SIAM J. MATRIX ANAL. APPL.
Vol. 38, No. 1, pp. 155–187

ERROR BOUNDS FOR THE KRYLOV SUBSPACE METHODS FOR COMPUTATIONS OF MATRIX EXPONENTIALS*

HAO WANG† AND QIANG YE‡

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 $\lambda_{\min}(\frac{A+A^*}{2})$, $\lambda_{\max}(\frac{A+A^*}{2})$ (the smallest and the largest eigenvalue of the Hermitian part of A), and $|\lambda_{\max}(\frac{A-A^*}{2})|$ (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 convergence behavior where the error may first stagnate for a certain number of 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.

Key words. matrix exponential, Krylov subspace method, Arnoldi method, Lanczos method

AMS subject classifications. 15A18, 65F15, 62B10

DOI. 10.1137/16M1063733

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 the reader to the classical paper [23] 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]

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

See [12, 17, 25] 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 involves a skew-Hermitian matrix A (i.e., $A = iH$ with H being Hermitian), for which (1.1) has a norm-conserving 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, 26, 28]. 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

*Received by the editors March 1, 2016; accepted for publication (in revised form) by V. Simoncini November 14, 2016; published electronically February 22, 2017.
http://www.siam.org/journals/simax/38-1/M106373.html
Funding: The work of the second author was supported in part by NSF under grants DMS-1317424, DMS-1318633, and DMS-1620082.
†Department of Biomedical Engineering, University of Kentucky, Lexington, KY 40506 (hao.wang@uky.edu).
‡Department of Mathematics, University of Kentucky, Lexington, KY 40506 (qiang.ye@uky.edu).

solving many large-scale linear algebra problems. For the matrix exponential problem

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

where $\tau \in \mathbb{R}$ is a fixed parameter typically representing a time step in applications, the Lanczos/Arnoldi approximations, introduced by Gallopoulos and Saad (see [14, 27]), have become very popular methods. Some earlier works on other Krylov subspace methods for matrix functions including the exponential are reviewed in [12]; see the references cited there. A comprehensive theory has been developed in the literature with error bounds demonstrating convergence of the approximation and its relation to certain properties of the matrix. For example, earlier results in [14, 27] relate convergence of the Lanczos/Arnoldi methods to the norm of the matrix τA. More refined error bounds have since been derived, which 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, 12, 11, 16, 18, 24, 27] and the references contained therein. For a real symmetric positive definite matrix A, it has been shown in a recent work [31] that the speed of convergence is also related to 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 [12, 16, 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 predefined region, and are not easy to apply or interpret. In general, there is an inherited theoretical difficulty in quantitatively characterizing the influence on the convergence by the field of values, a two-dimensional object. In [2], Beckermann and Reichel have derived a sharp convergent bound with an exponential factor determined from the inverse of a conformal mapping from the exterior of a region enclosing the field of values to the exterior of the unit disk in the complex plane. Although this reduces the dependence of convergence on the field of values to a single value as determined by the conformal map, this connection through the conformal map to be constructed is still indirect and not easy to interpret.

In this paper, we extend the previous works, [2] in particular, by relating the convergence of the Krylov subspace methods to the field of values through its bounding rectangle $[a, b] \times [-c, c]$, where $a = \lambda_{\min}(\frac{A+A^*}{2})$, $b = \lambda_{\max}(\frac{A+A^*}{2})$ (the smallest and the largest eigenvalue of the Hermitian part of A), and $c = |\lambda_{\max}(\frac{A-A^*}{2})|$ (the largest eigenvalue in absolute value of the skew-Hermitian part of A). With this approach, we will derive new a priori error bounds in terms of a, b, and c that relate the speed of convergence to the size and the shape of the rectangle. In particular, our bounds explain an interesting observed convergence behavior where the error may first stagnate for a certain number of iterations before it starts to converge. Simplified bounds will be presented for non-Hermitian positive definite matrices and skew-Hermitian matrices. 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. Our new a priori error bounds are derived from that of Beckermann and Reichel [2], as well as from our a posteriori bound combined with some decay bounds on the exponential of a Hessenberg matrix derived using the same technique as in the literature [2, 3, 6, 16, 18] by constructing Faber polynomial approximations of the exponential function in a region containing the field of values. The novelty in this work is using the Jacobi elliptic functions to construct a conformal mapping for the rectangular region that

tightly encloses the field of values and showing that this highly complicated mapping can be simplified to yield some simple and interesting 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 the exponential of a Hessenberg matrix. 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 to skew-Hermitian matrices and derive simpler a priori bounds. Numerical examples are presented in section 7, with some concluding remarks in section 8.

Throughout this paper, $\|\cdot\|$ denotes the 2-norm $\|\cdot\|_2$ unless otherwise stated. We will also assume throughout that v in (1.2) is normalized such that $\|v\| = 1$.

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. This starts with the Riemann mapping theorem [21, Theorem 1.2] that states that every simply 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 $\bar{\mathbb{C}} = \mathbb{C} \cup \{\infty\}$ be the extended complex plane, and let D be a bounded continuum in the complex plane with boundary Γ such that the complement of D is a simply connected domain in the extended plane and contains the point at ∞. A continuum is a nonempty, 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

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

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

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

at infinity [21, p. 104]. Then, we call the polynomial containing nonnegative powers of z in the expansion,

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

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 $|w| \le \rho$. Let Ψ be the inverse of Φ, and let C_R be the image under Ψ of the circle $|w| = R > \rho$. We denote by $I(C_R)$ the bounded region enclosed by C_R. By [21, 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

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

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,

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

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 Π_N. We first need to introduce the definition of total rotation of the boundary. For this, we assume that D is a closed Jordan region, i.e., its boundary Γ 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 Γ has bounded total rotation V if $V = \int_\Gamma |d\Theta(z)| < \infty$. Then $V \geq 2\pi$ and the equality holds if D is convex; see [13].

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

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

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 Φ 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 $\widehat{\Phi}$ 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 $\widehat{\Phi}$ through a simple normalization transformation. Namely, given $\widehat{\Phi}$, let $\rho = \lim_{z\to\infty} \frac{z}{\widehat{\Phi}(z)}$ and $\Phi(z) := \rho\widehat{\Phi}(z)$. Then Φ satisfies the normalization condition (2.1), but now maps the exterior of D onto the exterior of the disk $|w| = \rho$. Applying Theorem 2.1 to Φ, (2.4) holds for any $R > \rho$. Let $r := R/\rho > 1$. Let C_R be the inverse image under Φ of the circle $|w| = R$, and let $\widehat{C}_r$ be the inverse image under $\widehat{\Phi}$ of the circle $|w| = r$. It is easy to check that $C_R = \widehat{C}_r$, and then $M(R) := \max_{z\in C_R} |f(z)| = \max_{z\in \widehat{C}_r} |f(z)|$. Thus, (2.4) is reduced to

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

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 ρ as defined in the two normalizations is invariant and is called the 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

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

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 notation:

$$\begin{aligned} K &:= K(m), \quad K' := K(m_1) = K(1 - m); \\ E &:= E(m), \quad E' := E(m_1) = E(1 - m). \end{aligned} \tag{2.7}$$

We now introduce the Jacobi elliptic functions. There are a total of twelve Jacobi elliptic functions in the family, but we will discuss only the basic three 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

$$\begin{aligned} \mathrm{sn}(u|m) &:= \sin \phi, \\ \mathrm{cn}(u|m) &:= \cos \phi, \\ \mathrm{dn}(u|m) &:= \sqrt{1 - m \sin^2 \phi}. \end{aligned} \tag{2.8}$$

The notation $\mathrm{sn}(u|m)$, $\mathrm{cn}(u|m)$, and $\mathrm{dn}(u|m)$ indicates 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 $m \in (0, 1)$, $\mathrm{sn}(u) := \mathrm{sn}(u|m)$, $\mathrm{cn}(u) := \mathrm{cn}(u|m)$, and $\mathrm{dn}(u) := \mathrm{dn}(u|m)$ are doubly periodical meromorphic functions defined on $u \in \mathbb{C}$ [22, 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, 20, 22].

PROPOSITION 2.2. *The following properties hold:*

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*

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

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

2. *(See* [1, p. 591]*.)*

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

3. *(See* [1, p. 592]*.)*

$$E(u+2iK')=E(u)+2i(K'-E'). \tag{2.12}$$

4. sn, cn, *and* dn *satisfy*

$$\begin{aligned}\operatorname{sn}^2(u|m)+\operatorname{cn}^2(u|m)&=1,\\ m\cdot\operatorname{sn}^2(u|m)+\operatorname{dn}^2(u|m)&=1.\end{aligned}$$

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

$$\begin{aligned}\operatorname{sn}(u+2lK+2niK'|m)&=(-1)^l\operatorname{sn}(u|m),\\ \operatorname{cn}(u+2lK+2niK'|m)&=(-1)^{l+n}\operatorname{cn}(u|m),\\ \operatorname{dn}(u+2lK+2niK'|m)&=(-1)^n\operatorname{dn}(u|m).\end{aligned}$$

6. *(See* [1, Table 16.8, p. 572]*.)*

$$\begin{aligned}\operatorname{sn}(2iK'-\sigma|m)&=\operatorname{sn}(-\sigma|m)=-\operatorname{sn}(\sigma|m),\\ \operatorname{cn}(2iK'-\sigma|m)&=-\operatorname{cn}(-\sigma|m)=-\operatorname{cn}(\sigma|m),\\ \operatorname{dn}(2iK'-\sigma|m)&=-\operatorname{dn}(-\sigma|m)=-\operatorname{dn}(\sigma|m).\end{aligned} \tag{2.13}$$

7. *(See* [1, Table 16.16, p. 574]*.)*

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

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

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

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

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

Then

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

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

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

We will also need to use the signs of the real and imaginary parts of $\operatorname{sn}(u|m)$, $\operatorname{cn}(u|m)$, and $\operatorname{dn}(u|m)$ 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 [20, pp. 172–176], and we summarize it in Tables 1–3 for easy reference.

TABLE 1
Signs of $(\mathrm{Re}(\mathrm{sn}(u|m)), \mathrm{Im}(\mathrm{sn}(u|m)))$.

Im(u) \ Re(u)	$(-K, 0)$	$(0, K)$
$(K', 2K')$	$(-, -)$	$(+, -)$
$(0, K')$	$(-, +)$	$(+, +)$

TABLE 2
Signs of $(\mathrm{Re}(\mathrm{cn}(u|m)), \mathrm{Im}(\mathrm{cn}(u|m)))$.

Im(u) \ Re(u)	$(-K, 0)$	$(0, K)$
$(K', 2K')$	$(-, +)$	$(-, -)$
$(0, K')$	$(+, +)$	$(+, -)$

TABLE 3
Signs of $(\mathrm{Re}(\mathrm{sn}(u|m)), \mathrm{Im}(\mathrm{sn}(u|m)))$.

Im(u) \ Re(u)	$(-K, 0)$	$(0, K)$
$(K', 2K')$	$(-, +)$	$(-, -)$
$(0, K')$	$(+, +)$	$(+, -)$

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 $\|v\| = 1$, k iterations of the Arnoldi process with the initial vector $v_1 = v$ generate an orthonormal basis $\{v_1, v_2, \ldots, v_k, v_{k+1}\}$ for the Krylov subspace $K_{k+1}(A, v) = span\{v, Av, A^2v, \ldots, A^kv\}$ by

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

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

$$AV_k = V_kH_k + h_{k+1,k}v_{k+1}e_k^T, \tag{3.1}$$

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

$$h_{k+1,k}^2 = \|Av_k\|^2 - \sum_{i=1}^{k} h_{i,k}^2 \leq \|A\|^2. \tag{3.2}$$

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

$$V_kV_k^Te^{-\tau A}v = V_kV_k^Te^{-\tau A}V_ke_1 \approx V_ke^{-\tau V_k^TAV_k}e_1 = V_k^Te^{-\tau H_k}e_1.$$

We call

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

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

Let $W(A) := \{x^*Ax : x \in \mathbb{C}^n; \|x\| = 1\}$ be the field of values of A, and let $\mu(A) := \max\{\mathrm{Re}(z) : z \in W(A)\}$ be the logarithmic norm of A (associated with the Euclidean

inner product). We also define $\nu(A) := -\mu(-A) = \min\{\mathrm{Re}(z) : z \in W(A)\}$. Then we have

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

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and the smallest eigenvalue, respectively. In this notation, A is positive definite if and only if $\nu(A) > 0$. An important property associated with the logarithmic norm [9, 29] is that for $t \geq 0$,

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

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 let 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*

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

where

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

is the $(k,1)$ entry of the matrix e^{-tH_k} and $\nu(A)$ is defined in (3.4). *In particular, if $\nu(A) \geq 0$, we have an a posteriori error bound*

$$\|w(\tau) - w_k(\tau)\| \leq h_{k+1,k}\int_0^\tau |h(t)|dt. \tag{3.8}$$

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

$$\begin{aligned} w_k'(t) &= -V_k H_k e^{-tH_k} e_1 \\ &= -(AV_k - h_{k+1,k}v_{k+1}e_k^T)e^{-tH_k}e_1 \\ &= -Aw_k(t) + h_{k+1,k}h(t)v_{k+1}. \end{aligned}$$

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

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

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

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

Then the approximation error satisfies

$$\begin{aligned}\|E_k(\tau)\| &\le h_{k+1,k}\left\|\int_0^\tau h(t)e^{(t-\tau)A}v_{k+1}dt\right\| \\ &\le h_{k+1,k}\int_0^\tau |h(t)|\cdot\|e^{(t-\tau)A}\|dt \\ &\le h_{k+1,k}\int_0^\tau |h(t)|\cdot e^{(t-\tau)\nu(A)}dt.\end{aligned}$$

This proves (3.6). Now, if $\nu(A)\ge 0$, we have $e^{(t-\tau)\nu(A)}\le 1$, since $t-\tau\le 0$. Applying this bound to (3.6) 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, 6]). This provides an understanding of the convergence of the error. Indeed, in section 5, we shall extend the techniques introduced in [3, 6] 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 remark that (3.6) is an a posteriori bound if $\nu(A)$ or a lower bound is known. If $\nu(A)$ is unknown but the matrix is positive semidefinite, then (3.8) provides an a posteriori bound. Both bounds contain an 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 very sharp a posteriori error estimates; see the numerical examples in section 7. Note that several a posteriori error estimates presented in [27] are 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 the $\tilde z$-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 σ by

$$\begin{cases} z=\alpha-\dfrac{i}{\lambda}\{E(\sigma|m)-m_1\sigma\}, \\ w=\dfrac{1-\mathrm{dn}(\sigma|m)}{\sqrt{m}\,\mathrm{sn}(\sigma|m)}, \end{cases} \tag{4.2}$$

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

$$\frac{E(m)-m_1K(m)}{\beta}=\frac{E(m_1)-mK(m_1)}{\alpha}, \tag{4.3}$$

where K and E are functions of m defined in (2.7) and $m_1 := 1 - m$. The existence and uniqueness of m will be shown in Lemma 4.1 below. It is shown in [20, p. 178] that ϕ_2 conformally maps the exterior of the rectangle $[-\alpha, \alpha] \times [-\beta, \beta]$ to the upper half-plane $\{\mathrm{Im}(w) > 0\}$ and that the range of σ is in the rectangle $[-K(m), K(m)] \times [0, 2iK(m_1)]$.

- Step 3:

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

maps $\{\mathrm{Im}(w) > 0\}$ 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 is not readily available in the literature.

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

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

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

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

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

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

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

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) holds. ∎

The parameter m determined by (4.3) is defined by the aspect ratio β/α (or the shape) of the rectangle $[a,b] \times [-c,c]$. For example, from the proof, $m \approx 0$ if the rectangle is a long flat one 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 section 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 $\mathrm{Re}(\tilde{z})$ in C_r, i.e., the leftmost 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 $-K(m) < x < K(m)$ and $0 < y < 2K(m_1)$,*

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

Proof. By (2.18) and (2.19),

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

So,

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

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

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

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

$$\mathrm{dn}(x|m) > 0, \quad \mathrm{dn}(y|m_1) > 0. \tag{4.11}$$

Applying (4.10) and (4.11) to (4.9), we conclude that the imaginary parts of $\mathrm{cn}(u|m)$ and $\mathrm{dn}(u|m)$ have 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 4.3. *Let $\tilde{\Phi} : \tilde{z} \mapsto u$ be defined as in (4.5). Let $\tilde{\Psi} : u \mapsto \tilde{z}$ be its inverse mapping, and let 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 (4.1),

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

Recalling the definition $E(\sigma|m) = \int_0^\sigma \mathrm{dn}^2(z|m)dz$ and the identities $\mathrm{sn}^2 + \mathrm{cn}^2 \equiv 1$ and $m \cdot \mathrm{sn}^2 + \mathrm{dn}^2 \equiv 1$, we have from (4.2) that

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

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

$$\begin{aligned}\frac{dw}{d\sigma} &= \frac{-(-m\cdot\mathrm{sn}\cdot\mathrm{cn})\cdot\sqrt{m}\cdot\mathrm{sn} - (1-\mathrm{dn})\cdot\sqrt{m}\cdot\mathrm{cn}\cdot\mathrm{dn}}{m\cdot\mathrm{sn}^2}\\ &= \frac{\sqrt{m}\cdot\mathrm{cn}\cdot(m\cdot\mathrm{sn}^2 - \mathrm{dn} + \mathrm{dn}^2)}{m\cdot\mathrm{sn}^2}\\ &= \frac{\sqrt{m}\cdot\mathrm{cn}\cdot(1-\mathrm{dn})}{1-\mathrm{dn}^2}\\ &= \frac{\sqrt{m}\cdot\mathrm{cn}}{1+\mathrm{dn}}. \qquad (4.14)\end{aligned}$$

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

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

Combining (4.12)–(4.15), we have

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

Equation (4.4) also implies

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

On the other hand, by (4.2),

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

So,

$$\mathrm{dn} = \frac{1-w^2}{1+w^2} = \frac{(u+1)^2+(u-1)^2}{(u+1)^2-(u-1)^2} = \frac{1}{2}\left(u+\frac{1}{u}\right), \tag{4.17}$$

and hence

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

Substituting this into (4.16), we have

$$\frac{d\tilde{z}}{du} = \frac{\sqrt{m}\cdot \mathrm{cn}}{\lambda u}. \tag{4.18}$$

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

$$\frac{du}{d\theta} = re^{i\theta}\cdot i = iu. \tag{4.19}$$

Treating $\tilde{z} \in C_r$ as a function of θ, we have from (4.18) and (4.19) that

$$\frac{d\tilde{z}}{d\theta} = \frac{i\sqrt{m}}{\lambda}\cdot \mathrm{cn}(\sigma|m). \tag{4.20}$$

So

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

From (4.17) and $u = r\cos\theta + ir\sin\theta$, we write $\mathrm{dn}(\sigma|m)$ as a function of θ,

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

So $\mathrm{Im}(\mathrm{dn}(\sigma|m)) < 0$ when $\theta \in (-\pi, 0)$, and $\mathrm{Im}(\mathrm{dn}(\sigma|m)) > 0$ when $\theta \in (0, \pi]$. By Lemma 4.2, the imaginary part of $\mathrm{cn}(\sigma|m)$ always has the same sign as that of $\mathrm{dn}(\sigma|m)$. Thus, by (4.20), $\frac{d(\mathrm{Re}(\tilde{z}))}{d\theta} > 0$ 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 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*

$$\tilde{\Psi}(-r) = a - \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds, \tag{4.21}$$

where the parameter m is determined by (4.3) *and λ is the ratio in* (4.3)*.*

Proof. Recall that $\tilde{\Phi} = \phi_3 \circ \phi_2 \circ \phi_1$ with ϕ_1, ϕ_2, and ϕ_3 the three conformal mappings defined in (4.1), (4.2), and (4.4). Let

$$\Phi := \phi_3 \circ \phi_2, \tag{4.22}$$

and let Ψ be its inverse. Then obviously

$$\tilde{\Psi}(-r) = \phi_1^{-1} \circ \Psi(-r). \tag{4.23}$$

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

$$\Psi(r) = \alpha + \frac{1}{\lambda}\int_0^{\frac{1}{2}\left(r-\frac{1}{r}\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}ds. \tag{4.24}$$

By the same equation (4.17) that was derived from (4.2) and (4.4), w in the map can be eliminated to define Φ: $z \longleftrightarrow \sigma \longleftrightarrow u$ through the auxiliary parameter σ as

$$\begin{cases} z(\sigma) = \alpha - \dfrac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\}, \\ \mathrm{dn}(\sigma|m) = \dfrac{1}{2}\left(u + \dfrac{1}{u}\right). \end{cases} \tag{4.25}$$

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

$$\mathrm{dn}(\sigma|m) = \frac{1}{2}\left(r + \frac{1}{r}\right) > 1. \tag{4.26}$$

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

$$s = -i\sqrt{m} \cdot \mathrm{sn}(\xi|m), \tag{4.27}$$

where ξ is on the line segment connecting 0 and σ. By Tables 1–3, $\mathrm{sn}(\xi|m)$ is purely imaginary with positive imaginary part, and $\mathrm{cn}(\xi|m)$ and $\mathrm{dn}(\xi|m)$ are both real and positive. Then

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

By (4.27) and (2.14),

$$ds = -i\sqrt{m} \cdot \mathrm{cn}(\xi|m) \cdot \mathrm{dn}(\xi|m)d\xi;$$

then

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

By (4.26),

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

then

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

Thus, as ξ moves along the positive imaginary axis from 0 to σ, s as defined by (4.27) moves along the positive real axis from 0 to $\frac{1}{2}(r - \frac{1}{r})$. Then

$$\begin{aligned} \Psi(r) = z(\sigma) &= \alpha - \frac{i}{\lambda}\{E(\sigma|m) - m_1\sigma\} \\ &= \alpha - \frac{i}{\lambda}\left\{\int_0^\sigma \mathrm{dn}^2(\xi|m)ds - m_1\sigma\right\} \\ &= \alpha - \frac{i}{\lambda}\int_0^\sigma m \cdot \mathrm{cn}^2(\xi|m)ds \end{aligned}$$

$$= \alpha - \frac{i}{\lambda}\int_0^{\frac{1}{2}(r-\frac{1}{r})} (m+s^2)\frac{ds}{-i\sqrt{m+s^2}\sqrt{1+s^2}}$$
$$= \alpha + \frac{1}{\lambda}\int_0^{\frac{1}{2}(r-\frac{1}{r})} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}}dt.$$

This completes the proof of the first part (4.24). We next prove that for any $r > 1$,

$$\Psi(-r) = -\Psi(r). \tag{4.28}$$

Let σ and $\tilde{\sigma}$ be the auxiliary parameters in (4.25) corresponding to r and $-r$, respectively. Then

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

By (2.13), $\tilde{\sigma} = 2iK' - \sigma$. Thus, using (2.12) and (4.3), we get

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

Finally, applying ϕ_1^{-1} to $\Psi(-r)$ as in (4.23) and noting that $\alpha = \frac{b-a}{2}$, equation (4.21) is proved. ∎

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

LEMMA 4.5. *Let λ be the ratio in* (4.3). *We have*

$$\lim_{\tilde{z}\to\infty} \frac{\tilde{\Phi}(\tilde{z})}{\tilde{z}} = 2\lambda > 0. \tag{4.29}$$

Proof. First, by (4.17) and $m\cdot\mathrm{sn}^2(\sigma|m) + \mathrm{dn}^2(\sigma|m) = 1$, we have $\sqrt{m}\cdot\mathrm{sn}(\sigma|m) = \frac{i}{2}(u - \frac{1}{u})$. Applying this to (4.18), we have

$$\frac{d\tilde{z}}{du} = \frac{i}{2\lambda}\cdot\frac{\mathrm{cn}(\sigma|m)}{\mathrm{sn}(\sigma|m)}\left(1 - \frac{1}{u^2}\right). \tag{4.30}$$

As $\tilde{z}\to\infty$, $\sigma \to iK'$ and $u\to\infty$ (see [20, p. 178]). Since

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

we have $\lim_{\sigma\to iK'}\frac{\mathrm{cn}(\sigma|m)}{\mathrm{sn}(\sigma|m)} = -i$. Applying this to (4.30), $\frac{d\tilde{z}}{du}\to\frac{1}{2\lambda}$ 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:

$$\begin{cases} a = \min\limits_i \left\{ \lambda_i \left(\dfrac{A+A^*}{2} \right) \right\} = \nu(A), \\ b = \max\limits_i \left\{ \lambda_i \left(\dfrac{A+A^*}{2} \right) \right\} = \mu(A), \\ c = \max\limits_i \left\{ \left| \lambda_i \left(\dfrac{A-A^*}{2} \right) \right| \right\}, \end{cases} \tag{5.1}$$

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 section 3, as in [31]. As mentioned before, analytic functions of banded matrices have a decay property, i.e., their entries decrease away from the main diagonal. Sharp decay bounds were originally derived by Benzi and Golub [5] for Hermitian matrices; see [4, 7] and the references 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 [6] 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 values have been introduced in [3, 6] to bound the decay rate. Here we will follow the same approach of [3, 6], but we will use the conformal mapping that is constructed in section 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 $k \times k$ 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 $a_k = \min_i\{\lambda_i(\frac{H_k+H_k^*}{2})\}$, $b_k = \max_i\{\lambda_i(\frac{H_k+H_k^*}{2})\}$, and $c_k = \max_i\{|\lambda_i(\frac{H_k-H_k^*}{2})|\}$. 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 $Q = 11.08$,

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

and the parameter m is determined from a_k, b_k, c_k by (4.3) *and λ is the ratio in* (4.3).

Proof. Let $\tilde{\Phi} : \tilde{z} \mapsto u$ 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 $t \ge 0$, 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

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

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

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

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

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

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

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

by Lemmas 4.3 and 4.4. ∎

We remark that $Q = 11.08$ is called Crouzeix's constant, and it is conjectured that it can be reduced to 2 [8]. In [2], Crouzeix's theorem is also used to derive an error bound for the Arnoldi method, which contains Crouzeix's constant Q, but using some highly sophisticated technique from the approximation theory, the same result is proved with the constant reduced to its conjectured value 2. This is a very strong result that we will use to derive an a priori bound or to reduce Q in our bound (5.2) to 2. We first state Beckermann and Reichel's result.

THEOREM 5.2 (Beckermann and Reichel [2, Theorem 3.2 and Corollary 4.1]). *Let $\mathbb{E} \subset \mathbb{C}$ be a convex compact set symmetric with respect to the real axis that contains $W(A)$. Then for all $k > 1$ and $r > 1$, the error of the Arnoldi approximation satisfies*

$$\|e^{\tau A} v - V_k e^{\tau H_k} e_1\| \le 4 \frac{e^{\tau \psi(r)}}{r^k (1 - r^{-1})}, \tag{5.3}$$

where ψ is the inverse of the conformal mapping ϕ from $\bar{C}\backslash\mathbb{E}$ to $\bar{C}\backslash\mathbb{D}$ ($\mathbb{D}$ is the closed unit disk) satisfying $\lim_{z\to\infty} \frac{\phi(z)}{z} > 0$.

We remark that [2, Theorem 3.2] actually shows that $\|e^{\tau A} v - V_k e^{\tau H_k} e_1\| \le 4 \sum_{j=k}^{\infty} |f_j|$, where f_j are the Faber coefficients in the Faber series expansion of e^x. This generalizes an earlier result in the symmetric case [12]. Although this bound can be numerically computed and may be quite a bit sharper than (5.3) [19], it is not easy to interpret, and we will not consider it further. Indeed, the bound (5.3) can also be numerically computed for a rectangular region $\mathbb{E}$ [13, 19]. However, we can apply the conformal mapping $\hat{\Phi}$ that we construct in (4.5) to derive the following more explicit bound.

THEOREM 5.3. *Let $A \in \mathbb{C}^{n\times n}$ and $v \in \mathbb{C}^n$ with $\|v\| = 1$. 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 q with $0 < q < 1$,*

the approximation error satisfies

$$\|w(\tau) - w_k(\tau)\| \le \frac{4q^k}{1-q} e^{-\tau \tilde{z}},$$

where

$$\tilde{z} = a - \frac{1}{\lambda} \int_0^{\frac{1}{2}\left(\frac{1}{q} - q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds, \tag{5.4}$$

the parameter m is determined by (4.3) *from a, b, c of* (5.1)*, and λ is the ratio in* (4.3).

Proof. Let $\mathbb{E} = [a, b] \times [-c, c]$. Under our construction in section 4, $\hat{\Phi}$ in (4.5) maps $\bar{C}\backslash\mathbb{E}$ conformally onto $\bar{C}\backslash\mathbb{D}$, where $\mathbb{D}$ is the unit disk, satisfying (4.29). Let $B = -A$. Then $-\mathbb{E}$ contains the field of values of B. Then $\phi := (-\mathrm{id}) \circ \hat{\Phi} \circ (-\mathrm{id})$, where id is the identity mapping, maps $\bar{C}\backslash(-\mathbb{E})$ conformally onto $\bar{C}\backslash\mathbb{D}$, satisfying $\lim_{z\to\infty} \frac{\phi(z)}{z} > 0$. Clearly, the inverse mapping is $\psi = (-\mathrm{id}) \circ \hat{\Psi} \circ (-\mathrm{id})$. Thus,

$$\psi(r) = -\hat{\Psi}(-r) = -\tilde{z}, \tag{5.5}$$

where $\tilde{z}$ is given in (5.4). Now, the theorem follows from applying this to Theorem 5.2 and letting $q = \frac{1}{r}$. ∎

Applying Theorem 5.2 to the upper Hessenberg matrix H_k obtained after k steps of Arnoldi iterations gives us the following corollary, which reduces the constant Q to 2.

COROLLARY 5.4. *Theorem* 5.1 *holds with Q in* (5.2) *being* 2.

Proof. Applying $k-1$ Arnoldi iterations to H_k with initial vector e_1, we obtain

$$H_k E_{k-1} = E_{k-1} H_{k-1} + h_{k,k-1} e_k e_{k-1}^T,$$

where $E_{k-1} = [e_1, e_2, \ldots, e_{k-1}]$ and $e_j \in \mathbb{R}^k$ is the jth coordinate vector. Using $e_k^T E_{k-1} = 0$, we have

$$\begin{aligned} |h(t)| &= |e_k^T e^{-tH_k} e_1| = |e_k^T (e^{-tH_k} - E_{k-1} e^{-tH_{k-1}}) e_1| \\ &\le \|e^{-tH_k} e_1 - E_{k-1} e^{-tH_{k-1}} e_1\|. \end{aligned} \tag{5.6}$$

Applying Theorem 5.3 to (5.6) with A, Q_k, and H_k being, respectively, H_k, E_{k-1}, and H_{k-1}, we obtain

$$|h(t)| \le \|e^{-tH_k} e_1 - E_{k-1} e^{-tH_{k-1}} e_1\| \le \frac{4q^k}{1-q} e^{-t\tilde{z}}$$

with $\tilde{z}$ defined by a_k, b_k, c_k as in (5.4). This completes the proof. ∎

Although the above corollary supersedes Theorem 5.1, we have kept Theorem 5.1 as it provides a more direct proof. Finally, combining the results in Theorem 5.3 and Corollary 5.4 with Theorem 3.1 leads to the following a priori error bound.

THEOREM 5.5. *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*

$$\|w(\tau) - w_k(\tau)\| \le \frac{4q^{k-1}}{1-q} e^{-\tau\tilde{z}} \min\{\tau\|A\|, q\}, \tag{5.7}$$

where $\tilde{z}$ is given in (5.4).

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, Corollary 5.4 holds for $h(t) = e_k^T e^{-tH_k} e_1$, and indeed, from the inclusion relation above and following the same proof as for Theorem 5.3, it holds with a, b, c in place of a_k, b_k, c_k. Namely, $|h(t)| \leq \frac{4q^{k-1}}{1-q} e^{-t\tilde{z}}$, with $\tilde{z}$ defined as in (5.4) from a, b, c. Now, using this bound in the a posteriori error bound (3.6) in Theorem 3.1 and noting that $h_{k+1,k} \leq \|A\|$ (see (3.2)), we have that

$$\begin{aligned}
\|w(\tau) - w_k(\tau)\| &\leq h_{k+1,k} \int_0^\tau \frac{4q^{k-1}}{1-q} e^{-t\tilde{z}} e^{(t-\tau)a} dt \\
&\leq 4\|A\| \frac{q^{k-1}}{1-q} e^{-\tau a} \int_0^\tau e^{t(a-\tilde{z})} dt \\
&\leq 4\|A\| \frac{q^{k-1}}{1-q} e^{-\tau a} \tau e^{\tau(a-\tilde{z})} \\
&= 4\tau\|A\| \frac{q^{k-1}}{1-q} e^{-\tau\tilde{z}},
\end{aligned}$$

where we note that $a - \tilde{z} > 0$. Combining this with Theorem 5.3, the theorem is proved. ∎

The bound in the above theorem can be simplified for easy interpretation by bounding $\tilde{z}$ in the following corollary.

COROLLARY 5.6. *Under the assumptions of Theorem* 5.5*, for any q with $0 < q < 1$, the approximation error satisfies*

$$\|w(\tau) - w_k(\tau)\| \leq \frac{4q^{k-1-\frac{\tau\sqrt{m}}{\lambda}}}{1-q} \min\{\tau\|A\|, q\} e^{-\tau(a-L)}, \tag{5.8}$$

where $L = L(q) := \frac{1}{2\lambda}(\frac{1}{q} + q - 2)$. If A is positive definite (i.e., $a > 0$), for $q_0 := \frac{1}{a\lambda+1+\sqrt{(a\lambda+1)^2-1}}$, we have

$$\|w(\tau) - w_k(\tau)\| \leq \frac{4q_0^{k-1-\frac{\tau\sqrt{m}}{\lambda}}}{1-q_0} \min\{\tau\|A\|, q_0\}. \tag{5.9}$$

Proof. It is easy to check that

$$\sqrt{m+s^2} \leq \sqrt{m} + s.$$

Then

$$\begin{aligned}
\int_0^{\frac{1}{2}(\frac{1}{q}-q)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds &\leq \int_0^{\frac{1}{2}(\frac{1}{q}-q)} \left[\frac{\sqrt{m}}{\sqrt{1+s^2}} + \frac{s}{\sqrt{1+s^2}}\right] ds \\
&= \left[\sqrt{m}\sinh^{-1}(s) + \sqrt{1+s^2}\right] \Big|_{s=0}^{s=\frac{1}{2}(\frac{1}{q}-q)} \\
&= \sqrt{m}\sinh^{-1}\left(\frac{1}{2}\left(\frac{1}{q} - q\right)\right) + \frac{1}{2}\left(\frac{1}{q} + q\right) - 1 \\
&= \sqrt{m}\ln\left(\frac{1}{q}\right) + \frac{1}{2}\left(\frac{1}{q} + q - 2\right).
\end{aligned}$$

Noting (5.4), it follows that

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

and hence

$$e^{-\tau\tilde{z}} \le e^{-\tau a + \tau\frac{\sqrt{m}}{\lambda}\ln\left(\frac{1}{q}\right)+\tau L} = q^{-\frac{\tau\sqrt{m}}{\lambda}} e^{-\tau(a-L)}.$$

Using this in Theorem 5.5 yields (5.8). Finally, (5.9) follows from (5.8) by noting that $-a + L = 0$ if $q = q_0$. ∎

The above bounds show that the error starts to decrease at the rate of q only when

$$k \ge k_s := \frac{\tau\sqrt{m}}{\lambda}. \tag{5.10}$$

Namely, we may not expect convergence to take place until k_s steps; i.e., the iterations may stagnate for the first k_s steps. This theoretical prediction is nicely confirmed in our numerical testing; see section 7. On the other hand, the fixed rate convergent bound (5.9) is usually very pessimistic.

For a positive definite A, the bound (5.9) guarantees the convergence at least at the rate of q_0, up to initial k_s stagnation steps. This guaranteed overall linear rate may be expected to be pessimistic as superlinear convergence is generally observed. The superlinear convergence behavior can be explained by noting that optimal q value for the error bounds changes with k. 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, resulting 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 $e^{-\tau\tilde{z}}$.

The optimal q to be used in the bound (5.7) can be determined by minimizing at each step k

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

Taking the 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

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

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

$$(k-1)q + (2-k)q^2 - C(1-q)\sqrt{(1-q^2)^2+4mq^2} = 0. \tag{5.11}$$

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

We note that Beckermann and Reichel [2, Corollary 4.1] have also discussed choosing r for their bound (5.3) by optimizing $e^{\tau\psi(r)}/r^m$ (an upper bound of the Faber coefficient) with respect to r. Their optimal r is 1 if $k > \sqrt{m}\tau/\lambda$, and is given by the solution of $r\psi'(r) = k/\tau$ otherwise. We choose q by optimizing the entire upper bound $E(q)$. As a result, we have found numerically that using their optimal value of r with $q = 1/r$ gives clearly worse bounds than the one determined by (5.11).

Finally, we note that the bound in Corollary 5.6 is derived using $\sqrt{m+s^2} \le \sqrt{m}+s$, which is quite tight for most values of m as each of $\sqrt{m}$ and s is also a lower bound of $\sqrt{m+s^2}$. When $m \approx 1$, it is better to simply bound it as

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

to obtain

$$||w(\tau) - w_k(\tau)|| \le \frac{4q^{k-1}}{1-q} \min\{\tau\|A\|, q\}\, e^{-\tau\left\{a - \frac{1}{2\lambda}\left(\frac{1}{q}-q\right)\right\}}.$$

The special case that $m = 1$ and $a = 0$ will be discussed in the next section. Here, we finish this section with a discussion of the special case that A is nearly Hermitian positive definite, i.e., $m \approx 0$.

COROLLARY 5.7. *Under the assumptions of Theorem* 5.5*, A being positive definite, and $m \approx 0$, the approximation error satisfies*

$$\|w(\tau) - w_k(\tau)\| \le 4\frac{q_0^{k-1}}{1-q_0} \min\{\tau\|A\|, q_0\},$$

where $q_0 = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} + O(\sqrt{m})$ *and* $\kappa = \frac{b}{a}$.

Proof. $E(m_1) = E(1-m)$ and $K(m_1) = K(1-m)$ are both functions of m and have the following expansions at $m = 0$ [1, p. 591]:

$$E(m_1) = E(1-m) = 1 - \frac{1}{4} m \ln m + O(m),$$
$$K(m_1) = K(1-m) = -\frac{1}{2}\ln m + O(1).$$

Then $E(m_1) - mK(m_1)$ can be expanded at $m = 0$ as

$$E(m_1) - mK(m_1) = 1 + \frac{1}{4} m \ln m + O(m). \tag{5.12}$$

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

$$\lambda = \frac{E(m_1) - mK(m_1)}{\alpha} = \frac{2}{b-a}\left(1 + \frac{1}{4} m \ln m\right) + O(m).$$

Then

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

As we have proved,

$$\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds \le \sqrt{m} \ln\left(\frac{1}{q}\right) + \frac{1}{2}\left(\frac{1}{q}+q-2\right),$$

so

$$\int_0^{\frac{1}{2}\left(\frac{1}{q}-q\right)} \frac{\sqrt{m+s^2}}{\sqrt{1+s^2}} ds = \frac{1}{2}\left(\frac{1}{q}+q-2\right) + O(\sqrt{m}). \tag{5.14}$$

Let $q = q_0$ be the unique solution of

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

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 ∞ 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 (5.7), we have $\tilde{z} = 0$, and the theorem is proved. ∎

Note that m is determined by β/α. 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 $E(m_1) - mK(m_1)$ in (5.12). Then

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

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

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 a Hermitian matrix. In this case, the Arnoldi algorithm is theoretically equivalent to the Lanczos algorithm for H. As we will see, the error bound for computing

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

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 $\{v_1, v_2, \ldots, v_k, v_{k+1}\}$ and a $k \times k$ tridiagonal matrix T_k such that

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

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

$$w_k(\tau) := V_k e^{i\tau T_k} e_1, \tag{6.1}$$

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,

$$\|w(\tau) - w_k(\tau)\| \le \beta_{k+1} \int_0^\tau |h(t)|dt \le \|H\| \int_0^\tau |h(t)|dt. \tag{6.2}$$

Furthermore, slightly better bounds may be obtained by shifting the matrix and exploring the fact that such a shift only results in a multiplication by $e^{-i\tau\alpha}$ which has modulus 1. 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 (6.2) 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 $h(t, \alpha) := e_k^T e^{it(T_k - \alpha I)} e_1 = e^{-it\alpha} h(t, \alpha)$. Thus

$$\|w(\tau) - w_k(\tau)\| = \|w(\tau, \alpha) - w_k(\tau, \alpha)\| \le \|H - \alpha I\| \int_0^\tau |h(t)|dt. \tag{6.3}$$

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

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

$$\|w(\tau) - w_k(\tau)\| \le \frac{4 \min\{1/(1-q^2), \tau\rho/q\}}{1-q} q^k e^{\tau\rho\left(\frac{1}{q} - q\right)}, \tag{6.4}$$

where $\rho = (\lambda_{\max}(H) - \lambda_{\min}(H))/4$ with $\lambda_{\min}(H)$ and $\lambda_{\max}(H)$ being the smallest and 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 $[a, b]$, $z_2 = \phi_2(z_1) = \frac{2}{b-a}(z_1 - \frac{a+b}{2})$ maps the exterior of $[a, b]$ to the exterior of $[-1, 1]$, and $w = \phi_3(z_2) = i(z_2 + \sqrt{z_2^2 - 1})$ maps the exterior of $[-1, 1]$ to $\{|w| > 1\}$. In the definition of ϕ_3, we choose the branch of $\sqrt{z^2 - 1}$ such that $\lim_{z\mapsto\infty} \frac{\sqrt{z^2-1}}{z} = 1$. Then Φ 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 Φ and the Faber polynomial approximation Π_{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 Φ 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 $z \in C_r$, we write $z = \Phi^{-1}(w)$ with $w = re^{i\theta}$, where $\theta \in [0, 2\pi)$. Then, it follows from the definition of Φ that

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

Thus

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

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

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

Finally, using (6.3) with $\alpha = (a+b)/2$, we have $\|H - \alpha I\| = (b-a)/2$ and hence

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

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 $e^{\tau\rho\left(\frac{1}{q}-q\right)}$ is expected to be larger.

So, again, we study the value of q that minimizes the bound

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

Taking the derivative of $E(q)$ with respect to q, we 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

$$\tau\rho q^4 + (3-k)q^3 + q^2 + kq - \tau\rho = 0. \tag{6.5}$$

Note that it can be shown that the above equation has a unique solution $q_0 \in (0,1)$ (see [30] 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 $q \le 0.9$. To quantitatively interpret the bound, we can consider minimization of

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

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 $k \le k_s := 2\tau\rho$, $E_s(q)$ is monotonically decreasing with the minimum occurring at $q_0 = 1$. If $k > k_s$, $E_s(q)$ 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 $1 \le k \le k_s$, there is essentially no decrease in the error bound.
2. For $k > k_s$, 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 section 7), where the error initially stagnates for approximately k_s steps and then begins to decrease superlinearly. This k_s is the same as $k_s = \frac{\tau\sqrt{m}}{\lambda}$ defined in (5.10), as $m = 1$ and $\frac{1}{\lambda} = 2\rho$ here. Thus, the stagnation steps k_s can be explained from the optimal value of q here or a delayed convergence term q^{k-1-k_s} in (5.8).

Finally, we note that the convergence bound for skew-Hermitian matrices has also been studied by Hochbruck and Lubich [16, Theorem 4]. It is proved there that for $k \ge 2\rho\tau$,

$$\|w(\tau) - w_k(\tau)\| \le 12 e^{\frac{-(\rho\tau)^2}{k}} \left(\frac{e\rho\tau}{k}\right)^k. \tag{6.6}$$

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 (6.4). For $k \ge 2\rho\tau$, let $q = \frac{\tau\rho}{k} \le \frac{1}{2}$. Then our

bound (6.4), simply using $1/(1-q^2)$ for the minimum, reduces to (6.6) as follows:

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

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 were 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 posteriori error estimate (3.8) and a priori bounds (5.7) or (6.4). The convergence rate q in the bounds (5.7) and (6.4) was chosen to satisfy (5.11) and (6.5), respectively. The a posteriori error estimate (3.8) assumes A is positive semidefinite. When that is not the case, i.e., $\nu(A) < 0$, we use (3.6) with the value of $\nu(A)$ assumed known. The integrals in the a posteriori error estimates (3.6) and (3.8) are approximated using Simpson's rule with 10 subintervals on $[0, \tau]$.

We shall compare our bounds with the bounds by Saad [27] and those of Hochbruck and Lubich [16]. Specifically, we consider the bound [27, Theorem 4.5]

$$\|w(\tau) - w_k(\tau)\| \le \frac{2(\tau\rho_\alpha)^k e^{\tau(\rho_\alpha - \alpha)}}{k!}, \tag{7.1}$$

where $\rho_\alpha = \|A - \alpha I\|$, and the bound [16, Theorem 2]

$$\|w(\tau) - w_k(\tau)\| \le 12 e^{-\rho\tau} \left(\frac{e\rho\tau}{k}\right)^k, \tag{7.2}$$

which holds for $k \ge 2\rho\tau$ and with the assumption that the field of values $W(A)$ is contained in the disk $|z - \rho| < \rho$. When the latter assumption does not hold, we consider the circumscribing circle of the rectangle $[a, b] \times [-c, c]$ enclosing $W(A)$, and shift it by some α to $|z - \rho| < \rho$. Then (7.2) can be applied, from which a bound on the original error is obtained by multiplying $e^{-\tau\alpha}$. For Saad's bound, we set α to be the center of the rectangle $\alpha = (a + b)/2$ to minimize ρ_α.

In comparison with our bounds, (7.1) and (7.2) are based on enclosing $W(A)$ by a disk and are tight if $W(A)$ is known to be tightly enclosed in a disk. Note that our bounds are derived essentially by first mapping a rectangle to a disk and then bounding on the disk; so when $W(A)$ is already tightly enclosed by a disk, an additional step of mapping a circumscribing square to a larger disk is redundant, and using the original enclosing disk as in (7.1) or (7.2) is a better approach. We therefore consider examples where the enclosing rectangle is the best available information about $W(A)$.

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 2×2 matrices $B_{\ell,j}$ for $\ell = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, \frac{N-1}{2}$, where

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

Then, the eigenvalues of A are $x_\ell \pm iy_j$ (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 $[a,b]\times[-c,c]$.

We 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 we construct a matrix A as above such that the eigenvalues of A form a 31×31 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, our a posteriori error estimate (3.8) in the +-line, our a priori bound (5.7) in the dashed line, Hochbruck and Lubich's bound (7.2) in the dotted line, and Saad's bound (7.1) in the x-line. Note that Hochbruck and Lubich's bound is only valid for $k \geq 2\rho\tau$.

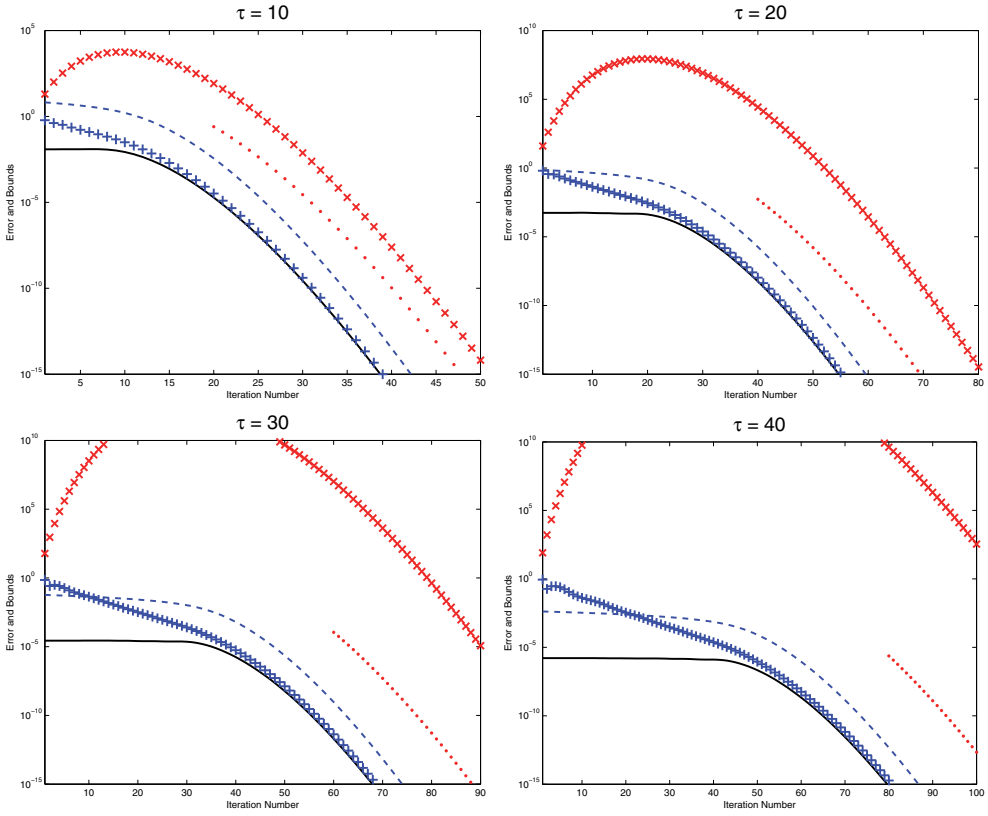

FIG. 1. *Example* 1. $W(A)$ *in* $|z-1|<1$ *and* $\tau = 10, 20, 30, 40$; $k_s = 12, 24, 35, 47$. *Error (solid), our a posteriori estimate (+), our a priori bound (dashed), Hochbruck and Lubich's bound (dotted), and Saad's bound (×).*

We observe that when τ is relatively small, our new a priori bound slightly outperforms the classical bounds, but as τ increases, our bound improves significantly. In particular, the error has an initial stagnation before the convergence takes place. Our theoretically predicted values for the stagnation steps $k_s = \frac{\tau\sqrt{m}}{\lambda}$ are 12, 24, 35, and 47 for the four corresponding τ values. They sharply capture the actual stagnation stage of iterations in all cases. Our a posteriori error estimate (3.8) is sharp at the

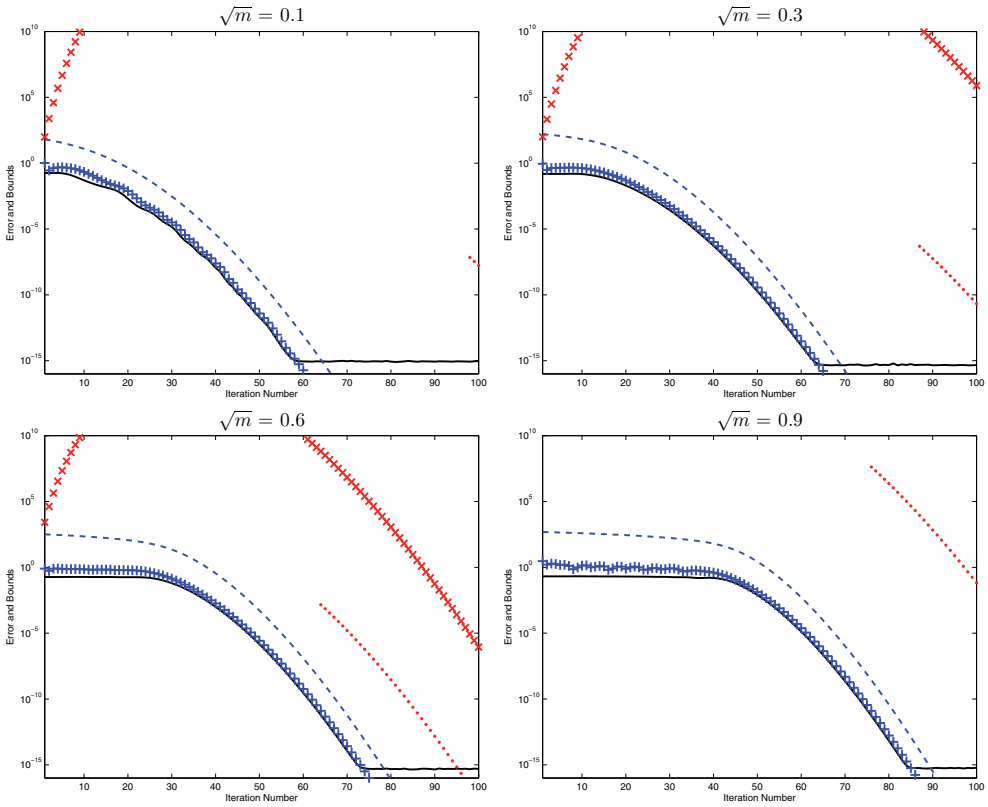

FIG. 2. *Example* 2. $\sqrt{m} = 0.1, 0.3, 0.6, 0.9$; $k_s = 5, 15, 30, 45$. *Error (solid), a posteriori estimate* (+)*, a priori bound (dashed), Hochbruck and Lubich's bound (dotted), and Saad's bound* (×).

convergence stage for all tests. If we assume $\nu(A)$ is known and use (3.6), the resulting a posteriori error estimate (not plotted here) will be sharp at all steps.

In the next example, we use the same construction as in Example 1, but consider the field of values contained in rectangles of different shapes. This is to investigate the influence on the convergence 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 α and β by $\alpha = E(m_1) - mK(m_1)$, $\beta = E(m) - m_1 K(m)$, which means $\lambda = 1$. We then construct as in Example 1 a matrix whose field of values is contained in the rectangle $[0, 2\alpha] \times [-\beta, \beta]$. We use $\tau = 50$ and $\sqrt{m} \in \{0.1, 0.3, 0.6, 0.9\}$, whose corresponding values of $k_s = \frac{\tau\sqrt{m}}{\lambda}$ are 5, 15, 30, and 45. Note from section 5 that $m \approx 0$ means that the matrix is close to being Hermitian, and that $m \approx 1$ means that 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. In Figure 2 we plot the error $\|w(\tau) - w_k(\tau)\|$ in the solid line, our a posteriori error estimate (3.8) in the +-line, our a priori bound (5.7) in the dashed line, Hochbruck and Lubich's bound (7.2) in the dotted line, and Saad's bound (7.1) in the x-line. In the last plot

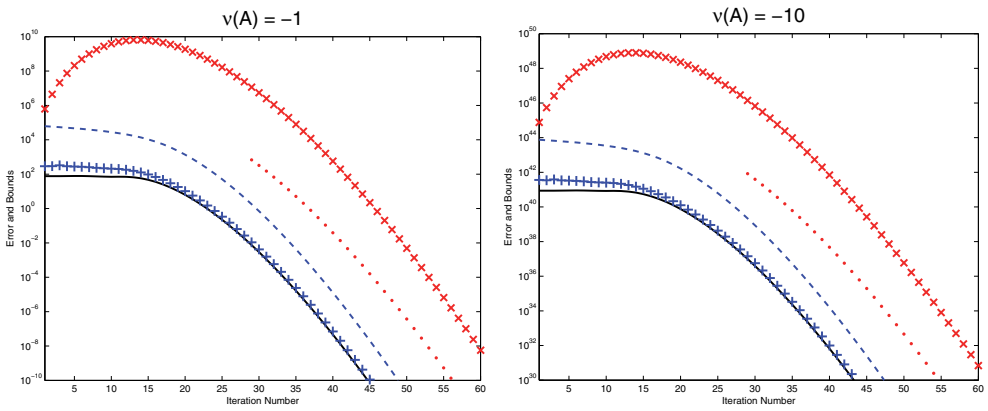

FIG. 3. *Example* 2. *Nonpositive definite matrix with negative* $\nu(A)$: $\sigma = -1$ *and* -10; $k_s = 17, 17$. *Error (solid), our a posteriori bound (+), our a priori bound (dashed), Hochbruck and Lubich's bound (dotted), and Saad's bound (×).*

($\sqrt{m} = 0.9$), Saad's bound is out of range and is not shown.

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, the convergence has an initial stagnation stage before the convergence occurs. Again, this behavior is captured in our new a priori bound (5.8). In particular, our theoretically predicted values for the stagnation steps k_s sharply capture the actual stagnation stage of iterations in all cases. Our bound also significantly improves both Hochbruck and Lubich's and Saad's bounds. Our a posteriori error estimate is sharp for all tests.

We further demonstrate our new bounds for nonpositive 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. For $\tau = 10$, we plot in Figure 3 the approximation error (solid), our a posteriori estimate (3.6) (+), our a priori bound (5.7) (dashed), Hochbruck and Lubich's bound (7.2) (dotted), and Saad's bound (7.1) (×). We see that our bounds are still valid and sharp when A is not positive definite. The predicted stagnation steps k_s are 17 for both cases, and they agree with the actual results.

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

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

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 20×20 grid in $[0, 1]^2$ scaled with h^2 so that $\|A\| \approx 8$. Then A is non-Hermitian but positive definite. We let v be a random vector with $\|v\| = 1$ and compute the matrix exponential $w(\tau) = e^{-\tau A}v$. We use various values of $\tau = 1, 2, 5, 10$ and apply the Arnoldi method to A and v. The results are presented in Figure 4, with $\|w(\tau) - w_k(\tau)\|$ in

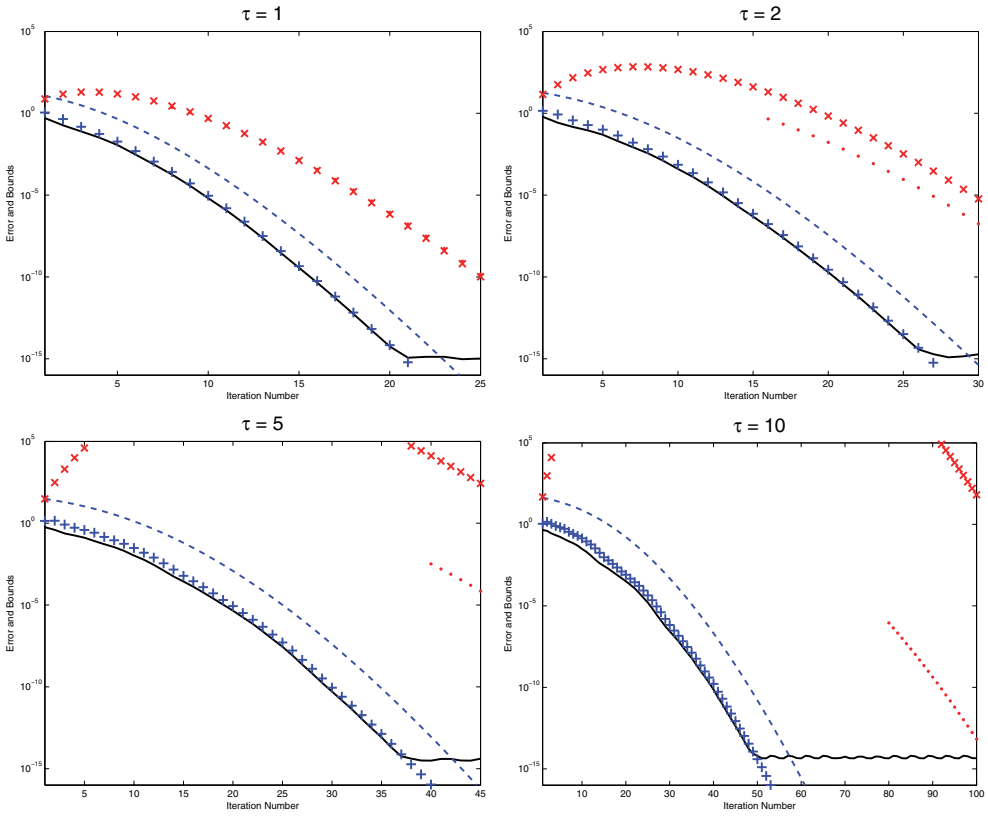

FIG. 4. *Example* 3. $\tau = 1, 2, 5, 10$; $k_s = 1, 1, 3, 6$. *Error (solid), a posteriori estimate (+), a priori bound (dashed), Hochbruck and Lubich's bound (dotted), and Saad's bound (×).*

the solid line, our a posteriori estimate (3.8) in the +-line, our a priori bound (5.7) in the dashed line, Hochbruck and Lubich's bound (7.2) in the dotted line, and Saad's bound (7.1) in the x-line. In the first plot, Hochbruck and Lubich's bound (the dotted line) and Saad's bound (x-line) coincide and are indistinguishable.

We observe that our a posteriori error estimate and our a priori bounds closely follow the convergence curve and are significant improvements on the classical bounds. The predicted stagnation steps k_s are $1, 1, 3, 6$. This is consistent with the actual convergence, where there is little initial stagnation shown.

Our final example concerns skew-Hermitian matrices.

Example 4. Let H be an $n \times n$ diagonal matrix whose jth diagonal entry is j/n. Let v be a random $n \times 1$ normalized vector. Then $\|H\| = 1$, and the spectral gap $4\rho = \lambda_{\max}(H) - \lambda_{\min}(H)$ is approximately 1. We apply k iterations of the Lanczos method to compute $w(\tau) = e^{i\tau H}v$. We will test $n = 1000$ with $\tau = 2, 10, 20, 50$; the results are presented in Figure 5, with $\|w(\tau) - w_k(\tau)\|$ in the solid line, our a posteriori error estimate (6.2) in the +-line, our a priori bound (6.4) in the dashed line, Hochbruck and Lubich's bound (6.6) in the dotted line, and Saad's bound (7.1) in the x-line.

We first observe that our bound only improves Hochbruck and Lubich's bound very slightly. It is significantly better than Saad's bound when τ is large. In all cases,

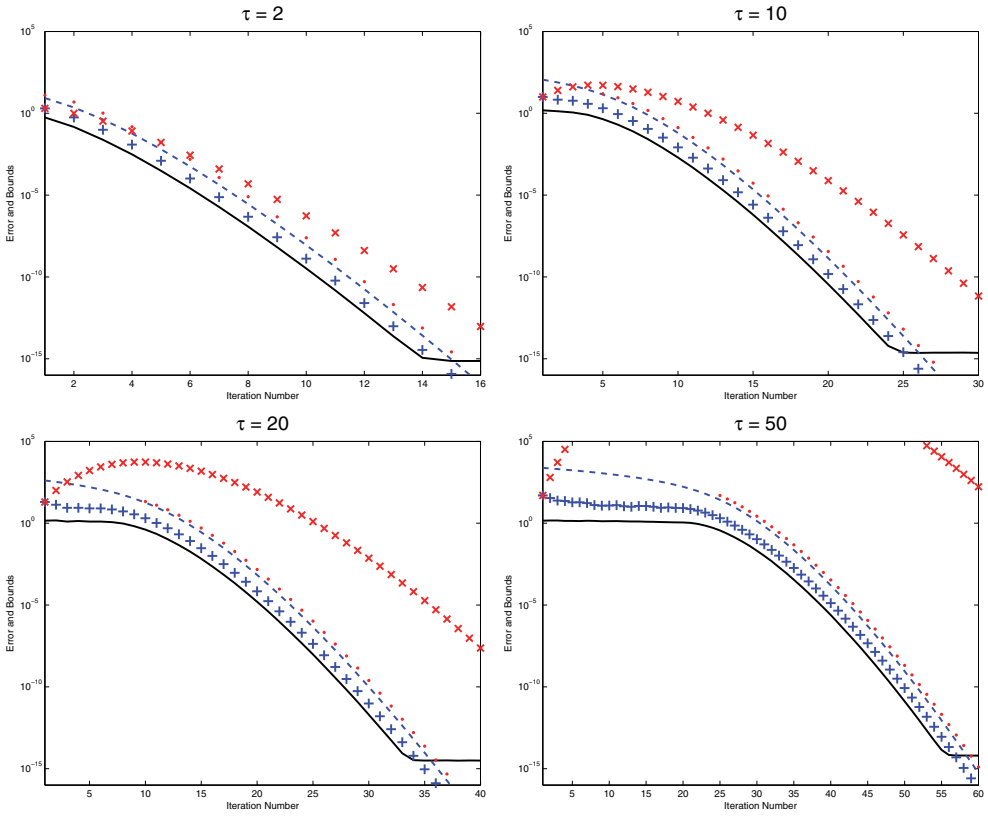

FIG. 5. *Example* 4. 1000×1000 *diagonal matrix with* $a_{jj} = j/1000$ *and* $\tau = 2, 10, 20, 50$. $k_s = 1, 5, 10, 25$. *Error (solid), a posteriori bound (+), a priori bound (dashed), Hochbruck and Lubich's bound (dotted), and Saad's bound (×).*

our bound and Hochbruck and Lubich's bound follow the actual error quite closely, and our a posteriori error estimate is sharp. For $\tau = 2, 10, 20, 50$, the corresponding predicted stagnation steps k_s are 1, 5, 10, and 25, respectively. This again sharply predicts the actual stagnation stage in Figure 5.

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 error. We have also derived from the bounds of Beckermann and Reichel [2] as well as from our a posteriori bound some 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 provide a sharp prediction of the initial stagnation of the convergence curve as shown in all numerical examples. Numerical comparisons with existing bounds also show that our new bounds may significantly improve the a priori bound by Hochbruck and Lubich [16] 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, our bounds agree with those of [31] 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. In future work, 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.

Acknowledgments. We would like to thank Prof. Michele Benzi for many valuable discussions and, in particular, for his suggestion to use the technique in [3], which has turned out to be very fruitful. We would also like to thank Dr. Leonid Knizhnerman and an anonymous referee for many constructive comments that have significantly improved the paper.

REFERENCES

[1] M. Abramowitz and I. A. Stegun, *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, Dover, New York, 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, https://doi.org/10.1137/080741744.

[3] M. Benzi and P. Boito, *Decay properties for functions of matrices over C*-algebras*, Linear Algebra 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 Rev., 55 (2013), pp. 3–64, https://doi.org/10.1137/100814019.

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

[7] 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, https://doi.org/10.1137/151006159.

[8] M. Crouzeix, *Numerical range and functional calculus in Hilbert space*, J. Funct. Anal., 244 (2007), pp. 668–690.

[9] G. Dahlquist, *Stability and Error Bounds in the Numerical Integration of Ordinary Differential Equations*, Almqvist & Wiksells, Uppsala, 1958.

[10] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997, https://doi.org/10.1137/1.9781611971446.

[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, https://doi.org/10.1137/S1064827596303661.

[12] V. L. Druskin and L. A. Knizhnerman, *Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic*, 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, https://doi.org/10.1137/0913071.

[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. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925, https://doi.org/10.1137/S0036142995280572.

[17] M. Ilić, I. W. Turner, and V. Anh, *A numerical solution using an adaptively preconditioned Lanczos method for a class of linear systems related with the fractional Poisson equation*, J. Appl. Math. Stoch. Anal., 2008 (2008), 104525, https://doi.org/10.1155/2008/104525.

[18] L. A. Knizhnerman, *Calculation of functions of unsymmetric matrices using Arnoldi's method*, Comput. Math. Math. Phys., 31 (1991), pp. 1–9.

[19] L. A. Knizhnerman, *private communication*, 2016.
[20] H. Kober, *Dictionary of Conformal Representations*, Dover, New York, 1957.
[21] A. I. Markushevich, *Theory of Functions of a Complex Variable*, Vol. III, revised English edition translated and edited by R. A. Silverman, Prentice-Hall, Englewood Cliffs, NJ, 1967.
[22] L. M. Milne-Thomson, *Jacobian Elliptic Function Tables*, Dover, New York, 1950.
[23] 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, https://doi.org/10.1137/S00361445024180.
[24] I. Moret and P. Novati, *On the convergence of Krylov subspace methods for matrix Mittag–Leffler functions*, SIAM J. Numer. Anal., 49 (2011), pp. 2144–2164, https://doi.org/10.1137/080738374.
[25] A. Nauts and R. Wyatt, *New approach to many state quantum dynamics: The recursive residue generation method*, Phys. Rev. Lett., 51 (1983), pp. 2238–2241.
[26] T. J. Park and J. C. Light, *Unitary quantum time evolution by iterative Lanczos reduction*, J. Chem. Phys., 85 (1986), pp. 5870–5876.
[27] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator*, SIAM J. Numer. Anal., 29 (1992), pp. 209–228, https://doi.org/10.1137/0729014.
[28] 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), pp. 1551–1558.
[29] G. Söderlind, *The logarithmic norm. History and modern theory*, BIT, 46 (2006), pp. 631–652.
[30] H. Wang, *The Krylov Subspace Methods for the Computation of Matrix Exponentials*, Ph.D. dissertation, Department of Mathematics, University of Kentucky, Lexington, KY, 2015.
[31] Q. Ye, *Error bounds for the Lanczos methods for approximating matrix exponentials*, SIAM J. Numer. Anal., 51 (2013), pp. 68–87, https://doi.org/10.1137/11085935X.