PADÉ APPROXIMATION FOR THE EXPONENTIAL OF A BLOCK TRIANGULAR MATRIX.

LUCA DIECI AND ALESSANDRA PAPINI

ABSTRACT. In this work we obtain improved error bounds for Padé approximations to e^A when A is block triangular. As a result, improved scaling strategies ensue which avoid some common overscaling difficulties.

1. INTRODUCTION

In this work, we are interested in computation of the exponential of a matrix $A \in \mathbb{R}^{n\times n}$ (trivial modifications are needed in the complex case), e^A. Throughout, and unless otherwise stated, $\|\cdot\|$ will be the 2-norm.

Approximation of e^A is a most important and frequently encountered task: in 1978, the one hundred plus references of [14] gave a clear indication of this, and in the 20 years since the importance of computing the matrix exponential has only increased. For a case in point, there are several new approaches put forward for solving time dependent differential equations which require computing matrix exponentials; e.g., see [9] or [2]. Yet, with the possible exception of normal matrices and near the identity approximations (i.e., A close to 0), reliable numerical approximation of e^A remains somewhat elusive. We are particularly concerned with approximating e^A in case A is not normal and not close to 0. Especially for these cases, in our opinion, the ongoing effort on development and analysis of new techniques is fully warranted; e.g., see [12], but see also [8] for issues related to the challenging case of large, sparse, A.

Probably, the most popular and successful technique for approximating e^A consists of diagonal Padé approximations along with so-called scaling and squaring. The technique exploits the *scaling and squaring* identity

$$e^A = \left(e^{A/2^k}\right)^{2^k} \tag{1.1}$$

as follows. First, a sufficiently large k is chosen so that $A/2^k$ is close to 0, then a diagonal Padé approximant is used to calculate $e^{A/2^k}$, and finally the result is squared k times to obtain the required approximation to e^A. This basic approach is implemented in the `Matlab` function to evaluate e^A: `expm`. The fundamental issue is

1991 *Mathematics Subject Classification.* 65F30, 65F35, 65F99, 15A24.

Key words and phrases. Exponential, scaling and squaring, Padé approximation.

This work was supported in part under NSF Grant DMS-9973266 and CNR-GNIM.

how to choose k; on one hand, we need k large enough because Padé approximations are accurate only if $\left\|A/2^k\right\|$ is sufficiently small (see [5]), but on the other hand even a very accurate approximation to $e^{A/2^k}$, after repeated squaring, may become a disappointing approximation to e^A. This is clearly shown in the next example.

Example 1.1. The following is a standard test problem (e.g., see [3], [12]). We have

$$A = \begin{bmatrix} \omega & x \\ 0 & \omega \end{bmatrix}, \qquad e^A = e^\omega \begin{bmatrix} 1 & x \\ 0 & 1 \end{bmatrix}, \qquad \text{and we fix} \qquad x = 1.0\text{e+}6\ .$$

In Table 1, we report on the errors for the `Matlab` function `expm` which, supposedly, computes e^A to machine precision EPS $\approx 1.0\text{e-}16$ (scientific notation is used throughout). The reported error is the relative error in norm: $\left\|F - \widehat{F}\right\| / \|F\|$, where F and $\widehat{F}$ denote exact and computed exponentials, respectively, for both $e^{A/2^k}$ and e^A. For later reference, `expm` chooses k so to bring $\left\|A/2^k\right\|$ below $1/2$, and then uses the $(6,6)$ diagonal Padé approximation. In the particular example, this means $k = 21$.

ω	`Matlab` $A/2^k$–error	`Matlab` A–error
0.1	2.5e-016	2.1e-011
0.3	6.2e-015	2.0e-010
0.5	0.0e+000	4.8e-012

TABLE 1. `Matlab` errors.

The previous example served as motivation for our work. In a similar way to the error estimates we proved for Padé approximations to $\log(A)$ (see [3]), we will derive improved error estimates for Padé approximations to e^A in case A is a 2×2 block upper triangular matrix[1],

$$A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}, \tag{1.2}$$

where A_{11} and A_{22} are square matrices. Clearly, also e^A in this case has the same block structure, and the following formula for e^A is well known (see [13])

$$\begin{aligned} e^A &= \begin{bmatrix} e^{A_{11}} & F_{12} \\ 0 & e^{A_{22}} \end{bmatrix}, \qquad \text{where} \\ F_{12} &= \int_0^1 e^{(1-u)A_{11}} A_{12} e^{uA_{22}} du\ . \end{aligned} \tag{1.3}$$

[1] we do not require that the diagonal blocks be themselves triangular matrices

Loosely speaking, our new error estimates for Padé approximation of e^A say that if the diagonal blocks are well scaled, then one obtains accurate (in both absolute and relative error sense) approximations to the diagonal blocks of e^A, as well as accurate (in a relative error sense) approximation to F_{12}. As a consequence, e^A will have been accurately computed. Further, the error estimates imply that one does not need to scale (i.e., choose k in (1.1)) based upon $\|A\|$ but only based upon $\|A_{ii}\|$, $i = 1, 2$. This will avoid some common overscaling pitfalls.

In the next section, we give some general results on Padé approximation of e^A and also discuss errors arising during the squaring phase of the scaling and squaring algorithm. In section 3, we explicitly deal with the case of block triangular matrices and derive new error estimates in this case. We exemplify the computational impact of the new estimates on the model problem Example 1.1 and on a larger problem.

2. Padé error estimates.

In what follows, we let $F(A)$ be e^A and $R(A) = P(A)Q^{-1}(A)$ be its (s, s) diagonal Padé approximation. In this section, A is not restricted to be of the form (1.2).

It is well known (see [6], [5], or [14, Appendix 1]) that

$$Q(A) = \sum_{j=0}^{s} \frac{(2s-j)!s!}{(2s)!(s-j)!} \frac{(-A)^j}{j!} \tag{2.1}$$

and

$$F(A) - R(A) = Q^{-1}(A)G(A), \tag{2.2}$$

with

$$G(A) = \frac{(-1)^s}{(2s)!} A^{2s+1} \int_0^1 u^s (1-u)^s e^{A(1-u)} du. \tag{2.3}$$

Several error bounds for $F(A) - R(A)$ are available in case all eigenvalues of A are inside a suitable circle. For example, let $\omega = \|A\|_1$ and let s be large enough so that $R(z)$ is analytic on (and inside) the circle C of center the origin and radius $\gamma = k\omega$, with $1 < k < 2$. Then, Fair and Luke in [4] obtain the following bound (the quantity $c(s)$ is defined below in Theorem 2.5):

$$\left\|e^A - R(A)\right\|_1 \leq \frac{k}{k-1} \sup_{\lambda \in C} r(\lambda), \qquad \text{where}$$

$$r(z) = |e^z - R(z)| = \left| c(s) e^{z + z^2/[4(2s+1)]} z^{2s+1} [1 + O(s^{-3})] \right|.$$

This bound does not necessarily require smallness assumptions on ω, but may force s to be very large in order to satisfy the requirement of analyticity of $R(z)$, and a high value of s presents some computational drawbacks. On the other hand, for small values of ω, one can get more accurate error estimates by using the approach of Moler and Van Loan, [14]. We follow their approach to derive the estimates in

Theorem 2.5. Our error estimates differ from those in [14] on two accounts: (i) we first derive bounds on the error in terms of the error in the scalar case, and then further provide bounds which only involve $\|A\|$, (ii) we do not a priori restrict $\|A\|$ to be bounded by 1/2 (cfr. [14]).

Lemma 2.1. *Suppose that for all eigenvalues z of A one has $|Q(z)-1| < 1$ [2]. Then $Q(A)$ is invertible. Next, let $\alpha = Q(-\|A\|) - 1$. If $\alpha < 1$, we then have*

$$\|Q^{-1}(A)\| \leq \frac{1}{1-\alpha}\,. \tag{2.4}$$

Further, if $\|A\| < \log 4$, then

$$\|Q^{-1}(A)\| \leq \frac{1}{2 - e^{\|A\|/2}}. \tag{2.5}$$

Proof. First, assume that A is diagonalizable: $T^{-1}AT =: \Lambda = \mathrm{diag}(\lambda_i, i = 1, \ldots, n)$. Using (2.1), we get

$$T^{-1}(Q(A) - I)T = \sum_{j=1}^{s} \frac{(2s-j)!s!}{(2s)!(s-j)!}\frac{(-\Lambda)^j}{j!}.$$

So, by letting z to be an eigenvalues of A, the statement on invertibility of Q follows. Moreover, notice that

$$\frac{(p+q-j)!q!}{(p+q)!(q-j)!} \leq \left[\frac{q}{p+q}\right]^j \tag{2.6}$$

so that we get $Q(z) - 1 \leq e^{z/2} - 1$ which justifies the claim we made in the footnote. If A is not diagonalizable, it is ϵ-close to a diagonalizable matrix, and standard norm estimates give the result on invertibility of Q.

Next, observe that

$$\|Q(A) - I\| \leq \sum_{j=1}^{s} \frac{(2s-j)!s!}{(2s)!(s-j)!}\frac{\|A\|^j}{j!} = Q(-\|A\|) - 1\,.$$

Thus, if $\alpha = Q(-\|A\|) - 1 < 1$, then (2.4) follows. Now, if $\|A\| < \log 4$, from the expression of $Q(-\|A\|) - 1$ and (2.6) we also have that

$$\alpha \leq \sum_{j=1}^{s} \left[\frac{\|A\|}{2}\right]^j \frac{1}{j!} \leq e^{\|A\|/2} - 1 < 1,$$

so that (2.5) follows. □

Lemma 2.2. *Let $G(A)$ be defined by (2.3). Then*

$$\|G(A)\| \leq |Q(\|A\|)|\ |e^{\|A\|} - R(\|A\|)|\ , \tag{2.7}$$

[2] this condition is satisfied if $|z| < \log 4$

$$\frac{\|G(A)\|}{\|e^A\|} \leq |Q(\|A\|)|\ |e^{\|A\|} - R(\|A\|)|\ . \tag{2.8}$$

Moreover, the following estimates also hold

$$\|G(A)\| \leq \frac{(s!)^2}{(2s+1)((2s)!)^2} e^{\|A\|} \|A\|^{2s+1}, \tag{2.9}$$

$$\frac{\|G(A)\|}{\|e^A\|} \leq \frac{(s!)^2}{(2s+1)((2s)!)^2} e^{\|A\|} \|A\|^{2s+1}. \tag{2.10}$$

Proof. Taking norms in (2.3) and using

$$\int_0^1 u^p (1-u)^q du = \frac{p!q!}{(p+q+1)!}\ , \tag{2.11}$$

(2.9) is immediate. Also (2.10) can be obtained in a similar way, rewriting $G(A)$ as

$$G(A) = \frac{(-1)^s}{(2s)!} A^{2s+1} e^A \int_0^1 u^s (1-u)^s e^{-Au} du. \tag{2.12}$$

To obtain (2.7) is enough to observe that expanding G in (2.3) in powers of A all coefficients have same sign, and thus from (2.2) one obtains (2.7). To obtain (2.8), rewrite $G(A) = -e^A G(-A)$, and thus

$$\|G(A)\|/\|e^A\| \leq \|G(-A)\| \leq |G(\|A\|)|\ .$$

□

Remark 2.3. More accurate estimates can be obtained by using the logarithmic norm of A, $\mu(A)$, or the Schur form of A, $Q^*AQ = \Lambda + N$, where N is strictly upper triangular and Q is unitary. In fact, it is well known (e.g., see [13]) that (in the 2-norm)

$$\|e^{At}\| \leq e^{\mu(A)t} \quad \text{and} \quad \|e^{At}\| \leq e^{a(A)t} \sum_{k=0}^{n-1} \frac{\|Nt\|^k}{k!}\ ,$$

for $t \geq 0$, where $a(A)$ denotes the spectral abscissa of A. Then, the factor $e^{\|A\|}$ can be replaced in (2.9) by $e^{\max\{\mu(A),0\}}$ or $e^{\max\{a(A),0\}} \sum_{k=0}^{n-1} \frac{\|N\|^k}{k!}$, and in (2.10) by $e^{\max\{\mu(-A),0\}}$ or $e^{\max\{a(-A),0\}} \sum_{k=0}^{n-1} \frac{\|N\|^k}{k!}$. Obviously, similar changes apply to later estimates as well.

Remark 2.4. Notice that our bounds for $\|G(A)\|$, and hence those in Theorem 2.5 below, are identical in an absolute and relative sense with respect to $\|e^A\|$. This is because we do not know if $\|e^A\|$ happens to be < 1 or > 1. If we could use this information, then of course the estimates could be trivially refined (and would be different) for the absolute and relative error cases; for example, if we knew that $\|e^A\| > 1$ (e.g., as when $a(A) > 0$, since $\|e^A\| \geq e^{a(A)}$), then we could divide the right hand sides of the relative error bounds (2.8) and (2.10) by $\|e^A\|$ (simply using

the absolute error estimates and dividing them by $\|e^A\|$). Moreover, for later use, notice that by expanding in series under the integrals in (2.3) and (2.12) and using (2.11) one gets

$$G(A) = \frac{(-1)^s}{(2s)!}A^{2s+1}\sum_{k=0}^{\infty}\beta_k\frac{A^k}{k!} = e^A\,\frac{(-1)^s}{(2s)!}A^{2s+1}\sum_{k=0}^{\infty}(-1)^k\beta_k\frac{A^k}{k!} \tag{2.13}$$

with $\beta_k = s!(s+k)!/(2s+k+1)!$.

We are now ready to state the following result on Padé error estimates.

Theorem 2.5. *Let $R(A)$ be the (s,s) diagonal Padé approximation to e^A. Let $\alpha = Q(-\|A\|) - 1$, and assume that $\alpha < 1$. Then*

$$\|e^A - R(A)\| \le \frac{|Q(\|A\|)|}{1-\alpha}\;|e^{\|A\|} - R(\|A\|)| \tag{2.14}$$

$$\frac{\|e^A - R(A)\|}{\|e^A\|} \le \frac{|Q(\|A\|)|}{1-\alpha}\;|e^{\|A\|} - R(\|A\|)|\,. \tag{2.15}$$

Further, if $\|A\| = \omega < \log 4$, then

$$\|e^A - R(A)\| \le c(s)\frac{e^\omega}{2-e^{\omega/2}}\omega^{2s+1} \tag{2.16}$$

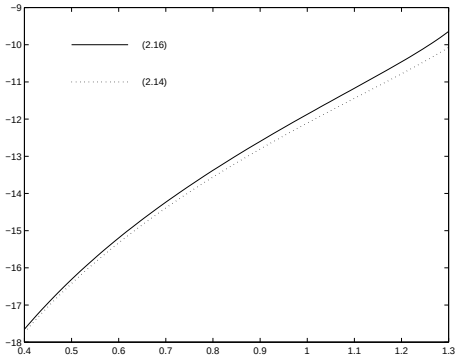
$$\frac{\|e^A - R(A)\|}{\|e^A\|} \le c(s)\frac{e^\omega}{2-e^{\omega/2}}\omega^{2s+1}\,, \tag{2.17}$$

with $c(s) = \frac{(s!)^2}{(2s+1)((2s)!)^2}$.

Proof. The statement is a consequence of (2.2) and the previous Lemmas. □

Remark 2.6. Observe that $|Q(\|A\|)| \le Q(-\|A\|) = 1+\alpha$, and thus $\frac{|Q(\|A\|)|}{1-\alpha}$ can be further bounded by $\frac{1+\alpha}{1-\alpha}$ which is a bound on the condition number of $Q(A)$. In fact, (2.14-2.15) make clear that there are two contributions to the error: one is the conditioning of the denominator, the other is the error one has in the scalar case.

Remark 2.7. We stress that the scalar estimates (2.14-2.15) are perfectly computable for given values of s and $\|A\|$, and are superior to the estimates (2.16-2.17). For example, for $s = 6$, computing the right hand side of (2.14) in extended precision and comparing it to the right hand side of (2.16), produces the figure on the right (in semi-logarithmic scale).



However, from the practical point of view, all estimates require $\|A\|$ to be sufficiently small, and differences between the right hand sides become minor for small values of $\|A\|$. Because of this, and the fact that the right hand side in the estimates (2.16-2.17) is more easily computable, we will henceforth only refer to the estimates (2.16-2.17). Nonetheless, it should be appreciated that later results such as Lemma 3.1 and Theorem 3.2 can be easily rephrased by using the scalar estimates presently derived.

As already remarked, to take advantage of the fact that the right hand sides in (2.16) and (2.17) are small for ω small, one may exploit (1.1). Let k be given, let $A_k := \frac{A}{2^k}$ and let $R(A_k)$ be the diagonal Padé approximation to e^{A_k}. Then, by repeatedly squaring $R(A_k)$, one approximates $F(A)$ with $\widehat{F}(A) := [R(A_k)]^{2^k}$. But, if an error of size η, say, is made by approximating e^{A_k} with $R(A_k)$, what error bound can we obtain for $\widehat{F}(A)$ as approximation of $F(A)$? This question was addressed in [14] by Moler and Van Loan. Here below we obtain a different error bound (cfr. (2.19) with [14, (pp. 809-810)]).

Let $\Delta_k := R(A_k) - e^{A_k}$ and $\Delta := \Delta_k e^{-A_k}$. Observe that all the matrices we are considering are functions of A and therefore commute with A and with each other, and so we have $\Delta = -Q^{-1}(A_k)e^{-A_k}G(A_k) = Q^{-1}(A_k)G(-A_k)$. Thus, following the same arguments used in Lemma 2.2 and in Theorem 2.5, for $\|A_k\|$ sufficiently small, we can assume to have both

$$\|\Delta_k\| \le \eta \quad \text{and} \quad \|\Delta\| \le \eta\,. \tag{2.18}$$

Now, we rewrite

$$\begin{aligned} F(A) - \widehat{F}(A) = [I - \widehat{F}(A)F^{-1}(A)]F(A) &= [I - (R(A_k)e^{-A_k})^{2^k}]F(A) \\ &= [I - (I+\Delta)^{2^k}]F(A)\,, \end{aligned}$$

and observe that $(I+\Delta)^{2^k}$ has an expansion in powers of Δ with positive coefficients. Then, we immediately get

$$\|I - (I+\Delta)^{2^k}\| \le (1+\eta)^{2^k} - 1\,,$$

and therefore

$$\frac{\|F(A) - \widehat{F}(A)\|}{\|F(A)\|} \;\le\; (1+\eta)^{2^k} - 1\,. \tag{2.19}$$

To gain some insight into the order of magnitude of the right-hand-side of (2.19), if $2^k\eta < 1$, one may use: $(1+\eta)^{2^k} - 1 = 2^k\eta e^{2^k\eta} + O(2^{k-1}\eta^2)$.

Remark 2.8. To interpret what we obtained, observe that (2.19), along with (2.17), show the potential benefits of the scaling and squaring technique. For example, for $s = 6$, if $\|A\| = 1$ and $k = 0$ the relative error bound is 1.3e-12, while if $\|A\| = 1$ and $k = 2$ the relative error bound is 1.5e-20. However, (2.19) makes it also clear that there is an advantage to avoid using large scaling factors: the error bound

(2.19) deteriorates quickly with k. For example, already with $k = 20$ and $\eta = \text{EPS}$ it predicts a loss of six digits. Precisely what we observed in Example 1.1.

Remark 2.9. In this section, we have not taken rounding errors into account. However, the accumulation of roundoff errors occurring during the squaring phase can severely affect the accuracy of the computed exponential matrix. Some interesting results about rounding errors accumulation can be found in [15] and [1].

3. Block triangular matrices and Padé approximations.

Let us now restrict attention to A as in (1.2) and $F(A) = e^A$ as in (1.3). (To avoid trivial cases, we will also assume that $A_{12} \neq 0$, as otherwise the computation reduces to separate computations for the diagonal blocks.) In this case, it is well known that Padé approximations are also block triangular matrices: $R(A) = \begin{bmatrix} R(A_{11}) & R_{12}(A) \\ 0 & R(A_{22}) \end{bmatrix}$. This special block structure hints that finding $R(A_{ii}),\ i = 1, 2,$ should not be affected by A_{12}, and that approximation of F_{12} may be done differently from that of $R(A_{ii})$.

Indeed, suppose we have approximated the diagonal blocks by $R(A_{ii}),\ i = 1, 2,$ and that A_{11} and A_{22} have no common eigenvalues. Then, $F(A)$ may be approximated by $R(A)$ exploiting the relation (*Parlett's method*) $R(A)A = AR(A)$. This gives the following equation for $R_{12}(A)$:

$$A_{11}R_{12}(A) - R_{12}(A)A_{22} = R(A_{11})A_{12} - A_{12}R(A_{22})\,. \tag{3.1}$$

As a matter of fact, if the blocks A_{11} and A_{22} are sufficiently separated so that (3.1) is well conditioned (see [7]), probably there is no simpler way to approximate F_{12} than using (3.1). For this reason, we will think of having to approximate e^A in case in which the blocks A_{11} and A_{22} are not sufficiently separated (say, they have close –or identical– eigenvalues).

The approach recently proposed by Kenney and Laub in [12] is of interest in the situation we just described, and we refer to [12] for details on their approach. Presently, we observe that Padé approximations for all of e^A can also be used in case A_{11} and A_{22} are not well separated, but they seem to suffer from the need to bring $\|A\|$ close to 0, and this may result in overscaling relative to the diagonal blocks (see Example 1.1). However, this is only a result of unrefined error estimates. Our main result, Theorem 3.2, will guarantee that, if $\|A_{11}\|$ and $\|A_{22}\|$ are sufficiently small, a single Padé approximant for all of e^A will give small relative errors in a block sense:

$$\frac{\|e^{A_{11}} - R(A_{11})\|}{\|e^{A_{11}}\|}\,,\quad \frac{\|e^{A_{22}} - R(A_{22})\|}{\|e^{A_{22}}\|}\,,\quad \frac{\|F_{12} - R_{12}(A)\|}{\|F_{12}\|}\,. \tag{3.2}$$

Therefore, we will avoid dealing separately with the term F_{12}. Moreover, in case in which $\|A_{ii}\|,\ i = 1, 2,$ are not sufficiently small, our result will tell that it suffices to scale A so to reduce the norm of the A_{ii}. In general, therefore, the value of k we

will select in order to approximate $F(A_k)$, $A_k = A/2^k$, will be smaller, often much smaller, than the value we would have needed to guarantee that $\|A\|/2^k$ had been sufficiently small. As a consequence, we will also need to perform fewer squarings of the obtained result: this will result in a gain with respect to the bound (2.19), reduce the impact of roundoff errors (we perform less arithmetic), and altogether give in an efficient algorithm.

We assume that $Q(A)$ is invertible, and begin by rewriting (2.2) in block form:

$$F(A) - R(A) = \begin{bmatrix} Q^{-1}(A_{11})G(A_{11}) & F_{12} - R_{12}(A) \\ 0 & Q^{-1}(A_{22})G(A_{22}) \end{bmatrix},$$

with

$$\begin{aligned} F_{12} - R_{12}(A) &= Q^{-1}(A_{11})G_{12}(A) + Q_{12}^{-1}(A)G(A_{22}) \\ &= Q^{-1}(A_{11})G_{12}(A) - Q^{-1}(A_{11})Q_{12}(A)Q^{-1}(A_{22})G(A_{22}) \\ &= Q^{-1}(A_{11})\left\{G_{12}(A) - Q_{12}(A)[e^{A_{22}} - R(A_{22})]\right\}. \end{aligned}$$

Then

$$\|F_{12} - R_{12}(A)\| \leq \|Q^{-1}(A_{11})\| \left\{\|G_{12}(A)\| + \|Q_{12}(A)\| \, \|e^{A_{22}} - R(A_{22})\|\right\}, \tag{3.3}$$

and we have to consider the extra-diagonal blocks of $Q(A)$ and $G(A)$.

Lemma 3.1. *Let A be partitioned as in (1.2), $Q(A) = \begin{bmatrix} Q(A_{11}) & Q_{12}(A) \\ 0 & Q(A_{22}) \end{bmatrix}$ be the s-degree matrix polynomial defined in (2.1), and $G(A) = \begin{bmatrix} G(A_{11}) & G_{12}(A) \\ 0 & G(A_{22}) \end{bmatrix}$ be the matrix function defined in (2.3). Then*

$$\|Q_{12}(A)\| \leq \|A_{12}\| \, \frac{1}{2} e^{\frac{s-1}{2s-1}\omega}, \tag{3.4}$$

$$\|G_{12}(A)\| \leq \|A_{12}\| \, \frac{(s!)^2}{((2s)!)^2} e^{\omega} \omega^{2s}, \tag{3.5}$$

with $\omega = \max(\|A_{11}\|, \|A_{22}\|)$.

Proof. Notice that $A^j = \begin{bmatrix} A_{11}^j & \sum_{k=0}^{j-1} A_{11}^{j-1-k} A_{12} A_{22}^k \\ 0 & A_{22}^j \end{bmatrix}$. Then, by (2.1) we get

$$Q_{12}(A) = \sum_{j=1}^{s} \frac{(2s-j)!s!}{(2s)!(s-j)!} \frac{(-1)^j}{j!} \sum_{k=0}^{j-1} A_{11}^{j-1-k} A_{12} A_{22}^k .$$

So, taking norms and using (2.6) we obtain

$$\begin{aligned}
\|Q_{12}(A)\| &\leq \sum_{j=1}^{s} \frac{(2s-j)!s!}{(2s)!(s-j)!}\frac{1}{j!}\sum_{k=0}^{j-1}\|A_{11}\|^{j-1-k}\|A_{22}\|^k\|A_{12}\| \\
&\leq \sum_{j=0}^{s-1} \frac{(2s-1-j)!(s-1)!s}{2s(2s-1)!(s-1-j)!}\frac{\omega^j}{j!}\|A_{12}\| \\
&\leq \frac{1}{2}\sum_{j=0}^{s-1}\left(\frac{s-1}{2s-1}\right)^j\frac{\omega^j}{j!}\|A_{12}\| \;\leq\; \frac{1}{2}e^{\frac{s-1}{2s-1}\omega}\|A_{12}\|,
\end{aligned}$$

which gives (3.4). To obtain (3.5), we first observe that by (2.13)

$$G_{12}(A) = \frac{(-1)^s}{(2s)!}\sum_{j=0}^{\infty}\frac{(s+j)!s!}{(2s+j+1)!}\frac{1}{j!}\sum_{k=0}^{2s+j}A_{11}^{2s+j-k}A_{12}A_{22}^k.$$

Then, taking norms and using (2.11) we get

$$\begin{aligned}
\|G_{12}(A)\| &\leq \frac{1}{(2s)!}\sum_{j=0}^{\infty}\frac{(s+j)!s!}{(2s+j+1)!}\frac{1}{j!}\sum_{k=0}^{2s+j}\|A_{11}\|^{2s+j-k}\|A_{22}\|^k\|A_{12}\| \\
&\leq \frac{s}{(2s)!}\sum_{j=0}^{\infty}\frac{(s+j)!(s-1)!}{(2s+j)!}\frac{1}{j!}\omega^{2s+j}\|A_{12}\| \\
&= \frac{s}{(2s)!}\sum_{j=0}^{\infty}\left(\int_0^1 u^{s+j}(1-u)^{s-1}du\right)\frac{1}{j!}\omega^{2s+j}\|A_{12}\| \\
&= \frac{s}{(2s)!}\omega^{2s}\left(\int_0^1 u^s(1-u)^{s-1}\sum_{j=0}^{\infty}\frac{(\omega u)^j}{j!}du\right)\|A_{12}\| \\
&= \frac{s}{(2s)!}\omega^{2s}\left(\int_0^1 u^s(1-u)^{s-1}e^{\omega u}du\right)\|A_{12}\| \\
&\leq \frac{s}{(2s)!}\omega^{2s}e^{\omega}\frac{s!(s-1)!}{(2s)!}\|A_{12}\|\,.
\end{aligned}$$

□

We are now ready to state our main result.

Theorem 3.2. *Let A be partitioned as in (1.2) and assume that $\|A_{ii}\| = \omega_i < \log 4$, $i = 1, 2$. Let $R(A) = \begin{bmatrix} R(A_{11}) & R_{12}(A) \\ 0 & R(A_{22}) \end{bmatrix}$ be the (s,s) diagonal Padé approximant for $F(A) = e^A = \begin{bmatrix} F(A_{11}) & F_{12} \\ 0 & F(A_{22}) \end{bmatrix}$. Then we have the following error*

bounds:

$$\|F(A_{ii}) - R(A_{ii})\| \leq c(s)\frac{e^{\omega_i}}{2-e^{\omega_i/2}}\omega_i^{2s+1}\,,\ i=1,2, \tag{3.6}$$

$$\frac{\|F(A_{ii}) - R(A_{ii})\|}{\|F(A_{ii})\|} \leq c(s)\frac{e^{\omega_i}}{2-e^{\omega_i/2}}\omega_i^{2s+1}\,,\ i=1,2, \tag{3.7}$$

$$\|F_{12} - R_{12}(A)\| \ \leq\ \|A_{12}\|\, c(s)\frac{e^{\omega}}{2-e^{\omega/2}}\omega^{2s}\left[2s+1+\frac{\omega}{2}\frac{e^{\frac{s-1}{2s-1}\omega}}{2-e^{\omega/2}}\right], \tag{3.8}$$

with $\omega = \max(\omega_1,\omega_2)$ *and* $c(s) = \frac{(s!)^2}{(2s+1)((2s)!)^2}$. *If* $\omega < \log 2$ *we also have*

$$\frac{\|F_{12} - R_{12}(A)\|}{\|F_{12}\|} \ \leq\ \frac{c(s)}{2-e^{\omega}}\,\frac{e^{\omega}}{2-e^{\omega/2}}\omega^{2s}\left[2s+1+\frac{\omega}{2}\frac{e^{\frac{s-1}{2s-1}\omega}}{2-e^{\omega/2}}\right]. \tag{3.9}$$

Proof. Easily, (3.7) and (3.6) are the block-diagonal versions of (2.17) and (2.16); (3.8) follows from (3.3)-(3.5), (2.16) and (2.5). To obtain (3.9) we observe that A is the principal logarithm[3] of $F(A)$. Moreover, if $\omega_i < \log 2$,

$$\|I - F(A_{ii})\| \leq e^{\|A_{ii}\|} - 1 \leq e^{\omega_i} - 1 < 1\ .$$

Then, we can apply [3, Theorem 4.6] and [3, (4.15)] to obtain $\|A_{12}\| \leq \frac{\|F_{12}\|}{2-e^{\omega}}$. □

Remark 3.3. If $\omega = 0.4, 0.45, 0.5$ (notice that $\log 2 \approx .69$) and $s = 6$, the bound (3.9) in Theorem 3.2 guarantees relative errors in the extradiagonal block less then 1.5e-16, 7.8e-16, 3.7e-15, respectively. If $\omega = 0.5, 1, 1.35$ and $s = 6$, (3.8) ensures relative errors with respect to $\|A_{12}\|$ less than 1.3e-15, 2.1e-11, 3.3e-8, respectively. Notice that, in theory, the relative error bounds (3.7) and (3.9) could be arbitrarily small, if ω_i are allowed to be arbitrarily small. Of course, because of finite precision, we are not going to observe relative errors arbitrarily small. The correct interpretation of the bounds, then, is that relative errors of $O(\text{EPS})$ are in principle attainable if ω is sufficiently small; as we just saw, for $s = 6$, it is enough that $\omega \leq 0.4$.

Example 3.4. Consider once more Example 1.1: $A = \left[\begin{smallmatrix}\omega & x\\ 0 & \omega\end{smallmatrix}\right]$ with $x = 1.0\text{e+}6$ and several values of $\omega > 0$ (results for negative values of ω were nearly identical to those for $|\omega|$). The results in Table 2 refer to computed and estimated errors for the (6,6) diagonal Padé to e^A. We have implemented the (6,6) Padé by using integer coefficients and arranging powers of A in ascending order. In Table 2, we adopted the following notation: $\texttt{est}_i$: absolute (and relative) error bound (3.6) (of course, for the present example, we got identical results for the case $i = 1, 2$), $\texttt{abs}_i$: $|e^{A_{ii}} - R(A_{ii})|$, $\texttt{rel}_i$: $\texttt{abs}_i/|e^{A_{ii}}|$, $\texttt{est}_{12}$: absolute error bound (3.8), $\texttt{abs}_{12}$: $|(e^A)_{12} - R_{12}(A)|$, $\texttt{rel}_{12}$: $\texttt{abs}_{12}/|(e^A)_{12}|$. Quite clearly, there is excellent agreement

[3]This is because if $\lambda = \alpha + i\beta$ is an eigenvalue of A with $|\beta| \geq \pi$, then either $\|A_{11}\| \geq \pi$ or $\|A_{22}\| \geq \pi$, contradicting that $\|A_{ii}\| \leq \log 4, i = 1, 2$

ω	$\texttt{est}_i$	$\texttt{abs}_i$	$\texttt{est}_{12}$	$\texttt{abs}_{12}$	$\texttt{rel}_{12}$
0.1	2.0e-026	0.0e+000	2.6e-018	0.0e+000	0.0e+000
0.3	4.5e-020	0.0e+000	2.0e-012	0.0e+000	0.0e+000
0.5	4.9e-017	0.0e+000	1.3e-009	1.2e-009	7.1e-016
0.7	5.8e-015	3.6e-015	1.2e-007	6.7e-008	3.3e-014
0.9	2.5e-013	1.1e-013	4.1e-006	1.7e-006	6.9e-013
1.1	6.8e-012	1.8e-012	1.0e-004	2.4e-005	7.9e-012
1.3	2.3e-010	2.0e-011	4.7e-003	2.2e-004	6.0e-011

TABLE 2. Estimated and computed errors for the (6,6) Padé

with the theoretical bounds. Entries of 0.0e+000 correspond to an identical finite precision representation of computed and "exact" values.

Finally, suppose we have a matrix A as in (1.2) whose diagonal blocks $\|A_{ii}\|$ do not satisfy the assumptions of Theorem 3.2. Then, Theorem 3.2 suggests the following strategy:

1. select k so that the diagonal blocks of the matrix $A_k = A/2^k$ satisfy the assumptions of Theorem 3.2,
2. approximate $F(A_k)$ with the Padé approximation $R(A_k)$,
3. approximate $F(A)$ with $\widehat{F}(A) = (R(A_k))^{2^k}$.

Our examples below show that this is an effective strategy, and it avoids some common overscaling pitfalls. To understand why, we need to work harder than how we did to reach (2.19), since (2.18) cannot be assumed for Δ (we have scaled only the diagonal blocks, so $\|A_k\|$ is not necessarily small, while the arguments leading to (2.18) required $\|A_k\|$ sufficiently small); of course, (2.18) can be assumed for Δ_k, since relative to A_k the estimates (3.6, 3.7, 3.8, 3.9) hold. Here below, we will give two types of error statements: forward and backward. In both cases, we will obtain error bounds which betray the dangers of overscaling, and are more refined than the standard ones, which require $\|A_k\|$ to be small.

Let us begin trying to obtain a bound on $\frac{\|F(A)-\widehat{F}(A)\|}{\|F(A)\|}$. Observe that

$$\|F(A) - \widehat{F}(A)\| \leq \| \begin{bmatrix} \|F(A_{11})-\widehat{F}_{11}(A)\| & \|F_{12}(A)-\widehat{F}_{12}(A)\| \\ 0 & \|F(A_{22})-\widehat{F}_{22}(A)\| \end{bmatrix} \| \,.$$

Now, let $\Delta = \left[\begin{smallmatrix} \Delta_{11} & \Delta_{12} \\ 0 & \Delta_{22} \end{smallmatrix}\right]$ be defined as before (2.18). Then, if $\Delta_{ii} \leq \eta_i$, $i = 1, 2$, it is immediate to get

$$\frac{\|F_{ii}(A) - \widehat{F}_{ii}(A)\|}{\|F_{ii}(A)\|} \leq (1 + \eta_i)^{2^k} - 1 \,. \tag{3.10}$$

Finally, notice that an upper bound for η_i is given by the right-hand-side of (3.7) ($i = 1, 2$). To bound $\|F_{12}(A) - \widehat{F}_{12}(A)\|$, we observe that

$$F_{12}(A)-\widehat{F}_{12}(A) = (I-(I+\Delta_{11})^{2^k})F_{12}(A)-\big[\sum_{j=0}^{2^k-1}(I+\Delta_{11})^{2^k-1-j}\Delta_{12}(I+\Delta_{22})^j\big]F_{22}(A)\,.$$

Therefore, with $\eta = \max(\eta_1, \eta_2)$, we have

$$\|F_{12}(A) - \widehat{F}_{12}(A)\| \le ((1+\eta_1)^{2^k} - 1)\|F_{12}(A)\| + 2^k(1+\eta)^{2^k-1}\|F(A_{22})\|\ \|\Delta_{12}\|\,.$$

Finally, since

$$\Delta_{12} = Q^{-1}(A_{11}/2^k)\big(G_{12}(-A_k) - Q_{12}(A_k)\Delta_{22}\big)\,,$$

following the same arguments used in Lemma 3.1 and Theorem 3.2, we obtain that for $\|\Delta_{12}\|$ the bound (3.8) holds as well:

$$\|\Delta_{12}\| \le \gamma\frac{\|A_{12}\|}{2^k}\,, \quad \gamma = c(s)\frac{e^\omega}{2-e^{\omega/2}}\omega^{2s}\left[2s+1+\frac{\omega}{2}\frac{e^{\frac{s-1}{2s-1}\omega}}{2-e^{\omega/2}}\right]\,, \tag{3.11}$$

where $\omega = \max(\|A_{11}\|/2^k\ ,\ \|A_{22}\|/2^k)$. In summary, we get

$$\begin{aligned}\frac{\|F_{12}(A) - \widehat{F}_{12}(A)\|}{\|F(A)\|} &\le ((1+\eta_1)^{2^k} - 1) + \gamma(1+\eta)^{2^k-1}\frac{\|F(A_{22})\|\ \|A_{12}\|}{\|F(A)\|} \\ &\le ((1+\eta_1)^{2^k} - 1) + \gamma(1+\eta)^{2^k-1}\|A_{12}\|\,.\end{aligned} \tag{3.12}$$

Putting together (3.10) and (3.12), we get a bound for the relative error $\frac{\|F(A)-\widehat{F}(A)\|}{\|F(A)\|}$, which is more refined than traditional error bounds (e.g., see [5, Section 11.3.2]).

Remark 3.5. To obtain a sharp bound for $\frac{\|F(A_{22})\|\ \|A_{12}\|}{\|F(A)\|}$ is not trivial at all, and the search for a sharp bound remains an open problem. We have used $\frac{\|F(A_{22})\|\ \|A_{12}\|}{\|F(A)\|} \le \|A_{12}\|$, but this is not necessarily sharp in cases of interest: e.g., for Example 1.1, regardless of ω, the quantity $\frac{\|F(A_{22})\|\ \|A_{12}\|}{\|F(A)\|}$ is $O(1)$, while $\|A_{12}\| = x = 1.\text{e+}6$.

Next, we provide a backward analysis of our new scaling & squaring strategy; this analysis will clearly show when and how the new strategy improves upon the traditional one. We have the following result

Theorem 3.6. *Let A be defined as in* (1.2)*. Let k be chosen so that for the matrix $A_k = A/2^k$ the assumptions of Theorem 3.2 hold, and let $R(A_k)$ be the (s, s) Padé approximant for e^{A_k}. Let $\widehat{F}(A) = (R(A_k))^{2^k}$ be the adopted approximant for $F(A) = e^A$, and let $\Delta = e^{-A_k}R(A_k)-I$ be partitioned conformally to A. Then, $\widehat{F}(A) = e^{A+E}$, where $E = \left[\begin{smallmatrix} E_{11} & E_{12} \\ 0 & E_{22}\end{smallmatrix}\right]$ and*

$$\|E_{11}\| \le c_{11}\|A_{11}\|\,,\ \|E_{22}\| \le c_{22}\|A_{22}\|\,,\ \|E_{12}\| \le c_{12}\|A_{12}\|\,. \tag{3.13}$$

Moreover, we have

$$c_{ii} \le \frac{|\log(1-\eta_i)|}{\omega_i}\ ,\ \ i=1,2\ ,\ c_{12} \le \frac{e^{\omega_1}}{1-\eta}\ (\gamma + \eta_2 e^{\omega_2})\,, \tag{3.14}$$

where ω_i are defined in Theorem 3.2, for η_i we can take the right-hand-side of (3.6)*, $i=1,2$, $\eta=\max(\eta_1,\eta_2)$, and γ is defined in* (3.11).

Proof. We first look for E_k such that $e^{-A_k}R(A_k)=e^{E_k}$, and then –since $e^{-A}\widehat{F}(A) = (e^{-A_k}R(A_k))^{2^k}$– we will have $E = 2^k E_k$. Now, E_k is nothing but the principal logarithm of $e^{-A_k}R(A_k)$. Partition

$$e^{-A_k} = \begin{bmatrix} e^{-A_{11}/2^k} & -e^{-A_{11}/2^k}F_{12}(A_k)e^{-A_{22}/2^k} \\ 0 & e^{-A_{22}/2^k} \end{bmatrix}\ ,\ R(A_k) = \begin{bmatrix} R(A_{11}/2^k) & R_{12}(A_k) \\ 0 & R(A_{22}/2^k) \end{bmatrix}\ ,$$

so that

$$e^{-A_k}R(A_k) = \begin{bmatrix} I+\Delta_{11} & e^{-A_{11}/2^k}[R_{12}(A_k)-F_{12}(A_k)(I+\Delta_{22})] \\ 0 & I+\Delta_{22} \end{bmatrix}\ .$$

Then, according to [3, (1.1) and (4.10)], E_k is given by $E_k = \begin{bmatrix} E_{11}^{(k)} & E_{12}^{(k)} \\ 0 & E_{22}^{(k)} \end{bmatrix}$ where

$$\begin{aligned} E_{ii}^{(k)} &= \log(e^{-A_{ii}/2^k}R(A_{ii}/2^k)) \ = \ \log(I+\Delta_{ii})\,,\ i=1,2,\ \text{and} \\ E_{12}^{(k)} &= \int_0^1 [\Delta_{11}t+I]^{-1}e^{-A_{11}/2^k}[R_{12}(A_k)-F_{12}(A_k)(I+\Delta_{22})][\Delta_{22}t+I]^{-1}dt\,. \end{aligned} \tag{3.15}$$

At this point, we observe that $\|\log(I+\Delta_{ii})\| \le |\log(1-\eta_i)|$, $i=1,2$, where (following the same arguments used in Theorem 3.2) for η_i we can take the right-hand-side of (3.6). Finally, since $E_{ii}=2^kE_{ii}^{(k)}$, we immediately get

$$\|E_{ii}\| \le 2^k|\log(1-\eta_i)| = \frac{|\log(1-\eta_i)|}{\omega_i}\|A_{ii}\|\,,\ i=1,2\,,$$

thereby obtaining the result about the diagonal blocks. For $E_{12}^{(k)}$ we have

$$\begin{aligned} \|E_{12}^{(k)}\| &\le \|e^{-A_{11}/2^k}[R_{12}(A_k)-F_{12}(A_k)(I+\Delta_{22})]\| \int_0^1 \frac{1}{(1-\eta_1 t)(1-\eta_2 t)}dt \\ &\le \frac{1}{1-\eta}\|e^{-A_{11}/2^k}[R_{12}(A_k)-F_{12}(A_k)(I+\Delta_{22})]\|\,. \end{aligned}$$

Now, from (3.8), we have

$$\|e^{-A_{11}/2^k}(R_{12}(A_k)-F_{12}(A_k))\| \le \frac{\gamma e^{\omega_1}}{2^k}\|A_{12}\|\,,$$

and from (1.3) relative to $F_{12}(A_k)$ we get

$$\|e^{-A_{11}/2^k}F_{12}(A_k)\Delta_{22}\| \le \frac{\eta_2}{2^k}\|A_{12}\|e^{\omega_1+\omega_2}\,.$$

Thus, we eventually obtain

$$\|E_{12}^{(k)}\| \leq \frac{e^{\omega_1}}{2^k(1-\eta)}(\gamma + \eta_2 e^{\omega_2})\|A_{12}\|$$

and the Theorem is proved. □

Remark 3.7. Theorem 3.6 shows precisely when the new scaling strategy improves upon the standard one. In practice, the η_i's are generally at best $O(\text{EPS})$, and this can be expected to be the case (see Remark 3.3). Therefore, with our strategy, as well as with the standard strategy, c_{12} is $O(\text{EPS})$. For us, also at least one of c_{11} and c_{22} is $O(\text{EPS})$ (in fact, both c_{11} and c_{22} are $O(\text{EPS})$ if $\|A_{11}\|$ and $\|A_{22}\|$ are of the same size). But in the standard strategy this is not true if $\|A_{12}\|$ is large compared to $\|A_{ii}\|$. In this case, one ends up overscaling with respect to the diagonal blocks and the constants c_{ii} become large because the denominator ω_i approaches 0. Indeed, our examples clearly show that in cases in which $\|A_{11}\| \approx \|A_{22}\| \ll \|A_{12}\|$ the standard strategy does not lead to a stable algorithm.

Remark 3.8. An intriguing phenomenon encountered in approximating e^A is the so called "hump", introduced in [14][4]. Although we do not fully understand this hump, it appears to be an issue caused by roundoff errors. For matrices like in (1.2), our results imply that fewer scalings and successive squarings of A are generally needed with respect to the standard implementation. Thus, we should expect less roundoff propagation and a likely reduction in the occurrence of the hump phenomenon.

Example 3.9. This is once more Example 1.1. In Table 3 we report on relative errors (in norm) obtained with different scaling strategies. We stress once more that in the "standard" scaling strategy it is the norm of A to dictate the exponent k in the scaling factor 2^k, whereas with our improved strategy based on Theorem 3.2 it is ω to dictate the value of k. In agreement with our error estimates, it is evident the loss of about 6 decimal digits when using a scaling factor $2^k = 2^{21}$ (see Remark 2.8). However, scaling only with respect to the diagonal elements of A, we recover a fully accurate approximation.

Example 3.10. This is similar to Example 1.1, except that the blocks have arbitrary dimension n. We have the matrix

$$A = \frac{1}{n}\begin{bmatrix} \omega\mathcal{E} & x\mathcal{E} \\ 0 & -\omega\mathcal{E} \end{bmatrix}, \qquad \text{where} \qquad \mathcal{E} \in \mathbb{R}^{n\times n}\ ,\ \mathcal{E} = \begin{bmatrix} 1 & \dots & 1 \\ \vdots & & \vdots \\ 1 & \dots & 1 \end{bmatrix}. \tag{3.16}$$

Notice that $\|A_{ii}\| = \omega$, $i = 1, 2$, and also

$$e^A = \begin{bmatrix} e^{\frac{\omega}{n}\mathcal{E}} & \frac{x}{n}\frac{e^{\omega}-e^{-\omega}}{2\omega}\mathcal{E} \\ 0 & e^{\frac{-\omega}{n}\mathcal{E}} \end{bmatrix}, \quad e^{\frac{\omega}{n}\mathcal{E}} = \frac{1}{n}\begin{bmatrix} a & b & \dots & b \\ b & a & \ddots & b \\ \vdots & \ddots & \ddots & \vdots \\ b & \dots & b & a \end{bmatrix}\ ,\ a = e^{\omega} + (n-1),\ b = e^{\omega} - 1\,.$$

[4] a referee asked us to relate our results to this phenomenon

ω	"standard" scaling (k)	no scaling ($k=0$)	improved scaling (k)
0.1	2.1e-011 (21)	0.0e+000	0.0e+000 (0)
0.5	4.8e-012 (21)	7.1e-016	7.1e-016 (0)
0.9	5.4e-011 (21)	6.9e-013	5.7e-016 (1)
1.3	2.2e-010 (21)	6.0e-011	2.5e-016 (2)
2.1	1.4e-010 (21)	2.1e-008	5.7e-016 (3)
4.1	7.6e-011 (21)	9.6e-005	1.9e-015 (4)
6.1	1.9e-010 (21)	1.9e-002	1.1e-015 (4)
8.1	8.0e-012 (21)	6.6e-001	1.7e-015 (5)

TABLE 3. Relative errors (used scaling factor: 2^k).

For $n = 10$, in Table 4 we report on relative errors for the (6,6) Padé approximation, coupled with different scaling strategies, in a similar way to what we did in Table 3.

ω	"standard" scaling (k)	no scaling ($k=0$)	improved scaling (k)
0.1	2.5e-010 (21)	7.7e-016	7.7e-016 (0)
0.3	5.8e-010 (21)	2.1e-016	2.1e-016 (0)
0.5	1.3e-009 (21)	2.4e-016	2.4e-016 (0)
0.7	1.2e-009 (21)	3.0e-015	3.6e-016 (1)
0.9	1.0e-009 (21)	6.3e-014	2.9e-016 (1)
1.1	2.6e-010 (21)	7.7e-013	9.5e-016 (2)
1.3	1.9e-009 (21)	6.3e-012	6.2e-016 (2)

TABLE 4. Relative errors Example 3.10; scaling factor: 2^k.

4. Conclusions and extensions.

In this work, we have revisited Padé approximation techniques to compute the exponential of a block triangular matrix. Our main result has been Theorem 3.2, which gives improved error bounds for a 2×2 block triangular matrix with well scaled diagonal blocks. As a consequence of this theorem, we have proposed a new scaling & squaring strategy for matrices of the form (1.2), and given an error analysis for the new strategy. We have exemplified how the new strategy can lead to accurate approximations by avoiding overscaling.

We have restricted to 2×2 block triangular matrices, since in our opinion this is the most important case one needs to understand. But, of course, our results can be used for a block triangular matrix A with any number of diagonal blocks;

for example, A may have been obtained by a prior Schur reduction. (Of course, in agreement with what we said at the beginning of Section 3, for us this is of interest when the diagonal blocks of A are not sufficiently separated – have close, or identical, eigenvalues). So, suppose that A is in the form

$$A = \begin{bmatrix} A_{11} & A_{12} & \dots & A_{1p} \\ 0 & A_{22} & \dots & A_{2p} \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & A_{pp} \end{bmatrix} . \tag{4.1}$$

Clearly, also $F(A)$ and a diagonal Padé approximation $R(A)$ to $F(A)$ have this same block structure. Now, assuming that all entries of $F(A)$ are of equal interest, and that **just one** Padé approximation $R(A)$ is computed for all of $F(A)$, we can use Theorem 3.2 for the 2×2 block partitioning of A associated to the most favorable error bounds predicted by the theorem. In this case, a little thought reveals that the best 2×2 block partitioning is that which achieves

$$\min_{1 \le j \le p} \text{Max} \left[\, \|A(1:j, 1:j)\|, \ \|A(j+1:p, j+1:p)\| \, \right], \tag{4.2}$$

where our notation is inherited from (4.1); e.g., $A(1:2, 1:2) = \left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix} \right]$. Therefore, in agreement with our discussion on scaling, one may want to scale with respect to the diagonal blocks of this "best" block partitioning.

This is how one can use the results in this paper if is willing to do just one Padé approximation for $F(A)$ when A is as in (4.1). Alternatively, one may want to proceed recursively from the diagonal of $F(A)$ upward, one superdiagonal at the time. This may be a useful way to proceed in case blocks close to the diagonal need to be found with greater accuracy, but one may end up computing the same quantities more than once. For example, suppose that $p = 3$ in (4.1). We can find F_{12} and F_{23} by using Theorem 3.2 on the matrices $\left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix} \right]$ and $\left[\begin{smallmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{smallmatrix} \right]$, respectively. To obtain F_{13}, we can use Theorem 3.2 with two different choices of blocking: $\left[\begin{smallmatrix} A_{11} & [A_{12} \ A_{13}] \\ \left[\begin{smallmatrix} 0 \\ 0 \end{smallmatrix} \right] & \left[\begin{smallmatrix} A_{22} & A_{23} \\ 0 & A_{33} \end{smallmatrix} \right] \end{bmatrix}$ or $\left[\begin{smallmatrix} \left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix} \right] & \left[\begin{smallmatrix} A_{13} \\ A_{23} \end{smallmatrix} \right] \\ [0 \ 0] & [A_{33}] \end{bmatrix}$. To fix ideas, suppose we use the latter choice. But then, we may end up having to rescale the block $\left[\begin{smallmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{smallmatrix} \right]$ more than how we had previously rescaled it to compute F_{12}, and thus we will end up recomputing quantities which had already been computed. With our present understanding, this seems unavoidable.

References

[1] M. Arioli, B. Codenotti, and C. Fassino. The Padé method for computing the matrix exponential. *Linear Algebra and Applic.*, 240:111–130, 1996.

[2] P.E. Crouch and R. Grossman. Numerical integration of ordinary differential equations on manifolds. *J. Nonlinear Sc.*, 3:1–33, 1993.

[3] L. Dieci and A. Papini. Conditioning and Padé approximation of the logarithm of a matrix. *SIAM J. Matrix Anal. Appl.*, 1999. To appear.

[4] W. Fair and Y. L. Luke. Padé approximations to the operator exponential. *Numer. Math.*, 14:379–382, 1970.
[5] G. H. Golub and C. F. Van Loan. *Matrix Computations*. The Johns Hopkins University Press, 2nd edition, 1989.
[6] E. Hairer and G. Wanner. *Solving Ordinary Differential Equations II*. Springer-Verlag, Berlin-Heidelberg, 1991.
[7] N. J. Higham. Perturbation theory and backward error for $AX - XB = C$. *BIT*, 33:124–136, 1993.
[8] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the exponential operator. *SIAM J. Numer. Anal.*, 34:1911–1925, 1997.
[9] M. Hochbruck, C. Lubich, and H. Selhofer. Exponential integrators for large systems of differential equations. *SIAM J. Sci. Comput.*, 19:1552–1574, 1998.
[10] B. Kagström. Bounds and perturbation bounds for the matrix exponential. *BIT*, 17:39–57, 1977.
[11] C. Kenney and A. J. Laub. Padé error estimates for the logarithm of a matrix. *Internat. J. Control*, 50(3):707–730, 1989.
[12] 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.
[13] C. Van Loan. The sensitivity of the matrix exponential. *SIAM J. Numer. Anal.*, 14:971–981, 1977.
[14] C. B. Moler and C. Van Loan. Nineteen dubious ways to compute the exponential of a matrix. *SIAM Rev.*, 20:801–836, 1978.
[15] R. C. Ward. Numerical computation of the matrix exponential with accuracy estimates. *SIAM J. Numer. Anal.*, 14:600–610, 1977.

School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332 U.S.A.
E-mail address: dieci@math.gatech.edu

Dep. Energetica S. Stecco, Univ. of Florence, via C. Lombroso 6/17, 50134 Florence, Italy
E-mail address: papini@de.unifi.it