Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

SIAM J. Matrix Anal. Appl.
Vol. 41, No. 4, pp. 1674–1703

© 2020 Society for Industrial and Applied Mathematics

COMPUTING ENCLOSURES FOR THE MATRIX EXPONENTIAL[*]

ANDREAS FROMMER[†] AND BEHNAM HASHEMI[‡]

Abstract. We present a review of old interval arithmetic techniques, and develop new ones, for computing enclosures for all entries of the *exact* exponential of a matrix. This means that all the rounding and truncation errors committed in the course of computation are rigorously taken into account, and the result is mathematically guaranteed to contain the correct matrix exponential. We consider algorithms relying on verified spectral decomposition, two variants relying on Taylor series expansion, a Padé approximation, and a contour integration approach, together with a Chebyshev approximation–based method which is designed for Hermitian matrices. Most of our methods use the scaling and squaring framework and are examined when applied to both the original matrix and an approximate diagonalization. In addition to a comparative study of algorithms, several illustrative numerical examples are given.

Key words. matrix exponential, interval arithmetic, automatic result verification, INTLAB, scaling and squaring

AMS subject classifications. 65F60, 65F30, 65G20

DOI. 10.1137/19M1263431

1. Introduction.

The task of computing the exponential $\exp(A)$ of a matrix $A \in \mathbb{C}^{n\times n}$ arises in a variety of applications such as in exponential integrators for ODEs and semidiscretizations of PDEs, in network analysis, or in continuous-time Markov models. The development of stable and efficient methods for computing $\exp(A)$ has thus been a topic of intensive research; see the survey paper [29] from 1978 and its update [30] from 2003. A Padé approximation–type method [15] combined with a scaling and squaring approach recently improved in [1] is implemented in the MATLAB function `expm`. Roughly speaking, the approach determines a scaling parameter s and an order q with the ultimate goal that the *backward* error in computing $\exp(A)$ via squaring a (q,q)-type Padé approximation of the scaled matrix is in the order of the unit round-off u.[1] Specifically, it is shown in [1] how to choose s and $q \in \{3,5,7,9,13\}$ to achieve this goal in double precision.

An important question is now how well the result of a computation indeed approximates $\exp(A)$. In this paper, we develop new approaches which, together with the approximation to $\exp(A)$, also compute mathematically guaranteed error bounds for each entry of the matrix. Such approaches will be termed *verified* computations. We compare them with existing ones with respect to the tightness of the bounds obtained and the computational complexity. Conceptually, all these verified computations rely on theoretical results on approximation errors as well as on the use of (machine) interval arithmetic to control the rounding errors related to the use of

[*]Received by the editors May 22, 2019; accepted for publication (in revised form) August 14, 2020; published electronically November 9, 2020.
https://doi.org/10.1137/19M1263431
Funding: The work of the second author was supported by Iran National Science Foundation (INSF) grant 98012590.

[†]Department of Mathematics, University of Wuppertal, 42097 Wuppertal, Germany (frommer@math.uni-wuppertal.de).
[‡]Department of Mathematics, Shiraz University of Technology, Modarres Blvd., Shiraz 71555-313, Iran (hashemi@sutech.ac.ir).

[1]Denoting by $\varepsilon_{\mathrm{mach}}$ the difference between the smallest floating point number > 1 and 1, the unit roundoff u is $\varepsilon_{\mathrm{mach}}/2$ when the rounding mode is rounding to nearest, and it is $\varepsilon_{\mathrm{mach}}$ for directed roundings.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

floating point arithmetic. Note that all the techniques developed in this paper computationally bound the *forward error* and are thus complementary to what is obtained when performing a backward error analysis as is done in [16], for instance.

The paper is organized as follows: Section 2 briefly reviews those properties of interval arithmetic which matter in our setting. Section 3 presents known and develops new approaches to the verified computation of $\exp(A)$. Section 4 shows how these approaches can be combined with approximate diagonalization. We then present a variety of numerical experiments and comparisons in section 5. Some conclusions are formulated in section 6.

2. Interval arithmetic.

This section summarizes those aspects of interval arithmetic that are most important for this paper. For a more thorough treatment we refer the reader to the textbooks [26, 31] and the review paper [42].

Let $\mathbb{IR}$ denote the set of all compact intervals $\boldsymbol{x} = [\underline{\boldsymbol{x}}, \overline{\boldsymbol{x}}]$ on the real line. (Interval quantities will always be denoted in boldface.) The (standard) interval arithmetic operations $+, -, \cdot, /$ on $\mathbb{IR}$ are defined in the set theoretic sense. They again yield an element from $\mathbb{IR}$, the bounds of which can be obtained from the bounds of the interval operands. One way to extend the interval concept to the complex plane is to take $\mathbb{IC}_{\text{disc}}$ as the set of all compact disks $\boldsymbol{z}$ in the complex plane with center $\text{mid}\,(\boldsymbol{z})$ and radius $\text{rad}\,(\boldsymbol{z})$ and to define the result of an arithmetic operation $\boldsymbol{z}_1 \circ \boldsymbol{z}_2$ as the disk with center $\text{mid}\,(\boldsymbol{z}_1) \circ \text{mid}\,(\boldsymbol{z}_2)$ and smallest radius such that it still contains $\{z_1 \circ z_2 : z_1 \in \boldsymbol{z}_1, z_2 \in \boldsymbol{z}_2\}$. This radius can be computed from the midpoints and radii of the operands. This *circular* interval arithmetic can also be used on $\mathbb{IR}$ by restriction to the real axis. The results of multiplication and division are then, in general, supersets of what one gets from the standard interval arithmetic on $\mathbb{IR}$.

In a floating point environment, it is important that for any arithmetic operation $\circ \in \{+, -, \cdot, /\}$ interval arithmetic preserves the very crucial *enclosure property*

$$\{x \circ y : x \in \boldsymbol{x}, y \in \boldsymbol{y}\} \subseteq \boldsymbol{x} \circ \boldsymbol{y}. \tag{2.1}$$

This means that in the floating point computation of the lower and upper bounds of the result (or its midpoint and radius), we have to use different directed rounding modes. On a given modern hardware, changing the rounding mode is a very time-consuming operation as compared to the floating point computation itself. Efficient implementations of machine interval arithmetic as in the MATLAB Toolbox INTLAB [40] or the C++ library C-XSC [21, 19] therefore try to do as few changes of the rounding mode as possible, and this can be achieved by using an operator concept which works on whole arrays in the same spirit as the well-known BLAS (basic linear algebra subprograms). On $\mathbb{IR}$, circular arithmetic must then be used; see [39]. It cannot be emphasized enough that these savings in switchings of the rounding modes affect run times very substantially: interval computations then perform comparably as fast as floating point computations, whereas without these techniques they are likely to be slower by at least two orders of magnitude.

Trivially, the enclosure property (2.1) carries over to any rational expression $r(x_1, \ldots, x_n)$,

$$\{r(x_1), \ldots, r(x_n) : x_i \in \boldsymbol{x}_i \text{ for } i = 1, \ldots, n\} \subseteq r(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n).$$

If any of the variables x_i appears several times in r, we typically encounter the phenomenon of *overestimation* inherent in the use of interval arithmetic, which treats each occurrence of a variable as being independent of its other occurrences. A very

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

simple case is the expression $r(x) = x \cdot x$, which for an interval $\boldsymbol{x} \in \mathbb{IR}$ with $0 \in \boldsymbol{x}$ gives

$$r(\boldsymbol{x}) = [-|\underline{\boldsymbol{x}}\overline{\boldsymbol{x}}|, \max\{\underline{\boldsymbol{x}}^2, \overline{\boldsymbol{x}}^2\}] \supsetneq [0, \max\{\underline{\boldsymbol{x}}^2, \overline{\boldsymbol{x}}^2\}] = \{r(x) : x \in \boldsymbol{x}\}.$$

We face a similar situation when we use interval arithmetic to compute the square $\boldsymbol{B} = \boldsymbol{A} \cdot \boldsymbol{A}$ of an *interval matrix* $\boldsymbol{A} = (\boldsymbol{a}_{ij})_{i,j=1}^n$. In the expression

$$\boldsymbol{b}_{ij} = \sum_{k=1}^{n} \boldsymbol{a}_{ik} \cdot \boldsymbol{a}_{kj}$$

the entry $\boldsymbol{a}_{ij}$ is the only one which occurs more than once, either in $\boldsymbol{a}_{ij}\boldsymbol{a}_{jj}$ and $\boldsymbol{a}_{ii}\boldsymbol{a}_{ij}$ if $i \neq j$ or in $\boldsymbol{a}_{ii}\boldsymbol{a}_{ii}$ if $i = j$. INTLAB as well as virtually any other interval software provides a function $(\cdot)^2$ for intervals which returns (up to roundings) the exact range of the second power for any interval argument. Using this and replacing $\boldsymbol{a}_{ij}\boldsymbol{a}_{jj} + \boldsymbol{a}_{ii}\boldsymbol{a}_{ij}$ by $(\boldsymbol{a}_{ii} + \boldsymbol{a}_{jj})\boldsymbol{a}_{ij}$ in case $i \neq j$ will thus give, in general, narrower intervals for the diagonal of $\boldsymbol{B}$, but for computational efficiency it is important that the whole computation can still be cast into operations on arrays without explicit loops and case distinctions. The following self-explaining MATLAB-INTLAB code shows how this can be achieved using pointwise multiplication:

```
function S = square(A)
n = size(A,1);
c = diag(A);
A(1:n+1:end) = intval(0); % A(i,i) = intval(0) = [0,0];

C = ones(n,1)*c' + c * ones(1,n);
C = C.*A;
C(1:n+1:end) = c.^2;      % C(i,i) = c(i)^2;

S = A*A + C;
end
```

As was observed in [22], proceeding this way we obtain, up to roundings, the *interval hull* $\boldsymbol{S}$ of the set $\mathcal{S} := \{A^2 : A \in \boldsymbol{A}\}$, i.e., the intersection of all interval matrices containing $\mathcal{S}$. Note that $\mathcal{S}$ itself is not an interval matrix. Thus, if we perform another squaring with $\boldsymbol{S}$, we will get the interval hull of all the squares of matrices from $\boldsymbol{S}$ which is in general *more* than the interval hull of the squares of the matrices from $\mathcal{S}$ and is thus larger than the interval hull of the set $\{A^4 : A \in \boldsymbol{A}\}$. It will be important to be aware of this *wrapping effect* when considering scaling and squaring approaches in this paper.

If the two end-points of an interval coincide, it is termed a *point interval.* Performing machine interval arithmetic with point intervals yields nonpoint intervals which contain the exact value of the computation. Interval arithmetic can thus be used as a tool for an automated forward error analysis yielding lower and upper bounds for (arithmetic) expressions involving point quantities. For a more involved computation, though, such a naive use of interval arithmetic will typically end up with quite wide intervals. To obtain narrow enclosures, specific interval methods have to be used. For example, rather than just performing Gaussian elimination in interval arithmetic to solve a linear system $Ax = b$, a narrow enclosure for the solution is obtained by a correction $\boldsymbol{x}$, an interval vector, to an approximate solution $\tilde{x}$, obtained via some

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

floating point computation. The vector $\boldsymbol{x}$ is determined in a such a way that

$$-R(A\tilde{x} - b) + (I - RA)\boldsymbol{x} \subseteq \mathrm{int}\boldsymbol{x}, \tag{2.2}$$

with R being an approximate inverse for A, again computed in standard floating point arithmetic. By a result from [23, 38], based on Brouwer's fixed point theorem, A is then nonsingular and $A^{-1}b \in \tilde{x} + \boldsymbol{x}$. As a side remark let us mention that it was recently shown in [2] that if one uses the restriction of circular arithmetic to the reals to evaluate the left-hand side in (2.2), a much simpler fixed point theorem than Brouwer's can be used to show that $A^{-1}b \in \tilde{x} + \boldsymbol{x}$. The outlined method is the basis of the INTLAB function `verifylss.m` (see also [44]), which will be heavily used in our algorithms.

As a final remark in this section, let us note that when computing a higher power $\boldsymbol{A}^k, k \geq 3$, of an interval matrix, the result will typically depend on the order that we choose for its evaluation. This means that in general we have

$$(\boldsymbol{A} \cdot \boldsymbol{A}) \cdot \boldsymbol{A} \neq \boldsymbol{A} \cdot (\boldsymbol{A} \cdot \boldsymbol{A}),$$

with both sets containing $\{A^3; A \in \boldsymbol{A}\}$. In order to reduce wrapping effects, higher powers of interval matrices should be computed in a way to minimize the number of matrix multiplications. For example, for $k = 2^s$ we need just s multiplications if we work recursively, $\boldsymbol{S} = \boldsymbol{A}, \boldsymbol{S} \leftarrow \boldsymbol{S}^2$ for $i = 1, \ldots, s$, and if k is not a power of 2, the Paterson–Stockmeyer approach [35] also aims at keeping the number of matrix multiplications small. In a similar spirit, one may aim at minimizing the number of matrix multiplications when evaluating general polynomials of degree d. Paterson and Stockmeyer [35] proved that one always needs at least $\sqrt{d}$ matrix multiplications and presented a method that requires about $2\sqrt{d}$ matrix multiplications [18]. For the special case of polynomials approximating the exponential, approaches which further reduce the number of matrix multiplications were developed in [46, 47, 48]; see section 3.7.

3. Enclosure methods for the matrix exponential.

We start this section with a detailed discussion in the scaling and squaring approach, which turns out to be crucial for the enclosure methods, too. We then proceed by introducing the different enclosure methods, grouping them by the respective approaches they use to approximate the exponential of the scaled matrix and discussing the variants resulting from different ways of performing the interval arithmetic operations involved.

3.1. The scaling and squaring framework.

It is easier to well approximate $\exp(A)$ when $\|A\|$ is small. This is why scaling and squaring is an ingredient of the majority of methods to compute $\exp(A)$. It relies on the simple identity

$$\exp(A) = \left(\exp\left(\frac{A}{2^s}\right)\right)^{2^s}.$$

Higham [17] notes that the main issue in the accuracy of a scaling and squaring method is the significant rounding errors which might occur as a result of severe numerical cancellation in the squaring phase. The fundamental problem can be seen in the result [14, sect. 3.5]

$$\|A^2 - fl(A^2)\|_p \leq \gamma_n \|A\|_p^2,$$

in which $p \in \{1, \infty, F\}$ and $\gamma_n := \frac{nu}{1-nu}$. Here u is the unit roundoff and fl denotes the result obtained in floating point arithmetic. This shows that the errors in the

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

computed squared matrix are small compared with the square of the norm of the original matrix but not necessarily small compared with the matrix being computed. It is therefore important to keep the number s of squaring steps small. The MATLAB function `expm` uses an algorithm from [1] called `expm_new`. While the classical approach chooses s based on $\|A\|$ alone, `expm_new` is an improvement that chooses s by also involving $\|A^k\|^{1/k}$ for modest powers k. Since $\rho(A) \leq \|A^k\|^{1/k} \leq \|A\|$ for $k = 1, \ldots, \infty$ this new approach tends to yield smaller values for s. Keeping s small remains an important research topic; see, e.g., [47] for the case when $\exp(A/2^s)$ is approximated by a polynomial.

From our experiments and the results in [12] it is apparent that enclosure methods for the matrix exponential have to rely on the scaling and squaring technique, too. For two reasons we use the classical scaling and squaring strategy, based solely on $\|A\|$, in our enclosure methods. First, in a guaranteed, interval arithmetic method we also have to involve bounds on the approximation error. The second reason is related to the fact that we consider two variants of each algorithm as explained in the beginning of section 4 below; in the first variant, each algorithm is applied directly to A, while in the second variant we apply our algorithms to a *transformation* $\tilde{\boldsymbol{A}}$ of A. We have observed that in the second variant, the classical and the new scaling strategies compute the *same* scaling factor s, and so the second variant of our algorithms already prevents overscaling.

3.2. Spectral decomposition for diagonalizable matrices. Assume that A is diagonalizable, i.e.,

$$A = VDW,$$

where $VW = I$ and $D = \operatorname{diag}(d_1, \ldots, d_n)$ is diagonal. The exponential of A is then given as

$$\exp(A) = V \exp(D) W.$$

An enclosure method results if we are able to compute interval matrices $\boldsymbol{V}$ and $\boldsymbol{W}$ such that $V \in \boldsymbol{V}, W \in \boldsymbol{W}$ and intervals $\boldsymbol{d}_i$ containing the eigenvalues d_i. We then have

$$\exp(A) \in \boldsymbol{V} \cdot \operatorname{diag}(\exp(\boldsymbol{d}_1), \ldots, \exp(\boldsymbol{d}_n)) \cdot \boldsymbol{W}, \tag{3.1}$$

where we assume that we are able to evaluate the exponential function on intervals in a way that the result is guaranteed to contain its range over that interval. This is possible with the standard function implementations for intervals present in INTLAB or C-XSC, for instance.

The approach outlined here is taken by the `vermatfun.m` routine of the VERSOFT package [37], which, by calling the `verifyeig.m` function of INTLAB [40], uses interval arithmetic to first compute enclosures $(\boldsymbol{v}_i, \boldsymbol{d}_i)$ for all eigenpairs $(v_i, d_i), i = 1, \ldots, n$, and then obtains $\boldsymbol{W}$ by computing an interval matrix $\boldsymbol{W}$ which is guaranteed to contain all solutions $\widetilde{W}$ to all linear systems of the form $\tilde{V}\tilde{W} = I$ for $\tilde{V} \in \boldsymbol{V} := [\boldsymbol{v}_1 \mid \cdots \mid \boldsymbol{v}_n]$. This is done using the INTLAB function `verifylss.m`. Note that `vermatfun.m` is applicable to general matrix functions, not just the exponential.

This approach has two drawbacks. First, since computing an enclosure for just one eigenpair has complexity $\mathcal{O}(n^3)$, its overall complexity is $\mathcal{O}(n^4)$. Second, if the eigenvector matrix is ill-conditioned, $\boldsymbol{W}$ will consist of relatively wide intervals such that the right-hand side of (3.1) will have wide interval entries, too. As illustrated in section 5, the same issue arises in the presence of eigenvalue clusters.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Recently, Miyajima [28] presented an enclosure method which requires an approximate spectral decomposition $A \approx VDV^{-1}$ only, and then constructs an interval matrix $\boldsymbol{M}$ which uses an enclosure $\boldsymbol{S}$ for the residual quantity $V^{-1}(AV - VA)$ obtained using interval arithmetic and additional bounds for other quantities to obtain the enclosure $\exp(A) \in V^{-1}\exp(D)\boldsymbol{M}V$. This algorithm has complexity $\mathcal{O}(n^3)$, and its accuracy crucially depends on the quality of the enclosure $\boldsymbol{S}$, for which evaluating $AV - VA$ in interval arithmetic is not sufficient. We will discuss this somewhat more in section 5. The paper [28] also presents an extension to defective matrices, where the spectral decomposition is replaced by what is called a *numerical Jordan decomposition*. The complexity then increases to $\mathcal{O}(n^4)$.

3.3. Taylor approximations. Since for any matrix $A \in \mathbb{C}^{n\times n}$ we have

$$\exp(A) = \sum_{k=0}^{\infty} \frac{1}{k!}A^k,$$

we can use the first $d+1$ terms of this Taylor expansion to obtain the approximation

$$T_d(A) := I + A + \frac{1}{2!}A^2 + \cdots + \frac{1}{d!}A^d. \tag{3.2}$$

The following results on bounds for the truncation error hold, where the first part is due to Liou [24] and the second to Suzuki; see [16, 51].

THEOREM 3.1. *Let* $\|\cdot\|$ *be the operator* 1-*,* 2-*, or* ∞*-norm and assume* $d+2 > \|A\|$. *Then we have*

$$\|\exp(A) - T_d(A)\| \le \vartheta(d, \|A\|) := \frac{\|A\|^{d+1}}{(d+1)!(1 - \frac{\|A\|}{d+2})}. \tag{3.3}$$

Moreover, $T_{d,s}(A) := \big(T_d(A/s)\big)^s$ *for* $s \in \mathbb{N}$ *satisfies*

$$\|\exp(A) - T_{d,s}(A)\| \le \frac{\|A\|^{d+1}}{s^d(d+1)!}\exp(\|A\|).$$

A consequence of (3.3) is

$$\exp(A) \in T_d(A) + \vartheta(d, \|A\|)\boldsymbol{E}, \tag{3.4}$$

where here and in what follows $\boldsymbol{E}$ denotes the interval matrix with all entries equal to $[-1.1]$.

In Oppenheimer's Ph.D. thesis [33], it was suggested to use the centered form of the truncated Taylor series (3.2) in order to enclose $\exp(A)$. The algorithm, published later in [34], bounds the truncation error by Liou's error bound (3.3). Taylor series are also used in [49] for the accurate computation of the exponential of essentially nonnegative matrices.

Goldsztejn and Neumaier [12] proposed an enclosure method using scaling and squaring based on the truncated Taylor series, the enclosure (3.4), and a variant of Horner's scheme to evaluate $T_d(A)$ in interval arithmetic according to

$$T_d(A) = I + A\bigg(I + \frac{1}{2}A\bigg(I + \cdots + \frac{1}{d-1}A\bigg(I + \frac{1}{d}A\bigg)\cdots\bigg)\bigg). \tag{3.5}$$

We formulate their method as Algorithm 3.1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Algorithm 3.1. Outline of the truncated Taylor series–based enclosure method [12].

```
1: Scale the matrix so that $\|\frac{1}{2^s}A\| \le 0.1$, i.e., $s = \max\{0, \lceil \log_2(10 \cdot \|A\|)\rceil\}$.
2: Determine the smallest integer $d$ such that the truncation error bound from Theorem 3.1 is less than $\varepsilon_{\text{mach}}$. ($d = 9$ in double precision.)
3: Obtain the interval matrix $\boldsymbol{T}_d$ by evaluating $T_d$ for the scaled matrix using interval arithmetic (to account for rounding errors).
4: Use interval arithmetic to compute an upper bound $\overline{\vartheta}$ for $\vartheta(d, \frac{1}{2^s}\|A\|)$ from (3.4) and compute $\boldsymbol{C} = \boldsymbol{T}_d + \overline{\vartheta}\boldsymbol{E}$.
5: Perform $s$ repeated squarings starting with $\boldsymbol{C}$. The final result is an enclosure for $\exp(A)$.
```

In [12], $\|\cdot\|$ is taken to be the ∞-norm and $\boldsymbol{T}_d$ is obtained via Horner's scheme (3.5). The squarings are done in an optimal way according to the function `square` from section 2. The choice for the ∞-norm is in particular motivated by the fact that for this norm one can show that the radii of the computed enclosures decrease monotonically with d, the degree of the truncated Taylor approximation. Interestingly, if the norm of the scaled matrix is less than 0.1, $d = 9$ already achieves $\overline{\vartheta} < \varepsilon_{\text{mach}}$ in double precision. It is also shown in [12] that the Horner scheme (3.5) yields substantially narrower intervals as compared to a "standard" interval arithmetic evaluation of $T_d(A)$ that first computes all powers of A and then their scaled sum.

In an attempt to obtain smaller radii for the computed enclosures, we implemented Algorithm 3.1 with the following two modifications:

$$\left\{\begin{array}{l}\text{replace the } \infty\text{-norm by the 2-norm,}\\ \text{evaluate } T_d(A) \text{ using the Paterson–Stockmeyer approach.}\end{array}\right. \tag{3.6}$$

Using the 2-norm is motivated by the fact that it is smallest possible for Hermitian matrices and that it also tends to be smaller than the ∞-norm for the many other classes of matrices we considered. So using $\|A\|_2$ is likely to require fewer scalings and also to yield a smaller value for $\overline{\vartheta}$ from (3.3). In INTLAB, an interval enclosure, and thus an upper bound, for $\|A\|_2$ is computed with $\mathcal{O}(n^3)$ operations (see [43]) and thus at a cost comparable to the other computations of the algorithm. Using the Paterson–Stockmeyer approach reduces the number of matrix-matrix multiplications and thus the number of wrappings in interval arithmetic. Details on the Paterson–Stockmeyer approach can be found in [35]; here we just give it for the case $d = 9$, where $T_9(A)$ is evaluated according to

$$T_9(A) = I + A + \frac{1}{2!}A^2 + A^3\left(\left(\frac{1}{3!}I + \frac{1}{4!}A + \frac{1}{5!}A^2\right) + A^3\left(\frac{1}{6!}I + \frac{1}{7!}A + \frac{1}{8!}A^2 + \frac{1}{9!}A^3\right)\right),$$

which requires just one squaring (for A^2) and three matrix-matrix multiplications (including one for $A^3 = A \cdot A^2$). Note that evaluation of $T_9(A)$ according to Horner's scheme (3.5) requires nine matrix-matrix multiplications.

3.4. Padé approximation. The type (k, m) Padé approximation to the scalar exponential function is given as

$$\exp(z) = \frac{p_k(z)}{q_m(z)} + r^{km}(z),$$

where $p_k(z)$ and $q_m(z)$ are polynomials of degree k and m, respectively, with $q_m(0) = 1$, and the remainder term $r^{km}(z)$ satisfies $r^{km}(x) = \mathcal{O}(x^{k+m+1})$, i.e., the first nonzero

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

term in the Taylor series of $r^{km}(x)$ is Cx^l for some $l \geq k+m+1$ and a constant C; see [16, p. 79].

Assuming $q_m(A)$ is nonsingular, in the matrix case, the type (k,m) Padé approximation to $\exp(A)$ is thus

$$\exp(A) \approx P_{km}(A) = q_m(A)^{-1} p_k(A),$$

and we have

$$q_m(A) \cdot \exp(A) = p_k(A) + t^{km}(A),$$

where $t^{km}(A) = q_m(A) r^{km}(A)$. The following theorem gives bounds for every entry of the matrix $T^{km} := t^{km}(A)$.

THEOREM 3.2. *Let $\|\cdot\|$ be any submultiplicative matrix norm. Then*

$$\|T^{km}\| \leq \pi(k,m,\|A\|) := \frac{k!\ m!}{(k+m)!\ (k+m+1)!} \|A\|^{k+m+1} \exp(\|A\|), \tag{3.7}$$

and if $\|\cdot\|$ is the 1-, 2-, or ∞-norm, then T^{km} satisfies

$$T^{km} \in \pi(k,m,\|A\|)\boldsymbol{E}.$$

Proof. A classical result from [55] (see also [16, p. 241]) establishes the representation

$$r^{km}(A) = \frac{(-1)^m}{(k+m)!} A^{k+m+1}\ q_m(A)^{-1} \int_0^1 \exp(tA)(1-t)^k t^m dt. \tag{3.8}$$

Multiplying both sides of (3.8) by $q_m(A)$ and using the fact that rational functions of the same matrix commute (see, e.g., [16, Thm. 1.13]), we get

$$T^{km} = q_m(A)\ r_{km}(A) = \frac{(-1)^m}{(k+m)!} A^{k+m+1} \int_0^1 \exp(tA)(1-t)^k t^m dt. \tag{3.9}$$

Since the Taylor series of the exponential has positive coefficients only, we have that $\|\exp(A)\| \leq \exp(\|A\|)$ for any submultiplicative matrix norm. Taking norms in (3.9) thus gives

$$\begin{aligned}
\|T\| &\leq \frac{\|A\|^{k+m+1}}{(k+m)!} \int_0^1 \exp(t\|A\|)(1-t)^k t^m dt \\
&= \frac{\|A\|^{k+m+1}}{(k+m)!} \exp(\theta\|A\|) \int_0^1 (1-t)^k t^m dt, \qquad \theta \in [0,1], \\
&= \frac{\|A\|^{k+m+1}}{(k+m)!} \exp(\theta\|A\|) \frac{k!\ m!}{(k+m+1)!} \\
&\leq \frac{k!\ m!}{(k+m)!\ (k+m+1)!} \|A\|^{k+m+1} \exp(\|A\|) = \pi(k,m,\|A\|).
\end{aligned}$$

The equality in the second line holds by the generalized mean value theorem. We thus have established the first part of the theorem, and its second part holds because the 1-, 2-, and ∞ norms of a matrix are all larger than or equal to the absolute value of any of the matrix entries. ∎

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

An important aspect of Padé approximation in comparison with methods based on Taylor series that work with the original coefficients of the expansion of the exponential function is the efficiency of the Padé approximation; see the discussion in [16], for instance. Also, see section 3.7 for comments regarding recent developments in this area. Since we work with IEEE double precision in all our numerical computations, we choose $m = k = 7$, which gives $\pi \approx 6.06 \times 10^{-16}$ for $\|A\| \le 1$, which is our target value for the scaling phase. The coefficients $b = (b_0, \ldots, b_7)$ of the polynomial $p_7(x) = \sum_{i=0}^{7} b_i x^i$ are the integers

$$b = \begin{bmatrix} 17{,}297{,}280 & 8{,}648{,}640 & 1{,}995{,}840 & 277{,}200 & 25{,}200 & 1{,}512 & 56 & 1 \end{bmatrix};$$

see [16, p. 246]. Moreover, $q_m(x) = p_m(-x)$ for all m. Our implementation represents an interval arithmetic extension of the method outlined in [16, p. 244] to compute the interval matrices $\boldsymbol{P} \ni p_k(A)$ and $\boldsymbol{Q} \ni q_m(A)$. To be specific, we take $\boldsymbol{P} = \boldsymbol{V} + \boldsymbol{U}$ and $\boldsymbol{Q} = \boldsymbol{V} - \boldsymbol{U}$, where

$$\left\{ \begin{array}{l} \boldsymbol{U} := A(b_7\boldsymbol{A}_6 + b_5\boldsymbol{A}_4 + b_3\boldsymbol{A}_2 + b_1 I), \\ \boldsymbol{V} := b_6\boldsymbol{A}_6 + b_4\boldsymbol{A}_4 + b_2\boldsymbol{A}_2 + b_0 \boldsymbol{I}, \end{array} \right. \tag{3.10}$$

and $\boldsymbol{A}_2$ is the (optimal) enclosure for A^2 obtained via the function square from section 2 applied to A, and, similarly, $\boldsymbol{A}_4$ is an enclosure for A^4 computed as the square of $\boldsymbol{A}_2$, while $\boldsymbol{A}_6$ is an enclosure for A^6 computed as $\boldsymbol{A}_2 \cdot \boldsymbol{A}_4$. In this way, $\boldsymbol{P}$ and $\boldsymbol{Q}$ are computed with only two interval matrix-matrix multiplications and two squarings.

Once $\boldsymbol{P}$ and $\boldsymbol{Q}$ are computed, we use Theorem 3.2, which shows that $\exp(A)$ is contained in the solution set of the interval linear system

$$\boldsymbol{Q}X = \boldsymbol{P} + \overline{\pi}\boldsymbol{E}, \tag{3.11}$$

where $\overline{\pi}$ is an upper bound for $\pi(k, m, \|A\|)$ from (3.7) obtained using an interval arithmetic evaluation. We take INTLAB's function `verifylss.m` to compute an interval enclosure for the solution set of (3.11).

Algorithm 3.2 summarizes the enclosure method based on Padé approximation when working in double precision where a $(7, 7)$ Padé approximation is sufficient to bound the approximation error by $\varepsilon_{\text{mach}}$ for matrices with norm ≤ 1.

Algorithm 3.2. Outline of the Padé approximation–based enclosure method.

```
1: Scale the matrix so that $\|\frac{1}{2^s}A\| \le 1$, i.e., $s = \max\{0, \lceil \log_2(\|A\|) \rceil\}$.
2: Compute enclosures $\boldsymbol{P} = \boldsymbol{U} + \boldsymbol{V}$ and $\boldsymbol{Q} = \boldsymbol{U} - \boldsymbol{V}$ for the two polynomials in the
   $(7,7)$ Padé approximation, with $\boldsymbol{U}, \boldsymbol{V}$ computed according to (3.10) applied to
   $2^{-s}A$ (instead of $A$).
3: Compute an upper bound $\overline{\pi}$ for $\pi(7, 7, \frac{1}{2^s}\|A\|)$ via an interval arithmetic evaluation
   of (3.7).
4: Use INTLAB's function verifylss.m to obtain an interval matrix $\boldsymbol{C}$ containing
   the solution set of the interval linear system (3.11).
5: Perform $s$ repeated squarings starting with $\boldsymbol{C}$. The final result is an enclosure for
   $\exp(A)$.
```

For the same reasons as in the Taylor series approach, we chose the norm to be the 2-norm in our computations.

It should be noted that Bochev [5] has already applied a Padé-based algorithm for enclosing $\exp(A)$, which also uses the representation (3.8) of the error. The Padé

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

approach we present here is different in at least two important aspects: Our approach relies directly on Theorem 3.2 and thus does not need to compute a rough enclosure for $\exp(A)$ which is required in [6, 5] in a preliminary step. Moreover, [6, 5] require the computationally expensive staggered correction format [50] to accurately bound the polynomials p_k and q_k and to enclose solutions of the interval linear system (3.11). Staggered correction formats or other (expensive) means to enhance floating point accuracy are not required by our approach.

3.5. Contour integration. For the exponential as for any other analytic function, Cauchy's formula

$$\exp(a) = \frac{1}{2\pi i}\int_\Gamma \frac{\exp(z)}{z-a}dz, \tag{3.12}$$

where Γ is a contour in the complex plane that encloses the point a, carries over to the matrix function case as

$$\exp(A) = \frac{1}{2\pi i}\int_\Gamma \exp(z)(zI-A)^{-1}dz, \tag{3.13}$$

provided the spectrum of A is enclosed by Γ; see [16, Def. 1.11]. We now develop an enclosure method based on quadrature for (3.12) and a rigorous bound for the remainder term. We will scale A such that $\|A\| < 1$ and therefore take the contour Γ to be the unit circle

$$e^{i\theta}, \theta \in [0, 2\pi].$$

Then (3.12) becomes

$$\exp(a) = \int_0^{2\pi} v(\theta)d\theta, \text{ where } v(\theta) = \frac{\exp(e^{i\theta})}{2\pi(e^{i\theta}-a)}e^{i\theta} \ \text{ (with } |a| < 1).$$

The function v is 2π-periodic, which is why the standard trapezoidal rule

$$\exp(a) = \underbrace{\frac{2\pi}{N}\sum_{k=1}^{N} v(\theta_k)}_{:=I_N(v)} + R_N(v)$$

with $N \in \mathbb{N}$ and $\theta_k = 2\pi k/N, k = 1, \ldots, N$, has a small error R_N. Indeed, the following result holds; see [54, Thm. 3.2], [57].

LEMMA 3.3. *Suppose v is 2π-periodic and analytic and satisfies $|v(\theta)| \le M(c)$ in the strip $-c < \Im(\theta) < c$ for some $c > 0$. Then for any $N \ge 1$,*

$$\left|\int_0^{2\pi} v(\theta)d\theta - I_N(v)\right| \le \frac{4\pi M(c)}{e^{cN}-1},$$

and the constant 4π is as small as possible.

In order to use this result for the matrix case, recall that by Banach's lemma (see [9, p. 33], for instance), we have that for any operator norm $\|\cdot\|$

$$|z| > \|A\| \Rightarrow \|(zI-A)^{-1}\| \le \frac{1}{|z| - \|A\|}. \tag{3.14}$$

If $\|\cdot\|$ is the 1-, 2-, or ∞-norm, this implies in particular

$$\left|[(zI-A)^{-1}]_{ij}\right| \le \frac{1}{|z|-\|A\|} \quad \text{for } |z| > \|A\|, \ i,j = 1,\ldots,n. \tag{3.15}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

THEOREM 3.4. *Let $\|A\| < 1$, where $\|\cdot\|$ is the 1-, 2-, or ∞-norm; let c be such that $e^{-c} > \|A\|$; and let $N \in \mathbb{N}$. Let $z_k = e^{2\pi ik/N}, k = 1, \ldots, N$. Then with*

$$\gamma(c, N, \|A\|) := \frac{2e^c \exp(e^c)}{(e^{cN} - 1)(e^{-c} - \|A\|)} \tag{3.16}$$

we have

$$\exp(A) \in \frac{1}{N} \sum_{k=1}^{N} z_k \exp(z_k)(z_k I - A)^{-1} + \gamma(c, N, \|A\|) \cdot \boldsymbol{E}. \tag{3.17}$$

In particular, if $e^{-c} > 2\|A\|$, we have

$$\exp(A) \in \frac{1}{N} \sum_{k=1}^{N} z_k \exp(z_k)(z_k I - A)^{-1} + \gamma(c, N) \cdot \boldsymbol{E}, \tag{3.18}$$

where

$$\gamma(c, N) := \frac{4e^{2c} \exp(e^c)}{e^{cN} - 1}. \tag{3.19}$$

Proof. We first note that (3.18) follows directly from (3.17), since $e^{-c} > 2\|A\|$ implies $\frac{1}{e^{-c} - \|A\|} \leq 2e^c$. With Γ the unit circle $z = e^{i\theta}, \theta \in [0, 2\pi]$, by Cauchy's formula (3.13), each entry of $\exp(A)$ can be expressed as

$$[\exp(A)]_{ij} = \int_0^{2\pi} \underbrace{\frac{1}{2\pi} \exp(e^{i\theta})[(e^{i\theta} I - A)^{-1}]_{ij} e^{i\theta}}_{=: v_{ij}(\theta)} \, d\theta.$$

Herein, by (3.14), v_{ij} is defined and analytic on an open superset of the strip $D^c = -c \leq \Im(\theta) \leq c$ for c with $e^{-c} > \|A\|$, and it is 2π-periodic. Using (3.15), one then obtains

$$\max_{\theta \in D^c} |v_{ij}(\theta)| \leq \frac{\exp(e^c)}{2\pi} \frac{1}{e^{-c} - \|A\|} e^c =: M(c).$$

By Lemma 3.3, we get

$$\left| [\exp(A)]_{ij} - \frac{1}{N} \sum_{k=1}^{N} z_k \exp(z_k) \left[(z_k I - A)^{-1}\right]_{ij} \right| \leq \frac{2M(c)}{e^{cN} - 1},$$

which gives (3.17). $\square$

The enclosure method based on contour integration will have to compute an enclosure for each of the inverses $(z_k I - A)^{-1}$. Using `verifylss.m` to that purpose will for each k give results where each entry has at least a relative width of $\varepsilon_{\text{mach}}$. Since we expect to have a few tens of these systems to solve, it is adequate to require the bound on the quadrature error to be approximately $10\varepsilon_{\text{mach}}$. In order to keep the computational cost low, in our algorithm we therefore choose c such that N is minimal under all pairs (N, c) that satisfy

$$\frac{4e^{2c} \exp(e^c)}{(e^{cN} - 1)} \leq 10\varepsilon_{\text{mach}} \approx 2.2 \times 10^{-15}.$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

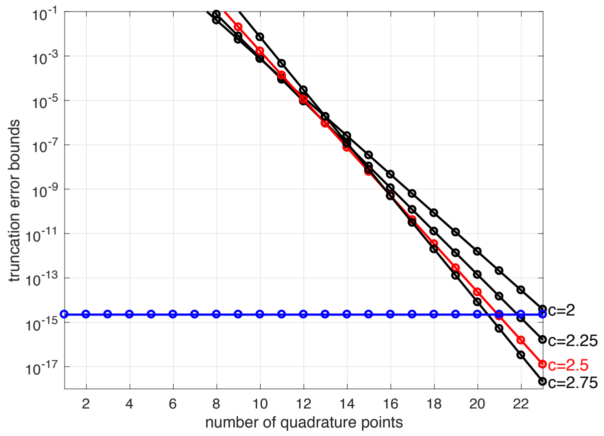

FIG. 1. *Bounds for the quadrature error of the periodic trapezoidal rule. The dash dotted line represents* $10\,\varepsilon_{\rm mach} \approx 2.2 \times 10^{-15}$.

Figure 1 illustrates that this is (approximately) achieved for $c = 2.5$ with $N = 21$, implying that A should be scaled so that $\|\frac{1}{2^s}A\| \le 0.03$. This might seem restrictive, but note that if we relax the scaling to just satisfy $\|\frac{1}{2^s}A\| \le 0.18$, for example, then we would need $c = 1$ and $N = 40$, thus doubling the computational cost. Also note that in our numerical experiments other choices than $c = 2.5, N = 21$ gave comparable if not larger radii for the enclosure obtained for $\exp(A)$.

Algorithm 3.3. Outline of the contour integration–based enclosure method.

```
1: Choose $c = 2.5$ and scale the matrix so that $\|\frac{1}{2^s}A\| \le \frac{1}{2}e^{-c}$, i.e., $s = \max\{0, \lceil\log_2(2e^c\|A\|)\rceil\}$.
2: Put $N = 21$ and compute interval enclosures $\boldsymbol{z}_k$ for the roots of unity $z_k = e^{2\pi ik/N}, k = 1, \dots, N$.
3: Use INTLAB's verifylss.m to compute an interval matrix $\boldsymbol{S}_k$ containing $\{(z_kI - 2^{-s}A)^{-1} : z_k \in \boldsymbol{z}_k\}$, $k = 1, \dots, N$.
4: Compute an upper bound $\overline{\gamma}$ for $\gamma(N, c)$ via an interval arithmetic evaluation of (3.19) to get the enclosure $\boldsymbol{C} = \frac{1}{N}\sum_{k=1}^{N} \boldsymbol{z}_k \exp(\boldsymbol{z}_k)\boldsymbol{S}_k + \overline{\gamma}\boldsymbol{E}$ for the exponential of the scaled matrix.
5: Perform $s$ repeated squarings starting with $\boldsymbol{C}$. The final result is an enclosure for $\exp(A)$.
```

Algorithm 3.3 summarizes our approach based on contour integration. In step 2 we use INTLAB's `verifypoly.m` to obtain as narrow as possible enclosures $\boldsymbol{z}_k$ for the roots z_k of the polynomial $z^N - 1$; see [32, 41] for an overview of relevant techniques. Let us also note that if A is real, we have that $z_kI - A = \overline{(z_{N-k-1}I - A)}$, so $(z_kI - A)^{-1} = \overline{(z_{N-k-1}I - A)^{-1}}$, which can be used to approximately halve the computational cost. Specifically, we then have to invert only 11 interval matrices rather than $N = 21$.

Like the Padé approach, the contour integration approach is based on a rational approximation. In the Padé scheme we have to enclose the solution of the interval linear system (3.11) once, where all entries of the system matrix and of the right-hand side are intervals. In the contour integration approach we have to enclose the inverses

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

of several matrices where only the diagonals contain nonpoint quantities, and we need more squaring steps.

3.6. Chebyshev approximation. The Chebyshev polynomials T_k for the interval $[-1,1]$ are the orthogonal polynomials with respect to the inner product $\langle f,g\rangle = \int_{-1}^{1}\frac{1}{\sqrt{1-x^2}}f(x)g(x)dx$ on the space of continuous functions on $[-1,1]$. They satisfy $T_k(x)=\cos(k\arccos x)$ for $x\in[-1,1]$ and obey the recurrence

$$\left\{\begin{array}{l} T_0=1,\ T_1=x, \\ T_k=2xT_{k-1}-T_{k-2},\ k=2,3,\dots, \end{array}\right. \tag{3.20}$$

and the (formal) Chebyshev series of a Lipschitz continuous function f on $[-1-1]$ is given by

$$\sum_{k=0}^{\infty}\frac{\langle f,T_k\rangle}{\langle T_k,T_k\rangle}T_k.$$

For $f=\exp$ the coefficients of the Chebyshev series are given via the modified Bessel functions of the first kind, $I_k(t)=\frac{1}{\pi}\int_0^{\pi}\exp(t\cos(\theta))\cos(k\theta)\,d\theta$, as

$$\frac{\langle \exp,T_0\rangle}{\langle T_0,T_0\rangle}=I_0(1),\quad \frac{\langle \exp,T_k\rangle}{\langle T_k,T_k\rangle}=2I_k(1),\quad k=1,2,\dots;$$

see, e.g., [25, p. 109] and [52, p. 23].

The use of Chebyshev series for approximating $\exp(A)b$, where A is Hermitian and b is a vector, was suggested by Druskin and Knizhnerman [11]. In [3], $\exp(A)$ was computed for sparse A using the degree 17 truncated Chebyshev series. Its advantage is that, typically, for the same degree, the polynomial approximation given by the truncated Chebyshev series will give a more accurate approximation than a Taylor polynomial if one considers the whole interval $[-1,1]$. For interval arithmetic–based enclosure methods this means that we can scale less and thus save some squaring steps and the associated wrappings as compared to the Taylor approach. This motivates our investigation of Chebyshev approximation in this work. Its use is restricted to Hermitian matrices A, however, because error bounds are difficult to get for general A. To state an enclosure result for the Hermitian case, recall that the Bernstein ellipse E_ρ is an ellipse with center at zero and foci at ±1 whose parameter $\rho>1$ is the sum of its semiaxis lengths. The following result can be found in, e.g., [52, Thm. 8.2].

LEMMA 3.5. *Let f be analytic in $[-1,+1]$ and analytically continuable to the open Bernstein ellipse E_ρ, where it satisfies $|f(z)|\le M(\rho)$. Then, for each $d\ge 0$, the truncated Chebyshev series $p_d=\sum_{k=0}^{d}\frac{\langle f,T_k\rangle}{\langle T_k,T_k\rangle}T_k$ satisfies*

$$|f(x)-p_d(x)|\le\frac{2M(\rho)\rho^{-d}}{\rho-1}\ \textit{for}\ x\in[-1,1].$$

THEOREM 3.6. *Let A be Hermitian with spectrum in $[-1,1]$, and let $p_d(A)$ be the degree d truncated Chebyshev series approximation*

$$p_d(A)=I_0(1)I+\sum_{k=1}^{d}2I_k(1)\cdot T_k(A) \tag{3.21}$$

for $\exp(A)$. Then, with $\tau(\rho,d)$ defined for $\rho\ge1$ as

$$\tau(\rho,d):=2e^{\frac{\rho+\rho^{-1}}{2}}\frac{\rho^{-d}}{\rho-1}, \tag{3.22}$$

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

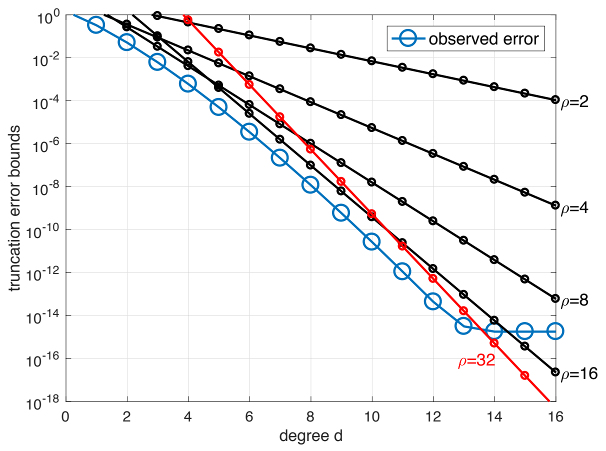

FIG. 2. *Truncation error when chopping the Chebyshev series. Observed error is the ∞-norm of* $\exp - p_d$ *on* $[-1, 1]$.

we have

$$\exp(A) \in p_d(A) + \tau(\rho, d)\boldsymbol{E}.$$

Proof. Since A is Hermitian we have $A = VDV^{-1}$, where V is orthonormal and $D = \operatorname{diag}(\lambda_i)$ is the diagonal matrix containing the eigenvalues. Also, $T_k(A) = VT_k(D)V^{-1}$ for all k, and $p_d(A) = Vp_d(D)V^{-1}$. Let $R = (r_{ij}) := \exp(A) - p_d(A)$. We thus have

$$\begin{aligned} |r_{ij}| \le \|R\|_2 &= \|V(\exp(D) - p_d(D))V^{-1}\|_2 \\ &= \|\exp(D) - p_d(D)\|_2 = \|\exp(D) - p_d(D)\|_\infty \\ &= \max_i |\exp(\lambda_i) - p_d(\lambda_i)|. \end{aligned}$$

The fact that the maximum value of the exponential on E_ρ is $e^{\frac{\rho+\rho^{-1}}{2}}$ (see [53]), together with Lemma 3.5, now completes the proof. ∎

In an algorithm, we want to choose ρ and d such that $\tau(\rho) \approx \varepsilon_{\text{mach}}$ for d as small as possible. Figure 2 reports an experiment using the Chebfun package [10]. It depicts the error bound $\tau(\rho, d)$ in (3.22) for different values of d and ρ together with the truncation error $\|\exp - p_d\|_\infty = \max_{-1\le x\le +1} |\exp(x) - p_d(x)|$. In Chebfun, it is particularly easy to compute the latter quantity if we replace exp with its adaptively computed Chebfun representation and then use the `chebfun/norm` command. The figure suggests that $d = 14$ (and $\rho = 32$) is an appropriate choice when working in double precision.

Also, in an interval arithmetic–based enclosure method we need to use intervals enclosing the exact values $I_k(1)$ of the Bessel functions, and these should be as narrow as possible. We used the Arb package [20], a C library for arbitrary-precision interval arithmetic, to obtain enclosures $\boldsymbol{I}_k$ of relative radii of about $\varepsilon_{\text{mach}}$ in double precision for the exact values of $I_k(1)$ for $k = 0, 1, \ldots, 14$. Note that those must be computed only once and can then be stored for later use.

Before summarizing our approach in Algorithm 3.4, the last algorithm of this section, we discuss how we evaluate the truncated Chebyshev sum $p_d(A)$. The direct way to evaluate $p_d(A)$ would be to precompute $T_k(A)$ for $k = 0, \ldots, d$ using the recurrence (3.20) and to subsequently evaluate the sum $I_0(1)I + \sum_{k=1}^{d} 2I_k(1)T_k(A)$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

For the scalar case, and a general Chebyshev approximation $p_d(x) = \sum_{k=0}^{d} a_k T_k(x)$, it is known that the use of the *Clenshaw recurrence*, which interleaves the evaluation of the Chebyshev polynomials with the summation,

$$\begin{cases} b_{d+2} = b_{d+1} = 0, \\ b_j = 2xb_{j+1} - b_{j+2} + a_j, \quad j = d, d-1, \ldots, 0, \end{cases}$$

and then $p_d(x) = b_0 - xb_1$, is more stable; see [36, p. 173], for instance. Both the direct method and the Clenshaw recurrence rely on three-term recurrences. When applied to matrices and using interval arithmetic, we thus experience not only the wrapping effect but also the fact that interval arithmetic treats the same variable occurring several times as being independent variables. This induces a tendency to further increase the width of the computed intervals. Analogous phenomena have been observed when enclosing scalar Chebyshev expansions of high degree [13]. To alleviate this problem we suggest using the matrix analogue of the product formulae

$$\begin{cases} T_{2k}(x) = 2T_k^2(x) - 1, \\ T_{2k+1}(x) = 2T_{k+1}(x)T_k(x) - x, \end{cases} \quad k = 1, 2, \ldots, \tag{3.23}$$

with $T_0 = 1, T_1 = x$ to evaluate $T_k(A)$ using interval arithmetic. Therein, the squares can be computed using the method from section 2. This approach makes the total number of multiplications, and thus of the associated wrappings, small. For example, $T_8(A)$ is computed with just 3 squarings (for T_8, T_4, and T_2), and T_{14} is computed with 3 squarings (for T_{14}, T_4, and T_2) and 2 matrix multiplications (for T_7 and T_3).

Algorithm 3.4. Outline of the Chebyshev-based enclosure algorithm (A Hermitian).

```
1: Scale the matrix so that $\|\frac{1}{2^s}A\| \leq 1$, i.e., $s = \max\{0, \lceil \log_2(\|A\|) \rceil\}$.
2: Use interval arithmetic to compute enclosures for $T_k(2^{-s}A)$ for $k = 0, \ldots, 14$
   via (3.23) and to then subsequently evaluate $\boldsymbol{S} = \boldsymbol{I}_0 + \sum_{k=1}^{14} 2\boldsymbol{I}_k \boldsymbol{T}_k(2^{-s}A)$, an
   enclosure for the value of the truncated Chebyshev series for the scaled matrix.
3: Compute an upper bound $\overline{\tau}$ for $\tau(32, 14)$ via an interval arithmetic evaluation of
   (3.22) to get the enclosure $\boldsymbol{C} = \boldsymbol{S} + \overline{\tau}\boldsymbol{E}$ for the exponential of the scaled matrix.
4: Perform $s$ repeated squarings starting with $\boldsymbol{C}$. The final result is an enclosure for
   $\exp(A)$.
```

3.7. Further polynomial approximation techniques. Whether we take a Taylor, a Chebyshev, or yet another polynomial approximation p to the matrix exponential, a crucial issue for interval arithmetic evaluations is always to reduce the number of matrix multiplications. We are grateful to an anonymous referee for pointing us to new results going beyond Paterson–Stockmeyer which were recently obtained in [47] for the Taylor approximation. There, a degree-16 Taylor-based floating-point method is proposed together with a technique in which only four matrix products are required. This involves a preprocessing stage which computes new intermediate polynomials from the coefficients of the Taylor approximation, through which then the evaluation of $p(A)$ is be obtained. The major benefit is that this allows the number of scalings to be less: We can make $\vartheta(16, \|2^{-s}A\|)$ in (3.3) as small as $\varepsilon_{\text{mach}}$ double precision with $\|2^{-s}A\| \leq 0.8$. This means that such a technique would involve three fewer scaling and squaring steps compared with `TayH` and `TayPS`. There is a similar approach for the degree-24 Taylor polynomial requiring only five matrix multiplications and which allows scaling such that $\|2^{-s}A\| \leq 1.68$.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

While these approaches are very appealing when working in floating point arithmetic, they present additional challenges when used within our general framework of bounding the forward error using interval arithmetic: The coefficients of the new polynomials are given as solutions of nonlinear systems of equations, so that in our context we would have to compute guaranteed enclosures for the solutions of these systems. The systems are quite involved, and they are, for example, not explicitly reproduced in [47], where computer algebra tools were used to obtain the final floating point approximations. While interval arithmetic provides, in principle, efficient tools to enclose solutions of (explicitly given) nonlinear systems as implemented in `verifynlss` in INTLAB, for example, the computed enclosures will necessarily be represented by nonpoint intervals. This introduces a new source for producing larger diameters in an interval arithmetic computation.

On the other hand, an alternative possibility for enclosing $p_d(A)$ in (3.21) could be the following. One may first rewrite $p_d(A)$ in the monomial basis, and then take advantage of efficient matrix polynomial evaluation techniques explained in [46]. The coefficients of the monomial expansion could be found via a technique similar to [48] where the matrix exponential is approximated in the basis of Hermite polynomials. Note that accurate approximations are not enough in our context; such a process should include sharp computable bounds on all possible errors.

Coping with the described challenges would go beyond the scope of the present work, so that we postpone an investigation along these interesting lines to the near future.

4. Approximate diagonalization.

If V is nonsingular and

$$D = V^{-1}AV \Longleftrightarrow A = VDV^{-1},$$

we have, in exact arithmetic, $\exp(A) = V\exp(D)V^{-1}$. In floating point arithmetic, if we compute an enclosure $\boldsymbol{W}$ for V^{-1} and then $\boldsymbol{D}$ as $\boldsymbol{D} = \boldsymbol{W}AV$ using interval arithmetic, we have

$$\exp(A) \in V\boldsymbol{F}\boldsymbol{W}, \tag{4.1}$$

where $\boldsymbol{F}$ is an enclosure for $\{\exp(D) : D \in \boldsymbol{D}\}$, and to compute $\boldsymbol{F}$, we can rely on any of the techniques presented in the previous section, replacing A by the interval matrix $\boldsymbol{D}$. This is a consequence of two facts: First, our algorithms take both rounding and truncation errors into account for any given *point* matrix D and therefore, as indicated by Theorems 3.1, 3.2, 3.4, and 3.6, our computed enclosures indeed contain the exact $\exp(D)$. Second, the fundamental enclosure property of interval arithmetic guarantees this remains correct for *every* point matrix D in a given *interval* matrix $\boldsymbol{D}$, and therefore $\{\exp(D) : D \in \boldsymbol{D}\} \subseteq \boldsymbol{F}$.

This observation, already pointed out in [12], can be used in an attempt to reduce the wrapping effect. Indeed, if $\boldsymbol{D}$ were diagonal, there would be no wrapping effect at all when computing powers of $\boldsymbol{D}$, and when the off-diagonal elements of $\boldsymbol{D}$ are small compared with those on the diagonal, the wrapping effect is also small. The price we pay is additional wrappings due to the multiplications with V and $\boldsymbol{W}$, and here the wrapping effect becomes large when V is ill-conditioned.

In our numerical examples we used two variants of this *transformation approach.* The first takes V as a computed approximation of the eigenvector matrix if we can expect A to be diagonalizable and V to have small condition number, e.g., when A is Hermitian. The second uses the MATLAB routine `bdschur` from the Control

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Table 1
Acronyms used in figures. `Cheb` *is the only method that is not applied to non-Hermitian matrices.*

Acronym	Corresponding enclosure method
`TayH`	Taylor–Horner: Alg. 3.1, Horner's scheme to evaluate the polynomial [12]
`TayPS`	Taylor–Paterson–Stockmeyer: Alg. 3.1 with the modifications from (3.6)
`Padé`	Alg. 3.2
`Cont`	Contour integration: Alg. 3.3
`Cheb`	Chebyshev: Alg. 3.4.
`SpecDec`	Miyajima's method relying on spectral decomposition [28]
`PadéBM`	Padé-based method of Bochev and Markov [6], $q = 7$ as implemented in [27]
`VER`	VERSOFT's routine `vermatfun.m` [37]
`Acronym-ad` is method `Acronym` using approximate diagonalization. Only the first five methods have an `-ad` variant; see section 4.	

System Toolbox, which, for a general matrix A, produces a *block diagonal* matrix D and a well-conditioned matrix V, computed in floating point arithmetic, such that $A \approx VDV^{-1}$; see [4]. In either case we use `verifylss.m` to compute an interval enclosure $\boldsymbol{W}$ for V^{-1}. Note that the matrix $\boldsymbol{D} = \boldsymbol{W}AV$ will in general have small nonzero entries outside its (block) diagonal.

In principle, one can consider other similarity transformations. For example, one could think of using the Schur decomposition which orthogonally transforms A to triangular form. In floating point arithmetic, exponentiation of a triangular matrix is done via the Parlett recurrence [8]. However, due to the overestimation caused by the dependency issue pointed out in section 2, we do not expect a straightforward interval arithmetic extension of the recurrence to be useful, unless the matrix size is very small.

To avoid confusion, let us end by noting that what we call approximate diagonalization here is unrelated to the approximate diagonalization considered in [7].

5. Numerical examples.

We compare the performance of the various algorithms for different classes of matrices with dimensions ranging from $n = 50$ to $n = 600$. They were chosen partly to represent a wide range of properties which potentially affect the quality of the computed enclosures and partly to reproduce results for matrices considered in the existing literature. Table 1 lists all the methods together with their acronyms used in the figures to come. For the methods `SpecDec` and `PadéBM` we use the MATLAB-INTLAB implementations of Miyajima [27].

We report two quantities for all methods. The first is the *average relative precision* (arp) of $\boldsymbol{X}$ defined by

$$\text{arp}(\boldsymbol{X}) := \Big(\prod_{i,j=1}^{n} (\text{rp}(\boldsymbol{X}_{ij})) \Big)^{1/n^2}, \tag{5.1}$$

where

$$\text{rp}(\boldsymbol{x}) := \min(\texttt{relerr}(\boldsymbol{x}), 1)$$

is the *relative precision* of an interval $\boldsymbol{x}$ with `relerr` defined as

$$\texttt{relerr}(\boldsymbol{x}) = \begin{cases} \frac{\text{rad}\,(\boldsymbol{x})}{|\text{mid}\,(\boldsymbol{x})|}, & 0 \notin \boldsymbol{x}, \\ \text{rad}\,(\boldsymbol{x}), & 0 \in \boldsymbol{x}, \end{cases}$$

as an indicator of the quality of the computed enclosures. Roughly speaking, the quantity $-\log_{10}(\text{arp}(\boldsymbol{X}))$ represents the average number of known correct digits of $\boldsymbol{X}$. For the largest matrix in every example, we include a table recording the average and the standard deviation of the number of known correct digits. The second

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

reported quantity is wall clock time (in seconds), averaged over three executions, as an indicator for the efficiency of the method. Note that we undertook quite some efforts in our implementations to obtain good (interval) arithmetic performance, using built-in INTLAB functions systematically and casting operations as matrix operations whenever possible.

All numerical results were obtained using INTLAB Version 11 and MATLAB R2017a on a Mac OS X with 2.5 GHz Intel Core i7 processor and 16 GB of RAM. The random number generator mode is always fixed by the command `rng(1,'twister')` for all the tests involving matrices with random entries. Most of our examples are from MATLAB's gallery of test matrices, accessible via `gallery.m`.

5.1. Non-Hermitian matrices. We compare the performance of all eleven methods from Table 1 which are applicable to nonsymmetric matrices.

Example 1. A is the $n \times n$ Helmert matrix, which is a permutation of a lower Hessenberg matrix, whose first row is `ones(1:n)/sqrt(n)`. It is in MATLAB's gallery as the `orthog` matrix of type 4.

The results are depicted in Figure 3. The narrowest enclosures are obtained by `TayPS`, and the second most accurate results are obtained by `Padé`. `TayPS` is not only the most accurate but also among the fastest. Methods applied to the original matrix are more accurate than the corresponding method applied to the approximately diagonalized matrix, and the difference is more pronounced the larger the dimension. The reason `TayH-ad` is faster than `TayH` is that the scaling parameter in `TayH` is $s = 8$, and that is reduced to $s = 4$ in `TayH-ad`. While `SpecDec` is ranked third with respect to accuracy, it is up to almost 700 times slower than `TayPS`. This drawback of `SpecDec` and `VER` as well as, to a lesser extent, `PadéBM` will be visible in all other experiments, just as the fact that `VER` typically yields the poorest enclosures. We will not repeat this observation explicitly for the other examples. Note that `SpecDec` has to use INTLAB's accurate dot product `AccDot.m` to obtain sufficiently narrow enclosures for

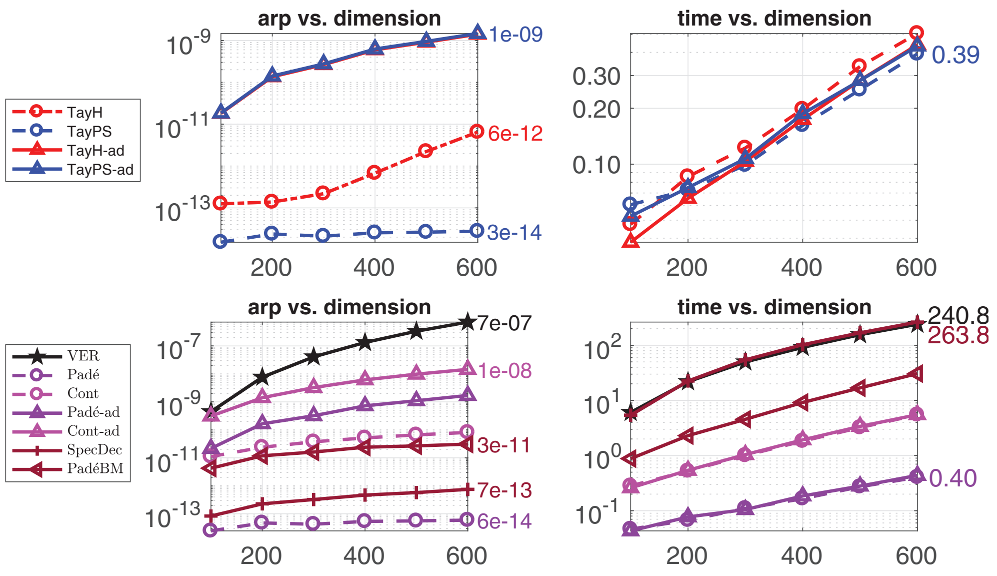

FIG. 3. *Average relative precision versus dimension (left) and time versus dimension (right) for the Helmert matrix, Example* 1.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

TABLE 2

Average number of correct digits and standard deviation for each method for the largest matrix of Example 1.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	11.2	13.6	8.8	8.8	6.1	13.2	10.1	8.8	7.8	12.1	10.5
Std. dev.	0.38	0.26	0.53	0.53	0.62	0.20	0.39	0.53	0.53	0.52	0.20

the entries of some matrix-matrix product, an approach which is not recommended in [45] because of the interpretation overhead. For `VER`, the long execution times result from its complexity being $\mathcal{O}(n^4)$ rather than, as in all the other methods, $\mathcal{O}(n^3)$. The bad timings for `PadéBM`, finally, result from the method requiring an "exponent safe bound evaluation" step that, as implemented in [27], has to enclose solutions for up to $\sqrt{n}$ linear systems with a matrix right-hand side using `verifylss.m`. Finally, we note that `Cont`, the contour integral approach, is about one order of magnitude slower than the best performing method, and that its accuracy is substantially less than that of `TayPS`, `TayH`, and `Padé`.

For the Helmert matrix of size $n = 600$, Table 2 reports $-\log_{10} \text{arp}(\boldsymbol{X})$, the average of the number of known correct digits, together with its standard deviation.

Example 2. A is the `forsythe` matrix [56] from MATLAB's gallery. A consists of one single $n \times n$ Jordan block with eigenvalue zero except that its $(n, 1)$ entry is equal to $\sqrt{\varepsilon_{\text{mach}}}$.

This time, `Padé` gives the narrowest enclosures. The second most accurate results are obtained via `TayH` and `TayPS`, which perform very similarly and are not easy to distinguish in the middle of Figure 4 (left). `Padé` gives about one more digit of accuracy compared with both `TayH` and `TayPS`; see also Table 3. Like Example 1, the methods applied to A generally give enclosures that are narrower than those obtained when applied to the approximately diagonalized matrix $\boldsymbol{D}$. The fastest method is `TayH`, but `TayPS` and `Padé` are comparable with respect to speed.

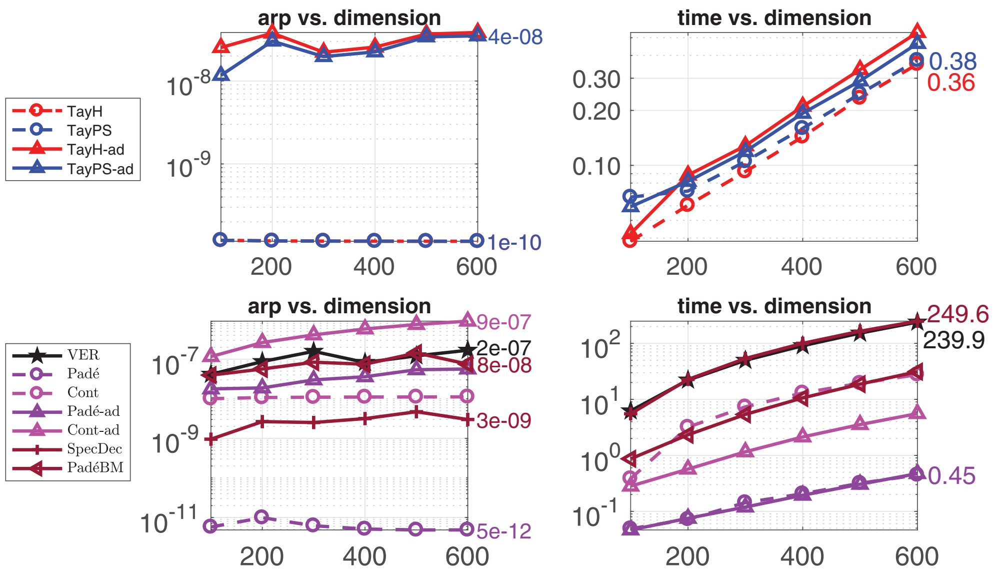

FIG. 4. *Average relative precision versus dimension (left) and time versus dimension (right) for the* `forsythe` *matrix, Example* 2.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

TABLE 3
Average number of correct digits and standard deviation for the largest matrix of Example 2.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	9.9	9.9	7.4	7.5	6.8	11.3	8.0	7.2	6.0	8.5	7.1
Std. dev.	4.36	4.36	3.59	3.59	3.28	4.65	3.89	3.59	3.23	4.00	3.89

Example 3. A is the `lesp` matrix from MATLAB's gallery. It has the property that the condition of its eigenvalues increases exponentially with the dimension n.

Because of the ill-conditioned eigenvalues, `SpecDec` and `VER` fail, returning NaNs for $n > 50$ and $n > 100$, respectively. Figure 5 and Table 4 show that the most accurate enclosures are obtained either with `Padé` or `PadéBM`, obtaining one to two more digits of accuracy compared with `TayH` and `TayPS`, which are the next most accurate approaches. `Padé-ad` loses about one digit in accuracy compared with `Padé`. The computing times of `Padé`, `TayH`, and `TayPS` are of the same order.

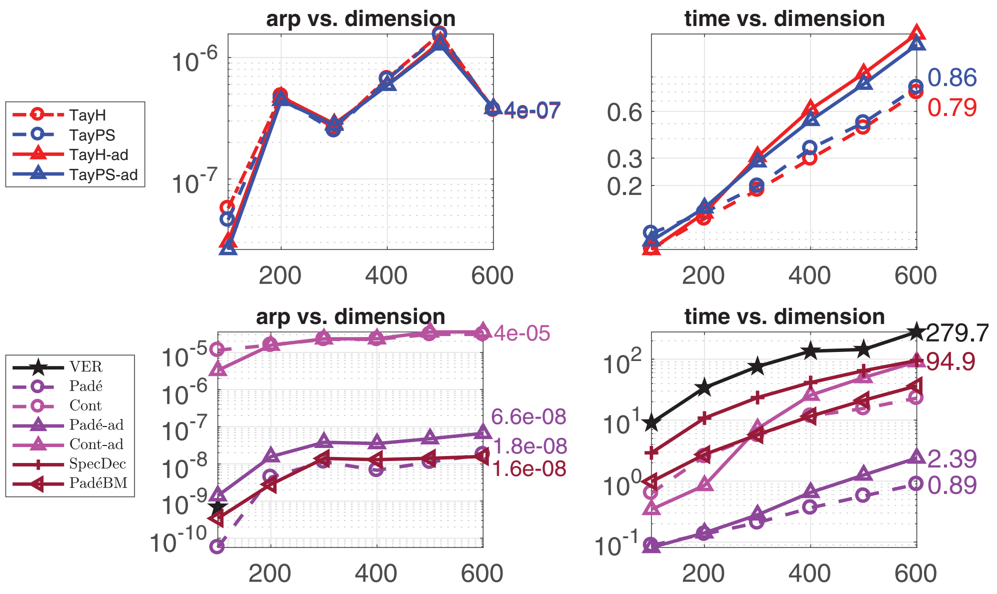

FIG. 5. *Average relative precision versus dimension (left) and time versus dimension (right) for the `lesp` matrix, Example 3.*

TABLE 4
Average number of correct digits and standard deviation for each method for the largest matrix of Example 3.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	6.4	6.4	6.4	6.4	NaN	7.7	4.5	7.2	4.4	NaN	7.8
Std. dev.	3.64	3.62	3.51	3.51	NaN	3.85	2.66	3.47	2.44	NaN	2.76

Example 4. We take $n \times n$ matrices A of the form $A := WDW^{-1}$, where W is a matrix with normally distributed random entries and D is the diagonal matrix with its diagonal entries taken as n equidistant points in the interval $[-1, 1]$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Much to the opposite of Example 3, the matrices of this example fit particularly well the `SpecDec` approach. Figure 6 and Table 5 show that this is indeed the most accurate algorithm. The much faster approaches, `TayPS-ad`, `Padé-ad`, and `Cont-ad`, are the second most accurate, but their accuracy is significantly lower (up to 7 decimal digits). Approaches without approximate diagonalization yield very poor accuracy. The reason is that W has a high condition number, so that $\|A\|_\infty$ and $\|A\|_2$ are large. This implies that the algorithms perform quite a few scaling steps ($s = 14, \ldots, 16$ for $n = 400$, for instance), whereas with approximate diagonalization this goes down to $s = 1, \ldots, 4$. Thus approximate diagonalization allows us to save a significant number of squaring steps and thus reduces the otherwise predominant wrapping effect.

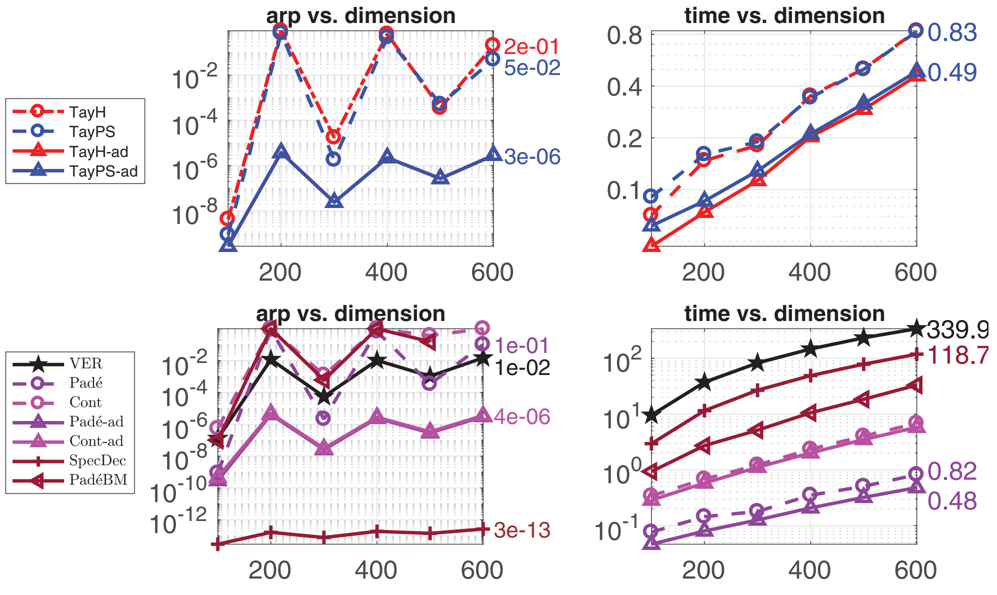

FIG. 6. *Average relative precision versus dimension (left) and time versus dimension (right) for the random diagonalizable matrix whose eigenvalues are equispaced points on* $[-1, 1]$. *See Example* 4.

TABLE 5

Average number of correct digits and standard deviation for each method for the largest matrix of Example 4.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	0.7	1.3	5.5	5.5	1.8	1.0	0.0	5.5	5.4	12.6	NaN
Std. dev.	0.25	0.33	0.47	0.47	0.44	0.30	NaN	0.47	0.47	0.47	NaN

Example 5. A is the `triw` matrix from MATLAB's gallery. A is upper triangular and ill-conditioned with respect to both inversion and eigenvalue computation.

Here, `Padé` gives the narrowest enclosures, and it is one of the fastest methods as well; see Figure 7. The quality of enclosures computed via approximate diagonalization is the same as that obtained when applied to the original matrix; see Table 6. `SpecDec` fails, returning NaNs, for all sizes n due to the ill-conditioning, and similarly for `VER` for $n = 100, \ldots, 400$. Since `VER` already takes more than 19 minutes for $n = 400$, we did not run it for $n = 500$ and $n = 600$. `VER` is therefore not reported in Figure 7, while we kept the run times for `SpecDec`.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

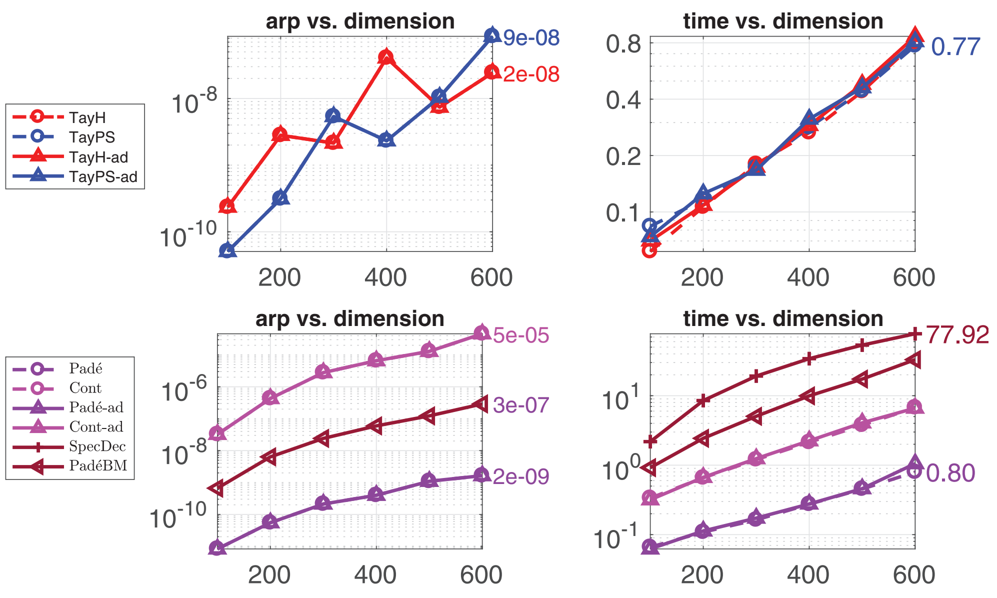

FIG. 7. *Average relative precision versus dimension (left) and time versus dimension (right) for the* `triw` *matrices, Example* 5.

TABLE 6
Average number of correct digits and standard deviation for each method for the largest matrix of Example 5.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	7.6	7.1	7.6	7.1	-	8.8	4.3	8.8	4.3	NaN	6.5
Std. dev.	0.78	0.72	0.78	0.72	-	1.02	0.72	1.02	0.72	NaN	1.09

Example 6. We take the point analogue of the matrices considered in [12] and define $A_n \in \mathbb{R}^{3n\times 3n}$ as the $3n\times 3n$ block diagonal matrix which for each $k = 1,\dots,n$ has one diagonal block of size 1 with entry $2k+1$ and one diagonal block of size 2 with entries $2k\cdot\left[\begin{smallmatrix}1 & -1\\ 1 & 1\end{smallmatrix}\right]$ and eigenvalues $2k(1\pm i)$. Then $A = P^{-1}A_nP$, where P is a random orthogonal matrix, obtained as the Q-factor in the QR decomposition of a random $3n\times 3n$ matrix with normally distributed entries.

The numerical results in Figure 8 and Table 7 show that the most accurate method is `Spec-Dec`. The `-ad` variants of Taylor-type algorithms, which are second with respect to accuracy, are (among) the fastest methods. This is quite an extreme example for larger dimensions n, since the moduli of the eigenvalues of $\exp(A)$ range from e to e^{2n}.

5.2. Symmetric matrices. If A is symmetric, then so is $\exp(A)$. In our algorithms, whenever we know that a point matrix for which we compute an enclosure is symmetric we can thus "symmetrize," and at the same time narrow, this enclosure by replacing it with the intersection with its adjoint. We do this whenever possible in our implementations involving symmetric matrices.

The tested methods for symmetric matrices now also include those based on the truncated Chebyshev series.

Example 7. A is the Hankel matrix `ris` from MATLAB's gallery. A is a normal matrix whose eigenvalues cluster around $-\pi/2$ and $\pi/2$.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

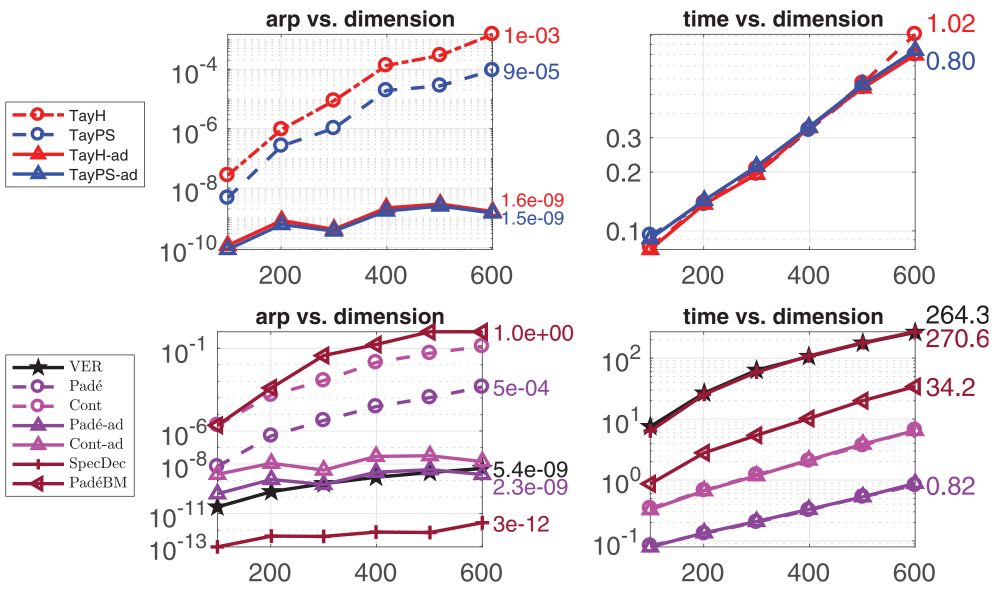

Fig. 8. *Average relative precision (left) and time (right) for the (point) matrix from* [12], *Example* 6.

Table 7

Average number of correct digits and standard deviation for each method for the largest matrix of Example 6.

	`TayH`	`TayPS`	`TayH-ad`	`TayPS-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	2.8	4.0	8.8	8.8	8.3	3.3	0.9	8.6	7.9	11.5	0
Std. dev.	0.46	0.46	0.46	0.46	0.49	0.46	0.33	0.46	0.45	0.45	NaN

Figure 9 shows that while the narrowest enclosures are obtained by `Cheb`, the fastest method is usually `Padé`, with the Taylor algorithms not far behind, `Cheb` being less than a factor of 2 off. See also Table 8. Even though A is symmetric, `VER` gives wide enclosures because, due to the clustering of the eigenvalues, the computed enclosures for the eigenvalues used in `VER` are already wide; see also Example 10. As a further illustration, for $n = 400$ we report the relative errors of all entries of the computed enclosing interval matrix for $\exp(A)$. This is depicted for six different methods in the left part of Figure 10, where the ordinate represents the 160,000 entries in ascending order of their relative errors. The figure shows that most of the entries have similar relative error and only a few have substantially larger or smaller radius. This is a quite typical situation, thus justifying that "arp" is indeed a good way of measuring the quality of enclosures; see (5.1).

The right part of Figure 10 gives a histogram reporting a comparison of the number of "known correct digits" of the floating point approximation $C = \texttt{expm}(A)$ obtained using MATLAB's `expm` function and of the midpoint $\operatorname{mid}\boldsymbol{C}$ of the interval matrix $\boldsymbol{C}$ obtained with `Cheb`. For an entry $\operatorname{mid}\boldsymbol{C}_{ij}$, the number of its known correct digits is the number to which the upper and lower bounds coincide. Similarly, the number of known correct digits of an entry C_{ij} is the number of digits which coincide with those of both the lower and the upper bound of $\boldsymbol{C}_{ij}$. If an entry C_{ij} of C is not contained in $\boldsymbol{C}_{ij}$, its number of correct digits is guaranteed to be smaller than or at most equal to that of $\operatorname{mid}\boldsymbol{C}_{ij}$, and for these cases we report the difference between

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

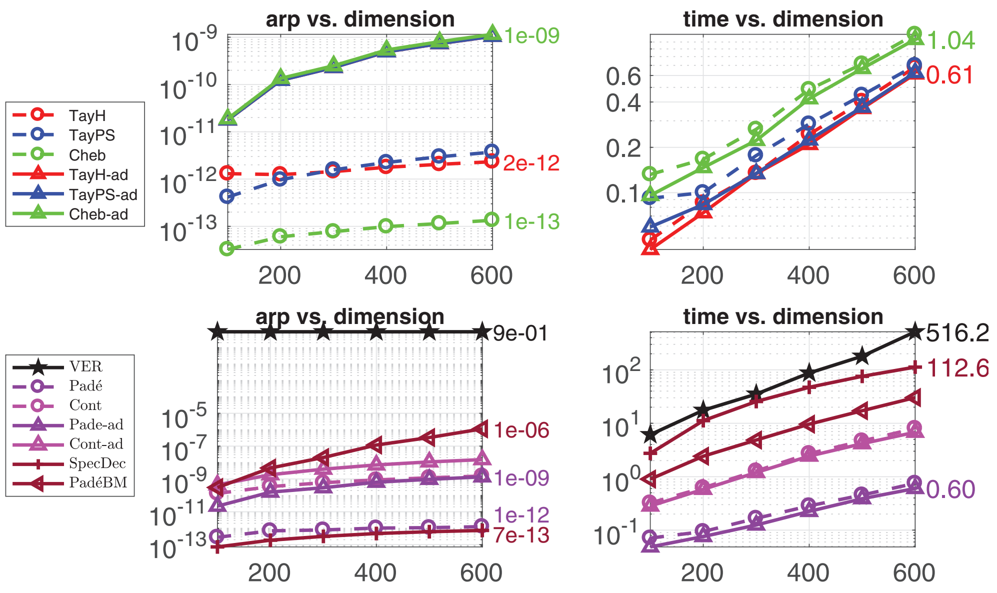

FIG. 9. *Average relative precision versus dimension (left) and time versus dimension (right) for the symmetric matrix* `ris`*. See Example 7.*

TABLE 8

Average number of correct digits and standard deviation for each method for the largest matrix of Example 7.

	TayH	TayPS	Cheb	TayH-ad	TayPS-ad	Cheb-ad	VER	Padé	Cont	Padé-ad	Cont-ad	SpecDec	PadéBM
Average	11.6	11.4	12.9	9.0	9.0	9.0	0.04	11.9	8.8	8.9	7.8	12.1	5.9
Std. dev.	0.52	0.59	0.54	0.62	0.62	0.62	0.00	0.52	0.61	0.62	0.63	0.61	0.60

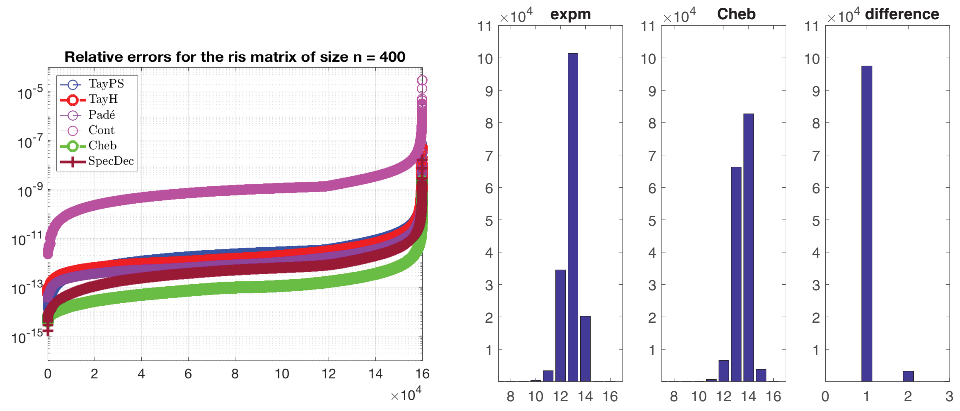

FIG. 10. *Relative errors of the* $160,000$ *entries of the computed enclosure for the symmetric matrix* `ris` *of size* 400×400 *for six different approaches (left), and histogram of known correct digits in MATLAB's* `expm`*, in midpoint of the interval matrix computed via* `Cheb`*, and increase in correct digits obtained by* `Cheb` *(right). See Example 7.*

the number of correct digits of mid $\boldsymbol{C}_{ij}$ and those of C_{ij} in the rightmost histogram. For this example, this difference is 1 or 2 for about 60% of the entries. So not only do the results obtained with the enclosure method give intervals which are guaranteed to

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

contain the exact values, but their midpoints are also (slightly) more accurate than the values obtained with `expm`.

Example 8. A is the symmetric `orthog` matrix of type 2 from MATLAB's gallery.

Figure 11 shows that for this example `Cheb` appears to be the best method. Even though `Cheb` takes almost twice the time needed for `Padé`, `TayH`, and `TayPS`, its accuracy is second best, only marginally lower than that of `SpecDec`, and slightly better than `Padé`; see Table 9. We also observe a better quality of the enclosures computed by `TayPS` compared with `TayH`.

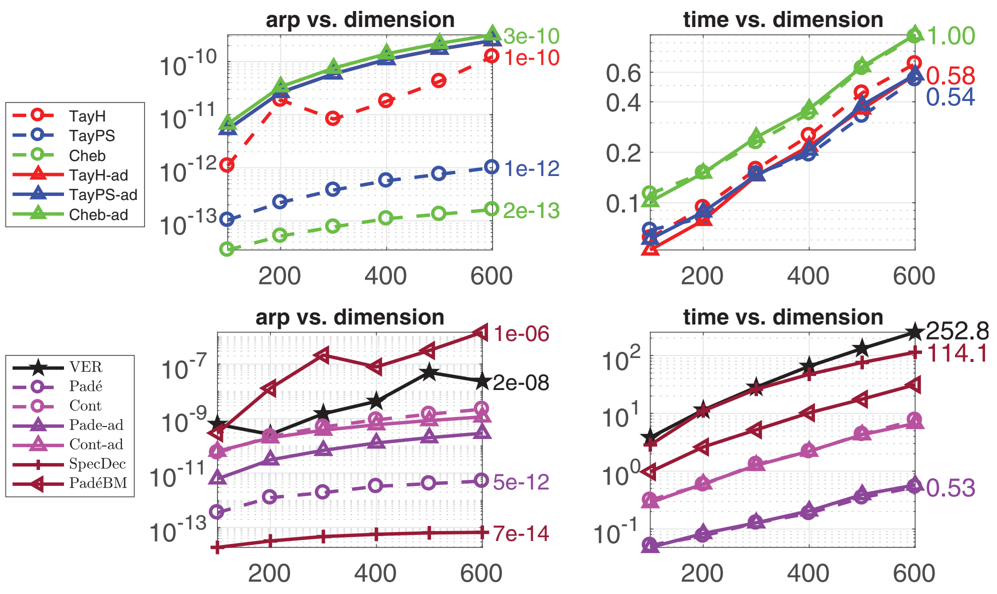

FIG. 11. *Average relative precision versus dimension (left) and time versus dimension (right) for the symmetric matrix* `orthog` *of type two, Example* 8.

TABLE 9
Average number of correct digits and standard deviation for each method for the largest matrix of Example 8.

	`TayH`	`TayPS`	`Cheb`	`TayH-ad`	`TayPS-ad`	`Cheb-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	9.9	12.0	12.8	9.6	9.6	9.5	7.6	11.3	8.7	9.5	8.9	13.2	5.8
Std. dev.	0.39	0.39	0.38	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.39	0.38	0.39

Example 9. A is generated as a symmetric random matrix by filling the upper triangle of a matrix with normally distributed random entries and complementing the lower triangle symmetrically.

Figure 12 shows that among the methods with acceptable run time, those with approximate diagonalization (`Pade-ad`, `TayH-ad`, `TayPS-ad`, and `Cheb-ad`) perform similarly and obtain about 2 to 3 additional digits of accuracy when compared with their counterparts without approximate diagonalization. See Table 10.

Example 10. A is the symmetric positive definite `prolate` matrix from MATLAB's gallery. It was also used as a test case in [28]. The matrix is Toeplitz and

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

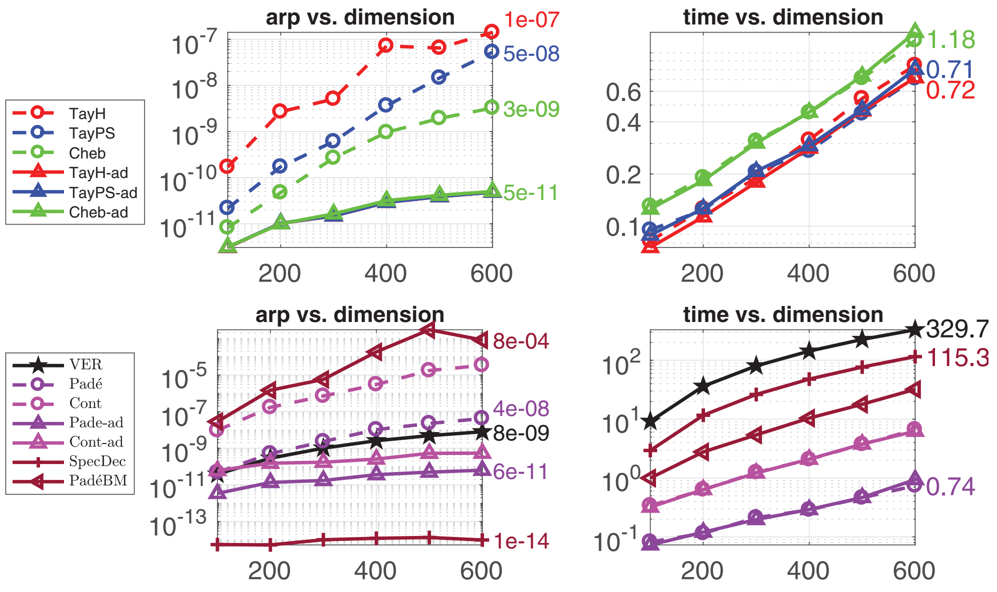

FIG. 12. *Average relative precision (left) and time (right) for the symmetric random matrix, Example 9.*

TABLE 10

Average number of correct digits and standard deviation for each method for the largest matrix of Example 9.

	TayH	TayPS	Cheb	TayH-ad	TayPS-ad	Cheb-ad	VER	Padé	Cont	Padé-ad	Cont-ad	SpecDec	PadéBM
Average	6.8	7.3	8.5	10.3	10.3	10.3	8.1	7.4	4.5	10.2	9.3	14.0	3.1
Std. dev.	0.48	0.48	0.48	0.48	0.48	0.48	0.49	0.48	0.48	0.48	0.47	0.40	0.48

perfectly well-conditioned with respect to the eigenvalues but ill-conditioned with respect to inversion or matrix multiplication.

The results in Figure 13 show that this time the most accurate results are obtained either by `TayPS` or `Cheb`, which show comparable speed. See Table 11. The eigenvalues of the `prolate` matrix tend to cluster around 0 and 1 which is why, as in Example 7, `VER` obtains poor enclosures. The cluster at 0 also explains why approximate diagonalization deteriorates the quality of the enclosures significantly: when we compute the almost diagonal matrix $\boldsymbol{D}$, the size of the off-diagonal entries is comparable to that of the eigenvalues clustering at 0. Then the computed enclosure for $\exp(\boldsymbol{D})$ will have off-diagonal entries which are not small relative to the diagonal elements, too, and this will spoil the relative accuracy when performing the two matrix-matrix multiplications in the back transformation (4.1).

Example 11. We take A as the symmetric and positive definite `poisson` matrix from MATLAB's gallery. It represents the finite difference discretization of the Laplace operator on an equispaced $N \times N$ grid with Dirichlet boundary conditions. This example is also considered in [28]. We took matrices of size $n = 100 = 10^2$, $196 = 14^2$, $289 = 17^2$, $400 = 20^2$, $484 = 22^2$, and $n = 625 = 25^2$.

Figure 14 shows that the most accurate results are obtained with `Padé`. The results with `SpecDec` are not as accurate as those for `Cheb`, `TayPS`, and `TayH`. See also Table 12. The fastest methods are `Padé` and Taylor-type techniques. `VER` returns NaNs for all dimensions.

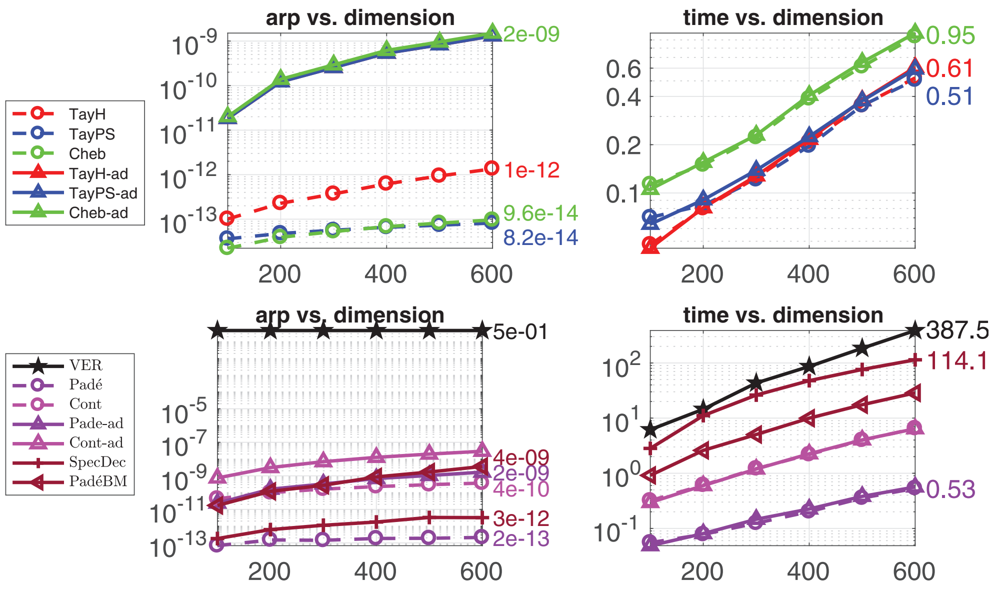

FIG. 13. *Average relative precision (left) and time (right) for the* `prolate` *matrix, Example* 10.

TABLE 11

Average number of correct digits and standard deviation for each method for the largest matrix of Example 10.

	`TayH`	`TayPS`	`Cheb`	`TayH-ad`	`TayPS-ad`	`Cheb-ad`	`VER`	`Padé`	`Cont`	`Padé-ad`	`Cont-ad`	`SpecDec`	`PadéBM`
Average	11.9	13.1	13.0	8.9	8.9	8.8	0.3	12.7	9.4	8.8	7.5	11.5	8.4
Std. dev.	0.58	0.54	0.56	0.64	0.64	0.64	0.00	0.58	0.62	0.64	0.64	0.65	0.61

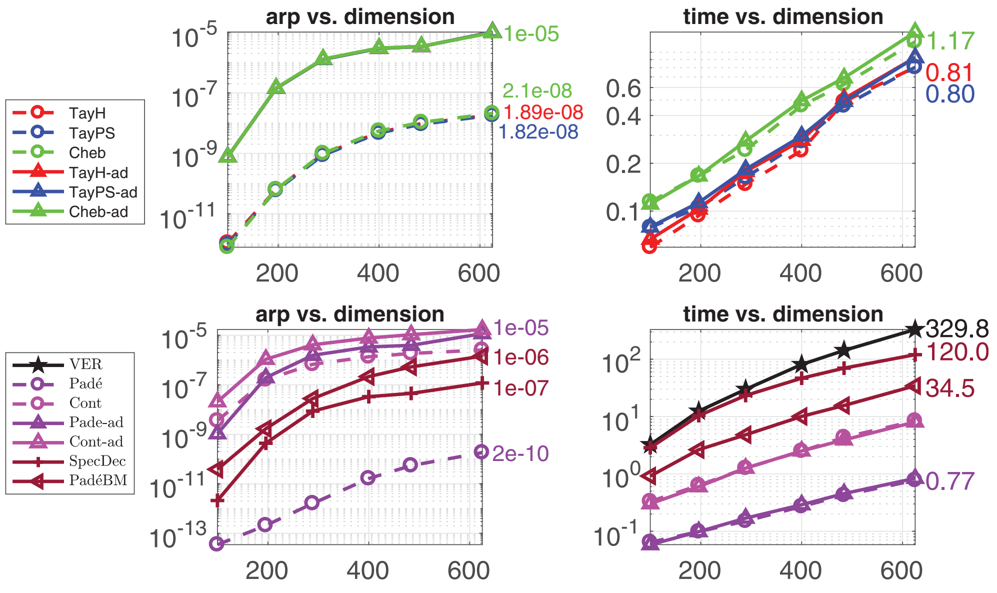

FIG. 14. *Average relative precision (left) and time (right) for the* `poisson` *matrix, Example* 11.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

TABLE 12
Average number of correct digits and standard deviation for each method for the largest matrix of Example 11.

	TayH	TayPS	Cheb	TayH-ad	TayPS-ad	Cheb-ad	VER	Padé	Cont	Padé-ad	Cont-ad	SpecDec	PadéBM
Average	7.7	7.7	7.7	5.0	5.0	5.0	NaN	9.7	5.6	4.9	4.8	6.9	5.8
Std. dev.	4.04	4.04	4.14	2.92	2.92	2.92	NaN	4.10	3.17	2.86	2.72	3.83	2.60

6. Conclusions and future work.

We have presented some new methods and improvements of several known methods for computing enclosures for the exponential of a matrix. The methods TayH, TayPS, Cheb rely exclusively on matrix-matrix multiplications so that, as a rule, the methods which require the fewest yield the tightest enclosures, since they reduce the wrapping effect. The methods Padé and Cont involve a linear system solve with an (interval) matrix right-hand side. For this task, state-of-the-art interval methods are available (implemented as verifylss.m of INTLAB, for instance), which compute tight enclosures for the solution set.

We performed a number of numerical experiments comparing the quality of the computed enclosure and the run time of these methods for a variety of test matrices. For general matrices, our new Padé is typically the best compromise in terms of accuracy and speed, closely followed by our Paterson–Stockmeyer variant TayPS of the Taylor approximation approach. For symmetric matrices, the new Cheb or Padé, followed by TayPS, are generally superior to the rest of the methods. The methods which rely on an either verified (VER) or approximate (SpecDec) spectral decomposition of the matrix suffer from a much higher wall clock time and do not, in general, provide tighter enclosures than Padé and Cheb. Moreover, these methods may fail completely for nondiagonalizable matrices. Method Cont requires the computation of enclosures for several linear systems which results in higher computational cost and less accurate overall enclosures than Padé. Finally, approximate diagonalization can be beneficial or detrimental to the quality of the computed enclosures, but it seems hard to characterize classes of matrices for which either of these observations would hold in general.

As mentioned in the introduction, the pure floating point Padé technique of [1] uses approximants of variable degree $m = q \in \{3, 5, 7, 9, 13\}$, while we simply work with $m = q = 7$. The latter is the result of a backward error analysis in which rounding errors in computing the Padé approximant are assumed to play no important role, and the truncation error is interpreted as equivalent to a perturbation in the original matrix A, while our algorithm gives automatic forward error bounds in which all the rounding errors are taken into account and the degree is determined solely based on the truncation error. One might also envision designing a variable degree Padé algorithm in interval arithmetic, which we leave for future work.

Acknowledgments. We are grateful to Holger Arndt, Anthony Austin, Bruno Lang, Yuji Nakatsukasa, Siegfried M. Rump, and Nick Trefethen for stimulating discussions. In particular it was Holger Arndt who ran the Arb package [20] and provided accurate enclosures for the values of the modified Bessel function of the first kind. We also thank the referees for their very careful reading of the paper and for offering several helpful comments.

REFERENCES

[1] A. H. Al-Mohy and N. J. Higham, *A new scaling and squaring algorithm for the matrix exponential*, SIAM J. Matrix Anal. Appl., 31 (2009), pp. 970–989, https://doi.org/10.1137/09074721X.

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

[2] G. Alefeld and G. Heindl, *A fixed point theorem based on a modified midpoint–radius interval arithmetic*, Reliab. Comput., 26 (2018), pp. 97–108.

[3] T. Auckenthaler, M. Bader, T. Huckle, A. Spörl, and K. Waldherr, *Matrix exponentials and parallel prefix computation in a quantum control problem*, Parallel Comput., 36 (2010), pp. 359–369.

[4] C. A. Bavely and G. W. Stewart, *An algorithm for computing reducing subspaces by block diagonalization*, SIAM J. Numer. Anal., 16 (1979), pp. 359–367, https://doi.org/10.1137/0716028.

[5] P. Bochev, *Simultaneous self-verified computation of* $\exp(a)$ *and* $\int_0^1 \exp(as)ds$, Computing, 45 (1990), pp. 183–191.

[6] P. Bochev and S. Markov, *A self-validating numerical method for the matrix exponential*, Computing, 43 (1989), pp. 59–72.

[7] E. B. Davies, *Approximate diagonalization*, SIAM J. Matrix Anal. Appl., 29 (2007), pp. 1051–1064, https://doi.org/10.1137/060659909.

[8] P. I. Davies and N. J. Higham, *A Schur–Parlett algorithm for computing matrix functions*, SIAM J. Matrix Anal. Appl., 25 (2003), pp. 464–485, https://doi.org/10.1137/S0895479802410815.

[9] J. W. Demmel, *Applied Numerical Linear Algebra*, SIAM, Philadelphia, 1997, https://doi.org/10.1137/1.9781611971446.

[10] T. A. Driscoll, N. Hale, and L. N. Trefethen, *Chebfun Guide*, Pafnuty Publications, Oxford, 2014.

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

[12] A. Goldsztejn and A. Neumaier, *On the exponentiation of interval matrices*, Reliab. Comput., 20 (2014), pp. 53–72.

[13] B. Hashemi, *Enclosing Chebyshev expansions in linear time*, ACM Trans. Math. Software, 45 (2019), pp. 1–33, https://doi.org/10.1145/3319395, http://bit.ly/2D9brjd.

[14] N. J. Higham, *Accuracy and Stability of Numerical Algorithms*, 2nd ed., SIAM, Philadelphia, 2002, https://doi.org/10.1137/1.9780898718027.

[15] N. J. Higham, *The scaling and squaring method for the matrix exponential revisited*, SIAM J. Matrix Anal. Appl., 26 (2005), pp. 1179–1193, https://doi.org/10.1137/04061101X.

[16] N. J. Higham, *Functions of Matrices: Theory and Computation*, SIAM, Philadelphia, 2008, https://doi.org/10.1137/1.9780898717778.

[17] N. J. Higham, *The scaling and squaring method for the matrix exponential revisited*, SIAM Rev., 51 (2009), pp. 747–764, https://doi.org/10.1137/090768539.

[18] N. Hoffman, O. Schwartz, and S. Toledo, *Efficient evaluation of matrix polynomials*, in Parallel Processing and Applied Mathematics. Part I, Lecture Notes in Comput. Sci. 10777, R. Wyrzykowski, J. Dongarra, E. Deelman, and K. Karczewski, eds., Springer, Cham, 2018, pp. 24–35.

[19] W. Hofschuster and W. Krämer, *C-XSC* 2.0*—a C++ library for extended scientific computing*, in Numerical Software with Result Verification, R. Alt, A. Frommer, R. B. Kearfott, and W. Luther, eds., Springer, New York, 2004, pp. 15–35.

[20] F. Johansson, *Arb: Efficient arbitrary-precision midpoint-radius interval arithmetic*, IEEE Trans. Comput., 66 (2017), pp. 1281–1292.

[21] R. Klatte, U. Kulisch, A. Wiethoff, and M. Rauch, *C-XSC: A C++ Class Library for Extended Scientific Computing*, Springer-Verlag, Berlin, 1993.

[22] O. Kosheleva, V. Kreinovich, G. Mayer, and H. T. Nguyen, *Computing the cube of an interval matrix is NP-hard*, in Proceedings of the 2005 ACM Symposium on Applied Computing, ACM, New York, 2005, pp. 1449–1453.

[23] R. Krawczyk, *Newton-Algorithmen zur Bestimmung von Nullstellen mit Fehlerschranken*, Computing, 4 (1969), pp. 187–201.

[24] M. Liou, *A novel method of evaluating transient response*, Proc. IEEE, 54 (1966), pp. 20–23.

[25] J. C. Mason and D. C. Handscomb, *Chebyshev Polynomials*, CRC Press, Boca Raton, FL, 2003.

[26] G. Mayer, *Interval Analysis and Automatic Result Verification*, de Gruyter, Berlin, 2017.

[27] S. Miyajima, *MATLAB-INTLAB Implementations of "Verified computation of the matrix exponential,"* 2019, http://web.cc.iwate-u.ac.jp/~miyajima/Mexp.zip.

[28] S. Miyajima, *Verified computation of the matrix exponential*, Adv. Comput. Math., 45 (2019), pp. 137–152.

[29] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a matrix*, SIAM Rev., 20 (1978), pp. 801–836, https://doi.org/10.1137/1020098.

[30] C. Moler and C. Van Loan, *Nineteen dubious ways to compute the exponential of a ma-*

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.

trix, twenty-five years later, SIAM Rev., 45 (2003), pp. 3–49, https://doi.org/10.1137/S00361445024180.

[31] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to Interval Analysis*, SIAM, Philadelphia, 2009, https://doi.org/10.1137/1.9780898717716.

[32] A. Neumaier, *Enclosing clusters of zeros of polynomials*, J. Comput. Appl. Math., 156 (2003), pp. 389–401.

[33] E. P. Oppenheimer, *Application of Interval Analysis to Problems of Linear Control Systems*, Ph.D. thesis, Electrical Engineering Department, Iowa State University, 1979.

[34] E. P. Oppenheimer and A. N. Michel, *Application of interval analysis techniques to linear systems* II: *The interval matrix exponential function*, IEEE Trans. Circuits Systems, 35 (1988), pp. 1230–1242.

[35] M. S. Paterson and L. J. Stockmeyer, *On the number of nonscalar multiplications necessary to evaluate polynomials*, SIAM J. Comput., 2 (1973), pp. 60–66, https://doi.org/10.1137/0202007.

[36] W. Press, S. A. Teukolsky, W. Vetterling, and B. P. Flannery, *Numerical Recipes in C*, 2nd ed., Cambridge University Press, Cambridge, UK, 1992.

[37] J. Rohn, *VERSOFT: Verification Software in MATLAB/INTLAB*, http://uivtx.cs.cas.cz/~rohn/matlab/.

[38] S. M. Rump, *Kleine Fehlerschranken bei Matrixproblemen*, Ph.D. thesis, Fakultät für Mathematik, Universität Karlsruhe, 1980.

[39] S. M. Rump, *Fast and parallel interval arithmetic*, BIT, 39 (1999), pp. 534–554.

[40] S. M. Rump, *INTLAB–INTerval LABoratory*, in Developments in Reliable Computing, T. Csendes, ed., Kluwer Academic, Dordrecht, 1999, pp. 77–104.

[41] S. M. Rump, *Ten methods to bound multiple roots of polynomials*, J. Comput. Appl. Math., 156 (2003), pp. 403–432.

[42] S. M. Rump, *Verification methods: Rigorous results using floating-point arithmetic*, Acta Numer., 19 (2010), pp. 287–449.

[43] S. M. Rump, *Verified bounds for singular values, in particular for the spectral norm of a matrix and its inverse*, BIT, 51 (2011), pp. 367–384.

[44] S. M. Rump, *Accurate solution of dense linear systems, Part* II: *Algorithms using directed rounding*, J. Comput. Appl. Math., 242 (2013), pp. 185–212.

[45] S. M. Rump, T. Ogita, and S. Oishi, *Accurate floating-point summation part* I: *Faithful rounding*, SIAM J. Sci. Comput., 31 (2008), pp. 189–224, https://doi.org/10.1137/050645671.

[46] J. Sastre, *Efficient evaluation of matrix polynomials*, Linear Algebra Appl., 539 (2018), pp. 229–250.

[47] J. Sastre, J. Ibáñez, and E. Defez, *Boosting the computation of the matrix exponential*, Appl. Math. Comput., 340 (2019), pp. 206–220.

[48] J. Sastre, J. Ibáñez, E. Defez, and P. Ruiz, *Efficient orthogonal matrix polynomial based method for computing matrix exponential*, Appl. Math. Comput., 217 (2011), pp. 6451–6463.

[49] M. Shao, W. Gao, and J. Xue, *Aggressively truncated Taylor series method for accurate computation of exponentials of essentially nonnegative matrices*, SIAM J. Matrix Anal. Appl., 35 (2014), pp. 317–338, https://doi.org/10.1137/120894294.

[50] H. J. Stetter, *Sequential defect correction for high-accuracy floating-point algorithms*, in Numerical Analysis, D. F. Griffiths, ed., Lecture Notes in Math. 1066, Springer, Berlin, 1984, pp. 186–202.

[51] M. Suzuki, *Generalized Trotter's formula and systematic approximants of exponential operators and inner derivations with applications to many-body problems*, Comm. Math. Phys., 51 (1976), pp. 183–190.

[52] L. N. Trefethen, *Approximation Theory and Approximation Practice*, SIAM, Philadelphia, 2013, https://doi.org/10.1137/1.9781611975949.

[53] L. N. Trefethen, *Convergence Bounds for Entire Functions*, Chebfun example, 2016, http://www.chebfun.org/examples/approx/EntireBound.html (accessed 2017-04-05).

[54] L. N. Trefethen and J. A. C. Weideman, *The exponentially convergent trapezoidal rule*, SIAM Rev., 56 (2014), pp. 385–458, https://doi.org/10.1137/130932132.

[55] R. S. Varga, *On higher order stable implicit methods for solving parabolic partial differential equations*, J. Math. Phys., 40 (1961), pp. 220–231.

[56] R. C. Ward, *Numerical computation of the matrix exponential with accuracy estimate*, SIAM J. Numer. Anal., 14 (1977), pp. 600–610, https://doi.org/10.1137/0714039.

[57] J. A. C. Weideman, *Numerical integration of periodic functions: A few examples*, Amer. Math. Monthly, 109 (2002), pp. 21–36.

Downloaded 01/01/21 to 59.64.129.2. Redistribution subject to SIAM license or copyright; see https://epubs.siam.org/page/terms

Copyright © by SIAM. Unauthorized reproduction of this article is prohibited.