PRECONDITIONING LANCZOS APPROXIMATIONS TO THE MATRIX EXPONENTIAL

JASPER VAN DEN ESHOF* AND MARLIS HOCHBRUCK*

Abstract. The Lanczos method is an iterative procedure to compute an orthogonal basis for the Krylov subspace generated by a symmetric matrix A and a starting vector v. An interesting application of this method is the computation of the matrix exponential $\exp(-\tau A)v$. This vector plays an important role in the solution of parabolic equations where A results from some form of discretization of an elliptic operator. In the present paper we will argue that for these applications the convergence behavior of this method can be unsatisfactory. We will propose a modified method that resolves this by a simple preconditioned transformation at the cost of an inner–outer iteration. A priori error bounds are presented that are independent of the norm of A. This shows that the worst case convergence speed is independent of the mesh width in the spatial discretization of the elliptic operator. We discuss, furthermore, a posteriori error estimation and the tuning of the coupling between the inner and outer iteration. We conclude with several numerical experiments with the proposed method.

1. Introduction. Using the matrix exponential operator in the numerical solution of large time-dependent problems has been an extensive area of research in the last few years. A key aspect of such exponential integrators is that they require the evaluation or approximation of the product of the exponential of the Jacobian with a vector. In this paper we consider the approximation of

$$y(\tau) = \exp(-\tau A)v, \qquad \|v\| = 1, \tag{1.1}$$

where A results from a finite difference or finite element discretization of an elliptic partial differential operator.

In 1978, Moler and van Loan [24] published their famous paper discussing nineteen dubious ways to compute the exponential of a matrix. Since then, Krylov subspace methods have been an important development towards tackling the problem (1.1) when the matrix A is very large and sparse. Hence the updated paper [23] mentions Krylov subspace methods as the "twentieth" method. For matrices A that stem from a self-adjoint elliptic partial differential equation, it has been shown [7, 8, 13] that Krylov approximations to $y(\tau)$ always yield superlinear error reduction with the superlinear decay starting after the number of steps exceeds $\|A\|^{1/2}$. Unfortunately, numerical experiments typically show that these error bounds are fairly sharp for these applications. This means that the computational complexity grows like $n^{1+1/d}$ for a uniform discretization of an elliptic operator on a d-dimensional cube with n spatial grid points. The aim of this paper is to present an algorithm which allows us to compute the approximation with $\mathcal{O}(n)$ operations. Speeding up the Lanczos process is achieved by a preconditioned operation. The emphasis in this paper is on the computation of the vector in (1.1) and we will not discuss details of the spatial discretization resulting in the matrix A. For ease of presentation, throughout this paper, we assume that A is symmetric and positive semi-definite, although most of the techniques apply to discretizations of sectorial operators as well. The demand that the matrix is positive definite is no essential restriction since indefinite matrices can easily be handled by shifting the matrix and multiplying the result with a suitable factor.

The paper is organized as follows. In Section 2 we discuss the motivation for this work. The central idea in this paper is to transform the spectrum in such a way that convergence is much faster. This leads to a new method which is proposed in Section 3, consisting of an inner and outer iteration. The convergence behavior of the new method is treated in Section 3.1 and termination strategies are the subject of Section 4. In Section 5 we discuss the tuning of the coupling between the inner and outer iteration in such a way that the overall procedure is as efficient as possible. We conclude by giving some examples of typical numerical experiments in Section 6.

We have learned recently that the method proposed in this paper has been studied earlier by Novati and Moret in [25]. In that work the authors derive the method from a different point of view and furthermore consider the more general problem of approximating the matrix exponential

* Mathematisches Institut, Heinrich-Heine Universität Düsseldorf, Universitätsstr. 1, D–40225 Düsseldorf, Germany, {eshof|marlis}@am.uni-duesseldorf.de. This work is supported by the Deutsche Forschungsgemeinschaft.

of sectorial operators. Unfortunately, this makes it much more challenging to give insight into the behavior and properties of the method. In that sense the results in this paper supplement the work in the report [25] since we consider Hermitian matrices which allows us to give a more thorough discussion. Moreover, the central question in the present paper is on how to include a preconditioner efficiently into Lanczos approximation methods for the matrix exponential whereas in [25] the linear systems are solved with a direct method

2. Motivation. When approximating the vector $y(\tau)$ in (1.1) for high dimensional matrices A, it becomes essential to exploit the sparsity of the matrix and the fact that only the action on a given vector is required. A standard idea is to use a polynomial approximation to the exponential function. In most cases this approach can be divided into two parts. First, there is the construction of a basis for the m-dimensional Krylov subspace. This part can often be summarized by the following matrix formulation:

$$AV_m = V_m T_m + \beta_m v_{m+1} e_m^*, \quad \text{where} \quad V_m e_1 = v. \tag{2.1}$$

The matrix T_m is upper Hessenberg and the vector e_j denotes the jth column of the identity matrix. The second ingredient is the approximation to the product of the matrix exponential and the vector v and is usually given by

$$y(\tau) \approx y_m(\tau) = V_m \exp(-\tau T_m) e_1. \tag{2.2}$$

Different choices for T_m lead to different polynomial approximations. The simplest choice is when T_m is defined by

$$T_m e_j = e_{j+1} \quad \text{for } j < m \quad \text{and} \quad T_m e_m = 0.$$

In this case the polynomial approximation (2.2) coincides with the standard Taylor series expansion of the matrix exponential, [23, Section 3]. Other attempts have been based on Chebyshev polynomials, e.g, [38] or, in the non-Hermitian case, on Faber polynomials in the complex plane, e.g., [29].

In this paper we focus on methods where (2.1) is constructed by means of the Lanczos method in which case the columns of the matrix V_m form an orthogonal basis for the Krylov subspace. The resulting method has been discussed by several authors and analyzed in various papers, see, for instance, [6, 7, 8, 10, 13, 33, 42] and [23] for a more extensive bibliography. We will refer to the vector $y_m(\tau)$ in this case as the *Lanczos* approximation. The advantage here is that these Lanczos approximations have a potential for exploiting the discrete nature of the spectrum of A: By interpolating the exponential in eigenvalues of A the required degree of the polynomial approximation can be much smaller than for Chebyshev approximations, for example, where a uniform approximation is constructed on an interval containing the spectrum of A. Moreover, no a priori knowledge is required about the spectral radius of the method.

Unfortunately, when A is a discrete representation of the elliptic operator, Lanczos approximations can have two distinct drawbacks. First of all, the Lanczos method often is unsuccessful in exploiting the discrete character of the spectrum as mentioned above. One example is when A is the usual finite difference/element discretization of a one dimensional Poisson operator with Dirichlet boundary conditions. In this case, the eigenvalues of A coincide with the roots of a shifted and scaled Chebyshev polynomial and the Lanczos method cannot identify any eigenvalue of A quickly, e.g, [21, Section 4.1]. This explains why we often observe for parabolic problems that Lanczos approximations do not require significantly fewer multiplications with the matrix than using a polynomial approximation based on Chebyshev polynomials, whereas they do have some overhead cost compared to this method. A related issue is that convergence analysis shows that the required number of steps is proportional to $\|A\|^{1/2}$. This means that the number of operations grows faster than linearly in the number n of spatial grid points. This should be compared with the solution of an elliptic problem which can be accomplished in $\mathcal{O}(n)$ flops using multigrid techniques.

An extension of the Lanczos approximation is to somehow precondition the iterative procedure as is done when solving linear systems. This is the aim of this paper. Other attempts in this direction have been made in [4]. The main idea there is to choose a matrix $M \approx A$ (the preconditioner) for which $\exp(-\tau M)v$ is cheap to compute and to combine this with a generalized Runge-Kutta method for the solution of the differential equation for the defect. It should also be mentioned that for the solution of parabolic partial differential equations the idea of preconditioning the time-differencing is not new. Two important developments in this area are parabolic multigrid methods [12] and methods based on waveform relaxation, e.g., [15].

3. Preconditioning Lanczos approximations. The exponential function is a quickly decaying function. This implies that the vector $\exp(-\tau A)v$ is mostly determined by the smallest eigenvalues and their corresponding invariant subspaces. Recall again that the Lanczos method does not necessarily utilize this since the first eigenvalues (and their eigenvectors) are difficult to find. This suggests that we should transform the spectrum in such a way that the Lanczos method can quickly find these eigenpairs. The simplest idea is to apply the Lanczos method to the matrix $(I + \gamma A)^{-1}$ (with $\gamma > 0$) which emphasizes the eigenvalues of importance. The Lanczos relation for the spectrally transformed matrix reads

$$(I + \gamma A)^{-1} V_m = V_m T_m + \beta_m v_{m+1} e_m^*, \quad \text{where} \quad V_m e_1 = v \text{ and } V_m^* V_m = I. \tag{3.1}$$

We define the function

$$f_\gamma^\tau(t) = \exp((1 - t^{-1})\tau/\gamma) \quad \text{for } t \in (0, 1], \quad f_\gamma^\tau(0) = 0,$$

and note that $f_\gamma^\tau((I + \gamma A)^{-1}) = \exp(-\tau A)$. As an approximation to $y(\tau)$ we use

$$y_m(\tau) = V_m f_\gamma^\tau(T_m) e_1 = V_m \exp(-\tau \widetilde{T}_m) e_1, \quad \widetilde{T}_m = \frac{1}{\gamma}(T_m^{-1} - I).$$

It is well known that the Lanczos method usually quickly finds approximations to the eigenvalues that are in some sense nicely separated from the other eigenvalues in the spectrum. A quantitative measure that expresses this separation is the ratio between the eigenvalue spread and gap. For the invariant subspace that corresponds to the smallest eigenvalue this quantity is given by

$$\frac{\lambda_n - \lambda_1}{\lambda_2 - \lambda_1}, \tag{3.2}$$

where λ_1, λ_2 and λ_n denote respectively the smallest, second smallest and largest eigenvalue of A. This quantity appears in upper bounds on the convergence speed of the Lanczos method for finding eigenpairs, see e.g., [17]. On the other hand, for the transformed system the ratio of the eigenvalue spread and gap of the invariant subspace corresponding to the smallest (positive) eigenvalue of the transformed system is bounded by

$$\frac{(1 + \gamma\lambda_1)^{-1} - (1 + \gamma\lambda_n)^{-1}}{(1 + \gamma\lambda_1)^{-1} - (1 + \gamma\lambda_2)^{-1}} \leq \frac{(1 + \gamma\lambda_1)^{-1}}{(1 + \gamma\lambda_1)^{-1} - (1 + \gamma\lambda_2)^{-1}} = \frac{1 + \gamma\lambda_2}{\gamma(\lambda_2 - \lambda_1)}.$$

The first expression follows directly from (3.2) by plugging in the transformed eigenvalues. When A stems from an appropriate discretization of an elliptic operator then the eigenvalues λ_1 and λ_2 approach the eigenvalues of the elliptic operator when the accuracy of the spatial discretization increases. As a consequence this ratio does not become arbitrarily large when the size of the matrix A increases as a result of decreasing the mesh width in the spatial discretization. For example, for the already mentioned one dimensional Poisson operator with Dirichlet boundary conditions the spread/gap ratio for the smallest invariant subspace of the continuous operator is unbounded since the spread of the spectrum is infinite. On the other hand, the spread/gap ratio for the transformed spectrum can be bounded by $(1 + \gamma 4\pi^2)/(3\gamma\pi^2)$.

The practical importance of these observations is that a good approximation to the leading invariant subspaces can in turn speed up the convergence of the method in a way that is analogous to what is often witnessed for the conjugate gradient method [41]. Our hope is that, due to the spectral transformation, the Lanczos method is more effectively able to exploit the discrete nature of an important part of the spectrum.

Finally, we mention that in our *shift-and-invert* Lanczos method we have to invert a fairly standard elliptic operator in every step. This can be done efficiently by very effective preconditioned methods that can be found in the literature. This is exploited in our numerical experiments in Section 6. The idea of this paper is to incorporate preconditioning in the Lanczos method by using it to accomplish a spectral transformation that results in more favorable convergence speed at the cost of an inner–outer iteration. We will further investigate this in the coming sections.

3.1. A priori error estimation. The Lanczos approximations are given by

$$V_m f_\gamma^\tau(T_m)e_1 = p((I+\gamma A)^{-1})v = (I+\gamma A)^{-(m-1)}q(A)v, \quad p,q \in \Pi_{m-1}.$$

The space Π_{m-1} is the space of all polynomials of degree $m-1$ or less. Our method can be alternatively characterized as constructing iterates from the class of *restricted* rational approximations defined by

$$\mathcal{R}_i^j = \left\{ p(t)(1+\gamma t)^{-i} \,|\, p \in \Pi_j \right\}.$$

This shows that the Lanczos approximations now are matrix rational approximations to $\exp(-\tau t)$ with all poles fixed at $-1/\gamma$. In the present section we will exploit this viewpoint to get more insight into the impact of our spectral transformation on the convergence speed and we will neglect the discrete nature of the spectrum. We define

$$E_i^j(\gamma) := \inf_{r \in \mathcal{R}_i^j} \sup_{t \geq 0} |r(t) - \exp(-t)|. \tag{3.3}$$

We can now easily derive the following result.

LEMMA 3.1. *Let μ be such that $A - \mu I$ is positive semi-definite. Then,*

$$\|V_m f_\gamma^\tau(T_m)e_1 - \exp(-\tau A)v\| \leq 2\exp(-\tau\mu)E_{m-1}^{m-1}(\widetilde{\gamma}) \quad \textit{with} \quad \widetilde{\gamma} = \frac{\gamma}{\tau(1+\gamma\mu)}.$$

Proof. The eigenvalues of A are all larger than μ and as a direct consequence of [33, Lemma 4.1] we have that

$$\begin{aligned}
\|V_m f_\gamma^\tau(T_m)e_1 - \exp(-\tau A)v\| &\leq 2 \inf_{p\in\Pi_{m-1}} \sup_{t\in(0,(1+\gamma\mu)^{-1}]} |f_\gamma^\tau(t) - p(t)| \\
&= 2 \inf_{p\in\Pi_{m-1}} \sup_{t\in(0,1]} |f_\gamma^\tau\left(\frac{t}{1+\gamma\mu}\right) - p(t)| \\
&= 2 \inf_{p\in\Pi_{m-1}} \sup_{t\in(0,1]} |\exp(-\mu\tau)\exp\left(\left(1-\frac{1}{t}\right)\frac{\tau(1+\gamma\mu)}{\gamma}\right) - p(t)| \\
&= 2\exp(-\mu\tau)E_{m-1}^{m-1}(\widetilde{\gamma}).
\end{aligned}$$

□

It is important to stress that this error estimate is independent of the norm of A and only the first (smallest) eigenvalue of A plays a modest role in the form of μ. We also see that the restriction of A to positive semi-definite matrices is too stringent since reasonable upper bounds are obtained if the spectrum is bounded from below by a modest constant.

A priori error bounds for the proposed method can be obtained by estimating $E_{m-1}^{m-1}(\widetilde{\gamma})$. A good choice for $\widetilde{\gamma}$ automatically leads to a good choice for γ (if τ and a reasonable estimate for μ is known). Therefore we want to derive upper bounds to (3.3) where we, without loss of generality, assume that $\tau = 1$ and $\mu = 0$ which leads to $\widetilde{\gamma} = \gamma$. Our attempts for bounding this quantity

have resulted in upper bounds which are too pessimistic. However, some insight can be obtained by considering the asymptotic situation for $m \to \infty$. The use of the class $\mathcal{R}_j^{j-1}$ to approximate $\exp(-t)$ for $t \in [0, \infty)$ is discussed by Saff et al. [34]. The authors' analysis implies that the best shift is given by $\gamma = 1/j$ which was shown to lead to geometric convergence with an asymptotic decay rate between $1/5.828$ and $1/2$. This result was later sharpened by Andersson [1] who showed the following result.

THEOREM 3.2 (Andersson [1]). *Asymptotically the optimal value for γ is given by $\sqrt{2}/j$ for which we have*

$$\lim_{j\to\infty} \left(E_j^j(\sqrt{2}/j)\right)^{1/j} = \frac{1}{\sqrt{2}+1}.$$

Hence, the value of γ should ideally be chosen to be inversely proportional to the number of Lanczos steps. Our shift-and-invert strategy requires a fixed a priori chosen value of γ. In this case we cannot expect linear convergence. Based on work in [34] we can show the following result.

THEOREM 3.3. *Let $0 < \gamma < 1$. Asymptotically we have*

$$\left(E_j^{j-1}(\gamma)\right)^{1/\sqrt{j}} \le \kappa_j, \quad \textit{where} \quad \lim_{j\to\infty} \kappa_j = \exp\left(-\sqrt{\frac{2(1-\gamma)}{\gamma}}\right).$$

Proof. Equations (3.4) and (3.14) in [34] are

$$E_j^{j-1}(\gamma) \le \sqrt{2}\rho_{j-1}(1/\gamma), \tag{3.4}$$

where

$$\rho_{j-1}(1/\gamma)^2 < \frac{1}{2}\exp(1/\gamma)\int_0^\infty \exp(-s/\gamma)\left(\frac{s}{s+1}\right)^{2j-2}\omega_j(s)ds,$$

and

$$\omega_j(t) = \frac{(j+1)(2j+1)t^2}{(1+2t)(1+t)} + \frac{(4j+5)t^4}{(1+2t)^2(1+t)} + \frac{4t^6}{(1+2t)^3(1+t)}.$$

We have

$$\omega_j(t) \le \frac{t^2}{(1+t)^2}\overline{\omega_j}(t), \qquad \overline{\omega_j}(t) := (j+1)(2j+1) + \frac{4j+5}{4}t + \frac{1}{2}t^2.$$

Using the identity $\exp(-s/\gamma) = \exp(-s/\gamma + s)\exp(-s)$ and the mean value theorem we obtain

$$\begin{aligned}
\rho_{j-1}(1/\gamma)^2 &< \frac{1}{2}\max_{t\ge 0}\left\{\exp\left(\frac{1-t}{\gamma}+t\right)\left(\frac{t}{t+1}\right)^{2j}\right\}\int_0^\infty \exp(-s)\overline{\omega_j}(s)ds \\
&\le \max_{t\ge 0}\exp(\psi_j(t))\left(j^2+2j+\frac{13}{8}\right) \\
&\le 5j^2\max_{t\ge 0}\exp(\psi_j(t))
\end{aligned}$$

with

$$\psi_j(t) = \frac{1}{\gamma}(1-t) + t + 2j(\log(t) - \log(t+1)).$$

We find that

$$\max_{t\ge 0}\psi_j(t) = \frac{1}{2\gamma}\left(3 - D_j - \gamma + 4j\gamma\log\left(-\frac{\gamma - 1 + D_j}{\gamma - 1 - D_j}\right)\right) =: \overline{\psi_j},$$

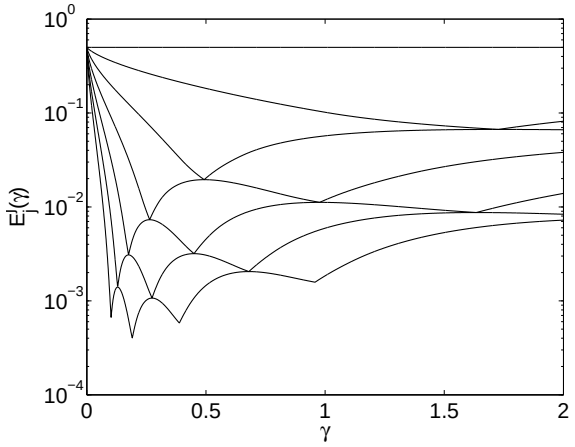

FIG. 3.1. *$E_j^j(\gamma)$ as function of γ for $j = 0, \ldots, 6$.*

j	$E_j^j(\gamma_{\text{opt}})$	γ_{opt}	j	$E_j^j(\gamma_{\text{opt}})$	γ_{opt}
1	$6.7 \cdot 10^{-2}$	$1.73 \cdot 10^{0}$	11	$4.0 \cdot 10^{-6}$	$9.90 \cdot 10^{-2}$
2	$2.0 \cdot 10^{-2}$	$4.93 \cdot 10^{-1}$	12	$1.6 \cdot 10^{-7}$	$1.19 \cdot 10^{-1}$
3	$7.3 \cdot 10^{-3}$	$2.64 \cdot 10^{-1}$	13	$6.1 \cdot 10^{-7}$	$1.00 \cdot 10^{-1}$
4	$3.1 \cdot 10^{-3}$	$1.75 \cdot 10^{-1}$	14	$2.5 \cdot 10^{-7}$	$8.64 \cdot 10^{-2}$
5	$1.4 \cdot 10^{-3}$	$1.30 \cdot 10^{-1}$	15	$1.0 \cdot 10^{-7}$	$7.54 \cdot 10^{-2}$
6	$4.0 \cdot 10^{-4}$	$1.91 \cdot 10^{-1}$	16	$4.0 \cdot 10^{-8}$	$8.67 \cdot 10^{-2}$
7	$1.6 \cdot 10^{-4}$	$1.44 \cdot 10^{-1}$	17	$1.6 \cdot 10^{-8}$	$7.63 \cdot 10^{-2}$
8	$6.5 \cdot 10^{-5}$	$1.90 \cdot 10^{-1}$	18	$6.6 \cdot 10^{-9}$	$6.78 \cdot 10^{-2}$
9	$2.4 \cdot 10^{-5}$	$1.47 \cdot 10^{-1}$	19	$2.7 \cdot 10^{-9}$	$7.62 \cdot 10^{-2}$
10	$9.7 \cdot 10^{-6}$	$1.19 \cdot 10^{-1}$	20	$1.1 \cdot 10^{-9}$	$6.82 \cdot 10^{-2}$

TABLE 3.1

Numerical approximations to the optimal value of γ, γ_{opt}, and the corresponding value $E_j^j(\gamma_{opt})$.

with $D_j = \sqrt{(1-\gamma)(8j\gamma - \gamma + 1)}$ (the maximum is attained at $t = \frac{1}{2}(-1 + \frac{D_j}{1-\gamma})$).

Taking the limit of $\overline{\psi_j}/\sqrt{j}$ for $j \to \infty$ and using (3.4) shows the desired result. □

In our context we are interested in approximations where the degree is modest. In the absence of insightful analytical estimates, we have decided to give numerical approximations to the values of $E_{m-1}^{m-1}(\gamma)$ and more interestingly the location of the optimal value of γ, that is, the value that minimizes $E_{m-1}^{m-1}(\gamma)$. These results can be obtained using the Remez method, e.g., [31, Section 1.3], by constructing the optimal polynomial approximation to $f_\gamma^1(t)$ on the interval $[0, 1]$, see the proof of Lemma 3.1 for more details.

In Figure 3.1 we have plotted $E_{m-1}^{m-1}(\gamma)$ as a function of γ for seven different values of m. This shows a very interesting pattern that is summarized by the following observation.

CONJECTURE 3.1. *The function $E_{j-1}^{j-1}(\gamma)$ for $j = \{2, 3, \ldots\}$ has $j-1$ local minima. These minima coincide with local maxima of $E_j^j(\gamma)$. The minima of $E_{j-1}^{j-1}(\gamma)$ interlace the minima of $E_j^j(\gamma)$.*

If the curve $E_j^j(\gamma)$ has a local minimum this appears to coincide with the fact that the error curve for the polynomial approximation at this point seems to have one additional alternation point. In Table 3.1 the optimal value of γ and the corresponding error have been reported. We used a straightforward search algorithm for this in combination with Remez approximations.

As mentioned previously we must choose our shift γ before we start our iterative procedure. Changing it during the process is too expensive. Although this appears to be a drawback from a theoretical point of view, in practice this is no problem since the required precision is known

beforehand. For instance, if one is interested in a precision of about 10^{-5} one can consider Table 3.1 and decide to choose $\gamma = 0.119\tau$ since this γ minimizes our upper bound from Lemma 3.1.

3.2. Discussion. One might wonder if there is anything to gain by using different real shifts in the Lanczos method, that is, does it pay to construct an approximation from the class

$$\widetilde{\mathcal{R}}_i^j = \left\{ p(t) \prod_{i=1}^{i} (1+\gamma_i t)^{-1} \,|\, p \in \Pi_j \right\} \quad \text{where } \gamma_i \in \mathbb{R}.$$

Using different shifts in a Lanczos method can be facilitated by rational Krylov methods [32]. Interestingly enough, it was conjectured in two independent papers [19, 22] that the optimal approximation from $\widetilde{\mathcal{R}}_i^j$ is also contained in $\mathcal{R}_i^j$. This conjecture was later shown to be true [2, Theorem 1]. If one is willing to work with quadratic factors, or linear factors with complex shift, one can consider using Chebyshev rational approximations. A comparison for solving linear parabolic equations between the Crank-Nicolson method and using a Chebyshev rational approximation can be found in [5].

Using optimal restricted rational approximations on the negative real axis to approximate the exponential function for solving semi-discretized parabolic differential equations is discussed in [34, p. 319]. The authors emphasize the computational advantages when only shifted systems with constant real shifts have to be solved. Many other rational approximations, for example, Chebyshev rational approximations, and the diagonal elements of the Padé table, cannot be factored into linear factors with real coefficients. Restricted rational approximations to the matrix exponential of Padé type have been used by Nørsett et al. [26, 27, 28], see also the paper by Van Iseghem [43]. These rational approximations play an important role in the development of so-called semi-implicit diagonal Runge-Kutta methods, which are attractive since in one step of the method only systems have to be inverted with a constant real shift. In [16] explicit expressions for the restricted rational approximations are derived that interpolate the exponential function in an equispaced mesh. It is shown by a numerical example that rational interpolations are better suited for approximating the matrix exponential than restricted Padé approximations when A originates from a Poisson operator. The main difference of all these techniques with the method considered in this paper is that the use of the Lanczos method allows us to exploit the favorable eigenvalue spectrum of the transformed matrix as discussed previously. This will become clear when we discuss numerical examples in Section 6.

The proposed method in this paper is also related to the work on *extended* Krylov subspaces in [9], where the authors are interested in Krylov subspaces for A extended by a Krylov subspace generated by A^{-1}. They show that these subspaces have a superior approximation quality (although they consider a different function class). One essential difference is that, in the present paper, we work with shifted inverses and, considering the results in the previous section, we mention that this is pivotal. Also, the a priori error bounds presented in [9] depend on the condition number of the matrix A. Finally, a difference is that we are interested in preconditioned solvers for the solution of linear systems. We will discuss this further in Section 5.

4. A posteriori error estimation. In this section we consider strategies for terminating the shift-and-invert Lanczos process as soon as $y_m(\tau)$ is within a predefined distance to the sought-after vector. To this purpose we first derive an explicit expression for the error. In practice this relation cannot be evaluated exactly. Subsequently we will discuss a few ideas to come to practical error estimators.

By rewriting (3.1) we find that

$$V_m T_m^{-1} = (I+\gamma A)V_m + \beta_m \widetilde{v}_{m+1} z_m^*, \quad \widetilde{v}_{m+1} = (I+\gamma A)v_{m+1}, \quad z_m = T_m^{-1} e_m.$$

Hence, we obtain the following relation

$$AV_m = V_m \widetilde{T}_m - \frac{\beta_m}{\gamma} \widetilde{v}_{m+1} z_m^*, \quad \widetilde{T}_m = \frac{1}{\gamma}(T_m^{-1} - I). \tag{4.1}$$

Using this relation we see that our approximation satisfies

$$y_m'(t) = -V_m \widetilde{T}_m \exp(-t\widetilde{T}_m)e_1 = -Ay_m(t) + g_m(t), \quad g_m(t) = -\frac{\beta_m}{\gamma}\widetilde{v}_{m+1} z_m^* \exp(-t\widetilde{T}_m)e_1,$$

with $y_m(0) = v$, whereas the exact solution fulfills

$$y'(t) = -Ay(t), \qquad y(0) = v.$$

We define $e_m(t) = y_m(t) - y(t)$ and get

$$e_m'(t) = -Ae_m(t) + g_m(t), \quad e_m(0) = 0. \tag{4.2}$$

Finally, an explicit expression for the error follows by writing down the solution of this ODE using variation of constants, which yields

$$e_m(\tau) = \int_0^\tau \exp(-(\tau - t)A) g_m(t)\,\mathrm{d}t = -\frac{\beta_m}{\gamma}\int_0^\tau \exp(-(\tau - t)A)\widetilde{v}_{m+1} z_m^* \exp(-t\widetilde{T}_m)e_1\,\mathrm{d}t \tag{4.3}$$

or, equivalently,

$$e_m(\tau) = -\frac{\beta_m}{\gamma} X_m v_{m+1} \tag{4.4}$$

where

$$X_m := \int_0^\tau e_m^*(I + \gamma\widetilde{T}_m)\exp(-t\widetilde{T}_m)e_1 \exp(-(\tau - t)A)(I + \gamma A)\,\mathrm{d}t \tag{4.5}$$

Hence, the error is determined by the norm of the operator X_m. Unfortunately, the precise size of the error is hard to assess through (4.3) since this quantity is difficult to evaluate. In the remainder of this section we focus on practical error estimators.

We substitute

$$\exp(-(\tau - t)A) = I - (\tau - t)A + \frac{1}{2}(\tau - t)^2 A^2 + \dots$$

in (4.3) and obtain

$$e_m(\tau) = -\frac{\beta_m}{\gamma}\int_0^\tau \big(I - (\tau - t)A + \frac{1}{2}(\tau - t)^2 A^2 + \dots\big)\widetilde{v}_{m+1} z_m^* \exp(-t\widetilde{T}_m)e_1\,\mathrm{d}t.$$

By defining the functions

$$\phi_j(-\tau A) = \frac{1}{(j-1)!}\int_0^\tau (\tau - t)^{j-1}\exp(-tA)\,\mathrm{d}t, \tag{4.6}$$

we get the following result.

LEMMA 4.1. *The error $e_m(\tau)$ satisfies the following expansion:*

$$e_m(\tau) = -\frac{\beta_m}{\gamma}\sum_{j=0}^{\infty} z_m^* \phi_{j+1}(-\tau\widetilde{T}_m)e_1(-A)^j(I + \gamma A)v_{m+1}. \tag{4.7}$$

By partial integration it can be shown that the ϕ-functions can be computed by means of a simple recursion. This series is similar to the one given in [33] for standard Lanczos approximations. A practical error estimator follows by taking only the first summand of the series:

$$\|e_m(\tau)\| \approx \frac{\beta_m}{\gamma}|z_m^*\phi_1(-\tau\widetilde{T}_m)e_1|\,\|(I + \gamma A)v_{m+1}\|. \tag{4.8}$$

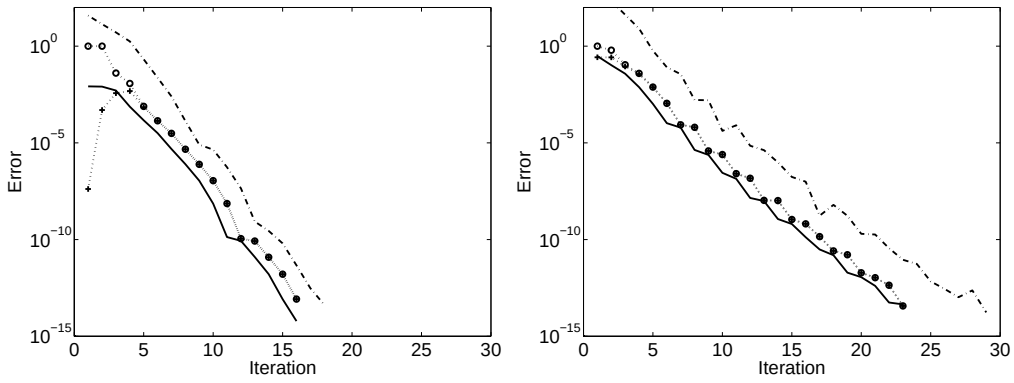

FIG. 4.1. *Error of the shift-and-invert Lanczos approximation (solid) and the error estimators (4.8) (dash-dot), (4.9) (o) and the difference of two iterates (+). In both pictures* $\gamma = \tau/10$. *Left picture:* $\tau = 1/5$ *and right picture:* $\tau = 1/100$.

More advanced strategies can be taken as in the software package [35] for the standard Lanczos approximation which is based on the analogous expansion in [33]. We will not follow this direction.

Our derivation of (4.7) immediately suggests alternative approximations to the error by using more advanced approximations than the Taylor series expansion, e.g., using Chebyshev series or again by Lanczos approximations. We can substitute for instance into (4.3) the approximation:

$$(I + \gamma A)\exp(-(\tau - t)A)v_{m+1} \approx V_{m+j}(I + \gamma\widetilde{T}_{m+j})\exp(-(\tau - t)\widetilde{T}_{m+j})e_{m+1}.$$

The obvious advantage of this approximation is that the error estimator only requires information that is already computed in the Lanczos process, so there is no additional work necessary. Not surprisingly, it is very easy to see that with this substitution the norm of (4.3) is equal to $\|y_{m+j}(\tau) - y_m(\tau)\|$.

Estimating the error by the difference of two (consecutive) approximations is straightforward and simple, however it has some less appealing properties. For example, stagnation of the process leads to excessively optimistic error estimates. In our practical experience with the presented method, no problems were encountered if γ is sensibly chosen and therefore convergence is fast enough. A different problem occurs in the first few steps where we see that $\|y_m(\tau)\| \ll \|y(\tau)\|$ which can be explained by the fact that the eigenvalues of $\widetilde{T}_m$ are far from the smallest eigenvalue of A. Numerical experience shows that an effective upper bound for the *relative* error in step m is given by

$$\delta_m = \frac{\|y_m(\tau) - y_{m-j}(\tau)\|}{\|y_m(\tau)\|}.$$

With this assumption we have that $\|y(\tau)\| \lesssim \delta_m\|y(\tau)\| + \|y_m(\tau)\|$ and we correct the error estimator as follows

$$\|e_m(\tau)\| \lesssim \frac{\delta_m}{1 - \delta_m}\|y_m(\tau)\|. \tag{4.9}$$

A trivial upper bound is $\|e_m(\tau)\| \leq 1 + \|y_m(\tau)\|$. In our experiments, we take the minimum of this bound and (4.9) as an error estimator to circumvent very large error estimates in the first steps. In all our experiments we, furthermore, have taken j to be equal to one.

The corrected estimator (4.9) turns out to be very effective in experiments and we now will demonstrate this with a simple example. The matrix A is a central finite difference discretization with 30 grid points in both directions of the Poisson operator on the unit square with homogeneous Dirichlet boundary conditions. Figure 4.1 illustrates that for $\tau = 1/5$ estimating the error by the difference of two iterates can lead to overly optimistic estimates in the start. This behavior is corrected by (4.9). Typically we see that the estimator starts in the neighborhood of one and after

a few iterations it quickly decreases as more information on the norm of the error becomes available. We remark that these two figures demonstrate some interesting properties of the convergence of the proposed method; we will discuss this in detail in Section 6.

5. Tuning the inner–outer iteration. We want to compute an approximation $y_m(\tau)$ which is at an absolute distance of about ϵ from the true vector. In step j of the shift-and-invert Lanczos process we have to solve a linear system involving an elliptic operator. If this is accomplished using an iterative solver we have to prescribe a constant η_j that determines the precision for this inner solve, that is

$$\|r_j\| \leq \eta_j \quad \text{with} \quad r_j = v_j - (I + \gamma A)c_j. \tag{5.1}$$

The vector c_j is the approximate solution of the linear system. In this section we look in more detail at the influence of these inaccuracies on the Lanczos approximation. Numerical experiments show that the sensitivity towards these errors in the Lanczos process is similar to that witnessed in other applications of the Lanczos method where matrix-vector products are perturbed. A precise error analysis, however, turns out to be very challenging and we restrict ourselves to more heuristical arguments here.

The errors in the solution of the linear systems in (5.1) result in a perturbed Lanczos relation (3.1) which after some manipulation can be recast as a perturbed form of (4.1):

$$AV_m = V_m\widetilde{T}_m - \frac{\beta_m}{\gamma}\widetilde{v}_{m+1}z_m^* - \frac{1}{\gamma}F_m T_m^{-1}. \tag{5.2}$$

Here, the jth column of the matrix F_m equals the residual of the linear system that is solved in the jth step of the Lanczos process, in other words we have $F_m e_j = r_j$. We warn the reader that the tridiagonal matrix T_m and the matrix V_m in this relation are different from the tridiagonal matrix generated by the Lanczos method in the error free case. For instance, there is no guarantee that the matrix V_m is orthogonal. Another source of errors is rounding errors resulting from the use of floating point arithmetic. Since these errors are typically many orders of magnitude smaller than the approximation errors introduced in the solution of the shifted systems, we will neglect rounding errors. Using the relation (5.2) and reasoning along the same lines as in the previous section, we find the following bound for the error:

$$\|e_m(\tau)\| \leq \frac{\beta_m}{\gamma}\|X_m v_{m+1}\| + \frac{1}{\gamma}\|\int_0^\tau \exp(-(\tau-t)A)F_m T_m^{-1}\exp(-t\widetilde{T}_m)e_1\,\mathrm{d}t\|.$$

The matrix X_m is again as defined in (4.5). If we assume that $\eta_j = \epsilon$, and therefore $\|r_j\| \leq \epsilon$, then a crude estimate for the second quantity is given by

$$\frac{1}{\gamma}\|\int_0^\tau \exp(-(\tau-t)A)F_m T_m^{-1}\exp(-t\widetilde{T}_m)e_1\,\mathrm{d}t\| \leq \|T_m^{-1}e_1\|\frac{\sqrt{m}\tau}{\gamma}\epsilon.$$

Here we have used the Cauchy-Schwarz inequality. This shows that

$$\|e_m(\tau)\| \leq \frac{\beta_m}{\gamma}\|X_m v_{m+1}\| + \|(I+\gamma\widetilde{T}_m)e_1\|\frac{\sqrt{m}\tau}{\gamma}\epsilon. \tag{5.3}$$

We see that in the first few iterations the first term is the dominating quantity and, in fact, determines the convergence behavior of the perturbed method. Standard practice would be to view the first term in (5.3) as the error computed by an *exact* Lanczos process on a nearby matrix. This is a common approach in the analysis of perturbed Lanczos methods, see for instance [6]. Unfortunately, this idea does not easily carry over to the method of the present paper since the matrix X_m defined by (4.5) not only depends on the tridiagonal matrix T_m, but also on the original matrix A.

As an alternative we focus on the expansion in Lemma 4.1 and in particular the quantities $z_m^*\phi_{j+1}(-\tau\widetilde{T}_m)e_1$ as a function of the iteration number m. We assume that for all m, the eigenvalues of T_m are contained in an interval $(0, \alpha]$. In practice the positive definiteness of the tridiagonal

matrix T_m can be guaranteed by using a version of the Lanczos method based on coupled recurrences in combination with solving the linear systems by the conjugate gradient method. Bounds on the eigenvalues of the tridiagonal matrix T_m generated by a perturbed Lanczos process can be found in [30]. For every polynomial $p \in \Pi_{m-2}$, we have

$$\begin{aligned} |z_m^* \phi_{j+1}(-\tau \widetilde{T}_m) e_1| &= |e_m^* \left(T_m^{-1} \phi_{j+1}(-\tau \widetilde{T}_m) e_1 - p(T_m) \right) e_1| \\ &\leq \inf_{p \in \Pi_{m-2}} \sup_{t \in (0,\alpha]} |\tfrac{1}{t} \phi_{j+1} \left(\tfrac{\tau}{\gamma}(1 - \tfrac{1}{t}) \right) - p(t)|. \end{aligned}$$

This suggests that the coefficients in the expansion (4.8) decrease as a function of m with our assumption on the eigenvalues of the tridiagonal matrices. This argument confirms the observation that the vector $\|X_m v_{m+1}\|$ gets many orders of magnitude smaller than ϵ before it stagnates. As a consequence this means that in the end the attainable precision of the method is essentially bounded by the second term in (5.3) and we have argued that the method can achieve a final precision of $\mathcal{O}(\epsilon)$ when all linear systems are solved with a fixed residual precision of ϵ.

There are some recent research efforts in the analysis of the effect of approximate matrix vector products on Krylov subspace methods, e.g., [3, 11, 36, 39]. These works show that when solving linear systems, or computing eigenvectors, by means of a Krylov subspace method, accurate approximations to the matrix-vector product are necessary in the first iterations, but this precision can be relaxed as the method converges. It is interesting to see if such strategies can be extended to the current context where inexact matrix-vector products result from the errors in the solution of the shifted systems.

A key issue of the analysis in [11, 36, 39] is that the computed vector in the approximation subspace has a decreasing pattern. Suppose that after m steps of the error free method we have computed an approximation that is sufficiently accurate, i.e., $\|y_m(\tau) - y(\tau)\| \leq \epsilon$. We have for the component of the solution in the direction of the vector $v_j = V_m e_j$:

$$|v_j^* y_m(\tau)| \leq |v_j^*(y_m(\tau) - y(\tau))| + |v_j^* y(\tau)| \leq \epsilon + \|(I - V_{j-1} V_{j-1}^*) y(\tau)\| \approx \epsilon + \|e_{j-1}(\tau)\|.$$

In the last step we have assumed that the iterate from step $j-1$ is approximately equal to the optimal approximation from the $(j-1)$st Krylov subspace. Intuitively, this shows that the errors in the Lanczos process in the first steps have a relatively large impact since the solution lies mainly in the direction of the first Lanczos vectors. For this reason, we propose the following strategy for controlling the tolerances of the linear systems:

$$\eta_j = \frac{\epsilon}{\|e_{j-1}(\tau)\| + \epsilon}. \tag{5.4}$$

Practical strategies result from replacing $\|e_{j-1}(\tau)\|$ in (5.4) with a suitable upper bound as discussed in the previous section, for instance (4.9). Following nomenclature from [3] we will refer to (5.4) as a *relaxation* strategy for the remainder of this paper. In the current context such a strategy turns out to be much harder to analyze than for linear systems due to the nonlinear character. Nevertheless, this strategy worked very satisfactorily in all our numerical experiments. We will give an example of this in Section 6.1.

6. Numerical experiments. In this section we discuss some numerical experiments that are typical of our experiences with the presented method. All experiments are conducted in Matlab. The purpose of our first experiment is to illustrate properties of the convergence behavior of the new method. The matrix A results from a simple central finite difference discretization of the one dimensional Poisson operator with periodic boundary conditions. Although this matrix is of little practical importance, it leads to representative results and, more importantly, it allows us to compute the true solution with a fast Fourier transform for comparison purposes. Moreover, the smallest eigenvalue of this matrix is zero, which means that the norms of the true solutions are not very sensitive to the choice of τ. The dimension of the matrix is 10^5 and in all experiments in this section we take $\gamma = \tau/10$. This latter choice is not necessarily optimal. However, using this fixed strategy in all experiments shows that the method is not very sensitive towards a very

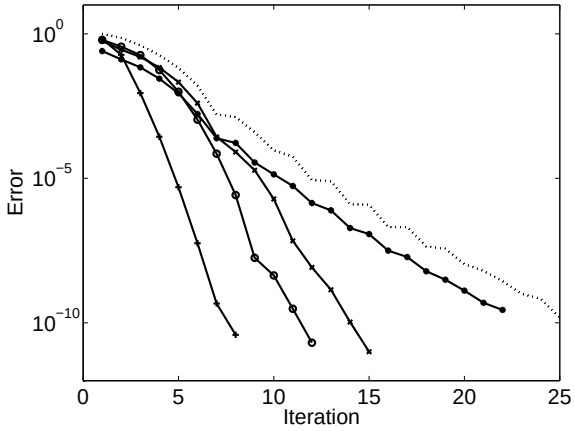

FIG. 6.1. *Error as a function of the number of iterations for $\tau = 1/2$ (+), $\tau = 1/20$ (o), $\tau = 1/50$ (×), $\tau = 1/2000$ (*) and the upper bound from Lemma 3.1 (dotted).*

precise choice for γ. Notice that the traditional Lanczos approximation is not an option here since upper bounds suggest, and experiments often show, that for one dimensional elliptic operators convergence starts after the number of iteration steps is almost equal to the number of spatial grid points. The results of our numerical experiments are reported in Figure 6.1 where we have included as a reference the upper bound from Lemma 3.1.

Interestingly, the number of Lanczos iterations increases when τ becomes smaller. An explanation lies in fact that for larger values of τ a relatively small portion of the eigenvalues essentially determines the final vector. Due to the nice separation of these eigenvalues, as discussed in Section 3, Krylov subspaces quickly come to contain approximations to the corresponding invariant subspaces and therefore only a few iteration steps are required. When τ decreases, loosely speaking, more and more eigenvalues determine the final result and a more uniform approximation is more effective then. This shows that the Lanczos approximation indeed takes advantage of the nice eigenvalue distribution of the transformed system and it shows that convergence might be much faster than is expected from treating the problem as a uniform approximation problem on an unbounded interval as we did in our upper bounds in Section 3.1.

6.1. The impact of errors in the solution of the linear systems. In the remainder of this section we take as model equation the parabolic partial differential equation in two spatial dimensions:

$$\frac{\partial}{\partial t}u(t,x,y) = \mathcal{L}u(t,x,y), \quad (x,y) \in \Omega = (0,1)^2, \quad 0 \le t \le \tau, \tag{6.1}$$

subject to the initial condition

$$u(0,x,y) = u_0(x,y).$$

The operator $\mathcal{L}$ is a self-adjoint linear second-order elliptic operator,

$$\mathcal{L} = \frac{\partial}{\partial x}a(x,y)\frac{\partial}{\partial x} + \frac{\partial}{\partial y}b(x,y)\frac{\partial}{\partial y} + c(x,y), \tag{6.2}$$

with time independent boundary conditions. The matrix A originates from discretizing the operator $\mathcal{L}$ using central finite differences on a square grid.

We investigate the impact of errors in the solution of the shifted systems (the inner iteration) on the final accuracy of the overall method. Therefore, we have experimented with two choices of η_j in (5.1): taking $\eta_j = \epsilon$ (fixed) and the relaxation strategy given by (5.4) in Section 5. For the latter strategy we have replaced the norm of the error by the approximation given by (4.9).

Moreover, the Lanczos process was terminated as soon as this quantity dropped below ϵ. The matrix A stems from discretizing the operator $\mathcal{L}$ in (6.2) with

$$a(x,y) = 1 + y - x, \quad b(x,y) = 1 + x + x^2, \quad c(x,y) = 0,$$

and homogenous Dirichlet boundary conditions on the western and eastern boundaries and a homogenous Neumann boundary condition on the northern and southern boundaries of the domain. For the discretization of the operator we have used 40 grid points in both directions which still allows us to compute the true solution with a dense method for comparison purposes. We are interested in a precision of $\epsilon = 10^{-6}$.

The results of our numerical experiments are summarized in Table 6.1. The shifted linear systems are solved with the conjugate gradient method (CG) preconditioned by a standard incomplete LU preconditioner. In the next section we use more advanced and effective preconditioners. As a measure for the total amount of work we have reported the total number of applications of the ILU preconditioner in the inner iterations and the number of outer iterations. Moreover, we have included the error of the computed approximation at termination as well as the error estimated at this point by (4.9).

	Fixed			Relaxed		
τ	in./out.	error	estimate	in./out.	error	estimate
1	228/5	$3.8 \cdot 10^{-8}$	$3.9 \cdot 10^{-7}$	152/5	$3.7 \cdot 10^{-8}$	$4.0 \cdot 10^{-7}$
1/5	348/10	$6.5 \cdot 10^{-8}$	$3.1 \cdot 10^{-7}$	230/10	$1.9 \cdot 10^{-7}$	$3.1 \cdot 10^{-7}$
1/10	326/12	$2.3 \cdot 10^{-7}$	$8.9 \cdot 10^{-8}$	202/12	$5.7 \cdot 10^{-7}$	$9.6 \cdot 10^{-8}$
1/50	182/13	$8.2 \cdot 10^{-7}$	$7.0 \cdot 10^{-7}$	114/13	$8.9 \cdot 10^{-7}$	$7.0 \cdot 10^{-7}$
1/100	140/14	$1.2 \cdot 10^{-6}$	$2.3 \cdot 10^{-7}$	82/14	$1.3 \cdot 10^{-6}$	$2.1 \cdot 10^{-7}$

TABLE 6.1
Numerical results for $\gamma = \tau/10$.

The table shows that in all cases we achieve the required precision of about 10^{-6}. The use of a relaxation strategy as opposed to using a fixed precision for the inner iterations has no influence on the total number of Lanczos (outer) iterations. The amount of work in the inversions of the linear systems, measured as the total number of applications of the ILU preconditioner, is reduced by about 30 to 40 percent. This is comparable with reductions that are seen in applications of inexact Krylov methods in other areas, see for a more detailed discussion [40, Section 3]. For the purpose of illustration we have included in Figure 6.2 and Figure 6.3 a visual representation of the results for $\tau = 1/10$ and $\tau = 1/100$ for both strategies.

Again, we have that the number of Lanczos iterations increases when τ becomes smaller as in the previous example. This is a combination of the fast convergence of the method for modest values of τ and, furthermore, is partially explained by the small norm of the true solution. On the other hand, we see that the number of CG steps per outer iteration decreases for smaller values of τ. This is a result of our choice $\gamma = \tau/10$ which makes the matrix more and more diagonally dominant when τ becomes smaller. For comparison purposes we mention that solving a linear system with this matrix and right-hand side to a precision of 10^{-6} requires 49 steps of the GMRES method preconditioned by an incomplete LU decomposition, a factor 4 less than for computing the exponential with $\tau = 1/10$.

From Figures 6.2 and 6.3 it seems that the convergence of the method stagnates now and then for one iteration. A possible explanation is that the Lanczos approximation mimics a property of the optimal uniform approximation. This is related to the observation in Conjecture 3.1 and can also be seen in Figure 3.1. Since, for smaller values of τ the Lanczos method constructs a more and more uniform approximation on an interval, this might explain why these plateaus occur more often for small values of τ, see Figures 6.2 and 6.3, and also Figure 4.1. We notice that in this experiment the error estimator (4.9) is a little too optimistic in the neighborhood of 10^{-5} for $\tau = 1/10$ which is caused by the temporary stagnation of the method at that point. Although such underestimation is not large and, in our experience, seldomly occurs, it can be resolved by switching to $j = 2$ in (4.9).

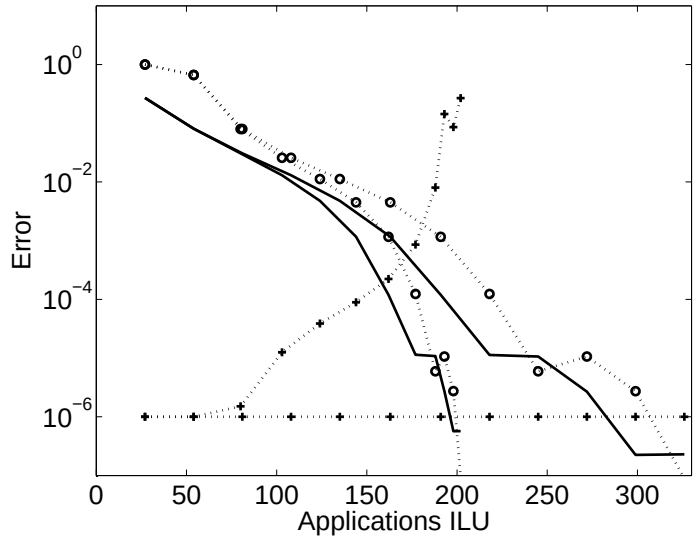

FIG. 6.2. *Error (solid), tolerance η_j (+) and error estimator (o) as a function of the total number of applications of the ILU preconditioner for $\eta_j = \epsilon$ and η_j as in (5.4) for $\tau = 1/10$. See also Table 6.1.*

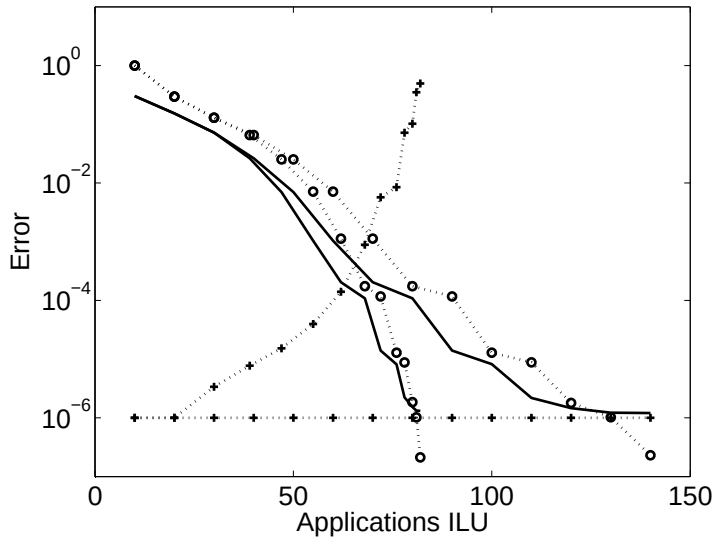

FIG. 6.3. *Error (solid), tolerance η_j (+) and error estimator (o) as a function of the total number of applications of the ILU preconditioner for $\eta_j = \epsilon$ and η_j as in (5.4) for $\tau = 1/100$. See also Table 6.1.*

6.2. The influence of mesh refinement. In our next experiment we explore the impact of the number of spatial grid points on the method. In order to be able to handle large matrices we have used the SAMG algebraic multigrid package [37] for solving the shifted systems. This package is written in Fortran and we have interfaced it with Matlab. For ease of illustration we have taken the standard Poisson operator on a square grid with Dirichlet boundary conditions, i.e., $a(x, y) = b(x, y) = 1$ and $c(x, y) = 0$ in (6.2). We aim at a precision of $\epsilon = 10^{-8}$. In Table 6.2 we have reported the total number of cycles done in the inner iterations and the number of Lanczos iterations necessary to reduce the error estimator below 10^{-8} for different combinations of n (number of spatial grid points) and τ. Again we have used two strategies for choosing η_j: $\eta_j = \epsilon$ and η_j as in (5.4) with the norm of the error replaced by the estimator (4.9). The matrices were unfortunately too large to check if the obtained solution fulfills our accuracy requirement by comparing it against the outcome of a dense method. Instead, we have verified the results via

another sparse method set to a smaller tolerance. In all cases the computed vectors have an error that is at most of the order of the required precision.

n	32^2	64^2	128^2	256^2	512^2	1024^2
			Fixed			
$\tau = 1$	20/4	24/4	24/4	27/4	28/4	28/4
$\tau = 1/10$	65/13	70/13	78/13	78/13	90/13	91/13
$\tau = 1/100$	104/15	48/16	89/18	90/18	108/18	108/18
$\tau = 1/1000$	23/8	52/11	123/14	64/16	76/16	80/16
			Relaxed			
$\tau = 1$	18/4	21/4	21/4	23/4	24/4	24/4
$\tau = 1/10$	43/13	48/13	52/13	53/13	58/13	60/13
$\tau = 1/100$	62/15	32/16	48/18	55/18	61/18	64/18
$\tau = 1/1000$	15/8	30/11	66/14	39/16	42/16	48/16

TABLE 6.2
Total number of cycles/number of Lanczos iterations as a function of τ and the size of the matrix.

The table shows that increasing the accuracy in the spatial discretization has no dramatic influence on the total number of Lanczos iterations. For example, for $\tau = 1$, in all tests we needed four steps. This shows that even in situations where convergence is heavily based on exploiting the eigenvalue spectrum, the convergence is grid independent. This is expected since the upper estimate of the spread/gap ratio does not depend on the norm of A. When τ is decreased we see again that more steps of the Lanczos method are required. The computational advantage of a relaxation strategy is modest for a large value of τ. The reason is that in this situation the upper bound is around one in the first iterations and then very quickly drops in the final two iterations. For smaller values of τ the gain is comparable to that for the example in the previous section. We warn the reader that the SAMG package detects diagonal dominance of a matrix and automatically switches to a different solution strategy if this is more appropriate. This means that one cannot straightforwardly compare the total number of cycles for different combinations of τ and n.

7. Summary and outlook. The aim of this paper is to generalize the concept of preconditioning for linear systems to the computation of the product of the matrix exponential and a vector. The idea explored is to apply the Lanczos process to a shifted and inverted matrix that better emphasizes the important eigenvalues. This results in an inner–outer iteration scheme where the solution of the shifted systems can be accomplished by preconditioned solvers. The worst case convergence behavior of the method can be bounded in terms of a rational approximation problem on the positive real axis. However, convergence speeds observed in practice are often much higher than these bounds suggest. This is due to the fact that the spectral transformation facilitates the exploitation of the discrete nature of the spectrum of the matrix. We have argued and demonstrated that the method shows convergence speeds independent of the norm of the matrix and, if an appropriate method is used for the solution of the linear systems, we can compute the sought-after vector in a time proportional to the number of spatial grid points. Furthermore, we have proposed an empirical strategy for choosing the tolerances for the errors in the solution of the shifted linear systems. This strategy can reduce the amount of work in the solution of the shifted systems with up to 40 percent compared with using a fixed tolerance on the error.

For clarity of this paper we have restricted our attention to the computation of the exponential function. However, the same approach seems to carry successfully over to the computation of the matrix functions related to (4.6) that play, beside the exponential function, an important role in exponentially based integrators, e.g., [14, 18, 20]. Also, we have assumed throughout this paper that the matrix is symmetric positive semi-definite. It is clear how our work can be extended to deal with so-called sectorial operators as well. Much of the theory and heuristics here are expected to hold in this case too. For example, it has been shown that there is a sequence of approximations from the classes $\mathcal{R}_m^{m-1}$ that converges geometrically to $\exp(-t)$ in a sector symmetrically around

the real axes, see [34].

In our future work we plan to investigate also other ideas for incorporating preconditioning in the computation of the matrix exponential. This is in particular important in applications where the matrix is skew-symmetric and, therefore, has purely imaginary eigenvalues. In this case it is not possible to exploit the rapid decay of the exponential functions and a different approach is required.

Acknowledgments. We like to thank Klaus Stüben and Tanja Clees of the GMD for providing us with the SAMG package and for their assistance in using the package. We appreciate the useful suggestions and comments of the referees that improved the presentation of this paper. We are thankful to Paolo Novati and Igor Moret for providing us with a preliminary version copy of [25]. We thank our colleague Jörg Niehoff for helpful discussions.

REFERENCES

[1] J.-E. Andersson. Approximation of e^{-x} by rational functions with concentrated negative poles. *J. Approx. Theory*, 32(2):85–95, 1981.

[2] P. B. Borwein. Rational approximations with real poles to e^{-x} and x^n. *J. Approx. Theory*, 38(3):279–283, 1983.

[3] A. Bouras and V. Frayssé. A relaxation strategy for inexact matrix-vector products for Krylov methods. Technical Report TR/PA/00/15, CERFACS, France, 2000.

[4] P. Castillo and Y. Saad. Preconditioning the matrix exponetial operator with applications. Technical Report UMSI 97/142, Supercomputer Institute, University of Minnesota, 1997.

[5] J. C. Cavendish, W. E. Culham, and R. S. Varga. A comparison of Crank-Nicolson and Chebyshev rational methods for numerically solving linear parabolic equations. *J. Computational Phys.*, 10:354–368, 1972.

[6] V. Druskin, A. Greenbaum, and L. Knizhnerman. Using nonorthogonal Lanczos vectors in the computation of matrix functions. *SIAM J. Sci. Comput.*, 19(1):38–54 (electronic), 1998. Special issue on iterative methods (Copper Mountain, CO, 1996).

[7] V L. Druskin and L. A. Knizhnerman. Two polynomial methods of calculating functions of symmetric matrices. *U.S.S.R Comput. Maths. Math. Phys.*, 29:112–121, 1989.

[8] V. L. Druskin and L. A. Knizhnerman. Krylov subspace approximations of eigenpairs and matrix functions in exact and computer arithmetic. *Numer. Lin. Alg. Appl.*, 2:205–217, 1995.

[9] 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(3):755–771 (electronic), 1998.

[10] E. Gallopoulos and Y. Saad. Efficient solution of parabolic equations by Krylov approximation methods. *SIAM J. Sci. Statist. Comput.*, 13(5):1236–1264, 1992.

[11] G. H. Golub, Z. Zhang, and H. Zha. Large sparse symmetric eigenvalue problems with homogeneous linear constraints: the Lanczos process with inner-outer iterations. *Linear Algebra Appl.*, 309(1-3):289–306, 2000.

[12] W. Hackbusch. Parabolic multigrid methods. In *Computing methods in applied sciences and engineering, VI (Versailles, 1983)*, pages 189–197. North-Holland, Amsterdam, 1984.

[13] M. Hochbruck and Ch. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34:1911–1925, 1997.

[14] M. Hochbruck, Ch. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Num. Anal.*, 19:1552–1574, 1998.

[15] G. Horton and S. Vandewalle. A space-time multigrid method for parabolic partial differential equations. *SIAM J. Sci. Comput.*, 16(4):848–864, 1995.

[16] A. Iserles. Rational approximations to $\exp(-x)$ with applications to certain stiff systems. *SIAM J. Numer. Anal.*, 18(1):1–12, 1981.

[17] S. Kaniel. Estimates for some computational techniques in linear algebra. *Math. Comp.*, 20:369–378, 1966.

[18] A.-K. Kassam and L. N. Trefethen. Fourth-order time stepping for stiff PDEs. *SIAM J. Sci. Comp.*, to appear.

[19] E. H. Kaufman, Jr. and G. D. Taylor. Best rational approximations with negative poles to e^{-x} on $[0, \infty)$. In *Padé and rational approximation (Proc. Internat. Sympos., Univ. South Florida, Tampa, Fla., 1976)*, pages 413–425. Academic Press, New York, 1977.

[20] S. Krogstad. Generalized integrating factor methods for stiff PDEs. *J. Comput. Phys.*, to appear.

[21] A. B. J. Kuijlaars. Which eigenvalues are found by the Lanczos method? *SIAM J. Matrix Anal. Appl.*, 22(1):306–321 (electronic), 2000.

[22] T. C.-Y. Lau. Rational exponential approximation with real poles. *BIT*, 17(2):191–199, 1977.

[23] C. Moler and Ch. Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Rev.*, 45(1):3–49 (electronic), 2003.

[24] C. Moler and Ch. F. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Review*, 20(4):801–836, 1978.

[25] I. Moret and P. Novati. RD rational approximations of the matrix exponential. *BIT*, to appear.

[26] S. P. Nørsett. Restricted Padé approximations to the exponential function. *SIAM J. Numer. Anal.*, 15(5):1008–1029, 1978.
[27] S. P. Nørsett and St. R. Trickett. Exponential fitting of restricted rational approximations to the exponential function. In *Rational approximation and interpolation (Tampa, Fla., 1983)*, volume 1105 of *Lecture Notes in Math.*, pages 466–476. Springer, Berlin, 1984.
[28] S. P. Nørsett and St. R. Trickett. Order-constrained uniform approximations to the exponential based on restricted rationals. *Constr. Approx.*, 2(2):189–195, 1986.
[29] P. Novati. Solving linear initial value problems by Faber polynomials. *Numer. Linear Algebra Appl.*, 10(3):247–270, 2003.
[30] C. C. Paige. Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem. *Linear Algebra Appl.*, 34:235–258, 1980.
[31] P. P. Petrushev and V. A. Popov. *Rational approximation of real functions*. Cambridge University Press, Cambridge, 1987.
[32] A. Ruhe. Rational Krylov sequence methods for eigenvalue computation. *Linear Alg. Appl.*, 58:391–405, 1984.
[33] Y. Saad. Analysis of some Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 29(1):209–228, 1992.
[34] E. B. Saff, A. Schönhage, and R. S. Varga. Geometric convergence to e^{-z} by rational functions with real poles. *Numer. Math.*, 25(3):307–322, 1975/76.
[35] R. B. Sidje. Expokit. A Software Package for Computing Matrix Exponentials. *ACM Trans. Math. Softw.*, 24(1):130–156, 1998.
[36] V. Simoncini and D. B. Szyld. Theory of inexact Krylov subspace methods and applications to scientific computing. *SIAM J. Sci. Comput.*, 25(2):454–477, 2003.
[37] K. Stüben and T. Clees. *USER's manual SAMG*. Fraunhofer Institute, Algorithms and Scientific Computing, Sankt Augustin, August 2003. Release 21c, available at `http://www.scai.fraunhofer.de/samg.htm`.
[38] H. Tal-Ezer. Spectral methods in time for parabolic problems. *SIAM J. Numer. Anal.*, 26(1):1–11, 1989.
[39] J. van den Eshof and G. L. G. Sleijpen. Inexact Krylov subspace methods for linear systems. Preprint 1224, Dep. Math., University Utrecht, Utrecht, the Netherlands, February 2002. To appear in SIMAX.
[40] J. van den Eshof, G. L. G. Sleijpen, and M. B. van Gijzen. Relaxation strategies for nested Krylov methods. Preprint 1268, Dep. Math., University Utrecht, Utrecht, the Netherlands, March 2003.
[41] A. van der Sluis and H. A. van der Vorst. The rate of convergence of conjugate gradients. *Numer. Math.*, 48(5):543–560, 1986.
[42] 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(2):249–263, 1987.
[43] J. Van Iseghem. Padé-type approximants of $\exp(-z)$ whose denominators are $(1 + z/n)^n$. *Numer. Math.*, 43(2):283–292, 1984.