CONDITIONING AND PADÉ APPROXIMATION OF THE LOGARITHM OF A MATRIX[*]

LUCA DIECI[†] AND ALESSANDRA PAPINI[‡]

Abstract. In this work we: (i) use theory of piecewise analytic functions to represent the Fréchet derivative of any primary matrix function, in particular of primary logarithms, (ii) propose an indicator to assess inherent difficulties to compute a logarithm, (iii) revisit Padé approximation techniques for the principal logarithm of block triangular matrices.

AMS subject classifications. 65F30, 65F35, 65F99, 15A24

Key words. Logarithm, conditioning, Fréchet derivative, Padé approximation

1. Introduction.

Given a matrix $A \in \mathbb{R}^{n\times n}$ or $\mathbb{C}^{n\times n}$, we call logarithm of A any matrix L such that $e^L = A$. It is well known that A has a logarithm L if and only if A is invertible, which we will henceforth assume. Amongst the (infinitely many) logarithms of A, some are *primary* matrix functions and some are not (see [6]). Primary functions are those which are true functions on $\Lambda(A)$; in this work, $\Lambda(A) = \{\lambda_i(A),\ i = 1,\dots,n\}$ indicates the spectrum of a matrix A. Thus, if L is a logarithm of A, $\mu_i \in \Lambda(L)$, and $\lambda_i \in \Lambda(A)$, then L is a primary function if there exists a (piecewise analytic) function log for which $\mu_i = \log(\lambda_i)$, $i = 1,\dots,n$; we will nonetheless always write $L = \log(A)$, even though "log" may fail to be a function on the spectrum of A.

In applications, it is often important to characterize under which conditions on real A we also have a real logarithm L. It is well known (e.g., see [6]) that

Given $A \in \mathbb{R}^{n\times n}$, a real logarithm L exists if and only if A has an even number of Jordan blocks of each size relative to every negative eigenvalue. If A has any negative eigenvalue, such L cannot be a primary matrix function.

Among the real primary logarithms of A, attention has been almost invariably restricted to the so-called *principal* logarithm; this is the one whose eigenvalues have imaginary parts in $(-\pi,\ \pi)$. Such principal logarithm enjoys a very useful integral representation:

$$\log(A) = \int_0^1 (A-I)((A-I)t+I)^{-1}dt\,. \tag{1.1}$$

Recently, there has been some interest in computation of logarithms of matrices, see [2], [8], [10] and references there. However, there are important questions still unanswered. In this work, we address the following three issues.

1. All studies so far have only characterized sensitivity of the principal logarithm. In §2, we characterize the sensitivity of any primary logarithm. We base our approach on the Fréchet derivative of piecewise analytic functions.
2. Arguably, a logarithm L has computational meaning only if one can reliably verify to what extent L satisfies $e^L = A$. Thus, in §3 we propose a criterion to assess the inherent difficulty to compute the log of a matrix, precisely by trying to quantify

[*] This work was supported in part under NSF Grant DMS-9625813 and Nato Grant CRG 950865.

[†] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 U.S.A. (dieci@math.gatech.edu)

[‡] Dep. Energetica S. Stecco, Univ. of Florence, via C. Lombroso 6/17, 50134 Florence, Italy (papini@de.unifi.it)

how reliably we can verify if $e^L = A$. This is similar in spirit to what happens with the square root of a matrix, $A^{1/2}$, see [1] and [4]. Indeed, we elucidate the relation between our indicator and the "square roots" indicator of [1] and [4] in the context of the *inverse scaling and squaring* procedure.

3. Recently, there has been some criticism of the inverse scaling and squaring technique coupled with Padé approximation (the so called *Briggs' method*) for computing principal logarithms of block triangular matrices; see [10]. In §4, we derive improved error bounds which show precisely the interplay between inverse scaling and squaring and accuracy of Padé approximants. As a result, we show that the appropriate implementation of Briggs' method is still a very viable technique for computing the logarithm of a block triangular matrix.

2. Fréchet derivative and conditioning.

Classical sensitivity considerations are statements about the variational problem. In the context of functions of a matrix, this means studying the Fréchet derivative of the function. Therefore, this type of sensitivity analysis is restricted to primary matrix functions. For the logarithm, this study has thus far been restricted to principal logarithms, see [8] and [2]. But –as made clear in [3]– there are many important applications where this restriction just cannot be made and it is desirable to extend these considerations to any primary logarithm. To this end, it is convenient to use the Cauchy integral form to represent a logarithm in the form put forward in [18] for piecewise analytic functions. For the sake of generality, at first, here below, f is not restricted to be the "log" function.

Let $G_k,\ k = 1, \dots, q$ be disjoint domains[1] of the complex plane, and let G be the union of these G_k; for example, the G_k may be disjoint disks. Let f be a single valued analytic function in the interior of the G_k's. Consider all matrices whose spectrum lies in G, and let A be a fixed one of them. Then, one can define the primary matrix function $f(A)$ as (see [18, sections I.2.2–2.6])

$$f(A) \ = \ \frac{1}{2\pi i} \sum_{k=1}^{q} \int_{\Gamma_k} f(z)(zI - A)^{-1} dz \,, \tag{2.1}$$

where $\Gamma_k,\ k = 1, \dots, q$ are rectifiable curves enclosing each and all eigenvalues $\lambda_j \in G_k$ only once, lying entirely in G_k, and where integration along each Γ_k is performed in the positive direction.

We can now use (2.1) to obtain a formula for the Fréchet derivative: (2.2) below. This formula may be known, but we did not find it anywhere. Attempts of a similar flavor for functions analytic on a domain enclosing all eigenvalues were made in [16], [15] and in [6, 6.6.15]. See also [13] for a different approach.

As usual, the Fréchet derivative of f at A in the direction of the matrix E is the linear map $f'(A)$ for which

$$\lim_{h \to 0} \left\| \tfrac{1}{h}(f(A + hE) - f(A)) - f'(A)E \right\| = 0$$

(h real). In this paper, $\|\cdot\|$ is the 2-norm; at times, we also use the Frobenius norm $\|\cdot\|_f$.

Lemma 2.1. *With previous notation, the Fréchet derivative of f at A is given by the map*

$$f'(A) : \ E \ \to \ \frac{1}{2\pi i} \sum_{k=1}^{q} \int_{\Gamma_k} f(z)(zI - A)^{-1} E (zI - A)^{-1} dz \,. \tag{2.2}$$

[1] Recall, a domain is open and simply connected.

Proof. Choose h sufficiently small, so that $\Lambda(A+hE) \subseteq \cup_k G_k$. Then, form the difference $f(A+hE)-f(A)$ and expand

$$(zI-(A+hE))^{-1}-(zI-A)^{-1} = (zI-A)^{-1}[(I+hE(zI-A)^{-1}+O(h^2))-I]\,. \qquad \square$$

Remark 2.2. If p is the number of distinct eigenvalues of A, in formulas (2.1) and (2.2) it is convenient to take $q=p$. That is, each Γ_k encloses precisely one of the distinct eigenvalues of A.

The norm of the Fréchet derivative is $\|f'(A)\| = \max_{\|Z\|=1} \|f'(A)Z\|$. This is a useful indicator of numerical conditioning of the evaluation of $f(A)$ (see [2, 4, 8, 11, 12]). In fact, let $X=f(A)$ and suppose $X \neq 0$, and let $X+\Delta X = f(A+E)$, where $\|E\|$ is small. Then, at first order we have

$$\frac{\|\Delta X\|}{\|X\|} \leq \|f'(A)\| \frac{\|A\|}{\|X\|}\frac{\|E\|}{\|A\|}\,. \tag{2.3}$$

Thus, $\|f'(A)\|$ acts as an absolute error magnification factor, whereas the quantity $\|f'(A)\| \frac{\|A\|}{\|X\|}$ acts as a relative error magnification factor, and it is called the condition number of f at A. Of course, this is a measure of conditioning based on worse type behavior and more refined estimates are possible if A has a particular structure and one restricts to matrices E for which $A+E$ has this same structure.

Unfortunately, to obtain a sharp estimate of $\|f'(A)\|$ from (2.2), is generally nontrivial. A typical line of action goes as follows. According to Remark 2.2, consider the case $q=p$ in (2.2). Let $V^{-1}AV = J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_p \end{bmatrix}$ be a grouping of the Jordan form of A, so that J_k comprises all Jordan blocks relative to the eigenvalue λ_k. Thus, we have

$$f'(A)E \;=\; V\frac{1}{2\pi i}\sum_{k=1}^{p}\int_{\Gamma_k} f(z)(zI-J)^{-1}(V^{-1}EV)(zI-J)^{-1}dzV^{-1}\,. \tag{2.4}$$

Next, since $(zI-J)^{-1}$ are easy to write down explicitly, repeated use of the Cauchy formula for analytic functions eventually gives a closed expression for the Fréchet derivative. But not a particularly useful one for computational purposes, since one typically gets stuck with the condition number of V. A great simplification occurs if A is normal, because then V is unitary and J is diagonal. This is the content of the following Lemma, whose proof is a simple application of the Cauchy formula on (2.4). A result like this Lemma 2.3 was known to be true for f analytic (see [8] and [12, (1.4)]).

LEMMA 2.3. *If A is a normal matrix, for the Fréchet derivative* (2.2) *of the primary matrix function $f(A)$ in* (2.1) *we have*

$$\|f'(A)\|_f \;=\; \max\left(\max_{1\leq j\leq p}|f'(\lambda_j)|\ ,\ \max_{1\leq j,k\leq p,\ j\neq k}\frac{|f(\lambda_j)-f(\lambda_k)|}{|\lambda_j-\lambda_k|}\right)\,. \qquad \square \tag{2.5}$$

Next, we restrict to the case of $f(A)$ being $\log(A)$, and specialize the study of the Fréchet derivative to this case. Of course, we must understand that $\log(z)$ makes sense in the complex plane slit along a ray going from 0 to infinity, and naturally such ray must not contain any eigenvalue λ_i of A. So, (2.1) is specialized to read as follows

$$\log(A) \;=\; \frac{1}{2\pi i}\sum_{k=1}^{p}\int_{\Gamma_k}\log(z)(zI-A)^{-1}dz\,, \tag{2.6}$$

where the Γ_k do not intersect one another, nor the ray along which we had slit the complex plane. It has to be stressed that $\log(z)$ is a single valued analytic function within each Γ_k, but it is not required to be so in a domain enclosing $\cup_k\Gamma_k$. To be precise, suppose the plane is slit along the ray $\arg z = \alpha$ and we fix $\alpha < \arg z < \alpha + 2\pi$ for all $z \in \cup_k\Gamma_k$, then

$$\log(z) = \log|z| + i(\arg z + 2\pi m), \;\; z \in \Gamma_k \; , \tag{2.7}$$

for some $m = 0, \pm 1, \pm 2, \ldots$, but the values of m are not required to be the same for different Γ_k. With this in mind, by construction, (2.6)–(2.7) characterize all primary logarithms. If A is real with no negative eigenvalues, then it has (in general, infinitely many) real primary $\log(A)$. To characterize these, (2.6) is too general. First, we slit along the negative real axis, so that $-\pi < \arg z < \pi$. Then, relative to a complex conjugate pair of eigenvalues of A, in (2.7) we must take $\log(\rho e^{\pm i\phi}) = \rho \pm i(\phi + 2\pi m)$. These choices characterize all real primary logarithms.

One can use Lemma 2.3 in the case of $\log(A)$, to classify primary logarithms of normal matrices in terms of which ones minimize $\|f'(A)\|_f$. For example, one easily gets:

(a) "if A and $\log(A)$ are complex, and we have chosen the same value of m in (2.7) in each Γ_k, then all logarithms give identical value of $\|f'(A)\|_f$";

(b) "if A is 2×2 (real or complex), then the logarithm giving the smallest value of $\|f'(A)\|_f$ is the one which minimizes the distance between the imaginary parts of its eigenvalues." Therefore, if A is real, normal, with real primary $\log(A)$, it is not necessarily true that the principal logarithm gives a smaller value of $\|f'(A)\|_f$ than a complex logarithm.

3. Inherent difficulties and scaling procedures.

Sensitivity assessment based on the Fréchet derivative is of course restricted to primary matrix functions. In essence, this type of sensitivity study tells us if we may (or may not) expect that the computed logarithm $\tilde{L} = L + F$ is close to the exact logarithm $L = \log(A)$. See (2.3) and also [2, 3, 8, 12]. For the sake of the present section, let us henceforth assume that F is of small norm. On the other hand, L in general has only an indirect meaning, as solution of the matrix equation $e^L = A$; therefore, if evaluation of the exponential of L is sensitive, calculation of L is intrinsically hard, since we cannot reliably verify to what extent we have solved $e^L = A$. For this reason, in this section we propose a simple indicator to assess the inherent difficulty of computing the logarithm, which attempts to measure how reliably we can verify if we have solved $e^L = A$. Our indicator is not restricted to primary logarithms, and is similar in spirit to the square root indicator of [1] and [4]; we take this similarity further by elucidating the interplay between our indicator and the square root indicator in the context of the popular inverse scaling and squaring procedure.

So, we let $e^{\tilde{L}} = \tilde{A} = A + E$, and ask when we can expect E to be of small norm. Or, in a somewhat backward way, is there a reasonable hope that we have verifiably found the logarithm of a matrix close to the original matrix A? To answer this question, we are thus led to consider

$$\frac{\|E\|}{\|A\|} = \frac{\|e^{L+F} - e^L\|}{\|A\|} \tag{3.1}$$

and the issue has shifted to how to obtain sharp bounds for (3.1). Several different bounds for (3.1) have been obtained by Kågström in [7] and Van Loan in [11]. We are interested in bounds which take the form

$$\frac{\|E\|}{\|A\|} \;\le\; g(\nu\,\|F\|)\,\beta(L)\,, \quad \text{where } \nu \text{ is a positive constant}\,, \tag{3.2}$$

and we require that

(i) g is monotone increasing and $g(x) = x + O(x^2)$ for $0 < x \ll 1$,

(ii) $\beta :\; L \in \mathbb{R}^{n\times n} \;\to \mathbb{R}^+$, $\beta(0) = 1$ and $\beta(L) \ge 1$.

Our idea is to regard β as an indicator of intrinsic difficulties in computing $\log(A)$. In practice, the success of this approach will depend on how tight is the bound in (3.2). Generally, g can be expressed as (see [7, (4.1)])

$$g(\nu\,\|F\|) \;=\; e^{\nu\|F\|} - 1\,, \tag{3.3}$$

and we have experimented with the choices (3.4), (3.5) and (3.8) below for β and ν in (3.2)–(3.3). From [6, Corollary 6.2.32], we get

$$\nu = 1\;,\;\text{ and }\;\beta(L) = \frac{e^{\|L\|}}{\|A\|}\,. \tag{3.4}$$

From [7, (4.9)], the following is obtained (a small refinement of [11, (3.2)])

$$\nu = 1\;,\;\beta(L) = \frac{e^{\mu(L)}}{\|A\|}\,, \tag{3.5}$$

where $\mu(L)$ is the logarithmic norm of L. Recall that, for a matrix B, the logarithmic norm is (see [17])

$$\mu(B) \;=\; \lim_{h\to 0^+} \frac{\|I + hB\| - 1}{h} \;=\; \lim_{h\to 0^+} \frac{\log(\left\|e^{hB}\right\|)}{h}\,, \tag{3.6}$$

and, in the 2–norm, $\mu(B)$ is the largest eigenvalue of $(B + B^*)/2$. For later purposes, it is useful to notice that the second of (3.6) gives

$$\mu(B) \;=\; \lim_{k\to\infty} 2^k\;\log(\left\|e^{B/2^k}\right\|)\,. \tag{3.7}$$

Finally, from [7, (4.15)] (a slight refinement of [11, (3.4)]) we also get

$$\nu = \sum_{k=0}^{n-1} \frac{\|N\|^k}{k!}\;,\;\;\beta(L) = \frac{e^{a(L)}\nu}{\|A\|}\,, \tag{3.8}$$

where $a(L)$ is the spectral abscissa of L (i.e., the largest real part of the eigenvalues of L), and N is the off diagonal part of a Schur form of L. That is, if $Q :\; Q^*LQ = D+N$ is a Schur form of L, then $D = \mathrm{diag}(\lambda_i)$, and N is the strictly upper triangular part. To be precise, in (3.8) ν is $\nu(L)$ and hence g also depends on L in this case; for simplicity we omit this dependence. Notice that β in (3.8) is trivial to obtain if computation of a primary logarithm is done after Schur reduction of A.

LEMMA 3.1. *For all choices of β in* (3.4), (3.5), *and* (3.8), *we always have $\beta \ge 1$ and $\beta(0) = 1$. For β given in* (3.5) *and* (3.8), *we also have $\beta(L + cI) = \beta(L)$ for any real number c, and if L is normal $\beta(L) = 1$.*

Proof. That $\beta(L) \ge 1$ in (3.4) is obvious. For (3.5), it follows from $e^{\mu(L)} \ge \left\|e^L\right\|$, and for (3.8) is similar (see [11]). When L is normal, a simple computation shows that $e^{\mu(L)} = e^{a(L)} = \|e^L\|$, and thus $\beta(L) = 1$ in (3.5), and (3.8). The statement about $L + cI$ is also an immediate verification. □

Remark 3.2. If the original matrix A is highly structured, and the algorithm used to compute L exploits this structure, it may be possible to provide refinement of the bound (3.2). An obvious case is if A is block–diagonal: $A = \operatorname{diag} A_{ii},\ i = 1, \ldots, p$, and one computes $L_{ii} = \log(A_{ii})$; then, in this case, it is more appropriate to look at the factors β relative to each diagonal block. A less trivial instance of this occurrence will be seen in §4.

Remark 3.3. Our numerical experiments showed that, in general, the bound (3.4) is not satisfactory, and (3.5) is often only a modest improvement over (3.4), whereas (3.8) gives the most satisfactory results. This is in agreement with the results of [11]. Nonetheless, the β values in (3.4) and (3.5) are so inexpensive to compute that we have included them in this study.

Next, we restrict to principal logarithms of real matrices. In such case, a popular procedure to compute the logarithm exploits the relation

$$\log A = 2^k \log(A^{1/2^k}) \ , \tag{3.9}$$

where the progressive square roots have been taken so that the arguments of their eigenvalues are in the strip $(-\pi/2^k, \pi/2^k)$. Since $A^{1/2^k}$ approaches the identity, for k sufficiently large it should be easy to compute $\log(A^{1/2^k})$ to full precision, and then obtain $\log A$. This approach, known as "inverse scaling and squaring," was first put forward in [8]. The crux of the technique is the computation of square roots of matrices. In [1] and [4], the authors propose an indicator to quantify the inherent difficulty of computation of a square root of A; this indicator is defined as

$$\alpha(A^{1/2}) \ = \ \frac{\|A^{1/2}\|^2}{\|A\|} \,, \tag{3.10}$$

and clearly $\alpha(A^{1/2}) \geq 1$. Also, $\alpha(A^{1/2}) = 1$ if A is normal and $A^{1/2}$ is a primary square root (see [4]). Next, we want to understand whether or not the inverse scaling and squaring procedure may have introduced additional difficulties (as detected by the factor α) with respect to the intrinsic difficulty of computation of the logarithm (as detected by the factor β).

Lemma 3.4. *With respect to the choices of* $\beta(L)$ *in* (3.4) *and* (3.5), *we have*

$$\beta(L) = (\beta(L/2))^2 \alpha(A^{1/2}) \,, \tag{3.11}$$

whereas with respect to (3.8) *we have*

$$\beta(L) = (\beta(L/2))^2 \alpha(A^{1/2}) \ \frac{\nu(L)}{\nu^2(L/2)} \,. \tag{3.12}$$

Proof. To prove (3.11) for β in (3.4) is simple:

$$\Big(\frac{e^{\|L/2\|}}{\|A^{1/2}\|}\Big)^2 = \frac{e^{\|L\|}}{\|A\|} \ \frac{\|A\|}{\|A^{1/2}\|^2} \ .$$

In a similar way, for β in (3.5), (3.11) is true since $\mu(L/2) = \mu(L)/2$. To show (3.12), we have

$$\big(\beta(L/2)\big)^2 \ = \ \frac{e^{a(L)}\nu(L)}{\|A\|} \frac{\|A\|}{\|A^{1/2}\|^2} \frac{[\nu(L/2)]^2}{\nu(L)} \,. \qquad □$$

COROLLARY 3.5. *Let k be a positive integer. For $\beta(L)$ in* (3.4) *and* (3.5)*, we have*

$$\beta(L) = (\beta(L/2^k))^{2^k} \prod_{j=1}^{k} \big(\alpha(A^{1/2^j})\big)^{2^{j-1}} . \tag{3.13}$$

For β in (3.8)*, instead, we have*

$$\beta(L) = (\beta(L/2^k))^{2^k} \prod_{j=1}^{k} \big(\alpha(A^{1/2^j})\big)^{2^{j-1}} \frac{\nu(L)}{(\nu(L/2^k))^{2^k}} . \tag{3.14}$$

□

COROLLARY 3.6. *For β given in* (3.8)*, in* (3.12) *we have*

$$\frac{[\nu(L/2)]^2}{\nu(L)} \geq 1 , \qquad (\beta(L/2))^2\alpha(A^{1/2}) \geq \beta(L) . \tag{3.15}$$

Proof. Some algebra gives

$$\nu^2(L/2) \;=\; \nu(L) + S,\; S := \sum_{k=n}^{2n-2} \frac{\|N\|^k}{2^k}\frac{1}{k!}[2^k - 2\sum_{i=0}^{k-n}\binom{k}{i}] .$$

Therefore, we have

$$\big(\beta(L/2)\big)^2 \;=\; \frac{\beta(L)}{\alpha(A^{1/2})}\big(1+\frac{S}{\nu(L)}\big) . \qquad \square$$

We are now ready to answer whether or not computation of $\log(A)$ through the inverse scaling and squaring procedure may have introduced additional difficulties with respect to the intrinsic difficulty of computation of $\log(A)$. The general situation is already clear after one square root, that is when we use $\log(A) = 2\log(A^{1/2})$. We have to compare $\beta(L)$, with the product $\beta(L/2)\ \alpha(A^{1/2})$. By putting together the results of Lemma 3.4 and Corollary 3.6, it is easy to obtain:
With respect to β in (3.4) *and* (3.5)*, taking square roots does not lead to a harder computational task, that is:*

$$\beta(L/2)\alpha(A^{1/2}) \leq \beta(L) . \tag{3.16}$$

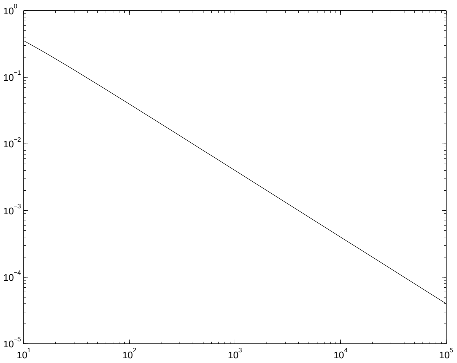
With respect to β in (3.8)*, we instead have*

$$\beta(L) = \beta(L/2)\alpha(A^{1/2})\ [\frac{e^{a(L/2)}}{\|A^{1/2}\|}\frac{\nu(L)}{\nu(L/2)}] . \tag{3.17}$$

The correct interpretation of (3.16) is more a statement about the inadequacy of the factors β in (3.4)–(3.5) to reveal computational difficulties, than a statement about real simplifications in computing $\log(A)$ which occurs when using inverse scaling and squaring. Instead, (3.17) lends itself to a more insightful and honest interpretation. In general, taking square roots may ease the computational task, or increase it, depending on whether the quantity in bracket in (3.17) is greater or less than 1. A complete characterization of matrices for which the quantity in bracket in (3.17) is greater than

1 seems out of reach, but we observe that for normal matrices it is 1; in fact, since both β and α equal 1, for normal matrices taking square roots does not change the inherent difficulties of computing $\log(A)$. Otherwise, the following examples show that computational difficulties, as measured by the ratio $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})}$ of (3.17), can be either increased or decreased by inverse scaling and squaring.

Example 3.7. Consider $A = \left[\begin{smallmatrix} e^c & be^c \\ 0 & e^c \end{smallmatrix}\right]$, with logarithm $L = \left[\begin{smallmatrix} c & b \\ 0 & c \end{smallmatrix}\right]$. For this example, taking square roots may make matters worse. In the figure on the right we show a "log-log" plot of the ratio $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})}$ for $b \in [10, 10^5]$.



Example 3.8. Consider now the matrix $A = \begin{bmatrix} 0.68 & 0.21 & 0.45 & 0.39 & 0.61 \\ 0.21 & 0.61 & 0.04 & 0.68 & 0.02 \\ 0.84 & 0.63 & 0.03 & 0.09 & 0.02 \\ 0.63 & 0.37 & 0.32 & 0.04 & 0.19 \\ 0.13 & 0.58 & 0.01 & 0.61 & 0.59 \end{bmatrix}$. In this case, we have $\frac{\beta(L)}{\beta(L/2)\alpha(A^{1/2})} \approx 6.24$, and hence the computational task is eased by taking square roots.

Clearly, no matter which factor β we are using, since $\beta(0) = 1$, and $\beta(L/2^k) \geq 1$, $\lim_{k\to\infty} \beta(L/2^k) = 1$. This tells us, if needed, that it is easy to compute the principal logarithm of a matrix close to the identity. However, it is of practical interest to study the rate at which $\beta(L/2^k) \to 1$. We need the following result on the logarithmic norm.

LEMMA 3.9. *Let $A \in \mathbb{R}^{n\times n}$ be invertible with no eigenvalues on the negative real axis, and let L be its principal logarithm. For each $k \in \mathbb{Z}^+$, let $A^{1/2^k} := e^{L/2^k}$ be the k–th root of A whose eigenvalues have arguments in $(-\pi/2^k,\ \pi/2^k)$, and let $\alpha(A^{1/2^k}) = \frac{\left\|A^{1/2^k}\right\|^2}{\left\|A^{1/2^{k-1}}\right\|}$. Then*

$$e^{\mu(L)} \;=\; \lim_{k\to\infty} \left\|A^{1/2^k}\right\|^{2^k} , \tag{3.18}$$

and

$$e^{\mu(L)} \;=\; \|A\| \lim_{k\to\infty} \prod_{j=1}^{k} \left(\alpha(A^{1/2^j})\right)^{2^{j-1}} . \tag{3.19}$$

Proof. Using (3.7) to characterize $\mu(L)$, (3.18) is obvious:

$$e^{\mu(L)} \;=\; e^{\lim_{k\to\infty} \log(\left\|e^{L/2^k}\right\|^{2^k})} \;=\; \lim_{k\to\infty} \left\|e^{L/2^k}\right\|^{2^k} .$$

Now, (3.19) follows from (3.18) and the identity

$$\frac{\left\|A^{1/2^k}\right\|^{2^k}}{\|A\|} = \alpha(A^{1/2})(\alpha(A^{1/4}))^2 \cdots (\alpha(A^{1/2^k}))^{2^{k-1}} . \qquad \square$$

LEMMA 3.10. *Under the same assumptions of* Lemma 3.9*, we have:*

(i) *for β in* (3.4)*:*

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; e^{\|L\|-\mu(L)}\,; \tag{3.20}$$

(ii) *for β in* (3.5)*:*

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; 1\,; \tag{3.21}$$

(iii) *for β in* (3.8)*:*

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} \;=\; e^{\|N\|+a(L)-\mu(L)}\,, \tag{3.22}$$

where N is the strictly upper triangular part in a Schur form of L. In particular, asymptotically,

$$\beta(L/2^k) \;\approx e^{(a(L)-\mu(L))/2^k}(1+\frac{\|N\|}{2^k})\,. \tag{3.23}$$

Proof. Let us begin showing (3.21). Upon using (3.19) and (3.13), we have

$$\beta(L) = \lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k}\, \beta(L)\,,$$

which is (3.21). (3.20) follows similarly from (3.19) and (3.13). To show (3.22), we pass to the limit as $k\to\infty$ in (3.14), and use (3.19), to obtain

$$\lim_{k\to\infty} \left(\beta(L/2^k)\right)^{2^k} = e^{a(L)}\; e^{-\mu(L)} \;\lim_{k\to\infty} \left(\nu(L/2^k)\right)^{2^k}\,.$$

Recalling that $\nu(L/2^k)=\sum_{j=0}^{n-1}\frac{\|N/2^k\|^j}{j!}$, we have

$$1+\frac{\|N\|}{2^k}\le \nu(L/2^k)\le e^{\|N\|/2^k}\,,$$

and therefore $\lim_{k\to\infty}\left(\nu(L/2^k)\right)^{2^k}=e^{\|N\|}$, and (3.22) and (3.23) follow. □

4. Block triangular matrices and Schur–Fréchet method. In this section, we revisit the familiar Briggs method (inverse scaling and squaring with Padé approximants) for computing the principal logarithm L of a (real) block triangular matrix R: $L=\log(R)$ (hereafter, the notation "log" always refers to the principal logarithm). We can think of the triangular matrix R as the result of Schur reduction of a matrix A. Thus, we have

$$R \;=\; \begin{bmatrix} R_{11} & R_{12} & \dots & R_{1p} \\ 0 & R_{22} & \dots & R_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & R_{pp} \end{bmatrix}\;,\quad L=\log(R)=\begin{bmatrix} L_{11} & L_{12} & \dots & L_{1p} \\ 0 & L_{22} & \dots & L_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & L_{pp} \end{bmatrix}\;. \tag{4.1}$$

Clearly, in (4.1), $L_{ii}=\log(R_{ii})$. The issue is how to determine the off diagonal blocks L_{ij}. Restrict to the case of $p=2$ in (4.1):

$$R=\left[\begin{smallmatrix} R_{11} & R_{12} \\ 0 & R_{22}\end{smallmatrix}\right]\;,\;\; L=\log(R)=\left[\begin{smallmatrix} L_{11} & L_{12} \\ 0 & L_{22}\end{smallmatrix}\right]\;. \tag{4.2}$$

Before discussing ways to obtain L_{12}, we must understand if computation of L_{12} presents some "inherent difficulties." We proceed along the lines of §3.

Let R and L be as in (4.2). In agreement with our previous Remark 3.2, we assume that the chosen algorithm of solution has respected this block structure, and thus may look at relative errors in a block sense. That is, with $F = \left[\begin{smallmatrix} F_{11} & F_{12} \\ 0 & F_{22} \end{smallmatrix}\right]$, where $L+F$ represents the computed logarithm, we look at

$$\frac{\left\| e^{L_{11}+F_{11}} - R_{11} \right\|}{\|R_{11}\|} \,, \quad \frac{\left\| e^{L_{22}+F_{22}} - R_{22} \right\|}{\|R_{22}\|} \,, \quad \frac{\left\| (e^{L+F})_{12} - R_{12} \right\|}{\|R_{12}\|} \,. \tag{4.3}$$

Then, there are three distinct inherent difficulties in computing L: one relative to L_{11}, one relative to L_{22} and one relative to L_{12}. The first two we treated in §3, here we deal with the third one. We will need the familiar formula (e.g., see [11]):

$$e^L \;=\; \left[\begin{smallmatrix} e^{L_{11}} & X \\ 0 & e^{L_{22}} \end{smallmatrix}\right] \,, \quad X = \int_0^1 e^{L_{11}(1-s)} L_{12} e^{L_{22}s} ds \,. \tag{4.4}$$

We let $(e^{L+F})_{12} = Y$, and with some algebra get

$$\begin{aligned} \frac{\|Y-X\|}{\|X\|} \;&\leq\; \frac{\|F_{12}\|}{\|X\|} \int_0^1 h(s) ds \;+\; \frac{\|L_{12}\| + \|F_{12}\|}{\|X\|} \\ \int_0^1 & h(s) \big[g(\nu_1 (1-s) \|F_{11}\|) g(\nu_2 s \|F_{22}\|) \beta((1-s)L_{11}) \beta(sL_{22}) \\ &+ g(\nu_1 (1-s) \|F_{11}\|) \beta((1-s)L_{11}) + g(\nu_2 s \|F_{22}\|) \beta(sL_{22}) \big] ds \;, \end{aligned} \tag{4.5}$$

where $h(s) := \left\| e^{(1-s)L_{11}} \right\| \; \left\| e^{sL_{22}} \right\|$ and g and β are the functions in (3.2), and (3.3). Recall that we assume that $g(x) = x + O(x^2)$ for $0 < x \ll 1$, and that the values of β and ν in which we are interested are given in (3.8). Now, we assume that $\|\nu_i F_{ii}\| \leq \eta$, $i = 1, 2$, with $\eta \ll 1$, and therefore from (4.5) we get

$$\frac{\|Y-X\|}{\|X\|} \leq \frac{\|F_{12}\|}{\|X\|} \int_0^1 h(s)(1 + \eta b(s)) ds + \eta \frac{\|L_{12}\|}{\|X\|} \int_0^1 h(s) b(s) ds + O(\eta^2) \;, \tag{4.6}$$

where $b(s) := \beta((1-s)L_{11}) + \beta(sL_{22})$. Since obviously $h(0) = \|e^{L_{11}}\|$ and $h(1) = \|e^{L_{22}}\|$, we arrive at the interesting conclusion that obtaining L_{12} may be an intrinsically hard computational task whenever $h(s)$, $0 < s < 1$, is much larger than the maximum of $\|e^{L_{11}}\|$ and $\|e^{L_{22}}\|$. This cannot happen if L_{ii} are normal (that is, if R_{ii} are). This claim can be verified as follows. Suppose that R_{ii}, $i = 1, 2$, are normal, and recall that for normal matrices $\beta(\cdot) = 1$. Then, with notation and results from Lemma 3.1, we have

$$\left\| e^{(1-s)L_{11}} \right\| = e^{(1-s)a(L_{11})} \;, \quad \left\| e^{sL_{22}} \right\| = e^{sa(L_{22})} \;,$$

where $a(L_{ii})$ indicates the spectral abscissa of the L_{ii}, $i = 1, 2$. Therefore, in this case, $h(s) = e^{(1-s)a(L_{11}) + sa(L_{22})}$, which is a monotone function, and the claim follows.

Example 4.1. Consider again Example 3.7. Obviously $R_{11} = R_{22}$ are normal, and there is no intrinsic difficulty in computing $L_{12} = b$ which is not already reflected in (and by) the computation of L_{11} and L_{22}.

Next, consider possible ways to approximate L_{12} in (4.2).

(a) *Parlett's method.* This well known general procedure rests on the identity $RL = LR$ (see [14]). From this, if $\Lambda(R_{11}) \cap \Lambda(R_{22}) = \emptyset$, one can uniquely find L_{12} by solving

$$R_{11}L_{12} \;-\; L_{12}R_{22} \;=\; L_{11}R_{12} \;-\; R_{12}L_{22}\,. \tag{4.7}$$

Remark 4.2. If the blocks R_{11} and R_{22} are sufficiently separated, so that the Sylvester equation (4.7) is well conditioned (see [5]), then it is hard for us to think of a better method to determine L_{12} than this one. The difficulty of this approach is to determine *a priori* if the Sylvester equation is well conditioned. In [2], we used a simple criterion based on the spectra of the diagonal blocks, and found it to be empirically reliable. However, if use of the identity (4.7) is not computationally feasible (say, the equation is singular), then some other way to obtain L_{12} must be found.

(b) *Padé approximation.* Here, one uses Padé rational functions to approximate $\log(R) = \log((R-I)+I)$. Therefore, L_{12} is obtained at once along with L_{11} and L_{22}. Good choices are diagonal Padé approximants.
Since Padé approximations are accurate only when R is "close to the identity," a standard practice is to exploit the relation (3.9) (with R there) and use Padé approximants for $\log(R^{1/2^k})$. The resulting approach is known as Briggs' method; it was first put forward in [8], and has enjoyed good success.
From a practical point of view, because of Remark 4.2, we think that Briggs' method is of greater appeal when solution of the system (4.7) is not advisable. Since our scope in this section is to revisit this Briggs' technique, we will assume that "*the matrix $R \in \mathbb{R}^{n\times n}$ in* (4.2) *does not have a blocking with R_{11} and R_{22} well separated.*" In particular, if R is in real Schur form, and we label its eigenvalues $\lambda_i,\ i = 1,\dots,k,\ k \le n$, writing only once a complex conjugate pair but repeating multiple eigenvalues, then for each fixed λ_i there exists a $\lambda_j,\ j \neq i$, such that $|\lambda_i - \lambda_j| \le \delta$.

(c) *Fréchet technique.* This is also a general procedure, which for the logarithm consists in exploiting the Fréchet identity

$$L \;=\; \begin{bmatrix} L_{11} & 0 \\ 0 & L_{22} \end{bmatrix} + \log'\!\left(\begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & R_{12} \\ 0 & 0 \end{bmatrix}\,, \tag{4.8}$$

where $\log'(B)Z$ is the Fréchet derivative of the logarithm at B in the direction Z. Therefore, one needs to get

$$L_{12} = \Big(\log'\!\left(\begin{bmatrix} R_{11} & 0 \\ 0 & R_{22} \end{bmatrix}\right)\begin{bmatrix} 0 & R_{12} \\ 0 & 0 \end{bmatrix}\Big)_{12}\,. \tag{4.9}$$

The *Fréchet method* proposed in [10] is based on approximating the Fréchet derivative of the logarithm in (4.9) by means of the hyperbolic tangent function.

Naturally, no matter what approach we use to obtain L_{12}, we must be at the same time solving (4.7) and satisfying (4.9).

THEOREM 4.3. *Let R and L be as in* (4.2). *Then we have*

$$\begin{aligned} L_{12} \;&=\; \int_0^1 ((R_{11}-I)t+I)^{-1}R_{12}((R_{22}-I)t+I)^{-1}\;dt \;= \\ &=:\; \int_0^1 L_1(t)R_{12}L_2(t)dt \;=:\; \int_0^1 F(t,R_{12})dt\,, \end{aligned} \tag{4.10}$$

where we have set $L_i(t) := ((R_{ii} - I)t + I)^{-1}$, $i = 1, 2$. L_{12} *in* (4.10) *is the only solution of* (4.7) *which eventually gives a principal logarithm* L. *Also, the expressions* (4.10) *and* (4.9) *are identical.*

Proof. To show (4.10), just use (1.1) and the block structure. To check that (4.10) solves (4.7), regardless of whether the spectra of R_{11} and R_{22} are disjoint, substitute L_{12} from (4.10) and use (1.1) for both L_{11} and L_{22}. That (4.10) and (4.9) give the same expression for L_{12} is verified using the analytic expression of the Fréchet derivative given in [2, (3.13)]. □

Theorem 4.3 suggests that to approximate the Fréchet derivative of the logarithm in (4.9) one might approximate the integral in (4.10) with a quadrature rule. For argument sake, suppose we do so, say with any of the standard polynomial rules. Then, it is simple to observe that to obtain good bounds for the quadrature error we generally need R_{11} and R_{22} close to the identity. However, R_{12} does not necessarily have to be close to 0 if we are interested in relative accuracy for L_{12}. This observation is at the basis of our revisitation of Briggs' method.

We must appreciate that in previous works on computation of the logarithm, in particular in [2, 9, 8], Padé approximations were used to approximate L in (4.2) by attempting to control the **absolute error**. That is, if $\hat{L}$ was the computed approximation, one tried to ensure that

$$\left\| L - \hat{L} \right\| \;\leq\; \eta\ , \tag{4.11}$$

where η is a small number, say the machine precision EPS. Since the original error estimates in [9], it has been clear that Padé approximations cannot obtain a small absolute error in (4.11), unless $\|I - R\|$ is sufficiently small. For this reason, people have been using (3.9) to work with $R^{1/2^k}$, where k was chosen so that $\|I - R^{1/2^k}\|$ was sufficiently small. E.g., in [2], we chose k so that $\|I - R^{1/2^k}\| \leq 0.35$ and then used the $(9, 9)$ diagonal Padé approximant, which guarantees –in absence of other errors– a value of η below EPS (see the first estimate in (4.14)). We also remarked that use of (3.9) may lead to undesired flattening (towards the identity) of the spectra of $R_{11}^{1/2^k}$ and $R_{22}^{1/2^k}$, and observed that (in some ill conditioned problems) this eventually produced loss of precision. In [10], the authors imputed this loss of precision on the possible loss of significance incurred when forming $I - R^{1/2^k}$. Indeed, in [10], the declared motivation for the Fréchet method and the **direct** approximation of the Fréchet derivative by the hyperbolic tangent, was the intent to avoid forming $I - R^{1/2^k}$. Kenney and Laub claimed that this technique allowed them to obtain better accuracy than with the standard Padé method. However, we believe that the real reason behind the reported success of the approach in [10] has nothing to do with avoiding the subtraction $I - R^{1/2^k}$, or the Padé method. To substantiate our claim, let us begin by looking at a simple Example.

Example 4.4. In Table 1 we report on some results relative to Example 3.7 (see also [10, Example 2]) for several values of c and constant b, $b = 10^6$. The results have been obtained exploiting the relation $\log(R) = 2^k \log(R^{1/2^k})$, and we used the $(9, 9)$ diagonal Padé approximation to obtain the logarithm of $R^{1/2^k}$. In Table 1, k is the number of square roots taken before using the Padé approximation for the logarithm, $\omega = \|I - R^{1/2^k}\|_\infty$, $\omega_i = \|I - R_{ii}^{1/2^k}\|_\infty$, $i = 1, 2$, "`abs`" refers to the matrix of absolute errors: $\mathtt{abs}_{ij} = \left| L_{ij} - \hat{L}_{ij} \right|$, and "`rel`" is the matrix of relative errors: $\mathtt{rel}_{ij} = \mathtt{abs}_{ij} / |L_{ij}|$. These numerical experiments were performed in `Matlab`, with machine

Table 1
Briggs-Padé method on Example 3.7: I.

c	k	ω	$\omega_1 = \omega_2$	`abs`	`rel`
0.1	0	1.105E6	0.105	$\begin{bmatrix} 8E{-}17 & 1E{-}10 \\ 0 & 8E{-}17 \end{bmatrix}$	$\begin{bmatrix} 8E{-}16 & 1E{-}16 \\ 0 & 8E{-}16 \end{bmatrix}$
0.1	22	0.22	1+2.4E-8	$\begin{bmatrix} 3E{-}10 & 6E{-}10 \\ 0 & 3E{-}10 \end{bmatrix}$	$\begin{bmatrix} 3E{-}9 & 6E{-}16 \\ 0 & 3E{-}9 \end{bmatrix}$
0.3	0	1.35E6	0.3498	$\begin{bmatrix} 5.5E{-}17 & 8E{-}10 \\ 0 & 5.5E{-}17 \end{bmatrix}$	$\begin{bmatrix} 2E{-}16 & 8E{-}16 \\ 0 & 2E{-}16 \end{bmatrix}$
0.9	2	3E5	0.25	$\begin{bmatrix} 2.2E{-}16 & 1.2E{-}10 \\ 0 & 2.2E{-}16 \end{bmatrix}$	$\begin{bmatrix} 2.5E{-}16 & 1.2E{-}16 \\ 0 & 2.5E{-}16 \end{bmatrix}$

precision EPS $\approx 2.2E - 16$ (exponential notation is used throughout). In Table 2, we instead report on the absolute and relative error matrices after exponentiating $\hat{L}$, "`ABS`" and "`REL`," where $\mathtt{ABS}_{ij} = \left|(R - e^{\hat{L}})_{ij}\right|$ and $\mathtt{REL}_{ij} = \mathtt{ABS}_{ij}/\,|R_{ij}|$ (see §3 and (4.3)), for the same values of c and k as in Table 1. The results highlight some interesting features: (i) taking as many square roots as would have been needed to obtain $\left\| I - R^{1/2^k} \right\| \leq 0.35$, as in the second row of Table 1, then we eventually lose seven digits along the diagonal of L, but the $(1, 2)$ entry has full accuracy (look at the relative errors in Table 1); (ii) taking k so to bring only the diagonal of R to within ω (rows 1, 3 and 4 of Table 1) gives full precision on the diagonal of L and maintain accuracy for the $(1, 2)$ entry. The loss of precision observed for $c = 0.1$ and $k = 22$ is easily explained as follows: write $(e^c)^{1/2^{22}} = 1 + x$. After 22 square roots the number e^c has the finite precision representation 1.000000023841858, so that only eight digits of x are retained. Now, since $\log(1 + x) = x - x^2/2 + x^3/3 - \dots$, in the double precision representation of the number $\log(1 + x)$ we should not expect anything better than eight digits accuracy (seemingly, we got nine). It is important to stress that the observed loss of digits is unavoidable given the finite precision representation of x, and no algorithm to approximate the logarithm can avoid it, subtracting 1 or not. Apparently, the only way to avoid it is to take fewer square roots, so that x has more (say, sixteen) significant digits. To explain why the $(1, 2)$ entry is always fully accurate, we need to wait until Theorem 4.6 below. Finally, the results of Table 2 are in striking agreement with the bound (4.6); in particular, for the second row of Table 2, with the notation from (4.6), $\eta \approx 10^{-10}$, $\frac{\|L_{12}\|}{\|X\|} \approx 1$, and $h(s) \approx 1$.

Table 2
Briggs–Padé method on Example 3.7: II.

c	k	`ABS`	`REL`
0.1	0	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$
0.1	22	$\begin{bmatrix} 3.3E{-}10 & 3.3E{-}4 \\ 0 & 3.3E{-}10 \end{bmatrix}$	$\begin{bmatrix} 3E{-}10 & 3E{-}10 \\ 0 & 3E{-}10 \end{bmatrix}$
0.3	0	$\begin{bmatrix} 0 & 9.3E{-}10 \\ 0 & 0 \end{bmatrix}$	$\begin{bmatrix} 0 & 6.9E{-}16 \\ 0 & 0 \end{bmatrix}$
0.9	2	$\begin{bmatrix} 4.4E{-}16 & 9.3E{-}10 \\ 0 & 4.4E{-}16 \end{bmatrix}$	$\begin{bmatrix} 1.8E{-}16 & 3.8E{-}16 \\ 0 & 1.8E{-}16 \end{bmatrix}$

Example 4.4 makes it evident that the approximation goal expressed by (4.11) (with $\eta \approx$ EPS) is not attainable, in general. Moreover, the legitimate suspicion is

that problems are caused by taking square roots. We quantify this next.

LEMMA 4.5. *Let $R^{1/2^k}$ and $\hat{R}^{1/2^k}$ be the exact and computed k-th square roots of R respectively, where k has been chosen so that $\|I - \hat{R}^{1/2^k}\| = \omega < 1$. Let $\hat{L}^{(k)}$ be the approximation obtained for $\log \hat{R}^{1/2^k}$, for example with the (N, N) diagonal Padé. Let $\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\| \leq \eta$ be the absolute error in approximating the logarithm, and $\frac{\|R^{1/2^k} - \hat{R}^{1/2^k}\|}{\|R^{1/2^k}\|} \leq \delta$ be the relative error in the computed k-th root. Then, we have*

$$\left\|\log R^{1/2^k} - \hat{L}^{(k)}\right\| \leq \eta + \frac{\delta}{1-\omega}\left\|R^{1/2^k}\right\| + O(\delta^2)\,, \tag{4.12}$$

and therefore with $\hat{L} = 2^k \hat{L}^{(k)}$ one has

$$\left\|\log R - \hat{L}\right\| \leq 2^k(\eta + \frac{\delta}{1-\omega}\left\|R^{1/2^k}\right\| + O(\delta^2))\,. \tag{4.13}$$

Proof. We have

$$\begin{aligned}\left\|\log R^{1/2^k} - \hat{L}^{(k)}\right\| &\leq \left\|\log R^{1/2^k} - \log \hat{R}^{1/2^k}\right\| + \left\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\right\| \\ &\leq \left\|\log'(\hat{R}^{1/2^k})\right\| \, \left\|R^{1/2^k} - \hat{R}^{1/2^k}\right\| + \left\|\log \hat{R}^{1/2^k} - \hat{L}^{(k)}\right\| \\ &\quad + O(\left\|R^{1/2^k} - \hat{R}^{1/2^k}\right\|^2)\,.\end{aligned}$$

Now we use [2, estimate after (3.2)] to get $\|\log'(\hat{R}^{1/2^k})\| \leq \frac{1}{1-\omega}$ and the proof is complete. □

Suppose we use a very accurate formula to approximate $\log \hat{R}^{1/2^k}$, so that (theoretically) $\eta \leq$ EPS. Then, the overall computational error is controlled by the relative accuracy in the k-th root, δ, and the magnification factor 2^k. In the best case, $\delta \approx$ EPS, an approximate equality which can be achieved in case R is normal, but even in this case these $O($EPS$)$ errors may be magnified by 2^k.

Another important feature exhibited by Example 4.4 is that L_{12} is accurate. This is a consequence of the following result.

THEOREM 4.6. *Let R be partitioned as in (4.2) with $R_{11} \in \mathbb{R}^{n_1 \times n_1}$, $R_{22} \in \mathbb{R}^{n_2 \times n_2}$, and $R_{12} \in \mathbb{R}^{n_1 \times n_2}$. Let $R^{(k)} := R^{1/2^k}$ and write $R^{(k)} = \begin{bmatrix} R_{11}^{(k)} & R_{12}^{(k)} \\ 0 & R_{22}^{(k)} \end{bmatrix}$, where k is the smallest integer such that $\left\|I - R_{11}^{(k)}\right\| \leq \omega_1 < 1$ and $\left\|I - R_{22}^{(k)}\right\| \leq \omega_2 < 1$ for preassigned values of ω_1 and ω_2. Let $\omega = \max(\omega_1,\ \omega_2)$, and let $P = \left[\begin{smallmatrix} P_{11} & P_{12} \\ 0 & P_{22} \end{smallmatrix}\right]$ be the approximation obtained with the (N, N) diagonal Padé approximant for $L^{(k)} = \log(R^{(k)})$. Then, we have the following error bounds*

$$\begin{aligned}&\left\|L_{ii}^{(k)} - P_{ii}\right\| \leq c(N)(2N)!(\frac{\omega_i}{1-\omega_i})^{2N+1}\ ,\ i = 1, 2\ , \\ &\left\|L_{12}^{(k)} - P_{12}\right\| \leq c(N)\left\|R_{12}^{(k)}\right\| \frac{(2N+1)!}{(1-\omega)^2}(\frac{\omega}{1-\omega})^{2N}\ ,\end{aligned} \tag{4.14}$$

where $c(N) = \frac{(N!)^4}{(2N+1)((2N)!)^3}$. Further, we have the following bounds for $\|R_{12}^{(k)}\|$

$$(1-\omega)\left\|L_{12}^{(k)}\right\| \leq \left\|R_{12}^{(k)}\right\| \leq \frac{1}{1-\omega}\ \left\|L_{12}^{(k)}\right\|\ , \tag{4.15}$$

which lead to a computable relative error bound in the second of (4.14).

Proof. The starting point of the proof is to recall that the (N, N) Padé approximant is the same as the N-point Gauss–Legendre quadrature rule for the integral in (1.1); therefore, with $M = I - R^{(k)}$, we have the following error estimate (see [2, Theorem 4.3 and Corollary 4.4])

$$L^{(k)} - P = c(N) \sum_{j=0}^{\infty} (2N+j) \dots (j+1) M^{2N+j+1} \eta_j^j \,, \tag{4.16}$$

where $0 \le \eta_j \le 1$. Now, partition $M = \left[\begin{smallmatrix} A & C \\ 0 & B \end{smallmatrix}\right]$, and notice that $M^p = \begin{bmatrix} A^p & \sum_{j=0}^{p-1} A^{p-j} C B^j \\ 0 & B^p \end{bmatrix}$. By assumption we have $\|A\| \le \omega_1 < 1$ and $\|B\| \le \omega_2 < 1$, so that (taking norms of sub-blocks on the right hand side of (4.16)) we immediately get

$$\begin{aligned} \left\| L_{ii}^{(k)} - P_{ii} \right\| &\le c(N) \textstyle\sum_{j=0}^{\infty} (2N+j) \dots (j+1) \omega_i^{2N+j+1} \ , \ i = 1, 2 \ , \\ \left\| L_{12}^{(k)} - P_{12} \right\| &\le c(N) \, \|C\| \textstyle\sum_{j=0}^{\infty} (2N+j) \dots (j+1) \textstyle\sum_{l=0}^{2N+j} \omega_1^{2N+j-l} \omega_2^l \\ &\le c(N) \, \|C\| \textstyle\sum_{j=0}^{\infty} (2N+j+1) \dots (j+1) \omega^{2N+j} \ . \end{aligned}$$

To complete the proof of (4.14), we observe that

$$\sum_{j=0}^{\infty} (2N+j) \dots (j+1) x^j = \frac{(2N)!}{(1-x)^{2N+1}} \quad ,$$

from which (4.14) follows. To obtain (4.15), we first use (4.10) to get

$$L_{12}^{(k)} = \int_0^1 ((R_{11}^{(k)} - I)t + I)^{-1} R_{12}^{(k)} ((R_{22}^{(k)} - I)t + I)^{-1} dt \,,$$

and we notice that if $\|R_{ii}^{(k)} - I\| \le \omega_i$ then $\|((R_{ii}^{(k)} - I)t + I)^{-1}\| \le \frac{1}{1 - \omega_i t}$, $i = 1, 2$. Therefore,

$$\left\| L_{12}^{(k)} \right\| \le \left\| R_{12}^{(k)} \right\| \int_0^1 \frac{dt}{(1 - \omega_1 t)(1 - \omega_2 t)} \,.$$

On the other hand, from (4.4), we also have

$$\left\| R_{12}^{(k)} \right\| \le \left\| L_{12}^{(k)} \right\| \int_0^1 \left\| e^{L_{11}^{(k)}(1-s)} \right\| \left\| e^{L_{22}^{(k)} s} \right\| ds \,.$$

Now, $\|e^{L_{ii}^{(k)} t}\| \le e^{\|L_{ii}^{(k)}\| t}$, and $\|L_{ii}^{(k)}\| \le \omega_i \int_0^1 \frac{dt}{1 - \omega_i t} = -\log(1 - \omega_i)$. Thus, $\|e^{L_{ii}^{(k)} t}\| \le \frac{1}{(1 - \omega_i)^t}$; from these, we get

$$\left\| R_{12}^{(k)} \right\| \le \left\| L_{12}^{(k)} \right\| \int_0^1 \frac{dt}{(1 - \omega_1)^{1-t} (1 - \omega_2)^t} \ .$$

Finally, to get (4.15), it is enough to observe that $(1 - \omega_1)^{1-t} (1 - \omega_2)^t \ge (1 - \omega)$, and also $\frac{1}{(1 - \omega_1 t)(1 - \omega_2 t)} \le \frac{1}{(1 - \omega t)^2}$, for $0 \le t \le 1$. □

Remark 4.7. From (4.14) and (4.15), if $\omega = \max(\omega_1, \omega_2) \le 0.30$, we should get a fully accurate approximation for $L_{12}^{(k)}$, in agreement with the results of Example 4.4.

The fundamental implication of Theorem 4.6 is that a (diagonal) Padé approximation for L produces an approximation for L_{12} accurate in a relative error sense, if the resulting approximations for L_{11} and L_{22} are accurate in an absolute error sense. To be precise, the estimates (4.15)–(4.14) predict a loss of precision: for $\omega = 0.3$, the relative eror bound for the approximation to L_{12} is about 50 times as large as the absolute error bound in the approximations of L_{11} and L_{22}. The situation is quite similar to what was obtained in [10]. However, we carry out the approximation of L_{12} *indirectly* by using Padé approximations for all of L at once, *rather than directly* as done in [10], without sacrificing accuracy. The need for the k-th square root in Theorem 4.6 is also present in the Schur–Fréchet algorithm of Kenney and Laub; this is what "Step 1" of the algorithm in [10, p. 651] accomplishes. However, it should be realized that in our Theorem 4.6 the value of k is not chosen so that $\|I - R^{1/2^k}\| < 1$, like in [10, Lemma 3.1 and later], but only so that $\|I - R_{ii}^{1/2^k}\| < 1$, $i = 1, 2$. Finally, as we had already remarked, there is no true issue associated to forming the subtraction $\|I - R^{1/2^k}\|$; regardless, the shift of emphasis from a global error bound as in (4.11) to such an estimate only for the diagonal blocks makes it quite likely that fewer square roots of R must be taken, and thus less likely that accuracy gets lost (see Lemma 4.5).

It is of course possible to extend the above considerations hinged on Theorem 4.6 to matrices as in (4.1). Although there are many subtle algorithmic issues which arise when we increase the number of blocks, Theorem 4.6 continues to hold for **all** possible choices of block-(2×2) submatrices. For example, suppose we have a block-(3×3) matrix: $\begin{bmatrix} R_{11} & R_{12} & R_{13} \\ 0 & R_{22} & R_{23} \\ 0 & 0 & R_{33} \end{bmatrix}$. Now, assume that $\|I - R_{ii}\| \le \omega_i < 1$, $i = 1, 2, 3$, and that $\|I - \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}\| \le \omega_{12} < 1$, or $\|I - \begin{bmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{bmatrix}\| \le \omega_{23} < 1$ (or both). Finally, take the (N, N) Padé approximant for $\log R$ (all of it). Theorem 4.6 applies and we can use the most favorable error estimates predicted by the theorem. We will think of the computed L_{12} and L_{23} as coming from the approximation to the logarithm of $\begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$ and $\begin{bmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{bmatrix}$, respectively; the approximation for L_{13}, instead, can be thought of as coming from two different block partitionings: (i) $R = \begin{bmatrix} R_{11} & \begin{bmatrix} R_{12} & R_{13} \end{bmatrix} \\ 0 & \begin{bmatrix} R_{22} & R_{23} \\ 0 & R_{33} \end{bmatrix} \end{bmatrix}$, or (ii) $R = \begin{bmatrix} \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix} & \begin{bmatrix} R_{13} \\ R_{23} \end{bmatrix} \\ 0 & R_{33} \end{bmatrix}$. We are free to pick whichever block partitioning is more convenient for us, in particular gives us better error estimates from (4.14).

5. Conclusions.

In this work we have:

(i) classified, and obtained conditioning information for, primary logarithms of matrices by using piecewise analytic functions theory;

(ii) proposed an indicator of the inherent difficulty to compute the logarithm of a matrix;

(iii) revisited Padé approximation techniques to compute principal logarithms of block triangular matrices.

From the practical point of view, the most important outcome of this work is that we have shown that the familiar Padé approximation technique with inverse scaling and squaring is still a viable way to compute the logarithm L of a triangular matrix R as in (4.2) with close eigenvalues; however, there are situations where one should not try to obtain an absolute error bound for the approximate L, but only for its diagonal blocks, while a relative error criterion is more appropriate for the off diagonal block. As a consequence, the number of square roots which we need to perform is not determined by the distance of the whole matrix from the identity, but rather by the distance from the identity of its diagonal blocks.

We believe that the above point (iii) has far reaching theoretical and practical implications, and we anticipate some work along similar lines also for the exponential of a matrix.

REFERENCES

[1] A. Bjorck and S. Hammarling. A Schur method for the square root of a matrix. *Linear Algebra Appl.*, 52-53:127–140, 1983.

[2] L. Dieci, B. Morini, and A. Papini. Computational techniques for real logarithms of matrices. *SIAM J. Matrix Anal. Appl.*, 17:570–593, 1996.

[3] L. Dieci, B. Morini, A. Papini, and A. Pasquali. On real logarithms of nearby matrices and structured matrix interpolation. *Applied Numerical Mathematics*, 1998. to appear.

[4] N. J. Higham. Computing real square roots of a real matrix. *Linear Algebra Appl.*, 88-89:405–430, 1987.

[5] N. J. Higham. Perturbation theory and backward error for $AX - XB = C$. *BIT*, 33:124–136, 1993.

[6] R.A. Horn and C.R. Johnson. *Topics in Matrix analysis.* Cambridge University Press, New York, 1991.

[7] B. Kagström. Bounds and perturbation bounds for the matrix exponential. *BIT*, 17:39–57, 1977.

[8] C. Kenney and A. J. Laub. Condition estimates for matrix functions. *SIAM J. Matrix Anal. Appl.*, 10:191–209, 1989.

[9] C. Kenney and A. J. Laub. Padé error estimates for the logarithm of a matrix. *Internat. J. Control*, 50(3):707–730, 1989.

[10] C. Kenney and A. J. Laub. A Schur-Fréchet algorithm for computing the logarithm and exponential of a matrix. *SIAM J. Matrix Anal. Appl.*, 19:640–663, 1998.

[11] C. Van Loan. The sensitivity of the matrix exponential. *SIAM J. Numer. Anal.*, 14:971–981, 1977.

[12] R. Mathias. Condition estimation for matrix functions via the Schur decomposition. *SIAM J. Matrix Anal. Appl.*, 16(2):565–578, 1995.

[13] R. Mathias. A chain rule for matrix functions and applications. *SIAM J. Matrix Anal. Appl.*, 17(3):610–620, 1996.

[14] B. N. Parlett. A recurrence among the elements of functions of triangular matrices. *Linear Algebra Appl.*, 14:117–121, 1976.

[15] D. L. Powers. The Fréchet differential of a primary matrix function. *Can. J. Math.*, 25:554–559, 1973.

[16] E. Stickel. On the Fréchet derivative of matrix functions. *Linear Algebra Appl.*, 91:83–88, 1987.

[17] T. Ström. On logarithmic norms. *SIAM J. Numer. Anal.*, 12(5):741–753, 1975.

[18] V.A. Yakubovich and V. M. Starzhinskii. *Linear Differential Equations with Periodic Coefficients*, volume 1&2. John-Wiley, New York, 1975.