# Error Bounds for the Krylov Subspace Methods for Computations of Matrix Exponentials

Hao Wang \* Qiang Ye<sup>†</sup>

#### Abstract

In this paper, we present new a posteriori and a priori error bounds for the Krylov subspace methods for computing  $e^{-\tau A}v$  for a given  $\tau > 0$  and  $v \in \mathbb{C}^n$ , where A is a large sparse non-Hermitian matrix. The *a priori* error bounds relate the convergence to  $\lambda_{\min}\left(\frac{A+A^*}{2}\right)$ ,  $\lambda_{\max}\left(\frac{A+A^*}{2}\right)$  (the smallest and the largest eigenvalue of the Hermitian part of A) and  $|\lambda_{\max}\left(\frac{A-A^*}{2}\right)|$  (the largest eigenvalue in absolute value of the skew-Hermitian part of A), which define a rectangular region enclosing the field of values of A. In particular, our bounds explain an observed superlinear convergence behavior where the error may first stagnate for certain iterations before it starts to converge. The special case that A is skew-Hermitian is also considered. Numerical examples are given to demonstrate the theoretical bounds.

#### 1 Introduction

The problem of computing matrix exponentials arises in many theoretical and practical problems. Numerous methods have been developed to efficiently compute  $e^{-A}$  or its product with a vector  $e^{-A}v$ , where A is an  $n \times n$  complex matrix and  $v \in \mathbb{C}^n$ . We refer to the classical paper [22] of Moler and Van Loan for a survey of a general theory and numerical methods for matrix exponentials. For matrix exponential problems involving a large and sparse matrix A, it is usually the product of the exponential with a vector that is of interest. This arises, for example, in solving the initial value problem ([14, 27])

$$\dot{x}(t) = -Ax(t) + b(t), \ x(0) = x_0.$$
 (1.1)

See [12, 16, 24] for some other applications.

A large number of matrix exponential problems concern a *positive definite* A (i.e.  $A + A^*$  is Hermitian positive definite), which defines a stable dynamical system (1.1) with a solution converging to a steady state. Another important class of problems involve a skew-Hermitian matrix A (i.e. A = iH with H being Hermitian), for which (1.1) has a norm-conserving

<sup>\*</sup>Department of Biomedical Engineering, University of Kentucky, Lexington, KY 40506, USA. E-mail: hao.wang@uky.edu. Research supported in part by NSF under Grant DMS-1318633.

<sup>&</sup>lt;sup>†</sup>Department of Mathematics, University of Kentucky, Lexington, KY 40506, USA. E-mail: qye3@uky.edu. Research supported in part by NSF under Grant DMS-1317424 and DMS-1318633.

solution. Such systems can be used to model a variety of physical problems where certain quantities such as energy are conserved. For example, a spectral method for solving the time-dependent Schrödinger equation modeling N electrons leads to (1.1) with a skew-Hermitian matrix; see [15, 25, 26]. While we will study a general non-Hermitian A, we are particularly interested in these two important classes of problems, where stronger theoretical results can be derived.

The Krylov subspace methods are a powerful class of iterative algorithms for solving many large scale linear algebra problems. Initially introduced by Gallopoulos and Saad [14, 27], they have also become a popular method for approximating

$$w(\tau) := e^{-\tau A} v, \tag{1.2}$$

where  $\tau \in \mathbb{R}$  is a fixed parameter typically representing a time step. For the ease of notation, we will assume throughout that  $||v||_2 = 1$ . A comprehensive theory has been developed in the literature with error bounds demonstrating convergence of the Krylov subspace methods and its relation to certain properties of the matrix. For example, earlier results in [14, 27] relate convergence of the Krylov subspace methods to the norm of the matrix  $\tau A$ . More refined error bounds have later been derived, that provide sharper estimates of the errors by considering additional spectral information such as enclosing regions of the field of values of A or positive definiteness of A; see [2, 11, 12, 17, 18, 23, 27] and the references contained therein. For a real symmetric positive definite matrix A, it has been shown in a recent work [30] that the speed of convergence is also determined by the condition number of A as in the conjugate gradient method. For positive definite matrices that are not necessarily Hermitian, stronger convergence bounds have also been obtained in [2, 12, 17, 18] in terms of the field of values. However, most of these bounds are derived by assuming the field of values lying in a certain pre-defined region, which are not easy to apply or interpret. There is an inherited theoretical difficulty in quantitatively characterizing the influence on the convergence by the field of values, a two dimensional object. This issue also arises in the theory of the Krylov subspace methods for solving linear systems.

In this paper, we study the relation between the convergence of the Krylov subspace methods and the field of values through its bounding rectangle  $[a, b] \times [-c, c]$  where  $a = \lambda_{\min}\left(\frac{A+A^*}{2}\right)$ ,  $b = \lambda_{\max}\left(\frac{A+A^*}{2}\right)$  (the smallest and the largest eigenvalue of the Hermitian part of A) and  $c = \left|\lambda_{\max}\left(\frac{A-A^*}{2}\right)\right|$  (the largest eigenvalue in absolute value of the skew-Hermitian part of A). With this approach, new *a priori* error bounds will be derived in terms of a, b and c. Simplified bounds will be presented for non-Hermitian positive definite matrices and skew-Hermitian matrices, which relate the speed of convergence to the size and the shape of the rectangular region. In particular, our bounds explain an interesting observed convergence behavior where the error may first stagnate for certain iterations before it starts to converge. Numerical examples will be presented to demonstrate the behavior of the new error bounds.

In developing our *a priori* error bounds, we also derive a new *a posteriori* error bound that is shown to provide a sharp and computable estimate of the error. The main technique used in deriving new *a priori* error bounds is the same as in the literature [3, 7, 2, 17, 18] by constructing Faber polynomial approximation of the exponential function in a region containing the field of values. The novelty in this work is to use the Jacobi elliptic functions

to construct a conformal mapping for the rectangular region that tightly encloses the fields of value and to show that this highly complicated mapping can be simplified to yield some simple final bounds.

The paper is organized as follows. In Section 2, we first present some preliminaries about the Faber polynomial approximation and the Jacobi elliptic functions. In Section 3, we present a new *a posteriori* error bound, which relates the convergence to the decay properties of functions of banded matrices. To study this decay behavior, we construct a conformal mapping in Section 4 and present our new *a priori* error bound in Section 5. In Section 6, we apply the same idea on skew-Hermitian matrices and derive simpler *a priori* bounds. Numerical examples are presented in Section 7 and some concluding remarks in Section 8.

# 2 Preliminaries

In this section, we briefly discuss some related results in complex analysis that will be needed.

#### 2.1 Faber polynomials

Faber polynomials extend the theory of power series to domains more general than a disk. It starts with the Riemann mapping theorem [20, Theorem 1.2] that 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. Let  $\overline{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$  be the extended complex plane and D be a bounded, closed continuum in the complex plane with boundary  $\Gamma$  such that the complement of D is simply connected in the extended plane and contains the point at  $\infty$ . A continuum is a non-empty, compact and connected subset of  $\mathbb{C}$ . Then there exists a function  $w = \Phi(z)$  which maps the complement of D conformally onto the exterior of a circle  $|w| = \rho > 0$  and satisfies the normalization conditions

$$\Phi(\infty) = \infty, \ \lim_{z \to \infty} \frac{\Phi(z)}{z} = 1.$$
(2.1)

Then, the function  $\Phi(z)$  has a Laurent expansion at infinity of the form

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

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

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

at infinity [20, p. 104]. Then, we call the following polynomial containing non-negative powers of z in the expansion

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

the Faber polynomials generated by D.

The Faber polynomials can be used to approximate analytic functions on D, essentially through the power series approximation of a transformed function on  $|w| \leq \rho$ . Let  $\Psi$  be the inverse of  $\Phi$  and let  $C_R$  be the image under  $\Psi$  of the circle  $|w| = R > \rho$ . We denote by  $I(C_R)$ the bounded region enclosed by  $C_R$ . By [20, Theorem 3.17], every function f(z) analytic on  $I(C_R)$  can be represented on  $I(C_R)$  as a series of the Faber polynomials

$$f(z) = \sum_{n=0}^{\infty} a_n \Phi_n(z)$$
(2.2)

with the coefficients  $a_n = \frac{1}{2\pi i} \int_{|w|=R} \frac{f[\Psi(w)]}{w^{n+1}} dw$ . The partial sum of the above series

$$\Pi_N(z) = \sum_{n=0}^N a_n \Phi_n(z)$$
(2.3)

is a polynomial of degree at most N that we can use to approximate f(z) on  $I(C_R)$ . The next theorem of [13] presents some approximation bounds concerning  $\Pi_N$ . We first need to introduce the definition of total rotation of the boundary. For this, we assume D is a closed Jordan region, i.e. its boundary  $\Gamma$  is rectifiable. Then there exists a tangent vector that makes an angle  $\Theta(z)$  with the positive real axis at almost all points  $z \in \Gamma$ . We say that  $\Gamma$ has bounded total rotation V if  $V = \int_{\Gamma} |d\Theta(z)| < \infty$ . Then  $V \ge 2\pi$  and the equality holds if D is convex; see [13].

**Theorem 2.1.** [13, Corollary 2.2] Assume D is a closed Jordan region whose boundary  $\Gamma$  has bounded total rotation V. For any  $R > \rho$ , let f be an analytic function in  $I(C_R)$ . We have for any  $N \ge 0$ ,

$$||f - \Pi_N||_{\infty} \le \frac{M(R)V}{\pi} \frac{\left(\frac{\rho}{R}\right)^{N+1}}{1 - \frac{\rho}{R}},\tag{2.4}$$

where  $M(R) = \max_{z \in C_R} |f(z)|$  and  $|| \cdot ||_{\infty}$  denotes the uniform norm on  $I(C_R)$ .

Theorem 2.1 is stated with  $C_R$  defined from the conformal map  $\Phi$  satisfying the normalization condition (2.1). In the literature (see [2] for example), another normalization has also been used and may be more convenient in our application. We may consider a conformal map  $\hat{\Phi}$  that maps the exterior of D onto the exterior of the unit disk (i.e. requiring  $\rho = 1$ rather than (2.1)). The above theorem can be adapted to  $\hat{\Phi}$  through a simple normalization transformation. Namely, given  $\hat{\Phi}$ , let  $\rho = \lim_{z \to \infty} \frac{z}{\hat{\Phi}(z)}$  and  $\Phi(z) := \rho \hat{\Phi}(z)$ , where we assume  $\rho$ is finite. Then  $\Phi$  satisfies the normalization condition (2.1) but now maps the exterior of Donto the exterior of the disk  $|w| = \rho$ . Applying Theorem 2.1 to  $\Phi$ , (2.4) holds for any  $R > \rho$ . Let  $r := R/\rho > 1$ . Let  $C_R$  be the inverse image under  $\Phi$  of the circle |w| = R and  $\hat{C}_r$  be the inverse image under  $\hat{\Phi}$  of the circle |w| = r. It is easy to check that  $C_R = \hat{C}_r$  and then  $M(R) := \max_{z \in C_R} |f(z)| = \max_{z \in \hat{C}_r} |f(z)|$ . Thus, (2.4) is reduced to

$$||f - \Pi_N||_{\infty} \le \frac{\widehat{M}(r)V}{\pi} \frac{\left(\frac{1}{r}\right)^{N+1}}{1 - \frac{1}{r}},$$
(2.5)

where  $\widehat{M}(r) := \max_{\widehat{\Phi}(z)=r} |f(z)|$ . Namely, Theorem 2.1 holds verbatim for a conformal map that is normalized to map the exterior of D onto the exterior of the unit disk. We note however that  $\rho$  as defined in the two normalizations is invariant and is called logarithmic capacity of D.

#### 2.2 Jacobi elliptic functions

In this subsection, we introduce the Jacobi elliptic functions, which will be used to construct a conformal mapping in Section 5. More details about the Jacobi elliptic functions can be found in [1].

Elliptic functions were first introduced as inverse functions of (incomplete) elliptic integrals. So before the introduction of the Jacobi elliptic functions, we first state the definition and properties of elliptic integrals. Given  $\phi \in \mathbb{C}$  and a real parameter m with 0 < m < 1, the (incomplete) Jacobi elliptic integral of the first kind is defined as

$$F(\phi, m) := \int_0^{\phi} (1 - m \sin^2 \theta)^{-\frac{1}{2}} d\theta.$$
 (2.6)

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

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

When  $\phi = \frac{\pi}{2}$ , the corresponding integrals

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

are called the complete Jacobi elliptic integrals of the first kind and the second kind. Let  $m_1 := 1 - m$ , the complementary parameter of m. Then,  $0 < m_1 < 1$ . For simplicity, we shall use the following notations.

$$K := K(m), \quad K' := K(m_1) = K(1-m); E := E(m), \quad E' := E(m_1) = E(1-m).$$
(2.7)

We now introduce the Jacobi elliptic functions. There are a total of twelve Jacobi elliptic functions in the family, but we will only discuss the basic three of them that will be used in this work. If  $u = F(\phi, m)$  where  $F(\phi, m)$  is the incomplete elliptic integral of the first kind defined in (2.6), three of the Jacobi elliptic functions are defined as

$$sn(u|m) := \sin \phi$$

$$cn(u|m) := \cos \phi$$

$$dn(u|m) := \sqrt{1 - m \sin^2 \phi}$$
(2.8)

The notations  $sn(\sigma|m)$ ,  $cn(\sigma|m)$  and  $dn(\sigma|m)$  indicate that sn, cn and dn are functions of two independent arguments: a complex argument u and a real parameter  $m \in (0, 1)$ . Furthermore, for a fixed  $m \in (0, 1)$ , sn(u) := sn(u|m), cn(u) := cn(u|m) and dn(u) :=dn(u|m) are doubly periodical meromorphic functions defined on  $u \in \mathbb{C}$  [21, p. 14].

In later sections, we will need some properties of the Jacobi elliptic integrals and Jacobi elliptic functions. We summarize them in the proposition below. For details, see [1], [19] and [21].

**Proposition 2.2.** 1. K = K(m) and E = E(m) are positive-valued functions of m. Moreover, they are differentiable with respect to the parameter  $m \in (0, 1)$ , and

$$\frac{dK}{dm} = \frac{E - m_1 K}{2mm_1},\tag{2.9}$$

$$\frac{dE}{dm} = \frac{E - K}{2m}.$$
(2.10)

2. [1, 17.3.26, p. 591]

$$\lim_{m \to 1} \left[ K - \frac{1}{2} \ln \left( \frac{16}{m_1} \right) \right] = 0 \tag{2.11}$$

3. [1, 17.4.5, p. 592]

$$E(u+2iK') = E(u) + 2i(K'-E')$$
(2.12)

4. sn, cn and dn satisfy

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

5. [1, Table 16.2, p. 570] sn, cn and dn are one-valued, doubly-periodic functions. For any  $l, n \in \mathbb{Z}$ ,

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

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

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

6. [1, Table 16.8, p. 572]

$$sn(2iK' - \sigma|m) = sn(-\sigma|m) = -sn(\sigma|m)$$
  

$$cn(2iK' - \sigma|m) = -cn(-\sigma|m) = -cn(\sigma|m)$$
  

$$dn(2iK' - \sigma|m) = -dn(-\sigma|m) = -dn(\sigma|m)$$
(2.13)

7. [1, Table 16.16, p. 574] Derivatives:

$$\frac{d}{du}sn(u|m) = cn(u|m) \cdot dn(u|m)$$
(2.14)

$$\frac{d}{du}cn(u|m) = -sn(u|m) \cdot dn(u|m)$$
(2.15)

$$\frac{d}{du}dn(u|m) = -m \cdot sn(u|m) \cdot cn(u|m)$$
(2.16)

8. [1, 16.21, p. 575] Let u = x + iy where  $x, y \in \mathbb{R}$  and denote

$$s = sn(x|m), c = cn(x|m), d = dn(x|m),$$
  

$$s_1 = sn(y|m_1), c_1 = cn(y|m_1), d_1 = dn(y|m_1),$$

Then

$$sn(x+iy|m) = \frac{s \cdot d_1 + ic \cdot d \cdot s_1 \cdot c_1}{c_1^2 + ms^2 \cdot s_1^2}$$
(2.17)

$$cn(x+iy|m) = \frac{c \cdot c_1 + is \cdot d \cdot s_1 \cdot d_1}{c_1^2 + ms^2 \cdot s_1^2}$$
(2.18)

$$dn(x+iy|m) = \frac{d \cdot c_1 \cdot d_1 + ims \cdot c \cdot s_1}{c_1^2 + ms^2 \cdot s_1^2}$$
(2.19)

We will also need to use the signs of the real and imaginary parts of sn(u|m), cn(u|m)and dn(u|m) when  $m \in (0, 1)$  and  $u \in \mathbb{C}$  is in the rectangular domain  $[-K, K] \times [0, 2iK']$ (i.e.  $\operatorname{Re}(u) \in [-K, K]$  and  $\operatorname{Im}(u) \in [0, 2K']$ ). This is discussed in [19, pp. 172-176] and we summarize it in Table 1, 2 and 3 for easy future references.

$\operatorname{Im}(u)$ $\operatorname{Re}(u)$	(-K, 0)	(0, K)
(K', 2iK')	(-, -)	(+, -)
(0,K')	(-, +)	(+, +)

Table 1: Signs of  $(\operatorname{Re}(sn(u|m)), \operatorname{Im}(sn(u|m)))$ 

$\boxed{\begin{array}{c} \operatorname{Re}(u)\\ \operatorname{Im}(u) \end{array}}$	(-K, 0)	(0, K)
(K', 2iK')	(-,+)	(-, -)
(0,K')	(+,+)	(+, -)

Table 2: Signs of  $(\operatorname{Re}(cn(u|m)), \operatorname{Im}(cn(u|m)))$ 

$\operatorname{Re}(u)$ $\operatorname{Im}(u)$	(-K, 0)	(0, K)
(K', 2iK')	(-,+)	(-, -)
(0,K')	(+, +)	(+, -)

Table 3: Signs of  $(\operatorname{Re}(sn(u|m)), \operatorname{Im}(sn(u|m)))$ 

### 3 A posteriori error bound

In this section, we first introduce the Arnoldi method for approximating  $w(\tau) = e^{-\tau A}v$  and then discuss an *a posteriori* error bound. Given  $A \in \mathbb{C}^{n \times n}$  and  $v \in \mathbb{C}^n$  with  $||v||_2 = 1$ , k iterations of the Arnoldi process generates an orthonormal basis  $\{v_1, v_2, \cdots, v_k, v_{k+1}\}$  for the Krylov subspace  $K_{k+1}(A, v) = span\{v, Av, A^2v, \cdots, A^kv\}$  by

$$h_{k+1,k}v_{k+1} = Av_k - \sum_{i=1}^k h_{i,k}v_i, \ h_{k+1,k} \ge 0.$$

Simultaneously, a k-by-k upper Hessenberg matrix  $H_k = [h_{ij}]$  is generated satisfying

$$AV_k = V_k H_k + h_{k+1,k} v_{k+1} e_k^T, (3.1)$$

where  $V_k = [v_1, v_2, \cdots, v_k]$  and  $e_k \in \mathbb{R}^n$  is the k-th coordinate vector. We note that

$$h_{k+1,k}^{2} = \|Av_{k}\|^{2} - \sum_{i=1}^{k} h_{i,k}^{2} \le \|A\|^{2}.$$
(3.2)

We can approximate  $w(\tau) = e^{-\tau A}v$  by its orthogonal projection on  $K_k(A, v)$ ,  $V_k V_k^T e^{-\tau A}v$ , which is further approximated as

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

We call

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

the Arnoldi approximation to  $w(\tau)$  in (1.2); see [14, 27].

Let  $W(A) := \{x^*Ax : x \in \mathbb{C}^n; \|x\|_2 = 1\}$  be the field of values of A and  $\mu(A) := \max\{\operatorname{Re}(z) : z \in W(A)\}$  be the logarithmic norm of A (associated with the Euclidean inner product). We also define  $\nu(A) := -\mu(-A) = \min\{\operatorname{Re}(z) : z \in W(A)\}$ . Then we have

$$\mu(A) = \lambda_{\max}\left(\frac{A+A^*}{2}\right) \text{ and } \nu(A) = \lambda_{\min}\left(\frac{A+A^*}{2}\right),$$
(3.4)

where  $\lambda_{\text{max}}$  and  $\lambda_{\text{min}}$  denote the largest and the smallest eigenvalues respectively. In this notation, A is positive definite if and only if  $\nu(A) > 0$ . An important property associated with the logarithmic norm [9, 28] is that for  $t \ge 0$ ,

$$||e^{tA}|| \le e^{t\mu(A)}.$$
 (3.5)

We now present a bound on the approximation error  $||w(\tau) - w_k(\tau)||$  in terms of the (k, 1) entry of the matrix  $e^{-tH_k}$ .

**Theorem 3.1.** Let  $A \in \mathbb{C}^{n \times n}$  and  $v \in \mathbb{C}^n$  with ||v|| = 1. Let  $V_k$  be the orthogonal matrix and  $H_k$  be the upper Hessenberg matrix generated by the Arnoldi process for A and v satisfying (3.1). Let  $w_k(\tau) = V_k e^{-\tau H_k} e_1$  be the Arnoldi approximation to  $w(\tau) = e^{-\tau A}v$ . Then the approximation error satisfies

$$||w(\tau) - w_k(\tau)|| \le h_{k+1,k} e^{-\min\{\nu(A),0\}\tau} \int_0^\tau |h(t)| dt,$$
(3.6)

where

$$h(t) := e_k^T e^{-tH_k} e_1 (3.7)$$

is the (k, 1) entry of the matrix  $e^{-tH_k}$  and  $\nu(A)$  is defined in (3.4).

*Proof.* First, we have  $w'(t) = -Ae^{-tA}v = -Aw(t)$  and

$$w'_{k}(t) = -V_{k}H_{k}e^{-tH_{k}}e_{1}$$
  
= -(AV\_{k} - h\_{k+1,k}v\_{k+1}e\_{k}^{T})e^{-tH\_{k}}e\_{1}  
= -Aw\_{k}(t) + h\_{k+1,k}h(t)v\_{k+1}.

Let  $E_k(t) := w(t) - w_k(t)$ . Then

$$E'_{k}(t) = -Aw(t) - (-Aw_{k}(t) + h_{k+1,k}h(t)v_{k+1})$$
  
=  $-AE_{k}(t) - h_{k+1,k}h(t)v_{k+1}.$ 

Note that  $E_k(0) = w(0) - w_k(0) = v - V_k e_1 = 0$ . Solving the initial value problem for  $E_k(t)$ , we have

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

Since  $\tau - t > 0$  in the integral, using (3.5), we have

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

Then the approximation error satisfies

$$||E_{k}(\tau)|| \leq h_{k+1,k} \left| \left| \int_{0}^{\tau} h(t)e^{(t-\tau)A}v_{k+1}dt \right| \right|$$
  
$$\leq h_{k+1,k} \int_{0}^{\tau} |h(t)| \cdot ||e^{(t-\tau)A}||dt$$
  
$$\leq h_{k+1,k} \int_{0}^{\tau} |h(t)| \cdot e^{(t-\tau)\nu(A)}dt$$

Thus, if  $\nu(A) \ge 0$ , we have  $||E_k(\tau)|| \le h_{k+1,k} \int_0^{\tau} |h(t)| dt$ . If  $\nu(A) < 0$ , then

$$||E_k(\tau)|| \le h_{k+1,k} \int_0^\tau |h(t)| e^{t\nu(A)} e^{-\tau\nu(A)} dt \le h_{k+1,k} e^{-\tau\nu(A)} \int_0^\tau |h(t)| dt.$$

This completes the proof.

h(t) in the above bound is computable *a posteriori* for any given *t*. Being the (k, 1) entry of the matrix  $e^{-tH_k}$ , it is expected to become small as *k* increases because of a decay property associated with functions of a banded matrix (see [3, 4, 5, 7]). This provides an understanding of the convergence of the error. Indeed, in §5, we shall extend the techniques introduced in [3, 7] to derive some sharp decay bounds on h(t), which will result in some new *a priori* bounds. Before we do that, we will need to construct some conformal mapping first in the next section.

We also remark that the *a posteriori* bound in the theorem contains the integral of h(t) that is not directly computable. For practical error estimates, we can approximate it using a quadrature rule, say, the Simpson's rule, by computing h(t) at some selected discrete points. This provides a fairly sharp *a posteriori* error estimates; see the numerical examples in §7. Note that there are several *a posteriori* error estimates presented in [27] derived from approximation of a different error expression, one of which is  $\tau h(\tau)$ .

#### 4 Conformal mapping

In this section, we construct a conformal mapping which maps the exterior of a rectangle onto the exterior of a unit disk and discuss some of its properties. Given a rectangle in  $\tilde{z}$ -plane whose vertices are  $a \pm ic$  and  $b \pm ic$  where b > a and c > 0, we map the exterior of this rectangle conformally onto |u| > 1. This can be done in the following three steps.

• Step 1:

$$z = \phi_1(\tilde{z}) = \tilde{z} - \frac{a+b}{2} \tag{4.1}$$

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  $\sigma$  by

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

$$(4.2)$$

where  $sn(\sigma|m)$ ,  $cn(\sigma|m)$  and  $dn(\sigma|m)$  are Jacobi elliptic functions and  $E(\sigma|m) := \int_0^{\sigma} dn^2(z|m)dz$ . The parameter *m* is determined from  $\alpha, \beta$  by the equation

$$\frac{E - m_1 K}{\beta} = \frac{E' - mK'}{\alpha},\tag{4.3}$$

here K, E, K' and E' are functions of m or  $m_1 := 1 - m$  defined in (2.7). The existence and uniqueness of m will be shown in Lemma 4.1 below. It is shown in [19, p. 178] that  $\phi_2$  conformally maps the exterior of the rectangle  $[-\alpha, \alpha] \times [-\beta, \beta]$  to the upper half plane {Im(w) > 0} and that the range of  $\sigma$  is in the rectangle  $[-K, K] \times [0, 2iK']$ .

• Step 3:

$$u = \phi_3(w) = \frac{i+w}{i-w} \tag{4.4}$$

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

Now let

$$\tilde{\Phi} := \phi_3 \circ \phi_2 \circ \phi_1 \tag{4.5}$$

be the composition of the above three conformal mappings defined in (4.1), (4.2) and (4.4). Then  $\tilde{\Phi}$  maps the exterior of the rectangle  $[a, b] \times [-c, c]$  conformally onto the exterior of the unit circle.

The rest of this section will present several results concerning  $\tilde{\Phi}$  that we will use in the next section, but first we give a proof of existence of a unique solution of (4.3) that appears not readily available in the literature.

**Lemma 4.1.**  $E(m) - (1-m)K(m) \in (0,1)$  is an increasing function and  $E'(m) - mK'(m) \in (0,1)$  is an decreasing function. For any  $0 < \alpha, \beta < +\infty$ , there exists a unique  $m \in (0,1)$ , as a function of  $\beta/\alpha$ , satisfying (4.3).

Proof. Let  $f(m) := E - m_1 K = E(m) - (1 - m)K(m)$  be a function of  $m \in (0, 1)$ . Then E'(m) - mK'(m) = f(1 - m). By the definition of K(m) and E(m),  $K(0) = \frac{\pi}{2}$ ,  $E(0) = \frac{\pi}{2}$ , and then

$$\lim_{m \to 0} f(m) = 0.$$
(4.6)

Moreover, by (2.11),

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

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

$$\lim_{m \to 1} f(m) = E(1) - \lim_{m \to 1} m_1 K(m) = 1.$$
(4.7)

By (2.9) and (2.10), 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

$$g(m) := \frac{f(m)}{f(1-m)} = \frac{E(m) - (1-m)K(m)}{E(1-m) - mK(1-m)}.$$
(4.8)

By (4.6) and (4.7), 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., (4.3).

The parameter m determined by (4.3) is defined by the aspect ratio  $\beta/\alpha$  (or the shape) of the rectangle  $[a, b] \times [-c, c]$ . For example, from the proof,  $m \approx 0$  if the rectangle is narrowly around the real axis, while  $m \approx 1$  if the rectangle is nearly a vertical line in the complex plane. When m = 1/2, the rectangle is a square.

As in §2, we denote by  $C_r$  in the  $\tilde{z}$ -plane the inverse image of the circle |u| = r under  $\tilde{\Phi}$  for a given r > 1. We need to determine the minimum of  $\operatorname{Re}(\tilde{z})$  in  $C_r$ , i.e. the left most point of  $C_r$ . First we prove a lemma about the Jacobi elliptic functions, which is a direct result of Proposition 2.2.

**Lemma 4.2.** For u = x + iy where -K < x < K and 0 < y < 2K',

$$\operatorname{sgn}(\operatorname{Im}(cn(u|m))) = \operatorname{sgn}(\operatorname{Im}(dn(u|m))).$$

*Proof.* By (2.18) and (2.19),

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

So,

$$\operatorname{sgn}(\operatorname{Im}(cn(u|m))) = \operatorname{sgn}(\operatorname{Im}(dn(u|m))) \cdot \operatorname{sgn}(cn(x|m) \cdot dn(x|m) \cdot dn(y|m_1))$$
(4.9)

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

$$cn(x|m) = \cos\phi > 0. \tag{4.10}$$

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

$$dn(x|m) > 0, \ dn(y|m_1) > 0.$$
 (4.11)

Applying (4.10) and (4.11) to (4.9), we conclude that the imaginary part of cn(u|m) and that of dn(u|m) have the same sign.

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

**Lemma 4.3.** Let  $\tilde{\Phi} : \tilde{z} \mapsto u$  be defined in (4.5). 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

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

Proof. By (4.1),

$$\frac{d\tilde{z}}{dz} = 1. \tag{4.12}$$

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 (4.2) that

$$\frac{dz}{d\sigma} = -\frac{i}{\lambda} \{ dn^2 - (1-m) \} = -\frac{i}{\lambda} \{ m - m \cdot sn^2 \} = -\frac{i}{\lambda} \cdot m \cdot cn^2.$$
(4.13)

Note that By (2.14) and (2.16), we have  $\frac{d(dn)}{d\sigma} = -m \cdot sn \cdot cn$  and  $\frac{d(sn)}{d\sigma} = cn \cdot dn$ . Then by (4.2),

$$\frac{dw}{d\sigma} = \frac{-(-m \cdot sn \cdot cn) \cdot \sqrt{m} \cdot cn - (1 - dn) \cdot \sqrt{m} \cdot cn \cdot dn}{m \cdot sn^2}$$

$$= \frac{\sqrt{m} \cdot cn \cdot (m \cdot sn^2 - dn + dn^2)}{m \cdot sn^2}$$

$$= \frac{\sqrt{m} \cdot cn \cdot (1 - dn)}{1 - dn^2}$$

$$= \frac{\sqrt{m} \cdot cn}{1 + dn}$$
(4.14)

By (4.4),  $w = i \frac{u-1}{u+1}$  and then

$$\frac{dw}{du} = \frac{2i}{(u+1)^2}.$$
(4.15)

Combining (4.12), (4.13), (4.14) and (4.15), we have

$$\frac{d\tilde{z}}{du} = \frac{d\tilde{z}}{dz} \cdot \frac{dz}{d\sigma} \cdot \frac{d\sigma}{dw} \cdot \frac{dw}{du} 
= -\frac{i}{\lambda} \cdot m \cdot cn^2 \cdot \frac{1+dn}{\sqrt{m} \cdot cn} \cdot \frac{2i}{(u+1)^2} 
= \frac{2\sqrt{m} \cdot cn(1+dn)}{\lambda(u+1)^2}.$$
(4.16)

(4.4) also implies

$$w^{2} = -\frac{(u-1)^{2}}{(u+1)^{2}}.$$
(4.17)

On the other hand, by (4.2),

$$w^{2} = \frac{(1-dn)^{2}}{m \cdot sn^{2}} = \frac{(1-dn)^{2}}{1-dn^{2}} = \frac{1-dn}{1+dn}.$$
(4.18)

So,

$$dn = \frac{1 - w^2}{1 + w^2} = \frac{(u+1)^2 + (u-1)^2}{(u+1)^2 - (u-1)^2} = \frac{1}{2} \left( u + \frac{1}{u} \right)$$
(4.19)

and hence

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

Substituting this into (4.16), we have

$$\frac{d\tilde{z}}{du} = \frac{\sqrt{m} \cdot cn}{\lambda u}.$$
(4.20)

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

$$\frac{du}{d\theta} = re^{i\theta} \cdot i = iu. \tag{4.21}$$

Treating  $\tilde{z} \in C_r$  as a function of  $\theta$ , we have from (4.20) and (4.21) that

$$\frac{d\tilde{z}}{d\theta} = \frac{i\sqrt{m}}{\lambda} \cdot cn(\sigma|m). \tag{4.22}$$

 $\operatorname{So}$ 

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

From (4.19) and  $u = r \cos \theta + ir \sin \theta$ , we write  $dn(\sigma|m)$  as a function of  $\theta$ ,

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

So  $\operatorname{Im}(dn(\sigma|m)) < 0$  when  $\theta \in (-\pi, 0)$ , and  $\operatorname{Im}(dn(\sigma|m)) > 0$  when  $\theta \in (0, \pi]$ . By Lemma 4.2, the imaginary part of  $cn(\sigma|m)$  always has the same sign as that of  $dn(\sigma|m)$ . Thus, by (4.22),  $\frac{d(\operatorname{Re}(\tilde{z}))}{d\theta} > 0$  when  $\theta \in (-\pi, 0)$ , and  $\frac{d(\operatorname{Re}(\tilde{z}))}{d\theta} < 0$  when  $\theta \in (0, \pi]$ . The minimum value of  $\operatorname{Re}(\tilde{z})$  is attained when  $\theta = \pi$ , i.e., u = -r.

Next, we find the explicit form for  $\tilde{\Psi}(-r)$  in Lemma 4.3.

**Lemma 4.4.** 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 (4.5), and let  $\tilde{\Psi} : u \mapsto \tilde{z}$  be its inverse. Then for any r > 1, we have

$$\tilde{\Psi}(-r) = a - \frac{1}{\lambda} \int_0^{\frac{1}{2}\left(r - \frac{1}{r}\right)} \frac{\sqrt{m + t^2}}{\sqrt{1 + t^2}} dt, \qquad (4.23)$$

where the parameters m is determined by (4.3) and  $\lambda$  is the ratio in (4.3).

*Proof.* Recall that  $\dot{\Phi} = \phi_3 \circ \phi_2 \circ \phi_1$  with  $\phi_1$ ,  $\phi_2$  and  $\phi_3$  the three conformal mappings defined in (4.1), (4.2) and (4.4). Let

$$\Phi := \phi_3 \circ \phi_2 \tag{4.24}$$

and  $\Psi$  be its inverse. Then obviously

$$\tilde{\Psi}(-r) = \phi_1^{-1} \circ \Psi(-r) \tag{4.25}$$

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

$$\Psi(r) = \alpha + \frac{1}{\lambda} \int_0^{\frac{1}{2}\left(r - \frac{1}{r}\right)} \frac{\sqrt{m + t^2}}{\sqrt{1 + t^2}} dt.$$
(4.26)

By the same equation (4.19) that was derived from (4.2) and (4.4), w in the map can be eliminated to define  $\Phi: z \longleftrightarrow \sigma \longleftrightarrow u$  through the auxiliary parameter  $\sigma$  as

$$\begin{cases} z(\sigma) = \alpha - \frac{i}{\lambda} \{ E(\sigma|m) - m_1 \sigma \} \\ dn(\sigma|m) = \frac{1}{2} \left( u + \frac{1}{u} \right) \end{cases}$$
(4.27)

To compute  $\Psi(r)$ , set u = r above. Then the corresponding  $\sigma$  satisfies

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

$$(4.28)$$

By Table 3,  $\sigma \in \mathbb{C}$  is on the line segment connecting 0 and iK'. Let

$$t = -i\sqrt{m} \cdot sn(s|m), \tag{4.29}$$

where s is on the line segment connecting 0 and  $\sigma$ . By Tables 1, 2 and 3, sn(s|m) is purely imaginary with positive imaginary part, and cn(s|m) and dn(s|m) are both real and positive. Then

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

By (4.29) and (2.14),

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

then

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

By (4.28),

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

then

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

Thus, as s moves along the positive imaginary axis from 0 to  $\sigma$ , t as defined by (4.29) moves along the positive real axis from 0 to  $\frac{1}{2}\left(r-\frac{1}{r}\right)$ . Then

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

This completes the proof of the first part (4.26).

We next prove for any r > 1,

$$\Psi(-r) = -\Psi(r). \tag{4.30}$$

Let  $\sigma$  and  $\tilde{\sigma}$  be the auxiliary parameters in (4.27) corresponding to r and -r respectively. 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 (2.13),  $\tilde{\sigma} = 2iK' - \sigma$ . Thus, using (2.12) and (4.3), we get

$$\Psi(-r) = z(\tilde{\sigma}) = \alpha - \frac{i}{\lambda} \{ E(2iK' - \sigma|m) - m_1(2iK' - \sigma) \}$$
  
$$= \alpha - \frac{i}{\lambda} \{ 2i(K' - E') - E(\sigma|m) - 2m_1iK' + m_1\sigma \}$$
  
$$= \alpha - \frac{i}{\lambda} \{ -2i(E' - mK') - [E(\sigma|m) - m_1\sigma] \}$$
  
$$= \alpha - \frac{i}{\lambda} \{ -2i \cdot \lambda\alpha - [E(\sigma|m) - m_1\sigma] \}$$
  
$$= -\alpha + \frac{i}{\lambda} \{ E(\sigma|m) - m_1\sigma \} = -z(\sigma) = -\Psi(r).$$

Finally, applying  $\phi_1^{-1}$  to  $\Psi(-r)$  as in (4.25) and noting that  $\alpha = \frac{b-a}{2}$ , (4.23) is proved.

Finally, we show that  $\tilde{\Phi}$  can be normalized according to (2.1).

**Lemma 4.5.** Let  $\lambda$  be the ratio in (4.3). We have

$$\lim_{\tilde{z}\to\infty}\frac{\tilde{\Phi}(\tilde{z})}{\tilde{z}} = 2\lambda > 0.$$

*Proof.* First, by (4.19) and  $m \cdot sn^2(\sigma|m) + dn^2(\sigma|m) = 1$ , we have  $\sqrt{m} \cdot sn(\sigma|m) = \frac{i}{2} \left(u - \frac{1}{u}\right)$ . Applying it to (4.20), we have

$$\frac{d\tilde{z}}{du} = \frac{i}{2\lambda} \cdot \frac{cn(\sigma|m)}{sn(\sigma|m)} \left(1 - \frac{1}{u^2}\right).$$
(4.31)

As  $\tilde{z} \to \infty$ ,  $\sigma \to iK'$  and  $u \to \infty$  (see [19, p. 178]). Since

$$\lim_{\sigma \to iK'} \frac{cn(\sigma|m)}{sn(\sigma|m)} = \lim_{\sigma \to iK'} \frac{cn'(\sigma|m)}{sn'(\sigma|m)} = \lim_{\sigma \to iK'} \frac{-sn(\sigma|m)dn(\sigma|m)}{cn(\sigma|m)dn(\sigma|m)} = -\left(\lim_{\sigma \to iK'} \frac{cn(\sigma|m)}{sn(\sigma|m)}\right)^{-1},$$

we have  $\lim_{\sigma \to iK'} \frac{cn(\sigma|m)}{sn(\sigma|m)} = -i$ . Applying it to (4.31),  $\frac{d\tilde{z}}{du} \to \frac{1}{2\lambda}$  or  $\frac{du}{d\tilde{z}} \to 2\lambda$  as  $\tilde{z} \to \infty$ . Then  $\frac{\tilde{\Phi}(\tilde{z})}{\tilde{z}} \to 2\lambda$  as  $\tilde{z} \to \infty$ .  $\lambda > 0$  follows from Lemma 4.1.

# 5 A priori error bound for non-Hermitian matrices

In this section, we derive new *a priori* error bounds for the Arnoldi approximations of  $e^{-\tau A}v$ . We shall bound the error in terms of the following spectral information of A:

$$\begin{cases} a = \min_{i} \left\{ \lambda_{i} \left( \frac{A + A^{*}}{2} \right) \right\} = \nu(A) \\ b = \max_{i} \left\{ \lambda_{i} \left( \frac{A + A^{*}}{2} \right) \right\} = \mu(A) \\ c = \max_{i} \left\{ \left| \lambda_{i} \left( \frac{A - A^{*}}{2} \right) \right| \right\} \end{cases}$$
(5.1)

where  $\lambda_i(M)$   $(1 \le i \le n)$  are the eigenvalues of M. These three numbers provide a region bounding W(A), the field of values of A, i.e. W(A) is contained in the rectangle  $[a, b] \times [-c, c]$ .

We shall study the convergence of the Arnoldi method through bounding |h(t)| (the (k, 1) entry of  $e^{-tH_k}$ ) in the *a posteriori* bound of §3 as in [30]. As mentioned before, analytic functions of banded matrices have a decay property, i.e. their entries decreases away from the main diagonal. Sharp decay bounds were originally derived by Benzi and Golub [5] for Hermitian matrices; see [4, 6] and the references contained therein for some further improvements. Generalizations to the non-Hermitian case, which is applicable to the Hessenberg matrix  $H_k$  here, have been obtained by Benzi and Razouk [7] and Benzi and Boito [3]. Specifically, for non-Hermitian matrices, the Faber polynomial approximation

and the conformal mappings on a circular region containing the field of value have been introduced in [3, 7] to bound the decay rate. Here we will follow the same approach of [3, 7], but we will use the conformal mapping that is constructed in §4 so as to utilize a more precise region  $[a, b] \times [-c, c]$  that encloses the field of values. By using a smaller bounding region, a stronger approximation result and hence a stronger bound are obtained as follows.

**Theorem 5.1.** Let  $H_k$  be a k-by-k upper Hessenberg matrix and let  $h(t) = e_k^T e^{-tH_k} e_1$  be the (k, 1) entry of the matrix  $e^{-tH_k}$ . Let  $a_k = \min_i \left\{ \lambda_i \left( \frac{H_k + H_k^*}{2} \right) \right\}$ ,  $b_k = \max_i \left\{ \lambda_i \left( \frac{H_k + H_k^*T}{2} \right) \right\}$  and  $c_k = \max_i \left\{ \left| \lambda_i \left( \frac{H_k - H_k^*}{2} \right) \right| \right\}$ . Then for any q with 0 < q < 1,

$$|h(t)| \le 2Q \, \frac{q^{k-1}}{1-q} e^{-t\tilde{z}},\tag{5.2}$$

where Q = 11.08,

$$\tilde{z} = a_k - \frac{1}{\lambda} \int_0^{\frac{1}{2} \left(\frac{1}{q} - q\right)} \frac{\sqrt{m + s^2}}{\sqrt{1 + s^2}} ds,$$
(5.3)

and the parameters m is determined from  $a_k$ ,  $b_k$ ,  $c_k$  by (4.3) and  $\lambda$  is the ratio in (4.3).

Proof. Let  $\Phi: \tilde{z} \mapsto u$  be the conformal mapping from the exterior of the rectangle  $[a_k, b_k] \times [-c_k, c_k]$  onto the exterior of the unit disk, as defined in (4.5). For a fixed  $t \geq 0$ , let  $f(z) = e^{-tz}$ . Since f is an analytic function, it can be approximated by the partial sum  $\Pi_{k-2}(z)$  of the series of Faber polynomials generated by  $\tilde{\Phi}$  as defined in (2.3). Let  $r = \frac{1}{q} > 1$  and consider  $C_r$ , the inverse image under  $\tilde{\Phi}$  of the circle |w| = r. Applying Theorem 2.1 or (2.5), the approximation error in  $I(C_r)$  is bounded as

$$||f - \Pi_{k-2}||_{\infty} = \max_{z \in I(C_r)} |f(z) - \Pi_{k-2}(z)| \le 2M(r) \frac{\left(\frac{1}{r}\right)^{k-1}}{1 - \frac{1}{r}},$$

where  $M(r) = \max_{z \in C_r} |f(z)|$  and we note that the total rotation around the rectangle is  $V = 2\pi$ . Since  $\Pi_{k-2}(z)$  is a polynomial of degree k-2,  $[\Pi_{k-2}(H_k)]_{k1} = e_k^T \Pi_{k-2}(H_k)e_1 = 0$ . Then

$$\begin{aligned} |h(t)| &= |[f(H_k)]_{k1}| = |[f(H_k)]_{k1} - [\Pi_{k-2}(H_k)]_{k1} \\ &\leq ||f(H_k) - \Pi_{k-2}(H_k)||_2 \\ &\leq Q \cdot \max_{z \in W(H_k)} |f(z) - \Pi_{k-2}(z)|, \end{aligned}$$

where  $W(H_k)$  is the field of values of  $H_k$  and the last inequality is by Crouzeix's Theorem [8]. Since  $W(H_k) \subseteq [a_k, b_k] \times [-c_k, c_k] \subseteq C_r$ , we have

$$|h(t)| \le Q \max_{z \in I(C_r)} |f(z) - \Pi_{k-2}(z)| \le 2 Q M(r) \frac{\left(\frac{1}{r}\right)^{k-1}}{1 - \frac{1}{r}}$$

Now, the theorem follows from  $M(r) = \max_{z \in C_r} e^{-tz} = \max_{z \in C_r} e^{-t\operatorname{Re}(z)} = e^{-t\tilde{z}}$ , where

$$\tilde{z} = \min\{\operatorname{Re}(z) : z \in C_r\} = \tilde{\Psi}(r) = a_k - \frac{1}{\lambda} \int_0^{\frac{1}{2}(\frac{1}{q}-q)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds$$

by Lemma 4.3 and Lemma 4.4.

_	_	1
		L
		L
		L

We remark that Q = 11.08 is called Crouzeix's constant and it is conjectured that it can be reduced to 2 [8]. Combining the above theorem with Theorem 3.1 leads to the following *a priori* error bound in the following theorem.

**Theorem 5.2.** Let  $A \in \mathbb{C}^{n \times n}$  and  $v \in \mathbb{C}^n$  with ||v|| = 1 and let  $w_k(\tau) = V_k e^{-\tau H_k} e_1$  be the Arnoldi approximation (3.3) to  $w(\tau) = e^{-\tau A}v$ . Then for any 0 < q < 1, the approximation error satisfies

$$||w(\tau) - w_k(\tau)|| \le 2 Q \tau ||A|| \frac{q^{k-1}}{1-q} e^{-\tau \min\{a,0\} - \tau \tilde{z}},$$
(5.4)

where Q = 11.08,

$$\tilde{z} = a - \frac{1}{\lambda} \int_0^{\frac{1}{2} \left(\frac{1}{q} - q\right)} \frac{\sqrt{m + s^2}}{\sqrt{1 + s^2}} ds,$$
(5.5)

the parameters m is determined by (4.3) from a, b, c of (5.1) and  $\lambda$  is the ratio in (4.3).

*Proof.* First note that  $H_k = V_k^T A V_k$  for an orthogonal  $V_k$ . Then

$$W(H_k) \subseteq W(A) \subseteq [a, b] \times [-c, c]$$

Now, Theorem 5.1 holds for  $h(t) = e_k^T e^{-tH_k} e_1$ , and indeed, from above and following the same proof, it holds with a, b, c in place of  $a_k, b_k, c_k$ . Namely,  $|h(t)| \leq 2Q \frac{q^{k-1}}{1-q} e^{-t\tilde{z}}$  with  $\tilde{z}$  defined as in (5.5) but from a, b, c. Now, using this bound in a posteriori error bound (3.6) in Theorem 3.1 and noting that  $h_{k+1,k} \leq ||A||_2$  (see (3.2)), we have that, if  $\tilde{z} \neq 0$ ,

$$\begin{aligned} ||w(\tau) - w_k(\tau)|| &\leq h_{k+1,k} e^{-\min\{\nu(A),0\}\tau} 2 Q \, \frac{q^{k-1}}{1-q} \int_0^\tau e^{-t\tilde{z}} dt \\ &\leq 2 Q \, ||A||_2 \frac{q^{k-1}}{1-q} e^{-\min\{a,0\}\tau} \frac{1-e^{-\tau\tilde{z}}}{\tilde{z}} \\ &= 2 Q \, ||A||_2 \frac{q^{k-1}}{1-q} e^{-\min\{a,0\}\tau} e^{-\tau\tilde{z}} \frac{e^{\tau\tilde{z}}-1}{\tilde{z}} \\ &\leq 2 Q \, \tau ||A||_2 \frac{q^{k-1}}{1-q} e^{-\tau\min\{a,0\}-\tau\tilde{z}} \end{aligned}$$

where we have used  $\frac{e^x-1}{x} \leq 1$  for any  $x \neq 0$ . If  $\tilde{z} = 0$ , the integration above gives  $\tau$  and the final bound holds for this case as well. So the theorem is proved.

For the rest of this section, we consider the case that A is positive definite (i.e. a > 0). In that case, the bound is simplified to

$$||w(\tau) - w_k(\tau)|| \le 2 Q \tau ||A|| \frac{q^{k-1}}{1-q} e^{-\tau \tilde{z}},$$
(5.6)

Bounding  $\tilde{z}$  of (5.5) using 0 < m < 1, we have

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

This leads to a simple but obviously crude bound. In particular, the bound can be further simplified by setting the exponent  $a - \frac{1}{2\lambda} \left(\frac{1}{q} - q\right)$  to 0, i.e.  $q = \frac{1}{\sqrt{a^2 \lambda^2 + 1} + a\lambda}$ . We state these as the following corollary.

**Corollary 5.3.** Under the the assumptions of Theorem 5.2 and that A is positive definite (i.e. a > 0), for any 0 < q < 1, the approximation error satisfies

$$||w(\tau) - w_k(\tau)|| \le 2Q\tau ||A|| \frac{q^{k-1}}{1-q} e^{-\tau \left\{a - \frac{1}{2\lambda}\left(\frac{1}{q} - q\right)\right\}}$$

In particular, for  $q = \frac{1}{\sqrt{a^2 \lambda^2 + 1} + a\lambda}$ , we have

$$||w(\tau) - w_k(\tau)|| \le 2Q\tau ||A|| \frac{q^{k-1}}{1-q},$$
(5.7)

*i.e.* the error converges at least at the rate of  $\frac{1}{\sqrt{a^2\lambda^2+1}+a\lambda}$ .

Note that  $a\lambda = 2a \frac{E'(m) - mK'(m)}{b-a} = 2 \frac{E'(m) - mK'(m)}{b/a-1}$ . Since *m* is a function of (b-a)/c (see Lemma 4.1) and b/a is the condition number of the Hermitian part of *A*, the bound relates the convergence to this condition number and the shape of the rectangle.

More generally, we can find  $q = q_0$  such that  $\tilde{z} = 0$ . Then (5.7) holds with this  $q_0$  and the error converges at the rate  $q_0$ . We call this  $q_0$  the threshold convergence rate. However, this  $q_0$  may not give the best bound possible among choices of q. Note that q influences the error bound through two opposing actions of  $q^k$  and  $e^{-\tau \tilde{z}}$ . Namely, choosing smaller q results in a faster geometrically decreasing term  $q^k$ , but  $e^{-\tau \tilde{z}}$  may be much larger to result in an overall larger bound. So the best choice of q should balance the two effects and will depend on k. For example, smaller q may be used for larger k so that the more significant decrease in  $q^k$  can offset the increase in  $e^{-\tau \tilde{z}}$ . This suggest a superlinear convergence behavior where, as k increases, the error is bounded with a smaller rate q.

In determining q to be used in the bound (5.6), we consider the minimization at each step k of

$$E(q) := \frac{q^{k-1}}{1-q} e^{-\tau \tilde{z}}.$$
(5.8)

Taking derivative of E with respect to q and using

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

we have

$$\frac{dE}{dq} = \frac{(k-1)q^{k-2}(1-q) - q^{k-1}(-1)}{(1-q)^2}e^{-\tau\tilde{z}} + \frac{q^{k-1}}{1-q}e^{-\tau\tilde{z}}(-\tau)\frac{d\tilde{z}}{dq}$$
$$= e^{-\tau\tilde{z}}\frac{q^{k-3}}{(1-q)^2}\left[(k-1)q + (2-k)q^2 - C(1-q)\sqrt{(1-q^2)^2 + 4mq^2}\right]$$

where  $C = \frac{\tau}{2\lambda}$ . Thus optimal q = q(k) can be found by solving

$$(k-1)q + (2-k)q^2 - C(1-q)\sqrt{(1-q^2)^2 + 4mq^2} = 0.$$
 (5.9)

Note that a solution  $q \in (0, 1)$  exists because the function in the equation is 1 when q = 1and -C < 0 when q = 0.

Finally, we discuss a special case, i.e.  $m \approx 0$ .

**Corollary 5.4.** Under the assumptions of Theorem 5.2, and  $m \approx 0$ , the approximation error satisfies

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

where

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

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

*Proof.* E' = E(1 - m) and K' = K(1 - m) are both functions of m and have the following expansions at m = 0 [1, 17.3.11-12, p. 591]

$$E' = E(m_1) = E(1-m) = 1 - \frac{1}{4}m\ln m + O(m)$$
(5.10)

$$K' = K(m_1) = K(1 - m) = -\frac{1}{2}\ln m + O(1)$$
(5.11)

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

$$E' - mK' = 1 + \frac{1}{4}m\ln m + O(m).$$
(5.12)

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

$$a\lambda = \frac{2}{\kappa - 1} \left( 1 + \frac{1}{4}m\ln m \right) + O(m).$$
 (5.13)

At the same time, for  $0 \le s \le \frac{1}{q} - q$ ,

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

 $\mathbf{SO}$ 

$$\int_{0}^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^{2}}}{\sqrt{1+s^{2}}} ds = \int_{0}^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{s}{\sqrt{1+s^{2}}} ds + O(\sqrt{m})$$
$$= \frac{1}{2}\left(\frac{1}{q}+q\right) - 1 + O(\sqrt{m}).$$
(5.14)

Let  $q = q_0$  be the unique solution of

$$a\lambda = \int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds,$$
(5.15)

where the existence of  $q_0$  and the uniqueness follow from the fact that the integral on the right is a function of q monotonically decreasing from  $\infty$  to 0 for 0 < q < 1. Using (5.13) and (5.14), the equation is written as

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

Solving this, the solution  $q_0$  with  $0 < q_0 < 1$  is

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

Using this  $q_0$  in the bound (5.4), we have  $\tilde{z} = 0$  and the theorem is proved.

Note that m is determined by  $\beta/\alpha$ . In particular, for  $m \approx 0$ , E(m) and K(m) have the expansions

$$E = E(m) = \frac{\pi}{2} - \frac{\pi}{8}m + O(m^2)$$
  

$$K = K(m) = \frac{\pi}{2} + \frac{\pi}{8}m + O(m^2).$$

We also have the expansion of E' - mK' in (5.12). Then

$$\frac{\beta}{\alpha} = \frac{E - m_1 K}{E' - mK'} = \frac{\pi}{2}m + O(m^2), \text{ or } c = \frac{(b - a)\pi}{4}m + O(m^2).$$

So the above theorem applies to the case when c/(b-a) is small or A is nearly Hermitian.

In an earlier paper [30], it is shown that for a symmetric positive definite matrix A, the approximation error satisfies

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

where  $\kappa = b/a$  is the condition number of the matrix A. This implies a conjugate gradient like convergence rate  $q = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$  regardless of the norm of the matrix. Then Theorem 5.4 shows that the same conclusion holds if A is nearly Hermitian.

### 6 A priori error bound for skew-Hermitian matrices

In this section, we consider the special case that A is skew-Hermitian which, as discussed in the introduction, arises in some interesting applications. We write A = -iH with H being an Hermitian matrix. In this case, the Arnoldi algorithm is theoretically equivalent to the Lanczos algorithm for H. As we will see, the error bounds for computing

$$w(\tau) := e^{i\tau H} v. \tag{6.1}$$

is also significantly simplified.

Applying k steps of the Lanczos method to H and  $v_1 = v$  with ||v|| = 1 (see [10]), we obtain an orthonormal basis  $\{v_1, v_2, \dots, v_k, v_{k+1}\}$  and a k-by-k tridiagonal matrix  $T_k$  such that

$$HV_{k} = V_{k}T_{k} + \beta_{k+1}v_{k+1}e_{k}^{T}, (6.2)$$

where  $V_k = [v_1, v_2, \dots, v_k]$ . This is equivalent to (3.1) for the Arnoldi algorithm for A = -iHwith  $H_k = -iT_k$  and  $h_{k+1,k} = \beta_{k+1}$ . Then, the corresponding approximation of  $w(\tau)$  is

$$w_k(\tau) := V_k e^{i\tau T_k} e_1, \tag{6.3}$$

which we call the Lanczos approximation. Then the same *a posteriori* error bound of Theorem 3.1 holds with  $h_{k+1,k} = \beta_{k+1}$  and  $h(t) := e_k^T e^{itT_k} e_1$ . Namely,

$$||w(\tau) - w_k(\tau)|| \le \beta_{k+1} \int_0^\tau |h(t)| dt \le ||H|| \int_0^\tau |h(t)| dt$$
(6.4)

Furthermore, slightly better bounds may be obtained by shifting the matrix. Specifically, for any  $\alpha \in \mathbb{R}$ , we can consider the shifted matrix  $H - \alpha I$  and correspondingly  $w(\tau, \alpha) := e^{i\tau(H-\alpha I)}v = e^{-i\tau\alpha}w(\tau)$  and  $w_k(\tau, \alpha) := V_k e^{i\tau(T_k-\alpha I)}e_1 = e^{-i\tau\alpha}w_k(\tau)$ . Since  $(H - \alpha I)V_k = V_k(T_k - \alpha I) + \beta_{k+1}v_{k+1}e_k^T$ , we can apply (6.4) to  $H - \alpha I$  to get

$$|w(\tau,\alpha) - w_k(\tau,\alpha)|| \le ||H - \alpha I|| \int_0^\tau |h(t,a)| dt$$

where  $h(t, \alpha) := e_k^T e^{it(T_k - \alpha I)} e_1 = e^{-it\alpha} h(t)$ . Thus

$$||w(\tau) - w_k(\tau)|| = ||w(\tau, \alpha) - w_k(\tau, \alpha)|| \le ||H - \alpha I|| \int_0^\tau |h(t)| dt.$$
(6.5)

We now bound h(t) as in the previous section to obtain the following a priori error bound.

**Theorem 6.1.** Let  $A = -iH \in C^{n \times n}$  be a skew-Hermitian matrix and  $v \in \mathbb{C}^n$  with ||v|| = 1. Then, for any q with 0 < q < 1, the error of the Lanczos approximation  $w_k(\tau) = V_k e^{i\tau T_k} e_1$ (6.3) satisfies

$$||w(\tau) - w_k(\tau)|| \le \frac{4\min\{1/(1-q^2), \tau\rho/q\}}{1-q} q^k e^{\tau\rho\left(\frac{1}{q}-q\right)},\tag{6.6}$$

where  $\rho = (\lambda_{\max}(H) - \lambda_{\min}(H))/4$  with  $\lambda_{\min}(H)$  and  $\lambda_{\max}(H)$  being the smallest and the largest eigenvalues of H respectively.

Proof. Let  $a = \lambda_{\min}(H)$  and  $b = \lambda_{\max}(H)$ . We first bound  $h(t) := e_k^T e^{itT_k} e_1$  as in Theorem 5.1 by constructing a conformal map and using the Faber polynomial approximation. Let  $\Phi := \phi_3 \circ \phi_2 \circ \phi_1$  where  $z_1 = \phi_1(z) = -iz$  maps the exterior of  $E := \{i\lambda : \lambda \in [a, b]\}$  to the exterior of  $[a, b], z_2 = \phi_2(z_1) = \frac{2}{b-a} \left(z_1 - \frac{a+b}{2}\right)$  maps the exterior of [a, b] to the exterior of  $[-1, 1], w = \phi_3(z_2) = i(z_2 + \sqrt{z_2^2 - 1})$  maps the exterior of [-1, 1] to  $\{|w| > 1\}$ . In the definition of  $\phi_3$ , we choose the branch of  $\sqrt{z^2 - 1}$  such that  $\lim_{z \to \infty} \frac{\sqrt{z^2 - 1}}{z} = 1$ . Then  $\Phi$  maps the exterior of E to the exterior of the unit circle  $\{|w| = 1\}$  with  $\rho := \lim_{z \to \infty} \frac{z}{\Phi(z)} = \frac{b-a}{4}$ . Construct the Faber polynomials from this conformal map  $\Phi$  and the Faber polynomial

approximation  $\Pi_{k-2}$  of  $f(z) := e^{tz}$  as defined in (2.3). Let  $r := \frac{1}{q} > 1$  and let  $C_r$  be the inverse image under  $\Phi$  of the circle |w| = r. Applying Theorem 2.1 or (2.5), the approximation error in  $I(C_r)$  is bounded as,

$$||f - \Pi_{k-2}||_{\infty} \le 2M(r) \frac{(\frac{1}{r})^{k-1}}{1 - \frac{1}{r}} = 2M(r) \frac{q^{k-1}}{1 - q},$$

where  $M(r) = \max_{z \in C_r} |f(z)|$  and we note that the total rotation of E (a line segment) is  $V = 2\pi$ . To find M(r) for any  $z \in C$ , we write  $z = \Phi^{-1}(w)$  with  $w = \pi e^{i\theta}$  where  $\theta \in [0, 2\pi)$ .

To find M(r), for any  $z \in C_r$ , we write  $z = \Phi^{-1}(w)$  with  $w = re^{i\theta}$  where  $\theta \in [0, 2\pi)$ . Then, it follows from the definition of  $\Phi$  that

$$z_{2} = \frac{1}{2} \left( -iw + \frac{1}{-iw} \right) = \frac{1}{2} \left( -i\frac{e^{i\theta}}{q} + \frac{iq}{e^{i\theta}} \right) = -\frac{i}{2} \left[ \left( \frac{1}{q} - q \right) \cos \theta + i \left( \frac{1}{q} + q \right) \sin \theta \right],$$
  

$$z_{1} = \frac{b-a}{2} z_{2} + \frac{b+a}{2} = \left[ \frac{b-a}{4} \left( \frac{1}{q} + q \right) \sin \theta + \frac{b+a}{2} \right] - i \left[ \frac{b-a}{4} \left( \frac{1}{q} - q \right) \cos \theta \right],$$
  

$$z = iz_{1} = \frac{b-a}{4} \left( \frac{1}{q} - q \right) \cos \theta + i \left[ \frac{b-a}{4} \left( \frac{1}{q} + q \right) \sin \theta + \frac{b+a}{2} \right].$$

Thus

$$M(r) = \max_{z \in C_r} |e^{tz}| = \max_{z \in C_r} e^{t \operatorname{Re}(z)} = e^{\frac{t(b-a)}{4} \left(\frac{1}{q} - q\right)}.$$

Now, let  $\lambda_j$   $(1 \leq j \leq n)$  be the eigenvalues of  $iT_k$ . Then  $\lambda_j \subset E$ . As in the proof of Theorem 5.1, we have

$$\begin{aligned} |h(t)| &= |[f(iT_k)]_{k1}| = |[f(iT_k)]_{k1} - [\Pi_{k-2}(iT_k)]_{k1}| \\ &\leq ||f(iT_k) - \Pi_{k-2}(iT_k)||_2 = \max_j |f(\lambda_j) - \Pi_{k-2}(\lambda_j)| \\ &\leq \max_{z \in E} |f(z) - \Pi_{k-2}(z)| \leq ||f - \Pi_{k-2}||_{\infty} \\ &\leq \frac{2q^{k-1}}{1-q} e^{\frac{t(b-a)}{4}(\frac{1}{q}-q)}. \end{aligned}$$

Finally, using (6.5) with  $\alpha = (a+b)/2$ , we have  $||H - \alpha I|| = (b-a)/2$  and hence

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

where we have used  $e^x - 1 \le \min\{1, x\}e^x$  for any  $x \ge 0$ .

As before, we have an error bound for any given  $q \in (0, 1)$ . Using smaller q results in a faster geometrically decreasing term  $q^k$ , but  $e^{\tau \rho(\frac{1}{q}-q)}$  is expected to be larger. So, again, we study the value of q that minimizes the bound

$$E(q) := \frac{q^k}{(1-q)(1-q^2)} e^{\tau \rho \left(\frac{1}{q}-q\right)},\tag{6.7}$$

Taking derivative of E(q) with respect to q to get

$$\frac{dE}{dq} = \frac{q^{k-2}e^{\tau\rho\left(\frac{1}{q}-q\right)}}{(1-q)^3(1+q)^2} \left[\tau\rho q^4 + (3-k)q^3 + q^2 + kq - \tau\rho\right].$$

With  $E(q) \to \infty$  as  $q \to 0$  or 1, the optimal value  $q_0 = q_0(k)$  that minimizes E(q) is given by the solution of the equation

$$\tau \rho q^4 + (3-k)q^3 + q^2 + kq - \tau \rho = 0.$$

Note that it can be shown that the above equation has a unique solution  $q_0 \in (0, 1)$  (see [29] for details).

Note that  $\frac{1}{1-q}$  in E(q) is a well bounded term unless  $q \approx 1$ . For example, it is bounded by 10 if  $q \leq 0.9$ . To quantitatively interpret the bound, we can consider minimization of

$$E_s(q) = q^k e^{\tau \rho\left(\frac{1}{q} - q\right)},\tag{6.8}$$

which is essentially the same as E(q) unless  $q \approx 1$ . Differentiate  $E_s$  to get

$$\frac{dE_s}{dq} = e^{\tau \rho \left(\frac{1}{q} - q\right)} q^{k-2} \left[ -\tau \rho q^2 + kq - \tau \rho \right].$$

The discriminant of the quadratic  $-\tau \rho q^2 + kq - \tau \rho$  is  $\Delta = k^2 - 4(\tau \rho)^2$ . So, if  $k \leq 2\tau \rho$ ,  $E_s(q)$  is monotonically decreasing with the minimum occurring at  $q_0 = 1$ . If  $k > 2\tau \rho$ ,  $E_s(q)$  is minimized at  $q_0 = \frac{k - \sqrt{k^2 - 4(\tau \rho)^2}}{2\tau \rho} < 1$ . Thus, the bound implies different convergence behavior at two stages of the Lanczos iterations.

- 1. When  $1 \le k \le 2\tau\rho$ , there is essentially no decrease in the error bound.
- 2. For  $k > 2\tau\rho$ , the error bounds for subsequent steps decrease at least at the rate of  $q_0$ .

The convergence behavior as implied from this theory is indeed what has been observed in the numerical examples (see §7), where the error initially stagnates for approximately  $2\tau\rho$ steps and then begins to decrease superlinearly. Thus our bound qualitatively explains this convergence property observed numerically.

Finally, we note that the convergence bound for skew-Hermitian matrices have also been studied by Hochbruch and Lubich [18, Theorem 4]. It is proved there that for  $k \ge 2\rho\tau$ ,

$$||w(\tau) - w_k(\tau)|| \le 12e^{\frac{-(\rho\tau)^2}{k}} \left(\frac{e\rho\tau}{k}\right)^k.$$
 (6.9)

Interestingly, the range of validity of the bound coincides with the point of initial convergence as implied by our bound. It turns out that this bound can be implied from a special case of our error bound (6.6). For  $k \ge 2\rho\tau$ , let  $q = \frac{\tau\rho}{k} \le \frac{1}{2}$ . Then our bound (6.6), simply using  $1/(1-q^2)$  for the minimum, reduces to (6.9) as follows:

$$||w(\tau) - w_k(\tau)|| \le \frac{4\left(\frac{\tau\rho}{k}\right)^k}{(1 - \frac{1}{2})(1 - \frac{1}{2})^2} e^{\tau\rho\left(\frac{k}{\tau\rho} - \frac{\tau\rho}{k}\right)} = \frac{32}{3} e^{-\frac{(\tau\rho)^2}{k}} \left(\frac{e\tau\rho}{k}\right)^k \le 12 e^{-\frac{(\tau\rho)^2}{k}} \left(\frac{e\tau\rho}{k}\right)^k$$

#### 7 Numerical examples

In this section, we present several numerical examples to demonstrate the error bounds obtained in this paper. All tests were carried out on a PC in MATLAB (R2013b) with the machine precision  $\approx 2e - 16$ . The Jacobi elliptic integrals that are needed for our bounds are computed using MATLAB built-in functions ellipticK and ellipticE.

We will construct several testing matrices with different spectral distributions and compare the actual approximation error with the new *a posterior* error estimate (3.6) and *a priori* bounds (5.4) or (6.6). The integral in the *a posterior* error estimate (3.6) is approximated using Simpson's rule with 10 subintervals on  $[0, \tau]$ .

We shall compare our bounds with the bounds by Saad [27] and where applicable with those of Hochbruck and Lubich [18] as well. For example, if the matrices are positive semidefinite, we consider the following bound of Saad [27, Cor. 2.2]:

$$||w(\tau) - w_k(\tau)|| \le \frac{2}{k!} (\tau ||A||)^k.$$
(7.1)

and the following bound of Hochbruck and Lubich [18, Theorem 2]:

$$||e^{-\tau A}v - V_k e^{-\tau H_k} e_1|| \le 12e^{-\rho\tau} \left(\frac{e\rho\tau}{k}\right)^k,$$
 (7.2)

which holds for  $k \ge 2\rho\tau$  and with the assumption that the field of values W(A) is contained in the disk  $|z - \rho| < \rho$ .

*Example 1.* Given an odd integer N and a rectangle  $[a, b] \times [-c, c]$  in the complex plane where a, b and c are all positive real numbers, let A be the  $N^2 \times N^2$  block diagonal matrix with the diagonal blocks being  $2 \times 2$  matrices  $B_{\ell,j}$  for  $\ell = 1, 2, \dots, N$  and  $j = 1, 2, \dots, \frac{N-1}{2}$ , where

$$B_{\ell,j} = \begin{bmatrix} x_{\ell} & y_j \\ -y_j & x_{\ell} \end{bmatrix}, \ x_{\ell} = a + \frac{(\ell - 1)(b - a)}{N - 1} \quad \text{and} \quad y_j = \frac{2jc}{N - 1}.$$

Then, the eigenvalues of A are  $x_l \pm iy_j$  with i being the imaginary unit, which are the grid points of the  $N \times N$  lattice on  $[a, b] \times [-c, c]$ . Clearly, A is a normal matrix, so the field of values of A is the convex hull of its eigenvalues, i.e., the rectangle  $[a, b] \times [-c, c]$ .

The primary purpose of this numerical test is to compare our *a priori* bound with Hochbruch and Lubich's bound (7.2). The latter is applicable when W(A) is contained in a disk  $|z-\rho| < \rho$ . We therefore choose  $[a,b] \times [-c,c]$  to be the square  $[1-\frac{\sqrt{2}}{2},1+\frac{\sqrt{2}}{2}] \times [-\frac{\sqrt{2}}{2},\frac{\sqrt{2}}{2}]$ 

which is enclosed in the circle |z - 1| < 1 and construct a matrix A as above such that the eigenvalues of A form a  $31 \times 31$  lattice in the square. We apply the Arnoldi method to compute  $e^{-\tau A}v$  where v is a random normalized vector and we use  $\tau = 10, 20, 30, 40$ . In Figure 1, we plot against the iteration number the actual error  $||w(\tau) - w_k(\tau)||$  in the solid line, the *a posteriori* error estimate (3.6) in the +-line, our *a priori* bound (5.4) in the dashed line, Hochbruch and Lubich's bound (7.2) in the dotted line, and Saad's bound (7.1) in the x-line. Note that Hochbruch and Lubich's bound is only valid for  $k \ge 2\rho\tau$ .



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

We observe that when  $\tau$  is relatively small, our new *a priori* bound is comparable to Hochbruch and Lubich's bound, but as  $\tau$  increases, our bound improves significantly. In particular, for larger  $\tau$  values, the error  $||w(\tau) - w_k(\tau)||$  first stagnates for certain number of iterations before it starts to converge. Our *a priori* bound nicely captures this behavior and the point where the convergence begins, while Hochbruch and Lubich's bound is pessimistic and is applicable to iterations long after the initial point of convergence. Our *a posteriori* error estimate is sharp at the convergence stage for all tests.

In the next example, we use the same construction as in Example 1, but consider the field of values contained in rectangles of different shape. This is to investigate the influence

on the convergence rate by the shape of the rectangle through the parameter m in (4.3).

Example 2. For a given parameter  $m \in (0,1)$ , we determine the dimensions of the rectangle  $\alpha$  and  $\beta$  by  $\alpha = E' - mK'$ ,  $\beta = E - m_1K$ . We then construct a matrix as in Example 1 whose field of values is contained in the rectangle  $[0, 2\alpha] \times [-\beta, \beta]$ . We use  $m \in \{0.01, 0.1, 0.9, 0.99\}$  whose corresponding values of  $\alpha, \beta$  are listed in Figure 2. Note from Section 3.3 that  $m \approx 0$  means that the matrix is close to being Hermitian, and that  $m \approx 1$  means the matrix is close to being skew-Hermitian with a real spectral shift. We apply the Arnoldi method to compute  $e^{-\tau A}v$  for a random normalized vector v and we use  $\tau = 30$  to give  $\tau A$  a moderate norm. In Figure 2 we plot the error  $||w(\tau) - w_k(\tau)||$  in the solid line, our a posteriori error estimate (3.6) in +-line, our a priori bound (5.4) in the dashed line and Saad's bound (7.1) in the x-line.



Figure 2: Example 2. m = 0.01, 0.1 (top) and m = 0.9, 0.99 (bottom). Error(solid), our *a posteriori* bound (+), our *a priori* bound (dashed), Saad's bound (x).

Figure 2 shows that the convergence is related to m. For smaller m when the eigenvalues lie close to the real axis, the convergence occurs at early iterations and at a faster rate. As m increases to 1, the convergence has an initial stagnation stage before the convergence occurs. Again, this behavior is captured in our new *a priori* bound. Our new bound also significantly improves Saad's, which is based on the norm of the matrix only. Our *a posteriori* error estimate is sharp for all tests.

We further demonstrate our new bounds for non-positive definite matrices. We construct as in Example 1 a matrix A whose field of values is contained in the square  $[\sigma, 2+\sigma] \times [-1, 1]$ with  $\sigma = -1$  and -10. We plot in Figure 3 the actual error (solid), a posteriori bound (+), a priori bound (dashed) and Saad's bound (x). We see that our bounds are still valid when A is not positive definite. They also demonstrate the initial stagnation of convergence. However, the bound becomes more pessimistic for larger  $\sigma$ .



Figure 3: Example 2. Non-positive definite matrix with negative  $\nu(A)$ . Error (solid), our *a* posteriori bound (+), our *a priori* bound (dashed), and Saad's bound (x).

In the next example, we consider matrices arising in the convection diffusion equation

$$\frac{\partial}{\partial t}u(x,y) = \Delta u(x,y) - u_x(x,y) - u_y(x,y), \quad u = 0 \quad \text{in} \quad \partial\Omega \tag{7.3}$$

where  $(x, y) \in \Omega = [0, 1]^2$ . The finite-difference discretization in x, y with a uniform mesh leads to an initial value problem (1.1) and hence the problem of computing  $w(\tau) = e^{-\tau A}v$ .

Example 3. Let -A be the finite-difference discretization of (7.3) in a  $20 \times 20$  grid in  $[0,1]^2$  scaled with  $h^2$  so that  $||A||_2 \approx 8$ . Then A is non-Hermitian but positive definite. 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  $\tau = 2, 10, 20, 50$  and apply the Arnoldi method to A and v and the results are presented in Figure 4 with  $||w(\tau) - w_k(\tau)||$  in the solid line, our *a posteriori* error estimate (3.6) in the +-line, our *a priori* bound (5.4) in the dashed line and Saad's bound (7.1) 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 modestly large values of  $\tau$ , Saad's bound becomes very pessimistic due to the large norm of  $\tau A$ , while our *a priori* bound still follows the convergence curve of the error. For the case when  $\tau = 50$  ( $\tau ||A||_2 \approx 400$ ) or larger, our *a priori* bound also becomes very pessimistic. In all the cases, our *a posteriori* error estimate remains sharp.

Our final example concerns skew-Hermitian matrices.

Example 4. Let H 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  $||H|||_2 = 1$  and the spectral gap  $4\rho = \lambda_{\max}(H) - \lambda_{\min}(H)$  is approximately 1. We apply k iterations of the Lanczos method to



Figure 4: Example 3.  $\tau = 2, 10, 20, 50$ . Error(solid), a posteriori bound (+), a priori bound (dashed), Saad's bound (x).

compute  $w(\tau) = e^{i\tau H}v$ . We will test n = 1000 with  $\tau = 2, 10, 20, 50$  and the results are presented in Figure 5 with  $||w(\tau) - w_k(\tau)||$  in the solid line, our *a posteriori* error estimate (6.4) in the +-line, our *a priori* bound (5.4) in the dashed line, Hochbruch and Lubich's bound (6.9) in the dotted line, and Saad's bound in the x-line.

We first observe that our bound only improves Hochbruch and Lubich's bound very slightly. It is significantly better than Saad's bound when  $\tau$  is large. In all cases, our and Hochbruch and Lubich's bound follow the actual error quite closely and our *a posteriori* error estimate is sharp. In addition, for larger  $\tau$ , the error typically stagnates first for some iterations before it starts to converge. An analysis of our bound has shown that the convergence may be expected to start at  $k = 2\tau\rho$ . For  $\tau = 2, 10, 20, 50$ , the corresponding k is 1, 5, 10 and 25, respectively. This basically matches the actual convergence curve in Figure 5, especially when  $\tau$  is relatively large and more iterations are needed for the convergence.

### 8 Concluding remarks

For the computation of  $e^{-\tau A}v$  with a non-Hermitian matrix A by the Krylov subspace methods, we have presented an *a posteriori* error bound that provides a sharp estimate of the



Figure 5: Example 4.  $1000 \times 1000$  diagonal matrix with  $a_{jj} = j/1000$ .  $\tau = 2, 10, 20, 50$ . Error (solid), *a posteriori* bound (+), *a priori* bound (dashed), Hochbruch and Lubich's bound (dotted), and Saad's bound (x).

error. We have also derive new *a priori* error bounds based on the largest and the smallest eigenvalues of the Hermitian and the skew-Hermitian parts of A. Using this simple spectral information, our bounds capture convergence characteristics of the Krylov subspace methods. They also explain often observed initial stagnation of the convergence curve. Numerical comparisons with existing bounds also show that our new bounds may significantly improve the *a priori* bound by Hochbruch and Lubich [18] that is based on a circular enclosing region of the field of values and the one by Saad [27] that is based on the norm. Finally, it agrees with the bound [30] for the symmetric positive definite case.

The technique developed in this paper provides a new way to analyze convergence of the Krylov subspace method for non-Hermitian matrices through the bounding rectangle for the field of values. It may be extended to other linear algebra problems. For the future works, we plan to study convergence bounds for linear systems based on the Hermitian and the skew-Hermitian parts of A, which may also add to the theory of the Krylov subspace method for linear systems.

Acknowledgement: We would like to thank Prof. Michele Benzi for many valuable discussions and in particular for his suggestion to use the technique in [3] that has turned out to be very fruitful.

#### References

- M. Abramowitz and I. A. Stegun, Handbook of Mathematical Functions, Dover Publications INC., 1965.
- [2] B. Beckermann and L. Reichel, Error estimates and evaluation of matrix functions via the Faber transform, SIAM J. Numer. Anal., 47 (2009), pp. 3849-3883.
- [3] M. Benzi and P. Boito, Decay Properties for Functions of Matrices over C\*-Algebras, Linear Alg. Appl., 456 (2014), pp. 174-198.
- [4] M. Benzi P. Boito and N. Razouk, Decay Properties of Spectral Projectors with Applications to Electronic Structure, SIAM Review, 55 (2013), pp. 3-64.
- [5] M. Benzi and G. H. Golub, Bounds for the entries of matrix functions with applications to preconditioning, BIT, 39 (1999), pp. 417-438.
- [6] M. Benzi and V. Simoncini, Decay Bounds for Functions of Hermitian Matrices with Banded or Kronecker Structure, SIAM J. Matrix Anal. Appl., 36 (2015), pp. 1263-1282.
- [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] M. Crouzeix, Numerical range and functional calculus in Hilbert space, Journal of Functional Analysis, 244 (2007), pp. 668-690.
- [9] 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.
- [10] J. Demmel, Applied Numerical Linear Algebra, SIAM, Philadelphia, 1997.
- [11] 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.
- [12] 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.
- [13] S. W. Ellacott, Computation of Faber series with application to numerical polynomial approximation in the complex plane, Math. Comp., 40 (1983), pp. 575-587.
- [14] E. Gallopoulos and Y. Saad, Efficient solution of parabolic equations by Krylov approximation methods, SIAM J. Sci. Statist. Comput., 13 (1992), pp. 1236-1264.

- [15] X. Guan, O. Zatsarinny, K. Bartschat, B.I. Schneider, J. Feist, and C.J. Noble, A general approach to few-cycle intense laser interactions with complex atoms, Phys. Rev. A 76 (2007), 053411.
- [16] M. Ilic, I. W. Turner, and V. Anh, Numerical solution of the fractional poisson equations using an adaptively preconditioned Lanczos methods, Journal of Applied Mathematics and Stochastic Analysis, vol. 2008, Article ID 104525, 2008. doi:10.1155/2008/104525.
- [17] L.A. Knizhnerman, Calculation of functions of unsymmetric matrices using Arnoldi's method, Comput. Math. and Math. Phys., 31 (1991), pp. 1-9.
- [18] M. Hochbruck and C. Lubich, On Krylov subspace approximations to the matrix exponential operator, SIAM J. Numer. Anal., 34 (1997), pp. 1911-1925.
- [19] H. Kober, Dictionary of Conformal Representations, Dover Publications INC., 1957.
- [20] 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.
- [21] L. M. Milne-Thomson, Jacobian Elliptic Function Tables, Dover Publications INC., 1950.
- [22] 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.
- [23] I. Moret and P. Novati, On the convergence of Krylov subspace methods for matrix Mittaga-Leffler functions, SIAM J. Numer. Anal., 49 (2011), pp. 2144-2164.
- [24] A. Nauts and R. Wyatt, New approach to many state quantum dynamics: The recurisive residue generation method, Phys. Rev. Lett., 51(1983), pp. 2238-2241.
- [25] T.J. Park and J.C. Light, Unitary quantum time evolution by iterative Lanczos reduction
   J. Chem. Phys. 85 (1986) 5870.
- B.I. Schneider and L.A. Collins, The discrete variable method for the solution of the time-dependent Schrödinger equation
   J. Non-Cryst. Solids 351 (2005) 1551.
- [27] Y. Saad, Analysis of some Krylov subspace approximations to the matrix exponential operator, SIAM J. Numer. Anal., 29 (1992), pp. 209-228.
- [28] G. Söderlind, The logarithmic norm. History and modern theory, BIT Numerical Mathematics, 46 (2006), pp. 631-652.
- [29] H. Wang, The Krylov Subspace Methods for the Computation of Matrix Exponentials, Ph.D. Thesis, Department of Mathematics, University of Kentucky, 2015.

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