Electronic Transactions on Numerical Analysis.
Volume 35, pp. 118-128, 2009.
Copyright © 2009, Kent State University.
ISSN 1068-9613.

ETNA
Kent State University
http://etna.math.kent.edu

MONOTONE CONVERGENCE OF THE LANCZOS APPROXIMATIONS TO MATRIX FUNCTIONS OF HERMITIAN MATRICES[*]

ANDREAS FROMMER[†]

Abstract. When A is a Hermitian matrix, the action $f(A)b$ of a matrix function $f(A)$ on a vector b can efficiently be approximated via the Lanczos method. In this note we use M-matrix theory to establish that the 2-norm of the error of the sequence of approximations is monotonically decreasing if f is a Stieltjes transform and A is positive definite. We discuss the relation of our approach to a recent, more general monotonicity result of Druskin for Laplace transforms. We also extend the class of functions to certain product type functions. This yields, for example, monotonicity when approximating sign$(A)b$ with A indefinite if the Lanczos method is performed for A^2 rather than A.

Key words. matrix functions, Lanczos method, Galerkin approximation, monotone convergence, error estimates

AMS subject classifications. 6530, 65F10, 65F50

1. Introduction.

Throughout the whole paper let $A \in \mathbb{C}^{n\times n}$ be a Hermitian matrix. Then there exists an othornormal set of eigenvectors of A which spans $\mathbb{C}^n$. We can express this via the spectral decomposition

$$A = Q\Lambda Q^H, \;\; \Lambda = \mathrm{diag}[\lambda_1, \ldots, \lambda_n], \tag{1.1}$$

the i-th column of Q being an eigenvector of A for the eigenvalue λ_i and $Q^HQ = I$.

Let spec$(A) = \{\lambda_1, \ldots, \lambda_n\}$ denote the set of all eigenvalues of A. Any function

$$f : z \in \text{spec}(A) \to f(z) \in \mathbb{C}$$

can be extended to a matrix function $f(A)$ as

$$f(A) = Qf(\Lambda)Q^H \text{ where } f(\Lambda) = \mathrm{diag}[f(\lambda_1), \ldots, f(\lambda_n)].$$

Other, equivalent, definitions are possible. For example, with the help of the polynomial p of degree at most $n-1$ which interpolates f on spec(A) we have

$$f(A) = p(A),$$

and for f analytic there is a representation as a contour integral for the resolvent; see, e.g., [13]. We will be particularly interested in cases where f is defined for $z > 0$ and can be represented as an (improper Riemann-Stieltjes) integral of the form

$$f(z) = \int_{t=0}^{\infty} \frac{1}{(t+z)^k} d\mu(t) \tag{1.2}$$

with k a natural number and $\mu(t) : \mathbb{R} \to \mathbb{R}$ a non-decreasing bounded function for which $\int_{t=1}^{\infty} 1/t^k d\mu(t)$ is finite. Using (1.1) we see that we then can represent $f(A)$ as

$$f(A) = \int_{t=0}^{\infty} (tI + A)^{-k} d\mu(t),$$

[*] Received June 3, 2008. Accepted for publication January 21, 2009. Published online June 24, 2009. Recommended by V. Druskin.

[†] Fachbereich Mathematik und Naturwissenschaften, Universität Wuppertal, Gauß-Straße 20, D-42097 Wuppertal, Germany (`frommer@math.uni-wuppertal.de`). Supported in part by Deutsche Forschungsgemeinschaft under grant DFG Fr 755/15-2 and SFB-TR 55 'Hadron Physics from Lattice QCD'.

the integral to be understood componentwise.

This paper deals with the situation where one wants to compute $u = f(A)b$ for some vector $b \in \mathbb{C}^n$. If A is large and sparse, computing $f(A)$ is prohibitive, since it usually is a dense matrix. The action of $f(A)$ on b may, however, still be computable at reasonable cost, and the Lanczos method has established itself as the standard way to do so.

Let us recall that given an initial vector $b \in \mathbb{C}^n$, which for notational consistency is now called $\tilde{v}^1, \tilde{v}^1 \neq 0$, the Lanczos process computes an orthonormal basis $v^1, v^2, \ldots, v^m$ of the Krylov subspace $K_m(A, \tilde{v}^1) = \text{span}\left\{\tilde{v}^1, A\tilde{v}^1, \ldots, A^{m-1}\tilde{v}^1\right\}$ up to a maximum stage m_{max} (which is the degree of the minimal polynomial of $\tilde{v}^1$ with respect to A) via the iteration (we put $v^0 = 0$), as follows.

```
for $m = 1, \ldots, m_{\max}$
    $\beta_m = \|\tilde{v}^m\|$
    $v^m = \tilde{v}^m / \beta_m$
    $\tilde{w}^{m+1} = Av^m - \beta_m v^{m-1}$
    $\alpha_m = \langle \tilde{w}^{m+1}, v^m \rangle$
    $\tilde{v}^{m+1} = \tilde{w}^{m+1} - \alpha_m v^m$
```

The process is stopped for $m = m_{\max}$ since this is the first index for which $\tilde{v}^{m+1} = 0$. The Lanczos process is usually summarized as

$$AV_m = V_m T_m + \beta_{m+1} v^{m+1} e_m^T, \tag{1.3}$$

where $V_m = [v^1 | \ldots | v^m] \in \mathbb{C}^{n \times m}$, e_m is the m-th Cartesian unit vector in $\mathbb{C}^m$ and T_m is the symmetric tridiagonal matrix

$$T_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-1} & \alpha_{m-1} & \beta_m \\ & & & \beta_m & \alpha_m \end{bmatrix} \in \mathbb{R}^{m \times m}.$$

Based on the Lanczos method, the following approach for obtaining approximations $u^m \in K_m(A, b)$ to $u = f(A)b$ has meanwhile established itself as standard:

$$u^m = V_m f(T_m) V_m^H b = \beta_1 V_m f(T_m) e_1. \tag{1.4}$$

This amounts to orthogonally project the matrix A onto the subspace $K_m(A, b)$ and to approximate $f(A)b$ by the matrix function evaluated on the subspace. In [7], to which we also refer for a detailed historic account including [14, 20, 24], this method is called the *spectral Lanczos decomposition method*. For brevity, let us call u^m just the (m-th) Lanczos approximation to $f(A)b$. Note that for $m = m_{\max}$ we have $AV_m = V_m T_m$. Since $f(T_m)$ can be represented as a polynomial in T_m we have that

$$f(A)b = \beta_1 V_{m_{\max}} f(T_{m_{\max}}) e_1.$$

Note also that (1.4) still requires to compute $f(T_m)$. But T_m will be of much smaller size than A and, in addition, it is tridiagonal. So various appropriate techniques may be applied to compute $f(T_m)$, including those using the spectral decomposition of T_m; see, e.g., [12] or [13].

Our purpose is to investigate the error

$$e^m = u^m - u$$

ETNA
Kent State University
http://etna.math.kent.edu

of the Lanczos approximations u^m and we will identify situations where $\|e^m\|$ decreases monotonically. Our basic result holds for the case where A is positive definite and f can be represented in the form (1.2). In this sense, we extend a well-known result for the Conjugate Gradient (CG) method for solving $Ax = b$; see, e.g., [21]. CG is mathematically equivalent to the Lanczos method described above with $f(z) = z^{-1}$ which can be expressed in the form (1.2) using the step function ω:

$$f(z) = \int_{t=0}^{\infty} \frac{1}{z+t} d\omega(t) \quad \text{with } \omega(t) = \begin{cases} 0 & \text{for } t = 0, \\ 1 & \text{for } t > 0. \end{cases}$$

In the CG method the residuals $r^m = b - Au^m$ are collinear to the Lanczos vectors, see [21]:

$$r^m = (-1)^{m+1} \|r^m\| \cdot v^m. \tag{1.5}$$

The presentation of the results in this paper will be greatly simplified if we flip the direction of every other Lanczos vector v^m just in the way suggested by (1.5). So let

$$V_m^{\pm} = [v^1 | -v^2 | \dots | (-1)^{m+1} v^m].$$

The basic relation (1.3) can then equivalently be expressed as

$$AV_m^{\pm} = V_m^{\pm} T_m^{\pm} + (-1)^{m+1} \beta_{m+1} v^{m+1} e_m^T \tag{1.6}$$

with

$$T_m^{\pm} = \begin{bmatrix} \alpha_1 & -\beta_2 & & & \\ -\beta_2 & \alpha_2 & -\beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -\beta_{m-1} & \alpha_{m-1} & -\beta_m \\ & & & -\beta_m & \alpha_m \end{bmatrix}.$$

Of course, $T_m^{\pm} = S^{-1} T_m S$ with the signature matrix $S = \text{diag}[1, -1, \dots, (-1)^{m-1}] \in \mathbb{R}^{m \times m}$. Since for any matrix function and any non-singular matrix X one has (see, e.g., [12], [13], or [15])

$$Xf(A)X^{-1} = f(XAX^{-1}),$$

we see that $f(T_m) = Sf(T_m^{\pm})S^{-1}$. It follows that the Lanczos approximation u^m from (1.4) is also given by

$$u^m = \beta_1 V_m^{\pm} f(T_m^{\pm}) e_1. \tag{1.7}$$

The remainder of this paper is organized as follows: In Section 2 we will study some properties of $T_m^{\pm}$ using M-matrix theory. In Section 3 we will use these to prove the monotone convergence for the given class of functions. Section 4 is devoted to a comparison with the recent results from [6]. In Section 5 we extend our results to a larger class of functions, thus including Lanczos type methods for approximating the action of the matrix sign function. The paper ends with a general discussion of the techniques used in Section 6 and some conclusions where we also address the impact of inexact arithmetic. Otherwise, exact arithmetic is assumed throughout.

2. Properties of T_m and $T_m^\pm$. In this section we assume A to be positive definite. Let us first note that from (1.3) and (1.6) we immeditaley see that

$$T_m = V_m^H A V_m, \;\; T_m^\pm = (V_m^\pm)^H A V_m^\pm = S T_m S^{-1}.$$

So, since A is positive definite, we have that T_m and $T_m^\pm$ are both positive definite, too. From the Lanczos process it is also clear that all non-zero entries of T_m are real and positive. Let $D_m = \mathrm{diag}[\alpha_1, \ldots, \alpha_m]$ be the diagonal part of T_m and let $B_m = T_m - D_m$. Then $B_m \geq 0$, where ”$\geq$” stands for the entrywise partial ordering. Since $T_m^\pm = D_m - B_m$, we see that the off-diagonal entries of $T_m^\pm$ are all nonpositive. A matrix with nonpositive off-diagonal entries whose inverse is (componentwise) nonnegative is called an *M-matrix*. The following lemma shows that $T_m^\pm$ is an M-matrix.

LEMMA 2.1. *If A is positive definite, then*

$$(T_m^\pm)^{-1} \geq 0.$$

Proof. A well-known result for M-matrices (see [3, Theorem 2.3, G20]) states that for $B \in \mathbb{R}^{n\times n}$ with nonpositive off-diagonal entries the relation $B^{-1} \geq 0$ is equivalent to that all eigenvalues of B have positive real parts. But $T_m^\pm = (V_m^\pm)^H A V_m^\pm$ has only nonpositive off-diagonal entries and its eigenvalues are all positive, since A is positive definite. □

M-matrices have plenty of useful properties. The two that we need are collected in the following lemma. For a proof see [3, Exercise 5.1], for example.

LEMMA 2.2. *Let $B, C \in \mathbb{R}^{m\times m}$ be two M-matrices and let $E \in \mathbb{R}^{m\times m}$ be such that $E \geq 0$.*

(i) If $B \leq C$, then $0 \leq C^{-1} \leq B^{-1}$.

(ii) If $B + E$ has all its off-diagonal entries nonpositive, then $B + E$ is an M-matrix.

3. Monotone convergence. Our approach to prove monotone convergence, which builds upon [6], starts from (1.7). Since the Lanczos basis vectors v^m are mutually orthogonal, if we can show that the coefficient vectors representing u^m from (1.7) in this basis, given as

$$s^m = \beta_1 f(T_m^\pm) e_1 \in \mathbb{R}^m,$$

satisfy

$$\mathbf{o} \leq \begin{bmatrix} s^{m-1} \\ 0 \end{bmatrix} \leq s^m \tag{3.1}$$

for $m = 1, \ldots, m_{\max}$, we have that the sequence $\|u^m\|$ is monotonically increasing. It is even strictly increasing if for one component, for example the last one, we have strict inequality in (3.1). Moreover, since $f(A)b = u^{m_{\max}}$, we also see that the norm of the errors

$$e^m = u^{m_{\max}} - u^m = V_{m_{\max}} \left(s^{m_{\max}} - \begin{bmatrix} s^m \\ \mathbf{o} \end{bmatrix} \right)$$

is monotonically decreasing. This is how we will prove our main result stated as follows.

THEOREM 3.1. *Let A be Hermitian and positive definite. Assume that the function $f : (0,\infty) \to \mathbb{R}$ can be expressed for all $z > 0$ as*

$$f(z) = \int_{t=0}^{\infty} \frac{1}{(t+z)^k} d\mu(t),$$

with $\mu(t)$ a non-decreasing function such that $\int_1^\infty \frac{1}{t^k} d\mu(t) < \infty$ and $k \in \mathbb{N}$. Let u^m be the Lanczos approximation defined in (1.4) or (1.7) and $e^m = f(A)b - u^m$ for $m = 1, \ldots, m_{\max}$. Then the following holds for the 2-norm $\|\cdot\|$:

(i) The sequence $\|u^m\|$ is monotonically increasing.
(ii) The sequence $\|e^m\|$ is monotonically decreasing.

Proof. As we just explained it is sufficient so show (3.1). To that purpose we use the representation

$$s^m = \beta_1 \cdot f(T_m^{\pm})e_1 = \beta_1 \int_{t=0}^{\infty} (tI + T_m^{\pm})^{-k} e_1 d\mu. \tag{3.2}$$

Note that this integral exists since $\mathrm{spec}(T_m^{\pm}) \subset (0,\infty)$. Denote by $\hat{T}_m^{\pm} \in \mathbb{R}^{m\times m}$ the matrix obtained from $T_m^{\pm}$ by setting the $(m-1,m)$ and $(m,m-1)$ entries to zero,

$$\hat{T}_m^{\pm} = \begin{bmatrix} \alpha_1 & -\beta_2 & & & \\ -\beta_2 & \alpha_2 & -\beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & -\beta_{m-1} & \alpha_{m-1} & 0 \\ & & & 0 & \alpha_m \end{bmatrix} = \begin{bmatrix} T_{m-1}^{\pm} & \mathbf{o} \\ \mathbf{o}^T & \alpha_m \end{bmatrix}. \tag{3.3}$$

Then

$$tI + \hat{T}_m^{\pm} = \begin{bmatrix} tI + T_{m-1}^{\pm} & \mathbf{o} \\ \mathbf{o}^T & t+\alpha_m \end{bmatrix}$$

and

$$tI + T_m^{\pm} \le tI + \hat{T}_m^{\pm} \quad \text{for all } t \ge 0.$$

But for all $t \ge 0$ the matrix $tI + T_m^{\pm}$ is an M-matrix by Lemmas 2.1 and 2.2(ii). Moreover, again by Lemma 2.2(ii), the matrix $tI + \hat{T}_m^{\pm}$ is an M-matrix for all $t \ge 0$, too. And since $tI + T_m^{\pm} \le tI + \hat{T}_m^{\pm}$, part (i) of that lemma gives us

$$0 \le (tI + \hat{T}_m^{\pm})^{-1} \le (tI + T_m^{\pm})^{-1} \text{ for all } t \ge 0.$$

Trivially, then, via repeated multiplication we get

$$0 \le (tI + \hat{T}_m^{\pm})^{-k} \le (tI + T_m^{\pm})^{-k} \text{ for all } t \ge 0,$$

which results in

$$0 \le \int_{t=0}^{\infty} (tI + \hat{T}_m^{\pm})^{-k} d\mu \le \int_{t=0}^{\infty} (tI + T_m^{\pm})^{-k} d\mu.$$

Given the block structure (3.3) and comparing the first columns, the inequality above finally yields

$$\mathbf{o} \le \begin{bmatrix} s^{m-1} \\ 0 \end{bmatrix} \le s^m. \qquad \square$$

COROLLARY 3.2. *Let A be Hermitian and positive definite. Then the norms of the Lanczos approximations u^m to $f(A)b$ increase monotonically, and the error norms $\|f(A)b - u^m\|$ decrease monotonically for the following functions f:*
(i) $f(z) = z^{-k}$, $k \in \mathbb{N}$,
(ii) $f(z) = \sum_{i=1}^{p} \frac{\alpha_i}{z+\beta_i}$ with $\alpha_i \ge 0, \beta_i > 0$ for $i = 1,\ldots,p$,
(iii) $f(z) = z^{-1/2}$,

(iv) $f(z) = z^{-\alpha}$ *for* $\alpha \in (0, 1)$,

(v) $f(z) = (z - 1)^{-1} \log z$,

(vi) $f(z) = z^{-\alpha}(1 + z)^{-\beta}, 0 < \alpha \le 1, \alpha + \beta \in [0, 1)$,

(vii) $f(z) = \sum_{i=1}^{\infty} \frac{\alpha_i}{z+\beta_i}$ *with* $\alpha_i \ge 0, \beta_i > 0$ *for* $i = 1, 2, \ldots$, *and* $\lim_{i\to\infty} z_i = \infty, \lim_{i\to\infty} |\alpha_i/\beta_i| < \infty$,

(viii) f *is the result of a Stieltjes transform, i.e.,*

$$f(z) = \int_{t=0}^{\infty} \frac{1}{z+t} d\mu(t),$$

where μ *is a non-decreasing real function such that* $\int_1^\infty \frac{1}{t} d\mu(t) < \infty$,

(ix) $f(z) = \sum_{i=1}^{\ell} \gamma_i f_i(z)$ *with* $\gamma_i \ge 0$ *for all* i *and* f_i *any function from (i)-(viii) or a constant.*

Proof. Part (i) follows by taking the step function,

$$\omega(t) = \begin{cases} 0 & \text{for } t = 0, \\ 1 & \text{for } t > 0, \end{cases}$$

so that $z^{-k} = \int_{t=0}^{\infty} \frac{1}{(t+z)^k} d\omega(t)$. The functions considered in (ii) to (vii) are all particular Stieltjes transforms, i.e., they are special cases of (viii) as we briefly outline now. For the rational function case (ii), assume that $0 \le \beta_1 < \cdots < \beta_p$ and define the step function ω as

$$\omega(t) = \begin{cases} 0 & \text{for } t \le \beta_1, \\ \sum_{j=1}^{i} \alpha_j & \text{for } \beta_i < t \le \beta_{i+1}, \\ \sum_{j=1}^{p} \alpha_j & \text{for } \beta_p < t, \end{cases}$$

to see that $f(z) = \int_{t=0}^{\infty} \frac{1}{t+z} d\omega(t)$. Part (iii) is contained in (iv) for which we observe that for $\alpha \in (0, 1)$,

$$z^{-\alpha} = \frac{\sin((1-\alpha)\pi)}{\pi} \int_0^\infty \frac{1}{t+z} d\mu(t),$$

with $\mu(t) = t^{-\alpha}$; see [4]. The fact that we are also in the presence of Stieltjes transforms in cases (v) and (vi) has been observed in [16], the case (vii) was treated in [7]. Finally, if f is of the form given in (ix) we have that $s^m = \beta_1 \cdot \sum_{i=1}^{\ell} f_i(T_m^{\pm})e_1$. Herein, each individual summand $f_i(T_m^{\pm})e_1$ fulfills a relation analogous to (3.1) which thus carries over to the whole sum. □

Let us remark that the set of Stieltjes transforms is a subset of the set of completely monotone functions. We refer to [11, Chapter 12] for a textbook treatment of Stieltjes transforms. Defining the Stieltjes cone as the set of all functions of the form

$$a + \int_0^\infty \frac{1}{t+z} d\mu(t),$$

with $a \ge 0$ and μ as before, it can be shown that the Stieltjes cone is exactly the restriction to the positive real axis of all functions g which are holomorphic in the cut plane $\mathbb{C} \setminus (-\infty, 0]$, nonnegative on $\mathbb{R}^+$ and which map the upper half plane to the lower half plane; see [1, Chapter 3, Addenda and Problems], [2] or [11, Chapter 12.10].

The importance of the Stieltjes cone for the analysis of matrix function methods has been realized by several authors, for example in [7, 17] for approximation in extended Krylov subspaces and in [16] (see also [8]) for an analysis of restarted variants.

4. Relation to the exponential.

For the matrix exponential we have the following result which has recently been proved in [6, Theorem 1 and Remark 1].

THEOREM 4.1. *Let A be Hermitian and*

$$g(z) = \int_{t=0}^{\infty} w(t)e^{tz}dt, \quad z \in [a,b] \supset \mathit{spec}(A),$$

with $w(t)$ real, nonnegative such that $g(z)$ exists and is bounded on $[a,b]$. Then the Lanczos approximations to $\exp(A)b$ as well as to $g(A)b$ converge monotonically.

Now, let A be positive definite, $(a,b) = (-\infty, 0)$ and take $f(z) = g(-z)$, i.e.,

$$f(z) = \int_{t=0}^{\infty} w(t)e^{-tz}dt, \quad z \in (0,\infty).$$

Then f can be interpreted as the Laplace transform (see, e.g., [11, Chapter 10]) of w, provided w is from what is called the 'original space' Ω in [11]. Laplace transforms are intimately related to Stieltjes transforms, since the latter ones arise as the result of two iterated Laplace transforms. Indeed, as is explained in detail in [11, Chapter 10.11], taking $\sigma(s) = \int_0^\infty w(t)e^{-st}dt$, and assuming that this integral converges absolutely for s in the closed right half plane, the following transformations are valid:

$$\begin{aligned}\int_0^\infty e^{-sz}\sigma(s)\,ds &= \int_0^\infty \int_0^\infty e^{-sz}e^{-st}w(t)\,dt\,ds \\ &= \int_0^\infty \int_0^\infty e^{-(z+t)s}ds\,w(t)\,dt \\ &= \int_0^\infty \frac{1}{z+t}w(t)dt.\end{aligned}$$

This shows that, at least for the case $k = 1$, our Theorem 3.1 with $\mu(t) = \int_{\tau=0}^{t} w(\tau)d\tau$ and 'standard' functions w is actually a special case of what has been proven in [6] in the context of the matrix exponential. The proof presented here, however, is quite different from that in [6], and may thus have some value by itself. In [6], an analog to a semidiscretized one-dimensional heat equation was built up from the Lanczos coefficients, and the monotonicity result was established considering the time stepping operator of an explicit Euler scheme. Our approach, in turn, highlights the role of M-matrices in this context and may be regarded more 'linear algebra oriented'.

5. Extensions.

Assume that the function f can be represented as

$$f(z) = g(z) \cdot p(z),$$

where p is a polynomial and g is of the form considered in Theorem 3.1, i.e.,

$$g(z) = \int_{t=0}^{\infty} \frac{1}{(t+z)^k}d\mu(t),$$

with $\mu(t)$ a non-decreasing real function, $k \in \mathbb{N}$, $\int_1^\infty \frac{1}{t^k}d\mu(k) < \infty$. An obvious way to approximate $f(A)b$ is to first compute $\tilde{b} = p(A)b$, e.g., using Horner's scheme or a known stable recurrence for p. This mainly requires only simple matrix vector multiplications. We then approximate $g(A)\tilde{b}$ using the Lanczos approach. Obviously, by Theorem 3.1 this approach leads to monotone convergence.

Approximating $f(A)b$ in this manner we considerably increase the class of functions for which the approximations to $f(A)b$ converge smoothly, i.e., monotonously. In the following example we explicitly list some functions which are from this class and which are important in practice.

EXAMPLE 5.1. (See [7].) The following matrix functions arise in the solution of elliptic boundary problems of the form

$$Aw - \frac{d^2w}{d\Theta^2} = g(\Theta)\varphi \tag{5.1}$$

using the method of lines:

(i) For $g \equiv 0$ and the boundary conditions, $w(0) = \varphi_0$, $w(\infty) = 0$, we have $w(\Theta) = \exp(-\Theta\sqrt{A})\varphi_0$, i.e., we have

$$f(z) = e^{-\Theta\sqrt{z}} = 1 - g(z)z,$$

with

$$g(z) = \frac{1-\exp(-\Theta\sqrt{z})}{z} = \int_0^\infty \frac{1}{z+t}d\mu, \quad \text{where } d\mu = \frac{\sin(\Theta\sqrt{t})}{\pi t}dt.$$

(ii) The matrix square root arises from the Dirichlet to Neumann problem for (5.1), i.e.,

$$f(z) = \sqrt{z} = g(z)z \quad \text{with } g(z) = z^{-1/2},$$

where $z^{-1/2}$ was considered in Corollary 3.2 (iii).

With a slight modification of the Lanczos approach, the discussion of this section also holds for the matrix sign function as we will explain now. Computing the action of the sign function $\mathrm{sign}(A)b$ for a Hermitian, indefinite matrix A is at the heart of very compute-intensive numerical simulations in lattice quantum chromodynamics with so-called overlap fermions; see, e.g., [18]. Since A is indefinite, the theory developed so far does not apply directly. Actually, numerical experiments reported in [23] show that there is no monotone decrease of the error norm if one computes the Lanczos approximations as given by (1.4). Based on numerical experiments and a partly heuristic explanation, the paper [23] therefore suggests to rather compute $\mathrm{sign}(A)b$ as $(A^2)^{-1/2}(Ab)$; see also [5]. This means that we use (1.4) for

$$f(B)\tilde{b} \ \text{ where } f(z) = z^{-1/2},\ B = A^2,\ \tilde{b} = Ab. \tag{5.2}$$

With Corollary 3.2 (iii), we now have a proof for the smooth convergence observed since it shows that the norm of the error of the Lanczos approximations for (5.2) is monotonically decreasing.

In the case of the matrix sign function, we know that $\|\mathrm{sign}(A)b\| = \|b\|$, because $\mathrm{sign}(A)$ is unitary. Together with the monotone convergence of the approximations via (1.4), we can thus even get bounds on the error of the approximations according to the following proposition.

PROPOSITION 5.2. *Assume that A is Hermitian and that approximations u^m for $u =$ sign$(A)b$ are computed by the Lanczos method for $B^{-1/2}\tilde{b}$ with $B = A^2, \tilde{b} = Ab$. Then the sequence $\|u^m\|$ is monotonically increasing, $\|u^m\| \le \|b\|$ for all m and*

$$\|b\| - \|u^m\| \le \|u - u^m\| \le \left(\|b\|^2 - \|u^m\|^2\right)^{1/2}.$$

Proof. We use the notation introduced in Section 3. Thus,

$$u = V^{\pm}_{m_{\max}} \cdot s^{m_{\max}} \quad \text{and } u^m = V^{\pm}_m \cdot s^m = V^{\pm}_{m_{\max}} \cdot \begin{bmatrix} s^m \\ \mathbf{o} \end{bmatrix}.$$

Defining $s^m_i = 0$ for $i = m+1, \dots, m_{\max}$, we extend s^m to a vector in $\mathbb{R}^{m_{\max}}$ and we know that

$$0 \le s^m_i \le s^{m_{\max}}_i \text{ for } i = 1, \dots, m_{\max}. \tag{5.3}$$

Our task is to bound the minimum and the maximum of

$$\begin{aligned} h(s^m) \;&=\; \|u - u^m\|^2 = \langle s^{m_{\max}} - s^m, s^{m_{\max}} - s^m\rangle \\ &= \underbrace{\langle s^{m_{\max}}, s^{m_{\max}}\rangle}_{=\|b\|^2} - 2 \cdot \langle s^{m_{\max}}, s^m\rangle + \underbrace{\langle s^m, s^m\rangle}_{=\|u^m\|^2} \end{aligned}$$

as a function of s^m under the constraints (5.3). From (5.3) we see that $\langle s^{m_{\max}}, s^m\rangle \ge \|s^m\|^2$, which gives the bound $h(s^m) \le \|b\|^2 - \|u^m\|^2$. On the other hand, the Cauchy-Schwarz inequality gives

$$\langle s^{m_{\max}}, s^m\rangle \le \underbrace{\|s^{m_{\max}}\|}_{=\|b\|} \cdot \underbrace{\|s^m\|}_{=\|u^m\|}$$

from which we deduce $h(s^m) \ge (\|b\| - \|u^m\|)^2$. □

6. Further discussion. Rational functions, which arise either directly or as approximations to other functions, have an important practical advantage in large scale computations if they allow for a partial fraction expansion as considered in Corollary 3.2 (ii): The Lanczos approximations can now be obtained by simultaneously performing the CG iterations for all p terms in the partial fraction expansion. Only one matrix-vector multiplication per iteration is needed for all systems together, and since CG relies on short recurrencies, it is not necessary to store all the Lanczos vectors. The storage requirements are thus determined by p, the number of poles, but they are independent of m, the iteration count. Details can be found in, e.g., [10].

As an example, consider the p pole Zolotarev rational approximation $Z_p(z)$ to $z^{-1/2}$ on an interval $[a, b]$ with $0 < a < b$. This approximation minimizes the *relative* ℓ_∞-error in $[a, b]$ over all rational functions with nominator and denominator of degree $\le p$. It has precisely the form considered in Corollary 3.2 (ii), and explicit formulae, involving the Jacobi elliptic function, are known for the all positive parameters α_i and β_i; see [19]. The use of $Z_p(z^2)z$ as an approximation to the sign function has been studied in [23]. As before we now have a proof that the Lanczos approximations for

$$Z_p(B)c \;\text{ with } B = A^2,\; c = Ab$$

have their errors decrease monotonically.

As a last contribution, let us turn back and consider the matrices T_m rather than $T^{\pm}_m$. Define

$$\hat{T}_m = \begin{bmatrix} \alpha_1 & \beta_2 & & & \\ \beta_2 & \alpha_2 & \beta_3 & & \\ & \ddots & \ddots & \ddots & \\ & & \beta_{m-1} & \alpha_{m-1} & 0 \\ & & & 0 & \alpha_m \end{bmatrix} = \begin{bmatrix} T_{m-1} & \mathbf{o} \\ \mathbf{o}^T & \alpha_m \end{bmatrix}.$$

If A is positive definite, we have

$$0 \leq \hat{T}_m \leq T_m \text{ for } m = 1, \dots, m_{\max},$$

and thus for $k = 0, 1, \dots$,

$$\hat{T}_m^k \leq T_m^k \text{ for } m = 1, \dots, m_{\max}.$$

Assume that $\text{spec}(A) \subseteq [0, b)$ and that f can be developed into a power series $f(z) = \sum_{i=0}^{\infty} \frac{f^{(i)}(0)}{i!} z^i$ that converges for $z \in [0, b]$ and that the derivatives satisfy $f^{(i)}(0) \geq 0$ for all $i = 1, 2, \dots$. From this power series representation, we immediately see that

$$0 \leq f(\hat{T}_m) \leq f(T_m).$$

Therefore, using the same argumentation as in Section 3, we obtain that for the Lanczos approximations u^m the norms $\|u^m\|$ increase monotonically, whereas the error norms $\|f(A)b - u^m\|$ are monotonically decreasing. This approach holds in particular for $f(z) = \exp(z)$, so that we are back to the results from [6] for A positive definite. Actually, we can easily generalize to A Hermitian but not necessarily positive definite.[1] We start from

$$\exp(A + \alpha I) = \exp(\alpha) \cdot \exp(A).$$

Together with the shift invariance of the Lanczos process (shifting the matrix from A to $A + \alpha I$ does not change the Lanczos vectors v^m and shifts the tridiagonal matrices from T_m to $T_m + \alpha I$) this shows that the Lanczos approximations for $\exp(A)$ are, up to the scalar scaling factor $\exp(\alpha)$, identical to those for $\exp(A + \alpha I)$. Taking α sufficiently large makes $A + \alpha I$ positive definite, from which the monotone decrease of the error norms can be deduced.

7. Conclusion. We have shown that the error of Lanczos approximations to the action of certain matrix functions on a vector is monotonically decreasing if the matrix is Hermitian and positive definite. This was done by showing that the moduli of the coefficients of the corresponding Lanczos vectors are monotonically increasing. Our results hold in particular for functions which arise as the result of a Stieltjes transform and thus for certain rational functions and for the inverse square root. The results can be extended to more general functions, in this manner including Lanczos-type approximations to the matrix sign function for indefinite matrices.

Our investigations assumed exact arithmetic throughout. It is well known that in actual numerical computations, inexact arithmetic due to rounding errors has a substantial effect on the quality of the Lanczos vectors v^i which will loose their theoretical orthogonality; see [22] for an analysis of error estimates for the CG method in this context. For our results, let us observe the following: Unless A has very small eigenvalues, the computed matrices T_m will usually still be positive definite if A is. By construction, they are also Hermitian. This implies that all what we have shown for the coefficient vectors s^m essentially remains valid in the presence of round-off. The only, but major, concern is that once the vectors v^i are not orthogonal any more, an increase (decrease) in the coefficients does not necessarily imply an increase (decrease) of the 2-norm. However, the Lanczos vectors tend to keep their orthogonality at least locally, and the coefficients in the Lanczos approximations tend to change significantly only in the last few places. These observations motivate that we can actually expect our monotonicity results to be also observed in computational pratice. At the very least they explain the *smooth* convergence behavior observed in practice.

[1] We thank Vladimir Druskin for pointing this out in a personal communication.

Acknowledgments. The author is very grateful to Michael Eiermann, Vladimir Druskin, and Valeria Simoncini for enligthening discussions on an early version of the manuscript.

REFERENCES

[1] N. I. Achiezer, *The Classical Moment Problem and Some Related Questions in Analysis (English Translation)*, Oliver and Boyd, Edinburgh, 1963.

[2] C. Berg, *Quelques remarques sur le cône de Stieltjes*, in Lecture Notes in Math., Vol. 814, Springer, Heidelberg, 1980.

[3] A. Berman and R. J. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic Press, New York, 1979. Updated edition, Classics in Applied Mathematics, Vol. 9, SIAM, Philadelphia, 1994.

[4] R. Bhatia, *Matrix Analysis*, Springer, Heidelberg, 1996.

[5] A. Boriçi, *Fast methods for computing the Neuberger operator*, in Frommer et al. [9], pp. 40–47.

[6] V. Druskin, *On monotonicity of the Lanczos approximation to the matrix exponential*, Linear Algebra Appl., 429 (2008), pp. 1679–1683.

[7] V. Druskin and L. Knizhnerman, *Extended Krylov subspaces: Approximation of the matrix square root and related functions*, SIAM J. Matrix Anal. Appl., 19 (1998), pp. 755–771.

[8] M. Eiermann and O. Ernst, *A restarted Krylov subspace method for the evaluation of matrix functions*, SIAM J. Numer. Anal., (2006), pp. 2481–2504.

[9] A. Frommer, T. Lippert, B. Medeke, and K. Schilling, eds., *Numerical Challenges in Lattice Quantum Chromodynamics*, Lect. Notes Comput. Sci. Eng., Vol. 15, Springer, Heidelberg, 2000.

[10] A. Frommer and V. Simoncini, *Matrix functions*, in Model Order Reduction: Theory, Research Aspects, and Applications, W. A. Schilders, H. A. van der Vorst, and J. Rommes, eds., Mathematics in Industry, Springer, Heidelberg, 2008, pp. 275–304.

[11] P. Henrici, *Applied and Computational Complex Analysis, Vol. 2, Special Functions–Integral Transforms–Asymptotics–Continued Fractions*, Wiley Classic Libraray, Wiley-VCH, Weinheim, 1991.

[12] N. J. Higham, *Functions of matrices*, in Handbook of Linear Algebra, L. Hogben, ed., Chapman & Hall/CRC, Boca Raton, 2007, pp. 11.1–11.13.

[13] ———, *Matrix Functions – Theory and Applications*, SIAM, Philadelphia, 2008.

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

[15] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge: Cambridge University Press, 1994.

[16] M. Ilić, I. W. Turner, and D. P. Simpson, *A restarted Lanczos approximation to functions of a symmetric matrix*, Tech. Report 8011, Queensland University of Technology, 2007. Available at http://eprints.qut.edu.au/archive//00008011/

[17] I. Moret, *Rational Lanczos approximations to the matrix square root and related functions*, Numer. Linear Algebra Appl., 16 (2009), pp. 431–445.

[18] H. Neuberger, *Overlap Dirac operator*, in Frommer et al. [9], pp. 1–17.

[19] P. P. Petrushev and V. A. Popov, *Rational Approximation of Real Functions*, Cambridge University Press, Cambridge, 1987.

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

[21] ———, *Iterative Methods for Sparse Linear Systems*, The PWS Publishing Company, Boston, 1996. Second edition, SIAM, Philadelphia, 2003.

[22] Z. Strakoš and P. Tichý, *On error estimation in the conjugate gradient method and why it works in finite precision computations*, Electron. Trans. Numer. Anal., 13 (2002), pp. 56–80. http://etna.math.kent.edu/vol.13.2002/pp56-80.dir

[23] J. van den Eshof, A. Frommer, T. Lippert, K. Schilling, and H. A. van der Vorst, *Numerical methods for the QCD overlap operator. I: Sign-function and error bounds*, Comput. Phys. Comm., 146 (2002), pp. 203–224.

[24] H. A. van der Vorst, *An iterative solution method for solving $f(A)x = b$, using Krylov subspace information obtained for the symmetric positive definite matrix A*, J. Comput. Appl. Math., 18 (1987), pp. 249–263.