Space-Time Spectral Methods for Partial Differential Equations

by

Avleen Kaur

A thesis submitted to the Faculty of Graduate Studies of
The University of Manitoba
in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics
University of Manitoba
Winnipeg

Abstract

Spectral methods for solving partial differential equations (PDEs) depict a high order of convergence, which is exponential when the solution is analytic. However, their applications to time-dependent PDEs typically enforce a finite difference scheme in time. The slower decay of error in time overwhelms the super-algebraic convergence of error in space. A relatively new class of techniques is space-time spectral methods converging spectrally in both space and time. We devise and analyze a space-time spectral method for the Stokes problem. The main objectives of the research are estimating the condition number of the global spectral operators and proving the spectral convergence of this scheme in space and time. Numerical experiments of this scheme verify the theoretical results. Furthermore, we discuss two space-time spectral methods for the Navier-Stokes problem.

The discrete systems resulting from classical space-time spectral methods are dense, ill-conditioned, and coupled in all time steps. A new class of spectral methods, called the ultraspherical spectral (US) methods, are applied to time-dependent PDEs, which along with spectral convergence, lead to the resultant discrete systems constituting sparse and well-conditioned matrices.

Additionally, we join the long tradition of estimating the eigenvalues of a sum of two symmetric matrices, say $P + Q$, in terms of the eigenvalues of P and Q. We derive two new lower bounds on $\lambda_{\min}(P + Q)$ in terms of the minimum positive eigenvalues of P and Q. The bounds incorporate geometric information by utilizing the Friedrichs angles between certain subspaces. Such estimates lead to new lower bounds on the minimum singular value of some full-rank block matrices in terms of the minimum positive singular value of their subblocks.

Acknowledgements

I sincerely thank my supervisor, Dr. Shaun Lui, with whom I have worked for the last seven years. He taught me to have perseverance and patience in research. He provided me with careful advice, encouragement, and feedback with kindness. He gave me a lot of flexibility and time in conducting research, which was crucial for my academic and personal development. I have taken a lot of his time in the past seven years, and I thank him for investing it in me so generously.

Dr. Craig Cowan was a co-supervisor in my M.Sc. program. I took my first course with him in Fall 2015, which he taught wonderfully, increasing my interest in partial differential equations. I will forever be grateful for his incredible support and understanding. I am thankful for the discussions on mathematics with Dr. Richard Mikaël Slevinsky, which were motivating. His passion for research is inspirational, and I thank him for his expert opinion and helpful comments. I thank Dr. Bing-Chen Wang for his reflections that have been beneficial for me and helped me explore applications. Some special thanks go to Dr. Raphaël Clouâtre, who uplifted me to have confidence in my abilities. I offer my thanks to my past supervisors, Drs. Gorakh Nath and Utpal Manna, for the much-needed guidance and support during my early years of graduate studies. I feel fortunate to have teachers from high school who introduced me to a Ph.D. program and stayed in touch with me throughout this long journey.

I thank Drs. Derek Krepski, Donald Trim, Michelle Davidson, and Michele Vir-

gilio for their help in my teaching journey, Dr. Darja Barr for offering me several teaching and outreaching opportunities, Clifford Allotey for his friendship and teaching support, and Carol Johnston and Tyla Turman for letting me serve Aboriginal students. I am grateful to the staff of the Mathematics Department for their support during my years in the program: Erin, Irene, John, Kristina, Leah and Sara. John went above and beyond in looking after me. I am thankful to Jodie, Golnaz, Megan, Natasha, Trevor, Faculty of Science, and Faculty of Graduate Studies, for their incredible support. I am grateful for the financial support provided by the Fields Institute, PIMS, UMGF, Department of Mathematics, Faculty of Science, and the Faculty of Graduate Studies at the University of Manitoba as well as my supervisor's NSERC grant.

I thank all of my friends for their continued love and care. In particular, I acknowledge Rohit for driving me towards my best since our high school days, Pratibha and Saraswati, for giving me a home away from home at the MNNIT Allahabad. I thank Brock and Eric for all the fun studying at the library, Hermie for teaching me the importance of equity, Susan for encouraging me to build a network, and Xiaohong for good times.

I express my regards for my guide and my godfather Gurdev Singh who has given me extraordinary inspiration. I will be forever grateful to my maternal grandparents and my mother for raising me and motivating me spiritually. I thank my late father, who dedicated his life to his career to be the best in his field, with whom I could not be in touch during his last days. I express my gratitude to my family for their love and care, especially Gian Kaur, my late great-grandmother, as this is a result of her blessings. I thank my pandemic baby nephew Armaanbir Singh, for being a little bundle of joy.

Post-Defence Acknowledgement: I would like to thank Dr. Dong Liang for his questions, comments, and suggestions which helped to improve this thesis.

I dedicate this thesis to my guide Gurdev Singh and to my family. For the blessings, encouragement, support, and values they have given me throughout their life.

Contributions of Authors

Chapters 3 and 5 are a version of journal articles submitted for publication co-authored with S.H. Lui. Some parts of Chapter 6 are a version of a manuscript in preparation co-authored with S.H. Lui.

Contents

List of Figures

List of Tables

List of Symbols

$\lvert a\rvert$	the absolute value of a real number a
$\lvert z\rvert$	the modulus of a complex number z
$\Re z$	the real part of a complex number z
$\Im z$	the imaginary part of a complex number z
$\mathbb{R}^n$	the space of real (column) vectors of length n
$\lvert x\rvert$	the 2-norm or Euclidean norm of a vector x
$\lvert x\rvert_1$	the 1-norm of a vector x
$\lvert x\rvert_\infty$	the ∞-norm of a vector x
x^T	the transpose of a vector x
$[x_1, x_2, \ldots, x_n]$	a row vector of length n
$[x_1; x_2; \ldots; x_n]$	a column vector of length n
$\mathbf{k}_n$	the constant vector of length n and components equal to $k \in \mathbb{R}$
$\mathbf{e}_k$	the elementary vector where the k-th component is 1
	and rest are zeros
$\mathbb{R}^{m\times n}$	the space of real matrices of size $m \times n$
A^T	the transpose of a matrix A
$\operatorname{rank} A$	the rank of a matrix A
$\mathcal{R}(A)$	the range space or column space of a matrix A
$\mathcal{N}(A)$	the null space of a matrix A
$\operatorname{diag}(a_1, a_2, \ldots, a_n)$	the diagonal $n \times n$ matrix with the stated diagonal entries

$[A$	the matrix obtained from A by deleting its first row
$A]$	the matrix obtained from A by deleting its first column
$\mathbf{[}A\mathbf{]}$	the matrix obtained from A by deleting its first row and column
$[\![A]\!]$	the matrix obtained from A by deleting its first and last rows and columns
I_n	the identity $n \times n$ matrix
$O_{m,n}$	the zero $m \times n$ matrix
$\Lambda(A)$	the spectrum of a square matrix A
$\rho(A)$	the spectral radius of a square matrix A
$\lambda_{\max}(A)$	the maximum eigenvalue of a square matrix A
$\lambda_{\min}(A)$	the minimum positive eigenvalue of a PSD matrix A
$\lvert\lambda\rvert_{\max}(A)$	the maximum absolute value of the eigenvalues of a square matrix A
$\lvert\lambda\rvert_{\min}(A)$	the minimum absolute value of the eigenvalues of a square matrix A
$\sigma(A)$	the set of all singular values of a matrix A
$\sigma_{\max}(A)$	the maximum singular value of a square matrix A
$\sigma_{\min}(A)$	the minimum positive singular value of matrix A
$\lVert A\rVert$	the 2-norm of a matrix A
$\lVert A\rVert_1$	the 1-norm of a matrix A
$\lVert A\rVert_\infty$	the ∞-norm of a matrix A
$A \otimes B$	the tensor product of matrices A and B
$A \oplus B$	the direct sum of matrices A and B
L_n	the Legendre polynomial of degree n
T_n	the Chebyshev polynomial of degree n
$J_n^{\alpha,\beta}$	the Jacobi polynomial of degree n and order α, β
$C_n^{(\lambda)}$	the ultraspherical polynomial of degree n and order λ
$\tilde{C}_n^{\left(\frac{3}{2}\right)}$	a rescaled ultraspherical polynomial of degree n and order $\frac{3}{2}$
δ_{jk}	the Kronecker delta function for a given set of indices j and k

1

Introduction

The numerical methods seeking the solution to a differential equation in terms of a series of known, smooth functions are called spectral methods. They have been influential in the field because of their easier implementation and super-algebraic convergence rate, which is exponential when the solution is analytic. This phenomenon of convergence is referred to as spectral convergence. Most spectral methods for time-dependent partial differential equations (PDEs) consider a finite difference scheme in time. The problem with such an implementation is that the error in time dominates the spectral convergence in space resulting in a low order convergence overall.

A series of recent studies has employed spectral methods to linear time dependent PDEs, demonstrating spectral convergence in both space and time. In [92], a space-time spectral method was proposed for the heat equation, which was analyzed in [65, 66]. Furthermore, these schemes were applied to other linear PDEs in [67]. In this thesis, this work is continued to analyze the space-time spectral methods to the Stokes problem, which leads to a saddle point problem. Furthermore, we devised two space-time spectral methods for one of the most significant non-linear PDEs, the Navier-Stokes problem, by virtue of schemes described in [12]. Such schemes employ spectral collocation in time, which leads to a system coupled in all times

with dense matrices.

The ultraspherical spectral methods represent a recent class of spectral method which lead to linear systems with sparse and well-conditioned matrices. In [74], two numerical experiments were described for the time dependent PDEs by using Chebfun. To illuminate this uncharted area, we examine space-time spectral methods schemes for linear PDEs by using the US method in both space and time. The linear systems arising from these schemes are sparse and lead to block almost banded global space-time spectral operators. Due to this special structure, a parallel-in-time solver for space time spectral methods now seem to be a foremost area of research.

While estimating the 2-norm condition number for the space-time spectral methods for the Stokes problem, we encountered the problem of estimating the minimum eigenvalue of a sum of two positive semi-definite (PSD) matrices thrice. Seminal contributions have been made on estimating the spectrum of a sum of two symmetric matrices, most well known being Weyl's inequalities, proofs for Horn's conjecture, arithmetic-geometric mean inequality for matrices, etc. A closer look to the literature, however, reveals a gap that they fail to provide a positive lower bound on the minimum eigenvalue, $\lambda_{\min}(P+Q)$, when $P+Q$ is SPD and P and Q are singular PSD matrices. The recurrence of this problem in the earlier chapters of this thesis gave us the rigor to approach it. Thus, two positive lower bounds on $\lambda_{\min}(P+Q)$ were derived, which further lead to lower bounds on the minimum singular value of some full rank block matrices up to the size 2×2. This is the principal aim of this thesis. The remainder of the document is arranged as follows.

In Chapter 2, we present the necessary mathematical preliminaries, and briefly review the literature in order to ground this work in context.

In Chapter 3, our focus is on developing and analyzing the space-time spectral methods for the Stokes problem. This problem is challenging to solve because it requires several prerequisites results, including studying the components of a spectral

method for the Stokes problem in the steady-state and estimating the 2-norm of a pseudospectral derivative matrix. They further involve approximating the spectrum and singular values of a saddle point matrix, which is solvable to some extent, whereas some of the observations motivate subsequent chapters.

In Chapter 4, we implement two space-time spectral methods on the Navier-Stokes problem. Numerical experiments verify the spectral convergence in both space and time.

In Chapter 5, our attention diverts to curating and investigating the spectral problem of sum of two PSD matrices. Although a myriad of results address this problem, our target is an improvement over a lower bound on the minimum eigenvalue for a specific case. Moreover, they aid in deriving more results on the minimum singular value of some full-rank matrices.

In Chapter 6, we implement the US method in both space and time, fulfilling desirable properties of sparse and well-structured linear systems. Thus, they eliminate the major drawback of space-time spectral methods by permitting a parallel solver. Here space-time US methods for the heat, Schrödinger, and wave equations are devised.

This thesis concludes in Chapter 7 with discussions of future directions.

Some alternative proofs for results used in this thesis are given in Appendix A for the interested reader.

2

Background

The fundamental concepts of a topic serve as the foundation on which one builds new results. This chapter summarizes the consequential results on some topics that, in conjunction, serve as a core for this thesis, such as classical linear algebra, angle between subspace, orthogonal polynomials, and spectral methods.

2.1 Fundamentals of linear algebra

A viewpoint of a *matrix* that facilitates the development and understanding of numerical algorithms is provided by a *block matrix*. Partitioning a matrix results in a block matrix whose elements are themselves matrices, which are called submatrices or sublocks (or simply as *blocks*). Suppose that matrices A and B are partitioned into blocks as follows,

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1r} \\ A_{21} & A_{22} & \dots & A_{2r} \\ \vdots & \vdots & \ddots & \vdots \\ A_{s1} & A_{s2} & \dots & A_{sr} \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} B_{11} & B_{12} & \dots & B_{1t} \\ B_{21} & B_{22} & \dots & B_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ B_{r1} & B_{r2} & \dots & B_{rt} \end{bmatrix}.$$

If the pairs (A_{ik}, B_{kj}) are conformable, then A and B are said to be *conformably partitioned*. For such matrices, the product AB is formed by combining the blocks exactly the same way as the scalars are combined in ordinary matrix multiplication. That is, the (i, j)-block in AB is $A_{i1}B_{1j} + A_{i2}B_{2j} + \ldots + A_{ir}B_{rj}$.

Another primary concept is that of a *vector space*. Recall that the *trivial subspace* of a vector space V contains only the zero vector, otherwise it is called as a *non-trivial subspace*. Let V be a non-trivial vector space, a *basis* for V is defined to be a linearly independent set $\mathcal{B} \subseteq V$ such that $\text{span}(\mathcal{B}) = V$. Moreover, the number of vectors in any basis for V is the *dimension* of a vector space V and is denoted by $\dim V$. For two vector spaces U and V such that $U \subseteq V$, $\dim U \leqslant \dim V$, and if $\dim U = \dim V$ then $U = V$. If X and Y are subspaces of a vector space V, then the *sum* of X and Y is, $X + Y = \{x + y \mid x \in X, y \in Y\}$ and $X + Y$ is again a subspace of V, so that

$$\dim(X + Y) = \dim X + \dim Y - \dim(X \cap Y). \tag{2.1.1}$$

Subspaces X , Y of a vector space V are said to be *complementary* whenever $V = X + Y$ and $X \cap Y = 0$, in which case V is said to be the *direct sum* of X and Y, and this is denoted by writing $V = X \oplus Y$. For a vector space V with subspaces X, Y having respective bases $\mathcal{B}_X$ and $\mathcal{B}_Y$, the following statements are equivalent.

1. $V = X \oplus Y$.

2. For each $v \in V$, there are unique vectors $x \in X$ and $y \in Y$ such that $v = x + y$.

3. $\mathcal{B}_X \cap \mathcal{B}_Y = \varnothing$ and $\mathcal{B}_X \cup \mathcal{B}_Y$ is a basis for V.

A *norm* for a vector space V is a function $\|\cdot\| : V \to \mathbb{R}$ that satisfies the following conditions:

1. $\|x\| \geqslant 0$ and $\|x\| = 0 \Leftrightarrow x = 0$, for all $x \in V$.

2. $\|kx\| = |k|\|x\|$ for all $k \in \mathbb{R}$, $x \in V$.

3. $\|x+y\| \leqslant \|x\| + \|y\|$, for all $x, y \in V$.

Throughout this thesis, a *column vector* x of size $n \times 1$ is denoted by $x = [x_1; x_2; \ldots; x_n] \in \mathbb{R}^n$, whereas the *row vector* of size $1 \times n$ is represented by $x = [x_1, x_2, \ldots, x_n]$. The *p-norm* of vector x is defined as follows,

$$\|x\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}}, \text{ for } 1 \leqslant p < \infty, \text{ and } \|x\|_\infty = \max_{1 \leqslant i \leqslant n} |x_i|. \tag{2.1.2}$$

Also, for convenience, $|x|$ will denote the *euclidean norm* of x, which is defined as $|x| = \|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$. A significant property of vector norms is the *parallelogram identity* which states that for all $x, y \in V$,

$$\|x+y\|^2 + \|x-y\|^2 = 2\left(\|x\|^2 + \|y\|^2\right). \tag{2.1.3}$$

We define the *standard inner product* on $\mathbb{R}^n$, for vectors $x = [x_1; x_2; \ldots; x_n], y = [y_1; y_2; \ldots; y_n] \in \mathbb{R}^n$, as $(x, y) = \sum_{i=1}^{n} x_i y_i = x^T y$, where $x^T = [x_1, x_2, \ldots, x_n]$. Moreover, x and y are said to be *orthogonal* to each other whenever $(x, y) = 0$. A subset $M = \{v_1, v_2, \ldots, v_r\}$ of $\mathbb{R}^n$ is said to be an *orthogonal set* if the vectors in M are pairwise orthogonal, that is, $(v_i, v_j) = 0$ for $i \neq j$ and $1 \leqslant i, j \leqslant r$. Moreover, M is called as an *orthonormal set* of vectors it is an orthogonal set and satisfies the condition that $(v_i, v_i) = 1$ or $|v_i| = 1$ for all $1 \leqslant i \leqslant r$. An *orthogonal matrix* $Q \in \mathbb{R}^{n \times n}$ is a matrix whose rows and columns are all mutually orthonormal, thus satisy $Q^T Q = I_n = Q Q^T$.

For a subspace $M \subseteq \mathbb{R}^n$, the *orthogonal complement* $M^\perp$ of M is defined to be the set of all vectors in $\mathbb{R}^n$ that are orthogonal to every vector in M. That is, $M^\perp = \{x \in V \,|\, (m, x) = 0, \forall\, m \in M\}$, then $\mathbb{R}^n = M \oplus M^\perp$ and $\dim M^\perp = n - \dim M$. Also, $M^{\perp\perp} = M$.

Proposition 2.1.1 (Orthogonality of subspaces, see [68]). *Let M and N be subspaces*

of $\mathbb{R}^n$*, then*

1. $M \subseteq N$ *implies* $N^\perp \subseteq M^\perp$.

2. $(M+N)^\perp = M^\perp \cap N^\perp$.

3. $(M \cap N)^\perp = M^\perp + N^\perp$.

Assume $V = X \oplus Y$ and $v = x + y$, for unique $x \in X$ and $y \in Y$, the *projection* operator P on V is defined so that the projection of V onto X along Y is $Pv = x$, whereas $(I - P)v = y$ describes the projection of V onto Y along X. Moreover, a linear operator P on a vector space V is a projection if and only if $P^2 = P$. Let $V = \mathbb{R}^n = X \oplus Y$, $\mathcal{B}_1 = \{x_1, x_2, \ldots, x_r\}$ and $\mathcal{B}_2 = \{y_{r+1}, y_{r+2}, \ldots, y_n\}$ be a basis for the subspace X and Y of $\mathbb{R}^n$, respectively. Then $\mathcal{B} = \mathcal{B}_1 \cup \mathcal{B}_2$ is a basis for $\mathbb{R}^n$, thus $z \in \mathbb{R}^n$ can be expressed as $z = \sum_{i=1}^{r} a_i x_i + \sum_{i=r+1}^{n} b_i y_i$, where $a_i, b_j \in \mathbb{R}$ for $1 \leqslant i \leqslant r$ and $r+1 \leqslant j \leqslant n$. Then, projection of z onto X along Y is $Pv = \sum_{i=1}^{r} a_i x_i$ and the projection of z onto Y along X is $\sum_{i=r+1}^{n} b_i y_i = v - Pv = (I - P)v$. In order to get a matrix form of P, define $B = [x_1, x_2, \ldots, x_r, y_{r+1}, \ldots, y_n] \in \mathbb{R}^{n \times n}$, which is a matrix whose columns form a basis for V. Then, the matrix form of projections is given as follows

$$P = B \left[\begin{array}{c|c} I_r & \\ \hline & O \end{array}\right] B^{-1}, \quad (I - P) = B \left[\begin{array}{c|c} O & \\ \hline & I_{n-r} \end{array}\right] B^{-1}.$$

The formulation is a result of the observations that $I = B^{-1}B$ implies $B^{-1}x_i = e_i$, $B^{-1}y_j = e_j$, for all $1 \leqslant i \leqslant r$ and $r+1 \leqslant j \leqslant n$. Thus, $Pv = \sum_{i=1}^{r} a_i x_i = \sum_{i=1}^{r} a_i B e_i + \sum_{i=r+1}^{n} b_i B(0)$. The matrix formulation for P implies the expression for $I - P$.

2.1.1 Results from matrix analysis

This section reviews the literature related to matrix analysis. Over time, an extensive literature has developed on this topic, we review some concepts that are imperative

for the results deduced in the later chapters. Firstly, recall that the maximal number of linearly independent columns of a matrix $A \in \mathbb{R}^{m\times n}$ is called *rank* of A, and it is denoted by $\text{rank}(A)$.

Proposition 2.1.2 (Properties of rank, see [44]). *Let $A \in \mathbb{R}^{m\times n}$, then*

1. $\text{rank}(A) \leqslant \min\{m, n\}$, *and a matrix that has rank* $\min\{m, n\}$ *is said to have full rank; otherwise the matrix is rank deficient.*

2. $\text{rank}(A) = 0$ *if and only if* $A = O$.

3. $\text{rank}(A) = \text{rank}(A^T) = \text{rank}(A^TA) = \text{rank}(AA^T)$.

4. $\text{rank}(A + B) \leqslant \text{rank}(A) + \text{rank}(B)$.

5. $\text{rank}(AB) \leqslant \min\{\text{rank}(A), \text{rank}(B)\}$.

The *range space* of a matrix $A \in \mathbb{R}^{m\times n}$ is defined to be the subspace $\mathcal{R}(A) = \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$. Similarly, the range of A^T is the subspace of $\mathbb{R}^n$ defined by $\mathcal{R}(A^T) = \{A^Ty | y \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$. Since $Ax = \sum_{i=1}^{n} a_ix_i$, where a_i are columns of A and x_i are components of x for $1 \leqslant i \leqslant n$, so that $x = [x_1; x_2; \ldots; x_n]$, thus $\mathcal{R}(A)$ is a linear combination of columns of A and is also called as the column space of A. Similarly, $\mathcal{R}(A^T)$ is called as row space of A. Also, the set $\mathcal{N}(A) = \{x \in \mathbb{R}^n \mid Ax = 0\} \subseteq \mathbb{R}^n$ is called the *null space* of A. In other words, $\mathcal{N}(A)$ is simply the set of all solutions to the homogeneous system $Ax = 0$. The set $\mathcal{N}(A^T) = \{y \in \mathbb{R}^m \mid A^Ty = \mathbf{0}\} \subseteq \mathbb{R}^m$ is called the *left-hand null space* of A because it is the set of all solutions to the left-hand homogeneous system $y^TA = \mathbf{0}^T$.

Proposition 2.1.3 (Properties of range and null spaces, see [68]). *Let $A \in \mathbb{R}^{m\times n}$ and $r = \text{rank}(A)$, then*

1. $\mathcal{N}(A) = \{0\}$ *if and only if* $r = n$.

2. $\mathcal{N}(A^T) = \{0\}$ *if and only if* $r = m$.

3. $\dim \mathcal{R}(A) = r = \dim \mathcal{R}(A^T)$.

4. $\dim \mathcal{N}(A) = n - r$.

5. $\dim \mathcal{R}(A) + \dim \mathcal{N}(A) = n$.

6. $\dim \mathcal{R}(A^T) = r$.

7. $\dim \mathcal{N}(A^T) = m - r$.

8. $\mathcal{R}(A^T A) = \mathcal{R}(A^T)$.

9. $\mathcal{R}(AA^T) = \mathcal{R}(A)$.

10. $\mathcal{N}(A^T A) = \mathcal{N}(A)$.

11. $\mathcal{N}(AA^T) = \mathcal{N}(A^T)$.

12. $\mathcal{N}(A)^\perp = \mathcal{R}(A^T)$.

13. $\mathbb{R}^n = \mathcal{R}(A^T) \oplus \mathcal{N}(A)$.

14. $\mathcal{N}(A + B) = \mathcal{N}(A) \cap \mathcal{N}(B)$.

Next, we introduce the concept of a *matrix norm*. It is denoted by $\|\cdot\|$ is a function from $\mathbb{R}^{m\times n} \to [0, \infty)$ satisfying the following properties,

1. $\|A\| \geqslant 0$, for all $A \in \mathbb{R}^{m\times n}$, and $\|A\| = 0$ if and only if $A = O$.

2. $\|\alpha A\| = |\alpha|\|A\|$, for all $\alpha \in \mathbb{R}$ and $A \in \mathbb{R}^{m\times n}$.

3. $\|A + B\| \leqslant \|A\| + \|B\|$, for all $A, B \in \mathbb{R}^{m\times n}$.

4. $\|AB\| \leqslant \|A\|\|B\|$, for all $A, B \in \mathbb{R}^{m\times n}$ with $m = n$.

Every vector norm $|\cdot|$ induces a matrix norm. We say that $\|\cdot\|_{(m,n)} : \mathbb{R}^{m\times n} \to [0,\infty)$ is the matrix norm induced by $\|\cdot\|_{(m)}$ and $\|\cdot\|_{(n)}$, such that

$$\|A\|_{(m,n)} = \sup_{x\in\mathbb{R}^n\setminus\{0\}} \frac{\|Ax\|_{(m)}}{\|x\|_{(n)}} = \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|_{(n)}=1}} \|Ax\|_{(m)},$$

for all $A \in \mathbb{R}^{m\times n}$. Moreover, for a vector $x \in \mathbb{R}^n$ and matrix $A \in \mathbb{R}^{m\times n}$,

$$\|Ax\|_{(n)} \leqslant \|A\|_{(m,n)}\|x\|_{(m)}.$$

However, not all matrix norms are induced by vector norms. An example of such a matrix norm is the *Frobenius matrix norm*, defined as $\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}$, for $A \in \mathbb{R}^{m\times n}$. The matrix norm induced by the p-norm of a vector is called as the *matrix p-norm*, and is denoted by $\|\cdot\|_p$. Throughout this thesis, we denote the matrix 2-norm as $\|\cdot\|$. Other useful p-norms are 1-norm and ∞-norm of matrices, defined as follows for matrices $A \in \mathbb{R}^{m\times n}$,

$$\begin{aligned} \|A\|_1 &= \text{ the maximum absolute column sum of } A, \\ \|A\|_\infty &= \text{ the maximum absolute row sum of } A. \end{aligned}$$

Another useful inequality between matrix norms is $\|A\|_2 \leqslant \sqrt{\|A\|_1\|A\|_\infty}$,, which is a special case of the *Hölder's inequality*.

A particular type of projection operator will play a significant role in performing analysis in this thesis, thus we conclude this section by discussing orthogonal projections. Recall that any subspace M of $\mathbb{R}^n$ allows the decomposition $\mathbb{R}^n = M \oplus M^\perp$. The projection operator P onto M along $M^\perp$ is called an *orthogonal projection*. In this case, $V = \mathcal{R}(P) \oplus \mathcal{N}(P)$, where $\mathcal{N}(P)^\perp = \mathcal{R}(P)$.

Definition 2.1.4. (Orthogonal projection, see [34]) Let $U \subseteq \mathbb{R}^n$ be a subspace. A

matrix $P \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto U if $\mathcal{R}(P) = U$, $P^2 = P$, and $P^T = P$. Moreover, if $x \in \mathbb{R}^n$, then $Px \in U$ and $(I - P)x \in U^\perp$. Also, $\|P\| = 1$.

The *Kronecker product* of the matrices $A = [a_{ij}] \in \mathbb{R}^{m\times n}$ and $B = [b_{ij}] \in \mathbb{R}^{p\times q}$ is denoted by $A \otimes B$ and is defined to be the block matrix

$$A \otimes B = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1n}B \\ a_{21}B & a_{22}B & \dots & a_{2n}B \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1}B & a_{m2}B & \dots & a_{mn}B \end{bmatrix} \in \mathbb{R}^{mp\times nq}.$$

It satisfies the properties $(A \otimes B)^T = A^T \otimes B^T$, and $(A \otimes B)(C \otimes D) = AC \otimes BD$, where $C \in \mathbb{R}^{n\times r}$, $D \in \mathbb{R}^{q\times s}$. Let $\text{rank}(A) = r_1$, $\text{rank}(B) = r_2$, then $\text{rank}(A \otimes B) = \text{rank}(B \otimes A) = r_1 r_2$. If $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{m\times n}$ are invertible, then $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$. Also, the *direct sum* of $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{p\times q}$ is defined as $A \oplus B = \left[\begin{smallmatrix} A & O \\ O & B \end{smallmatrix}\right] \in \mathbb{R}^{m+p\times n+q}$. Thus, $A \oplus B$ is a block diagonal matrix the main diagonal blocks of which are matrices A and B. Alternatively, a *block diagonal* square matrix A the main diagonal blocks of which are matrices square matrices $A_1, \dots, A_k$ is indicated as $\text{diag}(A_1, \dots, A_k)$. In particular, if all the main diagonal blocks are scalars, it is a diagonal matrix.

2.1.2 Spectrum of a matrix

The *characteristic polynomial* $\phi(A, t)$ of $A \in \mathbb{R}^{n\times n}$ is $\det(tI_n - A)$, for $t \in \mathbb{C}$. We call a root λ of $\phi(A, t)$ as an *eigenvalue* of A. The algebraic multiplicity of an eigenvalue λ of A refers to the multiplicity λ as a root of the characteristic polynomial of A. We say that λ is a *simple* eigenvalue of A if its algebraic multiplicity is one. We call the set of eigenvalues of A the *spectrum* of A, denoted $\Lambda(A)$. For an eigenvalue λ of A, any non-zero vector $w \in \mathbb{R}^n$ such that $Aw = \lambda w$ is called an *eigenvector* associated to λ, and we call the vector space of all eigenvectors corresponding λ the eigenspace

of the eigenvalue λ.

Theorem 2.1.5 (Spectral decomposition, see [44]). *Let $A \in \mathbb{R}^{n\times n}$ be a square matrix with n linearly independent eigenvectors $p_1, \ldots, p_n$. Then A can be factored as $A = P\mathcal{D}P^{-1}$, where the columns of P are the eigenvectors $p_1, \ldots, p_n$ and $\mathcal{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$.*

If the spectral decomposition of a square matrix A exists as described above then we say that A is *diagonalizable*. Moroever, every symmetric matrix is *orthogonally diagonalizable*. Let $M \in \mathbb{R}^{n\times n}$ be a symmetric matrix, then it is orthogonally diagonalized as $M = Q\mathcal{D}Q^T$, where Q is an orthogonal matrix whose columns are the orthonormal eigenvectors for M and $\mathcal{D} = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$, $\lambda_j \in \mathbb{R}$. Another result of great importance is the following, which estimates the spectrum of sum of two symmetric matrices as seen in [43, p. 239].

Theorem 2.1.6 (Weyl's inequalities, see [43]). *If A and E are $n \times n$ real symmetric matrices, then*

$$\lambda_k(A) + \lambda_n(E) \leqslant \lambda_k(A + E) \leqslant \lambda_k(A) + \lambda_1(E), \quad 1 \leqslant k \leqslant n.$$

A symmetric matrix with non-negative eigenvalues is called *positive semidefinite* (PSD) matrix, and with positive eigenvalues is called *positive definite* (SPD) matrix. The following properties are valuable for our work.

Theorem 2.1.7 (Properties of eigenvalues of a matrix, see [41]). *Let $A \in \mathbb{R}^{n\times n}$, and $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n(A)$ be the eigenvalues of A.*

1. *If A is symmetric then $\lambda_1(A) = \max\limits_{x\in\mathbb{R}^n\setminus\{0\}} \dfrac{x^TAx}{x^Tx}$ and $\lambda_n(A) = \min\limits_{x\in\mathbb{R}^n\setminus\{0\}} \dfrac{x^TAx}{x^Tx}$.*
2. *Let $A_i \in \mathbb{R}^{n_i\times n_i}$ and $A \in \mathbb{R}^{n\times n}$ is a block diagonal matrix, so that $A = \oplus_{i=1}^k A_i$, then $\Lambda(A) = \bigcup\limits_{i=1}^k \Lambda(A_i)$.*

3. *Let $A \in \mathbb{R}^{n\times n}$ be block upper triangular matrix so that,*

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1k} \\ & A_{22} & \dots & A_{2k} \\ & & \ddots & \vdots \\ & & & A_{kk} \end{bmatrix}.$$

Then $\operatorname{rank}(A) \geqslant \sum_{i=1}^{k} \operatorname{rank}(A_{ii})$, *and* $\Lambda(A) = \bigcup_{i=1}^{k} \Lambda(A_{ii})$.

4. *Let $B \in \mathbb{R}^{m\times m}$. If $\Lambda(A) = \{\lambda_1, \dots, \lambda_n\}$ and $\Lambda(B) = \{\mu_1, \dots, \mu_m\}$, then $\Lambda(A \otimes B) = \{\lambda_i\mu_j, 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$, including algebraic multiplicities.*

5. *Let $B \in \mathbb{R}^{m\times m}$, $\lambda \in \Lambda(A)$ with a corresponding eigenvector $x \in \mathbb{R}^n$ and $\mu \in \Lambda(B)$ with a corresponding eigenvector $y \in \mathbb{R}^m$, then $\lambda + \mu$ is an eigenvalue of $I_m \otimes A + B \otimes I_n$ for which $y \otimes x$ is a corresponding eigenvector. Or, $\Lambda(I_m \otimes A + B \otimes I_n) = \{\lambda_i + \mu_j \mid 1 \leqslant i \leqslant n, 1 \leqslant j \leqslant m\}$ including algebraic multiplicities.*

2.1.3 Singular value decomposition

The concept of spectral decomposition of a matrix is not applicable to a large class of matrices such as rectangular matrices. Thus, *singular value decomposition* plays a vital role. Let $A \in \mathbb{R}^{m\times n}$, then the singular value decomposition of A is given as $A = U\Sigma V^T$, where $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ are orthogonal matrices and $\Sigma \in \mathbb{R}^{m\times n}$ is a diagonal matrix with non-increasing singular values $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_{\min\{m,n\}}$. The columns of U are known as left singular vectors, while the columns of V are known as right singular vectors.

Theorem 2.1.8 (Properties of singular values of a matrix, see [41]). *Let $A \in \mathbb{R}^{m\times n}$, $\operatorname{rank}(A) = r$, and $\sigma_1 \geqslant \sigma_2 \geqslant \ldots \geqslant \sigma_{\min\{m,n\}} \geqslant 0$ be singular values of A.*

1. *$\sigma_{\min}(A) = \sigma_r > 0$ and $\sigma_i = 0$, for all $r + 1 \leqslant i \leqslant n$.*

2. $\sigma_{\max}(A) = \max\limits_{x \in \mathbb{R}^n \setminus \{0\}} \dfrac{\|Ax\|}{\|x\|}$.

3. $\sigma_{\min\{m,n\}}(A) = \min\limits_{x \in \mathbb{R}^n \setminus \{0\}} \dfrac{\|Ax\|}{\|x\|}$.

4. $\sigma_i(A) = \sigma_i(A^T) = \sqrt{\lambda_i(A^T A)} = \sqrt{\lambda_i(AA^T)}$, *for all* $1 \leqslant i \leqslant \min\{m, n\}$.

5. $\sigma_{\max}(A) = \|A\|$.

6. $\sigma_{\min}(A) = \|A^{-1}\|^{-1}$, *if* $m = n = r$.

7. *Let* $A_i \in \mathbb{R}^{m_i \times n_i}$ *and* $A \in \mathbb{R}^{m \times n}$ *is a block diagonal matrix, so that* $A = \bigoplus_{i=1}^{k} A_i$, *then* $\sigma(A) = \bigcup_{i=1}^{k} \sigma(A_i)$.

8. *Let* $A \in \mathbb{R}^{m \times n}$ *and define the symmetric matrix* $B = \left(\begin{smallmatrix} O & A \\ A^T & O \end{smallmatrix}\right) \in \mathbb{R}^{m+n, m+n}$. *The eigenvalues of* B *are* $\pm\sigma_1(A), \ldots, \pm\sigma_{\min\{m,n\}}(A)$ *along with* $|m - n|$ *zeros.*

9. *If* $m = n$ *and* $\lambda \in \Lambda(A)$, *then* $\sigma_{\min}(A) \leqslant |\lambda| \leqslant \sigma_{\max}(A)$.

10. *Let* $B \in \mathbb{R}^{p \times q}$ *and* $\operatorname{rank}(B) = s$. *The positive singular values of* $A \otimes B$ *are* $\{\sigma_i(A)\sigma_j(B) : 1 \leqslant i \leqslant r, 1 \leqslant j \leqslant s\}$, *including multiplicities. Zero is a singular value of* $A \otimes B$ *with multiplicity* $\min\{mp, nq\} - rs$.

2.1.4 Condition number

The *condition number* of a matrix $A \in \mathbb{R}^{n \times n}$ is a measure of the extent to which the relative error in the input is magnified to cause relative error in the output. It indicates how difficult it is to solve numerically the linear system $Ax = b$, where $x, b \in \mathbb{R}^n$. The problem is said to be ill-conditioned if the condition number is large depending on the precision, i.e., the number of digits, in the calculation. Some consequences of which include, the numerical solution is susceptible to round-off errors during the Gaussian elimination process and an iterative solution to the system is most likely to be slowly converging. Let us derive the expression for it, to this

end, let $\mathfrak{e}$ be the error in b. Assuming that A is a non-singular matrix, the error in the solution $x = A^{-1}b$ is $A^{-1}\mathfrak{e}$. The maximum of the ratio of the relative error in the solution to the relative error in b defines the condition number of A, as follows

$$\begin{aligned}\kappa(A) &:= \max_{\mathfrak{e},b\in\mathbb{R}^n\setminus\{0\}} \left(\frac{\|A^{-1}\mathfrak{e}\|}{\|\mathfrak{e}\|}\frac{\|b\|}{\|A^{-1}b\|}\right) \\ &= \max_{\mathfrak{e}\in\mathbb{R}^n\setminus\{0\}} \left(\frac{\|A^{-1}\mathfrak{e}\|}{\|\mathfrak{e}\|}\right) \max_{b\in\mathbb{R}^n\setminus\{0\}} \left(\frac{\|Ab\|}{\|b\|}\right) \\ &= \left\|A^{-1}\right\| \|A\|.\end{aligned}$$

Thus, the condition number of a non-singular matrix A is defined as $\kappa(A) = \|A\|\|A^{-1}\| = \dfrac{\sigma_{\max}(A)}{\sigma_{\min}(A)}$. Furthermore, the *spectral condition number* of a non-singular matrix $A \in \mathbb{R}^{n\times n}$ is defined as $\kappa_{sp}(A) = \dfrac{|\lambda|_{\max}(A)}{|\lambda|_{\min}(A)}$, where $|\lambda|_{\max}(A)$ and $|\lambda|_{\min}(A)$ represent the maximum and minimum absolute value of eigenvalues of A, respectively. Let $\lambda \in \Lambda(A)$, since $\sigma_{\min}(A) \leqslant |\lambda| \leqslant \sigma_{\max}(A)$, thus $\kappa_{sp}(A) \leqslant \kappa(A)$, and $\kappa_{sp}(A) = \kappa(A)$ when A is a normal matrix. The spectral condition number is widely used because it is usually easier to estimate. It is also useful for the analysis of preconditioned systems. Consider A and M to be SPD matrices, then Theorem 4.10 in [64] gives that $\kappa_M(M^{-1}A) = \kappa(M^{-1}A)$, where $\kappa_M(B) = |B|_M|B^{-1}|_M$ with $|x|_M := \sqrt{x^T Mx}$. Hence the spectral condition number of the preconditioned matrix $M^{-1}A$ is the same as its M-norm condition number.

2.2 Angles between subspaces

In previous sections, we discussed the concepts of subspaces, including a pair of orthogonal subspaces and complementary subspaces. In general, consider that we are studying a problem involving two non-trivial subspaces of $\mathbb{R}^n$, which requires us to somehow gauge the separation between them. One way to compute that is to measure the angle between them. A quote by C. Meyer is, "There is just too much *wiggle*

room in higher dimensions to make any one definition completely satisfying, and the *correct* definition usually varies with the specific application under consideration." The suitable angle for question considered in Chapter 5 of this thesis is the principal angle between subspaces.

In 1875, Jordan introduced the concept of principal angles and vectors, see [51]. It has been discussed in [34, 24, 68].

Definition 2.2.1 (Principal angles, see [32])**.** Let $U, V \subseteq \mathbb{R}^n$ be subspaces with $p = \dim(U) \geqslant \dim(V) = q \geqslant 1$. The *principal angles* $0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_q \leqslant \frac{\pi}{2}$ between U and V are recursively defined for $k = 1, 2, \ldots, q$ by

$$\cos(\theta_k) = \max_{\substack{u \in U,\ v \in V \\ |u|=|v|=1}} |u^T v| = u_k^T v_k, \tag{2.2.1}$$

subject to the constraints

$$u_i^T u = 0, \quad v_i^T v = 0, \quad i = 1, 2, \ldots, k-1.$$

The vectors $\{u_1, \ldots, u_q\}$, $\{v_1, \ldots, v_q\}$ are called *principal vectors* of the pair of spaces.

The principal angles are uniquely defined, while the principal vectors are not. Note that vectors $\{u_i\}_{i=1}^q$ and $\{v_i\}_{i=1}^q$ are orthonormal, so that $v_i^T u_j = \delta_{ij} \cos\theta_i$ for $1 \leqslant i \leqslant q$. Also, by definition it is observed that $\{v_i\}_{i=1}^q$ is an orthonormal basis for V or $\mathcal{R}(v_1, v_2, \ldots, v_q) = V$, and $U \cap \mathcal{R}^{\perp}(u_1, u_2, \ldots, u_q) \perp V$ or $\{u_i\}_{i=q+1}^p$ is orthogonal to V. So that, $\{u_i\}_{i=1}^p$ is an orthonormal basis of U such that its satisfies the following biorthogonality relation holds,

$$\left[u_i^T v_j\right]_{i,j=1}^{p,q} = \begin{bmatrix} \operatorname{diag}(\cos\theta_1, \ldots, \cos\theta_q) & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.2.2}$$

The angle θ_1 is also called the *minimal principal angle*. The minimal angle between

non-zero subspaces $U, V \subseteq \mathbb{R}^n$ is defined as the number $\theta_{\min} \in \left[0, \frac{\pi}{2}\right]$, so that

$$\cos\theta_{\min} = \max_{\substack{u\in U,\ v\in V \\ |u|=|v|=1}} |v^T u|. \tag{2.2.3}$$

The following properties of the minimal principal angle make it useful.

Proposition 2.2.2 (Properties of the minimal angle, see [68]). *Let $\theta_{\min}$ be the minimal angle between non-zero subspaces $U, V \subseteq \mathbb{R}^n$, then*

1. *$\theta_{\min} = 0$ if and only if $U \cap V \neq \{0\}$.*
2. *$\theta_{\min} = \dfrac{\pi}{2}$ if and only if $U \perp V$.*
3. *let $\theta^{\perp}_{\min}$ denote the minimal angle between $U^{\perp}$ and $V^{\perp}$. If $U \oplus V = \mathbb{R}^n$, then $\theta_{\min} = \theta^{\perp}_{\min}$.*
4. *U and V are complementary subspaces if and only if $P_1 - P_2$ is invertible, where $P_1, P_2 \in \mathbb{R}^{n\times n}$ are orthogonal projectors onto $\mathcal{R}(U), \mathcal{R}(V)$, respectively, and in this case $\sin\theta_{\min} = \sigma_{\min}(P_1 - P_2)$.*

The first property of the above implies that $\theta_{\min} = 0$ when $U \cap V \neq \{0\}$, although $U \neq V$. In fact, if $0 \leqslant \theta_i \leqslant \frac{\pi}{2}$, for $1 \leqslant i \leqslant \min\{\dim U, \dim V\}$, represent the principal angles between $U, V \subseteq \mathbb{R}^n$ such that $\dim(U \cap V) = k$, then $\theta_i = 0$, for all $1 \leqslant i \leqslant k$, and $\theta_{k+1} > 0$ if it exists. This is explained by the following.

The principal angles and vectors lead to natural subspace decompositions of U and V. Assume that $p = \dim U \geqslant \dim V = q$, $k = \dim(U \cap V)$, and the principal angle distribution is such that $\theta_i = 0$, for $1 \leqslant i \leqslant k$, $0 < \theta_{k+1} \leqslant \ldots \leqslant \theta_{k+r} \leqslant \frac{\pi}{2}$, and $\theta_i = \frac{\pi}{2}$, for all $k + r + 1 \leqslant i \leqslant q$.

Note that $k, r \geqslant 0$, and $k + r \leqslant q$. Let us consider the set of principal vector corresponding to the set principal angles for U as $U_1 = [u_1, u_2, \ldots, u_k]$, $U_2 = [u_{k+1}, \ldots, u_{k+r}]$, $U_3 = [u_{k+r+1}, \ldots, u_q]$, and the remaining basis vectors of U as

$U_4 = [u_{q+1}, \ldots, u_p]$. Similarly, define the set of principal vectors for V as $V_1 = U_1$, $V_2 = [v_{k+1}, \ldots, v_{k+r}]$, and $V_3 = [v_{k+r+1}, \ldots, v_q]$. According to Definition 2.2.1, $\mathcal{R}(U_i) \perp \mathcal{R}(U_j)$ and $\mathcal{R}(V_i) \perp \mathcal{R}(V_j)$, for $i \neq j$, thus the following orthogonal decompostion of U and V are obtained:

$$\begin{aligned} U &= \mathcal{R}(U_1) \oplus \mathcal{R}(U_2) \oplus \mathcal{R}(U_3) \oplus \mathcal{R}(U_4), \\ V &= \mathcal{R}(U_1) \oplus \mathcal{R}(V_2) \oplus \mathcal{R}(V_3). \end{aligned}$$

Also, eq. (2.2.2) implies that $\mathcal{R}(U_i) \perp \mathcal{R}(V_j)$ for $i \neq j$, $\mathcal{R}(U_3) \perp V$, $\mathcal{R}(U_4) \perp V$, and $\mathcal{R}(V_3) \perp U$. Let $u = \sum_{i=k+1}^{k+r} a_i u_i \in \mathcal{R}(U_2)$, where $a_i \in \mathbb{R}$, and assume that $u \perp \mathcal{R}(V_2)$, then $0 = v_j^T u = a_j \cos\theta_i$, implying $a_j = 0$ for all $k+1 \leqslant j \leqslant k+r$. Hence, $u = 0$, therefore there is no non-trivial vector in U_2 is orthogonal to V_2. Thus, these subspaces give the following decomposition for $U + V$ and $\mathbb{R}^n$,

$$\begin{aligned} U + V &= \mathcal{R}(U_1) \oplus \mathcal{R}(U_2, V_2) \oplus \mathcal{R}(U_3) \oplus \mathcal{R}(U_4) \oplus \mathcal{R}(V_3), \\ \mathbb{R}^n &= \mathcal{R}(U_1) \oplus \mathcal{R}(U_2, V_2) \oplus \mathcal{R}(U_3) \oplus \mathcal{R}(U_4) \oplus \mathcal{R}(V_3) \oplus (U + V)^{\perp}. \end{aligned}$$

Another expression is obtained from the facts that $U \cap V = \mathcal{R}(U_1)$, $U \cap V^{\perp} = \mathcal{R}(U_3) \oplus \mathcal{R}(U_4)$, $U^{\perp} \cap V = \mathcal{R}(V_3)$, and $U^{\perp} \cap V^{\perp} = (U + V)^{\perp}$. Thus, [32] provides the following decompositions,

$$\begin{aligned} U &= (U \cap V) \oplus \mathcal{R}(U_2) \oplus (U \cap V^{\perp}), \\ V &= (U \cap V) \oplus \mathcal{R}(V_2) \oplus (U^{\perp} \cap V), \\ U + V &= (U \cap V) \oplus \mathcal{R}(U_2, V_2) \oplus (U \cap V^{\perp}) \oplus (U^{\perp} \cap V), \\ \mathbb{R}^n &= (U \cap V) \oplus \mathcal{R}(U_2, V_2) \oplus (U \cap V^{\perp}) \oplus (U^{\perp} \cap V) \oplus (U^{\perp} \cap V^{\perp}). \end{aligned}$$

The above decompositions facilitate the proof of the following significant result.

Theorem 2.2.3 (See [98, 32]). *Let $U, V \subseteq \mathbb{R}^n$ be subspaces such that $p = \dim(U) \geqslant \dim(V) = q \geqslant 1$, with $k = \dim(U \cap V)$, $n_1 = \dim(U \cap V^\perp)$, $n_2 = \dim(U^\perp \cap V)$ and $n_3 = \dim(U^\perp \cap V^\perp)$. Let θ_i be the principal angles between U and V, defined by (2.2.1), and let r be the number of angles θ_i such that $0 < \theta_i < \frac{\pi}{2}$. If $\mathcal{P}, \mathcal{Q} \in \mathbb{R}^{n\times n}$ are orthogonal projections onto $\mathcal{R}(U), \mathcal{R}(V)$, respectively, then the following statements hold:*

1. *Subspaces U and V are in generic position, that is, $k = n_1 = n_2 = n_3 = 0$ if and only if $k = 0$, $p = q = r$ and $n = 2r$.*

2. *$\mathcal{P} + \mathcal{Q}$ is non-singular if and only if $n_3 = 0$.*

3. *$\mathcal{P} - \mathcal{Q}$ is non-singular if and only if $k = n_3 = 0$.*

4. *$\sigma(\mathcal{P} + \mathcal{Q}) = \{\mathbf{2}_k, 1 \pm \cos(\theta_{k+i})(i = 1, \ldots, r), \mathbf{1}_{n_1+n_2}, \mathbf{0}_{n_3}\}$.*

5. *$\sigma(\mathcal{P} - \mathcal{Q}) = \{\mathbf{1}_{n_1+n_2}, \mathbf{sin}(\theta_{k+i})_2(i = 1, \ldots, r), \mathbf{0}_{k+n_3}\}$.*

The concept of angles between subspaces of a Hilbert space is described in [26]. The Friedrichs angle between the subspaces M and N of a Hilbert space H is the angle $a(M, N)$ in $[0, \frac{\pi}{2}]$ whose cosine is given by

$$c(M, N) = \sup\{|(x, y)| \,|\, x \in M \cap (M \cap N)^\perp, \|y\| \leqslant 1, y \in N \cap (M \cap N)^\perp, \|y\| \leqslant 1\},$$

and the minimal angle between the subspaces M and N is the angle $a_0(M, N)$ in $[0, \frac{\pi}{2}]$ whose cosine is defined by

$$c_0(M, N) = \sup\{|(x, y)| \,|\, x \in M, \|x\| \leqslant 1, y \in N, \|y\| \leqslant 1\}.$$

The two definitions are different except for the case $M \cap N = \{0\}$ when they clearly agree. A notable property of the Friedrichs angle is $c(M, N) < 1$ or $a > 0$ if and only

if $M + N$ is closed. In this thesis, we work in finite dimensions, thus the following definition of the Friedrichs angle is used.

Definition 2.2.4. (Friedrichs angle, see [26]) The angle $\theta_F \in \left(0, \frac{\pi}{2}\right]$ between subspaces $U, V \subseteq \mathbb{R}^n$, whose cosine is defined by

$$\cos\theta_F := \sup\left\{|\langle x, y\rangle| \;\middle|\; x \in U \cap (U \cap V)^\perp, |x| \leqslant 1,\, y \in V \cap (U \cap V)^\perp, |y| \leqslant 1\right\},$$

is called the *Friedrichs angle*.

The following is a short list of properties of the Friedrichs angle, which will be used later to prove some results in this thesis.

Proposition 2.2.5 (Properties of Friedrichs angle, see [31])**.** *Let $U, V \subseteq \mathbb{R}^n$ be subspaces, as defined in Theorem 2.2.3. Let $\mathcal{P}, \mathcal{Q}$ be orthogonal projections onto U and V, respectively, and let θ_F denote the Friedrichs angle between subspaces U and V, then the following results hold.*

1. $\theta_F = \theta_1\left(U \cap (U \cap V)^\perp, V \cap (U \cap V)^\perp\right)$.
2. $\theta_F = \theta_1(U, V)$ *if and only if* $U \cap V = \{0\}$.
3. $\theta_F = \theta_1\left(U, V \cap (U \cap V)^\perp\right) = \theta_1\left(U \cap (U \cap V)^\perp, V\right)$.
4. $\cos\theta_F = \|\mathcal{P}\mathcal{Q} - \mathcal{P}_{U\cap V}\|$.
5. $\theta_F = \theta_{k+1}(U, V)$, *whenver* θ_{k+1} *exists.*

See [26, p. 110] for the first four properties of the above Proposition. A proof of Property 4 is also provided in [20, p. 1430], which along with $\|\mathcal{P}\mathcal{Q} - \mathcal{P}_{U\cap V}\| = \cos\theta_{k+1}$, proved in [32, p. 245], implies Property 5. Several more interesting results on the Friedrichs angle are stated in [31, p. 242].

We would like to emphasize that Property 5 is often missed in the standard references. Another simple proof for which can also be derived from [20], it mentions on Page 1129 that $\cos\theta_F = \sqrt{1-\lambda_{\min}(H)}$. Moreover, Page 1419 defines $\lambda_{\min}(H) =: \sin^2(\alpha_1)$, furthermore $\alpha_1 = \theta_{k+1}$, as seen on Page 1142. Therefore, $\cos\theta_F = \sqrt{1-\sin^2\theta_{k+1}} = \cos\theta_{k+1}$, hence the result. Note that, in this proof we have adapted the notations from the paper to facilitate easier understanding.

2.3 Orthogonal polynomials

Orthogonal polynomials play the role of building blocks in designing spectral methods, so it is necessary to discuss them. Given an open interval $I := (a, b)$, where $-\infty \leqslant a < b \leqslant \infty$, and a *weight function* ω such that $\omega(x) > 0$, for all $x \in I$ and $\omega \in L^1(I)$, two functions f and g are said to be *orthogonal* to each other in $L^2_\omega(a, b)$ or orthogonal with respect to ω if $\int_a^b f(x)g(x)\omega(x)dx = 0$. An algebraic polynomial of degree n is denoted by $p_n(x) = c_n x^n + c_{n-1}x^{n-1} + \ldots + c_1 x + c_0$,where $\{c_i\}$ are real constants, and $c_n \neq 0$ is the leading coefficient of p_n. A sequence of polynomials $\{p_n\}_{n=0}^{\infty}$ with $\deg(p_n) = n$ is said to be *orthogonal* if

$$\int_a^b p_n(x)p_m(x)\omega(x)dx = \delta_{mn}\gamma_n,$$

where $\gamma_n = \int_a^b p_n^2(x)\omega(x)dx$ is nonzero, and δ_{mn} is the Kronecker delta.

The most widely used orthogonal polynomials are the *classical orthogonal polynomials*. They consist of Jacobi for finite interval or $-\infty < a, b < \infty$, Hermite for infinite interval or $a = -\infty$ and $b = \infty$, and Laguerre polynomials for semi-infinite intervals. As we are dealing with finite intervals in this thesis, we elaborate the foremost.

The *Jacobi polynomials*, denoted by $J_n^{\alpha,\beta}(x)$, are orthogonal with respect to the

Jacobi weight function $\omega^{\alpha,\beta}(x) := (1-x)^{\alpha}(1+x)^{\beta}$ over $I := (-1,1)$, so that

$$\int_a^b J_n^{\alpha,\beta}(x) J_m^{\alpha,\beta}(x) \omega^{\alpha,\beta}(x) dx = \delta_{mn} \gamma_n^{\alpha,\beta},$$

where $\gamma_n^{\alpha,\beta} = \int_a^b (J_n^{\alpha,\beta}(x))^2 \omega^{\alpha,\beta}(x) dx$. The weight function $\omega^{\alpha,\beta}$ belongs to $L^1(I)$ if and only if $\alpha, \beta > -1$, thus it is assumed throughout. The Jacobi polynomial $J_n^{\alpha,\beta}$ is a solution of the second order linear homogeneous differential equation

$$\left(1-x^2\right) y'' + (\beta - \alpha - (\alpha+\beta+2)x) y' + n(n+\alpha+\beta+1) y = 0.$$

Also, the square of its norm, that is, $\gamma_n^{\alpha,\beta}$ is given as follows

$$\gamma_n^{\alpha,\beta} = \frac{2^{\alpha+\beta+1}}{2n+\alpha+\beta+1} \frac{\Gamma(n+\alpha+1)\Gamma(n+\beta+1)}{\Gamma(n+\alpha+\beta+1) n!}.$$

The polynomials have the symmetry relation $J_n^{(\alpha,\beta)}(-z) = (-1)^n J_n^{(\beta,\alpha)}(z)$; thus the terminal values are $J_n^{(\alpha,\beta)}(1) = \binom{n+\alpha}{n}$ and $J_n^{(\alpha,\beta)}(-1) = (-1)^n \binom{n+\beta}{n}$.

When $\alpha = \beta = 0$, the Jacobi polynomials $J_n^{0,0}(x)$ are called *Legendre polynomials* and are denoted by $L_n(x)$, for all $n \geqslant 0$. Thus, they are polynomials defined as an orthogonal system with respect to the weight function $\omega(x) = 1$ over the interval $[-1,1]$. That is, $L_n(x)$ is a polynomial of degree n, such that

$$\int_{-1}^{1} L_m(x) L_n(x)\, dx = \frac{2}{2n+1} \delta_{mn}.$$

Patently, the Legendre polynomials have definite parity. That is, they are even or odd, according to $L_n(-x) = (-1)^n L_n(x)$. Also, $L_n(1) = 1$ and $L_n(-1) = (-1)^n$. Another useful property is that they have zero average for all $n \geqslant 1$, that is,

$$\int_{-1}^{1} L_n(x)\, dx = 0 \text{ for } n \geqslant 1,$$

which follows from considering the orthogonality relation with $L_0(x) = 1$.

For $\alpha = \beta = -\frac{1}{2}$, the Jacobi polynomials, $J_n^{-\frac{1}{2},-\frac{1}{2}}(x)$, are called *Chebyshev polynomials* and are denoted by $T_n(x)$, for $n \geqslant 0$. Thus, they are polynomials defined as an orthogonal system with respect to the weight function $\omega(x) = \dfrac{1}{\sqrt{1-x^2}}$ over the interval $(-1, 1)$. That is, $T_n(x)$ is a polynomial of degree n, such that

$$\int_{-1}^{1} T_n(x)\, T_m(x)\, \frac{\mathrm{d}x}{\sqrt{1-x^2}} = \begin{cases} 0 & \text{if } \; n \neq m, \\ \pi & \text{if } \; n = m = 0, \\ \frac{\pi}{2} & \text{if } \; n = m \neq 0. \end{cases}$$

They also satisfy the properties, $T_n(-x) = (-1)^n T_n(x)$, $T_n(1) = 1$, and $T_n(-1) = (-1)^n$, for $n \geqslant 0$.

One of the most significant applications of orthogonal polynomials is approximation of analytic functions. Consider the set $\{\phi_n(x)\}_{n=0}^{\infty}$ of orthogonal polynomials on $(-1, 1)$, so that $\phi_n(x)$ is of degree n. Let u be an analytic function on $(-1, 1)$, then

$$u(x) = \sum_{k=0}^{\infty} u_k \phi_k(x).$$

The above expression of $u(x)$ in terms of the orthogonal basis polynomials is the fundamental result for spectral methods for solving PDEs numerically. In practice, it is undesirable to deal with an infinite number of modes u_k, thus an approximate $u(x)$ is considered. Such an approximation for u is given by its truncated series, given as $\Pi_n u(x) = \sum_{k=0}^{n} u_k \phi_k(x)$, which is a polynomial of degree n. This technique of considering a finite number of terms from the series expression for $u(x)$, is called as *truncation*. Many types of spectral methods, such as spectral coefficient, spectral Galerkin, ultraspherical spectral methods, etc., use this technique.

Another technique of approximating an analytic function in terms of a polynomial of degree n is called *interpolation*. For any analytic function $u(x)$ in $(-1, 1)$, denote

the interpolant of u by $\mathcal{I}_n u$, which is defined as the unique polynomial of degree n such that $(\mathcal{I}_n u)(x_j) = u(x_j)$, where $x_j \in (-1, 1)$ for $0 \leqslant j \leqslant n$. By uniqueness of an interpolating polynomial, a convenient expression is the Lagrange interpolating polynomial of u on the nodes x_j, $0 \leqslant j \leqslant n$. That is, $(\mathcal{I}_n u)(x) = \sum_{j=0}^{n} u(x_j)\ell_j(x)$, where $\ell_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{n} \frac{(x - x_k)}{(x_j - x_k)}$, for $0 \leqslant j \leqslant n$. This unique polynomial can also be derived in terms of classical orthogonal polynomials, such as in [64, p. 249], *Legendre interpolant* is alternatively defined as $\mathcal{I}_n u(x) = \sum_{k=0}^{n} \tilde{u}_k L_k(x)$, with $\tilde{u}_k = \tau_k [u, L_k]_m$, where $\tau_k = k + 0.5$ for $0 \leqslant k \leqslant n - 1$ and $\tau_n = 0.5n$. Also, $[\cdot, \cdot]_n$ represent the discrete L^2 inner product. Similarly, an expression for the Chebyshev interpolant of u is obtained in [64, p. 258].

Theorem 2.3.1 (Error in approximation, see [64])**.** *Let u be an analytic function in $(-1, 1)$, c and C be some positive constants, then the following are simplified forms for the stated errors:*

1. *the Legendre truncation error is given as $\|u - \Pi_n u\|_0 \leqslant ce^{-Cn}$.*
2. *the Legendre interpolation error is given as $\|u - \mathcal{I}_n u\|_0 \leqslant ce^{-Cn}$*
3. *the Chebyshev truncation error is given as $\|u - \Pi_n u\|_{0,\omega} \leqslant ce^{-Cn}$*
4. *the Chebyshev interpolation error is given as $\|u - \mathcal{I}_n u\|_{0,\omega} \leqslant ce^{-Cn}$.*

An excellent collection of recent applications of orthogonal polynomial is provided in [73].

2.4 Spectral methods

Spectral method is a fairly common term which may refer to various types of methods, but we refer to the ones that solve partial differential equations (PDEs) numerically.

Some well known classes of methods for solving PDEs include finite difference methods and finite elements methods (FEM), but the fastest converging methods belong to the class of *spectral methods*. They are quite similar to the FEM, both of them are based on approximating solutions by certain basis functions. They differ from each other in that FEM uses local basis functions, i.e., with a smaller support than the entire domain, whereas spectral methods use global basis functions living on the entire domain. If the solution is analytic, then spectral method converges exponentially quickly, which is called *spectral convergence*.

A considerable body of literature on spectral methods exists, including [64, 74, 80, 95, 94, 36]. A major drawback is that spectral methods are more difficult to apply on non-regular domains, whereas FEM is quite flexible and widely used in engineering and many other scientific studies. Progress is being made in recent years to incorporate spectral methods to a more general setting, such as problems with solutions containing discontinuities or on a more general domain. In general, the matrices arising from finite difference and finite element methods are sparse, whereas classical spectral methods lead to dense and ill-conditioned matrices. We continue by describing two well-known classifications of spectral methods, by considering a basic example of solving a Poisson problem given as follows:

$$\begin{aligned} -u''(x) &= f(x), \quad x \in (-1,1), \\ u(\pm 1) &= 0. \end{aligned}$$

2.4.1 Spectral Galerkin methods

For $0 \leqslant j \leqslant N-2$, define the basis functions $\phi_j = \dfrac{1}{\sqrt{4j+6}}\,(L_j - L_{j+2})$. Since $L_j(\pm 1) = (\pm 1)^j$, these basis functions satisfy the boundary conditions. Define the Galerkin subspace $V_h = \operatorname{span}\{\phi_j(x),\ 0 \leqslant j \leqslant N-2\}$, where $\{\phi_j\}_{j=0}^{\infty}$ is an L^2-

orthonormal basis. The Galerkin solution is defined as

$$u_h = \sum_{j=0}^{N-2} u_j \phi_j(x),$$

so that it satisfies the following Galerkin equations arising from the weak form of the Poisson problem,

$$\sum_{i=0}^{N-2} u_i \int_{-1}^{1} \phi_i'(x)\phi_j'(x)dx = \int_{-1}^{1} f(x)\phi_j(x)dx, \quad j = 0, ..., n-2,$$
$$u_j = \int_{-1}^{1} f(x)\phi_j(x)dx,$$

where the last equality is obtained by using the recurrence relation, for all $j \geqslant 0$, $L_{j+1} = \frac{1}{(2j+3)}\left(L'_{j+2} - L'_j\right)$. Thus, yielding the approximate solution u_h. See [64], for a proof of convergence which states that if the exact solution $u \in H_0^1(-1,1) \cap H^s(-1,1)$ then $\|u - u_h\|_0 \leqslant n^{-s}\|u\|_s$, $s \geqslant 1$.

2.4.2 Spectral collocation methods

For $0 \leqslant j \leqslant N$, let x_j denote the *Legendre Gauss-Lobatto nodes*, which are the zeros of $(1-x^2)L'_N(x)$ written in ascending order, i.e., $-1 = x_0 \leqslant x_1 \leqslant \ldots \leqslant x_N = 1$. Define the Lagrange basis functions $\ell_j(x) \in P_N$, where $\ell_j(x) = \prod_{\substack{k=0 \\ k \neq j}}^{N} \frac{(x - x_k)}{(x_j - x_k)}$, for all $0 \leqslant j \leqslant N$. Furthermore, consider the approximate solution

$$u_h = \sum_{i=0}^{N} u_i \ell_i(x).$$

Since $\ell_i(x_j) = \delta_{ij}$, it follows that $u_j = u(x_j)$, for all $0 \leqslant j \leqslant N$. Thus, $u(\pm 1) = 0$ gives $u_0 = u_N = 0$. For the remaining samples, the Poisson equation collocated at

$x = x_j$ leads to the following

$$-u''(x_j) = f(x_j), \quad 1 \leqslant j \leqslant N-1.$$

Define D_h to be a $(N+1) \times (N+1)$ matrix, so that $d_{ij} = \ell_j'(x_i)$, for $0 \leqslant i, j \leqslant N$, also known as the *pseudospectral Legendre Gauss-Lobatto derivative matrix*, then [64] the above equation can be written as the following linear system

$$-[\![D_h^2]\!]\mathbf{u}_h = \mathbf{f}_h,$$

where $\mathbf{u}_h = [u(x_1); u(x_2); \ldots; u(x_{N-1})]$ and $\mathbf{f}_h = [f(x_1); f(x_2); \ldots; f(x_{N-1})]$. Also, [64, p. 267] states that if $u \in H_0^1(-1,1) \cap H^s(-1,1)$, then for any positive integer N, $\|u - u_h\|_0 \leqslant cn^{-s}\|u\|_s$, $s \geqslant 1$. Therefore, proving a super-algebraic decay in error and implying spectral convergence.

There are several other spectral methods originating by incorporating various properties of orthogonal polynomials.

3

The Stokes problem

The Stokes equations are a linearized version of the Navier-Stokes equations and model incompressible viscous fluid flow with low Reynold's number. Several spectral methods, exhibiting exponential decay in error when the solution is analytic, are known to solve the steady-state Stokes problem numerically. A common strategy to solve such a problem in the time-dependent case involves extending the spectral scheme in spatial derivatives by implementing a low-order finite difference scheme for the time derivatives. We implement and analyze a space-time spectral method for the Stokes problem, which converges exponentially in both space and time. This numerical scheme imposes spectral collocation in time and the $P_N - P_{N-2}$ spectral Galerkin scheme in space by using a recombined Legendre polynomial basis, resulting in a global spectral operator that is a saddle point matrix. The main objectives of the research are estimating the condition number of the global spectral operators and proving the spectral convergence of this scheme in space and time. The analysis is not quite complete because two of the estimates are based on numerical evidence. However, some intermediate results, such as the 2-norm of the pseudospectral Chebyshev derivative matrix as well as condition number of the mass matrix and discrete Laplacian for a recombined Legendre basis, are proved to obtain the aforementioned

findings. Numerical experiments of this scheme verify the theoretical results.

3.1 Introduction

One of the topics investigated extensively in fluid dynamics is devising numerical schemes to solve the *Stokes problem*. The type of flow for which Reynold's number is low, say $R_e \ll 1$, i.e., the fluid velocity is extremely small, or the viscosity is very large, or an infinitesimal length scale is considered, is called the *Stokes flow* (or creeping flow). This type of flow is evident in many cases, such as swimming of a microorganism, flow of lava, flow of polymers, etc. The equations of motion for Stokes flow are called the *Stokes equations*, which along with suitable boundary and initial conditions are termed as the *Stokes problem*. In *steady state*, that is, independent of time, it is given as

$$\begin{aligned} -\Delta \mathbf{u} + \nabla p &= f \text{ in } \Omega := (-1,1)^2, \\ \nabla \cdot \mathbf{u} &= 0 \text{ in } \Omega, \\ \mathbf{u} &= 0 \text{ on } \partial\Omega, \end{aligned} \tag{3.1.1}$$

where velocity field and pressure are denoted by $\mathbf{u} = [u; v] \in V := (H_0^1(\Omega))^2$ and $p \in L_0^2(\Omega) := \left\{ q \in L^2(\Omega) \,\middle|\, \int_\Omega q = 0 \right\}$, respectively. It is defined in *unsteady state* as

$$\begin{aligned} \mathbf{u}_t - \Delta \mathbf{u} + \nabla p &= f \text{ in } \Omega_t := \Omega \times (-1,1), \\ \nabla \cdot \mathbf{u} &= 0 \text{ in } \Omega_t, \\ \mathbf{u} &= 0 \text{ on } \partial\Omega \times (-1,1), \\ \mathbf{u}(x,y,-1) &= \mathbf{u}_0(x,y) \text{ in } \Omega, \end{aligned} \tag{3.1.2}$$

where $\mathbf{u}(\cdot,t) \in (H_0^1(\Omega))^2$, $p(\cdot,t) \in L_0^2(\Omega)$, and $\mathbf{u}_0 \in (H_0^1(\Omega))^2$.

In [12], the authors describe three spectral methods for solving the Stokes prob-

lem. The first method is called single grid scheme, a collocation type using the same degree of polynomials for velocity and pressure; however, it provides spurious modes for pressure and, hence, is not used. The second method, the $P_N - P_{N-2}$ *scheme*, is a *mixed spectral Galerkin scheme* that uses polynomials of degree N for velocity and $N-2$ for pressure. The third one, the *staggered grid scheme*, is a spectral collocation method that uses staggered grids for velocity and pressure. In this research work, the last two methods will be applied. For all three methods, the inf-sup condition is not bounded independently of the discretization parameter of the scheme that decreases the accuracy of the error for pressure, which has been improved in [13] by proposing smaller discrete spaces for pressure.

In the past few years, space-time spectral methods, exhibiting spectral convergence in both space and time, are being used to solve time-dependent PDEs. A set practice was to implement a low-order finite difference approximation of the time derivative, which does not give spectral convergence for the whole scheme due to the dominance of the time discretization error. See [62, 39] for such schemes for linear PDEs, [69, 93, 19, 7, 18] and the references therein for problems related to the Stokes problem. Growing appeals for faster convergence in time generated this class of space-time spectral methods, some references for which are [47, 77, 87, 38, 81, 89, 88, 90, 91, 63, 101, 100]. A space-time spectral collocation method given in [92] was analyzed in [65] and [66] for Legendre and Chebyshev polynomials, respectively, based on which schemes for some linear PDEs were analyzed in [67], which serves as the motivation for this paper.

The aim of this work is to perform a condition number estimate for spectral method for the steady Stokes equations, and propose and analyze a space-time spectral method for the Stokes problem based upon an $P_N - P_{N-2}$ scheme; Stokes equations are more difficult to handle because it is a system of PDEs possessing different spaces for velocity and pressure. Thus, it requires the analysis for various terms ap-

pearing in the discrete problem, which includes proving condition number estimates for stiffness matrix, mass matrix and discrete Laplacian for a recombined Legendre basis derived in [79]. We also prove an estimate for the maximum singular value or the 2-norm for the Chebyshev-Gauss-Lobatto pseudospectral derivative matrix. This matrix is non-symmetric with an indefinite symmetric part, which makes the analysis more challenging. We believe our analysis of spectral convergence of the unsteady Stokes equations is new. We have also laid the ground work for a condition number estimate of the global space-time operator. A shortcoming of using this scheme is that it does not allow time stepping, the unknowns for all time need to be solved simultaneously. However, far fewer number of unknowns are required in comparison to finite difference discretizations in time. The results of the numerical experiments found clear support for the spectral convergence for these schemes for less than 20 spectral modes in each dimension.

This chapter is structured as follows. In Section 3.2, we define the notations being used in this chapter, define the spatial basis, and derive Proposition 3.2.4. In Section 3.3, we implement the $P_N - P_{N-2}$ scheme by using a recombined Legendre polynomial basis for the steady Stokes problem and prove the condition number estimates for the scheme in Sections 3.3.1 and 3.3.2, respectively. We extend the $P_N - P_{N-2}$ scheme to the unsteady Stokes problem by using the Chebyshev Gauss-Lobatto collocation in time in Section 3.4.1. Moreover, in Sections 3.4.2 and 3.4.3, we respectively prove the condition number estimates and spectral convergence in space and time.

3.2 Notations and fundamentals

We begin by summarizing some of the notations. Throughout this chapter, the discretization parameter is denoted by N, besides c and C denote some positive

constants independent of N. For an $n \times n$ matrix M, let $M]$ denote the $n \times (n-1)$ matrix obtained from M by deleting the first column, $[M$ denote the $(n-1) \times n$ matrix obtained from M by deleting the first row, $\mathbf{[}M\mathbf{]}$ denote the $(n-1) \times (n-1)$ matrix obtained from M by deleting the first column and row, and $[\![M]\!]$ denote the $(n-2) \times (n-2)$ matrix obtained from M by deleting the first and last columns and rows. The spectrum of M is denoted by $\Lambda(M)$.

Let P_N be the space of polynomials of degree less than or equal to N, and P_N^0 denote the polynomials in P_N that vanish at the end points $x = \pm 1$. Let $\mathbb{P}_{n_1,n_2}$ be the space of polynomials of degree less than equal to n_1, n_2 for x, y dimensions, receptively and $\mathbb{P}^0_{n_1,n_2} = \mathbb{P}_{n_1,n_2} \cap V$, i.e., they vanish on $\partial\Omega$. Let $\mathbb{P}_{n_1,n_2,m}$ be the space of polynomials of degree less than equal to n_1, n_2 for x, y dimensions and degree less than equal to m in time. Moreover, define $\mathbb{P}^0_{n_1,n_2,m}$ as the polynomials in $\mathbb{P}_{n_1,n_2,m}$ that vanish on the boundary of the spatial domain Ω.

The norm of $p \in L_0^2(\Omega)$ is given as,

$$\|p\|_0 = \left[\int_\Omega |p|^2\right]^{\frac{1}{2}},$$

and the inner product on the space $L_0^2(\Omega)$ is defined to be the same as that for $L^2(\Omega)$, which is defined as

$$\mathbf{(}f, g\mathbf{)} = \int_\Omega f(x)g(x)dx, \qquad f, g \in L^2(\Omega).$$

Finally, for our convenience we define the square of 2-norm of Legendre polynomials L_N of degree N, as $\gamma_N = \dfrac{2}{2N+1}$.

Now, we are ready to present some definitions that are the core of the spectral methods for the Stokes problem. The following definitions are given in [80, p. 145–146]. The first one defines the spatial basis for velocities.

Definition 3.2.1 (Recombined Legendre functions, see [80])**.** For some $i \in \mathbb{N} \cup \{0\}$,

define the *recombined Legendre functions* satisfying homogeneous Dirichlet boundary condition as

$$\phi_i(x) := L_i(x) - L_{i+2}(x), \qquad x \in [-1, 1]\,.$$

Therefore, $\phi_i(\pm 1) = 0$ implying $\phi_i \in P^0_{i+2}$.

Definition 3.2.2 (Stiffness Matrix, see [80])**.** The *stiffness matrix*, denoted by S, for the recombined Legendre basis functions ϕ_i, is defined as

$$s_{jk} := -\int_{-1}^{1} \phi_k''(x)\phi_j(x)dx.$$

It is a diagonal matrix with entries given as follows,

$$s_{kk} = (4k + 6). \tag{3.2.1}$$

Definition 3.2.3 (Mass Matrix, see [80])**.** The *mass matrix*, denoted by M, for the recombined Legendre basis functions ϕ_i, is defined as

$$m_{jk} = \int_{-1}^{1} \phi_j(x)\phi_k(x)dx.$$

It is a symmetric pentadiagonal matrix whose non-zero elements are given as follows,

$$m_{jk} = \begin{cases} \dfrac{2}{2k+1} + \dfrac{2}{2k+5}, & j = k, \\ -\dfrac{2}{2k+5}, & j = k+2. \end{cases} \tag{3.2.2}$$

Finally, we derive the following result for assistance in analysis performed in the next sections.

Proposition 3.2.4. *The matrix R, defined by $r_{mi} := \int_{-1}^{1} L_i(x)\phi_m'(x)dx$, and the*

matrix Q defined by $q_{nj} := \int_{-1}^{1} L_j(x)\phi_n(x)dx$, satisfy

$$r_{m,m+1} = -2, \qquad q_{nj} = \begin{cases} \gamma_n, & j = n, \\ -\gamma_{n+2}, & j = n+2, \end{cases}$$

where $\gamma_i = \|L_i\|_0^2 = \dfrac{2}{2i+1}$ for $i = 0, 1, \ldots$.

Proof. Since $\phi_m = L_m - L_{m+2}$,

$$r_{mi} = \int_{-1}^{1} L_i(x)\phi_m'(x)dx = \int_{-1}^{1} L_i(x)(L_m'(x) - L_{m+2}'(x))dx.$$

Using the recurrence relation,

$$(2n+1)L_n(x) = L_{n+1}'(x) - L_{n-1}'(x), \qquad n \in \mathbb{N},$$

the expression for r_{mi} becomes,

$$r_{mi} = -(2m+3)\int_{-1}^{1} L_i(x)L_{m+1}(x)dx = -(2m+3)\frac{2\delta_{i,m+1}}{2m+3}.$$

Similarly,

$$q_{nj} = \int_{-1}^{1} L_j(x)(L_n(x) - L_{n+2}(x))dx = \delta_{j,n}\gamma_n - \delta_{j,n+2}\gamma_{n+2}.$$

□

3.3 Steady state

The Stokes problem in *steady state* is given by (3.1.1), which on further simplification is expressed as:

$$-\Delta u + p_x = f_1 \text{ in } \Omega, \tag{3.3.1a}$$

$$-\Delta v + p_y = f_2 \text{ in } \Omega, \tag{3.3.1b}$$

$$u_x + v_y = 0 \text{ in } \Omega, \tag{3.3.1c}$$

$$u = 0,\ v = 0 \text{ on } \partial\Omega. \tag{3.3.1d}$$

3.3.1 Discretization

We implement the $P_N - P_{N-2}$ *scheme (a mixed spectral Galerkin scheme)* described in [12], by defining an *approximation for variables* as follows:

$$\begin{aligned}
u(x,y) &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\phi_i(x)\phi_j(y) \in \mathbb{P}^0_{N,N},\\
v(x,y) &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij}\phi_i(x)\phi_j(y) \in \mathbb{P}^0_{N,N},\\
p(x,y) &= \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}L_i(x)L_j(y) \in \mathbb{P}_{N-2,N-2} \cap L^2_0(\Omega),
\end{aligned}$$

so that $\int_\Omega p = 0$, i.e., it has zero average, since $\int_{-1}^{1} L_i(x)dx = 0$ for all $i \in \mathbb{N}$ and $L_0(x) = 1$ thus p_{00} is considered to be zero. Moreover, it is essential to implement this condition as (3.1.1) states that $p \in L^2_0(\Omega)$.

Define $\vartheta = (N-1)^2$, the number of unknowns for u and v each, and $\wp = (N-1)^2 - 1$, the number of unknowns for p. The total number of unknowns in the discrete Stokes equations are $2\vartheta + \wp = 3(N-1)^2 - 1$.

Define the *discrete unknowns* as $u_h = [u_{00}; u_{10}; \ldots; u_{N-2,0}; u_{01}; \ldots; u_{N-2,N-2}] \in \mathbb{R}^{\vartheta\times 1}$, similarly define $v_h \in \mathbb{R}^{\vartheta\times 1}$, and $p_h = [p_{10}; p_{20}; \ldots; p_{N-2,0}; p_{01}; \ldots; p_{N-2,N-2}] \in$

$\mathbb{R}^{\wp\times 1}$. Given the function $f_k(x,y)$ analytic in Ω, a forward discrete Legendre transform implies,

$$f_k(x,y) \approx \tilde{f}_k(x,y) = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^k L_i(x)L_j(y), \qquad k=1,2.$$

Define $F_k = [f_{00}^k; f_{10}^k; \ldots; f_{N-2,0}^k; f_{01}^k; \ldots; f_{N-2,N-2}^k] \in \mathbb{R}^{\vartheta\times 1}$, for $k=1,2$.

Consider (3.3.1a), then its weak form for $0 \leqslant m,n \leqslant N-2$, is given as follows

$$\big(-\Delta u + p_x, \phi_m(x)\phi_n(y)\big) = \big(f_1, \phi_m(x)\phi_n(y)\big). \tag{3.3.2}$$

Let us simplify the left hand side (LHS) of the above equation as follows,

$$\begin{aligned}
&\big(-\Delta u, \phi_m(x)\phi_n(y)\big) - \big(p, \phi_m'(x)\phi_n(y)\big)\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\big(-\phi_i''(x)\phi_j(y) - \phi_i(x)\phi_j''(y), \phi_m(x)\phi_n(y)\big)\\
&\quad - \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}\big(L_i(x)L_j(y), \phi_m'(x)\phi_n(y)\big)\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\Big(\int_{-1}^{1} -\phi_i''(x)\phi_m(x)dx \int_{-1}^{1}\phi_j(y)\phi_n(y)dy + \int_{-1}^{1}\phi_i(x)\phi_m(x)dx\int_{-1}^{1} -\phi_j''(y)\phi_n(y)dy\Big)\\
&\qquad - \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}\int_{-1}^{1} L_i(x)\phi_m'(x)dx\int_{-1}^{1} L_j(y)\phi_n(y)dy\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\left(s_{mi}m_{nj} + m_{mi}s_{nj}\right) - \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}r_{mi}q_{nj}.
\end{aligned}$$

The matrix form of the LHS of (3.3.2), for all $0 \leqslant m,n \leqslant N-2$, becomes

$$\left(M\otimes S + S\otimes M\right)u_h - (Q\otimes R)]p_h,$$

where $S=[s_{ij}]$, $M=[m_{ij}]$, $Q=[q_{ij}]$, and $R=[r_{ij}]$, for $0\leqslant i,j\leqslant N-2$. Also, the

RHS of (3.3.2) is approximated by $\tilde{f}_1$ as follows

$$\begin{aligned}
&\sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^k\big(L_i(x)L_j(y), \phi_m(x)\phi_n(y)\big),\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^k \int_{-1}^{1} L_i(x)\phi_m(x)dx \int_{-1}^{1} L_j(y)\phi_n(y)dy\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^k q_{mi}q_{nj}.
\end{aligned}$$

Thus, the matrix form the RHS of (3.3.2), for all $0 \leqslant m, n \leqslant N-2$, becomes

$$(Q \otimes Q)\, F_k.$$

Therefore, the equation (3.3.1a) implies

$$(M \otimes S + S \otimes M)u_h - (Q \otimes R)]p_h = (Q \otimes Q)\, F_1. \tag{3.3.3}$$

Similarly, the equation (3.3.1b) implies

$$(M \otimes S + S \otimes M)v_h - (R \otimes Q)]p_h = (Q \otimes Q)\, F_2. \tag{3.3.4}$$

Finally, the weak form of (3.3.1c), for $0 \leqslant m, n \leqslant N-2$ and $m+n > 0$, is given as

$$\begin{aligned}
0 &= \big(u_x + v_y, L_m(x)L_n(y)\big) = \big(u_x, L_m(x)L_n(y)\big) + \big(v_y, L_m(x)L_n(y)\big)\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\big(\phi_i'(x)\phi_j(y), L_m(x)L_n(y)\big) + \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij}\big(\phi_i(x)\phi_j'(y), L_m(x)L_n(y)\big)\\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}\int_{-1}^{1}\phi_i'(x)L_m(x)dx\int_{-1}^{1}\phi_j(y)L_n(y)dy\\
&\qquad + \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij}\int_{-1}^{1}\phi_i(y)L_m(y)dy\int_{-1}^{1}\phi_j'(y)L_n(y)dy\\
&= -\sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}r_{im}q_{jn} - \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij}q_{im}r_{jn}
\end{aligned}$$

$$= -\sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij} r_{mi}^T q_{nj}^T - \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij} q_{mi}^T r_{nj}^T,$$

which yields the following matrix form,

$$-[\left(Q^T \otimes R^T\right) u_h - [\left(R^T \otimes Q^T\right) v_h = O_{\wp,1}. \tag{3.3.5}$$

Thus, eqs. (3.3.3) to (3.3.5) together form the following *discrete Stokes problem*,

$$\begin{aligned}
(M \otimes S + S \otimes M)u_h - (Q \otimes R)]p_h &= (Q \otimes Q)\, F_1, \\
(M \otimes S + S \otimes M)v_h - (R \otimes Q)]p_h &= (Q \otimes Q)\, F_2, \\
-[(Q^T \otimes R^T)u_h - [(R^T \otimes Q^T)v_h &= O_{\wp,1}.
\end{aligned}$$

When written in matrix form, it gives the following saddle point linear system,

$$\begin{bmatrix} M \otimes S + S \otimes M & O_{\vartheta,\vartheta} & -(Q \otimes R)] \\ O_{\vartheta,\vartheta} & M \otimes S + S \otimes M & -(R \otimes Q)] \\ -[(Q^T \otimes R^T) & -[(R^T \otimes Q^T) & O_{\wp,\wp} \end{bmatrix} \begin{bmatrix} u_h \\ v_h \\ p_h \end{bmatrix} = \begin{bmatrix} (Q \otimes Q)\, F_1 \\ (Q \otimes Q)\, F_2 \\ O_{\wp,1} \end{bmatrix}.$$

Thus, the *global spectral operator of discrete Stokes problem* becomes,

$$G = \begin{bmatrix} \mathcal{A} & B \\ B^T & O_{\wp,\wp} \end{bmatrix} \in \mathbb{R}^{(2\vartheta+\wp)\times(2\vartheta+\wp)}, \tag{3.3.6}$$

where, $\mathcal{A} = A \oplus A \in \mathbb{R}^{2\vartheta\times 2\vartheta}$, that is,

$$\mathcal{A} = \begin{bmatrix} A & O_{\vartheta,\vartheta} \\ O_{\vartheta,\vartheta} & A \end{bmatrix}, \text{ with } A = M \otimes S + S \otimes M \in \mathbb{R}^{\vartheta\times\vartheta}, \tag{3.3.7}$$

and

$$B = \begin{bmatrix} B_1 \\ B_2 \end{bmatrix} \in \mathbb{R}^{2\vartheta \times \wp}, \text{ with } B_1 = -(Q \otimes R)], B_2 = -(R \otimes Q)] \in \mathbb{R}^{\vartheta \times \wp}. \quad (3.3.8)$$

Note that although for (3.1.1) the velocity $\mathbf{u}$ is divergence-free or $\nabla \cdot \mathbf{u} = 0$, that is not the case for the approximate solution obtained from the $P_N - P_{N-2}$ scheme. Observe that this method implements the weak form of (3.3.1d) for the approximate solution of velocity $\mathbf{u} = [u, v]^T$ given as $u_N, v_N \in \mathbb{P}^0_{N,N}$ as follows,

$$\left(q_{N-2}, u_{N_x} + v_{N_y}\right) = 0, \quad (3.3.9)$$

for all $q_{N-2} \in \mathbb{P}_{N-2,N-2}$. Thus, u_N, v_N are not divergence free, however, all divergence-free polynomials in $\mathbb{P}^0_{N,N}$ satisfy the above equation. For more details, see [12, p. 416].

This is not a major drawback as it is overpowered by the property that this scheme eliminates the presence of any *spurious modes* on pressure. A function $q_{N-2} \in \mathbb{P}_{N-2,N-2} \cap L^2_0(\Omega)$ is called a *spurious mode* if it satisfies the following condition:

$$\left(q_{N-2}, u_{N_x} + v_{N_y}\right) = 0,$$

for all $u_N, v_N \in \mathbb{P}^0_{N,N}$. If such a polynomial q_{N-2} exists then the solution for pressure is not unique, since for every solution p_{N-2}, the polynomial $p_{N-2} \pm q_{N-2}$ will also be a solution. Thus, it is imperative to not have any spurious modes. For this scheme, Theorem 25.1 given in [12] states that the set of spurious modes for this scheme is equal to $\{0\}$, thus there are no spurious modes. Hence, the following inf-sup condition holds,

$$\inf_{q_N \in \mathbb{P}_{N-2,N-2} \cap L^2_0(\Omega)} \sup_{v_n \in \mathbb{P}^0_{N,N}} \frac{\left(\nabla \cdot v_N, q_N\right)}{\|v_N\|_1 \|q_N\|_0} \geqslant \frac{c}{\sqrt{N}},$$

which by Theorem 23.8 in [12] implies the uniqueness of solution for the discrete Stokes problem obtained by the $P_N - P_{N-2}$ scheme.

As stated on [12, p. 424], the following are the main features of this scheme for the Stokes problem:

1. the velocity is not exactly divergence-free,
2. this method is a spectral-Galerkin scheme,
3. there are no spurious modes for pressure;
4. the best constant of the inf-sup condition on the pressure is of order $N^{-\frac{1}{2}}$.

3.3.2 Analysis

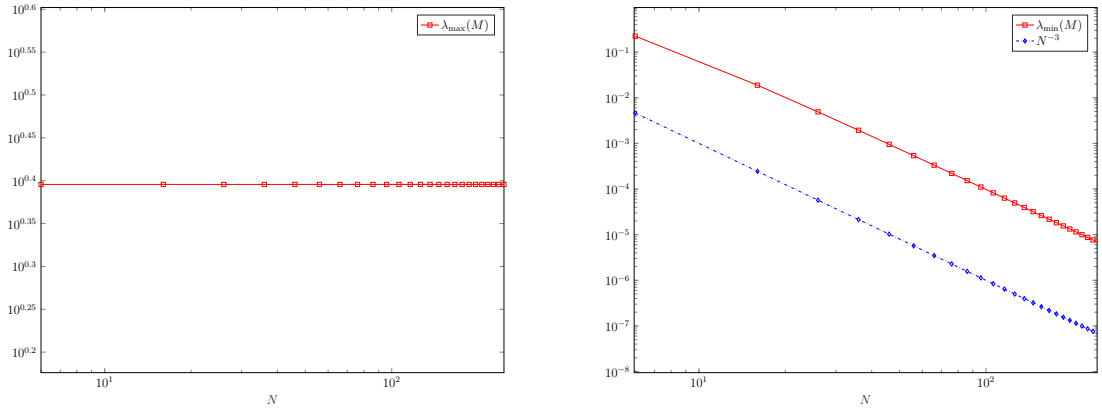
In this section, we estimate the condition number of the global matrix G for the discretized steady Stokes problem given by (3.3.6). Since G is a symmetric saddle point matrix with a SPD leading block and full rank matrix B, finding the bounds for spectrum of G is facilitated by the theory of spectrum of a symmetric saddle point matrix with the desired properties. This analysis requires bounds on the spectrum of the sub-blocks of G, thus we proceed as follows.

Lemma 3.3.1. *For $N \geqslant 2$, let $S \in \mathbb{R}^{(N-1)\times(N-1)}$ be the stiffness matrix defined by (3.2.1), then it is SPD with $\lambda_{\min}(S) = 6$ and $\lambda_{\max}(S) = 4N-2$, thus $\kappa(S) = \dfrac{2N-1}{3}$.*

Proof. By (3.2.1), S is a diagonal matrix with entries $s_{kk} = 4k+6$ for $0 \leqslant k \leqslant N-2$, thus $\lambda_{\min}(S) = 6$ and $\lambda_{\max}(S) = 4N-2$. Since the stiffness matrix S is an SPD, its condition number is $\kappa(S) = \frac{\sigma_{\max}(S)}{\sigma_{\min}(S)} = \frac{\lambda_{\max}(S)}{\lambda_{\min}(S)} = \frac{4N-2}{6}$. □

The following results give optimal condition number estimate for the mass matrix and discrete Laplacian matrix in two dimensions for the recombined Legendre basis considered in this scheme, derived in [80] for Dirichlet boundary conditions.

These results appear to be new. The optimality of these estimates is evident from Figures 3.1a and 3.1b.



(a) Maximum eigenvalue. (b) Minimum eigenvalue.

Figure 3.1: Spectrum of M.

Lemma 3.3.2. *For $N \geqslant 4$, let $M \in \mathbb{R}^{(N-1)\times(N-1)}$ be the mass matrix defined by (3.2.2). Then, it is SPD and $\frac{c}{N^3} \leqslant \Lambda(M) \leqslant C$, thus $\kappa(M) \leqslant cN^3$.*

Proof. Let $u(x) = \sum_{i=0}^{N-2} u_i\phi_i(x) \in \mathbb{P}_N^0$, where ϕ_i represent recombined Legendre basis functions and define $u_h := [u_0; u_1; \ldots; u_{N-2}] \in \mathbb{R}^{(N-1)\times 1}$. Then

$$\|u\|_0^2 = \int_{-1}^{1} u(x)^2 dx = \int_{-1}^{1} \left(\sum_{i=0}^{N-2} u_i\phi_i\right)^2 dx = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_i u_j \int_{-1}^{1} \phi_i(x)\phi_j(x)dx,$$

by using the definition of entries of the mass matrix,

$$\|u\|_0^2 = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_i u_j m_{ij} = u_h^T M u_h.$$

Hence M is SPD, for any $x \in \mathbb{R}^{(N-1)\times 1}\backslash\{0\}$, the bounds on the eigenvalues of M by estimating $x^T Mx$ are derived as follows.

$$x^T Mx = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} x_i m_{ij} x_j$$

$$
\begin{aligned}
&= \sum_{i=0}^{N-2} x_i^2 m_{ii} + 2\sum_{i=0}^{N-4} x_i x_{i+2} m_{i,i+2} && \text{(by using (3.2.2))}\\
&= 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) - 4\sum_{i=0}^{N-4}\frac{x_i x_{i+2}}{2i+5}, && (3.3.10)\\
&\leqslant 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) + 4\sum_{i=0}^{N-4}\frac{|x_i||x_{i+2}|}{2i+5}\\
&\leqslant 2\sum_{i=0}^{N-2} x_i^2\left(1+\frac{1}{5}\right) + \frac{4}{5}\sum_{i=0}^{N-4}|x_i||x_{i+2}|\\
&\leqslant \frac{12}{5}\sum_{i=0}^{N-2} x_i^2 + \frac{4}{5}\sqrt{\sum_{i=0}^{N-4}|x_i|^2\sum_{i=0}^{N-4}|x_{i+2}|^2}\\
&\leqslant \frac{12}{5}\sum_{i=0}^{N-2} x_i^2 + \frac{4}{5}\sum_{i=0}^{N-2} x_i^2 = \frac{16}{5}\sum_{i=0}^{N-2} x_i^2.
\end{aligned}
$$

Hence, $x^T Mx \leqslant C\|x\|_2^2$, therefore $\lambda_{\max}(M) \leqslant C$.

Note that

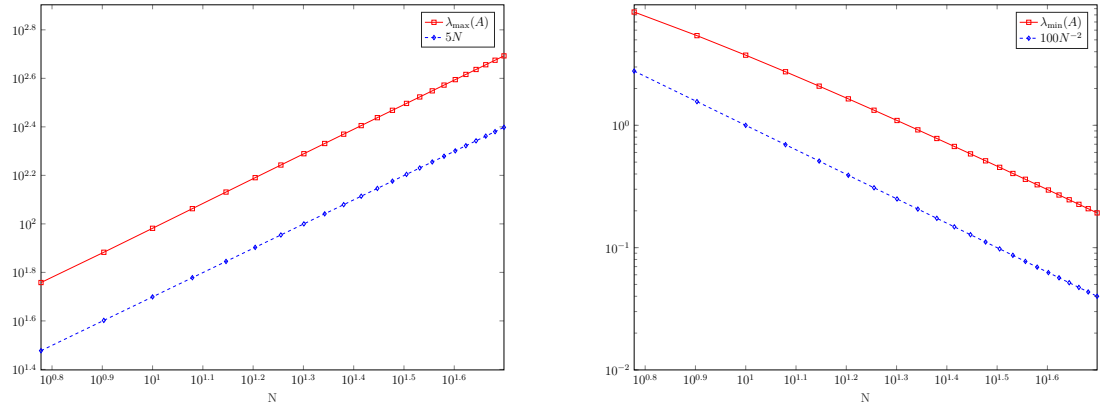
$$
4\sum_{i=0}^{N-4}\frac{x_i x_{i+2}}{2i+5} \leqslant 4\sum_{i=0}^{N-4}\frac{|x_i||x_{i+2}|}{2i+5}.
$$

By using the above in (3.3.10),

$$
\begin{aligned}
x^T Mx &\geqslant 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) - 4\sum_{i=0}^{N-4}\frac{|x_i||x_{i+2}|}{2i+5}\\
&= 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) - 4\sum_{i=0}^{N-4}\frac{\sqrt{2i+9}|x_i|}{2i+5}\cdot\frac{|x_{i+2}|}{\sqrt{2i+9}}\\
&\geqslant 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) - 2\sum_{i=0}^{N-4}\left(\frac{(2i+9)x_i^2}{(2i+5)^2}+\frac{x_{i+2}^2}{(2i+9)}\right)\\
&= 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}\right) - 2\sum_{i=0}^{N-4}\frac{(2i+9)x_i^2}{(2i+5)^2} - 2\sum_{i=2}^{N-2}\frac{x_i^2}{(2i+5)}\\
&\geqslant 2\sum_{i=0}^{N-2} x_i^2\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{(2i+9)}{(2i+5)^2}-\frac{1}{(2i+5)}\right)\\
&= 32\sum_{i=0}^{N-2} x_i^2\left(\frac{x_i^2}{(2i+1)(2i+5)^2}\right)\\
&\geqslant \frac{c}{N^3}|x|_2^2.
\end{aligned}
$$

Thus $\lambda_{\min}(M) \geqslant \dfrac{c}{N^3}$ and $\kappa(M) = \dfrac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \leqslant cN^3$. □

The following result provides a sharp condition number estimate for A, as shown in Figures 3.2a and 3.2b.



(a) Maximum eigenvalue. (b) Minimum eigenvalue.

Figure 3.2: Spectrum of A.

Theorem 3.3.3. *For $N \geqslant 4$, let the discrete Laplacian A be defined by* (3.3.7)*, then $\dfrac{c}{N^2} \leqslant \Lambda(A) \leqslant CN$, thus $\kappa(A) \leqslant cN^3$.*

Proof. Since $A \in \mathbb{R}^{\vartheta \times \vartheta}$ and is defined as $A = M \otimes S + S \otimes M$. Note that it is SPD, as both M and S are SPD, hence Theorem 2.1.6 yields

$$\begin{aligned}\lambda_{\max}(A) &\leqslant \lambda_{\max}(M \otimes S) + \lambda_{\max}(S \otimes M) \\ &= \lambda_{\max}(M)\lambda_{\max}(S) + \lambda_{\max}(S)\lambda_{\max}(M) = 2\lambda_{\max}(M)\lambda_{\max}(S),\end{aligned}$$

where we get the last equality by Result 4 of Theorem 2.1.7. Thus, Lemmas 3.3.1 and 3.3.2 give $\lambda_{\max}(A) \leqslant C(4N - 2) \leqslant CN$.

The definition of S and M, given by the equations (3.2.1) and (3.2.2) respectively, implies that $A \in \mathbb{R}^{\vartheta \times \vartheta}$ is a symmetric block matrix, with non-zero 0, 2 and -2 block

diagonals. Its blocks are defined by

$$A_{jk} = \begin{cases} s_{jj}M + m_{jj}S & j = k, \\ m_{jk}S & j = k+2, \end{cases} \tag{3.3.11}$$

for $0 \leqslant j, k \leqslant N$. Let $x = [x_0; x_1; \ldots; x_{N-2}] \in \mathbb{R}^{\vartheta \times 1} \backslash \{0\}$, where $x_i = \left[x_i^0; x_i^1; \ldots; x_i^{N-2}\right] \in \mathbb{R}^{(N-1)\times 1}$ for each $0 \leqslant i \leqslant N-1$.

$$\begin{aligned} x^T Ax &= \left[x_0^T, x_1^T, \ldots, x_{N-2}^T\right] Ax = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} x_i^T A_{ij} x_j \\ &= \sum_{i=0}^{N-2} x_i^T A_{ii} x_i + \sum_{i=0}^{N-2}\sum_{\substack{j=0 \\ j\neq i}}^{N-2} x_i^T A_{ij} x_j \\ &\qquad\qquad \text{(as A is symmetric and has non-zero 2, -2 block diagonals)} \\ &= \sum_{i=0}^{N-2} x_i^T A_{ii} x_i + 2\sum_{i=0}^{N-4} x_i^T A_{i,i+2} x_{i+2} \qquad \text{(by using (3.3.11))} \\ &= \sum_{i=0}^{N-2} x_i^T \left(s_{ii}M + m_{ii}S\right) x_i + 2\sum_{i=0}^{N-4} x_i^T m_{ii+2} S x_{i+2} \\ &= \sum_{i=0}^{N-2} x_i^T \left[(4i+6)\, M + \left[\frac{2}{2i+1} + \frac{2}{2i+5}\right] S\right] x_i + 2\sum_{i=0}^{N-4} \frac{-2}{2i+5} x_i^T S x_{i+2} \\ &\qquad\qquad \text{(by using (3.2.1) and (3.2.2))} \\ &= \sum_{i=0}^{N-2} (4i+6)\, x_i^T M x_i + \sum_{i=0}^{N-2}\left[\frac{2}{2i+1} + \frac{2}{2i+5}\right] x_i^T S x_i + 2\sum_{i=0}^{N-4} \frac{-2}{2i+5} x_i^T S x_{i+2}. \end{aligned}$$

As $x_i^T M x_i = \sum_{k=0}^{N-2}\sum_{j=0}^{N-2} x_i^k m_{kj} x_i^j$, then (3.3.10) gives

$$x_i^T M x_i = \sum_{j=0}^{N-2} (x_i^j)^2 \left[\frac{2}{2j+1} + \frac{2}{2j+5}\right] + 2\sum_{j=0}^{N-4} \frac{-2}{2j+5} x_i^j x_i^{j+2},$$

and similarly

$$x_i^T S x_i = \sum_{j=0}^{N-2} (x_i^j)^2 \,(4j+6)\,,$$
$$x_i^T S x_{i+2} = \sum_{j=0}^{N-2} x_i^j \,(4j+6)\, x_{i+2}^j.$$

The above three equations imply

$$\begin{aligned}
x^T A x &= \sum_{i=0}^{N-2} (4i+6) \left[\sum_{j=0}^{N-2} (x_i^j)^2 \left[\frac{2}{2j+1} + \frac{2}{2j+5} \right] + 2 \sum_{j=0}^{N-4} \frac{-2}{2j+5} x_i^j x_i^{j+2} \right] \\
&\quad + \sum_{i=0}^{N-2} \left[\frac{2}{2i+1} + \frac{2}{2i+5} \right] \sum_{j=0}^{N-2} (x_i^j)^2 \,(4j+6) + 2 \sum_{i=0}^{N-4} \frac{-2}{2i+5} \sum_{j=0}^{N-2} x_i^j \,(4j+6)\, x_{i+2}^j, \\
&= 2 \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} (x_i^j)^2 \left[(4i+6) \left[\frac{1}{2j+1} + \frac{1}{2j+5} \right] + (4j+6) \left[\frac{1}{2i+1} + \frac{1}{2i+5} \right] \right] \\
&\quad - 4 \sum_{i=0}^{N-2} (4i+6) \sum_{j=0}^{N-4} \frac{1}{2j+5} x_i^j x_i^{j+2} - 4 \sum_{i=0}^{N-4} \frac{1}{2i+5} \sum_{j=0}^{N-2} (4j+6)\, x_i^j x_{i+2}^j, \\
&\geqslant 2 \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} (x_i^j)^2 \left[(4i+6) \left[\frac{1}{2j+1} + \frac{1}{2j+5} \right] + (4j+6) \left[\frac{1}{2i+1} + \frac{1}{2i+5} \right] \right] \\
&\quad - 4 \sum_{i=0}^{N-2} (4i+6) \sum_{j=0}^{N-4} \left(\frac{\sqrt{2j+9}|x_i^j|}{2j+5} \right) \left(\frac{|x_i^{j+2}|}{\sqrt{2j+9}} \right) \\
&\quad - 4 \sum_{j=0}^{N-2} (4j+6) \sum_{i=0}^{N-4} \left(\frac{\sqrt{2i+9}|x_i^j|}{2i+5} \right) \left(\frac{|x_{i+2}^j|}{\sqrt{2i+9}} \right), \\
&\qquad\qquad \text{(as for } a, b \geqslant 0,\ a^2 + b^2 \geqslant 2ab) \\
&\geqslant 2 \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} (x_i^j)^2 \left[(4i+6) \left[\frac{1}{2j+1} + \frac{1}{2j+5} \right] + (4j+6) \left[\frac{1}{2i+1} + \frac{1}{2i+5} \right] \right] \\
&\quad - 2 \sum_{i=0}^{N-2} (4i+6) \sum_{j=0}^{N-4} \left[\frac{(2j+9)(x_i^j)^2}{(2j+5)^2} + \frac{(x_i^{j+2})^2}{(2j+9)} \right] \\
&\quad - 2 \sum_{j=0}^{N-2} (4j+6) \sum_{i=0}^{N-4} \left[\frac{(2i+9)(x_i^j)^2}{(2i+5)^2} + \frac{(x_{i+2}^j)^2}{(2i+9)} \right], \\
&\qquad\qquad \text{(by changing variable of summation)} \\
&\geqslant 2 \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} (x_i^j)^2 \left[(4i+6) \left[\frac{1}{2j+1} + \frac{1}{2j+5} \right] + (4j+6) \left[\frac{1}{2i+1} + \frac{1}{2i+5} \right] \right]
\end{aligned}$$

$$
\begin{aligned}
&- 2\sum_{i=0}^{N-2}(4i+6)\left[\sum_{j=0}^{N-4}\frac{(2j+9)(x_i^j)^2}{(2j+5)^2}+\sum_{j=2}^{N-2}\frac{(x_i^j)^2}{(2j+5)}\right]\\
&\quad - 2\sum_{j=0}^{N-2}(4j+6)\left[\sum_{i=0}^{N-4}\frac{(2i+9)(x_i^j)^2}{(2i+5)^2}+\sum_{i=2}^{N-2}\frac{(x_i^j)^2}{(2i+5)}\right],\\
&= 2\sum_{i=0}^{N-2}(4i+6)\left[\sum_{j=0}^{1}\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right)(x_i^j)^2\right.\\
&\quad \left.+\sum_{j=2}^{N-4}\left(\frac{1}{2j+1}-\frac{2j+9}{(2j+5)^2}\right)(x_i^j)^2+\sum_{j=N-3}^{N-2}\frac{(x_i^j)^2}{2j+1}\right]\\
&\qquad + 2\sum_{j=0}^{N-2}(4j+6)\left[\sum_{i=0}^{1}\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{2i+9}{(2i+5)^2}\right)(x_i^j)^2\right.\\
&\qquad\quad \left.+\sum_{i=2}^{N-4}\left(\frac{1}{2i+1}-\frac{2i+9}{(2i+5)^2}\right)(x_i^j)^2+\sum_{i=N-3}^{N-2}\frac{(x_i^j)^2}{2i+1}\right],\\
&= 2\sum_{i=0}^{N-2}(4i+6)\left[\sum_{j=0}^{1}\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right)(x_i^j)^2+\sum_{j=2}^{N-4}\frac{16(x_i^j)^2}{(2j+1)(2j+5)^2}\right.\\
&\quad \left.+\sum_{j=N-3}^{N-2}\frac{(x_i^j)^2}{2j+1}\right]+2\sum_{j=0}^{N-2}(4j+6)\left[\sum_{i=0}^{1}\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{2i+9}{(2i+5)^2}\right)(x_i^j)^2\right.\\
&\qquad \left.+\sum_{i=2}^{N-4}\frac{16(x_i^j)^2}{(2i+1)(2i+5)^2}+\sum_{i=N-3}^{N-2}\frac{(x_i^j)^2}{2i+1}\right].
\end{aligned}
$$

The above expression consists of six double summations, we split them into the combination of $i = 0,\ 1,\ 2 \leqslant i \leqslant N-4$ and $i = N-3,\ N-2$. This partitions each double summation into three parts and thus a total of eighteen double summations. On combining nine double summations from the first double summation to the respective nine from the second one, we get the following nine terms, precisely

$$
x^T Ax = 2\sum_{m=1}^{9} S_m, \tag{3.3.12}
$$

where

$$S_1 = \sum_{i=0}^{1}\sum_{j=0}^{1}(x_i^j)^2\Bigg[\,(4i+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right) + (4j+6)\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{2i+9}{(2i+5)^2}\right)\Bigg], \tag{3.3.13}$$

$$S_2 = \sum_{i=2}^{N-4}\sum_{j=0}^{1}(x_i^j)^2\left[(4i+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right)+\frac{16(4j+6)}{(2i+1)(2i+5)^2}\right], \tag{3.3.14}$$

$$S_3 = \sum_{i=N-3}^{N-2}\sum_{j=0}^{1}(x_i^j)^2\left[(4i+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right)+\frac{4j+6}{2i+1}\right], \tag{3.3.15}$$

$$S_4 = \sum_{i=0}^{1}\sum_{j=2}^{N-4}(x_i^j)^2\left[\frac{16(4i+6)}{(2j+1)(2j+5)^2}+(4j+6)\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{2i+9}{(2i+5)^2}\right)\right], \tag{3.3.16}$$

$$S_5 = \sum_{i=2}^{N-4}\sum_{j=2}^{N-4}(x_i^j)^2\left[\frac{16(4i+6)}{(2j+1)(2j+5)^2}+\frac{16(4j+6)}{(2i+1)(2i+5)^2}\right], \tag{3.3.17}$$

$$S_6 = \sum_{i=N-3}^{N-2}\sum_{j=2}^{N-4}(x_i^j)^2\left[\frac{16(4i+6)}{(2j+1)(2j+5)^2}+\frac{4j+6}{2i+1}\right], \tag{3.3.18}$$

$$S_7 = \sum_{i=0}^{1}\sum_{j=N-3}^{N-2}(x_i^j)^2\left[\frac{4i+6}{2j+1}+(4j+6)\left(\frac{1}{2i+1}+\frac{1}{2i+5}-\frac{2i+9}{(2i+5)^2}\right)\right], \tag{3.3.19}$$

$$S_8 = \sum_{i=2}^{N-4}\sum_{j=N-3}^{N-2}(x_i^j)^2\left[\frac{4i+6}{2j+1}+\frac{16(4j+6)}{(2i+1)(2i+5)^2}\right], \tag{3.3.20}$$

and, finally

$$S_9 = \sum_{i=N-3}^{N-2}\sum_{j=N-3}^{N-2}(x_i^j)^2\left[\frac{4i+6}{2j+1}+\frac{4j+6}{2i+1}\right]. \tag{3.3.21}$$

We claim that all of the above nine terms are bound below by $\frac{c}{N^2}$.

First of all, the sum S_1, given by (3.3.13), contains only constants independent of N and thus implies

$$S_1 \geqslant \frac{c}{N^2}\sum_{i=0}^{1}\sum_{j=0}^{1}(x_i^j)^2. \tag{3.3.22}$$

For the second one, (3.3.14) yields

$$\begin{aligned} S_2 &\geqslant \sum_{i=2}^{N-4}\sum_{j=0}^{1}(x_i^j)^2\,(4i+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right) \\ &\geqslant \sum_{i=2}^{N-4}\sum_{j=0}^{1}(4(2)+6)(x_i^j)^2\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right), \\ &\qquad\qquad \text{(note that, terms are constants independent of } N) \\ &\geqslant \frac{c}{N^2}\sum_{i=2}^{N-4}\sum_{j=0}^{1}(x_i^j)^2. \end{aligned} \tag{3.3.23}$$

Then (3.3.15) gives,

$$\begin{aligned} S_3 &\geqslant \sum_{i=N-3}^{N-2}\sum_{j=0}^{1}(x_i^j)^2\,(4i+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right) \\ &\geqslant \sum_{i=N-3}^{N-2}\sum_{j=0}^{1}(x_i^j)^2\,(4(N-3)+6)\left(\frac{1}{2j+1}+\frac{1}{2j+5}-\frac{2j+9}{(2j+5)^2}\right) \\ &\geqslant cN\sum_{i=N-3}^{N-2}\sum_{j=0}^{1}(x_i^j)^2 \geqslant \frac{c}{N^2}\sum_{i=N-3}^{N-2}\sum_{j=0}^{1}(x_i^j)^2. \end{aligned} \tag{3.3.24}$$

Note that the term S_4, given by the equation (3.3.16), is similar to the term S_2, given by the equation (3.3.14), and therefore

$$S_4 \geqslant \frac{c}{N^2}\sum_{i=0}^{1}\sum_{j=2}^{N-4}(x_i^j)^2. \tag{3.3.25}$$

Since for any $i,j \in \mathbb{N}$, $2j \leqslant (2j+1) \leqslant (2j+5)$ holds, thus

$$\frac{16(4i+6)}{(2j+1)(2j+5)^2} \geqslant \frac{16\cdot 4i}{(2i)^3} \geqslant \frac{i}{j^3},$$

and using the above relation in the equation (3.3.17) gives the following expression

$$S_5 \geqslant \sum_{i=2}^{N-4}\sum_{j=2}^{N-4}(x_i^j)^2\left[\frac{i}{j^3}+\frac{j}{i^3}\right]. \tag{3.3.26}$$

At this stage, we need to prove that for all $i, j \in \mathbb{N}$ and $2 \leqslant i, j \leqslant N-4$,

$$\frac{i}{j^3} + \frac{j}{i^3} \geqslant \frac{c}{N^2}. \tag{3.3.27}$$

To this end, we define the function

$$f(x, y) := \frac{x}{y^3} + \frac{y}{x^3}, \tag{3.3.28}$$

and minimize it over the domain $\Omega_f := 2 \leqslant x, y \leqslant N-4$. The boundary of the domain is given by $\Gamma_f := \bigcup_{m=1}^{4} \Gamma_m$, where

$$\begin{aligned} \Gamma_1 : &\quad x = 2, \\ \Gamma_2 : &\quad x = N-4, \\ \Gamma_3 : &\quad y = 2, \\ \Gamma_4 : &\quad y = N-4. \end{aligned}$$

We first find the critical points of f, as follows

$$0 = f_x(x, y) = \frac{1}{y^3} - \frac{3y}{x^4},$$

this implies, $3y^4 = x^4$ and since f is a symmetric function about x and y, hence $f_y = 0$ gives $3x^4 = y^4$. Solving these two equations, we get $x^4 = 9x^4$, and thus $x = 0$, which further gives $y = 0$. Hence the critical point is origin, which lies outside of the domain and is thus rejected. Let us now estimate the minimum of f on the boundary Γ_f. On Γ_1, we minimize

$$g_1(y) := f(2, y) = \frac{2}{y^3} + \frac{y}{8}, \tag{3.3.29}$$

for $2 \leqslant y \leqslant N-4$. The critical points of g_1 are

$$0 = g_1'(y) = -\frac{6}{y^4} + \frac{1}{8},$$

which implies $y^4 = 2^4 \cdot 3$, thus $y^2 = 2^2\sqrt{3}$ and $y = \pm 2\sqrt[4]{3}$, for which we reject the negative which lies outside of the domain of g_1. The possible points of minimum in domain are $y = 0, 2 \cdot \sqrt[4]{3}, N-4$, at which g_1 attains the following values

$$\begin{aligned} g_1(2) &= \frac{1}{2}, \\ g_1(2\sqrt[4]{3}) &= \sqrt[4]{\frac{1}{3^3}}, \\ g_1(N-4) &= \frac{2}{(N-4)^3} + \frac{(N-4)}{8} \geqslant \frac{N-4}{8} \geqslant \frac{c}{N^2}. \end{aligned}$$

Thus $g_1(y) \geqslant \frac{c}{N^2}$. Moreover, on Γ_2 we have minimize

$$g_2(y) := f(N-4, y) = \frac{(N-4)}{y^3} + \frac{y}{(N-4)^4}, \tag{3.3.30}$$

for $2 \leqslant y \leqslant N-4$. The critical points of g_2 are

$$0 = g_1'(y) = -\frac{3(N-4)}{y^4} + \frac{1}{(N-4)^3},$$

which implies $y^4 = 3(N-4)^4$, thus $y^2 = (N-4)^2\sqrt{3}$ and $y = \pm(N-4)\sqrt[4]{3}$. Both of the critical points lie outside of the domain. The possible points of minimum in domain are $y = 0, N-4$, at which g_2 attains the following values

$$\begin{aligned} g_2(2) &= \frac{N-4}{8} + \frac{2}{(N-4)^3} \geqslant \frac{c}{N^2}, \\ g_2(N-4) &= \frac{(N-4)}{(N-4)^3} + \frac{(N-4)}{(N-4)^3} = \frac{2}{(N-4)^2} \geqslant \frac{c}{N^2}. \end{aligned}$$

Hence it is concluded that $g_2(y) \geqslant \frac{c}{N^2}$. The bounds for g_1 and g_2 imply that $f \geqslant \frac{c}{N^2}$

on Γ_1 and Γ_2. Since the function f is symmetric about x and y, we get $f \geqslant \frac{c}{N^2}$ on Γ_f and hence on Ω_f, as it does not possess any critical point in Ω_f. By using this result we obtain the result (3.3.27), applying which on the inequality (3.3.26) gives

$$S_5 \geqslant \frac{c}{N^2} \sum_{i=2}^{N-4} \sum_{j=2}^{N-4} (x_i^j)^2. \tag{3.3.31}$$

Looking forward at the term S_6, given by (3.3.18),

$$\begin{aligned} S_6 &\geqslant \sum_{i=N-3}^{N-2} \sum_{j=2}^{N-4} (x_i^j)^2 \frac{4j+6}{2i+1} \geqslant \sum_{i=N-3}^{N-2} \sum_{j=2}^{N-4} (x_i^j)^2 \frac{4(2)+6}{2i+1} \\ &\geqslant \sum_{i=N-3}^{N-2} \sum_{j=2}^{N-4} (x_i^j)^2 \frac{14}{2(N-2)+1} \geqslant \frac{c}{N^2} \sum_{i=N-3}^{N-2} \sum_{j=2}^{N-4} (x_i^j)^2. \end{aligned} \tag{3.3.32}$$

Thus, the bound is true for S_6. On observing the term S_7 and S_8, defined by (3.3.19) and (3.3.20) respectively, it is deduced that they are respecticely similar to the terms S_3 and S_6, defined by (3.3.15) and (3.3.18) respectively, and hence

$$S_7 \geqslant \frac{c}{N^2} \sum_{i=0}^{1} \sum_{j=N-3}^{N-2} (x_i^j)^2, \tag{3.3.33}$$

$$S_8 \geqslant \frac{c}{N^2} \sum_{i=2}^{N-4} \sum_{j=N-3}^{N-2} (x_i^j)^2. \tag{3.3.34}$$

Since for any $i,j \in \mathbb{N}$,

$$\frac{4i+6}{2j+1} \geqslant \frac{4i}{2j} \geqslant \frac{i}{j},$$

and using it in the term S_9, defined by (3.3.21), yields

$$\begin{aligned} S_9 &\geqslant \sum_{i=N-3}^{N-2} \sum_{j=N-3}^{N-2} (x_i^j)^2 \left[\frac{i}{j} + \frac{j}{i}\right], \\ &= \sum_{i=N-3}^{N-2} (x_i^{N-3})^2 \left[\frac{i}{N-3} + \frac{N-3}{i}\right] + \sum_{i=N-3}^{N-2} (x_i^{N-2})^2 \left[\frac{i}{N-2} + \frac{N-2}{i}\right], \\ &= 2(x_{N-3}^{N-3})^2 + \left[\frac{N-2}{N-3} + \frac{N-3}{N-2}\right] (x_{N-2}^{N-3})^2 + \left[\frac{N-3}{N-2} + \frac{N-2}{N-3}\right] (x_{N-3}^{N-2})^2 + 2(x_{N-2}^{N-2})^2. \end{aligned}$$

Note that

$$\frac{N-2}{N-3}+\frac{N-3}{N-2}=\frac{(N-2)^2+(N-3)^2}{(N-3)(N-2)}\geqslant\frac{c}{N^2},$$

which implies

$$S_9\geqslant\frac{c}{N^2}\sum_{i=N-3}^{N-2}\sum_{j=N-3}^{N-2}(x_i^j)^2. \tag{3.3.35}$$

Thus, the results (3.3.22)-(3.3.25) and (3.3.31)-(3.3.35), in (3.3.12),

$$x^TAx\geqslant\frac{2c}{N^2}\sum_{i=0}^{N-2}\sum_{j=0}^{N-2}(x_i^j)^2=\frac{c}{N^2}x^Tx,$$

hence the Rayleigh quotient of the matrix A is bounded below by c/N^2, for all $x\in\mathbb{R}^{\vartheta\times1}$, which leads us to the desired result. □

Since $\Lambda(A\oplus A)=\Lambda(A)$, the above theorem gives the following estimate.

Corollary 3.3.4. *For $N\geqslant4$, the discrete vector Laplacian $\mathcal{A}$ defined by* (3.3.7) *satisfies $\frac{c}{N^2}\leqslant\lambda(\mathcal{A})\leqslant cN$, thus $\kappa(\mathcal{A})\leqslant cN^3$.*

Remark 3.3.5. Note that Theorem 2.1.6 can also be applied for estimating the $\lambda_{\min}(A)$, however, that does not provide an optimal lower bound. As by reiterating the process used for estimating $\lambda_{\max}(A)$,

$$\lambda_{\min}(A)\geqslant\lambda_{\min}(M\otimes S)+\lambda_{\min}(S\otimes M)=2\lambda_{\min}(S)\lambda_{\min}(M)\geqslant12\frac{c}{N^3},$$

whereas Figure 3.2b suggests the best lower bound of $\frac{c}{N^2}$, thus a detailed analysis was conducted.

We now proceed to analyze the matrix B, a sub-block of G, which is a rectangular matrix defined in terms of matrices R and Q given by Proposition 3.2.4. In order to prove estimates for G, the bounds on singular values of the matrix B are required, which are derived by the following results.

Lemma 3.3.6. *Let $N \geqslant 4$, $R, Q \in \mathbb{R}^{(N-1)\times(N-1)}$ be defined by Proposition 3.2.4, then $\sigma_{\max}(R) = 2$, $\sigma_{\min}(R) = 0$, $\sigma_{\max}(Q) \leqslant C$ and $\sigma_{\min}(Q) \geqslant \frac{c}{N^2}$.*

Proof. The definition of R implies that $R^T R$ is a diagonal matrix with entries $(R^T R)_{00} = 0$, and $(R^T R)_{ii} = 4$, for $1 \leqslant i \leqslant N-2$. Therefore, $\sigma_{\min}(R) = 0$ and $\sigma_{\max}(R) = \sqrt{\lambda_{\max}(R^T R)} = \sqrt{4} = 2$.

By definition,

$$Q_{ij} = \begin{cases} \gamma_i, & j = i, \\ -\gamma_{i+2}, & j = i+2, \end{cases},$$

where $\gamma_i = \frac{2}{2i+1}$ for $0 \leqslant i, j \leqslant N-2$. Since the $1-$norm of a matrix is its maximum absolute column sum,

$$\begin{aligned} \|Q\|_1 &= \max\{\gamma_0, \gamma_1, 2\gamma_2, 2\gamma_3, \ldots, 2\gamma_{N-2}\} \\ &= \gamma_0 = 2. \end{aligned}$$

Also, the maximum absolute row sum, $\|Q\|_\infty = \max\{\gamma_0 + \gamma_2, \gamma_1 + \gamma_3, \ldots, \gamma_{N-4} + \gamma_{N-2}, \gamma_{N-3}, \gamma_{N-2}\} = \gamma_0 + \gamma_2 = 2 + \frac{2}{5} = \frac{12}{5}$, hence

$$\sigma_{\max}(Q) = \|Q\|_2 \leqslant \sqrt{\|Q\|_1 \|Q\|_\infty} = \sqrt{2 \cdot \frac{12}{5}}.$$

We now estimate $\sigma_{\min}(Q) = \|Q^{-1}\|_2^{-1}$. It is easily verified that Q^{-1} is upper

triangular and is non-zero along every other diagonal:

$$Q^{-1} = \begin{bmatrix} \gamma_0^{-1} & 0 & \gamma_0^{-1} & 0 & \gamma_0^{-1} & 0 & \dots \\ & \gamma_1^{-1} & 0 & \gamma_1^{-1} & 0 & \gamma_1^{-1} & \dots \\ & & \gamma_2^{-1} & 0 & \gamma_2^{-1} & 0 & \dots \\ & & & \ddots & \ddots & \ddots & \\ & & & & \gamma_{N-4}^{-1} & 0 & \gamma_{N-4}^{-1} \\ & & & & & \gamma_{N-3}^{-1} & 0 \\ & & & & & & \gamma_{N-2}^{-1} \end{bmatrix} \in \mathbb{R}^{N-1\times N-1}.$$

Label the columns of Q^{-1} as $C_0, C_1, \dots, C_{N-2}$. Note that the maximum absolute column sum of Q^{-1} is attained at either C_{N-3} or C_{N-2}, denoted by $S_{C_{N-3}}$ or $S_{C_{N-2}}$, respectively, and are given as follows,

$$S_{C_{N-3}} = \begin{cases} \sum_{i=0}^{\lfloor \frac{N-2}{2} \rfloor} \gamma_{2i}^{-1}, & N \text{ is odd}, \\ \sum_{i=0}^{\lfloor \frac{N-3}{2} \rfloor} \gamma_{2i+1}^{-1}, & N \text{ is even}, \end{cases} \qquad S_{C_{N-2}} = \begin{cases} \sum_{i=0}^{\lfloor \frac{N-3}{2} \rfloor} \gamma_{2i+1}^{-1}, & N \text{ is odd}, \\ \sum_{i=0}^{\lfloor \frac{N-2}{2} \rfloor} \gamma_{2i}^{-1}, & N \text{ is even}. \end{cases}$$

Since

$$\sum_{i=0}^{\lfloor \frac{N-2}{2} \rfloor} \frac{1}{\gamma_{2i}} = \sum_{i=0}^{\lfloor \frac{N-2}{2} \rfloor} \frac{2(2i)+1}{2} = \frac{1}{2} \sum_{i=0}^{\lfloor \frac{N-2}{2} \rfloor} (4i+1) \leqslant cN^2,$$

and similarly, $\sum_{i=0}^{\lfloor \frac{N-3}{2} \rfloor} \frac{1}{\gamma_{2i+1}} \leqslant cN^2$, it follows that $\|Q^{-1}\|_1 \leqslant cN^2$.

For $0 \leqslant i \leqslant N-2$, the absolute sum of the ith row of Q^{-1} is $\frac{1}{\gamma_i} \left\lfloor \frac{N-i}{2} \right\rfloor$, thus

$$\begin{aligned} \|Q^{-1}\|_\infty &= \max_{0\leqslant i\leqslant N-2} \frac{1}{\gamma_i} \left\lfloor \frac{N-i}{2} \right\rfloor \leqslant \max_{0\leqslant i\leqslant N-2} \frac{1}{\gamma_i} \max_{0\leqslant i\leqslant N-2} \left\lfloor \frac{N-i}{2} \right\rfloor \\ &= \frac{1}{\gamma_{N-2}} \left\lfloor \frac{N}{2} \right\rfloor = \frac{2(N-2)+1}{2} \left\lfloor \frac{N}{2} \right\rfloor \leqslant cN^2. \end{aligned}$$

Therefore, $\|Q^{-1}\|_2 \leqslant \sqrt{\|Q^{-1}\|_1 \|Q^{-1}\|_\infty} \leqslant \sqrt{cN^2 \cdot cN^2} = cN^2$, hence the result. □

Lemma 3.3.7. *For $N \geqslant 4$, the matrix $B \in \mathbb{R}^{2\vartheta\times\wp}$ defined by* (3.3.6) *is full rank,*

that is, rank$(B) = \wp$*, and* $\dfrac{c}{N^2} \leqslant \sigma(B) \leqslant C$.

Proof. Let R_i be the rows of B for $1 \leqslant i \leqslant 2\vartheta$. On exchanging $R_{k(N-1)}$ with $R_{\vartheta+1+(k-1)(N-1)}$, for all $1 \leqslant k \leqslant N-2$, the first $\wp$ rows of B form an upper triangular matrix of size $\wp \times \wp$ with non-zero diagonal entries, hence $\operatorname{rank}(B) = \wp$.

We now estimate the singular values of B, which are the square-root of the eigenvalues of $B^TB \in \mathbb{R}^{\wp\times\wp}$. Note that $\operatorname{rank}(B^TB) = \operatorname{rank}(B) = \wp$, and $B^TB = B_1^TB_1 + B_2^TB_2$. So we consider the blocks B_1 and B_2.

Since $B_1 = -Q\otimes R]$ and $B_2 = -R\otimes Q]$, that is, their first column is deleted, which only contains zero, therefore $\operatorname{rank}(B_1) = \operatorname{rank}(B_2) = \operatorname{rank}(Q)\operatorname{rank}(R) = (N-1)(N-2) < \wp$. Thus, B_i are rank deficient, so that $\sigma_{\min}(B_i) = 0$ for $i = 1, 2$. Furthermore, $\sigma(B_i) = \sigma(Q) \times \sigma(R)$, so Lemma 3.3.6 implies that $\sigma_{\max}(B_i) = \sigma_{\max}(R)\sigma_{\max}(Q) \leqslant 2C$, for $i = 1, 2$. Therefore, by Theorem 2.1.6

$$\lambda_{\max}(B^TB) \leqslant \lambda_{\max}(B_1^TB_1) + \lambda_{\max}(B_2^TB_2) = \sigma^2_{\max}(B_1) + \sigma^2_{\max}(B_2) \leqslant 4C^2 + 4C^2,$$

thus, $\sigma_{\max}(B) \leqslant C$. However, Theorem 2.1.6 gives a trivial bound for the minimum singular value of B, but we need a positive value as B^TB is full rank. To this end, we do a detailed analysis as follows.

Let $\alpha_{ij} := (Q^TQ)_{ij}$ and $\beta_{ij} := (R^TR)_{ij}$, where $0 \leqslant i, j \leqslant N-2$. Note that Q^TQ is a symmetric matrix, so that for $0 \leqslant i, j \leqslant N-2$

$$\alpha_{ij} = (Q^TQ)_{ij} = \begin{cases} \gamma_i^2, & i = j = 0, 1, \\ 2\gamma_i^2, & 2 \leqslant i = j \leqslant N-2, \\ -\gamma_i\gamma_{i+2}, & j = i+2. \end{cases}$$

Since Lemma 3.3.6 implies that $\sigma_{\min}(Q) \geqslant \dfrac{c}{N^2}$, for $y \in \mathbb{R}^{N-1}\backslash\{0\}$

$$\frac{c}{N^4} y^T y \leqslant y^T Q^T Q y = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} y_i \alpha_{ij} y_j = \sum_{i=0}^{N-2} \alpha_{ii}(y_i)^2 + 2\sum_{i=0}^{N-4} \alpha_{ii+2} y_i y_{i+2}. \quad (3.3.36)$$

Recall that $R^T R$ is a diagonal matrix, with $\beta_{00} = 0$, and $\beta_{ii} = 4$ for $1 \leqslant i \leqslant N-2$. By a direct calculation, $B^T B = \mathbf{[}Q^T Q \otimes R^T R + R^T R \otimes Q^T Q\mathbf{]}$, that is, delete the first row and first column, then the (i,j)-th block of $B^T B$ is given as

$$(B^T B)_{ij} = \begin{cases} \alpha_{00}\mathbf{[}R^T R\mathbf{]}, & i = j = 0, \\ \alpha_{ii} R^T R + 4 Q^T Q, & 1 \leqslant i = j \leqslant N-2, \\ \alpha_{02}[R^T R, & i = 0, j = 2, \\ \alpha_{ii+2} R^T R, & 1 \leqslant i \leqslant N-4, j = i+2. \end{cases}$$

Let $x = [x_0; x_1; \ldots; x_{N-2}] \in \mathbb{R}^{\wp}\backslash\{0\}$, where $x_0 = [x_0^1; x_0^2; \ldots; x_0^{N-2}] \in \mathbb{R}^{N-2}$, and $x_i = [x_i^0; x_i^1; \ldots; x_i^{N-2}] \in \mathbb{R}^{N-1}$ for $1 \leqslant i \leqslant N-2$, then

$$\begin{aligned} x^T B^T B x &= \sum_{i=0}^{N-2} x_i^T (B^T B)_{ii} x_i + 2\sum_{i=0}^{N-4} x_i^T (B^T B)_{i,i+2} x_{i+2} \\ &= x_0^T \alpha_{00}\mathbf{[}R^T R\mathbf{]} x_0 + \sum_{i=1}^{N-2} x_i^T (\alpha_{ii} R^T R + \beta_{ii} Q^T Q) x_i + 2 x_0^T \alpha_{02}[R^T R \cdot x_2 \\ &\qquad + 2\sum_{i=1}^{N-4} x_i^T \alpha_{i,i+2} R^T R x_{i+2} \\ &= \alpha_{00} \sum_{j=1}^{N-2} 4(x_0^j)^2 + \sum_{i=1}^{N-2} \alpha_{ii} \sum_{j=1}^{N-2} 4(x_i^j)^2 + \sum_{i=1}^{N-2} \beta_{ii} x_i^T Q^T Q x_i \\ &\qquad + 2\alpha_{02} \sum_{\ell=1}^{N-2}\sum_{j=0}^{N-2} x_0^{\ell}([R^T R)_{\ell j} x_2^j + 2\sum_{i=1}^{N-4} \alpha_{i,i+2} \sum_{j=1}^{N-2} 4 x_i^j x_{i+2}^j \\ &= \alpha_{00} \sum_{j=1}^{N-2} 4(x_0^j)^2 + \sum_{i=1}^{N-2} \alpha_{ii} \sum_{j=1}^{N-2} 4(x_i^j)^2 + \sum_{i=1}^{N-2} 4 x_i^T Q^T Q x_i \\ &\qquad + 2\alpha_{02} \sum_{j=1}^{N-2} 4 x_0^j x_2^j + 2\sum_{i=1}^{N-4} \alpha_{i,i+2} \sum_{j=1}^{N-2} 4 x_i^j x_{i+2}^j \end{aligned}$$

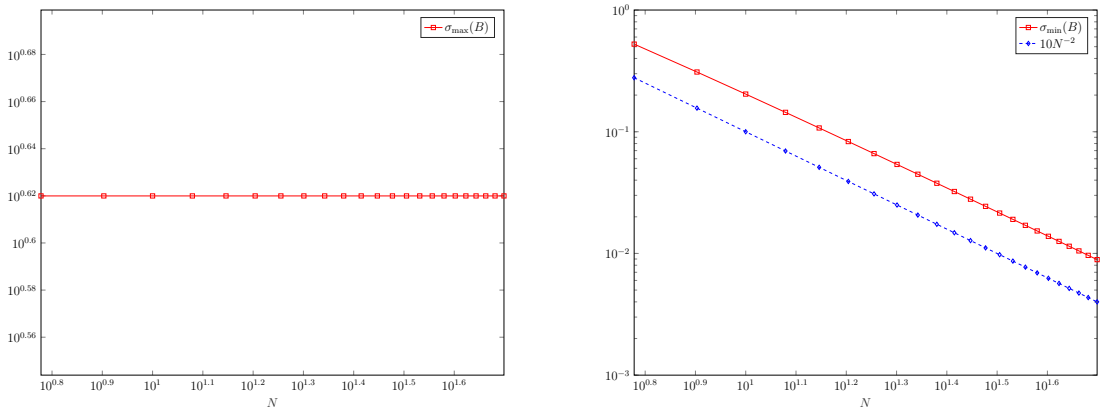
$$= 4\sum_{j=1}^{N-2}\left(\sum_{i=0}^{N-2}\alpha_{ii}(x_i^j)^2 + 2\sum_{i=0}^{N-4}\alpha_{i,i+2}x_i^j x_{i+2}^j\right) + \sum_{i=1}^{N-2} 4x_i^T Q^T Q x_i.$$

Define $\xi_j = [x_0^j; x_1^j; x_2^j; \ldots; x_{N-2}^j]$, then

$$\begin{aligned}
x^T B^T B x &= 4\sum_{j=1}^{N-2}\xi_j^T Q^T Q \xi_j + \sum_{i=1}^{N-2} 4x_i^T Q^T Q x_i \\
&\geqslant 4\frac{c}{N^4}\sum_{j=1}^{N-2}\xi_j^T\xi_j + 4\frac{c}{N^4}\sum_{i=1}^{N-2} x_i^T x_i \qquad \text{(by (3.3.36))} \\
&= 4\frac{c}{N^4}\sum_{j=1}^{N-2}\sum_{i=0}^{N-2}(x_i^j)^2 + 4\frac{c}{N^4}\sum_{i=1}^{N-2}\sum_{j=0}^{N-2}(x_i^j)^2 \\
&= 4\frac{c}{N^4}\sum_{j=1}^{N-2}(x_0^j)^2 + 4\frac{c}{N^4}\sum_{i=1}^{N-2}\left((x_i^0)^2 + \sum_{j=1}^{N-2}2(x_i^j)^2\right) \\
&\geqslant 4\frac{c}{N^4}\sum_{j=1}^{N-2}(x_0^j)^2 + 4\frac{c}{N^4}\sum_{i=1}^{N-2}\sum_{j=0}^{N-2}(x_i^j)^2 = \frac{4c}{N^4}x^T x.
\end{aligned}$$

Thus, $\lambda_{\min}(B^T B) \geqslant \dfrac{c}{N^4}$, which implies that $\sigma_{\min}(B) \geqslant \dfrac{c}{N^2}$. □

The above estimates are observed in Figures 3.3a and 3.3b. In order to get



(a) Maximum singular value. (b) Minimum singular value.

Figure 3.3: Singular values of B.

estimates on the singular values of the global spectral operator G, we need the following result that gives the spectrum of a symmetric saddle point matrix.

Theorem 3.3.8 (See [3]). *Let* $\mathcal{X} = \begin{pmatrix} A & B^T \\ B & O \end{pmatrix}$, *such that* $B \in \mathbb{R}^{m \times n}$ *is full rank and its Schur complement* $BA^{-1}B^T$ *is SPD, then*

$$\Lambda(\mathcal{X}) \subseteq \left[\frac{-\lambda_1}{\frac{1}{2}\left(1 + \sqrt{1 + 4\frac{\lambda_1}{\mu_1}}\right)}, \frac{-\lambda_m}{\frac{1}{2}\left(1 + \sqrt{1 + 4\frac{\lambda_m}{\mu_n}}\right)} \right] \cup \left[\mu_n, \mu_1 \frac{1 + \sqrt{1 + 4\frac{\lambda_1}{\mu_1}}}{2} \right], \tag{3.3.37}$$

where $0 < \mu_n \leqslant \ldots \leqslant \mu_1$ *denote the eigenvalues of* A *and* $0 < \lambda_m \leqslant \ldots \leqslant \lambda_1$ *are the eigenvalues of* $BA^{-1}B^T$.

In order to use the above result, we need to estimate the spectrum of the *Schur complement* for the discrete Stokes problem in steady state, with coefficient matrix (3.3.6), defined as $\Upsilon_h = B^T \mathcal{A}^{-1} B$. Let us first introduce the Schur complement for the continuous Stokes problem, which is known as the *Uzawa pressure operator*, denoted by $\Upsilon : L^2_0(\Omega) \to L^2_0(\Omega)$, is defined as $\Upsilon := \nabla \cdot \Delta^{-1} \nabla$. It is a self-adjoint, bounded, coercive and hence a bijective operator with $\lambda_{\max}(\Upsilon) = 1$. Also, $\Delta^{-1} : (H^{-1}(\Omega))^2 \to V$ denotes the *inverse Laplacian*. Let $u \in (H^{-1}(\Omega))^2$, we say $\Delta^{-1} u = v \in V$ if

$$\Delta v = u \text{ in } \Omega,$$
$$v = 0 \text{ on } \partial\Omega.$$

Note that Δ is the *vector Laplacian* as $v \in V$ is a vector having two components. From [12, p. 422], the following inf-sup condition holds for the $P_N - P_{N-2}$ scheme

$$\inf_{q_N \in \mathbb{P}_{N-2, N-2} \cap L^2_0(\Omega)} \sup_{v_n \in \mathbb{P}^0_{N,N}} \frac{(\nabla \cdot v_N, q_N)}{\|v_N\|_1 \|q_N\|_0} \geqslant \frac{c}{\sqrt{N}},$$

which, as stated in [28, p. 173], is equivalent to

$$\inf_{q \in \mathbb{R}^{\wp} \setminus 0} \sqrt{\frac{q^T B^T \mathcal{A}^{-1} B q}{q^T \mathfrak{M} q}} \geqslant \frac{c}{\sqrt{N}},$$

or

$$\inf_{q\in\mathbb{R}^{\wp}\setminus 0} \frac{q^T\Upsilon_h q}{q^T\mathfrak{M}q} \geqslant \frac{c}{N}. \tag{3.3.38}$$

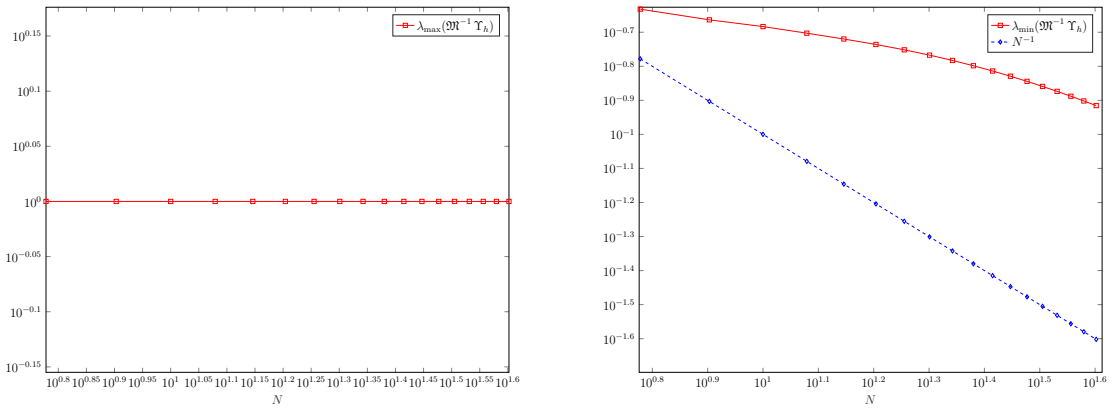
Here q is the vector of coefficients of $q_N = \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} q_{ij}L_i(x)L_j(y)$, and $\mathfrak{M}$ is the mass matrix, so that

$$q^T\mathfrak{M}q = \|q_N\|_0^2 = \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2}\sum_{m=0}^{N-2}\sum_{\substack{n=0\\ m+n>0}}^{N-2} q_{ij}q_{mn}\frac{2\delta_{ij}}{2i+1}\frac{2\delta_{mn}}{2m+1} = q^T\left[\tau\otimes\tau\right]q,$$

where $(\tau)_{ij} = \frac{2\delta_{ij}}{2i+1}$ for $0 \leqslant i,j \leqslant N-2$. Since τ is a diagonal matrix, then $\mathfrak{M} = [\tau\otimes\tau]$ is diagonal and $m+n, i+j > 0$,

$$\lambda_{\min}(\mathfrak{M}) = (L_{N-2}, L_{N-2})^2 = \left(\frac{2}{2(N-2)+1}\right)^2 \geqslant \frac{c}{N^2} \tag{3.3.39}$$

Figure 3.4b verifies (3.3.38), whereas the following result is depicted by Figure 3.4a.



(a) Maximum eigenvalue.

(b) Minimum eigenvalue.

Figure 3.4: Spectrum of $\mathfrak{M}^{-1}\Upsilon_h$.

Theorem 3.3.9. *For given* $N \geqslant 4$, $\lambda_{\max}(\mathfrak{M}^{-1}\Upsilon_h) \leqslant 1$.

Proof. For some $p = \sum_{i=0}^{N-2} \sum_{\substack{j=0 \\ i+j>0}}^{N-2} p_{ij} L_i(x) L_j(y) \in \mathbb{P}_{N-2,N-2} \cap L_0^2(\Omega)$,

$$(\Upsilon p, p) = (\nabla \cdot \Delta^{-1} \nabla (\sum_{i=0}^{N-2} \sum_{\substack{j=0 \\ i+j>0}}^{N-2} p_{ij} L_i(x) L_j(y)), p).$$

For $0 < i + j$ and $0 \leqslant i, j \leqslant N-2$, define $w^{ij} = \Delta^{-1} \nabla (L_i(x) L_j(y))$, and $w = \sum_{i=0}^{N-2} \sum_{\substack{j=0 \\ i+j>0}}^{N-2} p_{ij} w^{ij}$ so that the above problem becomes

$$(\Upsilon p, p) = (\nabla \cdot w, p) = (w_{1_x}, p) + (w_{2_y}, p), \tag{3.3.40}$$

and let $w^{ij} = \left[w_1^{ij}; w_2^{ij}\right]$ and $w = [w_1, w_2]$. By definition, w^{ij} is the solution of the following problem

$$\begin{aligned} \Delta w^{ij} &= \nabla \left(L_i(x) L_j(y)\right) \text{ in } \Omega, \\ w^{ij} &= 0 \text{ on } \partial\Omega, \end{aligned}$$

which is equivalent to the following two problems

$$\begin{aligned} \Delta w_1^{ij} &= L_i'(x) L_j(y) \text{ in } \Omega, \\ w_1^{ij} &= 0 \text{ on } \partial\Omega, \end{aligned} \qquad \text{and} \qquad \begin{aligned} \Delta w_2^{ij} &= L_i(x) L_j'(y) \text{ in } \Omega, \\ w_2^{ij} &= 0 \text{ on } \partial\Omega. \end{aligned} \tag{3.3.41}$$

Let us solve the first one by using the recombined Legendre basis functions, by considering $w_1^{ij} = \sum_{m=0}^{N-2} \sum_{n=0}^{N-2} w_{mn}^{ij} \phi_m(x) \phi_n(y)$ and let $w_{1,h}^{ij}$ be the vector of coefficients of w_1^{ij}. For $0 \leqslant r, s \leqslant N-2$,

$$\begin{aligned} (\Delta w_1^{ij}, \phi_r(x)\phi_s(y)) &= (L_i'(x) L_j(y), \phi_r(x)\phi_s(y)) \\ -A w_{1,h}^{ij} &= -R_{ri} Q_{sj}, \end{aligned}$$

hence $Aw_{1,h}^{ij} = R_{ri}Q_{sj}$ and similarly, $Aw_{2,h}^{ij} = Q_{ri}R_{sj}$ for $0 \leqslant r, s \leqslant N-2$, thus

$$Aw_1 = A\sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}w_{1,h}^{ij} = \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2} p_{ij}R_{ri}Q_{sj} = (Q\otimes R)]p_h = -B_1p_h,$$

similarly, $Aw_2 = -B_2p_h$, hence $w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -A^{-1}B_1p_h \\ -A^{-1}B_2p_h \end{bmatrix}$, which leads us to evaluating the final step (3.3.40).

Since $w_1 = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} (w_1)_{ij}\,\phi_i(x)\phi_j(y)$,

$$\begin{aligned}\big(w_{1_x}, p\big) &= \sum_{m=0}^{N-2}\sum_{\substack{n=0\\ m+n>0}}^{N-2} p_{mn}\sum_{i=0}^{N-2}\sum_{j=0}^{N-2}(w_1)_{ij}\,(\phi_i'(x)\phi_j(y), L_m(x)L_n(y)) \\ &= p_h^T\big[(Q^T\otimes R^T)\cdot w_1 = p_h^T(-B_1^T)(-A^{-1}B_1)p_h = p_h^TB_1^TA^{-1}B_1p_h.\end{aligned}$$

Similarly, $\big(w_{2_y}, p\big) = p_h^TB_2^TA^{-1}B_2p_h$, thus (3.3.40) gives $\big(\Upsilon p, p\big) = p_h^TB_1^TA^{-1}B_1p_h + p_h^TB_2^TA^{-1}B_2p_h = p_h^TB^T\,\mathcal{A}^{-1}\,Bp_h = p_h^T\Upsilon_hp_h$, and $\big(p, p\big) = p_h^T\mathfrak{M}p_h$, therefore

$$\begin{aligned}\lambda_{\max}(\mathfrak{M}^{-1}\Upsilon) &= \sup_{p_h\in\mathbb{R}^{\wp}\setminus 0}\frac{p_h^T\Upsilon_hp_h}{p_h^T\mathfrak{M}p_h} = \sup_{p\in\mathbb{P}_{N-2,N-2}\cap L_0^2(\Omega)}\frac{\big(\Upsilon p, p\big)}{\big(p, p\big)} \\ &\leqslant \sup_{p\in L_0^2(\Omega)}\frac{\big(\Upsilon p, p\big)}{\big(p, p\big)} = \lambda_{\max}(\Upsilon) = 1.\end{aligned}$$

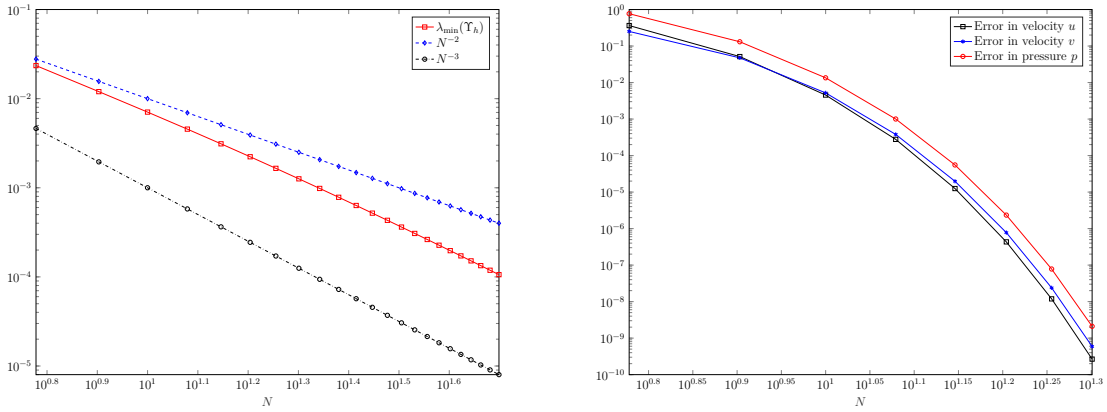
□

Figure 3.5a presents the numerical results for the following result.

Lemma 3.3.10. *For given $N \geqslant 4$, $\lambda_{\min}(\Upsilon_h) \geqslant \dfrac{c}{N^3}$ and $\lambda_{\min}(\Upsilon_h) \leqslant c\lambda_{\min}(\mathcal{A})$.*

Proof. Since Υ_h is symmetric,

$$\begin{aligned}\lambda_{\min}(\Upsilon_h) &= \inf_{p\in\mathbb{R}^{\wp}\setminus 0}\frac{p^T\Upsilon_hp}{p^T\mathfrak{M}p}\cdot\frac{p^T\mathfrak{M}p}{p^Tp} \\ &\geqslant \inf_{p\in\mathbb{R}^{\wp}\setminus 0}\frac{p^T\Upsilon_hp}{p^T\mathfrak{M}p}\cdot\inf_{p\in\mathbb{R}^{\wp}\setminus 0}\frac{p^T\mathfrak{M}p}{p^Tp}\end{aligned} \tag{3.3.42}$$



(a) Minimum eigenvalue of Υ_h. (b) Convergence.

Figure 3.5: Numerical results for global operators of the steady Stokes problem.

$$\geqslant \frac{c}{N} \cdot \frac{c}{N^2} = \frac{c}{N^3}. \qquad \text{(by (3.3.39))}$$

Moreover, by (3.3.42)

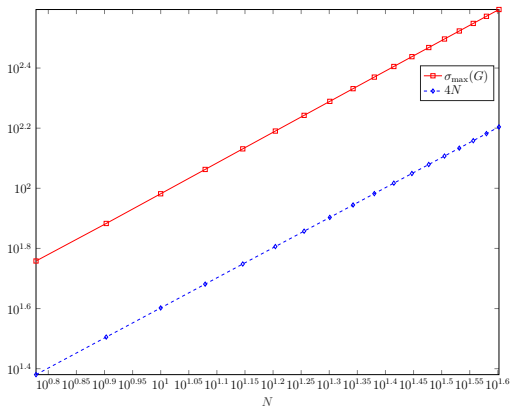
$$\begin{aligned}
\lambda_{\min}(\Upsilon_h) &\leqslant \sup_{p \in \mathbb{R}^{\wp} \setminus 0} \frac{p^T \Upsilon_h p}{p^T \mathfrak{M} p} \cdot \inf_{p \in \mathbb{R}^{\wp} \setminus 0} \frac{p^T \mathfrak{M} p}{p^T p} \\
&\leqslant 1 \cdot \left(\frac{2}{2(N-2)+1} \right)^2 = \frac{c}{N^2} \qquad \text{(by Theorem 3.3.9 and eq. (3.3.39))} \\
&\leqslant c \lambda_{\min}(\mathcal{A}).
\end{aligned}$$

□

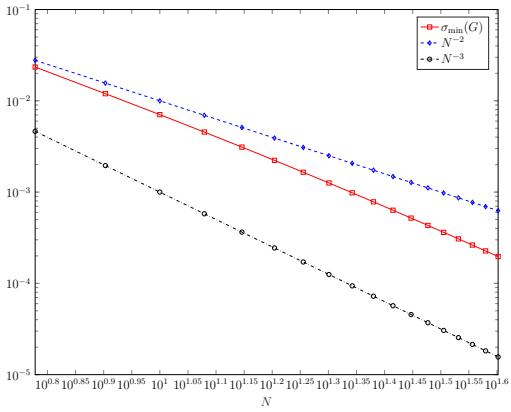
The above results aid us to prove our main goal of this section, that is, an optimal bound for the *condition number* of the global spectral operator G for the steady Stokes problem, as seen in Figures 3.6a and 3.6b.

Theorem 3.3.11. *For $N \geqslant 4$, let G be defined by (3.3.6), then $\frac{c}{N^2} \leqslant \sigma(G) \leqslant cN^2$, and $\kappa(G) \leqslant cN^4$.*

Proof. Note that $G = \begin{bmatrix} \mathcal{A} & O \\ O & O_{\wp,\wp} \end{bmatrix} + \begin{bmatrix} O & B \\ B^T & O_{\wp,\wp} \end{bmatrix} =: G_1 + G_2$, thus it is a sum of two



(a) Maximum eigenvalue of G.



(b) Minimum eigenvalue of G.

Figure 3.6: Spectrum of G.

symmetric matrices. Hence, by Theorem 2.1.6

$$\lambda_{\max}(G) \leqslant \lambda_{\max}(G_1) + \lambda_{\max}(G_2), \tag{3.3.43}$$

Since $\lambda_{\max}(G_1) = \lambda_{\max}(\mathcal{A})$, and $\lambda_{\max}(G_2) = \sigma_{\max}(B)$ by Result 8 of Theorem 2.1.8. Hence, by applying these results to (3.3.43), $\lambda_{\max}(G) \leqslant \lambda_{\max}(\mathcal{A}) + \lambda_{\max}(B) \leqslant cN + c \leqslant cN$, by Corollary 3.3.4 and Lemma 3.3.7.

Now, it remains to estimate the absolute minimum value of the eigenvalues of G, denoted by $|\lambda|_{\min}(G)$, for which Theorem 3.3.8 gives,

$$|\lambda|_{\min}(G) \geqslant \min\left\{\lambda_{\min}(\mathcal{A}), \frac{\lambda_{\min}(\Upsilon_h)}{\frac{1}{2}\left(1+\sqrt{1+\frac{4\lambda_{\min}(\Upsilon_h)}{\lambda_{\min}(\mathcal{A})}}\right)}\right\},$$

and by Lemma 3.3.10, $\frac{\lambda_{\min}(\Upsilon_h)}{\lambda_{\min}(\mathcal{A})} \leqslant c$, leading to $\frac{1}{2}\left(1+\sqrt{1+\frac{4\lambda_{\min}(\Upsilon_h)}{\lambda_{\min}(\mathcal{A})}}\right) \leqslant c$, thus

$$\frac{\lambda_{\min}(\Upsilon_h)}{\frac{1}{2}\left(1+\sqrt{1+\frac{4\lambda_{\min}(\Upsilon_h)}{\lambda_{\min}(\mathcal{A})}}\right)} \geqslant c\lambda_{\min}(\Upsilon_h).$$

Hence, the minimum absolute value of eigenvalues of G satisfies,

$$|\lambda|_{\min}(G) \geqslant \min\{\lambda_{\min}(\mathcal{A}), c\lambda_{\min}(\Upsilon_h)\} \geqslant \min\left\{\frac{c}{N^2}, \frac{c}{N^3}\right\} = \frac{c}{N^3}.$$

Since $\kappa(G) = \dfrac{|\lambda|_{\max}(G)}{|\lambda|_{\min}(G)}$, therefore $\kappa(G) \leqslant cN \cdot N^3 = cN^4$. □

To summarize, for the Stokes problem in the steady state given by (3.1.1), we implemented the proposed $P_N - P_{N-2}$ scheme in space by using a recombined Legendre basis functions on MATLAB®, see [52]. Take f_1, f_2 so that the exact solutions are $u(x,y) = (\cos(\pi x) + 1)\sin(2\pi y)$, $v(x,y) = (0.5)\sin(\pi x)(1 - \cos(2\pi y))$, and $p(x,y) = \sin(\pi x)\cos(\pi y)$, thus the boundary conditions are satisfied. It can easily be implemented for $P_N - P_{N-2}$ scheme with a recombined Chebyshev basis in space as mentioned in [80]. The spectral convergence of the $P_N - P_{N-2}$ scheme was proved in [12], and is evident from Figure 3.5b.

3.4 Unsteady state

Consider the *unsteady Stokes problem*, given by equation (3.1.2), which on further simplification is expressed as:

$$u_t - \Delta u + p_x = f_1 \text{ in } \Omega_t, \tag{3.4.1a}$$

$$v_t - \Delta v + p_y = f_2 \text{ in } \Omega_t, \tag{3.4.1b}$$

$$u_x + v_y = 0 \text{ in } \Omega \times (-1,1), \tag{3.4.1c}$$

$$u(x,y,-1) = u_0(x,y),\ v(x,y,-1) = v_0(x,y) \text{ in } \Omega, \tag{3.4.1d}$$

$$u = 0,\ v = 0 \text{ on } \partial\Omega \times (-1,1).$$

We extend the $P_N - P_{N-2}$ scheme of the last section to the unsteady case by applying Chebyshev Gauss-Lobatto spectral collocation in time. These particular polynomial

bases are chosen for simplicity of analysis of this scheme. In practice, Chebyshev recombined basis given in [80, p. 149] or Jacobi collocation can be chosen in place of Legendre recombined basis or Chebyshev collocation, respectively, without any difficulties. The goal is to show spectral convergence of a space-time spectral method and a condition number estimate of the scheme. The analysis of the latter is incomplete because two of the estimates are based on numerical evidence.

3.4.1 Discretization

For given $N \geqslant 4$, consider the Chebyshev Gauss-Lobatto nodes t_k for $0 \leqslant k \leqslant N$, so that $t_0 = -1$ and $t_N = 1$. Let ℓ_k denote the Lagrange basis polynomials for t_k, therefore $\ell_k(t_j) = \delta_{kj}$ for $0 \leqslant k, j \leqslant N$. Let D denote the Chebyshev Gauss-Lobatto pseudospectral derivative matrix of size $(N+1)\times(N+1)$. For this scheme, we define an *approximation* for the velocity u, v and the pressure p as follows,

$$\begin{aligned}
u_N(x,y,t) &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{k=0}^{N} u_{ijk}\phi_i(x)\phi_j(y)\ell_k(t) \in \mathbb{P}^0_{N,N,N},\\
v_N(x,y,t) &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{k=0}^{N} v_{ijk}\phi_i(x)\phi_j(y)\ell_k(t) \in \mathbb{P}^0_{N,N,N},\\
p_N(x,y,t) &= \sum_{i=0}^{N-2}\sum_{\substack{j=0\\ i+j>0}}^{N-2}\sum_{k=0}^{N} p_{ijk}L_i(x)L_j(y)\ell_k(t) \in \mathbb{P}_{N-2,N-2,N}.
\end{aligned} \tag{3.4.2}$$

The number of unknowns for u_N and v_N each are $N\vartheta$, and the number of unknowns for p_N are $N\wp$. The total number of unknowns in the discrete Stokes equations is $2N\vartheta + N\wp = 3N(N-1)^2 - N$. Define the *discrete unknowns* as $u_h = [u_h^1; u_h^2; \ldots; u_h^N]$, where $u_h^\ell = [u_{0,0,\ell}; u_{1,0,\ell}; \ldots u_{N-2,0,\ell}; u_{0,1,\ell}; \ldots u_{N-2,N-2,\ell}]$, similarly define v_h, p_h. Let $k = 1, 2$ and $t = t_r$, for a given $f_k(x,y,t)$ so that $f_k(x,y,t_r)$ is analytic in Ω for all $1 \leqslant r \leqslant N$, it can be approximated by a truncated Legendre series expansion as

follows,

$$f_k(x,y,t_r) \approx \tilde{f}_k^r(x,y) := \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^{k,r} L_i(x)L_j(y).$$

We define $\mathbf{F}_k = [\mathcal{F}_k^1; \mathcal{F}_k^2; \ldots; \mathcal{F}_k^N]$, where $\mathcal{F}_k^r = \left[{}^{k,r}_{00}; f_{10}^{k,r}; \ldots f_{N-2,0}^{k,r}; f_{01}^{k,r}; \ldots; f_{N-2,N-2}^{k,r}\right]$, for all $k = 1, 2$ and $1 \leqslant r \leqslant N$.

Let us begin with the initial condition $u(x,y,-1) = u_0(x,y)$. Assume that truncated Legendre series gives $u_0(x,y) \approx \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0\phi_i(x)\phi_j(y) = \sum_{i=0}^{N}\sum_{j=0}^{N} \mathfrak{u}_{ij}^0 L_i(x)L_j(y)$, then

$$\sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{k=0}^{N} u_{ijk}\phi_i(x)\phi_j(y)\ell_k(-1) = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0\phi_i(x)\phi_j(y)$$
$$\sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{k=0}^{N} u_{ij0}\phi_i(x)\phi_j(y)) = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0\phi_i(x)\phi_j(y)$$

which gives $u_{ij0} = u_{ij}^0$, for all $0 \leqslant i, j \leqslant N-2$.

Let $u_{0h} = \left[u_{0,0}^0; u_{1,0}^0; \ldots u_{N-2,0}^0; u_{0,1}^0; \ldots u_{N-2,N-2}^0\right]$. Since $\phi_i = L_i - L_{i+2}$, it follows that

$$\sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0\phi_i(x)\phi_j(y) = \sum_{i=0}^{N}\sum_{j=0}^{N} \mathfrak{u}_{ij}^0 L_i(x)L_j(y)$$
$$\sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0(L_i(x) - L_{i+2}(x))(L_j(y) - L_{j+2}(y)) = \sum_{i=0}^{N}\sum_{j=0}^{N} \mathfrak{u}_{ij}^0 L_i(x)L_j(y),$$

thus $(\mathcal{L}\otimes\mathcal{L})u_{ij}^0 = \mathfrak{u}_{ij}^0$, where $\mathcal{L}$ is an $(N+1)\times(N-1)$ Toeplitz matrix and is given as follows:

$$\mathcal{L} = \begin{bmatrix} 1 & 0 & -1 & & & \\ & 1 & 0 & -1 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & 0 & -1 \\ & & & & 1 & 0 \\ & & & & & 1 \end{bmatrix}. \tag{3.4.3}$$

Similarly, v_{0h} is obtained.

Now, the LHS of the weak form of (3.4.1a) collocated at time $t = t_r$, for $1 \leqslant r \leqslant N$, and $0 \leqslant m, n \leqslant N-2$, is equal to

$$
\begin{aligned}
&\big(u_t(x,y,t_r), \phi_m(x)\phi_n(y)\big) + \big(-\Delta u(x,y,t_r), \phi_m(x)\phi_n(y)\big) - \big(p(x,y,t_r), \phi_m'(x)\phi_n(y)\big) \\
&= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{k=1}^{N} u_{ijk}\big(\phi_i(x)\phi_j(y), \phi_m(x)\phi_n(y)\big)\ell_k'(t_r) \\
&\quad + \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}^0\big(\phi_i(x)\phi_j(y), \phi_m(x)\phi_n(y)\big)\ell_0'(t_r) \\
&\quad + \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ijr}\big(-\Delta(\phi_i(x)\phi_j(y)), \phi_m(x)\phi_n(y)\big) \\
&\quad - \sum_{i=0}^{N-2}\sum_{\substack{j=0 \\ i+j>0}}^{N-2} p_{ijr}\big(L_i(x)L_j(y), \phi_m'(x)\phi_n(y)\big)
\end{aligned}
$$

whereas, the RHS is equal to

$$
\big(f_1(x,y,t_r), \phi_m(x)\phi_n(y)\big) = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^{k,r}\big(L_i(x)L_j(y), \phi_m(x)\phi_n(y)\big).
$$

In order to write the discrete weak form of (3.4.1a), described above, in matrix form, we need the following definition.

Definition 3.4.1 (Chebyshev Gauss-Lobatto pseudospectral derivative matrix, see [64, 80])**.** For $N \geqslant 4$, let x_i be the Chebyshev Gauss-Lobatto quadrature nodes, defined as $x_i = -\cos(\frac{\pi i}{N})$ for $0 \leqslant i \leqslant N$. Let $\tilde{c}_0 = \tilde{c}_N = 2$ and $c_i = 1$ for $1 \leqslant i \leqslant N-1$. The *Chebyshev Gauss-Lobatto pseudospectral derivative matrix* is defined as

$D := [d_{i,j}]_{0 \leqslant i,j \leqslant N+1}$, where $d_{k,j} = \ell_j'(x_k)$ given as follows in [80, p. 109].

$$d_{k,j} = \begin{cases} -\dfrac{2N^2+1}{6}, & j = k = 0, \\ \dfrac{\tilde{c}_k(-1)^{k+j}}{\tilde{c}_j(x_k - x_j)}, & 0 \leqslant k \neq j \leqslant N, \\ -\dfrac{x_k}{2(1-x_k^2)}, & 1 \leqslant k = j \leqslant N-1, \\ \dfrac{2N^2+1}{6}, & k = j = N. \end{cases} \tag{3.4.4}$$

Additionally, we define $\mathbf{d}_{0h} := [d_{10}; d_{20}; \ldots; d_{N0}] \in \mathbb{R}^{N \times 1}$, which is the first column of D, except the entry d_{00}. Then, a matrix form of (3.4.1a) becomes,

$$(\mathbf{[}D\mathbf{]} \otimes \mathcal{M} + I_N \otimes A)\, u_h + (I_N \otimes B_1)\, p_h = (\mathbf{1}_N \otimes \mathcal{Q})\, \mathbf{F}_1 - \mathbf{d}_{0h} \otimes (\mathcal{M} u_{0h}).$$

Similarly, a matrix form of (3.4.1b) is given as

$$(\mathbf{[}D\mathbf{]} \otimes \mathcal{M} + I_N \otimes A)\, v_h + (I_N \otimes B_1)\, p_h = (\mathbf{1}_N \otimes \mathcal{Q})\, \mathbf{F}_1 - \mathbf{d}_{0h} \otimes (\mathcal{M} v_{0h}).$$

Finally, consider (3.4.1c), the following weak form of which collocated at time $t = t_r$, for $1 \leqslant r \leqslant N$ and $0 \leqslant m, n \leqslant N-2$ with $m + n > 0$,

$$\begin{aligned} \big(u_x(x,y,t_r), L_m(x)L_n(y)\big) + \big(v_y(x,y,t_r), L_m(x)L_n(y)\big) &= 0 \\ \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ijr}\big(\phi_i'(x)\phi_j(y), L_m(x)L_n(y)\big) + \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ijr}\big(\phi_i(x)\phi_j'(y), L_m(x)L_n(y)\big) &= 0, \end{aligned}$$

which gives the discrete (3.4.1c), as

$$\left(I_N \otimes B_1^T\right) u_h + \left(I_N \otimes B_2^T\right) v_h = \mathbf{0}_{N_\wp}.$$

Consequently, for given $N \geqslant 4$, the *discrete unsteady Stokes problem* becomes

$$\begin{aligned}
\left(\left[D\right] \otimes \mathcal{M} + I_N \otimes A\right) u_h + \left(I_N \otimes B_1\right) p_h &= \left(\mathbf{1}_N \otimes \mathcal{Q}\right) \mathbf{F}_1 - \mathbf{d}_{0h} \otimes \left(\mathcal{M} u_{0h}\right) \\
\left(\left[D\right] \otimes \mathcal{M} + I_N \otimes A\right) v_h + \left(I_N \otimes B_2\right) p_h &= \left(\mathbf{1}_N \otimes \mathcal{Q}\right) \mathbf{F}_2 - \mathbf{d}_{0h} \otimes \left(\mathcal{M} v_{0h}\right) \\
\left(I_N \otimes B_1^T\right) u_h + \left(I_N \otimes B_2^T\right) v_h &= \mathbf{0}_{N_\wp},
\end{aligned} \tag{3.4.5}$$

where $\mathcal{M} = M \otimes M$ and $\mathcal{Q} = Q \otimes Q$. In matrix form, the discrete unsteady Stokes problem becomes

$$\begin{bmatrix}
\left[D\right] \otimes \mathcal{M} + I_N \otimes A & O_{N\vartheta, N\vartheta} & I_N \otimes B_1 \\
O_{N\vartheta, N\vartheta} & \left[D\right] \otimes \mathcal{M} + I_N \otimes A & I_N \otimes B_1 \\
I_N \otimes B_1^T & I_N \otimes B_2^T & O_{N_\wp, N_\wp}
\end{bmatrix}
\begin{bmatrix} u_h \\ v_h \\ p_h \end{bmatrix} =
\begin{bmatrix}
\left(\mathbf{1}_N \otimes \mathcal{Q}\right) \mathbf{F}_1 - \mathbf{d}_{0h} \otimes \left(\mathcal{M} u_{0h}\right) \\
\left(\mathbf{1}_N \otimes \mathcal{Q}\right) \mathbf{F}_2 - \mathbf{d}_{0h} \otimes \left(\mathcal{M} v_{0h}\right) \\
\mathbf{0}_{N_\wp}
\end{bmatrix}.$$

Thus, the coefficient matrix of the discrete unsteady Stokes problem or the *global space-time spectral operator for the unsteady Stokes problem becomes*,

$$G_t = \begin{bmatrix} \mathcal{A}_t & \mathcal{B} \\ \mathcal{B}^T & O \end{bmatrix}, \tag{3.4.6}$$

where the sub-blocks are defined as

$$\mathcal{A}_t = A_t \oplus A_t \in \mathbb{R}^{2N\vartheta \times 2N\vartheta}, \text{ with } A_t = \left[D\right] \otimes \mathcal{M} + I_N \otimes A, \tag{3.4.7}$$

and

$$\mathcal{B} = \begin{bmatrix} I_N \otimes B_1 \\ I_N \otimes B_2 \end{bmatrix} \in \mathbb{R}^{2N\vartheta \times N_\wp}. \tag{3.4.8}$$

Analogous to the steady case, the following are the main features of this scheme for the unsteady Stokes problem:

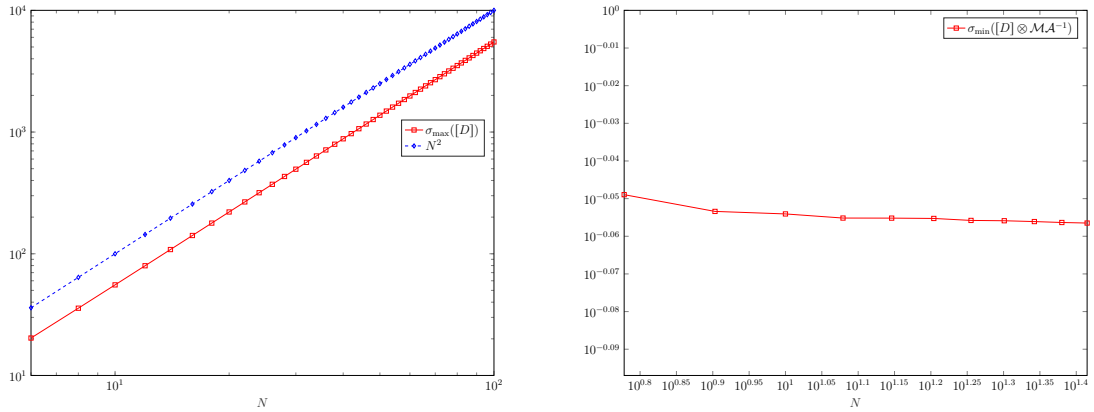
1. the velocity is not exactly divergence-free,

2. this method is a spectral-Galerkin scheme in space and collocation in time;

3. there are no spurious modes for pressure.

3.4.2 Analysis

We now analyze the proposed scheme for the unsteady Stokes problem, with the objective of formulating a condition number estimate for the global space-time spectral operator. We begin our analysis by giving the proof of a well-known numerical fact about the norm of Chebyshev derivative matrix, stated in [22, p. 499] and depicted by Figure 3.7a.



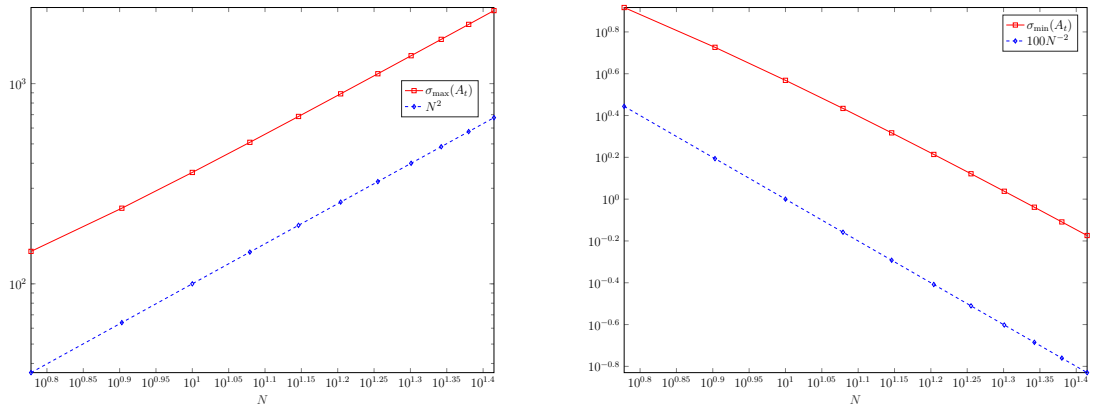
(a) Maximum singular value of $[\![D]\!]$. (b) $\sigma_{\min}([\![D]\!] \otimes \mathcal{M}\mathcal{A}^{-1} + I)$.

Figure 3.7: Some singular value estimates for the unsteady Stokes problem.

Lemma 3.4.2. *For $N \geqslant 2$, let $D \in \mathbb{R}^{(N+1)\times(N+1)}$ be the Chebyshev Gauss-Lobatto pseudospectral derivate matrix, then $\|[\![D]\!]\|_2 \leqslant cN^2$.*

Proof. Since $\|[\![D]\!]\|_2 \leqslant \sqrt{\|[\![D]\!]\|_1 \|[\![D]\!]\|_\infty}$, we evaluate the maximum absolute row and column sum of $[\![D]\!]$ by using Definition 3.4.1. Let C_i and R_i denote the absolute sum of i-th column and i-th row respectively, for $1 \leqslant i \leqslant N$, then

$$C_i = |d_{ii}| + \sum_{\substack{i=1 \\ i \neq j}}^{N} |d_{ij}|. \tag{3.4.9}$$



(a) Maximum singular value. (b) Minimum singular value.

Figure 3.8: Singular values of $\mathcal{A}_t$.

Note that

$$|d_{ii}| = \left|\frac{-x_i}{2(1-x_i^2)}\right| = \frac{\cos\frac{\pi i}{N}}{2\left(1-\cos^2\frac{\pi i}{N}\right)} \leqslant \frac{1}{2\sin^2\frac{\pi i}{N}}.$$

Note that $\dfrac{\pi i}{N} \leqslant \dfrac{\pi}{2}$ for $1 \leqslant i \leqslant \dfrac{N}{2}$, and since for $0 \leqslant x \leqslant \dfrac{\pi}{2}$,

$$\frac{2x}{\pi} \leqslant \sin x \leqslant x, \tag{3.4.10}$$

which implies for $1 \leqslant i \leqslant \dfrac{N}{2}$,

$$|d_{ii}| \leqslant \frac{1}{2\left(\frac{2}{N}\right)^2} \leqslant \frac{N^2}{8},$$

and for $\dfrac{N}{2} < i \leqslant N-1$, $\dfrac{\pi}{N} \leqslant \dfrac{\pi(N-i)}{N} < \dfrac{\pi}{2}$, by applying (3.4.10)

$$|d_{ii}| \leqslant \frac{1}{2\sin^2\frac{\pi i}{N}} = \frac{1}{2\sin^2\left(\frac{\pi}{N}(N-i)\right)} \leqslant \frac{1}{2\left(\frac{2}{N}\right)^2} = \frac{N^2}{4}.$$

Also, for $i = N$, $|d_{NN}| = \frac{2N^2+1}{6} \leqslant cN^2$, thus for all $1 \leqslant i \leqslant N$,

$$|d_{ii}| \leqslant cN^2. \tag{3.4.11}$$

For a fixed $1 \leqslant j \leqslant N$,

$$\begin{aligned}
\sum_{\substack{i=1 \\ i\neq j}}^{N} |d_{ij}| &= \sum_{\substack{i=1 \\ i\neq j}}^{N} \left| \frac{\tilde{c}_i(-1)^{i+j}}{\tilde{c}_j(x_i - x_j)} \right| \\
&\leqslant 2 \sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{1}{|x_i - x_j|} \\
&\leqslant 2 \sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{1}{\left|\cos\frac{\pi i}{N} - \cos\frac{\pi j}{N}\right|} \\
&= \sum_{\substack{i=1 \\ i\neq j}}^{N} \frac{1}{\left|\sin\frac{(j+i)\pi}{2N} \sin\frac{(j-i)\pi}{2N}\right|}.
\end{aligned}$$

For $1 \leqslant i \leqslant j-1$, $\frac{\pi}{2N} \leqslant \frac{(j-i)\pi}{2N} \leqslant \frac{\pi}{2}$, by applying (3.4.10)

$$\sin\left(\frac{(j-i)\pi}{2N}\right) \geqslant \frac{(j-i)}{N}.$$

For $j+1 \leqslant i \leqslant N$, $\frac{\pi}{2N} \leqslant -\frac{(j-i)\pi}{2N} \leqslant \frac{\pi}{2}$, which along with (3.4.10) implies

$$-\frac{(j-i)}{N} \leqslant \sin\left(-\frac{(j-i)\pi}{2N}\right) \leqslant -\frac{(j-i)\pi}{2N},$$

as $|j-i| = -(j-i)$ and $\left|\sin\left(\frac{(j-i)\pi}{2N}\right)\right| = -\sin\left(\frac{(j-i)\pi}{2N}\right)$,

$$\frac{|(j-i)|}{N} \leqslant \left|\sin\left(\frac{(j-i)\pi}{2N}\right)\right| \leqslant \frac{|(j-i)|\pi}{2N},$$

therefore, for all $i \neq j$ and $1 \leqslant i \leqslant N$, implies $\dfrac{1}{\left|\sin\left(\dfrac{(j-i)\pi}{2N}\right)\right|} \leqslant \dfrac{N}{|j-i|}$, thus

$$\sum_{\substack{i=1\\ i\neq j}}^{N} |d_{ij}| \leqslant N \sum_{\substack{i=1\\ i\neq j}}^{N} \frac{1}{\left|\sin\left(\dfrac{(j+i)\pi}{2N}\right)\right| |j-i|}.$$

Since $(j+i)\dfrac{\pi}{2N} \leqslant \dfrac{\pi}{2}$ implies $i \leqslant N-j$, split the above sum as follows

$$\begin{aligned}\sum_{\substack{i=1\\ i\neq j}}^{N} |d_{ij}| &\leqslant N \sum_{\substack{i=1\\ i\neq j}}^{N-j} \frac{1}{\left|\sin\left(\dfrac{(j+i)\pi}{2N}\right)\right| |j-i|} + N \sum_{\substack{i=N-j+1\\ i\neq j}}^{N} \frac{1}{\left|\sin\left(\dfrac{(j+i)\pi}{2N}\right)\right| |j-i|} \\ &=: NS_1 + NS_2. \end{aligned} \tag{3.4.12}$$

For S_1: Note that $1 \leqslant i \leqslant N-j$ gives $\dfrac{\pi}{N} \leqslant \dfrac{(i+j)\pi}{2N} \leqslant \dfrac{\pi}{2}$, by applying (3.4.10)

$$\sin\left(\frac{(i+j)\pi}{2N}\right) \geqslant \frac{2}{\pi}\frac{(i+j)\pi}{2N} = \frac{(i+j)}{N},$$

which implies

$$S_1 \leqslant N \sum_{\substack{i=1\\ i\neq j}}^{N-j} \frac{1}{(i+j)|j-i|} \leqslant N \sum_{\substack{i=1\\ i\neq j}}^{N-j} \frac{1}{|j-i|^2} \leqslant N\frac{\pi^2}{6},$$

as $|j-i| \leqslant i+j$ and $\displaystyle\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$.

For S_2 : Since $N-j+1 \leqslant i \leqslant N$ yields $N+1 \leqslant i+j \leqslant 2N-1$, as $i+j = 2N$ if and only if $i = j = N$, which does not hold since $i \neq j$. Therefore, $\dfrac{\pi}{2N} \leqslant \dfrac{(2N-(i+j))\pi}{2N} \leqslant \dfrac{(2N-1)\pi}{2N} < \dfrac{\pi}{2}$, so by using (3.4.10)

$$\sin\left(\frac{(i+j)\pi}{2N}\right) = \sin\left(\frac{(2N-(i+j))\pi}{2N}\right) \geqslant \frac{2}{\pi}\frac{(2N-(i+j))\pi}{2N} = \frac{(2N-(i+j))}{N},$$

therefore

$$S_2 \leqslant N \sum_{\substack{i=N-j+1 \\ i\neq j}}^{N} \frac{1}{|j-i|(2N-(i+j))}$$
$$\leqslant N \sqrt{\sum_{\substack{i=N-j+1 \\ i\neq j}}^{N} \frac{1}{|j-i|^2} \sum_{\substack{i=N-j+1 \\ i\neq j}}^{N} \frac{1}{(2N-(i+j))^2}} \leqslant N\sqrt{\frac{\pi^2}{6}\frac{\pi^2}{6}} = \frac{\pi^2}{6}N.$$

Thus, $\sum\limits_{\substack{i=N-j+1 \\ i\neq j}}^{N} |d_{ij}| \leqslant 2N^2\frac{\pi^2}{6} \leqslant cN^2$.

By using the results of the above two cases along with (3.4.12), $\sum_{\substack{i=1 \\ i\neq j}}^{N} |d_{ij}| \leqslant cN^2$, which along with (3.4.11) in (3.4.9) yields $C_i \leqslant cN^2$, for all $1 \leqslant i \leqslant N$. Hence,

$$\|[\![D]\!]\|_1 = \max_{1\leqslant i\leqslant N} C_i \leqslant cN^2.$$

Similarly, $\|[\![D]\!]\|_\infty = \max_{1\leqslant i\leqslant N} R_i \leqslant cN^2$, which gives the desired result. □

Remark 3.4.3. The above proof is easily extended to prove that $\sigma_{\max}(D) \leqslant cN^2$, since we only need to add the contribution of $|d_{0,i}| \leqslant cN^2$ to each C_i.

The analysis of the unsteady Stokes problem is much harder than in the steady state because of the presence of the Chebyshev derivative matrix D, which is a non-symmetric matrix with an indefinite symmetric part. These properties are inherited by the leading block $\mathcal{A}_t$ of the global space-time spectral operator G_t. There are no results in the literature for approximating spectrum of a saddle point matrix with the leading block of the form $\mathcal{A}_t$. Several results exist for estimating the spectrum of a symmetric saddle point matrix, thus creating scope for approximating the singular values of G_t, as they are the square-root of the eigenvalues of $G_t^T G_t$. However, the parameters for such a gram matrix could be difficult to analyze, thus we refer to the following result which is derived in Chapter 5.

Theorem 3.4.4 (See Corollary 5.6.6)**.** *For a non-singular saddle point matrix,* $\mathcal{X} =$

$\left[\begin{smallmatrix} A & B^T \\ B & O \end{smallmatrix}\right]$, *so that* A *and* B *are full rank, [54] gives the following estimate*

$$\sigma_{\min}(\mathcal{X}) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}(A), \sigma_{\min}(B)\right\}, \tag{3.4.13}$$

where θ *is the minimum principal angle between the range space* $\mathcal{R}\left(\left[\begin{smallmatrix} A & B^T \end{smallmatrix}\right]^T\right)$ *and* $\mathcal{R}\left(\left[\begin{smallmatrix} B & O \end{smallmatrix}\right]^T\right)$.

Note that (3.4.13) on G_t gives

$$\sigma_{\min}(G_t) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}(\mathcal{A}_t), \sigma_{\min}(\mathcal{B})\right\}. \tag{3.4.14}$$

We could not estimate the term $\sqrt{1-\cos\theta}$, for which a numerical evidence Figure 3.10a suggests

$$\sqrt{1-\cos\theta} \geqslant \frac{c}{N^2}. \tag{3.4.15}$$

Another estimate that has been difficult to show is

$$\sigma_{\min}\left(\llbracket D \rrbracket \otimes \mathcal{M}\,\mathcal{A}^{-1} + I_{2N\vartheta}\right) \geqslant c_1, \tag{3.4.16}$$

where $0 < c_1 < 1$ is a constant, as portrayed by numerical evidence Figure 3.7b. In the following result, we provide a *condition number* estimate for G_t by using computational and theoretical techniques.

Theorem 3.4.5. *For* $N \geqslant 4$, *let* G_t *be defined by* (3.4.6). *Assume eqs.* (3.4.15) *and* (3.4.16) *hold, then* $\kappa(G_t) \leqslant CN^6$.

Proof. We begin by estimating the maximum singular value of A_t,

$$\begin{aligned}
\sigma_{\max}(A_t) = \|A_t\|_2 &= \|\llbracket D \rrbracket \otimes \mathcal{M} + I_N \otimes A\|_2 \\
&\leqslant \|\llbracket D \rrbracket \otimes M \otimes M\|_2 + \|I_N \otimes A\|_2 \qquad \text{(since } \mathcal{M} = M \otimes M\text{)} \\
&= \|\llbracket D \rrbracket\|_2 \|M\|_2^2 + \|A\|_2
\end{aligned}$$

$$\leqslant cN^2 \cdot c + cN \leqslant cN^2,$$

which is obtained by using Lemmas 3.3.2 and 3.4.2, and Theorem 3.3.3. It remains to estimate the minimum singular value of $\mathcal{A}_t$.

$$\begin{aligned}\sigma_{\min}(A_t) &= \sigma_{\min}\left(\left([D] \otimes \mathcal{M}A^{-1} + I_{N\vartheta}\right)(I_N \otimes A)\right) \\ &\geqslant \sigma_{\min}\left([D] \otimes \mathcal{M}A^{-1} + I_{N\vartheta}\right)\sigma_{\min}\left(I_N \otimes \mathcal{A}\right) \\ &\geqslant c_1\sigma_{\min}(A),\end{aligned}$$

is obtained by using (3.4.16), thus Theorem 3.3.3 gives $\sigma_{\min}(A_t) \geqslant \dfrac{c}{N^2}$, and $\sigma(A_t \oplus A_t) = \sigma(A_t)$ implies $\sigma_{\min}(\mathcal{A}_t) \geqslant \dfrac{c}{N^2}$.

Next, we estimate the singular values of $\mathcal{B}$. Since $\mathcal{B}^T\mathcal{B} = I_N \otimes B^TB$, Lemma 3.3.7 gives $\operatorname{rank}(\mathcal{B}^T\mathcal{B}) = \operatorname{rank}(I_N)\cdot\operatorname{rank}(B^TB) = N\cdot\operatorname{rank}(B) = N\wp$. Hence, $\mathcal{B}$ is full rank. Also, $\Lambda(\mathcal{B}^T\mathcal{B}) = \Lambda(I_N)\Lambda(B^TB) = \Lambda(B^TB)$, hence $\sigma(\mathcal{B}) = \sigma(B)$, thus Lemma 3.3.7 implies $\sigma_{\max}(\mathcal{B}) \leqslant c$ and $\sigma_{\min}(\mathcal{B}) \geqslant \dfrac{c}{N^2}$.

Finally, for G_t, by following the proof of (3.3.43),

$$\sigma_{\max}(G_t) = \|G_t\|_2 \leqslant \sigma_{\max}(\mathcal{A}_t) + \sigma_{\max}(\mathcal{B}) \leqslant cN^2 + c \leqslant cN^2.$$

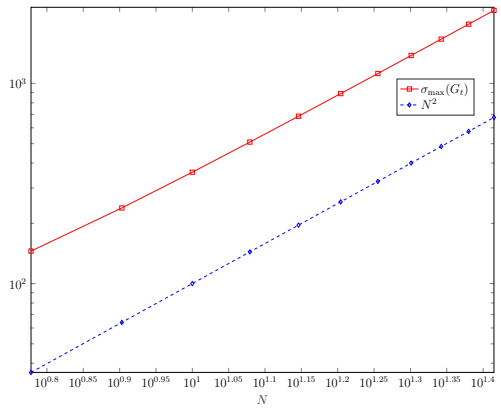
For the minimum singular value of G_t, eqs. (3.4.14) and (3.4.15) imply

$$\sigma_{\min}(G_t) \geqslant \frac{c}{N^2}\min\left(\frac{c}{N^2}, \frac{c}{N^2}\right) \geqslant \frac{c}{N^4}.$$

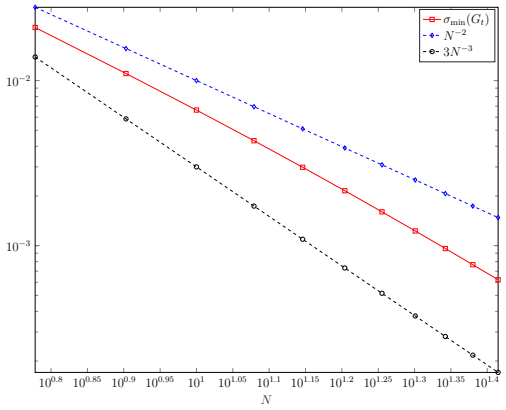
Thus $\kappa(G_t) = \dfrac{\sigma_{\max}(G_t)}{\sigma_{\min}(G_t)} \leqslant cN^6$. □

The above estimate is not sharp, as numerically Figure 3.9b hints that $\sigma_{\min}(G_t)$ behaves like $\mathcal{O}\left(N^{-2.5}\right)$, suggesting $\kappa(G_t) \approx \mathcal{O}(N^{4.5})$.

In the unsteady state, we implemented the scheme derived in Section 3.4.1 for the Stokes problem defined by (3.1.2). Based on our interest in analysis, we selected



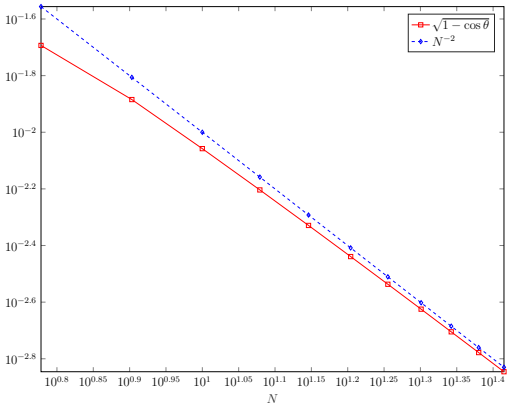
(a) Maximum singular value of G_t.



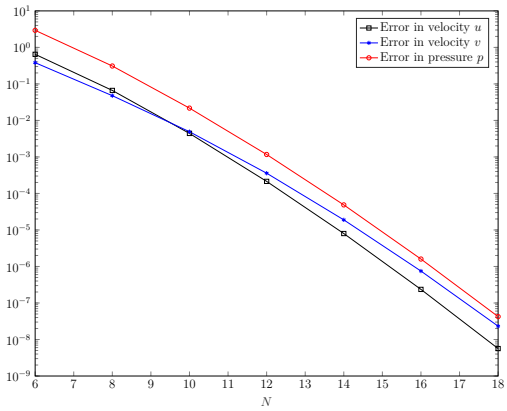
(b) Minimum singular value of G_t.

Figure 3.9: Singular values of G_t.

Chebyshev Gauss-Lobatto collocation in time, which can easily be replaced by other polynomials. For our implementation on MATLAB®, see [52], we take f_1, f_2, so that the exact solutions are $u(x, y, t) = (\cos(\pi x) + 1)\sin(2\pi y)\sin(0.5\pi t)$, $v(x, y, t) = (0.5)\sin(\pi x)(1 - \cos(2\pi y))\sin(0.5\pi t)$, $p(x, y, t) = \sin(\pi x)\cos(\pi y)\sin(0.5\pi t)$, satisfying the boundary and initial condition. The spectral convergence of this scheme is easily observed in Figure 3.10b.



(a) Value of $\sqrt{1 - \cos\theta}$.



(b) Convergence.

Figure 3.10: Numerical results for the unsteady Stokes problem.

3.4.3 Convergence

In this section, we discuss *space-time spectral convergence* of our method for the unsteady Stokes problem, as spectral convergence of the $P_N - P_{N-2}$ scheme for the Stokes problem in steady state was proved in [12].

Let $\|\cdot\|_{0,\omega}$ denote a weighted L^2 norm defined as

$$\|f\|_{0,\omega}^2 = \int_{\Omega_t} f(x,y,t)\frac{1}{\sqrt{1-t^2}}dxdydt.$$

The above norm is designed to incorporate the weight functions for the Legendre polynomials in space and Chebyshev polynomials in time. Recall that the velocity obtained by the scheme devised in this section for the unsteady Stokes problem is not exactly divergence-free, as implied by (3.3.9). Moreover, the uniqueness of solution for this scheme is a direct consequence of Theorem 3.4.5, thus we prove the following result infusing the conditions of the aforementioned result.

Theorem 3.4.6. *Let u, v, and p be the solution of* (3.1.2). *Assume u, v, and p are separately analytic in each variable. Let $N \geqslant 4$ and u_N, v_N, and p_N be the solution of the space-time method, of the form* (3.4.2), *with matrix defined by* (3.4.6). *If eqs.* (3.4.15) *and* (3.4.16) *hold, then for a large enough N*

$$\|u - u_N\|_{0,\omega} + \|v - v_N\|_{0,\omega} + \|p - p_{N-2}\|_{0,\omega} \leqslant cN^8 e^{-CN}.$$

Proof. Consider the exact solution and its truncation as follows,

$$u(x,y,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \hat{\mathrm{u}}_{ij}(t)\phi_i(x)\phi_j(y), \qquad \Pi_N u = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} \hat{\mathrm{u}}_{ij}(t)\phi_i(x)\phi_j(y),$$

$$v(x,y,t) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} \hat{\mathrm{v}}_{ij}(t)\phi_i(x)\phi_j(y), \qquad \Pi_N v = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} \hat{\mathrm{v}}_{ij}(t)\phi_i(x)\phi_j(y),$$

$$p(x,y,t) = \sum_{i=0}^{\infty}\sum_{\substack{j=0 \\ i+j>0}}^{\infty} \hat{\mathrm{p}}_{ij}(t)L_i(x)L_j(y), \qquad \Pi_{N-2} p = \sum_{i=0}^{N-2}\sum_{\substack{j=0 \\ i+j>0}}^{N-2} \hat{\mathrm{p}}_{ij}(t)L_i(x)L_j(y).$$

Let $\mathscr{T}_N u$ denote the truncation error in velocity u, that is, $\mathscr{T}_N u(x,y,t) = (u - \Pi_N u)(x,y,t)$, similarly, let $\mathscr{T}_N v$ and $\mathscr{T}_{N-2} p$ denote the truncation error of velocity v and pressure p, defined as $v - \Pi_N v$ and $p - \Pi_{N-2} p$, respectively.

Define semi-discrete solutions for (3.1.2) representing (3.4.2) as follows

$$\begin{aligned} u_N &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} u_{ij}(t)\phi_i(x)\phi_j(y), \\ v_N &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} v_{ij}(t)\phi_i(x)\phi_j(y), \\ p_{N-2} &= \sum_{i=0}^{N-2}\sum_{\substack{j=0 \\ i+j>0}}^{N-2} p_{ij}(t)L_i(x)L_j(y), \end{aligned}$$

where $u_{ij}(t) = \sum_{k=0}^{N} u_{ijk}\ell_k(t)$, and similarly $v_{ij}(t)$ and p_{ij} are defined, which implies $u_{ij}(t_k) = u_{ijk}$, $v_{ij}(t_k) = v_{ijk}$, and $p_{ij}(t_k) = p_{ijk}$, for t_k are Chebyshev Gauss-Lobatto nodes, $1 \leqslant k \leqslant N$. Also, define $t_h = [t_1; t_2; \ldots; t_N]$.

Define the error in truncated and approximated solutions as

$$\begin{aligned} e^u(x,y,t) &= (\Pi_N u - u_N)(x,y,t), \\ e^v(x,y,t) &= (\Pi_N v - v_N)(x,y,t), \\ e^p(x,y,t) &= (\Pi_{N-2} p - p_{N-2})(x,y,t). \end{aligned}$$

Also, define the error vectors as $E^u = [E_1^u; E_2^u; \ldots; E_N^u]$, $E^v = [E_1^v; E_2^v; \ldots; E_N^v]$, $E^p = [E_1^p; E_2^p; \ldots; E_N^p]$, where for $0 \leqslant i, j \leqslant N-2$ and $1 \leqslant k \leqslant N$, $E_k^u = [\hat{\mathrm{u}}_{ij}(t_k) - u_{ijk}]$, $E_k^v = [\hat{\mathrm{v}}_{ij}(t_k) - v_{ijk}]$, and only $E_k^p = [\hat{\mathrm{p}}_{ij}(t_k) - p_{ijk}]$ is considered along with the condition $i + j > 0$.

Recall that for given f_k in eqs. (3.4.1a) and (3.4.1b), for $k = 1, 2$, so that at time $t = t_r$, $f_k(x,y,t_r)$ is analytic in Ω, then it can be expressed as

$$f_k(x,y,t_r) = \sum_{i=0}^{\infty}\sum_{j=0}^{\infty} f_{ij}^{k,r}\phi_i(x)\phi_j(y), \qquad \Pi_N f_k^r = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} f_{ij}^{k,r}\phi_i(x)\phi_j(y),$$

where $\Pi_N f_k^r$ is the truncation for f_k and the truncation error is defined as $\mathscr{T}_N f_k^r = (f_k(x,y,t_r) - \Pi_N f_k^r)$ for $k = 1,2$ and $1 \leqslant r \leqslant N$.

For $w \in V$, the first equation of the Stokes problem implies that the exact solution u, p satisfy the following weak form, for all $t \in (-1,1)$, thus at time $t = t_k$, where $1 \leqslant k \leqslant N$

$$\big(u_t(x,y,t_k), w\big) + \big(-\Delta u(x,y,t_k), w\big) - \big(p(x,y,t_k), w_x\big) = \big(f_1(x,y,t_k), w\big) \tag{3.4.17}$$

and the approximated solution u_N, p_{N-2} satisfy

$$\begin{aligned}\big((u_N)_t(x,y,t_k), w_N\big) + \big(-\Delta u_N(x,y,t_k), w_N\big)\big) - \big(p_{N-2}(x,y,t_k), (w_N)_x\big) \\ = \big(\Pi_N f_1^k, w_N\big),\end{aligned} \tag{3.4.18}$$

for all $w_N \in \mathbb{P}_{N,N} \cap V$. Subtracting eqs. (3.4.17) and (3.4.18) for all $0 \leqslant m, n \leqslant N-2$ gives

$$\begin{aligned}&\big((u-u_N)_t(x,y,t_k), \phi_m(x)\phi_n(y)\big) + \big(-\Delta(u-u_N)(x,y,t_k), \phi_m(x)\phi_n(y)\big) \\ &\quad - \big((p-p_{N-2})(x,y,t_k), \phi_m'(x)\phi_n(y)\big) = \big((f_1 - \Pi_N f_1^k)(x,y,t_k), \phi_m(x)\phi_n(y)\big)\end{aligned}$$

which gives

$$\begin{aligned}&\big(e_t^u(x,y,t_k), \phi_m(x)\phi_n(y)\big) + \big(-\Delta e^u(x,y,t_k), \phi_m(x)\phi_n(y)\big) - \big(e^p(x,y,t_k), \phi_m'(x)\phi_n(y)\big) \\ &\qquad = \big(\mathscr{T}_N f_1^k, \phi_m(x)\phi_n(y)\big) - \big((\mathscr{T}_N u(x,y,t_k))_t, \phi_m(x)\phi_n(y)\big) \\ &\qquad\quad - \big(-\Delta(\mathscr{T}_N u)(x,y,t_k), \phi_m(x)\phi_n(y)\big) + \big((\mathscr{T}_{N-2} p)(x,y,t_k), \phi_m'(x)\phi_n(y)\big).\end{aligned} \tag{3.4.19}$$

Define

$$g(t) = \big(e^u(x,y,t), \phi_m(x)\phi_n(y)\big) = \sum_{i=0}^{N-2}\sum_{j=0}^{N-2} (\hat{u}_{ij}(t) - u_{ij}(t))\,\big(\phi_i(x)\phi_j(y), \phi_m(x)\phi_n(y)\big),$$

then (3.4.19) becomes

$$g'(t_k) + \big(-\Delta e^u(x,y,t_k), \phi_m(x)\phi_n(y)\big) - \big(e^p(x,y,t_k), \phi_m'(x)\phi_n(y)\big)$$
$$= \big(\mathscr{T}_N f_1^k, \phi_m(x)\phi_n(y)\big) - \big((\mathscr{T}_N u)_t(x,y,t_k), \phi_m(x)\phi_n(y)\big)$$
$$- \big(-\Delta(\mathscr{T}_N u)(x,y,t_k), \phi_m(x)\phi_n(y)\big) + \big((\mathscr{T}_{N-2} p)(x,y,t_k), \phi_m'(x)\phi_n(y)\big). \tag{3.4.20}$$

For any analytic z such that $z(-1) = 0$, recall the definition of the interpolant $\mathcal{I}_N z(t) = \sum_{i=1}^{N} z(t_i)\ell_i(t)$. For $0 \leqslant k \leqslant N-1$,

$$\begin{aligned} z'(t_k) &= \left(\mathcal{I}_N z\right)'(t_k) + \tilde{\epsilon}_k \\ &= \left(\llbracket D \rrbracket \left(\mathcal{I}_N(z)(t_h)\right)\right)_k + \tilde{\epsilon}_k \\ &= \left(\llbracket D \rrbracket \left(z(t_h)\right)\right)_k + \tilde{\epsilon}_k, \end{aligned}$$

where $\tilde{\epsilon}_k = (z - \mathcal{I}_N z)'(t_k)$, according to [78], satisfies

$$|\tilde{\epsilon}_k| \leqslant cN^2 e^{-CN}. \tag{3.4.21}$$

Since the initial condition is $u(x, y-1) = u_0(x,y)$, recall that $\hat{\mathrm{u}}_{ij}(-1) = u_{ij0} = u_{ij}^0$, therefore $g(-1) = \hat{\mathrm{u}}_{ij}(-1) - u_{ij}(-1) = u_{ij}^0 - u_{ij}^0 = 0$. Hence, the above expression implies

$$\begin{aligned} g'(t_k) &= \left(\llbracket D \rrbracket g(t_h)\right)_k + \epsilon_k^1 \\ &= \left(\llbracket D \rrbracket \cdot \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} \left(\hat{\mathrm{u}}_{ij}(t_h) - u_{ij}(t_h)\right) \big(\phi_i(x)\phi_j(y), \phi_m(x)\phi_n(y)\big) \right)_k + \epsilon_k^1, \end{aligned}$$

thus (3.4.20) gives the first $(N-1)^2$ equations for each time step t_k for$1 \leqslant k \leqslant N$

and $0 \leqslant m, n \leqslant N-2$ as follows,

$$\begin{aligned}(\llbracket D \rrbracket \cdot (e^u(x,y,t_h), \phi_m(x)\phi_n(y)))_k + (-\Delta e^u(x,y,t_k), \phi_m(x)\phi_n(y)) \\ - (e^p(x,y,t_k), \phi_m'(x)\phi_n(y)) = (\mathscr{T}_N f_1^k, \phi_m(x)\phi_n(y)) - \epsilon_k^1 \\ - (((\mathscr{T}_N u)_t - \Delta(\mathscr{T}_N u))(x,y,t_k), \phi_m(x)\phi_n(y)) \\ + ((\mathscr{T}_{N-2} p)(x,y,t_k), \phi_m'(x)\phi_n(y)).\end{aligned} \tag{3.4.22}$$

Thus, the $(N-1)^2$ equations together for all time steps $1 \leqslant k \leqslant N$ give

$$(\llbracket D \rrbracket \otimes \mathcal{M} + I_N \otimes \mathcal{A})\, E^u + (I_N \otimes B_1)\, E^p = -\epsilon_1 - R_1^u - R_2^u, \tag{3.4.23}$$

where we define $\epsilon_1 = [\epsilon_1^1; \epsilon_2^1; \ldots; \epsilon_N^1]$, and for $1 \leqslant i \leqslant 2$

$$\begin{aligned} R_i^u &= [r_i^u(t_1); r_i^u(t_2); \ldots; r_i^u(t_N)]\,, \\ r_1^u(t_k) &= [(\mathscr{T}_N f_1^k + ((\mathscr{T}_N u)_t - \Delta(\mathscr{T}_N u))(x,y,t_k), \phi_m(x)\phi_n(y))], \qquad 0 \leqslant m, n \leqslant N-2 \\ r_2^u(t_k) &= [((\mathscr{T}_{N-2} p)(x,y,t_k), \phi_m'(x)\phi_n(y))]\,, \qquad 0 \leqslant m, n \leqslant N-2. \end{aligned}$$

Similarly, the error equation for the velocity v in matrix form is given as

$$(\llbracket D \rrbracket \otimes \mathcal{M} + I_N \otimes \mathcal{A})\, E^v + (I_N \otimes B_2)\, E^p = -\epsilon_2 - R_1^v - R_2^v, \tag{3.4.24}$$

where $r_2^v(t_k) = [(\mathscr{T}_{N-2} p)(x,y,t_k), \phi_m(x)\phi_n'(y))]$, for $0 \leqslant m, n \leqslant N-2$,. The exact solution u, v satisfy the weak form of the third equation of the Stokes problem, for all $q \in L_0^2(\Omega)$ and time $t = t_k$,

$$(q, u_x(x,y,t_k)) + (q, v_y(x,y,t_k)) = 0, \tag{3.4.25}$$

also, the approximate solutions satisfy the following for all $q_{N-2} \in \mathbb{P}_{N-2,N-2} \cap L_0^2(\Omega)$,

$$\big(q_{N-2},(u_N)_x(x,y,t_k)\big)+\big(q_{N-2},(v_N)_y(x,y,t_k)\big)=0, \tag{3.4.26}$$

for all $q_{N-2}\in\mathbb{P}_{N-2,N-2}\cap L_0^2(\Omega)$. Since $q=\sum_{m=0}^{\infty}\sum_{\substack{n=0\\ m+n>0}}^{\infty} q_{mn}L_m(x)L_n(y)$ so that $q_{N-2}=\sum_{m=0}^{\infty}\sum_{\substack{n=0\\ m+n>0}}^{\infty} q_{mn}L_m(x)L_n(y)$, thus by subtracting eqs. (3.4.25) and (3.4.26) for all $0\leqslant m,n\leqslant N-2$ with $m+n>0$ and incorporating the truncated solution

$$\begin{aligned}-\big(L_m(x)L_n(y),e_x^u(x,y,t_k)\big)-\big(L_m(x)L_n(y),e_y^v(x,y,t_k)\big)\\ =\big(L_m(x)L_n(y),((\mathscr{T}_N u)_x+(\mathscr{T}_N v)_y)(x,y,t_k)\big),\end{aligned}. \tag{3.4.27}$$

Thus, the following linear system is obtained.

$$(I_N\otimes B_1^T)E^u+(I_N\otimes B_2^T)E^v=-R_2^p, \tag{3.4.28}$$

where $R_2^p=[r_2^p(t_1);r_2^p(t_2);\ldots;r_2^p(t_N)]$, and for $1\leqslant k\leqslant N$,

$$r_2^p(t_k)=-[\big(L_m(x)L_n(y),((\mathscr{T}_N u)_x+(\mathscr{T}_N v)_y)(x,y,t_k)\big)].$$

where $0\leqslant m,n\leqslant N-2, m+n>0$. Thus, eqs. (3.4.20), (3.4.24) and (3.4.28) imply

$$\begin{bmatrix}\mathcal{A}_t & O_{N\vartheta,N\vartheta} & I_N\otimes B_1\\ O_{N\vartheta,N\vartheta} & \mathcal{A}_t & I_N\otimes B_2\\ I_N\otimes B_1^T & I_N\otimes B_2^T & O_{\wp,\wp}\end{bmatrix}\begin{bmatrix}E^u\\E^v\\E^p\end{bmatrix}=-\begin{bmatrix}\epsilon_1\\\epsilon_2\\0\end{bmatrix}-\begin{bmatrix}R_1^u\\R_1^v\\0\end{bmatrix}-\begin{bmatrix}R_2^u\\R_2^v\\R_2^p\end{bmatrix}$$

which is expressed as the following linear system

$$G_tE=-\epsilon-\sum_{i=1}^{2}R_i,$$

First, we estimate $|G_tE|_\infty$. To this end, (3.4.21) implies $|\epsilon|_\infty\leqslant cN^2e^{-cN}$, it remains

to estimate the infinity-norm of R_i for $1 \leqslant i \leqslant 2$. For R_1, note that the non-zero entries of R_1^u, for $0 \leqslant m, n \leqslant N-2$, are of the form

$$
\begin{aligned}
&\big(\mathscr{T}_N f_1^k + ((\mathscr{T}_N u)_t - \Delta(\mathscr{T}_N u))(x,y,t_k), \phi_m(x)\phi_n(y)\big) \\
&= \big(\mathscr{T}_N f_1^k, \phi_m(x)\phi_n(y)\big) + \big(((\mathscr{T}_N u)_t(x,y,t_k), \phi_m(x)\phi_n(y)\big) - \big(\Delta(\mathscr{T}_N u))(x,y,t_k), \phi_m(x)\phi_n(y)\big) \\
&=: s_f + s_1(t_k) + s_2(t_k).
\end{aligned}
$$

Firstly, $|s_f| \leqslant \|\mathscr{T}_N f_1^k\|_0 \|\phi_m(x)\phi_n(y)\|_0$, by Theorem 5.12 in [64, p. 248] for the Legendre truncation error estimate,

$$|s_f| \leqslant ce^{-CN}.$$

Assume that $s_1(t_k) = z'(t_k)$, where $z(t) = \big((\mathscr{T}_N u)(x,y,t), \phi_m(x)\phi_n(y)\big)$, for some $0 \leqslant m, n \leqslant N-2$. The interpolant of $z(t)$ is given as $\mathcal{I}_N z(t) = \sum_{i=1}^{N} z(t_i)\ell_i(t) + z(-1)\ell_0(t)$, then

$$
\begin{aligned}
s_1(t_k) = z'(t_k) = \left(\mathcal{I}_N z\right)'(t_k) + \varepsilon_k &= \left(\llbracket D \rrbracket \left(\mathcal{I}_N(z)(t_h)\right)\right)_k + z(-1)\ell_0'(t_k) + \varepsilon_k \\
&= \left(\llbracket D \rrbracket \cdot z(t_h)\right)_k + z(-1)\ell_0'(t_k) + \varepsilon_k \\
&= \sum_{i=1}^{N} d_{ki} z(t_i) + z(-1)\ell_0'(t_k) + \varepsilon_k,
\end{aligned}
$$

where the error $|\varepsilon_k| \leqslant cN^2 e^{-CN}$ as derived in [78]. To estimate $s_1(t_k)$, note that for $1 \leqslant i \leqslant N$,

$$z(t_i) = \big((\mathscr{T}_N u)(x,y,t_i), \phi_m(x)\phi_n(y)\big) \leqslant c\|(\mathscr{T}_N u)(x,y,t_i)\|_0 \leqslant ce^{-CN},$$

where we have used Theorem 5.12 in [64, p. 248] for the Legendre truncation error estimate, i.e., $(\mathscr{T}_N u)(x,y,t_i)$. Also, $z(-1) = \big((u - \Pi_N u)(x,y,-1), \phi_m(x)\phi_n(y)\big) \leqslant c\|(u_0 - \Pi_N u_0)(x,y)\|_0 \leqslant ce^{-CN}$. Since $\|\llbracket D \rrbracket\|_\infty \leqslant cN^2$, thus $|d_{ki}| \leqslant cN^2$, and $d_{k0} \leqslant$

cN^2, therefore $|s_1(t_k)| \leqslant cN^3 e^{-CN}$.

Since $s_2(t_k) = \big((\mathscr{T}_N u)(x,y,t_k), -\Delta(\phi_m(x)\phi_n(y))\big)$, thus

$$|s_2(t_k)| \leqslant c\|(u - \Pi_N u)(x,y,t_k)\|_0 \leqslant ce^{-CN}.$$

Thus, $|R_1^u|_\infty \leqslant cN^3 e^{-CN} + 2ce^{-CN} \leqslant cN^3 e^{-CN}$, and similar estimate holds for R_1^v, hence $|R_1|_\infty \leqslant cN^3 e^{-CN}$.

For R_2, its components consist of as r_2^u, r_2^v, and r_2^p. We estimate the entries of r_2^u by using the same Legendre truncation error result, which gives

$$|r_2^u(t_k)| = |\big((\mathscr{T}_{N-2}p)(x,y,t_k), \phi_m'(x)\phi_n(y)\big)| \leqslant c\|(\mathscr{T}_{N-2}p)(x,y,t_k)\| \leqslant ce^{-CN},$$

similar result holds for r_2^v, and finally

$$\begin{aligned}
|r_2^p(t_k)| &= |\big(L_m(x)L_n(y), \left((\mathscr{T}_N u)_x + (\mathscr{T}_N v)_y\right)(x,y,t_k)\big)| \\
&\leqslant |\big(L_m'(x)L_n(y), (\mathscr{T}_N u)(x,y,t_k)\big)| + |\big(L_m(x)L_n'(y), (\mathscr{T}_N v)(x,y,t_k)\big)| \\
&\leqslant c\|\left(\mathscr{T}_N u\right)(x,y,t_k)\| \leqslant ce^{-CN},
\end{aligned}$$

hence, $|R_2|_\infty \leqslant ce^{-CN}$ and implying the following estimate

$$|G_t E|_\infty \leqslant |\epsilon|_\infty + \sum_{i=1}^{3} |R_i|_\infty \leqslant cN^3 e^{-CN}. \tag{3.4.29}$$

The next stage is to estimate the norm of error between the truncated and approximated solution defined e^u, e^v and e^p in the beginning of this proof. Since $\phi_i = L_i - L_{i+2}$,

$$e^u(x,y,t_k) = \sum_{i=0}^{N}\sum_{j=0}^{N} \mathfrak{c}_{ij}^k L_i(x)L_j(y),$$

where $\mathfrak{c}_{ij}^k = (\mathcal{L} \otimes \mathcal{L})\, E_k^u$, $1 \leqslant k \leqslant N$. From (3.4.3), it is easily proved that

$$\|\mathcal{L}\|_2 \leqslant \sqrt{\|\mathcal{L}\|_1 \|\mathcal{L}\|_\infty} \leqslant \sqrt{2 \cdot 2} = 2. \tag{3.4.30}$$

Let the Chebyshev Gauss-Lobatto quadrature weights be denoted by $\omega_i = \dfrac{\pi}{N d_i}$, where $d_0 = 2 = d_N$ and $d_i = 1$ for $1 \leqslant i \leqslant N-1$, and W denote the diagonal matrix containing the weights, $W_{ii} = \omega_i$, for $0 \leqslant i \leqslant N$, thus $\|[W]\|_2 \leqslant \frac{c}{N}$. The weighted norm of e^u is given as

$$\begin{aligned}
\|e^u\|_{0,\omega}^2 &= \int_\Omega |e^u|^2 \frac{1}{\sqrt{1-t^2}} dx dy dt \\
&\leqslant c \sum_{i=0}^{N} \sum_{j=0}^{N} \sum_{k=1}^{N} |\mathfrak{c}_{ij}^k|^2 \omega_k \\
&= c \left| \left([W]^{\frac{1}{2}} \otimes \mathcal{L} \otimes \mathcal{L} \right) E^u \right|_2^2 .
\end{aligned}$$

Similarly, the other two error estimates can be derived to get the following

$$\begin{aligned}
\|e^v\|_{0,\omega}^2 &\leqslant c \left| \left([W]^{\frac{1}{2}} \otimes \mathcal{L} \otimes \mathcal{L} \right) E^v \right|_2^2 \\
\|e^p\|_{0,\omega}^2 &\leqslant c \left| \left([W]^{\frac{1}{2}} \otimes I_{N_\wp} \right) E^p \right|_2^2
\end{aligned}$$

Define $W_h = \left([W]^{\frac{1}{2}} \otimes \mathcal{L} \otimes \mathcal{L} \right) \oplus \left([W]^{\frac{1}{2}} \otimes \mathcal{L} \otimes \mathcal{L} \right) \oplus \left([W]^{\frac{1}{2}} \otimes I_{N_\wp} \right)$, then

$$\|W_h\|_2 \leqslant \max \left\{ \|[W]^{\frac{1}{2}} \otimes \mathcal{L} \otimes \mathcal{L}\|_2, \|[W]^{\frac{1}{2}} \otimes I_{N_\wp}\|_2 \right\} \leqslant \frac{c}{\sqrt{N}}.$$

Define $\|e\| = \sqrt{\|e^u\|_{0,\omega}^2 + \|e^v\|_{0,\omega}^2 + \|e^p\|_{0,\omega}^2}$, then addition of the three estimates for weighted norms of e^u, e^v, and e^p yields,

$$\begin{aligned}
\|e\| &\leqslant c\, |W_h E|_2 \\
&\leqslant c \|W_h\|_2 |E_h|_2
\end{aligned}$$

$$
\begin{aligned}
&\leqslant \frac{c}{\sqrt{N}}\|G_t^{-1}\|_2|G_tE|_2 \\
&\leqslant \frac{c}{\sqrt{N}}\|G_t^{-1}\|_2\sqrt{N(2(N-1)^2-1)}|G_tE|_\infty \quad (\text{as } |x|_2 \leqslant \sqrt{m}|x|_\infty, \text{ for any } x \in \mathbb{R}^m) \\
&\leqslant cN^8e^{-CN},
\end{aligned}
$$

the last inequality results from Theorem 3.4.5 and eq. (3.4.29). Thus, $\|e\| \leqslant cN^8e^{-CN}$, which yields that for some big enough N, or $N > c$, the error in exact and approximate solution is

$$
\begin{aligned}
\|u-u_N\|_{0,\omega} + \|v-v_N\|_{0,\omega} + \|p-p_{N-2}\|_{0,\omega} &\leqslant \|\mathscr{T}_N u\|_{0,\omega} + \|e^u\|_{0,\omega} + \|\mathscr{T}_N v\|_{0,\omega} \\
&\quad + \|e^v\|_{0,\omega} + \|\mathscr{T}_{N-2}p\|_{0,\omega} + \|e^p\|_{0,\omega} \\
&\leqslant ce^{-CN} + c\|e\| \\
&\leqslant cN^8e^{-CN}.
\end{aligned}
$$

□

This concludes the proof of the spectral convergence in both space and time of the $P_N - P_{N-2}$ scheme in space and Chebyshev Gauss-Lobatto collocation in time. Thus, completing the analysis for a space-time spectral method for the Stokes problem.

4

The Navier-Stokes problem

In the former chapter, a space-time spectral method scheme was analyzed for the Stokes problem, which is a linearized version of the Navier-Stokes problem; momentous to the field of fluid dynamics. Consequently, we extend the $P_N - P_{N-2}$ scheme, introduced by [12], to the unsteady Navier-Stokes problem. Furthermore, we extend a staggered grid collocation scheme, derived in [11], to the unsteady Stokes and Navier-Stokes problems. This collocation scheme is implemented by using the quadrature nodes, such as the Jacobi Gauss, Jacobi Gauss-Lobatto and Jacobi Gauss along with $x = \pm 1$. Due to the presence of the staggered grid, we derive the expression for the pseudo-spectral derivative matrix for the Jacobi Gauss nodes on a closed interval for the convenience of application. It is an enormous challenge to analyze the schemes derived in this chapter, thus we only present numerical evidence of spectral convergence of this scheme in both space and time. However, the numerical experiments are conducted for the Jacobi polynomials $J^{\alpha,\beta}$ for any values of the parameters, $\alpha, \beta > -1$.

The Reynold's number, denoted by R_e, is considered to be equal to one for all of the problems considered in this chapter. It would be interesting to explore the the highest Reynolds number flows that can be accurately computed by these

numerical schemes. This is the limit where computations might be used to predict the breakdown of a laminar flow into turbulence. Since a spectral method has a super-algebraic decay in error, we may expect them to work for a flow with much higher Reynold's number. However, it is more tedious than it seems as there can be a problem differentiating between numerical and physical smoothing. High Reynold's number flows require an especially accurate estimate of viscous stresses.

4.1 Introduction

The *Navier-Stokes equations* model the conservation of momentum and conservation of mass for Newtonian fluids, thus describe the relationship between the velocity, pressure, temperature, and density of a moving fluid. It is apt to call them the most consequential problem in fluid dynamics, due to their extensive applications such as modeling water flow in a pipe, ocean currents, air flow around a wing, weather etc. Therefore, they help in design process of vehicles and airplanes, the study of blood flow, area of magneto-hydrodynamics, and in analysis of pollution, among others. They were derived over decades ranging roughly between 1822 to 1850 by Claude-Louis Navier and George Gabriel Stokes. The Clay Mathematics Institute designated the problem of proving the existence and smoothness of a solution of the Navier-Stokes problem in three dimensions as a Millennium Problem, one of seven mathematical problems, signifying its immense mathematical interest.

Recall that $\Omega = (-1,1)^2$, $\Omega_t = \Omega \times (-1,1)$, the velocity field and pressure are denoted by $\mathbf{u} = [u;v] \in V := (H_0^1(\Omega))^2$ and $p \in L_0^2(\Omega) := \left\{ q \in L^2(\Omega) \,\middle|\, \int_\Omega q = 0 \right\}$, respectively. The *Navier-Stokes problem in the unsteady state* is stated as follows,

$$u_t + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - \Delta u + p_x = f_1 \text{ in } \Omega_t, \tag{4.1.1a}$$

$$v_t + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} - \Delta v + p_y = f_2 \text{ in } \Omega_t, \tag{4.1.1b}$$

$$u_x + v_y = 0 \text{ in } \Omega_t, \tag{4.1.1c}$$

$$u = 0,\ v = 0 \text{ on } \partial\Omega,$$

$$u(x, y, -1) = u_0(x, y),\ v(x, y, -1) = v_0(x, y) \text{ in } \Omega.$$

Our goal is to devise space-time spectral method schemes for the above problem. In Section 4.2, we extend the $P_N - P_{N-2}$ scheme for the Stokes problem to (4.1.1). Additionally, a staggered grid collocation scheme using the Jacobi polynomials for the Stokes problem and the Navier-Stokes problem in the unsteady state are presented in Sections 4.3.1 and 4.3.2, respectively.

4.2 Mixed spectral Galerkin scheme

The space-time spectral method involving the $P_N - P_{N-2}$ *scheme*, a *mixed spectral Galerkin scheme* in space and spectral collocation in time, was applied to the unsteady Stokes problem in Section 3.4. This section extends the aforementioned scheme to the unsteady Navier-Stokes problem. Recall the approximation for the velocities u and v, and pressure p, by (3.4.2), which is given as follows

$$\begin{aligned}
u_N(x, y, t) &= \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} \sum_{k=0}^{N} u_{ijk} \phi_i(x) \phi_j(y) \ell_k(t) \in \mathbb{P}^0_{N,N,N}, \\
v_N(x, y, t) &= \sum_{i=0}^{N-2} \sum_{j=0}^{N-2} \sum_{k=0}^{N} v_{ijk} \phi_i(x) \phi_j(y) \ell_k(t) \in \mathbb{P}^0_{N,N,N}, \\
p_N(x, y, t) &= \sum_{i=0}^{N-2} \sum_{\substack{j=0 \\ i+j>0}}^{N-2} \sum_{k=0}^{N} p_{ijk} L_i(x) L_j(y) \ell_k(t) \in \mathbb{P}_{N-2,N-2,N},
\end{aligned}$$

along with their corresponding discrete vectors u_h, v_h, and p_h. Note that we only need to discuss the discretization of the non-linear term $(\mathbf{u} \cdot \nabla)\mathbf{u}$, which requires the following matrices:

1. For a given index $0 \leqslant k \leqslant N-2$, the matrix $\mathfrak{P}^k \in \mathbb{R}^{\vartheta \times \vartheta}$ is defined as

$$\mathfrak{P}^k_{ij} = \int_{-1}^{1} \phi_i(x)\phi_j'(x)\phi_k(x)dx, \qquad 0 \leqslant i,j \leqslant N-2. \tag{4.2.1}$$

2. For a given index $0 \leqslant k \leqslant N-2$, the matrix $\mathfrak{T}^k \in \mathbb{R}^{\vartheta \times \vartheta}$ is defined as

$$\mathfrak{T}^k_{ij} = \int_{-1}^{1} \phi_i(x)\phi_j(x)\phi_k(x)dx, \qquad 0 \leqslant i,j \leqslant N-2. \tag{4.2.2}$$

The above matrices are easily calculated by using Definition 3.2.1, that is, $\phi_j = L_j - L_{j+2}$ for $j \in \mathbb{N} \cup \{0\}$ and the following expression for the *triple product of the Legendre polynomials* given in [71],

$$\int_{-1}^{1} L_i(x)L_j(x)L_k(x)dx = 2\begin{pmatrix} i & j & k \\ 0 & 0 & 0 \end{pmatrix}^2,$$

where the special case of $3j$ symbol, when $2s = i+j+k$ is even, yields

$$\begin{pmatrix} i & j & k \\ 0 & 0 & 0 \end{pmatrix} = (-1)^s \sqrt{\frac{(2s-2i)!(2s-2j)!(2s-2k)!}{(2s+1)!}} \frac{s!}{(s-i)!(s-j)!(s-k)!},$$

whereas it is equal to zero whenever $i+j+k$ is odd.

Now, we are ready to discretize the four non-linear terms in (4.1.1).

Term 1: The first non-linear term in (4.1.1a) is $u\dfrac{\partial u}{\partial x}$. Its weak form collocated at time $t = t_r$, for $1 \leqslant r \leqslant N$ and $0 \leqslant m,n \leqslant N-2$, is given as

$$N_1^{m,n,r} = \Big(u\frac{\partial u}{\partial x}(x,y,t_r), \phi_m(x)\phi_n(y)\Big),$$

which on using the approximation of u defined by (3.4.2) becomes equal to

$$\begin{aligned} N_1^{m,n,r} &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{q=0}^{N-2}\sum_{s=0}^{N-2} u_{ijr}u_{qsr}\int_{-1}^{1}\phi_i(x)\phi_q'(x)\phi_m(x)dx\int_{-1}^{1}\phi_j(y)\phi_s(y)\phi_n(y)dy \\ &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{q=0}^{N-2}\sum_{s=0}^{N-2} u_{ijr}u_{qsr}\mathfrak{P}_{iq}^{m}\mathfrak{T}_{js}^{n} \\ &= (u_h^r)^T(\mathfrak{T}^n\otimes\mathfrak{P}^m)u_h^r. \end{aligned}$$

Now, we try to formulate it for a fixed $1 \leqslant r \leqslant N$, and all $0 \leqslant m, n \leqslant N-2$,

$$\left[(u_h^r)^T(\mathfrak{T}^0\otimes\mathfrak{P}^0)u_h^r; (u_h^r)^T(\mathfrak{T}^0\otimes\mathfrak{P}^1)u_h^r; \ldots; (u_h^r)^T(\mathfrak{T}^{N-2}\otimes\mathfrak{P}^{N-2})u_h^r\right],$$

the above $(N-1)^2\times 1$ vector can be written as

$$N_1^r = (I_\vartheta\otimes(u_h^r)^T)W_1u_h^r, \tag{4.2.3}$$

where we define $W_1 \in \mathbb{R}^{\vartheta^2\times\vartheta}$as s block column matrix with (m,n)-column as $(\mathfrak{T}^n\otimes\mathfrak{P}^m)\in\mathbb{R}^{\vartheta\times\vartheta}$, for all $0\leqslant m,n\leqslant N-2$.

Term 2: The second non-linear term in (4.1.1a) is $v\dfrac{\partial u}{\partial y}$. Its weak form collocated at time $t=t_r$, for $1\leqslant r\leqslant N$ and $0\leqslant m,n\leqslant N-2$, is given as

$$N_2^{m,n,r} = \Big(v\frac{\partial u}{\partial y}(x,y,t_r), \phi_m(x)\phi_n(y)\Big),$$

which on using the approximation of u and v defined by (3.4.2) becomes

$$\begin{aligned} N_2^{m,n,r} &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{q=0}^{N-2}\sum_{s=0}^{N-2} v_{ijr}u_{qsr}\int_{-1}^{1}\phi_i(x)\phi_q(x)\phi_m(x)dx\int_{-1}^{1}\phi_j(y)'\phi_s(y)\phi_n(y)dy \\ &= \sum_{i=0}^{N-2}\sum_{j=0}^{N-2}\sum_{q=0}^{N-2}\sum_{s=0}^{N-2} v_{ijr}u_{qsr}\mathfrak{T}_{iq}^{m}\mathfrak{P}_{js}^{n} \\ &= (v_h^r)^T(\mathfrak{P}^n\otimes\mathfrak{T}^m)u_h^r. \end{aligned}$$

Now, we try to formulate it for a fixed $1 \leqslant r \leqslant N$, and all $0 \leqslant m, n \leqslant N-2$,

$$\left[(v_h^r)^T(\mathfrak{P}^0 \otimes \mathfrak{T}^0)u_h^r; (v_h^r)^T(\mathfrak{P}^0 \otimes \mathfrak{T}^1)u_h^r; \ldots; (v_h^r)^T(\mathfrak{P}^{N-2} \otimes \mathfrak{T}^{N-2})u_h^r\right],$$

the above $(N-1)^2 \times 1$ vector can be written as

$$N_2^r = (I_\vartheta \otimes (v_h^r)^T)W_2 u_h^r, \tag{4.2.4}$$

where we define $W_2 \in \mathbb{R}^{\vartheta^2 \times \vartheta}$ as a block column matrix with (m,n)-column as $(\mathfrak{P}^n \otimes \mathfrak{T}^m \in \mathbb{R}^{\vartheta \times \vartheta}$, for all $0 \leqslant m, n \leqslant N-2$.

Term 3: The first non-linear term in (4.1.1b) is $u\dfrac{\partial v}{\partial x}$. Its weak form collocated at time $t = t_r$, for $1 \leqslant r \leqslant N$ and $0 \leqslant m, n \leqslant N-2$, is given as

$$N_3^{m,n,r} = \big(u\frac{\partial v}{\partial x}(x, y, t_r), \phi_m(x)\phi_n(y)\big),$$

so similar to (4.2.3):

$$N_3^r = (I_\vartheta \otimes (u_h^r)^T)W_1 v_h^r. \tag{4.2.5}$$

Term 4: The second non-linear term in (4.1.1b) is $v\dfrac{\partial v}{\partial y}$. Its weak form collocated at time $t = t_r$, for $1 \leqslant r \leqslant N$ and $0 \leqslant m, n \leqslant N-2$, is given as

$$N_4^{m,n,r} = \big(v\frac{\partial v}{\partial y}(x, y, t_r), \phi_m(x)\phi_n(y)\big),$$

so similar to (4.2.4):

$$N_4^r = (I_\vartheta \otimes (v_h^r)^T)W_2 v_h^r. \tag{4.2.6}$$

Thus, eqs. (4.2.3) and (4.2.4) give the discretization of the non-linear terms in

(4.1.1a), for all $1 \leqslant r \leqslant N$, and all $0 \leqslant m, n \leqslant N-2$, as

$$Wu_h, \qquad W = \oplus_{j=1}^{N} \left(\left(I_\vartheta \otimes u_h^j \right) W_1 + \left(I_\vartheta \otimes v_h^j \right) W_2 \right),$$

that is, W is a block diagonal matrix. Similarly, eqs. (4.2.5) and (4.2.6) give the discretization of the non-linear terms in (4.1.1b) for all $1 \leqslant r \leqslant N$, and all $0 \leqslant m, n \leqslant N-2$, as Wv_h. Since these terms are non-linear, we implement a simple fixed point iteration to solve the *discrete unsteady Navier-Stokes problem*, which is given by eq. (3.4.5) and the non-linear terms as follows,

$$\begin{aligned}
\left(W^{(k-1)} + [D] \otimes \mathcal{M} + I_N \otimes A \right) u_h^{(k)} + \left(I_N \otimes B_1 \right) p_h^{(k)} &= \left(\mathbf{1}_N \otimes \mathcal{Q} \right) \mathbf{F}_1 - D \otimes \mathcal{M} u_{0h}, \\
\left(W^{(k-1)} + [D] \otimes \mathcal{M} + I_N \otimes A \right) v_h^{(k)} + \left(I_N \otimes B_2 \right) p_h^{(k)} &= \left(\mathbf{1}_N \otimes \mathcal{Q} \right) \mathbf{F}_2 - D \otimes \mathcal{M} v_{0h}, \\
\left(I_N \otimes B_1^T \right) u_h^{(k)} + \left(I_N \otimes B_2^T \right) v_h^{(k)} &= O,
\end{aligned} \tag{4.2.7}$$

where the non-linear term $W^{(k-1)}$ is a block diagonal matrix with $N-1$ blocks with $W_{ii}^{(k-1)} = \left(I_\vartheta \otimes u_h^{i,(k-1)} \right) W_1 + \left(I_\vartheta \otimes v_h^{i,(k-1)} \right) W_2$. Here, $u_h^{i,(k-1)}$ and $v_h^{i,(k-1)}$ represent the component of u_h and v_h vectors for time $t = t_i$ at $(k-1)$st iteration, for $1 \leqslant i \leqslant N$. The decay of error in L^∞ norm at the final time step $t_N = 1$ for this scheme is shown in Figure 4.1. The schemes described in this chapter are implemented on MATLAB® and are given in [52]. The iteration is stopped whenever the infinity norm of the difference of two consecutive iterates is smaller than $\epsilon = 10^{-12}$. We take f_1 and f_2, so that the exact solutions are

$$\begin{aligned}
u(x, y, t) &= (\cos(\pi x) + 1) \sin(2\pi y) \sin(0.5\pi t), \\
v(x, y, t) &= (0.5) \sin(\pi x)(1 - \cos(2\pi y)) \sin(0.5\pi t), \\
p(x, y, t) &= \sin(\pi x) \cos(\pi y) \sin(0.5\pi t),
\end{aligned} \tag{4.2.8}$$

satisfying the boundary conditions, and giving initial conditions as $u(x, y, -1)$ and $v(x, y, -1)$.

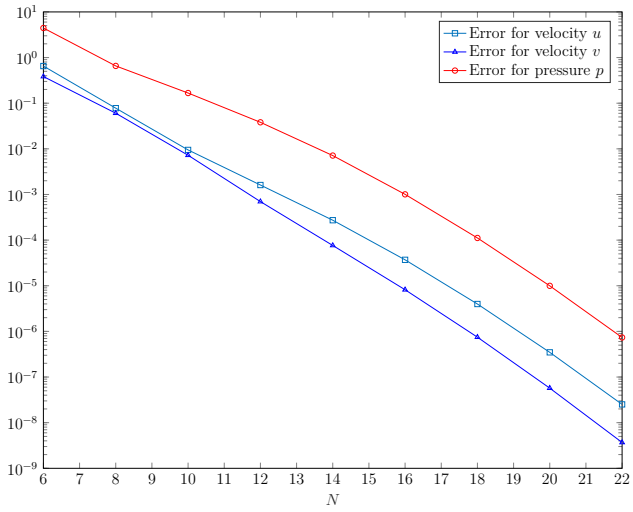


Figure 4.1: Convergence of the unsteady Navier-Stokes problem with the $P_N - P_{N-2}$ scheme.

4.3 Staggered-grid collocation scheme

In [11], the authors presented a collocation scheme for the steady Stokes problem that employs the use of different grids for velocity and pressure, thus the name *staggered grid*. We extend it to the unsteady state by using the Jacobi Gauss-Lobatto collocation in time. For given N, let $J_N^{\alpha,\beta}$ denote the Jacobi polynomial of degree N, orthogonal with respect to the Jacobi weight function $\omega(x) = (1-x)^\alpha(1+x)^\beta$. The *grids for velocity and pressure* are defined as follows.

1. The velocity u is defined on $\Xi_N^{(x)} := \{(\xi_m, \zeta_n, \xi_r) \,|\, 0 \leqslant m, r \leqslant N,\ 0 \leqslant n \leqslant N+1\}$.

2. The velocity v is defined on $\Xi_N^{(y)} := \{(\zeta_m, \xi_n, \xi_r) \,|\, 0 \leqslant m \leqslant N+1,\ 0 \leqslant n, r \leqslant N\}$.

3. The pressure p is defined on $\Xi_N^{(p)} := \{(\zeta_m, \zeta_n, \xi_r) \,|\, 1 \leqslant m, n \leqslant N,\ 1 \leqslant r \leqslant N\}$,

where $\{\xi_k\}_{k=0}^N$ denote the Jacobi-Gauss-Lobatto nodes, $\{\zeta_k\}_{k=1}^N$ represent the Jacobi-Gauss nodes, and we additionally define $\zeta_0 = -1$ and $\zeta_{N+1} = 1$. Thus, the approxi-

mations are defined as follows.

$$u(x,y,t) = \sum_{i=0}^{N}\sum_{j=0}^{N+1}\sum_{k=0}^{N} u_{ijk}\ell_i(x)\mathfrak{p}_j(y)\ell_k(t) \in \mathbb{P}^0_{N,N+1,N},$$
$$v(x,y,t) = \sum_{i=0}^{N+1}\sum_{j=0}^{N}\sum_{k=0}^{N} v_{ijk}\mathfrak{p}_i(x)\ell_j(y)\ell_k(t) \in \mathbb{P}^0_{N+1,N,N},$$
$$p(x,y,t) = \sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{N} p_{ijk}\mathfrak{L}_i(x)\mathfrak{L}_j(y)\ell_k(t) \in \mathbb{P}_{N-1,N-1,N-1}.$$

Since the problem definition contains homogeneous boundary conditions, $u = 0$ on $\partial\Omega$ implies $u_{0nr} = u_{Nnr} = u_{m0r} = u_{m,N+1,r} = 0$, for all $1 \leqslant m \leqslant N-1$, $1 \leqslant n \leqslant N$, and $0 \leqslant r \leqslant N$. Similarly, $v_{0nr} = v_{N+1,nr} = v_{m0r} = v_{mNr} = 0$, for all $1 \leqslant m \leqslant N$, $1 \leqslant n \leqslant N-1$, and $0 \leqslant r \leqslant N$. Also, pressure needs to have zero average, which is enforced by setting $p_{11k} = 0$, for all time steps $1 \leqslant k \leqslant N$. See Figure 4.2 for a plot of the grid of unknowns for the velocities u and v, and pressure p.

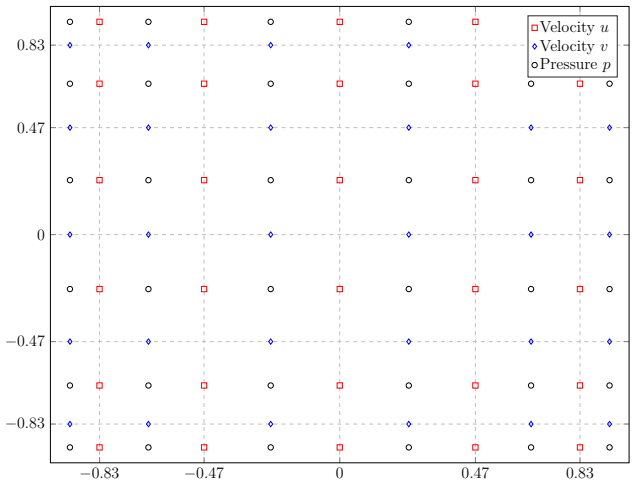


Figure 4.2: Staggered grid of unknowns in x and y axes for $N = 6$.

Therefore, on incorporating the homogenous boundary conditions and zero average condition for velocity and pressure, respectively, the following *variable approxi-*

mations are obtained

$$u(x,y,t) = \sum_{i=1}^{N-1}\sum_{j=1}^{N}\sum_{k=0}^{N} u_{ijk}\ell_i(x)\mathfrak{p}_j(y)\ell_k(t) \in \mathbb{P}^0_{N-2,N-1,N}, \tag{4.3.2a}$$

$$v(x,y,t) = \sum_{i=1}^{N}\sum_{j=1}^{N-1}\sum_{k=0}^{N} v_{ijk}\mathfrak{p}_i(x)\ell_j(y)\ell_k(t) \in \mathbb{P}^0_{N-1,N-2,N}, \tag{4.3.2b}$$

$$p(x,y,t) = \sum_{i=1}^{N}\sum_{\substack{j=1\\ i+j>2}}^{N}\sum_{k=1}^{N} p_{ijk}\mathfrak{L}_i(x)\mathfrak{L}_j(y)\ell_k(t) \in \mathbb{P}_{N-1,N-1,N-1}, \tag{4.3.2c}$$

where $\ell_k, \mathfrak{L}_k$, and $\mathfrak{p}_k$ denote the Lagrange polynomials for $\{\xi_k\}_{k=0}^N$, $\{\zeta_k\}_{k=1}^N$, and $\{\zeta_k\}_{k=0}^{N+1}$ nodes, respectively.

We define the discrete unknowns for the staggered grid scheme as follows.

1. For the velocity u, define $u_h = [u_h^1; u_h^2; \ldots; u_h^N] \in \mathbb{R}^{N^2(N-1)}$, where

$$u_h^k = [u_{11k}; u_{21k}; \ldots; u_{N-1,1,k}; u_{12k}; \ldots; u_{N-1,N,k}] \in \mathbb{R}^{N(N-1)}, \qquad 1 \leqslant k \leqslant N.$$

2. For the velocity v, define $v_h = [v_h^1; v_h^2; \ldots; v_h^N] \in \mathbb{R}^{N^2(N-1)}$, where

$$v_h^k = [v_{11k}; v_{21k}; \ldots; v_{N,1,k}; v_{12k}; \ldots; v_{N,N-1,k}] \in \mathbb{R}^{N(N-1)}, \qquad 1 \leqslant k \leqslant N.$$

3. For the pressure p, define $p_h = [p_h^1; p_h^2; \ldots; p_h^N] \in \mathbb{R}^{N(N^2-1)}$, where

$$p_h^k = [p_{21k}; p_{31k}; \ldots; p_{N,1,k}; p_{12k}; \ldots; p_{N,N,k}] \in \mathbb{R}^{N^2-1}, \qquad 1 \leqslant k \leqslant N.$$

Let $\mathfrak{D}$ represent the Jacobi Gauss-Lobatto pseudo-spectral derivative matrix, the expression for which is well-known, see [80, p. 89].

Definition 4.3.1 (Jacobi Gauss-Lobatto pseudospectral derivative matrix, see [80])**.** For given N, let x_i be the Jacobi Gauss-Lobatto quadrature nodes for $0 \leqslant i \leqslant N$. Let ℓ_i be the Lagrange polynomials for x_i, where $0 \leqslant i \leqslant N$ and $J(x) :=$

$\partial_x J_{N-1}^{\alpha+1,\beta+1}(x)$. The *Jacobi Gauss-Lobatto pseudospectral derivative matrix* is defined as $\mathfrak{D} := (\mathfrak{d}_{i,j})_{0\leqslant i,j\leqslant N}$, where $\mathfrak{d}_{i,j} = \ell_j'(x_i)$ given as follows.

1. For $j = 0$:

$$\mathfrak{d}_{i,0} = \begin{cases} \dfrac{\alpha - N(N+\alpha+\beta+1)}{2(\beta+2)}, & i = 0, \\ \dfrac{(-1)^{N-1}\Gamma(N)\Gamma(\beta+2)}{2\Gamma(N+\beta+1)}(1-x_i)J(x_i), & 1 \leqslant i \leqslant N-1, \\ \dfrac{(-1)^N\Gamma(\beta+2)\Gamma(N+\alpha+1)}{2\Gamma(\alpha+2)\Gamma(N+\beta+1)}, & i = N. \end{cases}$$

2. For $1 \leqslant j \leqslant N-1$:

$$\mathfrak{d}_{i,j} = \begin{cases} \dfrac{2(-1)^N\Gamma(N+\beta+1)}{\Gamma(N)\Gamma(\beta+2)(1-x_j)(1+x_j)^2J(x_j)}, & i = 0, \\ \dfrac{(1-x_i^2)J(x_i)}{(1-x_j^2)J(x_j)(x_i-x_j)}, & i \neq j,\ 1 \leqslant i \leqslant N-1, \\ \dfrac{\alpha-\beta+(\alpha+\beta)x_i}{2(1-x_i^2)}, & 1 \leqslant i = j \leqslant N-1, \\ \dfrac{-2\Gamma(N+\alpha+1)}{\Gamma(N)\Gamma(\alpha+2)(1-x_j)^2(1+x_j)J(x_j)}, & i = N. \end{cases}$$

3. For $j = N$:

$$\mathfrak{d}_{i,N} = \begin{cases} \dfrac{(-1)^{N+1}\Gamma(\alpha+2)\Gamma(N+\beta+1)}{2\Gamma(\beta+2)\Gamma(N+\alpha+1)}, & i = 0, \\ \dfrac{\Gamma(N)\Gamma(\alpha+2)}{2\Gamma(N+\alpha+1)}(1+x_i)J(x_i), & 1 \leqslant i \leqslant N-1, \\ \dfrac{N(N+\alpha+\beta+1)-\beta}{2(\alpha+2)}, & i = N. \end{cases}$$

Due the presence of staggered grid, the grid for velocity v requires the pseudo-spectral derivative matrix for $\{\zeta_k\}_{k=0}^{N+1}$. Such a quadrature employing Gauss quadrature nodes along with the end points is unusual. Thus, we derive the following expression.

Definition 4.3.2 (Jacobi Gauss pseudospectral derivative matrix on a closed interval)**.** For given N, let $x_0 = -1$, $x_{N+1} = 1$ and x_i be the Jacobi-Gauss quadrature nodes for $1 \leqslant i \leqslant N$. Let $\mathfrak{p}_i$ be the Lagrange polynomials for x_i, where $0 \leqslant i \leqslant N+1$. The *Jacobi Gauss pseudospectral derivative matrix on a closed interval* is defined as $\mathscr{D} := (\mathrm{d}_{i,j})_{0\leqslant i,j\leqslant N+1}$, where $\mathrm{d}_{i,j} = \mathfrak{p}_j'(x_i)$ given as follows.

1. For $j = 0$:

$$\mathrm{d}_{i,0} = \begin{cases} \dfrac{J_N^{\alpha,\beta}(-1) - 2\partial_x J_N^{\alpha,\beta}(-1)}{-2J_N^{\alpha,\beta}(-1)}, & i = 0, \\ \dfrac{(1-x_i)\partial_x J_N^{\alpha,\beta}(x_i)}{2J_N^{\alpha,\beta}(-1)}, & 1 \leqslant i \leqslant N, \\ \dfrac{J_N^{\alpha,\beta}(1)}{-2J_N^{\alpha,\beta}(-1)}, & i = N+1. \end{cases}$$

2. For $1 \leqslant j \leqslant N$:

$$\mathrm{d}_{i,j} = \begin{cases} \dfrac{-2J_N^{\alpha,\beta}(-1)}{(1+x_j)^2(1-x_j)\partial_x J_N^{\alpha,\beta}(x_j)}, & i = 0, \\ \dfrac{(1-x_i^2)\partial_x J_N^{\alpha,\beta}(x_i)}{(1-x_j^2)\partial_x J_N^{\alpha,\beta}(x_j)(x_i - x_j)}, & i \neq j,\ 1 \leqslant i \leqslant N, \\ \dfrac{\alpha - \beta + (\alpha+\beta-2)x_j}{2(1-x_j^2)}, & 1 \leqslant i = j \leqslant N, \\ \dfrac{-2J_N^{\alpha,\beta}(1)}{(1+x_j)(1-x_j)^2\partial_x J_N^{\alpha,\beta}(x_j)}, & i = N+1. \end{cases}$$

3. For $j = N+1$:

$$\mathrm{d}_{i,N+1} = \begin{cases} \dfrac{J_N^{\alpha,\beta}(-1)}{2J_N^{\alpha,\beta}(1)}, & i = 0, \\ \dfrac{(1+x_i)\partial_x J_N^{\alpha,\beta}(x_i)}{2J_N^{\alpha,\beta}(1)}, & 1 \leqslant i \leqslant N, \\ \dfrac{J_N^{\alpha,\beta}(1) + 2\partial_x J_N^{\alpha,\beta}(1)}{2J_N^{\alpha,\beta}(1)}, & i = N+1. \end{cases}$$

The proof for the derivation of the above formula is not presented here, as it is similar to that of other pseudospectral derivative matrices discussed in the literature

such as [80, 64].

4.3.1 The Stokes problem

The Stokes problem in unsteady state is given by (3.1.2). Firstly, the initial condition $u(x, y, -1) = u_0(x, y)$ and (4.3.2a) collocated on (ξ_m, ζ_n) for all $0 \leqslant m \leqslant N$ and $0 \leqslant n \leqslant N + 1$, gives

$$\sum_{i=1}^{N-1} \sum_{j=1}^{N} \sum_{k=0}^{N} u_{ijk} \ell_i(\xi_m) \mathfrak{p}_j(\zeta_n) \ell_k(-1) = u_0(\xi_m, \zeta_n)$$

$$u_{mn0} = u_0(\xi_m, \zeta_n),$$

thus $u_h^0 = u_{h0}$, where $u_{h0} = [u_0(\xi_0, \zeta_0); \ldots; u_0(\xi_N, \zeta_{N+1})]$.

Similarly, $v(x, y, -1) = v_0(x, y)$ and (4.3.2b) collocated on (ζ_m, ξ_n) for all $0 \leqslant m \leqslant N + 1$ and $0 \leqslant n \leqslant N$, gives $v_h^0 = v_{h0}$, where $u_{h0} = [u_0(\zeta_0, \xi_0); \ldots; u_0(\zeta_{N+1}, \xi_N)]$.

Define the collection of boundary and initial nodes as $\Xi_0^{(x)} = \{(\pm 1, \zeta_n, t_r) \,|\, 1 \leqslant n \leqslant N, 1 \leqslant r \leqslant N\} \cup \{(\xi_m, \pm 1, t_r) \,|\, 1 \leqslant m \leqslant N-1, 1 \leqslant r \leqslant N\} \cup \{(\xi_m, \zeta_n, -1) \,|\, 1 \leqslant m \leqslant N-1, 1 \leqslant n \leqslant N\}$, and $\Xi_0^{(y)} = \{(\pm 1, \xi_n, t_r) \,|\, 1 \leqslant n \leqslant N-1, 1 \leqslant r \leqslant N\} \cup \{(\zeta_m, \pm 1, t_r) \,|\, 1 \leqslant m \leqslant N, 1 \leqslant r \leqslant N\} \cup \{(\zeta_m, \xi_n, -1) \,|\, 1 \leqslant m \leqslant N, 1 \leqslant n \leqslant N-1\}$.

In order to perform staggered grid collocation on the unsteady Stokes problem, we need the following interpolation matrices.

1. The *pressure derivative interpolation matrix* $\mathfrak{B} \in \mathbb{R}^{N-1 \times N}$ is defined as

$$\mathfrak{B}_{ij} = \mathfrak{L}_j'(\xi_i), \qquad 1 \leqslant i \leqslant N-1, \qquad 1 \leqslant j \leqslant N. \tag{4.3.3}$$

2. The *velocity derivative interpolation matrix* $\mathfrak{C} \in \mathbb{R}^{N-1 \times N}$ is defined as follows

$$\mathfrak{C}_{ij} = \ell_j'(\zeta_i), \qquad 1 \leqslant i \leqslant N, \qquad 1 \leqslant j \leqslant N-1. \tag{4.3.4}$$

In this scheme, we perform the following collocation on the Stokes equations eqs. (3.4.1a) to (3.4.1c),

$$(u_t - \Delta u + p_x)(x,y,t) = f_1(x,y,t), \qquad (x,y,t) \in \Xi_N^{(x)} \backslash \Xi_0^{(x)}, \tag{4.3.5a}$$

$$(v_t - \Delta v + p_y)(x,y,t) = f_2(x,y,t), \qquad (x,y,t) \in \Xi_N^{(y)} \backslash \Xi_0^{(y)}, \tag{4.3.5b}$$

$$(u_x + v_y)(x,y,t) = 0, \qquad (x,y,t) \in \Xi_N^{(p)}. \tag{4.3.5c}$$

Consider (4.3.5a) for some $1 \leqslant m \leqslant N-1$, $1 \leqslant n \leqslant N$, and $1 \leqslant r \leqslant N$

$$\sum_{k=0}^{N} u_{mnk}\ell_k'(\xi_r) - \sum_{i=1}^{N-1} u_{inr}\ell_i''(\xi_m) - \sum_{j=1}^{N} u_{mjr}\mathfrak{p}_j''(\zeta_n) - \sum_{\substack{i=1 \\ i+n>2}}^{N} p_{inr}\mathfrak{L}_i'(\xi_m) = f_1(\xi_m,\zeta_n,\xi_r)$$

$$\sum_{k=1}^{N} u_{mnk}\ell_k'(\xi_r) - \sum_{i=1}^{N-1} u_{inr}\ell_i''(\xi_m) - \sum_{j=1}^{N} u_{mjr}\mathfrak{p}_j''(\zeta_n) - \sum_{\substack{i=1 \\ i+n>2}}^{N} p_{inr}\mathfrak{L}_i'(\xi_m) = f_1(\xi_m,\zeta_n,\xi_r) - u_0(\xi_m,\zeta_n)\ell_0'(\xi_r).$$

By (4.3.3), Definitions 4.3.1 and 4.3.2, the above equation gives the following matrix form,

$$\mathfrak{A}_1 u_h + (I_N \otimes (I_N \otimes \mathfrak{B})])\, p_h = f_{1h} - \mathfrak{d}_{0h} \otimes u_{h0}, \tag{4.3.6}$$

where

$$\mathfrak{A}_1 = [\mathfrak{D}] \otimes I_{N(N-1)} + I_{N^2} \otimes [\![\mathfrak{D}^2]\!] + I_N \otimes [\![\mathscr{D}^2]\!] \otimes I_{N-1}.$$

Similarly, (4.3.5b) gives,

$$\mathfrak{A}_2 v_h + (I_N \otimes (\mathfrak{B} \otimes I_N)])\, p_h = f_{2h} - \mathfrak{d}_{0h} \otimes v_{h0}, \tag{4.3.7}$$

where

$$\mathfrak{A}_2 = [\mathfrak{D}] \otimes I_{N(N-1)} + I_{N(N-1)} \otimes [\![\mathscr{D}^2]\!] + I_N \otimes [\![\mathfrak{D}^2]\!] \otimes I_N.$$

Finally, eqs. (4.3.4) and (4.3.5c) for some $1 \leqslant m,n \leqslant N$, $m+n>2$, and $1 \leqslant r \leqslant N$

gives

$$\sum_{i=1}^{N-1} u_{inr}\ell_i'(\zeta_m) + \sum_{j=1}^{N-1} v_{mjr}\ell_j'(\zeta_n) = 0,$$

which in matrix form becomes

$$(I_N \otimes [(I_N \otimes \mathfrak{C})) u_h + (I_N \otimes [(\mathfrak{C} \otimes I_N)) v_h = O. \tag{4.3.8}$$

On using eqs. (4.3.6) to (4.3.8), we obtain the following *discrete unsteady Stokes problem*,

$$\begin{aligned}
\mathfrak{A}_1 u_h + (I_N \otimes (I_N \otimes \mathfrak{B})]) p_h &= f_{1h} - \mathfrak{d}_{0h} \otimes u_{h0}, \\
\mathfrak{A}_2 v_h + (I_N \otimes (\mathfrak{B} \otimes I_N)]) p_h &= f_{2h} - \mathfrak{d}_{0h} \otimes v_{h0}, \\
(I_N \otimes [(I_N \otimes \mathfrak{C})) u_h + (I_N \otimes [(\mathfrak{C} \otimes I_N)) v_h &= O.
\end{aligned}$$

The spectral convergence for this scheme is evident from Figure 4.3, which depicts the decay of error in L^∞ norm at the final time step $t_N = 1$ for this scheme is shown in Figure 4.3. We take f_1 and f_2, so that the exact solutions are the functions defined by (4.2.8). They satisfy the boundary conditions and produce initial conditions as $u(x, y, -1)$ and $v(x, y, -1)$. An efficient way of solving such linear systems is by implementing Uzawa algorithms and augmented Lagrangian Uzawa methods for solving saddle point problems.

4.3.2 The Navier-Stokes problem

Let us extend the *staggered grid collocation scheme* to the unsteady Navier-Stokes problem as follows.

$$(u_t + u\frac{\partial u}{\partial x} + v\frac{\partial u}{\partial y} - \Delta u + p_x)(x, y, t) = f_1(x, y, t), \qquad (x, y, t) \in \Xi_N^{(x)} \backslash \Xi_0^{(x)} \tag{4.3.9a}$$

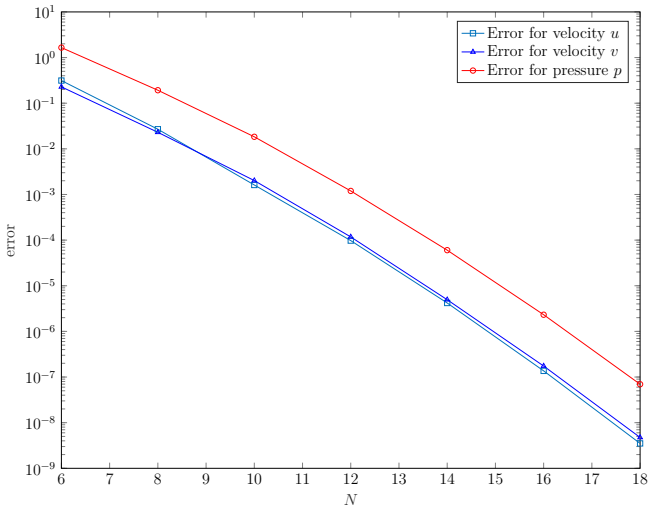


Figure 4.3: Convergence for the unsteady Stokes problem by staggered grid collocation scheme in space and collocation in time with $\alpha = -0.5$ and $\beta = 1.5$.

$$(v_t + u\frac{\partial v}{\partial x} + v\frac{\partial v}{\partial y} - \Delta v + p_y)(x, y, t) = f_2(x, y, t), \qquad (x, y, t) \in \Xi_N^{(y)} \backslash \Xi_0^{(y)}, \quad (4.3.9b)$$

$$(u_x + v_y)(x, y, t) = 0, \qquad (x, y, t) \in \Xi_N^{(p)}. \quad (4.3.9c)$$

Again, we only need to discuss the discretization of the non-linear term $(\mathbf{u} \cdot \nabla)\mathbf{u}$, which requires the following matrices:

1. The *Lobatto-Gauss bases interpolation matrix* $\mathfrak{U} \in \mathbb{R}^{N \times N-1}$ is defined as

$$\mathfrak{U}_{ij} = \ell_j(\zeta_i), \qquad 1 \leqslant i \leqslant N, 1 \leqslant j \leqslant N-1. \quad (4.3.10)$$

2. The *Gauss-Lobatto bases interpolation matrix* $\mathfrak{W} \in \mathbb{R}^{N-1 \times N}$ is defined as

$$\mathfrak{W}_{ij} = \mathfrak{p}_j(\xi_i), \qquad 1 \leqslant i \leqslant N-1, 1 \leqslant j \leqslant N. \quad (4.3.11)$$

Now, we discretize the four non-linear terms in (4.3.9).

Term 1: The first non-linear term in (4.3.9a) is $u\dfrac{\partial u}{\partial x}$. Its collocation on $(\xi_m, \zeta_n, \xi_r) \in \Xi_N^{(x)} \backslash \Xi_0^{(x)}$, that is, for $1 \leqslant m \leqslant N-1$, $1 \leqslant n \leqslant N$, and $1 \leqslant r \leqslant N$ along with

(4.3.2a) gives,

$$u_{mnr} \sum_{i=1}^{N-1} u_{inr} \ell_i'(\xi_m).$$

In matrix form, by Definition 4.3.1, the above equation is written as

$$\operatorname{diag}(u_h) \cdot (I_{N^2} \otimes [\![\mathfrak{D}]\!]) \,. \tag{4.3.12}$$

Term 2: The second non-linear term in (4.3.9a) is $v\dfrac{\partial u}{\partial y}$. Its collocation on $(\xi_m, \zeta_n, \xi_r) \in \Xi_N^{(x)} \backslash \Xi_0^{(x)}$, that is, for $1 \leqslant m \leqslant N-1$, $1 \leqslant n \leqslant N$, and $1 \leqslant r \leqslant N$ along with eqs. (4.3.2a) and (4.3.2b) gives,

$$\sum_{i=1}^{N} \sum_{j=1}^{N-1} v_{ijr} \mathfrak{p}_i(\xi_m) \ell_j(\zeta_n) \sum_{j=1}^{N} u_{mjr} \mathfrak{p}_j'(\zeta_n).$$

In matrix form, by Definition 4.3.2, eqs. (4.3.10) and (4.3.11), the above equation becomes

$$\operatorname{diag}\left((I_N \otimes \mathfrak{U} \otimes \mathfrak{W}) v_h\right) \cdot (I_N \otimes [\![\mathscr{D}]\!] \otimes I_{N-1}) \,. \tag{4.3.13}$$

Term 3: The first non-linear term in (4.3.9b) is $u\dfrac{\partial v}{\partial x}$. Its collocation on $(\zeta_m, \xi_n, \xi_r) \in \Xi_N^{(y)} \backslash \Xi_0^{(y)}$, that is, for $1 \leqslant m \leqslant N$, $1 \leqslant n \leqslant N-1$, and $1 \leqslant r \leqslant N$ along with eqs. (4.3.2a) and (4.3.2b) gives,

$$\sum_{i=1}^{N-1} \sum_{j=1}^{N} u_{ijr} \ell_i(\zeta_m) \mathfrak{p}_j(\xi_n) \sum_{i=1}^{N} v_{inr} \mathfrak{p}_i'(\zeta_m).$$

In matrix form, by Definition 4.3.2, eqs. (4.3.10) and (4.3.11), the above equation becomes

$$\operatorname{diag}\left((I_N \otimes \mathfrak{W} \otimes \mathfrak{U}) u_h\right) \cdot \left(I_{N(N-1)} \otimes [\![\mathscr{D}]\!]\right) \,. \tag{4.3.14}$$

Term 4: The second non-linear term in (4.3.9b) is $v\dfrac{\partial v}{\partial y}$. Its collocation on $(\zeta_m, \xi_n, \xi_r) \in \Xi_N^{(y)} \backslash \Xi_0^{(y)}$, that is, for $1 \leqslant m \leqslant N$, $1 \leqslant n \leqslant N-1$, and $1 \leqslant r \leqslant N$ along with (4.3.2b) gives,

$$v_{mnr} \sum_{j=1}^{N-1} v_{mjr} \ell_j'(\xi_n).$$

In matrix form, by Definition 4.3.1, the above equation becomes

$$\operatorname{diag}(v_h) \cdot (I_N \otimes [\![\mathfrak{D}]\!] \otimes I_N)\,. \tag{4.3.15}$$

Since the terms derived above are non-linear, we implement a simple fixed point iteration to set, thus (4.2.7) and the non-linear terms eqs. (4.3.12) to (4.3.15) yield the following *discrete unsteady Navier-Stokes problem*,

$$\begin{aligned}
\left(\mathfrak{N}_1^{(k)} + \mathfrak{A}_1\right) u_h^{(k+1)} + (I_N \otimes (I_N \otimes \mathfrak{B})])\, p_h^{(k+1)} &= f_{1h} - \mathfrak{d}_{0h} \otimes u_{h0}, \\
\left(\mathfrak{N}_2^{(k)} + \mathfrak{A}_2\right) v_h^{(k+1)} + (I_N \otimes (\mathfrak{B} \otimes I_N)])\, p_h^{(k+1)} &= f_{2h} - \mathfrak{d}_{0h} \otimes v_{h0}, \\
(I_N \otimes [(I_N \otimes \mathfrak{C})])\, u_h^{(k+1)} + (I_N \otimes [(\mathfrak{C} \otimes I_N)])\, v_h^{(k+1)} &= O,
\end{aligned}$$

where the matrices are defined as follows

$$\begin{aligned}
\mathfrak{A}_1 &= [\mathfrak{D}] \otimes I_{N(N-1)} + I_{N^2} \otimes [\![\mathfrak{D}^2]\!] + I_N \otimes [\![\mathscr{D}^2]\!] \otimes I_{N-1}, \\
\mathfrak{A}_2 &= [\mathfrak{D}] \otimes I_{N(N-1)} + I_{N(N-1)} \otimes [\![\mathscr{D}^2]\!] + I_N \otimes [\![\mathfrak{D}^2]\!] \otimes I_N, \\
\mathfrak{N}_1^{(k)} &= \operatorname{diag}(u_h^{(k)}) \cdot (I_{N^2} \otimes [\![\mathfrak{D}]\!]) + \operatorname{diag}\left((I_N \otimes \mathfrak{U} \otimes \mathfrak{W}) v_h^{(k)}\right) \cdot (I_N \otimes [\![\mathscr{D}]\!] \otimes I_{N-1})\,, \\
\mathfrak{N}_2^{(k)} &= \operatorname{diag}\left((I_N \otimes \mathfrak{W} \otimes \mathfrak{U}) u_h^{(k)}\right) \cdot \left(I_{N(N-1)} \otimes [\![\mathscr{D}]\!]\right) + \operatorname{diag}(v_h^{(k)}) \cdot (I_N \otimes [\![\mathfrak{D}]\!] \otimes I_N)\,.
\end{aligned}$$

The super-algerbaic decay of error in L^∞ norm at the final time step $t_N = 1$ for this scheme is shown in Figure 4.4. The iteration is stopped whenever the infinity norm of the difference of two consecutive iterates is smaller than $\epsilon = 10^{-12}$. We take f_1 and f_2, so that the exact solutions are functions defined by (4.2.8). They satisfy the

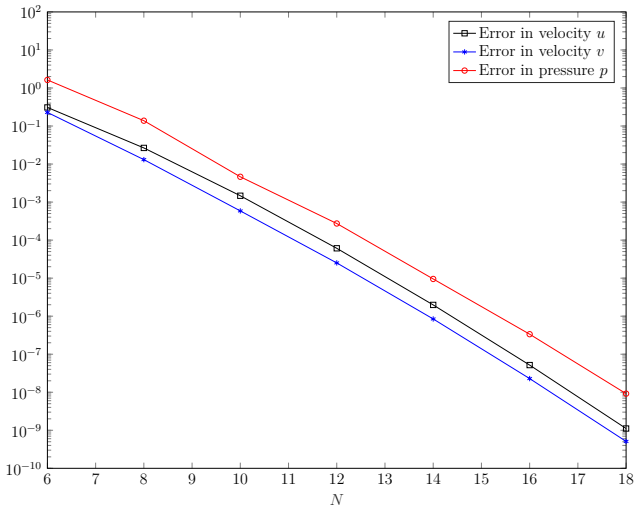


Figure 4.4: Convergence for the unsteady Navier-Stokes problem by staggered grid collocation scheme in space and collocation in time with $\alpha = \beta = 0$.

boundary conditions and yield initial conditions as $u(x, y, -1)$ and $v(x, y, -1)$. In addition to the fixed point iteration being used in case, we can implement other non-linear solvers such as the Newton's method, non-linear SOR, non-linear conjugate gradient, etc. The schemes derives in Sections 4.2 and 4.3 considered the Reynold's number to be equal to 1. It is of great interest to test the performance of these schemes for higher and lower Reynold's number flows.

5

New lower bounds on the minimum singular value

This chapter is a consequence of the analysis conducted in Chapter 3, in which estimates for eigenvalue or singular value were derived for some specific matrices. Such estimates enlightened us with the understanding of structure of matrices to derive more general results. The study of constraining the eigenvalues of the sum of two symmetric matrices, say $P + Q$, in terms of the eigenvalues of P and Q, has a long history. It is closely related to estimating a lower bound on the minimum singular value of a matrix, which has been discussed by a great number of authors. The question that originated the work presented in this chapter is a basic problem of classical linear algebra: "Can we derive a positive lower bound on the minimum eigenvalue, $\lambda_{\min}(P + Q)$, when $P + Q$ is symmetric positive definite with P and Q singular positive semi-definite?"

To the best of our knowledge, no study has yielded a positive lower bound on $\lambda_{\min}(P + Q)$. According to Sophie Germain, "Algebra is but written geometry and geometry is but figured algebra." One approach that provides an answer to the question mentioned above is a geometrical property of fundamental bases of linear

algebra, the Friedrichs angle between range spaces of matrices P and Q. It aids us to formulate new lower bounds on the minimum eigenvalue of a symmetric positive definite (SPD) matrix.

We derive two new lower bounds on $\lambda_{\min}(P+Q)$ in terms of the minimum positive eigenvalues of P and Q. The basic result is when P and Q are two non-zero singular positive semi-definite matrices such that $P+Q$ is non-singular, then $\lambda_{\min}(P+Q) \geqslant (1-\cos\theta_F)\min\{\lambda_{\min}(P), \lambda_{\min}(Q)\}$, where $\lambda_{\min}$ represents the minimum positive eigenvalue of the matrix, and θ_F is the Friedrichs angle between the range spaces of P and Q. Such estimates lead to new lower bounds on the minimum singular value of full rank 1×2, 2×1, and 2×2 block matrices. We provide some examples to further highlight the simplicity of applying the results in comparison to some existing lower bounds.

5.1 Introduction

The *spectral problem of a symmetric matrix sum* estimates the eigenvalues of a sum of two symmetric matrices $P+Q$, in terms of the eigenvalues of P and Q. Fundamental results, like Weyl's inequality in [44, p. 239], and several other works collected in [25], have addressed this problem. Another substantial contribution is Horn's conjecture proved in [55, 56]. The present work is focused on the case when P and Q are symmetric positive semi-definite (PSD) matrices, which impacts numerous areas-such as computational economics, graph theory, perturbation theory, semi-definite programming, spectrum of self-adjoint operators, among others. As variance-covariance matrices are PSD, this problem appears in statistics, and more recently in statistical machine learning and spectral methods for data science, discussed in [6] and [21], respectively.

Singular values have been investigated for more than a century. For a square real

matrix, its minimum singular value is less than or equal to its absolute minimum eigenvalue. Thus, formulation of a lower bound for the minimum singular value is an influential problem appearing in several studies including the condition number estimates of a matrix, resonant frequencies, population dynamics, principal component analysis, etc. Since singular values of a matrix are square-root of eigenvalues of its corresponding gram matrix, the singular values of a general block matrix are associated to the spectral problem of a sum of symmetric matrices. Although a myriad of research has been done on these topics, however, when the symmetric matrices are both singular PSD, we could not find a result providing a positive lower bound even if their sum is non-singular.

In practice, we often come across symmetric positive definite (SPD) matrices represented as a sum of two singular PSD matrices. To illustrate, let us estimate the minimum singular value of a full rank block column matrix, say $A = \left[\begin{smallmatrix} A_1 \\ A_2 \end{smallmatrix}\right]$ so that A_1 and A_2 are rank deficient. This problem is equivalent to finding the minimum eigenvalue of $A^TA = A_1^TA_1 + A_2^TA_2$, an SPD matrix which is a sum of two singular PSD matrices. We derive a positive lower bound on the minimum singular value of A in terms of the minimum positive singular values of A_1 and A_2 in Corollary 5.6.1. A similar problem was encountered in Lemma 3.3.7.

In this work, we desire a positive lower bound on the minimum eigenvalue of an SPD matrix $P + Q$, where $P, Q \in \mathbb{R}^{n\times n}$ are PSD matrices. Two positive lower bounds on the smallest eigenvalue of $P + Q$, framed in terms of the smallest positive eigenvalues of P and Q, are presented in Theorems 5.4.1 and 5.4.4. These estimates of the minimum eigenvalue employ the Friedrichs angle between certain subspaces, i.e., some principal angle between them, as shown in Proposition 2.2.5. A notable application of principal angles is canonical correlations of matrix pairs given in [35], and in many other areas, namely eigenspaces, functional analysis, matrix perturbation theory, statistics, etc., are found in [57, 86, 26, 23], respectively. The spectral

problem of a sum of two PSD matrices is closely related to the first aforementioned application of canonical correlations, and its dependence upon the Friedrichs angle elucidates geometric aspects of spectral theory. Moreover, the two new lower bounds lead to useful outcomes when applied to a 2×2 non-singular block matrix $M = \left[\begin{smallmatrix} A & B \\ C & D \end{smallmatrix}\right]$. Note that $M^T M$ can be calculated as follows,

$$M^T M = \begin{bmatrix} A^T & C^T \\ B^T & D^T \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A^T A & A^T B \\ B^T A & B^T B \end{bmatrix} + \begin{bmatrix} C^T C & C^T D \\ D^T C & D^T D \end{bmatrix}.$$

Also, $M^T M$ is a full rank matrix expressed as a sum of two PSD matrices. Therefore, the above expression admits a lower bound on the minimum singular value of M, in terms of the minimum positive singular values of its blocks A, B, C, and D (Theorem 5.6.3). One of these results were used to solve Theorem 3.4.5. Finally, the above expression and MM^T are used again to get two lower bounds on other singular values of M in Theorem 5.7.2.

This chapter is organized as follows. A brief survey of results related to the ones derived in this chapter is presented in Section 5.2, and Section 5.3 describes the notation and fundamental results used later. In Section 5.4, we prove some new positive lower bounds on the minimum eigenvalue of a matrix and discuss the origin of those estimates with the help of some theory and examples in Section 5.5. In Section 5.6, the new eigenvalue estimates are used to derive new lower bounds on the minimum singular value of some full rank block matrices. In addition, some examples and special cases for these results are discussed. Finally, estimates for some other singular values are mentioned in Section 5.7.

5.2 Literature review

There is an abundance of results related to these problems in the literature. We attempt to summarize some of the existing focal results related to the ones developed in Sections 5.4, 5.6 and 5.7.

5.2.1 Minimum eigenvalue of sum of two PSD matrices

The problem of estimating a lower bound on the minimum eigenvalue of sum of two PSD matrices has been investigated for many years. One of the most fundamental results is Weyl's inequalities given by Theorem 2.1.6 and more generally by Theorem 5.3.3, which estimates the eigenvalues of a sum of symmetric matrices.

R. Bhatia and F. Kittaneh established a lower bound on eigenvalues of sum of two PSD matrices (Theorem 5.3.5). They posed the question of a generalization of *arithmetic-geometric mean inequality* in [16], which stated that for two PSD matrices $P, Q \in \mathbb{R}^{n\times n}$, $\lambda_j(P+Q) \geqslant 2\sqrt{\sigma_j(PQ)}$, for all $j = 1, 2, \ldots, n$. It was proved by S. Drury in [27]. However, all these results give a trivial lower bound for the case in which $P+Q$ is non-singular, and both P and Q are rank deficient. Some additional properties for sum of two symmetric matrices are listed in [59].

5.2.2 Spectrum of saddle point matrices

One of the most commonly seen 2×2 block matrices of the form $M = \left[\begin{smallmatrix} A & B_1 \\ B_2^T & -C \end{smallmatrix}\right]$, where $A \in \mathbb{R}^{m\times m}$ and one or both of $B_1, B_2 \in \mathbb{R}^{m\times n}$ are non-zero, is called a *saddle point matrix*. See [8], for a good survey of results on saddle point matrices. In particular, see Theorem 3.5 in [8, p. 21], which estimates the spectrum for the case when A is SPD, $B_1 = B_2$ is full rank, and $C = O$. A noteworthy improvement was presented in the form of Theorem 1 in [4, p. 341] with a positive or negative semi-definite matrix C, also mentioned as Theorem 3.3.8. However, it can be difficult to estimate the

parameters defined in this theorem while being applied to some gram matrix of 2×2 block matrix. Another applicable result for the spectrum of a preconditioned saddle point matrix can be found in [83].

To formulate the spectrum of a more generalized saddle point matrix, several advancements have been considered, such as defining $B = B_1^T = -B_2$ in [9, 10, 82, 2]. Another step forward was to have a symmetric indefinite leading block A. The first of such case was proved in [37], by imposing the condition that A is SPD on $\mathcal{N}(B)$, which was eliminated in [5]. Recently, in [45], A has been considered to be a non-symmetric matrix with a positive definite symmetric part with $C = O$, which originates from discretized Navier-Stokes equations.

5.2.3 Lower bound on the minimum singular value

Several techniques are reported in the literature for formulating a lower bound on the minimum singular value of a particular type of matrices; however, we attempt to mention seminal contributions to this problem for a general non-singular matrix. An initial result for the special case of diagonally dominant matrices is derived in [96], and for a non-singular matrix a consequential approach is Gerschgorin-type lower bounds formulated in [76, 48]. The results evolved gradually into several stronger versions, as seen in [49, 60, 50, 105]. Also, [42] devised a lower bound in terms of the determinant, the 2-norms of the rows, and columns of the matrix. Some later advancements of this result include [102, 104, 61] and the references therein.

It is well-known that for a 2×2 block matrix, the maximum singular value is bounded above by 2-norm of the matrix consisting of 2-norms of its blocks, see Theorem 1(f) in [85, p. 2630]. Thus, for $\mathcal{X} \in \mathbb{R}^{n \times n}$,

$$\mathcal{X} = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \qquad \|\mathcal{X}\| \leqslant \left\| \begin{bmatrix} \|A\| & \|B\| \\ \|C\| & \|D\| \end{bmatrix} \right\|.$$

Since $\sigma_{\min}(\mathcal{X}) = \sigma_{\max}(\mathcal{X}^{-1})^{-1}$, so on applying this result to $\mathcal{X}^{-1}$ calculated in terms of its blocks, an estimate of $\sigma_{\min}(\mathcal{X})$ is obtained. One drawback of this method is that the expression for $\mathcal{X}^{-1}$ can be quite problematic.

Another estimate was given in [97] for a block matrix with non-singular diagonal blocks, which proved by using a block matrix technique that $\sigma_{\min}(\mathcal{X}) \geqslant \sigma_2(\mathcal{X}^c)$, where $\mathcal{X}^c$ represents the block comparison matrix of $\mathcal{X}$, which results in a trivial lower bound when $\mathcal{X}^c$ is singular.

5.3 Notations and fundamentals

We now summarize the notation used in this paper. Let $A \in \mathbb{R}^{n\times n}$ be a symmetric positive semi-definite (PSD) matrix and recall that $\Lambda(A)$ denote the spectrum of A, that is, the set of eigenvalue of A. Also, $\rho(A)$ denote the spectral radius of A. If $r = \text{rank}(A)$, then its eigenvalues $\lambda_1(A) \geqslant \lambda_2(A) \geqslant \ldots \geqslant \lambda_n(A)$, are such that $\lambda_r > 0$ and $\lambda_i = 0$ for all $r + 1 \leqslant i \leqslant n$.

Definition 5.3.1 (Minimum positive eigenvalue)**.** Let $A \in \mathbb{R}^{n\times n}$ be a PSD matrix of rank $r \leqslant n$, define the minimum positive eigenvalue of A as

$$\lambda_{\min}(A) := \begin{cases} \lambda_r(A), & \text{if } A \neq O, \\ \infty, & \text{if } A = O. \end{cases}$$

For a matrix $A \in \mathbb{R}^{m\times n}$, recall that $\text{rank}(A) = \text{rank}(A^TA) = \text{rank}(AA^T)$. The set of singular values of A is denoted by $\sigma(A)$. Let $r = \text{rank}(A)$, then its singular values $\sigma_1(A) \geqslant \sigma_2(A) \geqslant \ldots \geqslant \sigma_{\min(m,n)}(A)$ are such that $\sigma_r(A) > 0$ and $\sigma_i = 0$ for all $r + 1 \leqslant i \leqslant \min(m, n)$.

Definition 5.3.2 (Minimum positive singular value)**.** Let $A \in \mathbb{R}^{m\times n}$ of rank $r \leqslant n$,

define the minimum positive singular value of A as

$$\sigma_{\min}(A) := \begin{cases} \sigma_r(A), & \text{if } A \neq O, \\ \infty, & \text{if } A = O. \end{cases}$$

The above expressions for $\lambda_{\min}$ and $\sigma_{\min}$ are defined for the convenience of notation for results derived in the next sections and their value is set as infinity for zero matrices to ignore the zeros while calculating the minimum as it was required by the formulations derived for zero matrices.

The *Weyl's inequalities* are one the most consequential result providing a solution to the *spectral problem of a symmetric matrix sum.* s

Theorem 5.3.3 (Weyl's inequalities, see [43])**.** *Let $A, B \in \mathbb{R}^{n\times n}$ be symmetric matrices, then for every pair of integers j, k such that $1 \leqslant j, k \leqslant n$ and $j + k \leqslant n + 1$,*

$$\lambda_{j+k-1}(A + B) \leqslant \lambda_j(A) + \lambda_k(B),$$

and for every pair integers such that $1 \leqslant j, k \leqslant n$ and $j + k \geqslant n + 1$,

$$\lambda_{j+k-n}(A + B) \geqslant \lambda_j(A) + \lambda_k(B).$$

Another result which is significant to our analysis is the *spectrum of product* of two rectangular matrices, given in [103, p. 57] and stated as follows.

Lemma 5.3.4 (Spectrum of product of matrices, see [103])**.** *Let $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{n\times m}$. Then AB and BA have the same non-zero eigenvalues (multiplicities counted). If $m = n$, then eigenvalues of AB and BA are the same.*

Finally, we state the following estimate on some eigenvalues of sum of two PSD matrices is given in [14, p. 904] and [15]. It will be used to give a simpler proof for lower bound on some singular values.

Theorem 5.3.5 (See [15]). *Let $A \in \mathbb{R}^{m\times n}$ and $B \in \mathbb{R}^{\ell\times n}$, then*

$$2\sigma_j(AB^T) \leqslant \lambda_j(A^TA + B^TB), \qquad j = 1, 2, \ldots, \min(\ell, m, n).$$

In addition to the results mentioned in the previous section, we need Theorems 2.1.7, 2.1.8 and 2.2.3, Definitions 2.2.1 and 2.2.4, and Propositions 2.2.2 and 2.2.5. An alternative proof for two of these results are given in Appendix A.

5.4 Minimum eigenvalue estimates

In this section, we derive some new lower bounds on the *minimum eigenvalue* of a non-singular sum of two singular PSD matrices. As discussed in Section 5.1, a positive lower bound on the minimum eigenvalue of a non-singular sum of two PSD matrices, say $P, Q \in \mathbb{R}^{n\times n}$, is the key tool for the development of a positive lower bound on the minimum singular value of some full rank block matrices. Note that $\mathcal{N}(P) \cap \mathcal{N}(Q) = \{0\}$ when $P + Q$ is SPD, however, the range spaces of P and Q may intersect. Let $k = \dim(\mathcal{R}(P) \cap \mathcal{R}(Q))$, then the first k principal angles between $\mathcal{R}(P)$ and $\mathcal{R}(Q)$ vanish: $\theta_i = 0$ for $i = 1, 2, \ldots, k$. Therefore, if θ_{k+1} exists then it could contribute in estimating the minimum eigenvalue of $P + Q$ in terms of the minimum positive eigenvalues of P and Q. Even when θ_{k+1} does not exist, this idea serves as a motivation for the following theorem for a pair of *two PSD matrices with a non-singular sum.*

Theorem 5.4.1. *Let $P, Q \in \mathbb{R}^{n\times n}$ be PSD matrices of rank $p, q \leqslant n$, respectively, so that $P + Q$ is non-singular. Then*

$$\lambda_{\min}(P + Q) \geqslant c(P, Q) \min\left\{\lambda_{\min}(P), \lambda_{\min}(Q)\right\},$$

where $c(P,Q)$ is defined by

$$c(P,Q) = \begin{cases} 2, & r = 0,\ p+q = 2n, \\ 1, & r = 0,\ p+q < 2n, \\ 1-\cos(\theta_{k+1}), & r > 0, \end{cases} \tag{5.4.1}$$

where $k = p+q-n$, $0 \leqslant \theta_1 \leqslant \theta_2 \leqslant \ldots \leqslant \theta_{\min(p,q)} \leqslant \frac{\pi}{2}$ represent the principal angles between $\mathcal{R}(P), \mathcal{R}(Q) \subseteq \mathbb{R}^n$, and r is the number of angles θ_i so that $0 < \theta_i < \frac{\pi}{2}$, for $1 \leqslant i \leqslant \min(p,q)$.

Proof. Since P, Q are PSD matrices, there exist matrices $\mathscr{A}_1 \in \mathbb{R}^{p\times n}$ and $\mathscr{A}_2 \in \mathbb{R}^{q\times n}$, so that

$$P = \mathscr{A}_1^T \mathscr{A}_1, \qquad Q = \mathscr{A}_2^T \mathscr{A}_2.$$

Moreover, $\mathcal{N}(P) = \mathcal{N}(\mathscr{A}_1)$ and $\mathcal{N}(Q) = \mathcal{N}(\mathscr{A}_2)$. Define $M_1 := \mathcal{R}(P) = \mathcal{R}(P^T) = \mathcal{N}(P)^\perp = \mathcal{N}(\mathscr{A}_1)^\perp$ and similarly define $M_2 := \mathcal{R}(Q) = \mathcal{N}(\mathscr{A}_2)^\perp$. Let $P_i \in \mathbb{R}^{n\times n}$ be the orthogonal projection on M_i, for $i = 1, 2$. Therefore, $\mathcal{R}(P_i) = M_i$ and $\mathcal{R}(I - P_i) = M_i^\perp = \mathcal{N}(\mathscr{A}_i)$ for $i = 1, 2$. The variational characterization of the smallest eigenvalue of a symmetric matrix implies

$$\lambda_{\min}(P+Q) = \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{x^T\mathscr{A}_1^T\mathscr{A}_1 x + x^T\mathscr{A}_2^T\mathscr{A}_2 x}{|x|^2}.$$

Since any $x \in \mathbb{R}^n\setminus\{0\}$ can be represented as $x = (I-P_i)x + P_i x$, for $i = 1, 2$, thus

$$\begin{aligned} x^T\mathscr{A}_i^T\mathscr{A}_i x &= \left[\mathscr{A}_i((I-P_i)x + P_i x)\right]^T \left[\mathscr{A}_i((I-P_i)x + P_i x)\right] \\ &= (\mathscr{A}_i P_i x)^T (\mathscr{A}_i P_i x) \qquad\qquad (\text{as } (I-P_i)x \in \mathcal{N}(\mathscr{A}_i)) \\ &= (P_i x)^T \mathscr{A}_i^T \mathscr{A}_i (P_i x), \end{aligned}$$

therefore,

$$\lambda_{\min}(P+Q) = \inf_{x\in\mathbb{R}^n\backslash\{0\}} \frac{(P_1x)^T\mathscr{A}_1^T\mathscr{A}_1(P_1x) + (P_2x)^T\mathscr{A}_2^T\mathscr{A}_2(P_2x)}{|x|^2}. \tag{5.4.2}$$

Note that the minimum positive eigenvalue of $\mathscr{A}_i^T\mathscr{A}_i$ is identified by the variational characterization as follows,

$$\lambda_{\min}(\mathscr{A}_i^T\mathscr{A}_i) = \inf_{x\in\mathcal{N}(\mathscr{A}_i)^\perp} \frac{x^T\mathscr{A}_i^T\mathscr{A}_i x}{|x|^2}.$$

Since for any $x \in \mathbb{R}^n\backslash\{0\}$, $P_ix \in M_i = \mathcal{N}(\mathscr{A}_i)^\perp$, therefore the above expression gives

$$(P_ix)^T\mathscr{A}_i^T\mathscr{A}_i(P_ix) \geqslant \lambda_{\min}(\mathscr{A}_i^T\mathscr{A}_i)|P_ix|^2,$$

hence (5.4.2) provides the following estimate

$$\begin{aligned}
\lambda_{\min}(P+Q) &\geqslant \inf_{x\in\mathbb{R}^n\backslash\{0\}} \frac{\lambda_{\min}(\mathscr{A}_1^T\mathscr{A}_1)|P_1x|^2 + \lambda_{\min}(\mathscr{A}_2^T\mathscr{A}_2)|P_2x|^2}{|x|^2} \\
&= \inf_{x\in\mathbb{R}^n\backslash\{0\}} \frac{\lambda_{\min}(P)|P_1x|^2 + \lambda_{\min}(Q)|P_2x|^2}{|x|^2} \qquad (5.4.3) \\
&\geqslant \min\{\lambda_{\min}(P), \lambda_{\min}(Q)\} \inf_{x\in\mathbb{R}^n\backslash\{0\}} \frac{|P_1x|^2 + |P_2x|^2}{|x|^2} \\
&=: \min\{\lambda_{\min}(P), \lambda_{\min}(Q)\} \inf_{x\in\mathbb{R}^n\backslash\{0\}} \Delta(x), \qquad (5.4.4)
\end{aligned}$$

where $\Delta(x) := \dfrac{|P_1x|^2 + |P_2x|^2}{|x|^2}$, for $x \in \mathbb{R}^n\backslash\{0\}$. Since the parallelogram identity for inner-product spaces states that

$$|P_1x|^2 + |P_2x|^2 = \frac{1}{2}\left[|(P_1+P_2)x|^2 + |(P_1-P_2)x|^2\right], \tag{5.4.5}$$

leading to a lower bound,

$$\inf_{x\in\mathbb{R}^n\backslash\{0\}} \Delta(x) \geqslant \frac{1}{2}\left[\sigma_{\min}^2(P_1+P_2) + \sigma_{\min}^2(P_1-P_2)\right], \tag{5.4.6}$$

hence the set of singular values of $P_1 \pm P_2$ need to be analyzed. To this end, note that $M_1^{\perp} \cap M_2^{\perp} = \mathcal{N}(P) \cap \mathcal{N}(Q) = \{0\}$, as

$$x \in \mathcal{N}(P) \cap \mathcal{N}(Q) \quad \Leftrightarrow \quad (P+Q)x = 0 \quad \Leftrightarrow \quad x = 0,$$

since $P + Q$ is non-singular. Also, Proposition 2.1.1 gives

$$M_1 + M_2 = (M_1^{\perp} \cap M_2^{\perp})^{\perp} = \{0\}^{\perp} = \mathbb{R}^n,$$

consequently, Theorem 2.2.3 gives the set of singular values of $P_1 \pm P_2$ as

$$\begin{aligned} \sigma(P_1 + P_2) &= \{\mathbf{2}_k, 1 \pm \cos(\theta_{k+i})(i = 1, \ldots, r), \mathbf{1}_{n_1+n_2}\}, \\ \sigma(P_1 - P_2) &= \{\mathbf{1}_{n_1+n_2}, \mathbf{sin(\theta_{k+i})}_2(i = 1, \ldots, r), \mathbf{0}_k\}, \end{aligned} \tag{5.4.7}$$

where,

$$\begin{aligned} k &= \dim(M_1 \cap M_2) = \dim(M_1) + \dim(M_2) - \dim(M_1 + M_2) = p + q - n, \\ n_1 &= \dim(M_1 \cap M_2^{\perp}) = p - k - r, \\ n_2 &= \dim(M_1^{\perp} \cap M_2) = q - k - r, \\ n_3 &= \dim(M_1^{\perp} \cap M_2^{\perp}) = 0, \\ n &= n_1 + n_2 + k + 2r = p + q - k. \end{aligned} \tag{5.4.8}$$

Let us estimate (5.4.6), thus (5.4.4) in terms of the following cases.

Case 1: Suppose $r = 0$. Note that (5.4.8) implies $n_1 = p - k$, $n_2 = q - k$ for this case. Firstly, consider $k = 0$ and $n_1 = n_2 = 0$, then (5.4.8) implies $p = q = 0$, thus $P = Q = O$. This is rejected, since $P + Q$ is non-singular.

For $k = 0$ and $n_1 + n_2 > 0$, (5.4.7) and (5.4.8) yield $\sigma(P_1 \pm P_2) = \{\mathbf{1}_{n_1+n_2}\} = \{\mathbf{1}_n\}$, hence (5.4.6) gives

$$\inf_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) \geqslant \frac{1}{2}\left[1^2 + 1^2\right] = 1. \tag{5.4.9}$$

For $k > 0$ and $n_1 = n_2 = 0$, (5.4.8) implies that $k = p = q = n$ or both P and Q are non-singular. Therefore, $M_1 = M_2 = \mathbb{R}^n$, hence (5.4.4) gives

$$\inf_{x\in\mathbb{R}^n\setminus\{0\}} \Delta(x) = \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|x|^2 + |x|^2}{|x|^2} = 2. \tag{5.4.10}$$

For $k, n_1 > 0$, and $n_2 = 0$, (5.4.8) results in $p = n$ and $q = k$, that is, $M_2 \subseteq M_1 = \mathbb{R}^n$ or Q is non-singular. Hence, (5.4.4) becomes

$$\inf_{x\in\mathbb{R}^n\setminus\{0\}} \Delta(x) = \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|x|^2 + |P_2 x|^2}{|x|^2} \geqslant 1.$$

Similarly, for $k, n_2 > 0$ and $n_1 = 0$, the same lower bound as above is derived which also coincides with (5.4.9). Finally, consider $k, n_1, n_2 > 0$, since $\dim(M_1 \cap M_2) = k > 0$, therefore $M_1 \cap M_2$ is a non-trivial subspace. Thus, by Lemma A.1.1, $\mathcal{N}(P_1 - P_2) = M_1 \cap M_2$ is non-trivial implying $P_1 - P_2$ is singular which will give a weaker result. In order to improve it, we define $M_3 := M_2 \cap (M_1 \cap M_2)^\perp$, then

$$\begin{aligned}
\dim(M_3) &= n - \dim(M_3^\perp) \\
&= n - \dim(M_2^\perp + M_1 \cap M_2) && \text{(by Proposition 2.1.1)} \\
&= n - (n - \dim(M_2)) - \dim(M_1 \cap M_2) \\
&= n - (n - q) - k = q - k,
\end{aligned}$$

or $\dim(M_3) = q - k = n_2 > 0$, thus it is a non-trivial subspace. Let P_3 and P_U be the orthogonal projections onto the subspace M_3 and some subspace U of $\mathbb{R}^n$, respectively. The following was proved in [20, p. 1429],

$$\begin{aligned}
P_3 &= P_{M_2 \cap (M_1 \cap M_2)^\perp} \\
&= P_{M_2} P_{(M_1 \cap M_2)^\perp} \\
&= P_{M_2} \left(I - P_{M_1 \cap M_2}\right)
\end{aligned}$$

$$= P_{M_2} - P_{M_2} P_{M_1 \cap M_2}$$
$$= P_2 - P_{M_1 \cap M_2},$$

or $P_2 = P_{M_1 \cap M_2} + P_3$, which implies for any $x \in \mathbb{R}^n \backslash \{0\}$,

$$P_2 x = P_{M_1 \cap M_2} x + P_3 x. \tag{5.4.11}$$

By definition, $M_1 \cap M_2$ and M_3 are mutually orthogonal subspaces. Therefore, for any $x \in \mathbb{R}^n \backslash \{0\}$,

$$|P_2 x|^2 = |P_{M_1 \cap M_2} x|^2 + |P_3 x|^2, \tag{5.4.12}$$

and (5.4.4) becomes

$$\begin{aligned}
\inf_{x \in \mathbb{R}^n \backslash \{0\}} \Delta(x) &= \inf_{x \in \mathbb{R}^n \backslash \{0\}} \frac{|P_1 x|^2 + |P_{M_1 \cap M_2} x|^2 + |P_3 x|^2}{|x|^2} \\
&\geqslant \inf_{x \in \mathbb{R}^n \backslash \{0\}} \frac{|P_1 x|^2 + |P_3 x|^2}{|x|^2} \\
&= \frac{1}{2} \inf_{x \in \mathbb{R}^n \backslash \{0\}} \frac{|(P_1 + P_3)x|^2 + |(P_1 - P_3)x|^2}{|x|^2} \qquad \text{(by (5.4.5))} \\
&\geqslant \frac{1}{2} \left[\sigma_{\min}^2 (P_1 + P_3) + \sigma_{\min}^2 (P_1 - P_3) \right]. \qquad (5.4.13)
\end{aligned}$$

By Theorem 2.2.3,

$$\begin{aligned}
\sigma(P_1 + P_3) &= \{ \mathbf{2}_{\tilde{k}}, 1 \pm \cos(\alpha_{\tilde{k}+i}) (i = 1, \ldots, r), \mathbf{1}_{\tilde{n}_1 + \tilde{n}_2}, 0_{\tilde{n}_3} \}, \\
\sigma(P_1 - P_3) &= \{ \mathbf{1}_{\tilde{n}_1 + \tilde{n}_2}, \mathbf{sin}(\boldsymbol{\alpha}_{\tilde{\boldsymbol{k}}+\boldsymbol{i}})_2 (i = 1, \ldots, r), \mathbf{0}_{\tilde{k}+\tilde{n}_3} \},
\end{aligned} \tag{5.4.14}$$

where $0 \leqslant \alpha_1 \leqslant \alpha_2 \leqslant \ldots \leqslant \alpha_{\min(p,q-k)} \leqslant \frac{\pi}{2}$ represent the principal angles between the subspaces M_1 and M_3, and r is the number of principal angles that satisfy $0 < \alpha_i < \frac{\pi}{2}$. Observe that r is the same number for M_1 and M_2 by (2.2.1), or see

[32, p. 231] for more details. Thus, $r = 0$ and (5.4.8) gives the following parameters

$$\begin{aligned}
\tilde{k} &= \dim(M_1 \cap M_3) = \dim(M_1 \cap (M_2 \cap (M_1 \cap M_2)^{\perp})) = 0, \\
\tilde{n}_1 &= p - \tilde{k} - r = p, \\
\tilde{n}_2 &= (q-k) - \tilde{k} - r = q - k, \\
\tilde{n_3} &= n - p - (q-k) + \tilde{k} = n - (p + q - k) = 0.
\end{aligned} \tag{5.4.15}$$

Hence, $\sigma_{\min}(P_1 \pm P_3) = 1$, therefore (5.4.13) implies that $\inf\limits_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) \geqslant 1$, which coincides with (5.4.9).

In conclusion, for $r = 0$ and $p + q < 2n$, $\lambda_{\min}(P+Q) \geqslant \min\{\lambda_{\min}(P), \lambda_{\min}(Q)\}$ holds, or $c(P,Q) = 1$. Whereas, for $r = 0$ and $p + q = 2n$, $\lambda_{\min}(P+Q) \geqslant 2\min\{\lambda_{\min}(P), \lambda_{\min}(Q)\}$ holds, or $c(P,Q) = 2$.

Case 2: Suppose $r > 0$. For $k = 0$ and any $n_1, n_2 \geqslant 0$, (5.4.7) implies that $\sigma_{\min}(P_1 + P_2) = 1 - \cos\theta_1$ and $\sigma_{\min}(P_1 - P_2) = \sin\theta_1$. By (5.4.6),

$$\begin{aligned}
\inf_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) &\geqslant \frac{1}{2}\left[(1 - \cos\theta_1)^2 + \sin^2\theta_1\right] \\
&= 1 - \cos\theta_1.
\end{aligned}$$

Therefore, (5.4.4) gives $\lambda_{\min}(P+Q) \geqslant (1 - \cos\theta_1)\min\{\lambda_{\min}(P), \lambda_{\min}(Q)\}$.

For $k > 0$ and $n_1, n_2 \geqslant 0$, thus $M_1 \cap M_2$ is non-trivial. Consider the subspaces M_1 and M_3, as defined earlier for $k, n_1, n_2 > 0$ in Case 1. Note that M_3 is always non-trivial, as $n_1, n_2 \geqslant 0$, (5.4.8) implies $p \geqslant k + r$ and $q \geqslant k + r$, thus $\dim M_3 = q - k \geqslant r > 0$. Also, the set of parameters for M_1 and M_3 are given by (5.4.15) and

$r > 0$ as follows,

$$\begin{aligned} \tilde{k} &= \dim(M_1 \cap M_3) = 0, \\ \tilde{n}_1 &= p - \tilde{k} - r = p - r, \\ \tilde{n}_2 &= (q-k) - \tilde{k} - r = q - k - r, \\ \tilde{n}_3 &= n - p - (q-k) + \tilde{k} = 0. \end{aligned} \tag{5.4.16}$$

Let θ_F be the Friedrichs angle between M_1 and M_2, then by Property 3 of Proposition 2.2.5, it is equal to the minimal angle between M_1 and M_3 in (5.4.14), that is,

$$\alpha_1 = \theta_F = \theta_{k+1}, \tag{5.4.17}$$

where the last equality follows by Property 5 of Proposition 2.2.5. Therefore, (5.4.14) implies that $\sigma_{\min}(P_1 + P_3) = 1 - \cos\alpha_1 = 1 - \cos\theta_{k+1}$ and $\sigma_{\min}(P_1 - P_3) = \sin\alpha_1 = \sin\theta_{k+1}$, so (5.4.13) yields

$$\begin{aligned} \inf_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) &\geqslant \frac{1}{2}\left[(1-\cos\theta_{k+1})^2 + \sin^2\theta_{k+1}\right] \\ &= 1 - \cos\theta_{k+1}. \end{aligned} \tag{5.4.18}$$

Thus, (5.4.4) implies $\lambda_{\min}(P+Q) \geqslant (1-\cos\theta_{k+1}) \min\{\lambda_{\min}(P), \lambda_{\min}(Q)\}$, or $c(P,Q) = 1 - \cos\theta_{k+1}$, which is consistent for $r > 0$ and $k = 0$. □

Remark 5.4.2. Recall that θ_{k+1} is the *Friedrichs angle* between $\mathcal{R}(P)$ and $\mathcal{R}(Q)$ as stated in Property 5 of Proposition 2.2.5. It can be easily calculated by using a result by A. Björck and G. Golub given in [17]. Let $Q_1 \in \mathbb{R}^{n\times p}$ and $Q_2 \in \mathbb{R}^{n\times q}$ represent a orthogonal bases for $\mathcal{R}(P)$ and $\mathcal{R}(Q)$, respectively. Define $M = Q_1^T Q_2$, then $\cos\theta_k = \sigma_k(M)$, for $k = 1, 2, \ldots, q$. Consider the reduced SVD of $M = Y\Sigma Z^T$, then $\Sigma = \operatorname{diag}\{\sigma_1, \sigma_2, \ldots, \sigma_q\}$ gives the principal angles between $\mathcal{R}(A)$ and $\mathcal{R}(B)$, moreover, the principal vectors (u_k, v_k) corresponding to the principal angles θ_k are given by the columns of the matrices $U = Q_1 Y$ and $V = Q_2 Z$.

Definition of the measure $c(P, Q)$, given by (5.4.1), is also applicable to rectangular matrices as stated below.

Proposition 5.4.3. *For distinct non-zero matrices $A \in \mathbb{R}^{n\times p}$, $B \in \mathbb{R}^{n\times q}$*

1. $c(A, B) = c(AA^T, BB^T)$.
2. $c(A, B) = c(B, A)$.
3. $c(A, O_{n,k}) = 1$.
4. $c(A, B) = 1 - \cos\theta_F$, *when both A, B are rank-deficient, where θ_F is the Friedrichs angle between $\mathcal{R}(A)$ and $\mathcal{R}(B)$.*

The following result is derived to complete the analysis of $\lambda_{\min}(P + Q)$, the *two PSD matrices with a non-singular sum*. The new result reduces to Theorem 2.1.6 when at least one of P and Q is SPD. However, the result may be weaker or stronger than the estimate given by Theorem 5.4.1.

Theorem 5.4.4. *Let $P, Q \in \mathbb{R}^{n\times n}$ be PSD matrices of rank $p, q \leqslant n$, respectively, so that $P + Q$ is non-singular, then*

$$\lambda_{\min}(P + Q) \geqslant \psi(P, Q),$$

where, for $r = 0$

$$\psi(P, Q) := \begin{cases} a^2, & p = n, q < n, \\ b^2, & p < n, q = n, \\ a^2 + b^2, & p = q = n, \\ \min\{a^2, b^2\}, & \textit{otherwise}, \end{cases}$$

and for $r > 0$,

$$\psi(P,Q) := \frac{1}{2}\Bigg[a^2+b^2-\frac{1}{2}(a+b)\sqrt{(a+b)^2-4ab\sin^2\theta_{k+1}} - \frac{1}{2}|a-b|\sqrt{(a-b)^2+4ab\sin^2\theta_{k+1}}\Bigg],$$

where $a=\sqrt{\lambda_{\min}(P)}$, $b=\sqrt{\lambda_{\min}(Q)}$, $k=p+q-n$, $0\leqslant\theta_1\leqslant\theta_2\leqslant\ldots\leqslant\theta_{\min(p,q)}\leqslant\frac{\pi}{2}$ *represent the principal angles between* $\mathcal{R}(P),\mathcal{R}(Q)\subseteq\mathbb{R}^n$, *and* r *is the number of principal angles* θ_i *so that* $0<\theta_i<\frac{\pi}{2}$, *for* $1\leqslant i\leqslant\min(p,q)$.

Proof. Define $M_1 := \mathcal{R}(P)$, $M_2 := \mathcal{R}(Q)$ and $P_i \in \mathbb{R}^{n\times n}$ to be the orthogonal projection onto M_i, for $i=1,2$. On following the proof of Theorem 5.4.1, (5.4.3) gives

$$\begin{aligned}
\lambda_{\min}(P+Q) &\geqslant \inf_{x\in\mathbb{R}^n\setminus\{0\}}\frac{\lambda_{\min}(P)|P_1x|^2+\lambda_{\min}(Q)|P_2x|^2}{|x|^2}\\
&= \inf_{x\in\mathbb{R}^n\setminus\{0\}}\frac{a^2|P_1x|^2+b^2|P_2x|^2}{|x|^2} &(5.4.19)\\
&= \frac{1}{2}\inf_{x\in\mathbb{R}^n\setminus\{0\}}\frac{|(aP_1+bP_2)x|^2+|(aP_1-bP_2)x|^2}{|x|^2} &(\text{by (5.4.5)})\\
&\geqslant \frac{1}{2}\left[\sigma^2_{\min}(aP_1+bP_2)+\sigma^2_{\min}(aP_1-bP_2)\right]. &(5.4.20)
\end{aligned}$$

The results given in [32, p. 247] and [32, p. 234-235] give the following expression,

$$aP_1+bP_2 = Z\mathrm{diag}\left((a+b)I_k, aS+bE(\theta_{k+i})(i=1,\ldots,r), aI_{n_1}, bI_{n_2}, O_{n_3}\right)Z^T,$$

where Z is an orthogonal matrix, $S=\left[\begin{smallmatrix}1&0\\0&0\end{smallmatrix}\right]$, $E(\theta)=\left[\begin{smallmatrix}\cos^2\theta&\cos\theta\sin\theta\\\cos\theta\sin\theta&\sin^2\theta\end{smallmatrix}\right]$, and the expressions for k,n_1,n_2,n_3,n are given by (5.4.8). Thus, the set of singular values

of $aP_1 \pm bP_2$ are

$$\begin{aligned}\sigma(aP_1 + bP_2) = \Bigg\{ & \frac{1}{2}\left[(a+b) \pm \sqrt{(a+b)^2 - 4ab\sin^2\theta_{k+i}}\right](i = 1, \ldots, r), \\ & \boldsymbol{(a+b)}_k, \boldsymbol{a}_{n_1}, \boldsymbol{b}_{n_2}, \boldsymbol{0}_{n_3}\Bigg\}, \\ \sigma(aP_1 - bP_2) = \Bigg\{ & \frac{1}{2}\left[\sqrt{(a-b)^2 + 4ab\sin^2\theta_{k+i}} \pm |a-b|\right](i = 1, \ldots, r), \\ & \boldsymbol{|a-b|}_k, \boldsymbol{a}_{n_1}, \boldsymbol{b}_{n_2}, \boldsymbol{0}_{n_3}\Bigg\},\end{aligned} \tag{5.4.21}$$

resulting in the following cases.

Case 1: Suppose $r = 0$, then (5.4.8) yields $n_1 = p - k$ and $n_2 = q - k$ for this case. Firstly, consider $k = n_1 = 0$ and $n_2 > 0$, then (5.4.8) implies $p = 0$ and $q = n$. Thus, $P = O$ and $P + Q = Q$ is non-singular. Therefore, (5.4.19) leads to $\lambda_{\min}(P + Q) = \lambda_{\min}(Q) = b^2$. Similarly, $Q = O$ for $k = n_2 = 0$ and $n_1 > 0$, thus $\lambda_{\min}(P + Q) = \lambda_{\min}(P) = a^2$.

For $k = 0$ and $n_1, n_2 > 0$, (5.4.21) gives $\sigma_{\min}(aP_1 \pm bP_2) = \min\{a, b\}$. Thus, (5.4.20) gives $\lambda_{\min}(P + Q) \geqslant \min\{a^2, b^2\}$.

For $k > 0$ and $n_1 = n_2 = 0$, (5.4.8) implies $k = p = q = n$, that is, both P and Q are non-singular. By (5.4.21), $\sigma_{\min}(aP_1 + bP_2) = a + b$, $\sigma_{\min}(aP_1 - bP_2) = |a - b|$, therefore (5.4.20) gives

$$\lambda_{\min}(P + Q) \geqslant \frac{1}{2}\left[(a+b)^2 + (a-b)^2\right] = a^2 + b^2.$$

For $k, n_2 > 0$ and $n_1 = 0$, (5.4.8) gives $p = k$ and $q = n$, which imply $M_1 \subseteq M_2 = \mathbb{R}^n$, or Q is non-singular. Thus, (5.4.19) becomes

$$\lambda_{\min}(P + Q) = \inf_{x \in \mathbb{R}^n \setminus \{0\}} \frac{a^2|P_1 x|^2 + b^2|x|^2}{|x|^2} \geqslant b^2 = \lambda_{\min}(Q),$$

similarly $k, n_1 > 0$ and $n_2 = 0$ gives $\lambda_{\min}(P + Q) \geqslant a^2 = \lambda_{\min}(P)$. Finally, consider

$k, n_1, n_2 > 0$, since $\dim(M_1 \cap M_2) = k > 0$, then $M_1 \cap M_2$ is a non-trivial subspace. Consider $M_3 = M_2 \cap (M_1 \cap M_2)^\perp \neq \{0\}$ as defined in the proof of Theorem 5.4.1, then (5.4.12) in (5.4.19) gives

$$\begin{aligned}
\lambda_{\min}(P+Q) &= \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{a^2|P_1x|^2 + b^2|P_{M_1\cap M_2}x|^2 + b^2|P_3x|^2}{|x|^2} \\
&\geqslant \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{a^2|P_1x|^2 + b^2|P_3x|^2}{|x|^2} \\
&= \frac{1}{2}\inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|(aP_1+bP_3)x|^2 + |(aP_1-bP_3)x|^2}{|x|^2} \qquad \text{(by (5.4.5))} \\
&\geqslant \frac{1}{2}\left[\sigma_{\min}^2(aP_1+bP_3) + \sigma_{\min}^2(aP_1-bP_3)\right]. \qquad (5.4.22)
\end{aligned}$$

The set of singular values of $aP_1 \pm bP_3$ are given by (5.4.21) as follows

$$\begin{aligned}
\sigma(aP_1+bP_3) &= \left\{\frac{1}{2}\left[(a+b) \pm \sqrt{(a+b)^2 - 4ab\sin^2\alpha_{\tilde{k}+i}}\right](i=1,\ldots,r),\right. \\
&\qquad \left.(\boldsymbol{a+b})_{\tilde{k}}, \boldsymbol{a}_{\tilde{n}_1}, \boldsymbol{b}_{\tilde{n}_2}, \boldsymbol{0}_{\tilde{n}_3}\right\}, \\
\sigma(aP_1-bP_3) &= \left\{\frac{1}{2}\left[\sqrt{(a-b)^2 + 4ab\sin^2\alpha_{\tilde{k}+i}} \pm |a-b|\right](i=1,\ldots,r),\right. \\
&\qquad \left.|\boldsymbol{a-b}|_{\tilde{k}}, \boldsymbol{a}_{\tilde{n}_1}, \boldsymbol{b}_{\tilde{n}_2}, \boldsymbol{0}_{\tilde{n}_3}\right\},
\end{aligned} \qquad (5.4.23)$$

where the parameters are the same as (5.4.15) with $r = 0$. Thus, $\sigma_{\min}(aP_1 \pm bP_3) = \min\{a, b\}$. By (5.4.22), $\lambda_{\min}(P+Q) \geqslant \min\{a^2, b^2\}$.

Case 2: Suppose $r > 0$. Then for $k = 0$ and $n_1, n_2 \geqslant 0$, by (5.4.21)

$$\begin{aligned}
\sigma_{\min}(aP_1+bP_2) &= \frac{1}{2}\left[(a+b) - \sqrt{(a+b)^2 - 4ab\sin^2\theta_1}\right], \\
\sigma_{\min}(aP_1-bP_2) &= \frac{1}{2}\left[\sqrt{(a-b)^2 + 4ab\sin^2\theta_1} - |a-b|\right].
\end{aligned}$$

Thus, (5.4.20) gives

$$\lambda_{\min}(P+Q) \geqslant \frac{1}{2}\Bigg[a^2 + b^2 - \frac{1}{2}(a+b)\sqrt{(a+b)^2 - 4ab\sin^2\theta_1}$$

$$- \frac{1}{2}|a-b|\sqrt{(a-b)^2 + 4ab\sin^2\theta_1}\Bigg].$$

For $k > 0$ and $n_1, n_2 \geqslant 0$, (5.4.16), (5.4.17), and (5.4.23) yield

$$\sigma_{\min}(aP_1 + bP_3) = \frac{1}{2}\left[(a+b) + \sqrt{(a+b)^2 - 4ab\sin^2\theta_{k+1}}\right],$$
$$\sigma_{\min}(aP_1 - bP_3) = \frac{1}{2}\left[\sqrt{(a-b)^2 + 4ab\sin^2\theta_{k+1}} - |a-b|\right],$$

hence (5.4.22) implies

$$\lambda_{\min}(P+Q) \geqslant \frac{1}{2}\Bigg[a^2 + b^2 - \frac{1}{2}(a+b)\sqrt{(a+b)^2 - 4ab\sin^2\theta_{k+1}} - \frac{1}{2}|a-b|\sqrt{(a-b)^2 + 4ab\sin^2\theta_{k+1}}\Bigg].$$

□

In the proofs of Theorems 5.4.1 and 5.4.4, a technique similar to the case of $k = 0$ and $n_1, n_2 > 0$ can be applied to the cases $k = n_1 = 0$ and $n_2 > 0$, and $k = n_2 = 0$ and $n_1 > 0$, to get another positive lower bound; however, they turn out to be weaker than the stated results. On combining Theorems 5.4.1 and 5.4.4, another positive lower bound on $\lambda_{\min}(P+Q)$ is given as follows.

Corollary 5.4.5. *Let $P, Q \in \mathbb{R}^{n\times n}$ be PSD matrices of rank $p, q \leqslant n$, respectively, so that $P+Q$ is non-singular. Then*

$$\lambda_{\min}(P+Q) \geqslant \max\left[c(P,Q)\min\left\{\lambda_{\min}(P), \lambda_{\min}(Q)\right\}, \psi(P,Q)\right].$$

Example 5.4.6. Here, we consider four pairs of PSD matrices P, Q, so that $P+Q$ is SPD, to present simple illustrations of the above results. The exact value of $\lambda_{\min}(P+Q)$ is compared with the lower bounds given by Theorems 5.4.1 and 5.4.4. The existing results in the literature give a trivial lower bound. See Remark 5.4.2

for the definition of matrix M used in these examples.

1. Let $P = \operatorname{diag}(5,0,0)$, $Q = \operatorname{diag}(0,4,9)$, so that $\operatorname{rank} P = 1$ and $\operatorname{rank} Q = 2$, and $P+Q = \operatorname{diag}(5,4,9)$ is SPD. Note that $\mathcal{R}(P) = \mathcal{R}(Q)^{\perp}$, thus $k = r = 0$ and $p+q = 3 < 6$, so that $c(P,Q) = 1$ and thus, Theorems 5.4.1 and 5.4.4 give the same lower bound $4 = \lambda_{\min}(P+Q) \geqslant 4$.

2. Let $P = \operatorname{diag}(1,1,0)$, $Q = \operatorname{diag}(0,1,3)$, so that $\operatorname{rank} P = \operatorname{rank} Q = 2$, and $P+Q = \operatorname{diag}(1,2,3)$ is SPD. Note that principal angles between $\mathcal{R}(P)$ and $\mathcal{R}(Q)$ are $\theta_1 = 0$ and $\theta_2 = \frac{\pi}{2}$, so that $k = 1$, $r = 0$ and $p+q = 4 < 6$, thus $c(P,Q) = 1$. Therefore, Theorems 5.4.1 and 5.4.4 give the same lower bound, that is, $1 = \lambda_{\min}(P+Q) \geqslant 1 \cdot 1 = 1$.

3. Let $P = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$, $Q = \operatorname{diag}(6,0)$, so that they are PSD with $\operatorname{rank} P = \operatorname{rank} Q = 1$ and $P+Q = \begin{bmatrix} 8 & 2 \\ 2 & 2 \end{bmatrix}$ is SPD. Since the eigen-decomposition of $P = \mathcal{E}\Lambda\mathcal{E}^T$ and $Q = IQI^T$, where $\mathcal{E} = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ and $\Lambda = \operatorname{diag}(4,0)$, we get $M = \frac{\sqrt{2}}{2}\begin{bmatrix} 1 & 1 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = \frac{\sqrt{2}}{2}$. Note that $k = 0$ and $r > 0$, thus $\cos\theta_1 = \sigma_{\max}(M) = \frac{\sqrt{2}}{2}$ and Theorem 5.4.1 implies that $1.3944 \approx \lambda_{\min}(P+Q) \geqslant (1-\frac{\sqrt{2}}{2}) \cdot 4 \approx 1.1716$, which is stronger than the lower bound obtained by applying Theorem 5.4.4, $\lambda_{\min}(P+Q) \geqslant 1.1270$. Thus, Corollary 5.4.5 gives the former result, that is, $\lambda_{\min}(P+Q) \geqslant 1.1716$.

4. Let $P = \operatorname{diag}(10,5,0)$, $Q = \begin{bmatrix} 12 & & \\ & 3 & 9 \\ & 9 & 27 \end{bmatrix}$, so that $\operatorname{rank} P = \operatorname{rank} Q = 2$, and

$P+Q = \begin{bmatrix} 22 & & \\ & 8 & 9 \\ & 9 & 27 \end{bmatrix}$. Note that $k = r = 1 > 0$, so we use orthonormal bases for P and Q to find $M = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ \frac{1}{\sqrt{10}} & 0 \\ \frac{3}{\sqrt{10}} & 0 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ \frac{1}{\sqrt{10}} & 0 \end{bmatrix}$. Thus, $c(P,Q) = 1 - \cos\theta_2 = 1 - \frac{1}{\sqrt{10}}$. Therefore, Theorem 5.4.1 implies that $4.4137 \approx \lambda_{\min}(P+Q) \geqslant (1 - \frac{1}{\sqrt{10}}) \cdot 5 \approx 3.4189$, which is weaker than the lower bound obtained by applying Theorem 5.4.4, $\lambda_{\min}(P+Q) \geqslant 3.7770$. Thus, Corollary 5.4.5 gives $\lambda_{\min}(P+Q) \geqslant 3.7770$.

5.5 Importance of the Friedrichs angle

Although the *Friedrichs angle* is a vital quantity for studying the interaction between two given subspaces, its presence in the expression for lower bounds given by Theorems 5.4.1 and 5.4.4 limits its application as it could be difficult to evaluate it. In this section, we discuss an observation that indicated the presence of some angle in these estimates, its role in determining the interaction between certain subspaces, and our choice of subspaces in the proof of Theorem 5.4.1.

5.5.1 Motivation

An intuition for the presence of some angle came from the following toy example which we solved using the techniques of calculus.

Consider the following pair of matrices

$$P = \begin{bmatrix} 8 & 2 \\ 2 & 2 \end{bmatrix} \text{ and } Q = \begin{bmatrix} 6 & 0 \\ 0 & 0 \end{bmatrix}, \text{ then } P+Q = \begin{bmatrix} 8 & 2 \\ 2 & 2 \end{bmatrix}.$$

Note that P and Q are matrices of rank 1, whereas $P+Q$ is non-singular. By a direct calculation, we see that the bases for $\mathcal{R}(P)$ is $\{[1;1]\}$ and $\mathcal{R}(Q)$ is $[1;0]$, thus $c(P,Q) = 1 - \cos\theta_{\min} = 1 - \cos\frac{\pi}{4} \approx 0.2929$.

Define $P_1, P_2 \in \mathbb{R}^{2\times 2}$ to be orthogonal projections onto $\mathcal{R}(P)$ and $\mathcal{R}(Q)$, respectively. Let us evaluate the expression of $\inf_{x\in\mathbb{R}^2\setminus\{0\}} \Delta(x)$, defined by (5.4.4), explicitly. Clearly, for $x \in \mathbb{R}^2$,

$$P_1 x = \frac{1}{2}\begin{bmatrix} x_1 + x_2 \\ x_1 + x_2 \end{bmatrix}, \qquad P_2 x = \begin{bmatrix} x_1 \\ 0 \end{bmatrix}.$$

Then, $\inf_{\substack{x\in\mathbb{R}^2 \\ |x|=1}} \Delta(x) = \inf_{\substack{x\in\mathbb{R}^2 \\ |x|=1}} \frac{1}{2}(x_1+x_2)^2 + x_2^2$. By using calculus, this value is also equal to 0.2929 suggesting $1 - \dfrac{1}{\sqrt{2}}$.

5.5.2 A special case

Here we discuss the need of distinguishing the case $r = 0$ from $r > 0$ in Theorems 5.4.1 and 5.4.4. Note that $r = 0$ means the principal angles between M_1 and M_2 are either 0 or $\frac{\pi}{2}$. Recall that $n_1 = \dim(M_1 \cap M_2^\perp)$, $n_2 = \dim(M_1^\perp \cap M_2)$, $n_3 = \dim(M_1^\perp \cap M_2^\perp) = 0$, and $M_1 + M_2 = \mathbb{R}^n$. On substituting $r = 0$ in (5.4.8), we get $n_1 = p - k$, $n_2 = q - k$, so we need to analyze eight cases as follows.

k	n_1	n_2	Interpretation
0	0	0	$M_1 = M_2 = O$ or $P = Q = 0$
0	0	+	$M_1 = O$ or $P = O$
0	+	0	$M_2 = O$ or $Q = O$
0	+	+	$M_1 = M_2^\perp$, e.g. $P = \left[\begin{smallmatrix} 1 & \\ & 0 \end{smallmatrix}\right]$, $Q = \left[\begin{smallmatrix} 0 & \\ & 1 \end{smallmatrix}\right]$
+	0	0	$M_1 = M_2 = \mathbb{R}^n$, or both P and Q are non-singular
+	0	+	$M_1 \subseteq M_2 = \mathbb{R}^n$ or Q is non-singular
+	+	0	$M_2 \subseteq M_1 = \mathbb{R}^n$ or P is non-singular
+	+	+	$M_1 \cap M_2 \neq \{0\}$ and $M_3 = M_1^\perp$, e.g. $P = \left[\begin{smallmatrix} 1 & 1 & \\ & & 0 \end{smallmatrix}\right]$, $Q = \left[\begin{smallmatrix} 0 & & \\ & 1 & \\ & & 1 \end{smallmatrix}\right]$

Table 5.1: Cases for $r = 0$.

1. Suppose $k = n_1 = n_2 = 0$. Then $p = q = 0$, or M_1 and M_2 are zero subspaces. Thus, P and Q are zero matrices, and we do not consider this case.

2. Suppose $k = n_1 = 0$ and $n_2 > 0$. Then $p = 0$, thus M_1 is a zero subspace and P is a zero matrix. Therefore, all principal angles between M_1 and M_2 are $\frac{\pi}{2}$.

3. Suppose $k = n_2 = 0$ and $n_1 > 0$. Then $q = 0$, thus Q is a zero matrix. Therefore, all principal angles between M_1 and M_2 are $\frac{\pi}{2}$.

4. Suppose $k = 0$ and $n_1, n_2 > 0$. Then $n_1 = p$ and $n_2 = q$. Since $M_1 \cap M_2 = \{0\}$, $\dim(M_1 \cap M_2^{\perp}) = \dim(M_1)$, and $\dim(M_2 \cap M_1^{\perp}) = \dim(M_2)$, then $R^n = M_1 \oplus M_2$, thus $M_1 = M_2^{\perp}$. Therefore, all principal angles between them are $\frac{\pi}{2}$.

5. Suppose $k > 0$ and $n_1, n_2 = 0$. Then $p = q = k$, since $n = p + q - k$ or $n = k$, thus $p = q = n$. Therefore, $M_1 = M_2 = \mathbb{R}^n$ and P, Q are non-singular. Thus, all principal angles between them are zero.

6. Suppose $k, n_2 > 0$ and $n_1 = 0$. Then $p = k$, and $n = p + q - k$ implies $n = k + q - k$, then $n = q$. Thus, $M_2 = \mathbb{R}^n$ or Q is non-singular and clearly $M_1 \subseteq M_2$. Note that all principal angles are zero in this case as well.

7. Suppose $k, n_1 > 0$ and $n_2 = 0$. Similarly, to the above case $M_2 \subseteq M_1 = \mathbb{R}^n$, thus P is non-singular and all principal angles are zero in this case as well.

8. Suppose $k, n_1, n_2 > 0$. This is the most general case for the $r = 0$ setting. In this case, there is a non-trivial intersection between the subspaces, that is, $M_1 \cap M_2 \neq \{0\}$. Therefore, $r = 0$ means $M_3 = M_1^{\perp}$, so the first k principal angles are zero and rest of them are $\frac{\pi}{2}$.

For cases 5-7, the Friedrichs angle between M_1 and M_2 is 0, so that $\theta_F = 0$ which gives $1 - \cos\theta_F = 1 - \cos 0 = 0$. Therefore, we define $c(P, Q)$ separately for the cases $r = 0$ and $r > 0$ in the proof of the theorem. All the above cases are mentioned in

Table 5.1, where '+' represents that the corresponding parameter is considered to be positive.

5.5.3 The choice of subspaces

In this section, we discuss the choice of decomposition (5.4.12), particularly for the case $r, k > 0$, which states $|P_2 x|^2 = |P_{M_1 \cap M_2} x|^2 + |P_3 x|^2 = |P_4 x|^2 + |P_3 x|^2$.

Note that, this decomposition of M_2, in terms of M_3 and M_4, is considered for the case when $M_4 = M_1 \cap M_2 \neq \{0\}$. It is because when we try to estimate (5.4.6), which is

$$\inf_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) \geqslant \frac{1}{2} \left[\sigma_{\min}^2 (P_1 + P_2) + \sigma_{\min}^2 (P_1 - P_2) \right],$$

then since $k, r > 0$, Property 3 of Theorem 2.2.3 implies that $P_1 - P_2$ is singular, thus

$$\begin{aligned} \inf_{x \in \mathbb{R}^n \setminus \{0\}} \Delta(x) &\geqslant \frac{1}{2} \left[\sigma_{\min}^2 (P_1 + P_2) + 0 \right], \\ &= \frac{1}{2} \left(1 - \cos \theta_{k+1} \right)^2, \end{aligned}$$

where the last equality holds by (5.4.7). In order to improve this lower bound, we consider various combinations of two subspaces to estimate a lower bound on the infimum of $\Delta(x)$, as defined by (5.4.4). We restrict to two subspaces to facilitate analysis. The investigation of the case with three subspaces is a topic for future research. Note that $n_1, n_2 \geqslant 0$ and (5.4.8) imply $p, q \geqslant k + r$, thus $\dim(M_3) = q - k \geqslant r > 0$ and $\dim(M_5) = p - k \geqslant r > 0$, in other words, M_3 and M_5 are non-trivial whenever $r, k > 0$. Since (5.4.12) gives $|P_2 x|^2 = |P_4 x|^2 + |P_3 x|^2$, and similarly,

$|P_1x|^2 = |P_4x|^2 + |P_5x|^2$, thus the following approaches are established.

$$\begin{aligned}\inf_{x\in\mathbb{R}^n\setminus\{0\}} \Delta(x) &= \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|P_1x|^2 + |P_2x|^2}{|x|^2} \\ &= \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|P_1x|^2 + |P_4x|^2 + |P_3x|^2}{|x|^2} \\ &= \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|P_5x|^2 + 2|P_4x|^2 + |P_3x|^2}{|x|^2} \\ &= \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|P_5x|^2 + |P_4x|^2 + |P_2x|^2}{|x|^2}.\end{aligned}$$

By considering two terms at once in the above set of equations and applying (5.4.5), we can get a lower bound in terms of an expression of the form

$$\frac{1}{2}\left[\sigma^2_{\min}(P_i + P_j) + \sigma^2_{\min}(P_i - P_j)\right],$$

where $i \neq j$.

For subspaces $U, V \subseteq \mathbb{R}^n$, let P_U and P_V be orthogonal projections onto U and V, respectively. Properties 2 and 3 of Theorem 2.2.3 imply that $P_U + P_V$ is non-singular if and only if $\dim(U^\perp \cap V^\perp) = 0$, and $P_U - P_V$ is non-singular if and only if $\dim(U \cap V) = \dim(U^\perp \cap V^\perp) = 0$. Therefore, we construct Table 5.2, which describes the sum and difference of two orthogonal projections P_i, for $1 \leqslant i \leqslant 5$. Since $n = p$

U	V	$\dim(U \cap V)$	$\dim(U^\perp \cap V^\perp)$	Conclusion
M_1	M_3	0	0	both $P_1 \pm P_3$ are non-singular
M_1	M_4	k	$n-p$	both $P_1 \pm P_4$ are singular
M_2	M_4	k	$n-q$	both $P_2 \pm P_4$ are singular
M_2	M_5	0	0	both $P_2 \pm P_5$ are non-singular
M_3	M_4	0	$n-q$	both $P_3 \pm P_4$ are singular
M_3	M_5	0	k	both $P_3 \pm P_5$ are singular
M_4	M_5	0	$n-p$	both $P_4 \pm P_5$ are singular

Table 5.2: Combinations of two subspaces.

or $n = q$ occur when $r = 0$ as discussed in the cases 5-7 of the last section, therefore $n-p, n-q > 0$ whenever $r, k > 0$. Thus, only two of the combinations, M_1 and M_3,

and M_2 and M_5, give non-singular sum and difference of their orthogonal projections, that is, providing a non-trivial lower bound for both sum and difference of their orthogonal projections. The proof of Theorem 5.4.1 is completed by considering M_1 and M_3. It is easy to see that if we consider M_2 and M_5, then on following the proof to analyze

$$\inf_{x\in\mathbb{R}^n\setminus\{0\}} \Delta(x) \geqslant \inf_{x\in\mathbb{R}^n\setminus\{0\}} \frac{|P_5 x|^2 + |P_2 x|^2}{|x|^2},$$

we get the exact same lower bound as M_1 and M_4, that is,

$$\inf_{x\in\mathbb{R}^n\setminus\{0\}} \Delta(x) \geqslant 1 - \cos\theta_F = 1 - \cos\theta_{k+1},$$

which is optimal for our analysis.

5.6 Minimum singular value estimates

In this second, we use Theorems 5.4.1 and 5.4.4 for formulating new *lower bounds on the minimum singular value* of full rank 1×2, 2×1, and 2×2 matrices in terms of the minimum positive singular value of their sub-blocks. For convenience of notation, define the function Ψ for matrices $A\in\mathbb{R}^{n\times p}$, $B\in\mathbb{R}^{n\times q}$:

$$\Psi(A,B) = \sqrt{\psi(AA^T, BB^T)}, \tag{5.6.1}$$

where ψ is defined by Theorem 5.4.4. Note that $\psi(AA^T, BB^T)$ is a function defined in terms of $a = \sqrt{\lambda_{\min}(AA^T)} = \sigma_{\min}(A)$, $b = \sqrt{\lambda_{\min}(BB^T)} = \sigma_{\min}(B)$, and principal angles between $\mathcal{R}(AA^T) = \mathcal{R}(A)$ and $\mathcal{R}(BB^T) = \mathcal{R}(B)$. Thus, $\Psi(A,B)$ is a function defined in terms of positive singular values of A and B, and principal angles between $\mathcal{R}(A)$ and $\mathcal{R}(B)$.

A positive lower bound, defined by Corollary 5.4.5, could be useful in several circumstances, such as for a full rank block 2×1 matrix, with rank deficient sub-

blocks. Hence, the following applications to full rank *block column* and *block row matrix* are presented.

Corollary 5.6.1. *For $m \geqslant n$, let $A = \left[\begin{smallmatrix} A_1 \\ A_2 \end{smallmatrix}\right] \in \mathbb{R}^{m\times n}$ be full rank, then*

$$\sigma_{\min}(A) \geqslant \max\left[\sqrt{c(A_1^T, A_2^T)}\min\left\{\sigma_{\min}(A_1), \sigma_{\min}(A_2)\right\}, \Psi(A_1^T, A_2^T)\right].$$

Proof. Since A is a full rank matrix, $\sigma^2_{\min}(A) = \lambda_n(A^TA) = \lambda_{\min}(A_1^TA_1 + A_2^TA_2)$, thus Corollary 5.4.5 implies

$$\sigma^2_{\min}(A) \geqslant \max\left[c(A_1^TA_1, A_2^TA_2)\min\left\{\sigma^2_{\min}(A_1), \sigma^2_{\min}(A_2)\right\}, \psi\left(A_1^TA_1, A_2^TA_2\right)\right],$$

which gives the desired result after applying Property 1 of Proposition 5.4.3 and Equation (5.6.1). □

Corollary 5.6.2. *For $m \leqslant n$, let $A = \begin{bmatrix} A_1 & A_2 \end{bmatrix} \in \mathbb{R}^{m\times n}$ be full rank, then*

$$\sigma_{\min}(A) \geqslant \max\left[\sqrt{c(A_1, A_2)}\min\left\{\sigma_{\min}(A_1), \sigma_{\min}(A_2)\right\}, \Psi(A_1, A_2)\right].$$

After securing the above lower bounds, our subsequent aim is to extend them to a *non-singular* 2×2 *block matrix*. While it could be tedious to estimate the singular values of 2×2 block matrices, it is easier to find the singular values of its blocks which are of smaller size. Thus, another significant application of Theorems 5.4.1 and 5.4.4 is the following result, which give four estimates on the minimum singular value of a non-singular matrix.

Theorem 5.6.3. *For a non-singular matrix*

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \in \mathbb{R}^{n\times n},$$

where $A_{11} \in \mathbb{R}^{p\times k}$, $A_{22} \in \mathbb{R}^{q\times \ell}$, *for* $1 \leqslant p, q, k, \ell \leqslant n$, *the following hold*

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}\left([A_{11}, A_{12}]\right), \sigma_{\min}\left([A_{21}, A_{22}]\right)\right\}, \tag{5.6.2a}$$

$$\sigma_{\min}(A) \geqslant \Psi\left(\left[A_{11}, A_{12}\right]^T, \left[A_{21}, A_{22}\right]^T\right), \tag{5.6.2b}$$

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{r_1, r_2\right\}, \tag{5.6.2c}$$

where

$$r_1 := \max\left[c_1 \min\left\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{12})\right\}, \Psi\left(A_{11}, A_{12}\right)\right],$$
$$r_2 := \max\left[c_2 \min\left\{\sigma_{\min}(A_{21}), \sigma_{\min}(A_{22})\right\}, \Psi\left(A_{21}, A_{22}\right)\right],$$

where $c_1 = \sqrt{c(A_{11}, A_{12})}$, $c_2 = \sqrt{c(A_{21}, A_{22})}$, *and* $\theta \in \left(0, \frac{\pi}{2}\right]$ *is the minimum principal angle between* $\mathcal{R}\left(\left[A_{11}, A_{12}\right]^T\right), \mathcal{R}\left(\left[A_{21}, A_{22}\right]^T\right) \subseteq \mathbb{R}^n$. *Moreover,*

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{c_1, c_2\right\} \cdot \min_{1\leqslant i,j\leqslant 2}\left\{\sigma_{\min}(A_{ij})\right\}. \tag{5.6.3}$$

Proof. Since $A_{11} \in \mathbb{R}^{p\times k}$, $A_{22} \in \mathbb{R}^{q\times \ell}$, then $p+q = n = k+\ell$. Let $R_1 = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix} \in \mathbb{R}^{p\times n}$, and $R_2 = \begin{bmatrix} A_{21} & A_{22} \end{bmatrix} \in \mathbb{R}^{q\times n}$, then by a direct calculation

$$A^T A = \begin{bmatrix} R_1^T & R_2^T \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 \end{bmatrix} = R_1^T R_1 + R_2^T R_2. \tag{5.6.4}$$

Note that $\operatorname{rank}(R_1) \leqslant p$ and $\operatorname{rank}(R_2) \leqslant q$, thus $\operatorname{rank}(R_1) + \operatorname{rank}(R_2) \leqslant p+q = n$, also

$$\begin{aligned} \operatorname{rank}(R_1) + \operatorname{rank}(R_2) &= \operatorname{rank}(R_1^T R_1) + \operatorname{rank}(R_2^T R_2) \\ &\geqslant \operatorname{rank}(R_1^T R_1 + R_2^T R_2) \\ &= \operatorname{rank}(A^T A) = \operatorname{rank}(A) = n. \end{aligned}$$

Therefore, $\text{rank}(R_1) + \text{rank}(R_2) = n$, which implies that $\text{rank}(R_1) = p$ and $\text{rank}(R_2) = q$, that is, R_1 and R_2 are full rank matrices. And, A^TA is an SPD matrix expressed a sum of two singular PSD matrices, thus by Theorem 5.4.1

$$\begin{aligned}\sigma^2_{\min}(A) &= \lambda_{\min}(A^TA) \\ &\geqslant c(R_1^TR_1, R_2^TR_2)\min\left\{\lambda_{\min}(R_1^TR_1), \lambda_{\min}(R_2^TR_2)\right\} \\ &= c(R_1^T, R_2^T)\min\left\{\sigma^2_{\min}(R_1), \sigma^2_{\min}(R_2)\right\}, \end{aligned} \tag{5.6.5}$$

where the last equality results from Property 1 of Proposition 5.4.3. Let θ be the minimum principal angle between $\mathcal{R}(R_1^T)$ and $\mathcal{R}(R_2^T)$. Since A is non-singular, (5.6.4) gives $\mathcal{N}(R_1) \cap \mathcal{N}(R_2) = \{0\}$, thus Proposition 2.1.1 implies $\mathcal{R}(R_1^T) + \mathcal{R}(R_2^T) = \mathcal{N}(R_1)^\perp + \mathcal{N}(R_2)^\perp = (\mathcal{N}(R_1) \cap \mathcal{N}(R_2))^\perp = \{0\}^\perp = \mathbb{R}^n$. Therefore,

$$\begin{aligned}\dim(\mathcal{R}(R_1^T) \cap \mathcal{R}(R_2^T)) &= \dim(\mathcal{R}(R_1^T)) + \dim(\mathcal{R}(R_2^T)) - \dim(\mathcal{R}(R_1^T) + \mathcal{R}(R_2^T)) \\ &= \text{rank}(R_1) + \text{rank}(R_2) - \dim(\mathbb{R}^n) \\ &= p + q - n = 0,\end{aligned}$$

or, $\mathcal{R}(R_1^T) \cap \mathcal{R}(R_2^T) = \{0\}$. Hence, $\mathbb{R}^n = \mathcal{R}(R_1^T) \oplus \mathcal{R}(R_2^T)$, that is, $\mathcal{R}(R_1^T)$ and $\mathcal{R}(R_2^T)$ are complementary subspaces, thus by Lemma A.2.1 $0 < \theta \leqslant \frac{\pi}{2}$, which implies if $r = 0$ then $\theta = \frac{\pi}{2}$. Therefore, (5.4.1) can simply be expressed as $c(R_1^T, R_2^T) = 1 - \cos\theta$, thus (5.6.5) yields

$$\sigma^2_{\min}(A) \geqslant (1 - \cos\theta)\min\left\{\sigma^2_{\min}(R_1), \sigma^2_{\min}(R_2)\right\},$$

which leads to (5.6.2a). Also, applying Theorem 5.4.4 to (5.6.4) implies

$$\sigma^2_{\min}(A) = \lambda_{\min}\left(R_1^TR_1 + R_2^TR_2\right) \geqslant \psi\left(R_1^TR_1, R_2^TR_2\right),$$

which by (5.6.1) gives (5.6.2b). Since R_1 has full rank, Corollary 5.6.2 for estimating $\sigma_{\min}(R_1)$ leads to

$$\sigma_{\min}(R_1) \geqslant \max\left[c_1 \min\left\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{12})\right\}, \Psi\left(A_{11}, A_{12}\right)\right] =: r_1,$$

where $c_1 = \sqrt{c(A_{11}, A_{12})}$. Similarly, a lower bound on $\sigma_{\min}(R_2)$ is

$$\sigma_{\min}(R_2) \geqslant \max\left[c_2 \min\left\{\sigma_{\min}(A_{21}), \sigma_{\min}(A_{22})\right\}, \Psi\left(A_{21}, A_{22}\right)\right] =: r_2,$$

where $c_2 = \sqrt{c(A_{21}, A_{22})}$, and hence (5.6.2a) is expressed as

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{r_1, r_2\right\},$$

which on further simplification gives,

$$\begin{aligned}\sigma_{\min}(A) &\geqslant \sqrt{1-\cos\theta} \cdot \min\left[c_1 \min\left\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{12})\right\},\right. \\ &\qquad\qquad \left. c_2 \min\left\{\sigma_{\min}(A_{21}), \sigma_{\min}(A_{22})\right\}\right] \\ &\geqslant \sqrt{1-\cos\theta} \cdot \min\left\{c_1, c_2\right\} \cdot \min_{1 \leqslant i,j \leqslant 2}\left\{\sigma_{\min}(A_{ij})\right\}.\end{aligned}$$

□

As $\sigma_{\min}(A) = \sigma_{\min}(A^T)$, the above estimates yield the following result framed in terms of block columns of A.

Corollary 5.6.4. *For a non-singular matrix*

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right] \in \mathbb{R}^{n \times n},$$

where $A_{11} \in \mathbb{R}^{p\times k}$, $A_{22} \in \mathbb{R}^{q\times \ell}$, *for* $1 \leqslant p, q, k, \ell \leqslant n$,

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}\left(\left[\begin{smallmatrix} A_{11} \\ A_{21} \end{smallmatrix}\right]\right), \sigma_{\min}\left(\left[\begin{smallmatrix} A_{12} \\ A_{22} \end{smallmatrix}\right]\right)\right\}, \tag{5.6.6a}$$

$$\sigma_{\min}(A) \geqslant \Psi\left(\left[\begin{smallmatrix} A_{11} \\ A_{21} \end{smallmatrix}\right], \left[\begin{smallmatrix} A_{12} \\ A_{22} \end{smallmatrix}\right]\right), \tag{5.6.6b}$$

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{s_1, s_2\right\},$$

where

$$s_1 := \max\left[c_1 \min\left\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{21})\right\}, \Psi\left(A_{11}^T, A_{21}^T\right)\right],$$
$$s_2 := \max\left[c_2 \min\left\{\sigma_{\min}(A_{12}), \sigma_{\min}(A_{22})\right\}, \Psi\left(A_{12}^T, A_{22}^T\right)\right],$$

where $c_1 = \sqrt{c(A_{11}^T, A_{21}^T)}$, $c_2 = \sqrt{c(A_{12}^T, A_{22}^T)}$, *and* $\theta \in \left(0, \frac{\pi}{2}\right]$ *is the minimum principal angle between* $\mathcal{R}\left(\left[\begin{smallmatrix} A_{11} \\ A_{21} \end{smallmatrix}\right]\right), \mathcal{R}\left(\left[\begin{smallmatrix} A_{12} \\ A_{22} \end{smallmatrix}\right]\right) \subseteq \mathbb{R}^n$. *Moreover,*

$$\sigma_{\min}(A) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{c_1, c_2\right\} \cdot \min_{1\leqslant i,j\leqslant 2}\left\{\sigma_{\min}(A_{ij})\right\}.$$

Remark 5.6.5. The estimate given by (5.6.2a) is stronger than (5.6.2c) and (5.6.3), however, the sharpness of (5.6.2b) varies for different matrices (see Example 5.6.10). The inequality (5.6.3) gives a lower bound on the minimum singular value of a non-singular 2×2 block matrix in terms of the minimum positive singular value of its blocks. The estimates from Theorem 5.6.3 and Corollary 5.6.4 may differ, so in practice, one may use the maximum of all of the bounds obtained from both of them. A MATLAB® implementation for any 2×2 block matrix is given in [53].

Now, we simplify Theorem 5.6.3 for the special case of a saddle point matrix as follows.

Corollary 5.6.6. *For a non-singular saddle point matrix*

$$M = \begin{bmatrix} A & B \\ B^T & O \end{bmatrix} \in \mathbb{R}^{(m+n)\times(m+n)},$$

where $A \in \mathbb{R}^{m\times m}$ is non-singular and $B \in \mathbb{R}^{m\times n}$ is full rank,

$$\sigma_{\min}(M) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}(A), \sigma_{\min}(B)\right\},$$

where θ is the minimum principal angle between $\mathcal{R}\left([A,B]^T\right)$ and $\mathcal{R}\left([B^T,O]^T\right)$.

Proof. Since M is non-singular, according to (5.6.2a),

$$\sigma_{\min}(M) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}([A,B]), \sigma_{\min}([B^T,O])\right\}.$$

Since $\sigma^2_{\min}([A,B]) = \lambda_{\min}([A,B][A,B]^T) = \lambda_{\min}(AA^T + BB^T) \geqslant \sigma^2_{\min}(A)$, and similarly $\sigma^2_{\min}([B^T,O]) = \sigma^2_{\min}(B)$, hence the desired result. □

Example 5.6.7 (*Block diagonal matrix*)**.** For a non-singular block diagonal matrix $D = \left[\begin{smallmatrix} A & O \\ O & B \end{smallmatrix}\right]$, use (5.6.2a) to get

$$\sigma_{\min}(D) \geqslant \sqrt{1-\cos\theta} \cdot \min\left\{\sigma_{\min}([A,O]), \sigma_{\min}([O,B])\right\},$$

where θ is the minimum angle between $\mathcal{R}\left([A,O]^T\right)$ and $\mathcal{R}\left([O,B]^T\right)$. Note that $\sigma_{\min}([A,O]) = \sigma_{\min}(A)$ and $\sigma_{\min}([O,B]) = \sigma_{\min}(B)$, and it is straightforward to see that $\theta = \frac{\pi}{2}$. Therefore, the result becomes $\sigma_{\min}(D) \geqslant \min\left\{\sigma_{\min}(A), \sigma_{\min}(B)\right\}$. In fact, this inequality is an equality, thus the lower bound is sharp. Similar results can be obtained for a non-singular block anti-diagonal square matrix.

Example 5.6.8 (*Block triangular matrix*)**.** For a non-singular block upper triangular

matrix $U = \left[\begin{smallmatrix} A_{11} & A_{12} \\ O & A_{22} \end{smallmatrix}\right]$, the most simplified result of Theorem 5.6.3 is given by (5.6.3):

$$\sigma_{\min}(U) \geqslant \sqrt{1-\cos\theta} \cdot \min_{i=1,2}\{c_i\} \cdot \min\left\{\sigma_{\min}\left(A_{11}\right), \sigma_{\min}\left(A_{12}\right), \sigma_{\min}\left(O\right), \sigma_{\min}\left(A_{22}\right)\right\},$$

where θ is the minimum angle between $\mathcal{R}\left([A_{11}, A_{12}]^T\right)$ and $\mathcal{R}\left([O, A_{22}]^T\right)$, $c_1 = \sqrt{c(A_{11}, A_{12})}$, and $c_2 = \sqrt{c(O, A_{22})} = 1 \geqslant c_1$ by Property 3 of Proposition 5.4.3. Also, $\sigma_{\min}(O) = \infty$ by Definition 5.3.2. Hence,

$$\sigma_{\min}(U) \geqslant \sqrt{1-\cos\theta} \cdot c_1 \cdot \min\left\{\sigma_{\min}(A_{11}), \sigma_{\min}(A_{12}), \sigma_{\min}(A_{22})\right\}. \tag{5.6.7}$$

Similarly, an estimate for a non-singular block lower triangular matrix can be derived.

When every block is a square matrix, then [14, p. 352] gives the following expression

$$U^{-1} = \begin{bmatrix} A_{11}^{-1} & -A_{11}^{-1}A_{12}A_{22}^{-1} \\ O & A_{22}^{-1} \end{bmatrix},$$

whose maximum singular value is $\left(\sigma_{\min}(U)\right)^{-1}$. This task could be challenging to perform due to the presence of the term $A_{11}^{-1}A_{12}A_{22}^{-1}$. Note that the blocks need not to be square for (5.6.7).

Let us consider $U = \begin{bmatrix} 10 & 0 & 0 \\ 4 & 2 & 0 \\ 1 & 1 & 6 \end{bmatrix}$, then [105] gives $\sigma_{\min}(U) \geqslant 1.7087$. Whereas, on placing the partitions on U to make it a 2×2 block matrix so that its $(1,1)$ block is either of size 2×1 or 2×2, (5.6.2b) gives a stronger result $1.7473 \leqslant 1.8285 \approx \sigma_{\min}(U)$. A MATLAB® implementation is given in [53].

Example 5.6.9. Here, we explain the use of the new lower bounds with the help of two 2×2 block matrices denoted by A, where its (i,j)-th block is denoted by A_{ij} and its i-th block row is denoted by R_i, where $1 \leqslant i, j \leqslant 2$. Also, M represents the matrix defined in Remark 5.4.2. Most of the existing results do not provide a bound

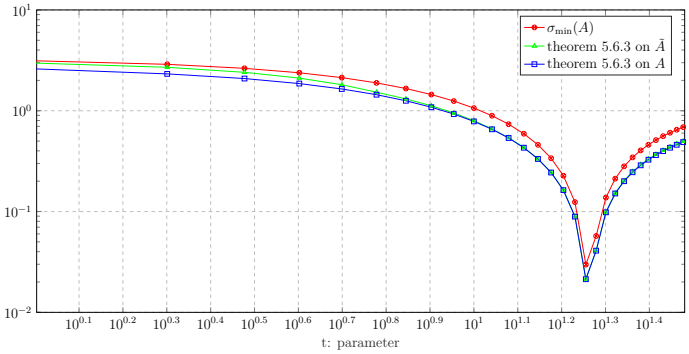
in terms of the blocks of the matrices considered in this example.

1. Consider $A = \left[\begin{array}{cc|cc} 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ \hline 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 \end{array}\right]$ which is non-singular such that its every block is singular. Note that $\sigma_{\min}(A) = 1 = \sigma_{\min}(A_{ij}) = \sigma_{\min}(R_i)$. It is straight forward to see that a basis for $\mathcal{R}(R_1^T)$ is $\left\{[0\ 1\ 0\ 0]^T, [0\ 0\ 0\ 1]^T\right\}$ and for $\mathcal{R}(R_2^T)$ is $\left\{[1\ 0\ 0\ 0]^T, [0\ 0\ 1\ 0]^T\right\}$. Since $\theta = \frac{\pi}{2}$ or $r = 0$, (5.6.2a) implies $1 = \sigma_{\min}(A) \geqslant (1-0) \cdot \min\{1,1\} = 1$, and (5.6.2b) yields $1 = \sigma_{\min}(A) \geqslant \sqrt{\min\{1,1\}} = 1$. Moreover, the same result is obtained on applying (5.6.2c), as $r = 0$ for both the pairs A_{1i} and A_{2i} for $i = 1, 2$, thus $1 = \sigma_{\min}(A) \geqslant (1-0)\min\{1,1\} = 1$. In order to use (5.6.3), note that a basis for $\mathcal{R}(A_{11}), \mathcal{R}(A_{21})$ is $\left\{[0\ 1]^T\right\}$, and for $\mathcal{R}(A_{12}), \mathcal{R}(A_{22})$ is $\left\{[1\,0]^T\right\}$, therefore $c_1 = c_2 = 1$. Thus, this inequality yields $1 = \sigma_{\min}(A) \geqslant (1-0)\min\{1,1\}\min\{1,1,1,1\} = 1$.

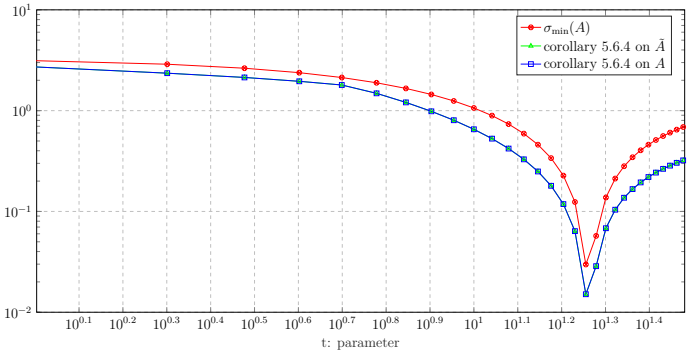
2. Consider $A = \left[\begin{array}{cc|c} 1 & 0 & 1 \\ 0 & -1 & 1 \\ \hline 1 & 0 & 0 \end{array}\right]$, which is a non-singular non-symmetric saddle point matrix with an indefinite matrix as its leading block. For using (5.6.2a), $M = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} & \frac{2}{\sqrt{3}} \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} \frac{1}{\sqrt{3}} \\ 1 \end{bmatrix}$. Thus, $\cos\theta = \sigma_{\max}(M) = \sqrt{\frac{2}{3}} \approx 0.8165$. Therefore, this inequality implies $0.4450 \approx \sigma_{\min}(A) \geqslant \sqrt{1-0.8165} \cdot \min\{1,1\} \approx 0.4284$. Moreover, the same lower bound is obtained on applying other results. Since $k = \operatorname{rank}(R_1^T) + \operatorname{rank}(R_2^T) - 3 = 0$, and $r = 1$ with $\cos\theta$ derived as above, (5.6.2b) implies $\sigma_{\min}(A) \geqslant \sqrt{\frac{1}{2}\left[2 - \sqrt{4(1-\sin^2\theta)}\right]} = \sqrt{1-0.8165} \approx 0.4284$. Also, (5.6.2c) results in $\sigma_{\min}(A) \geqslant \sqrt{1-0.8165} \cdot \min\{1,1\} \approx 0.4284$. For using (5.6.3), observe that $c_1 = c_2 = 1$ and $\sigma_{\min}(A_{11}) = 1$, $\sigma_{\min}(A_{12}) = \sqrt{2}$, $\sigma_{\min}(A_{21}) = 1$ and $\sigma_{\min}(A_{22}) = \infty$. Therefore, the inequal-

ity gives

$$0.4450 \approx \sigma_{\min}(A) \geqslant \sqrt{1-0.8165}\cdot\min\{1,1\}\cdot\min\{\min\{1,\sqrt{2}\},\min\{1,\infty\}\} \approx 0.4284.$$



(a) Graph for Theorem 5.6.3.



(b) Graph for Corollary 5.6.4.

Figure 5.1: Estimates of $\sigma_{\min}(A)$ for Example 5.6.10.

Example 5.6.10. Let us consider two different partitions on the same matrix as follows,

$$A = \left[\begin{array}{c|cc} t & 10 & 0 \\ \hline 3 & 2 & -2 \\ 2 & 0 & 6 \end{array}\right], \qquad \tilde{A} = \left[\begin{array}{cc|c} t & 10 & 0 \\ 3 & 2 & -2 \\ \hline 2 & 0 & 6 \end{array}\right], \qquad \text{where } t = 1, 2, \ldots, 30.$$

Figure 5.1 displays the result of best lower bounds from Theorem 5.6.3 and Corollary 5.6.4 for both partitions, along with the exact value of $\sigma_{\min}(A)$.

Figure 5.1a shows that Theorem 5.6.3 provides a decent estimate of $\sigma_{\min}(A)$. The best lower bound on $\sigma_{\min}(A)$ is given by (5.6.2b) for $t = 1$, and by (5.6.2a) for $t = 2, 3, \ldots 30$. For $\tilde{A}$, the largest lower bound is given by (5.6.2b) for $t = 1, 2, \ldots 5$, and by (5.6.2a) for $t = 6, 7, \ldots, 30$. On increasing the value of t, the lower bound obtained from (5.6.2a) improves up to $t = 18$, for which the absolute error in approximation for A is 0.008452 and for $\tilde{A}$ is 0.008389. The results from $\tilde{A}$ appear to be overall sharper than A. Thus, the sharpness of results may vary for distinct partitions of the same matrix.

Similarly, the trends for Corollary 5.6.4 are depicted by Figure 5.1b. It is observed that Corollary 5.6.4 gives identical results for both matrix A and $\tilde{A}$. The minimum absolute error in approximation is 0.01469, which occurs when $t = 18$. The best lower bound is given by the first inequality of Corollary 5.6.4 for all t except for $t = 1$, for which it was obtained from the second inequality of Corollary 5.6.4. A MATLAB® implementation is given in [53].

Example 5.6.11. (M- and H-matrices) The following matrices are considered in [70],

$$A = \begin{bmatrix} 8 & -2 & -1 \\ -5 & 7 & -3 \\ -3 & -4 & 5 \end{bmatrix}, \quad B = \begin{bmatrix} 7 & -3 & -2 \\ -2 & 5 & -1 \\ -3 & -4 & 9 \end{bmatrix}, \quad C = \begin{bmatrix} -5 & 2 & -4 \\ 3 & -6 & -2 \\ -1 & -4 & -8 \end{bmatrix},$$

where A and B are M-matrices and C is an H-matrix, for which upper bounds on the minimum singular values were devised. We calculate the best lower bounds secured from Theorem 5.6.3 and Corollary 5.6.4 in Table 5.3 through a MATLAB® implementation given in [53]. In Table 5.3, the size leading block refers to the size of $(1, 1)$ block of the matrix specifying the partition being placed, and more than one partition means that the same lower bound is obtained in all cases. It is evident that our results provide a good estimate for M- and H-matrices.

Matrix	$\sigma_{\min}$	Best lower bound	Size of leading block
A	0.7744	(5.6.6a):0.7354	1×2 or 2×2
B	1.8830	(5.6.6a):1.5855	1×2 or 2×2
C	0.9015	(5.6.6a):0.8770	1×1 or 2×1

Table 5.3: Lower bounds for M- and H- matrices.

Example 5.6.12. In this example, we compare our results to some well-known existing results that give a lower bound on the minimum singular value of a matrix. The following matrices are strictly diagonally dominant (SDD) matrices, for which several lower bounds were analyzed in [70],

$$D = \begin{bmatrix} 10 & 1 & 1 \\ 1 & 20 & 1 \\ 1 & 1 & 30 \end{bmatrix}, \; E = \begin{bmatrix} 10 & 1 & 1 \\ 1 & 20 & 1 \\ 10 & 1 & 30 \end{bmatrix}, \; F = \begin{bmatrix} 10 & 1 & 1 \\ 1 & 20 & 1 \\ 20 & 1 & 30 \end{bmatrix}, \; G = \begin{bmatrix} 10 & 1 & 1 \\ 10 & 20 & 1 \\ 20 & 1 & 30 \end{bmatrix}.$$

Also, some lower bounds for following matrices were compared in [61],

$$H = \begin{bmatrix} 3 & 2 & 0 \\ 1 & 9 & 5 \\ 0 & 5 & 7 \end{bmatrix}, \quad I = \begin{bmatrix} 2 & -1 & 0 \\ 2 & 1 & 0 \\ -4 & -4 & 5 \end{bmatrix}, \quad J = \begin{bmatrix} 5 & 0 & 0 \\ -4 & 9 & 4 \\ -1 & 7 & 9 \end{bmatrix}, \quad K = \begin{bmatrix} 4 & 0 & 0 \\ -1 & 5 & 0 \\ 0 & 5 & 4 \end{bmatrix}.$$

The third column of Table 5.4 states the best among all lower bounds evaluated for the above matrices in [70, 61]. The findings mentioned in Table 5.4 indicate that the results obtained from Theorem 5.6.3 and Corollary 5.6.4 provide a sharper lower bound on the minimum singular value of all SDD matrices considered in [70], albeit they may not be optimal for all SDD matrices. A MATLAB® implementation is given in [53]. In the above examples, we have listed the partitions that lead to best estimates, which may not be feasible if the matrix is large. Based on our numerical experiments, choosing the leading block of the matrix to be a square matrix of a suitable size often results in a partition that gives fine results.

Matrix	$\sigma_{\min}$	Existing	New Result	Size of leading block
D	9.8608	9.6389	(5.6.2b):9.6932	1×2 or 2×1 or 2×2
E	9.0409	8.0731	(5.6.2b):8.1814	2×1 or 2×2
F	7.6233	5.6070	(5.6.2a):6.0553	1×1 or 1×2
G	6.7547	5.2107	(5.6.2a):5.2728	1×1 or 1×2
H	1.9619	1.4142	(5.6.6b):1.8651	1×1 or 2×1
I	1.0677	0.7898	(5.6.2a):0.8996	1×1 or 1×2
J	3.0786	2.2303	(5.6.2b):2.8220	1×1 or 1×2
K	2.5146	2.2170	(5.6.6b):2.3847	1×1 or 2×1

Table 5.4: Comparison of new lower bounds with the existing results.

5.7 Some singular value estimates

After discussing lower bounds on the minimum singular value of a non-singular 2×2 block matrix, we divert our attention to constructing a lower bound on some other singular values. One such bounds is Theorem 5.7.1, stated as follows, which was given in [40].

Theorem 5.7.1 (See [40]). *Let $A_{ij} \in \mathbb{R}^{m \times n}$ for $i, j = 1, 2$, and*

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix},$$

then

$$\sigma_j(A) \geqslant \sqrt{2\sigma_j\left(A_{11}A_{12}^T + A_{21}A_{22}^T\right)}, \qquad j = 1, 2, \ldots, \min(m, n).$$

In the following theorem, we provide a simpler proof for two lower bounds similar to estimate provided by Theorem 5.7.1 with more general sizes for its sub-blocks.

Theorem 5.7.2. *For a block matrix*

$$A = \left[\begin{array}{c|c} A_{11} & A_{12} \\ \hline A_{21} & A_{22} \end{array}\right],$$

where $A_{11} \in \mathbb{R}^{m\times\ell}$ *and* $A_{22} \in \mathbb{R}^{n\times p}$*, then*

$$\sigma_j(A) \geqslant \sqrt{2\sigma_j\left(A_{11}A_{21}^T + A_{12}A_{22}^T\right)}, \qquad j = 1, 2, \ldots, \min(m, n, \ell + p),$$

also,

$$\sigma_j(A) \geqslant \sqrt{2\sigma_j\left(A_{11}^T A_{12} + A_{21}^T A_{22}\right)}, \qquad j = 1, 2, \ldots, \min(m + n, \ell, p).$$

Proof. Let $R_1 = \begin{bmatrix} A_{11} & A_{12} \end{bmatrix}$ and $R_2 = \begin{bmatrix} A_{21} & A_{22} \end{bmatrix}$, so that $A = \left[\begin{smallmatrix} R_1 \\ R_2 \end{smallmatrix}\right]$. For $j = 1, 2, \ldots, \min(m + n, \ell + p)$, $\sigma_j^2(A) = \lambda_j(A^T A) = \lambda_j(R_1^T R_1 + R_2^T R_2)$. Therefore, by theorem 5.3.5, for $j = 1, 2, \ldots, \min(m, n, \ell + p)$

$$\sigma_j^2(A) \geqslant 2\sigma_j(R_1 R_2^T) = 2\sigma_j\left(A_{11}A_{21}^T + A_{12}A_{22}^T\right).$$

Also, let $C_1 = \left[\begin{smallmatrix} A_{11} \\ A_{21} \end{smallmatrix}\right]$ and $C_2 = \left[\begin{smallmatrix} A_{12} \\ A_{22} \end{smallmatrix}\right]$, so that $A = \begin{bmatrix} C_1 & C_2 \end{bmatrix}$, then $\sigma_j^2(A) = \lambda_j(AA^T) = \lambda_j(C_1 C_1^T + C_2 C_2^T)$, for $j = 1, 2, \ldots, \min(m + n, \ell + p)$. Therefore, by theorem 5.3.5, for $j = 1, 2, \ldots, \min(m + n, \ell, p)$

$$\sigma_j^2(A) \geqslant 2\sigma_j(C_1^T C_2) = 2\sigma_j\left(A_{11}^T A_{12} + A_{21}^T A_{22}\right).$$

□

Corollary 5.7.3. *For a saddle point matrix*

$$M = \begin{bmatrix} A & B \\ B^T & O \end{bmatrix},$$

where $A \in \mathbb{R}^{m\times m}$ *is non-singular and* $B \in \mathbb{R}^{m\times n}$ *is full rank,*

$$\sigma_j(M) \geqslant \sqrt{2\sigma_j\left(AB\right)}, \qquad j = 1, 2, \ldots, \min(m, n).$$

To summarize, we formulated the lower bounds on $\lambda_{\min}(P+Q)$, for PSD matrices $P, Q \in \mathbb{R}^n$ described by Theorems 5.4.1 and 5.4.4, which are sharp for the case of $\mathcal{R}(P) \cap \mathcal{R}(Q) = \{0\}$. These gave lower bounds on the minimum singular value of some full rank matrices given in Corollaries 5.6.1 and 5.6.2, Theorem 5.6.3, and Corollary 5.6.4. Finally, an improvement on lower bounds on some other singular values were derived by Theorem 5.7.2, which concludes our analysis for general matrices.

6

Ultraspherical spectral methods in space and time

In Chapter 3, we analyzed a space-time spectral method for the Stokes problem, which utilizes collocation in time. A major disadvantage of such a scheme is that the resulting linear system is dense and is difficult to solve in parallel. We try to formulate space-time spectral methods leading to sparse linear systems, paving the way for robust numerical methods for solving time-dependent PDEs, possessing spectral convergence in both space and time.

In 2013, S. Olver and A. Townsend introduced a new class of spectral methods in [74], called the ultraspherical spectral (US) methods, which portray spectral convergence for analytic solutions, and the resultant discrete systems constitute sparse and well-conditioned matrices. However, the linear systems are non-symmetric for self-adjoint problems; [1] presents a technique on symmetrizing such schemes. Some authors have driven the further development of the US methods, such as [94, 84, 18, 30, 75, 29]. We explore the capabilities of the US method for solving unsteady problems. Firstly, we introduce the US method and analyze their performance for some ordinary differential equations. We proceed to design numerical

schemes for solving some time-dependent linear PDEs by utilizing the ultraspherical spectral method in both space and time. This chapter is concluded by discussing the scope of designing a robust solver for the schemes arising from the US method in space and time.

6.1 Introduction

Classical spectral methods converge spectrally, but they lead to dense and ill-conditioned matrices. The *ultraspherical spectral methods* exhibit spectral convergence and lead to almost banded and well-conditioned matrices. This new class of spectral method was developed in [74], which results from changing the basis of the solution by utilizing a differentiation expression for the ultraspherical polynomials defined later in this section.

In Chapter 2, we defined the Chebyshev polynomials, which are also termed as the Chebyshev polynomial of the first kind. Furthermore, the Chebyshev polynomial of the second kind are denoted by $C_n^{(1)}(x)$ or $U_n(x)$, where the former notation will be used in this work. They are the solution of the following problem:

$$(1-x^2)y'' - 3xy' + n(n+2)y = 0,\ x \in (-1,1), \qquad \forall\, n = 0, 1, \ldots,$$

and are orthogonal with respect to the $L^2_{\omega^{\frac{1}{2},\frac{1}{2}}}(-1,1)$ inner product with weight $\omega^{\frac{1}{2},\frac{1}{2}}(x) = \sqrt{1-x^2}$, that is,

$$\int_{-1}^{1} C_n^{(1)}(x) C_m^{(1)}(x) \sqrt{1-x^2}\, dx = \frac{\pi}{2}\delta_{mn}.$$

Moreover, they serve as a tool for demonstrating the ultraspherical spectral method

for a first order ODE. For $x \in [-1, 1]$, consider

$$u'(x) + a(x)u(x) = f(x),$$

with boundary condition $u(-1) = c$, where $a : [-1, 1] \to \mathbb{C}$, $f : [-1, 1] \to \mathbb{C}$ are continuous functions of bounded variation. Therefore, there exists a unique continuously differentiable solution. Let us consider the solution

$$u(x) = \sum_{k=0}^{\infty} u_k T_k(x),$$

where the coefficients satisfy

$$u_k = \frac{2 - \delta_{k0}}{\pi} \int_{-1}^{1} \frac{u(x)T_k(x)}{\sqrt{1 - x^2}} dx. \tag{6.1.1}$$

Since the *derivative* of the Chebyshev polynomial is given as,

$$\frac{dT_k}{dx} = \begin{cases} kC_{k-1}^{(1)}, & k \geqslant 1, \\ 0, & k = 0. \end{cases} \tag{6.1.2}$$

Thus,

$$u'(x) = \sum_{k=1}^{\infty} k u_k C_{k-1}^{(1)}(x).$$

In matrix form, the coefficients are given as

$$\mathcal{D}_0 \mathbf{u} = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 2 & & \\ & & 0 & 3 & \\ & & & \ddots & \ddots \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ \vdots \end{bmatrix}.$$

The sparse matrix $\mathcal{D}_0$ is called the *differentiation operator*. Expressing $a(x)$ in terms

of its Chebyshev series, i.e., $a(x) = \sum_{j=0}^{\infty} a_j T_j(x)$, we obtain

$$a(x)u(x) = \sum_{j=0}^{\infty}\sum_{k=0}^{\infty} a_j u_k T_j(x)T_k(x) = \sum_{k=0}^{\infty} c_k T_k(x),$$

where the coefficients c_k are given as

$$c_k = \begin{cases} a_0u_0 + \dfrac{1}{2}\displaystyle\sum_{l=1}^{\infty} a_l u_l, & k = 0 \\ \dfrac{1}{2}\displaystyle\sum_{l=0}^{k-1} a_{k-l}u_l + a_0u_k + \dfrac{1}{2}\sum_{l=1}^{\infty} a_l u_{l+k} + \dfrac{1}{2}\sum_{l=0}^{\infty} a_{l+k}u_l, & k \geqslant 1. \end{cases}$$

In matrix form, $\mathbf{c} = \mathcal{M}_0[\mathbf{a}]\mathbf{u}$, where

$$\mathcal{M}_0[\mathbf{a}] = \frac{1}{2}\left[\begin{pmatrix} 2a_0 & a_1 & a_2 & a_3 & \dots \\ a_1 & 2a_0 & a_1 & a_2 & \ddots \\ a_2 & a_1 & 2a_0 & a_1 & \ddots \\ \vdots & \ddots & \ddots & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} 0 & 0 & 0 & 0 & \dots \\ a_1 & a_2 & a_3 & a_4 & \iddots \\ a_2 & a_3 & a_4 & a_5 & \iddots \\ \vdots & \iddots & \iddots & \iddots & \iddots \end{pmatrix}\right].$$

The above Toeplitz plus an almost Hankel matrix $\mathcal{M}_0[a]$ is called the *multiplication operator*. We expressed $u(x) = \sum_{k=0}^{\infty} u_k T_k(x)$. Practically, we cannot find infinite number of coefficients. Therefore, we truncate the series and approximate $u(x) \approx \sum_{k=0}^{n-1} u_k T_k(x)$, i.e., we find n coefficients.

Since $a(x)$ is a continuous function with bounded variation, it has a unique representation as a uniformly convergent Chebyshev expansion. Therefore, for all $\epsilon > 0$ there exists $M(\epsilon) \in \mathbb{N}$ such that for all $m \geqslant M$

$$\left| a(x) - \sum_{k=0}^{m-1} a_k T_k(x) \right| < \epsilon.$$

Thus, we truncate the series expansion for a, i.e., $a(x) \approx \sum_{k=0}^{m-1} a_k T_k(x)$. Therefore,

for the infinite matrix $\mathcal{M}_0[a]$, terms $a_i = 0$ for $i \geqslant m$, hence it is a banded matrix. If $n > m$, then we will consider an $n \times n$ sub-matrix of $\mathcal{M}_0[a]$ which will also be banded.

The following is a relationship between the Chebyshev polynomial of first and second type,

$$T_k = \begin{cases} \frac{1}{2}(C_k^{(1)} - C_{k-2}^{(1)}), & k \geqslant 2, \\ \frac{1}{2}C_1^{(1)}, & k = 1, \\ C_0^{(1)}, & k = 0. \end{cases}$$

Then we can express $u(x)$ in terms of its series of Chebyshev polynomials of second type as,

$$u(x) = \sum_{k=0}^{\infty} u_k T_k(x) = \left(u_0 - \frac{1}{2}u_2\right) C_0^{(1)}(x) + \sum_{k=1}^{\infty} \frac{1}{2}(u_k - u_{k+2}) C_k^{(1)}(x).$$

Hence, $C^{(1)}$ coefficients of u are $\mathcal{S}_0\mathbf{u}$, where

$$\mathcal{S}_0 = \begin{bmatrix} 1 & -\frac{1}{2} & & & \\ & \frac{1}{2} & & -\frac{1}{2} & \\ & & \frac{1}{2} & & -\frac{1}{2} \\ & & & \ddots & & \ddots \end{bmatrix}.$$

The above banded and sparse matrix $\mathcal{S}_0$ is called the *conversion operator*.

The discrete version of the operator, $\frac{d}{dx} + a(x)$, is given as

$$\mathcal{L} := \mathcal{D}_0 + \mathcal{S}_0\mathcal{M}_0[a],$$

which corresponds to the second kind Chebyshev series of $u(x)$. Define $f = \sum_{k=0}^{\infty} f_k T_k(x)$, and $\mathbf{f}$ as the vector of coefficients f_k, for all $0 \leqslant k < \infty$. Thus, the differential equa-

tion, without its boundary conditions, is given as

$$\mathcal{L}\mathbf{u} = \mathcal{S}_0\mathbf{f}.$$

Note that this is an infinite system. In order to truncate this system to order n, we define a *projection operator* as, $\mathcal{P}_n = [\mathbb{I}_n \,|\, \mathbb{O}]_{n\times\infty}$. We truncate the differentiation operator to a system of order n as, $\mathcal{P}_n\mathcal{D}_0\mathcal{P}_n^T$. The operator $\mathcal{D}_0$ is hence recovered for the considered modes, but with the last row containing all zeros. Henceforth, we swap the first row with the last row of the system $\mathcal{P}_n\mathcal{L}\mathcal{P}_n^T$. Since the boundary condition is $u(-1) = c$,

$$c = u(-1) = \sum_{k=0}^{\infty} u_k T_k(-1) = \sum_{k=0}^{\infty} (-1)^k u_k.$$

Therefore, we impose the boundary condition in the first row, hence the discrete form of the problem is given as

$$A_n\mathbf{u} = \begin{bmatrix} T_0(-1) & T_1(-1) & \dots & T_{n-1}(-1) \\ & \mathcal{P}_{n-1}\mathcal{L}\mathcal{P}_n^T & & \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} c \\ \left(\mathcal{P}_{n-1}\mathcal{S}_0\mathcal{P}_n^T\right)\left(\mathcal{P}_n\mathbf{f}\right) \end{bmatrix}.$$

On solving this system we get the approximate solution: $\sum_{k=0}^{n-1} u_k T_k(x)$.

A similar approach is used for the higher order ordinary differential equations by using *ultraspherical polynomials*, also known as Gegenbauer polynomials, which are represented as $C^{(\lambda)}(x)$, for $x \in [-1, 1]$. For $n = 0, 1, \dots$ and $\lambda > 0$, $C_n^{(\lambda)}$, is the solution of the Gegenbauer differential equation:

$$(1-x)^2 y'' - (2\lambda+1)xy' + n(n+2\lambda)y = 0.$$

They are a family of polynomials orthogonal with respect to the weight

$$\left(1-x^2\right)^{\lambda-\frac{1}{2}}.$$

They can also be generated by the following recurrence relation:

$$\begin{aligned}
C_0^{(\lambda)}(x) &= 1,\\
C_1^{(\lambda)}(x) &= 2\lambda x,\\
C_{n+1}^{(\lambda)}(x) &= \frac{1}{n}\left[2x(n+\lambda-1)C_n^{(\lambda)} - (n+2\lambda-2)C_{n-1}^{(\lambda)}\right],\ \forall\, n\in\mathbb{N}.
\end{aligned}$$

For the case $\lambda = 1$, we recover the Chebyshev polynomials of second kind. For the case $\lambda = \dfrac{1}{2}$, $C_k^{(\lambda)}$ reduce to the Legendre polynomials. We use them only for $\lambda = 0, 1, \ldots$ and normalize the leading coefficient so that

$$C_k^{(\lambda)}(x) = \frac{2^k\,(\lambda)_k}{k\,!}x^k + \mathcal{O}(x^{k-1}),$$

where $(\cdot)_k$ denotes the Pochhammer symbol defined by $(\lambda)_k = \frac{(\lambda+k-1)\,!}{(\lambda-1)\,!}$.

This process can be generalized to higher order differential equations of the form,

$$\sum_{\lambda=0}^{N} a^{\lambda}(x)\frac{d^{\lambda}u(x)}{dx^{\lambda}} = \tilde{f}(x), \qquad x\in[-1,1],$$

with some $N\in\mathbb{N}$ number of general boundary conditions

$$\mathcal{B}\mathbf{u} = \mathbf{c}.$$

As the first order differentiation resulted in a linear combination $C^{(1)}$ functions, similarly, higher order differentiation results in higher order ultraspherical polynomials. We begin by considering the solution of the higher order ODE as $u(x) = \sum_{k=0}^{\infty} u_k T_k(x)$.

As for $\lambda \geqslant 1$,

$$\frac{dC_k^{(\lambda)}}{dx} = \begin{cases} 2\lambda C_{k-1}^{(\lambda+1)}, & k \geqslant 1, \\ 0, & k = 0. \end{cases} \tag{6.1.3}$$

Then for $\lambda \in \mathbb{N}$,

$$\frac{d^\lambda u(x)}{dx^\lambda} = 2^{\lambda-1}(\lambda-1)\,! \sum_{k=\lambda}^{\infty} k u_k C_{k-\lambda}^{(\lambda)}(x).$$

In matrix form, the coefficients are given as $\mathcal{D}_\lambda \mathbf{u}$. Thus, the λ-order *differentiation operator* for the Chebyshev series of the first kind is given as

$$\mathcal{D}_\lambda = 2^{\lambda-1}(\lambda-1)\,! \begin{bmatrix} 0 & \dots & 0 & \lambda & & & \\ & & & & \lambda+1 & & \\ & & & & & \lambda+2 & \\ & & & & & & \ddots \end{bmatrix},$$

where the first λ entries in the first row of $\mathcal{D}_\lambda$ are zero. A conversion relationship between $C^{(\lambda)}$ and $C^{(\lambda+1)}$ for $\lambda \in \mathbb{N}$ is given as,

$$C_k^{(\lambda)} = \begin{cases} \frac{\lambda}{\lambda+k}(C_k^{(\lambda+1)} - C_{k-2}^{(\lambda+1)}), & k \geqslant 2, \\ \frac{\lambda}{\lambda+1} C_1^{(\lambda+1)}, & k = 1, \\ C_0^{(\lambda+1)}, & k = 0. \end{cases}$$

Then, the *conversion operator* that converts the coefficients of $C^{(\lambda)}$ polynomials to $C^{(\lambda+1)}$ is denoted by $\mathcal{S}_\lambda$ and is given as

$$\mathcal{S}_\lambda = \begin{bmatrix} 1 & & -\frac{\lambda}{\lambda+2} & & & \\ & \frac{\lambda}{\lambda+1} & & -\frac{\lambda}{\lambda+3} & & \\ & & \frac{\lambda}{\lambda+2} & & -\frac{\lambda}{\lambda+4} & \\ & & & \ddots & & \ddots \end{bmatrix}. \tag{6.1.4}$$

For $a(x) = \sum_{j=0}^{\infty} a_j C_j^{(\lambda)}(x)$ and $u^{(\lambda)}(x) = \sum_{j=0}^{\infty} a_j C_j^{(\lambda)}(x)$,

$$a(x)u^{(\lambda)}(x) = \sum_{j=0}^{\infty} \left(\sum_{k=0}^{\infty} \sum_{s=\max(0,k-j)}^{k} a_{2s+j-k} c_s^{\lambda}(k, 2s+j-k) u_k \right) C_j^{(\lambda)}(x),$$

where

$$c_s^{\lambda}(j,k) = \frac{j+k+\lambda-2s}{j+k+\lambda-s} \frac{(\lambda)_s (\lambda)_{j-s} (\lambda)_{k-s}}{s!(j-s)!(k-s)!} \frac{(2\lambda)_{j+k-s}}{(\lambda)_{j+k-s}} \frac{(j+k-2s)!}{(2\lambda)_{j+k-2s}}.$$

Thus the *multiplication operator* for the product of $C^{(\lambda)}$ series is given as

$$\mathcal{M}_{\lambda}[a]_{j,k} = \sum_{s=\max(0,k-j)}^{k} a_{2s+j-k} c_s^{\lambda}(k, 2s+j-k) u_k.$$

The differential operator can be represented as

$$\mathcal{L} := \mathcal{M}_N[a^N]\mathcal{D}_N + \sum_{\lambda=1}^{N-1} \mathcal{S}_{N-1} \ldots \mathcal{S}_{\lambda} \mathcal{M}_{\lambda}[a^{\lambda}] \mathcal{D}_{\lambda} + \mathcal{S}_{N-1} \ldots \mathcal{S}_0 \mathcal{M}_0[a^0].$$

This system is expressed in terms of coefficients of $C^{(N)}$ series, thus the coefficients of the Chebyshev series of f must be converted. The discrete form of the concerned higher order problem can be written as

$$A_n \mathbf{u} = \begin{bmatrix} \mathcal{B}\mathcal{P}_n^T \\ \mathcal{P}_{n-K} \mathcal{L} \mathcal{P}_n^T \end{bmatrix} \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_{n-1} \end{bmatrix} = \begin{bmatrix} \mathbf{c} \\ \mathcal{P}_{n-K} \mathcal{S}_{N-1} \ldots \mathcal{S}_0 \mathbf{f} \end{bmatrix},$$

yielding the approximate solution of the problem: $\sum_{k=0}^{n-1} u_k T_k(x)$.

6.2 Some linear ODEs

In [74, p. 479], the following *right-preconditioner* was defined for a discrete linear ODE of order $d \geqslant 1$ with the leading coefficient $a_d(x) = 1$,

$$\mathfrak{R} = \frac{1}{2^{d-1}(d-1)!} \operatorname{diag}\left(\overbrace{1,\ldots,1}^{d \text{ times}}, \frac{1}{d}, \frac{1}{d+1}, \ldots\right).$$

The space $\ell^2_\lambda \subset \mathbb{C}^\infty$ is defined as the Hilbert space with norm

$$\|\mathbf{u}\|_{\ell^2_\lambda} = \sqrt{\sum_{k=0}^{\infty} |u_k|^2 (k+1)^{2\lambda}} < \infty,$$

where $\lambda = D - 1, D, \ldots$, so that $D = 1$ for Dirichlet boundary conditions, and each additional derivative used in the boundary condition will increase D by one. In [74, p. 480], the condition number of a discrete linear ODE right-preconditioned by $\mathfrak{R}_n = \mathcal{P}_n \mathfrak{R} \mathcal{P}_n^T$, that is, $A_n R_n$ is $\mathcal{O}(1)$. Out of curiosity, in this section we derive the spectral and 2-norm condition number of discrete linear ODEs obtained by applying the ultraspherical spectral method.

6.2.1 First order

Let us first consider solving the most basic problem, that is, *first order problem* by using the US method, given as follows

$$\begin{aligned} u'(x) &= f(x), \quad x \in (-1, 1) \\ u(-1) &= 0. \end{aligned} \tag{6.2.1}$$

Let $u(x) = \sum_{k=0}^{N} u_k T_k(x)$, then the boundary condition stated above implies

$$\sum_{k=0}^{N} u_k T_k(-1) = \sum_{k=0}^{N} u_k (-1)^k = 0. \tag{6.2.2}$$

Recall (6.1.2), then (6.2.1) gives,

$$\begin{aligned}
\sum_{k=0}^{N} u_k T_k'(x) = \tilde{f}(x) &= \sum_{k=0}^{N-1} f_k C_k^{(1)}(x) \\
\sum_{k=1}^{N} k u_k C_{k-1}^{(1)}(x) &= \sum_{k=0}^{N-1} f_k C_k^{(1)}(x) \\
\sum_{k=0}^{N-1} (k+1) u_{k+1} C_k^{(1)}(x) &= \sum_{k=0}^{N-1} f_k C_k^{(1)}(x).
\end{aligned}$$

Comparing the coefficients of $C_k^{(1)}(x)$ gives $(k+1)u_{k+1} = f_k$, for $0 \leqslant k \leqslant N-1$, which along with (6.2.2) gives the following linear system,

$$\begin{bmatrix}
1 & -1 & 1 & -1 & \dots & (-1)^N \\
 & 1 & & & & \\
 & & 2 & & & \\
 & & & 3 & & \\
 & & & & \ddots & \\
 & & & & & N
\end{bmatrix}_{(N+1)\times(N+1)}
\begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{bmatrix}_{(N+1)\times 1}
=
\begin{bmatrix} 0 \\ f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{bmatrix}_{(N+1)\times 1}, \tag{6.2.3}$$

written as $\mathcal{A}_1 u_h = f_h$, with $\mathcal{A}_1 \in \mathbb{R}^{(N+1)\times(N+1)}$ defined as the coefficient matrix of the above system. Now, we derive a condition number estimate for this system.

Theorem 6.2.1. *For $N \geqslant 4$, let $\mathcal{A}_1$ be defined by* (6.2.3)*. Then, $\kappa_{sp}(\mathcal{A}_1) \leqslant cN$ and $\kappa(\mathcal{A}_1) \leqslant cN$.*

Proof. Note that $\mathcal{A}_1$ is an upper triangular matrix, therefore its eigenvalues are its

diagonal entries. Clearly, $\lambda_{\min}(\mathcal{A}_1) = 1$ and $\lambda_{\max}(\mathcal{A}_1) = N$, thus

$$\kappa_{sp}(\mathcal{A}_1) = \frac{\lambda_{\max}(A_1)}{\lambda_{\min}(A_1)} = \frac{N}{1} = N.$$

For deriving the 2-norm condition number estimate, note that the first row of $\mathcal{A}_1$ gives $\|\mathcal{A}_1\|_\infty = (N+1)$. Also, $\|\mathcal{A}_1\|_1 = (N+1)$, is achieved by the $(N+1)$st column of $\mathcal{A}_1$ Thus,

$$\|\mathcal{A}_1\| \leqslant \sqrt{\|\mathcal{A}_1\|_1 \|\mathcal{A}_1\|_\infty} = (N+1) \leqslant cN. \tag{6.2.4}$$

It is easily derived that the inverse of $\mathcal{A}$ is the following matrix.

$$\mathcal{A}_1^{-1} = \begin{bmatrix} 1 & 1 & -2^{-1} & 3^{-1} & \dots & (-1)^{N-1}N^{-1} \\ & 1 & & & & \\ & & 2^{-1} & & & \\ & & & 3^{-1} & & \\ & & & & \ddots & \\ & & & & & N^{-1} \end{bmatrix}_{(N+1)\times(N+1)} =: \left[\begin{array}{c|c} 1 & B_{1_{1\times N}} \\ \hline & D_{1_{N\times N}} \end{array}\right].$$

In order to estimate $\sigma_{\max}(\mathcal{A}_1)$, let us consider the gram matrix of $\mathcal{A}_1^{-1}$,

$$\begin{aligned} \mathfrak{S}_1 &:= \mathcal{A}_1^{-T}\mathcal{A}_1^{-1} \\ &= \begin{bmatrix} 1 & \\ B_1^T & D_1 \end{bmatrix} \begin{bmatrix} 1 & B_1 \\ & D_1 \end{bmatrix} \\ &= \begin{bmatrix} 1 \\ B_1^T \end{bmatrix} \begin{bmatrix} 1 & B_1 \end{bmatrix} + \begin{bmatrix} 0 & \\ & D_1^2 \end{bmatrix} \\ &=: C_1^T C_1 + \mathfrak{D}_1. \end{aligned}$$

Thus, $\mathfrak{S}_1$ is the sum of two symmetric matrices given above. Since Lemma 5.3.4 im-

plies that the non-zero eigenvalues of $C_1^T C_1$ and $C_1 C_1^T$ are equal, we get the following result.

$$\begin{aligned} C_1 C_1^T &= 1 + B_1 B_1^T \\ &= 1 + (1 + 2^{-2} + 3^{-2} + \ldots + N^{-2}) \\ &\leqslant 1 + \sum_{i=1}^{\infty} i^{-2} \\ &= 1 + \frac{\pi^2}{6}. \end{aligned}$$

Thus, $\lambda_{\max}(C_1^T C_1) = \lambda_{\max}(C_1 C_1^T) \leqslant 1 + \frac{\pi^2}{6}$, leading to the following.

$$\begin{aligned} \lambda_{\max}(\mathfrak{S}_1) &\leqslant \lambda_{\max}(C_1^T C_1) + \lambda_{\max}(\mathfrak{D}_1) \\ &\leqslant 1 + \frac{\pi^2}{6} + \lambda_{\max}(D_1)^2 \\ &= 1 + \frac{\pi^2}{6} + 1 = 2 + \frac{\pi^2}{6}. \end{aligned}$$

Finally, $\|\mathcal{A}_1^{-1}\| = \sqrt{\lambda_{\max}(\mathfrak{S}_1)} \leqslant \sqrt{2 + \frac{\pi^2}{6}} = c$, which along with (6.2.4) gives $\kappa(\mathcal{A}_1) = \|\mathcal{A}_1\|\|\mathcal{A}_1^{-1}\| \leqslant cN$. □

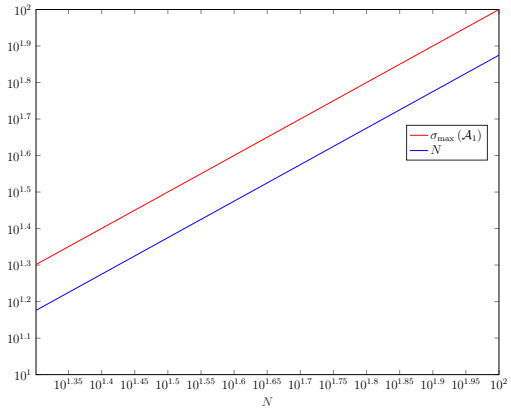
This method depicts spectral convergence, as seen in Figure 6.2a. The schemes derived in this chapter are implemented on **julia**. We take f so that the exact solution is given by

$$u(x) = \sin(\pi x), \tag{6.2.5}$$

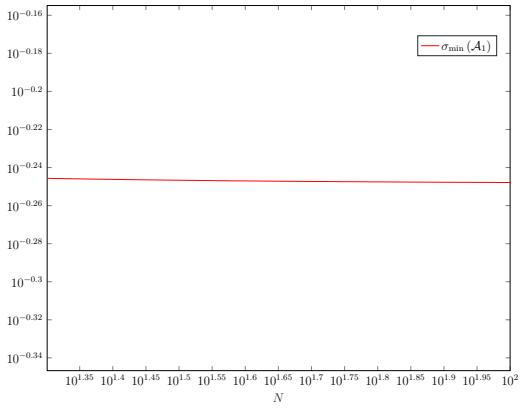
which satisfies the boundary condition. Moreover, he singular value estimates for $\mathcal{A}_1$ are sharp as seen in Figures 6.1a and 6.1b.

Now, consider the same *first order ODE*, with another boundary condition, defined as follows.

$$\begin{aligned} u'(x) &= f(x), \quad x \in (-1, 1) \\ u(1) &= 0. \end{aligned} \tag{6.2.6}$$



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.1: The first order ODE with boundary condition $u(-1) = 0$.

Consider the approximate solution as $\tilde{u}(x) = \sum_{k=0}^{N} u_k T_k(x)$, then the boundary condition stated above implies

$$\sum_{k=0}^{N} u_k T_k(1) = \sum_{k=0}^{N} u_k = 0. \tag{6.2.7}$$

On following the same procedure as before for (6.2.1), the following linear system is obtained by using $(k+1)u_{k+1} = f_k$, for $0 \leqslant k \leqslant N+1$,

$$\begin{bmatrix} 1 & 1 & 1 & 1 & \dots & 1 \\ & 1 & & & & \\ & & 2 & & & \\ & & & 3 & & \\ & & & & \ddots & \\ & & & & & N \end{bmatrix}_{(N+1)\times(N+1)} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ \vdots \\ u_N \end{bmatrix} = \begin{bmatrix} 0 \\ f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-1} \end{bmatrix}, \tag{6.2.8}$$

which we write as $\mathfrak{A}_1 u_h = f_h$, where $\mathfrak{A}_1 \in \mathbb{R}^{(N+1)\times(N+1)}$ is the coefficient matrix of the above system.

Theorem 6.2.2. *For $N \geqslant 4$, let $\mathfrak{A}_1$ be defined by* (6.2.8)*. Then, $\kappa_{sp}(\mathfrak{A}_1) \leqslant cN$ and*

$\kappa(\mathfrak{A}_1) \leqslant cN$.

Proof. Since $\mathfrak{A}_1$ is an upper triangular matrix,

$$\kappa_{sp}(\mathfrak{A}_1) = \frac{\lambda_{\max}(\mathfrak{A}_1)}{\lambda_{\min}(\mathfrak{A}_1)} = \frac{N}{1} = N.$$

On following the proof of Theorem 6.2.1, $\|\mathfrak{A}_1\| \leqslant cN$. Also, it is easy to see that

$$\mathfrak{A}_1^{-1} = \begin{bmatrix} 1 & -1 & -2^{-1} & -3^{-1} & \dots & -N^{-1} \\ & 1 & & & & \\ & & 2^{-1} & & & \\ & & & 3^{-1} & & \\ & & & & \ddots & \\ & & & & & N^{-1} \end{bmatrix}_{(N+1)\times(N+1)}.$$

Thus, following the same procedure as before, $\|\mathfrak{A}_1^{-1}\| \leqslant \sqrt{2 + \frac{\pi^2}{6}}$. Thus,

$$\kappa(\mathfrak{A}_1) = \|\mathfrak{A}_1\|\|\mathfrak{A}_1^{-1}\| \leqslant cN.$$

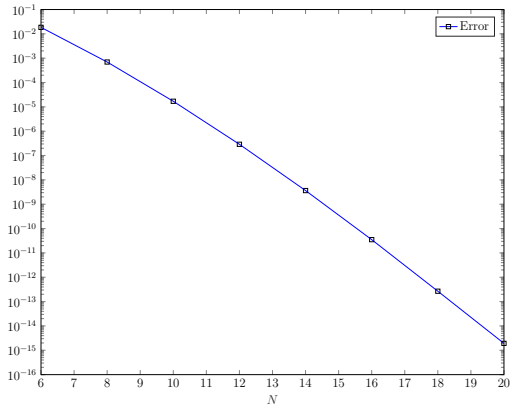
□

This method achieves spectral convergence, as seen in Figure 6.2b. We take f so that the exact solution is defined by (6.2.5), which satisfies the boundary condition. The singular value estimates for $\mathcal{A}_1$ are sharp as seen in Figures 6.3a and 6.3b.

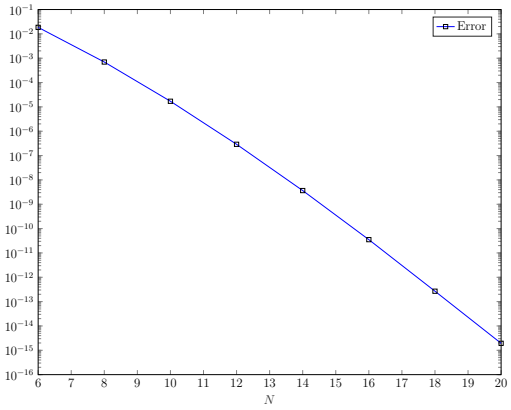
6.2.2 Poisson problem

In this section, we consider the one-dimensional *Poisson problem*:

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

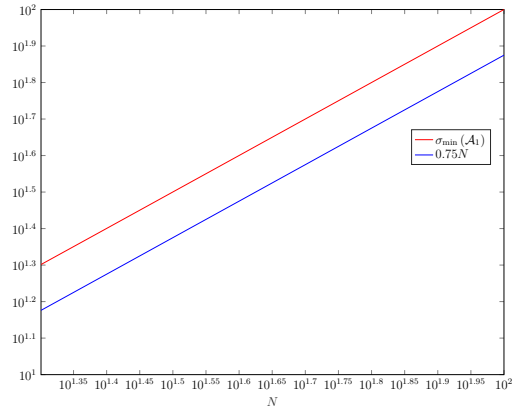


(a) With boundary condition $u(-1) = 0$.

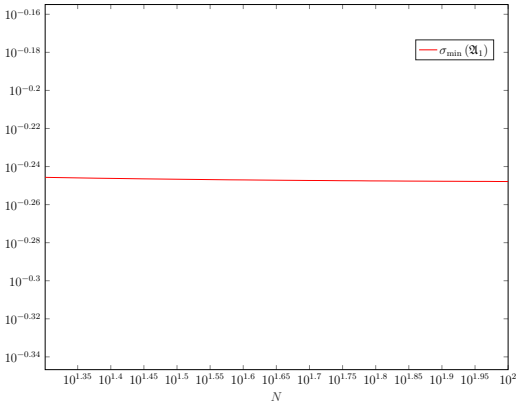


(b) With boundary condition $u(1) = 0$.

Figure 6.2: Convergence of the US method for the first order ODE.



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.3: The first order ODE with boundary condition $u(1) = 0$.

Consider the approximate solution $\sum_{k=0}^{N} u_k T_k(x)$, the boundary condition $u(-1) = 0$, gives

$$\sum_{k=0}^{N} u_k T_k(-1) = \sum_{k=0}^{N} u_k (-1)^k = 0. \tag{6.2.9}$$

Also, the boundary condition $u(1) = 0$ gives,

$$\sum_{k=0}^{N} u_k T_k(1) = \sum_{k=0}^{N} u_k = 0. \tag{6.2.10}$$

Now, we discretize the ODE, $-u'' = f$ by using eqs. (6.1.2) and (6.1.3) for $\lambda = 1$, as follows

$$\begin{aligned}
-\sum_{k=0}^{N} u_k T_k''(x) = \tilde{f}(x) &= \sum_{k=0}^{N-2} f_k C_k^{(2)}(x) \\
-\sum_{k=1}^{N} k u_k \frac{d}{dx} C_{k-1}^{(1)}(x) &= \sum_{k=0}^{N-2} f_k C_k^{(2)}(x) \\
-\sum_{k=2}^{N} k u_k 2 C_{k-2}^{(2)}(x) &= \sum_{k=0}^{N-2} f_k C_k^{(2)}(x) \\
-\sum_{k=0}^{N-2} 2(k+2) u_{k+2} C_k^{(2)}(x) &= \sum_{k=0}^{N-2} f_k C_k^{(2)}(x).
\end{aligned}$$

Comparison of coefficients of $C_k^{(2)}(x)$ yields $-2(k+2)u_{k+2} = f_k$, for $0 \leqslant k \leqslant N-2$, giving the following linear system:

$$\left[\begin{array}{cc|ccccc}
1 & -1 & 1 & -1 & 1 & \dots & (-1)^N \\
1 & 1 & 1 & 1 & 1 & \dots & 1 \\
\hline
 & & -4 & & & & \\
 & & & -6 & & & \\
 & & & & -8 & & \\
 & & & & & \ddots & \\
 & & & & & & -2N
\end{array}\right]_{(N+1)\times(N+1)} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ \vdots \\ u_N \end{bmatrix}_{(N+1)\times 1} = \begin{bmatrix} 0 \\ 0 \\ f_0 \\ f_1 \\ f_2 \\ \vdots \\ f_{N-2} \end{bmatrix}_{(N+1)\times 1}, \tag{6.2.11}$$

which we rewrite as $\mathcal{A}_2 u_h = f_h$.

Theorem 6.2.3. *For $N \geqslant 4$, let $\mathcal{A}_2$ be defined by* (6.2.11)*. Then, $\kappa_{sp}(\mathcal{A}_2) \leqslant cN$ and $\kappa(\mathcal{A}_2) \leqslant cN$.*

Proof. Note that $\mathcal{A}_2$ is a block upper triangular matrix, when partitioned as follows.

$$\mathcal{A}_2 =: \begin{bmatrix} A & B \\ & C \end{bmatrix}.$$

Thus, eigenvalues of $\mathcal{A}_2$ are the eigenvalues of its diagonal blocks, that is, $\Lambda(\mathcal{A}_2) = \Lambda(A) \cup \Lambda(C)$. By a direct calculation, $\Lambda(A) = \frac{2\pm\sqrt{4-8}}{2} = 1 \pm 1i$ and $\Lambda(C) = \{-4, -6, -8, \ldots, -2N\}$. Therefore, $\Lambda(\mathcal{A}_2) = \{1\pm i, -4, -6, -8, \ldots, -2N\}$, $|\lambda|_{\min}(\mathcal{A}_2) = \sqrt{2}$ and $|\lambda|_{\max}(\mathcal{A}_2) = 2N$, thus $\kappa_{sp}(\mathcal{A}_2) = \dfrac{|\lambda|_{\max}(\mathcal{A}_2)}{|\lambda|_{\min}(\mathcal{A}_2)} = \dfrac{2N}{\sqrt{2}} = \sqrt{2}N$.

Now, we estimate the 2-norm condition number of $\mathcal{A}_2$. To this end, note that $\|\mathcal{A}_2\|_\infty = 2N$ and $\|\mathcal{A}_2\|_1 = 2N + 2$, which are achieved by $(N+1)$st row and column, respectively. Thus, $\|\mathcal{A}_2\|_2 \leqslant \sqrt{\|\mathcal{A}_2\|_1\|\mathcal{A}_2\|_\infty} = \sqrt{4N(N+1)} \leqslant cN$.

In order to estimate the minimum singular value of $\mathcal{A}_2$, we calculate $\mathcal{A}_2^{-1}$. It is straightforward to see that for an even $N \geqslant 4$,

$$\mathcal{A}_2^{-1} = \left[\begin{array}{cc|cccccc} 0.5 & 0.5 & 4^{-1} & & 8^{-1} & \ldots & & (2N)^{-1} \\ -0.5 & 0.5 & & 6^{-1} & \ldots & (2N-2)^{-1} & & \\ \hline & & -4^{-1} & & & & & \\ & & & -6^{-1} & & & & \\ & & & & -8^{-1} & & & \\ & & & & & \ddots & & \\ & & & & & & & -(2N)^{-1} \end{array}\right]_{(N+1)\times(N+1)}$$

and for an odd $N \geqslant 4$,

$$\mathcal{A}_2^{-1} = \left[\begin{array}{cc|cccccc} 0.5 & 0.5 & 4^{-1} & & 8^{-1} & \ldots & (2N-2)^{-1} & \\ -0.5 & 0.5 & & 6^{-1} & \ldots & & (2N)^{-1} & \\ \hline & & -4^{-1} & & & & & \\ & & & -6^{-1} & & & & \\ & & & & -8^{-1} & & & \\ & & & & & \ddots & & \\ & & & & & & -(2N)^{-1} & \end{array}\right]_{(N+1)\times(N+1)}$$

For any $N \geqslant 4$, let

$$\mathcal{A}_2^{-1} := \left[\begin{array}{c|c} E_{2\times 2} & F_{2\times n-1} \\ \hline & G_{n-1\times n-1} \end{array}\right],$$

then its corresponding gram matrix is denoted by $\mathfrak{S}_2$ and is simplified as follows.

$$\mathfrak{S}_2 := \mathcal{A}_2^{-T}\mathcal{A}_2^{-1} = \begin{bmatrix} E^T \\ F^T \end{bmatrix} \begin{bmatrix} E & F \end{bmatrix} + \begin{bmatrix} 0 & \\ & G^2 \end{bmatrix} =: H^T H + \mathfrak{G}.$$

Thus $\mathfrak{S}_2$ is the sum of two symmetric matrices defined by the above equation. Now, the non-zero eigenvalues of $H^T H$ and HH^T are equal, so we analyze $HH^T = EE^T + FF^T$, which is facilitated by first calculating the spectrum of EE^T and FF^T. Since

$$EE^T = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix},$$

hence $\Lambda(EE^T) = \{0.5\}$, and it remains to estimate the spectrum of FF^T. For an even $N \geqslant 4$,

$$\begin{aligned} FF^T &= \begin{bmatrix} 4^{-1} & & 8^{-1} & & \dots & & (2N)^{-1} \\ & 6^{-1} & & 10^{-1} & \dots & (2N-2)^{-1} & \end{bmatrix} \begin{bmatrix} 4^{-1} & & \\ & & 6^{-1} \\ 8^{-1} & & \\ & & 10^{-1} \\ \vdots & & \vdots \\ & & (2N-2)^{-1} \\ (2N)^{-1} & & \end{bmatrix} \\ &= \begin{bmatrix} \sum_{i=1}^{\frac{N}{2}} (4i)^{-2} & \\ & \sum_{i=1}^{\frac{N}{2}-1} (2(2i+1))^{-2} \end{bmatrix}. \end{aligned}$$

Thus, for an even $N \geqslant 4$, $\Lambda(FF^T) = \left\{ \sum_{i=1}^{\frac{N}{2}} (4i)^{-2}, \sum_{i=1}^{\frac{N}{2}-1} (2(2i+1))^{-2} \right\}$, whereas for an odd $N \geqslant 4$,

$$FF^T = \begin{bmatrix} \sum_{i=1}^{\frac{N-1}{2}} (4i)^{-2} & \\ & \sum_{i=1}^{\frac{N-1}{2}} (2(2i+1))^{-2} \end{bmatrix}.$$

Thus, for an odd $N \geqslant 4$, $\Lambda(FF^T) = \left\{ \sum_{i=1}^{\frac{N-1}{2}} (4i)^{-2}, \sum_{i=1}^{\frac{N-1}{2}} (2(2i+1))^{-2} \right\}$. Note that for both case of even and odd $N \geqslant 4$, both of the summations in the spectrum of FF^T are less than or equal to $\sum_{i=1}^{\infty} i^{-2} = \frac{\pi^2}{6}$, we get that $\lambda_{\max}(FF^T) \leqslant \frac{\pi^2}{6}$, for all $N \geqslant 4$. Therefore,

$$\begin{aligned} \lambda_{\max}(\mathfrak{S}_2) &\leqslant \lambda_{\max}(H^T H) + \lambda_{\max}(\mathfrak{G}) \\ &= \lambda_{\max}(HH^T) + \lambda_{\max}(G^2) \\ &\leqslant \lambda_{\max}(EE^T) + \lambda_{\max}(FF^T) + 4^{-2} \\ &\leqslant 0.5 + \frac{\pi^2}{6} + 4^{-2} = \frac{\pi^2}{6} + \frac{9}{16}. \end{aligned}$$

Since $\|\mathcal{A}_2^{-1}\| = \sqrt{\lambda_{\max}(\mathfrak{S}_2)} \leqslant \sqrt{\frac{\pi^2}{6} + \frac{9}{16}} = c$. Thus, $\kappa(\mathcal{A}_2) = \|\mathcal{A}_2\|\|\mathcal{A}_2^{-1}\| \leqslant cN$. □

The spectral convergence of this method is evident in Figure 6.4, whereas the sharpness of the bounds is portrayed by Figures 6.5a and 6.5b. For its numerical implementation on julia, we take f so that the exact solution is defined by (6.2.5), which satisfies the boundary condition.

6.2.3 Biharmonic problem

Finally, we consider the *biharmonic problem*:

$$\frac{d^4u}{dx^4} = f, \qquad x \in (-1, 1)$$

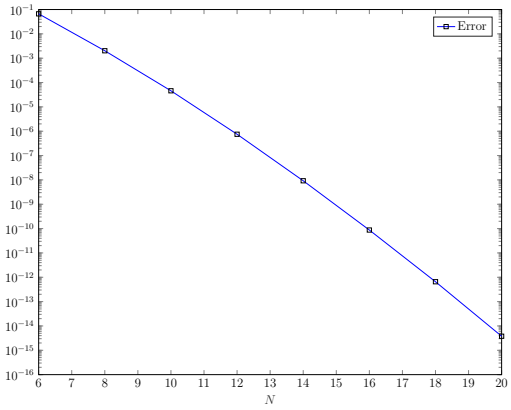
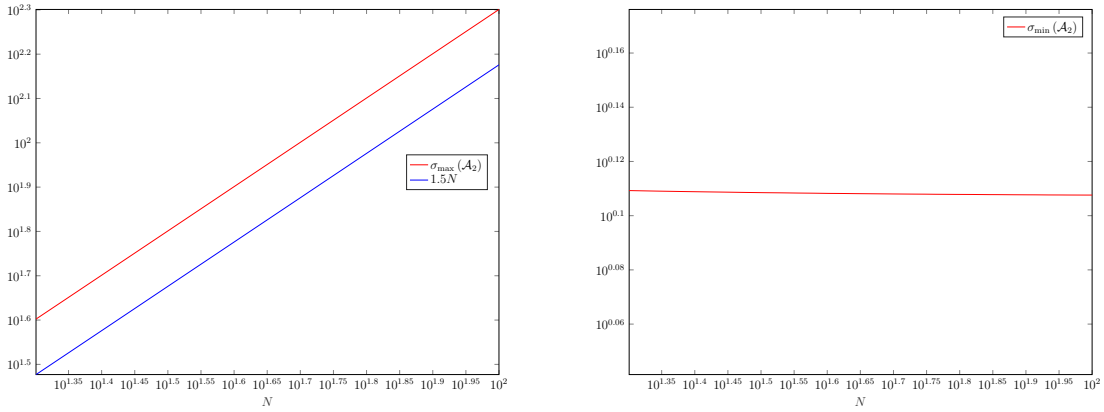


Figure 6.4: Spectral convergence of Poisson problem in one-dimension.



(a) Maximum singular value.

(b) Minimum singular value.

Figure 6.5: The Poisson problem in one-dimension

$$u(\pm 1) = 0,$$

$$u'(\pm 1) = a \in \mathbb{R}.$$

Consider the discrete solution: $\tilde{u}(x) = \sum_{k=0}^{N} u_k T_k(x),$. The boundary conditions $u(\pm 1) = 0$ yield the following two equations,

$$\sum_{k=0}^{N} u_k (\pm 1)^k = 0.$$

For $k \geqslant 1$, $T_k'(\pm 1) = kC_{k-1}^{(1)}(\pm 1) = k \cdot (\pm 1)^{k-1}(k-1+1) = k^2(\pm 1)^{k+1}$, thus the boundary conditions $u'(\pm 1) = a$ give the following two equations,

$$\sum_{k=1}^{N} u_k (\pm 1)^{k+1} k^2 = a.$$

As $\sum_{k=0}^{N} u_k \frac{d^4 T_k(x)}{dx^4} = \tilde{f}(x) = \sum_{k=0}^{N-4} f_k C_k^{(4)}(x)$,

$$\begin{aligned}
\sum_{k=1}^{N} k u_k \frac{d^3}{dx^3} C_{k-1}^{(1)}(x) &= \tilde{f}(x) \\
\sum_{k=2}^{N} k u_k 2 \frac{d^2 C_{k-2}^{(2)}}{dx^2}(x) &= \tilde{f}(x) \\
\sum_{k=3}^{N} k u_k 2^2 \cdot 2 \frac{dC_{k-3}^{(3)}}{dx}(x) &= \tilde{f}(x) \\
\sum_{k=4}^{N} k u_k 2^3 \cdot 2 \cdot 3 C_{k-4}^{(4)}(x) &= \tilde{f}(x) \\
\sum_{k=0}^{N-4} 2^3 \cdot 3!(k+4) u_{k+4} C_k^{(4)}(x) &= \sum_{k=0}^{N-4} f_k C_k^{(4)}(x).
\end{aligned}$$

Comparing coefficients of $C_k^{(4)}(x)$, yields $2^3 \cdot 3! \cdot (k+4) u_{k+4} = f_k$, for $0 \leqslant k \leqslant N-4$, which leads to the following linear system

$$\left[\begin{array}{cccc|cccc}
1 & -1 & 1 & -1 & 1 & -1 & \dots & (-1)^N \\
1 & 1 & 1 & 1 & 1 & 1 & \dots & 1 \\
0 & 1 & -2^2 & 3^2 & -4^2 & 5^2 & \dots & (-1)^{N+1} N^2 \\
0 & 1 & 2^2 & 3^2 & 4^2 & 5^2 & \dots & N^2 \\
\hline
 & & & & 2^3 \cdot 3! \cdot 4 & & & \\
 & & & & & 2^3 \cdot 3! \cdot 5 & & \\
 & & & & & & \ddots & \\
 & & & & & & & 2^3 \cdot 3! \cdot N
\end{array}\right]_{(N+1)\times(N+1)} \begin{bmatrix} u_0 \\ u_1 \\ u_2 \\ u_3 \\ u_4 \\ u_5 \\ \vdots \\ u_N \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ a \\ a \\ f_0 \\ f_1 \\ \vdots \\ f_{N-4} \end{bmatrix}, \tag{6.2.12}$$

which we rewrite as $\mathcal{A}_4 u_h = f_h$.

Theorem 6.2.4. *For $N \geqslant 4$, let $\mathcal{A}_4$ be defined by eq.* (6.2.12)*. Then, $\kappa_{sp}(\mathcal{A}_4) \leqslant cN$ and $\kappa(\mathcal{A}_4) \leqslant cN^4$.*

Proof. Note that $\mathcal{A}_4$ is a block upper triangular matrix, with the following partition

$$\mathcal{A}_4 =: \begin{bmatrix} A & B \\ & C \end{bmatrix}.$$

A straightforward calculation gives $\Lambda(A) = \{11.5250, 0.9259 + 1.0498i, 0.9259 - 1.0498i, -6.3768\}$, thus the set of absolute value of eigenvalues of A is $|\Lambda|(A) = \{11.5250, 1.3998, 1.3998, 6.3768\}$. Thus, $\Lambda(\mathcal{A}_4) = \{11.5250, 0.9259 \pm 1.0498i, -6.3768, 2^3 \cdot 3! \cdot 4, \ldots, 2^3 \cdot 3! \cdot N\}$. Therefore, $|\lambda|_{\min}(\mathcal{A}_4) = 6.3768$ and $|\lambda|_{\max}(\mathcal{A}_4) = 2^3 \cdot 3! \cdot N$, thus $\kappa_{sp}(\mathcal{A}_4) = \dfrac{|\lambda|_{\max}(\mathcal{A}_4)}{|\lambda|_{\min}(\mathcal{A}_4)} = \dfrac{2^3 \cdot 3! \cdot N}{6.3768} \leqslant cN$.

Now, we estimate the 2-norm condition number of $\mathcal{A}_4$. For estimating $\sigma_{\max}(\mathcal{A}_4)$, note that $\|\mathcal{A}_4\|_1 = 1 + 1 + N^2 + N^2 + 2 \cdot 3! \cdot N = 2 + 2N^2 + 2 \cdot 3! \cdot N \leqslant cN^2$, and $\|\mathcal{A}_4\|_\infty = \sum_{i=1}^{N} i^2 = \frac{N(N+1)(2N+1)}{6} \leqslant cN^3$. Therefore,

$$\|\mathcal{A}_4\| \leqslant \sqrt{\|\mathcal{A}_4\|_1 \|\mathcal{A}_4\|_\infty} \leqslant \sqrt{cN^5} = cN^{2.5}.$$

Finally, we estimate $\sigma_{\min}(\mathcal{A}_4)$, by utilizing the expression for $\mathcal{A}_4^{-1}$, given as follows:

$$\mathcal{A}_4^{-1} := \left[\begin{array}{c|c} E & F \\ \hline & G \end{array}\right],$$

where its blocks are defined as follows:

$$E = \begin{bmatrix} 0.5 & 0.5 & 0.125 & -0.125 \\ -0.5624 & 0.5625 & -0.0625 & -0.0625 \\ 0 & 0 & -0.125 & 0.125 \\ 0.0625 & -00625 & 0.0625 & 0.0625 \end{bmatrix},$$

for an even $N \geqslant 4$,

$$F = \begin{bmatrix} \frac{(2^2-1)}{2^3 3!4} & 0 & \frac{(3^2-1)}{2^3 3!6} & 0 & \dots & 0 & \frac{\left(\frac{N}{2}\right)^2-1}{2^3 3!N} \\ 0 & \frac{1}{2^3 3!5}\left[\frac{5^2-1}{8}-1\right] & 0 & \frac{1}{2^3 3!7}\left[\frac{7^2-1}{8}-1\right] & \dots & \frac{1}{2^3 3!(N-1)}\left[\frac{(N-1)^2-1}{8}-1\right] & 0 \\ \frac{-2^2}{2^3 3!4} & 0 & \frac{-3^2}{2^3 3!6} & 0 & \dots & 0 & \frac{-\left(\frac{N}{2}\right)^2}{2^3 3!N} \\ 0 & \frac{5^2-1}{2^3 3!5} & 0 & \frac{7^2-1}{2^3 3!7} & \dots & \frac{(N-1)^2-1}{2^3 3!(N-1)} & 0 \end{bmatrix},$$

for an odd $N \geqslant 4$,

$$F = \begin{bmatrix} \frac{(2^2-1)}{2^3 3!4} & 0 & \frac{(3^2-1)}{2^3 3!6} & 0 & \dots & \frac{\left\lfloor\frac{N}{2}\right\rfloor^2-1}{2^3 3!(N-1)} & 0 \\ 0 & \frac{1}{2^3 3!5}\left[\frac{5^2-1}{8}-1\right] & 0 & \frac{1}{2^3 3!7}\left[\frac{7^2-1}{8}-1\right] & \dots & 0 & \frac{1}{2^3 3!N}\left[\frac{N^2-1}{8}-1\right] \\ \frac{-2^2}{2^3 3!4} & 0 & \frac{-3^2}{2^3 3!6} & 0 & \dots & \frac{-\left\lfloor\frac{N}{2}\right\rfloor^2}{2^3 3!(N-1)} & 0 \\ 0 & \frac{5^2-1}{2^3 3!5} & 0 & \frac{7^2-1}{2^3 3!7} & \dots & 0 & \frac{N^2-1}{2^3 3!N} \end{bmatrix},$$

and for all $N \geqslant 4$,

$$H = \begin{bmatrix} \frac{1}{2^3 3!4} & & & & & \\ & \frac{1}{2^3 3!5} & & & & \\ & & \frac{1}{2^3 3!6} & & & \\ & & & \ddots & & \\ & & & & \frac{1}{2^3 3!N} \end{bmatrix}.$$

For any $N \geqslant 4$, denote the gram matrix of $\mathcal{A}_4^{-1}$ as $\mathfrak{S}_4$ and simplify it as follows.

$$\mathfrak{S}_4 := \mathcal{A}_4^{-T}\mathcal{A}_4^{-1} = \begin{bmatrix} E^T \\ F^T \end{bmatrix} \begin{bmatrix} E & F \end{bmatrix} + \begin{bmatrix} 0 & \\ & G^2 \end{bmatrix} =: H^T H + \mathfrak{G}.$$

Thus, $\mathfrak{S}_4$ is the sum of two symmetric matrices given above. Now, non-zero eigenvalues of $H^T H$ and $HH^T$are equal, so we analyze $HH^T = EE^T + FF^T$. Since $\Lambda(EE^T) = \{0.0060, 0.0293, 0.5332, 0.6502\}$, thus $\lambda_{\max}(EE^T) = 0.6502$. It remains to estimate $\lambda_{\max}(FF^T)$, consequently, for any $N \geqslant 4$, we define FF^T as follows.

$$FF^T = \mathfrak{D} + \mathfrak{A},$$

where

$$\mathfrak{D} = \frac{1}{(2^3 3!)^2} \begin{bmatrix} \sum_{i=2}^{\lfloor \frac{N}{2} \rfloor} \left(\frac{i^2-1}{2i}\right)^2 & 0 & 0 & 0 \\ 0 & \sum_{i=2}^{\lfloor \frac{N-1}{2} \rfloor} \left(\frac{\frac{(2i+1)^2-1}{8}-1}{2i+1}\right)^2 & 0 & 0 \\ 0 & 0 & \sum_{i=2}^{\lfloor \frac{N}{2} \rfloor} \frac{i^4}{(2i)^2} & 0 \\ 0 & 0 & 0 & \sum_{i=2}^{\lfloor \frac{N-1}{2} \rfloor} \frac{\left((2i+1)^2-1\right)^2}{64(2i+1)^2} \end{bmatrix},$$

and,

$$\mathfrak{A} = \begin{bmatrix} O & \mathfrak{F} \\ \mathfrak{F}^T & O \end{bmatrix}.$$

Furthermore, the expression for $\mathfrak{F}$ for any $N \geqslant 4$ is given as

$$\mathfrak{F} = \frac{1}{(2^3 3!)^2} \begin{bmatrix} -\sum_{i=2}^{\lfloor \frac{N}{2} \rfloor} \frac{i^2(i^2-1)}{(2i)^2} & 0 \\ 0 & \sum_{i=2}^{\lfloor \frac{N-1}{2} \rfloor} \frac{\frac{(2i+1)^2-1}{8}\left(\frac{(2i+1)^2-1}{8}-1\right)}{(2i+1)^2} \end{bmatrix}.$$

Since $\mathfrak{D}$ and $\mathfrak{A}$ are symmetric,

$$\lambda_{\max}(FF^T) \leqslant \lambda_{\max}(\mathfrak{D}) + \lambda_{\max}(\mathfrak{A}).$$

Note that $\mathfrak{D}$ is a diagonal matrix and its eigenvalues are bounded above by $\sum_{i=1}^{N} i^2 \leqslant cN^3$, thus $\lambda_{\max}(\mathfrak{D}) \leqslant cN^3$.

Note that the maximum eigenvalue of $\mathfrak{A}$ is $\sigma_{\max}(\mathfrak{F})$. As $\mathfrak{F}$ is diagonal, its singular values are its absolute diagonal entries, which are bounded above by $\sum_{i=1}^{N} i^2 \leqslant cN^3$, thus $\lambda_{\max}(\mathfrak{A}) \leqslant cN^3$. Hence, $\lambda_{\max}(FF^T) \leqslant cN^3$, giving

$$\begin{aligned}
\lambda_{\max}(\mathfrak{S}_4) &\leqslant \lambda_{\max}(H^T H) + \lambda_{\max}(\mathfrak{G}) \\
&= \lambda_{\max}(HH^T) + \lambda_{\max}(G^2) \\
&\leqslant \lambda_{\max}(EE^T) + \lambda_{\max}(FF^T) + (2^3 3!4)^{-2} \\
&\leqslant 0.6502 + cN^3 + (2^3 3!4)^{-2}.
\end{aligned}$$

Now, $\|\mathcal{A}_4^{-1}\| = \sqrt{\lambda_{\max}(\mathfrak{S}_2)} \leqslant \sqrt{cN^3} = cN^{1.5}$. Thus,

$$\kappa(\mathcal{A}_4) = \|\mathcal{A}_4\|\|\mathcal{A}_4^{-1}\| \leqslant cN^{2.5}N^{1.5} = cN^4.$$

□

The spectral convergence of this method is shown in Figure 6.6. We take f so that the exact solution is defined by (6.2.5), which satisfies the boundary condition. The bounds on the singular values of $\mathcal{A}_4$ are evident from Figures 6.7a and 6.7b.

6.3 Time-dependent PDEs

In this section, we extend the ultraspherical spectral method to unsteady problems. In order to analyze the linear systems arising from unsteady problems, the following

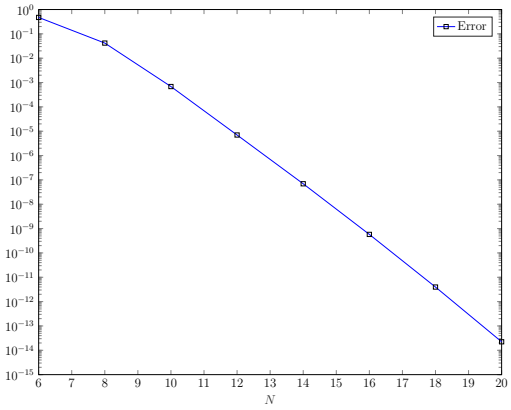
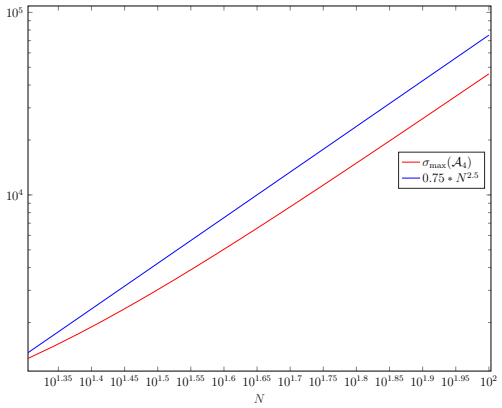
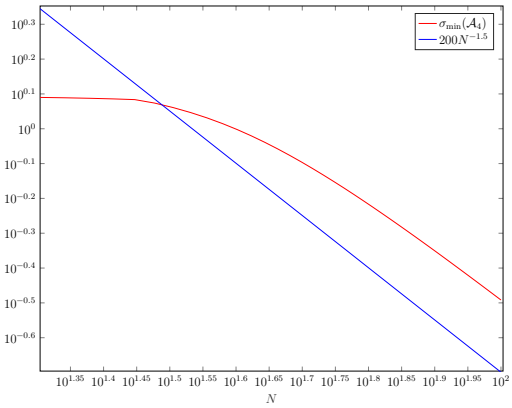


Figure 6.6: Convergence of the US method for the biharmonic equation in one-dimension.



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.7: The Biharmonic problem in one-dimension.

results will be required.

Lemma 6.3.1. *Let $U, V \in \mathbb{R}^{n\times n}$ and $W \in \mathbb{R}^{m\times m}$ be diagonalizable, then $\Lambda(I_m \otimes U + W \otimes V) = \bigcup_{k=1}^{m} \Lambda(U + \lambda_k V)$, where $\lambda_i \in \Lambda(W)$ for $1 \leqslant k \leqslant m$.*

Proof. Let $A = I_m \otimes U + W \otimes V$ and $W = X\mathcal{D}X^{-1}$, where $\mathcal{D}$ is diagonal. Then,

$$A = (X \otimes I_n)(I_m \otimes U + \mathcal{D} \otimes V)(X \otimes I_n)^{-1},$$

therefore $\Lambda(A) = \Lambda(I_m \otimes U + \mathcal{D} \otimes V)$. Note that the eigenvectors of $I_m \otimes U + \mathcal{D} \otimes V$

have the form $e_i \otimes x$, where e_i are elementary basis vectors of $\mathbb{R}^m$ and $x \in \mathbb{R}^n$. Then

$$(I_m \otimes U + \mathcal{D} \otimes V)(e_i \otimes x) = e_i \otimes ((U + \lambda_i V)x),$$

hence the result. □

The above lemma is a result which is often used in practice. The following result was a conjecture for many decades, until it was proved in [67].

Lemma 6.3.2 (See [67])**.** *Let $N \geqslant 1$. Then the real part of every eigenvalue of the pseudospectral Chebyshev derivative matrix $[D]$ is larger than some positive constant independent of N.*

6.3.1 Heat equation

Consider the linear *heat equation*,

$$u_t - u_{xx} = f(x,t) \text{ on } (-1,1)^2, \tag{6.3.1}$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x,-1) = u_0(x)$. We seek a numerical solution, a polynomial of degree N, at $t = 1$.

In [30], the authors presented a simpler way of discretizing the Poisson problem to avoid boundary bordering to implement homogeneous boundary conditions at the end points $x = \pm 1$. They are hard-coded by including a factor of $(1 - x^2)$ in the basis. Furthermore, the ultraspherical polynomials, say $\phi_j(x)$, are carefully chosen so that the following expression can be expressed in terms of $\phi_j(x)$.

$$\frac{d^2}{dx^2}\left((1-x^2)\phi_j(x)\right) = (1-x^2)\phi_j''(x) - 4x\phi_j'(x) - 2\phi_j(x).$$

In [72, Chap. 18], it is given that the normalized ultraspherical polynomial, denoted

by $\widetilde{C}_j^{(\frac{3}{2})}(x) = \sqrt{\dfrac{(j+\frac{3}{2})}{(j+1)(j+2)}} C_j^{(\frac{3}{2})}(x)$, of degree $j \geqslant 0$, satisfies the second-order differential equation, see [72, Table 18.8.1]

$$(1-x^2)\widetilde{C}_j^{(\frac{3}{2})''}(x) - 4x\widetilde{C}_j^{(\frac{3}{2})'}(x) + j(j+3)\widetilde{C}_j^{(\frac{3}{2})}(x) = 0, \quad x \in [-1,1].$$

In particular, this means that $\widetilde{C}^{(\frac{3}{2})}(x)$ is an eigenfunction of the differential operator

$$\frac{d^2}{dx^2}\left((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\right) = -(j(j+3)+2)\widetilde{C}_j^{(\frac{3}{2})}(x), \quad j \geqslant 0.$$

Thus, the appropriate choice is $\phi_j = \widetilde{C}_j^{(\frac{3}{2})}$. Note that [72, (18.9.8)] for $\lambda = \frac{1}{2}$ implies that $\widetilde{C}_n^{(\frac{3}{2})}(x)$ is a scalar multiple of the recombined Legendre basis $L_j - L_{j+2}$. Therefore, the *derivative matrix* is a diagonal matrix with entries, $\mathfrak{D}_{j,j} = -(j(j+3)+2)$. We denote $\mathfrak{D} = -\mathcal{D}$ and prove the following result.

Lemma 6.3.3. *For $N \geqslant 4$, $\lambda_{\max}(\mathfrak{D}) \leqslant cN^2$ and $\lambda_{\min}(\mathfrak{D}) \geqslant c$.*

Proof. Since $\mathfrak{D}$ is a diagonal matrix of size $(N+1)\times(N+1)$,

$$\lambda_{\max}(\mathfrak{D}) = \max_{0\leqslant j\leqslant N}(j(j+3)+2)) \leqslant cN^2,$$
$$\lambda_{\min}(\mathfrak{D}) = \min_{0\leqslant j\leqslant N}(j(j+3)+2)) \geqslant c.$$

□

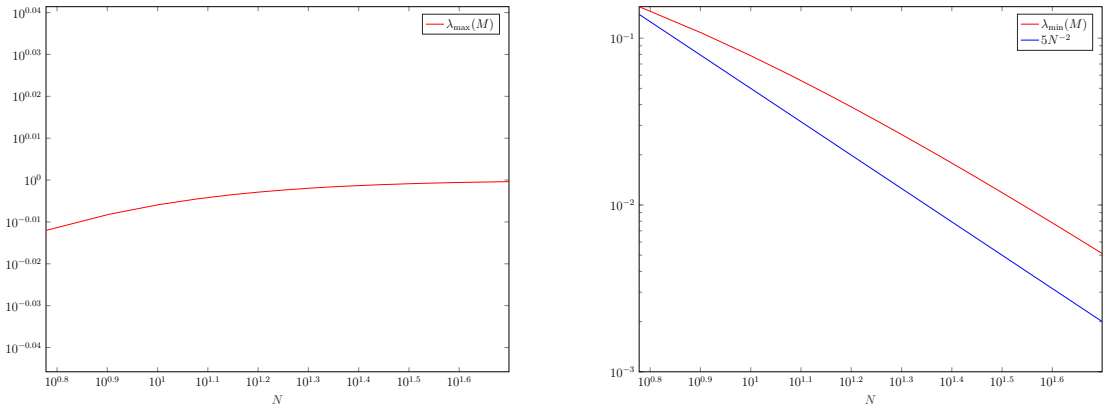
Define M to be the matrix that represents multiplication by $(1-x^2)$ in the $C^{(\frac{3}{2})}$ basis. Since the recurrence relation for the unnormalized ultraspherical polynomials, $C^{(\frac{3}{2})}$, is given by [72, (18.9.7) & (18.9.8)],

$$(1-x^2)C_j^{(\frac{3}{2})}(x) = -(2j+1)(2j+3)(2j+5)(2j+1)C_{j+2}^{(\frac{3}{2})}(x) - 2(2j+3)C_j^{(\frac{3}{2}))}(x) + (2j+5)C_{j-2}^{(\frac{3}{2})}(x).$$

Thus, M is symmetric with non-zero main diagonals and 2 and -2 diagonals, defined as follows.

$$M_{j,k} = \begin{cases} \dfrac{2(j+1)(j+2)}{(2j+1)(2j+5)}, & k = j, \\ \dfrac{-1}{(2j+3)(2j+5)}\sqrt{\dfrac{(j+4)!(2j+3)}{j!(2j+7)}}, & k = j+2. \end{cases} \tag{6.3.2}$$

We prove the following bounds on the spectrum of $M \in \mathbb{R}^{(N+1)\times(N+1)}$, which were observed through numerical experiments depicted in Figures 6.8a and 6.8b.



(a) Maximum eigenvalue.

(b) Minimum eigenvalue.

Figure 6.8: Bounds for the spectrum of M.

Lemma 6.3.4. *For $N \geqslant 4$, $\lambda_{\max}(M) \leqslant c$ and $\lambda_{\min}(M) \geqslant \dfrac{c}{N^2}$.*

Proof. Let $x \in \mathbb{R}^{N+1}\backslash\{0\}$, since M is symmetric, its eigenvalues are estimated as follows.

$$\begin{aligned} x^T M x &= \sum_{i=0}^{N}\sum_{j=0}^{N} x_i m_{ij} x_j \\ &= \sum_{j=0}^{N} x_j^2 m_{jj} + 2\sum_{j=0}^{N-2} x_j x_{j+2} m_{j,j+2} && \text{(by (3.2.2))} \\ &= \sum_{j=0}^{N} x_j^2 \frac{2(j+1)(j+2)}{(2j+1)(2j+5)} - 2\sum_{j=0}^{N-2} \frac{x_j x_{j+2}}{(2j+3)(2j+5)}\sqrt{\frac{(j+4)!(2j+3)}{j!(2j+7)}} && (6.3.3) \end{aligned}$$

$$\leqslant \sum_{j=0}^{N} x_j^2 + 2\sum_{j=0}^{N-2} x_j x_{j+2} \frac{(j+4)^2}{(2j+3)(2j+5)}$$
$$\leqslant \sum_{j=0}^{N} x_j^2 + 2\sum_{j=0}^{N-2} x_j x_{j+2} \leqslant 3\sum_{j=0}^{N} x_j^2, \tag{6.3.4}$$

where the last inequality is achieved by applying the Cauchy Schwarz inequality. Thus, $\lambda_{\max}(M) \leqslant C$. Note that

$$\begin{aligned}
&2\sum_{j=0}^{N-2} \frac{x_j x_{j+2}}{(2j+3)(2j+5)} \sqrt{\frac{(j+4)!(2j+3)}{j!(2j+7)}} \\
&\leqslant 2\sum_{j=0}^{N-2} \frac{\sqrt{(j+1)(j+2)(2j+9)}|x_j|}{\sqrt{(2j+3)}\cdot(2j+5)} \cdot \frac{\sqrt{(j+3)(j+4)}|x_{j+2}|}{\sqrt{(2j+7)(2j+9)}} \\
&\leqslant \sum_{j=0}^{N-2} \frac{(j+1)(j+2)(2j+9)|x_j|^2}{(2j+3)\cdot(2j+5)^2} + \sum_{j=0}^{N-2} \frac{(j+3)(j+4)|x_{j+2}|^2}{(2j+7)(2j+9)} \\
&= \sum_{j=0}^{N-2} \frac{(j+1)(j+2)(2j+9)x_j^2}{(2j+3)\cdot(2j+5)^2} + \sum_{j=2}^{N} \frac{(j+1)(j+2)x_j^2}{(2j+3)(2j+5)} \\
&\leqslant 2\sum_{j=0}^{N} (j+1)(j+2)\left(\frac{(2j+9)}{(2j+3)\cdot(2j+5)^2} + \frac{1}{(2j+3)(2j+5)}\right) x_j^2 \\
&= 2\sum_{j=0}^{N} \frac{(j+1)(j+2)(4j+14)}{(2j+3)(2j+5)^2} x_j^2
\end{aligned}$$

thus by (6.3.3),

$$\begin{aligned}
x^T M x &\geqslant 2\sum_{j=0}^{N} (j+1)(j+2)\left(\frac{1}{(2j+1)(2j+5)} - \frac{(4j+14)}{(2j+3)(2j+5)^2}\right) x_j^2 \\
&\geqslant 4\sum_{j=0}^{N} x_j^2 \frac{4j^2+12j+1}{(2j+1)(2j+3)(2j+5)^2} \\
&\geqslant 4\sum_{j=0}^{N} x_j^2 \frac{(i+1)^2}{(2j+1)(2j+3)(2j+5)^2} \\
&\geqslant \frac{c}{N^2}\sum_{j=0}^{N} x_j^2,
\end{aligned}$$

hence the desired result. □

First Scheme: US in both space and time

We first formulate an ultraspherical scheme in space and time for the *heat equation*, that is, we consider the approximate solution as follows

$$u \approx \sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k(t), \quad (x,t) \in (-1,1)^2, \tag{6.3.5}$$

where $L_k(x)$ represent the Legendre polynomial of degree k. Recall that $C^{(\frac{1}{2})}(x)$ are the Legendre polynomials. Thus, we have the following relation by (6.1.3),

$$\frac{d}{dt}L_k(t) = \frac{d}{dt}C_k^{(\frac{1}{2})}(t) = C_{k-1}^{(\frac{3}{2})}(t), \quad k \geqslant 1.$$

This allows us to find the discretization of the heat equation as follows. We begin with the projected initial condition:

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k(-1) = \sum_{j=0}^{N} u_j^0\widetilde{C}_j^{(\frac{3}{2})}(x),$$
$$\sum_{j=0}^{N}\sum_{k=0}^{N} (-1)^k u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x) = \sum_{j=0}^{N} u_j^0\widetilde{C}_j^{(\frac{3}{2})}(x).$$

Define $u_h = [u_{00}; u_{10}; \ldots; u_{N0}; u_{01}; \ldots; u_{NN}] \in \mathbb{R}^{(N+1)^2\times 1}$, $u_{oh} = [u_0^0; u_1^0; \ldots; u_N^0] \in \mathbb{R}^{(N+1)\times 1}$, and

$$\mathscr{B} = \begin{bmatrix} 1 & -1 & 1 & \ldots & (-1)^N \end{bmatrix} \in \mathbb{R}^{1\times(N+1)},$$

then the initial condition gives

$$(\mathscr{B}\otimes M)\, u_h = u_{oh}. \tag{6.3.6}$$

Now, recall (6.1.4), which for $\lambda = \frac{1}{2}$ gives the conversion operator for converting $C_j^{(\frac{1}{2})}$ or L_j basis to $C^{(\frac{3}{2})}$. We denote it by $\mathcal{S}_{\frac{1}{2}}$ and it is given as follows,

$$\mathcal{S}_{\frac{1}{2}} = \begin{bmatrix} 1 & -\frac{1}{5} & & & & \\ & \frac{1}{3} & -\frac{1}{7} & & & \\ & \ddots & & \ddots & & \\ & & \frac{1}{2N-3} & & -\frac{1}{2N+1} & \\ & & & \frac{1}{2N-1} & & \\ & & & & \frac{1}{2N+1} & \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \tag{6.3.7}$$

Let us now discretize the heat equation $u_t - u_{xx} = f(x,t)$ as follows,

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}\left((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k'(t) - \left((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\right)'' L_k(t)\right) = \tilde{f}(x,t)$$

$$\sum_{j=0}^{N}\sum_{k=0}^{N-1} u_{j,k+1}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)C_k^{(\frac{3}{2})}(t) + \sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(j(j+3)+2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k(t) = \tilde{f}(x,t),$$

where

$$\tilde{f}(x,t) = \sum_{j=0}^{N}\sum_{k=0}^{N} f_{jk}\widetilde{C}_j^{(\frac{3}{2})}(x)C_k^{(\frac{3}{2})}(t).$$

On changing the basis in time from L_k to $C_k^{(\frac{3}{2})}$ in the second summation of the above equation, the following linear system is obtained,

$$\left(J \otimes M + S_{\frac{1}{2}} \otimes \mathfrak{D}\right) u_h = f_h,$$

where J is a 1-diagonal matrix, defined as,

$$J = \begin{bmatrix} 0 & 1 & & & \\ & 0 & 1 & & \\ & & \ddots & \ddots & \\ & & & 0 & 1 \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)},$$

and $f_h = [f_{00}; f_{10}; \ldots; f_{N,0}; f_{01}; \ldots; f_{NN}] \in \mathbb{R}^{(N+1)^2\times 1}$.

Now, we need to incorporate the initial conditions. Multiplying (6.3.6) by $\mathbf{e}_{n+1} \otimes I_{N+1}$, where $\mathbf{e}_{n+1} = [0; 0; \ldots; 1] \in \mathbb{R}^{(N+1)\times 1}$, and adding or subtracting it to the above equation, we obtain the following *discrete heat equation*,

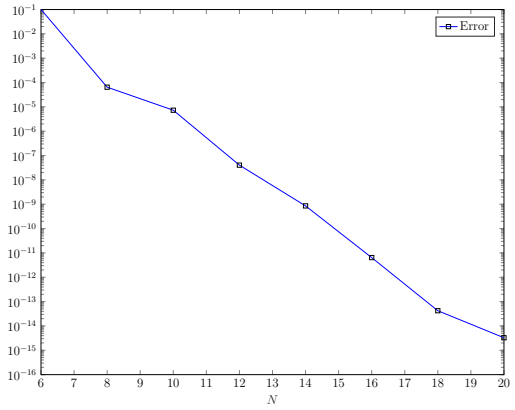
$$\left((J \pm \mathbf{e}_{n+1}\mathscr{B}) \otimes M + S_{\frac{1}{2}} \otimes \mathfrak{D}\right) u_h = f_h \pm (\mathbf{e}_{n+1} \otimes I_{N+1}) u_{oh}. \tag{6.3.8}$$

This scheme portrays spectral convergence in both space and time, as seen in Figures 6.9a and 6.9b for '$-$' and '$+$' sign, respectively. All the schemes in this section are also implemented on **julia**. For both of these schemes, we take f so that the exact solution is

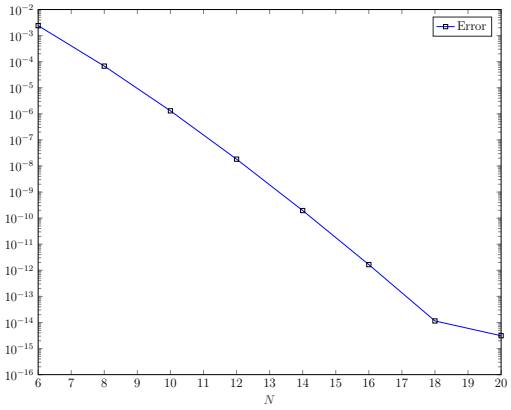
$$u(x, y, t) = e^t \sin(\pi t), \tag{6.3.9}$$

satisfying the boundary conditions, and giving initial condition as $u(x, y, -1) = e^{-1}\sin(\pi x)$. Denote coefficient matrix of the system given in (6.3.8), by considering '$+$' sign, i.e., the *global space-time spectral operator for the heat equation*, by $\mathcal{A}_h = (J + \mathbf{e}_{n+1}\mathscr{B}) \otimes M + S_{\frac{1}{2}} \otimes \mathfrak{D} \in \mathbb{R}^{(N+1)^2\times(N+1)^2}$. Figures 6.10a and 6.10b suggest that the 2-norm condition number of $\mathcal{A}_h$, $\kappa(\mathcal{A}_h) \leqslant cN^3$, for all $6 \leqslant N \leqslant 50$. Note that it is sufficient to consider $N \leqslant 50$, because typically for us $N \lesssim 20$.

Remark 6.3.5. Note that for the case of a k-th order ODE and discretization parameter $n = N + 1$, we get a $n - k \times n$ order linear system and define the first k

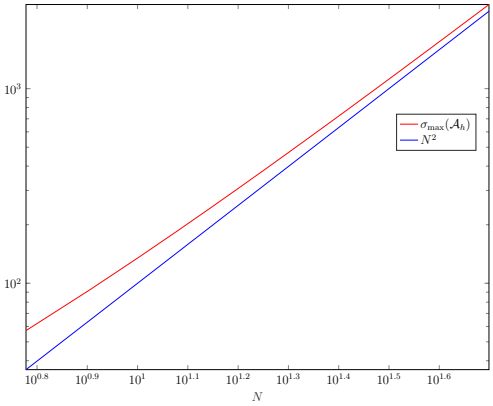


(a) Considering '−' sign.

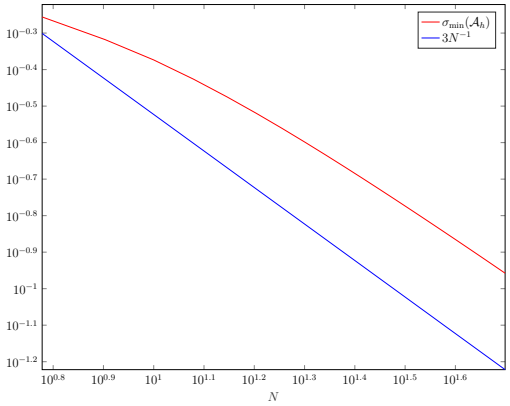


(b) Considering '+' sign.

Figure 6.9: Convergence of the US method in space and time for the heat equation.



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.10: Bounds for singular values of $\mathcal{A}_h$.

rows as the boundary condition to get a square linear system of order n. However, for the case of a time-dependent PDE such as (6.3.8), the equation $u_t - u_{xx} = f(x,t)$ on $(x,t) \in (-1,1)^2$ returns a square linear system of order $(N+1)^2$. Moreover, the initial condition $u(x,-1) = u_0(x)$ implies (6.3.6), which is $(N+1)$ equations in $(N+1)^2$ unknowns. Thus, we add or subtract their contribution to the last row of all zeros of J, which generates the last $(N+1)$ rows (containing all zeros) of the term corresponding to the discretization of u_t, i.e., $J \otimes M$.

Second scheme: Collocation in time and US in space

In this section, we present another scheme for solving the *heat equation*, given by (6.3.1), demonstrating spectral convergence in both space and time. In this method, we incorporate the US method in space and collocation in time, by defining an approximation for u as follows,

$$\tilde{u} = \sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\ell_k(t), \quad (x,t) \in (-1,1)^2, \tag{6.3.10}$$

where $\widetilde{C}_j^{(\frac{3}{2})}$ are the rescaled ultraspherical polynomials of order $\frac{3}{2}$ as defined in this section and ℓ_k represents the Lagrange polynomial of degree $N+1$ for the Chebyshev Gauss-Lobatto nodes given by $t_k = \cos\frac{(N-\pi)k}{N}$, for $0 \leqslant k \leqslant N$. Note that we considered these nodes for convenience of analysis. In practice, any other Gauss quadrature nodes can be used. Substituting the approximate solution (6.3.10) to (6.3.1) yields,

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}\left((1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\ell_k'(t) - \left((1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\right)''\ell_k(t)\right) = \tilde{f}(x,t).$$

On collocating at time $t = t_m$, for $1 \leqslant m \leqslant N$,

$$\begin{aligned}
&\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}\big((1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\ell_k'(t_m) - \left((1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\right)''\delta_{km}\big) = \tilde{f}(x,t_m)\\
&\sum_{j=0}^{N}\sum_{k=1}^{N} u_{jk}\big((1-x^2)\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\ell_k'(t_m) + (j(j+3)+2)\widetilde{C}_j^{(\frac{3}{2})}(x)\delta_{km}\big) = \tilde{f}(x,t_m)\\
&\qquad\qquad - \sum_{j=0}^{N}(1-x^2)u_{j0}\widetilde{C}_j^{\left(\frac{3}{2}\right)}(x)\ell_0'(t_m).
\end{aligned}$$

For $1 \leqslant k \leqslant N$, let $\tilde{f}(x,t_k) = \sum_{j=0}^{N} f_j^k \widetilde{C}_j^{(\frac{3}{2})}(x)$, and define $f_h = [f_h^1; f_h^2; \ldots; f_h^N]$, where $f_h^k = [f_0^k; f_2^k; \ldots; f_N^k]$. Thus, definition of $M \in \mathbb{R}^{(N+1)\times(N+1)}$ and $\mathfrak{D} \in \mathbb{R}^{(N+1)\times(N+1)}$

give the following *discrete heat equation*

$$\left(\left[\!\left[D\right]\!\right] \otimes M + I_N \otimes \mathfrak{D}\right) u_h = f_h - d_{0h} \otimes \left(M u_{0h}\right). \tag{6.3.11}$$

Denote coefficient matrix of the above system, i.e., the *global space-time spectral operator for the heat equation*, by $\mathscr{A}_h = \left[\!\left[D\right]\!\right] \otimes M + I_N \otimes \mathfrak{D} \in \mathbb{R}^{N(N+1)\times N(N+1)}$. Our next goal is to prove a spectral condition number estimate for $\mathscr{A}_h$, which is given by the following result.

Theorem 6.3.6. *Let $N \geqslant 4$, then $\kappa_{sp}(\mathscr{A}_h) \leqslant cN^2$.*

Proof. Observe that, Lemma 6.3.1 gives $\Lambda(\mathscr{A}_h) = \bigcup_{k=1}^{N} \Lambda(\mathfrak{D} + \lambda_k M)$. Let $\lambda \in \Lambda(\left[\!\left[D\right]\!\right]) \subseteq \mathbb{C}$, and M and $\mathfrak{D}$ are SPD, with the latter diagonal. For some $x \in \mathbb{R}^{N+1}$, so that $|x| = 1$, the absolute value of the eigenvalues can be estimated as follows.

$$\begin{aligned}
|x^T \mathfrak{D} x + \lambda x^T M x| &= |(x^T \mathfrak{D} x + \Re\lambda \cdot x^T M x) + i \cdot \Im\lambda \cdot (x^T M x)| \\
&\geqslant |(x^T \mathfrak{D} x + \Re\lambda \cdot x^T M x)| \\
&= x^T \mathfrak{D} x + \Re\lambda \cdot x^T M x \\
&\geqslant \lambda_{\min}(\mathfrak{D}) + c\lambda_{\min}(M) \\
&\geqslant c + \frac{c}{N^2} \geqslant c,
\end{aligned}$$

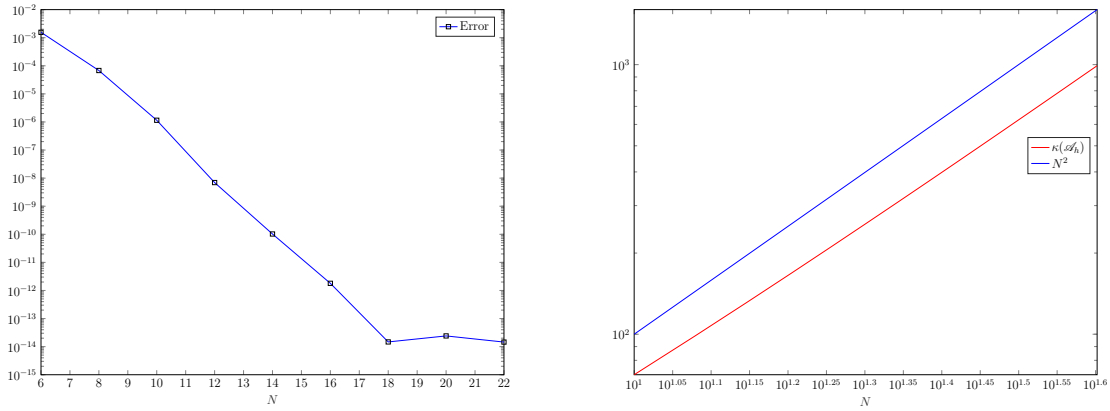
thus $|\lambda|_{\min}(\mathcal{A}_h) \geqslant c$. Also, since $|\lambda(\left[\!\left[D\right]\!\right])| \leqslant \sigma_{\max}(\left[\!\left[D\right]\!\right])$,

$$\begin{aligned}
|x^T \mathfrak{D} x + \lambda x^T M x| &\leqslant \lambda_{\max}(\mathfrak{D}) + \sigma_{\max}(\left[\!\left[D\right]\!\right])\lambda_{\max}(M) \\
&\leqslant cN^2 + cN^2 \leqslant cN^2,
\end{aligned}$$

implying $|\lambda|_{\max}(\mathscr{A}_h) \leqslant cN^2$, hence, $\kappa_{sp}(\mathscr{A}_h) \leqslant cN^2$. □

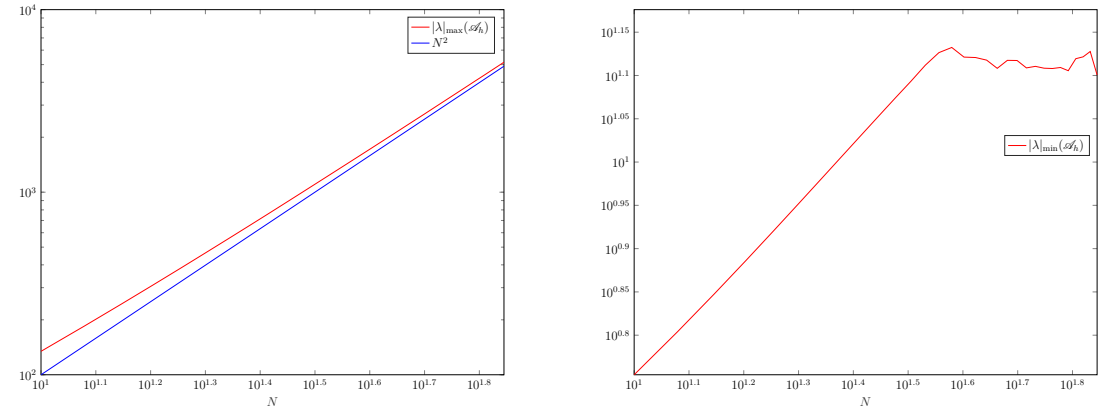
This scheme converges *spectrally* in both space and time, which is verified by result of a numerical experiment shown in Figure 6.11a. For this scheme, we take f so

that the exact solution is defined by (6.3.9), which satisfies the boundary conditions, and giving initial condition as $u(x, y, -1)$. Note that the sharpness of estimates derived in the above result is evident from Figures 6.12a and 6.12b, and is a sharp estimate for the 2-norm condition number estimate for $\mathscr{A}_h$, $\kappa(\mathscr{A}_h)$, as given by Figure 6.11b.



(a) Convergence. (b) 2-norm condition number of $\mathscr{A}_h$.

Figure 6.11: The US method in both space and collocation in time for the heat equation.



(a) Maximum absolute value of eigenvalue. (b) Minimum absolute value of eigenvalue.

Figure 6.12: Bounds for absolute value of eigenvalues of $\mathscr{A}_h$.

6.3.2 Schrödinger equation

The linear *Schrödinger equation* is

$$u_t - iu_{xx} = f(x,t), \quad (x,t) \in (-1,1)^2, \tag{6.3.12}$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x,-1) = u_0(x)$. Here $i = \sqrt{-1}$.

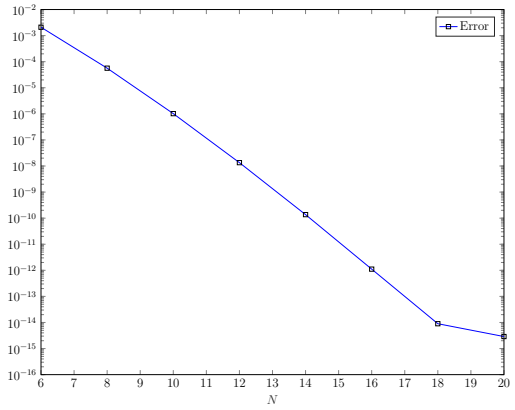
The two space-time spectral method schemes for the Schröndinger equation are analogous to the ones derived for the heat equation in the previous section. However, the presence of $i = \sqrt{-1}$, motivates our interest in analyzing the schemes for the Schrödinger equation.

First Scheme: US in both space and time

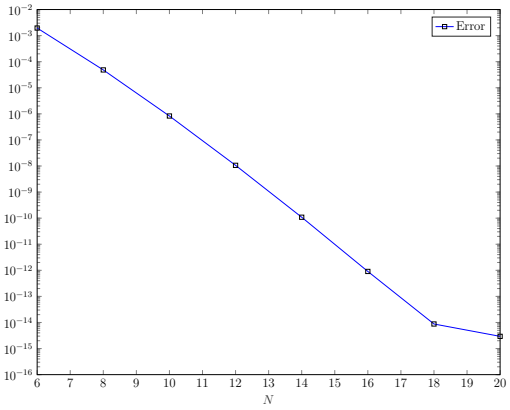
The approximation of u given by (6.3.5) in (6.3.12) leads to the following *discrete Schrödinger equation*,

$$\left((J \pm \mathbf{e}_{n+1}\mathscr{B}) \otimes M + iS_{\frac{1}{2}} \otimes \mathfrak{D}\right) u_h = f_h \pm (\mathbf{e}_{n+1} \otimes I_{N+1}) u_{oh},$$

where the constituting matrices and vectors are the same as those defined for (6.3.8). These schemes demonstrate spectral convergence in both space and time as seen in Figure 6.13b. For both of these schemes, we take f so that the exact solution is defined by (6.3.9), which satisfies the boundary conditions, and gives initial condition as $u(x,y,-1)$. Define the *global space-time spectral operator* $\mathcal{A}_s := (J \pm \mathbf{e}_{n+1}\mathscr{B}) \otimes M + iS_{\frac{1}{2}} \otimes \mathfrak{D} \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$. Figures 6.14a and 6.14b suggest that the condition number of $\mathcal{A}_s$, $\kappa(\mathcal{A}_s) \leqslant cN^{3.5}$, for all $6 \leqslant N \leqslant 40$.

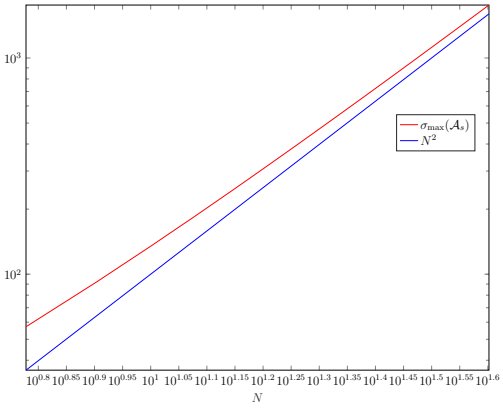


(a) Considering '−' sign.

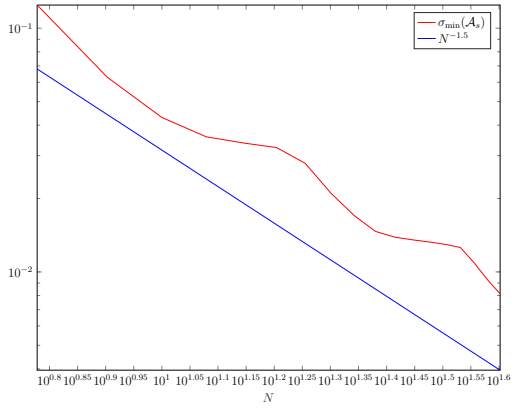


(b) Considering '+' sign.

Figure 6.13: Convergence of the US method in both space and time for the Schrödinger equation.



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.14: Bounds for singular values of $\mathcal{A}_s$.

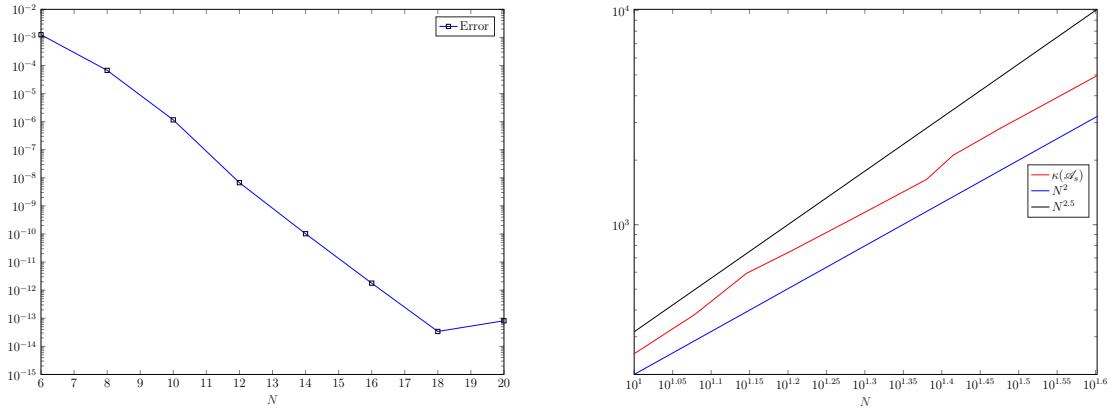
Second scheme: Collocation in time and US in space

For this scheme, the approximation of u defined by (6.3.10) in (6.3.12) yields the following *discrete Schrödinger equation*,

$$\left([D] \otimes M + iI_N \otimes \mathfrak{D}\right) u_h = f_h - d_{0h} \otimes (Mu_{0h}),$$

where the constituting matrices and vectors are the same as those defined for (6.3.11). The spectral convergence in space and time for this scheme is evident from Fig-

ure 6.15a. For this scheme, we take f so that the exact solution is defined by (6.3.9), which satisfies the boundary conditions, and gives initial condition as $u(x, y, -1)$. Define the *global space-time spectral operator* for this scheme as $\mathscr{A}_s = [D] \otimes M + iI_N \otimes \mathfrak{D} \in \mathbb{C}^{N(N+1) \times N(N+1)}$. Similar to the case of the heat equation, this scheme is slightly better conditioned than the previous one for the Schrödinger equation, as Figure 6.15b suggests that $\kappa(\mathscr{A}_s) \leqslant cN^{2.5}$, as compared to $\kappa(\mathcal{A}_s) \leqslant cN^{3.5}$.



(a) Convergence.

(b) 2-norm condition number of $\mathscr{A}_s$.

Figure 6.15: The US method in both space and collocation in time for the Schrödinger equation.

6.3.3 Wave equation

Consider the linear *wave equation*,

$$u_{tt} = u_{xx} + f(x, t), \qquad \text{on } (-1, 1)^2, \tag{6.3.13}$$

with boundary conditions $u(\pm 1, t) = 0$ and initial conditions $u(x, -1) = u_0(x)$ and $u_t(x, -1) = u_1(x)$.

First Scheme: US in both space and time

We first formulate an ultraspherical in space and time scheme for the wave equation, that is, we consider the approximate solution as defined by (6.3.5). Since the wave equation has second-order partial derivative w.r.t. time (t), (6.1.3) implies,

$$\frac{d^2}{dt^2}L_k(t) = 3C_{k-2}^{(\frac{5}{2})}(t), \quad k \geqslant 2,$$

and $\dfrac{d^2}{dt^2}L_k(t) = 0$ for $k = 0, 1$. This allows us to find the discretization of the wave equation as follows. The first initial condition is obtained analogously as (6.3.6), thus

$$(\mathscr{B} \otimes M)\, u_h = u_{oh}. \tag{6.3.14}$$

Since $u_t(x, -1) = u_1(x)$,

$$\sum_{j=0}^{N}\sum_{k=1}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)C_{k-1}^{(\frac{3}{2})}(-1) = \sum_{j=0}^{N} u_j^0 \widetilde{C}_j^{(\frac{3}{2})}(x),$$
$$\sum_{j=0}^{N}\sum_{k=0}^{N-1} u_{j,k+1}\frac{(-1)^k(k+1)(k+2)}{2}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x) = \sum_{j=0}^{N} u_j^0 \widetilde{C}_j^{(\frac{3}{2})}(x).$$

Define $u_h = [u_{00}; u_{10}; \ldots; u_{N0}; u_{01}; \ldots; u_{NN}] \in \mathbb{R}^{(N+1)^2\times 1}$, $u_{1h} = [u_0^1; u_1^1; \ldots; u_N^1] \in \mathbb{R}^{(N+1)\times 1}$, and

$$\mathscr{B}_t = \begin{bmatrix} 0 & 1 & -3 & 6 & -10 & \ldots & (-1)^{N-1}\dfrac{N(N+1)}{2} \end{bmatrix} \in \mathbb{R}^{1\times(N+1)},$$

then the initial condition gives

$$(\mathscr{B}_t \otimes M)\, u_h = u_{1h}. \tag{6.3.15}$$

Now, recall (6.1.4), which for $\lambda = \frac{3}{2}$ gives the conversion operator for converting $C_j^{(\frac{3}{2})}$ basis to $C_j^{(\frac{5}{2})}$. We denote it by $\mathcal{S}_{\frac{3}{2}}$ and it is given as follows,

$$\mathcal{S}_{\frac{3}{2}} = \begin{bmatrix} 1 & -\frac{3}{7} & & & & & \\ & \frac{3}{5} & -\frac{1}{3} & & & & \\ & & \frac{3}{7} & -\frac{3}{11} & & & \\ & & \ddots & & \ddots & & \\ & & & \frac{3}{2N-1} & & -\frac{3}{2N+3} & \\ & & & & \frac{3}{2N+1} & & \\ & & & & & \frac{3}{2N+3} & \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)}. \tag{6.3.16}$$

Let us now discretize the wave equation $u_{tt} - u_{xx} = f(x,t)$ as follows.

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}\left((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k''(t) - \left((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\right)'' L_k(t)\right) = \tilde{f}(x,t)$$

$$\sum_{j=0}^{N}\sum_{k=0}^{N-2} 3u_{j,k+2}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)C_k^{(\frac{5}{2})}(t) + \sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(j(j+3)+2)\widetilde{C}_j^{(\frac{3}{2})}(x)L_k(t) = \tilde{f}(x,t),$$

where

$$\tilde{f}(x,t) = \sum_{j=0}^{N}\sum_{k=0}^{N} f_{jk}\widetilde{C}_j^{(\frac{3}{2})}(x)C_k^{(\frac{5}{2})}(t).$$

On changing the basis in time from L_k to $C_k^{(\frac{5}{2})}$ in the second summation of the above equation, the following linear system is obtained,

$$\left(\mathcal{J}\otimes M + S_{\frac{3}{2}}S_{\frac{1}{2}}\otimes\mathfrak{D}\right)u_h = f_h,$$

where J is a 1-diagonal matrix, defined as,

$$\mathcal{J} = \begin{bmatrix} 0 & 0 & 3 & & & \\ & 0 & 0 & 3 & & \\ & & & \ddots & \ddots & \\ & & & 0 & 0 & 3 \\ & & & & 0 & 0 \\ & & & & & 0 \end{bmatrix} \in \mathbb{R}^{(N+1)\times(N+1)},$$

and $f_h = [f_{00}; f_{10}; \ldots; f_{N,0}; f_{01}; \ldots; f_{NN}] \in \mathbb{R}^{(N+1)^2\times 1}$. Now, we need to incorporate the initial conditions by using eqs. (6.3.14) and (6.3.15) and $\mathbf{e} \otimes I_{N+1}$, defined as

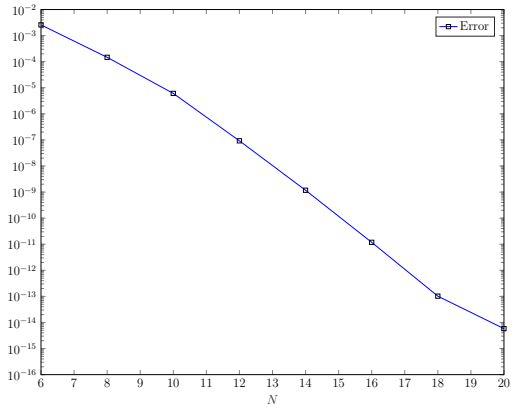
$$\mathbf{e} = \begin{bmatrix} O_{N-1,2} \\ \hline I_{2,2} \end{bmatrix} \in \mathbb{R}^{(N+1)\times 2},$$

and adding or subtracting their contribution above equation, we obtain the following *discrete wave equation*,

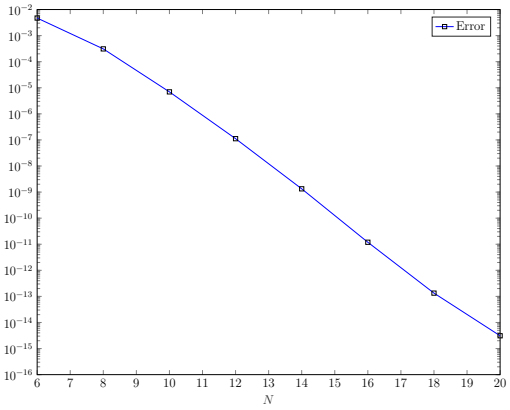
$$\Big((\mathcal{J} \pm \mathbf{e}\mathscr{B}_w) \otimes M + S_{\frac{3}{2}} S_{\frac{1}{2}} \otimes \mathfrak{D}\Big)\, u_h = f_h \pm (\mathbf{e} \otimes I_{N+1}) v_h, \tag{6.3.17}$$

where $\mathscr{B}_w := \left[\begin{smallmatrix} \mathscr{B} \\ \mathscr{B}_t \end{smallmatrix}\right]$ and $v_h := \left[\begin{smallmatrix} u_{oh} \\ u_{1h} \end{smallmatrix}\right]$. These scheme converge spectrally in both space and time as seen in Figure 6.16. For both of these schemes, we take f so that the exact solution is defined by (6.3.9), which satisfies the boundary conditions, and gives initial conditions as $u(x, y, -1)$ and $u_t(x, y, -1)$.

Consider (6.3.17) with '+' sign, that is, define the *global space-time spectral operator* $\mathcal{A}_w = (\mathcal{J} + \mathbf{e}\mathscr{B}) \otimes M + S_{\frac{3}{2}} S_{\frac{1}{2}} \otimes \mathfrak{D} \in \mathbb{R}^{(N+1)^2\times(N+1)^2}$. On observing the growth and decay of the maximum and minimum singular values of $\mathcal{A}_w$ from Figures 6.17a and 6.17b, respectively, it is deduced that $\kappa(\mathcal{A}_w) \leqslant cN^4$, for $6 \leqslant N \leqslant 40$.

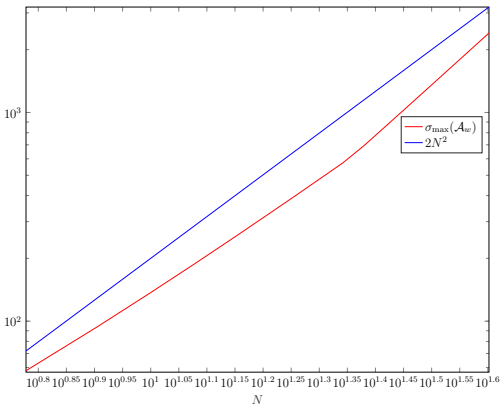


(a) Considering '−' sign.

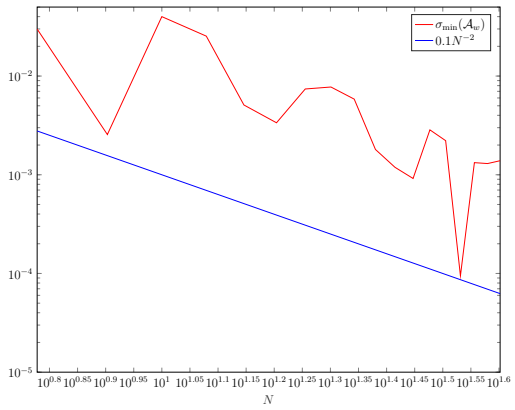


(b) Considering '+' sign.

Figure 6.16: Convergence of the US method in both space and time for the wave equation.



(a) Maximum singular value.



(b) Minimum singular value.

Figure 6.17: Bounds for singular values of $\mathcal{A}_w$.

Second scheme: Collocation in time and US in space

Now, we present another scheme for solving the *wave equation*, given by (6.3.13), demonstrating spectral convergence in both space and time. In this method, we incorporate the US method in space and collocation in time, by defining an approximation for u as (6.3.10). Since this equation has second order derivative w.r.t time, we also define $v(x,t) = u_t(x,t) \approx \sum_{j=0}^{N}\sum_{k=0}^{N} v_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k(t)$, where $\widetilde{C}_j^{(\frac{3}{2})}$ and ℓ_j

are the same as (6.3.10). Thus, the initial condition $u(x,-1)=u_0(x)$ implies

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k(-1) = \sum_{j=0}^{N} u_j^0\widetilde{C}_j^{(\frac{3}{2})}(x)$$
$$\sum_{j=0}^{N} u_{j0}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x) = \sum_{j=0}^{N} u_j^0\widetilde{C}_j^{(\frac{3}{2})}(x).$$

Define $u_h^0=[u_{00};u_{10};\ldots;u_{N0}]\in\mathbb{R}^{(N+1)}$ and $u_{0h}=[u_0^0;u_1^0;\ldots;u_N^0]\in\mathbb{R}^{(N+1)}$, then the above equation gives

$$Mu_h^0 = u_{0h}. \tag{6.3.18}$$

Similarly, define $v_h^0=[v_{00};v_{10};\ldots;v_{N0}]\in\mathbb{R}^{(N+1)}$ and $u_{1h}=[u_0^1;u_1^1;\ldots;u_N^1]\in\mathbb{R}^{(N+1)}$, so that $u_1(x)\approx\sum_{j=0}^{N} u_j^1\widetilde{C}_j^{(\frac{3}{2})}(x)$. Then, the second initial condition $u_t(x,-1)=u_1(x)$ gives

$$Mv_h^0 = u_{1h}. \tag{6.3.19}$$

Now, consider $v(x,t)=u_t(x,t)$, and collocate it on $t=t_m$, for some $1\leqslant m\leqslant N$. Recall that $\{t_m\}_{m=0}^N$ represent the Chebyshev Gauss-Lobatto nodes. Then, approximation of u and $v=u_t$ imply,

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_m) - \sum_{j=0}^{N}\sum_{k=0}^{N} v_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k(t_m) = 0,$$

or

$$\sum_{j=0}^{N}\sum_{k=0}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_m) - \sum_{j=0}^{N} v_{jm}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x) = 0.$$

Rearranging,

$$\sum_{j=0}^{N}\sum_{k=1}^{N} u_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_m) - \sum_{j=0}^{N} v_{jm}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x) = -\sum_{j=0}^{N} u_{j0}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_0'(t_m),$$

which yields the following linear system

$$(\llbracket D \rrbracket \otimes I_{N+1})u_h = v_h - d_{0h} \otimes u_h^0, \tag{6.3.20}$$

where d_{0h} represents the 0-th column of the matrix D, $v_h = [v_h^1; v_h^2; \ldots; v_h^N] \in \mathbb{R}^{N(N+1)}$, with $v_h^k = [v_{0k}; v_{1k}; \ldots; v_{Nk}] \in \mathbb{R}^{(N+1)}$, for $1 \leqslant k \leqslant N$. Now, we discretize the equation $u_{tt} - u_{xx} = f(x,t)$, or $v_t - u_{xx} = f(x,t)$, by collocating it at time $t = t_m$ as follows,

$$\sum_{j=0}^{N}\sum_{k=0}^{N} v_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_m) - \sum_{j=0}^{N} u_{jm}((1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x))'' = f(x,t_m)$$

$$\sum_{j=0}^{N}\sum_{k=1}^{N} v_{jk}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_m) + \sum_{j=0}^{N} u_{jm}(j(j+3)+2)\widetilde{C}_j^{(\frac{3}{2})}(x) = f(x,t_m)$$

$$- \sum_{j=0}^{N} v_{j0}(1-x^2)\widetilde{C}_j^{(\frac{3}{2})}(x)\ell_k'(t_0),$$

giving the following linear system,

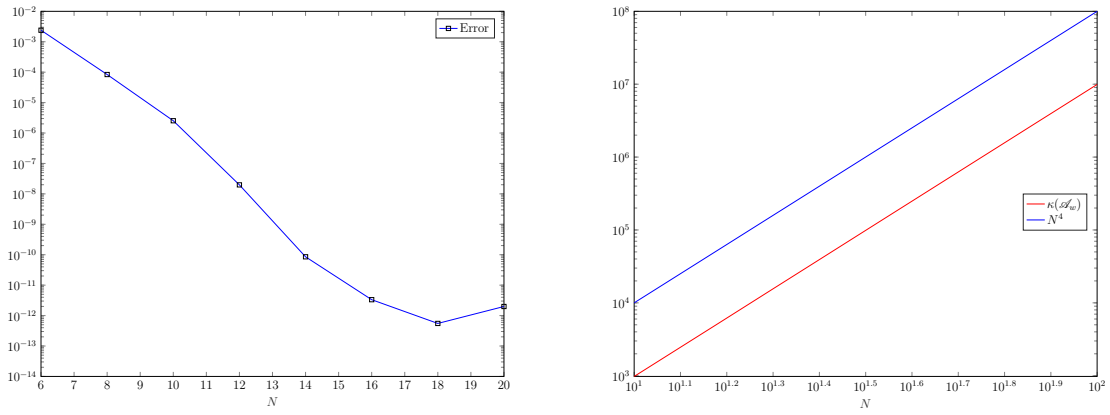
$$(\llbracket D \rrbracket \otimes M)v_h + (I_N \otimes \mathfrak{D})u_h = f_h - d_{0h} \otimes M v_h^0,$$

where $u_h = [u_h^1; u_h^2; \ldots; u_h^N] \in \mathbb{R}^{N(N+1)}$, with $u_h^k = [u_{0k}; u_{1k}; \ldots; u_{Nk}] \in \mathbb{R}^{(N+1)}$, for $1 \leqslant k \leqslant N$. From eqs. (6.3.18) to (6.3.20), the above equation becomes,

$$\left(\llbracket D \rrbracket^2 \otimes M + I_N \otimes \mathfrak{D}\right) u_h = f_h - d_{0h} \otimes (u_{0h}) - \llbracket D \rrbracket d_{0h} \otimes u_{1h},$$

which represents the *wave equation in discrete form*. This scheme converges spectrally in both space and time as seen in Figure 6.18a. For this scheme, we take f so that the exact solution is defined by (6.3.9), which satisfies the boundary conditions, and gives initial conditions as $u(x,y,-1)$ and $u_t(x,y,-1)$. Also, see Figure 6.18b for an estimate on condition number of the *global space-time spectral operator* for

this scheme, defined as $\mathscr{A}_w = ([D]^2 \otimes M + I_N \otimes \mathfrak{D}) \in \mathbb{R}^{N(N+1)\times N(N+1)}$. It is inferred that $\kappa(\mathscr{A}_w) \leqslant cN^4$, which is the same order as that of the previous scheme.



(a) Convergence. (b) 2-norm condition number of $\mathscr{A}_w$.

Figure 6.18: The US method in both space and collocation in time for the wave equation.

To summarize this section, we observe that the US method in space and time leads to sparse linear systems which are easier to formulate. However, incorporating spectral collocation in time and the US method in space gives better conditioned systems.

6.4 A fast solver for the space-time US method

The US method in space and time are revolutionary, as they lead to sparse global space time spectral operator. For instance, let us compare the spy graphs of the global space-time spectral operators. Figure 6.19 is the spy graph for $\mathscr{A}_h$, defined by (6.3.11), for $N = 9$, which results on applying the US method in space and spectral collocation in time for solving the heat equation. The density of pseudospectral derivative matrix $[D]$ makes $\mathscr{A}_h$ dense as well. Whereas, the spy graph of $\mathcal{A}_h$ for $N = 9$, defined by (6.3.8) and obtained on employing the US method in both space and time, is given by Figure 6.20a, is visibly sparse. Note that the initial condition

at the bottom of the global spectral operator. This can be easily moved to the top by using the QR decomposition of $J + \mathbf{e}_{n+1}\mathscr{B}$, defined by (6.3.8). Note that it is simply a permutation of rows, as a simple adjustment gives, $J + \mathbf{e}_{n+1}\mathscr{B} = Q_h R_h$, where

$$Q_h = \left[\begin{array}{c|c} & I_N \\ \hline 1 & \end{array}\right], \qquad R_h = \left[\begin{array}{cc} \multicolumn{2}{c}{\mathscr{B}} \\ \hline O_{N,1} & I_N \end{array}\right].$$

Define $\mathcal{Q}_h = (Q_h^T \otimes I_{N+1})$, then $\mathcal{Q}_h\mathcal{A}_h = R_h \otimes M + Q_h^T S_{\frac{1}{2}} \otimes \mathfrak{D}$, which is now a block almost banded structure as seen in Figure 6.20b.

Similarly, Figure 6.21a shows the sparsity of $\mathcal{A}_w$, defined by (6.3.17) for $N = 9$. For $\mathcal{A}_w \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$, we perform the same procedure as $\mathcal{J} + \mathbf{e}\mathscr{B}_w = Q_w R_w$, where

$$Q_w = \left[\begin{array}{c|c} & I_{N-1} \\ \hline I_2 & \end{array}\right], \qquad R_w = \left[\begin{array}{cc} \multicolumn{2}{c}{\mathscr{B}_w} \\ \hline O_{N-1,2} & 3I_{N-1} \end{array}\right].$$

Define $\mathcal{Q}_w = (Q_w^T \otimes I_{N+1})$, then $\mathcal{Q}_w\mathcal{A}_w = R_w \otimes M + Q_w^T S_{\frac{3}{2}} S_{\frac{1}{2}} \otimes \mathfrak{D}$. Figure 6.21b shows that its structure is also now similar to that of $\mathcal{Q}_h\mathcal{A}_h$, that is, a block almost banded structure.

Thus, the US method in space and time yield a block almost banded structure which can be easily exploited for solving these linear systems in parallel, thus constructing parallel-in-time (PinT) solvers. We introduce one of the simpler iterative techniques from [99], for solving tridiagonal linear systems in parallel. Note that the discrete heat equation allows the decomposition of $\mathcal{A}_h = L + \mathcal{D} + U$, where $\mathcal{D}$, L and U represent the block diagonal, block strictly lower and strictly upper triangular parts of $\mathcal{A}_h$, respectively. Since they are sparse, thus $\mathcal{A}_h$ can also be solved in parallel by using the hybrid blocked iterative solving algorithm (HBISA) given in [99, p. 1770]. However, for schemes arising from the US method in both space and

time such an algorithm generally leads to an iteration matrix with spectral radius greater than unity, thus implying divergence.

Our another aim is to formulate another direct parallel solver by using the specific structure of the matrices arising from applying the US method in both space and time. A step in this direction is to use ParaDIAG algorithm described in [33], which can be modified for the linear systems generated by the US method in space and time. To this end, let us consider the following linear system for sparse matrices $A, B, C, D \in \mathbb{R}^{n\times n}$

$$(A \otimes C + B \otimes D)u_h = f_h, \tag{6.4.1}$$

where $u_h, f_h \in \mathbb{R}^{n^2\times n^2}$ represent the discrete unknown vector and discrete forcing term vector. Numerical experiments imply that A and B for such linear systems possess a generalized eigenvalue decomposition $AV = BV\Lambda_{(A,B)}$, where $\Lambda_{(A,B)}$ is a diagonal matrix and V is invertible which transforms (6.4.1) to the following

$$(\Lambda_{(A,B)} \otimes C + I_n \otimes D)v_h = (V^{-1}B^{-1} \otimes I_n)f_h,$$

where $v_h = (V^{-1} \otimes I_n)u_h$, thus resulting in a block diagonal linear system which can be solved in parallel for $(v_h)_{1+kn:n+kn}$, where $0 \leqslant k \leqslant (n-1)$. Finally, $u_h = (V \otimes I_n)v_h$.

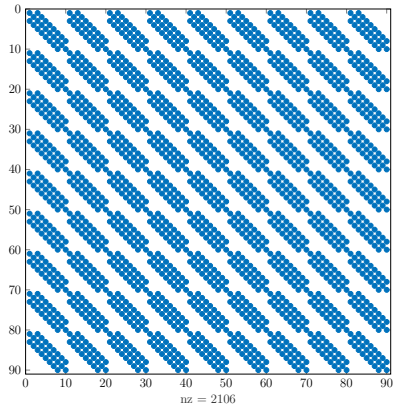
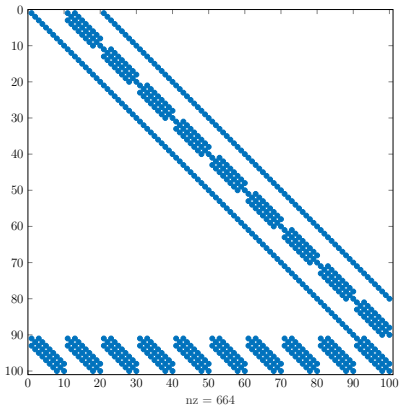
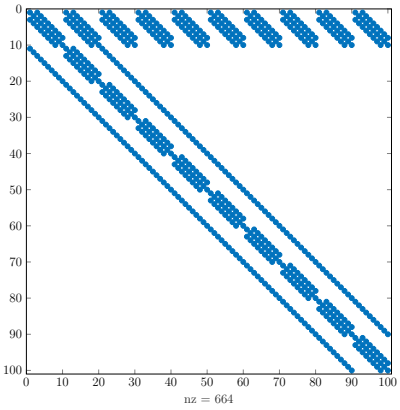


Figure 6.19: Spy graph of $\mathscr{A}_h$ for the heat equation with the US method in space and spectral collocation in time and $N = 9$.

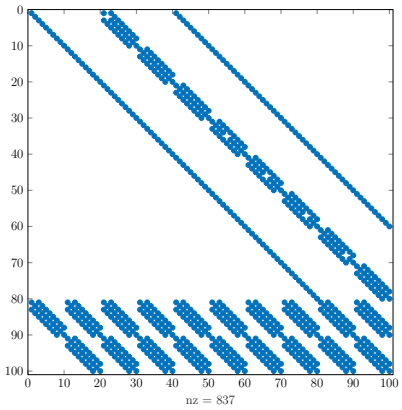


(a) $\mathcal{A}_h$

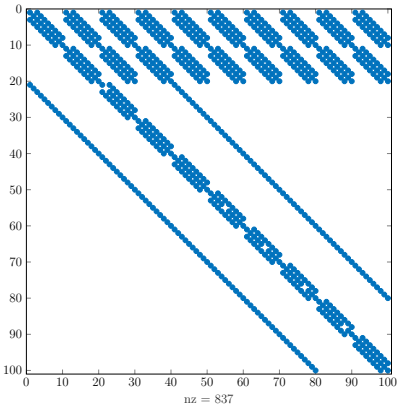


(b) $\mathcal{Q}_h\mathcal{A}_h$

Figure 6.20: Spy graphs for the heat equation with the US method in both space and time and $N = 9$.



(a) $\mathcal{A}_w$



(b) $\mathcal{Q}_w\mathcal{A}_w$

Figure 6.21: Spy graphs for the wave equation with the US method in space and time and $N = 9$.

7

Concluding remarks and future directions

In this chapter, we briefly describe some extensions of the work completed in Chapters 3–6.

In Chapter 3, we devised a space-time spectral method for the Stokes problem, which implements the $P_N - P_{N-2}$ scheme in space and spectral collocation in time. For simplicity of analysis, a recombined Legendre basis was considered in space and Chebyshev collocation was used in time. Note that, this scheme can easily be adapted to other orthogonal polynomial bases. A drawback of this scheme is that all time steps are coupled in the global spectral operator. Despite the limitations these methods are valuable in light of spectral convergence in both space and time, requiring far less number of unknowns for a highly accurate solution.

While voyaging through the goal of proving a condition number estimate for the global space-time spectral operator $\mathcal{G}_t$ for the unsteady Stokes problem, the first milestone was proving a condition number estimate for the $P_N - P_{N-2}$ scheme applied to the case independent of time, that is, in steady state. To this end, we proved estimates for the stiffness matrix, mass matrix, and the Laplacian in two di-

mension for a recombined Legendre basis functions, stated in form of Lemmas 3.3.1 and 3.3.2, and Theorem 3.3.3, respectively. Other intermediate results include estimates for the discrete gradient matrix B and the Schur complement Υ_h. Consequently, Theorem 3.3.11 proved the optimal condition number estimates for the sub-block appearing in the global spectral operator for the Stokes problem in steady state.

For the next step, the analysis of the scheme in unsteady state required a new estimate of maximum singular value (or 2-norm) of the Chebyshev derivative matrix, proved in Lemma 3.4.2, which was observed in existing literature through numerical experiments. The condition number estimate of the global space-time operator is still incomplete because we relied on numerical results for two estimates, namely eqs. (3.4.15) and (3.4.16). Nevertheless, we paved a proof for the minimum singular value of a non-symmetric saddle point matrix. This chapter was concluded with a proof of spectral convergence of this scheme in space and time is given in Theorem 3.4.6, which appears to be new. All of these results are verified through numerical experiments.

Recently, [46] present spectral distributions of the saddle point matrices arising from the discretization and linearization of the Navier–Stokes equations, where the leading block is nonsymmetric with a positive definite symmetric part. The global space-time spectral operator of the scheme analyzed in Theorem 3.4.5 is a non-symmetric saddle point matrix, so that the symmetric part of its leading block $\mathcal{A}_t$ is indefinite. As far as we know, there are no results in literature that estimates the spectrum of such a matrix, thus problem highlights a potential linear algebra problem. Approximations on the spectrum of this special case of a saddle point matrix will be significant for deriving spectral condition number for such schemes as seen in [67], which in turn is useful for the analysis of preconditioned systems as discussed in Section 2.1.4.

Research Direction 1. Sharper estimates for the spectrum of a saddle point matrix $\mathcal{X} = \left[\begin{smallmatrix} A & B^T \\ B & O \end{smallmatrix}\right]$, where $A \in \mathbb{R}^{n\times n}$ is non-symmetric matrix with an indefinite symmetric part and $B \in \mathbb{R}^{m\times n}$ is full rank.

Since the linear systems arising from space-time spectral methods are coupled at all times, a subsequent objective is to tailor the scheme as parallel in time. Although the Chebyshev derivative matrix $\mathbf{[}D\mathbf{]}$ is responsible for the dense nature of the discrete Stokes problem, however, it is diagonalizable and ParaDIAG algorithms mentioned in [33] result in a convenient parallel solver numerically. Albeit it is a tough task to formulate it theoretically, that is, to find an explicit expression of spectral decomposition of $\mathbf{[}D\mathbf{]}$. This problem can perhaps be more well-defined by considering exploring a suitable preconditioner which may resolve this issue.

Research Direction 2. Formulate a parallel-in-time (PinT) scheme for the linear system arising from spectral collocation in time.

In Chapter 4, we extended the $P_N - P_{N-2}$ from [12] scheme in space and collocation in time from Section 3.4 to the Navier-Stokes problem in Section 4.2. Furthermore, in Section 4.3, we applied a staggered grid collocation scheme in space, given in [11], and spectral collocation in time to both the unsteady Stokes and Navier-Stokes problem. Numerical experiments for all of these schemes validate a super-algebraic decay in error in L^∞ norm for the solution evaluated at the final time step, $t_N = 1$.

Research Direction 3. Study and estimate the high limit of the Reynold's number for the $P_N - P_{N-2}$ and staggered grid collocation schemes described in Sections 4.2 and 4.3, respectively, for the unsteady Stokes and Navier-Stokes problems.

In Chapter 5, we derived two new lower bounds on the minimum eigenvalue of a non-singular sum of two PSD matrices in form of Theorems 5.4.1 and 5.4.4, that is, $\lambda_{\min}(P + Q)$, where P and Q are PSD and $P + Q$ is SPD. To our knowledge, this is the first report of a positive lower bound for this case despite being a topic of

historic concern. These bounds incorporate the Friedrichs angle between the range spaces $\mathcal{R}(P)$ and $\mathcal{R}(Q)$. These findings provide a basis for lower bounds on the minimum singular value of some full rank block matrices presented in Corollaries 5.6.1 and 5.6.2, Theorem 5.6.3, and Corollary 5.6.4. Broadly translated our work indicates that the minimum singular value of a non-singular matrix can be derived in terms of its constituting matrices, P and Q for $P+Q$, or submatrices for block matrices even if they are singular. Importantly, by utilizing the projection on range spaces, the key component for bringing the minimum positive singular value of rank-deficient matrices into play. This may be considered a promising aspect of studying the singular values of matrices with a geometric approach.

The lower bounds described by Theorems 5.4.1 and 5.4.4 are sharp for the case of $\mathcal{R}(P) \cap \mathcal{R}(Q) = \{0\}$, so that the parameter $k = 0$. However, there is a scope for improvement in these results for the case when $r, k > 0$. It may be possible to incorporate $P_{M_1 \cap M_2}$ in (5.4.22) to improve the lower bound. In Example 5.6.10, it was observed that changing the partition of a matrix changes the lower bounds given by Theorem 5.6.3 and Corollary 5.6.4. Thus, one can try to determine the best partition for a certain class of matrices yielding optimal estimates on the minimum singular value. Some techniques exist for calculating the principal angles between two subspaces, see [58]. A more efficient algorithm can be designed for calculating the Friedrichs angle between two subspaces. Since it is defined for subspaces of a Hilbert space in Definition 2.2.4, it may allow us to extend the main results to a more general setting. This discussion is summarized in form of the following pointers.

Research Direction 4. Improve the results for the sum of two PSD matrices with a non-trivial intersection of their range spaces.

Research Direction 5. Determine a criteria for achieving an optimal partition for matrices yielding optimal estimates of the minimum singular value, as different partitions of a matrix yield different estimates.

Research Direction 6. Extend Theorems 5.4.1 and 5.4.4 to self-adjoint operators on a separable Hilbert space.

Research Direction 7. Devise a computational or theoretical procedure for calculating the Friedrichs angle between the range spaces of any two given matrices.

In Chapter 6, we implemented the US method in space and time for the heat, Schrödinger, and wave equations. Additionally, we imposed the US method in space and spectral collocation in time to collate the merits and demerits of employing the two schemes. This experiment adds to a growing corpus of research showing the effectiveness of the US methods and promise a state-of-art scheme for solving time dependent PDEs. As Section 6.4 indicates that we can fruitfully explore a PinT solver by exploiting the sparse block almost banded structure of the schemes arising from the US method in space and time. Thus, a number of recommendations for future research are given as follows.

Research Direction 8. Prove an estimate on 2-norm condition number of the heat, Schrödinger, and wave equations.

Research Direction 9. Devise a space-time US method for solving linear PDEs such as Airy, beam and Stokes problem.

Research Direction 10. Prove the spectral convergence of the US method in both space and time for linear time dependent PDEs.

Research Direction 11. Design a PinT solver for the numerical schemes resulting on employing the US method in space and time for linear time dependent PDEs.

Appendix A

Alternative proofs

The results mentioned in this chapter were derived prior to coming across [32] and solely relying on knowledge from [68].

A.1 Difference of orthogonal projections

The following is a simpler proof for the expression for the null space of difference of two projections onto subspaces $U, V \subseteq \mathbb{R}^n$ spanning $\mathbb{R}^n$, that is, $\mathbb{R}^n = U + V$.

Lemma A.1.1. *Let $U, V \subseteq \mathbb{R}^n$ be subspaces so that $\mathbb{R}^n = U + V$. If $P_1, P_2 \in \mathbb{R}^{n\times n}$ are orthogonal projections onto U and V, respectively, then $\mathcal{N}(P_1 - P_2) = U \cap V$.*

Proof. Let $x \in U \cap V$, then $P_1 x = x$ and $P_2 x = x$, thus $(P_1 - P_2)x = 0$, which implies $x \in \mathcal{N}(P_1 - P_2)$. Therefore, $U \cap V \subseteq \mathcal{N}(P_1 - P_2)$.

Let $x \in \mathcal{N}(P_1 - P_2) \subseteq \mathbb{R}^n$, then $P_1 x = P_2 x$. Let $B = \{\xi_1, \xi_2, \ldots, \xi_k\}$ be a basis of $U \cap V$. Consequently, we define $B_U = \{\xi_1, \ldots, \xi_k, \phi_{k+1}, \ldots, \phi_{n_1}\}$, $B_V = \{\xi_1, \ldots, \xi_k, \psi_{k+1}, \ldots, \psi_{n_2}\}$ as bases of U, V, respectively. Since $x \in \mathcal{N}(P_1 - P_2) \subseteq \mathbb{R}^n = U + V$, there exist some scalars a_i, b_j, where $1 \leqslant i \leqslant n_1$, $k+1 \leqslant j \leqslant n_2$, such

that

$$\begin{aligned} x &= \sum_{i=1}^{k} a_i\xi_i + \sum_{i=k+1}^{n_1} a_i\phi_i + \sum_{i=k+1}^{n_2} b_i\psi_i, \\ P_1x &= \sum_{i=1}^{k} a_i\xi_i + \sum_{i=k+1}^{n_1} a_i\phi_i, \\ P_2x &= \sum_{i=1}^{k} a_i\xi_i + \sum_{i=k+1}^{n_2} b_i\psi_i. \end{aligned}$$

Thus, $P_1x = P_2x$ implies

$$\sum_{i=k+1}^{n_1} a_i\phi_i = \sum_{i=k+1}^{n_2} b_i\psi_i \in (U\backslash V) \cap (V\backslash U) = \{0\},$$

thus $x = \sum_{i=1}^{k} a_i\xi_i \in U \cap V$. Hence, $\mathcal{N}(P_1 - P_2) \subseteq U \cap V$. □

A.2 A result on complementary subspaces

The following is a simpler proof for the minimum singular value of the sum of two orthogonal projections onto complementary subspaces $R_1, R_2 \subseteq \mathbb{R}^n$.

Lemma A.2.1. *Let $R_1, R_2 \in \mathbb{R}^n$ be two complementary subspaces, that is, $\mathbb{R}^n = R_1 \oplus R_2$, then $c(R_1^T, R_2^T) = 1 - \cos\theta$, where θ is the minimum principal angle between $\mathcal{R}(R_1^T)$ and $\mathcal{R}(R_2^T)$.*

Proof. Let $P_i \in \mathbb{R}^{n\times n}$, be orthogonal projections onto the subspaces $\mathcal{R}(R_i^T)$ for $i = 1, 2$. In order to prove that $c(R_1^T, R_2^T) = 1 - \cos\theta$, it suffices to prove that, for any $x \in \mathbb{R}^n$, $x \neq 0$,

$$\frac{|P_1x|^2 + |P_2x|^2}{|x|^2} \geqslant 1 - \cos\theta.$$

By Property 4 of Proposition 2.2.2,

$$\min_{\substack{x\in\mathbb{R}^{m+n}\\ x\neq 0}} \frac{|(P_1-P_2)x|}{|x|} = \sigma_{\min}(P_1-P_2) = \sin\theta.$$

Therefore, for $x \in \mathbb{R}^n\backslash\{0\}$, $|(P_1-P_2)x|^2 \geqslant \sin^2(\theta)|x|^2$, or

$$(P_1-P_2)^2 \geqslant \sin^2(\theta)I \tag{A.2.1}$$

Note that,

$$\begin{aligned}
(P_1+P_2-I)^2 &= (P_1+P_2-I)(P_1+P_2-I)\\
&= P_1P_2 - P_1 + P_2P_1 - P_2 + I\\
&= I - (P_1^2 - P_1P_2 - P_2P_1 + P_2^2)\\
&= I - (P_1-P_2)^2\\
&\leqslant (1-\sin^2(\theta))I = \cos^2(\theta)I. \qquad \text{(by (A.2.1))}
\end{aligned}$$

Thus,

$$-\cos(\theta)I \leqslant P_1+P_2-I \leqslant \cos(\theta)I,$$

or

$$(1-\cos(\theta))I \leqslant P_1+P_2 \leqslant (1+\cos(\theta))I.$$

Therefore, for $x \in \mathbb{R}^n\backslash\{0\}$,

$$|(P_1+P_2)x|^2 \geqslant (1-\cos(\theta))^2|x|^2.$$

By the parallelogram law and using the above inequality with (A.2.1),

$$\begin{aligned}
|P_1x|^2 + |P_2x|^2 &= \frac{1}{2}\left(|P_1x + P_2x|^2 + |P_1x - P_2x|^2\right) \\
&\geqslant \frac{1}{2}\left((1-\cos(\theta))^2 + \sin^2(\theta)\right)|x|^2 \\
&= \frac{1}{2}\left(1 + \cos^2(\theta) - 2\cos(\theta) + \sin^2(\theta)\right)|x|^2 \\
&= (1-\cos(\theta))\,|x|^2.
\end{aligned}$$

□

Index

M

N

O

P

S

U

V

W

References

[1] Jared Lee Aurentz and Richard Mikaël Slevinsky. On symmetrizing the ultraspherical spectral method for self-adjoint problems. *Journal of Computational Physics*, 410:109383, 2020. 149

[2] O. Axelsson. Unified analysis of preconditioning methods for saddle point matrices. *Numerical Linear Algebra with Applications*, 22(2):233–253, Mar 2015. 112

[3] O. Axelsson and M. Neytcheva. Eigenvalue estimates for preconditioned saddle point matrices. *Numerical Linear Algebra with Applications*, 13(4):339–360, May 2006. 58

[4] O. Axelsson and M. Neytcheva. Eigenvalue estimates for preconditioned saddle point matrices. *Numerical Linear Algebra with Applications*, 13(4):339–360, May 2006. 111

[5] Z. Z. Bai. Eigenvalue estimates for saddle point matrices of Hermitian and indefinite leading blocks. *Journal of Computational and Applied Mathematics*, 237(1):295–306, Jan 2013. 112

[6] D. Barber. *Bayesian Reasoning and Machine Learning*. Bayesian Reasoning and Machine Learning. Cambridge University Press, 2012. 108

[7] Guo Ben-Yu and Li Jian. Fourier-Chebyshev spectral method for the two-

dimensional Navier-Stokes equations. *SIAM Journal on Numerical Analysis*, 33(3):1169–1187, 1996. 30

[8] M. Benzi, G.H. Golub, and J. Liesen. Numerical solution of saddle point problems. *Acta Numerica*, 14(123):1–137, 2005. 111

[9] M. Benzi and V. Simoncini. On the eigenvalues of the class of saddle point matrices. *Numerische Mathematik*, 103(123):173–196, Mar 2006. 112

[10] L. Bergamaschi. On the eigenvalue distribution of constrained preconditioned symmetric saddle point matrices. *Numerical Linear Algebra with Applications*, 19(4):754–772, Aug 2012. 112

[11] Christine Bernardi and Yvon Maday. A collocation method over staggered grids for the Stokes problem. *International Journal for Numerical Methods in Fluids*, 8:537–557, 1988. 88, 95, 202

[12] Christine Bernardi and Yvon Maday. *Spectral methods, Techniques of Scientific Computing (Part 2), Handbook of Numerical Analysis Vol. V.* Elsevier, Amsterdam, 1997. 1, 29, 35, 39, 40, 58, 64, 78, 88, 202

[13] Christine Bernardi and Yvon Maday. Uniform inf-sup conditions for the spectral discretization of the Stokes problem. *Mathematical Models and Methods in Applied Sciences*, 9(03):395–414, 1999. 30

[14] Dennis S. Bernstein. *Scalar, Vector, and Matrix Mathematics: Theory, Facts, and Formulas - Revised and Expanded Edition.* Princeton University Press, Princeton, NJ, 2018. 114, 141

[15] R. Bhatia and F. Kittaneh. On singular values of product of operators. *SIAM Journal on Matrix Analysis and Applications*, 11(2):272–277, April 1990. 114, 115

[16] R. Bhatia and F. Kittaneh. The matrix arithmetic-geometric mean inequality revisited. *Linear Algebra and its Applications*, 428(123):2177–2191, Feb 2008. 111

[17] A. Björck and G. Golub. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594, July 1973. 122

[18] Nicolas Boullé, Jonasz Słomka, and Alex Townsend. An optimal complexity spectral method for Navier–Stokes simulations in the ball, 2021. 30, 149

[19] Chaima Bousbiat, Yasmina Daikh, and Sarra Maarouf. Spectral discretization of the time-dependent Stokes problem with mixed boundary conditions. *Mathematical Methods in the Applied Sciences*, 44(18):14517–14544, 2021. 30

[20] A. Böttcher and I.M. Spitkovsky. A gentle guide to the basics of two projections theory. *Linear Algebra and its Applications*, 432(6):1412–1459, 2010. 20, 21, 119

[21] Yuxin Chen, Yuejie Chi, Jianqing Fan, and Cong Ma. Spectral methods for data science: A statistical perspective. *Foundations and Trends in Machine Learning*, 14(5):566–806, 2021. 108

[22] R.M. Corless and N. Fillion. *A Graduate Introduction to Numerical Methods: From the Viewpoint of Backward Error Analysis*. SpringerLink: Bücher. Springer New York, 2013. 70

[23] J. Dauxois and G.M. Nkiet. Canonical analysis of two Euclidean subspaces and its applications. *Linear Algebra and its Applications*, 264:355–388, 1997. Sixth Special Issue on Linear Algebra and Statistics. 109

[24] Chandler Davis and W. M. Kahan. The rotation of eigenvectors by a perturbation. iii. *SIAM journal on numerical analysis*, 7(1):1–46, 1970. 16

[25] Jane M. Day, Wasin So, and Robert C. Thompson. The spectrum of a Hermitian matrix sum. *Linear Algebra and its Applications*, 280:289–332, 1998. 108

[26] Frank Deutsch. The angle between subspaces of a Hilbert space. In S. P. Singh, editor, *Approximation Theory, Wavelets and Applications*, pages 107–130. Springer Netherlands, Dordrecht, 1995. 19, 20, 109

[27] S.W. Drury. On a question of Bhatia and Kittaneh. *Linear Algebra and its Applications*, 437(7):1955–1960, Oct 2012. 111

[28] H.C. Elman, D.J. Silvester, and A.J. Wathen. *Finite Elements and Fast Iterative Solvers: with Applications in Incompressible Fluid Dynamics: with Applications in Incompressible Fluid Dynamics*. Numerical Mathematics and Scientific Computation. OUP Oxford, 2005. 58

[29] Daniel Fortunato, Nicholas Hale, and Alex Townsend. The ultraspherical spectral element method. *Journal of Computational Physics*, 436:110087, 2021. 149

[30] Daniel Fortunato and Alex Townsend. Fast Poisson solvers for spectral methods. *IMA Journal of Numerical Analysis*, 2019. 149, 176

[31] A. Galántai. *Projectors and Projection Methods*. Advances in Mathematics. Springer, Boston, MA, 2004. 20

[32] A. Galántai. Subspaces, angles and pairs of orthogonal projections. *Linear and Multilinear Algebra*, 56(3):227–260, May 2008. 16, 18, 19, 20, 121, 124, 205

[33] Martin J. Gander, Jun Liu, Shu-Lin Wu, Xiaoqiang Yue, and Tao Zhou. ParaDIAG: Parallel-in-Time algorithms based on the diagonalization technique. May 2020. 198, 202

[34] Gene H. Golub and Charles F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, Baltimore, 4th edition, 2013. 10, 16

[35] Gene H. Golub and Hongyuan Zha. The canonical correlations of matrix pairs and their numerical computation. In Adam Bojanczyk and George Cybenko, editors, *Linear Algebra for Signal Processing*, pages 27–49, New York, NY, 1995. Springer New York. 109

[36] David Gottlieb and Steven A. Orszag. *Numerical Analysis of Spectral Methods: Theory and Applications*. Society for Industrial and Applied Mathematics, Philadelphia,Pa., 1977. 25

[37] N.I.M. Gould and V. Simoncini. Spectral analysis of saddle point matrices with indefinite leading blocks. *SIAM Journal on Matrix Analysis and Applications*, 31(3):1152–1171, 2010. 112

[38] Ben Yu Guo and Zhong Qing Wang. Legendre–Gauss collocation methods for ordinary differential equations. *Adv. Comput. Math.*, 30(3):249–280, 2009. 30

[39] J.S. Hesthaven, S. Gottlieb, and D. Gottlieb. *Spectral Methods for Time-Dependent Problems*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007. 30

[40] O. Hirzallah. Inequalities for sums and products of operators. *Linear Algebra and its Applications*, 407(123):32–42, Sept 1990. 146

[41] L. Hogben. *Handbook of Linear Algebra, Second Edition*. Discrete Mathematics and Its Applications. Taylor & Francis, 2013. 12, 13

[42] Y.P. Hong and C.-T. Pan. A lower bound for the smallest singular value. *Linear Algebra and its Applications*, 172(123):27–32, July 1992. 112

[43] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 2012. 12, 114

[44] Roger A. Horn and Charles R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, 1991. 8, 12, 108

[45] N. Huang and C.-F. Ma. On the eigenvalues of the saddle point matrices discretized from Navier-Stokes equations. *Numerical Algorithms*, 79(1):41–64, Sept 2018. 112

[46] Na Huang and Chang Feng Ma. On the eigenvalues of the saddle point matrices discretized from Navier–Stokes equations. *Numerical Algorithms*, 2018. 201

[47] G. Ierley, B. Spencer, and R. Worthing. Spectral methods in time for a class of parabolic partial differential equations. *Journal of Computational Physics*, 102(1):88–97, 1992. 30

[48] C.R. Johnson. A Gersgorin-type lower bound for the smallest singular value. *Linear Algebra and its Applications*, 112(123):1–7, Jan 1989. 112

[49] C.R. Johnson and T. Szulc. Further lower bounds for the smallest singular value. *Linear Algebra and its Applications*, 272(123):169–179, Mar 1998. 112

[50] C.R. Johnson, T. Szulc, and D. Wojtera-Tyrakowska. Optimal Gersgorin-style estimation of extremal singular values. *Linear Algebra and its Applications*, 402(123):46–60, June 2005. 112

[51] Camille Jordan. Essai sur la géométrie à n dimensions. *Bulletin de la Société Mathématique de France*, 3:103–174, 1875. 16

[52] Avleen Kaur. `https://github.com/avleenk2312/Space-Time-Spectral`, 2022. 64, 77, 94

[53] Avleen Kaur and S. H. Lui. `https://github.com/avleenk2312/lb_minsingular`, 2022. 139, 141, 144, 145

[54] Avleen Kaur and S. H. Lui. New lower bounds on the minimum singular value of a matrix. `https://github.com/avleenk2312/lb_minsingular/blob/main/Draft.pdf`, 2022. 75

[55] Alexander A. Klyachko. Stable bundles, representation theory and hermitian operators. *Selecta Mathematica, New Series*, 4:419–445, 1998. 108

[56] Allen Knutson and Terence Tao. Honeycombs and sums of Hermitian matrices. *Notices of the American Mathematical Society*, 48(2):175–186, 2001. 108

[57] Andrew Knyazev. Toward the optimal preconditioned eigensolver: Locally optimal block preconditioned conjugate gradient method. *SIAM Journal on Scientific Computing*, 2001. 109

[58] Andrew Knyazev and Merico Argentati. Principal angles between subspaces in an A-based scalar product: Algorithms and perturbation estimates. *SIAM Journal on Scientific Computing*, 23(6):2009–2041, jan 2002. 203

[59] Chi-Kwong Li and Yiu-Tung Poon. Sum of Hermitian matrices with given eigenvalues: Inertia, rank, and multiple eigenvalues. *Canadian Journal of Mathematics*, 62(1):109–132, 2010. 111

[60] L. Li. Estimation for matrix singular value. *Computers and Mathematics with Applications*, 37(9):9–15, May 1999. 112

[61] M. Lin and M. Xie. On some lower bounds for smallest singular value of matrices. *Applied Mathematics Letters*, 121(123):107411, Nov 2021. 112, 145

[62] Wenjie Liu, Jiebao Sun, and Boying Wu. Galerkin-Chebyshev spectral method

and block boundary value methods for two-dimensional semilinear parabolic equations. *Numerical Algorithms*, 71:437–455, 2016. 30

[63] Wenjie Liu, Boying Wu, and Jiebao Sun. Space-time spectral collocation method for the one-dimensional Sine-Gordon equation. *Numerical Methods for Partial Differential Equations*, 31(3):670–690, 2015. 30

[64] S. H. Lui. *Numerical Analysis of Partial Differential Equations*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2012. 15, 24, 25, 26, 27, 67, 84, 100

[65] S. H. Lui. Legendre spectral collocation in space and time for PDEs. *Numerische Mathematik*, 136:75–99, 2017. 1, 30

[66] S. H. Lui and Sarah Nataj. Chebyshev spectral collocation in space and time for the heat equation. *Elect. Trans. Numer. Anal.*, 52:295–319, 2020. 1, 30

[67] S.H. Lui and Sarah Nataj. Spectral collocation in space and time for linear PDEs. *Journal of Computational Physics*, 424:109843, 2021. 1, 30, 176, 201

[68] Carl D. Meyer. *Matrix Analysis and Applied Linear Algebra*. SIAM, Philadelphia, 1st edition, 2000. 6, 8, 16, 17, 205

[69] Rajat Mittal. A Fourier-Chebyshev spectral collocation method for simulating flow past spheres and spheroids. *International journal for numerical methods in fluids*, 30(7):921–937, 1999. 30

[70] Nenad Morača. Bounds for norms of the matrix inverse and the smallest singular value. *Linear Algebra and its Applications*, 2008. 144, 145

[71] Matthew P. O'Donnell and Paul M. Weaver. Rapid analysis of variable stiffness beams and plates: Legendre polynomial triple-product formulation. *International Journal for Numerical Methods in Engineering*, 112(1):86–100, 2017. 91

[72] Frank W. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark. *NIST Handbook of Mathematical Functions*. Cambridge University Press, USA, 1st edition, 2010. 176, 177

[73] Sheehan Olver, Richard Mikaël Slevinsky, and Alex Townsend. Fast algorithms using orthogonal polynomials. *Acta Numerica*, 29:573–699, 2020. 24

[74] Sheehan Olver and Alex Townsend. A fast and well-conditioned spectral method. *SIAM Review*, 55(3):462–489, 2013. 2, 25, 149, 150, 158

[75] Sheehan Olver, Alex Townsend, and Geoffrey Vasil. A sparse spectral method on triangles. *SIAM Journal on Scientific Computing*, 41(6):A3728–A3756, 2019. 149

[76] L. Qi. Some simple estimates for singular value of a matrix. *Linear Algebra and its Applications*, 56(123):105–119, Jan 1984. 112

[77] Yonghui Qin and He Ping Ma. Legendre-tau-Galerkin and spectral collocation method for nonlinear evolution equations. *Applied Numerical Mathematics*, 153:52–65, 2020. 30

[78] S. C. Reddy and J. A. C. Weideman. The accuracy of the Chebyshev differencing method for analytic functions. *SIAM J. Numer. Anal.*, 42(5):2176–2187, 2005. 81, 84

[79] Jie Shen. Efficient Chebyshev-Legendre Galerkin Methods for Elliptic Problems. *International Conference on Spectral and High Order Methods*, 1996. 31

[80] Jie Shen, Tao Tang, and Li-Lian Wang. *Spectral Methods Algorithms, Analysis and Applications*. Springer-Verlag, Berlin Heidelberg, 2011. 25, 32, 33, 40, 64, 65, 67, 68, 97, 100

[81] Jie Shen and Li-Lian Wang. Fourierization of the Legendre–Galerkin method and a new space–time spectral method. *Applied Numerical Mathematics*, 57(5):710–720, 2007. 30

[82] S.-Q. Shen, L. Jian, W.-D. Bao, and T.-Z Huang. On the eigenvalue distribution of preconditioned nonsymmetric saddle point matrices. *Numerical Linear Algebra with Applications*, 21(4):557–568, Aug 2014. 112

[83] D. Silvester and A. Wathen. Fast iterative solution of stabilised stokes systems, part ii: Using general block preconditioners. *SIAM Journal on Numerical Analysis*, 31(5):1352–1367, Oct 1994. 112

[84] Richard Mikael Slevinsky and Sheehan Olver. A fast and well-conditioned spectral method for singular integral equations. *Journal of Computational Physics*, 332:290–315, 2017. 149

[85] A. Smoktunowicz. Block matrices and symmetric perturbations. *Linear Algebra and its Applications*, 429(10):2628–2635, 2008. 112

[86] G. W. Stewart and Ji guang Sun. *Matrix Perturbation Theory*. Computer Science and Scientific Computing. Elsevier Science, 1990. 109

[87] A. Y. Suhov. A spectral method for the time evolution in parabolic problems. *J. Sci. Comput.*, 29(2):201–217, 2006. 30

[88] H. Tal-Ezer. Spectral methods in time for hyperbolic equations. *SIAM Journal on Numerical Analysis*, 23(1):11–26, 1986. 30

[89] H. Tal-Ezer. Spectral methods in time for parabolic problems. *SIAM Journal on Numerical Analysis*, 26(1):1–11, 1989. 30

[90] Jian Guo Tang and He Ping Ma. Single and multi-interval Legendre τ-methods

in time for parabolic equations. *Advances in Computational Mathematics*, 17(4):349–367, 2002. 30

[91] Jian Guo Tang and He Ping Ma. A Legendre spectral method in time for first-order hyperbolic equations. *Applied Numerical Mathematics*, 57(1):1–11, 2007. 30

[92] Tao Tang and Xiang Xu. Accuracy enhancement using spectral postprocessing for differential equations and integral equations. In *Communications in Computational Physics*, 2009. 1, 30

[93] D. J. Torres and E. A. Coutsias. Pseudospectral solution of the two-dimensional Navier-Stokes equations in a disk. *SIAM Journal on Scientific Computing*, 21(1), 9 1999. 30

[94] Alex Townsend and Sheehan Olver. The automatic solution of partial differential equations using a global spectral method. *Journal of Computational Physics*, 299:106–123, 2015. 25, 149

[95] Lloyd N. Trefethen. *Spectral Methods in MATLAB*. Society for Industrial and Applied Mathematics, Philadelphia, 2000. 25

[96] J.M. Varah. A lower bound on the smallest singular value of a matrix. *Linear Algebra and its Applications*, 11(1):3–5, 1975. 112

[97] C.-L. Wang and S.-J. Zhang. The block lower bounds for the smallest singular value. *International Journal of Computer Mathematics*, 82(3):313–319, March 2005. 113

[98] Per Åke Wedin. On angles between subspaces of a finite dimensional inner product space. In Bo Kågström and Axel Ruhe, editors, *Matrix Pencils*, pages 263–285, Berlin, Heidelberg, 1983. Springer Berlin Heidelberg. 19

[99] Wangdong Yang, Kenli Li, and Keqin Li. A parallel solving method for block-tridiagonal equations on cpu—gpu heterogeneous computing systems. *J. Supercomput.*, 73(5):1760–1781, may 2017. 197

[100] Lijun Yi and Zhongqing Wang. Legendre Gauss type spectral collocation algorithms for nonlinear ordinary partial differential equations. *International Journal of Computer Mathematics*, 91(7):1434–1460, 2014. 30

[101] Lijun Yi and Zhongqing Wang. Legendre spectral collocation method for second-order nonlinear ordinary partial differential equations. *Discrete & Continuous Dynamical Systems–B*, 19:299–322, 2014. 30

[102] Y.-S. Yu and D.-H. Gu. A note on a lower bound for the smallest singular value. *Linear Algebra and its Applications*, 253(123):25–38, March 1997. 112

[103] Fuzhen Zhang. *Matrix Theory, Basic Results and Techniques*. Springer, New York, 2nd edition, 2011. 114

[104] L. Zou. A lower bound for the smallest singular value. *Journal of Mathematical Inequalities*, 6(4):625–629, Dec 2012. 112

[105] L. Zou and Y. Jiang. Estimation of the eigenvalues and the smallest singular value of matrices. *Linear Algebra and its Applications*, 433(6):1203–1211, Nov 2010. 112, 141