameter, and b, c, and f are $C^1(\omega_x)$-functions, with b and c satisfying

$$b(x) \geq \beta > 0, \quad c(x) \geq 0 \ \text{ for } x \in \bar{\omega}_x.$$

We preserve the notation of [115], i.e., $b(x)$ and $c(x)$ are the coefficients of the convection and reaction terms respectively. It is well known, see [50, 71] for instance, that (2.1) has a unique solution u in $C^3(\omega_x)$, and, in general, has an exponential boundary layer near $x = 0$.

We consider a finite difference discretization of (2.1), where the standard central 3-point finite difference is used for u'', and 2-point upwind scheme is for u', on the Shishkin mesh with N subintervals. It is shown in [90] that the condition number, in the maximum norm, of the matrix of the resulting system is of magnitude $\mathcal{O}(\varepsilon^{-1}(N/\ln N)^2)$. Since this is unsatisfactory when $\varepsilon \to 0$, a simple preconditioning is proposed in the same paper. This behavior of the condition number is contrasted in [90] to that of the singularly perturbed reaction-diffusion problem, which can be described as (2.1) with $b \equiv 0$ and $c > 0$ on $\bar{\omega}_x$. When the reaction-diffusion problem is discretized using the standard central scheme on the Shishkin mesh, there is no need for preconditioning because the condition number behaves like $\mathcal{O}((N/\ln N)^2)$. On the other hand, if a symmetrised finite difference schemes is used, the resulting linear system is ill-conditioned. These will be discussed in detail in Chapter 4.

The standard techniques used to prove ε-uniform convergence of numerical methods for convection-diffusion problems are quite different from those applied to reaction-diffusion problems. For example, for the reaction-diffusion problems, one can prove that a finite difference discretization yields ε-uniform convergence by using the following principle, which originated from non-perturbed problems:

Principle 2.1. ε-uniform stability and ε-uniform consistency, both in the maximum norm, imply ε-uniform pointwise convergence.

The ε-uniform convergence proofs which are based on the above principle for reaction-diffusion problems can be found in [77, Chapter 6], and [112] where a one dimensional semilinear reaction-diffusion is studied on the generalized Bakhvalov type meshes. In Section 2.5, we present a uniform convergence proof based on Principle 2.1 for a one-dimensional reaction-diffusion problem discretized on a Shishkin mesh.

Principle 2.1 does not, however, work for convection-diffusion problems (2.1) because ε-uniform pointwise consistency is not present, although it is easy to show that the upwind scheme is ε-uniformly stable in the maximum norm. For these problems, ε-uniform consistency can be proved in a discrete L^1-norm, and so the proofs based on the stability consistency principle have to rely on some kind of hybrid stability inequality [5, 61, 67], an approach that typically involves the use of discrete Green's functions.

Our main result is that we show that essentially the same preconditioning, which eliminates the difference in the condition numbers of simple finite difference discretiza-

tions for the convection-diffusion and reaction-diffusion problems, can also be used to eliminate the difference in the application of Principle 2.1 to the proofs of ε-uniform pointwise convergence for these two problem types. We do this by appropriately modifying the method from [90]. In other words, a suitable preconditioning technique enables the use of Principle 2.1 for the convection-diffusion problem. Using this approach, we prove almost first-order pointwise ε-uniform convergence for the upwind scheme discretizing the problem (2.1) on the Shishkin mesh. This result, however, is not the main contribution of this chapter, because the same has already been proved elsewhere (see the above references). However, this conceptually simple proof points out that there is a connection between preconditioning and ε-uniform pointwise convergence for convection-diffusion problems.

The rest of the chapter is organized as follows. We give the properties of the continuous solution to (2.1) in Section 2.2. Then, in Section 2.3, we introduce the simple upwind scheme on the Shishkin mesh and discuss the preconditioning of the discrete problem. Section 2.4 provides the proof of ε-uniform pointwise convergence. Then, in Section 2.5, we demonstrate how Principle 2.1 is used to prove ε-uniform convergence for the reaction-diffusion in one dimension. Finally, some concluding remarks are given in Section 2.6.

2.2 Solution decomposition

The solution, u, of (2.1) can be decomposed into the regular and boundary layer parts. We present here a version of such a decomposition taken from [61, Theorem 3.48]:

$$u(x) = s(x) + y(x), \tag{2.2}$$

$$|s^{(k)}(x)| \le C\left(1 + \varepsilon^{2-k}\right), \quad |y^{(k)}(x)| \le C\varepsilon^{-k} e^{-\beta x/\varepsilon}, \tag{2.3}$$

$$x \in \bar{\omega}_x, \quad k = 0, 1, 2, 3.$$

Details of the construction are given in [61, §3.4.1.2]. The regular component, s, satisfies $\mathcal{L}s = f, x \in \omega_x$, while the layer component, y, solves the problem

$$\mathcal{L}y(x) = 0, \quad x \in \omega_x, \quad y(0) = -s(0), \quad y(1) = 0, \tag{2.4}$$

We shall use this fact later on in the proof of Lemma 2.2.

2.3 The discrete problem and conditioning

Recall from Section 1.6 that $\omega_x^N := \{0 = x_0 < x_1 < \cdots < x_N = 1\}$. We discretize the problem (2.1) on ω_x^N using the upwind finite difference scheme:

$$\begin{aligned} U_0^N &= 0, \\ \mathcal{L}^N U_i^N := -\varepsilon D'' U_i^N - b_i D' U_i^N + c_i U_i^N &= f_i, \quad i = 1, 2, \ldots, N-1, \\ U_N^N &= 0, \end{aligned} \tag{2.5}$$

where

$$D' W_i^N = \frac{W_{i+1}^N - W_i^N}{h_{i+1}},$$

and

$$D'' W_i^N = \frac{1}{\bar{h}_i} \left(\frac{W_{i+1}^N - W_i^N}{h_{i+1}} - \frac{W_i^N - W_{i-1}^N}{h_i} \right).$$

The linear system (2.5) can be written down in matrix form,

$$A_N U^N = \hat{f}^N, \tag{2.6}$$

where $A_N = [a_{i,j}]$ is a tridiagonal matrix with $a_{0,0} = 1$ and $a_{N,N} = 1$ being the only nonzero elements in the 0th and Nth rows, respectively, and where $\hat{f}^N = [0, f_1, \ldots, f_{N-1}, 0]^T$. (Note that in this chapter, for convenience, we index the entries of vectors and matrices from 0).

It is easy to see that A_N is an L-matrix, i.e., $a_{i,i} > 0$ and $a_{i,j} \leq 0$ if $i \neq j$, for all $i, j = 0, 1, \ldots, N$. The matrix A_N is also inverse monotone, which means that it is non-singular and that $A_N^{-1} \geq 0$ (inequalities involving matrices and vectors should be understood component-wise), and therefore an M-matrix (inverse monotone L-matrix). This can be proved using the following M-criterion, see, e.g., [96, Theorem 2.7].

Theorem 2.1. *Let A be an L-matrix and let there exist a vector w such that $w > 0$ and $Aw \geq \gamma$ for some positive constant γ. A is then an M-matrix and it holds that $\|A^{-1}\| \leq \gamma^{-1}\|w\|$.*

To see that A_N is an M-matrix, just set $w_i = 2 - x_i$, $i = 0, 1, \ldots, N$, in Theorem 2.1 to get that $A_N w \geq \min\{1, \beta\}$. This also implies that the discrete problem (2.6) is stable uniformly in ε,

$$\|A_N^{-1}\| \leq \frac{2}{\min\{1, \beta\}} \leq C. \tag{2.7}$$

Of course, the system (2.6) has a unique solution U^N.

From this point on, we take ω_x^N to be the standard Shishkin mesh for these problems. However, our results equally hold true for the slightly generalized Shishkin mesh

considered in [113]. Let N be even and let $J = N/2$. While the solution of the reaction-diffusion problem (1.5) has two boundary layers at both ends of the unit interval with the layer width proportional to the square root of the diffusion parameter; for the convection-diffusion problem (2.1), the solution has only one boundary layer near $x = 0$ with the width of the layer proportional to the diffusion parameter. Hence, the Shishkin mesh is constructed as follows. Let

$$\tau_x = \min\left\{\frac{1}{2}, \frac{a\varepsilon \ln N}{\beta}\right\},$$

where a is a user-chosen parameter, and we take $a \geq 2$. The Shishkin mesh is constructed by forming a fine equidistant mesh with J mesh steps of size h in the interval $[0, \tau_x]$, and a coarse equidistant mesh with J mesh steps of size H in $[\tau_x, 1]$. We only consider the case when $\tau_x = a\varepsilon \ln N/\beta$, since N is otherwise unrealistically large. We have that

$$h = \frac{\tau_x}{J} \leq C\varepsilon \frac{\ln N}{N}, \quad \text{and} \quad H = \frac{1 - \tau_x}{J} \leq 2N^{-1},$$

and we define $\hbar = (h + H)/2$. In particular, $\bar{h}_J = \hbar$.

When the discrete problem (2.5) is formed on the Shishkin mesh, it is shown in [90] that the condition number of A_N,

$$\kappa(A_N) := \|A_N^{-1}\|\|A_N\|,$$

satisfies the following sharp estimate:

$$\kappa(A_N) \leq C\frac{N^2}{\varepsilon \ln^2 N}.$$

Therefore, the system is ill-conditioned when $\varepsilon \to 0$. This unpleasant behavior is eliminated in [90] using the preconditioning by the diagonal matrix $D := [\mathrm{diag}(A_N)]^{-1}$. When the system (2.5) is multiplied by D, the resulting matrix DA_N satisfies

$$\|DA_N\| \leq C, \quad \text{and} \quad \|(DA_N)^{-1}\| \leq C\frac{N^2}{\ln N},$$

so that

$$\kappa(DA_N) \leq C\frac{N^2}{\ln N}. \tag{2.8}$$

Note, however, that the matrix DA_N no longer satisfies $\|(DA_N)^{-1}\| \leq C$, thus the original stability estimate $\|A_N^{-1}\| \leq C$ in (2.7) is not preserved. Below we modify the preconditioning by a diagonal matrix so that the same estimate as in (2.8) holds true, while the stability of type (2.7) is retained.

Let $M = \mathrm{diag}\,(m_0, m_1, \ldots, m_N)$ be a diagonal matrix with the entries

$$m_i = \begin{cases} 1, & i = 0, \\ \dfrac{h}{H}, & 1 \le i \le J-1, \\ 1, & J \le i \le N. \end{cases}$$

When the system (2.6) is multiplied by M, this is equivalent to multiplying the equations 1, 2, ..., $J-1$, of the system (2.5) by h/H. The modified system is

$$\tilde{A}_N U^N = M \hat{f}^N, \tag{2.9}$$

where $\tilde{A}_N = MA_N$. Let the entries of $\tilde{A}_N$ be denoted by $\tilde{a}_{i,j}$, the nonzero ones being

$$l_i := \tilde{a}_{i,i-1} = \begin{cases} -\dfrac{\varepsilon}{hH}, & 1 \le i \le J-1, \\ -\dfrac{\varepsilon}{h\hbar}, & i = J, \\ -\dfrac{\varepsilon}{H^2}, & J+1 \le i \le N-1, \end{cases}$$

$$r_i := \tilde{a}_{i,i+1} = \begin{cases} -\dfrac{\varepsilon}{hH} - \dfrac{b_i}{H}, & 1 \le i \le J-1, \\ -\dfrac{\varepsilon}{H\hbar} - \dfrac{b_i}{H}, & i = J, \\ -\dfrac{\varepsilon}{H^2} - \dfrac{b_i}{H}, & J+1 \le i \le N-1, \end{cases}$$

and

$$d_i := \tilde{a}_{ii} = \begin{cases} 1, & i = 0 \\ -l_i - r_i + \dfrac{h}{H} c_i, & 1 \le i \le J-1, \\ -l_i - r_i + c_i, & J \le i \le N-1, \\ 1, & i = N. \end{cases}$$

It is easy to see that $\tilde{A}_N$ is an L-matrix. The next lemma shows that $\tilde{A}_N$ is an M-matrix and that the modified discretization (2.9) is stable uniformly in ε.

Lemma 2.1. *The matrix $\tilde{A}_N$ of the system* (2.9) *satisfies*

$$\left\| \tilde{A}_N^{-1} \right\| \le C.$$

Proof. We construct a vector $v = [v_0, v_1, \ldots, v_N]^T$ such that

(a) $v_i \geq \delta$, $i = 0, 1, \ldots, N$, where δ is a positive constant independent of both ε and N,

(b) $v_i \leq C$, $i = 0, 1, \ldots, N$,

(c) $l_i v_{i-1} + d_i v_i + r_i v_{i+1} \geq \delta$, $i = 1, 2, \ldots, N-1$.

Then, according to Theorem 2.1,

$$\|\tilde{A}_N^{-1}\| \leq \delta^{-1}\|v\| \leq C.$$

The vector v can be constructed as follows:

$$v_i = 2 + \beta - Hi + \lambda \min\left\{(1+\rho)^{J-i}, 1\right\}, \quad \text{for} \quad i = 0, 1, \ldots, N,$$

where λ is a fixed positive constant and $\rho = \beta H/\varepsilon$. This construction is motivated by the proof of Lemma 4 in [90].

Since $N^{-1} \leq H \leq 2N^{-1}$, we see that the conditions (a) and (b) are satisfied if we show that $\lambda \leq C$. We do this next at the same time as we verify condition (c).

When $1 \leq i \leq J-1$, we have

$$\begin{aligned} l_i v_{i-1} + d_i v_i + r_i v_{i+1} &= (l_i + d_i + r_i)v_i + l_i H - r_i H \\ &= \frac{h}{H} c_i v_i - \frac{\varepsilon}{h} + \frac{\varepsilon}{h} + b_i \\ &\geq \beta. \end{aligned}$$

For $i = J$, condition (c) is verified as follows:

$$\begin{aligned} l_J v_{J-1} + d_J v_J + r_J v_{J+1} &= c_J v_J + l_J H - r_J \left(H + \frac{\lambda\rho}{1+\rho}\right) \\ &\geq -r_J \frac{\lambda\rho}{1+\rho} + (l_J - r_J)H \\ &= \left(\frac{\varepsilon}{\hbar H} + \frac{b_J}{H}\right)\frac{\lambda\rho}{1+\rho} - \frac{\varepsilon H}{h\hbar} + \frac{\varepsilon}{\hbar} + b_J \\ &\geq \frac{\varepsilon + \beta\hbar}{\hbar H} \cdot \frac{\lambda\beta H}{\varepsilon + \beta H} - \frac{\varepsilon H}{h\hbar} + \beta \\ &= \frac{1}{\hbar}\left(\frac{\varepsilon + \beta\hbar}{\varepsilon + \beta H}\beta\lambda - \frac{\varepsilon H}{h}\right) + \beta \\ &\geq \frac{1}{\hbar}\left(\frac{\beta\lambda}{2} - \frac{\varepsilon H}{h}\right) + \beta \\ &\geq \beta, \end{aligned}$$

where in the last step we choose λ so that $\lambda \leq C$ and

$$\frac{\beta\lambda}{2} \geq \frac{\varepsilon H}{h}.$$

This is possible to do because

$$\frac{\varepsilon H}{h} \leq \frac{C}{\ln N} \leq C.$$

Finally, if $J+1 \leq i \leq N-1$, we have

$$\begin{aligned} l_i v_{i-1} + d_i v_i + r_i v_{i+1} &= c_i v_i + l_i H - r_i H + l_i \left[\frac{\lambda}{(1+\rho)^{i-1-J}} - \frac{\lambda}{(1+\rho)^{i-J}} \right] \\ &\quad + r_i \left[\frac{\lambda}{(1+\rho)^{i+1-J}} - \frac{\lambda}{(1+\rho)^{i-J}} \right] \\ &\geq b_i + \frac{\rho(1+\rho)l_i - \rho r_i}{(1+\rho)^{i+1-J}}\lambda \\ &\geq \beta + \frac{(l_i - r_i + l_i\rho)\rho}{(1+\rho)^{i+1-J}}\lambda \\ &= \beta + \left(\frac{b_i}{H} - \frac{\beta}{H} \right) \lambda\rho(1+\rho)^{J-i-1} \\ &\geq \beta. \end{aligned}$$

□

By examining the elements of the matrix $\tilde{A}_N$, we see that

$$\|\tilde{A}_N\| \leq C\frac{N^2}{\ln N}.$$

When we combine this with Lemma 2.1, we get the following result.

Theorem 2.2. *The matrix $\tilde{A}_N$ of the system (2.9) satisfies*

$$\kappa(\tilde{A}_N) \leq C\frac{N^2}{\ln N}.$$

To conclude this section, we reiterate that both discrete systems (2.6) and (2.9) are stable uniformly in ε. Their corresponding stability inequalities are

$$\|W^N\| \leq \|A_N^{-1}\|\|A_N W^N\|, \tag{2.10}$$

and

$$\|W^N\| \leq \|\tilde{A}_N^{-1}\|\|\tilde{A}_N W^N\|, \tag{2.11}$$

where both $\|A_N^{-1}\|$ and $\|\tilde{A}_N^{-1}\|$ are bounded from above by a constant independent of ε.

2.4 Uniform convergence for the convection-diffusion problem

Let σ_i, $i = 1, 2, \ldots, N-1$, be the consistency error of the finite difference operator $\mathcal{L}^N$,

$$\sigma_i = \sigma_i[u] := \mathcal{L}^N u_i - (\mathcal{L}u)_i, \tag{2.12}$$

that is,

$$\sigma_i = \mathcal{L}^N u_i - f_i = [A_N(u^N - U^N)]_i.$$

Convergence uniform in ε would follow from (2.10) if we could show that

$$|\sigma_i| \to 0 \text{ uniformly in } \varepsilon \text{ when } N \to \infty. \tag{2.13}$$

However, this does not hold true, as the following simple numerical experiment indicates.

Consider the test problem taken from [61, p.1],

$$-\varepsilon u'' - u' = 1, \quad x \in (0,1), \quad u(0) = u(1) = 0.$$

The exact solution is easily determined. Table 2.1 clearly shows that (2.13) is not satisfied.

ε	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$	$N = 1024$
10^{-2}	1.33e+1	9.71e+0	6.42e+0	3.95e+0	2.32e+0	1.32e+0
10^{-3}	1.33e+2	9.71e+1	6.42e+1	3.95e+1	2.32e+1	1.32e+1
10^{-4}	1.33e+3	9.71e+2	6.42e+2	3.95e+2	2.32e+2	1.32e+2
10^{-5}	1.33e+4	9.71e+3	6.42e+3	3.95e+3	2.32e+3	1.32e+3
10^{-6}	1.33e+5	9.71e+4	6.42e+4	3.95e+4	2.32e+4	1.32e+4
10^{-7}	1.33e+6	9.71e+5	6.42e+5	3.95e+5	2.32e+5	1.32e+5
10^{-8}	1.33e+7	9.71e+6	6.42e+6	3.95e+6	2.32e+6	1.32e+6

Table 2.1: The maximum norm of the consistency error $\left[A_N u^N - \hat{f}^N\right]$ on a Shishkin mesh.

However, for the preconditioned system (2.9), the consistency error is

$$\tilde{\sigma}_i[u] = \begin{cases} \dfrac{h}{H}\sigma_i[u], & 1 \le i \le J-1, \\ \\ \sigma_i[u], & J \le i \le N-1, \end{cases} \tag{2.14}$$

and it tends to 0, uniformly in ε, when $N \to \infty$, as Table 2.2 indicates. We prove this in the following lemma.

ε	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$	$N=1024$
10^{-2}	1.002	0.891	0.697	0.499	0.335	0.216
10^{-3}	0.938	0.823	0.635	0.448	0.297	0.188
10^{-4}	0.932	0.817	0.629	0.443	0.293	0.186
10^{-5}	0.932	0.816	0.629	0.443	0.293	0.186
10^{-6}	0.932	0.816	0.629	0.443	0.293	0.185
10^{-7}	0.932	0.816	0.629	0.443	0.293	0.185
10^{-8}	0.932	0.816	0.629	0.443	0.293	0.185

Table 2.2: The maximum norm of the preconditioned consistency error $\left[\tilde{A}_N u^N - M\hat{f}^N\right]$ on a Shishkin mesh.

Lemma 2.2. *The following estimate holds true for all $i = 1, 2, \ldots, N-1$:*

$$|\tilde{\sigma}_i[u]| \leq CN^{-1}\ln^2 N.$$

Proof. By a Taylor expansion

$$u(x\pm h_i) = u(x)\pm h_i u'(x)+h_i^2\frac{u''(x)}{2}\pm h_i^3\frac{u'''(x)}{6}+\int_x^{x\pm h_i}\left(u'''(\xi)-u'''(x)\right)\frac{(x\pm h_i-\xi)^2}{2}d\xi,$$

we have that

$$|\sigma_i[u]| \leq Ch_{i+1}(\varepsilon\|u'''\|_i + \|u''\|_i), \tag{2.15}$$

where $\|g\|_i := \max_{x_{i-1}\leq x\leq x_{i+1}}|g(x)|$ for any $C(\omega_x)$-function g, (see, e.g. [77, Lemma 1]). We use the decomposition (2.2) to get

$$\tilde{\sigma}_i[u] = \tilde{\sigma}_i[s] + \tilde{\sigma}_i[y].$$

Then (2.15) and the derivative estimates of s, given in (2.3), immediately imply that

$$|\tilde{\sigma}_i[s]| \leq CN^{-1}.$$

It remains to be proved that

$$|\tilde{\sigma}_i[y]| \leq CN^{-1}\ln^2 N. \tag{2.16}$$

For $1 \leq i \leq J-1$, we use (2.15) again, together with the derivative estimates of y, see (2.3):

$$|\tilde{\sigma}_i(y)| \leq C\frac{h^2}{H}(\varepsilon\|y'''\|_i + \|y''\|_i) \leq C\frac{h^2}{H}\varepsilon^{-2} \leq CN^{-1}\ln^2 N.$$

Therefore, (2.16) is proved in this case.

When $J+2 \leq i \leq N-1$, (2.15) and (2.3) give

$$\begin{aligned}|\tilde{\sigma}_i[y]| &\leq CH(\varepsilon\|y'''\|_i + \|y''\|_i) \leq CH\varepsilon^{-2}e^{-\beta x_{i-1}/\varepsilon} \leq CH\varepsilon^{-2}e^{-\beta(\tau_x+H)/\varepsilon}\\ &\leq CN\left(H\varepsilon^{-1}\right)^2 e^{-\beta H/\varepsilon}e^{-\beta\tau_x/\varepsilon}.\end{aligned}$$

The estimate (2.16) follows from here because

$$\left(H\varepsilon^{-1}\right)^2 e^{-\beta H/\varepsilon} \leq C,$$

and because the definition of τ_x and $a \geq 2$ imply that

$$e^{-\beta\tau_x/\varepsilon} \leq N^{-2}.$$

We finally prove (2.16) for $i = J, J+1$. In this case, we use similar arguments to those in [114, Lemma 5], and see also [96, Remark 2.98]. We also use the fact that $\mathcal{L}y = 0$ to work with

$$|\tilde{\sigma}_i[y]| = |\sigma_i[y]| \leq P_i + Q_i + c_i|y_i|,$$

where

$$P_i = \varepsilon|D''y_i|, \quad \text{and} \quad Q_i = b_i|D'y_i|.$$

We immediately have that

$$c_i|y_i| \leq Ce^{-\beta x_i/\varepsilon} \leq Ce^{-\beta\tau_x/\varepsilon} \leq CN^{-2}.$$

For P_i, since

$$|D''y_i| = \left|\hbar_i^{-1}\left(\frac{y_{i+1}-y_i}{h_{i+1}} - \frac{y_i - y_{i-1}}{h_i}\right)\right| \leq 2\hbar_i^{-1}\|y'\|_i,$$

so

$$P_i \leq \varepsilon\left(2\hbar_i^{-1}\|y'\|_i\right) \leq CNe^{-\beta(\tau_x-h)/\varepsilon} \leq CN^{-1}.$$

Analogously, we have

$$|D'y_i| = \left|h_{i+1}^{-1}(y_{i+1}-y_i)\right| \leq 2h_{i+1}^{-1}\|y\|_{[x_i,x_{i+1}]},$$

then

$$Q_i \leq 2h_{i+1}^{-1}\|y\|_{[x_i,x_{i+1}]} \leq CH^{-1}\|y\|_i \leq CNe^{-\beta(\tau_x-h)/\varepsilon} \leq CN^{-1}.$$

This completes the proof. □

Note that when the above proof technique is applied to the unpreconditioned consistency error σ_i as defined in (2.12), this quantity cannot be estimated uniformly in ε. We can only get that

$$|\sigma_i| \leq C\frac{\ln N}{\varepsilon N}.$$

It is because we multiply equations 1, 2,..., $J-1$ of the system (2.5) by h/H to give the preconditioned truncation error (2.14) that we get the extra ε-factor needed for the ε-uniform consistency on the fine part of the mesh.

When Lemmas 2.1 and 2.2 are combined, in the application of Principle 2.1, we obtain the following result.

Theorem 2.3. *The solution U^N of the discrete problem* (2.6) *on the Shishkin mesh described in Section 2.3 satisfies*

$$\left\|U^N - u\right\| \leq CN^{-1}\ln^2 N,$$

where u is the solution of the continuous problem (2.1).

Remark 2.1. *The result of Theorem 2.3 is the same as in [77, Theorem 8.4], proved by the barrier-function technique for the case $c \equiv 0$, but with the mesh parameter $a > 1$. A finer, but more complicated, analysis in [33, Theorem 3.6] improves the above estimate to*

$$\left\|U^N - u^N\right\| \leq CN^{-1}\ln N, \tag{2.17}$$

with $a \geq 1$ and still for $c \equiv 0$. The same result as in (2.17) *is proved in [61, Chapter 4] for the general problem* (2.1)*, by using a finite element approach to the discretization scheme, which is slightly different from $\mathcal{L}^N$, having h_{i+1} instead of $\bar{h}_i$ in D''. However, using the preconditioning arguments, it is not possible to remove the extra logarithmic factor.*

2.5 Uniform convergence for the reaction-diffusion problem

In this section, we outline the uniform pointwise convergence proof based on Principle 2.1 for the one-dimensional reaction-diffusion problem (1.5) discretized by a central finite difference scheme on the Shishkin mesh defined in Section 1.7.1. The main result is already well established in the literature. Also, unlike the convection-diffusion case, no special preconditioning is required. However, the inclusion of this result provides a link to the remaining chapters. First, we provide some details on the solution decomposition and bounds on derivatives discussed in Section 1.5.2 (see also, e.g., [61, §3.3.1.2]).

Theorem 2.4 ([61, Theorem 3.35])**.** *Suppose $b, f \in C^q(\bar{\omega}_x)$, with q is a positive integer. Then* (1.5) *possesses a unique solution $u \in C^{q+2}(\bar{\omega}_x)$. It can be decomposed as*

$$u = v + w_0 + w_1, \tag{2.18}$$

with

$$\mathcal{L}v = f, \quad \mathcal{L}w_0 = 0, \quad \textit{and} \quad \mathcal{L}w_1 = 0 \quad \textit{in} \quad \omega_x.$$

The regular part, v, satisfies

$$\|v^{(m)}\| \leq C(1 + \varepsilon^{q-m}), \tag{2.19}$$

while for the layer parts, w_0 and w_1, we have

$$|w_0^{(m)}(x)| \le C\varepsilon^{-m}e^{-\beta x/\varepsilon}, \quad \textit{and} \quad |w_1^{(m)}(x)| \le C\varepsilon^{-m}e^{-\beta(1-x)/\varepsilon}, \tag{2.20}$$

for $x \in \bar{\omega}_x$, and $m = 0, 1, \ldots, q$.

On the Shishkin mesh constructed in Section 1.7.1 for the one-dimensional reaction-diffusion problem (1.5), we discretize this problem by the following finite difference scheme

$$\begin{aligned} U_0^N &= 0, \\ \mathcal{L}^N U_i^N := -\varepsilon^2 D^- D^+ U_i^N + b_i U_i^N &= f_i, \quad i = 1, 2, \ldots, N-1, \\ U_N^N &= 0. \end{aligned} \tag{2.21}$$

It is very easy to see that the matrix of the linear system (2.21) is an M-matrix. Thus, the operator $\mathcal{L}^N$ is stable uniformly in ε. We now prove that the consistency error is convergent uniformly in ε.

Lemma 2.3. *Let $\sigma_i[u] := \mathcal{L}^N u_i - (\mathcal{L}u)_i$, $i = 1, 2, \ldots, N-1$, be the consistency error of the operator $\mathcal{L}^N$ defined in* (2.21) *on the Shishkin mesh described in Section 1.7.1. Then*

$$|\sigma_i[u]| \le \begin{cases} C(\varepsilon^2 N^{-1} + N^{-2}\ln^2 N), & x_i = \{\tau_x, 1-\tau_x\}, \\ CN^{-2}\ln^2 N, & \textit{otherwise.} \end{cases} \tag{2.22}$$

Proof. The consistency error is split in the same manner as (2.18), i.e., into a regular part, v, and layer parts, w_0 and w_1:

$$|\sigma_i[u]| \le |\sigma_i[v]| + |\sigma_i[w_0]| + |\sigma_i[w_1]|.$$

First for the regular part, when $x_i < \tau_x$, or $x_i > 1 - \tau_x$, by a Taylor expansion and derivative estimate (2.19), we have that

$$|\sigma_i[v]| \le C\varepsilon^2 h^2 \|v^{(4)}\|_i \le CN^{-2}\ln^2 N.$$

For $\tau_x < x_i < 1 - \tau_x$, we get

$$|\sigma_i[v]| \le C\varepsilon^2 H^2 \|v^{(4)}\|_i \le CN^{-2}.$$

For $x_i = \{\tau_x, 1-\tau_x\}$, we have

$$|\sigma_i[v]| \le C\varepsilon^2 (h+H) \|v^{(3)}\|_i \le C\varepsilon^2 N^{-1}.$$

Next for the layer part w_0, we use the derivative estimate (2.20) together with the Taylor's expansion to show that $|\sigma_i[w_0]| \le CN^{-2}\ln^2 N$. To this end, we first consider $x_i < \tau_x$,

$$|\sigma_i[w_0]| \le C\varepsilon^2 h^2 \|w_0^{(4)}\|_i \le CN^{-2}\ln^2 N e^{-\beta x_{i-1}/\varepsilon} \le CN^{-2}\ln^2 N.$$

For $x_i \geq \tau_x$, we use the fact that $|\sigma_i[w_0]| \leq C\varepsilon^2 \|w_0^{(2)}\|_i$ to get

$$|\sigma_i[w_0]| \leq C\varepsilon^2 \|w_0^{(2)}\|_i \leq Ce^{-\beta(\tau_x - h)/\varepsilon} \leq Ce^{-\beta\tau_x/\varepsilon} e^{\beta h/\varepsilon} \leq CN^{-2}\ln^2 N.$$

An analogous argument can be used to show that

$$|\sigma_i[w_1]| \leq CN^{-2}\ln^2 N, \quad x_i \in \omega_x^N.$$

Combining the above results, we get

$$|\sigma_i[u]| \leq \begin{cases} C(\varepsilon^2 N^{-1} + N^{-2}\ln^2 N), & x_i = \{\tau_x, 1-\tau_x\}, \\ CN^{-2}\ln^2 N, & \text{otherwise.} \end{cases}$$

This completes the proof. □

Invoking Principle 2.1, we obtain the main theorem of this section.

Theorem 2.5. *The solution U^N of the discrete problem* (2.21) *on the Shishkin mesh described in Section 1.7.1 satisfies*

$$\left\|U^N - u\right\| \leq \begin{cases} C(\varepsilon^2 N^{-1} + N^{-2}\ln^2 N), & x_i = \{\tau_x, 1-\tau_x\}, \\ CN^{-2}\ln^2 N, & \textit{otherwise}, \end{cases}$$

where u is the solution of the continuous problem (1.5).

Despite ε-dependency in the bound above, it can be interpreted as uniform convergence in the sense discussed in Section 1.1. The proof presented here is straightforward and simple. It only requires the standard truncation error estimate, together with the bounds on derivatives, and the use of the classical stability-consistency principle. Furthermore, in practice, we usually have $\varepsilon^2 \leq N^{-1}$, and in this case, we recover the usual almost second-order convergence as in (1.20).

2.6 Concluding remarks

Since $W^N = (A_N^{-1}M^{-1})(MA_NW^N)$, the stability inequality (2.11) can be represented as

$$\|W^N\| \leq \|A_N^{-1}\|'_M \|A_N W^N\|_M, \tag{2.23}$$

where for a matrix B, $\|B\|'_M = \|BM^{-1}\|$, and $\|W^N\|_M = \|MW^N\|$. Note that the matrix norm $\|\cdot\|'_M$ is not induced by the vector norm $\|\cdot\|_M$ (which is why we denote them differently), but the two norms are consistent in the sense that

$$\|BW^N\| \leq \|B\|'_M \|W^N\|_M.$$

The inequality (2.23) is a stability inequality of hybrid nature, having different vector norms on the two sides. However, the vector norm $\|\cdot\|_M$ is still essentially a maximum norm and this is completely different from the hybrid stability inequalities used in [37, 5, 67, 61], which have a discrete L^1-norm on the right-hand side. Moreover, these hybrid stability inequalities are derived by using the discrete Green's function, which we do not use here.

In conclusion, although the method presented here gives a slightly weaker result in some cases, it provides a straightforward proof, based on a simple principle, of ε-uniform pointwise convergence for the solution of the standard upwind scheme discretizing the singularly perturbed convection-diffusion problem (2.1). It is even more interesting that the proof is enabled by the preconditioning of the system arising from the discretization. Whether this can be used as a general approach when proving ε-uniform pointwise convergence for other types of singular perturbation problems, including multi-dimensional ones, remains to be seen, but the generalization to the semilinear problem of type (2.1) (with $c = c(x, u)$, $c_u \geq 0$) is straightforward.

Chapter 3

Direct solvers and their limitations

In this chapter we consider the solution of large linear systems of equations that arise when two-dimensional singularly perturbed reaction-diffusion differential equations are discretized by a standard central finite difference method. The system matrices are symmetric positive definite. The direct solvers of choice for such systems are based on Cholesky factorizations. However, as observed in [72], these solvers may exhibit poor performance for singularly perturbed problems. We provide a careful analysis of the distribution of entries in the Cholesky factors based on their magnitude that explains this phenomenon, and give bounds on the ranges of the perturbation and discretization parameters where poor performance is to be expected. Numerical experiments supporting the analysis are also reported.

The material in this chapter has been accepted for publication as [83]: Thái Anh Nhan and Niall Madden, *Cholesky factorization of linear systems coming from finite difference approximations of singularly perturbed problems*, Proceedings of BAIL 2014–Boundary and Interior Layers–Computational and Asymptotic Methods, Lecture Notes in Computational Science and Engineering, Springer, Berlin, 2015.

3.1 Introduction

The numerical solution of the two-dimensional reaction-diffusion problem (1.7) is analyzed in this chapter. For the sake of completeness, we briefly recall the model problem:

$$-\varepsilon^2\Delta u + b(x,y)u = f(x,y), \quad \Omega = (0,1)^2, \quad u(\partial\Omega) = g(x,y), \tag{3.1}$$

where the "perturbation parameter", ε, is a small and positive, and the functions g, b and f are given, with $b(x,y) \geq \beta^2 > 0$ and $\beta > 0$.

Using the same notation as in Section 1.6, for a mesh grid $\Omega^{N,N}$ on the unit square and the local mesh widths $h_i = x_i - x_{i-1}$, $k_j = y_j - y_{j-1}$, $\bar{h}_i = (x_{i+1} - x_{i-1})/2$, and $\bar{k}_j = (y_{j+1} - y_{j-1})/2$, the symmetrised 5-point second-order central difference operator is

$$\Delta^N := \begin{bmatrix} & \dfrac{\bar{h}_i}{k_{j+1}} & \\ \dfrac{\bar{k}_j}{h_i} & -\left(\bar{k}_j\left(\dfrac{1}{h_i}+\dfrac{1}{h_{i+1}}\right)+\bar{h}_i\left(\dfrac{1}{k_j}+\dfrac{1}{k_{j+1}}\right)\right) & \dfrac{\bar{k}_j}{h_{i+1}} \\ & \dfrac{\bar{h}_i}{k_j} & \end{bmatrix}. \tag{3.2}$$

The resulting linear system can be written as

$$AU^N = f^N, \tag{3.3}$$

where A is a banded, symmetric positive definite $(N-1)^2 \times (N-1)^2$ matrix whose entries are defined by

$$A = \left(-\varepsilon^2\Delta^N + \bar{h}_i\bar{k}_j b(x_i, y_j)\right). \tag{3.4}$$

The direct solvers of choice for the symmetric positive definite system (3.3) are variants on Cholesky factorization. This is based on the idea that there exists a unique lower-triangular matrix L (the "Cholesky factor") such that $A = LL^T$ (see, e.g., [42, Theorem 4.25]). One of the advantages of Cholesky factorization is that the computational cost required is half that of other direct solvers, such as Gaussian elimination and LU factorization, see, e.g., [26, §2.7] and also [42, Chapter 4]. That is, because Cholesky factorization only has to compute and store the lower triangular matrix L, instead of both L and U as in the LU factorization process, it is twice as fast as its alternatives.

The goal of this chapter is to investigate Cholesky factorization of matrices in which the magnitudes of diagonal entries are dominant compared to the off-diagonal entries. These types of matrices frequently arise from the discretization of differential equations in which reaction dominates diffusion, such as (3.1) when the perturbation parameter, ε, is small. It is observed by MacLachlan and Madden in [72], the Cholesky factors of the matrix A defined in (3.3) have many entries that are extremely small in magnitude. Those authors also observe the exponential decay in successive fill-in entries during the factorization process. For large N, this may produce *subnormal* and *underflow-zero* numbers (concepts explained in Section 3.3.1).

To demonstrate this phenomenon, let us consider the Cholesky factor L of the matrix A in (3.4) with $b \equiv 1$, and taking a uniform mesh with $N = 128$ intervals in each direction. The matrix L contains nonzero entries only between the N^{th} subdiagonal and the main diagonal. In Figure 3.1, we plot the absolute value of largest entry of a given diagonal of L. When $\varepsilon = 1$ (on the left), we observe a gradual decay in the magnitude of fill-in entries. By contrast, when $\varepsilon = 10^{-6}$ (on the right), it is easy to

see a rapid decay in these values. Moreover, there are entries with magnitudes less than `realmin` (the smallest normalized positive number can be represented by IEEE standard double precision). Note that, in practice, ε can be usually very small and discretization parameter N can be large. Thus, there are many fill-in entries, the magnitude of which is less than `realmin`. The operations on subnormal numbers have a considerable impact on the computational speed of floating-point calculations (see Section 3.3.1). Thus, the amount of time required to solve these linear systems depends badly on the perturbation parameter. As a result, it is important to understand the propagation of subnormal and underflow-zero numbers in the context of singularly perturbed differential equations.

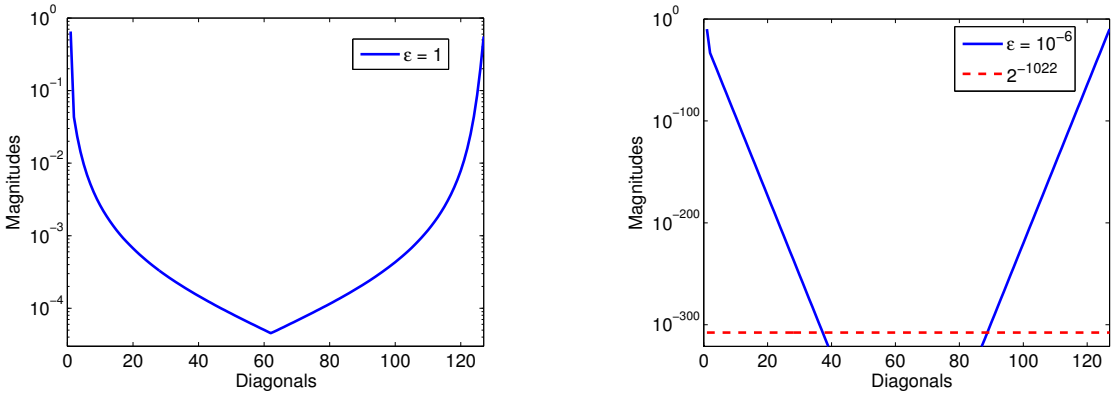

Figure 3.1: Semi-log plot of maximum absolute values of entries on diagonals of L with $\varepsilon = 1, N = 2^7$ (left), and $\varepsilon = 10^{-6}, N = 2^7$ (right).

To demonstrate the effect of the presence of subnormal numbers on computational efficiency, in Table 3.1 we show the time, in seconds, taken to compute the Cholesky factorization of A in (3.3) with a uniform mesh, and $N = 512$, on a single core of AMD Opteron 2427, 2200 MHz processor, using CHOLMOD (*supernodal sparse Cholesky factorization and update/downdate* [12, 22]); with "natural order", i.e., without a fill reducing ordering. Our programs that generate the results in Table 3.1, and others throughout this chapter, were coded in C and compiled using `gcc` version 4.7.1 with all optimizations enabled. Observe in Table 3.1 that the time-to-factorization increases from 52 seconds when ε is large, to nearly 500 seconds when $\varepsilon = 10^{-3}$, when over 1% of the entries are in the subnormal range. When ε is smaller again, the number of nonzero entries in L is further reduced, due to underflow, and so the execution time decreases as well.

We emphasize that this phenomenon is not due to the implementation of the solver. For example, the degradation of performance with respect to ε is also observed when the LU factorizations of A are computed using the Unsymmetric MultiFrontal method (UMFPACK) [23, 24], MA57 [29], and built-in MATLAB routines. Furthermore, we

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Time (s)	52.587	52.633	496.887	175.783	74.547	45.773
Nonzeros in L	133,433,341	133,433,341	128,986,606	56,259,631	33,346,351	23,632,381
Subnormals in L	0	0	1,873,840	2,399,040	1,360,170	948,600
Underflow-zeros	0	0	4,446,735	77,173,710	100,086,990	109,800,960

Table 3.1: Time taken (in seconds) to compute the Cholesky factor, L, of A in (3.3) on a uniform mesh with $N = 2^9$. The number of nonzeros, subnormals, and underflow-zeros in L are also shown.

have also observed this effect with our own implementation of Cholesky factorization based on Algorithm 3.1 below.

Motivated by this, in this chapter, we aim to give an analysis that fully explains the observations of Figure 3.1 and Table 3.1, and that can also be exploited in other solver strategies. We derive expressions, in terms of N and ε, for the magnitude of entries of L as determined by their location. Ultimately, we are interested in the analysis of systems that arise from the numerical solution of (3.1) on appropriate boundary layer-adapted meshes. Away from the boundary, such meshes are usually uniform. Therefore, we begin in Section 3.2 with studying a uniform mesh discretization, in the setting of exact arithmetic, which provides mathematical justification for observations in Figure 3.1. In Section 3.3.2, this analysis is used to quantify to number of entries in the Cholesky factors of a given magnitude. As an application of this, we show how to determine the number of subnormal numbers that will occur in L in a floating-point setting, and also determine a lower bound for ε for which the factors are free of subnormal numbers. Finally, the Cholesky factorization on a boundary layer-adapted mesh is discussed in Section 3.4, and our conclusions are summarized in Section 3.5.

3.2 Analysis of Cholesky factorization on a uniform mesh

We consider the discretization (3.2) of the model problem (3.1) on a uniform mesh with N intervals on each direction. The equally spaced mesh width is denoted by $h = N^{-1}$. We shall always assume that $\varepsilon \ll h$, which is typical for a singularly perturbed problem. More precisely, we shall assume that

$$\delta = \varepsilon/h \leq 0.1. \tag{3.5}$$

(See also remarks in Section 3.5). Then the system matrix in (3.3) can be written as the following 5-point stencil

$$A = \begin{bmatrix} & -\varepsilon^2 & \\ -\varepsilon^2 & 4\varepsilon^2 + h^2 b(x_i, y_j) & -\varepsilon^2 \\ & -\varepsilon^2 & \end{bmatrix} = \begin{bmatrix} & -\varepsilon^2 & \\ -\varepsilon^2 & \mathcal{O}(h^2) & -\varepsilon^2 \\ & -\varepsilon^2 & \end{bmatrix}, \tag{3.6}$$

since $(4\varepsilon^2 + h^2 b(x_i, y_j)) = \mathcal{O}(h^2)$.

Algorithm 3.1 presents a version of Cholesky factorization which was adapted from [42, page 143]. It computes a lower triangular matrix L such that $A = LL^T$ where A is an $n \times n$ real symmetric positive definite matrix. We will follow MATLAB notation by denoting $A = [a(i,j)]$ and $L = [l(i,j)]$.

Algorithm 3.1 Cholesky factorization:

for $j = 1 : n$
  if $j = 1$
    for $i = j : n$
      $l(i,j) = \dfrac{a(i,j)}{\sqrt{a(j,j)}}$
    end
  elseif $(j > 1)$
    for $i = j : n$
      $l(i,j) = \dfrac{a(i,j) - \sum_{k=1}^{j-1} l(i,k)l(j,k)}{\sqrt{a(j,j)}}$
    end
  end
end

We set $m = N - 1$, so A is a sparse, banded, $m^2 \times m^2$ matrix, with a bandwidth of m, and has no more than five nonzero entries per row. Its factor, L, is far less sparse: although it has the same bandwidth as A (see, e.g., [26, Prop. 2.4]), it has $\mathcal{O}(m)$ nonzeros per row. Figure 3.2 below shows the structure of the coefficient matrix A (on the left) and Cholesky factor L (on the right) when $N = 8$. We refer to the set of nonzero entries in L that are zero in the corresponding location in A as the *fill-in*. We want to find a recursive way to express the magnitude of these fill-in entries, in terms of ε and h.

To analyze the magnitude of the fill-in entries, we borrow notation from [97, Sec. 10.3.3], and form distinct sets denoted $L^{[0]}$, $L^{[1]}, \ldots, L^{[m]}$, where all entries of L that are of the same magnitude (in a sense explained carefully below) belong to the same set. We denote by $l^{[k]}$ the magnitude of entry in $L^{[k]}$, i.e., $l(i,j) \in L^{[k]}$ if and only if $l(i,j)$ is $\mathcal{O}(l^{[k]})$. $L^{[0]}$ is used to denote the set of nonzero entries in A, and entries of L

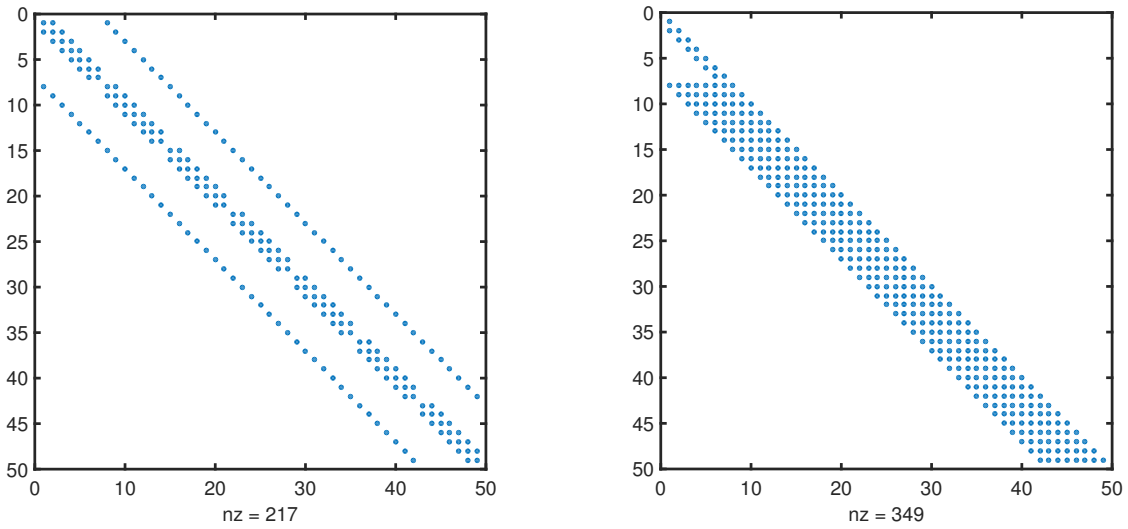

Figure 3.2: The matrix A (left), and Cholesky factor L (right) when $N = 2^3$.

that are zero are defined to belong to $L^{[\infty]}$. We shall see that all these sets are quite distinct, meaning that $l^{[k]} \gg l^{[k+1]}$ for $k \geq 1$.

In Algorithm 3.1, all the entries of L are initialized as a zero, and so belong to $L^{[\infty]}$. Suppose that $p_{i,j}$ is such that $l(i,j) \in L^{[p_{i,j}]}$, so, initially, each $p_{i,j} = \infty$. At each sweep through the algorithm, a new value of $l(i,j)$ is computed, and so $p_{i,j}$ is modified. From line 8 in Algorithm 3.1, then the $p_{i,j}$ is updated by

$$p_{i,j} = \begin{cases} 0, & \text{if } a(i,j) \neq 0, \\ \min\{p_{i,1} + p_{j,1} + 1, p_{i,2} + p_{j,2} + 1, \ldots, p_{i,j-1} + p_{j,j-1} + 1\}, & \text{otherwise.} \end{cases} \tag{3.7}$$

Remark 3.1. *From* (3.7)*, we see that the fill-in entries in the positions where the original entries of matrix A are nonzeros belong to $L^{[0]}$.*

As we shall explain in detail below, it can be determined that L has the following block structure, where, for brevity, the entries belonging to $L^{[k]}$ are denoted by $[k]$, except for the entries in $L^{[0]}$, which correspond to nonzero entries of the original matrix, and are written in terms of their magnitude:

$$L = \begin{pmatrix} M & & & & \\ P & Q & & & \\ & P & Q & & \\ & & \ddots & \ddots & \\ & & & P & Q \end{pmatrix}, \tag{3.8a}$$

$$M = \begin{pmatrix} \mathcal{O}(h) & & & & \\ \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & & & \\ & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & & \\ & & \ddots & \ddots & \\ & & & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) \end{pmatrix},$$

$$P = \begin{pmatrix} \mathcal{O}(\varepsilon^2/h) & [1] & [2] & [3] & \dots & [m-2] & [m-1] \\ & \mathcal{O}(\varepsilon^2/h) & [1] & [2] & \dots & [m-3] & [m-2] \\ & & \ddots & \ddots & \ddots & \vdots & \vdots \\ & & & \mathcal{O}(\varepsilon^2/h) & [1] & [2] & [3] \\ & & & & \mathcal{O}(\varepsilon^2/h) & [1] & [2] \\ & & & & & \mathcal{O}(\varepsilon^2/h) & [1] \\ & & & & & & \mathcal{O}(\varepsilon^2/h) \end{pmatrix}, \tag{3.8b}$$

$$Q = \begin{pmatrix} \mathcal{O}(h) & & & & & & \\ \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & & & & & \\ [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & & & & \\ [4] & [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & & & \\ \vdots & \vdots & \ddots & \ddots & \ddots & & \\ [m-1] & [m-2] & \dots & [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) & \\ [m] & [m-1] & \dots & [4] & [3] & \mathcal{O}(\varepsilon^2/h) & \mathcal{O}(h) \end{pmatrix}. \tag{3.8c}$$

We now explain why the entries of L, which are computed by column, have the structure shown in (3.8a)–(3.8c). According to Algorithm 3.1, the first column of L is computed by

$$l(i,1) = \frac{a(i,1)}{\sqrt{a(1,1)}}, \quad \text{for } i = 1, \dots, (N-1)^2,$$

which shows that there are no fill-in entries in this column. For the second column, the only fill-in entry is

$$l(m+1,2) = \frac{a(m+1,2) - l(m+1,1)l(2,1)}{\sqrt{a(2,2)}} = \frac{0 - \mathcal{O}(\varepsilon^2/h)\mathcal{O}(\varepsilon^2/h)}{\mathcal{O}(h)} = \mathcal{O}(\varepsilon^4/h^3),$$

where $l(m+1,1)$ and $l(2,1)$ belong to $L^{[0]}$, so $l(m+1,2)$ is in $L^{[1]}$. Similarly, there are two fill-ins in third column: $l(m+1,3)$ and $l(m+2,3)$. The entry $l(m+1,3)$ is computed as

$$l(m+1,3) = \frac{a(m+1,3) - \sum_{k=1}^{2} l(m+1,k)l(3,k)}{\sqrt{a(3,3)}} = \frac{-l(m+1,2)l(3,2)}{\sqrt{a(3,3)}},$$

which is $\mathcal{O}(\varepsilon^6/h^5)$; moreover, since $l(m+1,2) \in L^{[1]}$, and $l(3,2) \in L^{[0]}$, so $l(m+1,3) \in L^{[2]}$. Similarly, it is easy to see that $l(m+2,3) \in L^{[1]}$. We may now proceed by induction

to show that $l(m+1,j+1) = \mathcal{O}(\varepsilon^{2(j+1)}/h^{(2j+1)})$ belongs to $L^{[j]}$, for $1 \le j \le m-2$. Suppose $l(m+1,j) = \mathcal{O}(\varepsilon^{(2j)}/h^{(2j-1)}) \in L^{[j-1]}$. Then

$$\begin{aligned} l(m+1,j+1) &= \frac{a(m+1,j+1) - \sum_{k=1}^{j} l(m+1,k)l(j+1,k)}{\sqrt{a(j,j)}} \\ &= \frac{-l(m+1,j)l(j+1,j)}{\sqrt{a(j,j)}}, \quad (\text{since } l(j+1,k) = 0,\ \forall k \le j-1), \\ &= \frac{\mathcal{O}(\varepsilon^{(2j)}/h^{(2j-1)})\mathcal{O}(\varepsilon^2/h)}{\mathcal{O}(h)} = \mathcal{O}(\varepsilon^{(2j+2)}/h^{(2j+1)}). \end{aligned}$$

And, because $l(j+1,j) \in L^{[0]}$, we can deduce that $l(m+1,j+1) \in L^{[j]}$. The process is repeated from column 1 to column m, yielding the pattern for P shown in (3.8b). A similar process is used to show Q is as given in (3.8c). Its first fill-in entry is $l(m+3,m+1)$.

Note that $a(m+3,m+1) = l(m+1,1) = l(m+1,2) = 0$, that the magnitude of the entry in $L^{[j]}$ is $\mathcal{O}(\varepsilon^{2(j+1)}/h^{(2j+1)})$, and that the sum of two entries of the different magnitude has the same magnitude as larger one. Then

$$\begin{aligned} l(m+3,m+1) &= \frac{-\sum_{k=3}^{m} l(m+3,k)l(m+1,k)}{\sqrt{a(m+1,m+1)}} \\ &= \left[\mathcal{O}\left(\frac{\varepsilon^2}{h}\right)\mathcal{O}\left(\frac{\varepsilon^6}{h^5}\right) + \mathcal{O}\left(\frac{\varepsilon^4}{h^3}\right)\mathcal{O}\left(\frac{\varepsilon^8}{h^7}\right) + \dots \right. \\ &\qquad \left. + \mathcal{O}\left(\frac{\varepsilon^{2(m-2)}}{h^{(2(m-3)+1)}}\right)\mathcal{O}\left(\frac{\varepsilon^{2(m)}}{h^{(2(m-1)+1)}}\right)\right]\frac{1}{\mathcal{O}(h)} \\ &= \left[\mathcal{O}\left(\frac{\varepsilon^2}{h}\right)\mathcal{O}\left(\frac{\varepsilon^6}{h^5}\right)\right]\frac{1}{\mathcal{O}(h)} = \mathcal{O}\left(\frac{\varepsilon^8}{h^7}\right), \end{aligned}$$

and so $l(m+3,m+1)$ belongs to $L^{[3]}$. Proceeding inductively, as was done for P, shows that Q has the form given in (3.8c). Furthermore, the same process applies to each block of L in (3.8a). Summarizing, we have established the following result.

Theorem 3.1. *The fill-in entries of the Cholesky factor L of the matrix A defined in (3.6) are given in (3.8a)–(3.8c). Moreover, setting $\delta = \varepsilon/h$ as in (3.5), the magnitude $l^{[k]}$ is*

$$l^{[k]} = \mathcal{O}\big(\varepsilon^{2(k+1)}/h^{(2k+1)}\big) = \mathcal{O}\big(\delta^{2(k+1)}h\big) \quad \text{for} \quad k = 1,2,\dots,m. \tag{3.9}$$

Remark 3.2. *The formulation given in (3.9) tells us why the values of fill-in entries decay exponentially with respect to k. In practice, for a reaction-dominated problem the perturbation parameter is usually very small compared to h. Hence, when ε decreases and the mesh parameter N increases, the fill-in entries tend to 0 rapidly. This fact also suggests that an incomplete Cholesky factorization preconditioner would be very effective for this singularly perturbed problem.*

To conclude this section, in Figure 3.3 we plot the magnitude of the entries of the vector $L(m+1, 1:m)$ for various values of ε and $N = 128$. It clearly shows that the magnitude of the fill-in entries decays exponentially, as given in (3.9). Furthermore, when $\varepsilon = 10^{-6}$, only the first 41 entries of the vector $L(128, 1:127)$ can be plotted because the magnitudes of the other entries are less than the smallest non-normalized number in IEEE standard, and are thus flushed to zero.

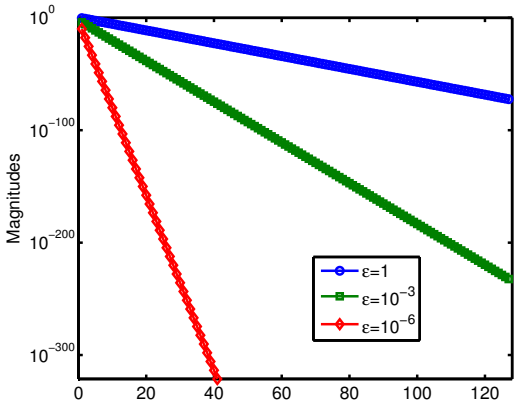

Figure 3.3: The magnitude of $L(128, 1:127)$ for various ε.

3.3 Analysis of Cholesky factorization in a floating-point setting

In many theoretical studies, it is assumed that computations, such as finding the Cholesky factors of a matrix, are done exactly. In practice, however, they are implemented on computers with finite precision. Therefore, it is important to understand the performance of such computations in a floating-point setting.

In this section, we investigate the Cholesky decomposition of the matrix (3.6) in the context of floating-point arithmetic. We briefly discuss the concept of subnormal and underflow-zero numbers in Section 3.3.1, and in Section 3.3.2 we analyze the distribution of such numbers in the Cholesky factors.

3.3.1 Subnormal and underflow-zero numbers: a short introduction

We follow closely [87, Chap. 4] and [81, Sec. 1.7] to describe subnormal and underflow-zero numbers. In the IEEE double precision format, floating point numbers are expressed as $\pm(1+f)\times 2^{Y-1023}$ where $0 \leq f < 1$ and 52 bits are used to store the *binary*

fraction f, with an implied 53^{rd} bit that is the leading 1; 11 bits are used to store the *binary exponent* Y; and the remaining bit stores the sign. So the smallest positive normalized number is when $Y = 1$, $f = 0$, i.e.,

$$N_{\min} = (1 + 0) \times 2^{-1022} \approx 2.2 \times 10^{-308}.$$

In the case when the exponent has a zero bit-string (then the implied bit is taken to be zero), but the fraction has a nonzero bit-string, the number represented is said to be *subnormal*. The smallest positive number can be stored in f is 2^{-52}. Thus, the smallest (non-normalized) positive number that can be stored is as small as $2^{-52} \times 2^{-1022} \approx 5 \times 10^{-324}$. The positive numbers less than this value are flushed to zero, and are called *underflow-zero* numbers.

Since subnormal numbers have leading zeros in the fraction, they have reduced precision compared to normal floating-point numbers. However, they allow for "graceful degradation" by allowing gradual underflow in computations involving *very* small numbers.

Although subnormal numbers are part of the official IEEE standard, most processors do not provide hardware support for arithmetic with these numbers. Instead, subnormal numbers are handled at the software level. Thus, execution time of computations that involves subnormal numbers is significantly slower than those that involves only normal numbers [55] as we have seen in Table 3.1. For an informal discussion, see Cleve Moler's blog post of July 7, 2014, [80],which describes the inclusion of subnormals in the IEEE standard, and the resulting controversy.

3.3.2 Distribution of fill-in entries in a floating-point setting

As discussed above, the time taken to compute these factorizations increases greatly if there are many subnormal numbers present. Moreover, even the underflow-zeros can also be expensive to compute, since, in the context of Cholesky factorization, they typically arise from intermediate calculations involving subnormal numbers. Therefore, in this section we use the analysis of Section 3.2, to estimate, in terms of ε and N, the number of fill-in entries in L that are of a given magnitude. From this, one can easily determine the number of subnormals and underflow-zeros that will be present.

Lemma 3.1. *Let A be the $m^2 \times m^2$ matrix in* (3.3) *where the mesh is uniform. Then the number of nonzero entries in the Cholesky factor L (i.e., $A = LL^T$)* computed using exact arithmetic *is*

$$L_{nz} = m^3 + m - 1. \tag{3.10}$$

Proof. A has bandwidth m, and so too does L ([26, Prop. 2.3]). By Algorithm 3.1, the fill-in entries only occur from row $(m+1)$. So, from row $(m+1)$, any row of L has $(m+1)$ nonzero entries and there are $m(m-1)$ such rows, plus $2m-1$ nonzero entries from top-left block M. Summing these values, we obtain (3.10). □

Let $|L^{[k]}|$ be the number of fill-in entries which belong to $L^{[k]}$. To estimate $|L^{[k]}|$, it is sufficient to evaluate the fill-in entries in the submatrices P and Q shown in (3.8). Table 3.2 describes the number of fill-in entries associated with their magnitude. Note

$L^{[k]}$	$\lvert L^{[k]}\rvert$ in P	$\lvert L^{[k]}\rvert$ in Q	$\lvert L^{[k]}\rvert$ in $[P,Q]$
$L^{[1]}$	$m-1$	0	$m-1$
$L^{[2]}$	$m-2$	0	$m-2$
$L^{[3]}$	$m-3$	$m-2$	$2m-5$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$L^{[k]}$	$m-k$	$m-k+1$	$2m-2k+1$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
$L^{[m-2]}$	2	3	5
$L^{[m-1]}$	1	2	3
$L^{[m]}$	0	1	1

Table 3.2: Number of fill-in entries in P and Q associated with their magnitude.

that there are $(m-1)$ blocks like $[P,Q]$ in L. Then, since $l^{[k]} \ll l^{[k-1]}$, and the smallest (exact) nonzero entries belong to $L^{[m]}$, we can use Table 3.2 to determine the number of entries that are at most $\mathcal{O}(l^{[p]})$, for some given p as:

$$\sum_{k=p}^{m} |L^{[k]}| = \begin{cases} (m-1)(2m-3) + (m-1)(m-2)^2 = (m-1)^3 & p=1, \\ (m-1)(m-2) + (m-1)(m-2)^2 = (m-2)(m-1)^2 & p=2, \\ (m-1)(m-p+1)^2 & p \geq 3. \end{cases}$$

These equations can be combined and summarized as follows.

Theorem 3.2. *Let A be the matrix of the form (3.6). Then, the number of fill-in entries in L that are at most $\mathcal{O}(l^{[p]})$ satisfies*

$$\sum_{k=p}^{m} |L^{[k]}| \leq (m-1)(m-p+1)^2, \quad p \geq 1. \tag{3.11}$$

Combining Theorems 3.1 and 3.2 enables us to accurately predict the total number and location of subnormal and underflow-zero entries in L, for given N and ε. For

example, recall Figure 3.1 where we took $\varepsilon = 10^{-6}$ and $N = 128$. To determine, the diagonals where entries are subnormal using Theorem 3.1, we solve

$$(\varepsilon N)^{2(k+1)} \approx 2^{-1022} N, \tag{3.12}$$

for $N = 128$ and $\varepsilon = 10^{-6}$, which yields $k \approx 38$. It clearly agrees with the observation in Figure 3.1; i.e., the maximal value of the entries on diagonals 38 and $N - 38 = 90$ are less than `realmin`. Similarly, all entries on diagonals between 40 and 88 are flushed to zero. Furthermore, for this example, by (3.11), the total number of underflow-zero and subnormal entries in L are, respectively,

$$\sum_{k=40}^{127} |L^{[k]}| = 953,694, \quad \text{and} \quad \sum_{k=38}^{39} |L^{[k]}| = \sum_{k=38}^{127} |L^{[k]}| - \sum_{k=40}^{127} |L^{[k]}| = 44,352.$$

Computer experiments verify that this is exactly what is observed in practice. Moreover, the total number of entries with magnitude less than `realmin` is 998,046 which is just less than a half of the exact nonzero entries in L, i.e., 2,048,509 (cf. Lemma 3.1). Such a predictable appearance of subnormals and underflow-zeros is important in the sense of choosing suitable linear solvers, i.e., direct or iterative ones.

More generally, we can use (3.12) to investigate ranges of N and ε for which subnormal entries occur. Since the largest possible value of k is m, a Cholesky factor will have subnormal entries if ε and N are such that $(\varepsilon N)^{2N} \lesssim 2^{-1022} N$. Rearranging, this gives that

$$\varepsilon \lesssim \frac{1}{N} \left(2^{-1022} N\right)^{1/(2N)} = 2^{-511/N} N^{(1/(2N)-1)} =: g(N). \tag{3.13}$$

The function $g(N)$ defined in (3.13) is informative because it gives the largest value of ε for a discretization with given N leading to a Cholesky factor with entries less than 2^{-1022}. For example, Figure 3.4 (on the left) shows $g(N)$ for $N \in [200, 500]$. It demonstrates that, for $\varepsilon \leq 1.05 \times 10^{-3}$ (determined numerically), subnormal entries are to be expected for some values of N (cf. Table 3.1). The line $\varepsilon = 10^{-3}$ intersects $g(N)$ at approximately $N = 263$ and $N = 484$, meaning that a discretization with $263 \leq N \leq 484$ yields entries with the magnitude less than 2^{-1022} in L for $\varepsilon = 10^{-3}$. On the right of Figure 3.4 we show that, for large N, $g(N)$ decays like $1/N$. Since we are interested in the regime where $\varepsilon \leq 1/N$, this shows that, for small ε, subnormals are to be expected for all but the smallest values of N.

As a final example, in Figure 3.5 on the left, we take $\varepsilon = 10^{-6}$, and show that subnormals will occur with N greater than 35 (compared with Figure 3.1). When $\varepsilon = 10^{-4}$, as seen in Figure 3.5 (on the right), the subnormals are expected for any $N \geq 70$. It clearly points out that for small values of ε, subnormal numbers occur in the Cholesky factorization process even for a relatively small N.

In summary, we emphasize that the presence of subnormals and underflow-zeros should be taken into account as solving the linear systems whose diagonal entries are

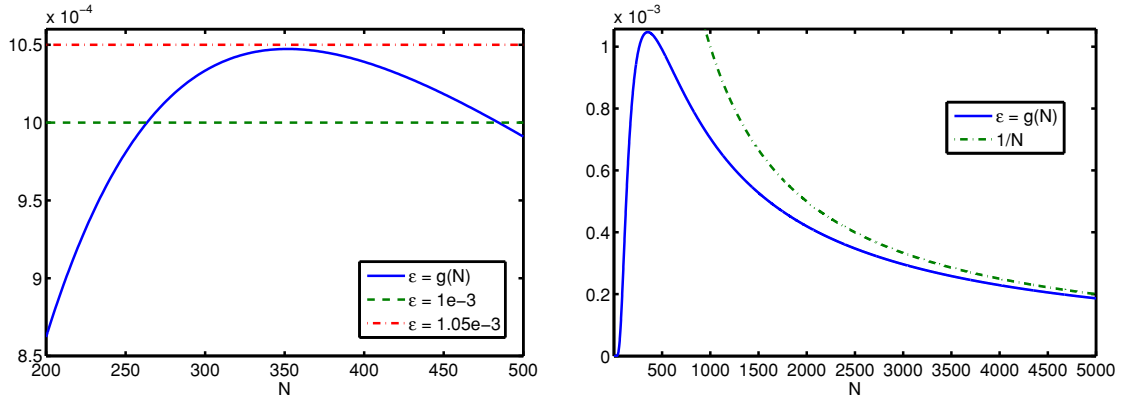

Figure 3.4: The function $g(N)$ defined in (3.13), with $N \in [200, 500]$ (left) and $N \in [1, 5000]$ (right).

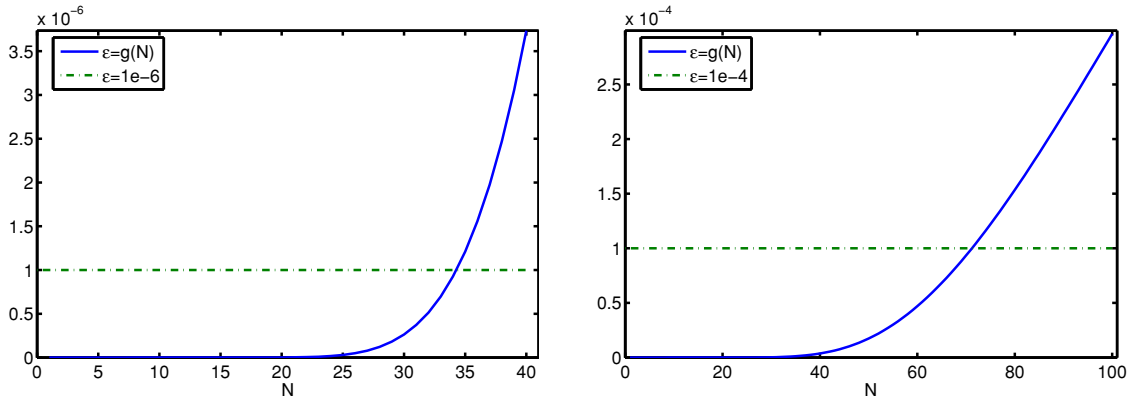

Figure 3.5: The function $g(N)$ defined in (3.13), with $N \in [1, 40]$ (left) and $N \in [1, 100]$ (right).

dominant in magnitude compared to off-diagonal entries, such as the system arising from finite difference discretization of the singularly perturbed problem (3.1) applied on a uniform mesh. Next section, we will discuss how the analysis of this section can be integrated to the case of finite difference discretization on a boundary layer-adapted mesh.

3.4 Cholesky factorization on boundary layer-adapted meshes

Our analysis so far has been for finite difference methods applied on uniform meshes. However, a scheme such as (3.4) for (3.1) is usually applied on a layer-adapted mesh, such as a Shishkin mesh. For these meshes, in the neighbourhood of the boundaries, and especially near corner layers, the local mesh width is $\mathcal{O}(\varepsilon N^{-1})$ in each direction, and so the entries of the system matrix are of the same order, and no issue with

subnormal numbers is likely to arise. However, away from layers, these fitted meshes are usually uniform, with a local mesh width of $\mathcal{O}(N^{-1})$, and so the analysis outlined above applies directly. Since roughly one quarter (depending on mesh construction) of all mesh points are located in this region, the influence on the computation is likely to be substantial.

The main complication in extending our analysis to, say, a Shishkin mesh, is in considering the "edge layers", where the mesh width may be $\mathcal{O}(\varepsilon N^{-1})$ in one coordinate direction, and $\mathcal{O}(N^{-1})$ in another. Although we have not analyzed this situation carefully, in practise it seems that the factorization behaves much like a uniform mesh. This is showed in Table 3.3 below. Comparing with Table 3.1, we see, for small ε, the number of entries flushed to zero is roughly three-quarters that of the uniform mesh case.

ε	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
Time (s)	52.580	58.213	447.533	179.540	101.507	73.250
Nonzeros in L	133,433,341	133,240,632	127,533,193	78,091,189	62,082,599	54,497,790
Subnormals in L	0	28,282	2,648,308	1,669,345	1,079,992	814,291
Underflow zeros	0	192,709	5,900,148	55,342,152	71,350,742	78,935,551

Table 3.3: Time taken (in seconds) to compute the Cholesky factor, L, of A in (3.3) on a Shishkin mesh with $N = 2^9$. The number of nonzeros, subnormals, and underflow-zeros in L are also shown.

3.5 Conclusions

The chapter addresses, in a comprehensive way, issues raised in [72] by showing how to predict the number and location of subnormal and underflow-zero entries in the Cholesky factors of A in (3.3) for given ε and N.

Further developments on this work are possible. In particular, the analysis shows that, away from the existing diagonals, the magnitude of fill-in entries decay exponentially, as seen in (3.9), a fact that could be exploited in the design of preconditioners of iterative solvers. For example, as shown in Lemma 3.1, the Cholesky factor of A, in exact arithmetic, has $\mathcal{O}(N^3)$ nonzero entries. However, Theorem 3.2 shows that, in practice (i.e., in a floating-point setting), there are only $\mathcal{O}(N^2)$ entries in L when ε is small and N is large. This suggests that, for a singularly perturbed problem, an incomplete Cholesky factorization may be a very good approximation for L. This is a topic considered in Chapter 4.

In this chapter we have restricted our study to Cholesky factorization of the coefficient matrix arising from a finite difference discretization of the model problem (3.1) on a uniform mesh and a boundary layer-adapted mesh. However, the same phenomenon is also observed in other settings, including the LU factorization of coefficient matrices coming from both finite difference and finite element methods applied to reaction-diffusion and convection-diffusion problems. For example, numerical evidence of this phenomenon in the case of a finite element discretization of a two-dimensional reaction-diffusion problem is given in Table 5.14. Further investigation is required to establish the details.

Other applications of this analysis could be also considered. For example, suppose that L_p denotes the lower triangular matrix formed by dropping all entries in L belonging in $L^{[k]}$, $k \geq p+1$. Then, our analysis can be used to estimate the difference $\|L - L_p\|$, or even $\|A - A_p\|$ where $A_p = L_p L_p^T$. This even opens up the possibility of employing inexact direct solver strategies based on A_p whose structure has been given.

Finally, we recall from this discussion leading to (3.5) we assume that $\delta \leq 0.1$. This means that, for example, when $N = 512$, then the analysis will hold only when $\varepsilon \leq 1.9 \times 10^{-4}$. Possible further refinements of the analysis are possible, which would allow the results to hold for larger values of ε. This is a topic currently under investigation.

Chapter 4

An analysis of simple preconditioners on a layer-adapted mesh

This chapter investigates the iterative solution of a two-dimensional reaction-diffusion problem when it is discritized by a finite difference method on a Shishkin mesh. As we have seen in previous chapter, there are difficulties in solving the resulting linear systems by direct methods when the perturbation parameter, ε, is small. Therefore, iterative methods are natural choices. The ultimate goal of any numerical method for solving linear systems in general, and for singularly perturbed problems in particular, is robustness and efficiency. As stated in [96, page 3]:

> It is important to realize that iterative solvers, like the underlying discretization, should be robust with respect to the singular perturbation parameter.

However, we show that the condition number of the coefficient matrix grows unboundedly when ε tends to zero, and so unpreconditioned iterative schemes, such as the conjugate gradient algorithm, perform poorly with respect to ε. We provide a careful analysis of diagonal and incomplete Cholesky preconditionings, and show that the condition number of the preconditioned linear system is independent of the perturbation parameter. We demonstrate numerically the surprising fact that these schemes are more efficient when ε is small, than when ε is $\mathcal{O}(1)$.

The material of this chapter is based on the article [84]: Thái Anh Nhan and Niall Madden, *An analysis of simple preconditioners for a singularly perturbed problem on a layer-adapted mesh*, submitted for publication, July 2015.

4.1 Introduction

As a natural progression from Chapter 3, in this chapter we consider iterative solvers, with a focus on preconditioning techniques. The linear systems in question come from the finite difference approximation (3.4) of the singularly perturbed two-dimensional reaction-diffusion differential equation:

$$-\varepsilon^2 \Delta u + B(x,y)u = f(x,y), \quad \text{on} \quad \Omega = (0,1)^2, \quad \text{and} \quad u(\partial\Omega) = g(x,y). \tag{4.1}$$

Note that, for the sake of convenience in this chapter, we denote the coefficient of reaction term by $B(x,y)$. This is because, here we follow conventions of standard numerical linear algebra textbooks, such as [44], and reserve $\boldsymbol{a}, \boldsymbol{b}$, etc., for the vectors of diagonal entries of matrices. As usual, it is assumed that there is a positive constant β such that $B(x,y) \geq \beta^2 > 0$.

Example 4.1. Recall the example of a two-dimensional reaction-diffusion problem given in Section 1.6, for which the problem data are chosen so that the exact solution is

$$u(x,y) = x^3(1+y^2) + \sin(\pi x^2) + \cos(\pi y/2) + (1+x+y)\left(e^{-2x/\varepsilon} + e^{-2y/\varepsilon}\right). \tag{4.2}$$

Its solution has two boundary layers along the edges $x = 0$ and $y = 0$, as well as one corner layer near $(0,0)$, as shown in Figure 4.1.

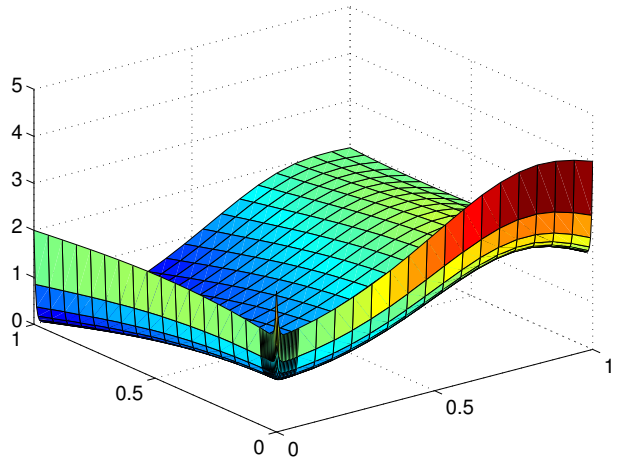

Figure 4.1: Example 4.1 with $\varepsilon = 10^{-2}$.

Applying the standard symmtrised 5-point, second-order, central difference operator defined in (3.2) to the problem (4.1), gives the system matrix

$$A := -\varepsilon^2 \Delta^N + \bar{h}_i \bar{k}_j B(x_i, y_j), \tag{4.3}$$

which is a banded, symmetric positive definite $(N-1)^2 \times (N-1)^2$ matrix. We write the corresponding linear system as

$$AU^N = f^N. \tag{4.4}$$

The analysis given in Chapter 3 has provided us with a useful observation: direct solvers are problematic for singularly perturbed problems (and in any case are of limited use for large systems), so it is natural to consider the application of iterative techniques. Surprisingly, as discussed in Section 1.2, little attention has been given to the development of (parameter) robust and reliable iterative solvers for these problems. Notable exceptions to this are [6, 34, 38, 39], which all focus on singularly perturbed convection-diffusion problems; see Section 1.2 and also [72, §1] for a detailed discussion of these.

The matrix, A, in (4.3) is symmetric positive definite, and so the natural iterative method for solving (4.4) is the Conjugate Gradient (CG) algorithm, which is based on Krylov subspace methods, see, e.g., [26, §6.6.3] and also [97, §6.7]. Its benefits include that it is easy to implement, and is very efficient for many well-conditioned problems. However, if the linear system is ill-conditioned, convergence may be slow unless a suitable preconditioning method is employed.

For the reaction-diffusion problem (4.1), a multigrid-based preconditioner was proposed in [72] and shown, in theory and practice, to be robust and efficient. That approach is quite sophisticated: to optimize efficiency, it applies different preconditioning techniques in different regions of the problem domain, depending on the nature of the solution's layers. In contrast, in this chapter we investigate the performance of standard preconditioning techniques whose application does not require any *a priori* knowledge of the location of boundary layers, with the goal of showing that they are robust, and surprisingly effective in the singularly perturbed regime.

We start in Section 4.2 with a detailed description of the layer-adapted Shishkin mesh constructed for problems which have two edge layers and one corner layer, like the one given in Example 4.1. In Section 4.3 we give an analysis of the (unpreconditioned) conjugate gradient algorithm, and show that, unsurprisingly, it performs poorly when ε is small, due to the unbounded growth of condition number of the discretization matrix. At the heart of the issue is not just the singularly perturbed nature of (4.1), but the highly refined meshes used to resolve the boundary layers.

The simplest approach for improving the performance of CG is to employ a diagonal preconditioner, which is the topic of Section 4.4. As we show, this preconditioner works very well in the context of fitted mesh methods, by which we mean that it can be proved

that the condition number of the diagonally preconditioned matrix is independent of ε.

Our main focus, however, is on the use of an incomplete Cholesky factorization as a preconditioner, which is analyzed in Section 4.5. This has several motivations, as we have briefly mentioned in Section 1.3. Here we provide more technical details for these motivations. Firstly, Ansari and Hegarty [6] have given an empirical study in a related setting (incomplete LU factorization applied to a convection-diffusion problem), though without detailed mathematical justification. They find that the number of iterations required for convergence of linear solvers, such as GMRES and BiCGStab, is adversely dependent on ε. In order to reduce the number of iterations of the linear solver to acceptable levels, they use preconditioners based on an incomplete LU factorization with drop tolerance. In particular, their numerical experiments show that, when the dropping tolerance is of $\mathcal{O}(\varepsilon)$, the number of iterations is substantially reduced. Secondly, and more significantly, the studies of direct solvers in [72] and in Chapter 3 have shown than the fill-in entries in Cholesky factors tend to zero *exponentially* as $\varepsilon \to 0$. So, in fact, in the singularly perturbed case, and in finite precision, the Cholesky factorization closely resembles an incomplete factorization, a statement we make precise in Section 4.5. It follows quickly that the incomplete Cholesky factorization without fill-in, IC(0), yields a preconditioned system matrix whose condition number is bounded in a parameter-robust way.

To highlight the efficiency of IC(0), we present a theoretical analysis of this preconditioner for two reaction-diffusion problems of the form (4.1). One is a standard case on a square domain, featuring both edge and corner layers, as in Example 4.1. The second case features an edge layer, but no corner layers, as would arise for a problem on a smooth domain, or on a square domain with suitable problem data. This allows us to distinguish the numerical complexities that are primarily due to the corner layer.

In Section 4.6 we present the results of numerical experiments that examine the number of iterations needed by the preconditioned conjugate gradient algorithm to yield a reasonable solution to our problem. It is shown that, not only are the two preconditioners we study parameter robust (in the sense that it is proven that the condition number does not degrade as $\varepsilon \to 0$), they are even more effective when the perturbation parameter is small, compared to when it is $\mathcal{O}(1)$. It is observed that the IC(0) is particularly efficient—a fact that is also demonstrated in diagrams showing convergence rates, and the distribution of the eigenvalues of the different preconditioned matrices.

4.2 The Shishkin mesh

The Shishkin mesh for the singularly perturbed problem in Example 4.1 has been briefly mentioned in Section 1.7.1. We now describe its construction in detail. Since the layers presented in the solution are all of width $\mathcal{O}(\varepsilon|\ln\varepsilon|)$, a *mesh transition point* is selected as

$$\tau_x := \min\left\{2\frac{\varepsilon}{\beta}\ln N, \frac{1}{2}\right\}. \tag{4.5}$$

Then the unit interval, $\bar{\omega}_x$, is partitioned into subintervals $[0,\tau_x]$ and $[\tau_x,1]$, each with $N/2+1$ equally spaced mesh points. More precisely, the mesh is $\omega_x^N = \{x_0, x_1, \ldots, x_N\}$, where

$$x_i = \begin{cases} ih_{\mathrm{f}}, & \text{for } i = 0,1,\ldots,N/2, \\ \tau_x + (i-N/2)h_{\mathrm{c}}, & \text{for } i = N/2+1,\ldots,N, \end{cases}$$

where, particular to this chapter, we use $h_{\mathrm{f}} := 2\tau_x/N$ to denote the "fine" mesh-widths in $[0,\tau_x]$, and $h_{\mathrm{c}} := 2(1-\tau_x)/N$ the "coarse" mesh-widths in $[\tau_x,1]$. (Later, we need $h_\tau := (h_{\mathrm{f}}+h_{\mathrm{c}})/2$).

We then construct a similar piecewise uniform mesh, $\omega_y^N = \{y_0,\ldots,y_N\}$, with transition point $\tau_y = \tau_x$. Taking the Cartesian product of ω_x^N and ω_y^N gives the two-dimensional Shishkin mesh, $\Omega_S^{N,N}$, as illustrated on the left of Figure 4.2.

Remark 4.1. *If ε is so large that $\tau_x = 1/2$, then the problem is not singularly perturbed, and the mesh is uniform. We are specifically interested in the case where the mesh is nonuniform and so, for the remainder of this chapter, we will assume that ε is small enough to guarantee that $\tau_x < 1/2$.*

For later analysis, we decompose $\bar{\Omega}$ into subregions associated with the corner layer at (0,0), the edge layers along $x=0$ and $y=0$, and the remainder of the region (see the right of Figure 4.2):

$$\Omega_{CC} := [0,\tau_x)\times[0,\tau_y), \quad \Omega_{II} := (\tau_x,1]\times(\tau_y,1],$$

and

$$\Omega_{EE}^x := (\tau_x,1]\times[0,\tau_y), \quad \Omega_{EE}^y := [0,\tau_x)\times(\tau_y,1].$$

The interfaces of these are subregions are

$$\omega_{CE}^x := \tau_x\times[0,\tau_y), \qquad \omega_{CE}^y := [0,\tau_x)\times\tau_y,$$
$$\omega_{EI}^x := (\tau_x,1]\times\tau_y, \qquad \omega_{EI}^y := \tau_x\times(\tau_y,1], \qquad \omega_{TT} := \{(\tau_x,\tau_y)\}.$$

In next section, we shall show that, in fact, the use of this highly anisotropic Shishkin mesh results in the condition number of unpreconditioned linear system coming from

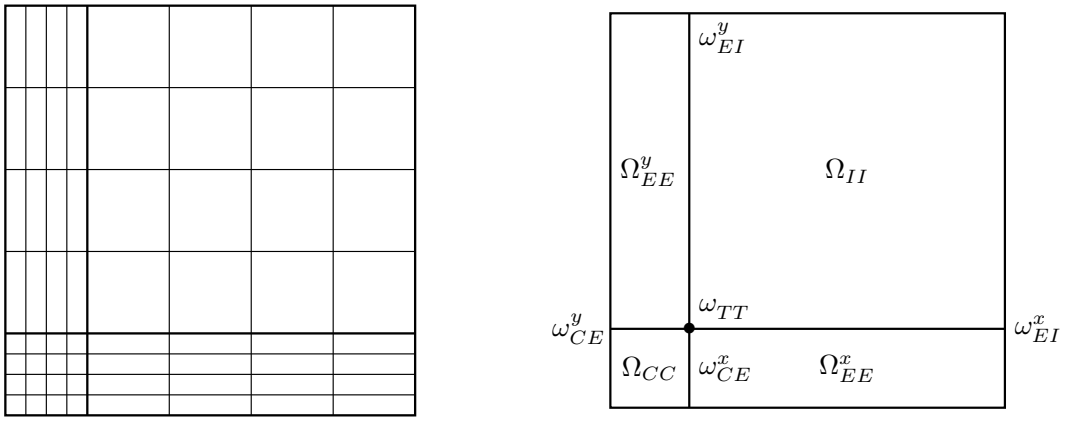

Figure 4.2: The Shishkin mesh, $\Omega_S^{N,N}$, and its decomposition for Example 4.1.

the finite difference discretization (4.3) being adversely dependent on ε. More precisely, when ε tends to zero, the condition number of the coefficient matrix A grows unboundedly, i.e., the system's conditioning is not robust with respect to the perturbation parameter.

4.3 The condition number estimate of the unpreconditioned matrix

It is well-known that the condition number of the coefficient matrix plays a key role in the convergence analysis of CG. Let $\kappa_2(A) := \|A\|_2\|A^{-1}\|_2$ be the condition number of A associated with the 2-norm, and $U^{(k)}$ be the approximation of U^N after k iterations of the CG algorithm. Then the error at iteration k is bounded as follows [44, Theorem 3.1.1]

$$\|U^N - U^{(k)}\|_A \leq 2\left(\frac{\sqrt{\kappa_2(A)}-1}{\sqrt{\kappa_2(A)}+1}\right)^k \|U^N - U^{(0)}\|_A, \tag{4.6}$$

where $\|x\|_A = (x^T Ax)^{1/2}$. Furthermore, since A is symmetric positive definite, $\|A\|_2 = \lambda_{\max} > 0$, where $\lambda_{\max}$ is the largest eigenvalue of the matrix A. Hence, $\kappa_2(A)$ can be also computed as

$$\kappa_2(A) = \|A\|_2\|A^{-1}\|_2 = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\min}$ is the smallest eigenvalue of the matrix A.

It is clear from (4.6) that, if $\kappa_2(A)$ is large, then convergence will be slow. It is an easy application of Geršgorin's Theorem (see Section 1.8.1) to verify that $\lambda_{\max}$ is bounded above by CN^{-2}. However, the same argument also gives that $\lambda_{\min}$ is at least $h_{\mathrm{f}}^2\beta^2$. This is summarised as follows.

Theorem 4.1. *The coefficient matrix A of the symmetrized finite difference discretization of* (4.1) *on the Shishkin mesh, $\Omega_S^{N,N}$, satisfies*

$$\kappa_2(A) \leq C(\varepsilon \ln N)^{-2}. \tag{4.7}$$

Proof. From the definition of h_{f} in Section 4.2, we have $h_{\mathrm{f}} = 2\tau_x/N = 4\varepsilon \ln N/(\beta N)$. This implies that $\lambda_{\min} \geq C(\varepsilon \ln N)^2/(N^2)$. Thus,

$$\kappa_2(A) = \frac{\lambda_{\max}}{\lambda_{\min}} \leq C(\varepsilon \ln N)^{-2},$$

which completes the proof. □

Theorem 4.1 suggests that the system (4.4) is ill-conditioned when ε approaches zero. The results of experiments, shown in Table 4.1, demonstrate that this is indeed the case: the constant in (4.7) is approximately 0.25. One clearly sees that $\kappa_2(A)$ is proportional to ε^{-2} for a fixed N, and slightly decreases when N grows for a fixed ε.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	1.43e+02	5.71e+02	2.29e+03	9.15e+03	3.66e+04	1.46e+05
10^{-2}	2.09e+01	8.08e+01	3.20e+02	1.28e+03	5.11e+03	2.04e+04
10^{-4}	2.42e+02	2.65e+02	6.21e+02	1.56e+03	4.17e+03	1.16e+04
10^{-6}	2.50e+04	1.80e+04	1.33e+04	1.69e+04	4.56e+04	1.29e+05
10^{-8}	2.51e+06	1.82e+06	1.34e+06	1.01e+06	7.85e+05	1.30e+06
10^{-10}	2.52e+08	1.82e+08	1.34e+08	1.01e+08	7.85e+07	6.25e+07
10^{-12}	2.52e+10	1.82e+10	1.34e+10	1.01e+10	7.86e+09	6.25e+09

Table 4.1: $\kappa_2(A)$ for the finite difference discretization (4.3) on $\Omega_S^{N,N}$.

To demonstrate that the poor scaling with ε is due to the mesh, rather than the singularly perturbed nature of the differential equation *per se*, Table 4.2 shows the condition number for a range of values of N and ε, but where the mesh is uniform. Somewhat surprisingly, perhaps, we see that these linear systems are extremely well-conditioned for small ε. This is because, when $\varepsilon \ll N^{-1}$, the reaction term in (4.4) dominates, and so A tends towards a diagonal matrix as $\varepsilon \to 0$. (This can also be reasoned by adapting the arguments of Theorem 4.1 to show that $\lambda_{\max} \leq CN^{-2}$ and $\lambda_{\min} \geq \beta^2 N^{-2}$).

The adverse ε-dependence of condition number of A, when a layer-adapted mesh is used, does indeed lead to poor performance of the unpreconditioned conjugate gradient algorithm; this is demonstrated later in Section 4.6. Therefore, our overall goal is to study preconditioning techniques that are robust with respect to the singular perturbation parameter. We will do that over the remainder of this chapter by analyzing the standard diagonal and incomplete Cholesky preconditioners.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	1.43e+02	5.71e+02	2.29e+03	9.15e+03	3.66e+04	1.46e+05
10^{-2}	2.09e+01	8.08e+01	3.20e+02	1.28e+03	5.11e+03	2.04e+04
10^{-4}	1.20e+00	1.82e+00	4.28e+00	1.41e+01	5.34e+01	2.11e+02
10^{-6}	1.00e+00	1.01e+00	1.03e+00	1.13e+00	1.52e+00	3.10e+00
10^{-8}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.01e+00	1.02e+00
10^{-10}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00
10^{-12}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00

Table 4.2: $\kappa_2(A)$ for the finite difference discretization (4.3) on a uniform mesh.

4.4 Diagonal preconditioner

As discussed in Section 1.8.3, in order to accelerate the performance of CG, one uses a preconditioner, which is usually represented as a matrix. We provide here some more technical details of preconditioners for CG, in order to carry out the analysis later on. Suppose M is a preconditioner which is symmetric positive definite. Then, instead of solving $AU^N = f^N$, preconditioned CG solves $M^{-1}AU^N = M^{-1}f^N$. This is equivalent to applying CG directly to solving the symmetric positive definite system

$$(M^{-1/2}AM^{-1/2})(M^{1/2}U^N) = M^{-1/2}f^N, \tag{4.8}$$

where $M^{1/2}$ is the principle square root of the symmetric positive definite matrix M. Furthermore, since the matrices $M^{-1}A$ and $M^{-1/2}AM^{-1/2}$ are similar, it suffices to analyze the condition number of the latter.

In this section, we study a simple diagonal preconditioner

$$D := \operatorname{diag}(a_{11}, a_{22}, \ldots, a_{nn}).$$

That is, we take $M = D$ in (4.8) where D is the diagonal matrix whose entries are taken from the main diagonal of A. As we shall show, the condition number of the resulting system is independent of ε.

Theorem 4.2. *Let $A_D = D^{-1/2}AD^{-1/2}$. Then*

$$\kappa_2(A_D) \le C\frac{N^2}{\ln^2 N}. \tag{4.9}$$

Proof. Because A_D and $D^{-1}A$ are similar, we have that

$$\kappa_2(A_D) = \lambda_{\max}(D^{-1}A)/\lambda_{\min}(D^{-1}A).$$

In order to bound $\lambda_{\max}(D^{-1}A)$, we apply Geršgorin's Theorem to the matrix $D^{-1}A$. Since A is strictly diagonally dominant, it is easy to see that $\lambda_{\max}(D^{-1}A) \le 2$. We

also use Geršgorin's Theorem to bound $\lambda_{\min}(D^{-1}A)$ from below. Indeed, if λ is an eigenvalue of $D^{-1}A$, then

$$\lambda \geq \min_{i,j} \left\{ \frac{\bar{h}_i \bar{k}_j B(x_i, y_j)}{\gamma_{ij} + \bar{h}_i \bar{k}_j B(x_i, y_j)} \right\}, \tag{4.10}$$

where γ_{ij} denotes the diagonal entry of $(-\varepsilon^2 \Delta^N)$ in the row corresponding to the (i, j)-node of the mesh. That is

$$\gamma_{ij} = \varepsilon^2 \left(\bar{k}_j \left(\frac{1}{h_i} + \frac{1}{h_{i+1}} \right) + \bar{h}_i \left(\frac{1}{k_j} + \frac{1}{k_{j+1}} \right) \right).$$

The minimum of (4.10) is achieved when $\bar{h}_i = \bar{k}_j = h_{\mathrm{f}} = 4\varepsilon \ln N/(\beta N)$. Hence, $\lambda_{\min}(D^{-1}A) \geq CN^{-2} \ln^2 N$. □

Theorem 4.2 shows that the condition number of the preconditioned linear system is robust with respect to singular perturbation parameter; Table 4.3 shows this bound is quite sharp, with the constant in (4.9) being approximately 0.5. Furthermore, notice that for a fixed N, the system is better conditioned when ε is small, compared to when $\varepsilon = 1$. It is verified in Section 4.6 that, in practice, diagonally-preconditioned CG is more efficient when the problem is singularly perturbed. However, even greater efficiencies are possible when an incomplete Cholesky preconditioner is used; this is the topic of Section 4.5.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	1.43e+02	5.71e+02	2.29e+03	9.15e+03	3.66e+04	1.46e+05
10^{-2}	2.09e+01	8.08e+01	3.20e+02	1.28e+03	5.11e+03	2.04e+04
10^{-4}	1.42e+01	3.95e+01	1.15e+02	3.54e+02	1.10e+03	3.49e+03
10^{-6}	1.38e+01	3.84e+01	1.12e+02	3.39e+02	1.44e+03	6.12e+03
10^{-8}	1.37e+01	3.82e+01	1.11e+02	3.34e+02	1.04e+03	3.79e+03
10^{-10}	1.37e+01	3.82e+01	1.11e+02	3.33e+02	1.03e+03	3.29e+03
10^{-12}	1.37e+01	3.82e+01	1.11e+02	3.32e+02	1.03e+03	3.28e+03

Table 4.3: $\kappa_2(A_D)$ for the finite difference discretization (4.3) on $\Omega_S^{N,N}$.

4.5 Incomplete Cholesky Preconditioner

In this section, we investigate in detail the use of *incomplete Cholesky factorization without fill-in*, IC(0), as a preconditioner to the singularly perturbed linear system (4.4) and show that it is particularly advantageous when ε is very small. Note that the matrix A defined in (4.4) is symmetric positive definite, so there exists a complete Cholesky

factorization such that $A = LL^T$, where L is a lower triangular matrix. The idea of incomplete Cholesky factorization is to set some entries of L to be zeros (or, rather, don't compute them at all), and so $A \approx \tilde{L}\tilde{L}^T$ where $\tilde{L}$ is the incomplete Cholesky factor. Then, $M = \tilde{L}\tilde{L}^T$ can be used as a preconditioner.

The motivation for this preconditioner comes from the fact that the usual Cholesky factor of A is quite dense: although A has typically five nonzero entries per row (assuming the usual lexicographic ordering), the factors have $\mathcal{O}(\mathcal{N})$ nonzeros per row. The fill-in entries in the factor are smaller in magnitude than the entries corresponding to nonzeros in A. Dropping these fill-ins yields an approximate factorization of A that can be used as a preconditioner. However, as analyzed in Chapter 3, when ε is small, the fill-in entries are extremely small, which leads us to propose that IC(0) should be an excellent preconditioner for our problem.

Furthermore, the IC(0) factorization has only the $\mathcal{O}(N^2)$ nonzero entries that are in the positions of the nonzeros of A, compared with the $\mathcal{O}(N^3)$ nonzeros found in the complete Cholesky factorization. Therefore it is *much* cheaper to compute. In addition, it is worth noting that no subnormal numbers are involved in computing the IC(0) factorization, so, unlike the complete factorization, its computational cost does not depend adversely on the perturbation parameter.

For our analysis, we first study IC(0) applied to the discretization of (4.1) on an arbitrary mesh. In Section 4.5.2, we consider the specific case when the mesh is the Shishkin mesh as given in Figure 4.2. Then, in Section 4.5.3, we outline the results of the analysis in the specialized case where the solution (and mesh) features an edge layer, but no corner layers.

4.5.1 Analysis of IC(0) on an arbitrary mesh

For convenience (to avoid square roots in the calculations), we write the incomplete factorization as $\tilde{L}\tilde{D}\tilde{L}^T$ where $\tilde{D}$ is a diagonal matrix. We follow the notation and description in [44, §11.1]. Let m be the bandwidth of the matrix A. Let $\boldsymbol{a}$ denote the main diagonal of A, $\boldsymbol{b}$ its first lower diagonal, and $\boldsymbol{c}$ its $(m-1)^{\text{st}}$ lower diagonal. We want to compute $\tilde{L}$ and $\tilde{D}$ in terms of $\boldsymbol{a}, \boldsymbol{b}$ and $\boldsymbol{c}$. Let $\tilde{\boldsymbol{a}}$ denote the main diagonal of $\tilde{L}$, $\tilde{\boldsymbol{b}}$ its first lower diagonal, and $\tilde{\boldsymbol{c}}$ its $(m+1)^{\text{st}}$ lower diagonal. Also, let $\tilde{\boldsymbol{d}}$ denote the main diagonal of $\tilde{D}$. Then, using lexicographical ordering with $p = i + m(j-1),\ i, j = 1, \ldots, N-1$, we have that

$$\tilde{\boldsymbol{b}} = \boldsymbol{b}, \quad \tilde{\boldsymbol{c}} = \boldsymbol{c}, \tag{4.11}$$

$$\tilde{a}_p = \frac{1}{\tilde{d}_p} = a_p - \tilde{b}_{p-1}^2 \tilde{d}_{p-1} - \tilde{c}_{p-m}^2 \tilde{d}_{p-m}, \qquad p = 1, 2, \ldots, m^2, \tag{4.12}$$

with the convention that $\tilde{b}_p = \tilde{c}_p = \tilde{d}_p = 0$ for $p \leq 0$. Then row p of the product $M = \tilde{L}\tilde{D}\tilde{L}^T$ has the form

$$\ldots \quad 0 \quad c_{p-m} \quad r_{p-m+1} \quad 0 \quad \ldots \quad 0 \quad b_{p-1} \quad a_p \quad b_p \quad 0 \quad \ldots \quad 0 \quad r_p \quad c_p \quad 0 \quad \ldots,$$

where

$$r_p = \frac{b_{p-1}c_{p-1}}{\tilde{a}_{p-1}}. \tag{4.13}$$

Note that M has two extra nonzero diagonals compared to A, which are denoted by $\boldsymbol{r}$. More precisely, if $M = \tilde{L}\tilde{D}\tilde{L}^T$ is the IC(0) preconditioner, then the only difference between M and A are the diagonals $\boldsymbol{r}$. That is, if we write $A = M - R$, then $R = [R_{i,j}]$ is the symmetric matrix whose nonzero entries are

$$R_{p,p+(m-2)} = r_p = R_{p+(m-2),p}, \quad p = 1, 2, \ldots, m^2.$$

Lemma 4.1. *Set $B_p := \bar{h}_i\bar{k}_j B(x_i, y_j)$, the scaled reaction term in (4.3). For $p = 1, 2, \ldots, m^2$, the entries $\tilde{a}_p$ and $\tilde{r}_p$ satisfy*

$$\tilde{a}_p \geq |b_p| + |c_p| + B_p, \tag{4.14a}$$

$$r_p \leq \frac{b_{p-1}c_{p-1}}{|b_{p-1}| + |c_{p-1}| + B_{p-1}}. \tag{4.14b}$$

Proof. The arguments are based on the ideas from [46, Lem. 4.1], but adapted for our case. From (4.3), the matrix A can be written in stencil notation as

$$A = \begin{bmatrix} & c_p & \\ b_{p-1} & a_p & b_p \\ & c_{p-m} & \end{bmatrix},$$

where $a_p \geq |b_p| + |b_{p-1}| + |c_p| + |c_{p-m}| + B_p$, and $B_p > 0$. First, we show by induction that $\tilde{a}_p \geq |b_p|$ and $\tilde{a}_p \geq |c_p|$. Indeed, when $p = 1$, then

$$\tilde{a}_1 = a_1 \geq |b_1| + |c_1| + B_1,$$

so $\tilde{a}_1 \geq |b_1|$ and $\tilde{a}_1 \geq |c_1|$. Now suppose that

$$\tilde{a}_q \geq |b_q|, \quad \text{and} \quad \tilde{a}_q \geq |c_q|, \qquad q = 1, \ldots, p-1.$$

From (4.11) and (4.12), we have

$$\begin{aligned}
\tilde{a}_p &= a_p - \frac{b_{p-1}^2}{\tilde{a}_{p-1}} - \frac{c_{p-m}^2}{\tilde{a}_{p-m}} \\
&\geq |b_p| + |c_p| + |b_{p-1}| + |c_{p-m}| + B_p - \frac{|b_{p-1}||b_{p-1}|}{\tilde{a}_{p-1}} - \frac{|c_{p-m}||c_{p-m}|}{\tilde{a}_{p-m}} \\
&\geq |b_p| + |c_p| + B_p \quad (\text{since } \tilde{a}_{p-1} \geq |b_{p-1}| \Rightarrow \frac{|b_{p-1}||b_{p-1}|}{\tilde{a}_{p-1}} \leq |b_{p-1}|).
\end{aligned}$$

Therefore, $\tilde{a}_p \geq |b_p|$ and $\tilde{a}_p \geq |c_p|$, for $p = 1, \ldots, m^2$. This establishes (4.14a), which along with (4.13), completes the proof. □

Applying (4.14b) to an arbitrary mesh, we get that

$$r_p \leq \frac{(-\varepsilon^2)\dfrac{\bar{k}_j}{h_{i+1}}(-\varepsilon^2)\dfrac{\bar{h}_i}{k_{j+1}}}{\varepsilon^2\dfrac{\bar{k}_j}{h_{i+1}} + \varepsilon^2\dfrac{\bar{h}_i}{k_{j+1}} + \bar{h}_i\bar{k}_j B(x_i, y_j)}. \tag{4.15}$$

In the following section, we apply this bound to particular fitted meshes.

4.5.2 Analysis of IC(0) for a corner layer problem

In Lemma 4.1, we gave an estimate for the entries of the matrix R in the general case. The next theorem gives a concrete upper bound for the entries r_p when we employ the Shishkin mesh, $\Omega_S^{N,N}$, shown in Figure 4.2, which features two edge layers, and a single corner layer adjacent to $(0, 0)$.

Theorem 4.3. *Let A be the coefficient matrix of symmetrised finite difference discretization on the Shishkin mesh $\Omega_S^{N,N}$, and let M be the associated $IC(0)$ preconditioner. If $A = M - R$, then,*

$$r_p \leq \frac{(\varepsilon N)^2}{2N^2 + 16\ln^2 N}, \quad \textit{for } p = 1, 2, \ldots, m^2. \tag{4.16}$$

Proof. We will show that

$$r_p \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}}^2\beta^2}, \quad \text{for } p = 1, 2, \ldots, m^2.$$

The proof is done by direct computation on the subregions of $\Omega_S^{N,N}$ (see Figure 4.2). Because of the symmetry of (4.15), it is sufficient to examine the regions Ω_{CC}, Ω_{EE}^x, Ω_{II}, ω_{CE}^x, ω_{EI}^x, and ω_{TT}.

Case 1: $(x_i, y_j) \in \Omega_{CC}$, then $h_{i+1} = k_{j+1} = \bar{h}_i = \bar{k}_j = h_{\mathrm{f}}$, so, from (4.15), we have that

$$r_{CC} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}}^2 B(x_i, y_j)} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}}^2\beta^2}.$$

Case 2: $(x_i, y_j) \in \Omega_{EE}^x$, then $h_{i+1} = \bar{h}_i = h_{\mathrm{c}}$, and $k_{j+1} = \bar{k}_j = h_{\mathrm{f}}$, so we get that

$$\begin{aligned} r_{EE} &\leq \frac{\varepsilon^2\dfrac{h_{\mathrm{f}}}{h_{\mathrm{c}}}\varepsilon^2\dfrac{h_{\mathrm{c}}}{h_{\mathrm{f}}}}{\varepsilon^2\left(\dfrac{h_{\mathrm{f}}}{h_{\mathrm{c}}} + \dfrac{h_{\mathrm{c}}}{h_{\mathrm{f}}}\right) + h_{\mathrm{f}}h_{\mathrm{c}}\beta^2} = \frac{\varepsilon^4}{\varepsilon^2\left(\dfrac{h_{\mathrm{f}}}{h_{\mathrm{c}}} + \dfrac{h_{\mathrm{c}}}{h_{\mathrm{f}}}\right) + h_{\mathrm{f}}h_{\mathrm{c}}\beta^2} \\ &\leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}}h_{\mathrm{c}}\beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}}^2\beta^2}. \end{aligned}$$

Case 3: $(x_i, y_j) \in \Omega_{II}$, then $h_{i+1} = k_{j+1} = \bar{h}_i = \bar{k}_j = h_\mathrm{c}$, so

$$r_{II} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{c}^2\beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{f}^2\beta^2}.$$

Case 4: $(x_i, y_j) \in \omega_{CE}^x$, then $h_{i+1} = h_\mathrm{c}, \bar{h}_i = h_\tau$, and $k_{j+1} = \bar{k}_j = h_\mathrm{f}$, so

$$\begin{aligned} r_{CE} &\leq \frac{\varepsilon^2\dfrac{h_\mathrm{f}}{h_\mathrm{c}}\varepsilon^2\dfrac{h_\tau}{h_\mathrm{f}}}{\varepsilon^2\left(\dfrac{h_\mathrm{f}}{h_\mathrm{c}} + \dfrac{h_\tau}{h_\mathrm{f}}\right) + h_\mathrm{f}h_\tau\beta^2} = \frac{\varepsilon^4}{\varepsilon^2\left(\dfrac{h_\mathrm{f}}{h_\tau} + \dfrac{h_\mathrm{c}}{h_\mathrm{f}}\right) + h_\mathrm{c}h_\mathrm{f}\beta^2} \\ &\leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{c}h_\mathrm{f}\beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{f}^2\beta^2}. \end{aligned}$$

Case 5: $(x_i, y_j) \in \omega_{EI}^x$, then $h_{i+1} = \bar{h}_i = k_{j+1} = h_\mathrm{c}$, and $\bar{k}_j = h_\tau$, so

$$r_{EI} \leq \frac{\varepsilon^2\dfrac{h_\tau}{h_\mathrm{c}}\varepsilon^2\dfrac{h_\mathrm{c}}{h_\mathrm{c}}}{\varepsilon^2\left(\dfrac{h_\tau}{h_\mathrm{c}} + 1\right) + h_\tau h_\mathrm{c}\beta^2} = \frac{\varepsilon^4}{\varepsilon^2\left(1 + \dfrac{h_\mathrm{c}}{h_\tau}\right) + h_\mathrm{c}^2\beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{f}^2\beta^2}.$$

Case 6: $(x_i, y_j) \in \omega_{TT}$, then $h_{i+1} = k_{j+1} = h_\mathrm{c}$, and $\bar{h}_i = \bar{k}_j = h_\tau$, so

$$r_{TT} \leq \frac{\left(\varepsilon^2\dfrac{h_\tau}{h_\mathrm{c}}\right)^2}{2\varepsilon^2\dfrac{h_\tau}{h_\mathrm{c}} + h_\tau^2\beta^2} = \frac{\varepsilon^4}{2\varepsilon^2\dfrac{h_\mathrm{c}}{h_\tau} + h_\mathrm{c}^2\beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{f}^2\beta^2}.$$

Combining all these cases, we get that

$$r_p \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_\mathrm{f}^2\beta^2} = \frac{(\varepsilon N)^2}{2N^2 + 16\ln^2 N}.$$

□

Remark 4.2. *Note that* $(N^2)/(2N^2 + 16\ln^2 N) < 1$*, so* (4.16) *can be simplified as*

$$\|R\|_2 \leq \|R\| \leq C\varepsilon^2, \tag{4.17}$$

since R *is symmetric. This highlights the fact that the difference between* A *and its IC(0) preconditioner is, in the maximum norm,* $\mathcal{O}(\varepsilon^2)$*. Furthermore, when* $\varepsilon \ll 1$*, the IC(0) preconditioner satisfies the conditions of being a good preconditioner, stated in Section 1.8.3. For example, with a small diffusion parameter* $\varepsilon = 10^{-6}$*, the difference in the maximum norm between* M *and* A *is*

$$\|M - A\| \approx 10^{-12},$$

showing that condition (P1) is satisfied. Furthermore, the condition (P2) is also satisfied because solving $Mx = b$ *just involves solving* $\tilde{L}y = b$*, where* $y = (\tilde{D}\tilde{L}^T)x$*. Since, by definition,* $\tilde{L}$ *is a lower tridiagonal matrix with only three nonzero entries per row, the system* $\tilde{L}y = b$ *can be efficiently solved by back-substitution. Then, a similar strategy is applied for the upper triangular matrix* $\tilde{D}\tilde{L}^T$ *to solve* $(\tilde{D}\tilde{L}^T)x = y$*.*

As shown in Theorem 4.1, the condition number of A depends badly on the perturbation parameter, ε. As with the diagonal preconditioner, we will prove that the condition number of the IC(0)-preconditioned system is independent of ε. In our analysis, we require the following standard property of a regular splitting. Related results on regular splitting are discussed in depth in standard textbooks, e.g., [44, Sec 10.4].

Definition 4.1 ([44, Definition 10.3.1])**.** For the $n \times n$ real matrices A, M and R, the splitting $A = M - R$ is a *regular splitting* if M is nonsingular with $M^{-1} \geq 0$ and $M \geq A$.

Lemma 4.2 ([44, Theorem 10.3.1])**.** *Let $A = M - R$ be a regular splitting of A, where $A^{-1} \geq 0$. Then*

$$\rho(M^{-1}R) = \frac{\rho(A^{-1}R)}{1 + \rho(A^{-1}R)} < 1,$$

where $\rho(A)$ is spectral radius of matrix A.

We use the standard finite difference scheme (4.3) to discretize (4.1). As we shall see, the resulting system matrix is strictly diagonally dominant. In our analysis, we will make use of the following result of Varah.

Lemma 4.3 ([109, Theorem 1])**.** *If the matrix A is strictly diagonally dominant by rows, and*

$$\alpha := \min_i \left(|a_{ii}| - \sum_{j \neq i} |a_{ij}| \right) > 0,$$

then $\|A^{-1}\| < 1/\alpha$.

Theorem 4.4. *Let A be the discretization matrix in* (4.3) *on the Shishkin mesh $\Omega_S^{N,N}$, and let M be its IC(0) preconditioner. If $A_M = \left(M^{-1/2} A M^{-1/2}\right)$, then*

$$\kappa_2(A_M) \leq C \frac{N^2}{\ln^2 N}. \tag{4.18}$$

Proof. We will first show that $\|A_M\|_2 \leq C$. The coefficient matrix A defined in (4.3) is symmetric positive definite. Furthermore, since its off-diagonal entries are nonpositive, it is an M-matrix [44, Theorem 10.3.3]. Hence, by [75, Theorem 2.4], the IC(0) preconditioner M exists uniquely and $A = M - R$ is a regular splitting. In addition, A being an M-matrix implies that $A^{-1} \geq 0$. Therefore, the premises of Theorem 4.2 are satisfied, so $\rho(M^{-1}R) < 1$. Then,

$$\begin{aligned} \|A_M\|_2 &= \|M^{-1/2}(M - R)M^{-1/2}\|_2 = \|I - M^{-1/2} R M^{-1/2}\|_2 \\ &\leq 1 + \|M^{-1/2} R M^{-1/2}\|_2 = 1 + \rho(M^{-1/2} R M^{-1/2}) \\ &= 1 + \rho(M^{-1}R) \leq 2. \end{aligned} \tag{4.19}$$

The next step is to show that $\|A_M^{-1}\|_2 \leq CN^2 \ln^{-2} N$. First, we use Lemma 4.3 to get that

$$\|A^{-1}\| \leq C\frac{N^2}{\varepsilon^2 \ln^2 N}. \tag{4.20}$$

Hence,

$$\begin{aligned}\left\|A_M^{-1}\right\|_2 = \rho\left(A^{-1}M\right) \leq \|A^{-1}M\| = \|A^{-1}(A+R)\| \\ \leq 1 + \|A^{-1}R\| \leq 1 + \|A^{-1}\|\|R\|.\end{aligned}$$

By (4.17) and (4.20), we get that

$$\|A_M^{-1}\|_2 \leq C\frac{N^2}{\ln^2 N}. \tag{4.21}$$

The proof is completed by combining (4.19) and (4.21). □

As shown by the results in Table 4.4, the bound given by Theorem 4.4 is sharp. Note that this is the same bound as was found for the diagonal preconditioner in Theorem 4.2. However, comparing the results in Table 4.4 with the corresponding results in Table 4.3, we see that $C \approx 0.04$ in (4.18), compared to $C \approx 0.5$ in (4.9). As we shall see in Section 4.6, it follows that IC(0) is a more efficient preconditioner.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	9.52	35.71	140.35	558.90	2232.98	8929.29
10^{-2}	2.37	6.89	25.02	97.50	387.46	1547.28
10^{-4}	1.75	3.53	9.06	29.74	106.99	361.05
10^{-6}	1.73	3.48	8.88	25.94	80.98	262.00
10^{-8}	1.73	3.47	8.86	25.85	80.59	260.31
10^{-10}	1.73	3.47	8.85	25.84	80.54	260.12
10^{-12}	1.73	3.47	8.85	25.84	80.54	260.10

Table 4.4: $\kappa_2(A_M)$ for the finite difference discretization (4.3) on $\Omega_S^{N,N}$.

4.5.3 Analysis of IC(0) for a problem without corner layers

Inspection of the proof of Theorem 4.3 shows that the final estimate is dominated by the terms associated with the corner region, Ω_{CC}, and its interfaces. To verify that this is indeed the case, we now consider a variant on (4.1) but where the problem data is chosen such that there are no corner layers at all. This could arise if, for example, problem (4.1) was posed on a smooth domain: see, e.g., [54] for an analysis of a semilinear reaction-diffusion problem on a Shishkin mesh for such a case. It could also arise for a problem on the unit square, but where the data are such that it has

only one edge layer, or two edge layers along parallel sides of the unit square. This can happen, for example, if $f(x,y)/B(x,y)$ agrees with $g(x,y)$ on the other boundary regions of Ω.

For simplicity, we consider such a problem with only one layer along the edge $x = 0$, as in Example 4.2 below.

Example 4.2. We modify the problem data of Example 4.1 so that the solution is

$$u(x,y) = x^3(1+y^2) + \sin(\pi x^2) + \cos(\pi y/2) + (1+x+y)\left(e^{-2x/\varepsilon}\right), \tag{4.22}$$

which is plotted in Figure 4.3.

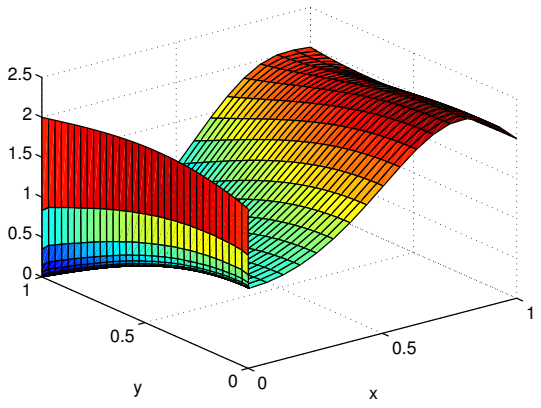

Figure 4.3: Example 4.2 with one boundary layer along y-axis when $\varepsilon^2 = 10^{-6}$.

Thus $\bar{\Omega}$ has two subregions: Ω^y_{EE} associated with the edge layer along $x = 0$, and Ω_{II} for the interior region (see the right of Figure 4.4)

$$\Omega_{II} := (\tau_x, 1] \times [0,1], \quad \Omega^y_{EE} := [0, \tau_x) \times [0,1].$$

The interface of these two subregions is

$$\omega^y_{EI} := \tau_x \times [0,1].$$

Then, a suitable Shishkin mesh, which we denote $\Omega^{N,N}_Q$, is formed by taking ω^N_x as is given in Section 1.7.1, but with ω^N_y uniform.

Applying the finite difference scheme (4.3) on the Shishkin mesh just described, we obtain the errors in the computed solution in Table 4.5, where almost second-order convergence is easily observed (see (1.21)).

Figure 4.5 shows the errors in the computed solution of Example 4.1 (left), and Example 4.2 (right). It highlights the fact that computing an accurate solution in the edge layer regions is at least as challenging as in the corner layer regions. However, as

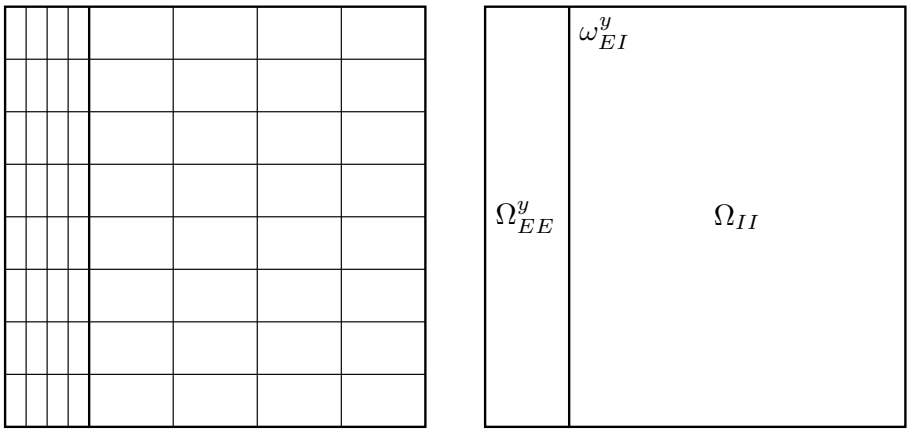

Figure 4.4: The Shishkin mesh, $\Omega_Q^{N,N}$, and its decomposition for Example 4.2.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	6.16157e-03	1.55254e-03	3.89910e-04	9.74913e-05	2.43787e-05	6.09475e-06
10^{-2}	6.23862e-02	1.67878e-02	4.27930e-03	1.07828e-03	2.69959e-04	6.75285e-05
10^{-4}	9.23588e-02	3.79423e-02	1.44510e-02	5.00241e-03	1.64905e-03	5.22641e-04
10^{-6}	9.38062e-02	3.88367e-02	1.47858e-02	5.11275e-03	1.68984e-03	5.35840e-04
10^{-8}	9.38703e-02	3.88759e-02	1.48182e-02	5.13590e-03	1.69490e-03	5.37153e-04
10^{-10}	9.38759e-02	3.88783e-02	1.48193e-02	5.13653e-03	1.69540e-03	5.37685e-04
10^{-12}	9.38765e-02	3.88785e-02	1.48194e-02	5.13656e-03	1.69542e-03	5.37694e-04

Table 4.5: $\|u - U^N\|_{\Omega_Q^{N,N}}$ for Example 4.2.

we shall now show, the edge layers are far less problematic than corner layers for linear solvers, if a suitable preconditioner is used.

The analysis of a result corresponding to Theorem 4.3 is simplified, since there are only three cases to consider.

Corollary 4.1. *Let A be the coefficient matrix of symmetrized finite-difference discretization on the simplified Shishkin mesh $\Omega_Q^{N,N}$, and let M be the associated $IC(0)$ preconditioner. If $A = M - R$, then,*

$$r_p \leq \frac{\varepsilon^3 N^2}{2\varepsilon N^2 + 4\beta \ln N}. \tag{4.23}$$

Proof. This follows from inspecting the arguments in the proof of Theorem 4.3. Indeed, let $H = N^{-1}$ be the equidistant mesh size along the y-direction. Since $N^{-1} \leq h_c \leq 2N^{-1}$ and $N^{-1}/2 \leq h_\tau \leq 2N^{-1}$, both h_c and h_τ are $\mathcal{O}(H)$. In this case, we have only to consider three cases, and, in the notation of the proof of Theorem 4.3, we have:

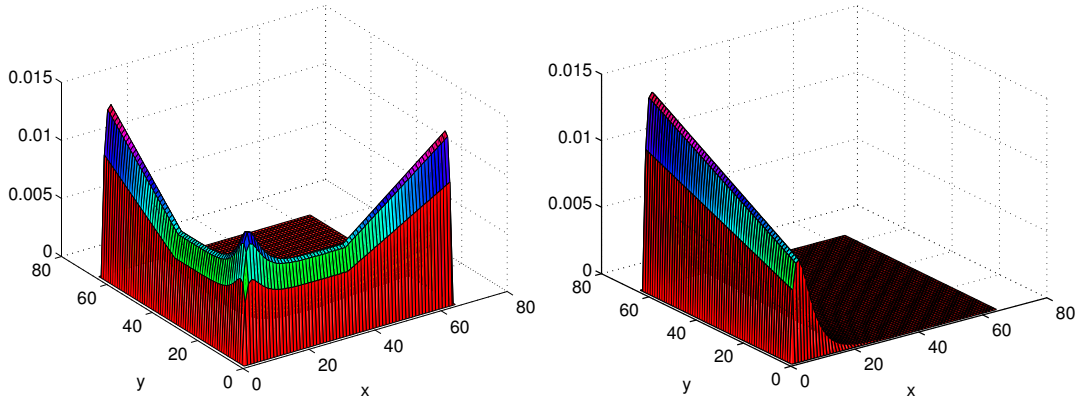

Figure 4.5: Errors for Example 4.1 (left) and Example 4.2 (right) with $N = 2^5$ and $\varepsilon^2 = 10^{-6}$.

Case 1: $(x_i, y_j) \in \Omega_{EE}^y$, then $h_{i+1} = \bar{h}_i = h_{\mathrm{f}}$, and $k_{j+1} = \bar{k}_j = h_{\mathrm{c}}$, so we get that

$$r_{EE} \leq \frac{\varepsilon^2 \dfrac{H}{h_{\mathrm{f}}} \varepsilon^2 \dfrac{h_{\mathrm{f}}}{H}}{\varepsilon^2 \left(\dfrac{H}{h_{\mathrm{f}}} + \dfrac{h_{\mathrm{f}}}{H} \right) + h_{\mathrm{f}} H \beta^2} = \frac{\varepsilon^4}{\varepsilon^2 \left(\dfrac{H}{h_{\mathrm{f}}} + \dfrac{h_{\mathrm{f}}}{H} \right) + h_{\mathrm{f}} H \beta^2} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}} H \beta^2}.$$

Case 2: $(x_i, y_j) \in \omega_{EI}^y$, then $h_{i+1} = h_{\mathrm{c}}, \bar{k}_j = k_{j+1} = H$, and $\bar{h}_i = h_\tau$, so (using $a + b \geq 2\sqrt{ab}$ and $h_{\mathrm{c}}/h_\tau \geq 1/2$)

$$r_{EI} \leq \frac{\varepsilon^2 \dfrac{H}{h_{\mathrm{c}}} \varepsilon^2 \dfrac{h_\tau}{H}}{\varepsilon^2 \left(\dfrac{H}{h_{\mathrm{c}}} + \dfrac{h_\tau}{H} \right) + h_\tau H \beta^2} = \frac{\varepsilon^4}{\varepsilon^2 \left(\dfrac{H}{h_\tau} + \dfrac{h_{\mathrm{c}}}{H} \right) + h_{\mathrm{c}} H \beta^2} \leq \frac{\varepsilon^4}{\sqrt{2}\varepsilon^2 + h_{\mathrm{c}}^2 \beta^2}.$$

Case 3: $(x_i, y_j) \in \Omega_{II}$, then $h_{i+1} = \bar{h}_i = h_{\mathrm{c}}$, and $k_{j+1} = \bar{k}_j = H$ so

$$r_{II} \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{c}} H \beta^2}.$$

Combining all these cases, we get that

$$r_p \leq \frac{\varepsilon^4}{2\varepsilon^2 + h_{\mathrm{f}} H \beta^2} = \frac{\varepsilon^3 N^2}{2\varepsilon N^2 + 4\beta \ln N}.$$

This completes the proof. □

In Theorem4.3, for sufficiently small ε, we found that r_p behaves like ε^2. Corollary 4.1 gives the much smaller bound $r_p \sim \varepsilon^3 N^2/(\varepsilon N^2 + \ln N)$. Moreover, it allows us to establish the following result.

Corollary 4.2. *Let M be the IC(0) preconditioner of A on the simplified Shishkin mesh $\Omega_Q^{N,N}$. If $A_M = \left(M^{-1/2} A M^{-1/2}\right)$, then*

$$\kappa_2(A_M) \leq C \left(1 + \frac{\varepsilon^2 N^4}{2\varepsilon N^2 \ln N + 4\beta \ln^2 N} \right). \tag{4.24}$$

Proof. The arguments of Theorem 4.4 can be adapted easily to give that $\|A_M\|_2 \leq C$, and that

$$\|A_\varepsilon^{-1}\| \leq C\frac{N^2}{\varepsilon \ln N},$$

from Lemma 4.3. Then, from (4.23), we get that

$$\begin{aligned}\left\|A_M^{-1}\right\|_2 = \rho\left(A^{-1}M\right) \leq \|A^{-1}M\| &= \|A^{-1}(A+R)\| \\ &\leq 1 + \|A^{-1}R\| \leq 1 + \|A^{-1}\|\|R\| \\ &\leq C\left(1 + \frac{N^2}{\varepsilon \ln N}\frac{\varepsilon^3 N^2}{2\varepsilon N^2 + 4\beta \ln N}\right) \\ &\leq C\left(1 + \frac{\varepsilon^2 N^4}{2\varepsilon N^2 \ln N + 4\beta \ln^2 N}\right).\end{aligned}$$

This completes the proof. □

The bound in (4.24) suggests that, for sufficiently small ε, we have that $\kappa_2(A_M)$ is bounded by a constant, which is verified for Example 4.2 in Table 4.6. It follows that IC(0) should be an excellent preconditioner for this problem. We can deduce two further facts from this example. First, for iterative procedures, the corner layer regions are the most troublesome. Secondly, a corresponding result does not hold for the diagonal preconditioner studied in Section 4.4: the bound given in Theorem 4.2 holds for the one-layer problem, with only a modest reduction in the associated constant.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	9.52e+00	3.57e+01	1.40e+02	5.59e+02	2.23e+03	8.93e+03
10^{-2}	2.37e+00	6.89e+00	2.50e+01	9.75e+01	3.87e+02	1.55e+03
10^{-4}	1.03e+00	1.20e+00	1.90e+00	4.65e+00	1.51e+01	5.50e+01
10^{-6}	1.00e+00	1.00e+00	1.01e+00	1.05e+00	1.22e+00	1.91e+00
10^{-8}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00
10^{-10}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00
10^{-12}	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00

Table 4.6: $\kappa_2(A_M)$ for the finite difference discretization (4.3) on $\Omega_Q^{N,N}$.

4.6 Numerical results

In this section we compare the performance of (unpreconditioned) CG, and the two approaches to preconditioning discussed in Sections 4.4 and 4.5. To make the comparisons meaningful, a carefully chosen stopping criterion must be used so that the

discretization accuracy is achieved, while taking care not to over-solve the linear system. Here we briefly discuss the necessity of a suitable stopping criterion for singularly perturbed problems. See [72, §4.6] for a detailed discussion. Let $e^{(k)} = U^N - U^{(k)}$, the error in the linear solver's solution after k iterations. This quantity is unknown, but is estimated from $e^{(k)} = A^{-1}r^{(k)}$, where $r^{(k)} = f^N - AU^{(k)}$ is the residual, leading to

$$\|e^{(k)}\| \leq \|A^{-1}\|\|r^{(k)}\|.$$

However, as seen in (4.20), $\|A^{-1}\|$ is $\mathcal{O}(\varepsilon^{-2})$. Therefore, to guarantee that $\|e^{(k)}\|$ has the same order as the error bound given in (1.21), $\|r^{(k)}\|$ must be bounded by

$$\|r^{(k)}\| \leq C\varepsilon^2 N^{-4} \ln^4 N.$$

However, this bound is not computationally practical since, for singularly perturbed problems, the value of ε is typically very small, and so $\varepsilon^2 N^{-4}$ is *extremely* small. Instead, the stopping criterion derived in [72, §4.6] uses the *preconditioned residual* approach with which we can approximate $\|e^{(k)}\|_A$ by the inner product of $r^{(k)}$ with the preconditioned residual, $z^{(k)} = M^{-1}r^{(k)}$, as follows

$$(z^{(k)})^T r^{(k)} \approx \|e^{(k)}\|_A^2,$$

when M is a good preconditioner of A.

If ε is so large so that $\tau_x = 1/2$ in (4.5), then the mesh is uniform, and classical analysis shows that the error, in the maximum norm, is $\mathcal{O}(N^{-2})$. If $\tau_x < 1/2$, then, as given in (1.21), the error is $\mathcal{O}(N^{-2}\ln^2 N)$. Using this preconditioned residual approach (up to the assumption relating $\|e^{(k)}\|$ and $\|e^{(k)}\|_2$, see details in [72, §4.6]), to achieve these bounds we iterate until

$$(z^{(k)})^T r^{(k)} \leq K \begin{cases} N^{-4}, & \tau_x = 1/2, \\ \varepsilon N^{-2} \ln^3 N, & \tau_x < 1/2, \end{cases} \tag{4.25}$$

where K is a user-chosen constant, determined experimentally for a particular preconditioner (but independent of ε and N). For example, for the experiments reported on below, we took $K = 0.5$ for diagonally-preconditioned CG, and $K = 1$ for IC(0)-preconditioned CG.

We first consider the application of unpreconditioned CG. The approach that leads to (4.25) makes sense only when a good preconditioner is used. Since we only wish to demonstrate that the unpreconditioned CG algorithm is not suitable of finite difference methods on a fitted mesh, we take the (artificial) approach of iterating until $\|e^{(k)}\|$ is less than discretization accuracy, or until 5000 iterations are reached. The resulting iteration counts are shown in Table 4.7. As expected, given Theorem 4.1,

ϵ^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	35	77	163	346	734	1547
10^{-2}	17	40	88	199	428	927
10^{-4}	25	34	65	122	235	452
10^{-6}	59	129	223	361	716	1441
10^{-8}	73	209	578	1317	2278	4150
10^{-10}	86	287	876	2602	5000	5000
10^{-12}	90	332	1110	3640	5000	5000

Table 4.7: Iteration counts for unpreconditioned CG applied to Example 4.1.

the unpreconditioned algorithm performs poorly when the diffusion parameter tends to zero.

Next we consider the diagonally preconditioned CG algorithm, as analyzed in Section 4.4. The iteration counts for this approach are shown in Table 4.8. As expected, given Theorem 4.2, this algorithm *is* robust with respect to ε. As ε decreases, and the mesh transitions from being uniform to piecewise uniform, the number of iterations needed decreases; thereafter, it is uniform in ε. The initial decrease in the number of iterations needed is due to several factors:

(a) as seen in Table 1.3, the discretization error initially increases due to reduction in the order of convergence from $\mathcal{O}(N^{-2})$ to $\mathcal{O}(N^{-2}\ln^2 N)$;

(b) the change in the termination criterion, as given in (4.25);

(c) and, for small ε, the submatrix of A associated with the interior region of the domain resembles the discretization on a uniform mesh with $\varepsilon \ll N^{-1}$. Recall from the discussion leading to Table 4.2 that this means it has a particularly simple structure.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	41	91	200	434	918	1926
10^{-2}	23	55	126	288	639	1376
10^{-4}	11	19	36	71	140	265
10^{-6}	11	19	35	69	136	258
10^{-8}	11	19	35	69	135	266
10^{-10}	11	19	35	69	147	287
10^{-12}	11	19	35	69	147	288

Table 4.8: Iteration counts for diagonal-preconditioned CG applied to Example 4.1.

Finally, we consider the IC(0)-preconditioned CG algorithm, as analyzed in Section 4.5. The bound for the condition number given in Theorem 4.4 is the same as for the diagonal preconditioner. In particular, it is robust with respect to ε. However, as shown in Table 4.9, IC(0) is considerably more efficient. Of course, the method is not as efficient (in terms of number of iterations) as the multigrid-based boundary-layer preconditioner devised in [72]; comparing the results in Table 4.9 with those in [72, Table 4.7], we see far fewer iterations are required by the boundary-layer preconditioner, and its performance is essentially independent of N. However, the IC(0) approach is far simpler to implement, and is widely supported on standard platforms. Furthermore, it does not require any *a priori* knowledge of the location of the layers, and so could be easily applied, for example, to a problem featuring interior layers.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	14	31	65	139	292	626
10^{-2}	8	17	39	85	188	405
10^{-4}	3	5	10	19	41	83
10^{-6}	3	5	9	17	34	69
10^{-8}	3	5	9	17	34	67
10^{-10}	3	5	9	17	34	67
10^{-12}	3	5	9	17	34	67

Table 4.9: Iteration counts for IC(0)-preconditioned CG applied to Example 4.1.

To verify that the algorithm is not under-solving the linear system, in Table 4.10 we give the maximum pointwise error in the approximation $U^{(k)}$ to u, where we have underlined the digits that agree with the discretization error, $\|u - U^N\|$. Arguably, the algorithm is over-solving the system for small ε and some further efficiency may be gained by refining the choice of K in (4.25).

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	6.17255e-03	1.55463e-03	3.90418e-04	9.76186e-005	2.44107e-05	6.10274e-06
10^{-2}	6.36057e-02	1.70981e-02	4.35619e-03	1.09850e-03	2.74954e-04	6.87776e-05
10^{-4}	9.03240e-02	3.77818e-02	1.43853e-02	4.96686e-03	1.64828e-03	5.61645e-04
10^{-6}	9.08187e-02	3.82459e-02	1.46879e-02	5.10491e-03	1.69172e-03	5.37035e-04
10^{-8}	9.08455e-02	3.82598e-02	1.47019e-02	5.11603e-03	1.69189e-03	5.37033e-04
10^{-10}	9.08480e-02	3.82613e-02	1.47027e-02	5.11639e-03	1.69209e-03	5.37164e-04
10^{-12}	9.08482e-02	3.82614e-02	1.47027e-02	5.11642e-03	1.69210e-03	5.37168e-04

Table 4.10: Errors corresponding to Table 4.9.

An optimally preconditioned matrix should have all its eigenvalues clustered around 1. To investigate how the matrices A, A_D and A_M differ from this ideal, in Figure 4.6

below we give semi-log plots of the normalised spectrum (i.e., with the eigenvalues ordered so that $\lambda_i \leq \lambda_{i+1}$, we plot $\kappa_i = \lambda_i/\lambda_1$ for $i = 1, \ldots, (N-1)^2$; note that $\kappa = \kappa_{(N-1)^2}$) for $N = 64$. When $\varepsilon^2 = 1$ (left), the diagonal of A is constant, and so the normalised spectra of A and A_D are the same. The normalised spectrum of A_M is noticeably smaller. When ε is small (right) the normalised spectrum of the unpreconditioned matrix is much larger than the other two.

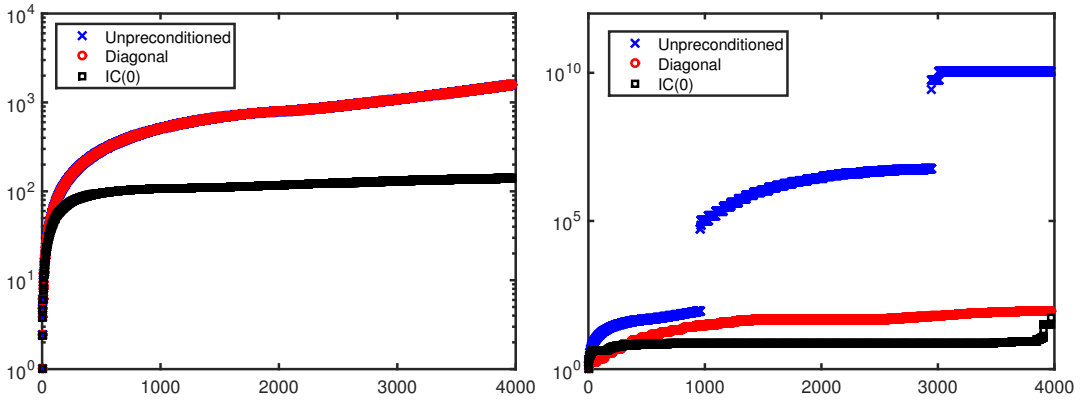

Figure 4.6: Normalized spectra of A, A_D and A_M when $N = 2^6$, $\varepsilon^2 = 1$ (left), and $\varepsilon^2 = 10^{-12}$ (right).

To compare the two preconditioned matrices more closely, in Figure 4.7 we show the normalised spectra of just A_D and A_M. In spite of the large aspect ratio for the mesh rectangles when ε is small (right) we see that the normalised spectra are smaller, and more uniform.

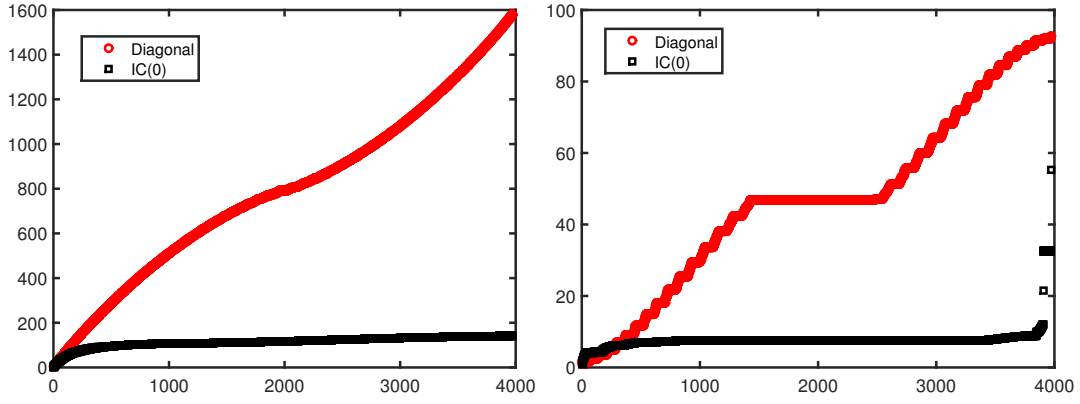

Figure 4.7: Normalized spectra of A_D and A_M when $N = 2^6$, $\varepsilon^2 = 1$ (left), and $\varepsilon^2 = 10^{-12}$ (right).

In Figure 4.8, we show the 2-norm of the residual at each iteration of unpreconditioned CG, and the diagonal and IC(0) preconditioned algorithms, taking $N = 64$ and $\varepsilon^2 = 10^{-6}$. From this, we can observe the slow and erratic convergence of unprecon-

ditioned CG. Although both preconditioners lead to fast and smooth convergence, the IC(0) is dramatically superior to diagonal preconditioning.

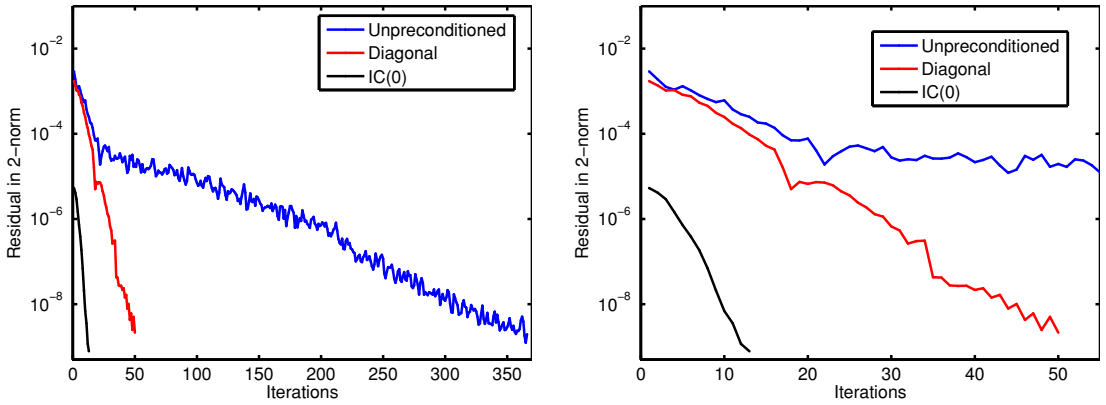

Figure 4.8: Residual reduction for $N = 2^6$ and $\varepsilon^2 = 10^{-6}$.

We conclude with results of experiments for Example 4.2, which has an edge layer, but no corner layer. As mentioned, diagonal preconditioning leads to the same bound for $\kappa_2(A_D)$ as in the two-layer case. In practice, the number of iterations required are only slightly less than as reported in Table 4.8. In contrast, Table 4.11 shows that very few iterations are required when the IC(0)-preconditioned CG algorithm is applied to the discretization on $\Omega_Q^{N,N}$ (compare with Table 4.9). This agrees with the theoretical result in Corollary 4.2.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	15	31	65	138	295	619
10^{-2}	8	17	38	85	186	401
10^{-4}	2	2	4	9	18	38
10^{-6}	1	2	2	3	4	6
10^{-8}	1	1	1	2	2	2
10^{-10}	1	1	1	1	2	2
10^{-12}	1	1	1	1	1	2

Table 4.11: Iteration counts for IC(0)-preconditioned CG applied to Example 4.2.

4.7 Conclusion

The numerical solution of linear systems arising from the discretization of singularly perturbed problems is not trivial: as shown in [72] even the performance of direct solvers degrades for small ε. Therefore, iterative schemes are needed. However, the

fitted meshes used to resolve boundary layers lead to linear systems that are poorly conditioned when ε is small, and so a good preconditioner is even more important than in the case when ε is $\mathcal{O}(1)$. We have proven that the diagonal and incomplete Cholesky preconditions are good choices in the sense that they are robust with respect to ε. Our numerical experiments demonstrate that they can be more efficient when ε is small, compared to when $\varepsilon = 1$. Finally, when the problem features no corner layers, we have shown that IC(0) is clearly the preconditioner of choice.

Arguably, however, when the corner layer is present, this scheme is suboptimal when compared with a multigrid-based preconditioner. Therefore, the final major component of this thesis will be the design and analysis of a multigrid-based boundary layer preconditioner.

Chapter 5

Boundary layer preconditioners for finite element discretizations applied to reaction-diffusion problems

In this chapter we consider the iterative solution of linear systems of equations arising from the discretization of singularly perturbed reaction-diffusion differential equations by a finite element method. As we have previously discussed, discretizations on layer-adapted meshes tend to lead to very ill-conditioned matrices. Therefore, careful design of a suitable preconditioner is necessary in order to solve these linear systems in a way that is robust, with respect to the perturbation parameter, and efficient. We propose a *boundary layer preconditioner* for a one-dimensional problem, in the style of that used for a finite difference method in [72]. We prove the optimality of this preconditioner, and establish suitable stopping criteria. Numerical results are presented which demonstrate that the ideas extend to problems in two dimensions.

The material in this chapter is based on: Scott MacLachlan, Niall Madden and Thái Anh Nhan, *Boundary layer preconditioners for finite element discretizations of reaction-diffusion problems in one and two dimensions*, in preparation.

5.1 Introduction

We consider the solution of linear systems of equations, by iterative methods, that arise in the discretizations of the singularly perturbed reaction-diffusion differential

equations (1.5) and (1.7) by *finite element* methods. For the completeness of this chapter, we briefly recall the model problems in one and two dimensions:

$$-\varepsilon^2 u'' + b(x)u = f(x), \quad \text{on} \quad \omega_x := (0,1), \qquad u(0) = u(1) = 0, \tag{5.1}$$

and

$$-\varepsilon^2 \Delta u + b(x,y)u = f(x,y), \quad \text{on} \quad \Omega := (0,1)^2, \qquad u(\partial\Omega) = 0. \tag{5.2}$$

For the analysis in this chapter, we will also assume that there are positive constants β_0 and β_1 such that, at all points in $\bar{\omega}_x$ and $\bar{\Omega}$, we have $0 < \beta_0^2 \leq b \leq \beta_1^2$.

The present chapter is motivated and inspired by the ideas in [72], which proposed and analyzed a robust *boundary layer preconditioner* for a finite difference discretization of reaction-diffusion problems in one and two dimensions on layer-adapted meshes. We wish to extend this approach to finite element discretizations. This presents a number of challenges, in particular

(a) In a finite difference discretization, the zero-order term contributes only to the diagonal entries of the system matrix. Away from boundaries, this term dominates, and so the application of a diagonal preconditioner in this region is both natural and easy to analyze. However, in the finite element method, the corresponding term contributes nonzero off-diagonal terms to the system matrix, so the choice of diagonal preconditioner for the interior region is not straightforward.

(b) For two-dimensional problems, the finite difference method has a five-point stencil, whereas the finite element method has a nine-point stencil, again complicating the method and its analysis.

(c) The condition number of the (unsymmetrised) linear systems yielded by finite difference methods on boundary layer-adapted meshes for Problems (5.1) and (5.2) is independent of ε (see [90, Remark 2]). In contrast, the condition number of the finite element discretization depends badly on ε.

For these reasons, in this chapter we focus our analyses on the one-dimensional problem in Section 5.2. We derive the estimate of the condition number of the finite element discretization matrix on a boundary layer-adapted mesh in Section 5.2.1. The analysis of the boundary layer preconditioner is given in Section 5.2.2. Section 5.2.3 contains the derivation of stopping criteria associated with the energy and maximum norms. The results of numerical experiments showing the optimality of the scheme are reported in Section 5.2.4.

In Section 5.3, we show how these ideas for the one-dimensional problem can be extended to the two-dimensional case. The condition number estimate of the system

matrix is given in Section 5.3.1. The design of a boundary layer preconditioner is discussed in Section 5.3.2. Section 5.3.3 presents a stopping criterion and numerical results. Finally, some concluding remarks are made in Section 5.4.

5.2 One-dimensional problems

We discretize (5.1) by a finite element method with linear elements, as described in Section 1.6.2. That is, we seek to find $u \in H_0^1(\omega_x)$ such that

$$B_\varepsilon(u, v) := \varepsilon^2(u', v') + (bu, v) = (f, v), \quad \text{for all } v \in H_0^1(\omega_x).$$

In this chapter we use u^N to denote the finite element solution to the problem (5.1). Recall from Section 1.6.2, the energy norm is

$$\|u\|_\varepsilon := \sqrt{\varepsilon^2 \|u'\|_0^2 + \beta_0^2 \|u\|_0^2}.$$

Then, using standard finite element analysis arguments, one can show that the following quasi-optimality result holds: there is a constant C, which is independent of ε, such that

$$\|u - u^N\|_\varepsilon \leq C\|u - v^N\|_\varepsilon, \quad \text{for all } v^N \in V^N, \tag{5.3}$$

where V^N is the space of piecewise linear functions on the mesh ω_x^N. Therefore, the error analysis is purely dependent on the approximation properties of the space V^N, and it is sufficient to prove a bound for $\|u^I - u^N\|_\varepsilon$, where u^I is the nodal interpolant to the solution of problem (5.1) in V^N.

Our main interest, however, is the resulting linear system coming from the above finite element discretization. This system is symmetric positive definite and can be written as

$$Au^N = f^N, \tag{5.4}$$

where $A = S + M$,

$$S = \begin{bmatrix} -\dfrac{\varepsilon^2}{h_i} & \dfrac{\varepsilon^2}{h_i} + \dfrac{\varepsilon^2}{h_{i+1}} & -\dfrac{\varepsilon^2}{h_{i+1}} \end{bmatrix}, \tag{5.5a}$$

and

$$M = \begin{bmatrix} \dfrac{h_i b_i}{6} & \dfrac{h_i b_i + h_{i+1} b_{i+1}}{3} & \dfrac{h_{i+1} b_{i+1}}{6} \end{bmatrix}, \tag{5.5b}$$

using the same notation as in Section 1.4, i.e., $h_i = x_i - x_{i-1}$, $i = 1, \ldots, N$.

Numerous fitted meshes had been proposed for this problem, the most commonly studied ones being the piecewise uniform mesh of Shishkin (see Section 1.7.1), and the graded mesh of Bakhvalov (see Section 1.7.2). These meshes have been discussed at length in Section 1.7, and so we do not repeat the details here. For the analysis later

in this chapter, we denote the equidistant mesh width of the interior region of these fitted meshes by h_I, which is $\mathcal{O}(N^{-1})$ when $\varepsilon \ll 1$, and set

$$h_{\min} = \min_{1\le i\le N}\{h_i\}.$$

Throughout this chapter, we assume that $\varepsilon \le N^{-1}$, which is usually the case of singularly perturbed problems.

A unified treatment of Shishkin and Bakhvalov meshes, and related meshes, is given by Linß [61]. In particular, in [61, §2.2], a mesh characterisation function is defined that, for the problem and method considered here, is

$$\vartheta_{rd}^{[p]}(\omega_x^N) := \max_{i=1,\dots,N} \int_{x_{i-1}}^{x_i} \left(1 + \varepsilon^{-1}e^{-\beta_0 s/(p\varepsilon)} + \varepsilon^{-1}e^{-\beta_0(1-s)/(p\varepsilon)}\right) ds, \tag{5.6}$$

where p is a parameter coming from the method and which depends on the formal order of the scheme. The characterisation function is the crucial quantity in the following approximation result.

Theorem 5.1 ([61, Theorem 6.2])**.** *If u^I denotes the piecewise linear interpolant of u on an arbitrary mesh ω_x^N to the solution of* (5.1)*, then there is a constant, C, independent of ε, such that*

$$\|u - u^I\|_\varepsilon \le C\big(\varepsilon^{1/2} + \vartheta_{rd}^{[p]}(\omega_x^N)\big)\vartheta_{rd}^{[p]}(\omega_x^N).$$

Using this result, the convergence properties of a specially constructed mesh can be established. For the Shishkin mesh, one can invoke (5.3), (5.6), and Theorem 5.1 to show that

$$\|u - u^N\|_\varepsilon \le C\big(\varepsilon^{1/2}N^{-1}\ln N + N^{-2}\ln^2 N\big). \tag{5.7}$$

Based on this estimate, one might expect that, if $\varepsilon \ll N^{-1}$, this bound would simplify to $N^{-2}\ln^2 N$. However, the $\varepsilon^{1/2}N^{-1}\ln N$ term stems from first-order term in the energy norm, and tends to dominate. To see this, let us consider the following example.

Example 5.1. We apply the finite element discretization to a one-dimensional reaction-diffusion problem:

$$-\varepsilon^2 u'' + u = e^x, \quad x \in \omega_x, \quad u(0) = u(1) = 0. \tag{5.8}$$

Table 5.1 shows computed errors, in the energy norm, by the finite element method on a Shishkin mesh. This agrees with the error estimate (5.7). Note that, for small ε, the error is $\mathcal{O}(\varepsilon^{1/2}N^{-1}\ln N)$.

For the Bakhvalov mesh, we have $\vartheta_{rd}^{[p]}(\omega_x^N) \le CN^{-1}$, and thus it follows that

$$\|u - u^N\|_\varepsilon \le C(\varepsilon^{1/2}N^{-1} + N^{-2}). \tag{5.9}$$

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	3.7559e-03	1.8779e-03	9.3897e-04	4.6948e-04	2.3474e-04	1.1737e-04
10^{-2}	1.4485e-02	7.2427e-03	3.6214e-03	1.8107e-03	9.0535e-04	4.5268e-04
10^{-4}	1.7910e-02	1.0241e-02	5.7617e-03	3.2012e-03	1.7607e-03	9.6038e-04
10^{-6}	5.6642e-03	3.2388e-03	1.8222e-03	1.0124e-03	5.5683e-04	3.0373e-04
10^{-8}	1.7913e-03	1.0242e-03	5.7623e-04	3.2015e-04	1.7609e-04	9.6048e-05
10^{-10}	5.6665e-04	3.2390e-04	1.8222e-04	1.0124e-04	5.5683e-05	3.0373e-05
10^{-12}	1.7985e-04	1.0250e-04	5.7632e-05	3.2016e-05	1.7609e-05	9.6048e-06

Table 5.1: $\|u - u^N\|_\varepsilon$ with u defined in (5.8) approximated by a FEM on a Shishkin mesh.

Theoretical analysis of the finite element methods in the maximum norm is outside the scope of this thesis. However, it is known that the method presented here is (almost) second-order convergent pointwise. That is, if $g_\infty(\omega_x^N)$ denotes the discretization error in the maximum norm, then

$$g_\infty(\omega_x^N) = \begin{cases} CN^{-2}\ln^2 N, & \omega_x^N \text{ is a Shishkin mesh,} \\ CN^{-2}, & \omega_x^N \text{ is a Bakhvalov mesh.} \end{cases} \tag{5.10}$$

For maximum norm estimates of finite element methods for reaction-diffusion problems, but on quasi-uniform meshes, see, e.g., [56, 98]. Related works for so-called balanced norm may be found in [1, 76, 93, 99]. In particular, in [76] it is shown how to extend certain results for balanced norms to the maximum norm. Table 5.2 shows the errors in the maximum norm computed by the finite element method on a Shishkin mesh. It clearly shows that the convergence is of $\mathcal{O}(N^{-2}\ln^2 N)$, independently of ε.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	8.8508e-07	2.2127e-07	5.5319e-08	1.3834e-08	3.4744e-09	9.3232e-10
10^{-2}	2.4960e-04	6.2399e-05	1.5598e-05	3.8996e-06	9.7489e-07	2.4372e-07
10^{-4}	3.8474e-03	1.2533e-03	3.9601e-04	1.2215e-04	3.6952e-05	1.0994e-05
10^{-6}	3.8483e-03	1.2536e-03	3.9610e-04	1.2218e-04	3.6961e-05	1.0997e-05
10^{-8}	3.8483e-03	1.2536e-03	3.9610e-04	1.2218e-04	3.6961e-05	1.0997e-05
10^{-10}	3.8483e-03	1.2536e-03	3.9610e-04	1.2218e-04	3.6961e-05	1.0997e-05
10^{-12}	3.8483e-03	1.2536e-03	3.9610e-04	1.2218e-04	3.6961e-05	1.0997e-05

Table 5.2: $\|u - u^N\|_{\omega_x^N}$ with u defined in (5.8) approximated by a FEM on a Shishkin mesh.

In next section, we propose and analyze a preconditioner for the linear system (5.4) arising from the finite element solution of the one-dimensional reaction-diffusion problem. As discussed above, special fitted meshes are required in order to resolve the boundary layers. Therefore, we show in Section 5.2.1 that the condition number of the matrix in (5.4) depends badly on ε. Consequently, an unpreconditioned iterative solver

is not robust, in the sense that, as $\varepsilon \to 0$, the number of iterations required increases (see also the discussion in Chapter 4).

So our goal is to propose a block-structured preconditioner that is suitable for layer-adapted meshes, which is motivated from the physical distribution of points in the meshes, and which is optimal, in the sense of spectral equivalence.

5.2.1 Condition number estimate

In general, layer-adapted meshes have intervals of width $\mathcal{O}(N^{-1}\varepsilon)$ near the boundaries. As we now show, the condition number of the unpreconditioned discrete system arising from the finite element discretization on such a mesh is unbounded as $\varepsilon \to 0$. This is similar to that of the unpreconditioned discrete system coming from the symmetrized finite difference discretization on boundary layer adapted meshes (see Section 4.3).

Recall that the condition number of the matrix A, associated with the 2-norm, is $\kappa_2(A) := \|A\|_2\|A^{-1}\|_2$. By examining the entries of A as defined in (5.5a), (5.5b), and applying Geršgorin's Theorem, we easily see that

$$\|A\|_2 = \lambda_{\max} \leq (\beta_1^2 h_I + 4\varepsilon^2/h_I) \leq CN^{-1}. \tag{5.11}$$

In addition, we can bound the smallest eigenvalue of A, $\lambda_{\min}$, from below by Geršgorin's Theorem, giving

$$\lambda_{\min} \geq \min_i \left\{ \frac{h_i b_i + h_{i+1} b_{i+1}}{6} \right\} \geq \frac{\beta_0^2 h_{\min}}{3}. \tag{5.12}$$

A combination of (5.11) and (5.12) implies the following estimate.

Theorem 5.2. *Let A be the matrix associated with the linear system* (5.4)*. Then, there is a constant C, independent of both N and ε, such that*

$$\kappa_2(A) \leq C/(Nh_{\min}).$$

In practice, one finds that this bound is quite sharp for small ε, and the associated constant is $\mathcal{O}(1)$. Therefore, as $\varepsilon \to 0$ the system (5.4) is ill-conditioned. In particular, for the Shishkin mesh, $h_{\min} \sim \varepsilon \ln N/(N\beta)$, implying that $\kappa_2(A) \leq C(\varepsilon \ln N)^{-1}$. In Table 5.3, it is shown that this bound is sharp, for sufficiently small ε, with $C \approx 3$. For the Bakhvalov mesh, $h_{\min} \sim \varepsilon/(N\beta)$, and so $\kappa_2(A) \leq C\varepsilon^{-1}$. As shown in Table 5.4, this bound is also sharp, with $C \approx 1$.

Remark 5.1 (Bound on $\kappa_\infty(A)$)**.** *The condition numbers associated with the maximum-norm of the linear system obtained by unsymmetrised finite difference discretizations of problems* (5.1) *and* (5.2) *are independent of ε (see [90, Remark 2]). In contrast,*

ε^2	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$
1	1.16e+02	4.64e+02	1.85e+03	7.42e+03	2.97e+04	1.19e+05
10^{-2}	1.04e+01	4.07e+01	1.62e+02	6.47e+02	2.59e+03	1.03e+04
10^{-4}	1.54e+01	2.07e+01	5.82e+01	1.72e+02	5.30e+02	1.68e+03
10^{-6}	1.59e+02	1.35e+02	1.16e+02	1.72e+02	5.30e+02	1.68e+03
10^{-8}	1.59e+03	1.35e+03	1.16e+03	1.01e+03	8.94e+02	1.68e+03
10^{-10}	1.59e+04	1.36e+04	1.17e+04	1.01e+04	8.95e+03	7.98e+03
10^{-12}	1.59e+05	1.36e+05	1.17e+05	1.01e+05	8.95e+04	7.98e+04

Table 5.3: $\kappa_2(A)$ of the problem (5.1) discretized by a FEM on a Shishkin mesh.

ε^2	$N=16$	$N=32$	$N=64$	$N=128$	$N=256$	$N=512$
1	1.16e+02	4.64e+02	1.85e+03	7.42e+03	2.97e+04	1.19e+05
10^{-2}	1.04e+01	4.36e+01	1.86e+02	7.67e+02	3.12e+03	1.26e+04
10^{-4}	1.06e+01	2.80e+01	1.19e+02	4.92e+02	2.00e+03	8.07e+03
10^{-6}	1.23e+02	1.22e+02	1.22e+02	4.92e+02	2.00e+03	8.07e+03
10^{-8}	1.26e+03	1.25e+03	1.25e+03	1.25e+03	2.00e+03	8.07e+03
10^{-10}	1.26e+04	1.25e+04	1.25e+04	1.25e+04	1.25e+04	1.25e+04
10^{-12}	1.26e+05	1.25e+05	1.25e+05	1.25e+05	1.25e+05	1.25e+05

Table 5.4: $\kappa_2(A)$ of the problem (5.1) discretized by a FEM on a Bakhvalov mesh.

the condition numbers of these linear systems arising from finite element discretizations are dependent on ε. To see this, by direct inspection of the entries of A, we get $\|A\| \leq CN^{-1}$. We next want to give an upper bound for $\|A^{-1}\|$. The argument used in [72] relies on A being an M-matrix, which is not the case here. Nevertheless, we can still bound $\|A^{-1}\|$ by exploiting the diagonal dominance property and using the Varah's Theorem (see Lemma 4.3). Then, for the matrix A as defined in (5.4)*, we easily see that*

$$\alpha = \min_i \left\{ \frac{h_i b_i + h_{i+1} b_{i+1}}{6} \right\} \geq \frac{\beta_0^2 h_{\min}}{3}.$$

It follows then that

$$\|A^{-1}\| \leq 3/(\beta_0^2 h_{\min}). \tag{5.13}$$

Hence, the following estimates hold true

$$\kappa_\infty(A) := \|A\|\|A^{-1}\| \leq \begin{cases} C(\varepsilon \ln N)^{-1}, & \omega_x^N \text{ is a Shishkin mesh,} \\ C\varepsilon^{-1}, & \omega_x^N \text{ is a Bakhvalov mesh.} \end{cases}$$

5.2.2 Boundary layer preconditioners

The difficulties associated with solving (5.4) arise from the transition from a mesh that is very fine in the region close to the boundary, to one that is course and uniform in

the interior. The scaling of the problems in these regions is very different, and so it is natural to precondition them differently. Close to the boundaries, the linear system resembles that of a classical problem (i.e., one that is not singularly perturbed), and so it is amenable to solution using standard techniques, such as multigrid methods. In the interior, the entries are dominated by the contribution from the reaction term, so we employ a diagonal scaling in this region.

Motivated by the ideas in [72], we partition the mesh into two pieces:

- the boundary region including the end points of the left and right layer regions, $\omega_B^N := \omega_x^N \cap ([0, \tau_x] \cup [1 - \tau_x, 1])$, where the subscript B denotes the boundary region, and
- the interior (not including the transition points), $\omega_I^N := \omega_x^N \backslash \omega_B^N$, where the subscript I denotes the interior region.

Then, the re-ordered mesh is denoted by $\omega^N := [\omega_B^N \;\; \omega_I^N]$. This ordering is also used to partition the matrix A as

$$A = \begin{pmatrix} A_{BB} & A_{BI} \\ A_{IB} & A_{II} \end{pmatrix}. \tag{5.14}$$

As we explain below, the matrix A_{II} may be approximated, in the spectral sense, by a suitably chosen diagonal matrix. Furthermore, the submatrices A_{BI} and A_{IB} contain only two nonzero entries each, and make only a very modest contribution to the system. Following from these observations, we approximate A in the following way. Recall the mass matrix, M, defined in (5.5b). Let

$$D_{II} = m\text{diag}(M_{II}), \tag{5.15}$$

where m is a positive parameter whose choice depends on the analysis. Define

$$A_D = \begin{pmatrix} A_{BB} & 0 \\ 0 & D_{II} \end{pmatrix}. \tag{5.16}$$

Recall from Section 5.1 that we have assumed that $0 < \beta_0^2 \leq b(x) \leq \beta_1^2$ for all $x \in [0, 1]$. Let

$$\delta_h = \left(\varepsilon/(h_I \beta_0)\right)^2,$$

and also set

$$\gamma := \frac{\beta_1^2}{\beta_0^2 + \beta_1^2}. \tag{5.17}$$

Since we are interested in the case where $\varepsilon \ll 1$, we also have that $\delta_h \ll 1$.

Theorem 5.3. *Let A by the system matrix in (5.4) for the finite element solution on a layer adapted mesh, and let A_D be as defined in (5.16). Let m be the parameter in (5.15) and q be any number such that $\gamma q/2 < m < q/2$. Then, for all vectors V,*

$$\left(\theta_q - \frac{3}{m}\sqrt{2\gamma}\delta_h - \frac{9}{m}\delta_h^2\right) V^T A_D V \leq V^T A V \leq \left(1 + \frac{3}{2m} + \frac{6}{m}\delta_h + \frac{9}{m}\delta_h^2\right) V^T A_D V, \tag{5.18}$$

where

$$\theta_q = \min\left\{1 - \frac{\gamma q}{2m}, \frac{1}{2m} - \frac{1}{q}\right\} > 0.$$

Proof. In the same way that we partitioned $\omega^N = [\omega_B^N \quad \omega_I^N]$, we partition a generic vector V as $V = [V_B \quad V_I]^T$, and note that

$$V^T A V^T = V_B^T A_{BB} V_B + 2V_B^T A_{BI} V_I + V_I^T A_{II} V_I,$$

and

$$V^T A_D V^T = V_B^T A_{BB} V_B + V_I^T D_{II} V_I.$$

Therefore, we require bounds for $V_I^T A_{II} V_I$ and $V_B^T A_{BI} V_I$.

Firstly, to bound $V_I^T A_{II} V_I$, we easily see that

$$\frac{1}{2m} V_I^T D_{II} V_I \leq V_I^T A_{II} V_I, \tag{5.19}$$

since $V_I^T \left(A_{II} - \frac{1}{2m} D_{II}\right) V_I \geq 0$ for all V_I.

Since $A = S + M$, the matrices S and M can be partitioned in the same way as (5.14), and so $A_{II} = S_{II} + M_{II}$. By Geršgorin's Theorem, S_{II} can be bounded, in the sense of spectral equivalence, by the diagonal matrix whose nonzero entries are $(4\varepsilon^2/h_I)$. Also,

$$\frac{4\varepsilon^2}{h_I} = \frac{4\varepsilon^2 h_I \beta_0^2}{h_I^2 \beta_0^2} \leq \delta_h \frac{6}{m} \frac{m h_I (b_i + b_{i+1})}{3}.$$

So, for any V_I,

$$V_I^T S_{II} V_I \leq \delta_h \frac{6}{m} V_I^T D_{II} V_I.$$

By Geršgorin's Theorem again, M_{II} can be bounded by the diagonal matrix whose entries are

$$\frac{h_I(b_i + b_{i+1})}{2} = \frac{3}{2m}\left(\frac{m h_I (b_i + b_{i+1})}{3}\right).$$

Hence,

$$V_I^T M_{II} V_I \leq \frac{3}{2m} V_I^T D_{II} V_I.$$

Thus, combing this with (5.19), we get that

$$\frac{1}{2m} V_I^T D_{II} V_I \leq V_I^T A_{II} V_I \leq \frac{1}{m}\left(\frac{3}{2} + 6\delta_h\right) V_I^T D_{II} V_I.$$

We proceed to finding upper and lower bounds for $V_B^T A_{BI} V_I$. Set $D_{BB} = m\text{diag}(M_{BB})$. By the Cauchy-Schwarz inequality,

$$\left|V_B^T A_{BI} V_I\right| \leq \left\|D_{BB}^{-1/2} A_{BI} V_I\right\|_2 \left\|D_{BB}^{1/2} V_B\right\|_2,$$

for any V_B and V_I. Now, since $V_B^T D_{BB} V_B \leq 2m V_B^T A_{BB} V_B$ for all V_B,

$$\left|V_B^T A_{BI} V_I\right| \leq \sqrt{2m} \left\|D_{BB}^{-1/2} A_{BI} V_I\right\|_2 \left(V_B^T A_{BB} V_B\right)^{1/2}, \tag{5.20}$$

for any V_B and V_I.

To bound $\left\|D_{BB}^{-1/2} A_{BI} V_I\right\|_2$, we use that

$$\left\|D_{BB}^{-1/2} A_{BI} V_I\right\|_2 = \left\|D_{BB}^{-1/2}(S_{BI} + M_{BI}) V_I\right\|_2 \leq \left\|D_{BB}^{-1/2} S_{BI} V_I\right\|_2 + \left\|D_{BB}^{-1/2} M_{BI} V_I\right\|_2.$$

There are only two nonzero entries in each of S_{BI} and S_{IB}. They are in the first and last columns, and on different rows. So $S_{IB} D_{BB}^{-1} S_{BI}$ has only two nonzero entries,

$$s_1 := \frac{3\varepsilon^4}{m(h_{N/4} b_{N/4} + h_{N/4+1} b_{N/4+1}) h_{N/4+1}^2},$$

and

$$s_2 := \frac{3\varepsilon^4}{m(h_{3N/4} b_{3N/4} + h_{3N/4+1} b_{3N/4+1}) h_{3N/4}^2},$$

which are the first and last entries on the diagonal, and with $h_{N/4+1} = h_{3N/4} = h_I$. Then, s_1 and s_2 can be bounded from above by

$$\frac{3\varepsilon^4}{h_I^3 m \beta_0^2} = \frac{9\varepsilon^4}{2m^2 h_I^4 \beta_0^4} \frac{2m h_I \beta_0^2}{3} \leq \delta_h^2 \frac{9}{2m^2} \frac{m h_I (b_{i+1} + b_{i+2})}{3},$$

for any i. Thus,

$$V_I^T S_{IB} D_{BB}^{-1} S_{BI} V_I \leq \delta_h^2 \frac{9}{2m^2} V_I^T D_{II} V_I,$$

and so,

$$\left\|D_{BB}^{-1/2} S_{BI} V_I\right\|_2 \leq \delta_h \frac{3}{m\sqrt{2}} \left\|D_{II}^{1/2} V_I\right\|_2.$$

We use a similar argument to bound the term involving M_{BI}. From the definition of γ in (5.17), we see that $1/2 \leq \gamma < 1$. Also, because

$$\frac{b_{i+1}}{b_i + b_{i+1}} \leq \frac{b_{i+1}}{\beta_0^2 + b_{i+1}} \leq \frac{\beta_1^2}{\beta_0^2 + \beta_1^2} = \gamma,$$

it follows that $b_{i+1} \leq \gamma(b_i + b_{i+1})$ for all i. Again we note that there are only two nonzero entries in each of M_{BI} and M_{IB}, so that there are only two nonzero entries in $M_{IB} D_{BB}^{-1} M_{BI}$, given by

$$m_1 := \frac{3 h_{N/4+1}^2 b_{N/4+1}^2}{36 m(h_{N/4} b_{N/4} + h_{N/4+1} b_{N/4+1})},$$

and

$$m_2 := \frac{3h_{3N/4}^2 b_{3N/4}^2}{36m(h_{3N/4}b_{N/4} + h_{3N/4+1}b_{3N/4+1})}.$$

Both m_1 and m_2 can be bounded from above by

$$\frac{h_I b_{i+1}}{12m} \leq \frac{\gamma}{4m^2}\frac{mh_I(b_{i+1}+b_{i+2})}{3}, \quad \text{for } i = N/4+1, 3N/4.$$

Thus,

$$V_I^T M_{IB} D_{BB}^{-1} M_{BI} V_I \leq \frac{\gamma}{4m^2} V_I^T D_{II} V_I,$$

and so,

$$\left\| D_{BB}^{-1/2} M_{BI} V_I \right\|_2 \leq \frac{\sqrt{\gamma}}{2m} \left\| D_{II}^{1/2} V_I \right\|_2.$$

Hence,

$$\left\| D_{BB}^{-1/2} A_{BI} V_I \right\|_2 \leq \left(\frac{3}{m\sqrt{2}}\delta_h + \frac{\sqrt{\gamma}}{2m} \right) \left\| D_{II}^{1/2} V_I \right\|_2.$$

Recalling (5.20), this gives that

$$\begin{aligned} \left| V_B^T A_{BI} V_I \right| &\leq \sqrt{2m} \left(\frac{3}{m\sqrt{2}}\delta_h + \frac{\sqrt{\gamma}}{2m} \right) \left(V_I^T D_{II} V_I \right)^{1/2} \left(V_B^T A_{BB} V_B \right)^{1/2}, \\ &= \frac{1}{\sqrt{m}} \left(3\delta_h + \frac{\sqrt{\gamma}}{\sqrt{2}} \right) \left(V_I^T D_{II} V_I \right)^{1/2} \left(V_B^T A_{BB} V_B \right)^{1/2}, \end{aligned}$$

for all V_B and V_I. Since

$$2ab \leq a^2/q + b^2 q, \tag{5.21}$$

for any real a, b, and $q > 0$, we have

$$2\left| V_B^T A_{BI} V_I \right| \leq \frac{1}{q} V_I^T D_{II} V_I + \frac{q}{2m} \left(3\sqrt{2}\delta_h + \sqrt{\gamma} \right)^2 V_B^T A_{BB} V_B, \tag{5.22}$$

for all V_B and V_I. Then, the lower bound for $V^T A V$ is

$$\begin{aligned} V^T A V &\geq V_B^T A_{BB} V_B - 2\left| V_B^T A_{BI} V_I \right| + V_I^T A_{II} V_I \\ &\geq V_B^T A_{BB} V_B - \frac{1}{q} V_I^T D_{II} V_I - \frac{q}{2m} \left(3\sqrt{2}\delta_h + \sqrt{\gamma} \right)^2 V_B^T A_{BB} V_B \\ &\quad + \frac{1}{2m} V_I^T D_{II} V_I \\ &\geq \left(1 - \frac{q}{2m} \left(3\sqrt{2}\delta_h + \sqrt{\gamma} \right)^2 \right) V_B^T A_{BB} V_B + \left(\frac{1}{2m} - \frac{1}{q} \right) V_I^T D_{II} V_I \\ &= \left(1 - \frac{q}{2m} \left(18\delta_h^2 + 6\sqrt{2\gamma}\delta_h + \gamma \right) \right) V_B^T A_{BB} V_B \\ &\quad + \left(\frac{1}{2m} - \frac{1}{q} \right) V_I^T D_{II} V_I \\ &\geq \left(\min\left\{ 1 - \frac{\gamma q}{2m}, \frac{1}{2m} - \frac{1}{q} \right\} - \frac{3}{m}\sqrt{2\gamma}\delta_h - \frac{9}{m}\delta_h^2 \right) V^T A_D V, \end{aligned} \tag{5.23}$$

in which q is chosen such that $\frac{\gamma q}{2} < m < \frac{q}{2}$ to guarantee $\theta_q > 0$ and maximize its value.

As for the corresponding upper bound, it is sufficient to choose $q = 1$ in (5.21), then

$$\begin{aligned} V^T AV &\leq V_B^T A_{BB} V_B + 2\left|V_B^T A_{BI} V_I\right| + V_I^T A_{II} V_I \\ &\leq V_B^T A_{BB} V_B + V_I^T D_{II} V_I + \frac{1}{2m}\left(3\sqrt{2}\delta_h + \sqrt{\gamma}\right)^2 V_B^T A_{BB} V_B \\ &\quad + \frac{1}{m}\left(\frac{3}{2} + 6\delta_h\right) V_I^T D_{II} V_I \\ &= \left(1 + \frac{\gamma}{2m} + \frac{3}{m}\sqrt{2\gamma}\delta_h + \frac{9}{m}\delta_h^2\right) V_B^T A_{BB} V_B \\ &\quad + \left[1 + \frac{3}{2m} + \frac{6}{m}\delta_h\right] V_I^T D_{II} V_I \\ &\leq \left[1 + \frac{3}{2m} + \frac{6}{m}\delta_h + \frac{9}{m}\delta_h^2\right] V^T A_D V. \end{aligned} \tag{5.24}$$

Combining (5.23) and (5.24) completes the proof. □

The maximum value of θ_q is achieved when

$$1 - \frac{\gamma q}{2m} = \frac{1}{2m} - \frac{1}{q}.$$

This gives

$$m = \frac{q(\gamma q + 1)}{2(q+1)}.$$

Hence,

$$\theta_{q,\max} = 1 - \frac{q+1}{2(\gamma q + 1)}.$$

In the next corollary, we give a particular lower bound for (5.18) when the reaction coefficient, b, is constant.

Corollary 5.1. *When b in (5.1) is constant, and so $\gamma = 1/2$, and thus, the maximum value of θ_q is achieved when*

$$m = \frac{q(q+2)}{4(q+1)}.$$

Taking $q = 1$, this gives $m = 3/8$, and from (5.18), we then obtain the following

$$\left(\frac{1}{3} - 8\sqrt{2\gamma}\delta_h - 24\delta_h^2\right) V^T A_D V \leq V^T AV \leq \left(5 + 16\delta_h + 24\delta_h^2\right) V^T A_D V. \tag{5.25}$$

Remark 5.2. *We have now established that A is spectrally equivalent to A_D, and so it follows that A_D is an excellent preconditioner for A, at least with respect to the condition (P1) of being a good preconditioner (see Section 1.8.3). However, it does not satisfy the condition (P2), since solving a system with A_{BB} as the coefficient matrix has the same computational complexity as one involving A. Nonetheless, this is a useful intermediate result towards obtaining an optimal preconditioner. This is because A_{BB},*

unlike the full matrix A, resembles that coming from a classical diffusion-dominated problem. There are many fast solvers for these problems, with multigrid methods being the most important. Therefore, it is quite reasonable to replace A_{BB} in A_D with a matrix that is spectrally equivalent to it, which we denote by M_{BB}. We make the idea precise in the following corollary.

Corollary 5.2. *Under the assumptions of Theorem 5.3, if M_{BB} is spectrally equivalent to A_{BB}, meaning that there exists constants c_0 and c_1 such that*

$$c_0 V_B^T M_{BB} V_B \leq V_B^T A_{BB} V_B \leq c_1 V_B^T M_{BB} V_B, \quad \text{for all } V_B,$$

then the matrix

$$A_M = \begin{pmatrix} M_{BB} & 0 \\ 0 & D_{II} \end{pmatrix}.$$

satisfies

$$C_0 V^T A_M V \leq V^T A V \leq C_1 V^T A_M V,$$

for all V where

$$C_1 = \max\left\{1 + \frac{3}{2m} + \frac{6}{m}\delta_h, c_1\left(1 + \frac{\gamma}{2m} + \frac{3}{m}\sqrt{2\gamma}\delta_h + \frac{9}{m}\delta_h^2\right)\right\},$$

and

$$C_0 = \min\left\{\frac{1}{2m} - \frac{1}{q}, c_0\left(1 - \frac{q}{2m}\left(18\delta_h^2 + 6\sqrt{2\gamma}\delta_h + \gamma\right)\right)\right\}.$$

5.2.3 Stopping criteria

Having constructed a boundary-layer preconditioner that is robust with respect to ε, we need to derive suitable stopping criteria for its application to problems posed on various layer-adapted meshes. The approach is similar in spirit to Section 4.6 (and thus, [72, §4.6]) which was concerned with finite difference approximations and maximum norm estimates. We now adapt that reasoning to the setting of finite element discretizations and energy norm estimates.

As in Section 4.6, we require any stopping criterion to be feasible, in the sense of not needing to compute a residual (say) that is comparable to machine epsilon. However, as we now show, this may not be possible for an unpreconditioned problem for the cases of interest, where $\varepsilon \ll N^{-1}$. This motivates the analysis of the preconditioned residual approach which, as our numerical experiments show, is effective.

Stopping criterion associated with the energy norm

Recall that u^N is the solution of the discrete problem (5.4). Let $u^{(k)}$ be the k^{th} iterate computed by the iterative procedure. Naturally, we wish to choose k so that $u^{(k)}$ is as good an approximation to u as u^N. That is, we would like

$$\|u - u^{(k)}\|_* \simeq \|u - u^N\|_*,$$

in some suitable norm. Since

$$\|u - u^{(k)}\|_* \leq \|u - u^N\|_* + \|u^N - u^{(k)}\|_*,$$

this means finding $u^{(k)}$ such that

$$\|u^N - u^{(k)}\|_* \leq \lambda \|u - u^N\|_*,$$

where λ is some moderately small positive constant, for example, $\lambda = 0.01$. We know that, for example on the Shishkin mesh:

$$\|u - u^N\|_\varepsilon \leq C(\varepsilon^{1/2} N^{-1} \ln N + N^{-2} \ln^2 N).$$

Of course, in practice, u^N is unknown, so we must estimate the solver error $e^{(k)} = u^N - u^{(k)}$. This can be done with the residual

$$r^{(k)} = b - Au^{(k)} = b - A(u^N - e^{(k)}) = Ae^{(k)},$$

giving

$$e^{(k)} = A^{-1} r^{(k)}.$$

Then,

$$\begin{aligned} \|e^{(k)}\|_\varepsilon = \|e^{(k)}\|_A &= \sqrt{(e^{(k)})^T A e^{(k)}} = \sqrt{(e^{(k)})^T A^T A^{-1} A e^{(k)}} \\ &= \|A^{-1/2} r^{(k)}\|_2 \leq \|A^{-1/2}\|_2 \|r^{(k)}\|_2 \leq \|A^{-1/2}\| \|r^{(k)}\|_2. \end{aligned} \tag{5.26}$$

This is because A is symmetric (and so is A^{-1}), implying that $\|A\|_2 \leq \|A\|$. $A^{1/2}$, a principle square root of A, is also symmetric and positive definite with $\|A^{-1/2}\|^2 = \|A^{-1}\|$. Recalling the bound for $\|A^{-1}\|$ in (5.13), the above calculation gives that,

$$\|e^{(k)}\|_A \leq \frac{\sqrt{3}\|r^{(k)}\|_2}{\beta_0 \sqrt{h_{\min}}}.$$

Let us denote the finite element discretization error the energy norm yielded by a particular mesh as $g_\varepsilon(\omega_x^N)$. Then, to ensure that this error is matched by the iterative solver, we need

$$\|r^{(k)}\|_2 \leq \frac{\beta_0 g_\varepsilon(\omega_x^N) \sqrt{h_{\min}}}{\sqrt{3}}.$$

In the context of the Shishkin and Bakhvalov meshes, this leads to

$$\|r^{(k)}\|_2 \leq \begin{cases} C(\varepsilon N^{-3/2}\ln^{3/2} N + \varepsilon^{1/2}N^{-5/2}\ln^{5/2} N), & \omega_x^N \text{ is a Shishkin mesh,} \\ C(\varepsilon N^{-3/2} + \varepsilon^{1/2}N^{-5/2}), & \omega_x^N \text{ is a Bakhvalov mesh.} \end{cases} \tag{5.27}$$

Since we are interested in the case where ε is very small, this demonstrates that the required residual reduction may not be feasible in a finite precision setting.

Instead, we shall use the standard stopping criterion for the preconditioned conjugate gradient algorithm. Let M be a good preconditioner for the matrix A in the sense that $MA \approx I$. Let

$$\tilde{r}^{(k)} = Mr^{(k)},$$

be the preconditioned residual. Then, the inner product of residual and preconditioned residual can be used to estimate $\|e^{(k)}\|_A$ because

$$\left(\tilde{r}^{(k)}\right)^T r^{(k)} = \left(e^{(k)}\right)^T AMAe^{(k)} \approx \|e^{(k)}\|_A^2. \tag{5.28}$$

Therefore, it is straightforward to see that the stopping criterion needed to guarantee $\|e^{(k)}\|_\varepsilon \leq g_\varepsilon(\omega_x^N)$ is

$$\sqrt{\left(\tilde{r}^{(k)}\right)^T r^{(k)}} \leq \begin{cases} C(\varepsilon^{1/2}N^{-1}\ln N + N^{-2}\ln^2 N), & \omega_x^N \text{ is a Shishkin mesh,} \\ C(\varepsilon^{1/2}N^{-1} + N^{-2}), & \omega_x^N \text{ is a Bakhvalov mesh.} \end{cases} \tag{5.29}$$

Stopping criterion associated with the maximum norm

To derive a suitable stopping criterion associated with the maximum norm, we make use of the standard stopping criterion for preconditioned conjugate gradient as discussed in [72, §4.6]. We use

$$\frac{1}{\sqrt{n}}\|e^{(k)}\|_2 \leq \|e^{(k)}\| \leq \|e^{(k)}\|_2,$$

for vectors of length n. We make an assumption similar to that of [72, §4.6], i.e., the error in the maximum norm can be approximated by that in the 2-norm,

$$\|e^{(k)}\| \approx \frac{c}{\sqrt{n}}\|e^{(k)}\|_2 \leq \frac{c}{\sqrt{n}}\|A^{-1/2}\|_2\|e^{(k)}\|_A,$$

with $n \approx N$ for the one-dimensional problem (5.1). Since A^{-1} is symmetric, we have $\|A^{-1}\|_2 \leq \|A^{-1}\|$. From (5.13), it is easy to see that $\|A^{-1/2}\|_2 \leq \sqrt{3}/(\beta\sqrt{h_{\min}})$. Then, we get $\|A^{-1/2}\|_2 \leq \sqrt{3}/(\beta_0\sqrt{h_{\min}}) \approx 1/(\beta_0\sqrt{h_{\min}})$. Thus, we have

$$\|e^{(k)}\| \approx \frac{c}{\sqrt{n}}\|e^{(k)}\|_2 \leq \frac{c}{\sqrt{N}}\|A^{-1/2}\|_2\|e^{(k)}\|_A \leq \frac{c}{\sqrt{N}}\frac{1}{\beta_0\sqrt{h_{\min}}}\|e^{(k)}\|_A.$$

Hence, up to the assumption relating $\|e^{(k)}\|$ and $\|e^{(k)}\|_2$, invoking (5.28), the stopping criterion to guarantee $\|e^{(k)}\| \leq g_\infty(\omega_x^N)$ is given by

$$\begin{aligned}\sqrt{\left(\tilde{r}^{(k)}\right)^T r^{(k)}} &\leq C\frac{\beta_0\sqrt{Nh_{\min}}}{c}g_\infty(\omega_x^N) \\ &= \begin{cases} C\varepsilon^{1/2}N^{-2}\ln^{5/2}N, & \omega_x^N \text{ is a Shishkin mesh,} \\ C\varepsilon^{1/2}N^{-2}, & \omega_x^N \text{ is a Bakhvalov mesh,} \end{cases}\end{aligned} \tag{5.30}$$

where we have used $g_\infty(\omega_x^N)$ as given in (5.10).

It should be noted that both stopping criteria (5.29) and (5.30) are dependent of ε. However, where the ε-dependency of the former comes from the error bound (5.7) and (5.9), that of the latter comes from the fact that $h_{\min}$ is $\mathcal{O}(\varepsilon N^{-1})$.

5.2.4 Numerical results

Implementation

We will implement the boundary layer preconditioner as part of a CG solver. So, we briefly recap the preconditioned CG algorithm in order to carefully illustrate the use of the boundary layer preconditioner, as well as the role of a multigrid algorithm in its implementation. Suppose that M is a preconditioner. Then, the preconditioned CG can be written as in Algorithm 5.1 (see, e.g., [26, Alg. 6.12]).

Algorithm 5.1 Preconditioned Conjugate Gradient algorithm:

$k = 0; x_0 = 0; r_0 = b; y_0 = M^{-1}r_0; p_1 = y_0$
repeat
  $k = k + 1$
  $z = Ap_k$
  $\nu_k = (y_{k-1}^T r_{k-1})/(p_k^T z)$
  $x_k = x_{k-1} + \nu_k p_k$
  $r_k = r_{k-1} - \nu_k z$
  $y_k = M^{-1}r_k$
  $\mu_{k+1} = (y_k^T r_k)/(y_{k-1}^T r_{k-1})$
  $p_{k+1} = y_k + \mu_{k+1}p_k$
until stopping criterion is reached

It is worth mentioning that we do not actually form the inverse of the preconditioner M in the procedures $y_0 = M^{-1}r_0$ (line 1), and $y_k = M^{-1}r_k$ (line 7), within Algorithm 5.1. Instead, we solve the residual equations $My_0 = r_0$, and $My_k = r_k$,

respectively. If M happens to diagonal, this is easily done. If M comes from an IC(0) preconditioner, then it just involves solving two triangular systems by back-substitutions. More generally, however, $y_k = M^{-1}r_k$ can be interpreted as a call to a subroutine that returns an approximate solution to $y_k = A^{-1}r_k$. Where this routine is based on a multigrid method, the algorithm is called *multigrid preconditioned CG*, see, e.g., [44, §12.1]. In our case, $M = A_D$ is the block matrix

$$A_D = \begin{pmatrix} A_{BB} & 0 \\ 0 & D_{II} \end{pmatrix}.$$

Due to this block nature, the components of the system can be solved independently. The interior component, with the system matrix D_{II}, is easily solved by a diagonal solver since it is inexpensive to invert a diagonal matrix. To verify Corollary 5.1, we apply a direct solver to solve the boundary component involving the matrix A_{BB}. We call this approach *boundary layer preconditioned CG* (BLPCG) in which A_{BB} is used as a preconditioner.

The BLPCG algorithm requires the use of a direct solver. Recalling Remark 5.2, Corollary 5.2 is verified when we replace A_{BB} by a spectrally equivalent matrix M_{BB}. In this implementation, we apply a (geometric) multigrid preconditioner to solve the boundary component of the residual systems. We call this approach *multigrid boundary layer preconditioned CG* (MG-BLPCG).

Results for unpreconditioned CG

We begin by studying the application of unpreconditioned CG to solve the linear system (5.4) which comes from the finite element discretization of Example 5.1 applied on a Shishkin mesh. Since the stopping criteria discussed in Section 5.2.3 only make sense if a good preconditioner is used, we terminate the unpreconditioned CG computation if $\|e^{(k)}\|$ is less than discretization accuracy multiplied by a user-chosen constant $\lambda = 0.01$, or 3,000 iterations are reached. Table 5.5 shows the iteration counts for the unpreconditioned CG applied to Example 5.1. It can be seen that, for a fixed N, the iteration counts depend badly on ε. For example, when $N = 2^{10}$, the number of iterations required for $\varepsilon^2 = 10^{-8}$ is 146, but it increases up to 470 for $\varepsilon^2 = 10^{-12}$. This agrees with Theorem 5.2 where the condition number is proportional to $(\varepsilon \ln N)^{-1}$ on a Shishkin mesh.

Results for BLPCG

As discussed above, in the light of verifying Corollary 5.1, we now apply the boundary layer preconditioner, A_D, to Algorithm 5.1. Our boundary layer preconditioner is

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	126	253	508	1019	2041	3000
10^{-2}	96	208	445	933	1931	3000
10^{-4}	22	43	86	173	348	701
10^{-6}	22	40	78	151	304	615
10^{-8}	57	69	73	146	291	575
10^{-10}	80	138	191	223	272	549
10^{-12}	78	159	296	470	654	730

Table 5.5: Iteration counts for unpreconditioned CG.

applicable only when ε is sufficiently small that the boundary layers actually exist. Therefore, we only report results for cases where δ_h is small enough for this to happen. In the following numerical experiments, we take $m = 3/8$ to form the preconditioner A_D, and report results only when $\delta_h \leq 0.1$, which is consistent with the spectral equivalence bounds in Theorem 5.3, and Corollary 5.1. Table 5.6 shows the iteration counts for the BLPCG, where the stopping criterion (5.29) associated with the energy norm is used with $C = 0.4$. It clearly shows the number of iterations is reduced significantly, compared to Table 5.5. More importantly, the iteration counts are *robust* with respect to ε. The iteration counts increase only slightly when ε is small. This is acceptable since the error bound, in the energy norm, is ε-dependent. Furthermore, as expected from a spectrally equivalent preconditioner, the iterations counts are optimal with respect to N. This is contrast to results in Theorems 4.2 and 4.4 where the estimates are dependent of N (see also corresponding Tables 4.3 and 4.4).

The corresponding errors in the energy norm are shown in Table 5.7, where we recover the accuracy of Table 5.1 to at least third decimal digit. This also verifies that the stopping criterion given in (5.29) is sharp.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	8	8	7	–	–	–
10^{-8}	9	9	9	9	9	8
10^{-10}	10	10	10	10	10	11
10^{-12}	10	11	11	11	11	12

Table 5.6: Iteration counts for BLPCG, using the energy norm stopping criterion.

In Table 5.8 we report the iteration counts for the BLPCG that yields the computed error in the maximum norm in Table 5.9, where we take C to be 0.5 in (5.30). For a fixed N, we observe only a slight increase in the number of iterations when ε becomes small.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	5.6658e-03	3.2393e-03	1.8223e-03	–	–	–
10^{-8}	1.7917e-03	1.0246e-03	5.7660e-04	3.2043e-04	1.7621e-04	9.6152e-05
10^{-10}	5.6676e-04	3.2400e-04	1.8231e-04	1.0132e-04	5.5754e-05	3.0377e-05
10^{-12}	1.8011e-04	1.0252e-04	5.7651e-05	3.2034e-05	1.7625e-05	9.6058e-06

Table 5.7: $\|u - u^N\|_\varepsilon$ by BLPCG.

This is because the maximum norm stopping criterion given in (5.30) is ε-dependent. Furthermore, the BLPCG algorithm requires a few more iterations compared to the corresponding iteration counts for the energy norm. This may be because, in this norm, we are over-solving: Table 5.9 agrees with Table 5.2 to almost all reported digits. The estimate leading to (5.30) is predicted on the assumption relation $\|e^{(k)}\|_2$ and $\|e^{(k)}\|$. In contrast, the corresponding criterion of the energy norm bound has no such assumption, and so is sharper.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	10	10	8	–	–	–
10^{-8}	11	11	12	12	12	11
10^{-10}	11	12	13	13	14	14
10^{-12}	12	13	13	14	15	15

Table 5.8: Iteration counts for BLPCG, using the maximum norm stopping criterion.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
10^{-6}	3.8493e-03	1.2537e-03	3.9616e-04	–	–	–
10^{-8}	3.8502e-03	1.2545e-03	3.9615e-04	1.2220e-04	3.6963e-05	1.0997e-05
10^{-10}	3.8669e-03	1.2548e-03	3.9617e-04	1.2221e-04	3.6963e-05	1.0997e-05
10^{-12}	3.8728e-03	1.2552e-03	3.9675e-04	1.2222e-04	3.6964e-05	1.0998e-05

Table 5.9: $\|u - u^N\|_{\omega_x^N}$ by BLPCG.

Results for MG-BLPCG

We now consider into a more realistic setting where, instead of resolving the component of the system involving A_{BB}, a multigrid-based solver is used. High frequency components of errors are damped using the Jacobi or Gauss-Seidel smoothers. Low frequency

components of errors are resolved by projecting to a lower-dimensional space, and applying a direct solver, or by applying multigrid methods recursively. Such multigrid algorithms can be used as solvers, or as preconditioners as described in Remark 5.2. However, they are best suited to diffusion-dominated problems: they are not effective on highly isotropic meshes, and too computationally expensive to use in regions where a diagonal preconditioner would suffice. For more details of multigrid methods, we refer the reader to textbooks, [9, 108, 116].

In this section, we focus on the use of multigrid as a preconditioner to verify Corollary 5.2. Recall that when $\delta_h > 0.1$, the boundary layers are not developed, and the problem is effectively diffusion-dominated. In this case we use multigrid to precondition the whole system (i.e., we do not decompose into boundary and interior regions). When $\delta_h \leq 0.1$, we employ a multigrid V-cycle, expressed as M_{BB}, to efficiently solve the boundary component. For the interior component where our preconditioner is diagonal, we simply apply a diagonal solver. Our multigrid implementation is based on the description of [9, Chapter 3]. We use three Gauss-Seidel sweeps (see Section 1.8.2) as our smoother. As soon as the number of grids in the restriction process reaches 8, we apply a direct solver and start the projection process.

Table 5.10 shows the iteration counts for the algorithm, where we underlined the counts for the cases when $\delta_h > 0.1$. For a fixed ε, we observe the optimality of the method: iteration counts are unchanged as N increases. For $\delta_h \leq 0.1$, we use the MG-BLPCG algorithm, invoking the stopping criterion (5.29) with $C = 0.4$ (see the computed error in Table 5.11). In this case, the iteration counts also are steady and show only very small dependency on both N and ε. The iteration counts are far less than that of Table 5.5. More importantly, they are robust with respect to ε and optimal in N. To verify that the algorithm is not under-solving the linear system, in Table 5.11 we underline the digits that agree with the discretization error, $\|u - u^N\|_\varepsilon$.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	3	3	3	3	3	3
10^{-2}	3	3	3	3	3	3
10^{-4}	2	2	2	2	2	2
10^{-6}	8	8	7	3	3	3
10^{-8}	9	9	10	10	10	9
10^{-10}	10	10	10	11	11	11
10^{-12}	11	11	11	11	12	12

Table 5.10: Iteration counts for MG-BLPCG, using the energy norm stopping criterion.

ε^2	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$	$N = 2^{12}$
1	3.7559e-03	1.8779e-03	9.3897e-04	4.6948e-04	2.3474e-04	1.1737e-04
10^{-2}	1.4485e-02	7.2427e-03	3.6214e-03	1.8107e-03	9.0535e-04	4.5268e-04
10^{-4}	1.7910e-02	1.0241e-02	5.7618e-03	3.2013e-03	1.7608e-03	9.6050e-04
10^{-6}	5.6658e-03	3.2393e-03	1.8223e-03	1.0124e-03	5.5685e-04	3.0375e-04
10^{-8}	1.7917e-03	1.0246e-03	5.7626e-04	3.2017e-04	1.7609e-04	9.6054e-05
10^{-10}	5.6676e-04	3.2400e-04	1.8231e-04	1.0125e-04	5.5689e-05	3.0378e-05
10^{-12}	1.7988e-04	1.0252e-04	5.7651e-05	3.2034e-05	1.7610e-05	9.6059e-06

Table 5.11: $\|u - u^N\|_\varepsilon$ by MG-BLPCG.

5.3 Two-dimensional problems

We now turn to a finite element discretization of the two-dimensional reaction-diffusion problem (5.2) on a tensor product mesh with bilinear elements. As usual, let ω_x^N and ω_y^N be arbitrary meshes, each with N intervals on $[0,1]$. Set $\Omega^{N,N} = \{(x_i, y_j)\}_{i,j=0}^N$ to be the Cartesian product of ω_x^N and ω_y^N. For any fixed ε, the discretization of $-\varepsilon^2\Delta u$ is straightforward; in order to avoid issues of quadrature in evaluating the weighted finite element mass matrix entries, we assume that b is approximated as a piecewise constant on each element, writing $b_{i,j} = b(x_{i+1/2}, y_{j+1/2})$, for $x_{i+1/2} = (x_i + x_{i+1})/2$ and $y_{j+1/2} = (y_j + y_{j+1})/2$. The matrix decomposes into three terms, $A = \varepsilon^2(S^{(x)} + S^{(y)}) + M$, whose stencils are

$$S^{(x)} = \begin{bmatrix} -\frac{k_j}{6h_{i-1}} & \frac{k_j}{6h_{i-1}} + \frac{k_j}{6h_i} & -\frac{k_j}{6h_i} \\ -\frac{k_j + k_{j-1}}{3h_{i-1}} & \frac{k_j + k_{j-1}}{3h_{i-1}} + \frac{k_j + k_{j-1}}{3h_i} & -\frac{k_j + k_{j-1}}{3h_i} \\ -\frac{k_{j-1}}{6h_{i-1}} & \frac{k_{j-1}}{6h_{i-1}} + \frac{k_{j-1}}{6h_i} & -\frac{k_{j-1}}{6h_i} \end{bmatrix},$$

$$S^{(y)} = \begin{bmatrix} -\frac{h_{i-1}}{6k_j} & -\frac{h_i + h_{i-1}}{3k_j} & -\frac{h_i}{6k_j} \\ \frac{h_{i-1}}{6k_j} + \frac{h_{i-1}}{6k_{j-1}} & \frac{h_i + h_{i-1}}{3k_j} + \frac{h_i + h_{i-1}}{3k_{j-1}} & \frac{h_i}{6k_j} + \frac{h_i}{6k_{j-1}} \\ -\frac{h_{i-1}}{6k_{j-1}} & -\frac{h_i + h_{i-1}}{3k_{j-1}} & -\frac{h_i}{6k_{j-1}} \end{bmatrix},$$

and

$$M = \begin{bmatrix} m_{11} & m_{12} & m_{13} \\ m_{21} & m_{22} & m_{23} \\ m_{31} & m_{32} & m_{33} \end{bmatrix}. \tag{5.31}$$

where

$$m_{11} = \frac{h_{i-1}k_j b_{i-1,j}}{36}, \quad m_{12} = \frac{k_j\,(h_{i-1}b_{i-1,j} + h_i b_{i,j})}{18}, \quad m_{13} = \frac{h_i k_j b_{i,j}}{36},$$

$$m_{21} = \frac{h_{i-1}\,(k_j b_{i-1,j} + k_{j-1} b_{i-1,j-1})}{18}, \quad m_{23} = \frac{h_i\,(k_j b_{i,j} + k_{j-1} b_{i,j-1})}{18},$$

$$m_{22} = \frac{h_{i-1}\left(k_j b_{i-1,j} + k_{j-1} b_{i-1,j-1}\right) + h_i\left(k_j b_{i,j} + k_{j-1} b_{i,j-1}\right)}{9},$$

and

$$m_{31} = \frac{h_{i-1} k_{j-1} b_{i-1,j-1}}{36}, \quad m_{32} = \frac{k_{j-1}\left(h_{i-1} b_{i-1,j-1} + h_i b_{i,j-1}\right)}{18}, \quad m_{33} = \frac{h_i k_{j-1} b_{i,j-1}}{36}.$$

The error analysis of reaction-diffusion problems in two dimensions on a Shishkin mesh can be found in, e.g., [70] and also [96, pages 404–406]. They prove that there exists a constant independent of both ε and N such that

$$\|u - u^N\|_\varepsilon \leq C(\varepsilon^{1/2} N^{-1} \ln N + N^{-2}). \tag{5.32}$$

5.3.1 Condition number estimate

As with the one-dimensional reaction-diffusion problem, we will show that the linear system of the two-dimensional problem (5.2) is ill-conditioned, when the problem is discretized by the finite element method on boundary layer-adapted meshes.

First, when $\varepsilon \ll 1$, by Geršgorin's Theorem, it is easy to show that

$$\|A\|_2 \leq CN^{-2}. \tag{5.33}$$

It is more difficult to find the lower bound for $\lambda_{\min}$ (or, equivalently, upper bound for $\|A^{-1}\|_2$). This is because unlike the one-dimensional case, A is not diagonally dominant. In order to use Geršgorin's Theorem, we will construct an intermediate symmetric positive definite matrix $\tilde{A}$ so that $V^T \tilde{A} V \leq V^T A V$ for all V, where we can easily bound $\|\tilde{A}^{-1}\|_2$. To bound $\|A^{-1}\|_2$, we use that

$$\|A^{-1}\|_2 \leq \|\tilde{A}^{-1}\|_2.$$

This can be explained as follows. Since $V^T \tilde{A} V \leq V^T A V$, then

$$\tilde{\lambda}_{\min} = \min_V \frac{V^T \tilde{A} V}{V^T V} \leq \min_V \frac{V^T A V}{V^T V} = \lambda_{\min},$$

where $\lambda_{\min}$, and $\tilde{\lambda}_{\min}$ denote the smallest eigenvalues of A and $\tilde{A}$, respectively. Both A and $\tilde{A}$ are symmetric positive definite, so

$$\|A^{-1}\|_2 = \frac{1}{\lambda_{\min}} \leq \frac{1}{\tilde{\lambda}_{min}} = \|\tilde{A}^{-1}\|_2.$$

One simple way to define $\tilde{A}$ is to take $\tilde{A}$ to be a diagonal matrix whose stencil is given below

$$\tilde{A} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \tilde{a} & 0 \\ 0 & 0 & 0 \end{bmatrix}.$$

where

$$\begin{aligned}\tilde{a} &= m_{22} - (m_{11} + m_{12} + m_{13} + m_{21} + m_{23} + m_{31} + m_{32} + m_{33}) \\ &= \frac{h_{i-1}\left(k_j b_{i-1,j} + k_{j-1} b_{i-1,j-1}\right) + h_i\left(k_j b_{i,j} + k_{j-1} b_{i,j-1}\right)}{36}.\end{aligned}$$

Then, we can see that $V^T(A - \tilde{A})V \geq 0$, so $V^T\tilde{A}V \leq V^T A V$ by the definition of $\tilde{A}$. For the matrix $\tilde{A}$, by Geršgorin's Theorem, we easily see that

$$\tilde{\lambda}_{\min} \geq \min_i \left\{ \frac{h_{i-1}\left(k_j b_{i-1,j} + k_{j-1} b_{i-1,j-1}\right) + h_i\left(k_j b_{i,j} + k_{j-1} b_{i,j-1}\right)}{36} \right\} \geq \frac{\beta_0^2 h_{\min}^2}{9}.$$

It follows then that

$$\|A^{-1}\|_2 \leq \|\tilde{A}^{-1}\|_2 \leq 9/(\beta_0^2 h_{\min}^2). \tag{5.34}$$

Combining (5.33) and (5.34), we obtain

$$\kappa_2(A) \leq \begin{cases} C\varepsilon^{-2}\ln^{-2} N, & \Omega^{N,N} \text{ is a Shishkin mesh,} \\ C\varepsilon^{-2}, & \Omega^{N,N} \text{ is a Bakhvalov mesh.} \end{cases} \tag{5.35}$$

Table 5.12 gives computed values of $\kappa_2(A)$, and shows that the estimate (5.35) is sharp for the Shishkin mesh with the constant $C \approx 0.5$. (We do not include the results for the Bakhvalov mesh, but they are consistent with (5.35), and we observe that the corresponding C is approximately 0.01). In particular, for a fixed N, $\kappa_2(A)$ is proportional to ε^{-2}.

ε^2	$N = 16$	$N = 32$	$N = 64$	$N = 128$	$N = 256$	$N = 512$
1	4.92e+01	1.97e+02	7.90e+02	3.16e+03	1.26e+04	5.06e+04
10^{-2}	8.83e+00	3.45e+01	1.37e+02	5.48e+02	2.19e+03	8.76e+03
10^{-4}	1.79e+02	2.08e+02	5.15e+02	1.34e+03	3.66e+03	1.03e+04
10^{-6}	1.94e+04	1.44e+04	1.09e+04	1.44e+04	4.02e+04	1.16e+05
10^{-8}	1.95e+06	1.45e+06	1.11e+06	8.62e+05	6.88e+05	1.17e+06
10^{-10}	1.95e+08	1.45e+08	1.11e+08	8.63e+07	6.88e+07	5.60e+07
10^{-12}	1.95e+10	1.45e+10	1.11e+10	8.63e+09	6.89e+09	5.60e+09

Table 5.12: $\kappa_2(A)$ of the problem (5.2) discretized by a FEM on a Shishkin mesh.

5.3.2 Boundary layer preconditioners

As in the finite difference case [72, §4.5], we partition A into a corner region, where the mesh is highly resolved in both directions, the edge regions, where the mesh is highly resolved in one direction but not both, and the interior region. Further, we assume

that the mesh spacing (in both directions) in the non-resolved portions of the grid is uniform, with spacing h_I. Thus, we write

$$A = \begin{pmatrix} A_{CC} & A_{CE} & A_{CI} \\ A_{EC} & A_{EE} & A_{EI} \\ A_{IC} & A_{IE} & A_{II} \end{pmatrix}, \tag{5.36}$$

where the subscripts C, E and I indicate the block structure of corners, edge layers, and interior points, respectively. The preconditioner, A_D, will be defined using the same partitioning:

$$A_D = \begin{pmatrix} A_{CC} & 0 & 0 \\ 0 & T_{EE} & 0 \\ 0 & 0 & D_{II} \end{pmatrix}. \tag{5.37}$$

Here D_{II} is the diagonal matrix with entries based on the scaled diagonal of the mass matrix, i.e., $D_{II} = m\text{diag}(M_{II})$. The tridiagonal matrix T_{EE} is constructed so that it is spectrally equivalent to A_{EE}. The choice of T_{EE} stems from the following observation. Along the edges, rectangles are long and thin with one side of length $\mathcal{O}(\varepsilon N^{-1})$, and other side of length $\mathcal{O}(N^{-1})$. Therefore, depending on the orientation of the rectangles, some entries in A_{EE} are very small compared to others. A preconditioner can be formed by either neglecting these terms, or aggregating them.

For our implementation, here we consider the approach that T_{EE} is constructed by summing the coefficients along the direction of the large mesh-width. Considering the block associated with the edge along x-axis, we'll decompose $A_{EE} = \varepsilon^2(S_{EE}^{(x)} + S_{EE}^{(y)}) + M_{EE}$ as the (column-wise) tridiagonal terms in $\varepsilon^2 S_{EE}^{(y)}$ and M_{EE} to give

$$T_{EE} = \begin{bmatrix} 0 & -\dfrac{h_I\varepsilon^2}{k_j} + \dfrac{h_I k_j(b_{i-1,j} + b_{i,j})}{12} & 0 \\ 0 & \left(\dfrac{h_I}{k_j} + \dfrac{h_I}{k_{j-1}}\right)\varepsilon^2 + \dfrac{h_I k_j(b_{i-1,j} + b_{i,j}) + h_I k_{j-1}(b_{i-1,j-1} + b_{i,j-1})}{6} & 0 \\ 0 & -\dfrac{h_I\varepsilon^2}{k_{j-1}} + \dfrac{h_I k_{j-1}(b_{i-1,j-1} + b_{i,j-1})}{12} & 0 \end{bmatrix}. \tag{5.38}$$

5.3.3 Numerical results

In this section, we only focus on the computed error in the energy norm (rather than the maximum norm) on a Shishkin mesh since it is the case for which rigorous analysis is available. We will discuss performance of direct solvers, and demonstrate the difficulties due to the presence of subnormal numbers, as analyzed in Chapter 3, but for the finite difference discretization. Then, numerical results for our MG-BLPCG algorithm are reported.

Example 5.2. We reconsider Example 4.1 in which the data is chosen so that

$$u(x,y) = x^3(1+y^2) + \sin(\pi x^2) + \cos(\pi y/2) + (1+x+y)\left(e^{-2x/\varepsilon} + e^{-2y/\varepsilon}\right). \tag{5.39}$$

The errors in the energy norm by the finite element method on a Shishkin mesh are given in Table 5.13. This agrees with the error estimate seen in (5.32).

ε^2	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
1	4.744032e-02	2.372374e-02	1.186232e-02	5.931216e-03	2.965615e-03	1.482808e-03
10^{-2}	5.914034e-02	2.964357e-02	1.483101e-02	7.416662e-03	3.708476e-03	1.854256e-03
10^{-4}	4.515961e-02	2.669820e-02	1.532879e-02	8.636071e-03	4.800213e-03	2.640514e-03
10^{-6}	1.434196e-02	8.478054e-03	4.867799e-03	2.742527e-03	1.524410e-03	8.385593e-04
10^{-8}	4.551908e-03	2.683635e-03	1.540147e-03	8.676500e-04	4.822686e-04	2.652895e-04
10^{-10}	1.485438e-03	8.534980e-04	4.875788e-04	2.744442e-04	1.525197e-04	8.389624e-05
10^{-12}	5.958432e-04	2.847030e-04	1.558249e-04	8.696908e-05	4.825158e-05	2.653275e-05

Table 5.13: $\|u - u^N\|_\varepsilon$ to Example 5.2 by a FEM on a Shishkin mesh.

Direct solvers

In order to make computations comparable between different discretizations, we use the same setting as that of Section 3.1 and [72], i.e., the program was coded in C and executed using a single core of a node with an AMD Opteron 2427, 2200 MHz processor and 32Gb of RAM. We use CHOLMOD Version 1.7.1 to solve the sparse symmetric positive definite linear systems; see [12, 22]. In Table 5.14, we show the time in seconds, averaged over three runs, required to solve the linear systems that correspond to the results in Table 5.13. For a fixed N, say $N = 2^{11}$, it is easily observed that the amount of time required to solve the linear system depends quite badly on the perturbation parameter.

ε^2	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
1	0.02	0.10	0.72	5.59	35.38	354.82
10^{-2}	0.02	0.10	0.72	5.58	35.32	354.91
10^{-4}	0.02	0.09	0.72	5.58	35.31	355.01
10^{-6}	0.02	0.10	0.72	6.64	102.81	1263.85
10^{-8}	0.01	0.10	0.72	6.26	100.23	1327.79
10^{-10}	0.01	0.10	0.72	6.47	102.83	1349.70
10^{-12}	0.01	0.10	0.72	6.83	106.47	1359.60

Table 5.14: Cholesky (CHOLMOD) solve times for linear systems generated by a FEM on a Shishkin mesh.

If the results of Table 5.14 are compared with the corresponding results for the finite difference method, as given in [72, Table 4.1] and also Table 3.3, two issues become apparent:

1. When ε is $\mathcal{O}(1)$, the solve times for the finite element discretization are about twice those for the finite difference discretization. This is because the finite element discretization has a 9-point stencil rather than 5-point stencil of the finite difference discretization, and so the system matrix has roughly twice the number of nonzero entries.

2. For the finite difference case, the solve times, as shown in [72, Table 4.1] and Table 3.3, initially increase, and then decrease when ε becomes smaller. In contrast, the solve times for the finite element case increase initially, but then stabilize. Although we do not offer an analysis of Cholesky factorization in this case, the framework of Chapter 3 could be used to investigate this.

In Table 5.15, we give the number of nonzero entries in the Cholesky factors produced by CHOLMOD for a range of values of N and ε, as well as the number of subnormal entries. This agrees completely with the results of Table 5.14. For small ε and large N, we observe a significant increase in the number of subnormal numbers arising in the Cholesky factors, as well as a decrease in the number of nonzero numbers, due to underflow-zeros.

ε^2	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
1	102124	573163	3239141	17011189	63549693	304900961
	0	0	0	0	0	0
10^{-2}	102124	573163	3239141	17011189	63549693	304900961
	0	0	0	0	0	0
10^{-4}	102124	573163	3239141	17011189	63549693	304900961
	0	0	0	0	0	0
10^{-6}	102124	573163	3239141	17011189	63166392	300992678
	0	0	0	0	74982	441440
10^{-8}	102123	573163	3239141	17011189	63276869	293538627
	0	0	0	0	69508	955686
10^{-10}	102124	573162	3239134	17011179	63263046	293598199
	0	0	0	0	71831	957773
10^{-12}	100011	573160	3239136	17011171	63234561	293268356
	0	0	0	0	75100	934242

Table 5.15: Number of nonzero entries (top) and subnormal numbers (bottom) in Cholesky factors generated by CHOLMOD.

Boundary layer preconditioned CG

We do not provide an analysis of the boundary layer preconditioner proposed in Section 5.3.2. Because the matrix has a 9-point stencil, this would require very detailed and intricate analysis. However, here we report results of the use of the boundary layer preconditioner which show that, in practice, the approach is robust and very promising when the perturbation parameter, ε, is small.

Based on the structure of the boundary layer preconditioner defined in (5.37), we apply different solver strategies to efficiently solve the linear systems.

- For the corner region, we use the black box multigrid method (BoxMG) of Dendy [2, 27], and see also [72, §4.4]. This is because the types of problem for which BoxMG is optimized include those with 9-point stencil, and are diffusion-dominated, as is the case in this region.
- For the edge region, T_{EE} is a tridiagonal matrix. Therefore, we use the tridiagonal solver *algorithm* `DPTTRS` [111] that is part of LAPACK library of subroutines for solving problems in numerical linear algebra [4].
- A diagonal solver is used for the interior region since the corresponding preconditioner in this region is a diagonal matrix.

In addition, to enhance the computational efficiency, we introduce three user-chosen parameters, c_1, c_2, and c_3 to appropriately scale the corner, edge, and interior components, respectively, of the preconditioned residual $\tilde{r}^{(k)}$ in MG-BLPCG algorithm.

Arguments similar to those in Section 5.2.3 can be used to derive the stopping criterion associated with the energy norm for the two-dimensional problems. In particular, for a Shishkin mesh, the stopping criterion based on the preconditioned residual is

$$\sqrt{\left(\tilde{r}^{(k)}\right)^T r^{(k)}} \leq C(\varepsilon^{1/2} N^{-1} \ln N + N^{-2}). \tag{5.40}$$

As in the analysis of Section 5.2.2, our boundary layer preconditioner is specially designed for singularly perturbed problems. Thus, we only report results for cases where $\delta_h \leq 0.1$. In Table 5.16, we give the CPU solve times of MG-BLPCG, together with the iteration counts. The user-chosen parameters are $c_1 = 1$, $c_2 = 0.7$, and $c_3 = 0.5$, and the stopping criterion (5.40) is used. We emphasize that the iteration counts are robust with respect to ε. Furthermore, the iteration counts are optimal with respect to N, and only slightly depend on ε. This can be explained by the fact that the stopping criterion (5.40) is ε-dependent.

To verify that the MG-BLPCG algorithm is not under-solving, Table 5.17 shows the corresponding errors where we underline the digits that agree with the error in Table 5.13.

ε^2	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
10^{-6}	0.00 (10)	0.02 (9)	0.09 (9)	0.44 (9)	–	–
10^{-8}	0.00 (11)	0.02 (11)	0.11 (11)	0.51 (11)	3.00 (11)	17.74 (14)
10^{-10}	0.01 (13)	0.02 (13)	0.13 (13)	0.60 (13)	3.48 (13)	16.60 (13)
10^{-12}	0.01 (15)	0.02 (15)	0.15 (15)	0.68 (15)	3.95 (15)	18.85 (15)

Table 5.16: CPU times (and iteration counts) for MG-BLPCG, on a Shishkin mesh, averaged over 3 runs.

ε^2	$N = 2^6$	$N = 2^7$	$N = 2^8$	$N = 2^9$	$N = 2^{10}$	$N = 2^{11}$
10^{-6}	1.434198e-02	8.478082e-03	4.867803e-03	2.742530e-03	–	–
10^{-8}	4.551925e-03	2.683645e-03	1.540152e-03	8.676531e-04	4.822693e-04	2.652899e-04
10^{-10}	1.485441e-03	8.534998e-04	4.875798e-04	2.744448e-04	1.525200e-04	8.389643e-05
10^{-12}	5.958436e-04	2.847033e-04	1.558251e-04	8.696920e-05	4.825166e-05	2.653280e-05

Table 5.17: $\|u - u^N\|_\varepsilon$ by MG-BLPCG.

Table 5.16 clearly shows that, for a small ε and a large N, the MG-BLPCG algorithm is far more efficient than the direct solver used to compute the results in Table 5.14. For example, when $N = 2^{10}$ and $\varepsilon^2 \leq 10^{-8}$, the MG-BLPCG is about 30 times faster. When $N = 2^{11}$, the algorithm is about 80 times faster. To compare the solve times by CHOLMOD and the MG-BLPCG, in Figure 5.1, we plot the solve times versus the degrees of freedom in the system. It can be seen that when the degrees of freedom increase, the solve times taken by CHOLMOD (marked by circle) increase more rapidly than that of MG-BLPCG. In fact, the solve times for CHOLMOD grow quadratically with the degrees of freedom, whereas MG-BLPCG is almost linear.

5.4 Conclusions

We have considered the topic of solving the linear systems arising from the finite element discretization of the singularly perturbed reaction-diffusion problems on boundary layer-adapted meshes. The use of such highly nonuniform meshes result in the unbounded growth of the system matrix condition number. We have derived the sharp bounds on the condition number of the system matrix arising from a finite element discretization on any layer-adapted mesh. We have proposed and analyzed the boundary layer preconditioner for a one-dimensional problem. We have also proposed an analogous preconditioner for a two-dimensional problem. Although we have not provided

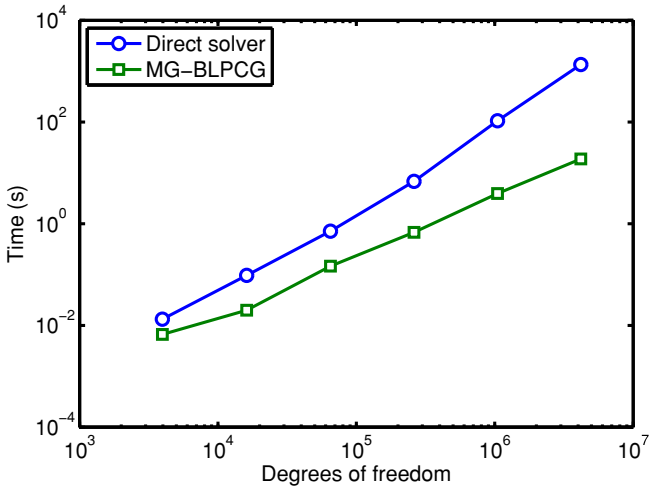

Figure 5.1: Semi-log plot of solve times taken by CHOlMOD and MG-BLPCG versus the degrees of freedom.

an analysis of boundary layer preconditioners in this two-dimensional case, numerical results have shown that it is robust with respect to ε, and very efficient. Appropriate stopping criteria for both energy and maximum norms have been carefully derived. Furthermore, our theoretical analysis of the boundary layer preconditioner applies to any layer-adapted mesh. Although we have chosen to report results for the Shishkin mesh, numerical experiments demonstrate that the method is successful when applied to a Bakhvalov mesh.

Chapter 6

Conclusions

Summary of the thesis

At the beginning of this project in September 2011, our primary goal was to investigate and develop robust numerical algorithms that can efficiently solve linear systems coming from finite difference and finite element discretizations applied to singularly perturbed problems. This presents several challenges. Unlike the topic of designing and analyzing ε-uniform numerical methods for singularly perturbed problems, which is very active, the topic of solving the resulting linear systems, in a manner that is parameter robust, is relatively new. Thus, in literature, there have been very few studies which focus on solving these linear systems, particularly in the context of layer-adapted meshes. This means, for example, that we have to establish very fundamental results, such as bounds on the condition number of unpreconditioned systems.

The thesis has successfully achieved the key target of the project by providing a detailed analysis of different strategies for solving linear systems of equations when the singularly perturbed problems are discretized on layer-adapted meshes. We have advanced the understanding of this topic, not only by showing the difficulties of solving such systems either both direct or iterative solvers, but also by suggesting appropriate strategies for improving the performance of iterative solvers.

More precisely, we have analyzed in detail a direct solver based on Cholesky factorization for symmetric positive definite systems by providing estimates for the magnitude of fill-in entries in a given location in the factors. Based on this analysis, we can determine the range of ε and N where the presence of the subnormal and underflow-zero numbers is expected. This fully explains the question of how the computational cost of solving a linear system coming from finite difference discretizations applied to singularly perturbed problems is influenced by ε and N.

For iterative solvers, we have analyzed the diagonal and incomplete Cholesky factorization preconditioners for reaction-diffusion problems discretized by a finite difference method. These preconditioners have been proven to be robust with respect to the perturbation parameter. They are not, however, optimal in N. Therefore, we have extended the ideas in [72] for spectrally equivalent preconditioners for finite difference methods to finite element methods.

Although the main interest of this dissertation is designing and developing robust strategies for solving linear systems of singularly perturbed problems, we have also provided a new uniform convergence proof technique, which is based on the preconditioning approach, for one-dimensional convection-diffusion problems on a Shishkin mesh. A simple proof of pointwise uniform convergence (without preconditioning) of one-dimensional reaction-diffusion problems has also been presented.

Further work

There are many possible extensions of the work of this thesis which can be considered in the future. Of course, we do not aim to discuss all possibilities of further work related to this thesis, rather we mention some directions which are inevitably personal and reflect our own interests.

The preconditioning technique used in Chapter 2 has been extended to a Bakhvalov-type mesh to prove ε-uniform convergence of a simple upwind scheme for one-dimensional convection-diffusion problems [85]. It is even more interesting to see whether this approach can be applied to finite element discretization cases, or to higher-dimensional problems. The major difference in finite element cases is that the error analysis does not rely on truncation error estimates, rather the method itself. Nonetheless, this would be very exciting, since it would establish new maximum norm results (rather than just giving new analyses of existing results). To extend these results to higher-dimensional problems, the main difficulty is to prove the stability after preconditioning.

The numerical results presented in Chapter 5 show that the performance of direct solvers for the finite element case is actually worse than that of the finite difference. Therefore, we would like to extend the analysis of Chapter 3 to the finite element discretization in which the primary difficulty is due to the complication of the 9-point stencil. Furthermore, the analysis has provided a useful tool for estimating the difference between the full and incomplete Cholesky factors which can be exploited in other direct solver-like approaches.

For convection-diffusion problems, on the other hand, direct solvers are based on LU factorizations since the problems are not self-adjoint. Nonetheless, the approach

of Chapter 3 could be carried over to this case, but it requires more detailed calculations because we have to compute both L and U, rather than just L as in Cholesky factorizations.

The extension of the diagonal and incomplete Cholesky preconditioners for finite difference discretizations on Shishkin meeshes in Chapter 4 to other types of layer-adapted meshes is straightforward, but detailed computations are required where the meshes are allowed to be graded in the layer regions. More interestingly, perhaps, the analysis of the IC(0) preconditioner can be applied to IC(k) preconditioners. In the notation of Chapter 3, IC(k) factors are formed by keeping all fill-in entries belonging to $L^{[p]}$ in the Cholesky factors, where p is a positive integer satisfying $p \leq k$ (see, e.g., [97, §10.3.3] for a detailed description of IC(k)). We expect IC(k) to be even more efficient than IC(0). This is because the fill-in entries corresponding to different $L^{[k]}$ are very distinct in magnitude (see Section 3.2) in the singularly perturbed regime: the larger k we have; the better approximation we get. Similarly, the incomplete Cholesky preconditioners with dropping tolerance, IC(μ), where μ denotes the dropping tolerance in Cholesky factors, could also be studied in the framework of Chapters 3 and 4.

Our numerical results in Chapter 5 show that our proposed boundary layer preconditioner is (almost) optimally efficient for two-dimensional problems. Proving that this is the case would be an achievable, but nontrivial goal. The extension of the method from two dimensions to three dimensions seems feasible in practice, but leads to numerous mathematical questions. Firstly, the condition number of unpreconditioned systems would have to be investigated. For example, we would like to know if the conditioning of three-dimensional problems is proportional to ε^{-3}. Secondly, since the use of direct solvers are very limited for higher-dimensional problems due to the exponential increase of computational cost, a preconditioned iterative scheme is alternative. Our primary interest is the question of how to design a robust preconditioner in three-dimensional case. Furthermore, we would also like to extend the idea of boundary layer preconditioners to design similarly structured preconditioners in which they could be robust and efficient for singularly perturbed problems discretized on different-shaped domains, rather than the unit square.

Looking further to the future, there are plenty of other related directions and applications in numerical linear algebra that may be relevant to singularly perturbed problems. Among the most obvious are domain decomposition approaches based on Schwarz and Schur methods. The applications and analysis of a Schwarz domain decomposition method on overlapping subdomains applied to a coupled system of one-dimensional reaction-diffusion problems can be found in [103, 104], and extended to a two-dimensional case in [54]. In this way, numerical solutions of some of subdomains can be solved simultaneously and in parallel. These domain decomposition methods,

which themselves are iterative, could be used as the basis for a robust preconditioner. They also have the potential to lead to a different proof-strategy for the boundary layer preconditioner of Chapter 5. This is because, similar to that decomposition, they separate the interior, boundary and edge layer regions where we can then apply suitable proof techniques.

The analysis of Chapter 3 has addressed limitations of direct solvers for singularly perturbed problems. Therefore, the search for new robust and fast direct solvers for these problems is a need. The partitioned matrix associated with the regions of the mesh suggests that solving the systems by a Schur complement method might be a reasonable approach for direct solvers. The studies of fast direct solvers by Martisson et al. (see, e.g., [41, 45]), which are based on the hierarchical construction of Schur complements, might be our starting point in this direction. Moreover, the Schur complement technique is successfully used to construct the sweeping preconditioner [31]. The crucial idea of this approach is to approximate the Schur complement matrix using moving perfectly matched layers. These preconditioners have been applied to Helmholtz [88] and Maxwell's equations [31]. We are interested in applying this technique to singularly perturbed problems in the near future.

The overall idea of complete and incomplete Cholesky factorizations is "approximate factorization". From that point of view, another avenue of investigation is Alternating Direction Implicit (ADI) techniques in which a time-dependent partial differential equation can be split into the temporal and spatial components. The analysis of the ADI for singularly perturbed problems can be found in, e.g., [21, 66]. For problems that are two-dimensional in space, at each time step, ADI allows to consider further *dimension splitting* into x- and y-directions. Therefore, we only need to solve tridiagonal systems of N unknowns, rather than a banded system of N^2 unknowns. Thus, from the view point of linear algebra, the methods could be used as preconditioners.

It is clear from the discussion of Section 1.9 that the topic of robust numerical methods for singularly perturbed problems is an important and active area of research. We hope that we have given the reader some ideas for potential mathematical investigations of the complementary topic of robust linear solvers, for which there are many interesting open questions.

Bibliography

[1] J. Adler, S. MacLachlan, and N. Madden. A first-order system Petrov-Galerkin discretization for a reaction-diffusion problem on a fitted mesh. *NUI Galway*, Preprint, 2014.

[2] R. E. Alcouffe, Achi Brandt, J. E. Dendy, Jr., and J. W. Painter. The multigrid method for the diffusion equation with strongly discontinuous coefficients. *SIAM J. Sci. Statist. Comput.*, 2(4):430–454, 1981.

[3] D. N. de G. Allen and R. V. Southwell. Relaxation methods applied to determine the motion, in two dimensions, of a viscous fluid past a fixed cylinder. *Quart. J. Mech. Appl. Math.*, 8:129–145, 1955.

[4] E. Anderson, Z. Bai, C. Bischof, L. S. Blackford, J. Demmel, Jack J. Dongarra, J. Du Croz, S. Hammarling, A. Greenbaum, A. McKenney, and D. Sorensen. *LAPACK Users' Guide (Third Ed.)*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, USA, 1999.

[5] V. B. Andreev and I. A. Savin. On the convergence, uniform with respect to the small parameter, of A. A. Samarskiĭ's monotone scheme and its modifications. *Zh. Vychisl. Mat. i Mat. Fiz.*, 35(5):739–752, 1995.

[6] Ali R. Ansari and Alan F. Hegarty. A note on iterative methods for solving singularly perturbed problems using non-monotone methods on Shishkin meshes. *Comput. Methods Appl. Mech. Engrg.*, 192(33-34):3673–3687, 2003.

[7] N. Bakhvalov. Towards optimization of methods for solving boundary value problems in the presence of boundary layers. *Zh. Vzchisl. Mat. i Mat fiz.*, 9:841–859, 1969.

[8] 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.

[9] William L. Briggs, Van Emden Henson, and Steve F. McCormick. *A multigrid tutorial*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2000.

[10] N. F. Britton. *Reaction-diffusion equations and their applications to biology*. Academic Press, Inc., London, 1986.

[11] B. Bujanda, C. Clavero, J. L. Gracia, and J. C. Jorge. A high order uniformly convergent alternating direction scheme for time dependent reaction-diffusion singularly perturbed problems. *Numer. Math.*, 107(1):1–25, 2007.

[12] Yanqing Chen, Timothy A. Davis, William W. Hager, and Sivasankaran Rajamanickam. Algorithm 887: CHOLMOD, supernodal sparse Cholesky factorization and update/downdate. *ACM Trans. Math. Software*, 35(3):Art. 22, 14, 2008.

[13] C. Clavero and J. L. Gracia. High order methods for elliptic and time dependent reaction-diffusion singularly perturbed problems. *Appl. Math. Comput.*, 168(2):1109–1127, 2005.

[14] C. Clavero and J. L. Gracia. On the uniform convergence of a finite difference scheme for time dependent singularly perturbed reaction-diffusion problems. *Appl. Math. Comput.*, 216(5):1478–1488, 2010.

[15] C. Clavero and J. L. Gracia. An improved uniformly convergent scheme in space for 1D parabolic reaction-diffusion systems. *Appl. Math. Comput.*, 243:57–73, 2014.

[16] C. Clavero, J. L. Gracia, and J. C. Jorge. High-order numerical methods for one-dimensional parabolic singularly perturbed problems with regular layers. *Numer. Methods Partial Differential Equations*, 21(1):148–169, 2005.

[17] C. Clavero, J. L. Gracia, and J. C. Jorge. A uniformly convergence alternating direction HODIE finite difference scheme for 2D time-dependent convection-diffusion problems. *IMA J. Numer. Anal.*, 26(1):155–172, 2006.

[18] C. Clavero, J. L. Gracia, and F. J. Lisbona. High order schemes for reaction-diffusion singularly perturbed systems. In *BAIL 2008—boundary and interior layers*, volume 69 of *Lect. Notes Comput. Sci. Eng.*, pages 107–115. Springer, Berlin, 2009.

[19] C. Clavero, J. L. Gracia, and F. J. Lisbona. An almost third order finite difference scheme for singularly perturbed reaction-diffusion systems. *J. Comput. Appl. Math.*, 234(8):2501–2515, 2010.

[20] C. Clavero, J.L. Gracia, and E. O'Riordan. A parameter robust numerical method for a two dimensional reaction-diffusion problem. *Math. Comput.*, 74(252):1743–1758, 2005.

[21] C. Clavero and J. C. Jorge. Another uniform convergence analysis technique of some numerical methods for parabolic singularly perturbed problems. *Computers and Mathematics with Applications*, In press, 2015.

[22] T. A. Davis. *Direct methods for sparse linear systems*, volume 2 of *Fundamentals of Algorithms*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2006.

[23] T.A. Davis. Algorithm 832: UMFPACK, an unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):196–199, 2004.

[24] T.A. Davis. A column pre-ordering strategy for the unsymmetric-pattern multifrontal method. *ACM Trans. Math. Softw.*, 30(2):165–195, 2004.

[25] C. de Falco and E. O'Riordan. A patched mesh method for singularly perturbed reaction-diffusion equations. In *BAIL 2008—boundary and interior layers*, volume 69 of *Lect. Notes Comput. Sci. Eng.*, pages 117–127. Springer, Berlin, 2009.

[26] James W. Demmel. *Applied numerical linear algebra*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[27] J. E. Dendy, Jr. Black box multigrid. *J. Comput. Phys.*, 48(3):366–386, 1982.

[28] E. P. Doolan, J. J. H. Miller, and W. H. A. Schilders. *Uniform numerical methods for problems with initial and boundary layers*. Boole Press, Dún Laoghaire, 1980.

[29] Iain S. Duff. MA57—a code for the solution of sparse symmetric definite and indefinite systems. *ACM Trans. Math. Software*, 30(2):118–144, 2004.

[30] Howard C. Elman, David J. Silvester, and Andrew J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics*. Numerical Mathematics and Scientific Computation. Oxford University Press, New York, 2005.

[31] Björn Engquist and Lexing Ying. Sweeping preconditioner for the Helmholtz equation: moving perfectly matched layers. *Multiscale Model. Simul.*, 9(2):686–710, 2011.

[32] Donald J. Estep, Mats G. Larson, and Roy D. Williams. Estimating the error of numerical solutions of systems of reaction-diffusion equations. *Mem. Amer. Math. Soc.*, 146(696):viii+109, 2000.

[33] 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.

[34] P. A. Farrell and G. I. Shishkin. On the convergence of iterative methods for linear systems arising from singularly perturbed equations. *in Proc. Copper Conference Conf. on Iterative Methods*, pages 1–7, 1998.

[35] Paul A. Farrell. Sufficient conditions for the uniform convergence of a difference scheme for a singularly perturbed turning point problem. *SIAM J. Numer. Anal.*, 25(3):618–643, 1988.

[36] S. Franz and H.-G. Roos. The capriciousness of numerical methods for singular perturbations. *SIAM Rev.*, 53(1):157–173, 2011.

[37] Eugene C. Gartland, Jr. Graded-mesh difference schemes for singularly perturbed two-point boundary value problems. *Math. Comp.*, 51(184):631–657, 1988.

[38] F. Gaspar, C. Clavero, and F. Lisbona. Some numerical experiments with multigrid methods on Shishkin meshes. *J. Comput. Appl. Math.*, 138(1):21–35, 2002.

[39] F. Gaspar, F. Lisbona, and C. Clavero. Multigrid methods and finite difference schemes for 2D singularly perturbed problems. In *Numerical analysis and its applications (Rousse, 2000)*, volume 1988 of *Lecture Notes in Comput. Sci.*, pages 316–324. Springer, Berlin, 2001.

[40] S. Geršgorin. Fehlerabschätzung für das differenzenverfahren zur lösung partieller differentialgleichungen. *Z. Angew. Math. Mech.*, 10:373–382, 1930.

[41] Adrianna Gillman and Per-Gunnar Martinsson. An $O(N)$ algorithm for constructing the solution operator to 2D elliptic boundary value problems in the absence of body loads. *Adv. Comput. Math.*, 40(4):773–796, 2014.

[42] Gene H. Golub and Charles F. Van Loan. *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition, 1996.

[43] J. L. Gracia and E. O'Riordan. Interior layers in a singularly perturbed time dependent convection-diffusion problem. *Int. J. Numer. Anal. Model.*, 11(2):358–371, 2014.

[44] Anne Greenbaum. *Iterative methods for solving linear systems*, volume 17 of *Frontiers in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997.

[45] Leslie Greengard, Denis Gueyffier, Per-Gunnar Martinsson, and Vladimir Rokhlin. Fast direct solvers for integral equations in complex three-dimensional domains. *Acta Numer.*, 18:243–275, 2009.

[46] Ivar Gustafsson. A class of first order factorization methods. *BIT*, 18(2):142–156, 1978.

[47] Wolfgang Hackbusch. *Multigrid methods and applications*, volume 4 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, 1985.

[48] A. M. Il′in. A difference scheme for a differential equation with a small parameter multiplying the highest derivative. *Mat. Zametki*, 6:237–248, 1969.

[49] R. B. Kellogg, T. Linß, and M. 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.

[50] R. B. Kellogg and A. Tsan. Analysis of some difference approximations for a singular perturbation problem without turning points. *Math. Comp.*, 32(144):1025–1039, 1978.

[51] R. B. Kellogg and C. Xenophontos. An enriched subspace finite element method for convection-diffusion problems. *Int. J. Numer. Anal. Model.*, 7(3):477–490, 2010.

[52] R.B. Kellogg, N. Madden, and M. Stynes. A parameter-robust numerical method for a system of reaction-diffusion equations in two dimensions. *Numer. Methods Partial Differ. Equations*, 24(1):312–334, 2008.

[53] N. Kopteva and E. O'Riordan. Shishkin meshes in the numerical solution of singularly perturbed differential equations. *Int. J. Numer. Anal. Model.*, 7(3):393–415, 2010.

[54] N. Kopteva and M. Pickett. A second-order overlapping Schwarz method for a 2D singularly perturbed semilinear reaction-diffusion problem. *Math. Comp.*, 81(277):81–105, 2012.

[55] O. Lawlor, H. Govind, I. Dooley, M. Breitenfeld, and L. Kale. Performance degradation in the presence of subnormal floating-point values. *in Proceedings of the International Workshop on Operating System Interference in High Performance Applications*, September 2005.

[56] D. Leykekhman. Uniform error estimates in the finite element method for a singularly perturbed reaction-diffusion problem. *Math. Comp.*, 77(261):21–39 (electronic), 2008.

[57] T. Linß. An upwind difference scheme on a novel Shishkin-type mesh for a linear convection-diffusion problem. *J. Comput. Appl. Math.*, 110(1):93–104, 1999.

[58] T. Linß. Layer-adapted meshes for convection-diffusion problems. *Comput. Methods Appl. Mech. Engrg.*, 192(9-10):1061–1105, 2003.

[59] T. Linß. Analysis of an upwind finite-difference scheme for a system of coupled singularly perturbed convection-diffusion equations. *Computing*, 79(1):23–32, 2007.

[60] T. Linß. Analysis of a system of singularly perturbed convection-diffusion equations with strong coupling. *SIAM J. Numer. Anal.*, 47(3):1847–1862, 2009.

[61] T. Linß. *Layer-adapted meshes for reaction-convection-diffusion problems*, volume 1985 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin, 2010.

[62] T. Linß and N. Madden. An improved error estimate for a numerical method for a system of coupled singularly perturbed reaction-diffusion equations. *Comput. Methods Appl. Math.*, 3(3):417–423, 2003. Dedicated to John J. H. Miller on the occasion of his 65th birthday.

[63] T. Linß and N. Madden. Accurate solution of a system of coupled singularly perturbed reaction-diffusion equations. *Computing*, 73(2):121–133, 2004.

[64] T. Linß and N. Madden. Parameter uniform approximations for time-dependent reaction-diffusion problems. *Numer. Methods Partial Differential Equations*, 23(6):1290–1300, 2007.

[65] T. Linß and N. Madden. Layer-adapted meshes for a linear system of coupled singularly perturbed reaction-diffusion problems. *IMA J. Numer. Anal.*, 29(1):109–125, 2009.

[66] T. Linß and N. Madden. Analysis of an alternating direction method applied to singularly perturbed reaction-diffusion problems. *Int. J. Numer. Anal. Model.*, 7(3):507–519, 2010.

[67] T. Linß, H.-G. Roos, and R. Vulanović. Uniform pointwise convergence on Shishkin-type meshes for quasi-linear convection-diffusion problems. *SIAM J. Numer. Anal.*, 38(3):897–912, 2000.

[68] T. Linß and M. Stynes. A hybrid difference scheme on a Shishkin mesh for linear convection-diffusion problems. *Appl. Numer. Math.*, 31(3):255–270, 1999.

[69] T. Linß and M. Stynes. Asymptotic analysis and Shishkin-type decomposition for an elliptic convection-diffusion problem. *J. Math. Anal. Appl.*, 261(2):604–632, 2001.

[70] F. Liu, N. Madden, M. Stynes, and A. Zhou. A two-scale sparse grid method for a singularly perturbed reaction-diffusion problem in two dimensions. *IMA J. Numer. Anal.*, 29(4):986–1007, 2009.

[71] Jens Lorenz. Stability and monotonicity properties of stiff quasilinear boundary problems. *Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser. Mat.*, 12:151–175, 1982.

[72] S. MacLachlan and N. Madden. Robust Solution of Singularly Perturbed Problems Using Multigrid Methods. *SIAM J. Sci. Comput.*, 35(5):A2225–A2254, 2013.

[73] N. Madden and M. Stynes. A uniformly convergent numerical method for a coupled system of two singularly perturbed linear reaction-diffusion problems. *IMA J. Numer. Anal.*, 23(4):627–644, 2003.

[74] N. Madden, M. Stynes, and G. P. Thomas. On the application of robust numerical methods to a complete-flow wave-current model. In *Boundary and Interior Layers (BAIL), Toulouse*, 2004.

[75] J. A. Meijerink and H. A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix. *Math. Comp.*, 31(137):148–162, 1977.

[76] J. M. Melenk and C. Xenophontos. A robust exponential convergence of hp-FEM in balanced norms for singularly perturbed reaction-diffusion equations. *Calcolo*, In press, 2015.

[77] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. *Fitted numerical methods for singular perturbation problems.* World Scientific Publishing Co. Inc., River Edge, NJ, 1996.

[78] 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.

[79] J. J. H. Miller, E. O'Riordan, G. I. Shishkin, and L. P. Shishkina. Fitted mesh methods for problems with parabolic boundary layers. *Math. Proc. R. Ir. Acad.*, 98A(2):173–190, 1998.

[80] C. B. Moler. Floating point numbers. *Cleve's Corner: Cleve Moler on Mathematics and Computing*, `http:blogs.mathworks.com/cleve/2014/07/07/floating-point-numbers`, 2014.

[81] Cleve B. Moler. *Numerical computing with MATLAB.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2004.

[82] K. W. Morton. *Numerical solution of convection-diffusion problems*, volume 12 of *Applied Mathematics and Mathematical Computation*. Chapman & Hall, London, 1996.

[83] T. A. Nhan and N. Madden. Cholesky factorization of linear systems coming from finite difference approximations of singularly perturbed problems. In *BAIL 2014—boundary and interior layers*, Lect. Notes Comput. Sci. Eng. Springer, Berlin, 2015, to appear.

[84] T. A. Nhan and N. Madden. An analysis of simple preconditioners for a singularly perturbed problem on a layer-adapted mesh. Submitted for publication, July 2015.

[85] T. A. Nhan and R. Vulanović. Uniform convergence on a Bakhvalov-type mesh using the preconditioning approach. Technical report, arXiv:1504.04283, 2015.

[86] Maxim A. Olshanskii and Eugene E. Tyrtyshnikov. *Iterative methods for linear systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2014.

[87] Michael L. Overton. *Numerical computing with IEEE floating point arithmetic*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001.

[88] Jack Poulson, Björn Engquist, Siwei Li, and Lexing Ying. A parallel sweeping preconditioner for heterogeneous 3D Helmholtz equations. *SIAM J. Sci. Comput.*, 35(3):C194–C212, 2013.

[89] Murray H. Protter and Hans F. Weinberger. *Maximum principles in differential equations*. Springer-Verlag, New York, 1984. Corrected reprint of the 1967 original.

[90] H.-G. Roos. A note on the conditioning of upwind schemes on Shishkin meshes. *IMA J. Numer. Anal.*, 16(4):529–538, 1996.

[91] H.-G. Roos. Robust numerical methods for singularly perturbed differential equations: a survey covering 2008–2012. *ISRN Appl. Math.*, pages Art. ID 379547, 30, 2012.

[92] H.-G. Roos and T. Linß. Sufficient conditions for uniform convergence on layer-adapted grids. *Computing*, 63(1):27–45, 1999.

[93] H.-G. Roos and M. Schopf. Convergence and stability in balanced norms of finite element methods on Shishkin meshes for reaction-diffusion problems. *ZAMM*, 95(6):551–565, 2015.

[94] H.-G. Roos and M. Schopf. The error of the Il′in scheme in 2d. To apprear BIT, 2015.

[95] H.-G. Roos and M. Stynes. Some open problems in the numerical analysis of singularly perturbed differential equations. *CMAM*, to apprear.

[96] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations*, volume 24 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, second edition, 2008.

[97] Yousef Saad. *Iterative methods for sparse linear systems*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, second edition, 2003.

[98] A. H. Schatz and L. B. Wahlbin. On the finite element method for singularly perturbed reaction-diffusion problems in two and one dimensions. *Math. Comp.*, 40(161):47–89, 1983.

[99] M. Schopf. *Error analysis of the Galerkin FEM in L_2-based norms for problems with layers*. PhD thesis, TU Dresden, 2014.

[100] G. Shishkin. *Grid Approximation of Singularly Perturbed Elliptic and Parapolic Equations*. Second doctoral thesis. Keldysh Institute, Moscow, 1990.

[101] G. I. Shishkin. *Discrete Approximation of Singularly Perturbed Elliptic and Parabolic Equations*. Russian Academy of Sciences, Ural Section, Ekaterinburg, 1992. In Russian.

[102] G. I. Shishkin and L. P. Shishkina. *Difference methods for singular perturbation problems*, volume 140 of *Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics*. CRC Press, Boca Raton, FL, 2009.

[103] M. Stephens and N. Madden. A parameter-uniform Schwarz method for a coupled system of reaction-diffusion equations. *J. Comput. Appl. Math.*, 230(2):360–370, 2009.

[104] M. Stephens and N. Madden. A Schwarz technique for a system of reaction diffusion equations with differing parameters. In *BAIL 2008—boundary and interior layers*, volume 69 of *Lect. Notes Comput. Sci. Eng.*, pages 247–255. Springer, Berlin, 2009.

[105] M. Stynes. Steady-state convection-diffusion problems. *Acta Numer.*, 14:445–508, 2005.

[106] M. Stynes and E. O'Riordan. An analysis of a singularly perturbed two-point boundary value problem using only finite element techniques. *Math. Comp.*, 56(194):663–675, 1991.

[107] M. Stynes and H.-G. Roos. The midpoint upwind scheme. *Appl. Numer. Math.*, 23(3):361–374, 1997.

[108] U. Trottenberg, C. W. Oosterlee, and A. Schüller. *Multigrid*. Academic Press, Inc., San Diego, CA, 2001.

[109] J. M. Varah. A lower bound for the smallest singular value of a matrix. *Linear Algebra and Appl.*, 11:3–5, 1975.

[110] Richard S. Varga. *Matrix iterative analysis*, volume 27 of *Springer Series in Computational Mathematics*. Springer-Verlag, Berlin, expanded edition, 2000.

[111] LAPACK version 3.5.0. *Subroutine* `DPRRTS`. www.netlib.org/lapack/.

[112] R. Vulanović. On a numerical solution of a type of singularly perturbed boundary value problem by using a special discretization mesh. *Univ. u Novom Sadu Zb. Rad. Prirod.-Mat. Fak. Ser. Mat.*, 13:187–201, 1983.

[113] R. Vulanović. A higher-order scheme for quasilinear boundary value problems with two small parameters. *Computing*, 67(4):287–303, 2001.

[114] R. Vulanović. A priori meshes for singularly perturbed quasilinear two-point boundary value problems. *IMA J. Numer. Anal.*, 21(1):349–366, 2001.

[115] R. Vulanović and T. A. Nhan. Uniform convergence via preconditioning. *Int. J. Numer. Anal. Model. Ser. B*, 5(4):347–356, 2014.

[116] Pieter Wesseling. *An introduction to multigrid methods*. Pure and Applied Mathematics (New York). John Wiley & Sons, Ltd., Chichester, 1992.