

Mixed-precision Paterson–Stockmeyer Method for Evaluating Matrix Polynomials

[Xiaobo Liu](#), Nicholas J. Higham

Department of Mathematics, The University of Manchester, UK

**29th Biennial Numerical Analysis Conference, Glasgow,
June 28, 2023**

Matrix Polynomials

We want to evaluate the matrix polynomial

$$p_m(X) = \sum_{i=0}^m b_i X^i = b_0 I + b_1 X + b_2 X^2 + \cdots + b_m X^m,$$

where

- $m \in \mathbb{N}$,
- $b_i \in \mathbb{C}$ and **mostly nonzero**,
- $X \in \mathbb{C}^{n \times n}$.

Motivation

- Computation of matrix functions
 - series expansion (Taylor series)
 - rational functions $q(X)^{-1}p(X)$
- Solution of matrix equations

Paterson–Stockmeyer Method

For a positive integer s , we can rewrite (Paterson and Stockmeyer, 1973)

$$p_m(X) = \sum_{i=0}^r B_i \cdot (X^s)^i, \quad r = \lfloor m/s \rfloor,$$

where

$$B_i = \begin{cases} \sum_{j=0}^{s-1} b_{si+j} X^j, & i = 0, \dots, r-1, \\ \sum_{j=0}^{m-sr} b_{sr+j} X^j, & i = r. \end{cases}$$

• $p_m(X)$ is a polyn. in X^s with coefficients B_i : e.g., ($s = 3$),

$$p_6(X) = \underbrace{b_6 I}_{B_2} (\underbrace{X^3}_{X^s})^2 + \underbrace{(b_5 X^2 + b_4 X + b_3 I)}_{B_1} X^3 + \underbrace{(b_2 X^2 + b_1 X + b_0 I)}_{B_0}$$

Paterson–Stockmeyer Method: Evaluation

$$p_m(X) = \left(((B_r X^s + B_{r-1})X^s + B_{r-2})X^s + \cdots + B_1 \right) X^s + B_0$$

Input : $X \in \mathbb{C}^{n \times n}$, $b_0, b_1, \dots, b_m \in \mathbb{C}$

Output: $Z = p_m(X)$

```
1  $\mathcal{X}_0 \leftarrow I, \mathcal{X}_1 \leftarrow X$ 
2 for  $i \leftarrow 2$  to  $s$  do
3   |  $\mathcal{X}_i \leftarrow X\mathcal{X}_{i-1} \triangleright X^2, \dots, X^s$  computed and stored
4 end
5  $Z \leftarrow \sum_{j=0}^{m-sr} b_{sr+j} \mathcal{X}_j$ 
6 for  $i \leftarrow r-1$  down to  $0$  do
7   |  $Z \leftarrow Z\mathcal{X}_s + \sum_{j=0}^{s-1} b_{si+j} \mathcal{X}_j$ 
8 end
9 return  $Z$ 
```

Paterson–Stockmeyer (PS) Method

$$p_m(X) = \left(((B_r X^s + B_{r-1})X^s + B_{r-2})X^s + \cdots + B_1 \right) X^s + B_0$$

- $(s - 1)n^2$ additional storage
- about $s + r - 1$ matrix products (recall that $r = \lfloor m/s \rfloor$)

Theorem (Hargreaves, 2005; Fasi, 2019)

The choice $s = \lfloor \sqrt{m} \rfloor$ or $s = \lceil \sqrt{m} \rceil$ minimizes the number of matrix products required to evaluate $p_m(A)$ over all choices of s . The minimized number of matrix products is about $2\sqrt{m}$.

Exploiting Mutiple Precisions in PS

Practical considerations:

- $\|X\|$ is usually small;
- b_i can decay quickly, e.g., the Taylor series of \exp , \cos .

For PS method

$$p_m(X) = \left(((B_r X^s + B_{r-1})X^s + B_{r-2})X^s + \cdots + B_1 \right) X^s + B_0,$$

can we have $\|B_i\| \|X^s\| \ll \|B_{i-1}\|, i = r: 1$?

Key idea: 1. If $|A| \leq |C|$, $|B| \leq |C|$, and $|A||B| \ll |C|$, computing the product in $AB + C$ in a lower precision than the addition:

$$\text{fl}_{\text{high}}(\text{fl}_{\text{low}}(AB) + C).$$

2. Apply the above idea recursively in the evaluation of p_m .

Exploiting Multi-Precisions in PS: Framework

$$p_m(X) = \underbrace{\left(\underbrace{\left(\underbrace{B_r X^s + B_{r-1}}_{u_r} \right) X^s + B_{r-2}}_{u_{r-1}} \right) X^s + \cdots + B_1}_{u_{r-2}} X^s + B_0.$$

where

$$\|\widehat{B}_i - B_i\| \leq u_i \|B_i\|, \quad i = r: 0, \quad \|\widehat{X}^s - X^s\| \leq u_1 \|X^s\|,$$

and the precisions u_i satisfy ($u = u_0 \ll u_1 \ll \cdots \ll u_r$)

$$u_i \approx \frac{\|B_{i-1}\|}{n \|B_i\| \|X^s\|} u_{i-1}, \quad u_0 = u,$$

such that $\|p_m - \widehat{p}_m\| \lesssim u \|p_m\|$.

The Matrix Multiply-Add in Two Precisions

Theorem 1.

If the relative forward errors in \hat{A} and \hat{B} are u_p , and that in $\hat{C} \in \mathbb{C}^{n \times n}$ is u_s , then for $E \equiv \text{fl}_s(\text{fl}_p(\hat{A}\hat{B}) + \hat{C}) - (AB + C)$,

$$\|E\| \leq \frac{(n+2)u_p + u_s}{1 - ((n+2)u_p + u_s)} \|A\| \|B\| + \frac{2u_s}{1 - 2u_s} \|C\|,$$

where the **matrix product** is done in precision u_p and the **matrix addition** in precision u_s .

- If $\|A\|\|B\| \ll \|C\|$ and $u_s \ll u_p$, to have $\|E\| \lesssim 3u_s \|C\|$,

$$u_s \approx \frac{\|A\| \|B\|}{\|C\|} nu_p,$$

so for moderate n we have $u_s \ll u_p$.

Explicit Powering for B_0 Using Two Precisions

Key idea: For the matrix sum $X_1 + X_2$ in u_h , where $\|X_2\| \ll \|X_1\|$. X_2 can be stored in a lower precision

$$u_\ell \leq \frac{u_h \|X_1 + X_2\|}{(1 + u_h) \|X_2\|} \approx \frac{u_h \|X_1\|}{\|X_2\|}.$$

\tilde{X}_2 : X_2 converted into precision $u_\ell > u_h$, we have

$$\text{fl}_h(X_1 + \tilde{X}_2) = (X_1 + X_2(1 + \delta_\ell))(1 + \delta_h), \quad |\delta_h| \leq u_h, \quad |\delta_\ell| \leq u_\ell,$$

and

$$E := \text{fl}_h(X_1 + \tilde{X}_2) - (X_1 + X_2) = \delta_h(X_1 + X_2) + \delta_\ell(1 + \delta_h)X_2$$

with

$$\|E\| \leq u_h \|X_1 + X_2\| + u_\ell(1 + u_h) \|X_2\| \leq 2u_h \|X_1 + X_2\|.$$

Explicit Powering for B_0 Using Two Precisions

Track the **norm** of $\text{fl}_h(q_j(X)) := \text{fl}_h(b_0I + b_1X + \dots + b_jX^j)$, until, for $j = t$,

$$\frac{u_\ell}{u_h} \lesssim \frac{\|q_t(X)\|}{\|b_{t+1}\| \|X^{t_1}\| \|X^{t_2}\|} \Rightarrow \frac{u_\ell}{u_h} \lesssim \frac{\|q_t(X)\|}{\|b_{t+1}X^{t+1}\|} \approx \frac{\|\text{fl}_h(q_t(X))\|}{\|b_{t+1}X^{t+1}\|},$$

where $t_1 + t_2 = t + 1$.

- Can find the best available t_1, t_2 in **t norm estimations**.

If $\|b_{t+2}X^{t+2}\| \lesssim \|b_{t+1}X^{t+1}\|$, next,

$$\frac{\|q_t(X) + b_{t+1}X^{t+1}\|}{\|b_{t+2}X^{t+2}\|} \gtrsim \frac{\|q_t(X)\| - \|b_{t+1}X^{t+1}\|}{\|b_{t+2}X^{t+2}\|} \gtrsim \frac{u_\ell}{u_h} - 1 \approx \frac{u_\ell}{u_h}.$$

- Can form the rest of the required powers X^{t+1}, \dots, X^{s-1} in precision $u_\ell > u_h$, if

$$\|b_{t+1}X^{t+1}\| \gtrsim \|b_{t+2}X^{t+2}\| \gtrsim \dots \gtrsim \|b_{s-1}X^{s-1}\|.$$

Taylor Approximant of the Matrix Exponential

Theorem 2.

If $\|X\|_1 \leq \sqrt[s]{s!} (\approx s/e + 1)$, for $i = 2:r$ and sufficiently large $s \geq 3$,

$$\frac{\|B_{i-1}\|_1}{\|B_i\|_1 \|X^s\|_1} \gtrsim \left(1 - \frac{1}{ei}\right) i^s.$$

Recall that we want, $\|B_i\| \|X^s\| \ll \|B_{i-1}\|$, $i = r:1$, for the used precisions u_i to be well separated in computing $p_m(X) = \left((B_r X^s + B_{r-1}) X^s + B_{r-2} \right) X^s + \cdots + B_1 \big) X^s + B_0$.

- the ratio $\|B_{i-1}\|_1 / (\|B_i\|_1 \|X^s\|_1)$ tends to increase polynomially as i increases, $i = 2:r$
- Bound not applicable for $\|B_0\|_1 / (\|B_1\|_1 \|X^s\|_1)$

For the Matrix Exponential: the Algorithm

Input : $X \in \mathbb{C}^{n \times n}$, $m \in \mathbb{N}^+$, $u > 0$

Output: A Taylor approximant P of order m for e^X

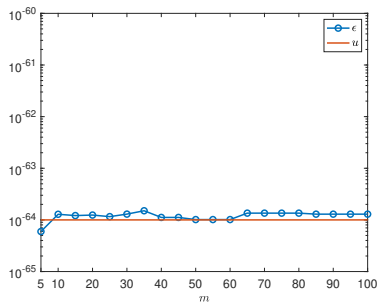
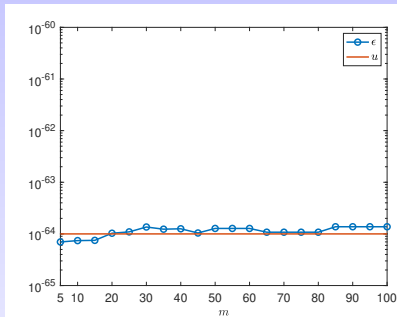
```
1  $s \leftarrow \lceil \sqrt{m} \rceil$ ,  $u_0 \leftarrow u$ ,  $\mathcal{X}_0 \leftarrow I$ ,  $\mathcal{X}_1 \leftarrow X$ 
2 Compute  $B_0$  and  $Y = X^s$  in  $u$  (and potentially  $u_\ell > u$ )
3 while  $(e-1)s! \|B_0\|_1 \leq e\tau \|Y\|_1$  and  $s < m$  do
4   |  $B_0 \leftarrow B_0 + Y/s!$ ,  $s \leftarrow s + 1$ 
5   | Update  $\mathcal{X}_s \leftarrow X\mathcal{X}_{s-1}$  and  $Y \leftarrow \mathcal{X}_s$ 
6 end
7 for  $i \leftarrow 1$  to  $r \leftarrow \lfloor m/s \rfloor$  do
8   | Compute  $B_i$  using elements in  $\mathcal{X}$  and estimate  $\|B_i\|_1$ 
9   | Downgrade  $B_i$  to  $u_i \leftarrow u_{i-1} \|B_{i-1}\|_1 / (n \|B_i\|_1 \|Y\|_1)$ 
10 end
11  $P = B_r$ 
12 for  $i \leftarrow r$  to 1 do
13   | Convert  $Y$  into  $u_i$  and compute  $P \leftarrow PY$  in  $u_i$ 
14   | Form  $P \leftarrow P + B_{i-1}$  in  $u_{i-1}$ 
15 end
16 return  $P$ 
```

The Parameter s and the Cost

- The matrix products in computing B_0 are most expensive: smaller s with larger $r = \lfloor m/s \rfloor$ benefits efficiency
- Smaller s more strict on $\|X\|_1 \leq \sqrt[s]{s!}$
- Larger s more likely being accepted by the algorithm

Overall cost: $\lceil \sqrt{m} \rceil - 1 \leq s - 1 \leq m - 1$ matrix multiplications in precision u and 1 matrix multiplication in each of $u_i > u, i = 1 : r$, where $1 \leq r = \lfloor m/s \rfloor \leq \lceil \sqrt{m} \rceil$.

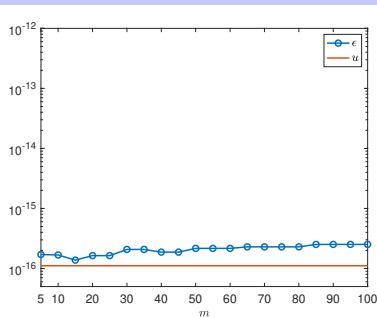
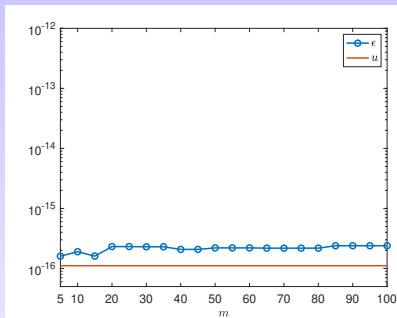
Numerical Experiment Using High Precisions



Left: $X = \text{rand}(n)$. Right: $X = \text{randn}(n)$. $n = 50$

$\|X\|_1 = 1$ and $u = 10^{-64}$ (Simulated by **Advanpix Multiprecision Computing Toolbox**).

Numerical Experiment Using Low Precisions



Left: $X = \text{rand}(n)$. Right: $X = \text{randn}(n)$. $n = 50$

$\|X\|_1 = 1$ and $u = 2^{-53} \approx 1.1 \times 10^{-16}$.

- Only **double**, **single**, and **half** (simulated by `chop`) (**Higham and Pranesh, 2019**) precisions are used.

Numerical Experiment: Approximating $\exp(X)$

Table: The minimal degree m such that the error in approximating the matrix exponential via a Taylor approximant is of order u . d_i represents the equivalent decimal digits of precision u_i .

(u, m)	(s, r)	(d_1, d_2, \dots, d_r)
$(10^{-32}, 32)$	$(6, 5)$	$(30, 26, 20, \mathbf{13}, \mathbf{6})$
$(10^{-64}, 54)$	$(8, 6)$	$(61, 54, 45, 35, \mathbf{24}, \mathbf{13})$
$(10^{-128}, 92)$	$(10, 9)$	$(123, 113, 101, 88, 73, \mathbf{58}, \mathbf{42}, \mathbf{25}, \mathbf{8})$
$(10^{-256}, 158)$	$(13, 12)$	$(248, 234, 217, 198, 178, 156, 133, \mathbf{110}, \mathbf{86}, \mathbf{62}, \mathbf{36}, \mathbf{11})$

$X = \text{gallery}('cauchy', n)$ for $n = 20$ with $\|X\|_1 \approx 2.65$

- The default $s = \lceil \sqrt{m} \rceil$ is chosen in all cases, and **20%** of the matrix products were performed in precision $u^{1/2}$ or much lower.

Numerical Experiment: Approximating $\exp(X)$

For $X_1 = \text{gallery}('forsythe', 10, 1e-10, 0)$ and

$$X_2 = \begin{bmatrix} -0.6 & 0 & 1.2 \\ 0 & -0.6 & 0.45 \\ -2.4 & 4.0 & 0.8 \end{bmatrix},$$

the degree $m = 20$ in double precision (**Fasi and Higham, 2019**).

- The fixed-precision PS: 8 matrix multiplications in double precision.
- The mixed-precision PS: 6 matrix multiplications in double precision and 1 matrix multiplication in single precision and 1 matrix multiplication in half precision.

Concluding Remarks

- Lower precisions can be used in the PS method if $\|X\|$ is small and the coefficients decay quickly.
- The key idea is to perform computations on data of small magnitude (norm) in low precision.

N. J. Higham and X. Liu. [Mixed-precision Paterson–Stockmeyer method for evaluating matrix polynomials](#). Working note.

Proof of Thm. 1

On defining

$$\Delta A = \hat{A} - A, \quad \Delta B = \hat{B} - B, \quad \Delta C = \hat{C} - C,$$

we have

$$\begin{aligned} E &= \text{fl}_s(\text{fl}_p(\hat{A}\hat{B}) + \hat{C}) - (AB + C) \\ &= A\Delta B + \Delta AB + \Delta A\Delta B + E_p + \Delta C + E_s, \end{aligned}$$

where, with $\gamma_n^p := nu_p/(1 - nu_p)$,

$$\begin{aligned} E_p &\leq \gamma_n^p \|A + \Delta A\| \|B + \Delta B\|, \\ E_s &\leq u_s \|(A + \Delta A)(B + \Delta B) + E_p + C + \Delta C\|. \end{aligned}$$

The result follows after straightforward calculation. □

Proof of Thm. 2: I

We have, with $\|X\|_1 =: \sigma \leq \sqrt[s]{s!}$, for $i = 2:r$ and $s \geq 3$,

$$\begin{aligned}
 \frac{\|B_{i-1}\|_1}{\|B_i\|_1 \|Y\|_1} &= \frac{\left\| \frac{1}{((i-1)s)!} I + \frac{1}{((i-1)s+1)!} X + \cdots + \frac{1}{((i-1)s+s-1)!} X^{s-1} \right\|_1}{\left\| \frac{1}{(is)!} I + \frac{1}{(is+1)!} X + \cdots + \frac{1}{(is+s-1)!} X^{s-1} \right\|_1 \|X^s\|_1} \\
 &\geq \frac{\frac{1}{((i-1)s)!} - \left(\frac{\sigma}{((i-1)s+1)!} + \frac{\sigma^2}{((i-1)s+2)!} + \cdots + \frac{\sigma^{s-1}}{((i-1)s+s-1)!} \right)}{\left(\frac{1}{(is)!} + \frac{\sigma}{(is+1)!} + \cdots + \frac{\sigma^{s-1}}{(is+s-1)!} \right) s!} \\
 &\geq \frac{\frac{1}{((i-1)s)!} - \frac{\sigma}{((i-1)s+1)!} \left(1 + \frac{\sigma}{(i-1)s+2} + \cdots + \frac{\sigma^{s-2}}{((i-1)s+2)^{s-2}} \right)}{\frac{1}{(is)!} \left(1 + \frac{\sigma}{is+1} + \cdots + \frac{\sigma^{s-1}}{(is+1)^{s-1}} \right) s!} \\
 &=: \gamma(s).
 \end{aligned}$$

Proof of Thm. 2: II

On the other hand, we know from Stirling's approximation

$$\frac{\sigma}{s} \leq \frac{\sqrt[s]{s!}}{s} \sim \frac{\sqrt[2s]{2\pi s}}{e} \rightarrow e^{-1}, \quad s \rightarrow \infty,$$

which says σ grows at most (linearly) like $e^{-1}s$ for sufficiently large s . Therefore, we have, for sufficiently large s ,

$$\begin{aligned} \gamma(s) &= \frac{\frac{1}{((i-1)s)!} - \frac{\sigma}{((i-1)s+1)!} \cdot \frac{1-(\sigma/((i-1)s+2))^{s-1}}{1-\sigma/((i-1)s+2)}}{\frac{s!}{(is)!} \cdot \frac{1-(\sigma/(is+1))^s}{1-\sigma/(is+1)}} \sim \frac{(is)! \left(1 - \frac{\sigma}{is+1}\right)}{s!(is-s)!} \\ &\gtrsim \left(1 - \frac{1}{ei}\right) \binom{is}{s} \geq \left(1 - \frac{1}{ei}\right) \frac{(is)^s}{s^s} = \left(1 - \frac{1}{ei}\right) i^s. \quad \square \end{aligned}$$

References I



Massimiliano Fasi.

Optimality of the Paterson–Stockmeyer method for evaluating matrix polynomials and rational matrix functions.

Linear Algebra Appl., 574:182–200, 2019.






Massimiliano Fasi and Nicholas J. Higham.

An arbitrary precision scaling and squaring algorithm for the matrix exponential.

SIAM. J. Matrix Anal. Appl., 39(1):472–491, 2018.

References II

-  Gareth Hargreaves.
Topics in matrix computations: Stability and efficiency of algorithms.
PhD thesis, University of Manchester, Manchester, England, August 2005, 204 pp.
-  Nicholas J. Higham and Srikara Pranesh.
Simulating low precision floating-point arithmetic
SIAM J. Sci. Comput., 41(5):C585–C602, 2019.
-  Michael S. Paterson and Larry J. Stockmeyer.
On the number of nonscalar multiplications necessary to evaluate polynomials
SIAM J. Comput., 2(1):60–66, 1973.