Available online at www.sciencedirect.com
SCIENCE DIRECT®
Applied Numerical Mathematics 56 (2006) 253–267
APPLIED NUMERICAL MATHEMATICS
www.elsevier.com/locate/apnum

Padé and Gregory error estimates for the logarithm of block triangular matrices

João R. Cardoso [a,*,1], F. Silva Leite [b,2]

[a] *Instituto Superior de Engenharia de Coimbra, Rua Pedro Nunes, 3030-199 Coimbra, Portugal*
[b] *Departamento de Matemática, Universidade de Coimbra, 3000 Coimbra, Portugal*

Available online 13 May 2005

Abstract

In this paper we give bounds for the error arising in the approximation of the logarithm of a block triangular matrix T by Padé approximants of the function $f(x) = \log[(1+x)/(1-x)]$ and partial sums of Gregory's series. These bounds show that if the norm of all diagonal blocks of the Cayley-transform $B = (T-I)(T+I)^{-1}$ is sufficiently close to zero, then both approximation methods are accurate. This will contribute for reducing the number of successive square roots of T needed in the inverse scaling and squaring procedure for the matrix logarithm.

Keywords: Matrix logarithm; Inverse scaling and squaring; Padé approximants and Gregory's series

1. Introduction

Given a nonsingular matrix $A \in \mathbb{R}^{n\times n}$, any solution of the matrix equation $e^X = A$, where e^X denotes the exponential of the matrix X, is called a *logarithm* of A. In general, a nonsingular real matrix may have an infinite number of real and complex logarithms. However, if A has no eigenvalues on the closed negative real axis then there exists a unique real logarithm of A whose eigenvalues lie on the open

* Corresponding author.
E-mail addresses: jocar@isec.pt (J.R. Cardoso), fleite@mat.uc.pt (F. Silva Leite).
[1] Work supported by a PRODEP grant – Program n. 4/5.3/PRODEP/2000 and ISR.
[2] Work supported in part by ISR and research network contract ERB FMRXCT-970137.

0168-9274/$30.00 © 2005 IMACS. Published by Elsevier B.V. All rights reserved.
doi:10.1016/j.apnum.2005.04.003

strip $\{z \in \mathbb{C}\colon -\pi < \operatorname{Im} z < \pi\}$ of the complex plane [10]. This unique logarithm is called the *principal* logarithm of A and will be denoted by $\log A$.

A well-known first step in algorithms to compute $\log A$ is an initial reduction of A to real Schur form,

$$A = QTQ^T,$$

where $T \in \mathbb{R}^{n\times n}$ is block upper triangular (each diagonal block of T is either a 1×1 matrix or a 2×2 having complex conjugate eigenvalues) and $Q \in \mathbb{R}^{n\times n}$ is orthogonal, and then reduce the problem to that of computing $\log T$ by means of the equation $\log A = Q(\log T)Q^T$. The real Schur decomposition may be computed by stable algorithms (see, for instance, Algorithm 7.5.2 in [6], which requires about $25n^3$ flops) and for any orthogonally invariant norm (e.g., 2-norm or Frobenius norm), the absolute error affecting $\log T$ is the same for $\log A$.

This motivates the interest in the computation of the principal logarithm of a certain block triangular matrices. However, this paper deals with a more general situation, when T is any block upper triangular real matrix with square diagonal blocks and no eigenvalues on the closed negative real axis.

In this paper we consider two well-known methods for computing the principal logarithm of T: diagonal Padé approximants of the function $\log[(1+x)/(1-x)]$ (see [2]) and partial sums of Gregory's series

$$\log T = 2\sum_{k=0}^{\infty} \frac{B^{2k+1}}{2k+1}, \tag{1}$$

where $B = (T-I)(T+I)^{-1}$ (see [13,5]). Both methods become effective if combined with the standard inverse scaling and squaring technique [12] which consists, first, in taking a certain number, say k, of consecutive square roots of T so that $\log T^{1/2^k}$ is accurately approximated, and then recover the original logarithm using the identity

$$\log T = 2^k \log T^{1/2^k}.$$

Here, $X^{1/2}$ stands for the principal square root of T, i.e., the unique square root of T with eigenvalues lying on the open right half plane. We refer the reader to [1,7,10] for details about matrix square roots.

To decide if the approximation for $\log T^{1/2^k}$ given by Padé approximants or partial sums has the required accuracy, one needs sharp bounds for the absolute error.

The most widely used bound for the Padé error was derived by Kenney and Laub [11, Corollary 4]:

$$\big\|\log A - S_m(I-A)\big\| \leqslant \big|S_m\big(\|I-A\|\big) - \log\big(1-\|I-A\|\big)\big|, \tag{2}$$

where $\|I-A\| < 1$, $S_m(x)$ is the (m,m) Padé approximant of $\log(1-x)$ and $\|\cdot\|$ denotes a consistent matrix norm.

An alternative upper bound was proposed in [2, Section 2]:

$$\big\|\log A - R_m(C)\big\| \leqslant \left|\log\left(\frac{1+\|C\|}{1-\|C\|}\right) - R_m\big(\|C\|\big)\right|, \tag{3}$$

where $C = (A-I)(A+I)^{-1}$, $\|C\| < 1$ and $R_m(x)$ is the (m,m) Padé approximant of $\log[(1+x)/(1-x)]$. Bound (3) is sharper than (2) and the condition required for A is in general less restrictive, since $\|I-A\| < 1$ implies $\|C\| < 1$, but the converse may not be true. Another interesting feature of (3) is that it can be used even when $\log A$ is approximated by the other approximant $S_m(I-A)$ because $R_m(C) = S_m(I-A)$.

Assume now that we are given a block triangular matrix T with no eigenvalues on the closed negative real axis. The Cayley-transform $B = (T - I)(T + I)^{-1}$ has the same block triangular structure. Our goal is to give bounds for the Padé and Gregory absolute errors that exploit the particular block structure of T and do not require so restrictive conditions such as $\|I - T\| < 1$ and $\|B\| < 1$. These bounds will be presented in Section 2 (Theorem 2.5) for diagonal Padé approximants and in Section 5 (Theorem 5.1) for Gregory series. One important feature of these estimates is that they allow us to conclude that if there exists w sufficiently close to zero such that the diagonal blocks of B satisfy

$$\|B_{ii}\| \leqslant w < 1, \quad \forall i, \tag{4}$$

then diagonal Padé approximants of $\log[(1 + x)/(1 - x)]$ and partial sums of Gregory's series (1) give accurate approximates.

As far as we know, the idea of estimating the Padé error using the norm of diagonal blocks was firstly presented by Dieci and Papini [5]. Among other results, they showed that for block-(2×2) triangular matrices, *"(diagonal) Padé approximation for* $\log T$ *produces an approximation for the* $(1, 2)$ *block accurate in a relative-error sense, if the resulting approximations for diagonal blocks are accurate in an absolute-error sense"* [5, p. 928]. This statement is a consequence of [5, Theorem 4.6], which will be generalized in Section 4 for block-$(p \times p)$ triangular matrices. However, here, one uses the Cayley transform, since it leads to an improvement on the sharpness of the error estimates.

2. Bounding the Padé error

We say that a $n \times n$ square matrix X is a block-$(p \times p)$ matrix if it is partitioned into p^2 matrices (blocks) X_{ij}, $1 \leqslant i, j \leqslant p$, where the diagonal blocks X_{ii}, $1 \leqslant i \leqslant p$, are square matrices. For such a matrix we write $X = [X_{ij}]_{1 \leqslant i,j \leqslant p}$. Two block-$(p \times p)$ matrices $X = [X_{ij}]_{1 \leqslant i,j \leqslant p}$ and $Y = [Y_{ij}]_{1 \leqslant i,j \leqslant p}$ of the same size are said to be *partitioned conformably* if, for all $i, j = 1, \ldots, p$, the blocks X_{ij} and Y_{ij} have the same size.

If $\|\cdot\|$ denotes a consistent matrix norm, for a block-$(p \times p)$ matrix $X = [X_{ij}]_{1 \leqslant i,j \leqslant p} \in \mathbb{R}^{n \times n}$ we define a new $p \times p$ matrix $|X|$ whose (i, j) entry equals $\|X_{ij}\|$, that is,

$$|X| = \left[\|X_{ij}\|\right]_{1 \leqslant i,j \leqslant p}.$$

It is clear that $|X|$ depends on the block partitioning of X. So, in order to avoid any dubious situation, we always fix a particular block partitioning of X before considering the corresponding matrix $|X|$. The norm of the matrix $|X|$ may also depend on the block partitioning of X. Such is the case for the norms $\|\cdot\|_1$ and $\|\cdot\|_\infty$. However, for the Frobenius norm $\|A\|_F = \sqrt{\sum_{i,j=1}^{n} |a_{ij}|^2}$, one has

$$\left\||X|\right\|_F = \|X\|_F,$$

for any blocking of X. Due to this property, some results in this paper are stated for the Frobenius norm only. However, with the appropriate modifications some of them may be extended to other matrix norms.

Definition 2.1. If $X = [X_{ij}]_{1 \leqslant i,j \leqslant p}$ and $Y = [Y_{ij}]_{1 \leqslant i,j \leqslant p}$ are block-$(p \times p)$ matrices partitioned conformably, we say that $|X| \leqslant |Y|$ if $\|X_{ij}\| \leqslant \|Y_{ij}\|$, for all $i, j = 1, \ldots, p$.

Lemma 2.2. *For X and Y block-$(p \times p)$ matrices partitioned conformably, the following holds*:

(i) $|X + Y| \leqslant |X| + |Y|$;
(ii) $|XY| \leqslant |X||Y|$;
(iii) $|X^k| \leqslant |X|^k$, $k = 1, 2, \ldots$;
(iv) $|X| \leqslant |Y| \Longrightarrow \|X\|_F \leqslant \|Y\|_F$.

The proof of this lemma is a straightforward consequence of the properties of a matrix norm.

Lemma 2.3. *Let T be a block-$(p \times p)$ upper triangular matrix decomposed as $T = D + N$, where $D = \operatorname{diag}(T_{11}, \ldots, T_{pp})$ is block-$(p \times p)$ diagonal and N is block-$(p \times p)$ strictly upper triangular. Assume that $w = \max\{\|T_{11}\|_F, \ldots, \|T_{pp}\|_F\}$ and f is given by the power series $f(x) = \sum_{k=0}^{\infty} a_k x^k$, with convergence radius R and $a_k \geqslant 0$, $\forall k$. If $w < R$, then*

$$\left\| f(T) \right\|_F \leqslant \left\| \sum_{i=0}^{p-1} \frac{f^{(i)}(w)}{i!} |N|^i \right\|_F \leqslant \max_{0 \leqslant i \leqslant p-1} \frac{f^{(i)}(w)}{i!} \left\| \left(I - |N|\right)^{-1} \right\|_F.$$

Proof. Firstly, we observe that from Lemma 2.2 we can write

$$\left|T^k\right| = \left|(D+N)^k\right| \leqslant \sum_{i=0}^{\min(k,p-1)} \binom{k}{i} w^{k-i} |N|^i. \tag{5}$$

Since $f(T) = \sum_{k=0}^{\infty} a_k T^k$, where a_k is nonnegative, from Lemma 2.2 and (5), it follows that

$$\begin{aligned}
\left|f(T)\right| &\leqslant \sum_{k=0}^{\infty} a_k |T^k| = \sum_{k=0}^{\infty} a_k \left|(D+N)^k\right| \\
&\leqslant \sum_{k=0}^{\infty} a_k \sum_{i=0}^{\min(k,p-1)} \binom{k}{i} w^{k-i} |N|^i = \sum_{i=0}^{p-1} \left(\sum_{k=i}^{\infty} a_k \binom{k}{i} w^{k-i} \right) |N|^i \\
&= \sum_{i=0}^{p-1} \frac{f^{(i)}(w)}{i!} |N|^i.
\end{aligned}$$

Since $\||X|\|_F = \|X\|_F$, for any block partitioning of X, the following holds:

$$\left\| f(T) \right\|_F \leqslant \left\| \sum_{i=0}^{p-1} \frac{f^{(i)}(w)}{i!} |N|^i \right\|_F.$$

The second inequality follows from the fact that

$$\left(I - |N|\right)^{-1} = I + |N| + \cdots + |N|^{p-1},$$

because $|N|$ is nilpotent of order p. □

This lemma may be viewed as a possible extension of Theorem 11.2.2 in [6] to block triangular matrices and the inequality (5) as a block version of Lemma 2.1 in [3].

The principal logarithm of a matrix $A \in \mathbb{R}^{n\times n}$ with no eigenvalues on the closed negative real axis enjoys the following integral representation in terms of $C = (A - I)(A + I)^{-1}$:

$$\log A = \int_{-1}^{1} C(I - Cs)^{-1}\,\mathrm{d}s. \tag{6}$$

Next lemma shows that a Gauss–Legendre quadrature rule applied to this integral is equivalent to Padé approximation for $\log A$. Since this lemma is basically the Cayley-transform version of Theorem 4.3 in [4], we omit its proof.

Lemma 2.4. *Let A and C be as above and $F(s) := C(I - Cs)^{-1}$. If*

$$\mathcal{Q} := \sum_{k=1}^{m} a_k F(s_k) \quad (a_k \in \mathbb{R},\ s_k \in [-1, 1])$$

is the m-point Gauss–Legendre quadrature rule applied to (6) *and $R_m(x)$ is the (m, m) Padé approximant of* $\log[(1 + x)/(1 - x)]$, *then*

$$\mathcal{Q} = R_m(C).$$

The main result of this section is stated in the following theorem.

Theorem 2.5. *Let $T \in \mathbb{R}^{n\times n}$ be a block-$(p \times p)$ upper triangular matrix with no eigenvalues on the closed negative real axis and $B = (T - I)(T + I)^{-1}$. Assume that $B = D + N$, where $D = \operatorname{diag}(B_{11}, \ldots, B_{pp})$ is block diagonal and N is block strictly upper triangular. If $R_m(x)$ is the (m, m) Padé approximant of* $\log[(1 + x)/(1 - x)]$ *and $w = \max\{\|B_{11}\|_F, \ldots, \|B_{pp}\|_F\} < 1$, then*

$$\left\|\log T - R_m(B)\right\|_F \leqslant \mu_{p,m} \frac{\|B^{2m+1}\|_F}{(1 - w)^{2m+p}} \left\|\left(I - |N|\right)^{-1}\right\|_F, \tag{7}$$

where $\mu_{p,m} = c_m \frac{2^{2m+1}(2m+p-1)!}{(p-1)!}$, with $c_m = \frac{(m!)^4}{(2m+1)((2m)!)^3}$.

Proof. The standard error formula for m-point Gauss–Legendre quadrature rules (in scalar case) states that

$$\int_a^b h(t)\,\mathrm{d}t - \sum_{i=1}^{m} a_k h(t_k) = c_m (b - a)^{2m+1} h^{(2m)}(\xi),$$

where h is $2m$-times differentiable in $[a, b]$ and $\xi \in [a, b]$. Since $\log T = \int_{-1}^{1} F(s)\,\mathrm{d}s$, where $F(s) = B(I - Bs)^{-1}$, and $F^{(2m)}(s) = (2m)![B(I - Bs)^{-1}]^{2m+1}$, from [14, Theorem 3] and Lemma 2.4 it is not hard to conclude that

$$\begin{aligned}\left\|\log T - R_m(B)\right\|_F &\leqslant c_m 2^{2m+1} \max_{s\in[-1,1]} \left\|F^{(2m)}(s)\right\|_F \\ &\leqslant c_m 2^{2m+1} \left\|B^{2m+1}\right\|_F \max_{s\in[-1,1]} \left\|(2m)!\left[B(I - Bs)^{-1}\right]^{2m+1}\right\|_F. \end{aligned}\tag{8}$$

Now one needs to find a bound for $\max_{s\in[-1,1]} \|(2m)![B(I-Bs)^{-1}]^{2m+1}\|_F$. Let $f(x) := -\log(1-x)$. To define the corresponding matrix function $f(X) = -\log(I-X)$ we must assume that the spectral radius of X satisfies $\rho(X) < 1$. Since $\rho(Bs) < 1$, $s \in [-1,1]$, we can write $f^{(2m+1)}(Bs) = (2m)![B(I-Bs)^{-1}]^{2m+1}$. If we let $g(X) := f^{(2m+1)}(X)$, we see that the coefficients of the Maclaurin series of g are nonnegative. From Lemma 2.3 it follows that

$$\begin{aligned}
\max_{s\in[-1,1]} \left\| f^{(2m+1)}(Bs) \right\|_F &= \max_{s\in[-1,1]} \left\| g(Bs) \right\|_F \\
&\leqslant \max_{s\in[-1,1]} \left\| I + g(ws)|Ns| + \cdots + \frac{g^{(p-1)}(ws)}{(p-1)!}|Ns|^{p-1} \right\|_F \\
&\leqslant \max_{0\leqslant k\leqslant p-1} \frac{g^{(k)}(w)}{k!} \left\| \left(I - |N|\right)^{-1} \right\|_F \\
&= \max_{0\leqslant k\leqslant p-1} \frac{f^{(2m+1+k)}(w)}{k!} \left\| \left(I - |N|\right)^{-1} \right\|_F \\
&\leqslant \frac{f^{(2m+p)}(w)}{(p-1)!} \left\| \left(I - |N|\right)^{-1} \right\|_F \\
&= \frac{(2m+p-1)!}{(p-1)!(1-w)^{2m+p}} \left\| \left(I - |N|\right)^{-1} \right\|_F,
\end{aligned}$$

and therefore the result follows. □

Remark 2.6. Consider the particular case $w = \|B\| \leqslant 1$. Taking $p = 1$, the bound (7) reduces to

$$\left\| \log T - R_m(B) \right\|_F \leqslant (2m)! c_m \left(\frac{2w}{1-w} \right)^{2m+1},$$

giving rise to an alternative bound to (2) and (3).

One important advantage of using bound (7) is that the conditions $\|I - T\| < 1$ or $\|B\| < 1$ are not required. We only need to ensure that the norm of each diagonal block of B is sufficiently close to zero for Padé approximants to give accurate approximations for $\log T$. The drawback of (7) is its computational cost: one needs to compute powers of the block triangular matrix B and the inverse of the upper $p \times p$ triangular matrix $I - |N|$. To derive a less costly version of (7), one needs to bound $\|B^{2m+1}\|_F$. We shall not use the trivial bound

$$\left\| B^{2m+1} \right\|_F \leqslant \|B\|_F^{2m+1},$$

because we may have $\|B\| \gg 1$. Instead, we may use (5) to obtain

$$\left\| B^{2m+1} \right\|_F \leqslant \binom{2m+1}{p-1} w^{2m+2-p} \left\| \left(I - |N|\right)^{-1} \right\|_F, \tag{9}$$

where $m \geqslant p-1$. This inequality also shows that $\|B^{2m+1}\|_F \to 0$ as $w \to 0$. Using (9) in (7) it follows that

$$\left\| \log T - R_m(B) \right\|_F \leqslant \theta_{p,m} \frac{w^{2m+2-p}}{(1-w)^{2m+p}} \left\| \left(I - |N|\right)^{-1} \right\|_F^2, \tag{10}$$

where $\theta_{p,m} = c_m \frac{2^{2m+1}(2m+p-1)!(2m+1)!}{[(p-1)!]^2(2m-p+2)!}$ and the number of blocks is restricted to $p \leqslant m+1$.

In both bounds (7) and (10) we can see that the nonnormal part of B, hidden inside the factor $\|(I-|N|)^{-1}\|_F$, may have an important influence on the quality of the approximates given by Padé approximants. Large values for $\|(I-|N|)^{-1}\|_F$ need to be compensated by large values for the order of Padé approximants m or by small values for w. Since it is not practical to take m large (Higham [9] suggests taking $m \leqslant 16$), an absolute error of order of the machine precision *eps* may not be attained (see Remark 1).

We finish this section with some comments on the stability of the computation of powers of B. Since we are assuming that $\|B_{ii}\| \leqslant w < 1$, $\forall i = 1, \ldots, p$, we have that $B^k \to O$, whenever $k \to \infty$. To understand the way as the powers of B converge to the zero matrix, the following relationship gives some insight:

$$|B|^k \leqslant \begin{bmatrix} O(w^k) & O(w^{k-1}) & \ldots & O(w^{k-p+1}) \\ 0 & O(w^k) & \ldots & O(w^{k-p+2}) \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & O(w^k) \end{bmatrix}.$$

Although all the entries of $|B|^k$ converge to zero, we can see that diagonal entries converge faster than the remaining ones. This becomes more clear when $|N|$ (the nonnormal part of $|B|$) has entries with large absolute values. In this case, the entries of the nondiagonal blocks of B^k are more sensitive to the effects of rounding errors, because they may grow rapidly before decay. This phenomenon, usually called "hump", may originate heavy cancellation in finite precision arithmetic (see, for instance, [8, Chapter 17]).

3. Numerical examples

In this section we report on some numerical examples to illustrate the behaviour of the bounds (7) and (10). All the experiments were performed in Matlab (with relative machine epsilon $\varepsilon \approx 2.2 \times 10^{-16}$) on a Pentium IV. We have used Padé approximants of order $m = 7$ and the Frobenius norm. Symbols w, p and B are as in Theorem 2.5 and `errex`, `errest1`, `errest2` denote, respectively, the exact values for the error, the estimate given by the bound (7) and the estimate given by the bound (10).

Table 1 compares the estimates for the absolute error with the exact values for some matrices of the form

$$T_a = \exp(a) \begin{bmatrix} 1 & b & b^2/2 + c \\ 0 & 1 & b \\ 0 & 0 & 1 \end{bmatrix}$$

for which the exact logarithm is

$$\log T_a = \begin{bmatrix} a & b & c \\ 0 & a & b \\ 0 & 0 & a \end{bmatrix}.$$

We denote $B_a = (T_a - I)(T_a + I)^{-1}$ and consider fixed values for $b = 10^3$, $c = 10^{-3}$ and four distinct values for a, as displayed in the first column of the table.

Table 1
Comparison of the error estimates with the exact values for T_a

a	w	$\|B_a\|_F$	errex	errest1	errest2
0.05	0.03	6.8×10^{3}	1.5×10^{-11}	3.4×10^{-10}	3.5×10^{-10}
0.1	0.05	1.3×10^{4}	1.5×10^{-11}	4.3×10^{-6}	4.6×10^{-6}
0.3	0.15	3.7×10^{4}	9.8×10^{-8}	4.2×10^{1}	4.9×10^{1}
0.5	0.24	5.8×10^{4}	6.5×10^{-5}	1.9×10^{5}	2.5×10^{5}

One interesting fact reported on the table is that the estimates based on the bounds (7) and (10) show that, although $\|B_a\|_F \gg 1$, Padé approximants of order 7 give high accuracy for the logarithm of the matrices $T_{0.05}$ and $T_{0.1}$. However, for matrices $T_{0.3}$ and $T_{0.5}$ we have obtained very poor estimates. The reason is that w is not sufficiently close to zero or m is not large enough.

One important issue in the inverse scaling and squaring procedure is to find the smallest number of successive square roots of T, say k, we need to take to guarantee that the approximation of $\log T^{1/2^k}$ given by $R_m(B^{(k)})$, where $B^{(k)} = (T^{1/2^k} - I)(T^{1/2^k} + I)^{-1}$, has the required accuracy. Computing unnecessary square roots not only increases the number of operations involved, but may lead to a loss of accuracy in the computed result (see [4]). Since the bound (7) does not require any of the conditions $\|I - T\| < 1$ or $\|B\| < 1$, but instead that $\|B_{ii}\| \leqslant w_i < 1$, $\forall i$, a reduction on the number of square roots involved will occur. This is more evident when T has diagonal block entries near one and other blocks with entries having large absolute values. This is the case of matrices $T_{0.05}$ and $T_{0.1}$ considered above, for which the smallest number of square roots k needed to have $\|I - T_{0.05}^{1/2^k}\| < 1$ (respectively, $\|B_{0.05}^{(k)}\| < 1$) is $k = 11$ (respectively $k = 10$); for $T_{0.1}$ we need $k = 11$ (respectively $k = 10$) to have $\|I - T_{0.1}^{1/2^k}\| < 1$ (respectively, $\|B_{0.1}^{(k)}\| < 1$). However, no square root is necessary to guarantee that diagonal blocks of B satisfy $\|(B_{0.05})_{ii}\| < 1$ or $\|(B_{0.1})_{ii}\| < 1$, for all i.

Although the estimates (7) and (10) give quite similar results for matrices T_a, this is in fact a coincidence and not a general rule. To illustrate the difference between both estimates we have tested a randomized 15×15 matrix T with real and nonreal pairs of eigenvalues, for which $\|B\|_F = 0.9$. Assuming that the number of blocks is $p = 5$, we have got the following results: $w = 0.4$, $\texttt{errex} = 1.5\times10^{-13}$, $\texttt{errest1} = 1.3 \times 10^{-5}$ and $\texttt{errest2} = 2.5 \times 10^{3}$. We shall observe that this significative difference between errest1 and errest2 is due to the very conservative estimate given by the bound (9) to $\|B^{2m+1}\|_F$.

Assume now that we are given a matrix T such that $B = (T - I)(T + I)^{-1}$ allows more than one blocking, with the diagonal blocks satisfying the conditions of Theorem 2.5. It is easy to see that the estimates given by the bounds (7) and (10) depend on the number p of blocks considered in B. Although it seems to be hard to find a general rule for choosing an optimal p, the nature of the bounds shows that we shall not take p large. For instance, if we have a 100×100 upper triangular matrix, it is not reasonable to take $p = 100$. Based on our experience with some numerical examples and taking into account that in practice Padé approximants of orders $m \leqslant 9$ are the most widely used, we also suggest to take $p \leqslant 9$ (check the tests above).

To analyse the effects of the blocking on the behaviour of bounds (7) and (10), we have tested a randomized 20×20 upper triangular matrix T satisfying $\|B\| = \|(T - I)(T + I)^{-1}\| < 1$. We have considered seven different block partitioning for T: $p = 2, 3, 4, 5, 6, 7, 8$. The result is displayed on Fig. 1, where we can see that the best estimates occur when $p = 3$. We have also tested several randomized

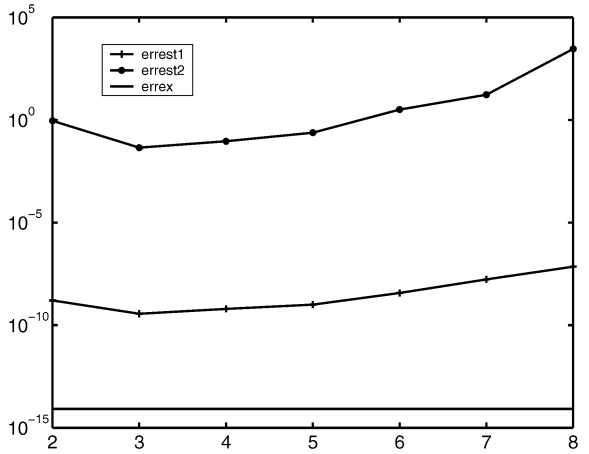

Fig. 1. Padé errors according to p.

100×100 upper triangular matrices with $\|B\| < 1$ and, curiously, the same observation holds, i.e., the polygonal lines corresponding to both bounds seem to have a minimum near $p = 3$.

4. Padé error bounds for diagonal and first superdiagonal blocks

Let

$$T = \begin{bmatrix} T_{11} & \cdots & T_{1p} \\ & \ddots & \vdots \\ 0 & & T_{pp} \end{bmatrix}$$

be a block-$(p \times p)$ triangular matrix with no eigenvalues on the closed negative real axis. The matrices $B = (T - I)(T + I)^{-1}$, $\log T$ and $R_m(B)$ have the same block triangular structure as T and we will use the following notation:

$$B = \begin{bmatrix} B_{11} & \cdots & B_{1p} \\ & \ddots & \vdots \\ 0 & & B_{pp} \end{bmatrix},$$

$$L = \log T = \begin{bmatrix} L_{11} & \cdots & L_{1p} \\ & \ddots & \vdots \\ 0 & & L_{pp} \end{bmatrix}, \qquad \bar{L} = R_m(B) = \begin{bmatrix} \bar{L}_{11} & \cdots & \bar{L}_{1p} \\ & \ddots & \vdots \\ 0 & & \bar{L}_{pp} \end{bmatrix},$$

where $B_{ii} = (T_{ii} - I)(T_{ii} + I)^{-1}$, $L_{ii} = \log T_{ii}$ and $\bar{L}_{ii} = R_m(B_{ii})$, for $i = 1, \ldots, p$.

The following result improves the error bounds given by Dieci and Papini (see [5, Theorem 4.6]) for the diagonal blocks $\bar{L}_{ii}$ and blocks $\bar{L}_{i,i+1}$ in the first superdiagonal of $\bar{L}$. These bounds reinforce an important fact already observed by those authors: we shall expect better accuracy in diagonal blocks (in general, full accuracy) than in superdiagonal ones. We note that Theorem 4.1 is stated for any consistent matrix norm.

Theorem 4.1. *With the above assumptions, the following holds*:

(i) *If* $\|B_{ii}\| \leqslant w_i < 1$, *for* $i = 1, \ldots, p$, *then*

$$\|L_{ii} - \bar{L}_{ii}\| \leqslant (2m)! c_m \left(\frac{2w_i}{1-w_i}\right)^{2m+1};$$

(ii) *If* $w = \max\{\|B_{11}\|, \ldots, \|B_{pp}\|\} < 0.9$, *then, for* $i = 1, \ldots, p-1$,

$$\|L_{i,i+1} - \bar{L}_{i,i+1}\| \leqslant (2m+1)! c_m \|B_{i,i+1}\| \left(\frac{2w}{1-w}\right)^{2m} \frac{2}{(1-w)^2}, \tag{11}$$

$$\frac{\|L_{i,i+1} - \bar{L}_{i,i+1}\|}{\|L_{i,i+1}\|} \leqslant c_m (2m+1)! \sec^2\left(\tanh^{-1} w\right) \left(\frac{2w}{1-w}\right)^{2m} \frac{1}{(1-w)^2},$$

where $c_m = \frac{(m!)^4}{(2m+1)((2m)!)^3}$.

Proof. (i) For the Frobenius norm, this is an immediate consequence of Theorem 2.5 (see also Remark 2.6). For any other consistent matrix norm, we can proceed as in the proof of part (ii) below.

(ii) Let $F(s) = B(I - Bs)^{-1} = [f_{ij}(s)]_{i,j=1,\ldots,n}$. From standard error formula for m-point Gauss–Legendre quadrature rules (in scalar case) and Lemma 2.4, we have

$$L - \bar{L} = c_m 2^{2m+1} \left[f_{ij}^{(2m)}(\xi_{ij})\right]_{i,j=1,\ldots,n}, \tag{12}$$

where $-1 \leqslant \xi_{ij} \leqslant 1$, $\forall i, j \in \{1, \ldots, n\}$. Partitioning $F(s)$ into $p \times p$ blocks conformably B, one may write

$$F(s) = \left[F_{ij}(s)\right]_{i,j=1,\ldots,p}.$$

Equating blocks $(i, i+1)$ in (12) and using (8), it follows that

$$\|L_{i,i+1} - \bar{L}_{i,i+1}\| \leqslant c_m 2^{2m+1} \max_{s\in[-1,1]} \left\|F_{i,i+1}^{(2m)}(s)\right\|, \quad i = 1, \ldots, p-1. \tag{13}$$

Since each block on the first superdiagonal of B^k, which will be denoted by $(B^k)_{i,i+1}$, can be written in the form

$$\left(B^k\right)_{i,i+1} = \sum_{j=0}^{k-1} B_{ii}^{k-1-j} B_{i,i+1} B_{i+1,i+1}^{j}, \tag{14}$$

$i = 1, \ldots, p-1$, one has

$$F_{i,i+1}^{(2m)}(s) = \sum_{k=0}^{\infty} \frac{(k+2m)!}{k!} \sum_{j=0}^{k+2m} B_{ii}^{k+2m-j} B_{i,i+1} B_{i+1,i+1}^{j} s^k.$$

Therefore

$$\begin{aligned}\max_{s\in[-1,1]} \left\|F_{i,i+1}^{(2m)}(s)\right\| &\leqslant \sum_{k=0}^{\infty} \frac{(k+2m)!}{k!} \sum_{j=0}^{k+2m} \|w\|^{k+2m-j} \|B_{i,i+1}\| w^j \\ &= \|B_{i,i+1}\| w^{2m} \sum_{k=0}^{\infty} \frac{(k+2m+1)!}{k!} w^k = \|B_{i,i+1}\| w^{2m} \frac{(2m+1)!}{(1-w)^{2m+2}},\end{aligned}$$

and from (13) the result follows. To obtain an upper bound for the relative error it suffices to bound $\|B_{i,i+1}\|$ in terms of $\|L_{i,i+1}\|$ and then use (11). Since $L = \log T = \log[(I+B)(I-B)^{-1}] = 2\tanh^{-1}(B)$, where $B = (T-I)(T+I)^{-1}$, we have that $B = \tanh(L/2)$ and hence

$$B_{i,i+1} = \big(\tanh(L/2)\big)_{i,i+1}.$$

The Taylor series for the hyperbolic tangent is given by

$$\tanh x = \sum_{k=1}^{\infty} a_{2k-1} x^{2k-1},$$

where $a_{2k-1} = \frac{2^{2k}(2^{2k}-1)}{(2k)!} b_{2k}$, with b_{2k} being the Bernoulli numbers, and $x^2 < \frac{\pi^2}{4}$. Thus, for L such that $\rho(L) < \pi$, we may write

$$\tanh\left(\frac{L}{2}\right) = \sum_{k=1}^{\infty} a_{2k-1}\left(\frac{L}{2}\right)^{2k-1}.$$

Since

$$\|L_{ii}\| \leqslant \log\left(\frac{1+w_i}{1-w_i}\right), \tag{15}$$

the assumption $\|B_{ii}\| \leqslant 0.9$, for all $i = 1, \ldots, p$, guarantees that L satisfies the spectral restriction $\rho(L) < \pi$. Therefore, by (14),

$$B_{i,i+1} = \big(\tanh(L/2)\big)_{i,i+1} = \sum_{k=1}^{\infty} a_{2k-1} \sum_{j=0}^{2k-2} \left(\frac{L_{ii}}{2}\right)^{2k-2-j} \left(\frac{L_{i,i+1}}{2}\right)\left(\frac{L_{i+1,i+1}}{2}\right)^{j},$$

and so

$$\|B_{i,i+1}\| \leqslant \sum_{k=1}^{\infty} |a_{2k-1}| \sum_{j=0}^{2k-2} \left\|\frac{L_{ii}}{2}\right\|^{2k-2-j} \left\|\frac{L_{i,i+1}}{2}\right\| \left\|\frac{L_{i+1,i+1}}{2}\right\|^{j}.$$

If $\ell := \max\{\|\frac{L_{11}}{2}\|, \ldots, \|\frac{L_{pp}}{2}\|\}$, then

$$\begin{aligned}\|B_{i,i+1}\| &\leqslant \left\|\frac{L_{i,i+1}}{2}\right\| \sum_{k=1}^{\infty} |a_{2k-1}| \sum_{j=0}^{2k-2} \ell^{2k-2} \\ &= \left\|\frac{L_{i,i+1}}{2}\right\| \sum_{k=1}^{\infty} |a_{2k-1}|(2k-1)\ell^{2k-2} = \left\|\frac{L_{i,i+1}}{2}\right\| \sec^2 \ell.\end{aligned}$$

Hence, from (15),

$$\ell \leqslant \frac{1}{2}\log\left(\frac{1+w}{1-w}\right) = \tanh^{-1} w,$$

and thus the result follows. □

Now we shall justify why the bounds in the previous theorem improve the ones given in [5] for block-(2×2) triangular matrices, which we recall in the following:

$$\|L_{ii}-\bar{L}_{ii}\| \leqslant (2m)!c_m\left(\frac{\alpha_i}{1-\alpha_i}\right)^{2m+1}, \tag{16}$$

with $\|I-T_{ii}\| \leqslant \alpha_i < 1,\ i=1,2$, and

$$\|L_{12}-\bar{L}_{12}\| \leqslant c_m\|T_{12}\|\left(\frac{\alpha}{1-\alpha}\right)^{2m}\frac{(2m+1)!}{(1-\alpha)^2}, \tag{17}$$

where $\alpha=\max\{\alpha_1,\alpha_2\}$. In both cases $c_m=\frac{(m!)^4}{(2m+1)((2m)!)^3}$.

To compare the bound in Theorem 4.1(i) with (16), it is enough to compare $\alpha_i/(1-\alpha_i)$ with $2w_i/(1-w_i)$, for $i=1,2$, and $\alpha_i<1$. Since $B_{ii}=(T_{ii}-I)(T_{ii}+I)^{-1}$ and $I+T_{ii}=2(I-\frac{I-T_{ii}}{2})$, we have

$$\|B_{ii}\| \leqslant \|T_{ii}-I\|\left\|(T_{ii}+I)^{-1}\right\| \leqslant \alpha_i\left\|\frac{1}{2}\left(I-\frac{I-T_{ii}}{2}\right)^{-1}\right\| \leqslant \frac{\alpha_i}{2(1-\frac{\|I-T_{ii}\|}{2})} \leqslant \frac{\alpha_i}{2-\alpha_i},$$

which implies that

$$\frac{2\|B_{ii}\|}{1-\|B_{ii}\|}=\frac{2w_i}{1-w_i} \leqslant \frac{2\frac{\alpha_i}{2-\alpha_i}}{1-\frac{\alpha_i}{2-\alpha_i}}=\frac{\alpha_i}{1-\alpha_i}. \tag{18}$$

This means that the bound in Theorem 4.1(i) is always smaller than or equal to (16).

For the bounds (ii) (Theorem 4.1) and (17), we need to compare

$$\|B_{12}\|\left(\frac{2w}{1-w}\right)^{2m}\frac{2}{(1-w)^2} \quad \text{with} \quad \|T_{12}\|\left(\frac{\alpha}{1-\alpha}\right)^{2m}\frac{1}{(1-\alpha)^2}.$$

Since $B_{12}=2(T_{11}+I)^{-1}T_{12}(T_{22}+I)^{-1}$ and $\|(T_{ii}+I)^{-1}\| \leqslant \frac{1}{2-\alpha_i}$, for $\alpha_i<1$ $(i=1,2)$, we have

$$\|B_{12}\|\frac{2}{(1-w)^2} \leqslant 2\left\|(T_{11}+I)^{-1}\right\|\|T_{12}\|\left\|(T_{22}+I)^{-1}\right\|\frac{2}{(1-w)^2} \leqslant \frac{2}{(2-\alpha)^2}\|T_{12}\|\frac{2}{(1-w)^2},$$

where $\alpha=\max\{\alpha_1,\alpha_2\}<1$. If we assume that $w\leqslant\alpha/2$, then

$$\frac{2}{(2-\alpha)^2}\|T_{12}\|\frac{2}{(1-w)^2} \leqslant \frac{2}{(2-\alpha)^2}\|T_{12}\|\frac{2}{(1-\alpha/2)^2}=\frac{16}{(2-\alpha)^4}\|T_{12}\|.$$

Since $\frac{16}{(2-\alpha)^4} \leqslant \frac{1}{(1-\alpha)^2},\ \forall\alpha\in[0,1[$, from (18) it follows that

$$\|B_{12}\|\left(\frac{2w}{1-w}\right)^{2m}\frac{2}{(1-w)^2} \leqslant \|T_{12}\|\left(\frac{\alpha}{1-\alpha}\right)^{2m}\frac{1}{(1-\alpha)^2}.$$

This means that the bound in Theorem 4.1(ii) is smaller than or equal to (17), provided that $w\leqslant\alpha/2$. We shall note that the condition $w\leqslant\alpha/2$ comes from the relationships $\|B_{ii}\| \leqslant \frac{\alpha_i}{2-\alpha_i}$ and $\frac{\alpha_i}{2-\alpha_i}\approx\frac{\alpha_i}{2}$, for all α_i sufficiently close to zero.

Remark 1. Under the assumptions of Theorem 4.1, let

$$\delta=(2m)!c_m\left(\frac{2w}{1-w}\right)^{2m+1}$$

be the maximum of the absolute errors affecting the diagonal blocks of $\bar{L}$. The following inequality relates the absolute error on diagonal blocks with the global error:

$$\left\|\log T - R_m(B)\right\|_F \leqslant \delta \frac{(2m+1)(2m+p-1)!}{[(p-1)!]^2(2m-p+2)![w(1-w)]^{p-1}} \left\|\left(I-|N|\right)^{-1}\right\|_F^2.$$

This shows that a small error δ in diagonal blocks may not avoid a large absolute error in the computed approximation $R_m(B)$. Thus, even when $\delta = eps$, it may be hard to have an approximation for $\log T$ with absolute error close to *eps*.

In the rest of this section, we briefly comment on some issues related with the Cayley transform B.

In this work we have been using $B = (T-I)(T+I)^{-1}$ and not simply $I-T$ because, as showed in our previous work [2], the Cayley transform seems to be an important tool to obtain sharper estimates for the error arising in the computation of $\log T$ via Padé or Gregory approximation. This means that using this transform (namely, the norm of its diagonal blocks B_{ii}) in the inverse scaling and squaring procedure requires, in general, fewer square roots than using the norm of $I - T_{ii}$. In some problems, this contributes to reduce the overscaling and consequently to avoid an eventual loss of precision. One disadvantage of using the Cayley transform is related with its computational cost. Since B is the solution of the linear matrix equation

$$(T+I)B = T - I,$$

it costs approximately the same as one block triangular matrix inversion.

5. Gregory error estimates

A similar analysis to the one made in the previous sections for Padé approximants may be extended to Gregory's series. In order to avoid repetition, we only mention the most important results, namely the Gregory versions of Theorems 2.5 and 4.1, and omit their proofs.

Theorem 5.1. *Under the assumptions of Theorem* 2.5, *if* $f(x) = \log[(1+x)/(1-x)]$ *then*

$$\left\|\log T - 2\sum_{k=0}^{q} \frac{B^{2k+1}}{2k+1}\right\|_F \leqslant \frac{\|B^{2q+3}\|_F}{(2q+3)!(p-1)!} \left\|\left(I-|N|\right)^{-1}\right\|_F f^{(2q+p+2)}(w), \tag{19}$$

for all q.

Remark 2. The factor $f^{(2q+p+2)}(w)$ involved in the bound (19) is given by the expression

$$f^{(2q+p+2)}(w) = (2q+p+1)!\left(\frac{(-1)^{2q+p+1}}{(1+w)^{2q+p+2}} + \frac{1}{(1-w)^{2q+p+2}}\right).$$

Since, for a sufficiently large q, $\frac{1}{(1+w)^{2q+p+2}} \ll \frac{1}{(1-w)^{2q+p+2}}$, we can estimate this factor using the relationship

$$f^{(2q+p+2)}(w) \approx \frac{(2q+p+1)!}{(1-w)^{2q+p+2}}.$$

Therefore, the inequality (19) may be rewritten in the form

$$\left\| \log T - 2\sum_{k=0}^{q} \frac{B^{2k+1}}{2k+1} \right\|_F \leqslant \frac{(2q+p+1)!\|B^{2q+3}\|_F}{(2q+3)!(p-1)!(1-w)^{2q+p+2}} \left\| \left(I - |N|\right)^{-1} \right\|_F.$$

Theorem 5.2. *Let T be a block-$(p \times p)$ upper triangular matrix with no eigenvalues on the closed negative real axis, $B = (T-I)(T+I)^{-1}$ and $\tilde{L} = 2\sum_{k=0}^{q} \frac{B^{2k+1}}{2k+1}$.*

(i) *If $\|B_{ii}\| \leqslant w_i < 1$, for $i = 1, \ldots, p$, then*

$$\|L_{ii} - \tilde{L}_{ii}\| \leqslant \frac{2}{2q+3} \left(\frac{w_i}{1-w_i} \right)^{2q+3};$$

(ii) *If $w := \max\{\|B_{11}\|, \ldots, \|B_{pp}\|\} \leqslant 0.9$, then, for $i = 1, \ldots, p-1$,*

$$\|L_{i,i+1} - \tilde{L}_{i,i+1}\| \leqslant 2\|B_{i,i+1}\| \frac{w^{2q+2}}{1-w^2},$$

$$\frac{\|L_{i,i+1} - \tilde{L}_{i,i+1}\|}{\|L_{i,i+1}\|} \leqslant \frac{w^{2q+2}}{1-w^2} \sec^2\left(\tanh^{-1}(w)\right).$$

6. Conclusion

In this work we have presented new estimates for the absolute error occurring whenever we approximate the logarithm of block triangular matrices using Padé approximants or partial sums of Gregory's series. These bounds exploit the block triangular structure of the given matrix and improve the existing estimates that treat the matrix as a whole. The error in the diagonal and superdiagonal blocks of the approximation was also addressed as well as some numerical issues concerning to the behaviour of the bounds.

References

[1] A. Björck, S. Hammarling, A Schur method for the square root of a matrix, Linear Algebra Appl. 52/53 (1983) 127–140.
[2] J.R. Cardoso, F. Silva Leite, Theoretical and numerical considerations about Padé approximants for the matrix logarithm, Linear Algebra Appl. 330 (2001) 31–42.
[3] P.A. Davies, N.J. Higham, A Schur–Parlett algorithm for computing matrix function, SIAM J. Matrix Anal. Appl. 25 (2) (2003) 464–485.
[4] L. Dieci, B. Morini, A. Papini, Computational techniques for real logarithms of matrices, SIAM J. Matrix Anal. Appl. 17 (1996) 570–593.
[5] L. Dieci, A. Papini, Conditioning and Padé approximation of the logarithm of a matrix, SIAM J. Matrix Anal. Appl. 21 (2000) 913–930.
[6] G. Golub, C. Van Loan, Matrix Computations, third ed., Johns Hopkins University Press, Baltimore, MD, 1996.
[7] N.J. Higham, Computing real square roots of a real matrix, Linear Algebra Appl. 88/89 (1987) 405–430.
[8] N.J. Higham, Accuracy and Stability of Numerical Algorithms, SIAM, Philadelphia, PA, 1996.
[9] N.J. Higham, Evaluating Padé approximants of the matrix logarithm, SIAM J. Matrix Anal. Appl. 22 (4) (2001) 1126–1135.

[10] R.A. Horn, C.R. Johnson, Topics in Matrix Analysis, paperback ed., Cambridge University Press, Cambridge, 1994.
[11] C. Kenney, A.J. Laub, Padé error estimates for the logarithm of a matrix, Internat. J. Control 50 (3) (1989) 707–730.
[12] C. Kenney, A.J. Laub, Condition estimates for matrix functions, SIAM J. Matrix Anal. Appl. 10 (1989) 191–209.
[13] G.J. Lastman, N.K. Sinha, Infinite series for logarithm of a matrix, applied to identification of linear continuous-time multivariate systems from discrete-time models, Electronics Lett. 27 (16) (1991) 1468–1470.
[14] R. Mathias, Approximation of matrix-valued functions, SIAM J. Matrix Anal. Appl. 14 (1993) 1061–1063.