Iterative Solution of Discretized Convection-Diffusion Problems

Theory and Practice

vorgelegt von Master of Science

Carlos Echeverría Serur

ORCID: 0000-0002-1670-1405

an der Fakultät II – Mathematik und Naturwissenschaften
der Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –

genehmigte Dissertation.

Promotionsausschuss:

Vorsitzender:	Prof. Dr. Jochen Blath
Gutachter:	Prof. Dr. Jörg Liesen
Gutachter:	Prof. Dr. Andreas Frommer (Bergische Universität Wuppertal)

Tag der wissenschaftlichen Aussprache: 7. Juli 2020

Berlin 2020

To the memory of my father, and to my mother, who is pure life

Abstract

Certain approximation techniques for the numerical solution of partial differential equations result in linear algebraic systems where the coefficient matrix is nonsymmetric, nonnormal and ill-conditioned. This is the case for the finite difference discretization of the convection-diffusion equation posed on a Shishkin mesh treated in this work. We present a convergence analysis of the (algebraic) multiplicative Schwarz method when it is used to solve linear systems arising from both the upwind and central finite difference discretization approaches to such problems. For one and two dimensional problems, we show that the iteration matrix of the method has low-rank, which allows us to bound its ∞-norm. These bounds lead to quantitative error bounds for the iterates of the method that are valid from the first step of the iteration process. For problems in one-dimension, we prove rapid convergence of the method for all parameter choices of the problem when the upwind discretization approach is used, while convergence can only be proven for certain parameter choices for the central difference approach. Only the upwind discretization is considered in the case of two-dimensional problems. Furthermore, we consider the method as a preconditioner to GMRES and prove the convergence of the preconditioned method in a small number of steps when the local subdomain problems are solved exactly. Numerical experiments show that, for problems in two-dimensions, the number of iterations either stays the same or decreases for the case of inexact local solves, achieving a speed up in computational time. We continue by generalizing our convergence results to the case where the coefficient matrix of the linear system possess a special block structure that arises, for example, when a partial differential equation is posed and discretized on a domain that consists of two subdomains that overlap. Our analysis does not assume that the system matrices resulting from the discretization process are symmetric (positive definite) or posses the M- or H-matrix property. Instead, our results are obtained by generalizing the theory of diagonal dominant matrices from the scalar to the block case. Based on this generalization we present bounds on the norms of the inverses of general block tridiagonal matrices and derive a variant of the Gershgorin Circle Theorem that provides eigenvalue inclusion regions in the complex plane that are potentially tighter than the usual sets derived from the classical definition.

Zusammenfassung

Bestimmte Approximationstechniken für die numerische Lösung partieller Differentialgleichungen führen zu linearen algebraischen Systemen, bei denen die Koeffizientenmatrix nicht symmetrisch, nicht normal und schlecht konditioniert ist. Dies ist der Fall zum Beispiel bei der Diskretisierung der Konvektions-Diffusions-Gleichung, die in dieser Arbeit auf einem Shishkin-gitter aufgestellt wurde. Wir stellen eine Konvergenzanalyse der (algebraischen) multiplikativen Schwarz-Methode vor, wenn sie zur Lösung linearer Systeme verwendet wird, die sich sowohl aus dem Upwind als auch aus dem zentralen Finite-Differenz-Diskretisierungsansatz für solche Probleme ergeben. Für ein- und zweidimensionale Probleme zeigen wir, dass die Iterationsmatrix der Methode einen niedrigen Rang hat, was uns erlaubt, ihre ∞-Norm abzuschätzen. Diese Schranken führen zu quantitativen Fehlerschranke für die Iterationen der Methode, die ab dem ersten Schritt des Iterationsprozesses gültig sind. Bei eindimensionalen Problemen beweisen wir eine schnelle Konvergenz der Methode für alle Parameterwahlen des Problems, wenn der Upwind-Diskretisierungsansatz verwendet wird, während die Konvergenz nur für bestimmte Parameterwahlen für den zentralen Differenzansatz nachgewiesen werden kann. Bei zweidimensionalen Problemen wird nur die Upwind-Diskretisierung berücksichtigt. Darüber hinaus betrachten wir die Methode als Vorkonditionierer von GMRES und weisen die Konvergenz der vorkonditionierten Methode in wenigen Schritten nach, wenn die lokalen Teilbereichsprobleme exakt gelöst werden. Numerische Experimente zeigen, dass bei zweidimensionalen Problemen die Anzahl der Iterationen entweder gleich bleibt oder bei ungenauen lokalen Lösungen abnimmt, wodurch eine Beschleunigung der Rechenzeit erreicht wird. Wir fahren fort, indem wir unsere Konvergenzergebnisse auf den Fall verallgemeinern, dass die Koeffizientenmatrix des linearen Systems eine spezielle Blockstruktur besitzt, die sich z.B. ergibt, wenn eine partielle Differentialgleichung auf einem Gebiet gestellt und diskretisiert wird, die aus zwei Untergebiete besteht, die sich überlappen. Unsere Analyse geht nicht davon aus, dass die aus dem Diskretisierungsprozess resultierenden Systemmatrizen symmetrisch (positiv definit) sind oder die Eigenschaft der M- oder H-Matrix besitzen. Stattdessen erhalten wir unsere Ergebnisse durch Verallgemeinerung der Theorie der diagonaldominanten Matrizen vom skalaren zum Blockfall. Basierend auf dieser Verallgemeinerung beschränken wir die Norm der Inversen von allgemeinen Blocktridiagonalmatrizen und leiten eine Variante des Satz von Gershgorin her. Die Eigenwert-Einschlussbereiche in der komplexen Ebene liefert, die potenziell kleiner sind als die klassischen Gershgorin-Kreise.

Acknowledgements

First and foremost I would like to thank my supervisor, Jörg Liesen. He was responsible for introducing me to the fascinating area of mathematics that gave rise to this work (and in a deeper sense to the entire universe of mathematics), for which I will always be grateful. I owe the completion of this thesis to his guidance, wisdom and quasi-infinite patience over the past six years. Given his support, inspiring discussions, great ideas, and total freedom of research it has been a pleasure to work with, and learn important values and ideals from such a wholesome person and exceptional scientist.

Following closely behind in importance I would like to thank Evelyn Schubert, for teaching me all the things about life that one can not learn in the books. For sharing her love with me throughout these years, helping me grow and become the person that I am today. I can only feel a sense of infinite gratitude for your comfort, patience and love.

I will also always feel a deep sense of gratitude for my colleagues at the Institute for Mathematics of the TU Berlin, my collaborators, supervisors, mentors, co-authors, and referees from other institutions: Reinhard Nabben, Petr Tichý, Daniel Szyld, Andreas Frommer, Michele Benzi, Kees Vuik, Volker Mehrmann, Mikhail Karow, Oliver Sète, Jan Zur, René Kehl, Candan Güdücü and everyone who shared insightful and great discussions with me throughout these years. Thank you for your sympathy, and your help! I will always remember the endless cups of coffee (*yerba mate* in my case) that we shared together.

I would also like to thank the institutions who have sustained my research throughout my time as a doctoral student: The Technische Universität Berlin, the Berlin Mathematical School, the Berlin Einstein Center for Mathematics, The Berlin Mathematics Research Center MATH+, the “Deutscher Akademischer Austauschdienst” (DAAD) in Germany, and the “Consejo Nacional de Ciencia y Tecnología” (CONACYT) in Mexico. This work could not have been achieved without their generous support.

Furthermore, I deeply wish to thank my close personal friends who include, but are not limited to: Dominik Dreiner, Maximillian-Johannes-Peter Geisthardt, Luis García Ramos, André Gaul, Pabla Fiorella Perez, Alexander Hamel, Christine Eberhardt, Judith Wolff, Héctor Andrade Loarca, Tobias Wieland, Rodrigo Chavezmontes Jato, and everyone else who kept close to me even from the other side of the ocean. Your friendship kept me going through the good and the bad times, where I needed it the most. Thank you for being there.

Lastly, but most importantly than anybody else, I would like to thank my family. I will be eternally grateful to my mother, Raquel Serur, and father, Bolivar Echeverría, to whom I owe my life. I thank them for the love, dedication, support and education that was given to me from the day I was born. And to Alberto Echeverría, my favorite person on this Earth, thank you for being my brother, both in flesh and soul.

To everyone else whom I have forgotten to mention: Thank you! .

Contents

1. Introduction

In a wide range of applications in the general field of scientific computing it is important to solve linear algebraic systems of the form

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{1.1}$$

where the matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ in (1.1) is nonnormal, ill-conditioned and possesses a specific block sparse structure. Very often, matrices of this class arise when discretizing boundary value problems (BVPs); mathematical models which try to describe the asymptotic (steady-state) solution of a particular partial differential equation (PDE) posed on a specific domain. In particular they appear in the discretization process of BVPs that describe the behavior of fluid flow, like the ones modelled by the steady-state convection-diffusion BVP:

$$-\epsilon\Delta u + \boldsymbol{\omega}\cdot\nabla u + \beta u = f \text{ in } \Omega, \quad u = g, \text{ on } \partial\Omega. \tag{1.2}$$

In the PDE of problem (1.2), the scalar-valued function u is commonly interpreted as the concentration of a transported quantity, $\boldsymbol{\omega}$ the velocity field or *wind* where the concentration is transported, ϵ the scalar diffusion coefficient, and β the scalar reaction coefficient - both measures of the amount of diffusion and production / destruction of the concentration throughout the domain Ω.

In most applications and relevant cases of convection-diffusion BVPs, it is important to study the *convection dominated* regime, i.e., the cases when $\|\boldsymbol{\omega}\| \gg \epsilon > 0$ in (1.2), leading to what is known in literature as a *singularly perturbed* boundary value problem and where ϵ is known as the *perturbation parameter*. One possible physical interpretation of this type of problems, as described by Elman, Silvester and Wathen in [25], may be the following: the unknown function u may represent the concentration of a pollutant, being transported, or "convected", along a river moving at velocity $\boldsymbol{\omega}$, while being subject to diffusive and reactive effects. In this context, the solution to the BVP would describe the final concentration of the pollutant at each point of the riverbed.

Singularly perturbed convection-diffusion BVPs of type (1.2) often exhibit the presence of *boundary layers*, small regions of the domain Ω where the solution u exhibits a sharp change in its gradient. In turn, the existence of boundary layers presents a challenge for numerical methods to finding an accurate numerical representation of u, and usually require special discretization techniques in order to guarantee the stability of a numerical solution method [60]. A very popular approach is to use finite difference methods for approximating the derivatives of (1.2) on a piecewise equidistant mesh, known as a *Shihskin mesh*, which decomposes the domain into subregions with different resolutions; typically allowing to emphasize computational attention in the region of interest inside of the boundary layers present in each particular problem; see e.g., [73]. A general overview of these types of problems, their difficulties and solution approaches can be found, e.g., in the excellent survey article [74].

The decomposition of the domain into various subregions, caused by the use of a Shishkin mesh to discretize the domain, is reflected on the resulting discretized convection-diffusion operators $\mathbf{A}$, by exhibiting a particular block sparse structure. Consequently, the property of block-sparsity in the entries of the operator suggests the implementation of iterative solution methods when solving linear systems of the form (1.1) with such coefficient matrices [69]. In particular, the use of domain decomposition methods like the multiplicative Schwarz method seems to be a natural choice of solution approach for these problems and, indeed, its efficiency is corroborated by numerical experiments (see Figures 3.1-3.5 which show the convergence of the multiplicative Schwarz method and compare them to Figures 2.8–2.8 which show the convergence of the unpreconditioned GMRES method). Most of the work presented in this thesis will concern with the analysis of the multiplicative Schwarz method for solving systems of type (1.1) coming from one- and two-dimensional finite difference discretizations of problems of type (1.2).

The (algebraic) multiplicative Schwarz method, which is often called the alternating Schwarz method (see [34] for a historical survey), is a stationary iterative method for solving large and sparse linear algebraic systems of the form (1.1). In each step of the method, the current iterate is multiplied by an iteration matrix that is the product of several factors, where each factor corresponds an inversion of only a restricted part of the matrix. In the context of interest of this thesis, i.e., the numerical solution of discretized convection-diffusion problems, the restrictions of the matrix correspond to different parts of the computational domain subdivided by the Shishkin mesh. This motivates the name “local solve”, which is a popular term used to describe each of these factors and is also used in a purely algebraic setting. The method is for the most part used as a preconditioner for a Krylov subspace method such as GMRES and many of its convergence results have been presented in that context; see, e.g., the treatment in the books [21, 76], and many references therein. When the method is considered from an algebraic point of view, as we do in this work, it is commonly treated as a s solution method; see, e.g., [5]. The convergence theory for the multiplicative Schwarz method is well established for important matrix classes including symmetric positive definite matrices and nonsingular M-matrices [5], symmetric indefinite matrices [32, 33], and H-matrices [14]. The derivation of convergence results for these matrix classes is usually based on splittings of $\mathbf{A}$ and no systematic convergence theory exists however for general nonsymmetric matrices.

Several authors have previously applied the alternating (or multiplicative) Schwarz method to the continuous problem (1.2) based on the partitioning of the domain into overlapping subdomains, and subsequently discretized by introducing uniform meshes on each subdomain; see, e.g., [28, 29, 55, 54, 56, 57, 59]. However, as clearly explained in [57], significant numerical problems including very slow convergence and accumulation of errors (up to the point of non-convergence of the numerical solution) can occur when layer-resolving mesh transition points are used in this setup. These problems are avoided in our approach, since we first discretize and then apply the multiplicative Schwarz method to the linear algebraic system. To the best of our knowledge, this approach has not been studied in the literature so far.

For problems in one spacial dimension, studied in Chapter 3, the structure of the coefficient matrices exhibits a tridiagonal structure and the main mathematical tool used in our analysis exploits the fact that such discrete operators are *diagonally dominant*. For problems in higher spatial dimensions, studied in Chapter 5, the discretized operator exhibits a block tridiagonal structure. In order to perform an analysis analogous to the one-dimensional case, we generalize the property of diagonal dominance from the scalar case to the case where the matrices possess a block structure and develop a new mathematical theory of *block diagonal dominant* matrices. The theory is presented in Chapter 4 and it is general enough that it can be applied to any matrix with a block structure, however it seems to be particularly useful for matrices coming from discretizations of PDEs.

As mentioned before, the system matrices we study in this work are nonsymmetric, nonnormal, ill-conditioned, and in particular not in one of the classes considered in [5, 14, 32, 33]. Moreover, our derivations are not based on matrix splittings, but on the off-diagonal decay of the matrix inverses, which in turn is implied by diagonal dominance. From a broader point of view our results show why a convergence theory for the multiplicative Schwarz method for "general" matrices will most likely remain elusive: Even in the simple model problem considered in Chapter 3 and in [23], the convergence of the method strongly depends on the problem parameters and on the chosen discretization, and while the method rapidly converges in some cases, it diverges in others.

1.1. Scope and Goals

The scope of this thesis aims to provide an analysis of the convergence behavior of the multiplicative Schwarz method when it is used to solve linear systems arising from special finite difference discretizations of singularly perturbed convection-diffusion problems posed on a Shishkin mesh. We restrict our attention to the cases where the domain Ω is one- or two-dimensional, and we analyze the method both as an algebraic solution approach as well as a preconditioner for the GMRES method.

The analysis presented in this work brings an understanding on why this solution technique is so effective for solving problems arising from the Shishkin mesh discretizations. We do this by providing convergence bounds for the norm of the error generated by the method at each iteration step. Moreover, the convergence bounds provided in this thesis, in the paper [22], and in the manuscript [24], shed light on an apparent contradiction: If the continuous problem becomes more difficult (a smaller diffusion coefficient is chosen), then the convergence of the multiplicative Schwarz method for the discretized problem becomes faster.

The mathematical tools developed in this work to achieve its main goal are, however, more general. The theory of block diagonal dominance of matrices presented in Chapter 4, with the bounds and eigenvalue inclusion sets resulting from our analysis, does not only apply to operators coming from discretizations of BVPs but is applicable to any matrix $\mathbf{A} \in \mathbb{C}^{N \times N}$ with a block structure.

1.2. Outline and Summary of Main Results

We begin by introducing a short review of the theoretical background material needed to understand the results presented in this thesis in Chapter 2. An experienced reader in these topics, might use this chapter for reference and directly visit Chapters 3 – 5 which encompass the main findings of this thesis: results for one-dimensional model problems, results for general matrices, and results for two-dimensional model problems. Finally, Chapter 6 provides a brief discussion and outlook on possible continuations of this work. The appendix A provides instructions for obtaining the computational code in order to perform and reproduce the numerical experiments presented throughout this thesis. In the following we provide a brief summary of the results obtained in each of the main chapters of this thesis:

Chapter 3: In this chapter, we analyze the convergence of the multiplicative Schwarz method applied to nonsymmetric linear algebraic systems obtained from discretizations of one-dimensional singularly perturbed convection-diffusion equations by upwind and central finite differences on a Shishkin mesh. Using the algebraic structure of the Schwarz iteration matrices we derive bounds on the infinity norm of the error that are valid from the first step of the iteration. Our bounds for the upwind scheme prove rapid convergence of the multiplicative Schwarz method for all relevant choices of parameters in the problem. The analysis for the central difference is more complicated, since the submatrices that occur are nonsymmetric and sometimes even fail to be M-matrices. Our bounds still prove the convergence of the method for certain parameter choices.

Chapter 4: Here we generalize the bounds on the inverses of diagonally dominant matrices obtained in [62] from scalar to block tridiagonal matrices. Our derivations are based on a generalization of the classical condition of block diagonal dominance of matrices given by Feingold and Varga in [30]. Based on this generalization, which was recently presented in [22] and a similar definition appearing first in [4], we also derive a variant of the Gershgorin Circle Theorem for general block matrices which can provide tighter spectral inclusion regions than those obtained by Feingold and Varga.

Chapter 5: Finally, we analyze the convergence of the multiplicative Schwarz method applied to linear algebraic systems with matrices having a special block structure that arises, for example, when a partial differential equation is posed and discretized on a domain that consists of two subdomains with an overlap. This is a basic situation in the context of domain decomposition methods. Again, our analysis is based on the algebraic structure of the Schwarz iteration matrices, and we derive error bounds that are based on the block diagonal dominance of the given system matrix. Our analysis does not assume that the system matrix is symmetric (positive definite), or has the M- or H-matrix property. Our approach in this chapter significantly generalizes the analysis for a special one-dimensional model problem treated in Chapter 3.

2. Background Material

"Accurate modelling of the interaction between convective and diffusive processes is the most ubiquitous and challenging task in numerical approximation of partial differential equations."

— K. W. Morton, *Numerical Solution of Convection-Diffusion Problems* [60]

2.1. Steady-State Convection-Diffusion Problems

Steady-state convection-diffusion problems arise in a wide range of mathematical models which aim to describe a variety of phenomena appearing in the natural world. In the area of fluid flow, which is the most attributed area of science where this equation plays an important role, they often appear in the linearization of the Navier-Stokes equations as well as the Oseen equations [7] and they typically aim to describe the transport of a certain concentration over a flow field. However, the range of application of these type of problems includes many other models of physical phenomena, like it is the case of the modeling of convective heat transport, the modeling of the concentrations of chemicals in a chemical reactor [38] as well as modeling the oil extraction from underground reservoirs [18]. Other physical phenomena, apparently unrelated to fluid flow, such as the modeling of electronic transport in semiconductor materials can also be described by such problems [58].

In their most simple form, steady-state convection-diffusion problems are described mathematically by the one-dimensional boundary value problem:

$$\begin{cases} -\epsilon \frac{\partial^2 u(x)}{\partial x^2} + \omega_x \frac{\partial u(x)}{\partial x} + \beta u(x) = f(x), \quad \text{in } (0,1) \\ \quad u(0) = g_0, \quad \text{and} \quad u(1) = g_1, \end{cases} \tag{2.1}$$

and more generally by its n-dimensional analogue:

$$\begin{cases} -\epsilon \Delta u(\mathbf{x}) + \boldsymbol{\omega} \cdot \nabla u(\mathbf{x}) + \beta u(\mathbf{x}) = f(\mathbf{x}), & \text{in } \Omega \in \mathbb{R}^n \\ \qquad\qquad\qquad\qquad u(\mathbf{x}) = g(\mathbf{x}). & \text{on } \partial\Omega. \end{cases} \tag{2.2}$$

The partial differential equation (PDE) in (2.1) or (2.2) is classified as a linear second-order elliptic PDE (see Section 2.1.1) where the term $\boldsymbol{\omega} \cdot \nabla u(\mathbf{x})$ in (2.2) models convection and the term $-\epsilon \Delta u(\mathbf{x})$ models diffusion. In most applications the parameter $\epsilon > 0$ is small, typically in the range $\mathcal{O}(10^{-2})$ to $\mathcal{O}(10^{-8})$ while the magnitude of the flow field $\|\boldsymbol{\omega}\|$ is typically of $\mathcal{O}(1)$, so that very often *convection dominates diffusion*. More specifically, in order to have a relative measure of which term is more dominant than the other it is common to introduce the *Peclet number*, Pe. If L is a length associated with the domain Ω (length of the interval in 1D, longest length inside the domain for 2D, etc.), then

$$\mathrm{Pe} = \frac{L}{\epsilon} \max\left\{|\omega_x|, |\omega_y|, \ldots, |\omega_d|\right\}, \tag{2.3}$$

where $\omega_x, \omega_y, \ldots \omega_d$ are the components of $\boldsymbol{\omega}$ and thus for most practical cases $\mathrm{Pe} \gg 1$, constituting what is known as a singularly perturbed PDE.

The region Ω can be any reasonable domain in n-dimensions and the velocity field $\boldsymbol{\omega}$, which throughout this work will be considered as incompressible (i.e., we assume that $\nabla \cdot \boldsymbol{\omega} = 0$), is used as reference to partition the boundary of the domain, $\partial\Omega$, as follows:

$$\begin{aligned} \partial\Omega_+ &= \{\mathbf{x} \text{ on } \partial\Omega \mid \boldsymbol{\omega}(\mathbf{x}) \cdot \boldsymbol{n}(\mathbf{x}) > 0\}, && \text{the } \textit{outflow boundary,} \\ \partial\Omega_0 &= \{\mathbf{x} \text{ on } \partial\Omega \mid \boldsymbol{\omega}(\mathbf{x}) \cdot \boldsymbol{n}(\mathbf{x}) = 0\}, && \text{the } \textit{characteristic boundary,} \\ \partial\Omega_- &= \{\mathbf{x} \text{ on } \partial\Omega \mid \boldsymbol{\omega}(\mathbf{x}) \cdot \boldsymbol{n}(\mathbf{x}) < 0\}, && \text{the } \textit{inflow boundary,} \end{aligned} \tag{2.4}$$

where $\boldsymbol{n}(\mathbf{x})$ denotes the unit outward pointing normal vector to the boundary at the point $\mathbf{x}$.

Since a typical value of ϵ in a convection dominated problem is of $\mathcal{O}(10^{-6})$, the influence of the diffusion term is very small and the solution u to (2.2) is usually very close to the solution $\hat{u}$ of the equation:

$$\boldsymbol{\omega} \cdot \nabla \hat{u}(\mathbf{x}) + \beta \hat{u}(\mathbf{x}) = f(\mathbf{x}), \tag{2.5}$$

commonly known as the *reduced problem*. It is important to note that the highest order derivative in equation (2.5) is of lower order than that of (2.2), and its solution $\hat{u}$ will most likely not be able to satisfy all the boundary conditions imposed on the original problem. In order to do so, the solution u to the convection-diffusion BVP usually presents sharp gradients close to the outflow boundary of the domain and u is said to have have an *exponential boundary layer* along the outflow boundary $\partial\Omega_+$. Sharp gradients of the solution may also appear in the interior portion of Ω if any number of discontinuities are present in the boundary conditions set on $\partial\Omega_-$. The diffusion term takes care of smoothing such discontinuities into a continuos but steep *characteristic/internal boundary layer*. The appearance of both types of boundary layers represent the main challenge when constructing numerical approximations to the solutions of (2.1) and (2.2) and constitutes the main reason why most numerical methods will not work satisfactorily when solving these type of problems; see [60]. Thus, the accurate solution of such problems requires the use of special discretization techniques and/or the use of specially *fitted methods* which modify the linear operator in (2.2) in order to guarantee the stability of a numerical method [73]. A general overview of accurate solution techniques can be found, e.g., in the survey article [74] or in the monographs [53, 67]. One widely accepted discretization technique in this context, and which represents the main focus of this work, is given by using upwind or central finite difference schemes posed on a Shishkin mesh as described, e.g., in [74, § 5] or [28, 47, 52, 59].

2.1.1. Elliptic Operators

In its most general form, a linear second-order partial differential equation can be written as

$$\hat{\mathcal{A}}\hat{u} = \hat{f}. \tag{2.6}$$

Usually the PDE holds on a domain $\Omega \in \mathbb{R}^2$ where the operator $\hat{\mathcal{A}}$ can be written as

$$\hat{\mathcal{A}} \equiv a\frac{\partial^2}{\partial x^2} + 2b\frac{\partial^2}{\partial x \partial y} + c\frac{\partial^2}{\partial y^2} + p\frac{\partial}{\partial x} + q\frac{\partial}{\partial y} + r, \tag{2.7}$$

where the coefficients a, b, c, p, q, r usually depend on both space variables x and y. The behavior of the solutions to (2.6) is usually determined by the highest order derivatives in the operator $\hat{\mathcal{A}}$ so for the sake of discussion and without loss of generality we will consider $p = q = r = 0$ and $a \geq 0$. There are three alternatives for classifying such operators (see [25, 27] for a detailed description of the classification of PDEs), which are

$$\begin{array}{ll} b^2 - 4ac > 0, & \textit{Hyperbolic-type,} \\ b^2 - 4ac = 0,\ a > 0, & \textit{Parabolic-type,} \\ b^2 - 4ac < 0, & \textit{Elliptic-type.} \end{array} \tag{2.8}$$

The convection-diffusion equation in (2.2) defines a linear second order differential operator of elliptic-type, given by:

$$\mathcal{A}u \equiv -\epsilon\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) + \omega_x\frac{\partial u}{\partial x} + \omega_y\frac{\partial u}{\partial y} + \beta u \tag{2.9}$$

where, in general, the coefficients $\epsilon, \omega_x, \omega_y$ and β represent functions of the space variables x and y. Very often in practice, and throughout this work, they will be considered constant unless otherwise specified. In order to have a uniquely defined solution an appropriate condition must be specified at each point of the boundary $\partial\Omega$. By partitioning the boundary into three disjoint sets Γ_i, $i = 1, 2, 3$ with $\partial\Omega = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$, and $\boldsymbol{n}(\mathbf{x})$ denotes the unit outward normal vector to $\partial\Omega$ at the point $\mathbf{x} = (x, y)$ then the boundary conditions are given by

$$\begin{array}{rll} u(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \Gamma_1, & \textit{Dirichlet,} \\ \boldsymbol{n}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \Gamma_2, & \textit{Neumann,} \\ \gamma(\mathbf{x})u(\mathbf{x}) + \delta(\mathbf{x})\boldsymbol{n}(\mathbf{x}) \cdot \nabla u(\mathbf{x}) = g(\mathbf{x}), & \mathbf{x} \in \Gamma_3, & \textit{Robin,} \end{array} \tag{2.10}$$

where γ, δ, and g are known functions. More generally, the boundary conditions can be written in compact form as $\mathcal{B}u = g$ so that the boundary value problem can be written as $\mathscr{A}u = \mathscr{F}$, where

$$\mathscr{A}u = \begin{cases} \mathcal{A}u & \text{in} \quad \Omega, \\ \mathcal{B}u & \text{on} \quad \partial\Omega, \end{cases}, \quad \text{and} \quad \mathscr{F} = \begin{cases} f & \text{in} \quad \Omega, \\ g & \text{on} \quad \partial\Omega. \end{cases} \tag{2.11}$$

In further sections of this work we will only treat problems with Dirichlet boundary conditions.

2.1.2. Finite Difference Methods

The aim of numerical approximation methods for solving PDEs is to construct accurate numerical solutions to BVPs of the form (2.11) that may or may not have analytical solutions. Of the available methods, those employing finite differences or finite elements are more frequently used and more universally applicable than any other, although finite elements are not considered in this work. For an excellent treatment of solving BVPs with finite elements please refer to the monograph [25], for a comprehensive survey on the theory of finite elements see [12].

In short, in a finite difference method a set of mesh of points is introduced within the domain of definition of the BVP and then any derivatives appearing in the PDE or boundary conditions are replaced by finite difference approximations at the mesh points. Thus, finite difference methods are approximate in the sense that derivatives *at a point* are approximated by difference quotients *over a small interval*, i.e., $u_x = \partial u/\partial x$ is replaced by $\Delta u/\Delta x$ where Δx is small and y is constant, but they are *not* approximate in the sense of being crude estimates. In the following we will describe the method for a two-dimensional model problem.

As much as the analytic solution of the BVP is a function u defined on a domain Ω, say $[0,1] \times [0,1]$, finite difference methods seek a numerical solution only at a finite number of grid points:

$$\Omega_D = \{(x_i, y_j) : i = 0, \dots, N; j = 0, \dots, M\}.$$

The grid points on the boundary are denoted by $\partial\Omega_D$, and the entire grid by $\overline{\Omega}_D = \Omega_D \cup \partial\Omega_D$. We can evaluate the exact solution of the BVP on the introduced points to obtain the values

$$\begin{matrix} u(x_0, y_M) & u(x_1, y_M) & \dots & u(x_N, y_M) \\ \vdots & \vdots & & \vdots \\ u(x_0, y_1) & u(x_1, y_1) & \dots & u(x_N, y_1) \\ u(x_0, y_0) & u(x_1, y_0) & \dots & u(x_N, y_0). \end{matrix}$$

The finite difference approximation scheme replaces the given BVP with a set of $(N+1) \times (M+1)$ finite difference equations in $(N+1) \times (M+1)$ unknowns. The function u^D, commonly referred to as a *grid function* is only defined on the set of grid points $\overline{\Omega}_D$ and its value at the grid point $\mathbf{x}_{i,j} = (x_i, y_j)$ is denoted by $u^D_{i,j}$.

$$\begin{matrix} u^D_{0,M} & u^D_{1,M} & \dots & u^D_{N,M} \\ \vdots & \vdots & & \vdots \\ u^D_{0,1} & u^D_{1,1} & \dots & u^D_{N,1} \\ u^D_{0,0} & u^D_{1,0} & \dots & u^D_{N,0} \end{matrix}$$

The solution to the finite difference equations can then be identified with the grid function u^D, whose value $u^D_{i,j}$ at a typical point $\mathbf{x}_{i,j} = (x_i, y_j) \in \Omega_D$ approximates the exact solution $u_{i,j} \equiv u(x_i, y_j)$ at that point. The values of u^D at the internal grid points are found by solving the system of finite difference equations while the values of u^D at the boundary nodes are known from the Dirichlet conditions. The premise of the method being that each element from the set $u^D_{0,0}, \dots, u^D_{N,M}$ should converge to its corresponding element in $u(x_0, y_0), \dots, u(x_N, y_M)$ as $N, M \to \infty$.

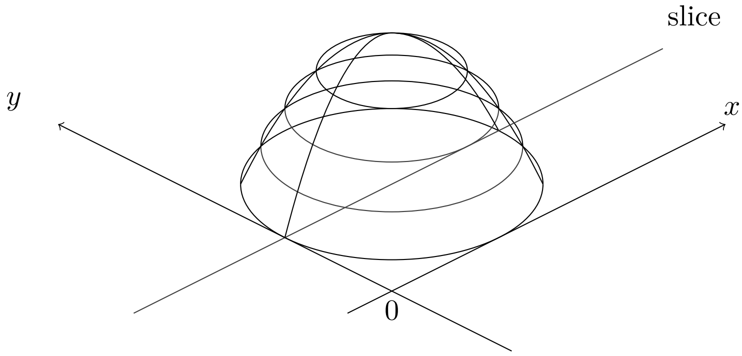


Figure 2.1.: Generic function of two variables with sufficiently many derivatives.

The interchange of differential equations and their boundary conditions by algebraic equations in the finite difference method is accomplished by analyzing the Taylor series expansions at each point. To exemplify this process we will assume that the function $v(\mathbf{x}) = v(x, y)$ possesses enough continuous derivatives for the Taylor series expansions to be well defined. An example of such a function is given in Figure 2.1

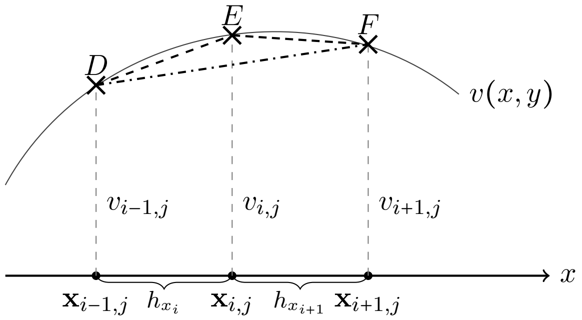


Figure 2.2.: One-dimensional slice of the function depicted in Figure 2.1. The chords DE, EF and DF are different possibilities of approximating the slope of the tangent at the point E.

Figure 2.2 represents a slice of the function pictured in Figure 2.1 for a constant value of y, the gradient of the function $v(\mathbf{x}) = v(x, y)$ at point E ($\mathbf{x}_{i,j} = (x_i, y_j)$) in the x-direction may be approximated by the gradients of any of the three chords DE, EF, or DF, each case with its own degree of approximation. The slope of the backward chord at the point E ($\mathbf{x}_{i,j} = (x_i, y_j)$) can be associated to the the following one-directional Taylor expansion in two-dimensions:

$$v(x - \Delta x, y) = v(x, y) - \frac{\partial v(x, y)}{\partial x}\Delta x + \frac{\partial^2 v(\xi, y)}{\partial x^2}\frac{\Delta x^2}{2}, \tag{2.12}$$

for some number $\xi \in ((x, y), (x + \Delta x, y))$, we choose the point $\mathbf{x}_{i,j} = (x_i, y_j)$ and rearrange the previous expression to give:

$$\frac{\partial v(x_i, y_j)}{\partial x} = h_{x_i}^{-1}(v_{i,j} - v_{i-1,j}) + \frac{h_{x_i}}{2}\frac{\partial^2 v(\xi_i, y_j)}{\partial x^2}, \quad \xi_i \in (x_{i-1}, x_i),$$

where the subscript i in ξ_i reflects its dependence on (x_i, y_j). We have

$$\frac{\partial v(x_i, y_j)}{\partial x} = h_{x_i}^{-1}(v_{i,j} - v_{i-1,j}) + R_{i,j} = h_{x_i}^{-1}(v_{i,j} - v_{i-1,j}) + \mathcal{O}(h_{x_i}).$$

The remainder term, commonly referred to as the *local truncation error* $R_{i,j}$ (see [64] for a good discussion of error distributions), corresponds to truncating the Taylor series (2.12). When the local truncation error is neglected, we obtain the *backward difference approxiamtion* of $\frac{\partial v(x_i,y_j)}{\partial x}$ at the point $\mathbf{x}_{i,j} = (x_i, y_j)$. The backward difference operator or *upwind finite difference operator* in the x-direction, Δ_x^-, is defined by

$$\Delta_x^- v_{i,j} \equiv v_{i,j} - v_{i-1,j}, \tag{2.13}$$

and so $\frac{\partial v(x_i,y_j)}{\partial x} \approx h_{x_i}^{-1}\Delta_x^- v_{i,j}$ with an error $\mathcal{O}(h_{x_i})$.

We chose the expansion (2.12), however we can proceed in a similar fashion using the forward version of the Taylor series instead:

$$v(x + \Delta x, y) = v(x, y) + \frac{\partial v(x, y)}{\partial x}\Delta x + \frac{\partial^2 v(\xi, y)}{\partial x^2}\frac{\Delta x^2}{2} \tag{2.14}$$

and by defining the *forward finite difference operator* in the x-direction, Δ_x^+,

$$\Delta_x^+ v_{i,j} \equiv v_{i+1,j} - v_{i,j}, \tag{2.15}$$

we obtain another approximation to the derivative of v at the point $\mathbf{x}_{ij}$, $\frac{\partial v(x_i,y_j)}{\partial x} \approx h_{x_{i+1}}^{-1}\Delta_x^+ v_{i,j}$ now with an error of $\mathcal{O}(h_{x_{i+1}})$.

Finally, by defining the *central finite difference operator* in the x-direction, Δ_x^0,

$$\Delta_x^0 v_{i,j} \equiv v_{i+1,j} - v_{i-1,j}, \tag{2.16}$$

we find that $\frac{\partial v(x_i,y_j)}{\partial x} \approx (h_{x_i} + h_{x_{i+1}})^{-1}\Delta_x^0 v_{i,j}$ now with an error $\mathcal{O}\left((h_{x_i} + h_{x_{i+1}})^2\right)$. It is possible to obtain more accurate approximations by including more terms in the Taylor series (2.12) and (2.14) before truncating it, obtaining higher order methods (see [20] for a discussion in this direction).

In order to approximate second order derivatives (and most even-order derivatives), it is common to introduce another “artificial" central approximation

$$\delta_x v_{i,j} \equiv v_{i+\frac{1}{2},j} - v_{i-\frac{1}{2},j}, \tag{2.17}$$

which makes use of intermediate points which are not on the grid. This approximation corresponds to the Δ approximation with a half-step $h_{x_i}/2$. Nevertheless, by computing $\delta_x(\delta_x)v_{i,j}$ we obtain

$$\delta_x(\delta_x v_{i,j}) = v_{i+1,j} - 2v_{i,j} + v_{i-1,j}. \tag{2.18}$$

In the exact fashion as above, by considering both Taylor approximations (2.12) and (2.14), adding them together adn evaluatng at the point $\mathbf{x}_{i,j}$ we obtain a centered finite difference approximation to the second derivative,

$$\frac{\partial^2 v(x_i, y_j)}{\partial x^2} = h_{x_i}^{-2}(v_{i+1,j} - 2v_{i,j} + v_{i-1,j}) - \mathcal{O}(h_{x_i}^2) \approx h_{x_i}^{-2}\delta^2 v_{i,j},$$

with a remainder term proportional to $h_{x_i}^2$.

This analysis can be repeated for approximating the partial derivative of the function v in the y-direction to obtain analogous results. Using the notation $v_x = \frac{\partial v}{\partial x}$, $v_{xx} = \frac{\partial^2 v}{\partial x^2}$, and where $h_{x_{\text{eff}}} = h_{x_i} + h_{x_{i+1}}$ we summarize the results in Table 2.19

$$\begin{aligned}
&\Delta_x^+ v_{i,j} = v_{i+1,j} - v_{i,j} = h_{x_{i+1}} v_x + \tfrac{1}{2} h_{x_i}^2 v_{xx} + \mathcal{O}(h_{x_{i+1}}^3) && \textit{Forward differences}\\
&\Delta_y^+ v_{i,j} = v_{i,j+1} - v_{i,j} = h_{y_{i+1}} v_y + \tfrac{1}{2} h_{y_i}^2 v_{yy} + \mathcal{O}(h_{y_{i+1}}^3) && \textit{Forward differences}\\
&\Delta_x^- v_{i,j} = v_{i,j} - v_{i-1,j} = h_{x_i} v_x - \tfrac{1}{2} h_{x_i}^2 v_{xx} + \mathcal{O}(h_{x_i}^3) && \textit{Upwind differences}\\
&\Delta_y^- v_{i,j} = v_{i,j} - v_{i,j-1} = h_{y_i} v_y - \tfrac{1}{2} h_{y_i}^2 v_{yy} + \mathcal{O}(h_{y_i}^3) && \textit{Upwind differences}\\
&\Delta_x^0 v_{i,j} = v_{i+1,j} - v_{i-1,j} = h_{x_{\text{eff}}} v_x + \tfrac{1}{6} h_{x_{\text{eff}}}^3 v_{xxx} + \mathcal{O}(h_{x_{\text{eff}}}^5) && \textit{Central differences}\\
&\Delta_y^0 v_{i,j} = v_{i,j+1} - v_{i,j-1} = h_{y_{\text{eff}}} v_y + \tfrac{1}{6} h_{y_{\text{eff}}}^3 v_{yyy} + \mathcal{O}(h_{y_{\text{eff}}}^5) && \textit{Central differences}\\
&\delta_x^2 v_{i,j} = v_{i+1,j} - 2v_{i,j} + v_{i-1,j} = h_{x_{\text{eff}}}^2 v_{xx} + \tfrac{1}{12} h_{x_{\text{eff}}}^4 v_{xxxx} + \mathcal{O}(h_{x_{\text{eff}}}^6) && \textit{Central differences}\\
&\delta_y^2 v_{i,j} = v_{i,j+1} - 2v_{i,j} + v_{i,j-1} = h_{y_{\text{eff}}}^2 v_{yy} + \tfrac{1}{12} h_{x_{\text{eff}}}^4 v_{yyyy} + \mathcal{O}(h_{y_{\text{eff}}}^6) && \textit{Central differences}
\end{aligned} \tag{2.19}$$

2.1.3. Shishkin Mesh in 1D

In order to obtain an accurate numerical solution of BVPs with singularly perturbed convection-diffusion problems described by equation (2.1), we require special discretization techniques. As mentioned in Section 2.1, we focus on the approach of using upwind or central finite difference schemes posed on a Shishkin mesh, an approach described described, e.g., in [74, § 5] or [28, 47, 52, 59]. In this subsection we introduce the one-dimensional Shishkin mesh and describe its construction. For a comprehensive treatment of layer-adapted meshes for convection-diffusion problems see the monograph [53]. Without loss of generality, we assume that $\omega_x \gg \epsilon > 0$ and $\beta \geq 0$ and that the parameters of the problem (2.1), i.e., $\epsilon, \omega_x, \beta, f, g_0$, and g_1, are chosen so that the solution $u(x)$ has one boundary layer close to the point $x = 1$.

In short, Shishkin meshes are formed by an overlapping union of piecewise uniform meshes, with their respective sizes and mesh transition (or interface) points adapted to the expected width of the boundary layers in the solution. Suppose that an even

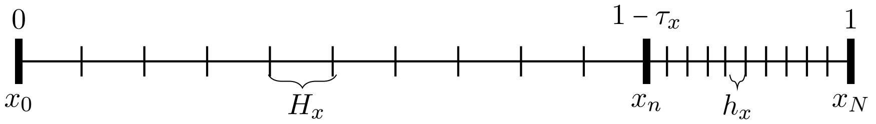


Figure 2.3.: Illustration of a one-dimensional Shishkin mesh.

integer $N \geq 4$ defining the number of intervals constituting the Shishkin mesh is given, and suppose that the *mesh transition parameter*, τ_x, fulfills

$$\tau_x \equiv 2\frac{\epsilon}{\omega}\ln N \leq \frac{1}{2}. \tag{2.20}$$

The inequality in (2.20) means that

$$\epsilon \leq \frac{\omega}{4\ln N}, \tag{2.21}$$

which is a natural assumption since $\omega \gg \epsilon$, and the number of mesh points usually is not exponentially large relative to ϵ. The *mesh transition point* $1-\tau_x$ then will be close to $x = 1$, and the boundary layer will be contained in the (small) interval $[1-\tau_x, 1]$. The idea of the Shishkin mesh discretization of the interval $[0,1]$ is to use the same number of equidistantly distributed mesh points in each of the subintervals $[0, 1-\tau_x]$ and $[1-\tau_x, 1]$ as can bee seen in Figure 2.3. Thus, if we denote

$$n \equiv \frac{N}{2}, \quad H_x \equiv \frac{1-\tau_x}{n}, \quad \text{and} \quad h_x \equiv \frac{\tau_x}{n}, \tag{2.22}$$

then the $N+1$ mesh points of the Shishkin mesh are given by

$$x_i \equiv iH_x, \quad i = 0,\dots,n, \qquad x_i \equiv 1-(N-i)h_x, \quad i = n+1,\dots,N.$$

Here $x_0 = 0$ and $x_N = 1$, so that the mesh consists of $N-1$ interior mesh points, where the mesh point x_n is exactly at the transition point $1-\tau_x$. The ratio between the mesh sizes in the two subdomains is

$$\frac{h_x}{H_x} = \frac{\tau_x}{1-\tau_x} = \tau_x + \mathcal{O}(\tau_x^2),$$

which is usually much less than 1.

Any Shishkin mesh discretization naturally leads to a decomposition of the given domain into overlapping subdomains. In our one-dimensional model problem the domain is the interval $[0,1]$, and the overlapping subdomains are the intervals $[0, 1-\tau_x+h_x]$ and $[1-\tau_x-H_x, 1]$. The width of the overlap is $H_x + h_x = 2/N$, and the mesh transition point $x_n = 1-\tau_x$ is the only mesh point in the overlap. An illustration of a Shishkin mesh is shown in Figure 2.3, and a plot of the (explicitly known) analytic solution of the problem (2.1) with $\epsilon = 0.03$, $\omega = 1$, $\beta = 0$, $f(x) \equiv 1$, and $g_0 = g_1 = 0$ is shown in Figure 2.4 cf. [74, Example 3.1]. Choosing, for example, $N = 48$ gives the mesh transition point $1-\tau_x = 0.7677$ (these parameters were chosen for presentation purposes only, if we choose $\epsilon = 0.01$ while keeping the other parameters constant we obtain $1-\tau_x = 0.9226$).

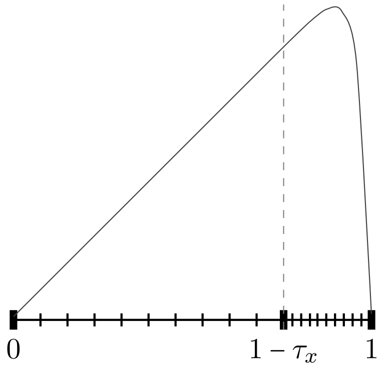


Figure 2.4.: Analytic solution of problem (2.1) with $\epsilon = 0.03$, $\omega = 1$, $\beta = 0$, $f(x) \equiv 1$, and $g_0 = g_1 = 0$. For $N = 48$ the mesh transition point is $1 - \tau_x = 0.7677$.

2.1.4. Shishkin Mesh in 2D

In the case of two spacial dimensions, we can create various types of Shishkin meshes depending on the number of outflow boundary layers in each coordinate direction for a given problem. In the case of one boundary layer in one direction and no layer in the other direction we will use a combination of a regular discretization in the coordinate direction without a boundary layer with a Shishkin mesh discretization in the coordinate direction with an outflow boundary layer. In the other case we will use two Shishkin meshes, one for each coordinate direction.

2.1.4.1. One Outflow Boundary Layer

The simplest generalization of the Shishkin mesh for the case of two spacial dimensions is to use a discretization approach that combines a regular mesh in one coordinate direction (say x) with a Shishkin mesh discretization in the other coordinate direction (say y).

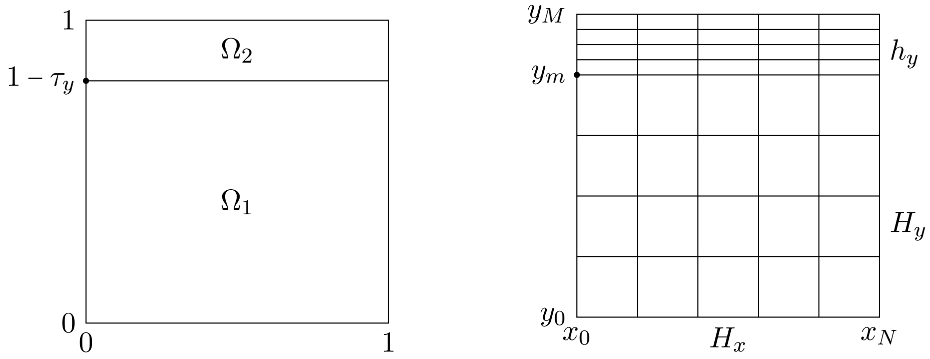


Figure 2.5.: Division of the domain and Shishkin mesh for equation (2.2) with one outflow exponential layer.

Given now two even positive integers $N \geq 4$ and $M \geq 4$ that denote the number of mesh intervals used in each coordinate direction, we let the transition parameter τ_y that will be used to specify where the mesh changes from coarse to fine in the y-direction, be defined by

$$\tau_y \;\equiv\; \min\left\{\frac{1}{2}, 2\frac{\epsilon}{\omega_y}\ln M\right\}. \tag{2.23}$$

Since we assume $\omega_y \gg \epsilon$ (or equivalently that $\epsilon \leq CM^{-1}$), we will have $\tau_y \;=\; 2\frac{\epsilon}{\omega_y}\ln M \;\ll\; 1$, so that in this case the *mesh transition point* $1-\tau_y$ will be very close to the boundary $y = 1$. By assuming that the parameters of the problem, i.e. $\epsilon, \boldsymbol{\omega}, \beta$, f, and g, are chosen so that the solution $u(x,y)$ has one boundary layer at $y = 1$. In particular, this is achieved by assuming that $\boldsymbol{\omega} = [0, \omega_y]^T$, and $\omega_y > 0$. The use of this Shishkin mesh discretization scheme will divide the domain Ω into two overlapping subdomains, $\overline{\Omega} = \Omega_1 \cup \Omega_2$, where

$$\Omega_1 = [0,1] \times [0, 1-\tau_y], \quad \Omega_2 = [0,1] \times [1-\tau_y, 1].$$

This subdivision is shown in the left side of Figure 2.5.

Let $m \equiv \frac{M}{2}$, if we denote by H_x the mesh width in the x- direction and by h_y and H_y the mesh widths inside and outside the boundary layer in the y-direction, i.e.,

$$H_x \equiv \frac{1}{N}, \qquad h_y \equiv \frac{\tau_y}{m}, \qquad \text{and} \qquad H_y \equiv \frac{(1-\tau_y)}{m}, \tag{2.24}$$

then the $(N+1) \times (M+1)$ nodes of the Shishkin mesh are given by

$$\Omega_D = \{(x_i, y_j) \in \overline{\Omega} : i = 0, \ldots, N, j = 0, \ldots, M\},$$

where

$$x_i \equiv iH_x, \text{ for } i = 0, \ldots, N, \qquad \text{and} \qquad y_j \equiv \begin{cases} jH_y & \text{for } j = 0, \ldots, m, \\ 1-(N-j)h_y & \text{for } j = m+1, \ldots, M. \end{cases} \tag{2.25}$$

The mesh is constructed by drawing lines parallel to the coordinate axes through these mesh points; see the right side of Figure 2.5. Here $x_0 = 0$, $y_0 = 0$ and $x_N = 1$, $y_M = 1$ so that the mesh consists of $N-1$ interior nodes in each direction and where the node y_m is exactly at the transition point $1-\tau_y$ in the y- direction. In contrast to the one-dimensional case where the overlapping subdomains intersect in exactly one grid point, in this two-dimensional case we have a whole row of grid points in common. The $N-1$ nodes with vertical coordinate equal to $y_n = 1-\tau_y$ are the grid points in the overlap. It is clear that the mesh widths on Ω_1 satisfy $1/N \leq H_x$, $H_y \leq 2/M$, so the mesh is coarse in this domain. On the other hand, h_y is $\mathcal{O}(\epsilon M^{-1}\log(M))$, so on Ω_2 the mesh is coarse in the x direction and fine in the y direction. The ratio between the different mesh sizes in the y direction is

$$\frac{h_y}{H_y} = \frac{\tau_y}{1-\tau_y} = \tau_y + \mathcal{O}(\tau_y^2) \ll 1.$$

Example 2.1. *We consider the BVP* (2.2) *defined on* $\Omega = (0,1) \times (0,1) \in \mathbb{R}^2$ *with* $\beta = 0$, $f = 0$ *and boundary conditions determined by the function*

$$u(x,y) = (2x-1)\left(\frac{1-e^{(2y-2)/\epsilon}}{1-e^{-2/\epsilon}}\right). \tag{2.26}$$

A three-dimensional rendering of the analytic solution to this problem is shown in Figure 2.6 for the case $\epsilon = 0.01$*. In most parts of the domain* Ω*, the values of the solution* $u(x,y)$ *closely resemble the ones given by the* inflow boundary *condition* $u(x,0) = 2x-1$*, except in a vicinity of the* outflow boundary *where they abruptly change to the constant values* $u(x,1) = 0$*. The boundary conditions at* $x = 0$ *and* $x = 1$ *satisfy* $u(0,y) \approx -1$ *and* $u(1,y) \approx 1$ *respectively (see [25, Example 6.1.1] where a similar problem is presented) and its values also change drastically as they approach the outflow boundary, where they change from* ≈ -1 *or* ≈ 1 *to* 0*. The portion of the domain where these changes occur is proportional to* ϵ *and it is determined by the function* $e^{(2-2y)/\epsilon}$*, thus, when small enough values of* ϵ *are chosen, the changes in the function* u *occur abruptly enough and in a portion of the domain that is small enough so that the solution of* (2.2) *presents an exponential boundary layer in this particular region of the domain.*

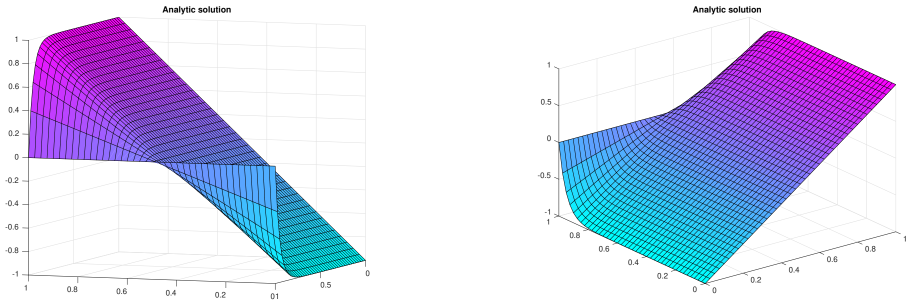


Figure 2.6.: Three-dimensional surface plot of the analytic solution of (2.2) with $n = 2$, $\omega_x, \beta = 0$ and $\epsilon = 10^{-1}$ for two different viewing angles.

This will be the main example used in future chapters to exemplify the theoretical results obtained for two-dimensional problems.

2.1.4.2. Two Outflow Boundary Layers

In the case of two outflow boundaries, we can now use a one dimensional Shishkin mesh in each coordinate direction. Again, given two even positive integers $N \geq 4$ and $M \geq 4$ that denote the number of mesh intervals used in each coordinate direction, let the transition parameters τ_x and τ_y, that will be used to specify where the mesh changes from coarse to fine, be defined by

$$\tau_x \equiv \min\left\{\frac{1}{2}, 2\frac{\epsilon}{\omega_x}\ln N\right\} \quad \text{and} \quad \tau_y \equiv \min\left\{\frac{1}{2}, 2\frac{\epsilon}{\omega_y}\ln M\right\}. \tag{2.27}$$

Once again, since we assume $\omega_x, \omega_y \gg \epsilon$, we will usually have $\tau_x = 2\frac{\epsilon}{\omega_x} \ln N \ll 1$ and $\tau_y = 2\frac{\epsilon}{\omega_y} \ln M \ll 1$, so that in both spacial directions the *mesh transition points* $1-\tau_x$ and $1-\tau_y$ will be located very close to the boundary with $x = 1$ in the x-direction and very close to the boundary with $y = 1$ in the y-direction.

In the case of this discretization scheme the domain Ω is divided into four overlapping subdomains, where $\overline{\Omega} = \Omega_1 \cup \Omega_2 \cup \Omega_3 \cup \Omega_4$. That is, by again letting $n \equiv \frac{N}{2}$ and $m \equiv \frac{M}{2}$, the mesh division divides Ω into a set of 4 subdomains, each of them consisting of $n \times m$ rectangles giving nm interior nodes in each subdomain. This subdivision is shown on the left side of Figure 2.7.

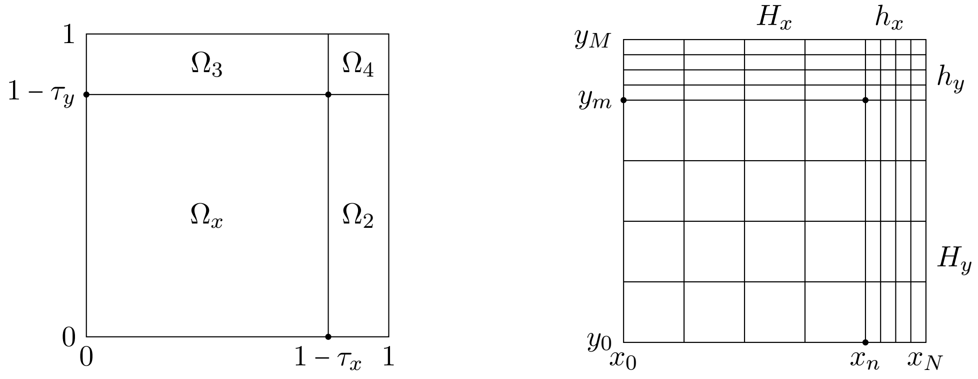


Figure 2.7.: Division of the domain and Shishkin mesh for equation (2.2) with two outflow exponential layers.

If we denote the mesh widths outside and inside the respective boundary layers by H_x, h_x and H_y, h_y, i.e.,

$$H_x \equiv \frac{(1-\tau_x)}{n}, \qquad h_x \equiv \frac{\tau_x}{n}, \qquad H_y \equiv \frac{(1-\tau_y)}{m}, \qquad h_y \equiv \frac{\tau_y}{m},$$

then the four overlapping subdomains are given by

$$\begin{aligned} \Omega_1 &= [0, 1-\tau_x + h_x) \times [0, 1-\tau_y + h_y), & \Omega_2 &= (1-\tau_x - H_x, 1] \times [0, 1-\tau_y + h_x), \\ \Omega_3 &= [0, 1-\tau_x + h_x) \times (1-\tau_y - H_y, 1], & \Omega_4 &= (1-\tau_x - H_x, 1] \times (1-\tau_y - H_y, 1], \end{aligned} \tag{2.28}$$

the $(N+1) \times (M+1)$ nodes of the two-dimensional Shishkin mesh are now given by

$$\Omega_D = \{(x_i, y_j) \in \overline{\Omega} : i = 0, \ldots, N, j = 0, \ldots, M\},$$

where

$$x_i \equiv \begin{cases} iH_x & \text{for } i = 0, \ldots, n \\ 1-(N-i)h_x & \text{for } i = n+1, \ldots, N, \end{cases},$$

and

$$y_j \equiv \begin{cases} jH_y & \text{for } j = 0, \ldots, m \\ 1-(M-j)h_y & \text{for } j = m+1, \ldots, M. \end{cases}$$

The mesh is constructed by drawing lines parallel to the coordinate axes through these mesh points, i.e., the mesh is obtained by a tensor product of two one-dimensional piecewise uniform meshes; see the right side of Figure 2.7. Here $x_0 = 0$, $y_0 = 0$ and $x_N = 1$, $y_M = 1$ so that the mesh consists of $N-1$ and $M-1$ interior nodes in each respective direction, where the node x_n is exactly at the transition point $1-\tau_x$ and the node y_n is exactly at the transition point $1-\tau_y$. It is clear that the mesh widths on Ω_1 satisfy $1/N \le H_x \le 2/N$ and $1/M \le H_y \le 2/M$, so the mesh is coarse in this domain. On the other hand, h_x and h_y are $\mathcal{O}(\epsilon N^{-1}\log(N))$ and $\mathcal{O}(\epsilon M^{-1}\log(M))$ respectively, so the mesh is very fine on Ω_4. On Ω_2 and Ω_3, the mesh is coarse in one direction and fine in the other direction. The ratio between the different mesh sizes in the x-coordinate direction is again

$$\frac{h_x}{H_x} = \frac{\tau_x}{1-\tau_x} = \tau_x + \mathcal{O}(\tau_x^2) \ll 1.$$

and similarly for the ratio $\frac{h_y}{H_y}$.

2.1.5. Approximation of BVPs

In the most general setting, in order to find a finite difference representation of a BVP, both the PDE and the boundary conditions need to be replaced with a suitable approximation. Here, the approximation process will be illustrated through the Dirichlet problem of the convection diffusion equation posed on the square domain $\Omega = (0,1)\times(0,1)$.

We first denote the internal grid points where the equation will be approximated by

$$\Omega_D = \{(x_i, y_j); i = 1,\ldots,N-1;\ j = 1,\ldots,M-1\},$$

and denote the grid points on the boundary by $\partial\Omega_D$ and the entire grid by $\overline{\Omega}_D = \Omega_D \cup \partial\Omega_D$. Since boundary conditions of Dirichlet-type are exact in the boundary nodes, an approximation of the boundary conditions is not needed and we can focus our attention on the approximation of the PDE. We proceed to evaluate the PDE of our two-dimensional BVP (2.2) at the internal grid points of the mesh to obtain

$$-\epsilon\Delta u(x_i,y_j) + \omega(x_i,y_j)\cdot\nabla u(x_i,y_j) + \beta(x_i,y_j)u(x_i,y_j) = f(x_i,y_j). \tag{2.29}$$

Using the standard finite difference operators (see Table 2.19 or e.g., [74, § 4]) to represent the first and second derivative terms and using the notation $\omega_{ij} = \omega(x_i,y_j)$, $\beta_{ij} = \beta(x_i,y_j)$, $f_{ij} = f(x_i,y_j)$, etc., leads to the the upwind scheme approximation

$$-\epsilon\left(h_{x_i}^{-2}\delta_x^2 u_{ij} + h_{y_j}^{-2}\delta_y^2 u_{ij}\right) + \omega_{ij}^x h_{x_i}^{-1}\Delta_x^- u_{ij} + \omega_{ij}^y h_{y_j}^{-1}\Delta_y^- u_{ij} + \beta_{ij}u_{ij} + \mathcal{O}(h_{x_i}) + \mathcal{O}(h_{y_j}) = f_{ij}. \tag{2.30}$$

This equation is satisfied exactly by the solution to the continuous BVP. By neglecting the higher order terms, equation (2.30) will no longer be satisfied by u but by a grid function u^D which we hope will be close to the solution u. The finite difference equations that approximate the PDE are obtained by discarding the remainder terms to yield, for each point $(x_i, y_i) \in \Omega_D$, an algebraic equation of the form

$$-\epsilon\left(h_{x_i}^{-2}\delta_x^2 + h_{y_j}^{-2}\delta_y^2\right)u_{ij} + \omega_{ij}^x h_{x_i}^{-1}\Delta_x^- u_{ij} + \omega_{ij}^y h_{y_j}^{-1}\Delta_y^- u_{ij} + \beta_{ij}u_{ij} = f_{ij}. \tag{2.31}$$

Since the values of the grid function are known on the boundary $\partial\Omega_D$, then equation (2.31) gives $(N-1)\times(M-1)$ linear equations to determine the unknown values of the grid function on Ω_D.

By repeating this process for each point $(x_i, y_i) \in \Omega_D$ we can obtain a finite difference approximation, $\mathcal{A}^D$, of the differential operator $\mathcal{A}$ (defined in (2.9)):

$$\mathcal{A}^D \equiv -\epsilon\left(h_{x_i}^{-2}\delta_x^2 + h_{y_j}^{-2}\delta_y^2\right) + \omega_{ij}^x h_{x_i}^{-1}\Delta_x^- + \omega_{ij}^y h_{y_j}^{-1}\Delta_y^- + \beta_{ij}, \quad {\scriptstyle i=1,\dots,N-1, \atop j=1,\dots,M-1,} \tag{2.32}$$

and we can thus write the (upwind) finite difference discretization of our BVP as:

$$\begin{cases} \mathcal{A}^D u_{ij}^D \equiv -\epsilon\left(h_{x_i}^{-2}\delta_x^2 + h_{y_j}^{-2}\delta_y^2\right)u_{ij}^D + \omega_x h_{x_i}^{-1}\Delta_x^- u_{ij}^D + \omega_y h_{y_j}^{-1}\Delta_y^- u_{ij}^D + \beta_{ij}u_{ij}^D = f_{ij}, & \text{in } \Omega^D, \\ \mathcal{B}^D u_{ij}^D \equiv u_{ij}^D = g_{ij}, & \text{on } \partial\Omega^D, \end{cases} \tag{2.33}$$

The $(N-1)\times(M-1)$ finite difference equations approximating the BVP (2.11) can be written compactly as $\mathscr{A}^D u^D = \mathscr{F}^D$ where

$$\mathscr{A}^D u^D = \begin{cases} \mathcal{A}^D u^D & \text{in } \Omega_D, \\ \mathcal{B}^D u^D & \text{on } \partial\Omega_D, \end{cases} \quad \text{and} \quad \mathscr{F}^D = \begin{cases} f & \text{in } \Omega_D, \\ g & \text{on } \partial\Omega_D. \end{cases} \tag{2.34}$$

2.1.6. The Shishkin Mesh Discretization of Convection-Diffusion BVPs

We can apply the concepts of the previous section to problems where the domain is n-dimensional. In this work we present the results for $n = 1, 2$, and we direct the reader to [41, § 6.2] for a detailed description of the case where $n = 1$.

2.1.6.1. 1D Problems

We first consider the one-dimensional convection-diffusion boundary value problem with constant coefficients and Dirichlet boundary conditions given by (2.1) and proceed to apply the finite difference procedure described in Section 2.1.2 on the one-dimensional Shishkin mesh presented in Section 2.1.3 and shown in Figure 2.3.

Applying the discrete operators given in Table 2.19 on the nodes of the Shishkin mesh, where $h_i = H_x$ for $i = 1, \dots, n$ and $h_i = h_x$ for $i = n+1, \dots, N-1$, we obtain for the upwind scheme:

$$\Delta_x^- u_i^D = \begin{cases} \frac{1}{H_x}\left(u_i^D - u_{i-1}^D\right) & \text{for } i = 1, \dots, n, \\ \frac{1}{h_x}\left(u_i^D - u_{i-1}^D\right) & \text{for } i = n+1, \dots, N-1, \end{cases} \tag{2.35}$$

and for the central finite difference scheme we obtain

$$\Delta_x^0 u_i^D = \begin{cases} \frac{1}{2H_x}\left(u_{i+1}^D - u_{i-1}^D\right) & \text{for } i = 1,\dots,n-1, \\ \frac{1}{h_x+H_x}\left(u_{i+1}^D - u_{i-1}^D\right) & \text{for } i = n, \\ \frac{1}{2h_x}\left(u_{i+1}^D - u_{i-1}^D\right) & \text{for } i = n+1,\dots,N-1. \end{cases} \tag{2.36}$$

For the second derivatives, we obtain

$$\delta_x^2 u_i^D = \begin{cases} \frac{1}{H_x^2}\left(u_{i-1}^D - 2u_i^D + u_{i+1}^D\right) & \text{for } i = 1,\dots,n-1, \\ \frac{2u_{i-1}^D}{(H_x+h_x)H_x} - \frac{2u_i^D}{H_x h_x} + \frac{2u_{i+1}^D}{(H_x+h_x)h_x} & \text{for } i = n, \\ \frac{1}{h_x^2}\left(u_{i-1}^D - 2u_i^D + u_{i+1}^D\right) & \text{for } i = n+1,\dots,N-1. \end{cases} \tag{2.37}$$

Thus, by including the boundary conditions and letting $\omega(x) = \omega_x > 0$ and $\beta(x) = \beta > 0$ be constant along the domain, the finite difference scheme applied to the continuous problem (3.2) results in the discrete version of our model problem:

$$\begin{cases} -\epsilon\delta^2 u_i^D + \omega_i \Delta^{-/0} u_i^D + \beta_i u_i^D = f_i, \quad i = 1,\dots,N-1, \\ u_0^D = g_0, \;\; u_N^D = g_1, \end{cases} \tag{2.38}$$

By collecting all equations for $i = 1,\dots,N-1$, both finite difference schemes yield a linear algebraic system (3.1) with the tridiagonal and nonsymmetric $(N-1)\times(N-1)$ coefficient matrix given by:

$$\mathbf{A} = \left[\begin{array}{cccc|c|cccc} a_H & b_H & & & & & & & \\ c_H & \ddots & \ddots & & & & & & \\ & \ddots & \ddots & b_H & & & & & \\ & & c_H & a_H & b_H & & & & \\ \hline & & & c & a & b & & & \\ \hline & & & & c_h & a_h & b_h & & \\ & & & & & c_h & \ddots & \ddots & \\ & & & & & & \ddots & \ddots & b_h \\ & & & & & & & c_h & a_h \end{array}\right]. \tag{2.39}$$

For the upwind scheme, the entries of $\mathbf{A}$ are given by

$$\begin{aligned} c_H &= -\frac{\epsilon}{H^2} - \frac{\omega_x}{H}, & a_H &= \frac{2\epsilon}{H^2} + \frac{\omega_x}{H} + \beta, & b_H &= -\frac{\epsilon}{H^2}, \\ c &= -\frac{2\epsilon}{H(H+h)} - \frac{\omega_x}{H}, & a &= \frac{2\epsilon}{hH} + \frac{\omega_x}{H} + \beta, & b &= -\frac{2\epsilon}{h(H+h)}, \\ c_h &= -\frac{\epsilon}{h^2} - \frac{\omega_x}{h}, & a_h &= \frac{2\epsilon}{h^2} + \frac{\omega_x}{h} + \beta, & b_h &= -\frac{\epsilon}{h^2}, \end{aligned} \tag{2.40}$$

and for the central difference scheme by

$$c_H = -\frac{\epsilon}{H^2} - \frac{\omega_x}{2H}, \qquad a_H = \frac{2\epsilon}{H^2} + \beta, \qquad b_H = -\frac{\epsilon}{H^2} + \frac{\omega_x}{2H},$$

$$\begin{aligned}
c &= -\frac{2\epsilon}{H(H+h)} - \frac{\omega_x}{h+H}, & a &= \frac{2\epsilon}{hH} + \beta, & b &= -\frac{2\epsilon}{h(H+h)} + \frac{\omega_x}{h+H}, \\
c_h &= -\frac{\epsilon}{h^2} - \frac{\omega_x}{2h}, & a_h &= \frac{2\epsilon}{h^2} + \beta, & b_h &= -\frac{\epsilon}{h^2} + \frac{\omega_x}{2h}.
\end{aligned} \tag{2.41}$$

If $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f} = [u_1^D, \dots, u_{N-1}^D]^T$ is the exact algebraic solution, and $u(x)$ is the solution of (3.2), then there exist constants $c_1, c_2 > 0$ such that

$$\max_{1 \le i \le N-1} |u(x_i) - u_i^D| \le c_1 \frac{\ln N}{N}$$

for the upwind scheme, and

$$\max_{1 \le i \le N-1} |u(x_i) - u_i^D| \le c_2 \left(\frac{\ln N}{N}\right)^2$$

for the central difference scheme. Thus, the convergence of both schemes is ϵ-uniform, and the central difference scheme is more accurate than the upwind scheme. As pointed out by Stynes [74, p. 470], the convergence proof for the central differences (originally due to Andreyev and Kopteva [1]) is complicated since the scheme does not satisfy a discrete maximum principle. We meet similar complications in our analysis in Section 3.2.3 below.

2.1.6.2. 2D Problems

We now consider the two-dimensional convection-diffusion boundary value problem with constant coefficients and Dirichlet boundary conditions given by (2.2) with $n = 2$ (see also (5.2)) and proceed to apply the finite difference procedure described in Section 2.1.2 on the two-dimensional Shishkin mesh presented in Section 2.1.4 and shown in Figure 2.5.

Using the standard upwind finite difference operators and the lexicographical line ordering of the unknowns of the Shishkin mesh, the finite difference scheme yields a linear algebraic system $\mathbf{A}\mathbf{u} = \mathbf{f}$ of the form (5.3) with the block-tridiagonal matrix $\mathbf{A}$ given by

$$\mathbf{A} = \left[\begin{array}{cccc|c|cccc}
\mathbf{A}_H & \mathbf{B}_H & & & & & & & \\
\mathbf{C}_H & \ddots & \ddots & & & & & & \\
 & \ddots & \ddots & \mathbf{B}_H & & & & & \\
 & & \mathbf{C}_H & \mathbf{A}_H & \mathbf{B}_H & & & & \\
\hline
 & & & \hat{\mathbf{C}} & \hat{\mathbf{A}} & \hat{\mathbf{B}} & & & \\
\hline
 & & & & \mathbf{C}_h & \mathbf{A}_h & \mathbf{B}_h & & \\
 & & & & & \mathbf{C}_h & \ddots & \ddots & \\
 & & & & & & \ddots & \ddots & \mathbf{B}_h \\
 & & & & & & & \mathbf{C}_h & \mathbf{A}_h
\end{array}\right] \in \mathbb{R}^{M^2 \times M^2}, \tag{2.42}$$

where $M \equiv N - 1$. The blocks $\mathbf{C}_H$, $\mathbf{A}_H$, $\mathbf{B}_H$, etc., each of dimension $M \times M$, in (2.42) are given by

$$\begin{aligned} \mathbf{C}_H &= \operatorname{diag}(d_H), \quad \mathbf{A}_H = \operatorname{tridiag}(c_H, a_H, b_H), \quad \mathbf{B}_H = \operatorname{diag}(e_H), \\ \hat{\mathbf{C}} &= \operatorname{diag}(d), \quad \hat{\mathbf{A}} = \operatorname{tridiag}(c, a, b), \quad \hat{\mathbf{B}} = \operatorname{diag}(e), \\ \mathbf{C}_h &= \operatorname{diag}(d_h), \quad \mathbf{A}_h = \operatorname{tridiag}(c_h, a_h, b_h), \quad \mathbf{B}_h = \operatorname{diag}(e_h), \end{aligned} \tag{2.43}$$

where the entries are given by

$$d_H = -\frac{\epsilon}{H_y^2} - \frac{\omega_y}{H_y}, \quad c_H = -\frac{\epsilon}{H_x^2} - \frac{\omega_x}{H_x}, \quad a_H = \frac{2\epsilon}{H_x^2} + \frac{2\varepsilon}{H_y^2} + \frac{\omega_x}{H_x} + \frac{\omega_y}{H_y} + \beta, \quad b_H = -\frac{\epsilon}{H_x^2}, \quad e_H = -\frac{\epsilon}{H_y^2}, \tag{2.44}$$

$$d = -\frac{2\epsilon}{H_y(H_y + h_y)} - \frac{\omega_y}{H_y}, \quad c = -\frac{\epsilon}{H_x^2} - \frac{\omega_x}{H_x}, \quad a = \frac{2\epsilon}{H_x^2} + \frac{2\varepsilon}{H_y h_y} + \frac{\omega_x}{H_x} + \frac{\omega_y}{H_y} + \beta, \quad b = -\frac{\epsilon}{H_x^2}, \quad e = -\frac{2\epsilon}{h_y(H_y + h_y)}, \tag{2.45}$$

$$d_h = -\frac{\epsilon}{h_y^2} - \frac{\omega_y}{h_y}, \quad c_h = -\frac{\epsilon}{H_x^2} - \frac{\omega_x}{H_x}, \quad a_h = \frac{2\epsilon}{H_x^2} + \frac{2\varepsilon}{h_y^2} + \frac{\omega_x}{H_x} + \frac{\omega_y}{h_y} + \beta, \quad b_h = -\frac{\epsilon}{H_x^2}, \quad e_h = -\frac{\epsilon}{h_y^2}. \tag{2.46}$$

That is, our matrix takes the form

$$\mathbf{A} = \left[\begin{array}{c|c|c} \widehat{\mathbf{A}}_H & e_m \otimes \mathbf{B}_H & \\ \hline e_m^T \otimes \hat{\mathbf{C}} & \hat{\mathbf{A}} & e_1^T \otimes \hat{\mathbf{B}} \\ \hline & e_1 \otimes \mathbf{C}_h & \widehat{\mathbf{A}}_h \end{array}\right], \tag{2.47}$$

which has the same structure as (5.3).[1]

2.1.7. Properties of Discrete Convection-Diffusion Operators

The matrix in a linear algebraic system obtained from a Shishkin mesh discretization of a singularly perturbed convection-diffusion equation is nonsymmetric, and often highly nonnormal and ill-conditioned; as it has been shown in [23] and is exemplified in the table below. Standard iterative solvers like the (unpreconditioned) GMRES method converge very slowly when applied to such a system; see Figures 2.8-2.9 in this chapter for examples. On the other hand, the Shishkin mesh discretization naturally leads to a *decomposition of the domain*, which suggests to solve the discretized problem by the multiplicative Schwarz method. This is the approach we explore in this chapter for one-dimensional model problems.

Both schemes lead to highly ill-conditioned matrices $\mathbf{A}$. The main reason is the large difference between the mesh sizes H and h, which implies large differences between the moduli of the nonzero entries of $\mathbf{A}$ corresponding to each subdomain. Thus, $\mathbf{A}$ is poorly scaled. As shown by Roos [66], a simple diagonal scaling reduces the order of the condition number for the matrix from the upwind scheme from $\mathcal{O}(\epsilon^{-1}(N/\ln N)^2)$ to $\mathcal{O}(N^2/\ln N)$. Although not shown by Roos, an analogous diagonal scaling appears to work well also for the central difference scheme. As it has been shown in [23] even with a proper scaling the solution methods exhibit

	upwind	upwind scaled	central	central scaled
$\kappa_2(\mathbf{A})$	4.0500×10^{10}	2.9569×10^{3}	6.2323×10^{10}	2.9514×10^{3}
$\kappa_2(\mathbf{Y})$	1.5143×10^{17}	1.2297×10^{19}	4.1070×10^{3}	1.8682×10^{2}

the poor solution behavior. The first row in the following table shows a numerical illustration for $\epsilon = 10^{-8}$, $\omega_x = 1$, $\beta = 0$ in (2.1), and $N = 198$.

The second row of the table shows the condition numbers of the eigenvector matrices from the decomposition $\mathbf{A} = \mathbf{YDY}^{-1}$ computed by `[Y,D]=eig(A)` in `MATLAB`. We observe that the upwind scheme yields matrices with very ill-conditioned eigenvectors, i.e., highly nonnormal matrices. Apparently, the eigenvector conditioning is not much affected by the diagonal scaling.

2.2. Iterative Solvers

As we have shown in previous sections, the numerical approximation of BVPs leads to linear algebraic systems of the form

$$\mathbf{Au} = \mathbf{f}. \tag{2.48}$$

In order to find a solution to (2.48), an *iterative solver* is a mathematical procedure that uses an initial guess to generate a sequence of *approximate* solutions, where the next approximation is obtained from the previous one. In contrast, *direct solvers* attempt to solve equation (2.48) by applying a finite sequence of operations and, in the absence of rounding errors, deliver an *exact* solution. In the following we only provide a brief description of the iterative type of solution methods, which are divided in two general classes: *stationary iterative methods* and the projection-based *Krylov subspace methods*.

Starting with an initial approximate solution vector, $\mathbf{u}^{(0)}$, stationary iterative methods (fixed-point iteration methods) modify individual or groups of components of the vector at each iteration step until a desired tolerance of approximation is reached. Although these methods are rarely used by themselves to obtain solutions to (2.48), when used as *preconditioners* to Kyrlov subspace methods, they can deliver fast and efficient results.

Krylov subspace methods are a more sophisticated type of iterative methods, which fall under the mathematical framework of *projection methods*, whose general idea is based on what is known in literature as the *Petrov-Galerkin* conditions. If the matrix $\mathbf{A}$ in (2.48) is an $N \times N$ real matrix, projection methods obtain its approximate solutions from a subspace of $\mathbb{R}^N$, say $\mathcal{K}$, commonly known as the *search space*. If $\mathcal{K}$ is of dimension n, then in order to obtain an approximate solution, $\hat{\mathbf{u}}$, to (2.48) from this subspace, n constraints must be imposed. Typically, the constrains consist of n orthogonality conditions on the residual vector, $\mathbf{r} \equiv \mathbf{f} - \mathbf{Au}$, with respect to n linearly independent vectors which define another subspace of $\mathbb{R}^N$ also with

[1] See (5.29) to see that the matrix $\mathbf{A}$ in (2.47) fulfills the conditions of block diagonal dominance.

dimension n, say $\mathcal{L}$, known as the *constraint space*. Thus the Petrov-Galerkin conditions lead to the general projection method:

$$\text{Find } \hat{\mathbf{u}} \in \mathbf{u}^{(0)} + \mathcal{K} \text{ such that } \mathbf{f} - \mathbf{A}\hat{\mathbf{u}} \perp \mathcal{L}. \tag{2.49}$$

The approach given by (2.49) is the most general formulation of a projection method which exploits the knowledge of the initial approximation vector $\mathbf{u}^{(0)}$, however, a detailed explanation of these methods falls outside of the scope of this thesis.

For a thorough description and treatment of projection methods for solving linear systems we point to the Ph.D. thesis [37]. For a very complete survey on the general area of iterative solvers for solving linear systems please see [69] or the very compact but excellently written [39]. For an in-depth analysis of Krylov subspace methods as well as their historical development and state of the art see the monograph [51].

2.2.1. Stationary Iterative Methods

Most stationary iterative methods (fixed-point iteration methods) begin by decomposing the coefficient matrix $\mathbf{A}$ in splittings of the form

$$\mathbf{A} = \mathbf{M} - \mathbf{N}, \tag{2.50}$$

where $\mathbf{M}$ is sometimes called a *splitting operator* (very often playing the role of a preconditioner), and $\mathbf{N}$ is the associated *error matrix*. Using this decomposition, we can equivalently write the linear system (2.48) as

$$\mathbf{u} = \mathbf{M}^{-1}\mathbf{N}\mathbf{u} + \mathbf{M}^{-1}\mathbf{f}. \tag{2.51}$$

Given an initial approximation $\mathbf{u}^{(0)}$ to the solution $\mathbf{u}$, we can then define a stationary iterative method that finds successive approximations to $\mathbf{u}$ by following the iteration

$$\mathbf{u}^{(k+1)} = \mathbf{M}^{-1}\mathbf{N}\mathbf{u}^{(k)} + \mathbf{M}^{-1}\mathbf{f}. \tag{2.52}$$

A variety of iterative methods is obtained by choosing different matrices $\mathbf{M}$ and $\mathbf{N}$ in the iteration (2.52). Consider now the matrix $\mathbf{A}$ written as

$$\mathbf{A} = \mathbf{L} + \mathbf{D} + \mathbf{U},$$

where $\mathbf{L}$ is the lower triangular part of $\mathbf{A}$, $\mathbf{U}$ the upper triangular part and $\mathbf{D}$ its diagonal. A short list of resulting iterative methods is given in the following table.

Richardson method,	$\mathbf{M}^{-1} = \alpha\mathbf{I}$,
Jacobi method,	$\mathbf{M}^{-1} = \mathbf{D}^{-1}$,
Weighted Jacobi method,	$\mathbf{M}^{-1} = \alpha\mathbf{D}^{-1}$,
Forward Gauss-Seidel method,	$\mathbf{M}^{-1} = (\mathbf{D} + \mathbf{L})^{-1}$,
Backward Gauss-Seidel method,	$\mathbf{M}^{-1} = (\mathbf{D} + \mathbf{U})^{-1}$,
Symmetric Gauss-Seidel method,	$\mathbf{M}^{-1} = (\mathbf{D} + \mathbf{U})^{-1}\mathbf{D}(\mathbf{D} + \mathbf{L})^{-1}$,
Successive Over-Relaxation (SOR),	$\mathbf{M}^{-1} = \alpha(\mathbf{D} + \alpha\mathbf{L})^{-1}$,
Symmetric SOR,	$\mathbf{M}^{-1} = \alpha(2 - \alpha)(\mathbf{D} + \alpha\mathbf{U})^{-1}\mathbf{D}(\mathbf{D} + \alpha\mathbf{L})^{-1}$,

Block variants of these methods are also possible, collecting groups of unknowns and modifying them collectively at each iteration step (instead of individually treating each entry); see for example [69] for a description and analysis in this direction. Furthermore, other types of stationary iterative methods can also be defined in the context of domain decomposition methods and will be treated in Section 2.2.3). For an in-depth treatment of stationary iterative methods, their analysis and implementation see [79]. Let us first turn to the issue of convergence of stationary iterative methods of the form (2.52).

In general we say that a sequence of vectors $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots$ in $\mathbb{C}^N$ converges to the vector $\mathbf{x} \in \mathbb{C}^N$ if every component $x_i^{(m)}$ satisfies

$$\lim_{k \to \infty} x_j^{(k)} = x_j, \text{ for all } 0 \leq j \leq N.$$

We say that an iteration of the form (2.52) is convergent if the iteration converges to a fixed point as $k \to \infty$, i.e., if for each component $u_j^{(k)}$ of the vector $\mathbf{u}^{(k)}$, the limit $\lim_{k\to\infty} u_j^{(k)}$ exists and its equal to u_j. If these conditions are fulfilled, we say that the iteration converges and thus, at the limit $k \to \infty$, the iterate from equation (2.52) satisfies (2.51), and therefore solves the linear system (2.48).

We can measure how close our approximate solution is to the exact solution at step $k+1$ by defining the error associated to the approximation,

$$\mathbf{e}^{(k+1)} \equiv \mathbf{u} - \mathbf{u}^{(k+1)}, \tag{2.53}$$

and noticing that by subtracting (2.52) from (2.51) we obtain the relation

$$\mathbf{e}^{(k+1)} = \mathbf{M}^{-1}\mathbf{N}\mathbf{e}^{(k)}, \tag{2.54}$$

commonly referred to as the *error equation* and where $\mathbf{M}^{-1}\mathbf{N}$ is known as the *iteration matrix*. We can thus provide conditions on the convergence of iterations of type (2.52) by studying the convergence of the error equation (2.54) instead; it is clear that for any component, $e_j^{(k)}$, of the error vector $\mathbf{e}^{(k)}$ at step k, the limit $\lim_{k\to\infty} e_j^{(k)}$ exists if and only if $\lim_{k\to\infty} u_j^{(k)}$ exists and if both of these limits exist, then $\lim_{k\to\infty} u_j^{(k)} = u_j$ if and only if $\lim_{k\to\infty} e_j^{(k)} = 0$. In other words, in order for the iteration (2.52) to be convergent, we seek that every component of the error vector vanishes as $k \to \infty$.

A closer look at (2.54) shows that we can use induction to obtain the relation

$$\mathbf{e}^{(k)} = \left(\mathbf{M}^{-1}\mathbf{N}\right)^k \mathbf{e}^{(0)}, \tag{2.55}$$

which shows that the error at step k will be related to the powers of the iteration matrix. Thus, in order to have convergence we seek conditions such that

$$\lim_{k\to\infty} \left(\mathbf{M}^{-1}\mathbf{N}\right)^k \mathbf{e}^{(0)} = 0, \tag{2.56}$$

is fulfilled for any starting vector $\mathbf{e}^{(0)}$. Since (2.56) should be fulfilled for any initial starting vector $\mathbf{e}^{(0)}$, an equivalent condition is reached by finding conditions such that

$$\lim_{k\to\infty} \left(\mathbf{M}^{-1}\mathbf{N}\right)^{k} = 0. \tag{2.57}$$

That is, the error of the iteration vanishes if and only if the iteration matrix $\mathbf{M}^{-1}\mathbf{N}$ is convergent[2]. In a similar way to the case of vectors, we define a similar concept of convergence for the case of matrix sequences as follows. We say that an infinite sequence of complex matrices $\mathbf{A}^{(0)}, \mathbf{A}^{(1)}, \ldots$, each in $\mathbb{C}^{N\times N}$, converges to the matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ if all components, $a_{i,j}^{(m)}$, of the matrix iterate, $\mathbf{A}^{(m)}$, converge as we take the limit $m \to \infty$, i.e.,

$$\lim_{m\to\infty} a_{i,j}^{(m)} = a_{i,j}, \text{ for all } 0 \le i, j \le N.$$

The previous discussion shows that in order to have convergence of the iteration (2.52), we seek conditions on the iteration matrix to be convergent. Furthermore, in order to compare or decide whether an iterative method is better than another, we need to compare their iteration matrices with some precise measure. The most common measures are the *spectral radius* and the *spectral norm* of a matrix which arise naturally from its eigenvlaues λ_i and by generalizing the concept of a vector norm; see for example [44] for a complete treatment of vector norms, matrix norms, spectral radii and for their connection to iterative methods see e.g. [72]. The spectral norm of an $N \times N$ complex matrix $\mathbf{A}$ is defined by

$$\|\mathbf{A}\| \equiv \sup_{\mathbf{x}\neq 0} \frac{\|\mathbf{A}\mathbf{x}\|}{\|\mathbf{x}\|},$$

while the *spectral radius*, $\rho(\cdot)$, of an $N \times N$ complex matrix $\mathbf{A}$ is defined by:

$$\rho(\mathbf{A}) \equiv \max_{1\le i\le N} |\lambda_i|.$$

On the one hand the spectral radius can be interpreted geometrically as the smallest circle in the complex plane with center at the origin and which includes all the eigenvalues of the matrix $\mathbf{A}$. The spectral norm, on the other hand, can be intuitively understood as the maximum 'scale' by which a matrix can 'stretch' a vector. Furthermore, we can relate both quantities; using the sumbultiplicativity property of the spectral norm it is easy to see that for a general $N \times N$ complex matrix $\mathbf{A}$ and any consistent pair of vector and matrix norms, we have that the condition

$$\|\mathbf{A}\| \ge \rho(\mathbf{A}),$$

is always fulfilled, and equality is achieved when $\mathbf{A}$ is Hermitian.

[2] For an $N \times N$ matrix $\mathbf{A}$, we say that $\mathbf{A}$ is convergent (to zero) if the sequences $\mathbf{A}, \mathbf{A}^2, \mathbf{A}^3, \ldots,$ converges to the zero matrix 0 and is divergent otherwise.

In summary, the convergence of the vector sequence $\mathbf{u}^{(k)}$, given by the iterative scheme (2.52), to the vector $\mathbf{u}$, solution to (2.48) is ensured if and only if we have

$$\|\mathbf{e}^{(k)}\| = \|\mathbf{u} - \mathbf{u}^{(k)}\| \to 0 \text{ as } k \to \infty.$$

In turn, this is the case if and only if the norm of each powers of the iteration matrix tend to zero as the iteration progresses. More precisely, we need:

$$\|(\mathbf{M}^{-1}\mathbf{N})^k\| \to 0 \text{ as } k \to \infty,$$

i.e., we will only achieve convergence to the solution of the linear system if the iteration matrix is convergent. In the proof of [79, Theorem 1.10] the author uses the Jordan decomposition of a general matrix $\mathbf{A} \in \mathbb{C}^{N\times N}$ to show the following result, which gives the exact conditions for this to happen:

$$\mathbf{A} \in \mathbb{C}^{N\times N} \text{ is convergent if and only if } \rho(\mathbf{A}) < 1. \tag{2.58}$$

Thus, an iteration of the form (2.52) is convergent if and only if the iteration matrix satisfies

$$\rho(\mathbf{M}^{-1}\mathbf{N}) < 1. \tag{2.59}$$

The previous condition is necessary since, in order to show that the mapping $\mathbf{u}^{(k)} \mapsto \mathbf{u}^{(k+1)}$ is indeed a contraction that leads to the fixed-point, the powers of the iteration matrix need to approach zero as the iteration progresses. In turn, the only way for this to happen - evident by considering the Jordan decomposition of $\mathbf{M}^{-1}\mathbf{N}$, is when all eigenvalues of the matrix are less than 1. The condition is also sufficient for convergence since the powers of a matrix with all of its eigenvalues less than 1 will tend to zero. It is also important to note that given the relation

$$\mathbf{M}^{-1}\mathbf{N} = \mathbf{M}^{-1}(\mathbf{M} - \mathbf{A}) = \mathbf{I} - \mathbf{M}^{-1}\mathbf{A},$$

the convergence condition is equivalent to

$$\rho(\mathbf{I} - \mathbf{M}^{-1}\mathbf{A}) < 1,$$

i.e., fast convergence can be expected if the “preconditioned matrix”, $\mathbf{M}^{-1}\mathbf{A}$, is “close” to the identity, or equivalently, if the preconditioner $\mathbf{M}^{-1}$ is a good approximation to the inverse matrix $\mathbf{A}^{-1}$.

In the main chapters of this work we will focus on finding expressions that bound the spectral radius as well as the norm of the powers of the iteration matrix of the multiplicative Schwarz method, a fixed point iteration method that will be described in Section 2.2.3.

2.2.2. Krylov Subspace Methods and GMRES

Krylov subspace methods are projection methods where the search space, $\mathcal{K}$ in (2.49) is a Krylov subspace, i.e., a subspace of the form:

$$\mathcal{K}_n(\mathbf{A}, \mathbf{v}) \equiv \text{span}\{\mathbf{v}, \mathbf{A}\mathbf{v}, \mathbf{A}^2\mathbf{v}, \dots, \mathbf{A}^{n-1}\mathbf{v}\}. \tag{2.60}$$

The plethora of solution methods that fall under the classification of Krylov subspace methods arise by choosing different subspaces for the constraint space $\mathcal{L}$. Moreover, given the construction of the search space $\mathcal{K}_n$, we can see that the approximations to the solution of (2.48) will be of the form

$$\mathbf{u}^{(n)} = \mathbf{u}^{(0)} + p_n(\mathbf{A})\mathbf{v}, \tag{2.61}$$

where $p_n(x)$ is a polynomial of degree at most n. Thus, all approximate solutions obtained with Krylov subspace methods will be of polynomial type and, as mentioned above, the choice of constraint space will have dramatic changes in how this polynomial is constructed by each method.

The generalized minimal residual method, introduced by Saad in [68] and known as the GMRES method in short, is defined by choosing $\mathcal{K} = \mathcal{K}_n$ with $\mathbf{v} = \mathbf{r}^{(0)}$ and $\mathcal{L} = \mathbf{A}\mathcal{K}_n$. This choice of search and constraint spaces minimizes the residual norm over all vectors in $\mathcal{K}_n$ [69], i.e., after computing the initial residual $\mathbf{r}^{(0)} = \mathbf{f} - \mathbf{A}\mathbf{u}^{(0)}$, using the initial guess $\mathbf{u}^{(0)}$, the GMRES method computes a sequence of iterates $\mathbf{u}^{(1)}, \dots, \mathbf{u}^{(n)}$ such that the n-th residual satisfies

$$\|\mathbf{r}^{(n)}\| = \|p_n(\mathbf{A})\mathbf{r}^{(0)}\| = \min_{p \in \pi_n} \|p(\mathbf{A})\mathbf{r}^{(0)}\|, \tag{2.62}$$

where π_n is the set of polynomials of degree at most n which are normalized at 0. For a very detailed description of this method see [51].

Even though this is the method of choice for solving large and sparse nonsymmetric linear systems, like the ones arising from the discretization of convection-diffusion problems studied in this thesis, the method presents a poor convergence behavior when used to solve such linear systems. In particular the method can exhibit an initial period of slow convergence followed by a faster decrease of the residual norm, as noted for example by Ernst in [26].

2.2.2.1. The Stagnation of GMRES for convection-diffusion problems

In the following, Figures 2.8–2.9 illustrate that linear algebraic systems resulting from discretizations of convection-dominated convection-diffusion problems represent a challenge for GMRES holds for the Shishkin mesh discretization of the model problem (2.1). These figures show the relative residual norms of the (unpreconditioned) GMRES method with zero initial vector applied to $\mathbf{A}\mathbf{u} = \mathbf{f}$ from the Shishkin mesh discretization of (2.1) with $\omega_x = 1$, $\beta = 0$, $f(x) \equiv 1$, $u_0 = u_1 = 0$, $N = 198$, and different values of ϵ. The GMRES convergence is virtually the same for both discretizations (upwind and central differences).

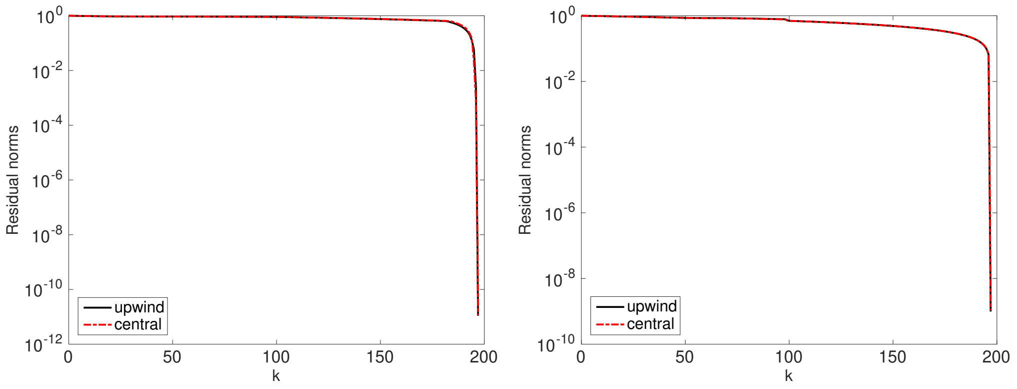


Figure 2.8.: GMRES convergence for $\epsilon = 10^{-2}$ and $\epsilon = 10^{-4}$ [r].

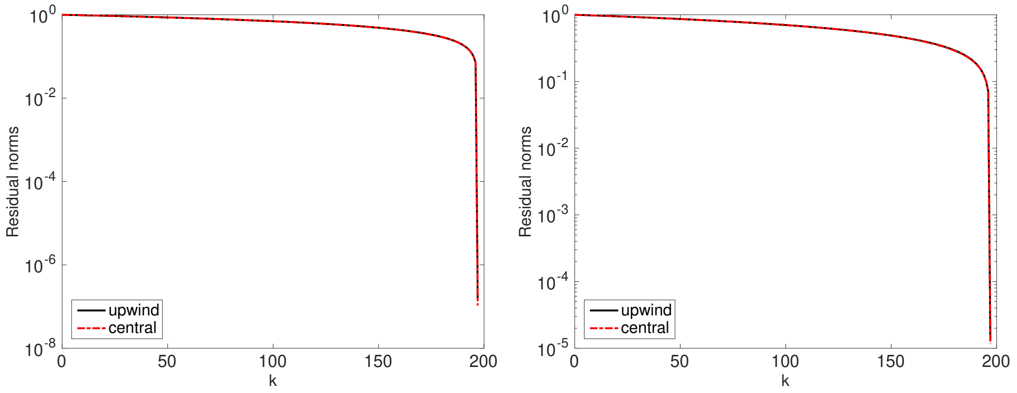


Figure 2.9.: Preconditioned GMRES convergence for $\epsilon = 10^{-6}$ and $\epsilon = 10^{-8}$ [r].

For the problems studied in this thesis, the eigenvector basis of the coefficient matrix $\mathbf{A}$ is very badly conditioned, making the matrix highly *nonnormal*. In such cases, the use of eigenvalues and eigenvectors in an analysis of convergence is not informative and very poorly descriptive, see for example [40] for an extreme case where this is true and [26] for an example of a convection-diffusion problem for which the eigenvalues alone give misleading information about convergence. Such analysis use the eigendecomposition of the coefficient matrix, $\mathbf{A} = \mathbf{YDY}^{-1}$, where $\mathbf{D}$ is a diagonal matrix whose elements contain the eigenvalues λ_k of $\mathbf{A}$, and is given by

$$\|\mathbf{r}^{(n)}\| = \|\mathbf{Y}p_n(\mathbf{D})\mathbf{Y}^{-1}\mathbf{r}^{(0)}\| = \min_{p\in\pi_n} \|\mathbf{Y}p(\mathbf{D})\mathbf{Y}^{-1}\mathbf{r}^{(0)}\| \tag{2.63}$$

$$\leq \|\mathbf{Y}\|\|\mathbf{Y}^{-1}\|\|\mathbf{r}^{(0)}\| \min_{p\in\pi_n} \max_{k} |p(\lambda_k)|. \tag{2.64}$$

This result is a worst case bound that does not take into account the fact that for some initial residuals GMRES may behave very differently than for others. It simplifies the analysis by separating the study of GMRES convergence behavior into optimizing the condition number of the eigenvector matrix $\mathbf{Y}$ and a polynomial minimization problem over the spectrum of $\mathbf{A}$, but it could potentially overestimate GMRES residuals. This is partly because, as observed by Liesen and Strakos in [50], possible cancellations of huge components in $\mathbf{Y}$ and/or $\mathbf{Y}^{-1}$ are artificially ignored for the sake of the convergence analysis.

The stagantion of GMRES for solving convection-diffusion problems of type (2.1) and (2.2) shows the need to either find a preconditioning strategy to accelerate the convergence of the method or the use of a completely different solution approach to solve these types of problems. In the following we will discuss another type of solution methods which fall under the category of domain decomposition methods, which seem to be a natural solution for the types of problems studied in this thesis.

2.2.3. Domain Decomposition and Schwarz Methods

The earliest know domain decomposition method is believed to have been discovered in 1869 by Hermann Amandus Schwarz in [70]. Schwarz devised the method for elliptic equations, to establish the existence of harmonic functions on regions with nonsmooth boundaries. Throughout the 20th century, the area of domain decomposition methods has grown extensively and a variety of methods have been introduced, both at the continuous level as well at the algebraic level; see [5] for the theory of algebraic methods and [71] for a complete monograph of domain decomposition methods. Our focus in this work will fall in the algebraic case, commonly known as the multiplicative Schwarz method, however we will describe Schwarz' original idea in the following (see [36] for a historical introduction to the method).

2.2.3.1. The Continuous Case: Schwarz's Alternating Procedure

When the alternating Schwarz method is considered at the continuous level, there are two main variants to be considered. The first one being the original method invented by Schwarz in [70] at the end of the 19th century as a mathematical tool for investigating the uniqueness of the solution to the Laplace equation when it is posed on a general complex domain, and the second one being the parallel Schwarz method used in the general area of parallel computing, which was introduced by Lions in the 1980's. We will briefly describe the first one and direct the reader to the complete review paper by Gander in [34] which gives a thorough analysis of Schwarz methods over the course of time.

Even though the scientific agenda at the time sought to find a proof of the uniqueness and existence of solutions for Laplace's equation posed on a general complex domain, Schwarz focused on first finding the solution on a domain that was composed of two simpler domains for which existence and uniqueness had already been proven. Figure 2.10 shows the original domain used by Schwarz for his analysis, which consisted of a disk Ω_1 (labeled T_1 in the figure) "stitched" together with a rectangle Ω_2 (labeled T_2 in the figure). The project was to show that the following

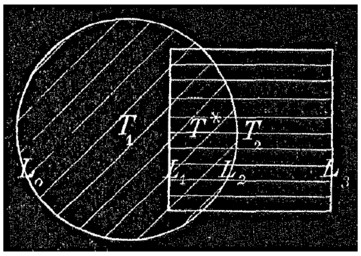


Figure 2.10.: Original domain used by Schwarz consisting of a disk and a rectangle.

BVP

$$\mathcal{A}u \equiv \Delta u = 0, \text{ in } \Omega, \quad u = g \text{ on } \partial\Omega, \tag{2.65}$$

holds for an arbitrary choice of Dirichlet boundary conditions. Since the solutions on each subdomain were known using Fourier series, Schwarz proposed an iterative solution scheme for the entire domain which only made use of the known solutions on the disk and the rectangle. Providing an initial guess u_0^2 along $\Gamma_1 \equiv \partial\Omega_1 \cap \Omega_2$, the iteration computes the approximations u_1^{n+1} and u_2^{n+1} on each subdomain as follows:

$$\begin{aligned} \Delta u_1^{n+1} &= 0 \text{ in } \Omega_1, & \Delta u_2^{n+1} &= 0 \text{ in } \Omega_2, \\ u_1^{n+1} &= u_2^n \text{ on } \Gamma_1, & u_2^{n+1} &= u_1^{n+1} \text{ on } \Gamma_2, \end{aligned} \tag{2.66}$$

where $\Gamma_2 \equiv \partial\Omega_2 \cap \Omega_1$ and both u_1^{n+1} and u_2^{n+1} satisfy the given Dirichlet conditions on the outer boundaries of each subdomain. The convergence of the previous algorithm to the solution of the problem was proven by Schwarz using a maximum principle and very loosely consisted in introducing artificial boundaries for each subdomain (given by the overlap) and using them as Dirichlet conditions for the second subdomain in order to solve each problem in an alternating fashion. For the specific proof, we direct the reader to the original paper by Schwarz in [70] or the survey paper [34]. We will next present the algebraic case, which is the main method of analysis in this thesis.

2.2.3.2. The Discrete (Algebraic) Case: The Multiplicative Schwarz Method

Schwarz methods have also been introduced directly at the algebraic level for solving the linear system $\mathbf{Au} = \mathbf{f}$, and there are several variants. We will now describe one of them: the multiplicative Schwarz method (see [5]). In short, the method uses restriction operators for constructing a multiplicative iteration matrix in which each factor corresponds to a local solve in one of the subdomains.

Just like the continuous domain is partitioned into subdomains, the unknowns in the vector $\mathbf{u}$ need to be subdivided into corresponding subsets, possibly overlapping each other. Without loss of generality, and referring to Figure 2.11, we consider one domain, Ω, subdivided into two contiguous local subdomains, Ω_1 and Ω_2, by one interface boundary Γ_{12}. The only assumption we make is that in each of the local subdomains we have the same number of unknowns. This assumption is made for

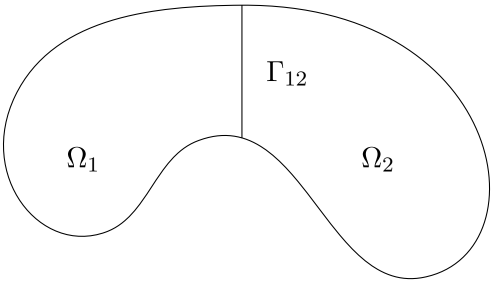


Figure 2.11.: Decomposition of a domain Ω into two local subdomains Ω_1 and Ω_2 with interface boundary Γ_{12}.

simplicity of the following exposition. Extensions to other block sizes and several subdomains are certainly possible, but would require even more technicalities. In this context, the restriction operators can be written as

$$\mathbf{R}_1 \equiv [\mathbf{I}_n \quad 0] \quad \text{and} \quad \mathbf{R}_2 \equiv [0 \quad \mathbf{I}_n],$$

both of size $n \times (N-1)$. Therefore the operation $\mathbf{R}_1\mathbf{u}$ yields the unknowns in the first subset while $\mathbf{R}_2\mathbf{u}$ delivers the unknowns of the second subset (subdomain). The corresponding unknowns in the matrix $\mathbf{A}$ can be obtained using the same restriction operators and thus, the restrictions of the matrix $\mathbf{A}$ in $\mathbf{Au} = \mathbf{f}$ to the two subdomains are given by the two $n \times n$ matrices

$$\widehat{\mathbf{A}}_1 \equiv \mathbf{R}_1\mathbf{A}\mathbf{R}_1^T \equiv \begin{bmatrix} \mathbf{A}_H & b_H\mathbf{e}_m \\ c\mathbf{e}_m^T & a \end{bmatrix} \quad \text{and} \quad \widehat{\mathbf{A}}_2 \equiv \mathbf{R}_2\mathbf{A}\mathbf{R}_2^T \equiv \begin{bmatrix} a & b\mathbf{e}_1^T \\ c_h\mathbf{e}_1 & \mathbf{A}_h \end{bmatrix}, \tag{2.67}$$

where $m \equiv n-1$, and $\mathbf{e}_1, \mathbf{e}_m \in \mathbb{R}^m$. In the following, the unit basis vectors $\mathbf{e}_j$ are always considered to be of appropriate length, which for simplicity is sometimes not explicitly stated. Note that $\mathbf{A}_H, \mathbf{A}_h \in \mathbb{R}^{m\times m}$ are tridiagonal Toeplitz matrices. The matrices corresponding to the solves on the two domains then are given by

$$\mathbf{P}_i \equiv \mathbf{R}_i^T\mathbf{A}_i^{-1}\mathbf{R}_i\mathbf{A}, \quad i = 1,2. \tag{2.68}$$

It is easy to see that $\mathbf{P}_i^2 = \mathbf{P}_i$, i.e., that these matrices are projections. Note also that since $\mathbf{P}_i$ is not symmetric, we have for the 2-norm, that $\|\mathbf{I} - \mathbf{P}_i\|_2 = \|\mathbf{P}_i\|_2 > 1$; see, e.g., [75]. Using the complimentary projections

$$\mathbf{Q}_i \equiv \mathbf{I} - \mathbf{P}_i \in \mathbb{R}^{(N-1)\times(M-1)}, \quad i = 1,2,$$

we define the multiplicative Schwarz iteration matrices

$$\mathbf{T}_{12} \equiv \mathbf{Q}_2\mathbf{Q}_1 \quad \text{and} \quad \mathbf{T}_{21} \equiv \mathbf{Q}_1\mathbf{Q}_2. \tag{2.69}$$

Thus, $\mathbf{T}_{ij}$ corresponds to first solving on Ω_i, and then on Ω_j. Both iteration matrices will be analyzed below.

Starting with an initial vector $\mathbf{u}^{(0)} \in \mathbb{R}^{N(2m+1)}$, the multiplicative Schwarz method is defined by

$$\mathbf{u}^{(k+1)} = \mathbf{T}_{ij}\mathbf{u}^{(k)} + \mathbf{v}, \quad k = 0, 1, 2, \dots, \tag{2.70}$$

where $\mathbf{T}_{ij} = \mathbf{T}_{12}$ or $\mathbf{T}_{ij} = \mathbf{T}_{21}$, and the vector $\mathbf{v} \in \mathbb{R}^{N-1}$ is defined to make the method consistent. For the iteration matrix $\mathbf{T} = \mathbf{T}_{ij}$ the consistency condition $\mathbf{u} = \mathbf{T}_{ij}\mathbf{u} + \mathbf{v}$ yields

$$\mathbf{v} = (\mathbf{I} - \mathbf{T}_{ij})\mathbf{u} = (\mathbf{P}_i + \mathbf{P}_j - \mathbf{P}_j\mathbf{P}_i)\mathbf{u},$$

which is (easily) computable since

$$\mathbf{P}_i\mathbf{u} = \mathbf{R}_i^T\mathbf{A}_i^{-1}\mathbf{R}_i\mathbf{A}\mathbf{u} = \mathbf{R}_i^T\mathbf{A}_i^{-1}\mathbf{R}_i\mathbf{f}, \quad i = 1, 2.$$

The error of the multiplicative Schwarz iteration (2.70) is given by

$$\mathbf{e}^{(k+1)} = \mathbf{u} - \mathbf{u}^{(k+1)} = (\mathbf{T}_{ij}\mathbf{u} + \mathbf{v}) - (\mathbf{T}_{ij}\mathbf{u}^{(k)} + \mathbf{v}) = \mathbf{T}_{ij}\mathbf{e}^{(k)}, \quad k = 0, 1, 2, \dots, \tag{2.71}$$

and hence $\mathbf{e}^{(k+1)} = \mathbf{T}_{ij}^{k+1}\mathbf{e}^{(0)}$ by induction. For any consistent norm $\|\cdot\|$, we therefore have the error bound

$$\|\mathbf{e}^{(k+1)}\| \le \|\mathbf{T}_{ij}^{k+1}\|\,\|\mathbf{e}^{(0)}\|. \tag{2.72}$$

Our main goal on the following chapters of this work is the derivation of quantitative convergence bounds for the error of the multiplicative Schwarz method, where we consider both $\mathbf{T}_{ij} = \mathbf{T}_{12}$ and $\mathbf{T}_{ij} = \mathbf{T}_{21}$.

3. Convergence of the Multiplicative Schwarz Method for Shihskin Mesh Discretizations of One-dimensional Convection-Diffusion Problems

Parts of this chapter have already been published in:

[23] C. Echeverría, J. Liesen, D. B. Szyld, and P. Tichý, **Convergence of the multiplicative Schwarz method for singularly perturbed convection-diffusion problems discretized on a Shishkin mesh**, Electron. Trans. Numer. Anal., 48 (2018), pp. 40–62.

3.1. Introduction

In this chapter we study the convergence behavior of the multiplicative Schwarz method when it is used to solve linear systems of the form

$$\mathbf{A}\mathbf{u} = \mathbf{f}, \tag{3.1}$$

where the coefficient matrix is obtained from finite difference discretizations of the following one-dimensional constant coefficient convection-diffusion BVP with Dirichlet boundary conditions posed on a Shishkin mesh:

$$\begin{cases} -\epsilon \dfrac{\partial^2 u(x)}{\partial x^2} + \omega_x \dfrac{\partial u(x)}{\partial x} + \beta u(x) = f(x), & \text{in } (0,1) \\ \quad u(0) = g_0, \quad \text{and} \quad u(1) = g_1, \end{cases} \tag{3.2}$$

We assume $\omega_x \gg \epsilon > 0$ and $\beta \geq 0$ and that the parameters of the problem, i.e., $\epsilon, \omega_x, \beta, f, g_0$, and g_1, are chosen so that the solution $u(x)$ has one boundary layer close to the point $x = 1$. We resolve the boundary layer using a finite difference discretization on a one-dimensional Shishkin mesh with transition point close to $x = 1$ and consider two different schemes on the mesh: upwind and central differences.

After the discretization process, the structure of the coefficient matrix, $\mathbf{A}$ in (3.1), takes the form:

$$\mathbf{A} = \left[\begin{array}{c|c|c} \mathbf{A}_H & b_H & \\ \hline c & a & b \\ \hline & c_h & \mathbf{A}_h \end{array}\right], \tag{3.3}$$

where the matrices $\mathbf{A}_H$ and $\mathbf{A}_h$ are given by:

$$\mathbf{A}_H = \text{tridiag}(c_H, a_H, b_H), \quad \text{and} \quad \mathbf{A}_h = \text{tridiag}(c_h, a_h, b_h), \tag{3.4}$$

and $c_H, a_H, b_H, c_h, a_h, b_h \in \mathbb{R}$. In turn, the structure of the coefficient matrix (3.3) induces a very special rank-one structure of the iteration matrices $\mathbf{T}_{ij}$, defined in (2.69) and used in the multiplicative Schwarz method. Using the resulting algebraic structure of the iteration matrices, we derive bounds on the infinity norm of the error produced by the method at each iteration step. Unlike asymptotic convergence results based on bounding the spectral radius of the iteration matrix, our results apply to the transient rather than the asymptotic behavior and thus our error bounds are valid from the first step of the multiplicative Schwarz iterations.

For linear systems obtained from the upwind scheme we prove rapid convergence of the multiplicative Schwarz iteration for all relevant parameters in the problem. The analysis of the central difference scheme is more complicated, since some of the submatrices that occur in this case are not only nonsymmetric, but also fail to be M-matrices. This reminds of the analysis in [1], which showed that in this case the difference scheme itself does not satisfy a discrete maximum principle. Nevertheless, we can prove the convergence of the multiplicative Schwarz method for problems discretized by central differences on a Shishkin mesh under the assumption that the number of discretization points in each of the local subdomains is even. If this assumption is not satisfied, then the method may diverge, which we also explain in our analysis.

Furthermore, we study the convergence of the preconditioned GMRES method when the multiplicatvie Schwarz method is used as a preconditioner and show that the low-rank structure of the iteration matrices $\mathbf{T}_{ij}$ is enough to prove convergence of the preconditioned GMRES method independently of the perturbation parameter ϵ.

The chapter is organized as follows. We immediately begin by presenting the convergence analysis of the multiplicative Schwarz method in Section 3.2; first for the upwind scheme and then for the central difference scheme. Background material including the Shishkin mesh discretization of the one-dimensional model problem is specified in previous chapters (see Chapter 2). In Section 3.3 we discuss the performance of GMRES when preconditioned with the multiplicative Schwarz method. Numerical examples are shown in Section 3.4.

3.2. Convergence Bounds for the Multiplicative Schwarz Method

As mentioned in Section 2.1.3, in this simple one-dimensional case, the Shishkin mesh divides the discretized domain into two local subdomains where the solution presents a different characteristic nature. Therefore, a solution approach based on domain decomposition methods seemed only natural. For the upwind scheme, we proved rapid convergence of the multiplicative Schwarz method for all relevant problem parameters. The convergence for the central difference scheme is slower, but still rapid, when $N^2\epsilon < \omega_x$ and if $N/2 - 1$ is even.

3.2.1. Structure of the iteration matrices

We start with a closer look at the structure of the iteration matrices $\mathbf{T}_{ij}$. Note that the matrices $\mathbf{P}_i$ from (2.68) satisfy

$$\mathbf{P}_1 = \mathbf{R}_1^T \mathbf{A}_1^{-1} \mathbf{R}_1 \mathbf{A} = \begin{bmatrix} \mathbf{I}_n \\ 0 \end{bmatrix} \mathbf{A}_1^{-1} \left[\begin{array}{c|c|c} \mathbf{A}_1^{-1} & b\mathbf{e}_n & 0 \end{array}\right] = \left[\begin{array}{c|c|c} \mathbf{I}_n & b\mathbf{A}_1^{-1}\mathbf{e}_n & 0 \\ \hline 0 & 0 & 0 \end{array}\right], \tag{3.5}$$

and

$$\mathbf{P}_2 = \mathbf{R}_2^T \mathbf{A}_2^{-1} \mathbf{R}_2 \mathbf{A} = \begin{bmatrix} 0 \\ \mathbf{I}_n \end{bmatrix} \mathbf{A}_2^{-1} \left[\begin{array}{c|c|c} 0 & c\mathbf{e}_1 & \mathbf{A}_2^{-1} \end{array}\right] = \left[\begin{array}{c|c|c} 0 & 0 & 0 \\ \hline 0 & c\mathbf{A}_2^{-1}\mathbf{e}_1 & \mathbf{I}_n \end{array}\right], \tag{3.6}$$

where $\mathbf{e}_1$, $\mathbf{e}_n \in \mathbb{R}^n$. We now denote

$$\left[\begin{array}{c} \mathbf{p}^{(1)} \\ \hline \pi^{(1)} \end{array}\right] \equiv b\mathbf{A}_1^{-1}\mathbf{e}_n \quad \text{and} \quad \left[\begin{array}{c} \pi^{(2)} \\ \hline \mathbf{p}^{(2)} \end{array}\right] \equiv c\mathbf{A}_2^{-1}\mathbf{e}_1, \tag{3.7}$$

where $\mathbf{p}^{(i)} = [p_1^{(i)}, \dots, p_m^{(i)}]^T \in \mathbb{R}^m$ for $i = 1, 2$, and $\pi^{(1)}$ and $\pi^{(2)}$ are scalars. Then

$$\mathbf{I} - \mathbf{P}_2 = \left[\begin{array}{c|c|c} \mathbf{I}_{m-1} & 0 & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & -\left[\begin{array}{c} \pi^{(2)} \\ \hline \mathbf{p}^{(2)} \end{array}\right] & 0 \end{array}\right], \qquad \mathbf{I} - \mathbf{P}_1 = \left[\begin{array}{c|c|c} 0 & -\left[\begin{array}{c} \mathbf{p}^{(1)} \\ \hline \pi^{(1)} \end{array}\right] & 0 \\ \hline 0 & 1 & 0 \\ \hline 0 & 0 & \mathbf{I}_{m-1} \end{array}\right],$$

which gives

$$\mathbf{T}_{12} = (\mathbf{I} - \mathbf{P}_2)(\mathbf{I} - \mathbf{P}_1) = \left[\begin{array}{c|c|c} 0 & -\mathbf{p}^{(1)} & 0 \\ \hline 0 & p_m^{(1)}\pi^{(2)} & 0 \\ \hline 0 & p_m^{(1)}\mathbf{p}^{(2)} & 0 \end{array}\right] = \left[\begin{array}{c} -\mathbf{p}^{(1)} \\ \hline p_m^{(1)}\pi^{(2)} \\ \hline p_m^{(1)}\mathbf{p}^{(2)} \end{array}\right] \mathbf{e}_{n+1}^T, \tag{3.8}$$

and

$$\mathbf{T}_{21} = (\mathbf{I} - \mathbf{P}_1)(\mathbf{I} - \mathbf{P}_2) = \left[\begin{array}{c} p_1^{(2)}\mathbf{p}^{(1)} \\ \hline p_1^{(2)}\pi^{(1)} \\ \hline -\mathbf{p}^{(2)} \end{array}\right] \mathbf{e}_{n-1}^T, \tag{3.9}$$

where $\mathbf{e}_{n+1}$, $\mathbf{e}_{n-1} \in \mathbb{R}^{N-1}$. Thus, both iteration matrices have rank one, and we can apply to them the following observation.

Proposition 3.1. *Let $\mathbf{T}$ be a square matrix of rank one, i.e., $\mathbf{T} = \mathbf{u}\mathbf{v}^T$ for some vectors $\mathbf{u}, \mathbf{v}$. Then $\mathbf{T}^2 = \rho\mathbf{T}$, with $\rho = \mathbf{v}^T\mathbf{u}$, and as a consequence $\mathbf{T}^{k+1} = \rho^k\mathbf{T}$, for $k \geq 0$.*

Proof. The proof follows by direct computation. □

Corollary 3.2. *In the notation established above, let* $\mathbf{T} = \mathbf{T}_{12}$ *or* $\mathbf{T} = \mathbf{T}_{21}$. *Then for any* $k \geq 0$ *we have*

$$\mathbf{T}^{k+1} = \rho^k \mathbf{T}, \quad \textit{where} \quad \rho \equiv p_m^{(1)} p_1^{(2)}. \tag{3.10}$$

Proof. Applying Proposition 3.1 to either (3.8) or (3.9) produces the desired result. □

Equation (3.10) shows, in particular, that $\|\mathbf{T}^{k+1}\| = |\rho|^k \|\mathbf{T}\|$ holds for any matrix norm $\|\cdot\|$. In order to obtain a convergence bound for the multiplicative Schwarz method we will bound $|\rho|$ and $\|\mathbf{T}\|_\infty$. The following lemma will be essential in our derivations.

Lemma 3.3. *In the notation established above,*

$$\left[\begin{array}{c} \mathbf{p}^{(1)} \\ \hline \pi^{(1)} \end{array}\right] = \pi^{(1)} \left[\begin{array}{c} -b_H \mathbf{A}_H^{-1} e_m \\ \hline 1 \end{array}\right], \qquad \pi^{(1)} = \frac{b}{a - c b_H \left(\mathbf{A}_H^{-1}\right)_{m,m}},$$

$$\left[\begin{array}{c} \pi^{(2)} \\ \hline \mathbf{p}^{(2)} \end{array}\right] = \pi^{(2)} \left[\begin{array}{c} 1 \\ \hline -c_h \mathbf{A}_h^{-1} e_1 \end{array}\right], \qquad \pi^{(2)} = \frac{c}{a - b c_h \left(\mathbf{A}_h^{-1}\right)_{1,1}}.$$

Proof. From (3.7) we know that $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, $\pi^{(1)}$, and $\pi^{(2)}$ solve the systems

$$\left[\begin{array}{c|c} \mathbf{A}_H & b_H e_m \\ \hline c e_m^T & a \end{array}\right] \left[\begin{array}{c} \mathbf{p}^{(1)} \\ \hline \pi^{(1)} \end{array}\right] = b e_n, \qquad \left[\begin{array}{c|c} a & b e_1^T \\ \hline c_h e_1 & \mathbf{A}_h \end{array}\right] \left[\begin{array}{c} \pi^{(2)} \\ \hline \mathbf{p}^{(2)} \end{array}\right] = c e_1.$$

Hence the expressions for $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)}$, $\pi^{(1)}$, and $\pi^{(2)}$ can be obtained using Schur complements. □

Combining (3.10) and Lemma 3.3 gives

$$\rho = \frac{b\, b_H \left(\mathbf{A}_H^{-1}\right)_{m,m}}{a - c\, b_H \left(\mathbf{A}_H^{-1}\right)_{m,m}} \cdot \frac{c\, c_h \left(\mathbf{A}_h^{-1}\right)_{1,1}}{a - b\, c_h \left(\mathbf{A}_h^{-1}\right)_{1,1}}. \tag{3.11}$$

In order to bound $|\rho|$ we thus need to bound certain entries of inverses of the tridiagonal Toeplitz matrices $\mathbf{A}_H$ and $\mathbf{A}_h$. The following Lemma shows that this is straightforward in the case of an M-matrix.

Recall that a nonsingular matrix $\mathbf{B} = [b_{i,j}]$ is called an *M-matrix* when $b_{i,i} > 0$ for all i, $b_{i,j} \leq 0$ for all $i \neq j$, and $\mathbf{B}^{-1} \geq 0$ (elementwise).

Lemma 3.4. *Let* $\mathbf{B}$ *be an* $\ell \times \ell$ *tridiagonal Toeplitz matrix,*

$$\mathbf{B} = \begin{bmatrix} \hat{a} & \hat{b} & & \\ \hat{c} & \ddots & \ddots & \\ & \ddots & \ddots & \hat{b} \\ & & \hat{c} & \hat{a} \end{bmatrix},$$

with $\hat{a} > 0$ and $\hat{b}, \hat{c} < 0$. Moreover, let $\mathbf{B}$ be diagonally dominant, i.e., $\hat{a} \geq |\hat{b}| + |\hat{c}|$. Then $\mathbf{B}$ is an M-matrix with $\mathbf{B}^{-1} > 0$ (elementwise),

$$\left(\mathbf{B}^{-1}\right)_{\ell,\ell} = \left(\mathbf{B}^{-1}\right)_{1,1} \leq \min\left\{\frac{1}{|\hat{b}|}, \frac{1}{|\hat{c}|}\right\}, \tag{3.12}$$

and the entries of $\mathbf{B}^{-1}$ decay along the columns away from the diagonal. In particular,

$$\begin{aligned} \left(\mathbf{B}^{-1}\right)_{1,1} &> \left(\mathbf{B}^{-1}\right)_{i,1} \quad \textit{for} \quad 1 < i \leq \ell, \\ \left(\mathbf{B}^{-1}\right)_{\ell,\ell} &> \left(\mathbf{B}^{-1}\right)_{i,\ell} \quad \textit{for} \quad 1 \leq i < \ell. \end{aligned}$$

Proof. The matrix $\mathbf{B}$ is an M-matrix since its entries satisfy the sign condition and $\mathbf{B}$ is irreducibly diagonally dominant; see, e.g., [11, Theorem 6.2.3, Condition M35] or [42, Criterion 4.3.10]. The elementwise nonnegativity of the inverse, $\mathbf{B}^{-1} > 0$, follows since the M-matrix $\mathbf{B}$ is irreducible; see, e.g., [11, Theorem 6.2.7] or [42, Theorem 4.3.11].

Since $\mathbf{B}$ is a tridiagonal Toeplitz matrix, its $(1,1)$ and (ℓ,ℓ) minors are equal. Therefore the classical formula $\mathbf{B}^{-1} = (\det(B))^{-1}\mathrm{adj}(\mathbf{B})$ implies that $\left(\mathbf{B}^{-1}\right)_{1,1} = \left(\mathbf{B}^{-1}\right)_{\ell,\ell}$. Moreover, since $\hat{a} \geq |\hat{b}| + |\hat{c}|$, we can apply [62, Lemma 2.1, equation (2.8)] to obtain

$$\left(\mathbf{B}^{-1}\right)_{1,1} \leq \frac{1}{\hat{a} - |\hat{b}|} \leq \frac{1}{|\hat{c}|}, \quad \left(\mathbf{B}^{-1}\right)_{\ell,\ell} \leq \frac{1}{\hat{a} - |\hat{c}|} \leq \frac{1}{|\hat{b}|}.$$

Finally, the bounds on the entries of $\mathbf{B}^{-1}$ are special cases of [62, Theorem 3.11], where it was shown that

$$\left(\mathbf{B}^{-1}\right)_{i,j} \leq \omega^{i-j}\left(\mathbf{B}^{-1}\right)_{j,j} \quad \text{for} \quad i \geq j \quad \text{and} \quad \left(\mathbf{B}^{-1}\right)_{i,j} \leq \tau^{j-i}\left(\mathbf{B}^{-1}\right)_{j,j} \quad \text{for} \quad i \leq j,$$

with some $\tau, \omega \in (0,1)$. (They can be expressed explicitly using the entries of $\mathbf{B}$.) □

As we will see later in Lemma 3.5 and Lemma 3.8, the matrix $\mathbf{A}_h$ is an M-matrix for both the upwind and the central difference scheme. However, while $\mathbf{A}_H$ is an M-matrix for the upwind scheme, it is not an M-matrix in the most common situation for the central difference scheme. We then have to use a different technique for bounding the entry $(\mathbf{A}_H^{-1})_{1,1}$; see Section 3.2.3. In the next two subsections we separately treat the upwind and the central difference schemes.

3.2.2. Bounds for the Upwind Difference Scheme

Using Lemma 3.4, which characterizes the inverse entries of a tridiagonal Toeplitz M-matrix, we can prove the following result for the upwind scheme.

Lemma 3.5. *For the upwind scheme both matrices $\mathbf{A}_H$ and $\mathbf{A}_h$ satisfy the assumptions of Lemma 3.4, and the related quantities from Lemma 3.3 satisfy*

$$|\pi^{(1)}| \leq 1, \quad \|\mathbf{p}^{(1)}\|_\infty = |p_m^{(1)}| \leq \frac{\epsilon}{\epsilon + \omega_x H}, \quad |\pi^{(2)}| \leq 1, \quad \|\mathbf{p}^{(2)}\|_\infty = |p_1^{(2)}| \leq 1.$$

Proof. It is easy to see from (2.40) that both matrices $\mathbf{A}_H$ and $\mathbf{A}_h$ resulting from the upwind scheme satisfy the assumptions of Lemma 3.4. Thus, from equation (3.12) we have

$$|b_H|\left(\mathbf{A}_H^{-1}\right)_{m,m} \le 1 \quad \text{and} \quad |c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1} \le 1.$$

Moreover, $a > 0$ and $b, c < 0$, as well as $a + b + c = \beta \ge 0$, so that

$$|\pi^{(1)}| = \frac{|b|}{a + c|b_H|\left(\mathbf{A}_H^{-1}\right)_{m,m}} \le \frac{|b|}{a+c} \le 1,$$
$$|\pi^{(2)}| = \frac{|c|}{a + b|c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1}} \le \frac{|c|}{a+b} \le 1.$$

Using these inequalities and the fact that the entries of $\mathbf{A}_h$ decay along a column away from the diagonal yields

$$\|\mathbf{p}^{(2)}\|_\infty = |p_1^{(2)}| = |\pi^{(2)}|\,|c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1} \le 1.$$

Using the decay of the entries of $\mathbf{A}_H$ and

$$|b_H|\left(\mathbf{A}_H^{-1}\right)_{m,m} \le \left|\frac{b_H}{c_H}\right|,$$

which follows from (3.12), as well as the definitions of the entries in (2.40), we obtain

$$\|\mathbf{p}^{(1)}\|_\infty = |p_m^{(1)}| = |\pi^{(1)}|\,|b_H|\left(\mathbf{A}_H^{-1}\right)_{m,m} \le \left|\frac{b_H}{c_H}\right| = \frac{\epsilon}{\epsilon + \omega_x H}. \qquad \square$$

We can now state our main result of this subsection.

Theorem 3.6. *For the upwind scheme we have*

$$|\rho| \le \frac{\epsilon}{\epsilon + \omega_x H} \tag{3.13}$$

and

$$\|\mathbf{T}_{12}\|_\infty \le \frac{\epsilon}{\epsilon + \omega_x H}, \qquad \|\mathbf{T}_{21}\|_\infty \le 1.$$

Thus, the error of the multiplicative Schwarz method (2.70) *satisfies*

$$\frac{\|\mathbf{e}^{(k+1)}\|_\infty}{\|\mathbf{e}^{(0)}\|_\infty} \le \begin{cases} \left(\frac{\epsilon}{\epsilon+\omega_x H}\right)^{k+1}, & \text{if } \mathbf{T} = \mathbf{T}_{12}, \\ \left(\frac{\epsilon}{\epsilon+\omega_x H}\right)^{k}, & \text{if } \mathbf{T} = \mathbf{T}_{21}. \end{cases}$$

Proof. For the bound on $|\rho|$ we apply Lemma 3.5 to the expression $\rho = p_m^{(1)} p_1^{(2)}$ from (3.10). From (3.8) and (3.9) we respectively see that

$$\|\mathbf{T}_{12}\|_\infty = \left\| \begin{bmatrix} -\mathbf{p}^{(1)} \\ \hline p_m^{(1)}\pi^{(2)} \\ \hline p_m^{(1)}\mathbf{p}^{(2)} \end{bmatrix} \right\|_\infty \quad \text{and} \quad \|\mathbf{T}_{21}\|_\infty = \left\| \begin{bmatrix} p_1^{(2)}\mathbf{p}^{(1)} \\ \hline p_1^{(2)}\pi^{(1)} \\ \hline -\mathbf{p}^{(2)} \end{bmatrix} \right\|_\infty .$$

Thus, using Lemma 3.5,

$$\|\mathbf{T}_{12}\|_\infty = \max\left\{|p_m^{(1)}|,\ |p_m^{(1)}\pi^{(2)}|,\ |p_m^{(1)}p_1^{(2)}|\right\} \le |p_m^{(1)}| \le \frac{\epsilon}{\epsilon + \omega_x H},$$
$$\|\mathbf{T}_{21}\|_\infty = \max\left\{|p_1^{(2)}p_m^{(1)}|,\ |p_1^{(2)}\pi^{(1)}|,\ |p_1^{(2)}|\right\} \le |p_1^{(2)}| \le 1.$$

Using these bounds and (3.10) in the first inequality in (2.72) yields the convergence bound for the multiplicative Schwarz method. □

Suppose that $\epsilon < \omega_x H$, which is a reasonable assumption in our context. Then

$$|\rho| = \frac{\epsilon}{\epsilon + \omega_x H} = \frac{\epsilon}{\omega_x H} + \mathcal{O}\left(\left(\frac{\epsilon}{\omega_x H}\right)^2\right).$$

This expression shows that the convergence of the multiplicative Schwarz method in case of the upwind scheme and a strong convection-dominance will be very rapid. Numerical examples are shown in Section 3.4.

Note that since $\frac{2}{N} = H + h \le 2H$, we have $\frac{1}{N} \le H$, and hence

$$|\rho| \le \frac{\epsilon}{\epsilon + \omega_x H} \le \frac{\epsilon}{\epsilon + \frac{\omega_x}{N}}. \tag{3.14}$$

Using the expression on the right hand of (3.14) in Theorem 3.6 would give (slightly) weaker convergence bounds for the multiplicative Schwarz method. However, the right hand side of (3.14) represents a more convenient bound on the convergence factor which directly depends on the parameters ϵ, ω_x and N of our problem.

3.2.3. Bounds for the Central Difference Scheme

We will now consider the discretization by the central difference scheme, i.e., the matrix $\mathbf{A}$ with the entries given by (2.41). It turns out that the analysis for this scheme is more complicated than for the upwind scheme since, as mentioned above, the matrix $\mathbf{A}_H$ need not be an M-matrix. Moreover, as we will see below, the multiplicative Schwarz method may not converge when the parameter m is odd.

The following result about the entries a, b, and c of $\mathbf{A}$ will be frequently used below.

Lemma 3.7. *For the central difference scheme we have*

$$a > 0, \quad c, b < 0, \quad -(c+b) = |c| + |b| = a - \beta \le a \quad \textit{and} \quad \left|\frac{b}{a}\right| < 1. \tag{3.15}$$

Proof. The inequalities $a > 0$ and $c < 0$ are obvious from (2.41). From (2.20)–(2.21) we have, since $N \ge 4$,

$$\omega_x h = 2\epsilon\,\frac{2\ln N}{N} < 2\epsilon, \tag{3.16}$$

and therefore

$$b = \frac{\omega_x h - 2\epsilon}{h(H+h)} < 0.$$

Moreover, $-(c+b) = a - \beta \leq a$, which yields

$$\left|\frac{b}{a}\right| = \left|\frac{b}{\beta - (c+b)}\right| < 1. \qquad \square$$

We next consider the matrix $\mathbf{A}_h$ from the central difference scheme.

Lemma 3.8. *The matrix $\mathbf{A}_h$ from the central difference scheme satisfies the assumptions of Lemma 3.4, and for the corresponding quantities from Lemma 3.3 we have*

$$\left|\pi^{(2)}\right| \leq 1 \quad \textit{and} \quad \|\mathbf{p}^{(2)}\|_\infty = \left|p_1^{(2)}\right| \leq 1.$$

Proof. The inequalities $a_h > 0$ and $c_h < 0$ are obvious from (2.41), and using (3.16) we obtain

$$b_h = \frac{\omega_x h - 2\epsilon}{2h^2} < 0.$$

Since also

$$|c_h| + |b_h| = \frac{2\epsilon}{h^2} \leq a_h,$$

the matrix $\mathbf{A}_h$ satisfies the assumptions of Lemma 3.4. Thus, in particular, $|c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1} \leq 1$. Using also (3.15) gives

$$|\pi^{(2)}| = \frac{|c|}{a + b|c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1}} \leq \frac{|c|}{a+b} = \frac{|c|}{|c| + \beta} \leq 1.$$

Finally, since the entries of $\mathbf{A}_h$ decay along a column away from the diagonal, we obtain $\|\mathbf{p}^{(2)}\|_\infty = \left|p_1^{(2)}\right| = |\pi^{(2)}|\,|c_h|\left(\mathbf{A}_h^{-1}\right)_{1,1} \leq 1$. □

We now concentrate on bounding the quantities from Lemma 3.3 related to the matrix $\mathbf{A}_H$ for the central difference scheme. We will distinguish the three cases $\omega_x H < 2\epsilon$, $\omega_x H = 2\epsilon$, and $\omega_x H > 2\epsilon$ or, equivalently, the cases that the entry

$$b_H = \frac{\omega_x H - 2\epsilon}{2H^2}$$

of A_H is negative, zero, or positive. It is clear from (3.11) that the sign of b_H is important for the value $|\rho|$.

A simple computation shows that $b_H \leq 0$ if and only if

$$\epsilon \geq \frac{\omega_x}{N + 2\ln N}\,.$$

If $\epsilon \ll \omega_x \approx 1$, then this condition means that $\epsilon(N + 2\ln N) = \mathcal{O}(1)$, which is an unrealistic assumption on the discretization parameter N. Nevertheless, we include the case $b_H \leq 0$ for completeness.

We first assume that

$$\omega_x H < 2\epsilon, \tag{3.17}$$

which means that $b_H < 0$.

Lemma 3.9. *If* (3.17) *holds, then the matrix* $\mathbf{A}_H$ *from the central difference scheme satisfies the assumptions of Lemma 3.4, and we have*

$$\left|\pi^{(1)}\right| \le 1 \quad \textit{and} \quad \left\|\mathbf{p}^{(1)}\right\|_\infty = \left|p_m^{(1)}\right| < \frac{\epsilon}{\epsilon + \frac{\omega_x}{N}}.$$

Proof. The inequalities $a_H > 0$ and $c_H < 0$ are obvious from (2.41), and $b_H < 0$ holds because of (3.17). Moreover,

$$|c_H| + |b_H| = \frac{\omega_x}{2H} + \frac{\epsilon}{H^2} + \frac{\epsilon}{H^2} - \frac{\omega_x}{2H} = \frac{2\epsilon}{H^2} \le a_H,$$

so that the matrix $\mathbf{A}_H$ satisfies the assumptions of Lemma 3.4. In particular,

$$|b_H| \left(\mathbf{A}_H^{-1}\right)_{m,m} \le 1.$$

Using (3.15) we obtain

$$|\pi^{(1)}| = \frac{|b|}{a + c|b_H| \left(\mathbf{A}_H^{-1}\right)_{m,m}} \le \frac{|b|}{a+c} \le 1.$$

Moreover, using that the entries of $\mathbf{A}_H$ decay along a column away from the diagonal as well as

$$|b_H| \left(\mathbf{A}_H^{-1}\right)_{m,m} \le \frac{|b_H|}{|c_H|},$$

which follows from (3.12), we see that

$$\begin{aligned}\left\|\mathbf{p}^{(1)}\right\|_\infty &= \left|p_m^{(1)}\right| = |\pi^{(1)}|\,|b_H| \left(\mathbf{A}_H^{-1}\right)_{m,m} \le \frac{|c_H|}{|b_H|} = \frac{2\epsilon - \omega_x H}{2\epsilon + \omega_x H} < \frac{2\epsilon - \omega_x H + \omega_x h}{2\epsilon + \omega_x H + \omega_x h} \\ &< \frac{2\epsilon}{2\epsilon + \omega_x(H+h)} = \frac{\epsilon}{\epsilon + \frac{\omega_x}{N}},\end{aligned}$$

where we used $h < H$ and $h + H = \frac{2}{N}$. □

Next we consider the (very) special case

$$\omega_x H = 2\epsilon, \tag{3.18}$$

which means that $b_H = 0$.

Lemma 3.10. *If* (3.18) *holds, then the matrix* $\mathbf{A}_H$ *from the central difference scheme is nonsingular, and we have* $|\pi^{(1)}| < 1$ *and* $\mathbf{p}^{(1)} = 0$.

Proof. If $b_H = 0$, then $\mathbf{A}_H$ is lower triangular and nonsingular since $a_H > 0$. Using the definitions of $\mathbf{p}^{(1)}$ and $\pi^{(1)}$ from Lemma 3.3 and the last inequality in (3.15) we obtain

$$\mathbf{p}^{(1)} = -\frac{b b_H \mathbf{A}_H^{-1} e_m}{a - b_H \left(\mathbf{A}_H^{-1}\right)_{m,m} c} = 0, \qquad |\pi^{(1)}| = \left|\frac{b}{a - c b_H \left(\mathbf{A}_H^{-1}\right)_{m,m}}\right| = \left|\frac{b}{a}\right| < 1. \qquad \square$$

The third case we consider is

$$\omega_x H > 2\epsilon, \tag{3.19}$$

which means that $b_H > 0$. This is the most common situation from a practical point of view, but now $\mathbf{A}_H$ does not satisfy the assumptions of Lemma 3.4. We therefore need a different approach for bounding the quantities from Lemma 3.3, and in particular the entries of the vector $\mathbf{A}_H^{-1}\mathbf{e}_m$. Note that because of (3.19) we have

$$-1 < \frac{2\epsilon - \omega_x H}{2\epsilon + \omega_x H} = \frac{b_H}{c_H} < 0.$$

Lemma 3.11. *If* (3.19) *holds, then the matrix* $\mathbf{A}_H$ *from the central difference scheme is a nonsingular tridiagonal Toeplitz matrix with the entries* $a_H, b_H > 0$ *and* $c_H < 0$. *Moreover,*

$$0 < \left|(\mathbf{A}_H^{-1})_{i,m}\right| \leq (\mathbf{A}_H^{-1})_{m,m} \frac{1 - \left(\frac{b_H}{c_H}\right)^i}{1 - \left(\frac{b_H}{c_H}\right)^m} \cdot \left|\frac{b_H}{c_H}\right|^{m-i}, \quad i = 1, \ldots, m, \tag{3.20}$$

where the second inequality in (3.20) is an equality if $\beta = 0$. *If* $m = N/2 - 1$ *is even, then*

$$b_H \left|(\mathbf{A}_H^{-1})_{i,m}\right| < 2, \quad i = 1, \ldots, m, \tag{3.21}$$

and

$$b_H (\mathbf{A}_H^{-1})_{m,m} \leq \frac{1 - \left|\frac{b_H}{c_H}\right|^m}{\left|\frac{c_H}{b_H}\right| + \left|\frac{b_H}{c_H}\right|^m} < \frac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}}. \tag{3.22}$$

Proof. The inequalities $a_H > 0$ and $c_H < 0$ are obvious from (2.41), and $b_H > 0$ holds because of (3.19).

In order to see that $\mathbf{A}_H$ is nonsingular, note that eigenvalues of the tridiagonal Toeplitz matrix $\mathbf{A}_H$ are given by

$$\lambda_i = a_H + 2\sqrt{b_H c_H} \cos\left(\frac{i\pi}{m+1}\right), \quad i = 1, \ldots, m.$$

Since $b_H c_H < 0$, the number $\sqrt{b_H c_H}$ is purely imaginary, and hence all eigenvalues are nonzero.

Adapting [77, Theorem 2] to our notation (and formulating this result in terms of columns instead of rows) shows that the entries of the vector $\boldsymbol{\xi} \equiv [\xi_1, \ldots, \xi_m]^T \equiv \mathbf{A}_H^{-1}\mathbf{e}_m$ can be written as

$$\xi_i = (-1)^{m-i} b_H^{m-i} \frac{\theta_{i-1}}{\theta_m}, \quad i = 1, \ldots, m,$$

where

$$\theta_i \equiv a_H \theta_{i-1} - b_H c_H \theta_{i-2}, \quad \theta_0 \equiv 1, \;\; \theta_1 \equiv a_H. \tag{3.23}$$

Since $b_H c_H < 0$ and $a_H > 0$, we have $\theta_i > 0$ for all $i \geq 0$, and $\xi_i \neq 0$. Since $b_H > 0$, ξ_i changes the sign like $(-1)^{m-i}$, and $\xi_m > 0$. Consequently, the first inequality in (3.20) holds.

If we define the sequence of positive numbers

$$\alpha_i \equiv \frac{\theta_{i-1}}{\theta_i}, \quad i = 1, 2, \ldots,$$

then

$$\xi_i = (-1)^{m-i} b_H^{m-i} \prod_{j=i}^{m} \alpha_j = \xi_m (-1)^{m-i} b_H^{m-i} \prod_{j=i}^{m-1} \alpha_j, \quad i = 1, \ldots, m. \tag{3.24}$$

We will prove by induction that

$$\alpha_i \leq -\frac{c_H^i - b_H^i}{c_H^{i+1} - b_H^{i+1}} \tag{3.25}$$

for all $i \geq 1$, with equality if $\beta = 0$. For $i = 1$ we have

$$-\frac{c_H - b_H}{c_H^2 - b_H^2} = \frac{1}{-(c_H + b_H)} = \frac{1}{a_H - \beta} \geq \frac{1}{a_H} = \alpha_1,$$

with equality if $\beta = 0$. Using the recurrence (3.23), the inequality $a_H \geq -(c_H + b_H)$, which is an equality if $\beta = 0$, and the induction hypothesis, we obtain

$$\frac{1}{\alpha_i} = a_H - \alpha_{i-1} b_H c_H \geq -(c_H + b_H) + \frac{c_H^{i-1} - b_H^{i-1}}{c_H^i - b_H^i} b_H c_H = -\frac{c_H^{i+1} - b_H^{i+1}}{c_H^i - b_H^i},$$

again with equality if $\beta = 0$.

Combining (3.24) and (3.25) yields

$$|\xi_i| \leq \xi_m b_H^{m-i} \left| \frac{c_H^i - b_H^i}{c_H^m - b_H^m} \right| = \xi_m \frac{1 - \left(\frac{b_H}{c_H}\right)^i}{1 - \left(\frac{b_H}{c_H}\right)^m} \cdot \left| \frac{b_H}{c_H} \right|^{m-i}, \tag{3.26}$$

showing the second inequality in (3.20), which is an equality if $\beta = 0$.

Now let m be even. Using (3.25) we obtain

$$b_H \xi_m \leq -b_H \frac{c_H^m - b_H^m}{c_H^{m+1} - b_H^{m+1}} = \frac{1 - \left|\frac{b_H}{c_H}\right|^m}{\left|\frac{c_H}{b_H}\right| + \left|\frac{b_H}{c_H}\right|^m} < 1 - \left|\frac{b_H}{c_H}\right|^m, \tag{3.27}$$

which contains the first inequality in (3.22). Using (3.26) and (3.27) we obtain

$$|\xi_i| < \xi_m \frac{2}{1 - \left|\frac{b_H}{c_H}\right|^m} < \xi_m \frac{2}{b_H \xi_m} = \frac{2}{b_H},$$

which shows (3.21). Let us write

$$\left|\frac{b_H}{c_H}\right| = \frac{\omega_x H - 2\epsilon}{\omega_x H + 2\epsilon} = 1 - \frac{2\epsilon}{\epsilon + \frac{\omega_x H}{2}} \equiv 1 - \nu.$$

Using (3.19) we have $0 < \nu < 1$, and by induction it can be easily shown that $1 - (1-\nu)^m < m\nu$ holds for every integer $m \geq 2$. Thus,

$$b_H \xi_m < 1 - \left|\frac{b_H}{c_H}\right|^m = 1 - (1-\nu)^m < m\nu = \frac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}} \ ,$$

which proves the second inequality in (3.22). □

Using Lemma 3.11 and the assumption that m is even, we can bound the quantities from Lemma 3.3 related to the matrix $\mathbf{A}_H$ from the central difference scheme as follows.

Lemma 3.12. *If* (3.19) *holds and if* $m = N/2 - 1$ *is even, then*

$$\left|\pi^{(1)}\right| < 1, \quad \left|p_m^{(1)}\right| < \frac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}}, \quad \left\|\mathbf{p}^{(1)}\right\|_\infty < 2.$$

Proof. From (3.15) we know that $c < 0$, and from Lemma 3.11 we know that $b_H > 0$ and $\left(\mathbf{A}_H^{-1}\right)_{m,m} > 0$. Therefore

$$\left|\pi^{(1)}\right| = \frac{|b|}{a + |c| b_H \left(\mathbf{A}_H^{-1}\right)_{m,m}} < \left|\frac{b}{a}\right| < 1,$$

where we have used (3.15). Thus, using also (3.22), we obtain

$$\left|p_m^{(1)}\right| = |\pi^{(1)}|\, b_H \left(\mathbf{A}_H^{-1}\right)_{m,m} < \frac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}} \ .$$

Finally, (3.21) implies that $\|\mathbf{p}^{(1)}\|_\infty = |\pi^{(1)}| b_H \|\mathbf{A}_H^{-1} e_m\|_\infty < 2$. □

Now we are ready to formulate an analogue of Theorem 3.6 for the central difference scheme.

Theorem 3.13. *For the central difference scheme we have*

$$|\rho| < \begin{cases} \dfrac{\epsilon}{\epsilon + \frac{\omega_x}{N}} & \text{if } \omega_x H \leq 2\epsilon, \\[2ex] \dfrac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}} & \text{if } \omega_x H > 2\epsilon \text{ and } m = N/2 - 1 \text{ is even.} \end{cases} \tag{3.28}$$

If $\omega_x H \leq 2\epsilon$, *we have*

$$\|\mathbf{T}_{12}\|_\infty \leq 1, \quad \|\mathbf{T}_{21}\|_\infty \leq 1,$$

and if $\omega_x H > 2\epsilon$*, we have*

$$\|\mathbf{T}_{12}\|_\infty < 2, \quad \|\mathbf{T}_{21}\|_\infty < 2.$$

Thus, the error of the multiplicative Schwarz method (2.70) *for both iteration matrices satisfies*

$$\frac{\|\mathbf{e}^{(k+1)}\|_\infty}{\|\mathbf{e}^{(0)}\|_\infty} < \begin{cases} \left(\dfrac{\epsilon}{\epsilon + \frac{\omega_x}{N}}\right)^k & \textit{if } \omega_x H \le 2\epsilon, \\ 2\left(\dfrac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}}\right)^k & \textit{if } \omega_x H > 2\epsilon \textit{ and } m = N/2 - 1 \textit{ is even.} \end{cases}$$

Proof. From (3.10) we know that $\rho = p_m^{(1)} p_1^{(2)}$, and hence the bounds on $|\rho|$ follow from $|p_1^{(2)}| \le 1$ (Lemma 3.8), and Lemmas 3.9–3.10 for the case $\omega_x H \le 2\epsilon$, as well as Lemma 3.12 for the case $\omega_x H > 2\epsilon$.

For the first iteration matrix we have

$$\begin{aligned} \|\mathbf{T}_{21}\|_\infty &= \left\| \begin{bmatrix} p_1^{(2)} \mathbf{p}^{(1)} \\ \hline p_1^{(2)} \pi^{(1)} \\ \hline -\mathbf{p}^{(2)} \end{bmatrix} \right\|_\infty \\ &= \max\{ |p_1^{(2)}| \, \|\mathbf{p}^{(1)}\|_\infty, |p_1^{(2)} \pi^{(1)}|, \|\mathbf{p}^{(2)}\|_\infty \}, \end{aligned}$$

and for the second iteration matrix we have

$$\begin{aligned} \|\mathbf{T}_{12}\|_\infty &= \left\| \begin{bmatrix} -\mathbf{p}^{(1)} \\ \hline p_m^{(1)} \pi^{(2)} \\ \hline p_m^{(1)} \mathbf{p}^{(2)} \end{bmatrix} \right\|_\infty \\ &= \max\{ \|\mathbf{p}^{(1)}\|_\infty, |p_m^{(1)} \pi^{(2)}|, |p_m^{(1)}| \, \|\mathbf{p}^{(2)}\|_\infty \}. \end{aligned}$$

The bounds on these matrices now follow from the Lemmas 3.8, 3.9, 3.10, 3.12, and the error bound for the multiplicative Schwarz method follows from (2.72) and (3.10). □

As in the discussion of Theorem 3.6 we could use $\frac{1}{N} \le H$ and $m = \frac{N}{2} - 1$, and thus obtain

$$|\rho| \le \frac{2m\epsilon}{\epsilon + \frac{\omega_x H}{2}} < \frac{N\epsilon}{\epsilon + \frac{\omega_x}{2N}},$$

where the right hand side again represents a bound on the convergence factor that directly depends of the parameters of our problem.

Because of the factor $2m \approx N$, the error bound for the central differences discretization can be significantly larger than for the upwind scheme. Thus, we expect that the multiplicative Schwarz method for the central differences discretization convergences slower than for the upwind scheme, at least when $\omega_x H > 2\epsilon$. An example with $\epsilon = 10^{-4}$ and $N = 198$, leading to $|\rho| = 8.3 \times 10^{-1}$ and a very slow convergence of the multiplicative Schwarz method is shown in Section 3.4. In this case, the bound (3.28) is even greater than one. It should be noted, however, that in a strongly convection-dominated case the situation $\epsilon N^2 = \mathcal{O}(1)$ is rather unrealistic.

Finally, let us discuss the situation when (3.19) holds, so that $-1 < b_H/c_H < 0$, but m is odd. For simplicity, let $\beta = 0$. Then (3.25) yields

$$b_H(A_H^{-1})_{m,m} = b_H\xi_m = -\frac{1 - \left(\frac{b_H}{c_H}\right)^m}{\frac{c_H}{b_H} - \left(\frac{b_H}{c_H}\right)^m} = \frac{1 + \left|\frac{b_H}{c_H}\right|^m}{\left|\frac{c_H}{b_H}\right| - \left|\frac{b_H}{c_H}\right|^m}.$$

The essential inequality in (3.27) then fails to hold, and we may have $b_H(A_H^{-1})_{m,m} > 1$, with significant consequences for the convergence factor $|\rho|$; see (3.11). It is then easy to find parameters for which $|\rho| > 1$, and for which the multiplicative Schwarz method in fact diverges.

Intuitively, the troubles with odd m correspond to the situation when the equation (3.2) is discretized using central differences on a uniform mesh. Consider for example the discrete solution of the problem (3.2) with $\omega_x = 1$, $\beta = 0$, $f(x) \equiv 1$, and $u_0 = u_1 = 0$, which can be found in [74, § 4]. If the number of the interior points of the uniform mesh is even, then the discrete solution oscillates, but with an amplitude bounded by one, so that some important information about the analytic solution is still preserved in the discrete solution. If the number of inner points is odd, the discrete solution is highly oscillating (cf. [74, Figure 4.1]) and does therefore not provide much useful information about the analytic solution. In our case of the Shishkin mesh, the multiplicative Schwarz method solves discrete problems on the coarse mesh and the fine mesh in an alternating way, and combines the solutions of the two subproblems. If m is odd, then the discrete solution on the coarse mesh is essentially useless because of high oscillations, and the multiplicative Schwarz method does not succeed to improve the approximation to the discrete solution.

3.3. Shishkin-Schwarz Preconditioning

As mentioned in the introduction, linear algebraic systems resulting from discretizations of convection-dominated convection-diffusion problems represent a challenge for iterative solvers. In particular the unpreconditioned GMRES method performs very poorly (see Section 2.2.2). In this section we present the results of using the multiplicative Schwarz method as a preconditioner for GMRES.

Based on (2.70), it is clear that the multiplicative Schwarz method can be seen as a *Richardson iteration* for the system

$$(\mathbf{I} - \mathbf{T}_{ij})\mathbf{u} = \mathbf{v}. \tag{3.29}$$

Furthermore, the iteration scheme (2.70) can be written in the form

$$\mathbf{u}^{(k+1)} = \mathbf{u}^{(k)} + (\mathbf{I} - \mathbf{T}_{ij})(\mathbf{u} - \mathbf{u}^{(k)}),$$

so that the multiplicative Schwarz method as well as GMRES applied to (3.29) obtain their approximations from the Krylov subspace $\mathcal{K}_k(\mathbf{I} - \mathbf{T}_{ij}, \mathbf{r}^{(0)})$. Consequently, in terms of the residual norm, GMRES applied to (3.29) will always converge faster than the multiplicative Schwarz method. Moreover, if one applies GMRES to (3.29), then the multiplicative Schwarz method can be seen as a *preconditioner* for the GMRES method; see, e.g., [46] where this approach is taken. The preconditioner $\mathbf{M}$ such that $\mathbf{M}^{-1}\mathbf{A}\mathbf{u} = \mathbf{M}^{-1}\mathbf{f}$ results in (3.29), can formally be written as $\mathbf{M} = \mathbf{A}(\mathbf{I} - \mathbf{T}_{ij})^{-1}$; see, e.g., [49, Lemma 2.3].

In general, if a matrix $\mathbf{T}$ satisfies $r = \operatorname{rank}(\mathbf{T})$, with $\mathbf{I} - \mathbf{T}$ nonsingular, then for any initial residual $\mathbf{r}^{(0)}$ we have

$$\dim\left(\mathcal{K}_k(\mathbf{I} - \mathbf{T}, \mathbf{r}^{(0)})\right) \leq r + 1,$$

so that GMRES applied to the system (3.29) converges to the solution in at most $r+1$ steps (in exact arithmetic). In the one-dimensional model problem studied in this chapter we have a matrix $\mathbf{T}$ with $r = 1$. Thus, GMRES applied to (3.29) converges in (at most) two steps (see Figures 3.6–3.7), even when the multiplicative Schwarz iteration itself converges slowly or diverges, which may happen for the central difference scheme and m odd. As it will be shown in Chapter 5, which presents a generalization of the approach presented in this chapter to two-dimensional problems, the low-rank structure of the iteration matrix assures the convergence of GMRES in a small number of steps. Hence, it is expected that for generalizations to three-dimensional problems more complicated Shishkin meshes with several transition points, one can possibly exploit a low rank structure of the iteration matrix as well.

It is important to point out that, typically, in practical implementations the local subdomain problems given by (2.67) will not be solved exactly, and thus, in the case of inexact local solves the bounds obtained in this work and the exact termination of GMRES in $r + 1$ steps will no longer hold. Nevertheless, the theory for the exact case presented here gives an indication for the behavior in the inexact case. This is a standard approach in the context of preconditioning. An example of this framework is given by the saddle point preconditioners for which GMRES terminates in a few steps; see [10]. In the context of domain decomposition methods, in particular for Schwarz methods, the concept of inexact subdomain solves was investigated, e.g., in [5, § 4]. See also [35], where a similar situation is described for algebraic optimized Schwarz methods. In Chapter 5, we present examples where the local subdomain problems are solved inexactly for two-dimensional problems. See the discussion in Section 5.3 and the numerical results in Section 5.4.

3.4. Numerical Illustrations

We exemplify the convergence behavior of the multiplicative Schwarz method applied to the Shishkin mesh discretization of the problem (3.2) with

$$\omega_x = 1, \quad \beta = 0, \quad f(x) \equiv 1, \quad u_0 = u_1 = 0.$$

The analytic solution of this problem with $\epsilon = 0.03$ is shown in Figure 2.4. All numerical experiments were computed on a 13-inch `Apple MacBook` computer model

Mid 2010 with a 2,4 GHz Intel Core 2 Duo processor equipped with `MATLAB` version R2015b.

We first consider $N = 198$, so that $m = N/2 - 1 = 98$ is even. Recall that the (unpreconditioned) GMRES method converges very slowly for both types of discretizations (upwind and central differences); see Figures 2.8–2.8.

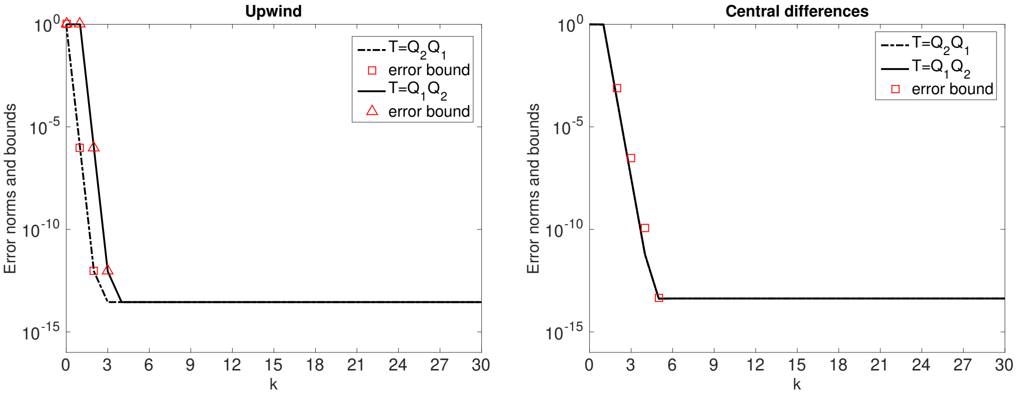


Figure 3.1.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-8}$, $N = 198$, and both discretization schemes.

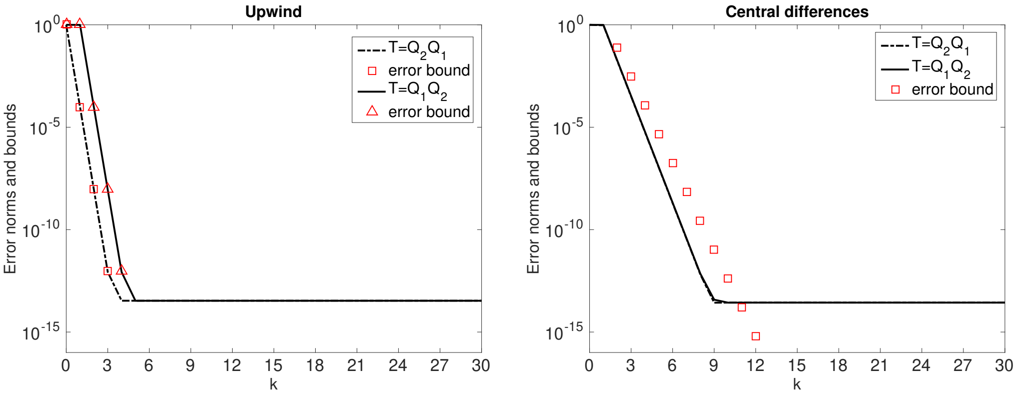


Figure 3.2.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-6}$, $N = 198$, and both discretization schemes.

All numerical results presented in this chapter were performed on a 13-inch Apple MacBook computer model Mid 2010 with a 2,4 GHz Intel Core 2 Duo processor. For our experiments we computed $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}$ using the backslash operator in `MATLAB` version R2015b. (Applying iterative refinement in order to improve the numerical solution obtained in this way yields virtually the same results, so we do not consider iterative refinement here.) Using the solution obtained by `MATLAB`'s backslash, we computed the error norms of the multiplicative Schwarz method by $\|\mathbf{e}^{(k)}\|_\infty = \|\mathbf{u}^{(k+1)} - \mathbf{u}\|_\infty$ with $\mathbf{u}^{(k+1)}$ as in (2.70) and $\mathbf{u}^{(0)} = 0$ (rather than using the

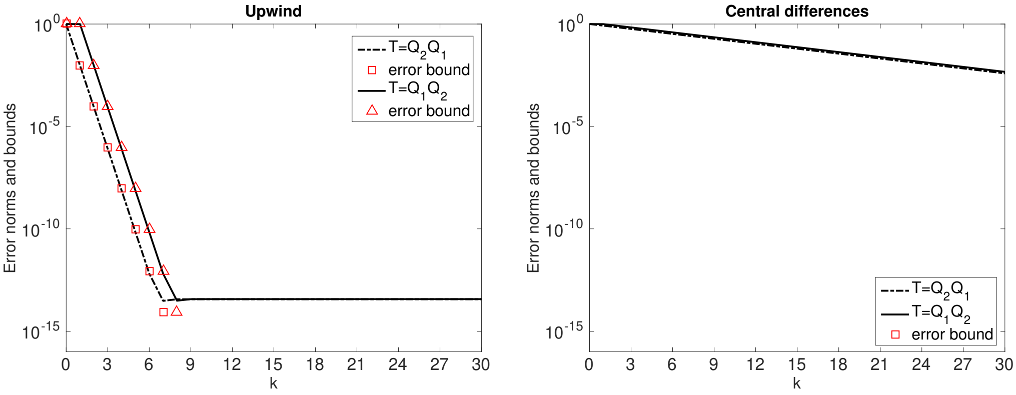


Figure 3.3.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-4}$, $N = 198$, and both discretization schemes.

update formula $\mathbf{e}^{(k)} = \mathbf{T}_{ij}\mathbf{e}^{(k-1)}$). Consequently, the computed error norms stagnate on the level of the maximal attainable accuracy of the method. On the other hand, an error bound of the form $|\rho|^k$ for some $|\rho| < 1$ becomes arbitrarily small for $k \to \infty$.

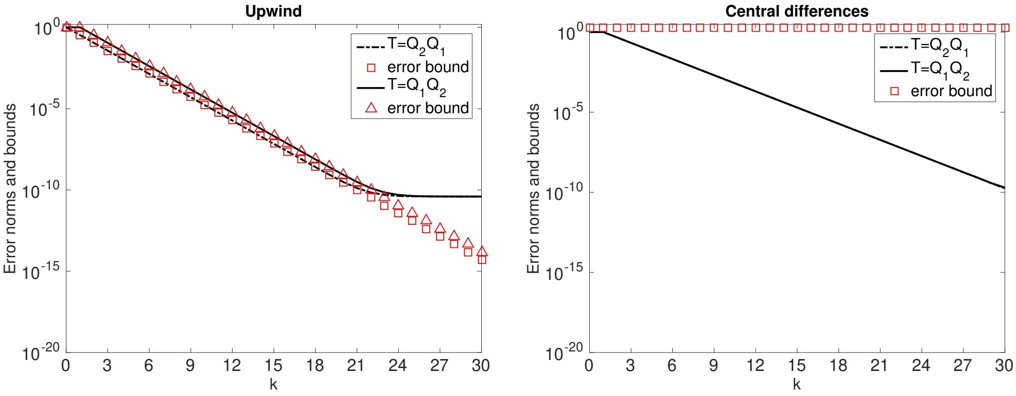


Figure 3.4.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-8}$, $N = 10002$ and both discretization schemes.

We start with the upwind discretization. The left parts of Figures 3.1–3.5 show the error norms

$$\frac{\|\mathbf{e}^{(k)}\|_\infty}{\|\mathbf{e}^{(0)}\|_\infty}, \quad k = 0, 1, 2 \ldots,$$

for the iteration matrices $\mathbf{T}_{12}$ (solid) and $\mathbf{T}_{21}$ (dashed) as well as the corresponding upper bounds from Theorem 3.6, for increasing values of ϵ. We observe that the bounds are quite close to the actual errors. Moreover, in each case the error norm for the multiplicative Schwarz method with the iteration matrix $\mathbf{T}_{21}$ almost stagnates in the first step, as predicted by the bound in Theorem 3.6.

On the right parts of Figures 3.1–3.5 we show the error norms of the multiplicative

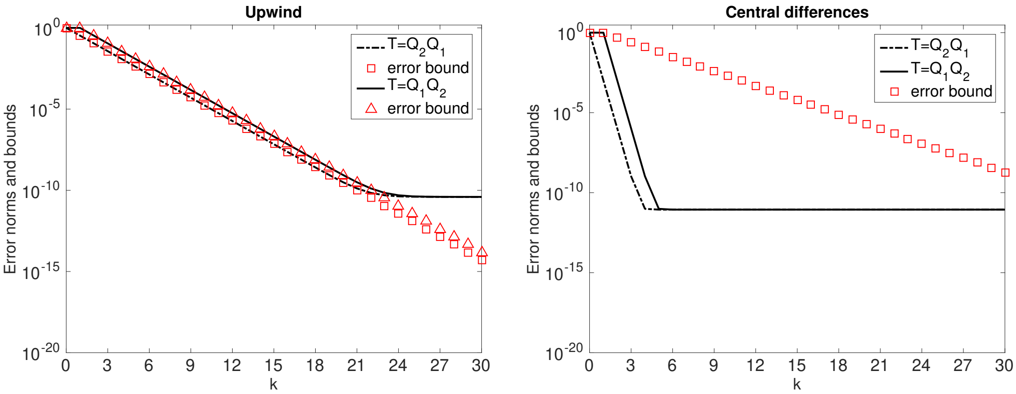


Figure 3.5.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-4}$, $N = 10002$ and both discretization schemes.

Schwarz method and the corresponding convergence bounds from Theorem 3.13 for the central difference scheme. For our choice of parameters we have $\omega_x H > 2\epsilon$. Note that the error norms are virtually the same for both iteration matrices. However, the bounds are not as tight as for the upwind scheme. For fixed N the bounds become weaker with increasing ϵ, i.e., decreasing convection- dominance. For our chosen parameters and $\epsilon = 10^{-4}$, giving $\epsilon N^2 = \mathcal{O}(1)$, the convergence of the multiplicative Schwarz method becomes very slow, and the bound (3.28) fails to predict convergence at all.

We also run the experiments for larger values of N. The values of $|\rho|$ and the corresponding bounds from Theorems 3.6 and 3.13 are shown in Table 3.1 for different values of N. We observe that for all cases the bounds on $|\rho|$ for the upwind scheme are tighter than for the central difference scheme.

To further illustrate our results, we also present the results for a larger value of N, namely $N = 10002$. We consider the special cases $\epsilon N^2 \approx 1$ (Figure 3.4) and $\epsilon N \approx 1$ (Figure 3.5) which are mainly of theoretical interest. While the bound (3.13) for the upwind scheme is still tight and descriptive, the bound (3.28) for the central difference scheme does not predict convergence well. Note that the parameters used in Figure 3.5 yield $\omega_x H \approx 1.9959 \times 10^{-4} < 2\epsilon$, and hence the right part of Figure 3.5 shows error norms and the convergence bound corresponding to the case $\omega_x H \leq 2\epsilon$ in Theorem 3.13.

We continue our numerical experiments by applying GMRES to the linear algebraic system *preconditioned with multiplicative Schwarz*, i.e., the linear algebraic system (3.29), in the case $N = 198$. Using the 2-norm, the (preconditioned) relative residual norms are shown in Figures 3.6–3.7 and were computed using `MATLAB`'s command `gmres` with a tolerance of 10^{-14}, a maximum number of iterations of $N - 1$ and initial approximation $\mathbf{u}^{(0)} = 0$. In all cases convergence is achieved in two iterations, which is explained in the previous section. These figures are the counterparts to Figures 2.8–2.9, where GMRES makes little progress until iteration 198.

$N = 66$				
10^{-8}	2.9×10^{-7}	3.3×10^{-7}	1.8×10^{-5}	4.2×10^{-5}
10^{-6}	2.9×10^{-5}	3.3×10^{-5}	1.8×10^{-3}	4.2×10^{-3}
10^{-4}	2.9×10^{-3}	3.3×10^{-3}	1.8×10^{-1}	4.2×10^{-1}
10^{-2}	2.3×10^{-1}	2.6×10^{-1}	1.3×10^{-1}	$2.7 \times 10^{+1}$
$N = 130$				
10^{-8}	6.0×10^{-7}	6.5×10^{-7}	7.7×10^{-5}	1.7×10^{-4}
10^{-6}	6.0×10^{-5}	6.5×10^{-5}	7.7×10^{-3}	1.7×10^{-2}
10^{-4}	6.0×10^{-3}	6.5×10^{-3}	5.9×10^{-1}	1.6×10^{0}
10^{-2}	3.8×10^{-1}	4.2×10^{-1}	1.6×10^{-1}	5.7×10^{-1}
$N = 258$				
10^{-8}	1.2×10^{-6}	1.3×10^{-6}	3.2×10^{-4}	6.6×10^{-4}
10^{-6}	1.2×10^{-4}	1.3×10^{-4}	3.2×10^{-2}	6.6×10^{-2}
10^{-4}	1.2×10^{-2}	1.3×10^{-2}	8.7×10^{-1}	6.4×10^{0}
10^{-2}	5.5×10^{-1}	5.9×10^{-1}	4.5×10^{-1}	7.2×10^{-1}
$N = 514$				
10^{-8}	2.5×10^{-6}	2.6×10^{-6}	1.3×10^{-3}	2.6×10^{-3}
10^{-6}	2.5×10^{-4}	2.6×10^{-4}	1.3×10^{-1}	2.6×10^{-1}
10^{-4}	2.4×10^{-2}	2.5×10^{-2}	8.6×10^{-1}	$2.5 \times 10^{+1}$
10^{-2}	7.2×10^{-1}	7.5×10^{-1}	6.8×10^{-1}	8.4×10^{-1}
$N = 1026$				
10^{-8}	5.1×10^{-6}	5.1×10^{-6}	5.2×10^{-3}	1.1×10^{-2}
10^{-6}	5.1×10^{-4}	5.1×10^{-4}	4.7×10^{-1}	1.0×10^{0}
10^{-4}	4.8×10^{-2}	4.9×10^{-2}	7.9×10^{-1}	$9.5 \times 10^{+1}$
10^{-2}	8.4×10^{-1}	8.6×10^{-1}	8.2×10^{-1}	9.1×10^{-1}
$N = 2050$				
10^{-8}	1.0×10^{-5}	1.0×10^{-5}	2.1×10^{-2}	4.2×10^{-2}
10^{-6}	1.0×10^{-3}	1.0×10^{-3}	9.5×10^{-1}	4.2×10^{0}
10^{-4}	9.2×10^{-2}	9.3×10^{-2}	6.5×10^{-1}	$3.5 \times 10^{+2}$
10^{-2}	9.1×10^{-1}	9.2×10^{-1}	9.1×10^{-1}	9.5×10^{-1}
$N = 4098$				
10^{-8}	2.0×10^{-5}	2.0×10^{-5}	8.3×10^{-2}	1.7×10^{-1}
10^{-6}	2.0×10^{-3}	2.0×10^{-3}	9.8×10^{-1}	$1.7 \times 10^{+1}$
10^{-4}	1.7×10^{-1}	1.7×10^{-1}	4.1×10^{-1}	$1.2 \times 10^{+3}$
10^{-2}	9.5×10^{-1}	9.6×10^{-1}	9.5×10^{-1}	9.8×10^{-1}
$N = 8194$				
10^{-8}	4.1×10^{-5}	4.1×10^{-5}	3.2×10^{-1}	6.7×10^{-1}
10^{-6}	4.1×10^{-3}	4.1×10^{-3}	9.8×10^{-1}	$6.7 \times 10^{+1}$
10^{-4}	2.9×10^{-1}	2.9×10^{-1}	9.8×10^{-2}	$3.7 \times 10^{+3}$
10^{-2}	9.8×10^{-1}	9.8×10^{-1}	9.8×10^{-1}	9.9×10^{-1}
ϵ	$\lvert\rho\rvert$ (3.11)	bound (3.13)	$\lvert\rho\rvert$ (3.11)	bound (3.28)
	upwind		central differences	

Table 3.1.: Values of $|\rho|$ computed using (3.11) and the corresponding bounds (3.13) and (3.28) for different values of ϵ and N.

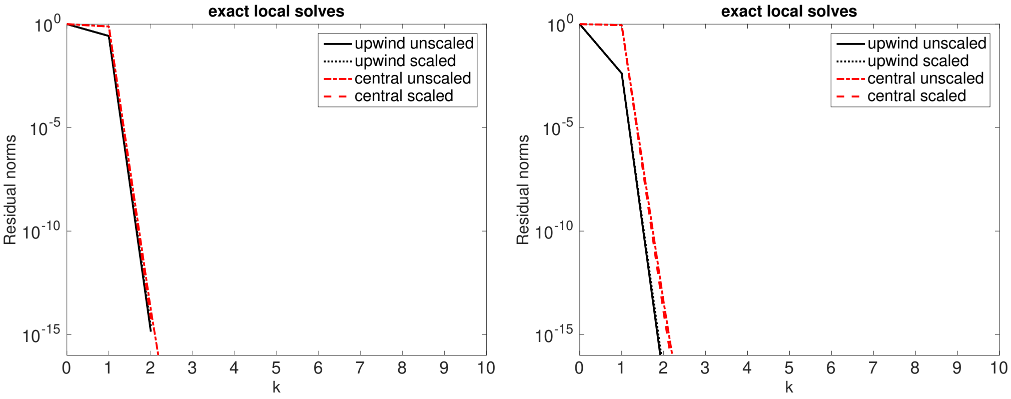


Figure 3.6.: Preconditioned GMRES convergence for $\epsilon = 10^{-2}$ [l.] and $\epsilon = 10^{-4}$ [r.]

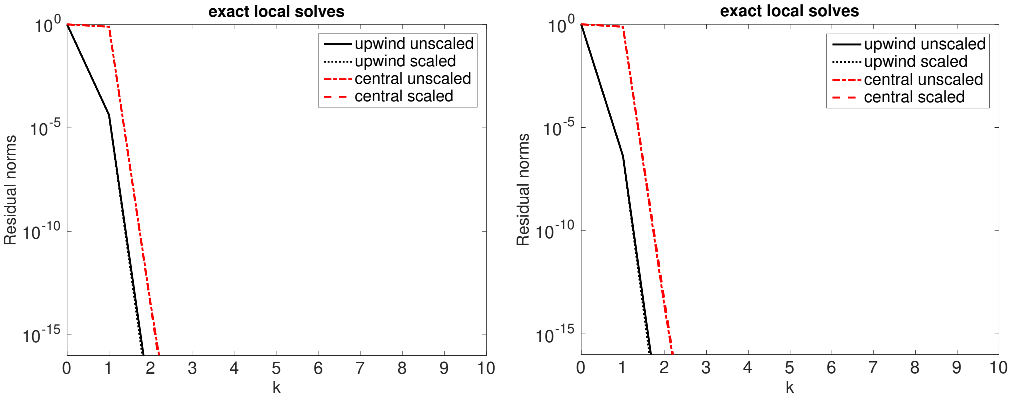


Figure 3.7.: Preconditioned GMRES convergence for $\epsilon = 10^{-6}$ [l.] and $\epsilon = 10^{-8}$ [r.].

4. The Theory of Block Diagonal Dominant Matrices

Parts of this chapter have already been published in:

[22] C. Echeverría, J. Liesen, and R. Nabben, **Block diagonal dominace of matrices revisited: Bounds for the norms of inverses and eigenvalue inclusion sets**, Linear Algebra Appl., 553 (2018), pp. 365–383.

4.1. Introduction

The main tool that is used for proving the convergence of the multiplicative Schwarz method in Theorems 3.6 and 3.13 is Lemma 3.4, which characterizes the decay away from the diagonal shown by the entries of the inverse of a tridiagonal Toeplitz matrix whenever it possesses the property of diagonal dominance and its sub- and super-diagonal entries have opposite signs. With the objective of deriving an equally powerful tool that can be applied to the matrices that arise in the discretization of higher dimensional convection-diffusion problems (treated in the Chapter 5), we now present a generalization of the classical theory of diagonal dominance of matrices from the scalar to the block case.

Matrices that are characterized by off-diagonal decay, or more generally "localization" of their entries, appear in applications throughout the mathematical and computational sciences. The presence of such localization can lead to computational savings, since it allows to approximate a given matrix by only using its significant entries, and discarding the negligible ones according to a pre-established criterion. In this context it is then of great practical interest to know a priori how many and which of these entries can be discarded as insignificant. Many authors have therefore studied decay rates for different matrix classes and functions of matrices; see, e.g., [3, 6, 8, 9, 15, 19, 48, 65]. For an excellent survey of the current state-of-the-art we refer to [2].

An important example in this context is given by the (nonsymmetric) diagonally dominant matrices, and in particular the diagonally dominant tridiagonal matrices, which were studied, e.g., in [61, 62]. As shown in these works, the entries of the inverse decay with an exponential rate along a row or column, depending on whether the given matrix is row or column diagonally dominant; see [2, § 3.2] for a more general treatment of decay bounds for the inverse and further references.

Our main goal in this chapter is to generalize results of [62] from scalar to block tridiagonal matrices. In order to do so, we use a generalization of the classical definition of block diagonal dominance of Feingold and Varga [30] to derive bounds and decay rates for the block norms of the inverse of block tridiagonal matrices of the form

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 & \mathbf{B}_1 & & & \\ \mathbf{C}_1 & \mathbf{A}_2 & \mathbf{B}_2 & & \\ & \ddots & \ddots & \ddots & \\ & & \mathbf{C}_{n-2} & \mathbf{A}_{n-1} & \mathbf{B}_{n-1} \\ & & & \mathbf{C}_{n-1} & \mathbf{A}_n \end{bmatrix}, \quad \text{where } \mathbf{A}_i, \mathbf{B}_i, \mathbf{C}_i \in \mathbb{C}^{m\times m}. \tag{4.1}$$

We also show how to improve these bounds iteratively (Section 4.2). Moreover, we obtain a new variant of the Gershgorin Circle Theorem for general block matrices of the form

$$\mathbf{A} = [\mathbf{A}_{ij}] \quad \text{with blocks } \mathbf{A}_{ij} \in \mathbb{C}^{m\times m} \text{ for } i, j = 1, \dots, n, \tag{4.2}$$

which can provide tighter spectral inclusion regions than those obtained by Feingold and Varga (Section 4.3). Throughout this chapter we assume that $\|\cdot\|$ is a given submultiplicative matrix norm.

4.2. Bounds on the Inverses of Block Tridiagonal Matrices

We start with a definition of block diagonally dominant matrices.

Definition 4.1. *Consider a matrix of the form* (4.2). *The matrix* $\mathbf{A}$ *is called* row block diagonally dominant *(with respect to the matrix norm* $\|\cdot\|$*) when the diagonal blocks* $\mathbf{A}_{ii}$ *are nonsingular, and*

$$\sum_{\substack{j=1\\ j\neq i}}^{n} \|\mathbf{A}_{ii}^{-1}\mathbf{A}_{ij}\| \le 1, \quad \textit{for } i = 1, \dots, n. \tag{4.3}$$

If a strict inequality holds in (4.3) *then* $\mathbf{A}$ *is called* row block strictly diagonally dominant *(with respect to the matrix norm* $\|\cdot\|$*).*

Obviously, an analogous definition of *column* block diagonal dominance is possible. Most of the results in this chapter can be easily rewritten for that case (see Definition 4.19 in Appendix 4.A and the discussion thereafter). A similar definition recently presented in [4] in an application of block diagonal preconditioning where the authors call a matrix block diagonally dominant when all its diagonal blocks are nonsingular, and (4.3) or the analogous conditions with $\mathbf{A}_{ij}\mathbf{A}_{ii}^{-1}$ replacing $\mathbf{A}_{ii}^{-1}\mathbf{A}_{ij}$ hold in the 1-norm, i.e., it does not consider both conditions to hold simultaneously and it is restricted to the 1-norm.

The above definition of row block diagonal dominance generalizes the one of Feingold and Varga given in [30, Definition 1], who considered a matrix as in (4.2) block diagonally dominant when the diagonal blocks $\mathbf{A}_{ii}$ are nonsingular, and

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} \|\mathbf{A}_{ij}\| \leq (\|\mathbf{A}_{ii}^{-1}\|)^{-1}, \quad \text{for } i = 1, \dots, n. \tag{4.4}$$

It is clear that if a matrix satisfies these conditions, then it also satisfies the conditions given in Definition 4.1. According to Varga [78, p. 156], the definition of block diagonal dominance given in [30] is one of the earliest, and it was roughly simultaneously and independently considered also by Ostrowski [63] and Fiedler and Pták [31]. Varga calls this a “Zeitgeist” phenomenon.

In the special case $m = 1$, i.e., all the blocks $\mathbf{A}_{ij}$ are of size 1×1 and $\|\mathbf{A}_{ij}\| = |\mathbf{A}_{ij}|$, the inequalities (4.3) and (4.4) are equivalent, and they can all be written as

$$\sum_{\substack{j=1 \\ j\neq i}}^{n} |\mathbf{A}_{ij}| \leq |\mathbf{A}_{ii}|, \quad \text{for } i = 1, \dots, n,$$

which is the usual definition of row diagonal dominance.

In the rest of this section we will restrict our attention to block tridiagonal matrices of the form (4.1). First Capovani for the scalar case in [17, 16] and later Ikebe for the block case in [45] (see also [62]), have shown that the inverse of a nonsingular block tridiagonal matrix can be described by four sets of matrices. The main result can be stated as follows.

Theorem 4.2. *Let* $\mathbf{A}$ *be as in* (4.1)*, and suppose that* $\mathbf{A}^{-1}$ *as well as* $\mathbf{B}_i^{-1}$ *and* $\mathbf{C}_i^{-1}$ *for* $i = 1, \dots, n-1$ *exist. If we write* $\mathbf{A}^{-1} = [\mathbf{Z}_{ij}]$ *with* $\mathbf{Z}_{ij} \in \mathbb{C}^{m\times m}$*, then there exist matrices* $\mathbf{U}_i, \mathbf{V}_i, \mathbf{X}_i, \mathbf{Y}_i \in \mathbb{C}^{m\times m}$ *with* $\mathbf{U}_i\mathbf{V}_i = \mathbf{X}_i\mathbf{Y}_i$ *for* $i = 1, \dots, n$*, and*

$$\mathbf{Z}_{ij} = \begin{cases} \mathbf{U}_i\mathbf{V}_j & \text{if } i \leq j, \\ \mathbf{Y}_i\mathbf{X}_j & \text{if } i \geq j. \end{cases} \tag{4.5}$$

Moreover, the matrices $\mathbf{U}_i, \mathbf{V}_i, \mathbf{X}_i, \mathbf{Y}_i$*,* $i = 1, \dots, n$*, are recursively given by*

$$\mathbf{U}_1 = \mathbf{I}, \quad \mathbf{U}_2 = -\mathbf{B}_1^{-1}\mathbf{A}_1\mathbf{U}_1, \tag{4.6}$$

$$\mathbf{U}_i = -\mathbf{B}_{i-1}^{-1}(\mathbf{C}_{i-2}\mathbf{U}_{i-2} + \mathbf{A}_{i-1}\mathbf{U}_{i-1}), \quad \text{for } i = 3, \dots, n, \tag{4.7}$$

$$\mathbf{V}_n = (\mathbf{A}_n\mathbf{U}_n + \mathbf{C}_{n-1}\mathbf{U}_{n-1})^{-1}, \quad \mathbf{V}_{n-1} = -\mathbf{V}_n\mathbf{A}_n\mathbf{B}_{n-1}^{-1}, \tag{4.8}$$

$$\mathbf{V}_i = -(\mathbf{V}_{i+1}\mathbf{A}_{i+1} + \mathbf{V}_{i+2}\mathbf{C}_{i+1})\mathbf{B}_i^{-1}, \quad \text{for } i = n-2, \dots, 1. \tag{4.9}$$

$$\mathbf{X}_1 = \mathbf{I}, \quad \mathbf{X}_2 = -\mathbf{X}_1\mathbf{A}_1\mathbf{C}_1^{-1}, \tag{4.10}$$

$$\mathbf{X}_i = -(\mathbf{X}_{i-2}\mathbf{B}_{i-2} + \mathbf{X}_{i-1}\mathbf{A}_{i-1})\mathbf{C}_{i-1}^{-1}, \quad \text{for } i = 3, \dots, n, \tag{4.11}$$

$$\mathbf{Y}_n = (\mathbf{X}_n\mathbf{A}_n + \mathbf{X}_{n-1}\mathbf{B}_{n-1})^{-1}, \quad \mathbf{Y}_{n-1} = -\mathbf{C}_{n-1}^{-1}\mathbf{A}_n\mathbf{Y}_n, \tag{4.12}$$

$$\mathbf{Y}_i = -\mathbf{C}_i^{-1}(\mathbf{A}_{i+1}\mathbf{Y}_{i+1} + \mathbf{B}_{i+1}\mathbf{Y}_{i+2}), \quad \text{for } i = n-2, \dots, 1. \tag{4.13}$$

The next result is a generalization of [61, Theorem 3.2].

Lemma 4.3. *Let* $\mathbf{A}$ *be a matrix as in Theorem 4.2. Suppose in addition that* $\mathbf{A}$ *is row block diagonally dominant, and that*

$$\|\mathbf{A}_1^{-1}\mathbf{B}_1\| < 1 \quad \textit{and} \quad \|\mathbf{A}_n^{-1}\mathbf{C}_{n-1}\| < 1. \tag{4.14}$$

Then the sequence $\{\|\mathbf{U}_i\|\}_{i=1}^n$ *is strictly increasing, and the sequence* $\{\|\mathbf{Y}_i\|\}_{i=1}^n$ *is strictly decreasing.*

Proof. First we consider the sequence $\{\|\mathbf{U}_i\|\}_{i=1}^n$. The definition of $\mathbf{U}_2$ in (4.6) implies that $\mathbf{U}_1 = -\mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{U}_2$. Taking norms and using the first inequality in (4.14) yields

$$\|\mathbf{U}_1\| \leq \|\mathbf{A}_1^{-1}\mathbf{B}_1\|\|\mathbf{U}_2\| < \|\mathbf{U}_2\|.$$

Now suppose that $\|\mathbf{U}_1\| < \|\mathbf{U}_2\| < \cdots < \|\mathbf{U}_{i-1}\|$ holds for some $i \geq 3$. The equation for U_i in (4.7) can be written as

$$-\mathbf{A}_{i-1}^{-1}\mathbf{B}_{i-1}\mathbf{U}_i = \mathbf{U}_{i-1} + \mathbf{A}_{i-1}^{-1}\mathbf{C}_{i-2}\mathbf{U}_{i-2}.$$

Rearranging terms and taking norms we obtain

$$\begin{aligned}\|\mathbf{U}_{i-1}\| &\leq \|\mathbf{A}_{i-1}^{-1}\mathbf{C}_{i-2}\|\|\mathbf{U}_{i-2}\| + \|\mathbf{A}_{i-1}^{-1}\mathbf{B}_{i-1}\|\|\mathbf{U}_i\| \\ &< \|\mathbf{A}_{i-1}^{-1}\mathbf{C}_{i-2}\|\|\mathbf{U}_{i-1}\| + \|\mathbf{A}_{i-1}^{-1}\mathbf{B}_{i-1}\|\|\mathbf{U}_i\|,\end{aligned}$$

where we have used the induction hypothesis, i.e., $\|\mathbf{U}_{i-2}\| < \|\mathbf{U}_{i-1}\|$, in order to obtain the strict inequality. Since $\mathbf{A}$ is row block diagonally dominant we have

$$\|\mathbf{A}_{i-1}^{-1}\mathbf{B}_{i-1}\| + \|\mathbf{A}_{i-1}^{-1}\mathbf{C}_{i-2}\| \leq 1.$$

Combining this with the previous inequality gives

$$\frac{\|\mathbf{U}_{i-1}\|}{\|\mathbf{U}_i\|} < \frac{\|\mathbf{A}_{i-1}^{-1}\mathbf{B}_{i-1}\|}{1 - \|\mathbf{A}_{i-1}^{-1}\mathbf{C}_{i-2}\|} \leq 1,$$

so that indeed $\|\mathbf{U}_{i-1}\| < \|\mathbf{U}_i\|$.

Next we consider the sequence $\{\|\mathbf{Y}_i\|\}_{i=1}^n$. The definition of $\mathbf{Y}_{n-1}$ in (4.12) implies that $-\mathbf{Y}_n = \mathbf{A}_n^{-1}\mathbf{C}_{n-1}\mathbf{Y}_{n-1}$. Taking norms and using the second inequality in (4.14) yields

$$\|\mathbf{Y}_n\| \leq \|\mathbf{A}_n^{-1}\mathbf{C}_{n-1}\|\|\mathbf{Y}_{n-1}\| < \|\mathbf{Y}_{n-1}\|.$$

Now suppose that $\|\mathbf{Y}_n\| < \|\mathbf{Y}_{n-1}\| < \cdots < \|\mathbf{Y}_{i+1}\|$ holds for some $i \leq n-2$. The equation for $\mathbf{Y}_i$ in (4.13) can be written as

$$-\mathbf{A}_{i+1}^{-1}\mathbf{C}_i\mathbf{Y}_i = \mathbf{Y}_{i+1} + \mathbf{A}_{i+1}^{-1}\mathbf{B}_{i+1}\mathbf{Y}_{i+2}.$$

Rearranging terms and taking norms we obtain

$$\begin{aligned}\|\mathbf{Y}_{i+1}\| &\leq \|\mathbf{A}_{i+1}^{-1}\mathbf{C}_i\|\|\mathbf{Y}_i\| + \|\mathbf{A}_{i+1}^{-1}\mathbf{B}_{i+1}\|\|\mathbf{Y}_{i+2}\| \\ &< \|\mathbf{A}_{i+1}^{-1}\mathbf{C}_i\|\|\mathbf{Y}_i\| + \|\mathbf{A}_{i+1}^{-1}\mathbf{B}_{i+1}\|\|\mathbf{Y}_{i+1}\|,\end{aligned}$$

where we have used the induction hypothesis, i.e., $\|\mathbf{Y}_{i+2}\| < \|\mathbf{Y}_{i+1}\|$, in order to obtain the strict inequality. Since $\mathbf{A}$ is row block diagonally dominant we have

$$\|\mathbf{A}_{i+1}^{-1}\mathbf{C}_i\| + \|\mathbf{A}_{i+1}^{-1}\mathbf{B}_{i+1}\| \le 1.$$

Combining this with the previous inequality gives

$$\frac{\|\mathbf{Y}_{i+1}\|}{\|\mathbf{Y}_i\|} < \frac{\|\mathbf{A}_{i+1}^{-1}\mathbf{C}_i\|}{1 - \|\mathbf{A}_{i+1}^{-1}\mathbf{B}_{i+1}\|} \le 1$$

so that indeed $\|\mathbf{Y}_{i+1}\| < \|\mathbf{Y}_i\|$. □

For the rest of this section we will assume that that $\mathbf{A}$ is a matrix as in Lemma 4.3. Then the inverse is given by $\mathbf{A}^{-1} = [\mathbf{Z}_{ij}]$ with $\mathbf{Z}_{ij} = \mathbf{Y}_i\mathbf{X}_j$ for $i \ge j$; see Theorem 4.2. Thus, for each *fixed* $j = 1, \dots, n$, the strict decrease of the sequence $\{\|Y_i\|\}_{i=1}^n$ suggests that the sequence $\{\|\mathbf{Z}_{ij}\|\}_{i=j}^n$ decreases as well, i.e., that the norms of the blocks of $\mathbf{A}^{-1}$ decay columnwise away from the diagonal. We will now study this decay in detail.

We set $\mathbf{C}_0 = \mathbf{B}_n = 0$, and define

$$\tau_i \equiv \frac{\|\mathbf{A}_i^{-1}\mathbf{B}_i\|}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|}, \quad \text{for } i = 1, \dots, n, \tag{4.15}$$

$$\mu_i \equiv \frac{\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|}{1 - \|\mathbf{A}_i^{-1}\mathbf{B}_i\|}, \quad \text{for } i = 1, \dots, n. \tag{4.16}$$

The row block diagonal dominance of $\mathbf{A}$ then implies that $0 \le \tau_i \le 1$ and $0 \le \mu_i \le 1$. Also note that, by assumption, $\tau_1 = \|\mathbf{A}_1^{-1}\mathbf{B}_1\| < 1$, $\mu_n = \|\mathbf{A}_n^{-1}\mathbf{C}_{n-1}\| < 1$, and $\tau_n = \mu_1 = 0$.

In order to obtain bounds on the norms of the block entries $\mathbf{A}^{-1}$, we will first derive alternative recurrence formulas for the matrices $\mathbf{U}_i$ and $\mathbf{Y}_i$ from Lemma 4.3. To this end, we introduce some intermediate quantities and give bounds on their norms in the following result.

Lemma 4.4. *The following assertions hold:*

(a) The matrices $\mathbf{L}_1 = \mathbf{T}_1 = \mathbf{A}_1^{-1}\mathbf{B}_1$, $\mathbf{T}_2 = \mathbf{I} - \mathbf{A}_2^{-1}\mathbf{C}_1\mathbf{T}_1$, *and*

$$\mathbf{L}_i = \mathbf{T}_i^{-1}\mathbf{A}_i^{-1}\mathbf{B}_i, \quad \textit{for } i = 2, \dots, n-1,$$
$$\mathbf{T}_i = \mathbf{I} - \mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}, \quad \textit{for } i = 3, \dots, n-1,$$

are all nonsingular, and $\|\mathbf{L}_i\| \le \tau_i$, *for* $i = 1, \dots, n-1$.

(b) The matrices $\mathbf{M}_n = \mathbf{W}_n = \mathbf{A}_n^{-1}\mathbf{C}_{n-1}$, $\mathbf{W}_{n-1} = \mathbf{I} - \mathbf{A}_{n-1}^{-1}\mathbf{B}_{n-1}\mathbf{W}_n$, *and*

$$\mathbf{M}_i = \mathbf{W}_i^{-1}\mathbf{A}_i^{-1}\mathbf{C}_{i-1}, \quad \textit{for } i = n-1, \dots, 2,$$
$$\mathbf{W}_i = \mathbf{I} - \mathbf{A}_i^{-1}\mathbf{B}_i\mathbf{M}_{i+1}, \quad \textit{for } i = n-2, \dots, 2,$$

are all nonsingular, and $\|\mathbf{M}_i\| \le \mu_i$, *for* $i = 2, \dots, n$.

Proof. We only prove *(a)*; the proof of *(b)* is analogous. The matrices $\mathbf{L}_1 = \mathbf{T}_1 = \mathbf{A}_1^{-1}\mathbf{B}_1$ are nonsingular since both $\mathbf{A}_1$ and $\mathbf{B}_1$ are. Moreover, (4.14) gives $\|\mathbf{L}_1\| = \|\mathbf{T}_1\| = \|\mathbf{A}_1^{-1}\mathbf{B}_1\| = \tau_1 < 1$. Now suppose that $\|\mathbf{L}_{i-1}\| \leq \tau_{i-1} \leq 1$ holds for some $i \geq 2$. Then

$$\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}\| \leq \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\|\mathbf{L}_{i-1}\| < 1,$$

where we have also used that $\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\| \leq 1 - \|\mathbf{A}_i^{-1}\mathbf{B}_i\| < 1$. Thus, $\mathbf{T}_i = \mathbf{I} - \mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}$ is nonsingular, and therefore $\mathbf{L}_i = \mathbf{T}_i^{-1}\mathbf{A}_i^{-1}\mathbf{B}_i$ is nonsingular. Using the Neumann series gives

$$\begin{aligned}\|\mathbf{T}_i^{-1}\| = \|(\mathbf{I} - \mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1})^{-1}\| &= \left\|\sum_{k=0}^{\infty}(\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1})^k\right\| \leq \sum_{k=0}^{\infty}\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}\|^k \\ &= \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}\|} \leq \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\|\mathbf{L}_{i-1}\|} \leq \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|},\end{aligned}$$

and $\|\mathbf{L}_i\| = \|\mathbf{T}_i^{-1}\mathbf{A}_i^{-1}\mathbf{B}_i\| \leq \frac{\|\mathbf{A}_i^{-1}\mathbf{B}_i\|}{1-\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|} = \tau_i \leq 1$, which finishes the proof. □

Using Lemma 4.4 we can now derive alternative recurrences for the matrices $\mathbf{U}_i$ and $\mathbf{Y}_i$ from Lemma 4.3.

Lemma 4.5. *If* $\mathbf{A}$ *is a matrix as in Lemma 4.3, then the corresponding matrices* $\mathbf{U}_i$ *and* $\mathbf{Y}_i$ *are given by*

$$\mathbf{U}_i = -\mathbf{L}_i\mathbf{U}_{i+1}, \quad \textit{for } i = 1, \dots n-1, \tag{4.17}$$

$$\mathbf{Y}_i = -\mathbf{M}_i\mathbf{Y}_{i-1}, \quad \textit{for } i = n, \dots 2, \tag{4.18}$$

where the matrices $\mathbf{L}_i$ *and* $\mathbf{M}_i$ *are defined as in Lemma 4.4.*

Proof. We only prove that (4.17) holds; the proof of (4.18) is analogous. From (4.6) and the definition of $\mathbf{T}_1$ in Lemma 4.3 we obtain

$$\mathbf{U}_1 = -\mathbf{A}_1^{-1}\mathbf{B}_1\mathbf{U}_2 = -\mathbf{L}_1\mathbf{U}_2.$$

We next write (4.7) for $i = 3$ as

$$-\mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{U}_3 = \mathbf{A}_2^{-1}\mathbf{C}_1\mathbf{U}_1 + \mathbf{U}_2 = -\mathbf{A}_2^{-1}\mathbf{C}_1\mathbf{T}_1\mathbf{U}_2 + \mathbf{U}_2 = (\mathbf{I} - \mathbf{A}_2^{-1}\mathbf{C}_1\mathbf{T}_1)\mathbf{U}_2 = \mathbf{T}_2\mathbf{U}_2,$$

and hence

$$\mathbf{U}_2 = -\mathbf{T}_2^{-1}\mathbf{A}_2^{-1}\mathbf{B}_2\mathbf{U}_3 = -\mathbf{L}_2\mathbf{U}_3.$$

Now suppose that $\mathbf{U}_{i-1} = -\mathbf{L}_{i-1}\mathbf{U}_i$ holds for some $3 \leq i \leq n-1$. Then from (4.7) we obtain

$$-\mathbf{A}_i^{-1}\mathbf{B}_i\mathbf{U}_{i+1} = \mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{U}_{i-1} + \mathbf{U}_i = (\mathbf{I} - \mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1})\mathbf{U}_i = \mathbf{T}_i\mathbf{U}_i,$$

and hence

$$\mathbf{U}_i = -\mathbf{T}_i^{-1}\mathbf{A}_i^{-1}\mathbf{B}_i\mathbf{U}_{i+1} = -\mathbf{L}_i\mathbf{U}_{i+1},$$

which completes the proof. □

We are now ready to state and prove our bounds on the norms of the blocks of $\mathbf{A}^{-1}$, which generalize [62, Theorems 3.1 and 3.2] from the scalar to the block case.

Theorem 4.6. *If* $\mathbf{A}$ *is a matrix as in Lemma 4.3, then* $\mathbf{A}^{-1} = [\mathbf{Z}_{ij}]$ *with*

$$\|\mathbf{Z}_{ij}\| \leq \|\mathbf{Z}_{jj}\| \prod_{k=i}^{j-1} \tau_k, \qquad \textit{for all } i < j, \tag{4.19}$$

$$\|\mathbf{Z}_{ij}\| \leq \|\mathbf{Z}_{jj}\| \prod_{k=j+1}^{i} \mu_k, \quad \textit{for all } i > j. \tag{4.20}$$

for τ_k *and* μ_k *given by* (4.15) *and* (4.16) *. Moreover, for* $i = 1, \dots, n$,

$$\frac{\|\mathbf{I}\|}{\|\mathbf{A}_i\| + \tau_{i-1}\|\mathbf{C}_{i-1}\| + \mu_{i+1}\|\mathbf{B}_i\|} \leq \|\mathbf{Z}_{ii}\| \leq \frac{\|\mathbf{I}\|}{\|\mathbf{A}_i^{-1}\|^{-1} - \tau_{i-1}\|\mathbf{C}_{i-1}\| - \mu_{i+1}\|\mathbf{B}_i\|}, \tag{4.21}$$

provided that the denominator of the upper bound is larger than zero, and where we set $\mathbf{C}_0 = \mathbf{B}_n = 0$*, and* $\tau_0 = \mu_{n+1} = 0$.

Proof. From Lemma 4.5 we know that $\mathbf{U}_i = -\mathbf{L}_i\mathbf{U}_{i+1}$ holds for $i = 1, \dots, (n-1)$. Thus, for all $i < j$,

$$\mathbf{Z}_{ij} = \mathbf{U}_i\mathbf{V}_j = -\mathbf{L}_i\mathbf{U}_{i+1}\mathbf{V}_j = (-1)^{j-i}\left(\prod_{k=i}^{j-1}\mathbf{L}_k\right)\mathbf{U}_j\mathbf{V}_j = (-1)^{j-i}\left(\prod_{k=i}^{j-1}\mathbf{L}_k\right)\mathbf{Z}_{jj}.$$

Taking norms and using Lemma 4.4 yields

$$\|\mathbf{Z}_{ij}\| \leq \|\mathbf{Z}_{jj}\| \prod_{k=i}^{j-1}\|\mathbf{L}_k\| \leq \|\mathbf{Z}_{jj}\| \prod_{k=i}^{j-1}\tau_k.$$

The expression for $i > j$ follows analogously using the two lemmas.

Since $\mathbf{A}\mathbf{A}^{-1} = \mathbf{A}\mathbf{Z} = \mathbf{I}$ we have

$$\mathbf{C}_{i-1}\mathbf{Z}_{i-1,i} + \mathbf{A}_i\mathbf{Z}_{ii} + \mathbf{B}_i\mathbf{Z}_{i+1,i} = \mathbf{I}, \quad \text{for } i = 1, \dots, n,$$

where we set $\mathbf{C}_0 = \mathbf{Z}_{0,1} = \mathbf{B}_n = \mathbf{Z}_{n+1,n} = 0$. Using (4.5) and Lemma 4.5,

$$\mathbf{Z}_{i-1,i} = \mathbf{U}_{i-1}\mathbf{V}_i = -\mathbf{L}_{i-1}\mathbf{U}_i\mathbf{V}_i = -\mathbf{L}_{i-1}\mathbf{Z}_{ii},$$
$$\mathbf{Z}_{i+1,i} = \mathbf{Y}_{i+1}\mathbf{X}_i = -\mathbf{M}_{i+1}\mathbf{Y}_i\mathbf{X}_i = -\mathbf{M}_{i+1}\mathbf{Z}_{ii},$$

where we set $\mathbf{U}_0 = \mathbf{L}_0 = \mathbf{Y}_{n+1} = \mathbf{M}_{n+1} = 0$. Combining this with the previous equation yields

$$-\mathbf{C}_{i-1}\mathbf{L}_{i-1}\mathbf{Z}_{ii} - \mathbf{B}_i\mathbf{M}_{i+1}\mathbf{Z}_{ii} + \mathbf{A}_i\mathbf{Z}_{ii} = \mathbf{I}, \quad \text{for } i = 1, \dots, n, \tag{4.22}$$

Taking norms and using again Lemma 4.4 now gives

$$\|\mathbf{I}\| = \|-\mathbf{C}_{i-1}\mathbf{L}_{i-1}\mathbf{Z}_{ii} - \mathbf{B}_i\mathbf{M}_{i+1}\mathbf{Z}_{ii} + \mathbf{A}_i\mathbf{Z}_{ii}\|$$

$$\begin{aligned}&\le (\|\mathbf{C}_{i-1}\|\|\mathbf{L}_{i-1}\| + \|\mathbf{A}_i\| + \|\mathbf{B}_i\|\|\mathbf{M}_{i+1}\|)\|\mathbf{Z}_{ii}\|\\ &\le (\tau_{i-1}\|\mathbf{C}_{i-1}\| + \|\mathbf{A}_i\| + \mu_{i+1}\|\mathbf{B}_i\|)\|\mathbf{Z}_{ii}\|, \quad \text{for } i = 1, \dots, n,\end{aligned}$$

where we set $\tau_0 = \mu_{n+1} = 0$, and which shows the lower bound in (4.21). In order to show the upper bound we write (4.22) as

$$\mathbf{I} - \mathbf{A}_i\mathbf{Z}_{ii} = -(\mathbf{C}_{i-1}\mathbf{L}_{i-1}\mathbf{Z}_{ii} + \mathbf{B}_i\mathbf{M}_{i+1}\mathbf{Z}_{ii}), \quad \text{for } i = 1, \dots, n.$$

This yields

$$\begin{aligned}\|\mathbf{A}_i\mathbf{Z}_{ii}\| - \|\mathbf{I}\| \le \|\mathbf{I} - \mathbf{A}_i\mathbf{Z}_{ii}\| &= \|\mathbf{C}_{i-1}\mathbf{L}_{i-1}\mathbf{Z}_{ii} + \mathbf{B}_i\mathbf{M}_{i+1}\mathbf{Z}_{ii}\|\\ &\le (\tau_{i-1}\|\mathbf{C}_{i-1}\| + \mu_{i+1}\|\mathbf{B}_i\|)\|\mathbf{Z}_{ii}\|.\end{aligned}$$

From $\|\mathbf{Z}_{ii}\| = \|\mathbf{A}_i^{-1}\mathbf{A}_i\mathbf{Z}_{ii}\| \le \|\mathbf{A}_i^{-1}\|\|\mathbf{A}_i\mathbf{Z}_{ii}\|$ we get $\|\mathbf{A}_i\mathbf{Z}_{ii}\| \ge \|\mathbf{Z}_{ii}\|/\|\mathbf{A}_i^{-1}\|$, and combining this with the previous inequality yields

$$(\tau_{i-1}\|\mathbf{C}_{i-1}\| + \mu_{i+1}\|\mathbf{B}_i\|)\,\|\mathbf{Z}_{ii}\| \ge \frac{1}{\|\mathbf{A}_i^{-1}\|}\|\mathbf{Z}_{ii}\| - \|\mathbf{I}\|.$$

When $\|\mathbf{A}_i^{-1}\|^{-1} - \tau_{i-1}\|\mathbf{C}_{i-1}\| - \mu_{i+1}\|\mathbf{B}_i\| > 0$ holds, we get the upper bound in (4.21). □

Note that the positivity assumption on the denominator of the upper bound in (4.21) is indeed necessary. A simple example for which the denominator is equal to zero is given by the matrix $\mathbf{A} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ with 1×1 blocks, which satisfies all assumptions of Lemma 4.3.

Both the off-diagonal bounds (4.19)–(4.20) and the diagonal bounds (4.21) depend on the values τ_i and μ_i, which bound $\|\mathbf{L}_i\|$ and $\|\mathbf{M}_i\|$, respectively. We will now show that by modifying the proof of Lemma 4.4 the bounds can be improved in an iterative fashion. This is analogous to the iterative improvement for the case when the blocks of $\mathbf{A}$ are scalars, which was considered in [62].

We have shown in the inductive proof of Lemma 4.4 that

$$\|\mathbf{T}_i^{-1}\| \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}\|} \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\|\mathbf{L}_{i-1}\|} \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|}.$$

This bound can be improved by making use of Lemma 4.4 itself, i.e.,

$$\|\mathbf{T}_i^{-1}\| \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\mathbf{L}_{i-1}\|} \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\|\mathbf{L}_{i-1}\|} \le \frac{1}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\tau_{i-1}},$$

and this yields

$$\|\mathbf{L}_i\| = \|\mathbf{T}_i^{-1}\mathbf{A}_i^{-1}\mathbf{B}_i\| \le \frac{\|\mathbf{A}_i^{-1}\mathbf{B}_i\|}{1 - \|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\tau_{i-1}}.$$

If we denote the expression on the right hand side by $\tau_{i,2}$, then we obtain a modified version of Lemma 4.4, where $\|\mathbf{L}_i\| \le \tau_{i,2} \le \tau_2 \le 1$. Iteratively we now define, for all

$i = 1, \ldots, n$ and $t = 1, \ldots, n-1$,

$$\tau_{i,t} \equiv \begin{cases} \tau_i & \text{if } t = 1, \\ \tau_{i,t-1} & \text{if } t > i, \\ \frac{\|\mathbf{A}_i^{-1}\mathbf{B}_i\|}{1-\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|\tau_{i-1,t-1}} & \text{else.} \end{cases} \tag{4.23}$$

Analogously we can proceed for the values $\|\mathbf{M}_i\|$, and here we define, for all $i = 1, \ldots, n$ and $t = 1, \ldots, n-1$,

$$\mu_{i,t} \equiv \begin{cases} \mu_i & \text{if } t = 1, \\ \mu_{i,t-1} & \text{if } n-t+1 < i, \\ \frac{\|\mathbf{A}_i^{-1}\mathbf{C}_{i-1}\|}{1-\|\mathbf{A}_i^{-1}\mathbf{B}_i\|\mu_{i+1,t-1}} & \text{else.} \end{cases} \tag{4.24}$$

Using these definitions we can easily prove the following modified version of Theorem 4.6, which refines the bounds (4.19), (4.20) and (4.21) as t increases, and which generalizes [62, Theorems 3.4 and 3.5] from the scalar to the block case.

Theorem 4.7. *If* $\mathbf{A}$ *is a matrix as in Lemma 4.3 with* $\mathbf{A}^{-1} = [\mathbf{Z}_{ij}]$*, then for each* $t = 1, \ldots, n-1$,

$$\|\mathbf{Z}_{ij}\| \le \|\mathbf{Z}_{jj}\| \prod_{k=i}^{j-1} \tau_{k,t}, \quad \textit{for all } i < j, \tag{4.25}$$

$$\|\mathbf{Z}_{ij}\| \le \|\mathbf{Z}_{jj}\| \prod_{k=j+1}^{i} \mu_{k,t}, \quad \textit{for all } i > j \tag{4.26}$$

with $\tau_{k,t}$ *and* $\mu_{k,t}$ *given by* (4.24) *and* (4.23)*. Moreover, for* $i = 1, \ldots, n$,

$$\frac{\|\mathbf{I}\|}{\|\mathbf{A}_i\| + \tau_{i-1,t}\|\mathbf{C}_{i-1}\| + \mu_{i+1,t}\|\mathbf{B}_i\|} \le \|\mathbf{Z}_{ii}\| \le \frac{\|\mathbf{I}\|}{\|\mathbf{A}_i^{-1}\|^{-1} - \tau_{i-1,t}\|\mathbf{C}_{i-1}\| - \mu_{i+1,t}\|\mathbf{B}_i\|},$$

provided that the denominator of the upper bound is larger than zero, and where we set $\mathbf{C}_0 = \mathbf{B}_n = 0$*, and* $\tau_{0,t} = \mu_{n+1,t} = 0$.

Note that the statements of Theorem 4.7 with $t = 1$ are the same as those in Theorem 4.6. By construction, the sequences $\{\tau_{i,t}\}_{t=1}^{n-1}$ and $\{\mu_{i,t}\}_{t=1}^{n-1}$ are decreasing, and hence the bounds (4.25), (4.26) and (4.27) become tighter as t increases. However, since we have used the submultiplicativity property of the matrix norm in the derivation, it is not guaranteed that the bounds in Theorem 4.7 with $t = n-1$ will give the exact norms of the blocks of $\mathbf{A}^{-1}$. This is a difference to the scalar case, where in the last refinement step one obtains the exact inverse; see [62].

Finally, let us define

$$\rho_{1,t} \equiv \max_i \tau_{i,t}, \quad \mu_{2,t} \equiv \max_i \mu_{i,t}, \quad \text{for } t = 1, \ldots, n-1.$$

Then the off-diagonal bounds (4.25) and (4.26) of Theorem 4.7 immediately give the following result about the decay of the norms $\|\mathbf{Z}_{ij}\|$; cf. [62, Corollary 3.7]

Corollary 4.8. *If* $\mathbf{A}$ *is a matrix as in Theorem 4.7, then*

$$\begin{aligned} \|\mathbf{Z}_{ij}\| &\le \rho_{1,t}^{j-i}\|\mathbf{Z}_{jj}\|, \quad && \textit{for all } i<j, \\ \|\mathbf{Z}_{ij}\| &\le \rho_{2,t}^{i-j}\|\mathbf{Z}_{jj}\|, \quad && \textit{for all } i>j, \end{aligned}$$

and for each $t = 1,\dots,n-1$.

4.3. Eigenvalue Inclusion Regions

In this section we generalize a result of Feingold and Varga on eigenvalue inclusion regions of block matrices. We start with the following generalization of [30, Theorem 1]; also cf. [78, Theorem 6.2].

Lemma 4.9. *If a matrix* $\mathbf{A}$ *as in* (4.2) *is row block strictly diagonally dominant, then* $\mathbf{A}$ *is nonsingular.*

Proof. The proof closely follows the proof of [30, Theorem 1]. Suppose that $\mathbf{A}$ is row block strictly diagonally dominant but singular. Then there exists a nonzero block vector $\mathbf{X}$, partitioned conformally with respect to the partition of $\mathbf{A}$ in (4.2), such that

$$\mathbf{A}\begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix} = 0.$$

This is equivalent to

$$\mathbf{A}_{ii}\mathbf{X}_i + \sum_{\substack{j=1\\ j\neq i}}^{n} \mathbf{A}_{ij}\mathbf{X}_j = 0, \quad 1 \le i \le n,$$

and, since the diagonal blocks $\mathbf{A}_{ii}$ are nonsingular,

$$\mathbf{X}_i = -\sum_{\substack{j=1\\ j\neq i}}^{n} \mathbf{A}_{ii}^{-1}\mathbf{A}_{ij}\mathbf{X}_j, \quad 1 \le i \le n,$$

Without loss of generality we can assume that $\mathbf{X}$ is normalized such that $\|\mathbf{X}_i\| \le 1$ for all $1 \le i \le n$, with equality for some $i = r$. For this index we obtain

$$1 = \|\mathbf{X}_r\| = \Bigg\| \sum_{\substack{j=1\\ j\neq i}}^{n} \mathbf{A}_{rr}^{-1}\mathbf{A}_{rj}\mathbf{X}_j \Bigg\| \le \sum_{\substack{j=1\\ j\neq r}}^{n} \|\mathbf{A}_{rr}^{-1}\mathbf{A}_{rj}\| \|\mathbf{X}_j\| \le \sum_{\substack{j=1\\ j\neq r}}^{n} \|\mathbf{A}_{rr}^{-1}\mathbf{A}_{rj}\|,$$

which contradicts the assumption that $\mathbf{A}$ is row block strictly diagonally dominant. Thus, $\mathbf{A}$ must be nonsingular. □

If λ is an eigenvalue of $\mathbf{A}$, then $\mathbf{A}-\lambda\mathbf{I}$ is singular, and hence $\mathbf{A}-\lambda\mathbf{I}$ cannot be block strictly diagonally dominant. This immediately gives the following result, which generalizes [30, Theorem 2]; also cf. [78, Theorem 6.3].

Corollary 4.10. *If a matrix* $\mathbf{A}$ *is as in* (4.2)*, and* λ *is an eigenvalue of* $\mathbf{A}$*, then there exists at least one* $i \in \{1, \dots, n\}$ *with*

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \|(\mathbf{A}_{ii} - \lambda \mathbf{I})^{-1} \mathbf{A}_{ij}\| \geq 1. \tag{4.27}$$

If all the blocks $\mathbf{A}_{ij}$ of $\mathbf{A}$ are of size 1×1, and $\|\mathbf{A}_{ij}\| = |\mathbf{A}_{ij}|$, then this result reduces to the classical Gershgorin Circle Theorem.

Corollary 4.10 shows that each eigenvalue λ of $\mathbf{A}$ must be contained in the union of the sets

$$G_i^{\text{new}} = \Big\{ z \in \mathbb{C} : \sum_{\substack{j=1 \\ j \neq i}}^{n} \|(\mathbf{A}_{ii} - z\mathbf{I})^{-1} \mathbf{A}_{ij}\| \geq 1 \Big\},$$

for $i = 1, \dots, n$. Due to the submultiplicativity property of the matrix norm, the sets G_i^{new} are potentially smaller than the ones proposed in [30, Definition 3],

$$G_i^{\text{FV}} = \Big\{ z \in \mathbb{C} : \sum_{\substack{j=1 \\ j \neq i}}^{n} \|(\mathbf{A}_{ii} - z\mathbf{I})^{-1}\| \|\mathbf{A}_{ij}\| \geq 1 \Big\},$$

i.e., we have $G_i^{\text{new}} \subseteq G_i^{\text{FV}}$. We will illustrate this fact with numerical examples.

4.4. Numerical Illustrations

In the following we provide a set of numerical examples that illustrate the main results presented in this chapter, mainly, the bounds in Theorem 4.7 and the new eigenvalue inclusion sets, G_i^{new}, consequence of Corollary 4.10.

We begin by presenting a set of numerical illustrations of the bounds in Theorem 4.7 for different values of t. We consider different matrices $\mathbf{A} = [\mathbf{A}_{ij}]$ which are row block diagonally dominant, and we compute the corresponding matrices $\mathbf{Z} = [\mathbf{Z}_{ij}]$ using the recurrences stated in Theorem 4.2. In all experiments we use the matrix 2-norm, $\| \cdot \|_2$. For each given pair i, j, we denote by u_{ij} the value of a computed upper bound (i.e., (4.25), (4.26) or (4.27)) on the value $\|\mathbf{Z}_{ij}\|_2$ and for each i we denote by l_i the value of the computed lower bound for the corresponding diagonal entry (i.e., (4.27)). Then the relative errors in the upper and lower bounds are given by

$$\mathbf{E}_{ij}^{\text{u}} = \frac{u_{ij} - \|\mathbf{Z}_{ij}\|_2}{u_{ij}} \quad \text{and} \quad \mathbf{E}_i^{\text{l}} = \frac{\|\mathbf{Z}_{ii}\|_2 - l_i}{\|\mathbf{Z}_{ii}\|_2}, \tag{4.28}$$

respectively. (Thus, both $\mathbf{E}_{ij}^{\text{u}}$ and $\mathbf{E}_i^{\text{l}}$ are between 0 and 1.)

Example 4.11. *We start with the symmetric block Toeplitz matrix*

$$\mathbf{A} = \mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T} \in \mathbb{R}^{81 \times 81}, \tag{4.29}$$

where $\mathbf{T} = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{9 \times 9}$*, i.e.,* $\mathbf{A}$ *is of the form* (4.1) *with* $\mathbf{A}_i = \text{tridiag}(-1, 4, -1)$*, and* $\mathbf{B}_i = \mathbf{C}_i = \text{diag}(-1)$ *for all* i*. We have* $\kappa_2(\mathbf{A}_1) = 58.4787$*, i.e., the matrix* $\mathbf{A}_1$ *is quite well conditioned. For the computed matrix* $\mathbf{Z} = [\mathbf{Z}_{ij}]$ *we obtain*

$\|\mathbf{ZA} - \mathbf{I}\|_2 = 2.7963 \times 10^{-10}$, *suggesting that* $\mathbf{Z}$ *is a reasonably accurate approximation of the exact inverse* $\mathbf{A}^{-1}$.

In the top row of Figure 4.1 *we show the relative errors* $\mathbf{E}^{\mathrm{u}}_{ij}$ *for the refinement step* $t = 1$ *(no refinement) and* $t = 8$ *(maximal refinement). We observe that the upper bounds are quite tight already for* $t = 1$, *and that for* $t = 8$ *the maximal relative error is on the order* 10^{-13}, *i.e., the value of the upper bound is almost exact. In the bottom row of* Figure 4.1 *we show the values* $\|\mathbf{Z}_{ii}\|_2$ *for* $i = 1, \dots, 9$, *and the corresponding upper and lower bounds* (4.27) *for the refinement steps* $t = 1$ *and* $t = 8$.

We observe that while the upper bounds on $\|\mathbf{Z}_{ii}\|_2$ *for* $t = 8$ *almost exactly match the exact values, the lower bounds do not improve by the iterative refinement. The maximal error of the lower bounds for the diagonal block entries of* $\mathbf{Z}$ *in the maximal refinement step is on the order* 10^{-1}. *The maximal relative errors in the upper and lower bounds and all refinement steps are shown in the following table:*

t	1	2	3	4	5	6	7	8
$\max_{ij} \mathbf{E}^{u}_{ij}$	0.84478	0.63381	0.39537	0.20899	0.09596	0.03780	0.01109	7.141×10^{-13}
$\max_{i} \mathbf{E}^{l}_{i}$	0.91039	0.90877	0.90765	0.90529	0.90529	0.90529	0.90529	0.90529

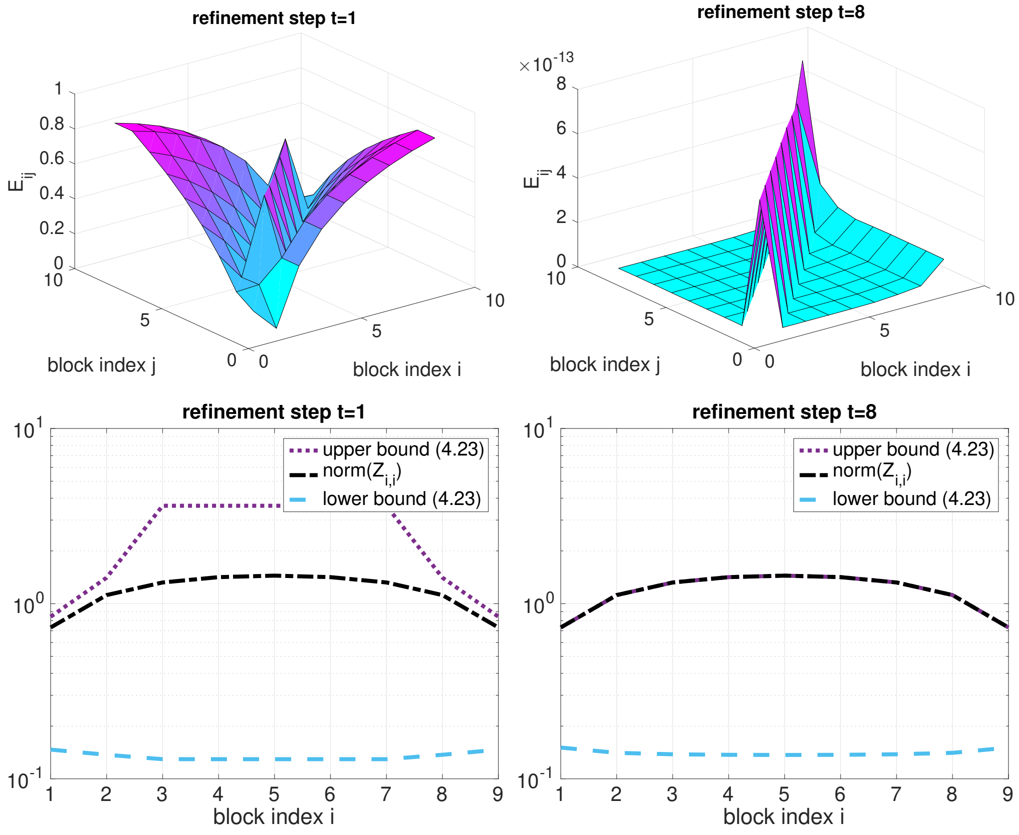


Figure 4.1.: Relative errors $\mathbf{E}^{\mathrm{u}}_{ij}$ (top row), upper and lower bounds on $\|\mathbf{Z}_{ii}\|_2$ (bottom row) for the matrix A of Example 4.11.

Example 4.12. *Let* $\mathbf{A}$ *be the nonsymmetric block Toeplitz matrix of the form* (4.29) *with* $\mathbf{T} = \mathrm{tridiag}(-110, 209.999, -99.999) \in \mathbb{R}^{9\times 9}$*, i.e.,* $\mathbf{A}$ *again takes the form* (4.1) *with* $\mathbf{A}_i = \mathrm{tridiag}(-110, 419.999, -99.999)$*,* $\mathbf{B}_i = \mathrm{diag}(-110)$*, and* $\mathbf{C}_i = \mathrm{diag}(-99.999)$*. The condition number in this case is* $\kappa_2(\mathbf{A}) = 57.5725$*, and for the computed matrix* $\mathbf{Z}$ *we obtain* $\|\mathbf{ZA} - \mathbf{I}\|_2 = 1.5151 \times 10^{-10}$*.*

The top row of Figure 4.2 *shows the relative errors for the refinement steps* $t = 1$ *and* $t = 8$*. We observe that for this nonsymmetric example the upper bounds are not as accurate as those given in the symmetric case, producing a maximal relative error at refinement step* $t = 8$ *on the order* 10^{-3}*. The bottom row of* Figure 4.2 *shows the upper and lower bounds* (4.27) *as well as the values* $\|\mathbf{Z}_{ii}\|_2$ *for* $i = 1, \ldots, 9$*, and refinement steps* $t = 1$ *and* $t = 8$*. Again we can observe that while we obtain a reasonable approximation in the upper bounds on* $\|\mathbf{Z}_{ii}\|_2$ *for* $t = 8$*, the lower bounds almost do not improve by the iterative refinement process. The maximal relative errors in the upper and lower bounds and all refinement steps is shown in the following table:*

t	1	2	3	4	5	6	7	8
$\max_{ij} \mathbf{E}^u_{ij}$	0.88856	0.70640	0.46700	0.25859	0.12442	0.05378	0.02140	0.00824
$\max_i \mathbf{E}^l_i$	0.90934	0.90768	0.90652	0.90411	0.90411	0.90411	0.90411	0.90411

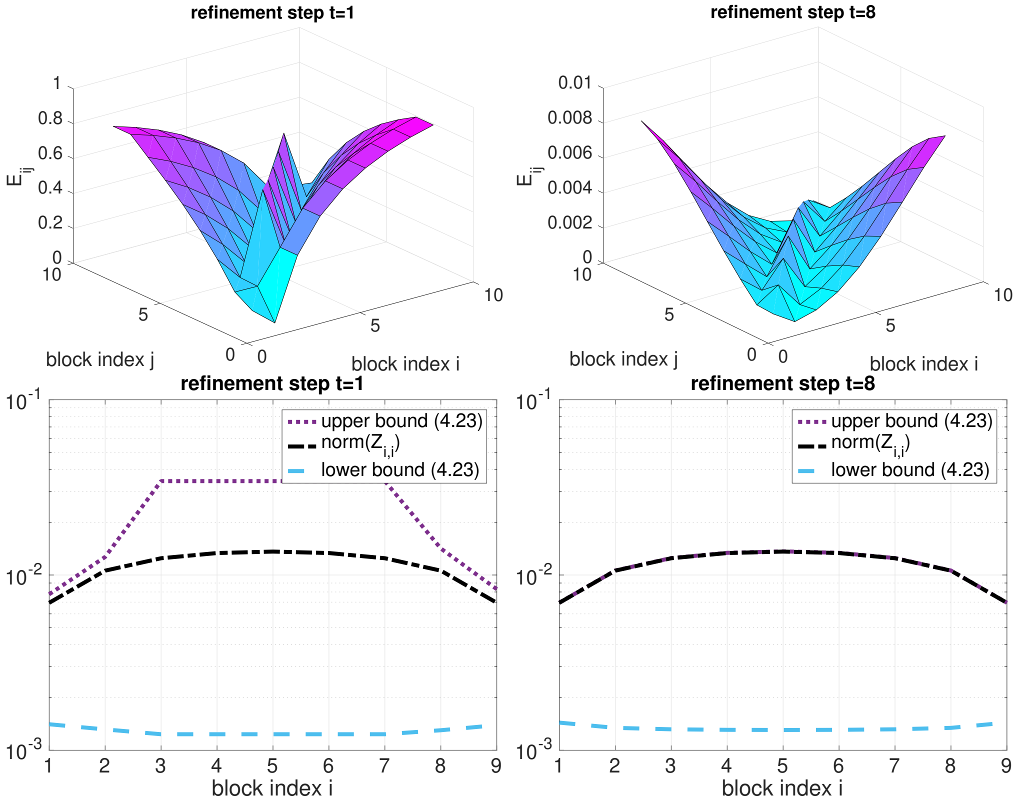


Figure 4.2.: Relative errors $\mathbf{E}^{\mathrm{u}}_{ij}$ (top row), and upper and lower bounds on $\|\mathbf{Z}_{ii}\|_2$ (bottom row) for the matrix $\mathbf{A}$ of Example 4.12.

Example 4.13. *We now consider the nonsymmetric block tridiagonal matrix*

$$\mathbf{A} = (\mathbf{R} \otimes \mathbf{I})(\mathbf{T} \otimes \mathbf{I} + \mathbf{I} \otimes \mathbf{T}) \in \mathbb{R}^{81\times 81},$$

where $\mathbf{T}$ *is given as in* Example 4.11, *and* $\mathbf{R} \in \mathbb{R}^{9\times 9}$ *is a random diagonal matrix with nonzero integer entries between* 0 *and* 10 *and constructed in* `MATLAB` *with the command* $\mathbf{R}$`= diag(ceil(10*rand(9,1)))`. *Thus,* $\mathbf{A}$ *is of the form* (4.1) *with random tridiagonal Toeplitz matrices* $\mathbf{A}_i$*, and random constant diagonal matrices* $\mathbf{B}_i$ *and* $\mathbf{C}_i$ *for all* i*. For this matrix we have* $\kappa_2(\mathbf{A}) = 489.7595$*, and the computed matrix* $\mathbf{Z}$ *yields* $\|\mathbf{ZA} - \mathbf{I}\|_2 = 2.8328 \times 10^{-10}$*. The relative errors in the bounds are shown in* Figure 4.3 *and in the following table:*

t	1	2	3	4	5	6	7	8
$\max_{ij} \mathbf{E}^u_{ij}$	0.84477	0.63381	0.39537	0.20898	0.09595	0.03780	0.01109	9.739×10^{-13}
$\max_i \mathbf{E}^l_i$	0.91039	0.90877	0.90765	0.90529	0.90529	0.90529	0.90529	0.90529

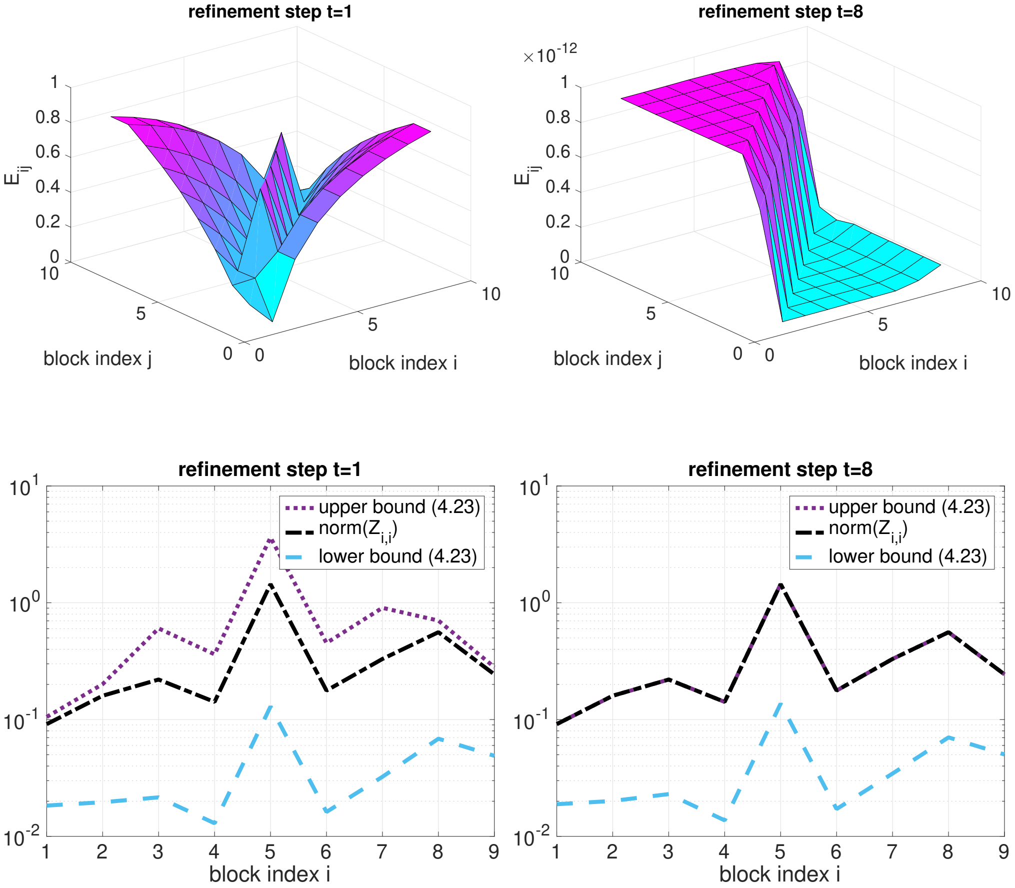


Figure 4.3.: Relative errors $\mathbf{E}^{\mathrm{u}}_{ij}$ (top row), and upper and lower bounds on $\|\mathbf{Z}_{ii}\|_2$ (bottom row) for the matrix $\mathbf{A}$ of Example 4.13.

Example 4.14. *Finally, we consider the nonsymmetric block tridiagonal matrix*

$$\mathbf{A} = (\mathbf{R} \otimes \mathbf{I})\,\mathrm{tridiag}(\mathrm{tridiag}(-0.01, -2, 1), \mathrm{tridiag}(-2, 10, -2), \mathrm{tridiag}(-0.01, -2, 1)),$$

with $\mathbf{A} \in \mathbb{R}^{81\times 81}$*, and where* $\mathbf{R} \in \mathbb{R}^{9\times 9}$ *is a random diagonal matrix constructed as in* Example 4.13*. In this case* $\mathbf{A}$ *takes the form* (4.1) *with* $\mathbf{A}_i$*,* $\mathbf{B}_i$ *and* $\mathbf{C}_i$ *random tridiagonal Toeplitz matrices with integer entries for all* i*. For this matrix we have* $\kappa_2(\mathbf{A}) = 58.478$*, and* $\|\mathbf{ZA} - \mathbf{I}\|_2 = 2.7962 \times 10^{-10}$*. The relative errors in the bounds are shown in* Figure 4.4 *and the following table:*

t	1	2	3	4	5	6	7	8
$\max_{ij} \mathbf{E}^u_{ij}$	0.84477	0.63381	0.39537	0.20898	0.09595	0.03780	0.01109	7.142×10^{-13}
$\max_i \mathbf{E}^l_i$	0.91039	0.90877	0.90765	0.90529	0.90529	0.90529	0.90529	0.90529

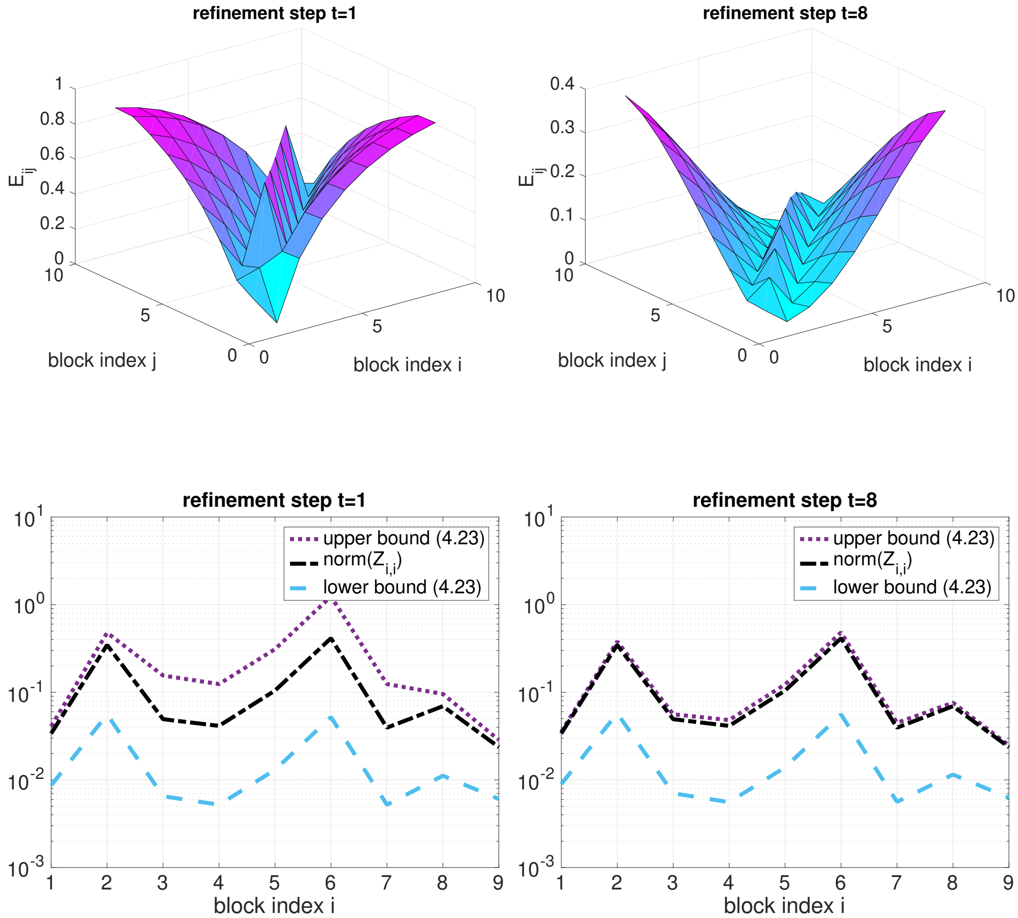


Figure 4.4.: Relative errors $\mathbf{E}^{\mathrm{u}}_{ij}$ (top row), and upper and lower bounds on $\|\mathbf{Z}_{ii}\|_2$ (bottom row) for the matrix $\mathbf{A}$ of Example 4.14.

We continue by providing a set of numerical illustrations of of the newly proposed eigenvalue inclusion sets, G_i^{new}, which are a consequence of Corollary 4.10. We consider different matrices $\mathbf{A} = [\mathbf{A}_{ij}]$, and we compute the boundaries of the sets G_i^{new} and G_i^{FV} for all $i \leq n$, i.e., the curves for $z \in \mathbb{C}$ where

$$\sum_{\substack{j=1 \\ j \neq i}}^{n} \|(\mathbf{A}_{ii} - z\mathbf{I})^{-1}\mathbf{A}_{ij}\| = 1, \text{ and } \sum_{\substack{j=1 \\ j \neq i}}^{n} \|(\mathbf{A}_{ii} - z\mathbf{I})^{-1}\|\|\mathbf{A}_{ij}\| = 1, \quad i,j \in \{1,\dots,n\}, \quad i \neq j,$$

respectively.

Example 4.15. *We first consider the symmetric matrix*

$$\mathbf{A} = \left[\begin{array}{rr|rr} 4 & -2 & -1 & 1 \\ -2 & 4 & 0 & -1 \\ \hline -1 & 0 & 4 & -2 \\ 1 & -1 & -2 & 4 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right], \tag{4.30}$$

which has the eigenvalues 1.4586, 2.3820, 4.6180, *and* 7.5414 *(computed in* `MATLAB` *and rounded to five significant digits).* Figure 4.5 *shows the boundaries of the corresponding sets G_i^{new} and G_i^{FV} for $i = 1, 2$, i.e., the curves for $z \in \mathbb{C}$ where*

$$\|(\mathbf{A}_{ii} - z\mathbf{I})^{-1}\mathbf{A}_{ij}\| = 1, \textit{ and } \|(\mathbf{A}_{ii} - z\mathbf{I})^{-1}\|\|\mathbf{A}_{ij}\| = 1, \quad i,j \in \{1,2\}, \quad i \neq j,$$

respectively. Clearly, the sets G_i^{new} give tighter inclusion regions for the eigenvalues than the sets G_i^{FV} as well as the usual Gershgorin circles for the matrix $\mathbf{A}$*, which are given by the two circles centered at $z = 4$ of radius 3 and 4.*

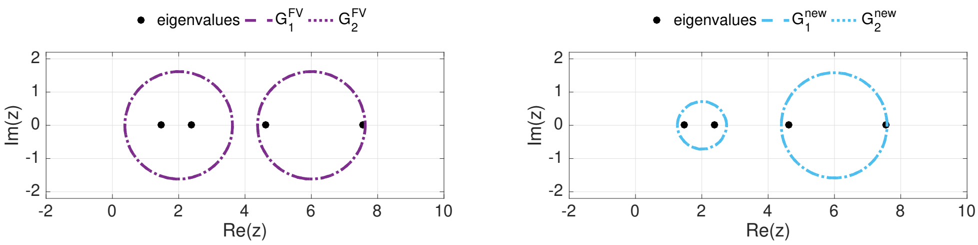


Figure 4.5.: Eigenvalue inclusion regions obtained from the sets G_i^{new} and G_i^{FV} for the matrix (4.30) of Example 4.15.

We next consider the nonsymmetric matrix

$$\mathbf{A} = \left[\begin{array}{rr|rr} 4 & -2 & -0.5 & 0.5 \\ -2 & 5 & -1.4 & -0.5 \\ \hline -0.5 & 0 & 4 & -2 \\ 0.5 & -0.5 & -2 & 4 \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A}_{11} & \mathbf{A}_{12} \\ \hline \mathbf{A}_{21} & \mathbf{A}_{22} \end{array}\right], \tag{4.31}$$

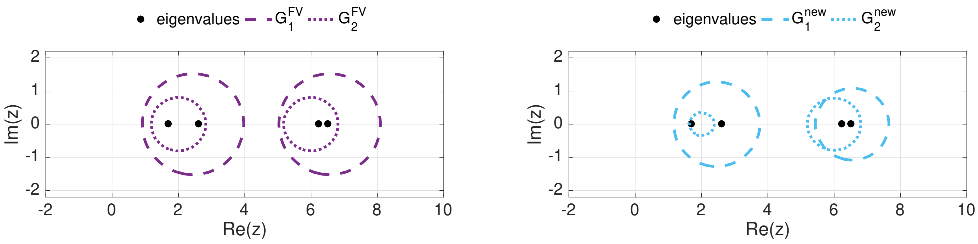


Figure 4.6.: Eigenvalue inclusion regions obtained from the sets G_i^{new} and G_i^{FV} for the matrix (4.31) of Example 4.15.

which has the eigenvalues 1.6851, 2.5959, 6.2263, *and* 6.4927. *As shown in* Figure 4.6, *the sets G_i^{new} again give tighter inclusion regions than the sets G_i^{FV} as well as the usual Gershgorin circles.*

We now present an example of the inclusion sets for the eigenvalues of a matrix

$$\mathbf{A} = \begin{bmatrix} \widehat{\mathbf{A}}_H & \mathbf{e}_m \otimes \mathbf{B}_H & 0 \\ \mathbf{e}_m^T \otimes \mathbf{C} & \widehat{\mathbf{A}} & \mathbf{e}_1^T \otimes \mathbf{B} \\ 0 & \mathbf{e}_1 \otimes \mathbf{C}_h & \widehat{\mathbf{A}}_h \end{bmatrix} \in \mathbb{R}^{N(2m+1)\times N(2m+1)}, \tag{4.32}$$

obtained, for example, from a Shishkin mesh discretization of the 2D convection-diffusion problems of type (5.2) studied in Chapter 5.

Example 4.16. *Consider the nonsymmetric block tridiagonal matrix*

$$\mathbf{A} = \begin{bmatrix} \widehat{\mathbf{A}} & \mathbf{B} & 0 \\ \mathbf{C} & \widehat{\mathbf{A}} & \mathbf{B} \\ 0 & \mathbf{C} & \widehat{\mathbf{A}} \end{bmatrix} \in \mathbb{R}^{15\times 15}, \tag{4.33}$$

where $\widehat{\mathbf{A}} = \mathrm{tridiag}(-36, 108, -36) \in \mathbb{R}^{5\times 5}$, $\mathbf{B} = \mathrm{diag}(-16) \in \mathbb{R}^{5\times 5}$ *and* $\mathbf{C} = \mathrm{diag}(-20) \in \mathbb{R}^{5\times 5}$, *i.e., we have a matrix of type* (4.32) *with* $N = 5$, $m = 1$ *and* $\widehat{\mathbf{A}}_H = \widehat{\mathbf{A}}_h = \widehat{\mathbf{A}}$, $\mathbf{B}_H = \mathbf{B}$, *and* $\mathbf{C}_h = \mathbf{C}$. Figure 4.7 *shows the* 15 *eigenvalues computed in* `MATLAB` *and rounded to five significant digits as well as the boundaries of the corresponding sets G_i^{new}, and G_i^{FV} for $i = 1, 2, 3$.*

The figure shows that sets G_i^{new} give the same inclusion regions for the eigenvalues than the sets G_i^{FV} (both inclusion regions are tighter than the usual Gershgorin circles); this is to be expected since the off-diagonal blocks are multiples of the identity matrix and the sets G_i^{new} reduce to the sets G_i^{FV}. However, by changing the structure of the off-diagonal blocks (we make the upper right corner of each upper off-diagonal block and the lower left entry of each lower off-diagonal block equal to 10*), the eigenvalues do not shift much, nevertheless, the sets are no longer equal and once again the new sets G_i^{new} present much tighter inclusion regions than the sets G_i^{FV} - see the right side of* Figure 4.7

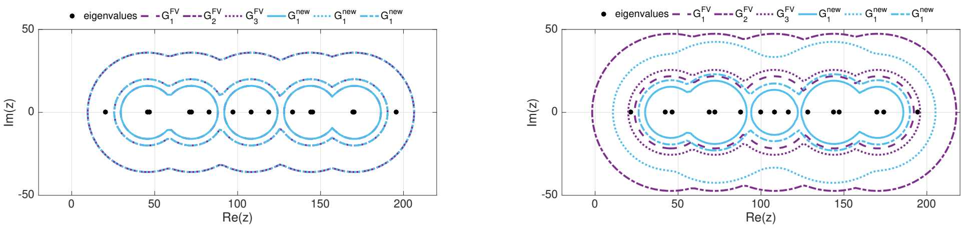


Figure 4.7.: Eigenvalue inclusion regions obtained from the sets G_i^{new} and G_i^{FV} for the matrix (4.33) of Example 4.16.

Example 4.17. *Consider the nonsymmetric block tridiagonal matrix* (4.33) *from* Example 4.16 *where* $\widehat{\mathbf{A}} = \text{tridiag}(-36, 108, -36) \in \mathbb{R}^{5\times5}$*, but now we choose the off-diagonal blocks* $\mathbf{B}$ *and* $\mathbf{C}$ *as* 5×5 *diagonal matrices with random positive integer entries between* 0 *and* 9 *using the* MATLAB *command* `diag(floor(10*rand(5,1)))`. Figure 4.8 *once again shows the* 15 *eigenvalues as well as the boundaries of the corresponding sets* G_i^{new}*, and* G_i^{FV} *for* $i = 1, 2, 3$*.*

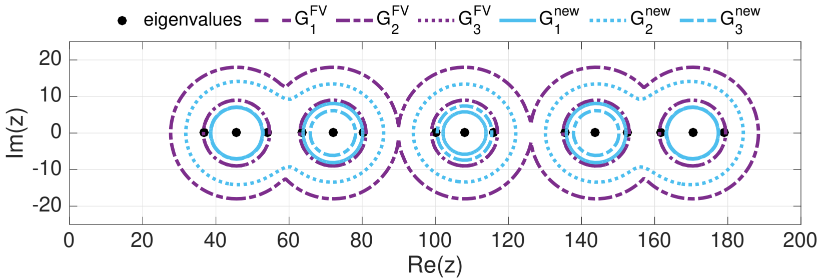


Figure 4.8.: Eigenvalue inclusion regions obtained from the sets G_i^{new} and G_i^{FV} for the matrix (4.33) of Example 4.17.

Figure 4.8 *once again shows that sets* G_i^{new} *present much tighter incusion regions than the sets* G_i^{FV}*. In the context of convection diffusion equations, a coefficient matrix with the structure given by this example might correspond to having variable coefficients in equation* (5.2) *instead of constant ones, like it is the case for the matrix in Example 4.16.*

Example 4.18. *Consider the nonsymmetric block tridiagonal matrix arising from the Shishkin mesh discretization of the convection-diffusion model problem* (5.2) *with* $\epsilon = 10^{-4}$*, using* 16 *intervals in the x-direction and* 8 *intervals in the y-direction (see next chapter) yielding a matrix of type* (4.32) *with* $N = 15$ *and* $m = 7$*.* Figure 4.9 *shows the eigenvalues of the matrix as well as the boundaries of the corresponding sets* G_i^{new}*, and* G_i^{FV} *for* $i = 1, \dots, 105$*. Just like it is the case for* Example 4.16, *the sets* G_i^{new} *present the same inclusion regions than the sets* G_i^{FV}*. As we have discussed before, this is due to the constant-coefficient nature of the problem. Even though the eigenvalues appear clustered together in 4 tight clusters, the difference in magnitude between the eigenvalues, caused by the convection-domitated characteristic of the problem, makes the inclusion regions cover a much larger part of the complex*

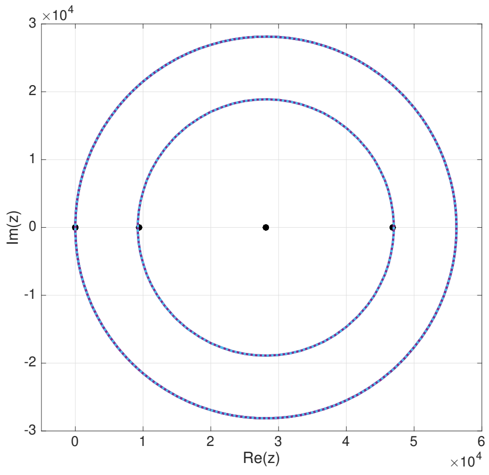


Figure 4.9.: Eigenvalue inclusion regions obtained from the sets G_i^{new} and G_i^{FV} for the matrix (4.33) of Example 4.18.

plane than expected (notice the scale in Figure 4.9*). Moreover, the regions grow as the problem becomes more convection dominated, making the inclusion sets not very informative in the case of real world problems. It is important to note, however, that a different subdivision of the blocks of* (4.32) *might lead to tighter inclusion regions for these type of problems, a task that remains to be explored.*

To complete this chapter, we present the definition of *column* block diagonal dominance of matrices in the following Appendix.

4.A. Column Block Diagonal Dominance of Matrices

According to Definition 4.1 in Section 4.2, the property of *column* block diagonal dominance of matrices is defined as follows:

Definition 4.19. *Consider a matrix of the form*

$$\mathbf{A} = [\mathbf{A}_{ij}] \quad \textit{with blocks } \mathbf{A}_{ij} \in \mathbb{C}^{m\times m} \textit{ for } i,j = 1,\dots,n. \tag{4.34}$$

The matrix $\mathbf{A}$ *is called* column block diagonally dominant *(with respect to the matrix norm* $\|\cdot\|$*) when the diagonal blocks* $\mathbf{A}_{jj}$ *are nonsingular, and*

$$\sum_{\substack{i=1\\ i\neq j}}^{n} \|\mathbf{A}_{ij}\mathbf{A}_{jj}^{-1}\| \le 1, \quad \textit{for } j = 1,\dots,n. \tag{4.35}$$

If strict inequality holds in (4.35) *then* $\mathbf{A}$ *is called* column block strictly diagonally dominant *(with respect to the matrix norm* $\|\cdot\|$*).*

Now, restricting our attention to block tridiagonal matrices of the form (4.1) and following the notation of that chapter, in order to obtain analogous bounds for the norms of the inverses of a column block diagonally dominant matrix we fist need to set $\mathbf{B}_0 = \mathbf{C}_n = 0$, and define the *new* quantities

$$\tilde{\tau}_i \equiv \frac{\|\mathbf{C}_i\mathbf{A}_i^{-1}\|}{1-\|\mathbf{B}_{i-1}\mathbf{A}_i^{-1}\|}, \quad \text{for } i=1,\dots,n,$$

$$\tilde{\mu}_i \equiv \frac{\|\mathbf{B}_{i-1}\mathbf{A}_i^{-1}\|}{1-\|\mathbf{C}_i\mathbf{A}_i^{-1}\|}, \quad \text{for } i=1,\dots,n.$$

The column block diagonal dominance of $\mathbf{A}$ then implies that $0 \le \tilde{\tau}_i \le 1$ and $0 \le \tilde{\mu}_i \le 1$. Using these quantities we obtain the following result.

Theorem 4.20. *Let A be as in* (4.1) *and suppose that $\mathbf{A}_i^{-1}$ as well as $\mathbf{B}_i^{-1}$ and $\mathbf{C}_i^{-1}$ for $i=1,\dots,n-1$ exist. Suppose in addition that $\mathbf{A}$ is column block diagonally dominant, and that*

$$\|\mathbf{C}_1\mathbf{A}_1^{-1}\| < 1 \quad \textit{and} \quad \|\mathbf{B}_{n-1}\mathbf{A}_n^{-1}\| < 1. \tag{4.36}$$

Then $\mathbf{A}^{-1} = [\mathbf{Z}_{ij}]$ with

$$\|\mathbf{Z}_{ij}\| \le \|\mathbf{Z}_{ii}\| \prod_{k=i+1}^{j} \tilde{\mu}_k, \quad \textit{for all } i<j, \tag{4.37}$$

$$\|\mathbf{Z}_{ij}\| \le \|\mathbf{Z}_{ii}\| \prod_{k=j}^{i-1} \tilde{\tau}_k, \quad \textit{for all } i>j, \tag{4.38}$$

with μ_k and τ_k given by (4.16) *and* (4.15). *Moreover, for $i=1,\dots,n$,*

$$\frac{\|\mathbf{I}\|}{\|\mathbf{A}_i\| + \tilde{\tau}_{i-1}\|\mathbf{B}_{i-1}\| + \tilde{\mu}_{i+1}\|\mathbf{C}_i\|} \le \|\mathbf{Z}_{ii}\| \le \frac{\|\mathbf{I}\|}{\|\mathbf{A}_i^{-1}\|^{-1} - \tilde{\tau}_{i-1}\|\mathbf{B}_{i-1}\| - \tilde{\mu}_{i+1}\|\mathbf{C}_i\|}, \tag{4.39}$$

provided that the denominator of the upper bound is larger than zero, and where we set $\mathbf{B}_0 = \mathbf{C}_n = 0$, and $\tilde{\tau}_0 = \tilde{\mu}_{n+1} = 0$.

Proof. The proof of this theorem is completely analogous to the one of Theorem 4.6 for row block diagonally dominant matrices when the necessary adaptations are made, i.e., by performing the following changes:

- Formulate Lemma 4.3 for the matrices $\mathbf{X}_i$ and $\mathbf{V}_i$, showing that the sequence $\{\|\mathbf{X}_i\|\}_{i=1}^n$ is strictly increasing while the sequence $\{\|\mathbf{V}_i\|\}_{i=1}^n$ is strictly decreasing.
- Formulate Lemma 4.4 for the matrices $\tilde{\mathbf{L}}_1 = \tilde{\mathbf{T}}_1 = \mathbf{C}_1\mathbf{A}_1^{-1}$, $\tilde{\mathbf{T}}_2 = \mathbf{I} - \tilde{\mathbf{T}}_1\mathbf{B}_1\mathbf{A}_2^{-1}$, $\tilde{\mathbf{L}}_i = \mathbf{C}_i\mathbf{A}_i^{-1}\tilde{\mathbf{T}}_i^{-1}$, and $\tilde{\mathbf{T}}_i = \mathbf{I} - \tilde{\mathbf{L}}_{i-1}\mathbf{B}_{i-1}\mathbf{A}_i^{-1}$. Analogously for $\tilde{\mathbf{M}}_i$ and $\tilde{\mathbf{W}}_i$, etc.
- Formulate Lemma 4.5 for the matrices $\mathbf{X}_i$ and $\mathbf{V}_i$, in particular showing that

$$\mathbf{X}_i = -\mathbf{X}_{i+1}\tilde{\mathbf{L}}_i, \quad \text{and} \quad \mathbf{V}_i = -\mathbf{V}_{i-1}\tilde{\mathbf{M}}_i.$$

- In the proof of Theorem 4.6 use the the aforementioned results and use equation $\mathbf{ZA} = \mathbf{I}$ instead of $\mathbf{AZ} = \mathbf{I}$.

Following these changes and proceeding analogously to the proof of Theorem 4.6 yields the desired result. □

5. Convergence of the Multiplicative Schwarz Method for Shihskin Mesh Discretizations of Two-dimensional Convection-Diffusion Problems

Parts of this chapter are expected to be published in:

[24] C. Echeverría, J. Liesen, and P. Tichý, **Analysis of the multiplicative Schwarz method for matrices with a special block structure.** *[Submitted].*

5.1. Introduction

We analyze the convergence behavior of the multiplicative Schwarz method for solving linear algebraic systems of the form

$$\mathbf{Au} = \mathbf{f}, \tag{5.1}$$

where the coefficient matrix $\mathbf{A}$ is obtained from the upwind finite difference discretization of the two-dimensional constant coefficient convection-diffusion equation posed on a domain Ω with Dirichlet boundary conditions

$$\begin{cases} -\epsilon\left(\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2}\right) + \omega_x \frac{\partial u(x,y)}{\partial x} + \omega_y \frac{\partial u(x,y)}{\partial y} + \beta u(x,y) = f(x,y), & \text{in } \Omega \\ \qquad u(x,y) = g(x,y), \quad \text{on } \partial\Omega. \end{cases} \tag{5.2}$$

We assume that the domain of definition of the BVP is the unit square, i.e., $\Omega = (0,1) \times (0,1)$ and further assume that the parameters of the problem are chosen such that the the problem is *convection dominated*, i.e., that $\epsilon \ll \|\boldsymbol{\omega}\|$, and that the solution, $u(x,y)$, presents *one boundary layer* near $y = 1$ In particular we assume that the components of the velocity field fulfill $\boldsymbol{\omega} = [0, \omega_y]^T$ with $\omega_y > 0$, and that the scalar reaction parameter, β, is nonnegative, i.e., $\beta \geq 0$.

In order to obtain a satisfactory approximation to the solution of (5.2), we discretize Ω using a Shishkin mesh that is refined inside the layer; a very similar approach to the one used in the one-dimensional case (see Chapter 3). The mesh is constructed by using a uniform mesh in the x-direction and a one-dimensional Shishkin mesh in the y-direction. This technique has been described in detail in Chapter 2, for external sources see the articles [74, § 5] and [47], as well as the book [59]. After the discretization process, the coefficient matrix exhibits the general structure:

$$\mathbf{A} = \begin{bmatrix} \widehat{\mathbf{A}}_H & \mathbf{e}_m \otimes \mathbf{B}_H & 0 \\ \mathbf{e}_m^T \otimes \mathbf{C} & \widehat{\mathbf{A}} & \mathbf{e}_1^T \otimes \mathbf{B} \\ 0 & \mathbf{e}_1 \otimes \mathbf{C}_h & \widehat{\mathbf{A}}_h \end{bmatrix} \in \mathbb{R}^{N(2m+1)\times N(2m+1)}, \tag{5.3}$$

with the blocks $\widehat{\mathbf{A}}_H, \widehat{\mathbf{A}}_h \in \mathbb{R}^{Nm\times Nm}$, $\widehat{\mathbf{A}}, \mathbf{B}, \mathbf{C}, \mathbf{B}_H, \mathbf{C}_h \in \mathbb{R}^{N\times N}$, and the canonical basis vectors $\mathbf{e}_1, \mathbf{e}_m \in \mathbb{R}^m$. We will think of $\widehat{\mathbf{A}}_H, \widehat{\mathbf{A}}_h \in \mathbb{R}^{Nm\times Nm}$ as matrices consisting of m blocks of size $N \times N$.

It is important to note that a structure such as the one given by (5.3) is *not* exclusive to the discretization of convection-diffusion problems of type (5.2). The coefficient matrices of linear algebraic systems with the structure (5.3) arise naturally when a general second-order partial differential equation is posed and discretized inside a domain Ω that is divided by one interface boundary into two local subdomains, Ω_1 and Ω_2 such as the one shown in Figure 2.5. In this context the first m block rows in the matrix $\mathbf{A}$ correspond to the unknowns in the domain Ω_1, the last m block rows correspond to the unknowns in the domain Ω_2, and the middle block row corresponds to the unknowns in the interface boundary. The underlying assumption here is that in each of the two domains we have the same number of unknowns. This assumption is made for simplicity of the following exposition. Extensions to other block sizes are certainly possible, but would require even more technicalities.

In this chapter, after deriving general expressions for the norms of the multiplicative Schwarz iteration matrices for systems of the form (5.1)–(5.3), we derive quantitative error bounds only for the case when the blocks $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ of $\mathbf{A}$ are block tridiagonal. We point out that the model problem studied in Chapter 3 is of the form (5.1)–(5.3) with $N = 1$. The transition to the higher dimensional cases is reflected in the block structure exhibited by the coefficient matrices. While results that exploit the classical property of diagonal dominance of tridiagonal matrices are the main tools used in Chapter 3 to obtain quantitative convergence results, the derivation of error bounds in this context relies on recent results on the theory of block diagonal dominance of block tridiagonal matrices presented in Chapter 4.

The chapter is organized as follows. In Section 5.2 we state the multiplicative Schwarz method for linear algebraic systems of the form (5.1)–(5.3). We continue by studying the algebraic structure and the norm of the iteration matrices and present the main differences to the one-dimensional case in Section 5.2.1. In Section 5.2.2 we present a general expression for the convergence factor of the method when used to solve systems with matrices of type (5.3). We proceed by deriving quantitative error bounds for the method when the matrix $\mathbf{A}$ is both block tridiagonal and block diagonally dominant in Sections 5.2.3–5.2.4. Theoretical results specific to the case of convection-diffusion problems of type (5.2) are given in Section 5.2.5 and numerical experiments for specific cases are found in Section 5.4. Finally, a summary of the main results of the chapter and a brief discussion of possible generalizations and alternative applications of our approach is given in Chapter 6.

5.2. Convergence Bounds for the Multiplicative Schwarz Method

The multiplicative Schwarz method for solving linear algebraic systems of the form (5.1)–(5.3) can naturally be based on two local solves using the top and the bottom $N(m+1) \times N(m+1)$ blocks of $\mathbf{A}$, respectively. More precisely, the restriction operators of the method are given by

$$\mathbf{R}_1 \equiv \begin{bmatrix} \mathbf{I}_{N(m+1)} & 0 \end{bmatrix} \quad \text{and} \quad \mathbf{R}_2 \equiv \begin{bmatrix} 0 & \mathbf{I}_{N(m+1)} \end{bmatrix},$$

both of size $N(m+1) \times N(2m+1)$. The corresponding restrictions of the matrix $\mathbf{A}$ to each local subdomain (commonly refered to as *local subdomain problems*)are then given by

$$\mathbf{A}_1 \equiv \mathbf{R}_1 \mathbf{A} \mathbf{R}_1^T = \begin{bmatrix} \widehat{\mathbf{A}}_H & \mathbf{e}_m \otimes \mathbf{B}_H \\ \mathbf{e}_m^T \otimes \mathbf{C} & \widehat{\mathbf{A}} \end{bmatrix}, \quad \mathbf{A}_2 \equiv \mathbf{R}_2 \mathbf{A} \mathbf{R}_2^T = \begin{bmatrix} \widehat{\mathbf{A}} & \mathbf{e}_1^T \otimes \mathbf{B} \\ \mathbf{e}_1 \otimes \mathbf{C}_h & \widehat{\mathbf{A}}_h \end{bmatrix}, \tag{5.4}$$

both of size $N(m+1) \times N(m+1)$. Analogous to the one dimensional case we define the projection matrices

$$\mathbf{P}_i \equiv \mathbf{R}_i^T \mathbf{A}_i^{-1} \mathbf{R}_i \mathbf{A} \in \mathbb{R}^{N(2m+1) \times N(2m+1)}, \quad i = 1, 2, \tag{5.5}$$

and note that they are now of size $N(2m+1) \times N(2m+1)$. Once again, using the complimentary projections

$$\mathbf{Q}_i \equiv \mathbf{I} - \mathbf{P}_i \in \mathbb{R}^{N(2m+1) \times N(2m+1)}, \quad i = 1, 2,$$

we define the multiplicative Schwarz iteration matrices

$$\mathbf{T}_{12} \equiv \mathbf{Q}_2 \mathbf{Q}_1 \quad \text{and} \quad \mathbf{T}_{21} \equiv \mathbf{Q}_1 \mathbf{Q}_2. \tag{5.6}$$

Using these iteration matrices, the method is then given by (2.70)-(2.72), i.e., the transition to higher dimensional cases is reflected only by the specific block structure of the matrices (5.6). Using the theory developed in Chapter 4 allows us to present an analogous analysis to the one given in Chapter 3 for the one-dimensional case.

5.2.1. Structure of the iteration matrices

We begin by taking a closer look at the structure of the iteration matrices $\mathbf{T}_{ij}$ in (5.6). A direct computation based on (5.5) shows that

$$\mathbf{P}_1 = \begin{bmatrix} \mathbf{I}_{N(m+1)} \\ 0 \end{bmatrix} \mathbf{A}_1^{-1} \left[\begin{array}{c|c|c} \mathbf{A}_1 & \mathbf{e}_{m+1} \otimes \mathbf{B} & 0 \end{array} \right] = \begin{bmatrix} \mathbf{I}_{N(m+1)} & \mathbf{A}_1^{-1}(\mathbf{e}_{m+1} \otimes \mathbf{B}) & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

and

$$\mathbf{P}_2 = \begin{bmatrix} 0 \\ \mathbf{I}_{N(m+1)} \end{bmatrix} \mathbf{A}_2^{-1} \left[\begin{array}{c|c|c} 0 & \mathbf{e}_1 \otimes \mathbf{C} & \mathbf{A}_2 \end{array} \right] = \begin{bmatrix} 0 & 0 & 0 \\ 0 & \mathbf{A}_2^{-1}(\mathbf{e}_1 \otimes \mathbf{C}) & \mathbf{I}_{N(m+1)} \end{bmatrix},$$

where $\mathbf{e}_1, \mathbf{e}_{m+1} \in \mathbb{R}^{m+1}$. We see that both $\mathbf{P}_1$ and $\mathbf{P}_2$ have exactly $N(m+1)$ linearly independent columns, and hence

$$\operatorname{rank}(\mathbf{P}_1) = \operatorname{rank}(\mathbf{P}_2) = N(m+1).$$

Moreover, the complementary projections are

$$\mathbf{Q}_1 = \begin{bmatrix} 0 & -\mathbf{A}_1^{-1}(\mathbf{e}_{m+1} \otimes \mathbf{B}) & 0 \\ 0 & \mathbf{I}_N & 0 \\ 0 & 0 & \mathbf{I}_{N(m-1)} \end{bmatrix}, \quad \mathbf{Q}_2 = \begin{bmatrix} \mathbf{I}_{N(m-1)} & 0 & 0 \\ 0 & \mathbf{I}_N & 0 \\ 0 & -\mathbf{A}_2^{-1}(e_1 \otimes \mathbf{C}) & 0 \end{bmatrix},$$

and we have

$$\operatorname{rank}(\mathbf{Q}_1) = \operatorname{rank}(\mathbf{Q}_2) = Nm.$$

In order to simplify the notation we write

$$\begin{bmatrix} \mathbf{P}^{(1)} \\ \mathbf{\Pi}^{(1)} \end{bmatrix} \equiv \mathbf{A}_1^{-1}(\mathbf{e}_{m+1} \otimes \mathbf{B}) \quad \text{and} \quad \begin{bmatrix} \mathbf{\Pi}^{(2)} \\ \mathbf{P}^{(2)} \end{bmatrix} \equiv \mathbf{A}_2^{-1}(\mathbf{e}_1 \otimes \mathbf{C}), \tag{5.7}$$

where $\mathbf{\Pi}^{(i)} \in \mathbb{R}^{N \times N}$, and

$$\mathbf{P}^{(i)} = \left[\left(\mathbf{P}_1^{(i)}\right)^T, \dots, \left(\mathbf{P}_m^{(i)}\right)^T \right]^T \in \mathbb{R}^{Nm \times N} \quad \text{with} \quad \mathbf{P}_j^{(i)} \in \mathbb{R}^{N \times N}, \text{ for } j = 1, \dots, m.$$

Then

$$\mathbf{Q}_1 = \begin{bmatrix} 0_{Nm} & & -\mathbf{P}^{(1)} & \\ & 0_N & -\mathbf{\Pi}^{(1)} & \\ & & \mathbf{I}_N & \\ & & & \mathbf{I}_{N(m-1)} \end{bmatrix}, \quad \mathbf{Q}_2 = \begin{bmatrix} \mathbf{I}_{N(m-1)} & & & \\ & \mathbf{I}_N & & \\ & -\mathbf{\Pi}^{(2)} & 0_N & \\ & -\mathbf{P}^{(2)} & & 0_{Nm} \end{bmatrix},$$

and these matrices yield

$$\begin{aligned} \mathbf{T}_{12} = \mathbf{Q}_2\mathbf{Q}_1 &= \begin{bmatrix} 0 & -\mathbf{P}^{(1)} & 0 \\ 0 & \mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)} & 0 \\ 0 & \mathbf{P}^{(2)}\mathbf{P}_m^{(1)} & 0 \end{bmatrix} \\ &= \begin{bmatrix} -\mathbf{P}^{(1)} \\ \mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)} \\ \mathbf{P}^{(2)}\mathbf{P}_m^{(1)} \end{bmatrix} \left[\begin{array}{c|c|c} -0_{N(m+1)} & \mathbf{I}_N & 0_{N(m-1)} \end{array} \right] \equiv \mathbf{V}_1(\mathbf{e}_{m+2}^T \otimes \mathbf{I}_N), \end{aligned} \tag{5.8}$$

and

$$\begin{aligned} \mathbf{T}_{21} = \mathbf{Q}_1\mathbf{Q}_2 &= \begin{bmatrix} 0 & \mathbf{P}^{(1)}\mathbf{P}_1^{(2)} & 0 \\ 0 & \mathbf{\Pi}^{(1)}\mathbf{P}_1^{(2)} & 0 \\ 0 & -\mathbf{P}^{(2)} & 0 \end{bmatrix} \\ &= \begin{bmatrix} \mathbf{P}^{(1)}\mathbf{P}_1^{(2)} \\ \mathbf{\Pi}^{(1)}\mathbf{P}_1^{(2)} \\ -\mathbf{P}^{(2)} \end{bmatrix} \left[\begin{array}{c|c|c} 0_{N(m-1)} & \mathbf{I}_N & 0_{N(m+1)} \end{array} \right] \equiv \mathbf{V}_2(\mathbf{e}_m^T \otimes \mathbf{I}_N), \end{aligned} \tag{5.9}$$

where $\mathbf{e}_m, \mathbf{e}_{m+2} \in \mathbb{R}^{2m+1}$ and $\mathbf{V}_1, \mathbf{V}_2 \in \mathbb{R}^{N(2m+1) \times N}$.

Using these representations of the matrices $\mathbf{T}_{ij}$, we can obtain the following result, which is a generalization of Proposition 3.1 and Corollary 3.2 given in Chapter 3.

Lemma 5.1. *In the notation established above we have* $\operatorname{rank}(\mathbf{T}_{ij}) \leq N$, *and*

$$\mathbf{T}_{12}^{k+1} = \mathbf{V}_1 \left(\mathbf{P}_1^{(2)}\mathbf{P}_m^{(1)}\right)^k (\mathbf{e}_{m+2}^T \otimes \mathbf{I}_N), \quad \mathbf{T}_{21}^{k+1} = \mathbf{V}_2 \left(\mathbf{P}_m^{(1)}\mathbf{P}_1^{(2)}\right)^k (\mathbf{e}_m^T \otimes \mathbf{I}_N). \tag{5.10}$$

for all $k \geq 0$.

Proof. We only consider the matrix $\mathbf{T}_{12}$; the proof for $\mathbf{T}_{21}$ is analogous. The result about the rank is obvious from (5.8). We denote $\mathbf{E}_{m+2} \equiv \mathbf{e}_{m+2}^T \otimes \mathbf{I}_N$, then $\mathbf{T}_{12} = \mathbf{V}_1\mathbf{E}_{m+2}$, and it is easy to see that

$$\mathbf{T}_{12}^{k+1} = \mathbf{V}_1 \left(\mathbf{E}_{m+2}\mathbf{V}_1\right)^k \mathbf{E}_{m+2}, \quad \text{for all } k \geq 0.$$

Now

$$\mathbf{E}_{m+2}\mathbf{V}_1 = (\mathbf{e}_{m+2}^T \otimes \mathbf{I}_N)\begin{bmatrix} -\mathbf{P}^{(1)} \\ \mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)} \\ \mathbf{P}_1^{(2)}\mathbf{P}_m^{(1)} \\ \mathbf{P}_{2:m}^{(2)}\mathbf{P}_m^{(1)} \end{bmatrix} = \mathbf{P}_1^{(2)}\mathbf{P}_m^{(1)}$$

shows the first equality in (5.10). □

The next result generalizes Lemma 3.3 given in Chapter 3 and gives expressions for some block entries of the matrices $\mathbf{T}_{ij}$, which will be essential in our derivations of error bounds in the following sections.

Lemma 5.2. *Suppose that the matrices* $\widehat{\mathbf{A}}_H, \widehat{\mathbf{A}}_h \in \mathbb{R}^{Nm \times Nm}$ *in* (5.3) *are nonsingular, and denote* $\widehat{\mathbf{A}}_H^{-1} = [\mathbf{Z}_{ij}^{(H)}]$ *and* $\widehat{\mathbf{A}}_h^{-1} = [\mathbf{Z}_{ij}^{(h)}]$ *with* $\mathbf{Z}_{ij}^{(H)}, \mathbf{Z}_{ij}^{(h)} \in \mathbb{R}^{N \times N}$. *Then, in the notation established above,*

$$\begin{bmatrix} \mathbf{P}^{(1)} \\ \mathbf{\Pi}^{(1)} \end{bmatrix} = \begin{bmatrix} -\mathbf{Z}_{1:m,m}^{(H)}\mathbf{B}_H \\ \mathbf{I}_N \end{bmatrix}\mathbf{\Pi}^{(1)}, \quad \mathbf{\Pi}^{(1)} = \left(\widehat{\mathbf{A}} - \mathbf{C}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\right)^{-1}\mathbf{B},$$

and

$$\begin{bmatrix} \mathbf{\Pi}^{(2)} \\ \mathbf{P}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_N \\ -\mathbf{Z}_{1:m,1}^{(h)}\mathbf{C}_h \end{bmatrix}\mathbf{\Pi}^{(2)}, \quad \mathbf{\Pi}^{(2)} = \left(\widehat{\mathbf{A}} - \mathbf{B}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\right)^{-1}\mathbf{C}.$$

Proof. From (5.7) we know that $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, $\mathbf{\Pi}^{(1)}$, and $\mathbf{\Pi}^{(2)}$ solve the linear algebraic saddle-point systems

$$\begin{bmatrix} \widehat{\mathbf{A}}_H & \mathbf{e}_m \otimes \mathbf{B}_H \\ \mathbf{e}_m^T \otimes \mathbf{C} & \widehat{\mathbf{A}} \end{bmatrix}\begin{bmatrix} \mathbf{P}^{(1)} \\ \mathbf{\Pi}^{(1)} \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{B} \end{bmatrix}, \quad \begin{bmatrix} \widehat{\mathbf{A}} & \mathbf{e}_1^T \otimes \mathbf{B} \\ \mathbf{e}_1 \otimes \mathbf{C}_h & \widehat{\mathbf{A}}_h \end{bmatrix}\begin{bmatrix} \mathbf{\Pi}^{(2)} \\ \mathbf{P}^{(2)} \end{bmatrix} = \begin{bmatrix} \mathbf{C} \\ 0 \end{bmatrix}.$$

Hence the expressions for $\mathbf{P}^{(1)}$, $\mathbf{P}^{(2)}$, $\mathbf{\Pi}^{(1)}$, and $\mathbf{\Pi}^{(2)}$ can be obtained using Schur complements; see, e.g., [43, § 0.7.3]. □

5.2.2. Bounds for General Matrices

In order to bound the norms of the iteration matrices $\mathbf{T}_{12}$ and $\mathbf{T}_{21}$, we first have to decide which matrix norm should be taken. In the following we use a general induced matrix norm $\|\cdot\|$ which can be considered for square as well as for rectangular matrices. Note that an induced matrix norm for square matrices is submultiplicative and satisfies $\|\mathbf{I}\| = 1$ where $\mathbf{I}$ is the identity matrix.

Lemma 5.3. *In the notation established above, for any induced matrix norm we have*

$$\|\mathbf{T}_{ij}^{k+1}\| \leq \rho_{ij}^{k} \|\mathbf{T}_{ij}\|, \quad \textit{for all } k \geq 0, \tag{5.11}$$

where

$$\rho_{12} \equiv \|\mathbf{Z}_{11}^{(h)} \mathbf{C}_h \mathbf{\Pi}^{(2)} \mathbf{Z}_{mm}^{(H)} \mathbf{B}_H \mathbf{\Pi}^{(1)}\| \quad \textit{and} \quad \rho_{21} \equiv \|\mathbf{Z}_{mm}^{(H)} \mathbf{B}_H \mathbf{\Pi}^{(1)} \mathbf{Z}_{11}^{(h)} \mathbf{C}_h \mathbf{\Pi}^{(2)}\|. \tag{5.12}$$

Proof. We only consider the matrix $\mathbf{T}_{12}$; the proof for $\mathbf{T}_{21}$ is analogous. Taking norms in (5.10) yields

$$\|\mathbf{T}_{12}^{k+1}\| = \|\mathbf{V}_1 (\mathbf{P}_1^{(2)} \mathbf{P}_m^{(1)})^k \mathbf{E}_{m+2}\| \leq \rho_{12}^{k} \|\mathbf{V}_1\| \|\mathbf{E}_{m+2}\| \quad \text{with } \rho_{12} \equiv \|\mathbf{P}_1^{(2)} \mathbf{P}_m^{(1)}\|,$$

and where $\mathbf{E}_{m+2}$ is defined in the same way as in the proof of Lemma 5.1. Noting that $\|\mathbf{E}_{m+2}\| = \|\mathbf{I}_N\| = 1$ and

$$\|\mathbf{T}_{12}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{T}_{12}\mathbf{x}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{V}_1 \mathbf{E}_{m+2}\mathbf{x}\| = \max_{\|\mathbf{y}\|=1} \|\mathbf{V}_1 \mathbf{y}\| = \|\mathbf{V}_1\|,$$

yields the bound on $\|\mathbf{T}_{12}^{k+1}\|$ in (5.11).

Finally, the equality $\|\mathbf{p}_1^{(2)} \mathbf{p}_m^{(1)}\| = \|\mathbf{Z}_{11}^{(h)} \mathbf{C}_h \mathbf{\Pi}^{(2)} \mathbf{Z}_{mm}^{(H)} \mathbf{B}_H \mathbf{\Pi}^{(1)}\|$ in (5.12) follows directly from Lemma 5.2. □

So far our analysis considered general (nonsingular) blocks $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ in the matrix $\mathbf{A}$ in (5.3), and combining (2.72) and (5.11) gives a general error bound for the multiplicative Schwarz method in terms of certain blocks of $\mathbf{A}$ and the inverses of $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$. Note that using the submultiplicativity of the matrix norm $\|\cdot\|$, which at this point is still a general induced norm, both convergence factors ρ_{12} and ρ_{21} can be bounded by

$$\rho_{ij} \leq \|\mathbf{Z}_{11}^{(h)} \mathbf{C}_h\| \, \|\mathbf{Z}_{mm}^{(H)} \mathbf{B}_H\| \, \|\mathbf{\Pi}^{(1)}\| \, \|\mathbf{\Pi}^{(2)}\|. \tag{5.13}$$

In order to derive a *quantitative error bound* from the terms on the right hand side, we have to make additional assumptions on $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$. One possible choice of such assumptions is considered in the next section.

5.2.3. Bounds for Row Block Diagonally Dominant Block Tridiagonal Matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$

We are most interested in the analysis of the multiplicative Schwarz method for linear algebraic systems that arise in certain discretizations of partial differential equations, in particular the convection-diffusion problem (5.2), and we will therefore consider the matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ be given by

$$\widehat{\mathbf{A}}_H = \text{tridiag}(\mathbf{C}_H, \mathbf{A}_H, \mathbf{B}_H) \quad \text{and} \quad \widehat{\mathbf{A}}_h = \text{tridiag}(\mathbf{C}_h, \mathbf{A}_h, \mathbf{B}_h). \tag{5.14}$$

Additionally, we will assume that the matrices

$$\mathbf{A}_H, \mathbf{B}_H, \mathbf{C}_H, \widehat{\mathbf{A}}, \mathbf{B}, \mathbf{C}, \mathbf{A}_h, \mathbf{B}_h, \mathbf{C}_h \in \mathbb{R}^{N\times N} \text{ are nonsingular,} \tag{5.15}$$

and that the matrix $\mathbf{A}$ is *row block diagonally dominant* in the sense of Definition 4.1 given in Chapter 4, i.e., that

$$\begin{aligned} \|\mathbf{A}_H^{-1}\mathbf{B}_H\| + \|\mathbf{A}_H^{-1}\mathbf{C}_H\| &\leq 1, \\ \|\widehat{\mathbf{A}}^{-1}\mathbf{B}\| + \|\widehat{\mathbf{A}}^{-1}\mathbf{C}\| &\leq 1, \\ \|\mathbf{A}_h^{-1}\mathbf{B}_h\| + \|\mathbf{A}_h^{-1}\mathbf{C}_h\| &\leq 1. \end{aligned} \tag{5.16}$$

Note that because of (5.15) each of the norms on the left hand sides of these inequalities is *strictly* less than one.

Both $\mathbf{A}_H$ and $\mathbf{A}_h$ satisfy all assumptions of Theorem 4.6 given in Chapter 4. A minor modification of the first equation in the proof of Theorem 4.6 (namely multiplying both sides of this equation by $\mathbf{C}_h$ or $\mathbf{B}_H$ before taking norms) shows that the blocks of the inverses, i.e., $\widehat{\mathbf{A}}_H^{-1} = [\mathbf{Z}_{ij}^{(H)}]$ and $\widehat{\mathbf{A}}_h^{-1} = [\mathbf{Z}_{ij}^{(h)}]$, satisfy

$$\|\mathbf{Z}_{i1}^{(h)}\mathbf{C}_h\| \leq \|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\| \quad \text{and} \quad \|\mathbf{Z}_{im}^{(H)}\mathbf{B}_H\| \leq \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|, \quad i = 1, \ldots, m. \tag{5.17}$$

Moreover, as shown in the proof of Theorem 4.6, the equations

$$\begin{aligned} \mathbf{Z}_{11}^{(h)} &= (\mathbf{A}_h - \mathbf{B}_h\mathbf{M}_h)^{-1} = (\mathbf{I} - \mathbf{A}_h^{-1}\mathbf{B}_h\mathbf{M}_h)^{-1}\mathbf{A}_h^{-1}, \\ \mathbf{Z}_{mm}^{(H)} &= (\mathbf{A}_H - \mathbf{C}_H\mathbf{L}_H)^{-1} = (\mathbf{I} - \mathbf{A}_H^{-1}\mathbf{C}_H\mathbf{L}_H)^{-1}\mathbf{A}_H^{-1} \end{aligned}$$

hold for some matrices $\mathbf{M}_h, \mathbf{L}_H \in \mathbb{R}^{N\times N}$ with $\|\mathbf{M}_h\| \leq 1$ and $\|\mathbf{L}_H\| \leq 1$; see (4.22) in Chapter 4. The precise definition of $\mathbf{M}_h$ and $\mathbf{L}_H$ is not important here.

The four matrices that appear on the right hand side of (5.13) are now given by

$$\begin{aligned} \mathbf{Z}_{11}^{(h)}\mathbf{C}_h &= (\mathbf{I} - \mathbf{A}_h^{-1}\mathbf{B}_h\mathbf{M}_h)^{-1}\mathbf{A}_h^{-1}\mathbf{C}_h, \\ \mathbf{Z}_{mm}^{(H)}\mathbf{B}_H &= (\mathbf{I} - \mathbf{A}_H^{-1}\mathbf{C}_H\mathbf{L}_H)^{-1}\mathbf{A}_H^{-1}\mathbf{B}_H, \\ \mathbf{\Pi}^{(2)} &= (\mathbf{I} - \widehat{\mathbf{A}}^{-1}\mathbf{B}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h)^{-1}\widehat{\mathbf{A}}^{-1}\mathbf{C}, \\ \mathbf{\Pi}^{(1)} &= (\mathbf{I} - \widehat{\mathbf{A}}^{-1}\mathbf{C}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H)^{-1}\widehat{\mathbf{A}}^{-1}\mathbf{B}. \end{aligned}$$

Since $\|\mathbf{A}_h^{-1}\mathbf{B}_h\mathbf{M}_h\| \le \|\mathbf{A}_h^{-1}\mathbf{B}_h\| < 1$, we can use the Neumann series to obtain

$$\begin{aligned}\|(\mathbf{I} - \mathbf{A}_h^{-1}\mathbf{B}_h\mathbf{M}_h)^{-1}\| &= \left\|\sum_{k=0}^{\infty}(\mathbf{A}_h^{-1}\mathbf{B}_h\mathbf{M}_h)^k\right\| \\ &\le \sum_{k=0}^{\infty}\|\mathbf{A}_h^{-1}\mathbf{B}_h\|^k = \frac{1}{1 - \|\mathbf{A}_h^{-1}\mathbf{B}_h\|}.\end{aligned}$$

Similarly, $\|\mathbf{A}_H^{-1}\mathbf{C}_H\mathbf{L}_H\| \le \|\mathbf{A}_H^{-1}\mathbf{C}_H\| < 1$ implies that

$$\|(\mathbf{I} - \mathbf{A}_H^{-1}\mathbf{C}_H\mathbf{L}_H)^{-1}\| \le \frac{1}{1 - \|\mathbf{A}_H^{-1}\mathbf{C}_H\|},$$

and hence

$$\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\| \le \frac{\|\mathbf{A}_h^{-1}\mathbf{C}_h\|}{1 - \|\mathbf{A}_h^{-1}\mathbf{B}_h\|} \equiv \eta_h \le 1, \tag{5.18}$$

$$\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\| \le \frac{\|\mathbf{A}_H^{-1}\mathbf{B}_H\|}{1 - \|\mathbf{A}_H^{-1}\mathbf{C}_H\|} \equiv \eta_H \le 1. \tag{5.19}$$

Using (5.18) and (5.19) yields

$$\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\| \le \eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\| < 1 \quad \text{and} \quad \|\widehat{\mathbf{A}}^{-1}\mathbf{C}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\| \le \eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\| < 1,$$

and another application of the Neumann series shows that

$$\|\mathbf{\Pi}^{(2)}\| \le \frac{\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}{1 - \eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|} \le 1 \quad \text{and} \quad \|\mathbf{\Pi}^{(1)}\| \le \frac{\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}{1 - \eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|} \le 1. \tag{5.20}$$

In summary, we have the following result.

Lemma 5.4. *In the notation established above, the convergence factors of the multiplicative Schwarz method given in* (5.12) *satisfy*

$$\rho_{ij} \le \frac{\eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}{1 - \eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}\,\frac{\eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}{1 - \eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}, \tag{5.21}$$

where each of the factors on the right hand side is less than or equal to one.

We now illustrate the bound from Lemma 5.4 with a simple example.

Example 5.5. *Let $N \ge 2$ and $m \ge 1$ be given, and consider the matrix*

$$\mathbf{A} \equiv \mathrm{tridiag}(-\mathbf{I}, \mathbf{W}, -\mathbf{I}) \in \mathbb{R}^{N(2m+1)\times N(2m+1)},$$

where $\mathbf{W} \equiv \mathrm{tridiag}(-1, 4, -1) \in \mathbb{R}^{N\times N}$ and $\mathbf{I} \in \mathbb{R}^{N\times N}$. It is well known that $\mathbf{A}$ is the result of a standard finite difference discretization of the 2D Poisson equation on the unit square and with Dirichlet boundary conditions. In our notation, $\mathbf{A}$ is of the form (5.3) *and* (5.14) *with*

$$\widehat{\mathbf{A}}_H = \widehat{\mathbf{A}}_h = \mathrm{tridiag}(-\mathbf{I}, \mathbf{W}, -\mathbf{I}) \in \mathbb{R}^{Nm\times Nm},$$

$\mathbf{B}_H = \mathbf{B} = \mathbf{B}_h = \mathbf{C}_H = \mathbf{C} = \mathbf{C}_h = -\mathbf{I}$, *and* $\mathbf{A}_H = \widehat{\mathbf{A}} = \mathbf{A}_h = \mathbf{W}$. *The eigenvalues of the symmetric positive definite matrix* $\mathbf{W}$ *are given by*

$$\lambda_k = 4 - 2\cos\frac{\pi k}{N+1} > 2, \quad k = 1, \ldots, N.$$

For the 2-norm $\|\cdot\|_2$ *we have*

$$\|\mathbf{W}^{-1}\|_2 = \frac{1}{\lambda_1} = \frac{1}{4 - 2\cos\frac{\pi}{N+1}} < \frac{1}{2},$$

and hence $\mathbf{A}$ *is strictly row block diagonally dominant with respect to the 2-norm; see the conditions* (5.16). *Note that* $\mathbf{A}$ *is only weakly row diagonally dominant in the classical (scalar) sense.*

Using the definitions (5.18) *and* (5.19) *we obtain*

$$\eta_h = \eta_H = \frac{\|\mathbf{W}^{-1}\|}{1 - \|\mathbf{W}^{-1}\|} = \frac{1}{\lambda_1 - 1},$$

Now (5.21) *yields the following bound on the convergence factor of the multiplicative Schwarz method:*

$$\rho_{ij} \le \left(\frac{\frac{\|\mathbf{W}^{-1}\|_2}{1-\|\mathbf{W}^{-1}\|_2}\|\mathbf{W}^{-1}\|_2}{1 - \frac{\|\mathbf{W}^{-1}\|_2}{1-\|\mathbf{W}^{-1}\|_2}\|\mathbf{W}^{-1}\|_2} \right)^2 = \left(\frac{1}{\lambda_1^2 - \lambda_1 - 1} \right)^2.$$

Thus, the convergence factor of the multiplicative Schwarz method for this problem is less than one, regardless of the choices of N *and* m. *But note that for* $N \to \infty$ *we have* $\lambda_1 \to 2$ *and hence* $\rho_{ij} \to 1$. □

To bound the norm of the error of the multiplicative Schwarz method, see (2.72), (5.11), and (5.21), it remains to bound $\|\mathbf{T}_{ij}\|$. Let us first realize that because of the equivalence of matrix norms, there exists a constant c such that

$$\|\mathbf{T}_{ij}\| \le c\,\|\mathbf{T}_{ij}\|_\infty,$$

where c can depend on the size of $\mathbf{T}_{ij}$.

Now we bound $\|\mathbf{T}_{ij}\|_\infty$. From (5.8) and (5.9) we see that

$$\|\mathbf{T}_{12}\|_\infty = \max\{\|\mathbf{P}^{(1)}\|_\infty,\ \|\mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)}\|_\infty,\ \|\mathbf{P}^{(2)}\mathbf{P}_m^{(1)}\|_\infty\}, \tag{5.22}$$

$$\|\mathbf{T}_{21}\|_\infty = \max\{\|\mathbf{P}^{(2)}\|_\infty,\ \|\mathbf{\Pi}^{(1)}\mathbf{P}_1^{(2)}\|_\infty,\ \|\mathbf{P}^{(1)}\mathbf{P}_1^{(2)}\|_\infty\}, \tag{5.23}$$

and Lemma 5.2 yields

$$\begin{aligned}
\mathbf{P}^{(1)} &= -\mathbf{Z}_{1:m,m}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}, & \mathbf{P}^{(2)} &= -\mathbf{Z}_{1:m,1}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)},\\
\mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)} &= -\mathbf{\Pi}^{(2)}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}, & \mathbf{\Pi}^{(1)}\mathbf{P}_1^{(2)} &= -\mathbf{\Pi}^{(1)}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)},\\
\mathbf{P}^{(2)}\mathbf{P}_m^{(1)} &= \mathbf{Z}_{1:m,1}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}, & \mathbf{P}^{(1)}\mathbf{P}_1^{(2)} &= \mathbf{Z}_{1:m,m}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}.
\end{aligned}$$

Therefore, using (5.17) we can bound the ∞-norms of these matrices as follows:

$$\begin{aligned}
\|\mathbf{P}^{(1)}\|_\infty &= \max\{\|\mathbf{Z}_{1m}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\|_\infty,\dots,\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\|_\infty\}\\
&\le \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty,\\
\|\mathbf{\Pi}^{(2)}\mathbf{P}_m^{(1)}\|_\infty &= \|\mathbf{\Pi}^{(2)}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\|_\infty\\
&\le \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty,\\
\|\mathbf{P}^{(2)}\mathbf{P}_m^{(1)}\|_\infty &= \max\{\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\|_\infty,\dots,\|\mathbf{Z}_{m1}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\|_\infty\}\\
&\le \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty,
\end{aligned}$$

and

$$\begin{aligned}
\|\mathbf{P}^{(2)}\|_\infty &= \max\{\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\|_\infty,\dots,\|\mathbf{Z}_{m1}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\|_\infty\}\\
&\le \|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty,\\
\|\mathbf{\Pi}^{(1)}\mathbf{P}_1^{(2)}\| &= \|\mathbf{\Pi}^{(1)}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\|_\infty\\
&\le \|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty,\\
\|\mathbf{P}^{(1)}\mathbf{P}_1^{(2)}\|_\infty &= \max\{\|\mathbf{Z}_{1m}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\|_\infty,\dots,\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\mathbf{\Pi}^{(1)}\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\mathbf{\Pi}^{(2)}\|_\infty\}\\
&\le \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty.
\end{aligned}$$

The individual terms on the right hand sides of previous inequalities are all less of equal than one. Therefore, the maximum of the first three bounds is $\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty$, and the maximum of the second three bounds is $\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty$. Hence, using (5.22) and (5.23), and (5.18), (5.19), (5.20) we obtain

$$\|\mathbf{T}_{12}\|_\infty \le \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|_\infty\|\mathbf{\Pi}^{(1)}\|_\infty \le \frac{\eta_{H,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty}{1-\eta_{H,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}$$

and

$$\|\mathbf{T}_{21}\|_\infty \le \|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|_\infty\|\mathbf{\Pi}^{(2)}\|_\infty \le \frac{\eta_{h,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}{1-\eta_{h,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty},$$

where $\eta_{h,\infty}$ and $\eta_{H,\infty}$ are defined as in (5.18) and (5.19) using the ∞-norm,

$$\eta_{h,\infty} \equiv \frac{\|\mathbf{A}_h^{-1}\mathbf{C}_h\|_\infty}{1-\|\mathbf{A}_h^{-1}\mathbf{B}_h\|_\infty}, \qquad \eta_{H,\infty} \equiv \frac{\|\mathbf{A}_H^{-1}\mathbf{B}_H\|_\infty}{1-\|\mathbf{A}_H^{-1}\mathbf{C}_H\|_\infty}. \tag{5.24}$$

Combining these bounds with Lemma 5.3 and Lemma 5.4 gives the following convergence result.

Theorem 5.6. *Suppose that* $\mathbf{A}$ *as in* (5.3) *has blocks as in* (5.14) *that satisfy* (5.15)–(5.16)*. Then the errors of the multiplicative Schwarz method* (2.70) *applied to the linear algebraic system* (5.1) *satisfy*

$$\frac{\|\mathbf{e}^{(k+1)}\|}{\|\mathbf{e}^{(0)}\|} \le \left(\frac{\eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}{1-\eta_h\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}\,\frac{\eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}{1-\eta_H\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}\right)^k \|\mathbf{T}_{ij}\|, \quad k=0,1,2,\dots,$$

where η_h and η_H are defined in (5.18)–(5.19). *Moreover,*

$$\|\mathbf{T}_{12}\| \le c\,\frac{\eta_{H,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty}{1-\eta_{H,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}, \qquad \|\mathbf{T}_{21}\| \le c\,\frac{\eta_{h,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}{1-\eta_{h,\infty}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty},$$

where $\eta_h^{(\infty)}$ and $\eta_H^{(\infty)}$ are given by (5.24), *and c is a constant such that $\|\mathbf{T}_{ij}\| \le c\,\|\mathbf{T}_{ij}\|_\infty$.*

We will now present an analysis for the case of the matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ are not only row block diagonal dominant but also column block diagonal dominant.

5.2.4. Bounds for Row and Column Block Diagonally Dominant Block Tridiagonal Matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$

In many practical applications, like the one we are interested in, the matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ of the form (5.14) are not only row block diagonally dominant, see (5.16), but also *column block diagonally dominant*, i.e., they satisfy the conditions

$$\|\mathbf{B}_H\mathbf{A}_H^{-1}\| + \|\mathbf{C}_H\mathbf{A}_H^{-1}\| \le 1, \quad \|\mathbf{B}_h\mathbf{A}_h^{-1}\| + \|\mathbf{C}_h\mathbf{A}_h^{-1}\| \le 1. \tag{5.25}$$

Suppose now that the conditions (5.25) are satisfied. Then, using a reformulation of the results of Chapter 4 for column block diagonally dominant matrices (see Appendix 4.A), the equations

$$\mathbf{Z}_{11}^{(h)} = (\mathbf{A}_h - \tilde{\mathbf{L}}_h\mathbf{C}_h)^{-1} = \mathbf{A}_h^{-1}(\mathbf{I} - \tilde{\mathbf{L}}_h\mathbf{C}_h\mathbf{A}_h^{-1})^{-1},$$
$$\mathbf{Z}_{mm}^{(H)} = (\mathbf{A}_H - \tilde{\mathbf{M}}_H\mathbf{B}_H)^{-1} = \mathbf{A}_H^{-1}(\mathbf{I} - \tilde{\mathbf{M}}_H\mathbf{B}_H\mathbf{A}_H^{-1})^{-1}$$

hold for some matrices $\tilde{\mathbf{L}}_h, \tilde{\mathbf{M}}_H \in \mathbb{R}^{N\times N}$ with $\|\tilde{\mathbf{L}}_h\| \le 1$ and $\|\tilde{\mathbf{M}}_H\| \le 1$. Analogously, using the Neumann series we obtain the bounds

$$\|\mathbf{Z}_{11}^{(h)}\| \le \frac{\|\mathbf{A}_h^{-1}\|}{1-\|\mathbf{C}_h\mathbf{A}_h^{-1}\|}, \qquad \|\mathbf{Z}_{mm}^{(H)}\| \le \frac{\|\mathbf{A}_H^{-1}\|}{1-\|\mathbf{B}_H\mathbf{A}_H^{-1}\|},$$

so that

$$\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\| \le \frac{\|\mathbf{A}_h^{-1}\|\|\mathbf{C}_h\|}{1-\|\mathbf{C}_h\mathbf{A}_h^{-1}\|}, \qquad \|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\| \le \frac{\|\mathbf{A}_H^{-1}\|\|\mathbf{B}_H\|}{1-\|\mathbf{B}_H\mathbf{A}_H^{-1}\|}.$$

Therefore, if $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ satisfy both, the conditions (5.16) as well as (5.25), then

$$\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\| \le \min\left\{\frac{\|\mathbf{A}_h^{-1}\mathbf{C}_h\|}{1-\|\mathbf{A}_h^{-1}\mathbf{B}_h\|}, \frac{\|\mathbf{A}_h^{-1}\|\|\mathbf{C}_h\|}{1-\|\mathbf{C}_h\mathbf{A}_h^{-1}\|}\right\} \equiv \eta_h^{\min}, \tag{5.26}$$

and

$$\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\| \le \min\left\{\frac{\|\mathbf{A}_H^{-1}\mathbf{B}_H\|}{1-\|\mathbf{A}_H^{-1}\mathbf{C}_H\|}, \frac{\|\mathbf{A}_H^{-1}\|\|\mathbf{B}_H\|}{1-\|\mathbf{B}_H\mathbf{A}_H^{-1}\|}\right\} \equiv \eta_H^{\min}. \tag{5.27}$$

The value of $\eta_H^{\min}$ can be much smaller than η_H for example if $\|\mathbf{B}_H\| \ll \|\mathbf{C}_H\|$. Since we only improved bounds (5.18) and (5.19) on $\|\mathbf{Z}_{11}^{(h)}\mathbf{C}_h\|$ and $\|\mathbf{Z}_{mm}^{(H)}\mathbf{B}_H\|$, we can formulate a version of Theorem 5.6 where we just replace η_h and η_H with $\eta_h^{\min}$ and $\eta_H^{\min}$.

Theorem 5.7. *Suppose that* **A** *as in* (5.3) *has blocks as in* (5.14) *that satisfy* (5.15)–(5.16), and (5.25). *Then the errors of the multiplicative Schwarz method* (2.70) *applied to the linear algebraic system* (5.1) *satisfy*

$$\frac{\|\mathbf{e}^{(k+1)}\|}{\|\mathbf{e}^{(0)}\|} \leq \left(\frac{\eta_h^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|}{1-\eta_h^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|} \, \frac{\eta_H^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|}{1-\eta_H^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|} \right)^k \|\mathbf{T}_{ij}\|, \quad k=0,1,2,\ldots,$$

where $\eta_h^{\min}$ *and* $\eta_H^{\min}$ *are defined in* (5.26) *and* (5.27). *Moreover,*

$$\|\mathbf{T}_{12}\| \leq c\,\frac{\eta_{H,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty}{1-\eta_{H,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}, \qquad \|\mathbf{T}_{21}\| \leq c\,\frac{\eta_{h,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}{1-\eta_{h,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty},$$

where $\eta_{h,\infty}^{\min}$ *and* $\eta_{H,\infty}^{\min}$ *are given by* (5.26) *and* (5.27) *using the* ∞*-norm, and c is a constant such that* $\|\mathbf{T}_{ij}\| \leq c\,\|\mathbf{T}_{ij}\|_\infty$.

In the next section we will explicitly state convergence results for the case when the matrices **A** come from the Shihskin mesh discretization of convection-diffusion model problems.

5.2.5. Bounds for Convection-Diffusion Problems

In order to obtain quantitative error bounds for the case of matrices arising in the discretization of BVPs of type (5.2), we need to make further assumptions on the parameters of the problem. In particular we assume that on $\overline{\Omega}$ the components of the velocity field are such that $\boldsymbol{\omega} = [0,1]^T$, leading to the problem:

$$\begin{cases} -\epsilon\left(\frac{\partial^2 u(x,y)}{\partial x^2} + \frac{\partial^2 u(x,y)}{\partial y^2}\right) + \frac{\partial u(x,y)}{\partial y} + \beta u(x,y) = f(x,y), & \text{in } \Omega = (0,1)\times(0,1) \\ \qquad u(x,y) = g(x,y), \quad \text{on } \partial\Omega. \end{cases} \tag{5.28}$$

Following the discretization procedure described in Section 2.1.6 leads to a linear system (5.1), where the matrix **A** exhibits the structure (5.3) with matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ of the form (5.14) which are both row *and* column block diagonally dominant. The entries of **A** are then given by (2.45) by setting $\omega_x = 0$ and $\omega_y = 1$. We will now show that for this model problem the assumptions of Theorem 5.7 are satisfied.

Lemma 5.8. *All nonzero blocks of the matrix* **A** *described above are nonsingular. Moreover, for the matrix* ∞*-norm the matrix* **A** *satisfies the conditions* (5.16), *i.e., it is row block diagonally dominant, and the submatrices* $\widehat{\mathbf{A}}_H$ *and* $\widehat{\mathbf{A}}_h$ *satisfy the conditions* (5.25), *i.e., they are column block diagonally dominant.*

Proof. Note that all (nonzero) off-diagonal entries of **A** are negative, and the diagonal entries a_H, a, a_h are positive. Moreover,

$$a_H + b_H + c_H + d_H + e_H = a + b + c + d + e = a_h + b_h + c_h + d_h + e_h = \beta \geq 0$$

It is thus easy to see that all nonzero blocks of **A** are nonsingular.

To prove (5.16) and (5.25) for the ∞-norm, we just need to show that

$$|e_H + d_H| \|\mathbf{A}_H^{-1}\|_\infty \le 1, \quad |e + d| \|\mathbf{A}^{-1}\|_\infty \le 1, \quad |e_h + d_h| \|\mathbf{A}_h^{-1}\|_\infty \le 1, \tag{5.29}$$

and hence we need to bound the ∞-norms of matrices $\mathbf{A}_H^{-1}$, $\widehat{\mathbf{A}}^{-1}$, and $\mathbf{A}_h^{-1}$.

First note that for any nonsingular matrix $\mathbf{M}$ and an induced matrix norm we have

$$\|\mathbf{M}^{-1}\| = \max_{\|\mathbf{v}\|=1} \left\| \mathbf{M}^{-1} \left(\frac{\mathbf{M}\mathbf{v}}{\|\mathbf{M}\mathbf{v}\|} \right) \right\| = \frac{1}{\min_{\|\mathbf{v}\|=1} \|\mathbf{M}\mathbf{v}\|}.$$

Therefore, if $\|\mathbf{M}\mathbf{v}\| \ge \gamma > 0$ for any unit norm vector $\mathbf{v}$, then $\|\mathbf{M}^{-1}\| \le \gamma^{-1}$.

Second, suppose that $\mathbf{M}$ is a strictly diagonally dominant tridiagonal Toeplitz matrix $\mathbf{M} = \text{tridiag}(\hat{c}, \hat{a}, \hat{b})$, where $\hat{a} > 0$, $\hat{b} < 0$, $\hat{c} < 0$, and $\hat{a} + \hat{b} + \hat{c} > 0$. We would like to bound $\|\mathbf{M}\mathbf{v}\|_\infty$ for any unit norm vector $\mathbf{v}$ from below. If $\|\mathbf{v}\|_\infty = 1$, then there is an index i such that $|v_i| = 1$. Without loss of generality we can assume that $v_i = 1$, because changing the sign of the vector does not change $\|\mathbf{M}\mathbf{v}\|_\infty$. Defining $v_0 = 0$ and $v_{n+1} = 0$ we obtain

$$\|\mathbf{M}\mathbf{v}\|_\infty \ge |v_{i-1}\hat{c} + \hat{a} + v_{i+1}\hat{b}| \ge \hat{a} + \hat{b} + \hat{c},$$

and therefore

$$\|\mathbf{M}^{-1}\|_\infty \le \frac{1}{\hat{a} + \hat{b} + \hat{c}}. \tag{5.30}$$

In order to prove (5.29), we now apply the bound (5.30) to matrices $\mathbf{A}_H$, $\widehat{\mathbf{A}}$, and $\mathbf{A}_h$, which are strictly diagonally dominant tridiagonal Toeplitz matrices with the required sign pattern. For $\mathbf{A}_H$ we get

$$|e_H + d_H| \|\mathbf{A}_H^{-1}\|_\infty \le \frac{|e_H + d_H|}{a_H + b_H + c_H} = \frac{|e_H + d_H|}{|e_H + d_H| + \beta} \le 1,$$

and the other inequalities in (5.29) follow analogously. □

Lemma 5.8 ensures that the assumptions of Theorem 5.7 are satisfied for matrices arising from the discretization of the problem (5.28) using upwind differences on a Shishkin mesh. We therefore obtain the following convergence result for the multiplicative Schwarz method.

Corollary 5.9. *Consider the lienar algebraic system* (5.1) *obtained from the upwind discretization of the convection-diffusion boundary value problem* (5.28) *on a Shishkin mesh given by* (2.25). *Then the errors of the multiplicative Schwarz method* (2.70) *applied to the linear system satisfy*

$$\frac{\|\mathbf{e}^{(k+1)}\|_\infty}{\|\mathbf{e}^{(0)}\|_\infty} \le \rho^k \|\mathbf{T}_{ij}\|_\infty, \quad k = 0, 1, 2, \ldots. \tag{5.31}$$

where,

$$\rho \equiv \frac{\epsilon}{\epsilon + H_y}, \qquad \|\mathbf{T}_{12}\|_\infty \le \rho \quad \textit{and} \quad \|\mathbf{T}_{21}\|_\infty \le 1\,. \tag{5.32}$$

Proof. To prove the bounds (5.31)–(5.32) we apply Theorem 5.7 with the ∞-norm and bound the factors $\eta_{h,\infty}^{\min}$ and $\eta_{H,\infty}^{\min}$ that correspond to the discretization scheme (2.45).

Since $|d_h| > |e_h|$, we obtain

$$\eta_{h,\infty}^{\min} = \min\left\{\frac{|d_h|\|\mathbf{A}_h^{-1}\|_\infty}{1 - |e_h|\|\mathbf{A}_h^{-1}\|_\infty}, \frac{|d_h|\|\mathbf{A}_h^{-1}\|_\infty}{1 - |d_h|\|\mathbf{A}_h^{-1}\|_\infty}\right\} = \frac{|d_h|\|\mathbf{A}_h^{-1}\|_\infty}{1 - |e_h|\|\mathbf{A}_h^{-1}\|_\infty} \le 1.$$

Similarly, from $|d_H| > |e_H|$ it follows that

$$\eta_{H,\infty}^{\min} = \min\left\{\frac{|e_H|\|\mathbf{A}_H^{-1}\|_\infty}{1 - |d_H|\|\mathbf{A}_H^{-1}\|_\infty}, \frac{|e_H|\|\mathbf{A}_H^{-1}\|_\infty}{1 - |e_H|\|\mathbf{A}_H^{-1}\|_\infty}\right\} = \frac{|e_H|\|\mathbf{A}_H^{-1}\|_\infty}{1 - |e_H|\|\mathbf{A}_H^{-1}\|_\infty} \le \left|\frac{e_H}{d_H}\right|.$$

Hence, we obtain an upper bound for the convergence factors ρ_{12} and ρ_{21}:

$$\frac{\eta_{h,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty}{1 - \eta_{h,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty} \frac{\eta_{H,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{B}\|_\infty}{1 - \eta_{H,\infty}^{\min}\|\widehat{\mathbf{A}}^{-1}\mathbf{C}\|_\infty} \le \eta_{h,\infty}^{\min}\eta_{H,\infty}^{\min} \le \left|\frac{e_H}{d_H}\right| \equiv \rho.$$

Lastly, for the norms $\|\mathbf{T}_{ij}\|_\infty$ we obtain

$$\|\mathbf{T}_{12}\|_\infty \le \eta_{H,\infty} \le \left|\frac{e_H}{d_H}\right|, \qquad \|\mathbf{T}_{21}\|_\infty \le \eta_{h,\infty}^{\min} \le 1\,.$$

Substituting the values of e_H and d_H given in (2.45) yields the desired result. □

Note that the bound (5.31) does not depend on the choice of N, i.e., the size of the mesh in the x-direction. Moreover, for a fixed choice of M, and hence of H_y, the value of $\epsilon/(\epsilon + H_y)$ decreases with decreasing ϵ. Similar to the one-dimensional model problem studied in Chapter 3, this indicates a faster convergence of the multiplicative Schwarz method for smaller ϵ, meaning larger convection-dominance. In Section 5.4 we will apply these results in the numerical study of the model problem (5.2), which can be considered a two-dimensional generalization of the one-dimensional problem (3.2) studied in Chapter 3.

5.3. Shishkin-Schwarz Preconditioning

As we have seen in Chapter 3, the multiplicative Schwarz method as well as GMRES applied to the preconditioned system

$$(\mathbf{I} - \mathbf{T}_{ij})\mathbf{u} = \mathbf{v} \tag{5.33}$$

obtain their approximations from the Krylov subspace $\mathcal{K}_k(\mathbf{I} - \mathbf{T}_{ij}, \mathbf{r}^{(0)})$. Consequently, if a matrix $\mathbf{T}$ satisfies $\operatorname{rank}(\mathbf{T}) = r$, with $\mathbf{I} - \mathbf{T}$ nonsingular, then for any initial residual $\mathbf{r}^{(0)}$ we have that GMRES applied to the system (5.33) converges to the solution in at most $r + 1$ steps (in exact arithmetic). For the two-dimensional model problem studied in this chapter we have a matrix $\mathbf{T}$ with $r \leq N$, where N is the number of gridpoints in the transition layer of the mesh, thus

$$\dim\left(\mathcal{K}_k(\mathbf{I} - \mathbf{T}_{ij}, \mathbf{r}^{(0)})\right) \leq N + 1, \tag{5.34}$$

and, GMRES applied to (5.33) converges in (at most) $N + 1$ steps. This result is valid even when the multiplicative Schwarz iteration itself converges slowly or might even diverge, which may happen for some special cases if the the central difference scheme is used; as we have shown for the one-dimensional problem.

As mentioned in Chapter 3, in most practical applications one is interested in the case where the local subdomain problems (5.4) are solved inexactly, and thus, the bounds obtained in this work and the exact termination of GMRES in $r + 1$ steps will no longer hold (see the numerical results presented in Table 5.2). For the theory behind inexact local solves for multiplicative schwarz methods see, for example, [5] or the monograph [71]. For an example of the use of inexact local solves for non-overlapping domain decomposition methods see [13].

In order to solve the local subdomain problems inexactly, we have to utilize two nested iterative processes to reach the solution of the linear system. The first process consists in applying GMRES to solve the system (5.33) (outer iteration) and the second being the usage of an iterative method to solve each of the subsystems (5.4) (inner iterations). In this work we use the GMRES method both as the outer as well as the inner iteration processes. For each of the outer iteration steps, we solve the subsystems up to a desired tolerance, referred to as the *inexact local solve tolerance*. However, it is also natural to ask about the effect of convergence of the outer iteration when only a finite number of inner iterations are taken. In this work we only explore the first approach in the numerical experiments present in the next section.

It is important to note that the solution of the local subdomain problems is only needed to construct the projection operators (5.5), whose complementary projections form the factors of the iteration matrices $\mathbf{T}_{ij}$. In a general case, all the columns of the matrices $\mathbf{A}_i^{-1}$ would be needed to construct the projections. However, for the particular case of the iteration matrices $\mathbf{T}_{ij}$ studied in this work, we have explicitly shown that the projection operators only make use of the last N columns of the matrix $\mathbf{A}_1$ and the first N columns of the matrix $\mathbf{A}_2$, as shown by the equations (5.8) and (5.9). Thus a specially efficient implementation can be achieved in this case. In particular, we only use the inner iteration to approximate the subsystems given by (5.7) and directly construct each of the factors $\mathbf{Q}_1$ and $\mathbf{Q}_2$ with the computed approximations. Furthermore, the author of this work did not find a general convention in existent literature regarding the implementation of inexact local solves in the context of Schwarz methods (for the theory see [5]), and thus, the results obtained and presented in the next section concerning the inexact local solves are highly dependent on the particular implementation used and can not be considered to hold in general. A more efficient implementation might be possible.

5.4. Numerical Illustrations

In this section, we set forth a set of numerical experiments that showcase the theoretical results obtained throughout the chapter. We present experiments that illustrate the obtained error bounds for the case when the multiplicative Schwarz method is used as a solution method as well as normwise relative residual bounds for the case where the method is used as a preconditioner to GMRES.

We begin by studying the convergence behavior of the multiplicative Schwarz method applied to the Shishkin mesh discretization of the model problem (5.2) with

$$\omega_x = 0, \quad \omega_y = 1 \quad \beta = 0, \quad f(x) \equiv 0,$$

using upwind finite difference operators and boundary conditions described in Section 2.1.4. The analytic solution of this problem with $\epsilon = 0.01$ is shown in Figure 2.6. We use a self implemented version of the multiplicative Schwarz method which can be found in the `Github` repository which contains the software developed for this thesis (see Appendix A) and `MATLAB`'s implementation of the GMRES method callable with the command `gmres`.

We first consider $N = 30$ and $M = 40$ intervals in the corresponding $x-$ and $y-$coordinate directions of the mesh, obtaining a linear system with a coefficient matrix $\mathbf{A} \in \mathbb{R}^{(N-1)\times(M-1)}$ of size 1131 × 1131. Figures 5.1–5.2 show the error norms

$$\frac{\|\mathbf{e}^{(k)}\|_\infty}{\|\mathbf{e}^{(0)}\|_\infty}, \ k = 0, 1, 2, \dots,$$

for the iteration matrices $\mathbf{T}_{12} = \mathbf{Q}_2\mathbf{Q}_1$ and $\mathbf{T}_{21} = \mathbf{Q}_1\mathbf{Q}_2$ (solid lines) as well as the corresponding error bounds of Corollary 5.9 for increasing values of epsilon (markers). Once again, for our experiments we compute (an approximation to) the exact solution $\mathbf{u} = \mathbf{A}^{-1}\mathbf{f}$ using `MATLAB`'s backslash operator and apply the multiplicative Schwarz method with initial approximation $\mathbf{u}^{(0)} = 0$. Using the exact solution, we calculated the error norms of the multiplicative Schwarz method by $\|\mathbf{e}^{(k)}\|_\infty = \|\mathbf{u}^{(k+1)} - \mathbf{u}\|_\infty$ with $\mathbf{u}^{(k+1)}$ as in (2.70).

We observe that the bounds are extremely close to the actual errors produced by the method. Just like the bounds for the one-dimensional case, the bounds for this two-dimensional case also predict the initial stagnation phase of the multiplicative Schwarz method for the iteration matrix $\mathbf{T}_{21} = \mathbf{Q}_1\mathbf{Q}_2$, making the bounds not only good quantitative measures for the method's behavior but also provide a very good qualitative description of it (see Table 5.1). Although the quality of the bounds decreases as the problems become less convection dominated, the bounds given by (5.31) and (5.32) are still tight and descriptive for perturbation parameters up to $\epsilon \leq 10^{-2}$. We also run the experiments for larger values of N and M. The values of the convergence factor $|\rho|$ as given in (5.12) with $\|\cdot\| = \|\cdot\|_\infty$ and the corresponding bound from Corollary 5.9 are shown in Table 5.1 for different values of ϵ.

We continue our numerical experiments by applying GMRES to the linear algebraic system *preconditioned with multiplicative Schwarz*, i.e., the linear algebraic system

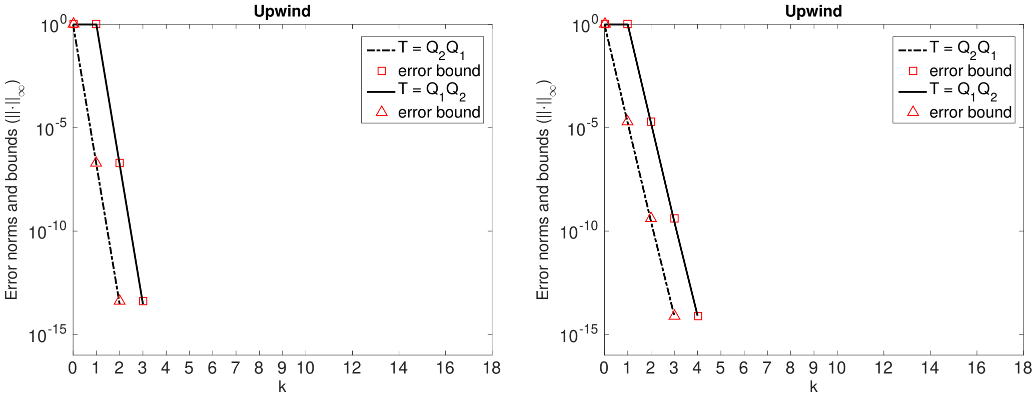


Figure 5.1.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-8}$ [l.], $\epsilon = 10^{-6}$ [r.], and $\mathbf{A} \in \mathbb{R}^{1131\times1131}$.

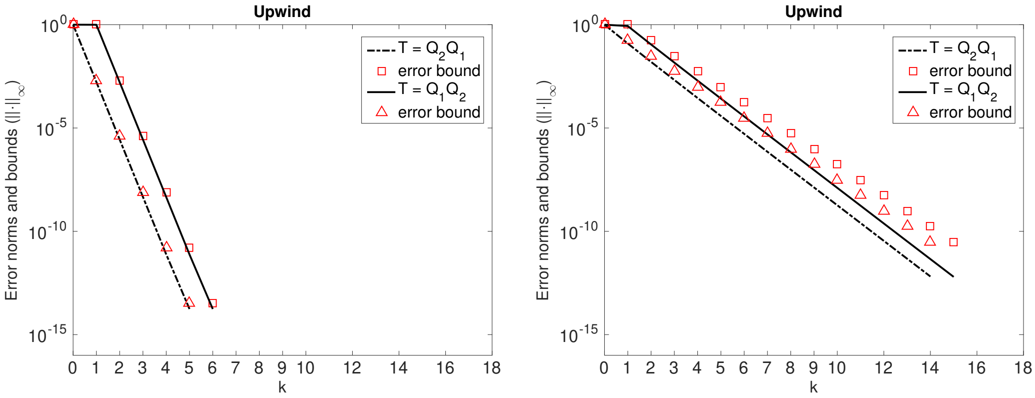


Figure 5.2.: Convergence of multiplicative Schwarz and error bounds for $\epsilon = 10^{-4}$ [l.], $\epsilon = 10^{-2}$ [r.], and $\mathbf{A} \in \mathbb{R}^{1131\times1131}$.

$N = 20, M = 20,\ \mathbf{A} \in \mathbb{R}^{361\times361}$		
ϵ	ρ_{12} in (5.12)	ρ in (5.32)
10^{-8}	7.5×10^{-8}	1.0×10^{-7}
10^{-6}	7.5×10^{-6}	1.0×10^{-5}
10^{-4}	7.5×10^{-4}	1.0×10^{-3}
10^{-2}	7.0×10^{-2}	9.6×10^{-2}
$N = 30, M = 30,\ \mathbf{A} \in \mathbb{R}^{841\times841}$		
10^{-8}	1.2×10^{-7}	1.5×10^{-7}
10^{-6}	1.2×10^{-5}	1.5×10^{-5}
10^{-4}	1.2×10^{-3}	1.5×10^{-3}
10^{-2}	1.1×10^{-1}	1.4×10^{-1}
$N = 40, M = 40,\ \mathbf{A} \in \mathbb{R}^{1521\times1521}$		
10^{-8}	1.7×10^{-7}	2.0×10^{-7}
10^{-6}	1.7×10^{-5}	2.0×10^{-5}
10^{-4}	1.7×10^{-3}	2.0×10^{-3}
10^{-2}	1.4×10^{-1}	1.8×10^{-1}
$N = 50, M = 50,\ A \in \mathbb{R}^{2401\times2401}$		
10^{-8}	2.2×10^{-7}	2.5×10^{-7}
10^{-6}	2.2×10^{-5}	2.5×10^{-5}
10^{-4}	2.2×10^{-3}	2.5×10^{-3}
10^{-2}	1.8×10^{-1}	2.1×10^{-1}

$N = 20, M = 30,\ \mathbf{A} \in \mathbb{R}^{551\times551}$		
ϵ	ρ_{12} in (5.12)	ρ in (5.32)
10^{-8}	1.2×10^{-7}	1.5×10^{-7}
10^{-6}	1.2×10^{-5}	1.5×10^{-5}
10^{-4}	1.2×10^{-3}	1.5×10^{-3}
10^{-2}	1.1×10^{-1}	1.4×10^{-1}
$N = 30, M = 40,\ \mathbf{A} \in \mathbb{R}^{1131\times1131}$		
10^{-8}	1.7×10^{-7}	2.0×10^{-7}
10^{-6}	1.7×10^{-5}	2.0×10^{-5}
10^{-4}	1.7×10^{-3}	2.0×10^{-3}
10^{-2}	1.4×10^{-1}	1.8×10^{-1}
$N = 40, M = 50,\ \mathbf{A} \in \mathbb{R}^{1911\times1911}$		
10^{-8}	2.2×10^{-7}	2.5×10^{-7}
10^{-6}	2.2×10^{-5}	2.5×10^{-5}
10^{-4}	2.2×10^{-3}	2.5×10^{-3}
10^{-2}	1.8×10^{-1}	2.1×10^{-1}
$N = 50, M = 60,\ \mathbf{A} \in \mathbb{R}^{2891\times2891}$		
10^{-8}	2.6×10^{-7}	3.0×10^{-7}
10^{-6}	2.6×10^{-5}	3.0×10^{-5}
10^{-4}	2.6×10^{-3}	3.0×10^{-3}
10^{-2}	2.1×10^{-1}	2.5×10^{-1}

Table 5.1.: Values of $|\rho_{12}|$ computed using (5.12) with $\|\cdot\| = \|\cdot\|_\infty$ and the corresponding bound (5.32) for different values of N, M and ϵ.

(5.33), in the case $N = 30$ and $M = 40$. On the right side of Figures 5.3–5.6 the preconditioned relative residual norms are shown for a specific value of ϵ and incresing values of the iteration step k. Convergence within a tolerance of 10^{-14} is always achieved in less than N iterations for the preconditioned systems, sometimes reaching the tolerance in only 2 iterations. This is a dramatic improvement compared to the behavior of the unpreconditioned GMRES method (shown in the left side of each figure) which does not converge to the same tolerance in less than 320 iterations for the chosen set of parameters.

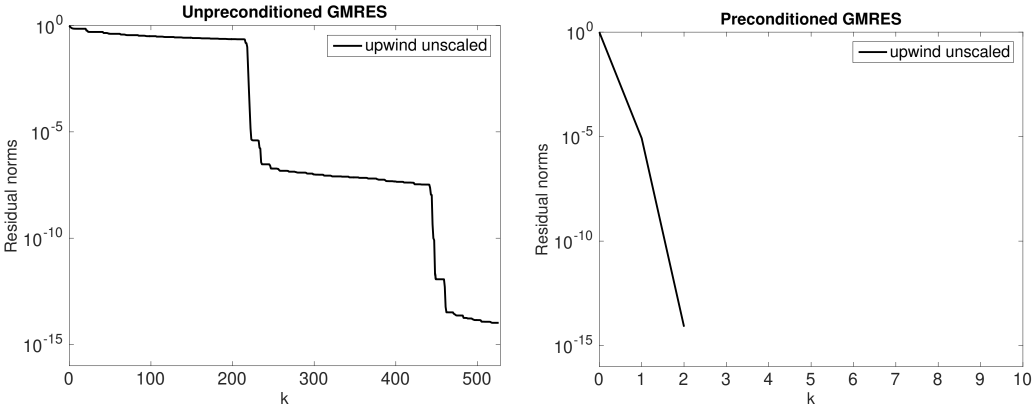


Figure 5.3.: Unpreconditioned [l.] and preconditioned [r.] GMRES convergence for $\epsilon = 10^{-8}$ and $\mathbf{A} \in \mathbb{R}^{1131 \times 1131}$.

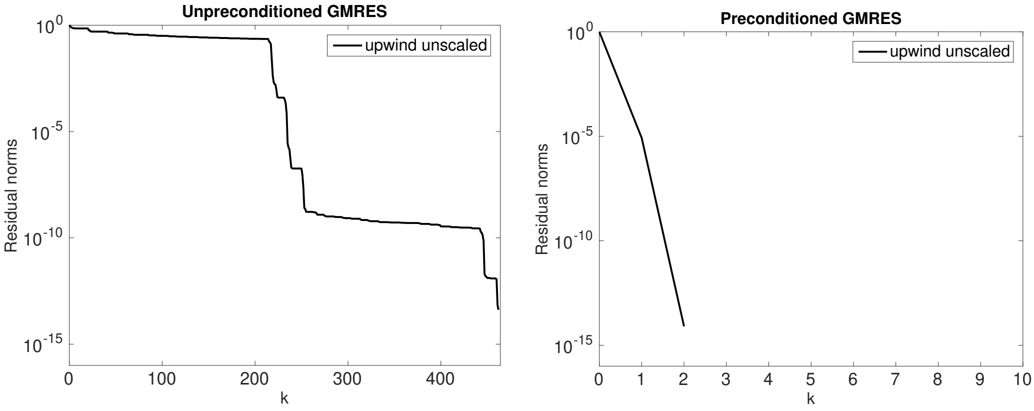


Figure 5.4.: Unpreconditioned [l.] and preconditioned [r.] GMRES convergence for $\epsilon = 10^{-6}$ and $\mathbf{A} \in \mathbb{R}^{1131 \times 1131}$.

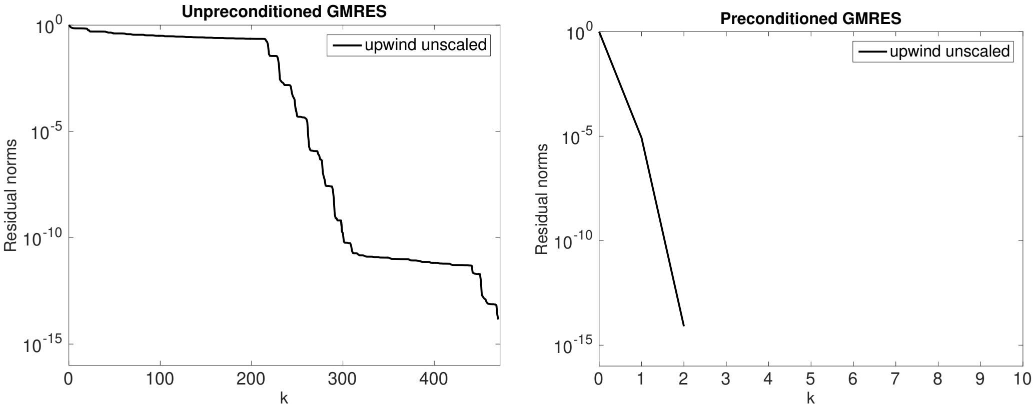


Figure 5.5.: Unpreconditioned [l.] and preconditioned [r.] GMRES convergence for $\epsilon = 10^{-4}$ and $\mathbf{A} \in \mathbb{R}^{1131 \times 1131}$.

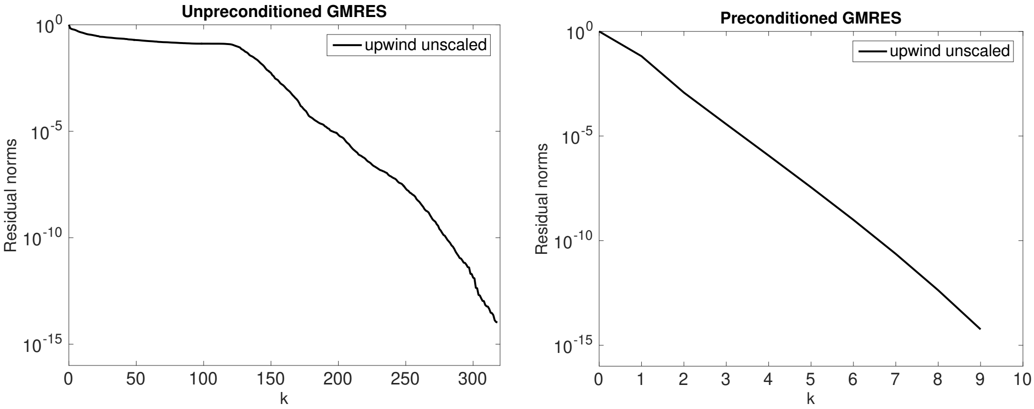


Figure 5.6.: Unpreconditioned [l.] and preconditioned [r.] GMRES convergence for $\epsilon = 10^{-2}$ and $\mathbf{A} \in \mathbb{R}^{1131 \times 1131}$.

In the numerical experiments presented so far, the local subdomain problems (5.4) were assumed to be solved exactly, since the inverses of the matrices $\mathbf{A}_i$ were calculated using the backslash operator in `MATLAB`. Nevertheless, as we have discussed previously (see Sections 5.3 and 3.3), very often in practice the solutions to linear systems with the matrices $\mathbf{A}_i$ are only solved approximately. We conclude our numerical experiments by presenting results for the preconditioned GMRES method for the case of inexact local solves. Table 5.2 shows the total number of iterations needed for the outer iteration to achieve a relative residual norm of 10^{-7} when the local subdomain problems are both solved inexactly as well as exactly for different values of ϵ, local solve tolerance and problem size.

outer iterations [exact/inexact(inner)]						
$\mathbf{A} \in \mathbb{R}^{1131\times1131}$						
ϵ \ tol	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10^{-8}	1/1(2)	1/1(2)	1/1(2)	1/1(2)	1/1(2)	1/1(2)
10^{-6}	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/2(129)
10^{-4}	2/1(2)	2/1(2)	2/1(2)	2/2(132)	2/2(175)	2/2(195)
10^{-2}	5/1(2)	5/5(290)	5/5(388)	5/5(534)	5/5(647)	5/5(651)
$\mathbf{A} \in \mathbb{R}^{2891\times2891}$						
10^{-8}	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)
10^{-6}	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(189)
10^{-4}	2/1(2)	2/1(2)	2/1(2)	2/2(196)	2/2(282)	2/2(288)
10^{-2}	6/1(2)	6/5(390)	6/6(573)	6/6(790)	6/6(966)	6/6(1117)
$\mathbf{A} \in \mathbb{R}^{4071\times4071}$						
10^{-8}	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)
10^{-6}	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/1(2)	2/2(219)
10^{-4}	2/1(2)	2/1(2)	2/1(2)	2/2(231)	2/1(327)	2/2(362)
10^{-2}	6/1(2)	6/5(435)	6/6(637)	6/6(843)	6/6(1067)	6/6(1240)

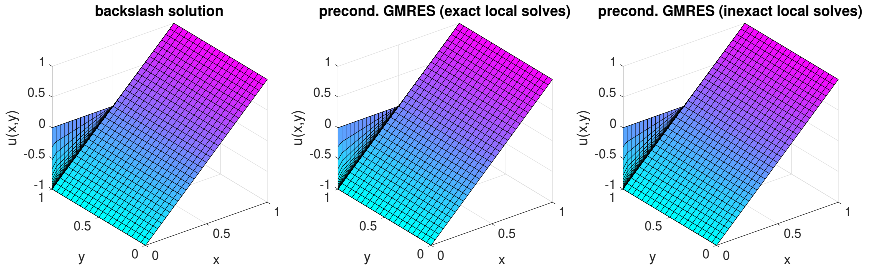
Table 5.2.: Total iteration count for solving problem (5.2) using GMRES with exact and inexact Shishkin-Schwarz preconditioning to reach a relative residual norm of 10^{-7} for different sizes, perturbation parameters and local solve tolerances.

The results of Table 5.2 show a surprising relation between the level of convection dominance and the accuracy of solution of the local subdomain problems. When the perturbation parameter is smaller than the local solve tolerance, the solution to the preconditioned system is reached in equal or less number of iterations than the case with exact local solves. When the local solve tolerance is chosen at the same level or smaller than the perturbation parameter the outer iterations remain less than or equal to the case with exact local solves, however, the number of inner iterations needed to solve the system increases greatly. Nevertheless, they may converge in less computational time if the saving from the inexact local solve is sufficiently large to offset the loss in convergence rate (the total computational time needed to achieve the desired solutions is shown in Table 5.3). Another surprising feature that can be seen in Table 5.2 is the fact that for all cases, the preconditioned GMRES method converges in at most 6 steps, when according to (5.34) we would expect it to converge in the order of $N - 1$ steps. Although, (5.34) is an upper bound, an exploration of the singular values of the matrix $\mathbf{T}_{ij}$ shows that indeed the first N singular values are comparable in size, however they are all one order of magnitude smaller than the the order of the perturbation parameter. Further experiments and analysis are needed to understand this phenomenon. We continue our analysis by showing the total computational time needed to reach the aforementioned solutions.

time [s.]									
$\mathbf{A} \in \mathbb{R}^{1131\times1131}$									
ϵ	$\backslash$	unprec	exact	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	10^{-6}
10^{-8}	0.0099	2.5303	0.0251	0.0795	0.0775	0.0668	0.0626	0.0779	0.0683
10^{-6}	0.0110	2.6841	0.0264	0.0602	0.0719	0.0716	0.0561	0.0797	0.3162
10^{-4}	0.0107	2.5351	0.0265	0.0718	0.0765	0.0602	0.3244	0.3952	0.4441
10^{-2}	0.0111	1.1483	0.0371	0.0536	0.6477	0.8750	1.1159	1.3260	1.3568
$\mathbf{A} \in \mathbb{R}^{2891\times2891}$									
10^{-8}	0.0384	25.2496	0.3096	0.3108	0.2579	0.2747	0.3008	0.3258	0.2869
10^{-6}	0.0432	30.6092	0.1846	0.2579	0.2447	0.2833	0.2490	0.2553	1.8440
10^{-4}	0.0285	25.1235	0.2372	0.2640	0.2680	0.2472	1.5442	2.0936	2.1142
10^{-2}	0.0299	5.3121	0.2377	0.2427	2.9197	4.0689	5.5994	6.7012	7.9111
$\mathbf{A} \in \mathbb{R}^{4071\times4071}$									
10^{-8}	0.0531	99.6026	0.6738	0.5405	0.5991	0.6309	0.6762	0.6097	0.5066
10^{-6}	0.0450	92.4812	0.4906	0.4900	0.4881	0.5883	0.4675	0.4564	2.9675
10^{-4}	0.0554	90.0034	0.4219	0.8435	0.4765	0.5237	3.2154	4.3034	4.7171
10^{-2}	0.0721	13.2551	0.6031	0.4286	5.5991	7.9189	10.2299	12.9283	15.1270

Table 5.3.: Total CPU time for solving problem (5.2) with `MATLAB`'s backslash and GMRES with and without Shishkin-Schwarz preconditioning for different problem sizes. Timings are reported for GMRES to reach a relative residual norm of 10^{-7} with exact and inexact preconditioning.

Table 5.3 shows that for all the problem sizes chosen for the experiments, the backslash solution is obtained very fast while the unpreconditioned GMRES method is very slow with increasing time as the convection dominance increases. As the problem size increases, the performance of GMRES deteriorates greatly, showing that for a much larger number of unknowns (when the backslash solution is no longer available) and a very small perturbation parameter, the unpreconditioned GMRES will very likely be innefficient. For the preconditioned system with exact local solves this is not the case, although still one order of magnitude slower than the backslash solution, we can see a more or less constant solution time for different values of the perturbation parameter. The benefits of the inexact local solve approach is shown by comparing the time it takes for the unpreconditioned GMRES method with any of the inexact local solve times - the speed up is always greater than three orders of magnitude, even for the case of a very low local solve tolerance.


backslash solution

precond. GMRES (exact local solves)

precond. GMRES (inexact local solves)

Figure 5.7.: Comparison of obtained solutions for $N = 30$, $M = 40$, $\epsilon = 10^{-8}$ and inexact local solve tolerance of 10^{-1}.

For the specific case of $\epsilon = 10^{-8}$, $N = 30$ and $M = 40$, Figure 5.7 shows a comparison of the obtained solutions for the exact case (backslash solution of (5.1)), and the cases of solving the preconditioned system with exact and inexact local solves (GMRES solution of (5.33)) with a tolerance of 10^{-1}. The accuracy of the obtained solutions is presented in Table 5.4, which shows the relative normwise error of the obtained solutions with respect to the exact solution obtained with the backslash operator for all the experiments presented in Tables 5.2 and 5.3.

relative error $[\|\mathbf{u}_* - \mathbf{u}_\backslash\|/\|\mathbf{u}_\backslash\|]$								
$\mathbf{A} \in \mathbb{R}^{1131\times1131}$								
ϵ	$\mathbf{u}_{unprec}$	$\mathbf{u}_{exact}$	$\mathbf{u}_{10^{-1}}$	$\mathbf{u}_{10^{-2}}$	$\mathbf{u}_{10^{-3}}$	$\mathbf{u}_{10^{-4}}$	$\mathbf{u}_{10^{-5}}$	$\mathbf{u}_{10^{-6}}$
10^{-8}	3.5×10^{-7}	8.4×10^{-8}	1.2×10^{-7}	1.2×10^{-7}	1.2×10^{-7}	1.2×10^{-7}	1.2×10^{-7}	1.2×10^{-7}
10^{-6}	2.8×10^{-9}	7.2×10^{-15}	1.2×10^{-5}	1.2×10^{-5}	1.2×10^{-5}	1.2×10^{-5}	1.2×10^{-5}	2.8×10^{-10}
10^{-4}	3.5×10^{-11}	7.7×10^{-9}	1.2×10^{-3}	1.2×10^{-3}	1.2×10^{-3}	2.8×10^{-6}	2.8×10^{-6}	2.3×10^{-6}
10^{-2}	3.5×10^{-13}	3.6×10^{-8}	8.8×10^{-2}	2.1×10^{-2}	2.7×10^{-3}	2.8×10^{-4}	2.4×10^{-5}	2.1×10^{-6}
$\mathbf{A} \in \mathbb{R}^{2891\times2891}$								
10^{-8}	1.9×10^{-6}	1.0×10^{-14}	1.8×10^{-7}	1.8×10^{-7}	1.8×10^{-7}	1.8×10^{-7}	1.8×10^{-7}	1.8×10^{-7}
10^{-6}	9.1×10^{-9}	2.8×10^{-14}	1.8×10^{-5}	1.8×10^{-5}	1.8×10^{-5}	1.8×10^{-5}	1.8×10^{-5}	6.6×10^{-10}
10^{-4}	8.5×10^{-11}	2.7×10^{-8}	1.8×10^{-3}	1.8×10^{-3}	1.8×10^{-3}	6.6×10^{-6}	6.6×10^{-6}	5.7×10^{-8}
10^{-2}	1.1×10^{-12}	1.7×10^{-8}	1.3×10^{-1}	3.1×10^{-2}	3.4×10^{-3}	4.0×10^{-4}	3.2×10^{-5}	4.9×10^{-6}
$\mathbf{A} \in \mathbb{R}^{4071\times4071}$								
10^{-8}	3.6×10^{-6}	1.0×10^{-14}	2.1×10^{-7}	2.1×10^{-7}	2.1×10^{-7}	2.1×10^{-7}	2.1×10^{-7}	2.1×10^{-7}
10^{-6}	1.3×10^{-8}	4.6×10^{-14}	2.1×10^{-5}	2.1×10^{-5}	2.1×10^{-5}	2.1×10^{-5}	2.1×10^{-5}	9.1×10^{-10}
10^{-4}	1.8×10^{-10}	4.1×10^{-8}	2.1×10^{-3}	2.1×10^{-3}	2.1×10^{-3}	9.1×10^{-6}	9.1×10^{-6}	9.5×10^{-8}
10^{-2}	1.9×10^{-12}	4.2×10^{-8}	1.4×10^{-1}	4.0×10^{-2}	5.2×10^{-3}	4.7×10^{-4}	5.2×10^{-5}	6.1×10^{-6}

Table 5.4.: Relative error $\|\mathbf{u}_* - \mathbf{u}_\backslash\|/\|\mathbf{u}_\backslash\|$, with $\|\cdot\| = \|\cdot\|_2$ for each solution approach

For a fixed level of convection dominance (ϵ), Table 5.4 shows that, on the one hand, solving the preconditioned system with exact local solves delivers a more accurate solution than solving the unpreconditioned system using GMRES when the convection dominance is high (less than 10^{-4}) and the opposite is true for low convection dominance (comparison of columns 1 & 2). Furthermore, solving the system with inexact local solves always delivers a less accurate solution (at least one order of magnitude worse) than using exact local solves, neverthless, the difference diminishes as the convection dominance is increased (comparison of column 2 with columns 3-7).

6. Discussion and Outlook

Motivated by the challenge of understanding and analyzing the convergence of the multiplicative Schwarz method for solving linear algebraic systems arising from a “simple” and problem-specific one-dimensional model problem, various results and contributions presented in this thesis are valid in a much broader and general context. This concluding section provides a very brief discussion on the consequences brought forth by the main results found in this work and points out possibilities for further investigations that naturally arise from the analysis and experiments present in this thesis.

In the context of Schwarz solution methods, our main contributions include expressions for the convergence factor of the multiplicative Schwarz method when used to solve linear algebraic systems with a special block structure arising from problems in one- and two-dimensions (see Corollary 3.2 and Lemma 5.3). The specific structure present in the coefficient matrix is brought about when the finite difference or the finite element method are used to discretize any second-order constant-coefficient elliptic partial differential equation which is posed on a very broad and common domain decomposition context: that of a domain being subdivided into two smaller subdomains with an overlap between them. By making further assumptions on the system matrix, mainly that it possesses a block-tridiagonal structure, we provide quantitative convergence bounds for the error of the method in terms of the norms of the off-diagonal blocks of the original system matrix (see Theorems 5.6 and 5.7). The provided bounds are valid from the first step of the iteration process, presenting a grand improvement in contrast with the classic convergence theory for multiplicative Schwarz methods, which is based on asymptotic convergence rates and is not able to describe the transient phase of the iteration scheme, where most problems like stagnation of the method might occur. Furthermore, our analysis does not lean on any of the usual assumptions put on the matrices needed to prove convergence in the classical domain decomposition convergence theory, such as symmetry, or the M- or H-matrix properties, making our approach much more general in its range of applications.

Contributions to the general area of preconditioning are also present in this work. Results on the convergence theory of the preconditioned GMRES method, given in the form of bounds on the residual norm of the iterates of the method, are provided for the case when the multiplicative Schwarz method is used as a preconditioner. The bounds are given in terms of the rank of the iteration matrix of the multiplicative Schwarz method, which in the domain decomposition context, we have proven to be of low-rank (see Lemma 5.1). Moreover this result might be extended to the case when other fixed point iteration methods are used as preconditioners in particular scenarios, since the convergence of the GMRES method in a small number of steps is only linked to the low rank structure of the iteration matrix of the method. In the specific case presented in this thesis, bounding the rank of the iteration matrix of the multiplicative Schwarz method allows us to prove rapid convergence of the preconditioned GMRES method, an approach that can be useful whenever the iteration matrices of other fixed point iteration methods posses a low rank structure.

The specific domain decomposition setting used to present our theory, which mimics the one fist studied by Schwarz at the end of the 19th century and consists of only two overlapping subdomains, reduces the technicalities present in the analysis needed to obtain the results, nevertheless, the arguments used to construct the convergence theory can be applied in a recursive fashion. It is easy to imagine each of the subdomains being subdivided once again into two smaller subdomains and for this process to be repeated until a desired finite number of subdivisions is reached in each of the original subdomains. This, of course, has immediate potential in the implementation of the method for parallel computing systems when a low memory storage prevents the modeling of large domains in a single computing station.

In the specific case where the multiplicative Schwarz method is used to solve systems coming from discretizations of the the convection-diffusion equation we have shown that the approach of using a one- or two-dimensional Shishkin mesh together with a finite difference approximation of the derivatives is very effective in terms of the number of iterations that the method needs to converge to an accurate solution, since in this context the iteration matrix has low rank. Not only that, but our analysis clarifies an apparent contradiction in the behavior of the method when used to solve these type of problems, where we observe that as the problem becomes harder (we choose a smaller perturbation parameter) the convergence of the multiplicative Schwarz method becomes faster and more effective (see Theorems 3.6 and 3.13 and Corollary 5.9). The effectivity of the bounds was illustrated numerically on one- and two-dimensional convection-diffusion model problems, showing that the bounds are extremely close to the actual errors produced by the method in both cases, becoming more accurate as the problems become more convection dominated.

The most general results presented in this work fall under the broad study area of linear algebra and include a generalization of the classical definition of block diagonal dominance of matrices given by Feingold and Varga in [30] (note that a very similar definition to the one presented in this thesis appeared on the same year in [4]), which brings new understanding of the concept of diagonal dominance when dealing with operators with a block structure (see Definitions 4.1 and 4.19). In the classical definitions of block diagonal dominance, the blocks are “treated as scalars”, however in our approach the block diagonal dominance is based on the influence of the blocks as matrices, i.e., the action of the block is taken into account not just their norms. This new definition and its consequences have proven to be useful tools in the analysis of the multiplicative Schwarz method and might very well be useful in analyzing other iterative methods, specially in the context of fixed point iteration methods. Moreover, the generalization of block diagonal dominance provided in this work implies new results in the spectral theory of eigenvalues, for example, it infers new eigenvalues inclusion sets for general block matrices which are potentially tighter that the classical ones given by the previous definitions of diagonal dominance or the Gershgorin's Circle Theorem (see Corollary 4.10).

Furthermore, based on our definition we established upper and lower bounds and decay rates on the block norms of the inverse of block diagonally dominant block tridiagonal matrices (see Theorems 4.6 and 4.7 and Corollary 4.8), which are a direct generalization of the bounds presented in [62] for the scalar case. The bounds may be extremely useful for the creation of preconditioners for large linear systems when the system matrices are block tridiagonal. The off-diagonal decay of the block norms of its inverse matrix implied by the bounds allows for the creation of approximate inverses by neglecting blocks that fall bellow a desired norm threshold.

Although the results provided in this thesis focus on the study and generalization of simple model problems, natural extensions of our approach to more general settings are still possible. We will now briefly mention a number of generalizations and alternative applications to our analysis that may be explored in future research.

Our analysis stems from on a very specific domain decomposition problem setting which translates into solving linear systems with coefficient matrices exhibiting a specific block structure. The problem setting which was chosen is the simplest one possible and the main assumption that guided our analysis was the fact that in each of the two local subdomains, we use the same number of unknowns. This assumption was helpful in reducing technicalities present in our analysis, however, by relaxing this and other different constraints, we can envision the same approach being applied to matrices $\mathbf{A}$ with more general structures.

It would be only natural to apply our approach in a setting where the local subdomains have a different number of unknowns. This would be reflected in a coefficient matrix $\mathbf{A}$ with blocks $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ of variable size in (5.3). An analysis of the problem in this more general context would be filled with more technicalities, however it would be certainly possible to carry out, specially when the sizes of the matrices are multiples of N, i.e. for matrices $\widehat{\mathbf{A}}_H \in \mathbb{R}^{Ns \times Ns}$ and $\widehat{\mathbf{A}}_h \in \mathbb{R}^{Nt \times Nt}$. It is also easy to think of applying our analysis to the solution of discretized of PDEs with variable coefficients. This change would be reflected in the structure of the matrices $\widehat{\mathbf{A}}_H$ and $\widehat{\mathbf{A}}_h$ by exhibiting general tridiagonal blocks of the form

$$\widehat{\mathbf{A}}_H = \text{tridiag}(C_{H,i}, A_{H,i}, B_{H,i}) \quad \text{and} \quad \widehat{\mathbf{A}}_h = \text{tridiag}(C_{h,i}, A_{h,i}, B_{h,i}), \tag{6.1}$$

instead of tridiagonal Toeplitz matrices of the form (5.14). The analysis would still be possible since the theory of block diagonal dominance presented in Chapter 4 used to obtain the bounds allows for relaxing this constraint. A generalization of 5.7 to $\mathbf{A}$ with blocks (6.1) would require that the conditions (5.16) hold in every block row, and then analogously to (5.18)–(5.19), every block row in $\widehat{\mathbf{A}}_H$ or $\widehat{\mathbf{A}}_h$ would give a parameter $\eta_{H,i}$ or $\eta_{h,i}$, respectively. For convection-diffusion problems, the structure of the matrix would also be affected by the number of boundary layers present in the problem. We would have more than one overlap region in the domain and the iteration matrixes would present different more complicated nonzero block structures affecting the rank of the iteration matrix. However, preliminary numerical experiments show that the iteration matrices still possess a low numerical rank (close to N). Thus, we believe that it is still analyzable along the lines of Section 5.2.1.

A. Documentation of the Software

All numerical experiments presented in this thesis can be reproduced with the freely available software developed by the author specifically for this work. The LaTeX source code used to compile this document as well as the `MATLAB` `m`-files to run the numerical experiments can be obtained from the `GitHub` repository found in:

`https://github.com/carlos-echeverria/phd-thesis`

To reproduce the figures and tables please refer to the following guidelines. All `m`-files are individually documented; see the header of each file for in-detail information.

- To reproduce Figures 2.8–2.9 use the script:
  `MATLAB_CODE/CH3/produce_figures_gmres_without_preconditioning.m`

- To reproduce Figures 3.1–3.3 use the script:
  `MATLAB_CODE/CH3/produce_figures_mSm_upwind_and_central.m`

- To reproduce Figures 3.4–3.5 use the above mentioned script with the parameter `N` changed from `N=198` to `N=10002`.

- Table 3.1 can be reproduced by choosing the corresponding value of the parameter `N` in the previous script and reading off the values in `MATLAB`'s console.

- To reproduce Figures 3.6–3.7 use the script:
  `MATLAB_CODE/CH3/produce_figures_gmres_with_preconditioning.m`

- To reproduce the images and tables from Examples 4.11–4.14 use the script:
  `MATLAB_CODE/CH4/SCRIPT.m`

  The different examples can be computed by choosing the parameter `matA` from 1 to 4. The tables will be given as output to the console once the parameter `refinement_step` is chosen to be 8.

- To reproduce the images and tables from Examples 4.15–4.18 use the script:
  `MATLAB_CODE/CH4/SCRIPT_INCLUSION_SETS.m`

  The different examples can be computed by choosing the parameters `Achoice`, `N` and `M` accordingly (see header of m-files).

- To reproduce Figures 5.1–5.2 use the script:
  `MATLAB_CODE/CH5/produce_figures_2D_mSm.m`

- To reproduce Figures 5.3–5.6 use the script:
  `MATLAB_CODE/CH5/produce_figures_2D_gmres.m`

- To reproduce Table 5.1 use the script:
  `MATLAB_CODE/CH5/produce_table_2D_rho_upwind.m`

  The values of the table can then be read off `MATALB`'s console window.

- To reproduce Tables 5.2 5.3 and 5.4 as well as Figure 5.7 use the script:
  `MATLAB_CODE/CH5/produce_tables_2D_inexact_upwind.m`

  The values of each table can then be read off `MATALB`'s console window.

WARNING: since maintaining code is an ever evolving endeavor, the above guidelines might be outdated. As stated above, please refer to the header of each individual file for up-to-date information. If any errors, bugs, or typos are found in the provided source code, please make a pull request in the `Github` repository after fixing, or send an email to: `echeverriacarlos@gmail.com` describing the error/bug/typo. Thank you.

Bibliography

[1] V. B. Andreyev and N. V. Kopteva. “A study of difference schemes with the first derivative approximated by a central difference ratio”. In: *Comput. Math. Math. Phys.* 36.8 (1996), pp. 1065–1078.

[2] M. Benzi. “Localization in Matrix Computations: Theory and Applications”. In: *Exploiting Hidden Structure in Matrix Computations: Algorithms and Applications.* Vol. 2173. Lecture Notes in Math. Springer, Cham, 2016, pp. 211–317.

[3] M. Benzi and P. Boito. “Decay properties for functions of matrices over C^*-algebras”. In: *Linear Algebra Appl.* 456 (2014), pp. 174–198.

[4] M. Benzi, T. M. Evans, S. P. Hamilton, M. Lupo Pasini, and S. R. Slattery. “Analysis of Monte Carlo accelerated iterative methods for sparse linear systems”. In: *Numer. Linear Algebra Appl.* 24.3 (2017), e2088, 18.

[5] M. Benzi, A. Frommer, R. Nabben, and D. B. Szyld. “Algebraic theory of multiplicative Schwarz methods”. In: *Numer. Math.* 89.4 (2001), pp. 605–639.

[6] M. Benzi and G. H. Golub. “Bounds for the entries of matrix functions with applications to preconditioning”. In: *BIT* 39.3 (1999), pp. 417–438.

[7] M. Benzi and M. A. Olshanskii. “An augmented Lagrangian-based approach to the Oseen problem”. In: *SIAM J. Sci. Comput.* 28.6 (2006), pp. 2095–2113.

[8] M. Benzi and N. Razouk. “Decay bounds and $O(n)$ algorithms for approximating functions of sparse matrices”. In: *Electron. Trans. Numer. Anal.* 28 (2007/08), pp. 16–39.

[9] M. Benzi and V. Simoncini. “Decay bounds for functions of Hermitian matrices with banded or Kronecker structure”. In: *SIAM J. Matrix Anal. Appl.* 36.3 (2015), pp. 1263–1282.

[10] M. Benzi and A. J. Wathen. “Some preconditioning techniques for saddle point problems”. In: *Model Order Reduction: Theory, Research Aspects and Applications.* Vol. 13. Math. Ind. Springer, Berlin, 2008, pp. 195–211.

[11] A. Berman and R. J. Plemmons. *Nonnegative Matrices in the Mathematical Sciences.* Vol. 9. Classics in Applied Mathematics. Revised reprint of the 1979 original. SIAM, Philadelphia, PA, 1994, pp. xx+340.

[12] D. Braess. *Finite elements.* Third. Theory, fast solvers, and applications in elasticity theory, Translated from the German by Larry L. Schumaker. Cambridge University Press, Cambridge, 2007, pp. xviii+365.

[13] J. H. Bramble, J. E. Pasciak, and A. T. Vassilev. “Analysis of non-overlapping domain decomposition algorithms with inexact solves”. In: *Math. Comp.* 67.221 (1998), pp. 1–19.

[14] R. Bru, F. Pedroche, and D. B. Szyld. “Overlapping additive and multiplicative Schwarz iterations for H-matrices”. In: *Linear Algebra Appl.* 393 (2004), pp. 91–105.

[15] C. Canuto, V. Simoncini, and M. Verani. “On the decay of the inverse of matrices that are sum of Kronecker products”. In: *Linear Algebra Appl.* 452 (2014), pp. 21–39.

[16] M. Capovani. “Su alcune proprietà delle matrici tridiagonali e pentadiagonali”. In: *Calcolo* 8 (1971), pp. 149–159.

[17] M. Capovani. “Sulla determinazione della inversa delle matrici tridiagonali e tridiagonali a blocchi”. In: *Calcolo* 7 (1970), pp. 295–303.

[18] A. Chang et al. “A time fractional convection-diffusion equation to model gas transport through heterogeneous soil and gas reservoirs”. In: *Phys. A* 502 (2018), pp. 356–369.

[19] S. Demko, W. F. Moss, and P. W. Smith. “Decay rates for inverses of band matrices”. In: *Math. Comp.* 43.168 (1984), pp. 491–499.

[20] J. W. Demmel. *Applied Numerical Linear Algebra.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, pp. xii+419.

[21] V. Dolean, P. Jolivet, and F. Nataf. *An Introduction to Domain Decomposition Methods: Algorithms, Theory, and Parallel Implementation.* SIAM, Philadelphia, PA, 2015, pp. x+238.

[22] C. Echeverría, J. Liesen, and R. Nabben. “Block diagonal dominance of matrices revisited: bounds for the norms of inverses and eigenvalue inclusion sets”. In: *Linear Algebra Appl.* 553 (2018), pp. 365–383.

[23] C. Echeverría, J. Liesen, D. B. Szyld, and P. Tichý. “Convergence of the multiplicative Schwarz method for singularly perturbed convection-diffusion problems discretized on a Shishkin mesh”. In: *Electron. Trans. Numer. Anal.* 48 (2018), pp. 40–62.

[24] C. Echeverría, J. Liesen, and P. Tichý. “Analysis of the multiplicative Schwarz method for matrices with a special block structure”. In: *arXiv e-prints*, arXiv:1912.09107 (Dec. 2019), arXiv:1912.09107. arXiv: `1912.09107 [math.NA]`.

[25] H. C. Elman, D. J. Silvester, and A. J. Wathen. *Finite elements and fast iterative solvers: with applications in incompressible fluid dynamics.* 2nd edition. Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2014, pp. xiv+479.

[26] O. G. Ernst. “Residual-minimizing Krylov subspace methods for stabilized discretizations of convection-diffusion equations”. In: *SIAM J. Matrix Anal. Appl.* 21.4 (2000), pp. 1079–1101.

[27] L. C. Evans. *Partial Differential Equations.* 2nd edition. Vol. 19. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, 2010, pp. xxii+749.

[28] P. A. Farrell, A. F. Hegarty, J. J. H. Miller, E. O’Riordan, and G. I. Shishkin. *Robust Computational Techniques for Boundary Layers.* Vol. 16. Applied Mathematics (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, 2000, pp. xvi+254.

[29] P. A. Farrell and G. I. Shishkin. “Schwartz Methods for singularly perturbed convection-diffusion problems”. In: *Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems.* Ed. by J. J. H. Miller, G. I. Shishkin, and L. Vulkov. Nova Science Publishers New York, 2000, pp. 33–42.

[30] D. G. Feingold and R. S. Varga. “Block diagonally dominant matrices and generalizations of the Gerschgorin circle theorem”. In: *Pacific J. Math.* 12 (1962), pp. 1241–1250.

[31] M. Fiedler and V. Pták. “Generalized norms of matrices and the location of the spectrum”. In: *Czechoslovak Math. J.* 12 (87) (1962), pp. 558–571.

[32] A. Frommer, R. Nabben, and D. B. Szyld. “Convergence of stationary iterative methods for Hermitian semidefinite linear systems and applications to Schwarz methods”. In: *SIAM J. Matrix Anal. Appl.* 30.2 (2008), pp. 925–938.

[33] A. Frommer and D. B. Szyld. “On necessary conditions for convergence of stationary iterative methods for Hermitian semidefinite linear systems”. In: *Linear Algebra Appl.* 453 (2014), pp. 192–201.

[34] M. J. Gander. “Schwarz methods over the course of time”. In: *Electron. Trans. Numer. Anal.* 31 (2008), pp. 228–255.

[35] M. J. Gander, S. Loisel, and D. B. Szyld. “An optimal block iterative method and preconditioner for banded matrices with applications to PDEs on irregular domains”. In: *SIAM J. Matrix Anal. Appl.* 33.2 (2012), pp. 653–680.

[36] M. J. Gander and G. Wanner. “The origins of the alternating Schwarz method”. In: *Domain decomposition methods in science and engineering XXI.* Vol. 98. Lect. Notes Comput. Sci. Eng. Springer, Cham, 2014, pp. 487–495.

[37] A. Gaul. “Recycling Krylov subspace methods for sequences of linear systems: Analysis and applications”. Doctoral Thesis. Berlin: Technische Universität Berlin, Fakultät II - Mathematik und Naturwissenschaften, 2014.

[38] W. Gray and M. Kostin. “Natural convection, diffusion and chemical reaction in a catalytic reactor: numerical results”. In: *The Chemical Engineering Journal* 8.1 (1974), pp. 1–10.

[39] A. Greenbaum. *Iterative methods for solving linear systems.* Vol. 17. Frontiers in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1997, pp. xiv+220.

[40] A. Greenbaum, V. Pták, and Z. Strakoš. "Any nonincreasing convergence curve is possible for GMRES". In: *SIAM J. Matrix Anal. Appl.* 17.3 (1996), pp. 465–469.

[41] D. F. Griffiths, J. W. Dold, and D. J. Silvester. *Essential Partial Differential Equations.* Springer Undergraduate Mathematics Series. Analytical and Computational Aspects. Springer, Cham, 2015, pp. xi+368.

[42] W. Hackbusch. *Elliptic Differential Equations.* English. Vol. 18. Springer Series in Computational Mathematics. Theory and numerical treatment, Translated from the 1986 corrected German edition by Regine Fadiman and Patrick D. F. Ion. Springer-Verlag, Berlin, 2010, pp. xiv+311.

[43] R. A. Horn and C. R. Johnson. *Matrix Analysis.* 2nd edition. Cambridge University Press, Cambridge, 2013, pp. xviii+643.

[44] R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis.* Cambridge University Press, Cambridge, 1991, pp. viii+607.

[45] Y. Ikebe. "On inverses of Hessenberg matrices". In: *Linear Algebra Appl.* 24 (1979), pp. 93–97.

[46] G. A. A. Kahou, E. Kamgnia, and B. Philippe. "An explicit formulation of the multiplicative Schwarz preconditioner". In: *Appl. Numer. Math.* 57.11-12 (2007), pp. 1197–1213.

[47] N. Kopteva and E. O'Riordan. "Shishkin meshes in the numerical solution of singularly perturbed differential equations". In: *Int. J. Numer. Anal. Model.* 7.3 (2010), pp. 393–415.

[48] I. Krishtal, T. Strohmer, and T. Wertz. "Localization of matrix factorizations". In: *Found. Comput. Math.* 15.4 (2015), pp. 931–951.

[49] P. J. Lanzkron, D. J. Rose, and D. B. Szyld. "Convergence of nested classical iterative methods for linear systems". In: *Numer. Math.* 58.7 (1991), pp. 685–702.

[50] J. Liesen and Z. Strakoš. "GMRES convergence analysis for a convection-diffusion model problem". In: *SIAM J. Sci. Comput.* 26.6 (2005), pp. 1989–2009.

[51] J. Liesen and Z. Strakoš. *Krylov Subspace Methods: Principles and Analysis.* Numerical Mathematics and Scientific Computation. Oxford University Press, Oxford, 2013, pp. xvi+391.

[52] T. Linß and M. Stynes. "Numerical methods on Shishkin meshes for linear convection-diffusion problems". In: *Comput. Methods Appl. Mech. Engrg.* 190.28 (2001), pp. 3527–3542.

[53] T. Linß. *Layer-adapted meshes for reaction-convection-diffusion problems.* Vol. 1985. Lecture Notes in Mathematics. Springer-Verlag, Berlin, 2010, pp. xii+320.

[54] H. MacMullen, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. "A second-order parameter-uniform overlapping Schwarz method for reaction-diffusion problems with boundary layers". In: *J. Comput. Appl. Math.* 130.1-2 (2001), pp. 231–244.

[55] H. MacMullen, J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. "Overlapping Schwartz Method for singularly perturbed convection-diffusion problems with boundary layers". In: *Analytical and Numerical Methods for Convection-Dominated and Singularly Perturbed Problems.* Ed. by J. J. H. Miller, G. I. Shishkin, and L. Vulkov. Nova Science Publishers New York, 2000, pp. 33–42.

[56] H. MacMullen, E. O'Riordan, and G. I. Shishkin. "Schwarz methods for convection-diffusion problems". In: *Proceedings of the 2nd International Conference on Numerical Analysis and Its Applications.* Ed. by L. Vulkov, J. Wasniewski, and P. Yamalov. Springer. 2000, pp. 544–551.

[57] H. MacMullen, E. O'Riordan, and G. I. Shishkin. "The convergence of classical Schwarz methods applied to convection-diffusion problems with regular boundary layers". In: *Appl. Numer. Math.* 43.3 (2002), pp. 297–313.

[58] P. A. Markowich and P. Szmolyan. "A system of convection-diffusion equations with small diffusion coefficient arising in semiconductor physics". In: *J. Differential Equations* 81.2 (1989), pp. 234–254.

[59] J. J. H. Miller, E. O'Riordan, and G. I. Shishkin. *Fitted Numerical Methods for Singular Perturbation Problems: Error Estimates in the Maximum Norm for Linear Problems in One and Two Dimensions.* Revised. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ, 2012, pp. xiv+176.

[60] K. W. Morton. *Numerical solution of convection-diffusion problems.* Vol. 12. Applied Mathematics and Mathematical Computation. Chapman & Hall, London, 1996, pp. xii+372.

[61] R. Nabben. "Decay rates of the inverse of nonsymmetric tridiagonal and band matrices". In: *SIAM J. Matrix Anal. Appl.* 20.3 (1999), pp. 820–837.

[62] R. Nabben. "Two-sided bounds on the inverses of diagonally dominant tridiagonal matrices". In: *Linear Algebra Appl.* 287.1-3 (1999). Special issue celebrating the 60th birthday of Ludwig Elsner, pp. 289–305.

[63] A. M. Ostrowski. "On some metrical properties of operator matrices and matrices partitioned into blocks". In: *J. Math. Anal. Appl.* 2 (1961), pp. 161–209.

[64] J. Papež, J. Liesen, and Z. Strakoš. "Distribution of the discretization and algebraic error in numerical solution of partial differential equations". In: *Linear Algebra Appl.* 449 (2014), pp. 89–114.

[65] R. Peluso and T. Politi. “Some improvements for two-sided bounds on the inverse of diagonally dominant tridiagonal matrices”. In: *Linear Algebra Appl.* 330.1-3 (2001), pp. 1–14.

[66] H.-G. Roos. “A note on the conditioning of upwind schemes on Shishkin meshes”. In: *IMA J. Numer. Anal.* 16.4 (1996), pp. 529–538.

[67] H.-G. Roos, M. Stynes, and L. Tobiska. *Robust numerical methods for singularly perturbed differential equations*. 2nd edition. Vol. 24. Springer Series in Computational Mathematics. Convection-diffusion-reaction and flow problems. Springer-Verlag, Berlin, 2008, pp. xiv+604.

[68] Y. Saad and M. H. Schultz. “GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems”. In: *SIAM J. Sci. Statist. Comput.* 7.3 (1986), pp. 856–869.

[69] Y. Saad. *Iterative Methods for Sparse Linear Systems*. 2nd edition. Philadelphia, U.S.A.: Society for Industrial and Applied Mathematics, 2003.

[70] H. A. Schwarz. “Ueber einige Abbildungsaufgaben”. In: *J. Reine Angew. Math.* 70 (1869), pp. 105–120.

[71] B. F. Smith, P. E. Bjørstad, and W. D. Gropp. *Domain decomposition*. Parallel multilevel methods for elliptic partial differential equations. Cambridge University Press, Cambridge, 1996, pp. xii+224.

[72] G. W. Stewart. *Matrix algorithms. Vol. II*. Eigensystems. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2001, pp. xx+469.

[73] M. Stynes. “Numerical methods for convection-diffusion problems or The 30 years war”. In: *arXiv e-prints*, arXiv:1306.5172 (June 2013), arXiv:1306.5172. arXiv: `1306.5172 [math.NA]`.

[74] M. Stynes. “Steady-state convection-diffusion problems”. In: *Acta Numer.* 14 (2005), pp. 445–508.

[75] D. B. Szyld. “The many proofs of an identity on the norm of oblique projections”. In: *Numer. Algorithms* 42.3-4 (2006), pp. 309–323.

[76] A. Toselli and O. Widlund. *Domain Decomposition Methods—Algorithms and Theory*. Vol. 34. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2005, pp. xvi+450.

[77] R. A. Usmani. “Inversion of Jacobi’s tridiagonal matrix”. In: *Comput. Math. Appl.* 27.8 (1994), pp. 59–66.

[78] R. S. Varga. *Gershgorin and His Circles*. Vol. 36. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2004, pp. x+226.

[79] R. S. Varga. *Matrix Iterative Analysis*. Expanded edition. Vol. 27. Springer Series in Computational Mathematics. Springer-Verlag, Berlin, 2000, pp. x+358.