Matrix Exponential Computational Algorithm

Kenneth C. Johnson
KJ Innovation

(Posted 13-Aug-2026.)
https://vixra.org/abs/2203.0072

Abstract

A numerical algorithm for the matrix exponential is developed, based on the scale-and-square method applied to a Padé approximant for small-norm matrices.

1. Introduction

A generalized Padé approximation method for numerically solving nonhomogeneous, coupled linear differential equations with non-constant coefficients was developed in [1, 2]. This paper refines and simplifies that work for the homogeneous, constant-coefficient case:

$$\frac{d}{dx} F[x] = D\, F[x] \tag{1}$$

where $F[x]$ is a vector function of scalar argument x and D is a constant, square matrix. (In this paper square braces "$[\ldots]$" delimit function arguments while round braces "$(\ldots)$" are reserved for grouping.) The solution of Eq. (1)

$$F[x] = \exp[x\, D]\, F[0] \tag{2}$$

The matrix exponential is calculated by the scale-and-square method:

$$\exp[X] = \exp[2^{-p}\, X]^{2^p} = (\ldots(\exp[2^{-p}\, X]\overbrace{^2)^2\ldots)^2}^{p\text{ times}} \tag{3}$$

where $\exp[2^{-p}\, X]$ approximated by a rational polynomial (Padé approximation)

$$\exp[2^{-p}\, X] \cong P[-2^{-1-p}\, X]^{-1}\, P[2^{-1-p}\, X] \tag{4}$$

(P is a polynomial function.)

The scaling power p is chosen to achieve a specified relative tolerance limit ϵ in the approximation error. Denoting the $F[x]$ approximation error as $\delta F[x]$, the tolerance condition is

$$|\delta F[x]| \le \epsilon\, |F[x]| \tag{5}$$

Formulas for the P polynomial coefficients, approximation error bound, and choice of scaling power p are developed. The squaring operation in Eq. (3) can be susceptible to numerical roundoff error in the matrix diagonal, but an alternative squaring operation is used to avoid the precision loss.

The algorithm improves upon the accuracy and runtime performance of the MATLAB® **`expm`** function [3, 4]. An implementation is posted on the MathWorks® File Exchange [5].

2. The Padé approximation

Following the method described by Gautschi [6], an order-n Padé approximant to $\exp[x\,D]$ is derived by applying $2n+1$ integration-by-parts operations to the following expression,

$$D^{2n+1}\int_{-h}^{h}\exp[x\,D](x^2-h^2)^n\,dx=\sum_{j=0}^{2n}(-1)^j\,D^{2n-j}\exp[x\,D]\frac{d^j}{dx^j}(x^2-h^2)^n\Bigg|_{x=-h}^{x=h} \tag{6}$$

The derivative factor is expanded via the general Leibniz rule,

$$\begin{aligned}&\frac{d^j}{dx^j}(x^2-h^2)^n=\frac{d^j}{dx^j}\left((x+h)^n\,(x-h)^n\right)\\&=\sum_{k=0}^{j}\frac{j!}{k!(j-k)!}\left(\frac{d^k}{dx^k}(x+h)^n\right)\left(\frac{d^{j-k}}{dx^{j-k}}(x-h)^n\right)\\&=\sum_{k=0}^{j}\frac{j!}{k!(j-k)!}\left(\frac{n!}{(n-k)!}(x+h)^{n-k}\right)\left(\frac{n!}{(n-j+k)!}(x-h)^{n-j+k}\right)\end{aligned} \tag{7}$$

At the lower integration limit ($x=-h$) the $(x+h)^{n-k}$ factor is nonzero only if $k=n$, and at the upper limit ($x=h$) the factor $(x-h)^{n-j+k}$ is nonzero only if $k=j-n$. With $j\le k$, neither condition holds when $j<n$; hence the summation terms $j=0,\ldots n-1$ vanish in Eq. (6) and the sum reduces to

$$\begin{aligned}&\sum_{j=0}^{2n}(-1)^j\,D^{2n-j}\exp[x\,D]\frac{d^j}{dx^j}(x^2-h^2)^n\Bigg|_{x=-h}^{x=h}=\\&\quad\sum_{j=n}^{2n}\frac{n!\,j!}{(j-n)!(2n-j)!}\left((-2\,h\,D)^{2n-j}\exp[h\,D]-(2\,h\,D)^{2n-j}\exp[-h\,D]\right)\end{aligned} \tag{8}$$

The summation index j is replaced by $2n-j$ and the result is substituted back into Eq. (6) to obtain

$$\frac{D^{2n+1}}{(2n)!}\int_{-h}^{h}\exp[x\,D](x^2-h^2)^n\,dx=P[-h\,D]\exp[h\,D]-P[h\,D]\exp[-h\,D] \tag{9}$$

where[1]

$$P[X]=\sum_{j=0}^{n}c_j\,X^j \tag{10}$$

$$c_j=\frac{n!(2n-j)!2^j}{(2n)!\,j!(n-j)!}=\frac{1}{j!}\prod_{k=0}^{j-1}\left(1-\frac{k}{2n-k}\right) \tag{11}$$

[1] The function $P[h\,D]$ corresponds to $Q[-h]$ in [1] and corresponds to "$P[n,n](2\,h\,D)$" in [5].

Denoting the P function for Padé order n as P_n, the polynomial coefficients c_j for P_n can be efficiently calculated from the following recursion formula. ($\mathbf{I}$ is an identity matrix.)

$$\left.\begin{aligned} &P_0[X] = \mathbf{I}, \\ &P_1[X] = \mathbf{I} + X, \\ &P_{n+1}[X] = P_n[X] + \frac{X^2}{4n^2-1} P_{n-1}[X] \end{aligned}\right\} \tag{12}$$

An algorithm for calculating $P[X]$ with a minimal number of matrix multiplies is outlined in the Appendix.

A factor of $\exp[h\,D]$ is multiplied into both sides of Eq. (9),

$$\frac{D^{2n+1}}{(2n)!}\int_{-h}^{h} \exp[(x+h)\,D](x^2-h^2)^n\,dx = P[-h\,D]\exp[2\,h\,D] - P[h\,D] \tag{13}$$

For sufficiently small h the left side of Eq. (13), which is of order h^{2n+1}, can be neglected. This leads to the Padé approximation,

$$\exp[2\,h\,D] \cong \Phi[2\,h\,D] = P[-h\,D]^{-1}\,P[h\,D] \tag{14}$$

$\exp[2^{p+1}\,h\,D]$ is approximated as $\Phi[2\,h\,D]^p$,

$$\exp[2^{p+1}\,h\,D] = \exp[2\,h\,D]^{2^p} \cong \Phi[2\,h\,D]^{2^p} = (\ldots(\Phi[2\,h\,D]\overbrace{^{2})^{2}\ldots)^{2}}^{p\text{ times}} \tag{15}$$

The right side of Eq. (15) is an approximation to $\exp[x\,D]$ for $x = 2^{p+1}\,h$. Note that even with the approximation, the relation $\exp[-x\,D] = \exp[x\,D]^{-1}$ holds exactly (from Eq. (14)):

$$\Phi[-2\,h\,D]^{2^p} = \left(\Phi[2\,h\,D]^{2^p}\right)^{-1} \tag{16}$$

For small h, $\Phi[2\,h\,D]$ is close to an identity matrix ($\mathbf{I}$) and the matrix diagonal's precision is limited by the dominant $\mathbf{I}$ term. To avoid precision loss, $\exp[x\,D]$ can be calculated with the identity function subtracted off. Eq. (14) is modified as

$$\Phi[2\,h\,D] - \mathbf{I} = P[-h\,D]^{-1}\,(P[h\,D] - P[-h\,D]) \tag{17}$$

The squaring operation ($\Phi \leftarrow \Phi^2$) can be implemented using the relation $\Phi^2 - \mathbf{I} = (\Phi - \mathbf{I})^2 + 2(\Phi - \mathbf{I})$: With $\Phi' = \Phi - \mathbf{I}$ apply the operation $\Phi' \leftarrow \Phi'^2 + 2\Phi'$ p times; then add $\mathbf{I}$ to the result. However, this approach does not work when D has large, negative eigenvalues resulting in and exponential below the numeric precision limit ($\|\exp[x\,D]\| << 1$). In this case, after the squaring steps Φ' would be equal to $-\mathbf{I}$ within numerical precision and Φ would evaluate to zero.

To minimize precision loss in the Φ diagonal at both the beginning and end of the squaring process, the squaring operation is applied with its diagonal approximately subtracted off at each step as follows:

$$
\left.\begin{array}{l}
\text{Initialize } \Phi = \Phi[2\,h\,D],\ d = \begin{pmatrix} 1 \\ 1 \\ \vdots \end{pmatrix} \big(\mathbf{I} = \mathrm{diag}[d]\big),\ \Phi' = \Phi - \mathrm{diag}[d]. \\
\text{Repeat } p \text{ times:} \\
\quad d'_j = d_j + \Phi'_{j,j} \quad \big(d'_j \cong \Phi_{j,j}\big) \\
\quad d'_j \ \leftarrow\ \mathrm{prec\_trunc}[d'_j] \\
\quad \Phi'_{j,j} \ \leftarrow\ \Phi'_{j,j} - (d'_j - d_j) \quad \big(\Phi'_{j,j} \cong 0,\ \Phi' = \Phi - \mathrm{diag}[d']\big) \\
\quad d \ \leftarrow\ d' \quad \big(\Phi' = \Phi - \mathrm{diag}[d]\big) \\
\quad \Phi' \ \leftarrow\ \Phi'^2 + \mathrm{diag}[d]\Phi' + \Phi'\mathrm{diag}[d] \quad \big(\Phi' \ \leftarrow\ \Phi^2 - \mathrm{diag}[d]^2;\ \Phi \ \leftarrow\ \Phi^2\big) \\
\quad d_j \ \leftarrow\ d_j^{\,2} \quad \big(\Phi' = \Phi - \mathrm{diag}[d]\big) \\
\Phi = \Phi' + \mathrm{diag}[d]
\end{array}\right\} \tag{18}
$$

Note that the operation $d'_j = d_j + \Phi'_{j,j}$ can result in d'_j being numerically equal to d_j when $\Phi'_{j,j}$ is very small. But the operation $\Phi'_{j,j} \ \leftarrow\ \Phi'_{j,j} - (d'_j - d_j)$ accurately preserves the relation $\Phi' = \Phi - \mathrm{diag}[d']$ while keeping the Φ' diagonal small ($\Phi'_{j,j} \cong 0$). The $\mathrm{prec\_trunc}[\ldots]$ function ("precision-truncate") truncates its argument's mantissa to limit the number of significant bits to at most half the maximum number of bits. (For double precision, 27 low-order mantissa bits are cleared, and for single precision, 12 bits are cleared.) The precision truncation operation ensures that the squaring operation ($d_j^{\,2}$) does not lose precision.

3. Error analysis

The relative error in the $\exp[2\,h\,D]$ approximation, denoted as $\delta^{[\mathrm{rel}]}[2\,h]$, is defined by

$$\Phi[2\,h\,D] = (\mathbf{I} + \delta^{[\mathrm{rel}]}[2\,h])\exp[2\,h\,D] \tag{19}$$

The relative error in $\exp[2^{p+1}\,h\,D]$ (i.e., $\exp[2\,h\,D]^{2^p}$) is denoted as $\delta^{[\mathrm{rel}]}[2^{p+1}\,h]$ and is similarly defined by

$$\Phi[2\,h\,D]^{2^p} = (\mathbf{I} + \delta^{[\mathrm{rel}]}[2^{p+1}\,h])\exp[2\,h\,D]^{2^p} \tag{20}$$

These definitions imply the following relation between $\delta^{[\mathrm{rel}]}[2\,h]$ and $\delta^{[\mathrm{rel}]}[2^{p+1}\,h]$,

$$\mathbf{I} + \delta^{[\mathrm{rel}]}[2^{p+1}\,h] = (\mathbf{I} + \delta^{[\mathrm{rel}]}[2\,h])^{2^p} \tag{21}$$

The error $\delta^{[\mathrm{rel}]}[2\,h]$ is required to be within the bound

$$\left\|\delta^{[\mathrm{rel}]}[2\,h]\right\| \le \kappa \tag{22}$$

where $\|\ldots\|$ is the Frobenius norm (the root-sum-square of the matrix elements). This implies the bound

$$\begin{aligned}\left\|\delta^{[\mathrm{rel}]}[2^{p+1} h]\right\| &= \left\|(\mathbf{I}+\delta^{[\mathrm{rel}]}[2h])^{2^p} - \mathbf{I}\right\| \le (1+\left\|\delta^{[\mathrm{rel}]}[2h]\right\|)^{2^p} - 1 \\ &\le (1+\kappa)^{2^p} - 1 \le \exp[\kappa]^{2^p} - 1 = \exp[2^p\,\kappa] - 1\end{aligned} \tag{23}$$

κ is defined by

$$\exp[2^p\,\kappa] - 1 = \epsilon, \quad \kappa = 2^{-p}\log[\epsilon+1] \tag{24}$$

(The $\log[\epsilon+1]$ factor can be calculated accurately with MATLAB's **`log1p`** function.) $\delta^{[\mathrm{rel}]}[2^{p+1} h]$ is thus within the bound

$$\left\|\delta^{[\mathrm{rel}]}[2^{p+1} h]\right\| \le \epsilon \tag{25}$$

With $x = 2^{p+1} h$, this implies Eq. (5)

$$\begin{aligned}\left|\delta F[2^{p+1} h]\right| &= \left|(\Phi[2\,h\,D]^{2^p} - \exp[2^{p+1} h\,D])\,F[0]\right| = \left|\delta^{[\mathrm{rel}]}[2^{p+1} h]\exp[2^{p+1} h\,D]\,F[0]\right| \\ &= \left|\delta^{[\mathrm{rel}]}[2^{p+1} h]\,F[2^{p+1} h]\right| \le \left\|\delta^{[\mathrm{rel}]}[2^{p+1} h]\right\| \cdot \left|F[2^{p+1} h]\right| \le \epsilon \left|F[2^{p+1} h]\right|\end{aligned} \tag{26}$$

(from Eq's. (2), (14), (20), and (25)).

The following formula for $\delta^{[\mathrm{rel}]}[2h]$ is obtained by eliminating $P[h\,D]$ between Eq's. (13) and (14), and substituting Eq. (19):

$$\delta^{[\mathrm{rel}]}[2h] = -P[-h\,D]^{-1}\,\frac{D^{2n+1}}{(2n)!}\int_{-h}^{h}\exp[(x-h)\,D]\,(x^2-h^2)^n\,dx \tag{27}$$

A bound on $\left\|\delta^{[\mathrm{rel}]}[2h]\right\|$ is determined by separating Eq. (27) into three factors and establishing a separate bound for each factor,

$$\delta^{[\mathrm{rel}]}[2h] = -(P[h\,D]\,P[-h\,D])^{-1}\,(P[h\,D]\exp[-h\,D])\,\Delta \tag{28}$$

where

$$\Delta = \frac{D^{2n+1}}{(2n)!}\int_{-h}^{h}\exp[x\,D]\,(x^2-h^2)^n\,dx \tag{29}$$

The divisor in the first factor of Eq. (28) is approximately quadratic in h: $P[h\,D]\,P[-h\,D] = \mathbf{I} - h^2\,D^2/(2n-1) + O\,h^4$. The following bounding condition is applied to the divisor, with R representing the two right-hand factors in Eq. (28),

$$\begin{aligned}&\left\|P[h\,D]\,P[-h\,D] - \mathbf{I}\right\| < 1 \;\rightarrow \\ &\quad \left\|(P[h\,D]\,P[-h\,D])^{-1}\,R\right\| \le (1 - \left\|P[h\,D]\,P[-h\,D] - \mathbf{I}\right\|)^{-1}\,\|R\|\end{aligned} \tag{30}$$

This (30) follows from the general relation $\left\|(\mathbf{I}+A)^{-1}\,R\right\| \le (1-\|A\|)^{-1}\,\|R\|$ when $\|A\| < 1$:

$$\left\|(\mathbf{I}+A)^{-1}\,R\right\| = \left\|R + \sum_{j=1}^{\infty}(-A)^j\,R\right\| \le \|R\| + \sum_{j=1}^{\infty}\|A\|^j\,\|R\| = (1-\|A\|)^{-1}\,\|R\| \tag{31}$$

The matrix product $P[X]P[-X]$ has a Taylor series of the form[2]

$$P[X]P[-X]=\sum_{j=0}^{n}a_j\,X^{2j},\quad a_j=\frac{(-1)^j\,j!(2n-2j)!}{(2n-j)!}c_j^{\;2} \tag{32}$$

The following bound is obtained from this series,

$$\begin{aligned}&\|P[X]P[-X]-\mathbf{I}\|=\left\|\sum_{j=1}^{n}a_j\,X^{2j}\right\|\le\sum_{j=1}^{n}|a_j|\cdot\|X^2\|^j=\sum_{j=1}^{n}a_j\cdot(i\sqrt{\|X^2\|})^{2j}\\&=P[i\sqrt{\|X^2\|}]P[-i\sqrt{\|X^2\|}]-1\end{aligned} \tag{33}$$

(The i factor cancels the $(-1)^j$ factor in a_j.) Eq's. (30) and (33) are combined to obtain the following condition,

$$\begin{aligned}&P[i|h|\sqrt{\|D^2\|}]P[-i|h|\sqrt{\|D^2\|}]<2\to\\&\quad\|(P[h\,D]P[-h\,D])^{-1}R\|\le\left(2-P[i|h|\sqrt{\|D^2\|}]P[-i|h|\sqrt{\|D^2\|}]\right)^{-1}\|R\|\end{aligned} \tag{34}$$

This condition can be reformulated by separating P into even and odd parts,

$$P^{[\text{even}]}[X]=\tfrac{1}{2}(P[X]+P[-X]),\quad P^{[\text{odd}]}[X]=\tfrac{1}{2}(P[X]-P[-X]) \tag{35}$$

$$P[X]P[-X]=(P^{[\text{even}]}[X])^2-(P^{[\text{odd}]}[X])^2 \tag{36}$$

$$\begin{aligned}&(P^{[\text{even}]}[i|h|\sqrt{\|D^2\|}])^2-(P^{[\text{odd}]}[-i|h|\sqrt{\|D^2\|}])^2<2\to\\&\quad\|(P[h\,D]P[-h\,D])^{-1}R\|\le\left(2-(P^{[\text{even}]}[i|h|\sqrt{\|D^2\|}])^2+(P^{[\text{odd}]}[-i|h|\sqrt{\|D^2\|}])^2\right)^{-1}\|R\|\end{aligned} \tag{37}$$

h is constrained to satisfy the premise (first inequality) of Eq. (37). In practice, the constraint $\ldots<2$ can be replaced by a somewhat tighter limit, e.g., $\ldots<1.9$, to ensure that the right-hand reciprocal factor is not very large.

The second factor in Eq. (28), $P[h\,D]\exp[-h\,D]$, is separated into even and odd functions,

$$\begin{aligned}&P[h\,D]\exp[-h\,D]=\\&\quad\tfrac{1}{2}(P[h\,D]\exp[-h\,D]+P[-h\,D]\exp[h\,D])+\\&\quad\tfrac{1}{2}(P[h\,D]\exp[-h\,D]-P[-h\,D]\exp[h\,D])\end{aligned} \tag{38}$$

The odd function is equal to Eq. (9) times $-\tfrac{1}{2}$,

$$\tfrac{1}{2}(P[h\,D]\exp[-h\,D]-P[-h\,D]\exp[h\,D])=-\tfrac{1}{2}\Delta \tag{39}$$

[2] Eq. (32) can be verified with the following Mathematica script:

```
c[n_, j_] := n! (2 n - j)! 2^j/(2 n)!/j!/(n - j)!
P[x_, n_] := Sum[c[n, j] x^j, {j, 0, n}]
a[n_, j_] := (-1)^j j! (2 n - 2 j)! c[n, j]^2/(2 n - j)!
PP[x_, n_] := Sum[a[n, j] x^(2 j), {j, 0, n}]
FullSimplify[PP[x, n] - P[x, n] P[-x, n]]
```

(cf. Eq. (29)). The even function is bounded by taking advantage of the fact that $P[\pm h\,D]$ is close to $\exp[\pm h\,D]$ for small h: $\exp[\pm h\,D]-P[\pm h\,D]=h^2\,D^2\,/\,(4\,n-2)+O\,h^3$. These differences are separated out in the even function, and the differences are then further separated into even and odd terms:

$$\begin{aligned}&P[h\,D]\exp[-h\,D]+P[-h\,D]\exp[h\,D]\\&=\mathbf{I}+P[h\,D]\,P[-h\,D]-(\exp[h\,D]-P[h\,D])\,(\exp[-h\,D]-P[-h\,D])\\&=\mathbf{I}+P[h\,D]\,P[-h\,D]-(\cosh[h\,D]-P^{[\mathrm{even}]}[h\,D])^2+(\sinh[h\,D]-P^{[\mathrm{odd}]}[h\,D])^2\end{aligned}\tag{40}$$

The function $\exp[X]-P[X]$ has a Taylor series expansion with non-negative coefficients,

$$\exp[X]-P[X]=\sum_{j=2}^{n}\left(1-\prod_{k=1}^{j-1}\left(1-\frac{k}{2\,n-k}\right)\right)\frac{X^j}{j\,!}+\sum_{j=n+1}^{\infty}\frac{X^j}{j\,!}\tag{41}$$

The even and odd parts of Eq. (41), $\cosh[X]-P^{[\mathrm{even}]}[X]$ and $\sinh[X]-P^{[\mathrm{odd}]}[X]$, also have non-negative Taylor series coefficients, and so do the squares of these functions. Furthermore, the squared terms are even functions of X comprising monomial powers $X^{2\,j}$, which are bounded by $\|X^{2\,j}\|\le\|X^2\|^j=\sqrt{\|X^2\|}^{\,2\,j}$; hence

$$\left\|(\cosh[X]-P^{[\mathrm{even}]}[X])^2\right\|\le(\cosh[\sqrt{\|X^2\|}]-P^{[\mathrm{even}]}[\sqrt{\|X^2\|}])^2\tag{42}$$

$$\left\|(\sinh[X]-P^{[\mathrm{odd}]}[X])^2\right\|\le(\sinh[\sqrt{\|X^2\|}]-P^{[\mathrm{odd}]}[\sqrt{\|X^2\|}])^2\tag{43}$$

The squared terms in Eq. (40) are bounded using Eq's. (42) and (43),

$$\begin{aligned}&\left\|-(\cosh[h\,D]-P^{[\mathrm{even}]}[h\,D])^2+(\sinh[h\,D]-P^{[\mathrm{odd}]}[h\,D])^2\right\|\\&\le\left\|(\cosh[h\,D]-P^{[\mathrm{even}]}[h\,D])^2\right\|+\left\|(\sinh[h\,D]-P^{[\mathrm{odd}]}[h\,D])^2\right\|\\&\le(\cosh[h\sqrt{\|D^2\|}]-P^{[\mathrm{even}]}[h\sqrt{\|D^2\|}])^2+(\sinh[h\sqrt{\|D^2\|}]-P^{[\mathrm{odd}]}[h\sqrt{\|D^2\|}])^2\end{aligned}\tag{44}$$

Eq's. (38)-(40) are combined and substituted in Eq. (28),

$$\delta^{[\mathrm{rel}]}[2\,h]=-\tfrac{1}{2}\left(\mathbf{I}+(P[h\,D]\,P[-h\,D])^{-1}\begin{pmatrix}\mathbf{I}-(\cosh[h\,D]-P^{[\mathrm{even}]}[h\,D])^2\\+(\sinh[h\,D]-P^{[\mathrm{odd}]}[h\,D])^2-\Delta\end{pmatrix}\right)\Delta\tag{45}$$

Eq's. (37) and (44) are combined with Eq. (45) to obtain the bound

$$\begin{aligned}&(P^{[\mathrm{even}]}[i|h|\sqrt{\|D^2\|}])^2-(P^{[\mathrm{odd}]}[-i|h|\sqrt{\|D^2\|}])^2<2\rightarrow\\&\left\|\delta^{[\mathrm{rel}]}[2\,h]\right\|\le\tfrac{1}{2}\begin{pmatrix}1+\left(2-(P^{[\mathrm{even}]}[i|h|\sqrt{\|D^2\|}])^2+(P^{[\mathrm{odd}]}[-i|h|\sqrt{\|D^2\|}])^2\right)^{-1}\\\begin{pmatrix}1+(\cosh[h\sqrt{\|D^2\|}]-P^{[\mathrm{even}]}[h\sqrt{\|D^2\|}])^2\\+(\sinh[h\sqrt{\|D^2\|}]-P^{[\mathrm{odd}]}[h\sqrt{\|D^2\|}])^2+\|\Delta\|\end{pmatrix}\end{pmatrix}\|\Delta\|\end{aligned}\tag{46}$$

A bound on Δ is obtained from Eq. (29). The integral is unchanged when the integrand factor $\exp[x\,D]$ is replaced by $\cosh[x\,D]$, and the product $D^{2n+1}\cosh[x\,D]$ has the bound

$$\begin{aligned}&\left\|D^{2n+1}\cosh[x\,D]\right\| = \left\|\sum_{j=0}^{\infty}\frac{x^{2j}\,D^{2n+1+2j}}{(2\,j)!}\right\| \le \sum_{j=0}^{\infty}\frac{x^{2j}\left\|D^{2n+1+2j}\right\|}{(2\,j)!}\\&\le \left\|D^{2n+1}\right\|\sum_{j=0}^{\infty}\frac{x^{2j}\left\|D^2\right\|^j}{(2\,j)!} = \left\|D^{2n+1}\right\|\cosh[x\sqrt{\left\|D^2\right\|}]\end{aligned} \tag{47}$$

Δ thus has the bound

$$\begin{aligned}&\left\|\Delta\right\| = \left\|\frac{D^{2n+1}}{(2\,n)!}\int_{-h}^{h}\exp[x\,D](x^2-h^2)^n\,dx\right\| = \left\|\frac{D^{2n+1}}{(2\,n)!}\int_{-h}^{h}\cosh[x\,D](x^2-h^2)^n\,dx\right\|\\&\le \frac{\left\|D^{2n+1}\right\|}{(2\,n)!}\int_{-|h|}^{|h|}\cosh[x\sqrt{\left\|D^2\right\|}](h^2-x^2)^n\,dx \le \frac{\left\|D^{2n+1}\right\|}{(2\,n)!}\cosh[h\sqrt{\left\|D^2\right\|}]\int_{-|h|}^{|h|}(h^2-x^2)^n\,dx\end{aligned} \tag{48}$$

The integral in Eq. (48) reduces to[3]

$$\begin{aligned}&\int_{-|h|}^{|h|}(h^2-x^2)^n\,dx = \int_{-|h|}^{|h|}\sum_{j=0}^{n}\frac{(-1)^j\,n!}{j!(n-j)!}x^{2j}\,h^{2n-2j}\,dx\\&= 2\left(\sum_{j=0}^{n}\frac{(-1)^j\,n!}{(2\,j+1)\,j!(n-j)!}\right)|h|^{2n+1} = \frac{2\,(2\,n)!!}{(2\,n+1)!!}|h|^{2n+1}\end{aligned} \tag{49}$$

With this substitution Eq. (48) simplifies to

$$\left\|\Delta\right\| \le \frac{2\left\|(h\,D)^{2n+1}\right\|\cosh[h\sqrt{\left\|D^2\right\|}]}{(2\,n+1)\,(2\,n-1)!!^2} \tag{50}$$

Eq. (50) can be replaced by a slightly looser bound that does not require explicit calculation of $(h\,D)^{2n+1}$, as outlined in the Appendix (Eq. (59)).

The right side of relation (50) is substituted for $\left\|\Delta\right\|$ in Eq. (46) to determine a bound on $\left\|\delta^{[\mathrm{rel}]}[2\,h]\right\|$, from which a bound on $\left\|\delta^{[\mathrm{rel}]}[x]\right\|$ is determined, with $x = 2^{p+1}\,h$, via Eq. (21). The minimum p required to achieve the specified tolerance bound (Eq. (25)) is found using a bracket-and-bisection algorithm.

The algorithm performance could potentially be improved, in some cases, by applying a balancing similarity transformation to D:

$$D = T\,B\,T^{-1}, \quad \exp[x\,D] = T\exp[x\,B]\,T^{-1} \tag{51}$$

where B has closely-matched row and column norms. (The transformation T can be determined using MATLAB's **`balance`** function [7].)

---

[3] The last equality in Eq. (49) can be verified with the following Mathematica script:

```
s = Sum[(-1)^j n!/(2 j + 1)/j!/(n - j)!, {j, 0, n}];
FullSimplify[(2 n)!!/(2 n + 1)!!/s]
```

References

[1] K. Johnson, *Numerical Solution of Linear, Nonhomogeneous Differential Equation Systems via Padé Approximation* (v8, January 15, 2017). https://vixra.org/abs/1611.0002.

[2] Linear differential equation solver (lde.m) , MATLAB Central File Exchange, 15 Jun 2021. https://www.mathworks.com/matlabcentral/fileexchange/60475-linear-differential-equation-solver-lde-m

[3] **`expm`**, Matrix exponential, 2020. https://www.mathworks.com/help/matlab/ref/expm.html

[4] Al-Mohy, A. H. and N. J. Higham, "A new scaling and squaring algorithm for the matrix exponential," SIAM J. Matrix Anal. Appl., 31(3) (2009), pp. 970–989.

[5] **`pageexpm`**, https://www.mathworks.com/matlabcentral/fileexchange/107959-pageexpm

[6] Gautschi, Walter. *Numerical analysis*. Springer Science & Business Media, 2011, section 5.9.2.

[7] **`balance`**, https://www.mathworks.com/help/matlab/ref/balance.html

Appendix: Polynomial evaluation

The polynomial $P[X]$ (Eq. (10)) is separated into even and odd parts (Eq. (35)),

$$\left.\begin{array}{l} P[X] = P^{[\text{even}]}[X] + P^{[\text{odd}]}[X] \\ P^{[\text{even}]}[X] = \tfrac{1}{2}(P[X] + P[-X]) = \displaystyle\sum_{j=0}^{j=\text{floor}[n/2]} c_j^{[\text{even}]}\,(X^2)^j \\ P^{[\text{odd}]}[X] = \tfrac{1}{2}(P[X] - P[-X]) = X \displaystyle\sum_{j=0}^{j=\text{floor}[(n-1)/2]} c_j^{[\text{odd}]}\,(X^2)^j \end{array}\right\} \tag{52}$$

where

$$c_j^{[\text{even}]} = c_{2j}, \quad c_j^{[\text{odd}]} = c_{2j+1} \tag{53}$$

If n is even, the number of matrix multiplies in Eq. (52) is unchanged when n is incremented by 1, so n is limited to being odd,

$$n = 2m + 1 \tag{54}$$

$$\left.\begin{array}{l} P^{[\text{even}]}[X] = \displaystyle\sum_{j=0}^{j=m} c_j^{[\text{even}]}\,(X^2)^j \\ P^{[\text{odd}]}[X] = X \displaystyle\sum_{j=0}^{j=m} c_j^{[\text{odd}]}\,(X^2)^j \end{array}\right\} \tag{55}$$

The number of matrix multiplies needed to evaluate $P[X]$ via Eq. (10), e.g., using Horner's method, is $2m$. Eq's. (55) would also require $2m$ matrix multiplies if Horner's method is used, but the number of matrix multiplies can be reduced to $m+1$ by precomputing the powers $(X^2)^j$

and using them in both sums. (Horner's method loses its advantage for this application because the scalar-vector products are not computationally expensive.)

The number of matrix multiplies can be further reduced by reorganizing the polynomials as follows,

$$\left.\begin{aligned} P^{[\text{even}]}[X] &= \sum_{k=0}^{k=M-1} \left(\sum_{j=0}^{j=\min(N-1,\, m-N\,k)} c_{j+N\,k}^{[\text{even}]} (X^2)^j \right) ((X^2)^N)^k \\ P^{[\text{odd}]}[X] &= X \sum_{k=0}^{k=M-1} \left(\sum_{j=0}^{j=\min(N-1,\, m-N\,k)} c_{j+N\,k}^{[\text{odd}]} (X^2)^j \right) ((X^2)^N)^k \end{aligned}\right\} \tag{56}$$

where

$$1 < N \le m+1, \quad M = \text{ceil}[(m+1)/N] \ge 1 \tag{57}$$

Horner's method is not used in the inner sums in Eq's. (56), but it is used in the outer sums because the outer sum coefficients are generally matrices. The number of matrix multiplies required for Eq's. (56), denoted as $count$, is

$$\begin{array}{ll} count = & \\ \quad 1 & \text{for } X^2 \text{ factor} \\ \quad +(M>1)\,N & \text{for } (X^2)^2,\, (X^2)^3,\, \ldots,\, (X^2)^N \\ \quad +(M=1)(m-1) & \text{for } (X^2)^2,\, (X^2)^3,\, \ldots,\, (X^2)^m \\ \quad +(M>1)\,2\,(M-1-(m=N\,(M-1))) & \text{for outer sum, Horner's method} \\ \quad +1 & \text{for leading } X \text{ factor in } P^{[\text{odd}]} \end{array} \tag{58}$$

(The computation cost additionally includes the matrix division in Eq. (17) and p multiplies from Eq. (18).) The logical terms in Eq. (58) (e.g., $M>1$) evaluate to 1 or 0 depending on whether the condition is true or false. The $(m=N\,(M-1))$ term is subtracted because under this condition the first step in calculating $P^{[\text{even}]}$ and $P^{[\text{odd}]}$ via Horner's method (for $k=M-1$) multiplies $(X^2)^N$ by a scalar factor (the j sum), not a matrix.

For large m, $count \cong N+2\,(M-1)$ from Eq. (58) with $M \cong m/N$ from Eq. (57). $count$ is approximately minimized with $N \cong 2\,M \cong \sqrt{2\,m}$ and $count \cong 2\sqrt{2\,m}-2$. This is a reduction by a factor of order $\sqrt{m/2}$ in $count$ relative to a straightforward implementation of Eq. (10) (or by a factor of $\frac{1}{2}\sqrt{m/2}$ relative to Eq's. (55)).

N and M are selected to satisfy conditions (57) and to minimize $count$. Within these constraints, N and M are selected to maximize N. A higher N is advantageous for estimating and upper bound on $\|X^{2\,n+1}\|$, as discussed below. Following is a tabulation of m, n, N, M, and $count$ for m up to 15. It would be advantageous to select m from the set $\{1,2,3,4,6,8,10,12,15\}$ because any other m selection in the range $1,\ldots,15$ could be incremented without increasing $count$.

m	n	N	M	*count*
1	3	2	1	2
2	5	3	1	3
3	7	4	1	4
4	9	5	1	5
5	11	6	1	6
6	13	3	3	6
7	15	4	2	7
8	17	4	3	7
9	19	5	2	8
10	21	5	3	8
11	23	6	2	9
12	25	6	3	9
13	27	7	2	10
14	29	7	3	10
15	31	5	4	10

The $\|(h\,D)^{2n+1}\|$ factor in Eq. (50) can be replaced by a looser bound that does not require explicit calculation of $(h\,D)^{2n+1}$. With $X = h\,D$, this factor has the following bound,

$$\|X^{2n+1}\| \le \|X\|\,\|(X^2)^{j_1}\|\,\|(X^2)^{j_2}\|\ldots; \quad j_1 + j_2 + \ldots = n, \quad j_k \le N \tag{59}$$

The powers $(X^2)^{j_1}, (X^2)^{j_2}, \ldots$ are selected from the precomputed matrices X^2, $(X^2)^2$, …, $(X^2)^N$ for Eq's. (56). Each j_k index in Eq. (59) is set to

$$j_k = \min[N, n - \sum_{i=1}^{k-1} j_i] \tag{60}$$