Mixed-precision Paterson–Stockmeyer Method for Evaluating Matrix Polynomials

Xiaobo Liu, Nicholas J. Higham

Department of Mathematics, The University of Manchester, UK

10th International Congress on Industrial and Applied Mathematics, Waseda University, Tokyo, August 22, 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

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

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

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

Paterson–Stockmeyer Method: Evaluation

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

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

Output: $Z = p_m(X)$

$\mathcal{X}_0 \leftarrow I$, $\mathcal{X}_1 \leftarrow X$

for $i \leftarrow 2$ **to** s **do**

  $\mathcal{X}_i \leftarrow X \mathcal{X}_{i-1}$ ▷ *$X^2, \ldots, X^s$ computed and stored*

end

$Z \leftarrow \sum_{j=0}^{m-sr} b_{sr+j} \mathcal{X}_j$

for $i \leftarrow r - 1$ **down to** 0 **do**

  $Z \leftarrow Z \mathcal{X}_s + \sum_{j=0}^{s-1} b_{si+j} \mathcal{X}_j$

end

return Z

Paterson–Stockmeyer (PS) Method

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

- $(s-1)n^2$ 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

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

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

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

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

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

Exploiting Multi-Precisions in PS: Framework

$$p_m(X) = \Big(\big(\overbrace{(\overbrace{B_r X^s}^{u_r} + B_{r-1})}^{u_{r-1}} X^s + B_{r-2} \big) X^s + \cdots + B_1 \Big) X^s + B_0.$$

$$\underbrace{(\underbrace{B_r X^s + B_{r-1}}_{u_{r-1}}) X^s + B_{r-2}}_{u_{r-2}}$$

where we require

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

and the precisions u_i are chosen by

$$u_i = \frac{\|B_0\|}{\|B_i\| \, \|X^s\|^i} u_0, \quad i = 1:r,$$

which means $u = u_0 \ll u_1 \ll \cdots \ll u_r$ since

$$\frac{u_i}{u_{i-1}} = \frac{\|B_{i-1}\|}{\|B_i\| \, \|X^s\|} \gg 1, \quad i = 1:r.$$

Exploiting Multi-Precisions in PS: Framework

If $\|B_i\| \, \|X^s\| \le \tau \, \|B_{i-1}\|$, $i = r\colon 1$ for some $\tau \ll 1$ (by choosing a suitable s), we then have **(Higham and L, Working note)**

$$\|\widehat{p}_m - p_m(X)\| \lesssim rnu \, \|p_m(X)\| \,,$$

where $r = \lfloor m/s \rfloor$.

• Do we have $\|\widehat{B}_i - B_i\| \le u_i \, \|B_i\| \,, \; i = r\colon 0$ and $\|\widehat{X^s} - X^s\| \le u_1 \, \|X^s\|$?

1. Form $\mathcal{X} = \{X^2, X^3, \dots, X^s\}$ in u_0 (note $u_0 \ll u_1$).

2. Compute B_i using the powers in $\mathcal{X}$ and downgrade B_i to u_i (after estimating $\|B_i\|$).

Question: Is it possible to use u_0 and a lower precision $u_\ell > u_0$ in forming the powers in $\mathcal{X}$?

Explicit Powering for B_0 Using Two Precisions

Key idea: For the matrix sum $X_1 + X_2$ in u_h (in our case $u_h = u_0$), where $\|X_2\| \ll \|X_1\|$. X_2 can be stored in a lower precision

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

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

$$\mathrm{fl}_h(X_1 + \widetilde{X}_2) = (X_1 + X_2(1+\delta_\ell))(1+\delta_h), \ |\delta_h| \le u_h, \ |\delta_\ell| \le u_\ell,$$

and

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

with

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

Explicit Powering for B_0 Using Two Precisions

Track the norm of $\mathrm{fl}_h(q_j(X)) := \mathrm{fl}_h(b_0 I + b_1 X + \cdots + b_j X^j)$, until, for $j = t$,

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

where $t_1 + t_2 = t + 1$.

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

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

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

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

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

Taylor Approximant of the Matrix Exponential

Theorem 1.

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

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

Recall that we need to choose s such that $\|B_i\| \, \|X^s\| \le \tau \, \|B_{i-1}\|$, $i = r\colon 1$ **for some** $\tau \ll 1$ in computing

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

- For a fixed $s \ge 3$, the ratio $\|B_{i-1}\|_1/(\|B_i\|_1 \|X^s\|_1)$ tends to increase polynomially as i increases, $i = 2\colon r$.

- 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

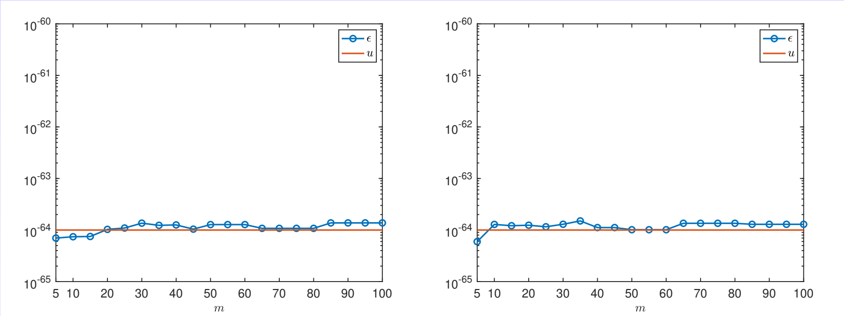
$s \leftarrow \lceil \sqrt{m} \rceil$, $u_0 \leftarrow u$, $\mathcal{X}_0 \leftarrow I$, $\mathcal{X}_1 \leftarrow X$
Compute B_0 and $Y = X^s$ in u (and potentially $u_\ell > u$)
while $(\mathrm{e}-1)s!\|B_0\|_1 \le \mathrm{e}\tau\|Y\|_1$ **and** $s < m$ **do**
    $B_0 \leftarrow B_0 + Y/s!$, $s \leftarrow s+1$
    Update $\mathcal{X}_s \leftarrow XY$ and $Y \leftarrow \mathcal{X}_s$
end
for $i \leftarrow 1$ **to** $r \leftarrow \lfloor m/s \rfloor$ **do**
    Compute B_i using elements in $\mathcal{X}$ and estimate $\|B_i\|_1$
    Downgrade B_i to $u_i \leftarrow u_{i-1}\|B_{i-1}\|_1/(\|B_i\|_1\|Y\|_1)$
end
$P = B_r$
for $i \leftarrow r$ **to** 1 **do**
    Convert Y into u_i and compute $P \leftarrow PY$ in u_i
    Form $P \leftarrow P + B_{i-1}$ in u_{i-1}
end
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 puts a more strict requirement: $\|X\|_1 \le \sqrt[s]{s!}$
- A larger s is more likely to be accepted by the algorithm

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

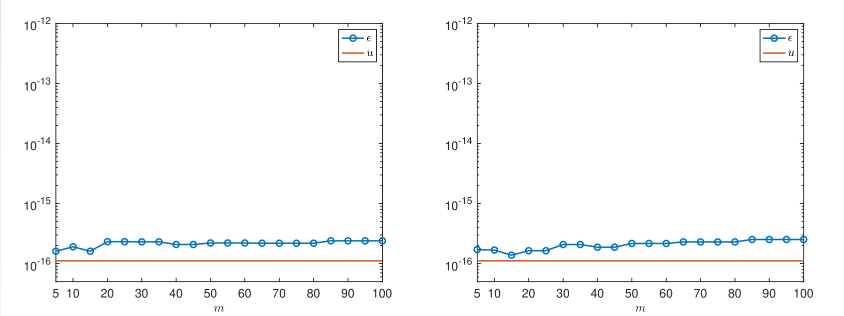
Numerical Experiment Using High Precisions



Left: $X =$ `rand(n)`. Right: $X =$ `randn(n)`. $n = 50$

$\|X\|_1 = 1$, $u = 10^{-64}$ (Simulated by **Advanpix Multiprecision Computing Toolbox**), and $\epsilon = \left\|\widehat{p}_m - p_m(X)\right\| / \|p_m(X)\|$.

Numerical Experiment Using Low Precisions



Left: $X =$ `rand(n)`. Right: $X =$ `randn(n)`. $n = 50$

$\|X\|_1 = 1$, $u = 2^{-53} \approx 1.1 \times 10^{-16}$, and
$\epsilon = \left\|\widehat{p}_m - p_m(X)\right\| / \|p_m(X)\|$.

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

$X =$ `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.

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: I

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

$$\frac{\|B_{i-1}\|_1}{\|B_i\|_1\|Y\|_1} = \frac{\left\| \frac{1}{((i-1)s)!}I + \frac{1}{((i-1)s+1)!}X + \cdots + \frac{1}{((i-1)s+s-1)!}X^{s-1} \right\|_1}{\left\| \frac{1}{(is)!}I + \frac{1}{(is+1)!}X + \cdots + \frac{1}{(is+s-1)!}X^{s-1} \right\|_1 \|X^s\|_1}$$

$$\ge \frac{\frac{1}{((i-1)s)!} - \left(\frac{\sigma}{((i-1)s+1)!} + \frac{\sigma^2}{((i-1)s+2)!} + \cdots + \frac{\sigma^{s-1}}{((i-1)s+s-1)!} \right)}{\left(\frac{1}{(is)!} + \frac{\sigma}{(is+1)!} + \cdots + \frac{\sigma^{s-1}}{(is+s-1)!} \right) s!}$$

$$\ge \frac{\frac{1}{((i-1)s)!} - \frac{\sigma}{((i-1)s+1)!} \left(1 + \frac{\sigma}{(i-1)s+2} + \cdots + \frac{\sigma^{s-2}}{((i-1)s+2)^{s-2}} \right)}{\frac{1}{(is)!} \left(1 + \frac{\sigma}{is+1} + \cdots + \frac{\sigma^{s-1}}{(is+1)^{s-1}} \right) s!}$$

$$=: \gamma(s).$$

Proof of Thm. 1: II

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

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

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

$$\gamma(s) = \frac{\frac{1}{((i-1)s)!} - \frac{\sigma}{((i-1)s+1)!} \cdot \frac{1-(\sigma/((i-1)s+2))^{s-1}}{1-\sigma/((i-1)s+2)}}{\frac{s!}{(is)!} \cdot \frac{1-(\sigma/(is+1))^{s}}{1-\sigma/(is+1)}} \sim \frac{(is)!\left(1 - \frac{\sigma}{is+1}\right)}{s!(is-s)!}$$

$$\gtrsim \left(1 - \frac{1}{\mathrm{e}i}\right)\binom{is}{s} \geq \left(1 - \frac{1}{\mathrm{e}i}\right)\frac{(is)^s}{s^s} = \left(1 - \frac{1}{\mathrm{e}i}\right)i^s. \qquad \square$$

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.

- 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.

References II

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.