1-2-2013

Error Bounds for the Lanczos Methods for Approximating Matrix Exponentials

Qiang Ye
University of Kentucky, qiang.ye@uky.edu

SIAM J. NUMER. ANAL.
Vol. 51, No. 1, pp. 68–87

ERROR BOUNDS FOR THE LANCZOS METHODS FOR APPROXIMATING MATRIX EXPONENTIALS*

QIANG YE†

Abstract. In this paper, we present new error bounds for the Lanczos method and the shift-and-invert Lanczos method for computing $e^{-\tau A}v$ for a large sparse symmetric positive semidefinite matrix A. Compared with the existing error analysis for these methods, our bounds relate the convergence to the condition numbers of the matrix that generates the Krylov subspace. In particular, we show that the Lanczos method will converge rapidly if the matrix A is well-conditioned, regardless of what the norm of τA is. Numerical examples are given to demonstrate the theoretical bounds.

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

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

DOI. 10.1137/11085935X

1. Introduction.

In this paper, we are concerned with Lanczos-type methods for approximating the product of a matrix exponential and a vector of the form

$$w(\tau) = e^{-\tau A}v, \tag{1.1}$$

where $A \in \mathbb{R}^{n\times n}$ is a large, sparse, and symmetric positive semidefinite matrix, $v \in \mathbb{R}^n$ with $\|v\|_2 = 1$, and τ is a fixed positive constant. This problem arises in the initial value problem for a time-dependent ODE,

$$\frac{dv(t)}{dt} = -Av(t) + r(t), \qquad v(0) = v_0. \tag{1.2}$$

Often, τ is a time step parameter in a finite difference discretization of (1.2), which is typically based on an approximation of the formula

$$v(t+\tau) = e^{-\tau A}v(t) + \int_0^\tau e^{-(\tau-\delta)A}r(t+\delta)d\delta. \tag{1.3}$$

The calculation of (1.3) with the integral approximated by a quadrature rule involves matrix-vector products of form (1.1). There are many other practical applications where the problem (1.1) arises directly; see [10, 16, 17, 22] for examples. We also refer to Moler and Van Loan [19] for a discussion on general theory and numerical methods for matrix exponentials.

The Lanczos method and more generally the Krylov subspace methods introduced by Saad [24] and Gallopoulos and Saad [14] are some of the most efficient methods for computing $\exp(A)v$; see Sidje [25] for a robust implementation. The methods have found applications in a variety of problems; see [5, 12, 17, 21, 23], for example. With the matrix A used to form matrix-vector products only, they are theoretically equivalent to a polynomial approximation of $\exp(A)$ and implicitly define a high-order explicit-type scheme for solving (1.2). Since their introduction [14, 24],

*Received by the editors December 19, 2011; accepted for publication (in revised form) October 31, 2012; published electronically January 2, 2013. This research was supported in part by NSF grant DMS-0915062.

http://www.siam.org/journals/sinum/51-1/85935.html

†Department of Mathematics, University of Kentucky, Lexington, KY 40506 (qye3@uky.edu).

several generalizations of the original Krylov subspace methods have been proposed in [1, 9, 10, 11, 13, 15, 18, 29]. Several other methods have also been proposed as a preconditioning technique for problem (1.1) in [4, 17, 20, 27]. In particular, the shift-and-invert Lanczos method that uses projections on a Krylov subspace generated by a shift-and-invert matrix can significantly accelerate the convergence of the Lanczos method; see van den Eshof and Hochbruck [27] and Moret and Novati [20].

Simultaneously, error bounds that aim at explaining convergence properties of the Krylov subspace methods have been extensively studied. Some a priori error bounds and a posteriori error estimates were first presented in [14, 24]. More sophisticated and refined error bounds have been obtained in [9, 10, 15, 21]. The existing error bounds suggest that the speed of convergence depends on the norm of τA. This may limit the use of the Krylov subspace methods to problems where $\tau\|A\|$ is not too large. In the context of time stepping (1.3) for solving the initial value problem, a small time step τ may be required, which may significantly increase overall cost and makes the method less attractive. We note that one way for increasing the step size is to use implicit schemes based on some rational approximations of the exponential (see [26], for example), which typically requires inverting a certain matrix. When an inexact inverse with an iterative solver is used in an implicit scheme, the Krylov subspace methods appear more competitive overall; see [14, 23].

We observe that the matrix spectral distribution, such as the condition number and spectral gaps, plays a dominant role in determining convergence behavior of the Krylov subspace methods for other linear algebra problems. For example, if a symmetric positive definite matrix A has a small condition number, then the Krylov subspace method (i.e., the conjugate gradient method) for the linear system $Ax = b$ converges rapidly. This property can be understood by observing that the eigenvalues of a well-conditioned matrix A are clustered near some point λ_0 or $A = \lambda_0 I + E$ for some small E. For such a matrix, computing the exponential $\exp(-\tau A)v = e^{-\tau\lambda_0}\exp(-\tau E)v$ is also reduced to the easier problem for $\exp(-\tau E)v$. Therefore, the spectral distribution, in addition to the norm, can be expected to influence the convergence of the Krylov subspace method for $e^{-\tau A}v$ as well.

In this paper, we consider symmetric matrices and present new a posteriori and a priori error bounds for the Lanczos method and the shift-and-invert Lanczos method. Our a priori bound demonstrates dependence of convergence of these two methods on the condition numbers of the related matrices. Indeed, for the Lanczos method, we show that it converges at least at the same convergence rate as the conjugate gradient method for A, regardless of what the norm of τA is. Our numerical tests confirm this convergence behavior. We remark that as in the Krylov subspace methods for other linear algebra problems, such a convergence property may potentially have implications in preconditioning, i.e., transforming the matrix exponential problem (1.1) to another one for accelerated convergence. However, at the moment, it is not clear what transformation can accomplish this.

The paper is organized as follows. In section 2, we first present some decay bounds obtained in [2] on entries of functions of banded matrices, which are used in this paper to derive new a priori bounds. We then present new error bounds for the Lanczos method in section 3 and for the shift-and-invert Lanczos method in section 4. We present some numerical examples to illustrate our bounds in section 5, followed by some concluding remarks in section 6.

Notation. Throughout, e_i denotes the ith coordinate vector, the dimension of which is determined from context. I denotes an identity matrix and I_n specifies the $n \times n$ identity matrix. For a symmetric matrix A, $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote its

largest and smallest eigenvalues, respectively. $\|\cdot\|$ always denotes the 2-norm for both vectors and matrices.

2. Decay bounds for entries of matrix functions. For certain functions of a banded matrix, there is an interesting property that the entries that are away from the main diagonal decrease very rapidly. Earlier results concerning the inverse function can be found in [6, 7] and results for more general matrix functions were obtained in [2, 3]. In this paper, this peculiar decay property will be used to explain convergence of the Krylov subspace method for approximating the matrix exponential. In this section, we present the general decay bound of Benzi and Golub [2] on the elements away from the main diagonal for certain functions of banded matrices. It will be the basis of several bounds related to approximation errors of the Krylov subspace methods in the next section.

Let F be an analytic function on a simply connected region of the complex plane that contains the interval $[-1, 1]$. Then, there exist ellipses with foci in -1 and 1 such that F is analytic in the interiors of the ellipses. Let $\alpha > 1$ be the major half axis and $\beta > 0$ be the minor half axis of such an ellipse. Since we have $\alpha^2 - \beta^2 = 1$, the ellipse can be defined from one parameter, say,

$$\chi = \alpha + \beta > 1.$$

We denote the ellipse so defined by $\mathcal{E}_\chi$. Then $\mathcal{E}_\chi$ has the major and minor half axes given by

$$\alpha = \frac{1}{2}\left(\chi + \frac{1}{\chi}\right) \quad \text{and} \quad \beta = \sqrt{\alpha^2 - 1}. \tag{2.1}$$

In particular, α and β are increasing functions of χ for $\chi > 1$. Therefore, we have

$$\mathcal{E}_\chi \subset \mathcal{E}_{\bar{\chi}} \ \text{ if } \ 1 < \chi < \bar{\chi}. \tag{2.2}$$

Below, we say that a matrix $B = (b_{ij})$ is k-banded if $b_{ij} = 0$ whenever $|i-j| > k/2$. For a k-banded matrix B, the following theorem from [2] bounds the elements of $F(B)$.

THEOREM 2.1 (Benzi and Golub [2, Theorem 2.2]). *Let F be an analytic function in the interior of the ellipse $\mathcal{E}_\chi$ and continuous on $\mathcal{E}_\chi$ for some $\chi > 1$. Assume that $F(z)$ is real for real z. Let B be symmetric, k-banded, and such that $[-1, 1]$ is the smallest interval containing the spectrum of B. Let $q = \frac{1}{\chi}$, $\rho = q^{\frac{2}{k}}$, and*

$$K = \max\{K_0, \|F(B)\|_2\}$$

with $K_0 = \frac{2\chi M(\chi)}{\chi - 1}$, where $M(\chi) = \max_{z\in\mathcal{E}_\chi} |F(z)|$. Then we have

$$|(F(B))_{ij}| \le K\rho^{|i-j|}.$$

By definition, a tridiagonal matrix is 2-banded. Then, if B in the theorem is a tridiagonal matrix, the above bound is simplified to

$$|(F(B))_{ij}| \le Kq^{|i-j|}. \tag{2.3}$$

Therefore, the bound shows that an entry of $F(B)$ decreases in absolute value at the rate of ρ (or q in the tridiagonal case) as it moves away from the main diagonal. In particular, we can expect that the $(m, 1)$ entry of $F(B)$ is tiny compared to its norm.

This is the property that we will use to explain convergence of the Lanczos method or the shift-and-invert Lanczos method for approximating $e^{-\tau A}v$.

We finally note that the rate of decay ρ is determined by the size of ellipse $\mathcal{E}_\chi$. If F is analytic on a larger ellipse $\mathcal{E}_\chi$, we have a smaller q, but $M(\chi)$ and hence K may be larger as well.

3. Error bounds for the Lanczos method. In this section, we consider the Lanczos method that was originally introduced in [14, 24] for approximating (1.1), i.e., $w(\tau) = e^{-\tau A}v$, where A is a symmetric positive semidefinite matrix, $v \in \mathbb{R}^n$, and $\tau > 0$. Without loss of generality, we assume that $\|v\| = 1$.

With $v_1 = v$, the Lanczos algorithm applied to A and v_1 generates an orthonormal basis $v_1, v_2, \ldots, v_m, v_{m+1}$ for the Krylov subspace

$$K_{m+1}(A, v) := \text{span}\{v, Av, A^2v, \ldots, A^m v\}$$

and an $m \times m$ tridiagonal matrix T_m such that

$$AV_m = V_m T_m + \beta_{m+1} v_{m+1} e_m^T, \tag{3.1}$$

where $V_m = [v_1, v_2, \ldots, v_m]$; see [8, Algorithm 6.10] for a detailed algorithm. Recall that $e_i \in \mathbb{R}^n$ is the ith coordinate vector. The vector $V_m V_m^T e^{-\tau A}v$ is the orthogonal projection of $e^{-\tau A}v$ on $K_m(A, v)$, which is the closest approximation to $e^{-\tau A}v$ from $K_m(A, v)$. The Lanczos method further approximates it as

$$V_m V_m^T e^{-\tau A}v = V_m V_m^T e^{-\tau A} V_m e_1 \approx V_m e^{-\tau T_m} e_1.$$

We call

$$w_m(\tau) := V_m e^{-\tau T_m} e_1 \tag{3.2}$$

the Lanczos approximation to $w(\tau) = e^{-\tau A}v$. The following is an a priori bound on the error due to Saad [24, Corollary 4.6]:

$$\|w(\tau) - w_m(\tau)\| \le \frac{2}{m!}\left(\frac{\tau\|A\|}{2}\right)^m. \tag{3.3}$$

The bound suggests that the Lanczos method converges rapidly if $\tau\|A\|$ is not too large. When $\tau\|A\|$ is large, the bound actually increases initially, although its limit as $m \to \infty$ is 0. Several more refined bounds have been obtained in [9, 10, 15] but they all suggest similar dependence of convergence on $\tau\|A\|$. This appears to limit applicability of the Lanczos method to problems where $\tau\|A\|$ is not too large. Noting that the Lanczos method has the finite termination property, i.e., $w_n = w$, we have used terms like convergence to refer to reduction of the error as m increases to n.

We shall show that convergence of the Lanczos method also depends on the condition number of A. We first present the following a posteriori error bound on the Lanczos method, which relates the error to the $(m, 1)$ entry of e^{-tT_m}.

THEOREM 3.1. *Let A be a symmetric positive semidefinite matrix and let $w_m(\tau)$ be the Lanczos approximation to $w(\tau)$ as defined in* (3.2) *and* (3.1). *Then, for any α with $0 \le \alpha \le \tau$, the error satisfies*

$$\|w(\tau) - w_m(\tau)\| \le \beta_{m+1}\left(h_{0,\alpha} e^{(\alpha-\tau)\lambda_{\min}(A)}\alpha + h_{\alpha,\tau}(\tau - \alpha)\right), \tag{3.4}$$

where

$$h_{t_1,t_2} = \max_{t_1 \le t \le t_2} |h(t)| \quad \textit{for} \quad 0 \le t_1 \le t_2 \le \tau \tag{3.5}$$

and $h(t) = e_m^T e^{-tT_m} e_1$.

Proof. First, $w(t) = e^{-tA}v$ is the solution to

$$w'(t) = -Aw(t), \qquad w(0) = v.$$

Since $w_m'(t) = -V_m T_m e^{-tT_m} e_1$, we obtain from (3.1) that

$$\begin{aligned} w_m'(t) &= -(AV_m - \beta_{m+1} v_{m+1} e_m^T) e^{-tT_m} e_1 \\ &= -AV_m e^{-tT_m} e_1 + \beta_{m+1} v_{m+1} e_m^T e^{-tT_m} e_1 \\ &= -Aw_m(t) + \beta_{m+1}(e_m^T e^{-tT_m} e_1) v_{m+1}. \end{aligned}$$

Subtracting the above two equations and writing $E_m(t) = w(t) - w_m(t)$, we have

$$E_m'(t) = -AE_m(t) - \beta_{m+1}(e_m^T e^{-tT_m} e_1) v_{m+1}.$$

Solving this initial value problem with $E_m(0) = w(0) - w_m(0) = 0$, we obtain

$$\begin{aligned} E_m(\tau) &= \int_0^\tau e^{(t-\tau)A} \left(-\beta_{m+1}(e_m^T e^{-tT_m} e_1) v_{m+1} \right) dt \\ &= -\beta_{m+1} \int_0^\tau h(t) e^{(t-\tau)A} v_{m+1} dt. \end{aligned}$$

Separating the integral into two subintervals and bounding them separately, we have

$$\begin{aligned} \|E_m(\tau)\| &\le \beta_{m+1} \left\| \int_0^\alpha h(t) e^{(t-\tau)A} dt + \int_\alpha^\tau h(t) e^{(t-\tau)A} dt \right\| \\ &\le \beta_{m+1} \left(h_{0,\alpha} \int_0^\alpha \|e^{(t-\tau)A}\| dt + h_{\alpha,\tau} \int_\alpha^\tau \|e^{(t-\tau)A}\| dt \right) \\ &\le \beta_{m+1} \left(h_{0,\alpha} \int_0^\alpha e^{(t-\tau)\lambda_{\min}(A)} dt + h_{\alpha,\tau} \int_\alpha^\tau e^{(t-\tau)\lambda_{\min}(A)} dt \right) \\ &= \beta_{m+1} \left(h_{0,\alpha} \frac{e^{(\alpha-\tau)\lambda_{\min}(A)} - e^{-\tau\lambda_{\min}(A)}}{\lambda_{\min}(A)} + h_{\alpha,\tau} \frac{1 - e^{(\alpha-\tau)\lambda_{\min}(A)}}{\lambda_{\min}(A)} \right) \\ &= \beta_{m+1} \left(h_{0,\alpha} e^{(\alpha-\tau)\lambda_{\min}(A)} \frac{1 - e^{-\alpha\lambda_{\min}(A)}}{\lambda_{\min}(A)} + h_{\alpha,\tau} \frac{1 - e^{(\alpha-\tau)\lambda_{\min}(A)}}{\lambda_{\min}(A)} \right). \end{aligned}$$

By noting that $1 - e^{-x} \le x$ for any $x \ge 0$, (3.4) is proved. ∎

An optimal bound can be obtained by minimizing (3.4) with respect to α, but a sufficiently good one can be derived by using the minimum of (3.4) over some equally spaced points in $[0, \tau]$. To use the bound as a practical estimate of errors, we may replace $\lambda_{\min}(A)$ in (3.4) by a lower bound, say, 0. The other terms used in the bound are all computable at the end of step m of the Lanczos algorithm, although $\max_{t_1 \le t \le t_2} |h(t)|$ can only be approximated.

If $\alpha = \tau$, the bound above reduces to

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

This is similar to a posteriori error estimates derived by Saad [24, section 5]. However, since the values of $h(t)$ may be very small for small t, separately bounding $h(t)$ over two intervals in (3.4) may improve significantly over the above bound. Indeed, (3.4) with an optimal α provides a quite sharp estimate of the actual error.

From our discussions in section 2, $e_m^T e^{-tT_m} e_1$, the $(m,1)$ entry of e^{-tT_m}, is expected to have a decay property as m increases. Indeed, it can be bounded as follows.

LEMMA 3.1. *Let $T_m \neq 0$ be an $m \times m$ symmetric positive semidefinite tridiagonal matrix and let $a = \lambda_{\min}(T_m)$ and $b = \lambda_{\max}(T_m)$ be the smallest and the largest eigenvalues of T_m, respectively. Then for any $t \geq 0$ and any q with $0 < q < 1$, we have*

$$|e_m^T e^{-tT_m} e_1| \leq \frac{2}{1-q} e^{-\frac{t\gamma}{4q}} q^{m-1}, \tag{3.6}$$

where $\gamma = (b-a)(q-q_0)(q_0^{-1}-q)$ and $q_0 = \frac{\sqrt{b}-\sqrt{a}}{\sqrt{b}+\sqrt{a}}$. In particular, if $a \neq 0$, we have

$$|e_m^T e^{-tT_m} e_1| \leq (\sqrt{\kappa}+1)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m-1}, \tag{3.7}$$

where $\kappa = b/a$.

Proof. Let $f(\lambda) = e^{-t\lambda}$ and $F = f \circ \psi^{-1}$, where $\psi : \mathbb{C} \to \mathbb{C}$ is defined by

$$\psi(\lambda) = \frac{2\lambda - (a+b)}{b-a}.$$

Then $\psi([a,b]) = [-1,1]$. Let $B = \psi(T_m)$. Then B is symmetric tridiagonal and its spectrum is contained in $[-1,1]$. Let $\chi = \frac{1}{q}$. Since F is analytic on $\mathbb{C}$, F is analytic in the interior of the ellipse $\mathcal{E}_\chi$ (as defined in (2.1)) and continuous on $\mathcal{E}_\chi$. Applying Theorem 2.1 with $k = 2$ to the function F and the matrix B, we have the following bound on the $(m,1)$ entry of the matrix e^{-tT_m}:

$$|e_m^T e^{-tT_m} e_1| = |e_m^T F(B) e_1| \leq K q^{m-1}, \tag{3.8}$$

where

$$K = \max\{K_0, \|F(B)\|\}, \quad K_0 = \frac{2\chi M(\chi)}{\chi - 1}, \quad \text{and } M(\chi) = \max_{z \in \mathcal{E}_\chi} |F(z)|.$$

We bound K now. For any $z = x + iy \in \mathcal{E}_\chi$, set

$$u = \frac{(b-a)x + a + b}{2}, \qquad v = \frac{b-a}{2} y,$$

i.e., $u + iv = \psi^{-1}(z)$. Then

$$|F(z)| = |e^{-t(u+iv)}| = e^{-tu}.$$

Note that the major half axis of $\mathcal{E}_\chi$ is $\alpha := \frac{1}{2}(\chi + \frac{1}{\chi}) = \frac{1}{2}(q + \frac{1}{q})$; see (2.1). Then we have $-\alpha \leq x \leq \alpha$. Furthermore, it can be checked that

$$\frac{b-a}{4}\left(q + \frac{1}{q}\right) - \frac{a+b}{2} = -\frac{b-a}{4q}(q-q_0)(q_0^{-1}-q) = -\frac{\gamma}{4q}.$$

Thus, we have

$$M(\chi) = \max_{z\in\mathcal{E}_\chi} e^{t((a-b)x-a-b)/2} = \max_{-\alpha\le x\le\alpha} e^{t((a-b)x-a-b)/2}$$
$$= e^{t\left((b-a)\left(q+\frac{1}{q}\right)-2(a+b)\right)/4} = e^{-t\gamma/(4q)}.$$

It follows that

$$K_0 = \frac{2\chi M(\chi)}{\chi-1} = \frac{2}{1-q}e^{-t\gamma/(4q)}.$$

Furthermore,

$$\|F(B)\| = \|e^{-tT_m}\| = e^{-ta} \le e^{-t(a+b)/2} \le e^{-t\gamma/(4q)}.$$

Thus

$$K = \frac{2}{1-q}e^{-t\gamma/(4q)}. \tag{3.9}$$

Substituting this into (3.8), the first bound is proved. If $a \neq 0$, $q_0 = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Then, the second bound (3.7) follows from substituting $q = q_0$ into the first bound (3.8). ∎

The lemma above shows that $e_m^T e^{-tT_m} e_1$ is reduced at least at the rate of q_0 as m increases. However, we can choose q in the first bound (3.6) to be smaller than q_0, resulting in a faster decreasing factor q^{m-1}, but the coefficient $e^{-\frac{t\gamma}{4q}}$ (with $\gamma < 0$ now) may be very large, offsetting any decrease in q^m. However, as long as $q < q_0$ is such that $e^{-\frac{t\gamma}{4q}}$ is not too large, this may still lead to a better bound. Specifically, given any $\delta > 0$, setting

$$q = \left(\frac{1}{q_0} + \frac{4\delta}{t(b-a)}\right)^{-1} < q_0$$

leads to $-\frac{t\gamma}{4q} \le t(b-a)(q_0 q^{-1} - 1)q_0^{-1}/4 = \delta$ and hence

$$|e_m^T e^{-tT_m} e_1| \le \frac{2}{1-q}e^{\delta} q^{m-1}. \tag{3.10}$$

Using a modest value for δ here (say, less than 10) may result in an overall stronger bound (3.10) with a modest e^δ but much reduced q if $t(b-a)$ is not too large.

Lemma 3.1 demonstrates the influence of the condition number of T_m on the $(m,1)$ entry of e^{-tT_m}, but if tT_m has a small norm, e^{-tT_m} is close to I and then $e_m^T e^{-tT_m} e_1$ is close to 0 whatever its condition number. Therefore, $e_m^T e^{-tT_m} e_1$ also depends on the magnitude of $t\|T_m\|$. The next lemma demonstrates this dependence.

LEMMA 3.2. *Let T_m be an $m \times m$ symmetric positive semidefinite tridiagonal matrix. For any $t \ge 0$, we have*

$$|e_m^T e^{-tT_m} e_1| \le \frac{1}{(m-1)!}\left(\frac{tb}{2}\right)^{m-1}, \tag{3.11}$$

where $b = \lambda_{\max}(T_m)$ is the largest eigenvalue of T_m.

Proof. For any $m \times m$ tridiagonal matrix $\hat{T}_m$, by Lemma 3.1 of [28], $e_m^T \hat{T}_m^j e_1 = 0$ for $1 \le j \le m-2$. Then

$$\begin{aligned}
|e_m^T e^{-t\hat{T}_m} e_1| &= \left| \sum_{j=0}^{\infty} \frac{1}{j!} e_m^T (-t\hat{T}_m)^j e_1 \right| \\
&= \left| \sum_{j=m-1}^{\infty} \frac{1}{j!} e_m^T (-t\hat{T}_m)^j e_1 \right| \\
&\le \frac{1}{(m-1)!} \sum_{j=0}^{\infty} \frac{1}{j!} (t\|\hat{T}_m\|)^{j+m-1} \\
&= \frac{1}{(m-1)!} (t\|\hat{T}_m\|)^{m-1} e^{t\|\hat{T}_m\|}.
\end{aligned}$$

Applying this bound to $\hat{T}_m = T_m - \frac{b}{2} I$ as in [24], we have

$$\begin{aligned}
|e_m^T e^{-tT_m} e_1| &= e^{-t\frac{b}{2}} |e_m^T e^{-t\hat{T}_m} e_1| \le e^{-t\frac{b}{2}} \frac{1}{(m-1)!} (t\|\hat{T}_m\|)^{m-1} e^{t\|\hat{T}_m\|} \\
&= \frac{1}{(m-1)!} \left(\frac{tb}{2} \right)^{m-1},
\end{aligned}$$

where we note that $\|\hat{T}_m\| = b/2$. ∎

Now, we can obtain various a priori bounds by applying the bounds on $|e_m^T e^{-tT_m} e_1|$ to Theorem 3.1. In particular, for small α, $h_{0,\alpha}$ is small due to small $\alpha\|T_m\|$ and we use (3.10) or (3.11) to bound it. We use (3.7) to bound $h_{\alpha,\tau}$.

THEOREM 3.2. *Let A be an $n \times n$ symmetric positive definite matrix and let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be its eigenvalues. Let $w_m(\tau)$ be the Lanczos approximation to $w(\tau)$ as defined in* (3.2) *and* (3.1). *Then, for any α with $0 \le \alpha \le \tau$ and for any $\delta > 0$, the error of the Lanczos method satisfies*

$$\|w(\tau) - w_m(\tau)\| \le \alpha e^{(\alpha-\tau)\lambda_1} \|A\| \epsilon_1(m) + (\tau - \alpha)\|A\| \epsilon_2(m), \tag{3.12}$$

where

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

$q = (\frac{1}{q_0} + \frac{4\delta}{\alpha(\lambda_n - \lambda_1)})^{-1}$, $q_0 = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$, *and* $\kappa = \frac{\lambda_n}{\lambda_1}$.

Proof. For β_{m+1} that is defined from the Lanczos algorithm (3.1), we have

$$|\beta_{m+1}| = \|\beta_{m+1} v_{m+1} e_m^T\| = \|AV_m - V_m T_m\| \le \|AV_m\| \le \|A\|,$$

where we notice that $(V_m T_m)^T (AV_m - V_m T_m) = 0$ and hence

$$\|AV_m\|^2 = \|(AV_m - V_m T_m) + V_m T_m\|^2 \ge \|AV_m - V_m T_m\|^2.$$

Let $a = \lambda_{\min}(T_m)$ and $b = \lambda_{\max}(T_m)$ be the smallest and the largest eigenvalues of T_m, respectively, and let $\kappa_0 = \frac{b}{a}$. It follows from $T_m = V_m^T A V_m$ that $b \le \lambda_n$ and $a \ge \lambda_1$. Thus, $\kappa_0 \le \kappa$.

Now, applying Theorem 3.1, the error satisfies (3.4). From (3.11), we obtain $h_{0,\alpha} \le \frac{1}{(m-1)!}(\alpha b/2)^{m-1} \le \frac{1}{(m-1)!}(\alpha\lambda_n/2)^{m-1}$. We also obtain from (3.10) that $|e_m^T e^{-tT_m} e_1| \le \frac{2}{1-\hat{q}} e^{\delta} \hat{q}^{m-1} \le \frac{2}{1-q} e^{\delta} q^{m-1}$ for $0 \le t \le \alpha$, where $\hat{q} := (\frac{1}{q_0} + \frac{4\delta}{t(b-a)})^{-1} \le q$. Therefore, $h_{0,\alpha} \le \epsilon_1(m)$. On the other hand, it follows from (3.7) that

$$h_{\alpha,\tau} \le (\sqrt{\kappa_0}+1)\left(\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}\right)^{m-1} \le (\sqrt{\kappa}+1)\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{m-1}.$$

Substituting these into (3.4), we obtain (3.12). ∎

The parameter δ in the above bound should be chosen to be a modest number, say, $\delta = 5$ or 10, to balance the fast growth of e^{δ} with the decrease in q. The bound also allows choosing a parameter α. An optimal bound requires minimizing the bound (3.12) with respect to α, but a sufficiently good bound can be obtained by using minimum over a few equally spaced points in $[0, \tau]$. Indeed, for most problems we tested, the best bound is essentially given by either $\alpha = 0$ or $\alpha = \tau$, which yields ϵ_2 or ϵ_1 as a bound, respectively. The two bounds reflect two independent factors affecting the convergence of the Lanczos method. We discuss two situations:

1. If A is well-conditioned, ϵ_2 decreases rapidly at the rate of $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Then choosing $\alpha = 0$, the bound is given entirely by ϵ_2 as

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

   Therefore, regardless of how large the norm of τA is, the Lanczos method converges at least at the same rate as the conjugate gradient method. This is a property that cannot be inferred from previous bounds such as (3.3).
2. If A is not well-conditioned, then by choosing $\alpha = \tau$, the bound is given by $\tau\|A\|\epsilon_1$. Inspecting ϵ_1, we have good convergence if $\tau\lambda_n$ or $\tau(\lambda_n - \lambda_1)$ is small. Note that asymptotically for very large m, ϵ_1 always gives a better bound.

Thus, for a general matrix, our bound (3.12) combines two different forces driving the convergence of the Lanczos method. On the one hand, as a high-order polynomial approximation scheme, its convergence depends on the norm of the matrix and τ. On the other hand, as a projection method, which captures most important spectral information of A in the time propagation, its convergence also depends on the spectral distribution, i.e., the condition number. The method itself will achieve an optimal combination of the two factors through a weighted average, as indicated in our bound (3.12). Our numerical examples in section 5 confirm this behavior.

Finally, Theorem 3.2 considers a symmetric positive definite A, but Theorem 3.1, Lemma 3.1, and Lemma 3.2 are all valid for symmetric positive semidefinite matrices. When A is symmetric positive semidefinite, it is straightforward to see that the bound by ϵ_1 still holds and we have $\|w(\tau) - w_m(\tau)\| \le \tau\|A\|\epsilon_1$. However, ϵ_2 can no longer be used because the condition number may be undefined.

4. Shift-and-invert Lanczos method. One implication of Theorem 3.2 is on preconditioning, i.e., to accelerate convergence of the Lanczos method by transforming the problem into an equivalent one with a well-conditioned matrix. For the Krylov subspace method for solving linear systems, the condition number is reduced using the transformation $M^{-1}A$ for some preconditioner matrix $M \approx A$. Unfortunately, it is not clear whether and how the transformation $M^{-1}A$ can be used for $\exp(A)v$,

although there are some related works [4, 17] to indirectly use a preconditioner matrix M to compute $e^{-\tau A}v$.

Shifting the matrix A by a positive σ is another transformation that reduces the condition number. With the transformation $A+\sigma I$, we may use

$$e^{-\tau A}v = e^{\tau\sigma}e^{-\tau(A+\sigma I)}v. \tag{4.1}$$

Then, applying the Lanczos method to $e^{-\tau(A+\sigma I)}v$ will indeed result in faster convergence, which is confirmed in our numerical experiments. However, to compute $e^{-\tau A}v$, the corresponding approximate solution and hence the associated error for $e^{-\tau A}v$ need to be multiplied by $e^{\tau\sigma} > 1$, which turns out to cancel the reduction in the error achieved by using the shift. Indeed, the larger the shift σ is, the better the condition number of $A+\sigma I$ is and hence the faster convergence to $e^{-\tau(A+\sigma I)}v$, but also the larger e^{σ} is. The final approximation is unfortunately not improved by this simple transformation.

A somewhat related approach is to consider $(A+\sigma I)^{-1}$, which also has a smaller condition number for $\sigma > 0$. However, there is no simple relation like (4.1) for the shift-and-invert matrix $(A+\sigma I)^{-1}$. Instead one can construct an approximation of $e^{-\tau A}v$ from the Krylov subspace generated by $(A+\sigma I)^{-1}$. This is basically the shift-and-invert Lanczos method introduced by van den Eshof and Hochbruck [27] and by Moret and Novati [20], which we describe now.

For some $\sigma \geq 0$, applying m steps of the Lanczos algorithm to $(A+\sigma I)^{-1}$ and $v_1 = v$, we obtain

$$(A+\sigma I)^{-1}V_m = V_m T_m + \beta_{m+1}v_{m+1}e_m^T, \tag{4.2}$$

where the columns of $V_m = [v_1, v_2, \ldots, v_m]$ form an orthonormal basis of $K_m((A+\sigma I)^{-1}, v)$, and T_m is a tridiagonal matrix. Then, $V_m V_m^T e^{-\tau A}v$ is the projection of $e^{-\tau A}v$ on $K_m((A+\sigma I)^{-1}, v)$, and the shift-and-invert Lanczos method further approximates it as

$$V_m V_m^T e^{-\tau A}v = V_m V_m^T e^{-\tau(B^{-1}-\sigma I)}V_m e_1 \approx V_m e^{-\tau(T_m^{-1}-\sigma I)}e_1,$$

where $B = (A+\sigma I)^{-1}$. We call

$$w_m^{SIL}(\tau) := V_m e^{-\tau(T_m^{-1}-\sigma I)}e_1 \tag{4.3}$$

the shift-and-invert Lanczos approximation to $w(\tau) = e^{-\tau A}v$.

The shift-and-invert Lanczos method has been derived primarily from the points of view of a special rational approximation in [20, 27] but has also been called a preconditioning scheme in [27]. Its convergence has been analyzed in [20, 27] by bounding $E_m^{SIL}(\tau)$ in terms of the error of a certain rational approximation to e^{-t} with σ as a parameter. Some a posteriori error estimates have also been discussed in [27]. We note that the method (4.3) cannot be formulated as a standard Lanczos method and therefore its convergence property cannot be analyzed or inferred from the existing theory for the Lanczos method.

Here, in light of the dependence of convergence of the Lanczos method on the condition number of A and the fact that the shift-and-invert transformation reduces the condition number, we shall analyze the shift-and-invert Lanczos method by considering it as a way to reduce the condition number. Specifically, we shall relate the error to the condition number of $(A+\sigma I)^{-1}$ or equivalently of $A+\sigma I$. We first present the following a posteriori error bound.

THEOREM 4.1. *Let A be an $n \times n$ symmetric positive semidefinite matrix and let $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_n$ be its eigenvalues. Let $w_m^{SIL}(\tau)$ be the shift-and-invert Lanczos approximation to $w(\tau) = e^{-\tau A}$ as defined in (4.2) and (4.3) and let $h(t) = e_m^T T_m^{-1} e^{-t(T_m^{-1}-\sigma I)} e_1$. For any α with $0 \le \alpha \le \tau$, we have*

$$\|w(\tau) - w_m^{SIL}(\tau)\| \le \beta_{m+1}(\lambda_n + \sigma)\left(h_{0,\alpha} e^{(\alpha-\tau)\lambda_1}\alpha + h_{\alpha,\tau}(\tau-\alpha)\right), \tag{4.4}$$

where

$$h_{t_1,t_2} = \max_{t_1 \le t \le t_2} |h(t)| \quad \text{for } 0 \le t_1 < t_2 \le \tau. \tag{4.5}$$

Proof. We rewrite (4.2) as

$$V_m(T_m^{-1} - \sigma I) = AV_m + \beta_{m+1}(A+\sigma I)v_{m+1}e_m^T T_m^{-1}.$$

Then, we have

$$\begin{aligned}
\frac{d}{dt}w_m^{SIL}(t) &= -V_m(T_m^{-1}-\sigma I)e^{-t(T_m^{-1}-\sigma I)}e_1 \\
&= -(AV_m + \beta_{m+1}(A+\sigma I)v_{m+1}e_m^T T_m^{-1})e^{-t(T_m^{-1}-\sigma I)}e_1 \\
&= -AV_m e^{-t(T_m^{-1}-\sigma I)}e_1 - \beta_{m+1}(A+\sigma I)v_{m+1}h(t) \\
&= -Aw_m^{SIL}(t) - \beta_{m+1}h(t)(A+\sigma I)v_{m+1}.
\end{aligned}$$

Let $E_m^{SIL}(t) = w(t) - w_m^{SIL}(t)$. We have

$$\frac{d}{dt}E_m^{SIL}(t) = -AE_m^{SIL}(t) + \beta_{m+1}h(t)(A+\sigma I)v_{m+1}.$$

Solving the above ODE with the initial condition $E_m^{SIL}(0) = 0$, we obtain

$$E_m^{SIL}(\tau) = \beta_{m+1}\left(\int_0^\tau h(t)e^{(t-\tau)A}(A+\sigma I)dt\right)v_{m+1}. \tag{4.6}$$

For a fixed α, we bound this error as

$$\begin{aligned}
\|E_m^{SIL}(\tau)\| &\le \beta_{m+1}\left\|\int_0^\alpha h(t)e^{(t-\tau)A}(A+\sigma I)dt + \int_\alpha^\tau h(t)e^{(t-\tau)A}(A+\sigma I)dt\right\| \\
&\le \beta_{m+1}\left(h_{0,\alpha}\int_0^\alpha \|e^{(t-\tau)A}(A+\sigma I)\|dt + h_{\alpha,\tau}\int_\alpha^\tau \|e^{(t-\tau)A}(A+\sigma I)\|dt\right) \\
&\le \beta_{m+1}(\lambda_n+\sigma)\left(h_{0,\alpha}\frac{e^{(\alpha-\tau)\lambda_1} - e^{-\tau\lambda_1}}{\lambda_1} + h_{\alpha,\tau}\frac{1-e^{(\alpha-\tau)\lambda_1}}{\lambda_1}\right) \\
&= \beta_{m+1}(\lambda_n+\sigma)\left(h_{0,\alpha}e^{(\alpha-\tau)\lambda_1}\frac{1-e^{-\alpha\lambda_1}}{\lambda_1} + h_{\alpha,\tau}\frac{1-e^{(\alpha-\tau)\lambda_1}}{\lambda_1}\right) \\
&\le \beta_{m+1}(\lambda_n+\sigma)\left(h_{0,\alpha}e^{(\alpha-\tau)\lambda_1}\alpha + h_{\alpha,\tau}(\tau-\alpha)\right).
\end{aligned}$$

The theorem is proved. ∎

As before, we can derive a near optimal bound by minimizing (4.4) with respect to α over some equally spaced discrete points in $[0,\tau]$. We also note that (4.6) has been presented in [27], from which an a posteriori error estimate is derived by replacing

$e^{(t-\tau)A}$ by the leading term of its Taylor series. The bound obtained here follows a different approach by relating the error to $h(t)$. We bound $h(t)$ now.

LEMMA 4.1. *Let T_m be an $m \times m$ symmetric positive definite tridiagonal matrix with $a = \lambda_{\min}(T_m)$ and $b = \lambda_{\max}(T_m)$ being its smallest and largest eigenvalue, respectively. Let $\kappa = \frac{b}{a}$ and $q_0 = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$. Then, for any fixed q with $q_0 < q < 1$, we have*

$$|e_m^T T_m^{-1} e^{-tT_m^{-1}} e_1| \le \frac{8}{(1-q)\gamma} e^{-\frac{t\gamma}{4q(a+b)^2}} q^m, \tag{4.7}$$

where $\gamma = (b-a)(q-q_0)(q_0^{-1}-q) > 0$.

Proof. Let $\psi(\lambda) = \frac{2\lambda-(a+b)}{b-a}$ and $B = \psi(T_m)$. Then B is a symmetric tridiagonal matrix with the spectrum contained in $[-1,1]$. Let $\mathcal{E}_{\bar{\chi}}$ be the ellipse that has foci at -1 and 1 with major half axis $\alpha_0 = (b+a)/(b-a)$ (see (2.1)). Then

$$\bar{\chi} = \alpha_0 + \sqrt{\alpha_0^2 - 1} = \frac{b+a}{b-a} + \sqrt{\left(\frac{b+a}{b-a}\right)^2 - 1} = \frac{\sqrt{\kappa}+1}{\sqrt{\kappa}-1}.$$

Let $f(\lambda) = \lambda^{-1} e^{-t\lambda^{-1}}$ and $F = f \circ \psi^{-1}$. Then

$$F(z) = \left(\frac{b-a}{2} z + \frac{a+b}{2}\right)^{-1} e^{-t\left(\frac{b-a}{2} z + \frac{a+b}{2}\right)^{-1}}.$$

Clearly, F is analytic in the interior of the ellipse $\mathcal{E}_{\bar{\chi}}$. Let $\chi = 1/q$. Then, $1 < \chi < \bar{\chi}$ and hence $\mathcal{E}_\chi \subset \mathcal{E}_{\bar{\chi}}$; see (2.2). Thus, the function F is analytic inside $\mathcal{E}_\chi$ and continuous on $\mathcal{E}_\chi$. It follows from Theorem 2.1 and $f(T_m) = F(B)$ that

$$|e_m^T T_m^{-1} e^{-tT_m^{-1}} e_1| = |e_m^T F(B) e_1| \le K q^{m-1}, \tag{4.8}$$

where $K = \max\{\frac{2\chi M(\chi)}{\chi-1}, \|F(B)\|_2\}$ and $M(\chi) = \max_{z \in \mathcal{E}_\chi} |F(z)|$.

We bound K now. Let $z = x + iy \in \mathcal{E}_\chi$. Set $u + iv = \psi^{-1}(x+iy)$, i.e.,

$$u = \frac{(b-a)x + a + b}{2}, \qquad v = \frac{b-a}{2} y.$$

Since the major half axis of $\mathcal{E}_\chi$ is $\alpha := \frac{1}{2}(\chi + \frac{1}{\chi}) < \frac{1}{2}(\bar{\chi} + \frac{1}{\bar{\chi}}) = \frac{b+a}{b-a}$, it follows from $-\alpha \le x \le \alpha$ that

$$u \ge -\frac{b-a}{4}\left(\chi + \frac{1}{\chi}\right) + \frac{a+b}{2} = \frac{b-a}{4q}(q-q_0)(q_0^{-1}-q) = \frac{\gamma}{4q}, \tag{4.9}$$

where we note that $q = 1/\chi$. Then

$$|F(z)| = |(u+iv)^{-1} e^{-t(u+iv)^{-1}}| = \frac{1}{\sqrt{u^2+v^2}} |e^{-t\frac{u-iv}{u^2+v^2}}| = \frac{1}{\sqrt{u^2+v^2}} e^{-t\frac{u}{u^2+v^2}}.$$

Since $u+iv$ is contained in the ellipse $\psi^{-1}(\mathcal{E}_{\bar{\chi}})$, which has foci at a and b with the major half axis equal to $\frac{1}{2}(a+b)$, and since this ellipse $\psi^{-1}(\mathcal{E}_{\bar{\chi}})$ is contained in the disk centered at $\frac{1}{2}(a+b)$ with the radius $\frac{1}{2}(a+b)$, we have $u^2+v^2 \le (a+b)^2$. Thus,

$$|F(z)| \le \frac{1}{u} e^{-t\frac{u}{(a+b)^2}} \le \frac{4q}{\gamma} e^{-\frac{t\gamma}{4q(a+b)^2}}.$$

Hence,

$$\frac{2\chi M(\chi)}{\chi-1}=\frac{2}{1-q}\max_{z\in\mathcal{E}_\chi}|F(z)|\le\frac{8q}{(1-q)\gamma}e^{-\frac{t\gamma}{4q(a+b)^2}}.$$

On the other hand, we have

$$\|F(B)\|=\|T_m^{-1}e^{-tT_m^{-1}}\|\le\frac{1}{a}e^{-t\frac{1}{b}}.$$

Finally, noting that $\gamma/(4q)=-\frac{b-a}{4}(\chi+\frac{1}{\chi})+\frac{a+b}{2}\le-\frac{b-a}{2}+\frac{a+b}{2}=a$ and $\frac{\gamma}{4q(a+b)^2}\le\frac{a}{(a+b)^2}\le\frac{1}{b}$, we have $\frac{8q}{(1-q)\gamma}e^{-\frac{t\gamma}{4q(a+b)^2}}\ge\frac{1}{a}e^{-t\frac{1}{b}}$ and hence

$$K=\max\left\{\frac{2\chi M(\chi)}{\chi-1},\|F(B)\|_2\right\}\le\frac{8q}{(1-q)\gamma}e^{-\frac{t\gamma}{4q(a+b)^2}},$$

which together with (4.8) proves (4.7). ∎

Finally, as in section 3, combining Theorem 4.1 with Lemma 4.1, we obtain an a priori error bound.

THEOREM 4.2. *Let A be an $n\times n$ symmetric positive semidefinite matrix and let $\lambda_1\le\lambda_2\le\cdots\le\lambda_n$ be its eigenvalues. For any $\sigma>0$, let*

$$\kappa=\frac{\lambda_n+\sigma}{\lambda_1+\sigma}\quad\textit{and}\quad q_0=\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}.$$

For any fixed q with $q_0<q<1$, the error of the shift-and-invert Lanczos method to the matrix exponential $e^{-\tau A}v$ as defined in (4.2) *and* (4.3) *satisfies*

$$\|w(\tau)-w_m^{SIL}(\tau)\|\le\frac{8(\lambda_n+\sigma)^2\mu}{(1-q)\gamma}q^m, \tag{4.10}$$

where $\gamma=(\lambda_n-\lambda_1)(q-q_0)(q_0^{-1}-q)$ and

$$\mu=\min_{0\le\alpha\le\tau}\left(\alpha e^{(\alpha-\tau)\lambda_1+\alpha\sigma}+(\tau-\alpha)e^{-\frac{\alpha\gamma}{16q\kappa}+\tau\sigma}\right)\le\tau e^{\tau\sigma}.$$

Proof. By using (4.2), β_{m+1} can be bounded as

$$|\beta_{m+1}|=\|\beta_{m+1}v_{m+1}e_m^T\|=\|(A+\sigma I)^{-1}V_m-V_mT_m\|\le\|(A+\sigma I)^{-1}V_m\|\le\frac{1}{\lambda_1+\sigma},$$

where we note that $(V_mT_m)^T\left((A+\sigma I)^{-1}V_m-V_mT_m\right)=0$ and hence

$$\|(A+\sigma I)^{-1}V_m\|^2=\|(A+\sigma I)^{-1}V_m-V_mT_m+V_mT_m\|^2\ge\|(A+\sigma I)^{-1}V_m-V_mT_m\|^2.$$

Let $a=\lambda_{\min}(T_m)$ and $b=\lambda_{\max}(T_m)$ be the smallest and the largest eigenvalues of T_m, respectively. It follows from $T_m=V_m^T(A+\sigma I)^{-1}V_m$ that $b\le(\lambda_1+\sigma)^{-1}$ and $a\ge(\lambda_n+\sigma)^{-1}$. Let $\kappa_0=b/a$ and $\hat{q}_0=\frac{\sqrt{\kappa_0}-1}{\sqrt{\kappa_0}+1}$. Then $\kappa_0\le\kappa$ and hence $q_0\ge\hat{q}_0$. Now, applying Lemma 4.1 to T_m with $q>q_0\ge\hat{q}_0$, we obtain

$$|e_m^TT_m^{-1}e^{-tT_m^{-1}}e_1|\le\frac{8}{(1-q)\gamma_0}e^{-\frac{t\gamma_0}{4q(a+b)^2}}q^m,$$

where $\gamma_0 = (\lambda_n - \lambda_1)(q - \hat{q}_0)(\hat{q}_0^{-1} - q)$. It follows that

$$h_{0,\alpha} := \max_{0\le t\le\alpha} |e_m^T T_m^{-1} e^{-tT_m^{-1}} e_1| e^{t\sigma} \le \frac{8}{(1-q)\gamma_0} e^{\alpha\sigma} q^m$$

and

$$\begin{aligned} h_{\alpha,\tau} &:= \max_{\alpha\le t\le\tau} |e_m^T T_m^{-1} e^{-tT_m^{-1}} e_1| e^{t\sigma} \\ &\le \frac{8}{(1-q)\gamma_0} e^{-\frac{\alpha\gamma_0}{4q(a+b)^2}} e^{\tau\sigma} q^m \\ &\le \frac{8}{(1-q)\gamma_0} e^{-\frac{\alpha\gamma_0(\lambda_1+\sigma)^2}{16q} + \tau\sigma} q^m. \end{aligned}$$

Now, as in (4.9), it can be checked that $\frac{\gamma_0}{4q} = (\frac{a-b}{4}(\frac{1}{q} + q) + \frac{a+b}{2})$. Thus, we have

$$\begin{aligned} \frac{\gamma_0}{4q} &= \frac{a}{2}\left(1 + \frac{1}{2}\left(\frac{1}{q} + q\right)\right) + \frac{b}{2}\left(1 - \frac{1}{2}\left(\frac{1}{q} + q\right)\right) \\ &\ge \frac{1}{2(\lambda_n + \sigma)}\left(1 + \frac{1}{2}\left(\frac{1}{q} + q\right)\right) + \frac{1}{2(\lambda_1 + \sigma)}\left(1 - \frac{1}{2}\left(\frac{1}{q} + q\right)\right) \\ &= \frac{1}{2(\lambda_1 + \sigma)(\lambda_n + \sigma)}\left(\frac{\lambda_1 - \lambda_n}{2}\left(\frac{1}{q} + q\right) + \lambda_1 + \lambda_n + 2\sigma\right) \\ &= \frac{(\lambda_n - \lambda_1)(q - q_0)(q_0^{-1} - q)}{4q(\lambda_1 + \sigma)(\lambda_n + \sigma)} = \frac{\gamma}{4q(\lambda_1 + \sigma)(\lambda_n + \sigma)}. \end{aligned}$$

Substituting these into (4.4) of Theorem 4.1, the theorem is proved. ∎

The bound in the above theorem involves choosing q. It can be chosen to be as close to q_0 as one wishes, but as $q \to q_0$, $\gamma \to 0$. However, choosing, for example, $q = 1.01q_0$ will result in a moderate factor $1/\gamma$ with $\gamma = 0.01(\lambda_n - \lambda_1)(1 - 1.01q_0^2)$, while $q^{m-1} = 1.01^m q_0^{m-1}$ decreases effectively at the same rate as q_0. The parameter α can again be chosen as some equally spaced points in $[0, \tau]$ and we use the best corresponding bound.

Our bound directly relates convergence of the shift-and-invert Lanczos method to σ through the condition number κ of $A - \sigma I$, but σ also affects the bound through the coefficient μ. We do not have a simple upper bound for μ other than $\mu \le \tau e^{\tau\sigma}$, which may become very large if $\tau\sigma$ is not small. In that case, (4.10) results in a very pessimistic bound of the actual error; see the numerical examples in section 5. Qualitatively, the shift σ affects the bound and hence convergence in two ways. That is, it improves the condition number on the one hand, but it also increases the coefficient μ of the bound quickly on the other hand. Our numerical examples confirm such convergence behavior. It is difficult to determine an optimal σ from our current bound, unfortunately.

5. Numerical examples. In this section, we present some numerical examples to demonstrate error bounds obtained in this paper. All numerical tests were carried out on a PC with an AMD Athlon processor in MATLAB (R2012a) with machine precision $\approx 2 \cdot 10^{-16}$.

We shall use in our examples diagonal matrices and discretized Laplacian matrices for which $e^{-\tau A}$ is readily available. We shall compare the approximation errors with the new a posteriori and a priori bounds. In our a posteriori bounds, we need to

compute $h_{\alpha_1,\alpha_2} = \max_{\alpha_1 \le t \le \alpha_2} |h(t)|$, where $h(t) = e_m^T e^{-tT_m} e_1$ for the Lanczos method and $h(t) = e_m^T T_m^{-1} e^{-t(T_m^{-1}-\sigma I)} e_1$ for the shift-and-invert Lanczos method. They can be computed by first computing the eigenvalue decomposition of T_m to obtain an algebraic expression for $h(t)$ and $h'(t)$ and then applying an optimization algorithm to find its extremum. However, since $h(t)$ is a fairly smooth function, they can also be easily approximated by its maximum at some densely distributed discrete points, i.e., $h_{\alpha_1,\alpha_2} \approx \max\{|h(\frac{i}{k}\tau)| : 0 \le i \le k \text{ and } \alpha_1 \le \frac{i}{k}\tau \le \alpha_2\}$, where k is some positive integer. We have used this approach with $k = 1000$ in our tests and we have used the same discrete points for the parametric value α in our a posteriori and a priori bounds (i.e., $\alpha = \frac{i}{k}\tau$ for $0 \le i \le k$). Also, the a priori bounds (3.12) and (4.10) involve choosing parametric values for δ and q, respectively. We have used $\delta = 10$ and $q = 1.01q_0$ in our tests; see the remarks after Theorems 3.2 and 4.2.

We first consider two well-conditioned diagonal matrices to illustrate the influence of the condition number on the convergence of the Lanczos method.

Example 1. Let A be the $n \times n$ diagonal matrix with the diagonal entries equal to $a_{ii} = 1 - \zeta\frac{i-1}{n-1}$ for $1 \le i \le n$, i.e., $A = \text{diag}\{1, 1-\zeta h, 1-2\zeta h, \ldots, 1-\zeta\}$, where $h = 1/(n-1)$ and $0 < \zeta < 1$. Then $\|A\| = 1$ and $\kappa = \lambda_{\max}(A)/\lambda_{\min}(A) = 1/(1-\zeta)$. We apply m steps of the Lanczos method to compute $w(\tau) = e^{-\tau A}v$, where v is a random vector with $\|v\| = 1$. We test various values of τ and we compare the error $\|w - w_m\|$ with the new bounds as well as the classical bound of Saad (3.3) by plotting them against m with the error in the solid lines, our a posteriori bound (3.4) in the +-lines, our a priori bound (3.12) in the dashed lines, and Saad's a priori bound (3.3) in the dash-dotted lines.

In our first test, we use $n = 10^3$ and $\zeta = 0.9$, resulting in a modestly well-conditioned matrix with $\kappa = 10$. We present the results for $\tau = 0.1,\ 1,\ 10,\ 100$ in Figure 5.1(a)–(d), respectively. We observe that when τ is very small ($\tau = 0.1$), the classical bound of Saad and our a priori bound are comparable. In this case, the convergence of the Lanczos method can be attributed to the small norm of τA. As τ increases, the classical bound deteriorates, while ours remains sharp. For $\tau = 10$, our bound is already much better than the classical bound, and for $\tau = 100$, the classical bound increases dramatically (out of range in the figure), while our bound follows the actual convergence curve quite closely. In this situation, the convergence of the Lanczos method is due to modest conditioning of A. For all cases, the a posteriori bound follows the actual error very closely.

In our second test, we use $n = 10^3$ and $\zeta = 0.1$, resulting in $\kappa = 10/9$. We present the results for $\tau = 0.1$ and 100 in Figure 5.2(a) and (b), respectively. For this very well-conditioned matrix, even when τ is very small ($\tau = 0.1$), our bound reflects the actual convergence rate more accurately; see Figure 5.2(a). Namely, the convergence is more driven by the small condition number even for $\tau = 0.1$. As τ increases from 0.1, $\tau\|A\|$ increases and the classical bound (3.3) further deteriorates with similar behavior, as observed in Example 1. We present the result for $\tau = 100$ only. For this case, since the smallest diagonal entry for the matrix is 0.9, the solution $w = e^{-90}e^{-\tau(A-0.9I)}v$ has effectively a scaling factor $e^{-90} \approx 10^{-40}$ and, with $w_m(\tau) = V_m e^{-\tau T_m} e_1 = e^{-90} V_m e^{-\tau(T_m - 0.9I)} e_1$, the Lanczos approximation can implicitly capture this scaling in the solution. Therefore, the error converges from about 10^{-40} to approximately 10^{-55}. The a posteriori error bound is derived from the maximum of $h(t) = |e_m^T e^{-tT_m} e_1|$ for $0 \le t \le \tau$, which removes this scaling for t near 0, and is therefore pessimistic in such an extreme situation.

Our next example uses an ill-conditioned matrix with a large τ for which we consider the shift-and-invert Lanczos method.

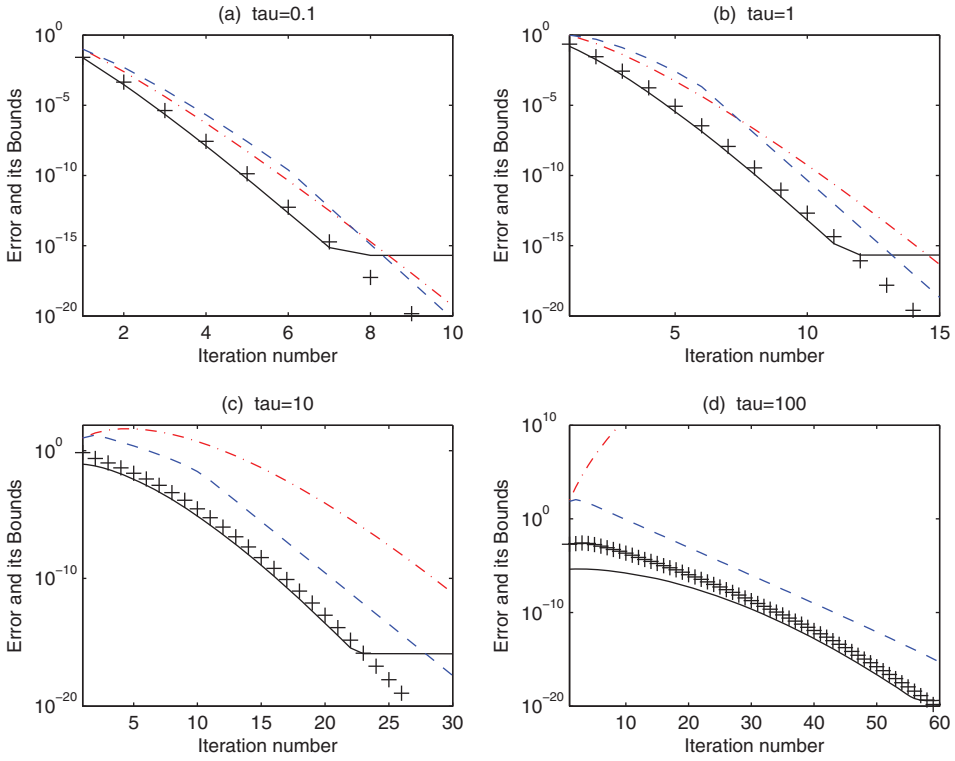

FIG. 5.1. *Example* 1, *case* $\zeta = 0.9$. *Error (solid), a posteriori bound (+), new a priori bound (dashed), Saad's bound (dash-dotted).*

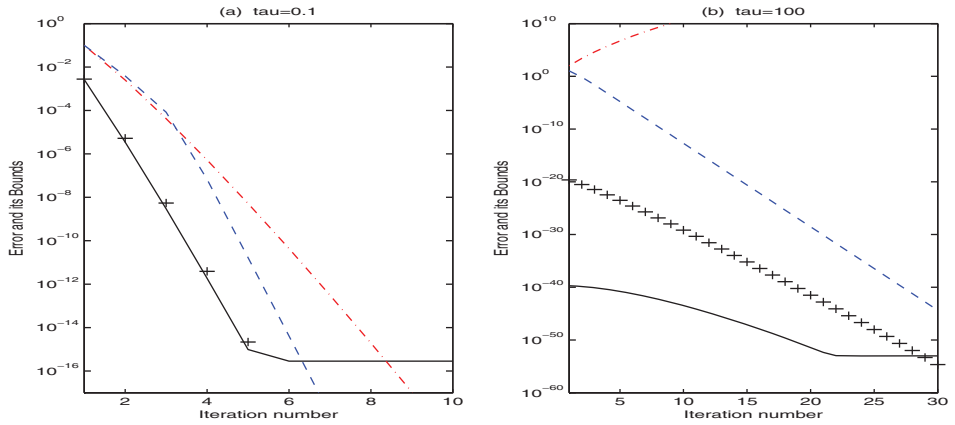

FIG. 5.2. *Example* 1, *case* $\zeta = 0.1$. *Error (solid), a posteriori bound (+), new a priori bound (dashed), Saad's bound (dash-dotted).*

Example 2. Consider computing $w = e^{-\tau A}v$ for a random vector v with $\|v\| = 1$, where $\tau = 1000$ and A is the $n \times n$ diagonal matrix with the diagonal entries equal to $a_{ii} = \frac{i}{n}$ for $1 \le i \le n$, i.e., $A = \text{diag}\{1/n, 2/n, \ldots, 1\}$. We use $n = 10^4$ in our test and both $\tau\|A\|$ and the condition number of A are large. We approximate w using w_m^{SIL} obtained by m steps of the shift-and-invert Lanczos method (4.3) with various

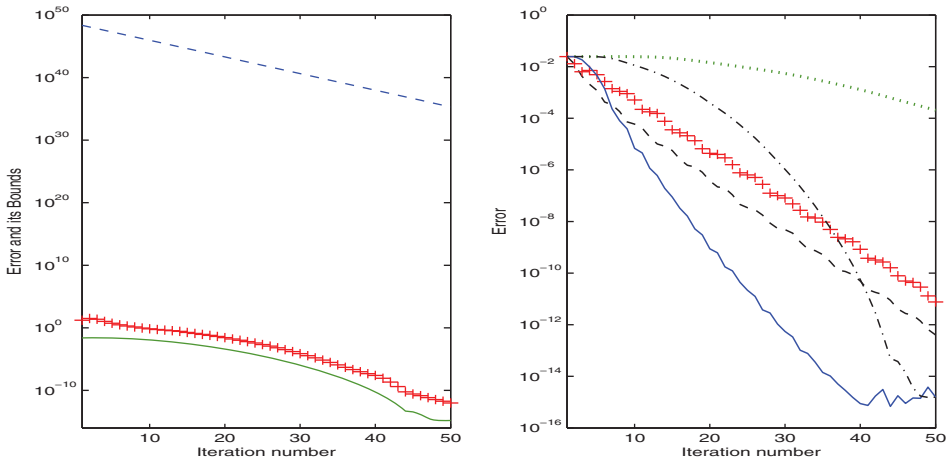

FIG. 5.3. *Example* 2. *Left:* $\sigma = 0.1$ *(error, solid; a posteriori bound,* + *line; a priori bound, dashed). Right: Error (*$\sigma = 1$*, dotted;* $\sigma = 0.1$*, dash-dotted;* $\sigma = 0.01$*, solid;* $\sigma = 0.001$*, dashed; and* $\sigma = 0$*,* + *line).*

shifts σ (i.e., $\sigma = 0,\ 0.001,\ 0.01,\ 0.1,\ 1$). In Figure 5.3, left, we present the results for the shift $\sigma = 0.1$ by plotting the error and its bounds against m with the error $\|w - w_m^{SIL}\|$ in the solid lines, the a priori bound (3.12) in the dashed line, and the a posteriori bound (4.4) in the + lines. In Figure 5.3, right, we compare convergence results for different shifts by plotting the error against m for $\sigma = 1$ (the dotted line), $\sigma = 0.1$ (the dash-dotted line), $\sigma = 0.01$ (the solid line), $\sigma = 0.001$ (the dashed line), and $\sigma = 0$ (the + line).

From Figure 5.3, left, we observe that the a posteriori bound gives a fairly good estimate of the error. In our experiments, however, the bounds tend to deteriorate when a smaller shift σ is used. (See Example 3 for such a result.) The a priori bound, however, overestimates the actual error by several orders of magnitude. We attribute this mainly to the pessimistic bound of the factor μ. In spite of this, it still seems to roughly capture the rate of convergence. With respect to influence of different shifts σ on the convergence, we observe from Figure 5.3, right, that increasing σ accelerates the convergence up to a certain point ($\sigma = 0.01$ in this case), after which the overall convergence actually decelerates. In terms of our bound, q_0 decreases as σ increases, but the factor μ increases very rapidly when $\tau\sigma$ is sufficiently large, overwhelming any decrease in q_0. Interestingly, even when the overall convergence is much slower for $\sigma = 0.1$ than for $\sigma = 0.01$, the asymptotic convergence (iterations 35 to 45) is actually faster, which appears to be a reflection of a smaller q_0 for $\sigma = 0.1$.

Our final example is the discrete Laplacian matrix that arises in space discretization of the heat equation

$$\frac{\partial}{\partial t}u(x,y,t) = \Delta u(x,y,t) \quad \text{for} \quad (x,y) \in R = [0\ 1]^2 \tag{5.1}$$

with the boundary condition $u = 0$ on ∂R.

Example 3. Let $A = T_N \otimes I_N + I_N \otimes T_N$, where $\otimes$ denotes the Kronecker product and T_N is the $N \times N$ tridiagonal matrix with the diagonal elements equal to 2 and the off-diagonal elements equal to -1. Since T_N has a known eigenvalue decomposition $T_N = Z\Lambda Z^T$ with $Z = [\sqrt{\frac{2}{N+1}} \sin \frac{jk\pi}{N+1}]_{j,k=1}^N$ and $\Lambda = \text{diag}\{2(1 - \cos \frac{j\pi}{N+1})\}$ (see [8,

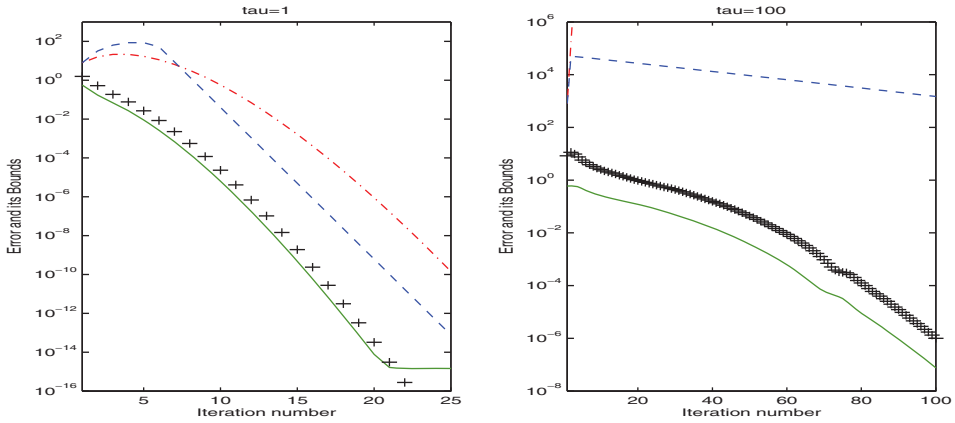

FIG. 5.4. *Example* 3 *(the Lanczos method). Error (solid), a posteriori bound (+), new a priori bound (dashed), Saad's bound (dash-dotted).*

p. 268]), we have $\exp(-\tau T_n) = Z\exp(-\tau\Lambda)Z^T$. It can be checked that $\exp(-\tau A) = \exp(-\tau T_N) \otimes \exp(-\tau T_N)$ and therefore $e^{-\tau A}v$ can be efficiently computed using this formula.

Computing $w(\tau) = e^{-\tau A}v$ arises in discretizing the heat equation (5.1) by the finite difference methods where $\tau = \Delta t/\Delta x^2$, $\Delta x = 1/(N+1)$ is the space mesh size, and Δt is the time step. In this context, even a modestly small Δt will result in a large τ. We test both the Lanczos method and the shift-and-invert Lanczos method for $e^{-\tau A}v$ with various values of τ. We use $N = 10^2$, resulting in a $10^4 \times 10^4$ matrix with norm ≈ 8 and condition number $\approx 4\cdot 10^3$. v is chosen to be a unit random vector.

We first apply m steps of the Lanczos method to compute $w(\tau)$ and we compare the error $\|w - w_m\|$ with our bounds as well as the classical bound of Saad (3.3). In Figure 5.4, we present the results for $\tau = 1$ and $\tau = 100$ by plotting against m the error in the solid lines, our a posteriori bound (3.4) in the + lines, our a priori bound (3.12) in the dashed lines, and Saad's a priori bound (3.3) in the dash-dotted lines. We point out that if $\tau = 1000$ (not shown here), the error of the Lanczos method does not converge in any meaningful way. We observe again, as in Example 1, that the a posteriori bound provides a fairly good estimate of the actual error and our a priori bound significantly improves the classical a priori bound, although it is also very pessimistic in the case $\tau = 100$.

We next apply m steps of the shift-and-invert Lanczos method (4.3) to the problem with $\tau = 100$ for which the Lanczos method converges slowly. We consider various shifts σ (i.e., $\sigma = 0.001,\ 0.01,\ 0.1,\ 1$). In Figure 5.5, left, we present the results for the shift $\sigma = 0.1$ by plotting against m the error $\|w - w_m^{SIL}\|$ in the solid lines, our a priori bound (3.12) in the dashed line, and our a posteriori bound (4.4) in the + lines. In Figure 5.5, right, we compare convergence results for different shifts by plotting the error $\|w - w_m^{SIL}\|$ against m for $\sigma = 1$ (the dotted line), $\sigma = 0.1$ (the solid line), $\sigma = 0.01$ (the dashed line), and $\sigma = 0.001$ (the + line).

From Figure 5.5, left, we see that the a posteriori bound is rather pessimistic in this case. This appears to be due to bounding the integral form of the error (4.6) in the proof of Theorem 4.1 by the maximum of $|h(t)|$. In our experiments, however, the bound improves for larger shift (e.g., $\sigma = 1$); see similar behavior discussed in Example 3. The a priori bound is inherently weak in this case as it is based on the a posteriori bound. With respect to influence of different shifts σ on convergence, we

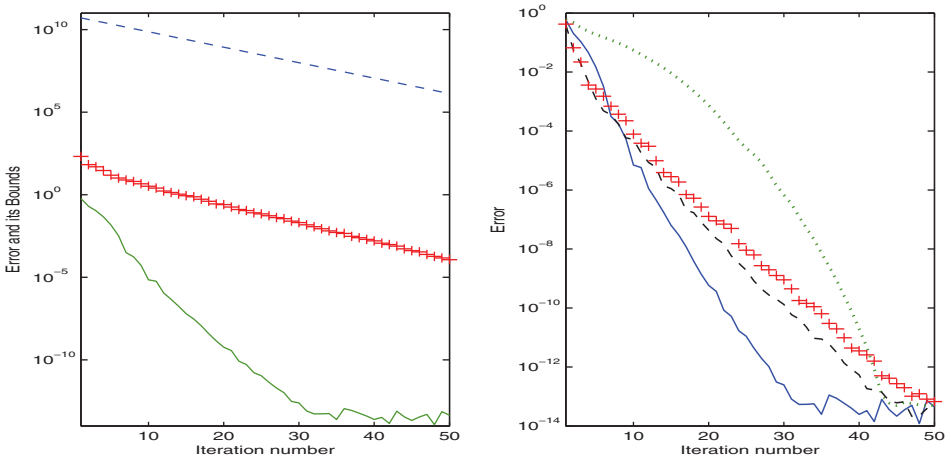

FIG. 5.5. *Example* 3 *(the shift-and-invert Lanczos method). Left:* $\sigma = 0.1$ *(error, solid; a posteriori bound,* + *line; a priori bound, dashed). Right: Error (*$\sigma = 1$*, dotted;* $\sigma = 0.1$*, solid;* $\sigma = 0.01$*, dashed;* $\sigma = 0.001$*,* + *line).*

observe from Figure 5.5, right, that increasing σ accelerates the convergence up to $\sigma = 0.1$, after which the overall convergence actually deteriorates, but we also note that the asymptotic convergence for the case $\sigma = 1$ (around iteration 40) appears slightly better. The best shift $\sigma = 0.1$ here is larger than the one in Example 2. These results and those of Example 3 confirm the influence of the condition number of $A + \sigma I$ on the asymptotic convergence rate. They also suggest that the factor μ plays an equally important role in the overall convergence and an optimal σ is problem dependent.

6. Concluding remarks. We have presented new error bounds for the Lanczos method and the shift-and-invert Lanczos method for computing $e^{-\tau A}v$. The bounds relates the error to the $(m, 1)$ entry of the exponential of the tridiagonal matrix which is known to have a decay property. Furthermore, the bounds demonstrate the dependence of convergence on the condition numbers of the related matrices. Numerical examples confirm the theoretical results.

For future work, it will be interesting to further investigate if the convergence property revealed here can be used for preconditioning. Another interesting question is how to choose an optimal shift in the shift-and-invert Lanczos method. Our a priori bound is still too pessimistic for this. It would be interesting to see if a sharper bound can be derived that more precisely reflects the two effects of increasing the shift.

Acknowledgments. I would like to thank Dr. Ping Zhang for carrying out some related earlier investigations in her thesis work [30]. I would also like to thank two anonymous referees for their careful reading and constructive comments.

REFERENCES

[1] M. AFANASJEW, M. EIERMANN, AND O. G. ERNST, *Implementation of a restarted Krylov subspace method for the evaluation of matrix functions,* Linear Algebra Appl., 429 (2008), pp. 2293–2314.

[2] M. BENZI AND G. H. GOLUB, *Bounds for the entries of matrix functions with applications to preconditioning,* BIT, 39 (1999), pp. 417–438.

[3] 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.

[4] P. Castillo and Y. Saad, *Preconditioning the matrix exponential operator with applications,* J. Sci. Comput., 13 (1999), pp. 275–302.
[5] S. Chen and Y. T. Zhang, *Krylov implicit integration factor methods for spatial discretization on high dimensional unstructured meshes: Application to discontinuous Galerkin methods,* J. Comput. Phys., 230 (2011), pp. 4336–4352.
[6] S. Demko, *Inverses of band matrices and local convergence of spline projections,* SIAM J. Numer. Anal., 14 (1977), pp. 616–619.
[7] S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverses of band matrices,* Math. Comp., 43 (1984), pp. 491–499.
[8] J. Demmel, *Applied Numerical Linear Algebra,* SIAM, Philadelphia, 1997.
[9] 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.
[10] 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.
[11] V. L. Druskin and L. A. Knizhnerman, *Extended Krylov subspaces: Approximation of the matrix square root and related functions,* SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.
[12] W. S. Edwards, L. S. Tuckerman, R. A. Friesner, and D. C. Sorensen, *Krylov methods for the incompressible Navier-Stokes equations,* J. Comput. Phys., 110 (1994), pp. 82–102.
[13] M. Eiermann and O. G. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions,* SIAM J. Numer. Anal., 44 (2006), pp. 2481–2504.
[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] M. Hochbruck and C. Lubich, *On Krylov subspace approximations to the matrix exponential operator,* SIAM J. Numer. Anal., 34 (1997), pp. 1911–1925.
[16] M. Ilic, I. W. Turner, and A. N. Pettitt, *Bayesian computations and efficient algorithms for computing functions of large, sparse matrices,* ANZIAM J., 45(E) (2004), pp. C504–C518.
[17] M. Ilic, I. W. Turner, and V. Anh, *Numerical solution of the fractional poisson equations using an adaptively preconditioned Lanczos methods,* J. Appl. Math. Stochastic Anal., 2008 (2008), 104525.
[18] L. Knizhnerman and V. Simoncini, *A new investigation of the extended Krylov subspace method for matrix function evaluations,* Numer. Linear Algebra Appl., 17 (2010), pp. 615–638.
[19] 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.
[20] I. Moret and P. Novati, *Rd-rational approximations of the matrix exponential,* BIT, 44 (2004), pp. 595–615.
[21] 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.
[22] 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.
[23] C. K. Newman, *Exponential Integrators for the Incompressible Navier-Stokes Equations*, Ph.D. thesis, Department of Mathematics, Virginia Tech, Blacksburg, 2003.
[24] Y. Saad, *Analysis of some Krylov subspace approximations to the matrix exponential operator,* SIAM J. Numer. Anal., 29 (1992), pp. 209–228.
[25] R. B. Sidje, *Expokit: A software package for computing matrix exponentials,* ACM Trans. Math. Software, 24 (1998), pp. 130–156.
[26] R. S. Varga, *On higher order stable implicit methods for solving parabolic partial differential equations,* J. Math. Phys., 40 (1961), pp. 220–231.
[27] J. van den Eshof and M. Hochbruck, *Preconditioning Lanczos approximations to the matrix exponential,* SIAM J. Sci. Comput., 27 (2005), pp. 1438–1457.
[28] Q. Ye, *A convergence analysis of nonsymmetric Lanczos algorithms,* Math. Comp., 56 (1991), pp. 677–691.
[29] A. Zafer, *Calculating the matrix exponential of a constant matrix on time scales,* Appl. Math. Letters, 21 (2008), pp. 612–616.
[30] P. Zhang, *Iterative Methods for Computing Eigenvalues and Exponentials of Large Matrices,* Ph.D. thesis, Department of Mathematics, University of Kentucky, Lexington, 2009.