2015

The Krylov Subspace Methods for the Computation of Matrix Exponentials

Hao Wang
University of Kentucky, hao.wang@uky.edu

STUDENT AGREEMENT:

I represent that my thesis or dissertation and abstract are my original work. Proper attribution has been given to all outside sources. I understand that I am solely responsible for obtaining any needed copyright permissions. I have obtained needed written permission statement(s) from the owner(s) of each third-party copyrighted matter to be included in my work, allowing electronic distribution (if such use is not permitted by the fair use doctrine) which will be submitted to UKnowledge as Additional File.

I hereby grant to The University of Kentucky and its agents the irrevocable, non-exclusive, and royalty-free license to archive and make accessible my work in whole or in part in all forms of media, now or hereafter known. I agree that the document mentioned above may be made available immediately for worldwide access unless an embargo applies.

I retain all other ownership rights to the copyright of my work. I also retain the right to use in future works (such as articles or books) all or part of my work. I understand that I am free to register the copyright to my work.

REVIEW, APPROVAL AND ACCEPTANCE

The document mentioned above has been reviewed and accepted by the student's advisor, on behalf of the advisory committee, and by the Director of Graduate Studies (DGS), on behalf of the program; we verify that this is the final, approved version of the student's thesis including all changes required by the advisory committee. The undersigned agree to abide by the statements above.

Hao Wang, Student

Dr. Qiang Ye, Major Professor

Dr. Peter Perry, Director of Graduate Studies

The Krylov Subspace Methods for the Computation of Matrix Exponentials

---

DISSERTATION

A dissertation submitted in partial
fulfillment of the requirements for
the degree of Doctor of Philosophy
in the College of Arts and Sciences
at the University of Kentucky

By
Hao Wang
Lexington, Kentucky

Director: Dr. Qiang Ye, Professor of Mathematics
Lexington, Kentucky 2015

ABSTRACT OF DISSERTATION

The Krylov Subspace Methods for the Computation of Matrix Exponentials

The problem of computing the matrix exponential e^{tA} arises in many theoretical and practical problems. Many methods have been developed to accurately and efficiently compute this matrix function or its product with a vector, i.e., $e^{tA}v$. In the past few decades, with the increasing need of the computation for large sparse matrices, iterative methods such as the Krylov subspace methods have proved to be a powerful class of methods in dealing with many linear algebra problems. The Krylov subspace methods have been introduced for computing matrix exponentials by Gallopoulos and Saad, and the corresponding error bounds that aim at explaining the convergence properties have been extensively studied. Many of those bounds show that the speed of convergence depends on the norm of the matrix, while some others emphasize the important role played by the spectral distribution. For example, it is shown in a recent work by Ye that the speed of convergence is also determined by the condition number for a symmetric negative definite matrix. Namely the convergence is fast for a well-conditioned matrix no matter how large the norm is.

In this dissertation, we derive new error bounds for computing $e^{tA}v$ for non-symmetric A, using the spectral information of A. Our result is based on the assumption that A is negative definite, i.e., the field of values of A lies entirely in the left half of the complex plane, such that the underlying dynamic system is stable. The new bounds show that the speed of convergence is related to the size and shape of the rectangle containing the field of values, and they agree with the existing results when

A is symmetric. Furthermore, we also derive a simpler error bound for the special case when A is skew-Hermitian. This bound explains an observed convergence behavior where the approximation error initially stagnates for certain number of iterations before it starts to converge. In deriving our new error bounds, we use sharper estimates of the decay property of exponentials of Hessenberg matrices, by constructing Faber polynomial approximating exponential function in the region containing the field of values. The Jacobi elliptic functions are used to construct the conformal mappings and generate the Faber polynomials. We also present numerical tests to demonstrate the behavior of the new error bounds.

KEYWORDS: matrix exponential, Krylov subspace methods, numerical range, Faber polynomials, Jacobi elliptic functions

Author's signature: Hao Wang

Date: December 9, 2015

The Krylov Subspace Methods for the Computation of Matrix Exponentials

By

Hao Wang

Director of Dissertation: Qiang Ye

Director of Graduate Studies: Peter Perry

Date: December 9, 2015

To Ying

ACKNOWLEDGMENTS

First of all, I would like to gratefully and sincerely thank Dr. Qiang Ye for his guidance during my graduate studies at University of Kentucky. His knowledge, insights, understanding and patience helped me to finish this dissertation.

Secondly, I would also like to thank all my advisory committee members, Dr. Russell Carden, Dr. Lawrence Harris and Dr. Mai Zhou for their valuable comments.

Finally and most importantly, I would like to thank my parents and my wife for their encouragement and support, without which I would never have made it this far.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1 Background and introduction

The classical problem of solving systems of linear ordinary differential equations

$$\dot{x}(t) = Ax(t) \tag{1.1}$$

arises in many physical and economic problems. Here A is a given fixed n-by-n matrix. With the initial condition

$$x(0) = x_0,$$

the solution of (1.1) is

$$x(t) = e^{tA}x_0.$$

The matrix exponential e^{tA} is defined by the convergent power series

$$e^{tA} = I + tA + \frac{t^2A^2}{2!} + \cdots .$$

Thus, the accurate and efficient computation of the matrix exponential e^{tA} or its product with a vector $e^{tA}v$ has both theoretical and practical importance.

Many methods have been studied to efficiently compute this matrix function. The classical work of *Nineteen dubious ways to compute the exponential of a matrix* by C. Moler and C. Van Loan provides a thorough survey of the existing methods, see [30] for more details. For relatively small and dense matrices, the Padé approximation method with the scaling and squaring techniques is widely used, as in the MATLAB function expm(A). For large and sparse matrices which have become more and more common in practice, the Krylov subspace iterative methods are proved to be a powerful class of methods in dealing with many linear algebra computations. Very good approximations are often obtained within a relatively small number of iterations, and computable error bounds exist for the approximations.

The Krylov subspace methods for computing the matrix exponentials were introduced by Saad [33] and Gallopoulos and Saad [20]. They are some of the most efficient

methods for computing $e^{\tau A}v$. Since their introduction, many error bounds have been studied to explain the convergence properties of the Krylov subspace methods. Some *a posteriori* and *a priori* error bounds were first presented by Saad in [33]. More refined error bounds were later presented in [15, 16, 22, 31]. These bounds show that the speed of convergence depends on the norm of τA. This is natural since the Krylov approximation can be treated as a polynomial approximation, but it will limit the use of the Krylov subspace methods to the problems where the norm of τA is not too large. Meanwhile, treated as a projection method, the eigenvalue distribution also plays an important role in the convergence of the Krylov subspace methods. Under the assumption that A is negative definite guaranteeing the stability of the underlying dynamic system, Ye presented stronger bounds in [42] showing that the speed of convergence is determined by the condition number. Therefore, for a well conditioned matrix A, the convergence is fast no matter how large the norm of τA is.

This dissertation focuses on the influence of the eigenvalue distribution on the convergence of the Krylov subspace methods for computing $e^{\tau A}v$ for a non-symmetric A. To generalize the result in [42] to non-symmetric matrices, we make the assumption that the field of values of A lies entirely in the left half of the complex plane, i.e., A is negative definite. To be precise, we consider a rectangle in the left half of the complex plane containing the field of values of A. We derive error bounds by considering polynomial approximations of e^{tz} on the rectangular domain. Conformal mappings using the Jacobi elliptic functions are constructed that maps the exterior of the rectangle onto the exterior of the unit circle, and then the Faber polynomials are generated to find a sharper bound of exponentials of Hessenberg matrices. Our new error bounds show that the speed of convergence is related to the shape and the size of that rectangle, i.e., the eigenvalue distribution. The new bounds also agree with the bound in [42] when A is symmetric.

A special case of the computation of $e^{\tau A}v$ for a non-Hermitian A is that when A is skew-Hermitian. One physical application of this computation is in the solution of the time-dependent Schrödinger equation

$$ih\frac{\partial}{\partial t}\Psi(r,t) = \hat{H}\Psi(r,t), \tag{1.2}$$

where i is the imaginary unit, h is the Planck constant, Ψ is the wave function of the quantum system and $\hat{H}$ is the Hamiltonian operator. See [34] for more details. In this case, writing (1.2) in the form of (1.1), we have that $A = -\frac{i}{h}\hat{H}$ is a skew-Hermitian matrix. Then the eigenvalues of A are purely imaginary. This is a special case of the discussion above when the rectangle there containing the field of values degenerates into a line segment on the imaginary axis. For this problem, the solution has a very different behavior from the symmetric case in the sense that the approximation error first stagnates for certain number of iterations before it actually starts to converge. We will present new error bounds for this simpler case showing that the iteration number at which the actual convergence begins can be calculated before hand. This behavior is also demonstrated in our numerical tests.

This dissertation is organized as follows. In chapter 2, we discuss some basic properties, classical methods and existing error bounds for computing matrix exponentials. The field of values, the Faber polynomials, the Jacobi elliptic functions and the conformal mappings are also discussed, for the preparation of our deductions in the next two chapters. In chapter 3, we generalize the result in [42] to non-symmetric A and present the new *a posteriori*, *a priori* and numerically optimized error bounds for the computation of $e^{\tau A}v$. We then make the same approach to the case when A is skew-Hermitian in chapter 4 and present our new *a posteriori*, *a priori* and optimized error bound. Numerical tests are presented at the end of the chapter.

Chapter 2 Preliminaries and earlier results

In this chapter, we provide some preliminary results needed for the discussion in the next two chapters. Section 1 gives the formal definition and some basic properties of matrix exponentials. In Sections 2 and 3, we discuss some existing methods and error bounds for computing matrix exponentials. Since our work focuses on the role of the spectral information in the convergence of the Krylov subspace methods, we discuss the numerical range in Section 4 and the logarithmic norm in Section 5. In Section 6, we discuss the Faber polynomials, as a polynomial approximation to the exponential function. Finally, as a preparation for the next chapter, the Jacobi elliptic functions are discussed in Section 7.

2.1 Basic properties of matrix exponentials

In this section, we discuss some fundamentals of matrix exponentials, starting with a formal definition.

Definition 2.1. *Let A be an $n \times n$ real or complex matrix. The exponential of A, denoted by e^A or $exp(A)$, is the $n \times n$ matrix given by the power series*

$$e^A := \sum_{k=0}^{\infty} \frac{1}{k!} A^k.$$

In many applications, we are more interested in e^{tA} where t is usually a small positive scalar for time steps. Formally,

$$e^{tA} = I + tA + \frac{t^2 A^2}{2!} + \cdots . \tag{2.1}$$

The next theorem shows that the power series in (2.1) is uniformly convergent, thus e^{tA} is well defined for all t and A.

Theorem 2.2. *[4, Theorem 2, p. 170] The matrix series defined in* (2.1) *exists for all A for any fixed value of t, and for all t for any fixed A. It converges uniformly in any finite region of the complex t plane.*

Proof. Note that

$$\frac{||t^n A^n||}{n!} \leq \frac{|t|^n ||A||^n}{n!}.$$

So the series in (2.1) is dominated by the uniformly convergent series expansion of $e^{|t|\,||A||}$, and hence is itself uniformly convergent in any finite region of the complex t plane. □

For the convenience of future uses, we list without proof some basic rules of this matrix function. See [4] for more details.

Proposition 2.3. *Let A and B be $n \times n$ complex matrices, t and s be arbitrary complex numbers. Denote the $n \times n$ identity matrix by I and the zero matrix by 0. The matrix exponential satisfies the following properties.*

1. $e^0 = I$
2. $e^{(s+t)A} = e^{sA}e^{tA}$
3. $e^{tA}e^{-tA} = I$, *so* e^{tA} *is never singular.*
4. $e^{t(A+B)} = e^{tA}e^{tB}$ *if A and B commute, i.e.,* $AB = BA$.
5. *If B is invertible, then* $e^{B^{-1}AB} = B^{-1}e^A B$.
6. $\frac{d}{dt}e^{tA} = Ae^{tA}$
7. $det(e^A) = e^{trace(A)}$

For the classical problem of solving homogeneous systems of ordinary differential equations

$$\dot{x}(t) = Ax(t)$$

with the initial condition

$$x(0) = x_0,$$

the solution is given by

$$x(t) = e^{tA}x_0.$$

For the non-homogeneous problem

$$\dot{x}(t) = Ax(t) + b(t), \tag{2.2}$$

with the initial condition

$$x(0) = x_0,$$

we can construct the solution with above properties. Starting with

$$\dot{x}(t) - Ax(t) = b(t),$$

we have

$$e^{-tA}(\dot{x}(t) - Ax(t)) = e^{-tA}b(t),$$

which is

$$\frac{d}{dt}(e^{-tA}x(t)) = e^{-tA}b(t).$$

Integrated over a small time step τ,

$$\int_t^{t+\tau} \frac{d}{ds}(e^{-sA}x(s))ds = \int_t^{t+\tau} e^{-sA}b(s)ds,$$

so

$$e^{-\tau A}x(t+\tau) - x(t) = \int_t^{t+\tau} e^{-sA}b(s)ds.$$

The solution is

$$x(t+\tau) = e^{\tau A}x(t) + \int_t^{t+\tau} e^{(\tau-s)A}b(s)ds.$$

In a finite difference discretization of (2.2),

$$x(t+\tau) = e^{\tau A}x(t) + \int_0^{\tau} e^{(\tau-s)A}b(t+s)ds, \tag{2.3}$$

where τ is a time step parameter. This involves the calculation of the product of the matrix exponential with a vector for some small τ. The integral in (2.3) can be computed using some quadrature rule, which also involves computing $e^{\tau A}v$.

2.2 Classical methods for computing matrix exponentials

Dozens of methods have been studied for computing the matrix exponential e^A. The classical work [30] by Moler and Van Loan presents a thorough survey. In this chapter, we will briefly discuss some of the classical methods, among which the Padé approximation method with a proper scaling and squaring technique is one of the most efficient ways for small dense matrices.

Taylor series method

The Taylor series method is one of the most fundamental series methods for computing e^A. The class of series methods comes from the ideas of approximating the scalar function e^z. These methods intrinsically treat the matrix exponential purely as a matrix function analogous to the scalar exponential function. Therefore the specific information of the matrix, such as the order and the eigenvalues, will not play a direct role in the computation.

The Taylor series method is a straightforward application to the definition

$$e^A = \sum_{j=1}^{\infty} \frac{A^j}{j!} = I + A + \frac{A^2}{2!} + \cdots . \tag{2.4}$$

Let $P_K(A)$ be the partial sum of the series (2.4). Liou [25] presents an error bound which can serve as a truncation criterion

$$||P_K(A) - e^A|| \leq \left(\frac{||A||^{K+1}}{(K+1)!} \right) \left(\frac{1}{1 - \frac{||A||}{K+2}} \right).$$

This is the most fundamental method for computing the matrix exponential, but not satisfactory. Extreme examples have been constructed in [30] to show the catastrophic cancellation and illustrate its serious shortcoming in accuracy. For example, let

$$A = \begin{bmatrix} -49 & 24 \\ -64 & 31 \end{bmatrix},$$

then the Taylor series method gives

$$e^A \approx \begin{bmatrix} -22.25880 & -1.432766 \\ -61.49931 & -3.474280 \end{bmatrix}.$$

However, the matrix A was initially constructed as

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} -1 & 0 \\ 0 & -17 \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1},$$

so

$$e^A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} e^{-1} & 0 \\ 0 & e^{-17} \end{bmatrix} \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}^{-1} \approx \begin{bmatrix} -0.735759 & 0.551819 \\ -1.471518 & 1.103638 \end{bmatrix}.$$

For some special matrices, however, better results can be achieved. In a recent study of Xue and Ye, the Taylor series method is shown to be competitive for computing the exponentials of essentially non-negative matrices. A matrix is called essentially non-negative if all of its off-diagonal entries are non-negative. An entrywise perturbation analysis in [40] shows that if E is a small perturbation to A such that $|E| \leq \epsilon|A|$, then

$$|e^{A+E} - e^A| \leq \kappa_{\exp}(A) e^{\kappa_{\exp}\epsilon/(1-\epsilon)} \frac{\epsilon}{1-\epsilon} |e^A|,$$

wheres $\kappa_{\exp}(A)$ is determined by the spectral radius of A. Later in [41], Xue and Ye implemented the Taylor series method with shifting to achieve this entrywise relative accuracy. They derived a new criterion to truncate the series and presented an entrywise error analysis. The analysis shows that when carefully implemented, the entrywise relative error of the new algorithm based on the Taylor series method is comparable to the error made in rounding the matrix.

Padé approximation method

In mathematics a Padé approximant is the an approximation of a function by a rational function of given order, in the sense that the power series of the approximant agrees with that of the function it is approximating. The (p, q) Padé approximation to e^A is defined by

$$R_{pq}(A) = [D_{pq}(A)]^{-1} N_{pq}(A),$$

where

$$N_{pq}(A) = \sum_{j=0}^{p} \frac{(p+q-j)!p!}{(p+q)!j!(p-j)!} A^j$$

and

$$D_{pq}(A) = \sum_{j=0}^{q} \frac{(p+q-j)!q!}{(p+q)!j!(q-j)!} (-A)^j.$$

The non-singularity of $D_{pq}(A)$ is guaranteed if p and q are large enough.

The diagonal approximants where $p = q$ are usually preferred over the off-diagonal approximants. To see this, suppose $p < q$. Then the amount of flops required to compute an off-diagonal approximant $R_{pq}(A)$ is qn^3, which is the same amount of work for computing the diagonal $R_{qq}(A)$ with a higher order $2q > p + q$. So the diagonal approximants can be expected to be more accurate with the same amount of work.

Scaling and squaring method

When the norm of A is large, both the round off errors and the computing costs will make the above two methods less attractive. This difficulty can be controlled by the following scaling and squaring technique. First note the property

$$e^A = \left(e^{\frac{A}{2^k}}\right)^{2^k}.$$

We can choose the smallest integer k such that $\frac{||A||}{2^k}$ is smaller than a modest value, say, 1. Then $e^{\frac{A}{2^k}}$ can be efficiently computed by the Padé approximation method and e^A can be obtained by k repeated squarings. This approach is the one of the most effective methods we know to compute the exponential of a matrix. The implementation and error analysis has been fully discussed in many works, such as Ward [39]. In the field of applications, both MATLAB and GNU Octave use Padé approximants with the scaling and squaring technique. The MATLAB function expm is based on the algorithm in [21] by Higham. Since the accuracy and the efficiency are affected by both the norm of $\frac{A}{2^k}$ and the order q of the Padé approximant $R_{qq}(\frac{A}{2^k})$, different

choices and the corresponding error analysis have been studied to improve the behavior of the algorithm. In [21], Higham identified the most efficient choice for IEEE double precision arithmetic: $m = 13$ and $||\frac{A}{2^k}|| < 5.4$. The scheme of overscaling, which results in a value of k much larger than necessary, is also studied in another paper of Higham [2].

2.3 Krylov subspace methods

Existing studies show that the Padé approximation method with the scaling and squaring technique is effective in computing exponentials of small dense matrices. For large scale problems, the iterative methods are preferred over the traditional direct methods. Over the recent decades, the Krylov subspace methods become popular in dealing with many large scale linear algebra problems, such as solving linear systems and computing eigenvalues. As this is the method we study, we will discuss the basic ideas and the algorithms of Krylov subspace methods in this section.

It is first noticed that in many applications we do not really need the full matrix e^A, only its product $e^A v$ with some given vector v. For example, the solution to the homogeneous initial value problem

$$\dot{x}(t) = Ax(t),\ x(0) = x_0$$

is $x(t) = e^{tA}x_0$, in the form of the product of a matrix exponential and a vector. Here A is a large sparse matrix. We also note that in this situation, e^A is typically dense even if A itself is sparse.

The idea of the Krylov subspace methods is to approximately project the exponential of the large matrix onto a small Krylov subspace. After this, the only matrix exponential operation performed is therefore with a much smaller matrix. Specifically, we are interested in approximations to the matrix exponential operation $e^A v$ of the form

$$e^A v \approx p_{m-1}(A)v,$$

where A is a matrix of dimension n, v is a normalized vector, and p_{m-1} is a polynomial of degree $m-1$. So $p_{m-1}(A)v$ is an element of the Krylov subspace

$$K_m = span\{v, Av, \cdots, A^{m-1}v\}.$$

For the general non-symmetric case, we can use the usual Arnoldi algorithm or non-symmetric Lanczos algorithm. Both reduce to the symmetric Lanczos algorithm when the matrix A becomes symmetric. The following algorithm was presented in [33].

Algorithm 2.4. ***(Arnoldi Algorithm)***

1. *Initialize: Compute* $v_1 := v/||v||_2$.
2. *Iterate: Do* $j = 1, 2, \cdots, m$
   a) *Compute* $w := Av_j$
   b) *Do* $i = 1, 2, \cdots, j$
      i. *Compute* $h_{i,j} := (w, v_i)$
      ii. *Compute* $w := w - h_{i,j}v_i$
   c) *Compute* $h_{j+1,j} = ||w||_2$ *and* $v_{j+1} = w/h_{j+1,j}$.

This Arnoldi algorithm is applied to a non-symmetric A and a random vector v. Then $\{v_1, v_2, \cdots, v_m\}$ is an orthonormal basis of the Krylov subspace K_m and $V_m := [v_1, v_2, \cdots, v_m]$ is an orthogonal matrix of dimensions $n \times m$. Let $H_m := [h_{ij}]$ be the $m \times m$ upper Hessenberg matrix, then by our construction in Algorithm 2.4, we have

$$AV_m = V_m H_m + h_{m+1,m} v_{m+1} e_m^T.$$

By the orthogonality of V_m, we have $H_m = V_m^T A V_m$, which represents the projection of A onto the Krylov subspace K_m, with respect to the basis V_m. Then the Arnoldi

approximation was introduced as

$$e^A v \approx V_m e^{H_m} e_1.$$

The above method was introduced by Saad in [33]. An *a priori* error bound was also presented in [33, Theorem 4.5, p. 13] as

$$||e^A v - \beta V_m e^{H_m} e_1||_2 \leq 2\beta \frac{\rho_\alpha^m e^{\rho_\alpha + \alpha}}{m!},$$

where $\rho_\alpha = ||A - \alpha I||_2$ with any real scalar α, and $\beta = ||v||_2$. If A is symmetric negative definite and $\rho = ||A||_2$, a sharper error bound was given in [33, Corollary 4.6, p. 13] as

$$||e^A v - \beta V_m e^{H_m} e_1||_2 \leq \beta \frac{\rho^m}{m! 2^{m-1}}.$$

More sophisticated and refined error bounds for approximating the matrix exponential with the Arnoldi method have been studied later. In [22], Hochbruck and Lubich presented several bounds for the error $\epsilon_m := ||e^{\tau A} v - V_m e^{\tau H_m} e_1||$ where $||v|| = 1$. If A is a Hermitian negative semi-definite matrix with its eigenvalues in the interval $[-4\rho, 0]$, the error bound satisfies

$$\epsilon_m \leq 10 e^{-\frac{m^2}{5\rho\tau}}, \text{ if } \sqrt{4\rho\tau} \leq m \leq 2\rho\tau,$$
$$\epsilon_m \leq 10(\rho\tau)^{-1} \left(\frac{e\rho\tau}{m}\right)^m, \text{ if } m \geq 2\rho\tau.$$

If A is skew-Hermitian with its eigenvalue in an interval on the imaginary axis of length 4ρ, the error satisfies

$$\epsilon_m \leq 12 e^{-\frac{(\rho\tau)^2}{m}} \left(\frac{e\rho\tau}{m}\right)^m, \text{ if } m \geq 2\rho\tau.$$

For a non-symmetric A whose field of values contained in the disk $|z + \rho| \leq \rho$, the error satisfies

$$\epsilon_m \leq 12 e^{-\rho\tau} \left(\frac{e\rho\tau}{m}\right)^m, \text{ if } m \geq 2\rho\tau.$$

The error bounds above show that the speed of the convergence depends on the norm of τA. It is natural since the Arnoldi approximation $V_m e^{\tau H_m} e_1$ is after all a

polynomial approximation and is more accurate when the norm of τH_m is not too large. When dealing with the time stepping discretization as in (2.3), this may limit the time step parameter τ to be too small. At the same time, as a projection scheme, we also expect the eigenvalue distribution to affect the speed of the convergence. In [42], Ye showed that for symmetric matrices, the speed of the convergence is directly related to the condition number. Specifically, let $w_m(\tau) = V_m e^{-\tau T_m} e_1$ be the Lanczos approximation to $w(\tau) = e^{-\tau A} v$, where A is positive definite and $||v|| = 1$. The approximation error is then related to an element of the matrix e^{-tT_m} by the *a posteriori* bound

$$||w(\tau) - w_m(\tau)|| \leq \tau \beta_{m+1} \max_{0 \leq t \leq \tau} |h(t)|,$$

where $h(t) := e_m^T e^{-tT_m} e_1$ and $|\beta_{m+1}| \leq ||A||$. The convergence of the error comes from the decay property of functions of banded matrices and the decay rate depends on the condition number κ of T_m by

$$||w(\tau) - w_m(\tau)|| \leq \tau ||A|| (\sqrt{\kappa} + 1) \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^{m-1}. \tag{2.5}$$

This bound shows that the convergence rate of the Lanczos method in computing the matrix exponential is at least the same as that of the conjugate gradient method. A more general *a priori* bound is also presented in [42] as follows

$$||w(\tau) - w_m(\tau)|| \leq \alpha e^{(\alpha - \tau)\lambda_1} ||A|| \epsilon_1(m) + (\tau - \alpha) ||A|| \epsilon_2(m), \tag{2.6}$$

where

$$\epsilon_1(m) = \min \left\{ \frac{(\alpha \lambda_n / 2)^{m-1}}{(m-1)!}, \frac{2e^{\delta}}{1-q} q^{m-1} \right\}, \ \epsilon_2(m) = (\sqrt{\kappa} + 1) q_0^{m-1},$$

$q = \left(\frac{1}{q_0} + \frac{4\delta}{\alpha(\lambda_n - \lambda_1)} \right)^{-1}$, $q_0 = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}$, $\kappa = \frac{\lambda_n}{\lambda_1}$ and $0 \leq \alpha \leq \tau$. A proper weighted average of (2.6) will achieve an optimal error bound and best describe the actual behavior of the Lanczos algorithm. Our main goal in the next chapter is to generalize this result to non-symmetric matrices. That is, the computation of $e^{-\tau A} v$ where A is non-symmetric whose field of values is on the left half of the complex plane and $||v|| = 1$. We will relate the convergence rate to the field of values and show our result agrees with (2.5) when A degenerates to symmetric.

2.4 Numerical range

We start with the formal definition.

Definition 2.5. *In linear algebra, the numerical range or the field of values of a complex $n \times n$ matrix A is the set*

$$W(A) = \left\{ \frac{x^*Ax}{x^*x} : x \in \mathbb{C}^n, x \neq 0 \right\},$$

where x^* denotes the conjugate transpose of the vector x.

Immediately from the definition, the numerical range of a matrix A is the set of Rayleigh quotient. When A is Hermitian, the numerical range is a line segment which coincides with the spectral range. For a non-Hermitian A, the numerical range still contains all the eigenvalues of A. The next theorem provides a characterization of the numerical range.

Theorem 2.6. *[38]* ***(Hausdorff-Toeplitz Theorem)*** *The numerical range is convex and compact.*

For the computation of the matrix exponential $e^{-\tau A}v$ where $\tau > 0$, we are usually more interested in the case when $W(A)$ lies entirely on the right half of the complex plane. So,

Proposition 2.7. *The numerical range $W(A)$ is a subset of the closed right half plane if and only if $A + A^*$ is positive semidefinite.*

Proof. Note that

$$x^*Ax = x^*\frac{A + A^*}{2}x + x^*\frac{A - A^*}{2}x.$$

Since $\frac{A+A^*}{2}$ is Hermitian and $\frac{A-A^*}{2}$ is skew-Hermitian, the real and imaginary part of x^*Ax comes from $x^*\frac{A+A^*}{2}x$ and $x^*\frac{A-A^*}{2}x$, respectively. The real part is non-negative if and only if the matrix $A + A^*$ is positive semidefinite. □

The next theorem plays an important role in our work. It is an inequality proved by Michel Crouzeix, related to polynomial functions of a square matrix, involving the numerical range of the matrix.

Theorem 2.8. *[10]* ***(Crouzeix Theorem)*** *For a square matrix A and a polynomial p, the following inequality holds*

$$||p(A)||_2 \leq 11.08 \sup_{z \in W(A)} |p(z)|.$$

Note that by the maximum modulus principle, the maximum on the right hand side of the inequality must be attained on the boundary of $W(A)$. Crouzeix pointed out that the constant 11.08 is not optimal. For some special matrix A, it can be improved drastically to 2. Crouzeix also conjectured that

$$||p(A)||_2 \leq 2 \sup_{z \in W(A)} |p(z)|$$

is still generally true, but it is not proved.

2.5 Logarithmic norm

The logarithmic norm of a matrix was introduced in [11] by G. Dahlquist, in order to derive error bounds in initial value problems. The name logarithmic norm originates from estimating the logarithm of the norm of solutions of ordinary differential equations.

In this section we will discuss its original definition for matrices, but note that it can also be extended to bounded linear operators.

Definition 2.9. *Let A be a square matrix and $||\cdot||$ be a matrix norm. The associated logarithmic norm μ of A is defined as*

$$\mu(A) = \lim_{h \to 0^+} \frac{||I + hA|| - 1}{h}. \tag{2.7}$$

Here I is the identity matrix of the same dimension as A, and h is a real positive number.

Note that the limit in the definition is taken as $h \to 0^+$. When $h \to 0^-$ instead, the limit equals $-\mu(-A)$, which is generally smaller than $\mu(A)$. Furthermore, the logarithmic norm, despite its name, is not a matrix norm, since $\mu(A)$ may take negative values, e.g., when A is negative definite.

The next proposition sets up a direct link between the logarithmic norm and the spectral information of a matrix. In this sense, it may also serve as an alternative definition of the logarithmic norm.

Proposition 2.10. *Let A be a square matrix and $\langle\cdot\rangle$ be an inner product. If $||\cdot||$ is the induced matrix norm in* (2.7)*, the associated logarithmic norm μ of A is*

$$\mu(A) = \sup_{x\neq 0} \frac{Re\langle x, Ax\rangle}{\langle x, x\rangle}.$$

If $\langle\cdot\rangle$ is the Euclidean inner product and $||\cdot||$ is the associated 2-norm,

$$\mu(A) = \sup_{x\neq 0} Re\left\{\frac{x^*Ax}{x^*x}\right\} = \lambda_{\max}\left(\frac{A+A^*}{2}\right),$$

and

$$-\mu(-A) = -\sup_{x\neq 0} \mathrm{Re}\left\{-\frac{x^*Ax}{x^*x}\right\} = \inf_{x\neq 0} \mathrm{Re}\left\{\frac{x^*Ax}{x^*x}\right\} = \lambda_{\min}\left(\frac{A+A^*}{2}\right). \tag{2.8}$$

The next interesting proposition was also introduced in [11, p. 14]. The logarithmic norm is used to bound the norm of the matrix exponential. See [35] for more details of the proof.

Proposition 2.11. *Let $A \in \mathbb{C}^{n\times n}$ and $t \geq 0$. The matrix exponential is bounded by*

$$||e^{tA}|| \leq e^{t\mu(A)}.$$

We will use this proposition in the proof of Theorem 3.1, for an *a posteriori* error bound of the computation $e^{-\tau A}v$.

2.6 Faber polynomials

In 1903, G. Faber extended the theory of power series to domains more general than a disk. The polynomials he introduced have been since proved useful in analysis and known as Faber polynomials. It starts with a fundamental result in analysis.

Riemann's mapping theorem [27, Theorem 1.2, p. 8] states that every connected domain in the extended complex plane whose boundary contains more than one point can be mapped conformally onto a disk with its center at the origin. Now let $\bar{\mathbb{C}} =$

$\mathbb{C} \cup \{\infty\}$ be the extended complex plane and F be a continuum containing more than one point. A continuum is a non-empty, compact and connected subset of $\mathbb{C}$. If G_∞ is the complement or a component of the complement of F containing ∞, then G_∞ is a simply connected domain in $\bar{\mathbb{C}}$. Then [27, Theorem 3.14, p. 104] shows that there exists a function $w = \Phi(z)$ which maps G_∞ conformally onto the exterior of a circle of the form $|w| > \rho > 0$. Furthermore, the conformal mapping Φ also satisfies the normalization conditions

$$\Phi(\infty) = \infty, \ \lim_{z \to \infty} \frac{\Phi(z)}{z} = 1. \tag{2.9}$$

Under those conditions, the function $\Phi(z)$ has a Laurent expansion of the form

$$\Phi(z) = z + \alpha_0 + \frac{\alpha_{-1}}{z} + \cdots$$

at infinity. Moreover, given any integer $n > 0$, the function $[\Phi(z)]^n$ has a Laurent expansion of the form

$$[\Phi(z)]^n = z^n + \alpha_{n-1}^{(n)} z^{n-1} + \cdots + \alpha_0^{(n)} + \frac{\alpha_{-1}^{(n)}}{z} + \cdots$$

at infinity [27, Corollary, p. 104].

The Faber polynomials are defined as

$$\Phi_n(z) = z^n + \alpha_{n-1}^{(n)} z^{n-1} + \cdots + \alpha_0^{(n)}$$

consisting of the non-negative powers of z in the expansion above. We call them the Faber polynomials generated by the continuum F, or simply the Faber polynomials for F. The following two examples discussed in [27] show Faber polynomials generated by different subsets of $\mathbb{C}$.

Example 2.12. ***(Disk)***

If F is the closed disk $|z - z_0| \leq \rho$, then the Riemann mapping is

$$w = \Phi(z) = z - z_0,$$

and then

$$[\Phi(z)]^n = (z - z_0)^n.$$

Thus, the Faber polynomials consisting of the non-negative powers will be power functions

$$\Phi_n(z) = (z - z_0)^n,$$

same as $[\Phi(z)]^n$.

Example 2.13. ***(Line segment)***

Let $F = [-1, 1]$ be a line segment of the real axis in $\mathbb{C}$. Then

$$w = \Phi(z) = \frac{1}{2}\left(z + \sqrt{z^2 - 1}\right)$$

maps G_∞ conformally onto the domain $|w| > \frac{1}{2}$. Note that here we choose the branch of $\sqrt{z^2 - 1}$ such that

$$\lim_{z\to\infty} \frac{\sqrt{z^2 - 1}}{z} = 1.$$

It is observed that

$$\frac{1}{4\Phi(z)} = \frac{1}{2}\left(z - \sqrt{z^2 - 1}\right)$$

is finite when z is at infinity, thus the Laurent expansion of $\frac{1}{4\Phi(z)}$ at infinity contains no non-negative powers of z. So $\frac{1}{[4\Phi(z)]^n}$ does not have non-negative powers either. As a consequence, $[\Phi(z)]^n$ has the same Laurent expansion at infinity as the function

$$[\Phi(z)]^n + \frac{1}{[4\Phi(z)]^n} = \left[\frac{1}{2}\left(z + \sqrt{z^2 - 1}\right)\right]^n + \left[\frac{1}{2}\left(z - \sqrt{z^2 - 1}\right)\right]^n.$$

Since the above equation is a polynomial of degree n, the Faber polynomials are

$$\Phi_n(z) = [\Phi(z)]^n = \frac{1}{2^n}\left[\left(z + \sqrt{z^2 - 1}\right)^n + \left(z - \sqrt{z^2 - 1}\right)^n\right].$$

Set $z = \cos t$, then

$$\Phi_n(\cos t) = \frac{1}{2^n}[(\cos t + i \sin t)^n + (\cos t - i \sin t)^n] = \frac{1}{2^{n-1}} \cos nt,$$

or equivalently

$$\Phi_n(z) = \frac{1}{2^{n-1}} \cos(n \arccos z).$$

So the Faber polynomials associated to the line segment $[-1, 1]$ are actually the classical Chebyshev polynomials.

The Faber polynomials can be used to approximate analytic functions. Let Ψ be the inverse of Φ and the circular image C_R be the inverse image under $w = \Phi(z)$ of a circle $|w| = R > \rho$. The (Jordan) region with boundary C_R is then denoted by $I(C_R)$. By [27, Theorem 3.17, p. 109], every function $f(z)$ analytic on $I(C_{R_0})$, where $R_0 > \rho$, can be represented on $I(C_{R_0})$ as the sum of a series of the form

$$f(z) = \sum_{n=0}^{\infty} a_n \Phi_n(z) \tag{2.10}$$

with the coefficients

$$a_n = \frac{1}{2\pi i} \int_{|w|=R} \frac{f[\Psi(w)]}{w^{n+1}} dw = \frac{1}{2\pi i} \int_{C_R} \frac{f(z)\Phi'(z)}{[\Phi(z)]^{n+1}} dz.$$

The partial sum of the above series

$$\Pi_N(z) = \sum_{n=0}^{N} a_n \Phi_n(z) \tag{2.11}$$

is a polynomial of degree at most N, since each Φ_n is of degree n. Immediately from the construction of a_n, we have

$$|a_n| \leq \frac{M(R)}{R^n} \ (n = 0, 1, 2, \cdots), \tag{2.12}$$

by [27, Corollary, p. 109], where

$$M(R) := \max_{z \in C_R} |f(z)|.$$

More quantitative estimates for certain choices of the continuum F are presented in [17]. Assume that F is a closed Jordan region. By a Jordan region we mean a region F that is bounded and whose boundary Γ consists of pairwise disjoint closed Jordan curves. If Γ is rectifiable, there exists at most every point $z \in \Gamma$ a tangent vector that makes an angle $\Theta(z)$ with the positive real axis. We say that Γ has bounded total rotation V if

$$V = \int_{\Gamma} |d\Theta(z)| < \infty.$$

We note that $V \geq 2\pi$ and the equality holds if F is convex.

Theorem 2.14. *[17, Corollary 2.2] Let F be a Jordan region whose boundary Γ is of bounded total rotation V.*

1. *For any $N \geq 1$,*

$$||\Phi_N||_\infty \leq \frac{\rho^N V}{\pi}.$$

This bound is best possible in the sense that when $D \equiv [-1, 1]$, equality holds.

2. *Let f be an analytic function in the interior of C_R for any $R > \rho$, we have for any $N \geq 0$,*

$$||f - \Pi_N||_\infty \leq \frac{M(R)V}{\pi} \frac{\left(\frac{\rho}{R}\right)^{N+1}}{1 - \frac{\rho}{R}},$$

where $M(r) = \max_{z \in C_R} |f(z)|$ and V is the total rotation of the boundary of C_R. Here $||\cdot||_\infty$ denotes the uniform norm on C_R.

Back to Example 2.12, the Faber polynomials of the disk $|z - z_0| \leq \rho$ are

$$\Phi_n(z) = (z - z_0)^n,$$

and the circular images C_R are the circles $|z - z_0| = R$. The Faber expansion reduces to the Taylor series

$$\sum_{n=0}^{\infty} a_n (z - z_0)^n.$$

For Example 2.13, the Faber polynomials of the line segment $[-1, 1]$ are

$$\Phi_n(z) = \frac{1}{2^{n-1}} \cos(n \arccos z),$$

and the circular images are the ellipses

$$\frac{x^2}{\left(r + \frac{1}{4r}\right)^2} + \frac{y^2}{\left(r - \frac{1}{4r}\right)^2} = 1.$$

2.7 Jacobi elliptic functions

In this section, we have a brief discussion of the Jacobi elliptic functions, which will be used to construct a conformal mapping in the next section. For a more complete and strict theory, see [1]. Let us start with the general elliptic functions.

Elliptic functions

In complex analysis, an elliptic function is a meromorphic function that is periodical in two directions. A meromorphic function is a function that is holomorphic on an open set except a set of isolated points.

Definition 2.15. ***(Elliptic function)*** *An elliptic function is a function f meromorphic on $\mathbb{C}$ for which exist two non-zero complex numbers w_1 and w_2 with $\frac{w_1}{w_2} \notin \mathbb{R}$, such that $f(z) = f(z + w_1) = f(z + w_2)$ for all $z \in \mathbb{C}$.*

In this definition, the ratio $\tau = \frac{w_1}{w_2}$ must not be purely real, because if it is, the function reduces to a single periodic function if τ is rational, and a constant if τ is irrational. The periods w_1 and w_2 are usually labeled such that $\mathrm{Im}(\frac{w_1}{w_2}) > 0$. Just as a periodic function of a real variable is defined by its value on an interval, an elliptic function is determined by its values on a fundamental parallelogram, which then repeat in a lattice. Such a lattice is called a cell of an elliptic function.

Elliptic integrals

As indicated by the name, elliptic functions were first introduced as inverse functions of (incomplete) elliptic integrals. This theory was later improved by Carl Gustav Jakob Jacobi (1829) and widely used in many practical problems as they do not require notions of complex analysis to be defined and/or understood. So before the introduction of the Jacobi elliptic functions, we first state the definition and properties of elliptic integrals.

Definition 2.16. ***(Incomplete elliptic integrals)*** *Given a real parameter m with $0 < m < 1$, the (incomplete) Jacobi elliptic integral of the first kind is defined as*

$$F(\phi, m) = \int_0^{\phi} (1 - m \sin^2 \theta)^{-\frac{1}{2}} d\theta. \tag{2.13}$$

The (incomplete) Jacobi elliptic integral of the second kind is defined as

$$E(\phi, m) = \int_0^{\phi} (1 - m \sin^2 \theta)^{\frac{1}{2}} d\theta.$$

Note from the above definition that in general incomplete elliptic integrals are functions of two independent arguments: a real parameter $m \in (0,1)$ and an argument $\phi \in \mathbb{C}$. With $\phi = \frac{\pi}{2}$, the incomplete integrals become the complete integrals as defined below.

Definition 2.17. ***(Complete elliptic integrals)*** *Given a real parameter m with $0 < m < 1$, the complete Jacobi elliptic integrals of the first kind and the second kind are defined respectively as*

$$K(m) := F(\frac{\pi}{2}, m) = \int_0^{\frac{\pi}{2}} (1 - m\sin^2\theta)^{-\frac{1}{2}} d\theta,$$
$$E(m) := E(\frac{\pi}{2}, m) = \int_0^{\frac{\pi}{2}} (1 - m\sin^2\theta)^{\frac{1}{2}} d\theta.$$

By $m_1 := 1-m$ we denote the complementary parameter of m. Hence $0 < m_1 < 1$. For simplicity, we always use the following shorter version notations.

$$K := K(m) \tag{2.14}$$
$$E := E(m) \tag{2.15}$$
$$K' := K(m_1) = K(1-m) \tag{2.16}$$
$$E' := E(m_1) = E(1-m) \tag{2.17}$$

It is observed that K, E, K' and E' are all functions of $m \in (0,1)$. Here are some basic properties of the elliptic integrals. For more details, see [1], [29] and [24].

Proposition 2.18. *1. Directly from Definition 2.17, both K and K are positive-valued functions of m. Moreover, K and E are differentiable with respect to the parameter $m \in (0,1)$, and*

$$\frac{dK}{dm} = \frac{E - m_1 K}{2mm_1} \tag{2.18}$$
$$\frac{dE}{dm} = \frac{E - K}{2m} \tag{2.19}$$

2. By (2.18) and (2.19), K' and E' are also differentiable functions of $m \in (0,1)$ and

$$\frac{dK'}{dm} = -\frac{E' - mK'}{2mm_1}$$

$$\frac{dE'}{dm} = -\frac{E' - K'}{2m_1}$$

3. *[1, 17.3.13, p. 591]* ***(Legendre's relation)*** *For any* $m \in (0, 1)$,

$$KE' + K'E - KK' = \frac{\pi}{2}.$$

4. *[1, 17.3.11-12, p. 591] Infinite series:*

$$K(m) = \frac{\pi}{2} \sum_{n=0}^{\infty} \left[\frac{(2n)!}{2^{2n}(n!)^2} \right]^2 m^n = \frac{\pi}{2} \sum_{n=0}^{\infty} \left[\frac{(2n-1)!!}{(2n)!!} \right]^2 m^n \tag{2.20}$$

$$E(m) = \frac{\pi}{2} \sum_{n=0}^{\infty} \left[\frac{(2n)!}{2^{2n}(n!)^2} \right]^2 \frac{m^n}{1-2n} = \frac{\pi}{2} \sum_{n=0}^{\infty} \left[\frac{(2n-1)!!}{(2n)!!} \right]^2 \frac{m^n}{1-2n} \tag{2.21}$$

5. *[1, 17.3.25, p. 591]*

$$\lim_{m \to 0} [K'(E - K)] = 0$$

6. *[1, 17.3.26, p. 591]*

$$\lim_{m \to 1} \left[K - \frac{1}{2} \ln \left(\frac{16}{m_1} \right) \right] = 0 \tag{2.22}$$

7. *[1, 17.3.27, p. 591]*

$$\lim_{m \to 0} [m^{-1}(K - E)] = \lim_{m \to 0} [m^{-1}(E - m_1 K)] = \frac{\pi}{4}$$

8. *[1, 17.4.5, p. 592]*

$$E(u + 2iK') = E(u) + 2i(K' - E') \tag{2.23}$$

Jacobi elliptic functions

Now we are well prepared for the introduction of the Jacobi elliptic functions. There are a total of twelve Jacobi elliptic functions in the family, but we are only going to discuss the basic three of them.

Definition 2.19. ***(Jacobi elliptic functions)*** *If $u = F(\phi, m)$ where $F(\cdot, m)$ is the incomplete elliptic integral of the first kind defined in* (2.13)*, three of the Jacobi elliptic functions are defined as*

$$sn(u|m) = \sin\phi \tag{2.24}$$

$$cn(u|m) = \cos\phi \tag{2.25}$$

$$dn(u|m) = \sqrt{1 - m\sin^2\phi} \tag{2.26}$$

For a fixed $m \in (0,1)$, $sn(u|m)$, $cn(u|m)$ and $dn(u|m)$ are doubly periodical meromorphic functions defined on $u \in \mathbb{C}$. The following table lists the periods, zeros, poles and residues of the three Jacobi elliptic functions [29, p. 14].

	$sn(u)$	$cn(u)$	$dn(u)$
Periods	$4K, 2iK'$	$4K, 2K + 2iK'$	$2K, 4iK'$
Zeros	$0, 2K$	$K, 3K$	$K + iK', K + 3iK'$
Poles	$iK', 2K + iK'$	$iK', 2K + iK'$	$iK', 3iK'$
Residues	$\sqrt{m}, -\sqrt{m}$	$-i\sqrt{m}, i\sqrt{m}$	$-i, i$

Table 2.1: Jacobi elliptic functions: periods, zeros, poles, residues

In addition, we list some fundamental facts about the functions $sn(u|m)$, $cn(u|m)$ and $dn(u|m)$.

Proposition 2.20. *Assume $m \in (0,1)$ and $u \in \mathbb{C}$, then we have the following properties.*

1. *Directly from Definition 2.19,*

$$sn^2(u|m) + cn^2(u|m) = 1$$
$$m \cdot sn^2(u|m) + dn^2(u|m) = 1$$

2. *[1, Table 16.2, p. 570] Periods: sn, cn and dn are one-valued, doubly-periodic functions. For any $l, n \in \mathbb{Z}$,*

$$sn(u + 2lK + 2niK'|m) = (-1)^l sn(u|m)$$
$$cn(u + 2lK + 2niK'|m) = (-1)^{l+n} cn(u|m)$$
$$dn(u + 2lK + 2niK'|m) = (-1)^n dn(u|m)$$

3. [1, Table 16.8, p. 572]

$$\begin{aligned} sn(2iK' - \sigma|m) &= sn(-\sigma|m) = -sn(\sigma|m) \\ cn(2iK' - \sigma|m) &= -cn(-\sigma|m) = -cn(\sigma|m) \\ dn(2iK' - \sigma|m) &= -dn(-\sigma|m) = -dn(\sigma|m) \end{aligned} \tag{2.27}$$

4. [1, Table 16.16, p. 574] Derivatives:

$$\frac{d}{du}sn(u|m) = cn(u|m) \cdot dn(u|m) \tag{2.28}$$

$$\frac{d}{du}cn(u|m) = -sn(u|m) \cdot dn(u|m) \tag{2.29}$$

$$\frac{d}{du}dn(u|m) = -m \cdot sn(u|m) \cdot cn(u|m) \tag{2.30}$$

5. [1, 16.21, p. 575] Write $u = x + iy$ where $x, y \in \mathbb{R}$. For simplicity, denote

$$\begin{aligned} &s = sn(x|m), c = cn(x|m), d = dn(x|m), \\ &s_1 = sn(y|m_1), c_1 = cn(y|m_1), d_1 = dn(y|m_1), \end{aligned}$$

then

$$sn(x + iy|m) = \frac{s \cdot d_1 + ic \cdot d \cdot s_1 \cdot c_1}{c_1^2 + ms^2 \cdot s_1^2} \tag{2.31}$$

$$cn(x + iy|m) = \frac{c \cdot c_1 + is \cdot d \cdot s_1 \cdot d_1}{c_1^2 + ms^2 \cdot s_1^2} \tag{2.32}$$

$$dn(x + iy|m) = \frac{d \cdot c_1 \cdot d_1 + ims \cdot c \cdot s_1}{c_1^2 + ms^2 \cdot s_1^2} \tag{2.33}$$

In our future discussion in Chapter 2, we will work on the three Jacobi elliptic functions $sn(u|m)$, $cn(u|m)$ and $dn(u|m)$ where the parameter $m \in (0, 1)$ and $u \in \mathbb{C}$ is in the rectangular domain $[-K, K] \times [0, 2iK']$. So, it is illustrative to figure out the range of these three functions in this specific domain. As a matter of fact, in our future discussion, it suffices to know the signs of the real and imaginary part of $sn(u|m)$, $cn(u|m)$ and $dn(u|m)$ when $\mathrm{Re}(u) \in [-K, K]$ and $\mathrm{Im}(u) \in [0, 2K']$. This is discussed in [24, pp. 172-176] and we summarize it in the following Table 2.2, 2.3 and 2.4 for future references.

$(K', 2iK')$	$- -$	$+ -$
$(0, K')$	$- +$	$+ +$
Re / Im	$(-K, 0)$	$(0, K)$

Table 2.2: Sign of values of $sn(u)$

$(K', 2iK')$	$- +$	$- -$
$(0, K')$	$+ +$	$+ -$
Re / Im	$(-K, 0)$	$(0, K)$

Table 2.3: Sign of values of $cn(u)$

$(K', 2iK')$	$- +$	$- -$
$(0, K')$	$+ +$	$+ -$
Re / Im	$(-K, 0)$	$(0, K)$

Table 2.4: Sign of values of $dn(u)$

Chapter 3 Error bounds for computing $e^{-\tau A}v$

In this chapter, we will discuss the computation of

$$w(\tau) := e^{-\tau A}v \tag{3.1}$$

with the Arnoldi method, where A is a non-symmetric positive semi-definite matrx, v is a real normalized vector and τ is a positive scalar. The chapter is organized as follows. In Section 1, the Arnoldi approximation to $w(\tau) = e^{-\tau A}v$ is defined and an *a posteriori* error bound is presented, which relates the error to an entry of the exponential of an upper Hessenberg matrix. To investigate the decay property of that entry, in Section 2, we discuss the conformal mapping needed for the construction of the Faber polynomial approximation to the exponential function. Then a new *a priori* error bound is presented in Section 3. In Section 4, we further optimize our new bound numerically to better describe the actual convergence of the Arnoldi method. Numerical examples are presented in Section 5.

3.1 *A posteriori* error bound

Let A be an n-by-n real non-symmetric matrix and v be an n-dimensional real normalized vector. We apply the Arnoldi method in Algorithm 2.4 to A and v. The first k iterations of the Arnoldi process generates a Krylov subspace

$$K_{k+1}(A, v) = span\{v, Av, A^2v, \cdots, A^kv\}$$

with an orthonormal basis $\{v_1, v_2, \cdots, v_k, v_{k+1}\}$. Meanwhile, a k-by-k upper Hessenberg matrix H_k is generated satisfying

$$AV_k = V_kH_k + \beta_{k+1}v_{k+1}e_k^T, \tag{3.2}$$

where $V_k = [v_1, v_2, \cdots, v_k]$ and $e_k \in \mathbb{R}^n$ is the k-th coordinate vector. Then for $w(\tau) = e^{-\tau A}v$ in the n-dimensional space, we can use $V_kV_k^Te^{-\tau A}v$, the orthogonal

projection of $e^{-\tau A}v$ on $K_k(A, v)$, as the best approximation of $e^{-\tau A}v$ from the k-dimensional subspace $K_k(A, v)$. By the orthogonality of the columns of V_k and (3.2),

$$V_k^T A V_k = V_k^T V_k H_k + \beta_{k+1} V_k^T v_{k+1} e_k^T = H_k,$$

then we have our approximation

$$V_k V_k^T e^{-\tau A} v = V_k V_k^T e^{-\tau A} V_k e_1 \approx V_k e^{-\tau V_k^T A V_k} e_1 = V_k^T e^{-\tau H_k} e_1.$$

We call

$$w_k(\tau) := V_k^T e^{-\tau H_k} e_1 \tag{3.3}$$

the Arnoldi approximation to $w(\tau)$ in (3.1). The next theorem is the first result of this chapter. It relates the approximation error of he Arnoldi method to the $(k, 1)$ entry of the matrix e^{-tH_k}. We denote the quantity defined by (2.8) that

$$\nu(A) := -\mu(-A) = \lambda_{\min}\left(\frac{A + A^*}{2}\right).$$

Theorem 3.1. ***(A posteriori error bound)*** *Assume that $A \in \mathbb{R}^{n\times n}$ with $\nu(A) = \lambda_{min}\left(\frac{A+A^*}{2}\right) > 0$, $v \in \mathbb{R}^n$ with $||v|| = 1$. Let V_k be the orthogonal matrix and H_k be the upper Hessenberg matrix generated by the Arnoldi process satisfying* (3.2)*. Let $w_k(\tau) = V_k e^{-\tau H_k} e_1$ in* (3.3) *be the Arnoldi approximation to $w(\tau) = e^{-\tau A}v$ in* (3.1)*. Then the approximation error satisfies*

$$||w(\tau) - w_k(\tau)|| \le \tau \beta_{k+1} \max_{0\le t\le \tau} |h(t)|, \tag{3.4}$$

where

$$h(t) := e_k^T e^{-tH_k} e_1 \tag{3.5}$$

is the $(k, 1)$ entry of the matrix e^{-tH_k}.

Proof. First, $w(t) = e^{-tA}v$, so $w'(t) = -Ae^{-tA}v = -Aw(t)$. Since $w_k(t) = V_k e^{-tH_k} e_1$, we have

$$w_k'(t) = -V_k H_k e^{-tH_k} e_1$$

$$
\begin{aligned}
&= -(AV_k - \beta_{k+1} v_{k+1} e_k^T) e^{-tH_k} e_1 \\
&= -AV_k e^{-tH_k} e_1 + \beta_{k+1} v_{k+1} e_k^T e^{-tH_k} e_1 \\
&= -Aw_k(t) + \beta_{k+1} h(t) v_{k+1},
\end{aligned}
$$

where $h(t) = e_k^T e^{-tH_k} e_1$. Let $E_k(t) := w(t) - w_k(t)$ be the approximation error. Then

$$
\begin{aligned}
E_k'(t) &= w'(t) - w_k'(t) \\
&= -Aw(t) - (-Aw_k(t) + \beta_{k+1} h(t) v_{k+1}) \\
&= -AE_k(t) - \beta_{k+1} h(t) v_{k+1}.
\end{aligned}
$$

Note that the initial condition

$$
E_k(0) = w(0) - w_k(0) = v - V_k e_1 = 0,
$$

and solve the initial value problem for $E_k(t)$, then we have

$$
E_k(\tau) = -\beta_{k+1} \int_0^\tau h(t) e^{(t-\tau)A} v_{k+1} dt.
$$

Since $\tau - t > 0$, we have

$$
||e^{(t-\tau)A}|| = ||e^{(\tau-t)(-A)}|| \leq e^{(\tau-t)\mu(-A)} = e^{(t-\tau)\nu(A)}. \tag{3.6}
$$

by Proposition 2.11. Then using (3.6), the approximation error satisfies

$$
\begin{aligned}
||E_k(\tau)|| &\leq \beta_{k+1} \left|\left| \int_0^\tau h(t) e^{(t-\tau)A} v_{k+1} dt \right|\right| \\
&\leq \beta_{k+1} \int_0^\tau |h(t)| \cdot ||e^{(t-\tau)A}|| \cdot |v_{k+1}| dt \\
&\leq \beta_{k+1} \cdot \max_{0 \leq t \leq \tau} |h(t)| \cdot \int_0^\tau e^{(t-\tau)\nu(A)} dt \\
&= \beta_{k+1} \cdot \max_{0 \leq t \leq \tau} |h(t)| \cdot \frac{1 - e^{-\tau\nu(A)}}{\nu(A)} \\
&\leq \tau \beta_{k+1} \max_{0 \leq t \leq \tau} |h(t)|.
\end{aligned}
$$

Note that $\nu(A) > 0$, then the last inequality comes from $1 - e^{-x} \leq x$ for $x > 0$. $\square$

Our next objective is to bound $h(t)$ in (3.5) with the spectral information of A. We consider an analytic function $f(z) = e^{-tz}$, then $h(t) = [f(H_k)]_{k1}$. In the next section, we will construct a conformal mapping which maps the exterior of the set containing the field of values of A conformally to the exterior of the unit circle.

3.2 Conformal mapping

In this section, it will be our sole interest to construct a proper conformal mapping which maps the exterior of a rectangle onto the exterior of a unit disk. For our practical purposes, it suffices to consider the rectangles which lie entirely in the right half of the extended complex plane and have symmetry with respect to the positive real axis. Formally speaking, given a rectangle in $\tilde{z}$-plane whose vertices are $a \pm ic$ and $b \pm ic$ where $b > a > 0$ and $c > 0$, we map the exterior of this rectangle conformally onto $|u| > 1$. Some necessary preparations are needed before the construction.

Recall in Section 2.7 that the complete elliptic integrals K, K', E and E' are all real functions of the parameter $m \in (0, 1)$ or its complementary parameter $m_1 = 1 - m$. First we have the following lemma.

Lemma 3.2. *For any $0 < \alpha, \beta < +\infty$, there exists a unique $m \in (0, 1)$ satisfying*

$$\frac{E - m_1 K}{\beta} = \frac{E' - mK'}{\alpha}, \tag{3.7}$$

where K, K', E and E' are complete elliptic integrals as in (2.14), (2.15), (2.16) *and* (2.17) *in Section 2.7.*

Proof. Let

$$f(m) := E - m_1 K = E(m) - (1 - m)K(m)$$

be a function of $m \in (0, 1)$. By the definition of $K(m)$ and $E(m)$ in Definition 2.17, $K(0) = \frac{\pi}{2}$ and $E(0) = \frac{\pi}{2}$, then

$$\lim_{m \to 0} f(m) = 0. \tag{3.8}$$

Moreover, by (2.22),

$$\lim_{m \to 1} m_1 \left[K(m) - \frac{1}{2} \ln \left(\frac{16}{m_1} \right) \right] = 0,$$

and therefore

$$\lim_{m \to 1} m_1 K(m) = \lim_{m \to 1} m_1 \ln \left(\frac{16}{m_1} \right) = \lim_{m_1 \to 0} m_1 \ln \left(\frac{16}{m_1} \right) = 0.$$

Again by the definition of $E(m)$, $E(1) = 1$. Then

$$\lim_{m\to 1} f(m) = E(1) - \lim_{m\to 1} m_1 K(m) = 1. \tag{3.9}$$

By (2.18) and (2.19), $f(m)$ is differentiable in $(0, 1)$ and

$$\frac{d}{dm} f(m) = \frac{K(m)}{2} > 0.$$

So f is an increasing function of m over $(0, 1)$. Now consider

$$g(m) := \frac{f(m)}{f(1-m)} = \frac{E(m) - (1-m)K(m)}{E(1-m) - mK(1-m)}. \tag{3.10}$$

By (3.8) and (3.9), $g(m)$ is an increasing function of m over $(0, 1)$ with

$$\lim_{m\to 0} g(m) = 0, \ \lim_{m\to 1} g(m) = +\infty.$$

Then for any $0 < \alpha, \beta < +\infty$, there exists a unique $m \in (0, 1)$ such that $g(m) = \frac{\beta}{\alpha}$, i.e., (3.7). □

Now in the following three steps, we can construct the conformal mapping from the exterior of the rectangle $[a, b] \times [-c, c]$ to the exterior of the unit circle.

- Step 1:

$$z = \phi_1(\tilde{z}) = \tilde{z} - \frac{a+b}{2} \tag{3.11}$$

  shifts the original rectangle to a new rectangle with vertices $\pm\alpha \pm i\beta$, where $\alpha = \frac{b-a}{2}$ and $\beta = c$.

- Step 2: $\phi_2 : z \mapsto w$ is defined through an auxiliary variable σ by

$$\begin{cases} z = \alpha - \dfrac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\ w = \dfrac{1 - dn(\sigma|m)}{\sqrt{m}sn(\sigma|m)} \end{cases} \tag{3.12}$$

  where the parameter m is determined by (3.7), and λ is defined to be the ratio in (3.7). ϕ_2 maps the exterior of the rectangle $[-\alpha, \alpha] \times [-\beta, \beta]$ to the upper half plane $\{\mathrm{Im}(w) > 0\}$. This mapping was presented in [24]. It also shows that the domain of σ is in the rectangle $[-K, K] \times [0, 2iK']$ in the complex plane, where K and K' are complete elliptic integrals of the first kind in (2.14).

- Step 3:

$$u = \phi_3(w) = \frac{i+w}{i-w} \tag{3.13}$$

maps $\{\mathrm{Im}(w) > 0\}$ onto $\{|u| > 1\}$.

Now let

$$\tilde{\Phi} := \phi_3 \circ \phi_2 \circ \phi_1 \tag{3.14}$$

be the composition of the above three conformal mappings defined in (3.11), (3.12) and (3.13). Then $\tilde{\Phi}$ maps the exterior of the rectangle $[a,b] \times [-c,c]$ conformally onto the exterior of the unit circle.

We denote by C_r in the $\tilde{z}$-plane the inverse image of $|u| = r > 1$ under $\tilde{\Phi}$. In our future discussion, we are particularly interested in the minimum of $\mathrm{Re}(\tilde{z})$ in C_r for a given $r > 1$. First we prove a lemma about the Jacobi elliptic functions. It is a direct result of Proposition 2.20 and will be needed later.

Lemma 3.3. *For $u = x + iy$ where $-K < x < K$ and $0 < y < 2K'$,*

$$\mathrm{sgn}(\mathrm{Im}(cn(u|m))) = \mathrm{sgn}(\mathrm{Im}(dn(u|m))),$$

where cn and dn are Jacobi elliptic functions defined in (2.25) *and* (2.26).

Proof. By (2.32) and (2.33),

$$\mathrm{Im}(cn(u|m)) = \frac{sn(x|m)dn(x|m)sn(y|m_1)dn(y|m_1)}{1 - dn^2(x|m)sn^2(y|m_1)}$$
$$\mathrm{Im}(dn(u|m)) = \frac{m \cdot sn(x|m)cn(x|m)sn(y|m_1)}{1 - dn^2(y|m)sn^2(y|m_1)}.$$

So,

$$\mathrm{sgn}(\mathrm{Im}(cn(u|m))) = \mathrm{sgn}(\mathrm{Im}(dn(u|m))) \cdot \mathrm{sgn}(cn(x|m) \cdot dn(x|m) \cdot dn(y|m_1)) \tag{3.15}$$

Write $x = F(\phi, m)$. When $-K < x < K$, we have $\phi \in (-\frac{\pi}{2}, \frac{\pi}{2})$. So,

$$cn(x|m) = \cos\phi > 0. \tag{3.16}$$

By the definition of $dn(u|m)$, for any $x, y \in \mathbb{R}$,

$$dn(x|m) > 0,\ dn(y|m_1) > 0. \tag{3.17}$$

Applying (3.16) and (3.17) to (3.15), we conclude that the imaginary part of $cn(u|m)$ and that of $dn(u|m)$ always share the same sign. □

The following lemma shows that the minimum of $\mathrm{Re}(\tilde{z})$ in C_r is attained at the inverse of $u = -r$.

Lemma 3.4. *Let $\tilde{\Phi} : \tilde{z} \mapsto u$ be the conformal mapping from the exterior of the rectangle $[a, b] \times [-c, c]$ onto the exterior of the unit disk, as defined in* (3.14)*. Let $\tilde{\Psi} : u \mapsto \tilde{z}$ be its inverse mapping and C_r be the image of $|u| = r > 1$ under $\tilde{\Psi}$. Then*

$$\min\{\mathrm{Re}(\tilde{z}) : \tilde{z} \in C_r\} = \tilde{\Psi}(-r).$$

Proof. By (3.11),

$$\frac{d\tilde{z}}{dz} = 1. \tag{3.18}$$

Here and below we will write sn, cn and dn in short of $sn(\sigma|m)$, $cn(\sigma|m)$ and $dn(\sigma|m)$, respectively. Recall the definition $E(\sigma|m) = \int_0^\sigma dn^2(z|m)dz$, the identities $sn^2 + cn^2 \equiv 1$ and $m \cdot sn^2 + dn^2 \equiv 1$, we have from (3.12) that

$$\frac{dz}{d\sigma} = -\frac{i}{\lambda}\{dn^2 - (1-m)\} = -\frac{i}{\lambda}\{m - m \cdot sn^2\} = -\frac{i}{\lambda} \cdot m \cdot cn^2. \tag{3.19}$$

Note that By (2.28) and (2.30), we have $\frac{d(dn)}{d\sigma} = -m \cdot sn \cdot cn$ and $\frac{d(sn)}{d\sigma} = cn \cdot dn$. Then by (3.12),

$$\begin{aligned}
\frac{dw}{d\sigma} &= \frac{-(-m \cdot sn \cdot cn) \cdot \sqrt{m} \cdot cn - (1 - dn) \cdot \sqrt{m} \cdot cn \cdot dn}{m \cdot sn^2} \\
&= \frac{\sqrt{m} \cdot cn \cdot (m \cdot sn^2 - dn + dn^2)}{m \cdot sn^2} \\
&= \frac{\sqrt{m} \cdot cn \cdot (1 - dn)}{1 - dn^2} \\
&= \frac{\sqrt{m} \cdot cn}{1 + dn}
\end{aligned} \tag{3.20}$$

By (3.12), $w = i\frac{u-1}{u+1}$ and then

$$\frac{dw}{du} = \frac{2i}{(u+1)^2}. \tag{3.21}$$

Combining (3.18), (3.19), (3.20) and (3.21), we have

$$\begin{aligned}\frac{d\tilde{z}}{du} &= \frac{d\tilde{z}}{dz} \cdot \frac{dz}{d\sigma} \cdot \frac{d\sigma}{dw} \cdot \frac{dw}{du} \\ &= -\frac{i}{\lambda} \cdot m \cdot cn^2 \cdot \frac{1+dn}{\sqrt{m} \cdot cn} \cdot \frac{2i}{(u+1)^2} \\ &= \frac{2\sqrt{m} \cdot cn(1+dn)}{\lambda(u+1)^2}. \end{aligned} \tag{3.22}$$

By (3.13), we have

$$w^2 = -\frac{(u-1)^2}{(u+1)^2}.$$

On the other hand, by (3.12),

$$w^2 = \frac{(1-dn)^2}{m \cdot sn^2} = \frac{(1-dn)^2}{1-dn^2} = \frac{1-dn}{1+dn}.$$

So,

$$dn = \frac{1-w^2}{1+w^2} = \frac{(u+1)^2 + (u-1)^2}{(u+1)^2 - (u-1)^2} = \frac{1}{2}\left(u + \frac{1}{u}\right) \tag{3.23}$$

and hence

$$1 + dn = \frac{(u+1)^2}{2u}. \tag{3.24}$$

Applying (3.24) to (3.22), we have

$$\frac{d\tilde{z}}{du} = \frac{\sqrt{m} \cdot cn}{\lambda u}. \tag{3.25}$$

Now let u be on the circle of radius r on the complex u-plane, then we can write $u = re^{i\theta}$ where $-\pi < \theta \leq \pi$, then

$$\frac{du}{d\theta} = re^{i\theta} \cdot i = iu. \tag{3.26}$$

Treating $\tilde{z} \in C_r$ as a function of θ, we have from (3.25) and (3.26) that

$$\frac{d\tilde{z}}{d\theta} = \frac{i\sqrt{m}}{\lambda} \cdot cn(\sigma|m). \tag{3.27}$$

So

$$\frac{d(\mathrm{Re}(\tilde{z}))}{d\theta} = \mathrm{Re}\left(\frac{d\tilde{z}}{d\theta}\right) = -\frac{\sqrt{m}}{\lambda}\,\mathrm{Im}(cn(\sigma|m)).$$

From (3.23) and $u = r\cos\theta + ir\sin\theta$, we write $dn(\sigma|m)$ as a function of θ,

$$dn(\sigma|m) = \frac{1}{2}\left(r + \frac{1}{r}\right)\cos\theta + \frac{i}{2}\left(r - \frac{1}{r}\right)\sin\theta.$$

So $\mathrm{Im}(dn(\sigma|m)) < 0$ when $\theta \in (-\pi, 0)$, and $\mathrm{Im}(dn(\sigma|m)) > 0$ when $\theta \in (0, \pi]$. By Lemma 3.3, the imaginary part of $cn(\sigma|m)$ always has the same sign as that of $dn(\sigma|m)$. Thus, by (3.27), $\frac{d(\mathrm{Re}(\tilde{z}))}{d\theta} > 0$ when $\theta \in (-\pi, 0)$, and $\frac{d(\mathrm{Re}(\tilde{z}))}{d\theta} < 0$ when $\theta \in (0, \pi]$. The minimum value of $\mathrm{Re}(\tilde{z})$ is attained when $\theta = \pi$, i.e., $u = -r$. □

Next, we find the explicit form for $\tilde{\Psi}(-r)$ in Lemma 3.4.

Lemma 3.5. *Let $\tilde{\Phi} : \tilde{z} \mapsto u$ be the conformal mapping from the exterior of the rectangle $[a, b] \times [-c, c]$ onto the exterior of the unit disk, as defined in* (3.14)*. Let $\tilde{\Psi} : u \mapsto \tilde{z}$ be its inverse mapping and C_r be the image of $|u| = r > 1$ under $\tilde{\Psi}$. Then for any $r > 1$,*

$$\tilde{\Psi}(-r) = a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r - \frac{1}{r}\right)} \frac{\sqrt{m + t^2}}{\sqrt{1 + t^2}}\,dt,$$

where the parameters m and λ are determined by (3.7)*.*

Proof. Before proving the lemma, we first recall in the construction of $\tilde{\Phi}$ and its inverse $\tilde{\Psi}$ in (3.14) that

$$\tilde{\Phi} = \phi_3 \circ \phi_2 \circ \phi_1$$

is the composition of the three conformal mappings defined in (3.11), (3.12) and (3.13). In addition to this, we further denote the composition of two mappings by

$$\Phi := \phi_3 \circ \phi_2 \tag{3.28}$$

and its inverse by Ψ. Then obviously

$$\tilde{\Psi}(-r) = \phi_1^{-1} \circ \Psi(-r) \tag{3.29}$$

The proof of this lemma consists of two parts. First, we prove that for any $r > 1$,

$$\Psi(r) = \alpha + \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+t^2}}{\sqrt{1+t^2}}dt. \tag{3.30}$$

By (3.12) and (3.23), Φ: $z \longleftrightarrow \sigma \longleftrightarrow u$ has an explicit form through the auxiliary parameter σ that

$$\begin{cases} z(\sigma) = \alpha - \dfrac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\ dn(\sigma|m) = \dfrac{1}{2}\left(u + \dfrac{1}{u}\right) \end{cases} \tag{3.31}$$

In (3.31), let $u = r$ where $r > 1$. Then

$$dn(\sigma|m) = \frac{1}{2}\left(r + \frac{1}{r}\right) > 1. \tag{3.32}$$

By Table 2.4, $\sigma \in \mathbb{C}$ is on the line segment connecting 0 and iK'. Let

$$t = -i\sqrt{m} \cdot sn(z|m), \tag{3.33}$$

where z is on the line segment connecting 0 and σ. By Table 2.2, 2.3 and 2.4, $sn(z|m)$ is purely imaginary with positive imaginary part, $cn(z|m)$ and $dn(z|m)$ are both real and positive. Then

$$\begin{aligned} &m \cdot sn^2(z|m) = -t^2, \\ &m \cdot cn^2(z|m) = m - m \cdot sn^2(z|m) = m + t^2 \Longrightarrow \sqrt{m} \cdot cn(z|m) = \sqrt{m+t^2}, \\ &dn^2(z|m) = 1 - m \cdot sn^2(z|m) = 1 + t^2 \Longrightarrow dn(z|m) = \sqrt{1+t^2}. \end{aligned}$$

By (3.33) and (2.28),

$$dt = -i\sqrt{m} \cdot cn(z|m) \cdot dn(z|m)dz,$$

then

$$dz = \frac{dt}{-i\sqrt{m} \cdot cn(z|m) \cdot dn(z|m)} = \frac{dt}{-i\sqrt{m+t^2}\sqrt{1+t^2}}.$$

By (3.32),

$$m \cdot sn^2(\sigma|m) = 1 - dn^2(\sigma|m) = -\frac{1}{4}\left(r - \frac{1}{r}\right)^2,$$

then

$$\sqrt{m} \cdot sn(\sigma|m) = \frac{i}{2}\left(r - \frac{1}{r}\right).$$

By (3.33), t moves along the positive real axis from 0 to $\frac{1}{2}\left(r - \frac{1}{r}\right)$, as z moves along the positive imaginary axis from 0 to σ. Then

$$\begin{aligned} z(\sigma) &= \alpha - \frac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\ &= \alpha - \frac{i}{\lambda}\left\{\int_0^\sigma dn^2(z|m)dz - m_1\sigma\right\} \\ &= \alpha - \frac{i}{\lambda}\int_0^\sigma m \cdot cn^2(z|m)dz \\ &= \alpha - \frac{i}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} (m+t^2)\frac{dt}{-i\sqrt{m+t^2}\sqrt{1+t^2}} \\ &= \alpha + \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+t^2}}{\sqrt{1+t^2}}dt. \end{aligned}$$

So,

$$\Psi(r) = \alpha + \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+t^2}}{\sqrt{1+t^2}}dt.$$

Secondly, we want to prove for any $r > 1$,

$$\Psi(-r) = -\Psi(r). \tag{3.34}$$

In (3.31), let the auxiliary parameter σ and $\tilde{\sigma}$ be such that

$$\begin{aligned} \Psi(r) &\longleftrightarrow \sigma \longleftrightarrow r \\ \Psi(-r) &\longleftrightarrow \tilde{\sigma} \longleftrightarrow -r \end{aligned}$$

Then

$$dn(\tilde{\sigma}|m) = \frac{1}{2}\left(-r + \frac{1}{-r}\right) = -\frac{1}{2}\left(r + \frac{1}{r}\right) = -dn(\sigma|m).$$

By (2.27), $\tilde{\sigma} = 2iK' - \sigma$. Then by (2.23),

$$\begin{aligned} \Psi(-r) = z(\tilde{\sigma}) = z(2iK' - \sigma) &= \alpha - \frac{i}{\lambda}\{E(2iK' - \sigma|m) - m_1(2iK' - \sigma)\} \\ &= \alpha - \frac{i}{\lambda}\{2i(K' - E') - E(\sigma|m) - 2m_1iK' + m_1\sigma\} \end{aligned}$$

$$
\begin{aligned}
&= \alpha - \frac{i}{\lambda}\{-2i(E' - mK') - [E(\sigma|m) - m_1\sigma]\} \\
&= \alpha - \frac{i}{\lambda}\{-2i \cdot \lambda\alpha - [E(\sigma|m) - m_1\sigma]\} \\
&= \alpha - 2\alpha + \frac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\
&= -\alpha + \frac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\
&= -z(\sigma) = -\Psi(r).
\end{aligned}
$$

Now we have by (3.29) that

$$
\begin{aligned}
\tilde{\Psi}(-r) &= \frac{a+b}{2} - \alpha - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+t^2}}{\sqrt{1+t^2}}dt \\
&= a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+t^2}}{\sqrt{1+t^2}}dt,
\end{aligned}
$$

noting that $\alpha = \frac{b-a}{2}$. □

3.3 *A priori* error bound

In this section, we derive a new *a priori* error bound for computing $e^{-\tau A}v$ with the Arnoldi algorithm. Throughout this section, for the non-symmetric matrix $A \in \mathbb{R}^{n\times n}$, we write

$$
\begin{cases}
a = \min\limits_i \left\{\lambda_i\left(\dfrac{A+A^T}{2}\right)\right\} = \nu(A) \\
b = \max\limits_i \left\{\lambda_i\left(\dfrac{A+A^T}{2}\right)\right\} = \mu(A) \\
c = \max\limits_i \left\{\left|\lambda_i\left(\dfrac{A-A^T}{2}\right)\right|\right\}
\end{cases}
\tag{3.35}
$$

where $\lambda_i(M), 1 \le i \le n$ are the eigenvalues of M. Then as in the proof of Proposition 2.7, we have the following lemma.

Lemma 3.6. *The field of values of A is contained in the rectangle $[a,b]\times[-c,c]$, where a, b and c are defined in* (3.35).

Proof. Note that

$$
x^*Ax = x^*\frac{A+A^*}{2}x + x^*\frac{A-A^*}{2}x,
$$

where $\frac{A+A^*}{2}$ is Hermitian and $\frac{A-A^*}{2}$ is skew-Hermitian. Then

$$\mathrm{Re}(x^*Ax) = x^*\frac{A+A^*}{2}x \in [a,b]$$
$$\mathrm{Im}(x^*Ax) = \mathrm{Im}\left(x^*\frac{A-A^*}{2}x\right) \in [-c,c].$$

□

Let $\tilde{\Phi}$ be the conformal mapping from the exterior of the rectangle $[a,b]\times[-c,c]$ to the exterior of unit disk and Ψ be its inverse mapping. The next theorem presents a bound of $e_k^T e^{-tH_k} e_1$, which is a key part in getting an *a priori* error bound. We first prove a simple lemma about numerical range, which will be needed soon.

Lemma 3.7. *Let $Q \in \mathbb{C}^{n\times k}$ be an orthogonal matrix. Then*

$$W(Q^*AQ) \subseteq W(A).$$

Proof. For any $x \in \mathbb{C}^{k\times 1}$ with $||x||_2 = 1$, $x^*Q^*AQx \in W(Q^*AQ)$. Let $y = Qx \in \mathbb{C}^{n\times 1}$, then $||y||_2 = 1$, since $Q \in \mathbb{C}^{n\times k}$ is orthogonal. Therefore, $x^*Q^*AQx = y^*Ay \in W(A)$. □

The next theorem gives a bound of $|h(t)| = |e_k^T e^{-tH_k} e_1|$, which is a key part in deducing our new *a priori* error bound. The same idea has been discussed in [6] by Benzi and Golub, and used in [42] to achieve a new *a priori* bound for symmetric matrices. For non-symmetric matrices, the techniques of the conformal mappings and the Faber polynomials have been carried out in [3] and [32]. Here we will get a sharper bound by constructing a new conformal mapping which captures more information about the eigenvalue distribution of A.

Theorem 3.8. *Let Φ be the conformal mapping defined in* (3.14). *Let $f(z) = e^{-tz}$ and H_k be a k-by-k upper Hessenberg matrix. Let $|h(t)| = e_k^T e^{-tH_k} e_1 = [f(H_k)]_{k1}$ be the $(k,1)$ entry of the matrix e^{-tH_k}. Then for any $r > 1$,*

$$|h(t)| \le 2\,Q\,M(r)\,\frac{\left(\frac{1}{r}\right)^{k-1}}{1-\frac{1}{r}}, \tag{3.36}$$

with the constant $Q = 11.08$*, and*

$$M(r) = \max_{z \in C_r} |f(z)| \tag{3.37}$$

where C_r *is the inverse image of* $\tilde{\Phi}$ *under* $|u| = r$*.*

Proof. Let $\{\Phi_j\}$ be the Faber polynomials generated by $\tilde{\Phi}$. Since f is an analytic function, it can be expanded as a series of Faber polynomials

$$f(z) = \sum_{j=0}^{\infty} \alpha_j \Phi_j(z).$$

Let

$$\Pi_N(z) = \sum_{j=0}^{N} \alpha_j \Phi_j(z)$$

be the partial sums of the above series. By Theorem 2.14,

$$||f - \Pi_N||_\infty \leq 2\, M(r)\, \frac{\left(\frac{1}{r}\right)^{N+1}}{1 - \frac{1}{r}},$$

where $M(r) = \max_{z \in C_r} |f(z)|$ and the total rotation $V = 2\pi$. Let H_k be the upper Hessenberg matrix generated in the Arnoldi process. Then $[p(H_k)]_{k1} = 0$ for all polynomials p of degree $\leq k - 2$. Then for $N \leq k - 2$,

$$\begin{aligned} |h(t)| = |[f(H_k)]_{k1}| &= |[f(H_k)]_{k1} - [\Pi_N(H_k)]_{k1}| \\ &\leq ||f(H_k) - \Pi_N(H_k)||_2 \\ &\leq Q \cdot \sup_{z \in W(H_k)} |f(z) - \Pi_N(z)|, \end{aligned}$$

where $W(H_k)$is the field of values of H_k and the constant $Q = 11.08$ by Theorem 2.8. Since $H_k = V_k^T A V_k$ for an orthogonal V_k, by Lemma 3.7,

$$W(H_k) \subseteq W(A) \subseteq F \subseteq C_r,$$

we have

$$|h(t)| \leq Q \cdot ||f - \Pi_N||_\infty.$$

Therefore,

$$|h(t)| \leq 2\, Q\, M(r)\, \frac{\left(\frac{1}{r}\right)^{k-1}}{1 - \frac{1}{r}}.$$

□

Now we can present the new *a priori* error bound for computing $e^{-\tau A}v$ with the Arnoldi method.

Theorem 3.9. ***(A priori error bound)*** *Assume that $A \in \mathbb{R}^{n\times n}$ with $\nu(A) = \lambda_{min}\left(\frac{A+A^*}{2}\right) > 0$, $v \in \mathbb{R}^n$ with $||v|| = 1$. Let $w_k(\tau) = V_k e^{-\tau H_k} e_1$ in (3.3) be the Arnoldi approximation to $w(\tau) = e^{-\tau A}v$ in (3.1). Then for any $0 < q < 1$, the approximation error satisfies*

$$||w(\tau) - w_k(\tau)|| \leq 2\,Q\,\tau\,||A||\,\frac{q^{k-1}}{1-q}\,e^{-\tau\tilde{z}}, \tag{3.38}$$

where

$$\tilde{z} = a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds, \tag{3.39}$$

the parameters m and λ are determined in (3.7).

Proof. Since $f(z) = e^{-tz}$ with $t > 0$, $|f(z)|$ has its maximum when z has the smallest real part. Let $q := \frac{1}{r}$ in (3.36). By Lemma 3.4 and Lemma 3.5, $M\left(\frac{1}{q}\right) = e^{-t\tilde{z}}$ where

$$\tilde{z} = a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds.$$

Then apply (3.36) to the *a posteriori* error bound (3.4) in Theorem 3.1. □

In observation of Theorem 3.9, the error depends on both the decay term q^{k-1} and the exponential term $e^{-\tau\tilde{z}}$. Since $e^{-\tau\tilde{z}} \leq 1$ if $\tilde{z} \geq 0$, there is a threshold convergence rate q when $\tilde{z} = 0$, i.e., the error converges at the rate q no matter how large $||\tau A||$ is. In the rest of this section, we will discuss this rate in two extreme cases. When $m \approx 0$, the matrix is close to symmetric and its eigenvalues lie close to the real axis. When $m \approx 1$, the eigenvalues lie close to a vertical line segment in the complex plane.

3.3.1. The case when A is close to symmetric ($m \approx 0$)

Recall the function $g(m)$ in (3.10). When $m \approx 0$,

$$g(m) = \frac{\beta}{\alpha} = \frac{2c}{b-a} \approx 0,$$

where the field of values of A is contained in the rectangle $[a,b]\times[-c,c]$. So the field of values lies close to the real axis and therefore the matrix A is nearly symmetric. We intend to get a threshold convergence rate at this extreme case and compare it with the error bounds in [42].

Theorem 3.10. *Under the assumptions of Theorem 3.9, and $m\approx 0$, the approximation error satisfies*

$$||w(\tau)-w_k(\tau)||\leq 2\,Q\,\tau\,||A||\frac{q_0^{k-1}}{1-q_0}$$

where

$$q_0=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}+O(\sqrt{m}),$$

and $\kappa=\frac{b}{a}$.

Proof. We want to find q such that $\tilde{z}$ in (3.39) is 0, by solving

$$a\lambda=\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)}\frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds. \tag{3.40}$$

By (2.20) and (2.21), $E'=E(1-m)$ and $K'=K(1-m)$ are both functions of m and have the following expansions at $m=0$

$$E'=E(m_1)=E(1-m)=1-\frac{1}{4}m\ln m+O(m) \tag{3.41}$$

$$K'=K(m_1)=K(1-m)=-\frac{1}{2}\ln m+O(1) \tag{3.42}$$

Then $E'-mK'$ can be expanded at $m=0$ as

$$E'-mK'=1+\frac{1}{4}m\ln m+O(m).$$

Since $\alpha=\frac{b-a}{2}$,

$$\lambda=\frac{E'-mK'}{\alpha}=\frac{2}{b-a}\left(1+\frac{1}{4}m\ln m\right)+O(m).$$

Then

$$a\lambda=\frac{2a}{b-a}\left(1+\frac{1}{4}m\ln m\right)+O(m)$$

$$= \frac{2}{\kappa - 1}\left(1 + \frac{1}{4} m \ln m\right) + O(m) \tag{3.43}$$

where $\kappa := \frac{b}{a}$. At the same time,

$$\frac{\sqrt{m + s^2}}{\sqrt{1 + s^2}} = \frac{s}{\sqrt{1 + s^2}} + O(\sqrt{m}),$$

so

$$\begin{aligned}\int_0^{\frac{1}{2}\left(\frac{1}{q} - q\right)} \frac{\sqrt{m + s^2}}{\sqrt{1 + s^2}} ds &= \int_0^{\frac{1}{2}\left(\frac{1}{q} - q\right)} \frac{s}{\sqrt{1 + s^2}} ds + O(\sqrt{m}) \\ &= \frac{1}{2}\left(\frac{1}{q} + q\right) - 1 + O(\sqrt{m}). \end{aligned} \tag{3.44}$$

Equating the two sides of (3.40) with (3.43) and (3.44), we get

$$\frac{2}{\kappa - 1} = \frac{1}{2}\left(\frac{1}{q} + q\right) - 1 + O(\sqrt{m}).$$

For m sufficiently small, the equation has two real roots and the one smaller than 1 is

$$q = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} + O(\sqrt{m}).$$

□

In an earlier paper [42] by Ye, it is shown that for a positive semi-definite matrix A, the approximation error of the Lanczos method satisfies

$$||w(\tau) - w_m(\tau)|| \leq \tau ||A|| (\sqrt{\kappa} + 1)\left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1}\right)^{m-1},$$

where κ is the condition number of the matrix A. So the convergence is fast even if the norm of τA is large, and the convergence rate $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$ is directly related to the condition number κ. Theorem 3.10 shows that our new bound for non-symmetric matrices agrees to the result in [42].

3.3.2 The case when A is close to skew-symmetric ($m \approx 1$)

Similarly, when $m \approx 1$,

$$g(m) = \frac{\beta}{\alpha} = \frac{2c}{b - a} \to +\infty,$$

where the field of values of A is contained in the rectangle $[a,b]\times[-c,c]$. So $b-a\approx 0$ and the field of values lies close to a vertical line segment on the right half of the complex plane. This will provide a good comparison with the error bound of $e^{i\tau A}v$ for a symmetric A discussed in the next chapter.

Theorem 3.11. *Under the assumptions of Theorem 3.9, and $m\approx 1$, the approximation error satisfies*

$$||w(\tau)-w_k(\tau)||\leq 2\,Q\,\tau\,||A||\frac{q_0^{k-1}}{1-q_0}$$

where

$$q_0=\frac{1}{\sqrt{\mu^2+1}+1}+O(\sqrt{1-m}).$$

and $\mu=\frac{c}{a}$.

Proof. As in the previous case, $\tilde{z}=0$ in (3.39) implies that

$$a\lambda=\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)}\frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds. \tag{3.45}$$

Since $m\approx 1$, the complementary parameter $m_1:=1-m\approx 0$. Now we expand the equation above at $m_1=0$. By (3.41) and (3.42),

$$\begin{aligned}E=E(m)=E(1-m_1)&=1-\frac{1}{4}m_1\ln m_1+O(m_1)\\ K=K(m)=K(1-m_1)&=-\frac{1}{2}\ln m_1+O(1)\end{aligned}$$

Expand $E-m_1K$ at $m_1=0$, we have

$$\begin{aligned}E-m_1K&=\left[1-\frac{1}{4}m_1\ln m_1+O(m_1)\right]-m_1\left[-\frac{1}{2}\ln m_1+O(1)\right]\\ &=1+\frac{1}{4}m_1\ln m_1+O(m_1)\end{aligned}$$

Since $\beta=c$,

$$\lambda=\frac{E-m_1K}{\beta}=\frac{1}{c}\left(1+\frac{1}{4}m_1\ln m_1\right)+O(m_1).$$

Then

$$a\lambda=\frac{E-m_1K}{\beta}=\frac{a}{c}\left(1+\frac{1}{4}m_1\ln m_1\right)+O(m_1)$$

$$= \frac{1}{\mu}\left(1 + \frac{1}{4} m_1 \ln m_1\right) + O(m_1) \tag{3.46}$$

where $\mu := \frac{c}{a}$. At the same time,

$$\frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} = \frac{\sqrt{1+s^2-m_1}}{\sqrt{1+s^2}} = 1 + O(\sqrt{m_1}).$$

So

$$\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds = \frac{1}{2}\left(\frac{1}{q} - q\right) + O(\sqrt{m_1}). \tag{3.47}$$

Equating two sides of (3.45) with (3.46) and (3.47),

$$\frac{1}{\mu} = \frac{1}{2}\left(\frac{1}{q} - q\right) + O(\sqrt{m_1}).$$

The root smaller than 1 is

$$q = \frac{1}{\sqrt{\mu^2+1}+1} + O(\sqrt{m_1}).$$

□

3.4 Optimized error bound

In the previous subsection, we analyzed the threshold convergence rates at two extreme cases. In other words, the actual convergence rate at those two extreme cases can be as good, if not better, as discussed above. In general, the threshold error rate needs not to be optimal. Unfortunately, there is no simple formula for the value of q that optimize the bound. Here we numerically find an optimal bound for convergence rate by minimizing the overall bound. To find q that minimizes the bound, we define

$$E(q) := \frac{q^{k-1}}{1-q} e^{-\tau \tilde{z}} \tag{3.48}$$

where $\tilde{z}$ is as in Theorem 3.9. Then for each k, we look for $q = q(k)$ which minimizes E. Take derivative of E with respect to q,

$$\frac{dE}{dq} = \frac{(k-1)q^{k-2}(1-q) - q^{k-1}(-1)}{(1-q)^2} e^{-\tau\tilde{z}} + \frac{q^{k-1}}{1-q} e^{-\tau\tilde{z}}(-\tau)\frac{d\tilde{z}}{dq}.$$

By (3.39), we have

$$\tilde{z} = a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds.$$

Differentiating $\tilde{z}$ with respect to q,

$$\begin{aligned}\frac{d\tilde{z}}{dq} &= -\frac{1}{\lambda}\frac{\sqrt{m+\frac{1}{4}\left(\frac{1}{q}-q\right)^2}}{\sqrt{1+\frac{1}{4}\left(\frac{1}{q}-q\right)^2}}\frac{1}{2}\left(-\frac{1}{q^2}-1\right)\\ &= \frac{\sqrt{m+\frac{1}{4}\left(\frac{1}{q}-q\right)^2}}{\lambda q}.\end{aligned}$$

Then

$$\begin{aligned}\frac{dE}{dq} &= e^{-\tau\tilde{z}}\left[\frac{(k-1)q^{k-2}(1-q)-q^{k-1}(-1)}{(1-q)^2}+\frac{q^{k-1}}{1-q}(-\tau)\frac{\sqrt{m+\frac{1}{4}\left(\frac{1}{q}-q\right)^2}}{\lambda q}\right]\\ &= e^{-\tau\tilde{z}}\frac{q^{k-3}}{(1-q)^2}\left[(k-1)q+(2-k)q^2-C(1-q)\sqrt{(1-q^2)^2+4mq^2}\right],\end{aligned}$$

where $C = \frac{\tau}{2\lambda}$. Setting $\frac{dE}{dq} = 0$ and solving for q,

$$(k-1)q+(2-k)q^2-C(1-q)\sqrt{(1-q^2)^2+4mq^2} = 0. \tag{3.49}$$

Let

$$h(q) := (k-1)q+(2-k)q^2-C(1-q)\sqrt{(1-q^2)^2+4mq^2}.$$

Then $h(0) = -C < 0$ and $h(1) = 1 > 0$. Thus there exists a $q \in (0,1)$ such that $h(q) = 0$. Since the error $E \to \infty$ as $q \to 0$, we can take q to be the smallest real root of (3.49) in $(0,1)$ and the error (3.48) will be locally minimized at q.

3.5 Numerical examples

In this section, we present several numerical examples to demonstrate the error bounds obtained in this chapter. All numerical tests were carried out on a PC with an Intel Core 2 Duo P8400 in MATLAB (R2013b) with the machine precision $\approx 2e-16$.

We will construct several testing matrices and plot the approximation error against our new *a posterior* bound (3.4), *a priori* bound (3.38). In our *a posteriori* bound, we need to compute $\max_{0\le t\le\tau}|h(t)|$ where $h(t) = e_k^T e^{-tT_k} e_1$. It can be approximated by its maximum at some densely distributed discrete points, i.e., $\max_{0\le t\le\tau}|h(t)| \approx \max\{|h(\frac{i}{N}\tau)| : 0 \le i \le N\}$ where N is some large positive integer, say 1000. We also plot the classical bound by Saad for comparison

$$||w(\tau) - w_k(\tau)|| \le \frac{2}{k!}(\tau||A||)^k. \tag{3.50}$$

The first test is on a randomly generated non-symmetric matrix. We want to show how the norm of τA affect the convergence of the Arnoldi method and the comparison of our bounds with (3.50) when τ is relatively large.

Example 1. Let A be a 1000×1000 dense non-symmetric matrix whose elements are uniformly distrubuted in $(0, 1)$. Then A is scaled such that $||A||_2 = 1$. Let v be a 1000×1 random vector with $||v||_2 = 1$. We apply 100 iterations of the Arnoldi method to A and v, to compute $w(\tau) = e^{-\tau A}v$, for various values of $\tau = 2, 5, 10, 20, 50, 100$. In Figure 3.1, we plot against the iteration number k the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* error bound (3.4) in the +-line, the *a priori* error bound (3.38) in the dashed line and Saad's classical bound (3.50) in the x-line.

From Figure 3.1 we observe that when τ is small ($\tau = 2, 5$), our new *a priori* bound and the classical bound of Saad are comparable. In this case, the convergence of the Arnoldi method is attributed to the small norm of τA. As τ increases, our new bound is proved to be much better than Saad's bound. For $\tau = 50$, our bound still follows the actual error while Saad's classical bound increases out of range of the figure. For all cases in Figure 3.1, our *a posteriori* error bound follows the actual error very closely.

In the following example, we will compare our new *a priori* bound (3.38) with the error bound obtained by Hochbruck and Lubich [22]. Theorem 5 in [22] states that if A is a matrix whose field of values is contained in the disk $|z - \rho| < \rho$ in the complex plane, then the error of $V_k e^{-\tau H_k} e_1$ for approximating $e^{-\tau A}v$ satisfies

$$||e^{-\tau A}v - V_k e^{-\tau H_k} e_1|| \le 12e^{-\rho\tau}\left(\frac{e\rho\tau}{k}\right)^k, \tag{3.51}$$

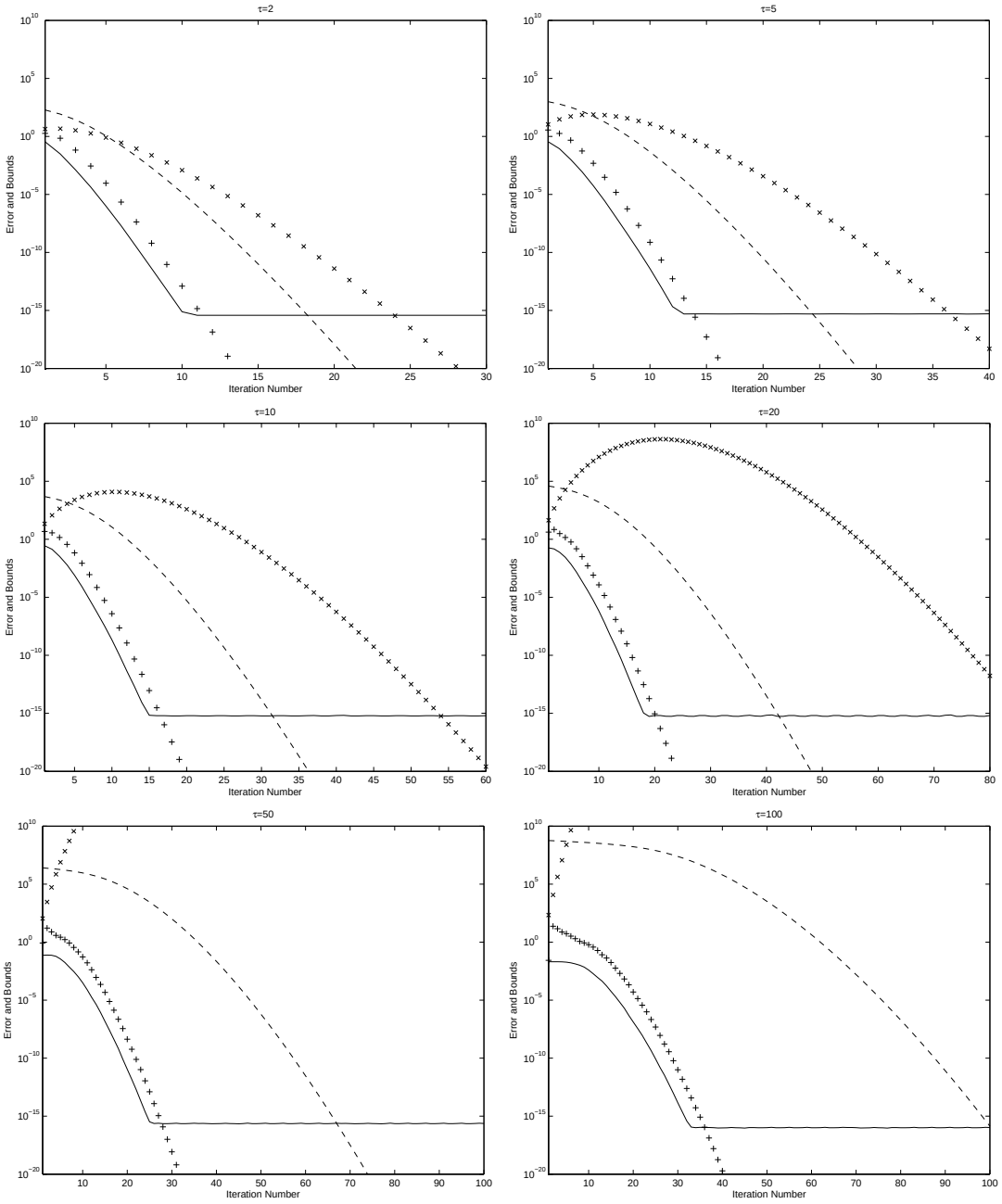

Figure 3.1: Example 1. 1000×1000 uniformly random matrix. $\tau = 2, 5, 10, 20, 50, 100$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

for $k \geq 2\rho\tau$.

Example 2. Given a rectangle $[a, b] \times [-c, c]$ in the complex plane where a, b and c are all positive real numbers. We set up an evenly distributed $N \times N$ lattice in $[a, b] \times [-c, c]$. To be precise, counting from the top left corner, the (l, j)-th node is

specified by the complex number $a + \frac{(l-1)(b-a)}{N-1} + i\left(c - \frac{2(j-1)c}{N-1}\right)$, where $1 \leq l, j \leq N$ and i is the imaginary unit. We want to construct a matrix A which has an eigenvalue at each nodes of the lattice. Since the lattice is symmetric with respect to the real axis, for each conjugate pair of eigenvalues, we have a 2×2 block

$$B = \begin{bmatrix} x & y \\ -y & x \end{bmatrix}$$

with eigenvalues $x \pm iy$ for real x and y. Then let A be $N^2 \times N^2$ block diagonal matrix with diagonal blocks like B. Then the eigenvalues of A fill the lattice in the rectangle $[a, b] \times [-c, c]$. Also note that by this construction, A is a normal matrix so the field of values of A is the convex hull of its eigenvalues, i.e., the field of values of A is also contained in the rectangle $[a, b] \times [-c, c]$.

In this numerical test, we want to compare our *a priori* bound with Hochbruch and Lubich's bound (3.51). So we take $\rho = 1$ and consider the disk $|z - 1| < 1$ containing the field of values. We choose the square $[1 - \frac{\sqrt{2}}{2}, 1 + \frac{\sqrt{2}}{2}] \times [-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$ enclosed in the circle $|z - 1| < 1$ and construct a matrix A in the ways described above such that the eigenvalues of A form a 31×31 lattice in $[1 - \frac{\sqrt{2}}{2}, 1 + \frac{\sqrt{2}}{2}] \times [-\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}]$. We use various values of $\tau = 10, 20, 30, 40$. We apply 120 Arnoldi iterations to compute $e^{-\tau A}v$ where v is a random normalized vector. In Figure 3.2, we plot against the iteration number the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* bound (3.4) in the +-line, the *a priori* bound (3.38) in the dashed line, Saad's bound (3.50) in the x-line and Hochbruch and Lubich's bound (3.51) in the dash-dotted line.

We observe from Figure 3.2 that when τ is relatively small ($\tau = 10, 20$), our new *a priori* bound is comparable to the bound by Hochbruch and Lubich. As τ increases, our new *a priori* improves Hochbruch and Lubich's bound by several orders of magnitude. For the case when $\tau = 40$, the actual error $||w(\tau) - w_k(\tau)||$ first stagnates for certain iterations before it starts to converge and our *a priori* bound captures the same behavior while Hochbruch and Lubich's bound is pessimistic.

In the previous example, we constructed the matrix A such that the field of values is contained in a square rectangle. In our discussion in this chapter, we have shown

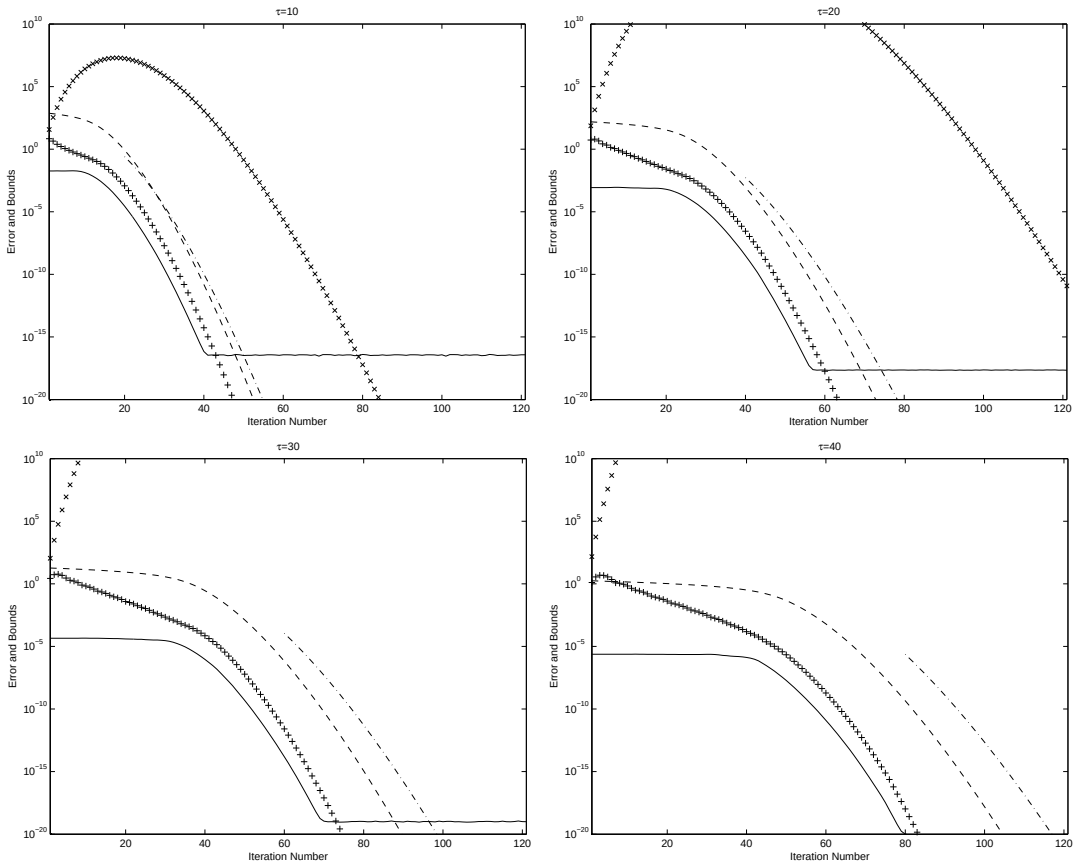

Figure 3.2: Example 2. Field of values in $|z-1| < 1$. $\tau = 10, 20, 30, 40$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x), Hochbruck and Lubich's bound (dash-dotted).

that the convergence rate depends on the shape of the rectangle, and the shape is determined by the parameter m in (3.7).

Example 3. In this example, we will manipulate the shape of the rectangle by taking different m in (3.7) and check how the actual convergence is related to the spectral information. For a given parameter $m \in (0, 1)$, we determine the dimensions of the rectangle α and β by

$$\alpha = E' - mK', \ \beta = E - m_1 K.$$

Then using the same technique we introduced in Example 2, we can construct a matrix whose field of values is contained in the rectangle $[0, 2\alpha] \times [-\beta, \beta]$. We will use various parameters $m \in \{0.01, 0.1, 0.9, 0.99\}$. Note from Section 3.3 that $m \approx 0$ means the matrix is close to symmetric, and that $m \approx 1$ means the matrix is close

to a skew-Hermitian matrix with a real spectral shift. We pick $\tau = 30$ to give τA a moderate norm. In Figure 3.3 we plot the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* bound (3.4) in +-line, the *a priori* bound (3.38) in the dashed line and the bound by Saad (3.50) in the x-line.

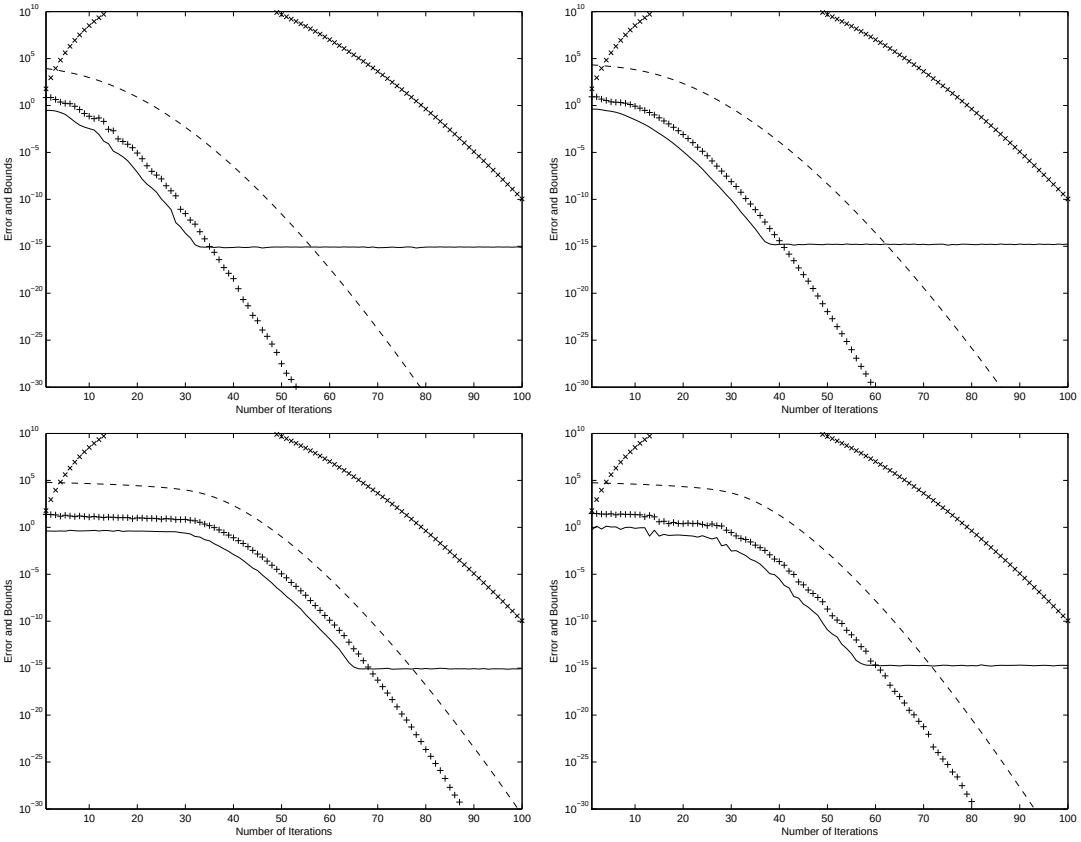

Figure 3.3: Example 3. Top two plots: $m = 0.01, 0.1$ where A is close to symmetric. Bottom two plots: $m = 0.9, 0.99$ where A is close to shifted skew-symmetric. Error(solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

Figure 3.3 shows that the convergence of the error $||w(\tau) - w_k(\tau)||$ is related to m, i.e., the eigenvalue distribution of A. The top two plots show that for a smaller m when the eigenvalues lie close to the real axis, the convergence is faster. The bottom plots show that when the eigenvalues of A have a large imaginary part, the error will not converge in the first few iterations. Compared to the classical bound by Saad, our new bound describes this behavior in a much better way.

In our final example, we consider the finite-difference discretization of the convec-

tion diffusion operator

$$-\triangle u + u_x + u_y = \lambda u, \tag{3.52}$$

where $(x, y) \in [0, 1]^2$.

Example 4. Let A be the finite-difference discretization of (3.52) in a 20×20 grid in $(0, 1)^2$. Then $||A||_2 \approx 8$. Let v be a random vector with $||v||_2 = 1$ and we compute the matrix exponential $w(\tau) = e^{-\tau A}v$. We use various values of $\tau = 2, 5, 10, 20, 50, 100$ and apply 100 Arnoldi iterations to A and v and the results are presented in Figure 3.4 with the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* bound (3.4) in the +-line, the *a priori* bound (3.38) in the dashed line and Saad's bound (3.50) in the x-line.

We observe that for $\tau = 2$, our *a priori* bound is already a significant improvement on the classical bound by Saad. For even larger values of τ, Saad's bound is very pessimistic due to the large norm of τA, while our *a priori* bound still follows the actual error. For the case when $\tau = 100$, i.e. $\tau||A||_2 \approx 800$, our *a priori* bound converges slowly. Again in all the cases, our a posteriori bound remains sharp.

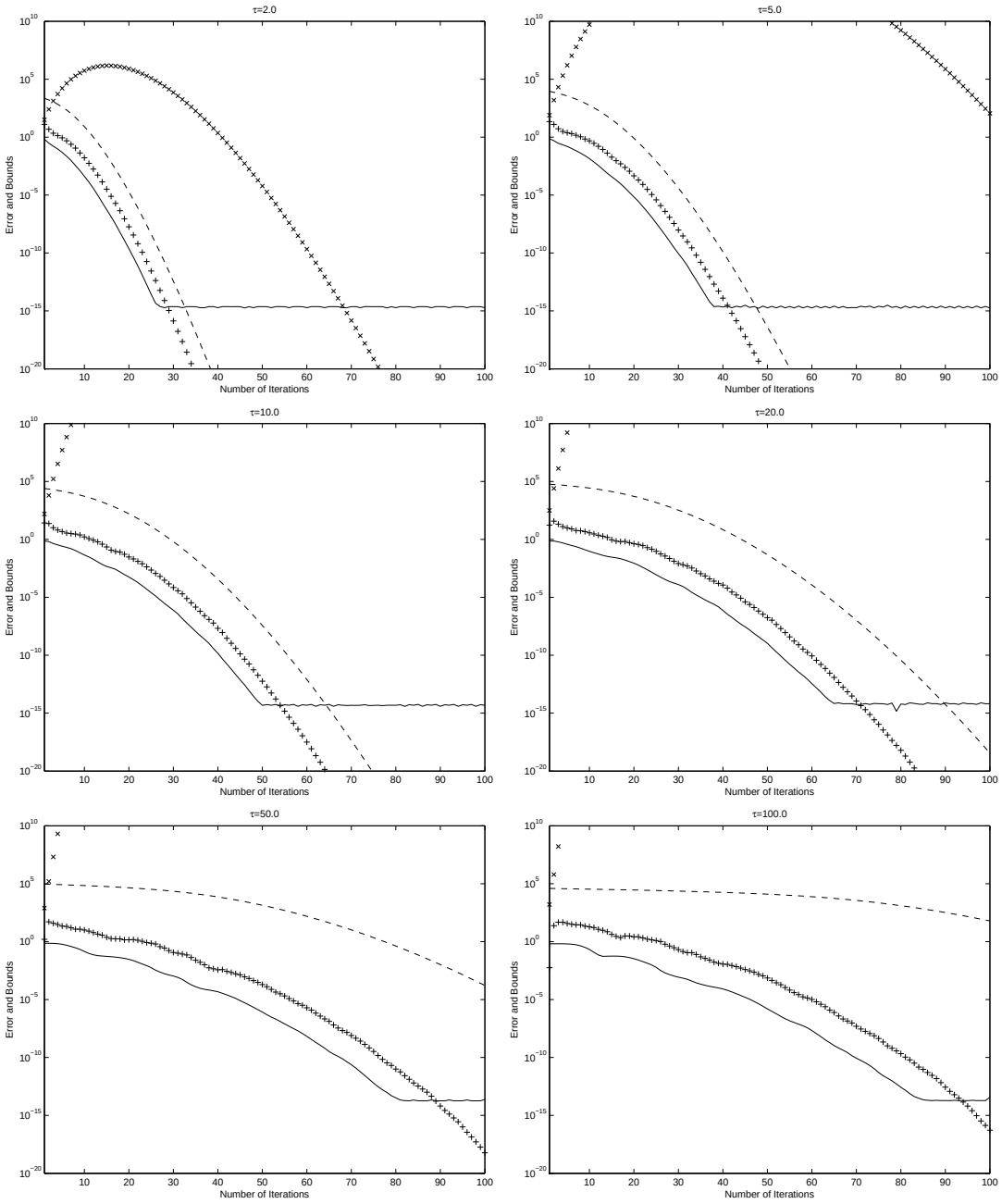

Figure 3.4: Example 4. $\tau = 2, 5, 10, 20, 50, 100$. Error(solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

Chapter 4 Error bounds for computing $e^{i\tau A}v$

In this chapter, we will discuss the error bounds for computing $w(\tau) = e^{i\tau A}v$ with the Lanczos method, where A is a real symmetric matrix, v is a real normalized vector, τ is a real scalar and i is the imaginary unit. It can be treated as a special case of the computation of $e^{-\tau A}v$ for a non-Hermitian positive semi-definite A. One application of this is in the time-dependent Schrödinger equation (TDSE) for the N-electron wavefunction $\Psi(r_1, \cdots, r_N)$ which satisfies

$$i\frac{\partial}{\partial t}\Psi(r_1, \cdots, r_N; t) = [H_0(r_1, \cdots, r_N) + V(r_1, \cdots, r_N; t)]\,\Psi(r_1, \cdots, r_N; t)$$

where $H_0(r_1, \cdots, r_N)$ is the field-free Hamiltonian containing the kinetic energy of the N electrons, and $V(r_1, \cdots, r_N; t)$ represents the interaction of the electrons with the eletromagnetic field.

We will take the same path as in the previous chapter for the computation of $e^{-\tau A}v$. The approximation error is firstly related to one entry of the exponential of a tridiagonal matrix, from which we can get an *a posteriori* error bound. The decay property of that entry is then fully studied to achieve an *a priori* error bound. This bound will be numerically optimized in the consequent section to describe the behavior of the actual convergence of the Lanczos method. Numerical examples are presented at the end of the chapter.

4.1 *A posteriori* error bound

Assume that A is an n-by-n real symmetric matrix, v is an n-dimensional real normalized vector and $\tau > 0$ is a scalar. We consider the computation of

$$w(\tau) := e^{i\tau A}v. \tag{4.1}$$

The following Lanczos method is applied to A and v.

Algorithm 4.1. ***(Lanczos Algorithm)***

1. *Initialize:*

   - $v_1 \leftarrow v$
   - $v_0 \leftarrow 0$
   - $\beta_1 \leftarrow 0$

2. *Iterate: for* $j = 1, 2, \cdots, k-1$

   a) $w_j \leftarrow Av_j$

   b) $\alpha_j \leftarrow w_j \cdot v_j$

   c) $w_j \leftarrow w_j - \alpha_j v_j - \beta_j v_{j-1}$

   d) $\beta_{j+1} \leftarrow ||w_j||$

   e) $v_{j+1} \leftarrow w_j/\beta_{j+1}$

   end for

3. $w_k = Av_k$

4. $\alpha_k = w_k \cdot v_k$

return

After k iterations, the Krylov subspace

$$K_{k+1}(A, v) = span\{v, Av, A^2v, \cdots, A^kv\}$$

is generated with an orthonormal basis $\{v_1, v_2, \cdots, v_k, v_{k+1}\}$. Let $V_k = [v_1, v_2, \cdots, v_k]$ be the n-by-k orthogonal matrix whose columns form the orthonormal basis of the Krylov subspace K_k, then there exists an k-by-k tridiagonal matrix T_k such that

$$AV_k = V_kT_k + \beta_{k+1}v_{k+1}e_k^T, \tag{4.2}$$

where $e_k \in \mathbb{R}^n$ is the k-th coordinate vector. We can use $V_k V_k^T e^{i\tau A} v$ as the best approximation to $e^{i\tau A} v$ from the Krylov subspace $K_k(A, v)$, since it is the orthogonal projection of $e^{i\tau A} v$ on $K_k(A, v)$. Applying the orthogonality of V_k to (4.2), we have

$$V_k^T A V_k = V_k^T V_k T_k + \beta_{k+1} V_k^T v_{k+1} e_k^T = T_k,$$

then

$$V_k V_k^T e^{i\tau A} v = V_k V_k^T e^{i\tau A} V_k e_1 \approx V_k e^{i\tau V_k^T A V_k} e_1 = V_k e^{i\tau T_k} e_1.$$

We now call

$$w_k(\tau) := V_k e^{i\tau T_k} e_1 \tag{4.3}$$

the Lanczos approximation to $w(\tau)$ in (4.1). The following theorem relates the Lanczos approximation error to the $(k, 1)$ entry of the matrix e^{itT_k} and therefore serves as an *a posteriori* error bound. It is the first main result of this chapter.

Theorem 4.2. ***(A posteriori error bound)*** *Assume that A is an n-by-n real symmetric matrix, v is a n-dimensional real vector with $||v|| = 1$. The orthogonal matrix $V_k \in \mathbb{R}^{n\times k}$ and tridiagonal matrix $T_k \in \mathbb{R}^{k\times k}$ are generated in the Lanczos process satisfying* (4.2)*. Let $w_k(\tau) = V_k e^{i\tau T_k} e_1$ in* (4.3) *be the Lanczos approximation to $w(\tau) = e^{i\tau A} v$ in* (4.1)*. Then the approximation error satisfies*

$$\begin{aligned} ||w(\tau) - w_k(\tau)|| &\leq \beta_{k+1} \int_0^\tau |h(t)| dt \\ &\leq \tau \beta_{k+1} \max_{0\leq t\leq \tau} |h(t)|, \end{aligned} \tag{4.4}$$

where

$$h(t) := e_k^T e^{itT_k} e_1 \tag{4.5}$$

is defined as the $(k, 1)$ entry of the matrix e^{itT_k}.

Proof. First, $w(t) = e^{itA} v$, then $w'(t) = iAe^{itA} v = iAw(t)$. Since $w_k(t) = V_k e^{itT_k} e_1$, we have

$$w_k'(t) = iV_k T_k e^{itT_k} e_1$$

$$
\begin{aligned}
&= i(AV_k - \beta_{k+1} v_{k+1} e_k^T) e^{itT_k} e_1 \\
&= iAV_k e^{itT_k} e_1 - i\beta_{k+1} v_{k+1} e_k^T e^{itT_k} e_1 \\
&= iAw_k(t) - i\beta_{k+1} h(t) v_{k+1},
\end{aligned}
$$

where $h(t) = e_k^T e^{itT_k} e_1$. Let $E_k(t) := w(t) - w_k(t)$ be the approximation error, then

$$
\begin{aligned}
E_k'(t) &= w'(t) - w_k'(t) \\
&= iAw(t) - (iAw_k(t) - i\beta_{k+1} h(t) v_{k+1}) \\
&= iAE_k(t) + i\beta_{k+1} h(t) v_{k+1}.
\end{aligned}
$$

Now we solve the ordinary differential equation with the initial condition

$$
E_k(0) = w(0) - w_k(0) = v - V_k e_1 = 0,
$$

then

$$
E_k(\tau) = i\beta_{k+1} \int_0^\tau h(t) e^{i(\tau - t)A} v_{k+1} dt.
$$

Since $||e^{iA}|| = 1$ for any real matrix A, we have the *a posteriori* error bound

$$
\begin{aligned}
||E_k(\tau)|| &\le \beta_{k+1} \left|\left| \int_0^\tau h(t) e^{i(\tau - t)A} v_{k+1} dt \right|\right| \\
&\le \beta_{k+1} \int_0^\tau |h(t)| \cdot ||e^{i(\tau - t)A}|| \cdot ||v_{k+1}|| dt \\
&= \beta_{k+1} \int_0^\tau |h(t)| dt \\
&\le \tau \beta_{k+1} \max_{0 \le t \le \tau} |h(t)|.
\end{aligned}
$$

□

To get an *a priori* error bound from Theorem 4.2, we need a bound of $|h(t)|$ in (4.5). It is based on the decay properties of functions of banded matrices. To be precise, we will consider the analytic function $f(z) = e^{tz}$ and $B = iT_k$, then $h(t) = [f(B)]_{k1}$. In the next section, we will use the Faber polynomials discussed in Section 2.6 to approximate the function $f(z)$.

4.2 *A priori* error bound

For the real symmetric matrix A, all of its eigenvalues are real. Let a and b be the smallest and the largest eigenvalues of A, respectively. Then $\sigma(T_k)$, the spectrum of T_k, is contained in the interval $[a, b]$ on the real axis, where T_k is obtained in (4.2). Write $B = iT_k$, then the spectrum $\sigma(B) \subseteq \{z = i\lambda : \lambda \in [a, b]\}$.

Let the set $E := \{z = i\lambda : \lambda \in [a, b]\}$. We are looking for a conformal mapping Φ which maps the exterior of E to the exterior of $|w| = \rho$ for some ρ, satisfying the normalization condition (2.9). This can be done in the following four successive steps.

- Step 1:

$$z_1 = \phi_1(z) = -iz \tag{4.6}$$

  maps the exterior of E to the exterior of $[a, b]$.

- Step 2:

$$z_2 = \phi_2(z_1) = \frac{2}{b-a}\left(z_1 - \frac{a+b}{2}\right) \tag{4.7}$$

  maps the exterior of $[a, b]$ to the exterior of $[-1, 1]$.

- Step 3:

$$z_3 = \phi_3(z_2) = z_2 + \sqrt{z_2^2 - 1} \tag{4.8}$$

  maps the exterior of $[-1, 1]$ to $\{|z_3| > 1\}$. Note that we choose the branch of $\sqrt{z^2-1}$ such that $\lim\limits_{z\mapsto\infty} \frac{\sqrt{z^2-1}}{z} = 1$.

- Step 4:

$$w = \phi_4(z_3) = \frac{i(b-a)}{4} z_3 \tag{4.9}$$

  maps $\{|z_3| > 1\}$ to $\{|w| > \frac{b-a}{4}\}$.

Let

$$\Phi := \phi_4 \circ \phi_3 \circ \phi_2 \circ \phi_1 \tag{4.10}$$

be the composition of the above four mappings, where ϕ_1, ϕ_2, ϕ_3 and ϕ_4 are defined in (4.6), (4.7), (4.8) and (4.9), respectively. Then $\Phi : z \mapsto w$ will map the exterior of E to $\{|w| > \frac{b-a}{4}\}$ conformally. We further verify that Φ satisfies the normalization condition (2.9) that

$$\Phi(\infty) = \infty, \ \lim_{z\to\infty} \frac{\Phi(z)}{z} = 1.$$

Therefore Φ is eligible for the construction of the Faber polynomials and the logarithmic capacity

$$\rho = \frac{b-a}{4}. \tag{4.11}$$

The following construction process of the Faber polynomials is similar to that in Example 2.13. First,

$$[\Phi(z)]^k = (i\rho z_3)^k = (i\rho)^k \left(z_2 + \sqrt{z_2^2 - 1}\right)^k.$$

Same as in Example 2.13, since the Laurent expansion at ∞ of

$$\left(z_2 - \sqrt{z_2^2 - 1}\right)^k$$

contains no non-negative powers of z_2,

$$\left(z_2 + \sqrt{z_2^2 - 1}\right)^k$$

has the same non-negative powers as

$$\left(z_2 + \sqrt{z_2^2 - 1}\right)^k + \left(z_2 - \sqrt{z_2^2 - 1}\right)^k.$$

Furthermore,

$$z_2 = \frac{2}{b-a}\left(z_1 - \frac{a+b}{2}\right) = \frac{2}{b-a}\left(-iz - \frac{a+b}{2}\right),$$

is a linear function of z, so the non-negative powers of z can only be produced from those of z_2. Then the Faber polynomials in z, i.e., the non-negative powers of the expansion of $[\Phi(z)]^k$ are

$$\Phi_k(z) = (i\rho)^k \left[\left(z_2 + \sqrt{z_2^2 - 1}\right)^k + \left(z_2 - \sqrt{z_2^2 - 1}\right)^k\right].$$

Since $z_2 \in [-1, 1]$, let $z_2 = \cos(\theta)$ where $\theta \in [0, 2\pi)$. Then

$$\Phi_k(z) = (i\rho)^k \left[(\cos(\theta) + i\sin(\theta))^k + (\cos(\theta) - i\sin(\theta))^k\right] = 2(i\rho)^k \cos(k\theta).$$

So the norm of the Faber polynomials

$$||\Phi_k||_\infty = 2\rho^k, \tag{4.12}$$

where the logarithmic capacity ρ is already determined in (4.11).

Now we can present the new *a priori* error bound for computing $w(\tau) = e^{i\tau A}v$ with the Lanczos method.

Theorem 4.3. ***(A priori error bound)*** *Assume that A is an n-by-n real symmetric matrix, v is a n-dimensional real vector with $||v|| = 1$. The orthogonal matrix $V_k \in \mathbb{R}^{n\times k}$ and tridiagonal matrix $T_k \in \mathbb{R}^{k\times k}$ are generated in the Lanczos process satisfying (4.2). Let $w_k(\tau) = V_k e^{i\tau T_k} e_1$ in (4.3) be the Lanczos approximation to $w(\tau) = e^{i\tau A}v$ in (4.1). Then, for any $0 < q < 1$, the approximation error satisfies*

$$||w(\tau) - w_k(\tau)|| \leq \frac{8b}{b-a} \frac{q^k}{(1-q)(1-q^2)} e^{\frac{\tau(b-a)}{4}\left(\frac{1}{q} - q\right)}, \tag{4.13}$$

where a and b are the smallest and the largest eigenvalues of A, respectively.

Proof. First recall (4.4) and (4.5) in Theorem 4.2 that

$$||w(\tau) - w_k(\tau)|| \leq \beta_{k+1} \int_0^\tau |h(t)|dt,$$

where

$$h(t) = e_k^T e^{itT_k} e_1.$$

Now consider the analytic function $f(z) = e^{tz}$ and $B = iT_k$, then $h(t) = [f(B)]_{k1}$. Let $E = \{z = i\lambda : \lambda \in [a, b]\}$ be the compact set containing the spectrum $\sigma(B)$ and Φ be defined as in (4.10). By (2.10) and (2.11), the Faber polynomials $\{\Phi_j\}$ are generated and f can be expanded as

$$f(z) = \sum_{j=0}^{\infty} \alpha_j \Phi_j(z)$$

with the partial sum

$$\Pi_N(z) = \sum_{j=0}^{N} \alpha_j \Phi_j(z).$$

By (2.12), the coefficients α_j satisfies

$$|\alpha_j| \leq \frac{M(R)}{R^j},$$

where

$$M(R) = \max_{z \in C_R} |f(z)|$$

and C_R is the inverse image of Φ under $|w| = R$. Since $f(z) = e^{tz}$ is an analytic function, for any $R > \rho$, the polynomial approximation error satisfies

$$\begin{aligned}
|f(z) - \Pi_N(z)| &= \left| \sum_{j=N+1}^{\infty} \alpha_j \Phi_j(z) \right| \\
&\leq \sum_{j=N+1}^{\infty} |\alpha_j| \cdot ||\Phi_j||_\infty \\
&\leq \sum_{j=N+1}^{\infty} \frac{M(R)}{R^j} \cdot 2\rho^j \\
&= 2M(R) \sum_{j=N+1}^{\infty} \left(\frac{\rho}{R}\right)^j \\
&= 2M(R) \frac{\left(\frac{\rho}{R}\right)^{N+1}}{1 - \frac{\rho}{R}}.
\end{aligned}$$

It is observed that $B = iT_k$ is an $k \times k$ tridiagonal matrix, so $[\Pi_N(B)]_{k1} = 0$ for $N \leq k - 2$. Then the decay entry

$$\begin{aligned}
|[f(B)]_{k1}| &= |[f(B)]_{k1} - [\Pi_N(B)]_{k1}| \\
&\leq ||f(B) - \Pi_N(B)||_2 \\
&= \max_{z \in \sigma(B)} |f(z) - \Pi_N(z)| \\
&\leq ||f - \Pi_N||_\infty \\
&\leq 2M(R) \frac{\left(\frac{\rho}{R}\right)^{N+1}}{1 - \frac{\rho}{R}}
\end{aligned}$$

$$\leq 2M(R)\frac{\left(\frac{\rho}{R}\right)^{k-1}}{1-\frac{\rho}{R}}.$$

Let the convergence rate $q := \frac{\rho}{R} < 1$ with $\rho = \frac{b-a}{4}$ in (4.11). So,

$$|h(t)| = |[f(B)]_{k1}| \leq 2M(R)\frac{q^{k-1}}{1-q} \tag{4.14}$$

Our next objective is to find the explicit form of z for $|w| = R$. Let $w = Re^{i\theta}$ where $\theta \in [0, 2\pi)$, then

$$\begin{aligned}
z_3 &= \frac{w}{i\rho} = \frac{Re^{i\theta}}{i\rho} = \frac{e^{i\theta}}{iq}, \\
z_2 &= \frac{1}{2}\left(z_3 + \frac{1}{z_3}\right) = \frac{1}{2}\left(\frac{e^{i\theta}}{iq} + \frac{iq}{e^{i\theta}}\right) = -\frac{i}{2}\left[\left(\frac{1}{q} - q\right)\cos\theta + i\left(\frac{1}{q} + q\right)\sin\theta\right], \\
z_1 &= \frac{b-a}{2}z_2 + \frac{b+a}{2} = \left[\frac{b-a}{4}\left(\frac{1}{q} + q\right)\sin\theta + \frac{b+a}{2}\right] - i\left[\frac{b-a}{4}\left(\frac{1}{q} - q\right)\cos\theta\right], \\
z &= iz_1 = \frac{b-a}{4}\left(\frac{1}{q} - q\right)\cos\theta + i\left[\frac{b-a}{4}\left(\frac{1}{q} + q\right)\sin\theta + \frac{b+a}{2}\right].
\end{aligned}$$

Moreover, for $f(z) = e^{tz}$ where $z = u + iv$ with $u, v \in \mathbb{R}$, we notice that

$$M(R) = \max_{z \in C_R} |e^{tz}| = \max_{z \in C_R} |e^{tu}|.$$

In order to get the maximum of the real part of z, we take $\theta = 0$. Therefore

$$u = \frac{b-a}{4}\left(\frac{1}{q} - q\right)$$

and immediately

$$M(R) = e^{\frac{t(b-a)}{4}\left(\frac{1}{q} - q\right)}. \tag{4.15}$$

Applying (4.15) to (4.14), we have

$$|h(t)| \leq \frac{2q^{m-1}}{1-q}e^{\frac{t(b-a)}{4}\left(\frac{1}{q} - q\right)}. \tag{4.16}$$

By noting that $\beta_{m+1} \leq ||A|| = b$ and applying (4.16) in the *a posteriori* error bound (4.4) in Theorem 4.2, we have the following *a priori* error bound

$$\begin{aligned}
||w(\tau) - w_k(\tau)|| &\leq b\int_0^\tau \frac{2q^{k-1}}{1-q}e^{\frac{t(b-a)}{4}\left(\frac{1}{q} - q\right)}dt \\
&= \frac{8b}{b-a}\frac{q^k}{(1-q)(1-q^2)}\left(e^{\frac{\tau(b-a)}{4}\left(\frac{1}{q} - q\right)} - 1\right) \\
&\leq \frac{8b}{b-a}\frac{q^k}{(1-q)(1-q^2)}e^{\frac{\tau(b-a)}{4}\left(\frac{1}{q} - q\right)}.
\end{aligned}$$

□

4.3 Optimized error bound

In observation of the *a priori* error bound (4.13) in Theorem 4.3, for a fixed τ and A, the magnitude of the approximation error depends on two contributing factors: the decay term q^k and the exponential term $e^{\frac{\tau(b-a)}{4}\left(\frac{1}{q}-q\right)}$. In this section, we will optimize the *a priori* error by treating it as a function of the convergence rate q. Before stating the result, we first give a lemma which will later serve as a part of the proof.

Lemma 4.4. *Given any $C > 0$ and a polynomial of q*

$$f(q) = Cq^4 + (3-k)q^3 + q^2 + kq - C, \tag{4.17}$$

there exists a unique real $q_0 \in (0,1)$ such that $f(q_0) = 0$. Furthermore, $f(q) < 0$ in the interval $(0, q_0)$, and $f(q) > 0$ in the interval $(q_0, 1)$.

Proof. Note that

$$f(0) = -C < 0$$
$$f(1) = C + 3 - k + 1 + k - C = 4 > 0.$$

By the continuity and the degree of the polynomial f, the graph of f intersects $(0,1)$ either once or three times. Now assume that there are three real roots in $(0,1)$. Since all the coefficients of f are real, the fourth root of f is also real. Note again that $f(0) < 0$ and $\lim_{q \mapsto -\infty} f(q) = -\infty$. Then all four roots of f are positive. However, the product of all roots of f is equal to $\frac{-C}{C} = -1$. This contradiction implies that f has exactly one real root q_0 within $(0,1)$. Again by the fact that $f(0) < 0$ and $f(1) > 0$, the continuity of f implies that $f(q) < 0$ in $(0, q_0)$ and $f(q) > 0$ in $(q_0, 1)$. □

Now we can state the main result of this section.

Theorem 4.5. ***(Optimized error bound)*** *Under the assumptions of Theorem 4.3, we have*

$$||w(\tau) - w_k(\tau)|| \leq \frac{8b}{b-a} \frac{q_0^k}{(1-q_0)(1-q_0^2)} e^{C\left(\frac{1}{q_0}-q_0\right)}, \tag{4.18}$$

where $q_0 = q_0(k)$ is the root of (4.17), with $C := \frac{\tau(b-a)}{4}$.

Proof. Considering the *a priori* error bound (4.13), we let

$$E(q) = \frac{q^k}{(1-q)(1-q^2)} e^{C\left(\frac{1}{q}-q\right)}, \tag{4.19}$$

where $C = \frac{\tau(b-a)}{4}$. Then in each step of the Lanczos process with different k, it is possible to find $q_0 = q_0(k)$ which minimizes $E(q)$. Take derivative of $E(q)$ with respect to q,

$$\begin{aligned}\frac{dE}{dq} &= \frac{q^{k-1}}{(1-q)^3(1+q)^2}\left[(3-k)q^2+q+k\right]e^{C\left(\frac{1}{q}-q\right)} - \frac{q^k}{(1-q)^2(1+q)}\frac{C(1+q^2)}{q^2}e^{C\left(\frac{1}{q}-q\right)} \\ &= e^{C\left(\frac{1}{q}-q\right)}\frac{q^{k-2}}{(1-q)^3(1+q)^2}\left[Cq^4+(3-k)q^3+q^2+kq-C\right].\end{aligned}$$

Setting $\frac{dE}{dq} = 0$ and solving for q, we have

$$Cq^4 + (3-k)q^3 + q^2 + kq - C = 0.$$

By Lemma 4.4, there exists a unique solution q_0 of the equation (4.17) and it is where the overall error $E(q)$ attains its minimum over the interval $(0,1)$. □

In observation of the polynomial in Theorem 4.5, the optimized convergence rate q_0, treated as a function of the iteration number k, will be close to 1 for small values of k. This implies that the convergence will be slow at the first steps of the Lanczos iterations, and the actual convergence does not begin until q_0 starts to take a fairly small value. So it is illustrative to figure out approximately at which step of the Lanczos process does the convergence actually begin. To achieve that, we use an adjusted version of (4.19) as

$$E_s(q) = q^k e^{C\left(\frac{1}{q}-q\right)}. \tag{4.20}$$

Here $E_s(q)$ is constructed to behave the same as $E(q)$ does when q is away from 1, and simple enough to work on. Since our sole interest here is to find out when $E(q)$ takes its minimum at a fairly small q, the simpler version $E_s(q)$ will serve this purpose well. Differentiate E_s with respect to q,

$$\frac{dE_s}{dq} = kq^{k-1}e^{C\left(\frac{1}{q}-q\right)} + q^k e^{C\left(\frac{1}{q}-q\right)}C\left(-\frac{1}{q^2}-1\right)$$

$$= e^{C\left(\frac{1}{q}-q\right)} q^{k-2} \left[-Cq^2 + kq - C\right],$$

and the discriminant of the quadratic $-Cq^2 + kq - C$ is $\Delta = k^2 - 4C^2$. So only for large values of k, or to be precise when $k >> 2C$, the equation $-Cq^2 + kq - C = 0$ has a root $q = \frac{k-\sqrt{k^2-4C^2}}{2C}$ away from 1. So the actual convergence process does not begin until approximately at step $k = 2C$. We mark this observation here and it will be verified by numerical tests in the next section.

Corollary 4.6. *The actual convergence of the Lanczos process for the computation of $w(\tau) = e^{i\tau A}v$ starts approximately at the iteration number $k = 2C$, where $C = \frac{\tau(b-a)}{4}$.*

A similar result was presented in an earlier paper [22, Theorem 4]. If A is a skew-Hermitian matrix with its eigenvalues in an interval on the imaginary axis of length 4ρ, then the error of the Arnoldi approximation of $e^{\tau A}v$ is bounded by

$$\epsilon_k \leq 12e^{\frac{-(\rho\tau)^2}{k}} \left(\frac{e\rho\tau}{k}\right)^k, \ k \geq 2\rho\tau. \tag{4.21}$$

Now we compare the bounds (4.18) and (4.21). Since $C = \frac{\tau(b-a)}{4} = \rho\tau$, the bound (4.21) is equivalent to

$$\epsilon_k \leq 12e^{-\frac{C^2}{k}} \left(\frac{eC}{k}\right)^k, \ k \geq 2C.$$

Let $q = \frac{C}{k}$ in our new *a priori* error bound (4.13). With the constraint $k \geq 2C$, we have $q \leq \frac{1}{2}$. So

$$\begin{aligned} ||w(\tau) - w_k(\tau)|| &\leq \frac{4\left(\frac{C}{k}\right)^k}{(1-\frac{1}{2})(1-\frac{1}{2})^2} e^{C\left(\frac{k}{C}-\frac{C}{k}\right)} \\ &= \frac{32}{3} e^{-\frac{C^2}{k}} \left(\frac{eC}{k}\right)^k \\ &\leq 12e^{-\frac{C^2}{k}} \left(\frac{eC}{k}\right)^k. \end{aligned}$$

So the *a priori* error bound (4.13) presented in Theorem 4.3 for one particular $q = \frac{C}{k}$ is sharper than the bound (4.21). With the optimization in Theorem 4.5, the optimized error bound (4.18) is expected to be even better than the bound (4.21).

4.4 Numerical examples

In this section, we present several numerical examples to demonstrate the error bounds obtained throughout this chapter. All numerical tests were carried out on a PC with an Intel Core 2 Duo P8400 in MATLAB (R2013b) with the machine precision $\approx 2e-16$.

We will construct diagonal and random matrices in our numerical tests and compare the approximation error $||w(\tau)-w_k(\tau)||$ with our new *a posteriori* error bound (4.4) and *a priori* error bound (4.18). In our *a posteriori* error bound, we need to compute $\max_{0\le t\le\tau}|h(t)|$ where $h(t)=e_k^T e^{itT_k}e_1$. It can be approximated by its maximum at some densely distributed discrete points, i.e., $\max_{0\le t\le\tau}|h(t)|\approx\max\{|h(\frac{i}{N}\tau)|:0\le i\le N\}$ where N is some large positive integer, say 1000. We will also plot the classical bound by Saad for comparison

$$||w(\tau)-w_k(\tau)||\le\frac{2}{k!}\left(\tau||A||\right)^k. \tag{4.22}$$

In the first example, we construct a diagonal matrix A such that the eigenvalues are evenly distributed in the interval $[0,1]$. We want to illustrate the influence of the spectral gap (difference between the smallest and the largest eigenvalues) on the convergence of the Lanczos method.

Example 1. Let A be an $n\times n$ diagonal matrix whose j-th diagonal entry is j/n. Let v be a random $n\times 1$ normalized vector. Then $|||A|||_2=1$ and the spectral gap $\lambda_{\max}(A)-\lambda_{\min}(A)$ is approximately 1. We apply k iterations of the Lanczos method to compute $w(\tau)=e^{i\tau A}v$. We will test various values of τ and compare the approximation error $||w(\tau)-w_k(\tau)||$ with our new bounds as well as the classical bound of Saad (4.22). We plot them against the iteration number k with the error in the solid line, the *a posteriori* error bound (4.4) in the +-line, the optimized *a priori* error bound (4.18) in the dashed line and Saad's bound (4.22) in the x-line.

In this test, we take the size $n=1000$ and the iteration number $k=100$. We present the results for $\tau=2,5,10,20,50,100$ in Figure 4.1. We observe that when τ is relatively small ($\tau=2$), the classical bound of Saad and our *a priori* bound are comparable. When $\tau=10$, our bound is already much better than the classical

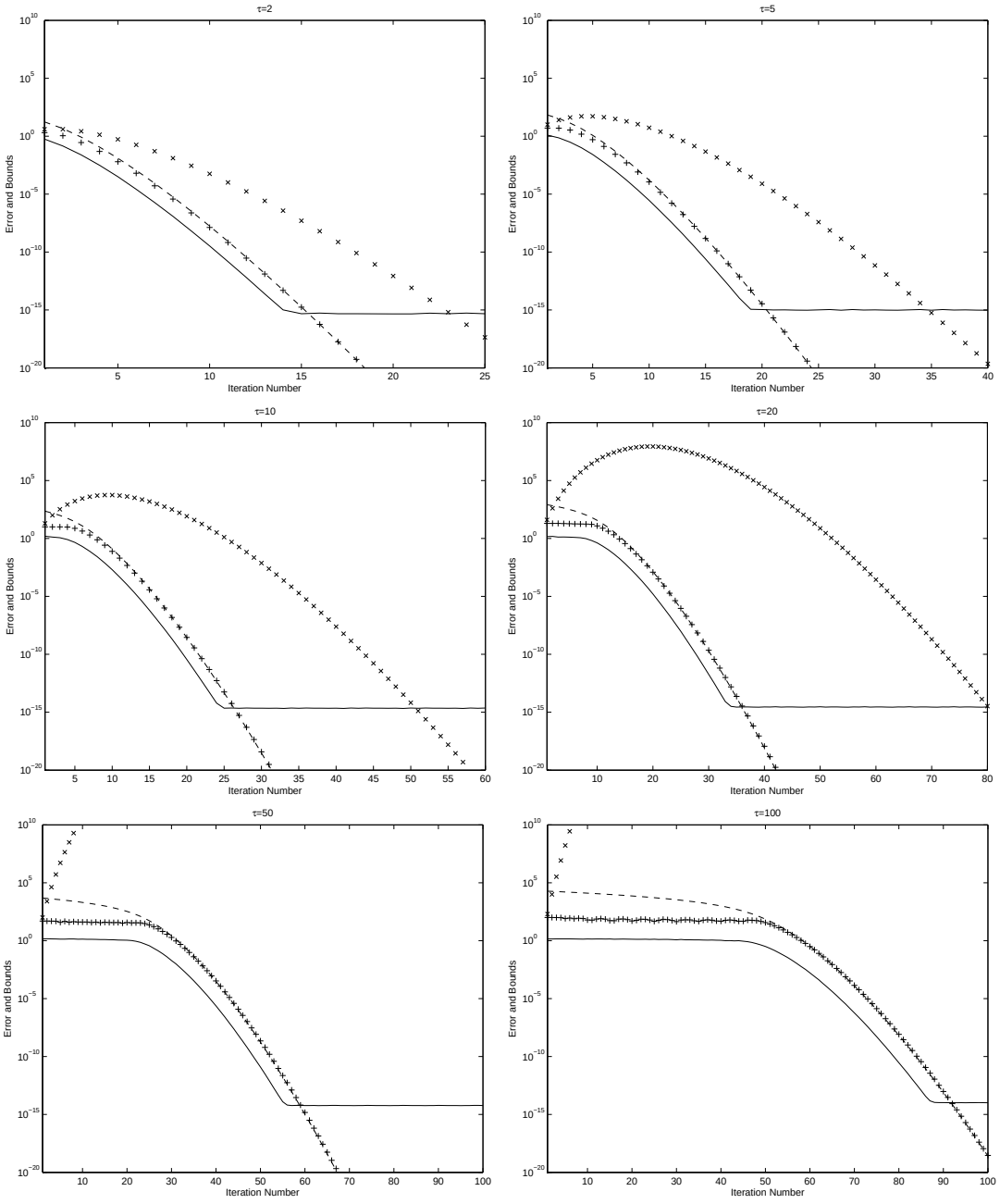

Figure 4.1: Example 1. 1000×1000 diagonal matrix with $a_{jj} = j/1000$. $\tau = 2, 5, 10, 20, 50, 100$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

bound of Saad. For $\tau > 50$, Saad's bound increase dramatically while our bound follows the actual error quite closely. Also note that for all cases, our *a posteriori* bound follows the actual convergence closely.

In addition, these tests show that the approximation error $||w(\tau) - w_k(\tau)||$ first

stagnates for certain number of iterations before it starts to converge. In the last section, we deduce that the convergence will not start until the iteration number $k = 2C$, where $C = \frac{\tau(b-a)}{4}$, and a, b are the smallest and the largest eigenvalues of A, respectively. For $\tau = 2, 5, 10, 20, 50, 100$, the corresponding k should be 1, 2.5, 5, 10, 25 and 50, respectively. They basically match our observations in Figure 4.1, especially when τ is relatively large and more iterations are needed for the convergence.

The above example shows that for different values of τ, the spectral gap of τA affects the error of $e^{i\tau A}v$ dramatically, and the starting time of convergence depends on this spectral gap. In our next example, we want to keep the same spectral gaps of τA while altering the eigenvalue distributions of A. Our tests show that our new *a posteriori* bound is still very sharp while the *a priori* bound is less optimistic.

Example 2. The construction of A is similar to that in Example 1. Let A be an $n \times n$ diagonal matrix whose j-th diagonal entry if $1/j$. Let v be a random n-dimensional normalized vector. We want to compute $w(\tau) = e^{i\tau A}v$ by applying k iterations of the Lanczos method to A and v. We will choose various values of τ and plot the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* bound (4.4) in the +-line, the optimized *a priori* bound (4.18) in the dashed line and Saad's bound (4.22) in the x-line.

In this test, we take the size $n = 1000$ and the iteration number $k = 100$. We also use the same values of $\tau = 2, 5, 10, 20, 50, 100$ so that we can have the same spectral gaps as in Example 1. The results are presented in Figure 4.2. The observations of Figure 4.2 are similar to those of Example 1. This similarity is expected since our new *a priori* bound (4.18) depends on the spectral gap and the matrices in two examples are constructed to have the same spectral gaps. It also shows that τ is small, our *a priori* error bound is comparable to the classical bound of Saad. As τ increases, our bound fits the actual convergence much better.

However, as we can see especially for large τ where more iterations are needed for the convergence, the actual convergence of the error $||w(\tau) - w_k(\tau)||$ is faster than that in Example 1. The reason is that, compared to the evenly spread eigenvalues in Example 1, the eigenvalues in this example are clustered at zero. After a few itera-

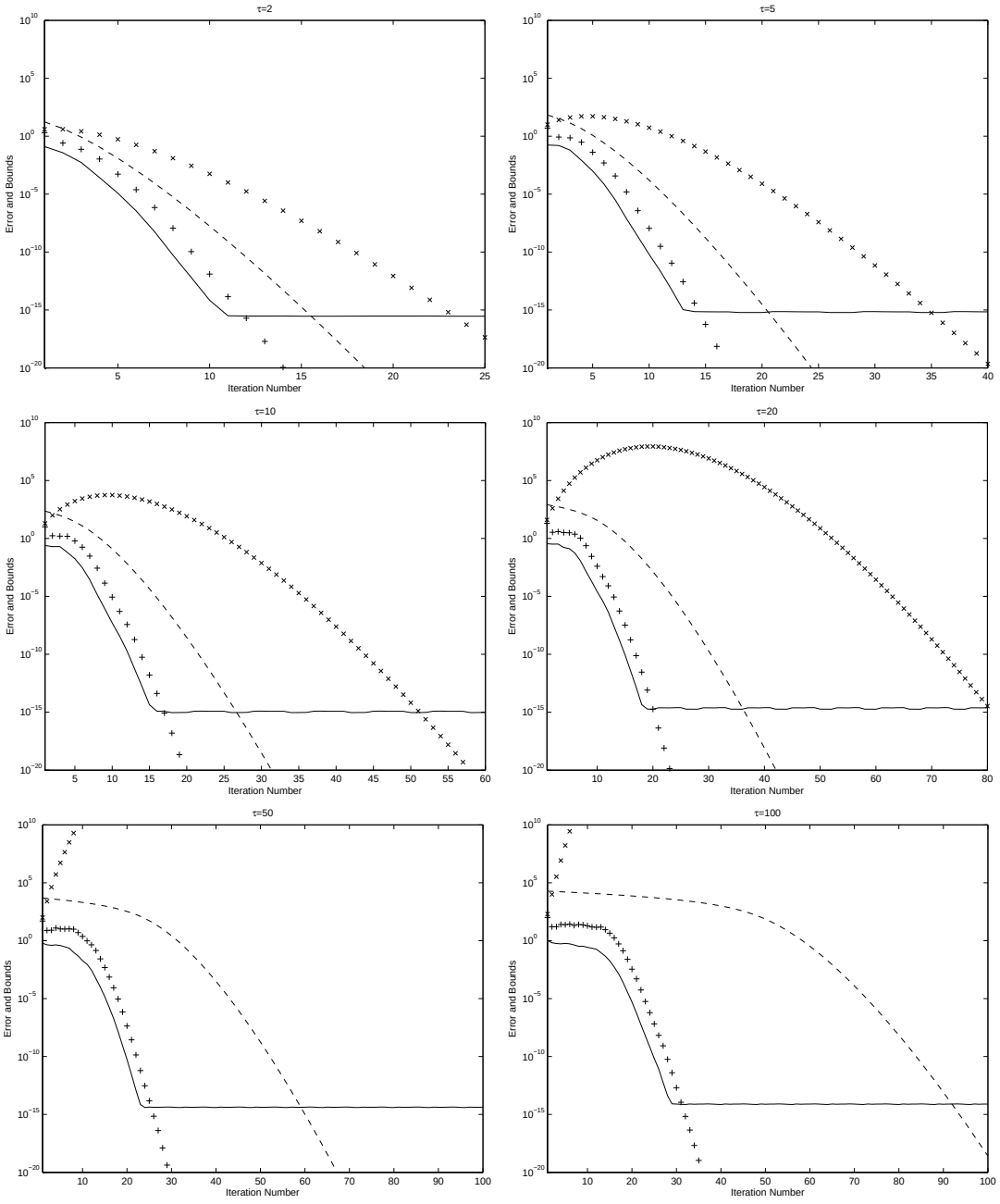

Figure 4.2: Example 2. 1000×1000 diagonal matrix with $a_{jj} = 1/j$. $\tau = 2, 5, 10, 20, 50, 100$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

tions, the length of the spectral gap of the matrix T_k will be significantly smaller than that of the original matrix A, due to the removal of separate eigenvalues Since both our *a priori* bound and its estimate of the beginning step of the actual convergence are based on the spectral gap of A, they will not match the actual convergence as

well as with a uniform eigenvalue distribution.

In the two examples above, the testing matrices are constructed to be sparse and diagonal, while $e^{-\tau A}$ are readily available. We also want to test our bounds on some smaller randomly generated dense matrices. In the next example, we generate a random symmetric matrix and use the MATLAB function expm for $e^{-\tau A}$.

Example 3. Let A be a uniformly random 500×500 symmetric matrix with $||A||_2 = 1$ and v be a random normalized vector. We apply 100 iterations for $\tau \in \{2, 5, 10, 20, 50, 100\}$ and plot the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* error bound (4.4) in the +-line, the optimized *a priori* error bound (4.18) in the dashed line and Saad's bound (4.22) in the x line in Figure 4.3.

From Figure 4.3, we observe that our *a posteriori* bound follows the error closely, for both small and large values of τ. Our new *a priori* bound, however, overestimates the actual error for several orders of magnitudes. By our construction of A, $||A||_2 = 1$ and there is an isolated eigenvalue at 1 while most eigenvalues are clustered near zero. So the convergence of the actual error behaves similarly as in Example 2. This again verifies the role of the spectral gaps in the convergence of the Lanczos method for computing $e^{i\tau A}$. Although our new *a priori* bound is pessimistic for large values of τ, it still improves significantly over the classical bound by Saad.

Example 4. In our final example, we consider a Laplacian matrix generated by a random graph. Let there be a graph containing 500 nodes. For each pair of nodes, there is a 50% chance that they are connected by an edge. The Laplacian matrix A of that graph is generated accordingly. By this construction, the norm of A is expected to be in hundreds so we take relatively small values of $\tau \in \{0.01, 0.02, 0.05, 0.1, 0.2, 0.5\}$. In Figure 4.4, we plot the actual error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* error bound (4.4) in the +-line, the optimized *a priori* bound (4.18) in the dashed line and the bound of Saad (4.22) in the x-line.

We first observe from Figure 4.4 that our *a posteriori* bound is very sharp in bounding the actual error. For small values of τ, our new *a priori* bound is comparable to Saad's classical bound. We observe again that our *a priori* bound significantly improves the classical *a priori* bound, although it is also pessimistic in the case

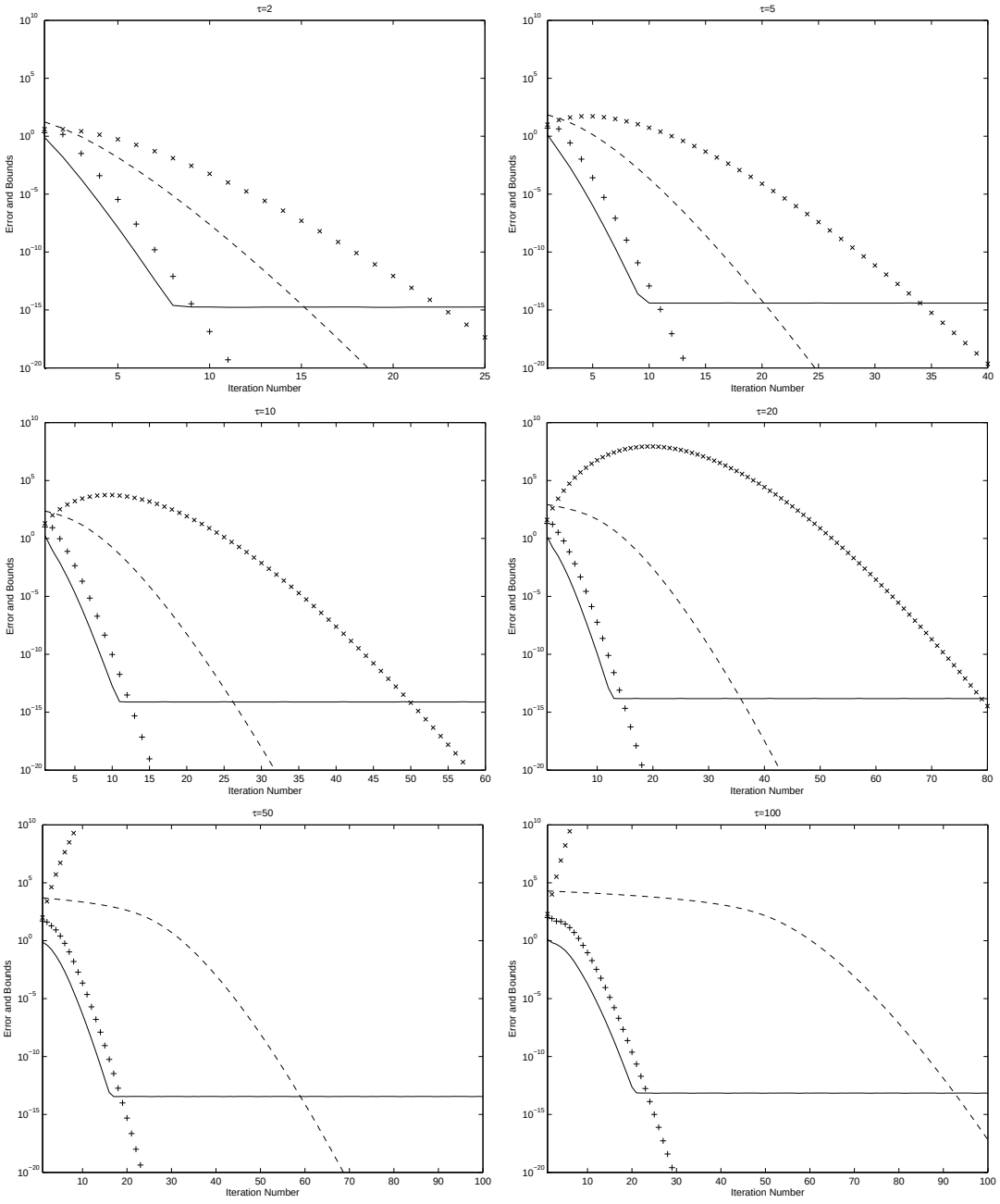

Figure 4.3: Example 3. Uniformly random matrix. $\tau = 2, 5, 10, 20, 50, 100$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

$\tau = 0.5$, which results in a value of a few hundreds for the norm of τA.

In all our numerical tests above, we have also compared the actual error $||w(\tau) - w_k(\tau)||$, our a priori bound (4.18) with Hochbruck and Lubich's bound (4.21). As our discussion in the last section illustrates, our new *a priori* bound is theoretically

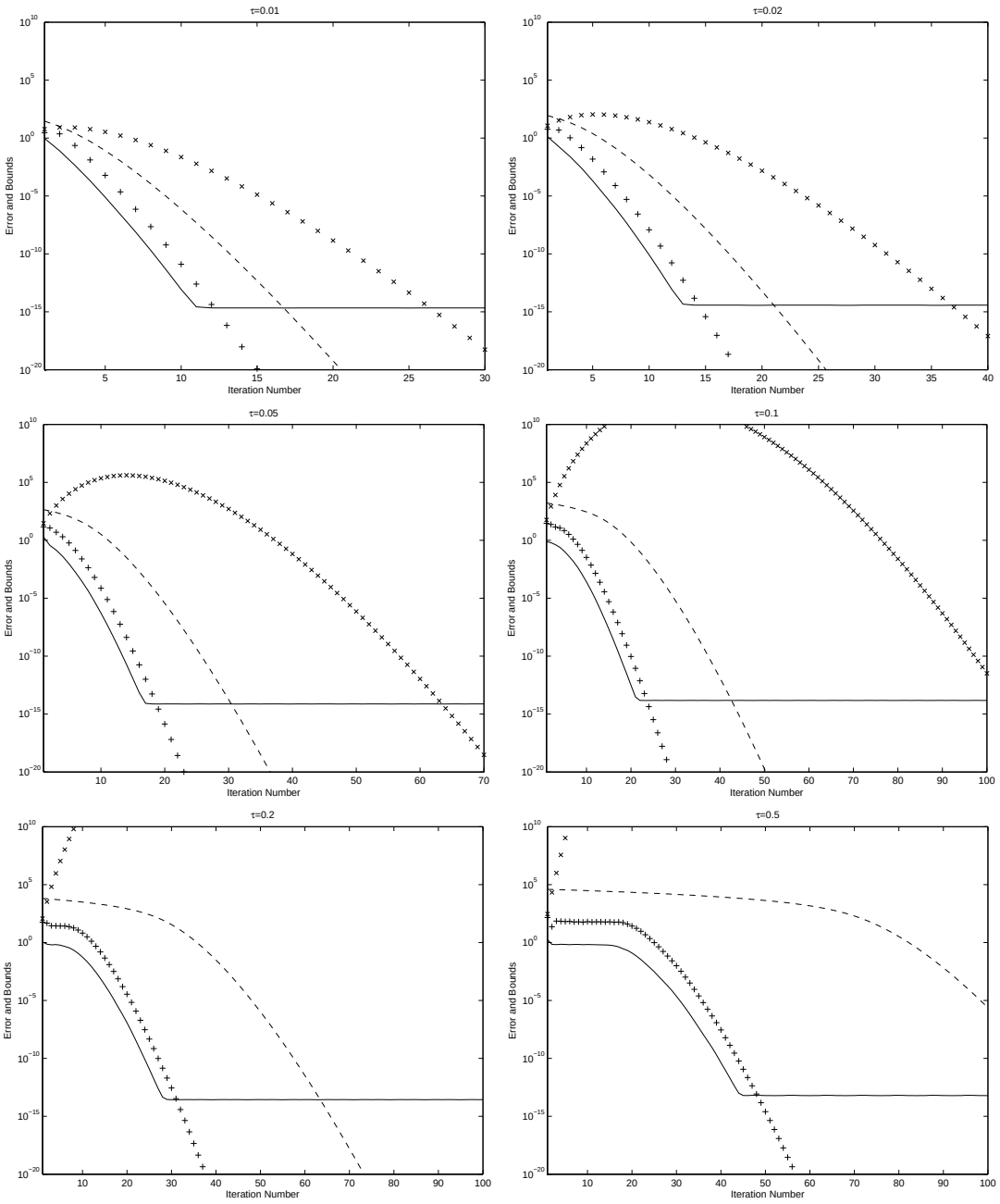

Figure 4.4: Example 4. 500×500 Laplacian matrix. $\tau = 0.01, 0.02, 0.05, 0.1, 0.2, 0.5$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Saad's bound (x).

better than Hochbruck and Lubich's bound, but they should be in almost the same order of magnitude after a large number of iterations. In all cases, the two bounds look indistinguishable after some iterations. At the beginning of the iteration, our bound is slightly better. We present the case in Example 1 with $\tau = 100$. In Figure

4.5 we plot the error $||w(\tau) - w_k(\tau)||$ in the solid line, the *a posteriori* bound (4.18) in the +-line, the *a priori* bound (4.18) in the dashed line, Hochbruch and Libuch's bound (4.21) in the dash-dotted line.

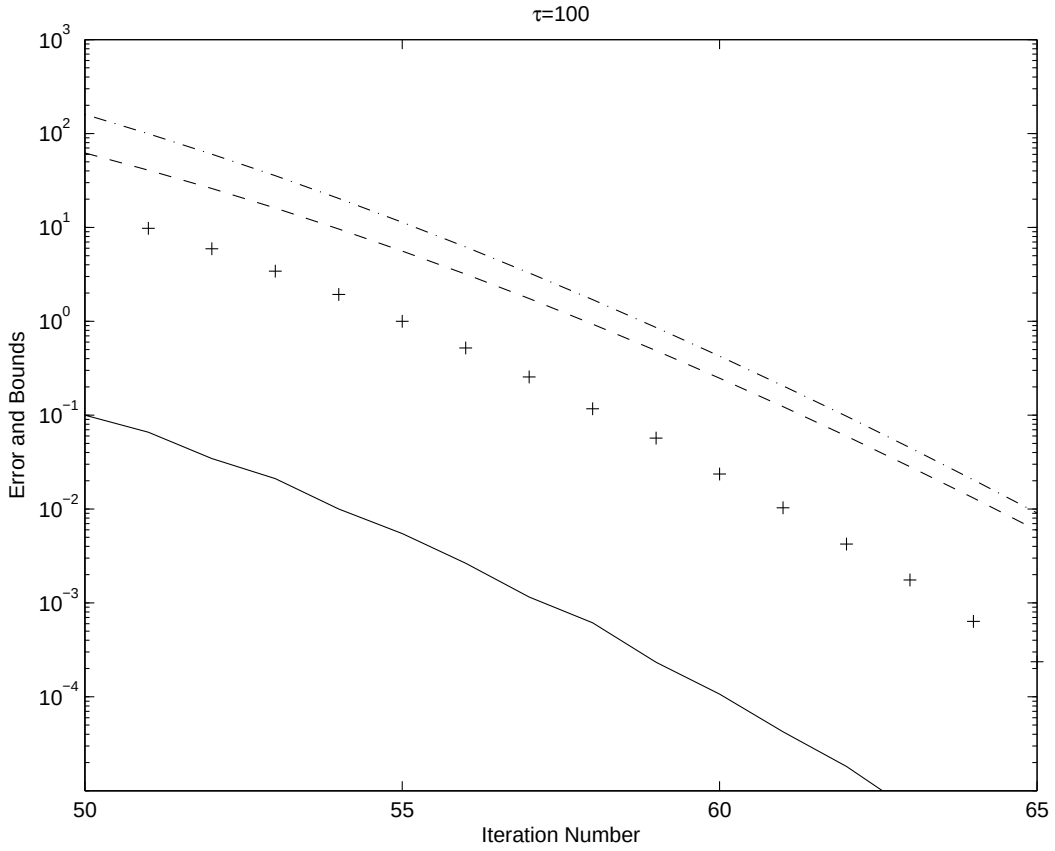

Figure 4.5: Example 1 with $\tau = 100$. Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Hochbruch and Lubich's bound (dash-dotted).

We observe from Figure 4.5 that in the first few iterations when the error starts to converge, our new *a priori* bound improves the bound by Hochbruch and Lubich by nearly one order of magnitude. After several steps, the two bounds become comparable.

Chapter 5 Conclusions

In this dissertation, we have discussed the application of the Krylov subspace methods in the computation of matrix exponentials. For the computation of $e^{-\tau A}v$ where A is a non-symmetric matrix whose eigenvalues are on the right half of the complex plane, we presented an *a posteriori* error bound related to the entry of the exponential of a Hessenberg matrix. We have also investigated the decay properties of the exponentials of Hessenberg matrices and presented a new *a priori* error bound, with the help of Faber polynomials. This bound is numerically optimized and proved sharper than the exiting *a priori* bound by Saad [33]. Our new bound shows that the convergence of the Krylov subspace methods is determined by the distribution of the eigenvalues and it agrees with the existing bound by Ye [42].

As a special case, we are also interested in the computation where A is skew-Hermitian, or $e^{i\tau A}$ where A is symmetric. We presented the new *a posteriori* and *a priori* error bounds. The a priori bound is also optimized showing that the convergence is determined by the spectral gap of the matrix A. Furthermore, our new bound also shows that the approximation error of the Lanczos method firstly stagnates for certain number of iterations before it starts to converge. It is then verified in several numerical examples.

Bibliography

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions*, Dover Publications INC., 1965.

[2] A. H. Al-Mohy and N. J. Higham, *A new scaling and squaring algorithm for the matrix exponential*, SIAM J. Matrix Anal. Appl., 31(3) (2009), pp. 970989.

[3] B. Beckermann and L. Reichel, *Error estimates and evaluation of matrix functions via the Faber transform*, SIAM J. Numer. Anal., Vol. 47, No. 5, pp. 3849-3883.

[4] R. Bellman, *Introduction to Matrix Analysis*, McGraw-Hill, New York, 1969.

[5] M. Benzi, P. Boito, *Decay properties of functions of matrices over C*-algebras*, Linear Algebra and its Applications, Vol. 456, pp. 174-198.

[6] M. Benzi and G. H. Golub, *Bounds for the entries of matrix functions with applications to preconditioning*, BIT, 39 (1999), pp. 417-438.

[7] M. Benzi and N. Razouk, *Decay bounds and O(n) algorithms for approximating functions of sparse matrices*, Electron. Trans. Numer. Anal., 28 (2007), pp. 16-39.

[8] T. A. Bickart, *Matrix exponential: Approximation by truncated power series*, Proc. IEEE, 56 (1968), pp. 372-373.

[9] P. Castillo and Y. Saad, *Preconditioning the matrix exponential operator with applications*, J. Sci. Comput., 13 (1999), pp. 275302.

[10] M. Crouzeix, *Numerical range and functional calculus in Hilbert space*, Journal of Functional Analysis, vol. 244, 2007, pp. 668-690.

[11] G. Dahlquist, *Stability and error bounds in the numerical integration of ordinary differential equations*, Almqvist & Wiksells, Uppsala, 1958; Transactions of the Royal Institute of Technology, Stockholm, 1959.

[12] C. Davis, *Explicit functional calculus*, J. Linear Algebra Appl., 6 (1973), pp. 193199.

[13] J. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997.

[14] E. Deutsch, *On matrix norms and logarithmic norms*, Numer. Math., 24 (1975), pp. 4951.

[15] V. Druskin, A. Greenbaum, and L. Knizhnerman, *Using nonorthogonal Lanczos vectors in the computation of matrix functions*, SIAM J. Sci. Comput., 19 (1998), pp. 38-54.

[16] V. L. Druskin and L. A. Knizhnerman, *Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithemetic*, Numer. Linear Algebra Appl., 2 (1995), pp. 205-217.

[17] S. W. Ellacott, *Computation of Faber series with application to numerical polynomial approximation in the complex plane*, Math. Comp., 40 (1983), No. 162, pp. 575-587.

[18] W. Everling, *On the evaluation of e^{At} by power series*, Proc. IEEE, 55 (1967), p. 413.

[19] V. N. Faddeeva, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[20] E. Gallopoulos and Y. Saad, *Efficient solution of parabolic equations by Krylov approximation methods*, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236-1264.

[21] N. J. Higham, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Appl., 26(4) (2005), pp. 11791193.

[22] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911-1925.

[23] L. Knizhnerman and V. Simoncini, *A new investigation of the extended Krylov subspace method for matrix function evaluations*, Numer. Linear Algebra Appl., 17 (2010), pp. 615-638.

[24] H. Kober, *Dictionary of Conformal Representations*, Dover Publications INC., 1957.

[25] M. L. Liou, *A novel method of evaluating transient response*, Proc. IEEE, 54 (1996), pp. 20-23.

[26] M. A. Malcolm and C. B. Moler, *Computer methods for mathematical computations*, Prentice-Hall, Englewood Cliffs, NJ, 1997.

[27] A. I. Markushevich, *Theory of functions of a complex variable, Vol. III*, Revised English edition translated and edited by Richard A. Silverman, Prenticd-Hall Inc., Englewood Cliffs, N.J., 1967.

[28] N. Mastronardi, M. Ng, AND E. E. Tyrtyshnikov, *Decay in functions of multiband matrices*, SIAM J. Matrix Anal. Appl., Vol. 31, No. 5, pp. 2721-2737.

[29] L. M. Milne-Thomson, *Jacobian Elliptic Function Tables*, Dover Publications INC., 1950.

[30] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Rev., 45 (2003), pp. 3-49.

[31] I. Moret and P. Novati, *On the convergence of Krylov subspace methods for matrix MittagLeffler functions*, SIAM J. Numer. Anal., 49 (2011), pp. 2144-2164.

[32] N. Razouk, *Localization phenomena in matrix functions: theory and algorithms*, Ph.D. thesis.

[33] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209-228.

[34] R. Shankar, *Principles of Quantum Mechanics (2nd Ed.)*, Kluwer Academic / Plenum Publishers, ISBN 978-0-306-44790-7.

[35] G. Söderlind, *The logarithmic norm. History and modern theory*, BIT Numerical Mathematics, 46 (2006), pp. 631-652.

[36] G. W. Stewart, *Matrix Algorithms: Volume I: Basic Decompositions*, SIAM, 1998, ISBN 978-0-89871-414-2.

[37] G. W. Stewart, *Matrix Algorithms: Volume II: Eigensystems*, SIAM, 2001, ISBN 978-0-89871-503-3.

[38] O. Toeplitz, *Das algebraische Analogon zu einem Satz von Fejer*, Math. Z. 2 (1918), 187-197.

[39] R. C. Ward, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1997), pp. 600-610.

[40] J. Xue and Q. Ye, *Entrywise relative perturbation bounds for exponentials of essentially nonnegative matrices*, Numer. Math., 110 (2008), pp. 393-403.

[41] J. Xue and Q. Ye, *Computing exponentials of essentially non-negative matrices entrywise to high relative accuracy*, Math. Comp., 82 (2013), pp. 1577-1596.

[42] Q. Ye, *Error bounds for the Lanczos methods for approximating matrix exponentials*, SIAM J. Numer. Anal., 51 (2013), pp. 66-87.

Vita

Education

- University of Kentucky, Lexington, Kentucky
  M. A. in Mathematics, 2011.

- University of Science and Technology of China, Hefei, Anhui, China
  B. S. in Mathematics, 2009.

Experience

- Teaching Assistant, Department of Mathematics, University of Kentucky, August 2009 - December 2015.

- Research Assistant under Dr. Qiang Ye, Department of Mathematics, University of Kentucky, Summer 2011, Summer 2012, Fall 2013, Summer 2015.

- Software Engineer Intern, Siemens PLM, Cypress, CA, May 2014 - August 2014.