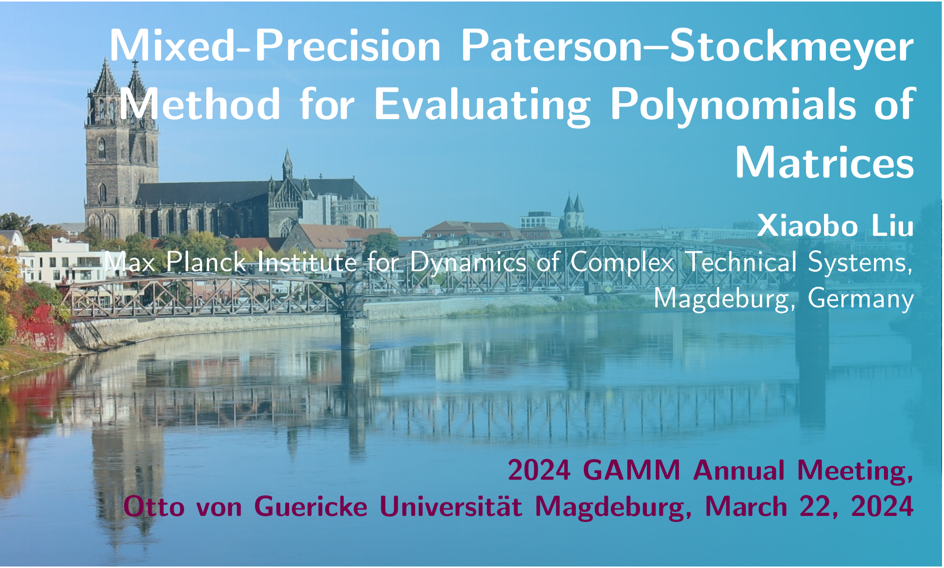
MAX PLANCK INSTITUTE
FOR DYNAMICS OF COMPLEX
TECHNICAL SYSTEMS
MAGDEBURG

CSC

COMPUTATIONAL METHODS IN
SYSTEMS AND CONTROL THEORY



Mixed-Precision Paterson–Stockmeyer Method for Evaluating Polynomials of Matrices

Xiaobo Liu
Max Planck Institute for Dynamics of Complex Technical Systems,
Magdeburg, Germany

2024 GAMM Annual Meeting,
Otto von Guericke Universität Magdeburg, March 22, 2024

Matrix Polynomials

The object and Motivation

The goal is to evaluate the matrix polynomial

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

It often results from truncated series expansions (with $\|b_m X^m\| \leq \epsilon \ll 1$) in computation of matrix functions and solution of matrix equations:

- series expansion (e.g., Taylor series)
- rational functions $q(X)^{-1} p(X)$
- rational matrix equations $r(X) = A$

So, practically,

- $m \in \mathbb{N}$,
- $b_i \in \mathbb{C}$ and $|b_i|$ can decay quickly, e.g., the Taylor series of $\exp$, $\cos$
- $X \in \mathbb{C}^{n\times n}$ with $\|X\|$ usually being small.

Matrix Polynomials

The object and Motivation

The goal is to evaluate the matrix polynomial

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

It often results from truncated series expansions (with $\|b_m X^m\| \leq \epsilon \ll 1$) in computation of matrix functions and solution of matrix equations:

- series expansion (e.g., Taylor series)
- rational functions $q(X)^{-1}p(X)$
- rational matrix equations $r(X) = A$

So, practically,

- $m \in \mathbb{N}$,
- $b_i \in \mathbb{C}$ and $|b_i|$ can decay quickly, e.g., the Taylor series of $\exp$, $\cos$
- $X \in \mathbb{C}^{n\times n}$ with $\|X\|$ usually being small.

Paterson–Stockmeyer Method

For $s \in \mathbb{N}^{+}$, we can rewrite $p_m(X)$ as a polynomial in X^s with matrix coefficients B_i **(Paterson and Stockmeyer, 1973)**

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

where

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

- For example, with $m = 6$ and $s = 3$,

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

Paterson–Stockmeyer Method

For $s \in \mathbb{N}^+$, we can rewrite $p_m(X)$ as a polynomial in X^s with matrix coefficients B_i **(Paterson and Stockmeyer, 1973)**

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

where

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

- For example, with $m = 6$ and $s = 3$,

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

Paterson–Stockmeyer Method

Evaluation

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

```
Input  : $X \in \mathbb{C}^{n\times n}$, $b_0, b_1, \dots, b_m \in \mathbb{C}$
Output: $Z = p_m(X)$
$\mathcal{X}_0 \leftarrow I$, $\mathcal{X}_1 \leftarrow X$
for $i \leftarrow 2$ to $s$ do
    $\mathcal{X}_i \leftarrow X\mathcal{X}_{i-1}$        ▷ $X^2, \dots, X^s$ computed and stored
$Z \leftarrow \sum_{j=0}^{m-sr} b_{sr+j}\mathcal{X}_j$
for $i \leftarrow r-1$ down to $0$ do
    $Z \leftarrow Z\mathcal{X}_s + \sum_{j=0}^{s-1} b_{si+j}\mathcal{X}_j$
return $Z$
```

- Two extreme cases: (i) $s = 1$: (plain) Horner's method
  (ii) $s = m$: evaluation via explicit powers.

Paterson–Stockmeyer Method

Evaluation

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

```
Input  : $X \in \mathbb{C}^{n\times n}$, $b_0, b_1, \dots, b_m \in \mathbb{C}$
Output: $Z = p_m(X)$
$\mathcal{X}_0 \leftarrow I$, $\mathcal{X}_1 \leftarrow X$
for $i \leftarrow 2$ to $s$ do
    $\mathcal{X}_i \leftarrow X\mathcal{X}_{i-1}$        ▷ $X^2, \dots, X^s$ computed and stored
$Z \leftarrow \sum_{j=0}^{m-sr} b_{sr+j}\mathcal{X}_j$
for $i \leftarrow r-1$ down to $0$ do
    $Z \leftarrow Z\mathcal{X}_s + \sum_{j=0}^{s-1} b_{si+j}\mathcal{X}_j$
return $Z$
```

- Two extreme cases: (i) $s = 1$: (plain) Horner's method
  (ii) $s = m$: evaluation via explicit powers.

Paterson–Stockmeyer Method

Evaluation

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

```
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
$Z \leftarrow \sum_{j=0}^{m-sr} b_{sr+j}\mathcal{X}_j$
for $i \leftarrow r-1$ down to $0$ do
    $Z \leftarrow Z\mathcal{X}_s + \sum_{j=0}^{s-1} b_{si+j}\mathcal{X}_j$
return $Z$
```

- Two extreme cases: (i) $s = 1$: (plain) Horner's method
  (ii) $s = m$: evaluation via explicit powers.

Paterson–Stockmeyer Method

Storage Requirement and Cost

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

- $(s+2)n^2$ elements of storage
- about $s-1+r$ matrix products, incl. $r = \lfloor m/s \rfloor$ products in the Horner's stage

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 Multi-Precisions in Paterson–Stockmeyer

Observation and Key Idea

For $p_m(X) = \Big(\big((B_r X^s + B_{r-1})X^s + B_{r-2}\big)X^s + \cdots + B_1\Big)X^s + B_0$,
$\|B_i\|\,\|X^s\| \ll \|B_{i-1}\|$ can hold for some $i = v\colon r$,

$$\Big\|b_{si}I + b_{si+1}X + \cdots + b_{si+s-1}X^{s-1}\Big\|\,\|X^s\| \ll \Big\|b_{si-s}I + b_{si-s+1}X + \cdots + b_{si-1}X^{s-1}\Big\|.$$

Intuition: dominant terms in B_i and B_{i-1} have scalar coefficients being s indices apart from $\{b_i\}$. Consider $X = \left[\begin{smallmatrix} -1 & 1 \\ 2 & 1 \end{smallmatrix}\right]$ with $b_i = 1/i!$ and $s = 6$,

$$\|B_2\|_1 \|X^s\|_1 \approx \left\|\frac{1}{12!}I + \frac{1}{13!}X\right\|_1 \|X^s\|_1 = 6.5 \times 10^{-8}$$
$$\ll 1.8 \times 10^{-3} = \left\|\frac{1}{6!}I + \frac{1}{7!}X\right\|_1 \approx \|B_1\|_1.$$

Idea for Utilizing Multi-Precisions

$\mathrm{fl}(AB + C) = \mathrm{fl}_{high}(\mathrm{fl}_{low}(AB) + C)$ for $|A||B| \ll |C|$ and do this recursively in the evaluation of p_m.

Exploiting Multi-Precisions in Paterson–Stockmeyer
Observation and Key Idea

For $p_m(X) = \Big(\big((B_r X^s + B_{r-1})X^s + B_{r-2}\big)X^s + \cdots + B_1\Big)X^s + B_0$,
$\|B_i\|\,\|X^s\| \ll \|B_{i-1}\|$ can hold for some $i = v\colon r$,

$$\left\|b_{si}I + b_{si+1}X + \cdots + b_{si+s-1}X^{s-1}\right\|\,\|X^s\| \ll \left\|b_{si-s}I + b_{si-s+1}X + \cdots + b_{si-1}X^{s-1}\right\|.$$

Intuition: dominant terms in B_i and B_{i-1} have scalar coefficients being s indices apart from $\{b_i\}$. Consider $X = \left[\begin{smallmatrix} -1 & 1 \\ 2 & 1 \end{smallmatrix}\right]$ with $b_i = 1/i!$ and $s = 6$,

$$\|B_2\|_1\|X^s\|_1 \approx \left\|\frac{1}{12!}I + \frac{1}{13!}X\right\|_1 \|X^s\|_1 = 6.5\times 10^{-8}$$
$$\ll 1.8\times 10^{-3} = \left\|\frac{1}{6!}I + \frac{1}{7!}X\right\|_1 \approx \|B_1\|_1.$$

Idea for Utilizing Multi-Precisions

$\mathrm{fl}(AB + C) = \mathrm{fl}_{high}(\mathrm{fl}_{low}(AB) + C)$ for $|A||B| \ll |C|$ and do this recursively in the evaluation of p_m.

Exploiting Multi-Precisions in Paterson–Stockmeyer
Observation and Key Idea

For $p_m(X) = \Big(\big((B_r X^s + B_{r-1})X^s + B_{r-2}\big)X^s + \cdots + B_1\Big)X^s + B_0$,
$\|B_i\|\,\|X^s\| \ll \|B_{i-1}\|$ can hold for some $i = v\colon r$,

$$\left\|b_{si}I + b_{si+1}X + \cdots + b_{si+s-1}X^{s-1}\right\|\,\|X^s\| \ll \left\|b_{si-s}I + b_{si-s+1}X + \cdots + b_{si-1}X^{s-1}\right\|.$$

Intuition: dominant terms in B_i and B_{i-1} have scalar coefficients being s indices apart from $\{b_i\}$. Consider $X = \left[\begin{smallmatrix} -1 & 1 \\ 2 & 1 \end{smallmatrix}\right]$ with $b_i = 1/i!$ and $s = 6$,

$$\|B_2\|_1\|X^s\|_1 \approx \left\|\frac{1}{12!}I + \frac{1}{13!}X\right\|_1 \|X^s\|_1 = 6.5 \times 10^{-8}$$
$$\ll 1.8 \times 10^{-3} = \left\|\frac{1}{6!}I + \frac{1}{7!}X\right\|_1 \approx \|B_1\|_1.$$

Idea for Utilizing Multi-Precisions

$\mathrm{fl}(AB + C) = \mathrm{fl}_{high}(\mathrm{fl}_{low}(AB) + C)$ for $|A||B| \ll |C|$ and do this recursively in the evaluation of p_m.

Exploiting Multi-Precisions in Paterson–Stockmeyer Framework

Given precisions $u_r \geq u_{r-1} \geq \cdots \geq u_v \geq u$, we compute

$$q_v(X) := \Big(\big(\big(\underbrace{\overbrace{\underbrace{\overbrace{B_r X^s}^{u_r} + B_{r-1}}_{u_{r-1}} \big) X^s}^{u_{r-1}} + B_{r-2}}_{u_{r-2}} \big) X^s + \cdots + B_v\Big) X^s$$

in the lower-than-working precisions and

$$p_m(X) = \Bigg(\Big(\big((q_v(X) + B_{v-1})X^s + B_{v-2}\big)X^s + \cdots + B_1\Big)X^s + B_0$$

in the working precision u.

Exploiting Multi-Precisions in Paterson–Stockmeyer
Error Bound

Evaluation: $$q_v(X) = \Big(\underbrace{\overbrace{\big((\underbrace{\overbrace{B_r X^s}^{u_r} + B_{r-1}}_{u_{r-1}})X^s}^{u_{r-1}} + B_{r-2}}_{u_{r-2}}\big)X^s + \cdots + B_v\Big)X^s.$$

Theorem (Error bound for $q_v(X)$)

Given $\|B_i\|\,\|X^s\| = \tau_i\,\|B_{i-1}\|$ for some $\tau_i \ll 1$, $\|\widehat{B}_i - B_i\| \le u_i\,\|B_i\|$ for $i = v\colon r$, and $\|\mathrm{fl}(X^s) - X^s\| \le u_v\,\|X^s\|$, then by setting the precisions $u_{v-1} \equiv u$ and

$$u_i = u_{i-1}/\tau_i, \quad i = v\colon r,$$

(so $u \ll u_v \ll \cdots \ll u_r$) we have

$$\|\widehat{q}_v - q_v(X)\| \lesssim (r - v + 1)nu\,\|q_v(X)\|\,,$$

where $r = \lfloor m/s \rfloor$ (assuming $((1 + \max_i \tau_i)n + 2)\,\|q_v(X)\|\,u \ll 1$).

Theorem (Error bound for $q_v(X)$)

Given $\|B_i\|\,\|X^s\| = \tau_i \|B_{i-1}\|$ for some $\tau_i \ll 1$, $\|\widehat{B}_i - B_i\| \le u_i \|B_i\|$[i] for $i = v\colon r$, and $\|\mathrm{fl}(X^s) - X^s\| \le u_v \|X^s\|$[ii], then by setting the precisions $u_{v-1} \equiv u$ and

$$u_i = u_{i-1}/\tau_i, \quad i = v\colon r,$$

(so $u \ll u_v \ll \cdots \ll u_r$) we have

$$\|\widehat{q}_v - q_v(X)\| \lesssim (r - v + 1)nu \,\|q_v(X)\|\,,$$

where $r = \lfloor m/s \rfloor$ (assuming $((1 + \max_i \tau_i)n + 2)\,\|q_v(X)\|\,u \ll 1$).

- If $v = 1$ and $\|\widehat{B}_0 - B_0\| \le cnu\,\|B_0\|$, $\|\widehat{p}_m - p_m(X)\| \lesssim rnu\,\|p_m(X)\|$.

[i] The required powers $X^2, \ldots, X^s$ are formed in the working precision u for the accuracy of $\widehat{B}_0$.

[ii] From standard analysis $|\,\mathrm{fl}(X^s) - X^s| \lesssim snu|X|^s$, so the condition holds if $sn\tau_v\,\|X\|^s \lesssim \|X^s\|$, or, $\|X^s\|$ not much less than $\|X\|^s$.

Exploiting Multi-Precisions in Paterson–Stockmeyer
Error Bound

Theorem (Error bound for $q_v(X)$)

Given $\|B_i\| \, \|X^s\| = \tau_i \|B_{i-1}\|$ for some $\tau_i \ll 1$, $\|\widehat{B}_i - B_i\| \le u_i \|B_i\|$[i] for $i = v\colon r$, and $\|\mathrm{fl}(X^s) - X^s\| \le u_v \|X^s\|$[ii], then by setting the precisions $u_{v-1} \equiv u$ and

$$u_i = u_{i-1}/\tau_i, \quad i = v\colon r,$$

(so $u \ll u_v \ll \cdots \ll u_r$) we have

$$\|\widehat{q}_v - q_v(X)\| \lesssim (r - v + 1)nu \, \|q_v(X)\| \, ,$$

where $r = \lfloor m/s \rfloor$ (assuming $((1 + \max_i \tau_i)n + 2) \, \|q_v(X)\| \, u \ll 1$).

- If $v = 1$ and $\|\widehat{B}_0 - B_0\| \le cnu \, \|B_0\|$, $\|\widehat{p}_m - p_m(X)\| \lesssim rnu \, \|p_m(X)\|$.

i The required powers $X^2, \dots, X^s$ are formed in the working precision u for the accuracy of $\widehat{B}_0$.

ii From standard analysis $|\mathrm{fl}(X^s) - X^s| \lesssim snu|X|^s$, so the condition holds if $sn\tau_v \, \|X\|^s \lesssim \|X^s\|$, or, $\|X^s\|$ not much less than $\|X\|^s$.

Theorem (Error bound for $q_v(X)$)

Given $\|B_i\|\,\|X^s\| = \tau_i \|B_{i-1}\|$ for some $\tau_i \ll 1$, $\|\widehat{B}_i - B_i\| \le u_i \|B_i\|$[i] for $i = v\colon r$, and $\|\mathrm{fl}(X^s) - X^s\| \le u_v \|X^s\|$[ii], then by setting the precisions $u_{v-1} \equiv u$ and

$$u_i = u_{i-1}/\tau_i, \quad i = v\colon r,$$

(so $u \ll u_v \ll \cdots \ll u_r$) we have

$$\|\widehat{q}_v - q_v(X)\| \lesssim (r - v + 1)nu \|q_v(X)\|,$$

where $r = \lfloor m/s \rfloor$ (assuming $((1 + \max_i \tau_i)n + 2)\|q_v(X)\| u \ll 1$).

- If $v = 1$ and $\|\widehat{B}_0 - B_0\| \le cnu \|B_0\|$, $\|\widehat{p}_m - p_m(X)\| \lesssim rnu \|p_m(X)\|$.

[i] The required powers $X^2, \dots, X^s$ are formed in the working precision u for the accuracy of $\widehat{B}_0$.

[ii] From standard analysis $|\,\mathrm{fl}(X^s) - X^s| \lesssim snu|X|^s$, so the condition holds if $sn\tau_v \|X\|^s \lesssim \|X^s\|$, or, $\|X^s\|$ not much less than $\|X\|^s$.

Mixed-Precision Paterson–Stockmeyer

Bounds for Taylor Approximants of e^X

- For the error in $\widehat{B}_0 \approx B_0(X) = \sum_{j=0}^{s-1} b_j X^j$, standard error analysis implies

$$\left\| \widehat{B}_0 - B_0(X) \right\| \le \gamma_{(s-2)n+2} B_0(\|X\|) \approx \gamma_{(s-2)n+2} \mathrm{e}^{\|X\|}, \quad \gamma_n := \frac{nu}{1-nu},$$

then using $1 \le \left\|\mathrm{e}^X\right\| \left\|\mathrm{e}^{-X}\right\| \le \left\|\mathrm{e}^X\right\| \mathrm{e}^{\|X\|}$,

$$\|\widehat{B}_0 - B_0(X)\| \lesssim \gamma_{(s-2)n+2} \mathrm{e}^{\|X\|} \mathrm{e}^{\|X\|} \left\|\mathrm{e}^X\right\| \approx \mathrm{e}^{2\|X\|} snu \left\|B_0(X)\right\|.$$

- A sufficient condition for $\|\mathrm{fl}(X^s) - X^s\| \le u_v \|X^s\|$ is $sn\tau_v \|X\|^s \lesssim \|X^s\|$, one can show

$$\frac{sn\tau_v \|X\|_1^s}{\|X^s\|_1} = \frac{sn\|B_v\|_1 \|X\|_1^s}{\|B_{v-1}\|_1} \lesssim \begin{cases} sn\mathrm{e}^{\|X\|_1}, & v = 1, \\ sn/\binom{vs}{s}, & v > 1, \end{cases}$$

with the asumption $\|X\|_1 \le s/\mathrm{e}$.

Mixed-Precision Paterson–Stockmeyer

Bounds for Taylor Approximants of e^X

- For the error in $\widehat{B}_0 \approx B_0(X) = \sum_{j=0}^{s-1} b_j X^j$, standard error analysis implies

$$\left\|\widehat{B}_0 - B_0(X)\right\| \le \gamma_{(s-2)n+2} B_0(\|X\|) \approx \gamma_{(s-2)n+2} \mathrm{e}^{\|X\|}, \quad \gamma_n := \frac{nu}{1-nu},$$

then using $1 \le \left\|\mathrm{e}^X\right\| \left\|\mathrm{e}^{-X}\right\| \le \left\|\mathrm{e}^X\right\| \mathrm{e}^{\|X\|}$,

$$\|\widehat{B}_0 - B_0(X)\| \lesssim \gamma_{(s-2)n+2} \mathrm{e}^{\|X\|} \mathrm{e}^{\|X\|} \left\|\mathrm{e}^X\right\| \approx \mathrm{e}^{2\|X\|} snu \left\|B_0(X)\right\|.$$

- A sufficient condition for $\|\mathrm{fl}(X^s) - X^s\| \le u_v \|X^s\|$ is $sn\tau_v \|X\|^s \lesssim \|X^s\|$, one can show

$$\frac{sn\tau_v \|X\|_1^s}{\|X^s\|_1} = \frac{sn\|B_v\|_1 \|X\|_1^s}{\|B_{v-1}\|_1} \lesssim \begin{cases} sn\mathrm{e}^{\|X\|_1}, & v = 1, \\ sn/\binom{vs}{s}, & v > 1, \end{cases}$$

with the asumption $\|X\|_1 \le s/\mathrm{e}$.

Mixed-Precision Paterson–Stockmeyer

The General Algorithm

```
Input   : $X \in \mathbb{C}^{n\times n}$, $\{b_i\}_{i=0}^{m} \subset \mathbb{C}$
Output: $P \approx p_m(X)$
$s \leftarrow \lceil \sqrt{m} \rceil,\ r \leftarrow \lfloor m/s \rfloor,\ v \leftarrow r+1$
Compute $\mathcal{X} := \{X^i\}_{i=2}^{s}$ and $B_0$ in precision $u \equiv u_0$
for $i \leftarrow 1$ to $r$ do
    Assemble $B_i$ using elements in $\mathcal{X} \cup \{I, X\}$ and estimate $\|B_i\|_1$
    $u_i \leftarrow \|B_{i-1}\|_1 u_{i-1} / (\|B_i\|_1 \|X^s\|_1)$        ▷ $u_i = u_{i-1}/\tau_i,\ \tau_i \ll 1$
$v \leftarrow \min\{i \colon u_i \geq \delta u\},\ u_{v-1}, u_{v-2}, \dots, u_1 \leftarrow u,\ P \leftarrow B_r$
for $i \leftarrow r$ down to $1$ do
    Compute $P \leftarrow PX^s$ in precision $u_i$
    Form $P \leftarrow P + B_{i-1}$ in precision $u_{i-1}$
return $P$
```

- need store $\{X^i\}_{i=1}^{s}$ and $\{B^i\}_{i=0}^{r}$: about $2sn^2$ elements of storage
- $s+v-2$ matrix products in u and 1 in each of $u_v, u_{v+1}, \dots, u_r$.

- How practical is the algorithm (are the conditions $\tau_i \ll 1$, $i = v \colon r$)?

Mixed-Precision Paterson–Stockmeyer

The General Algorithm

```
Input   : $X \in \mathbb{C}^{n\times n}$, $\{b_i\}_{i=0}^{m} \subset \mathbb{C}$
Output: $P \approx p_m(X)$
$s \leftarrow \lceil \sqrt{m} \rceil$, $r \leftarrow \lfloor m/s \rfloor$, $v \leftarrow r+1$
Compute $\mathcal{X} := \{X^i\}_{i=2}^{s}$ and $B_0$ in precision $u \equiv u_0$
for $i \leftarrow 1$ to $r$ do
    Assemble $B_i$ using elements in $\mathcal{X} \cup \{I, X\}$ and estimate $\|B_i\|_1$
    $u_i \leftarrow \|B_{i-1}\|_1 u_{i-1} / (\|B_i\|_1 \|X^s\|_1)$        $\triangleright$ $u_i = u_{i-1}/\tau_i$, $\tau_i \ll 1$
$v \leftarrow \min\{i \colon u_i \geq \delta u\}$, $u_{v-1}, u_{v-2}, \dots, u_1 \leftarrow u$, $P \leftarrow B_r$
for $i \leftarrow r$ down to $1$ do
    Compute $P \leftarrow PX^s$ in precision $u_i$
    Form $P \leftarrow P + B_{i-1}$ in precision $u_{i-1}$
return $P$
```

- need store $\{X^i\}_{i=1}^{s}$ and $\{B^i\}_{i=0}^{r}$: about $2sn^2$ elements of storage
- $s+v-2$ matrix products in u and 1 in each of $u_v, u_{v+1}, \dots, u_r$.

- How practical is the algorithm (are the conditions $\tau_i \ll 1$, $i = v \colon r$)?

Mixed-Precision Paterson–Stockmeyer

Bounds for Taylor Approximants of e^X

Theorem (Decay of τ_i)

If $\|X\|_1 \le s/\mathrm{e}$, for $i = 2\colon r$,

$$\tau_i = \frac{\|B_i\|_1 \|X^s\|_1}{\|B_{i-1}\|_1} \lesssim \frac{\mathrm{e}}{\mathrm{e}-1} i^{-s} \approx 1.58 i^{-s}.$$

- τ_i decreases at least polynomially as i increases and at least exponentially as s increases.
- Bound not applicable to τ_1 $\Rightarrow$ we have the bound

$$\tau_1 = \frac{\|B_1\|_1 \|X^s\|_1}{\|B_0\|_1} \lesssim \frac{\|X\|_1^s}{s!\|B_0\|_1} \cdot \frac{\|X^s\|_1}{\|X\|_1^s} \lesssim \frac{1}{\|\mathrm{e}^X\|_1} \cdot \frac{\|X^s\|_1}{\|X\|_1^s} \le 1.$$

- A special treatment for $\|X\|_1 \le s/\mathrm{e}$ is possible: choose s sufficiently large s.t. $\tau_i \ll 1$, $i = 1\colon r$.
- Insight for the general case (?): larger s makes v in $\tau_i \ll 1$, $i = v\colon r$ smaller. (Recall $s+v-2$ matrix products in u and 1 in $u_v, u_{v+1}, \dots, u_r$).

Theorem (Decay of τ_i)

If $\|X\|_1 \le s/\mathrm{e}$, for $i = 2\colon r$,

$$\tau_i = \frac{\|B_i\|_1 \|X^s\|_1}{\|B_{i-1}\|_1} \lesssim \frac{\mathrm{e}}{\mathrm{e}-1} i^{-s} \approx 1.58 i^{-s}.$$

- τ_i decreases at least polynomially as i increases and at least exponentially as s increases.
- Bound not applicable to τ_1 $\Rightarrow$ we have the bound

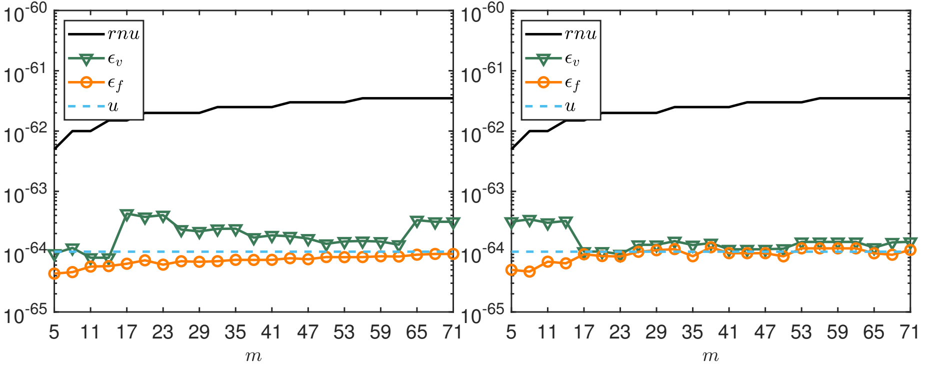
$$\tau_1 = \frac{\|B_1\|_1 \|X^s\|_1}{\|B_0\|_1} \lesssim \frac{\|X\|_1^s}{s!\|B_0\|_1} \cdot \frac{\|X^s\|_1}{\|X\|_1^s} \lesssim \frac{1}{\|\mathrm{e}^X\|_1} \cdot \frac{\|X^s\|_1}{\|X\|_1^s} \le 1.$$

• A special treatment for $\|X\|_1 \le s/\mathrm{e}$ is possible: choose s sufficiently large s.t. $\tau_i \ll 1$, $i = 1\colon r$.

• Insight for the general case (?): larger s makes v in $\tau_i \ll 1$, $i = v\colon r$ smaller. (Recall $s + v - 2$ matrix products in u and 1 in $u_v, u_{v+1}, \dots, u_r$).

Numerical Experiments

p_m from Taylor Approximant of $\exp$, Varying m



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

$\|X\|_1 = \lceil\sqrt{m}\rceil/\mathrm{e}$, Variable-precision environment with $u = 10^{-64}$ (Simulated by **Advanpix**), and $\epsilon = \|\widehat{p}_m - p_m(X)\|_1/\|p_m(X)\|_1$.

Numerical Experiments

Complexity Reduction in Variable Precisions

Table: m: minimal degree such that $\|e^X - p_m(X)\|_1 \le u$. d_i: equivalent decimal digits of precision u_i. C_p: approximate complexity reduction in percentage (assuming complexity is linearly proportional to the number of digits used).

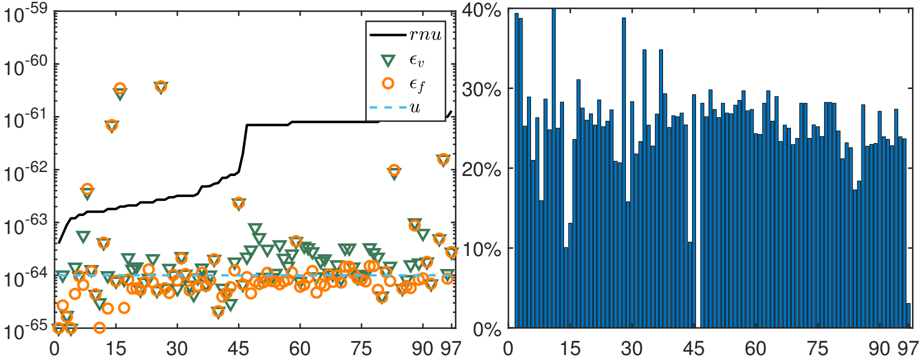
(u, m)	(s, r)	$(d_1, d_2, \ldots, d_r)$	C_p
$(10^{-32}, 37)$	$(7, 5)$	$(30, 25, 18, 11, 3)$	20.7%
$(10^{-64}, 60)$	$(8, 7)$	$(61, 55, 47, 38, 28, 18, 7)$	21.6%
$(10^{-128}, 99)$	$(10, 9)$	$(124, 115, 104, 92, 78, 64, 49, 34, 18)$	20.6%
$(10^{-256}, 169)$	$(13, 13)$	$(249, 237, 221, 203, 184, 164, 143, 121, 99, 75, 52, 28, 3)$	24.2%

$X =$ `gallery('cauchy',n)` for $n = 100$ with $\|X\|_1 \approx 4.20$

- $\tau_i = u_{i-1}/u_i = 10^{d_i - d_{i-1}}$ is in general decreasing (w.r.t. i), 20% of the matrix products were performed in precision $u^{1/2}$ or much lower.

Numerical Experiments

p_m from Taylor Approximant of $\exp$, $u = 10^{-64}$

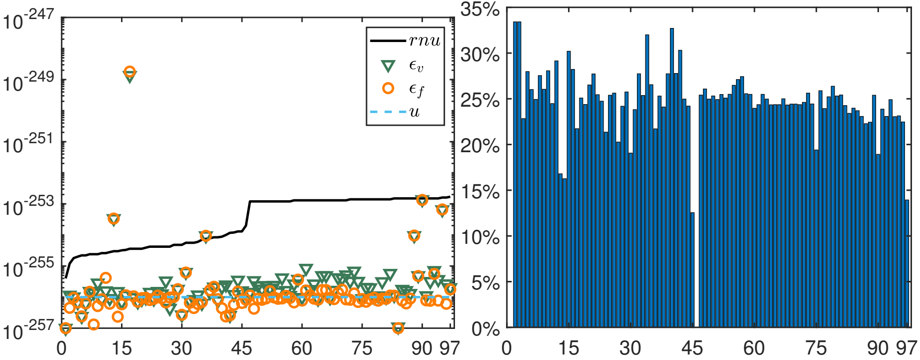


97 non-Hermitian matrices from **(Fasi and Higham, 2018)**, $2 \leq n \leq 100$. The degree m and scaling ℓ are from $\mathrm{e}^A \equiv \mathrm{e}^{2^\ell X} \approx p_m(X)^{2^\ell}$. $u = 10^{-64}$.

Left: $\epsilon = \|\widehat{p}_m - p_m(X)\|_1 / \|p_m(X)\|_1$. Right: the approximate percentages of complexity reduction.

Numerical Experiments

p_m from Taylor Approximant of $\exp$, $u = 10^{-256}$

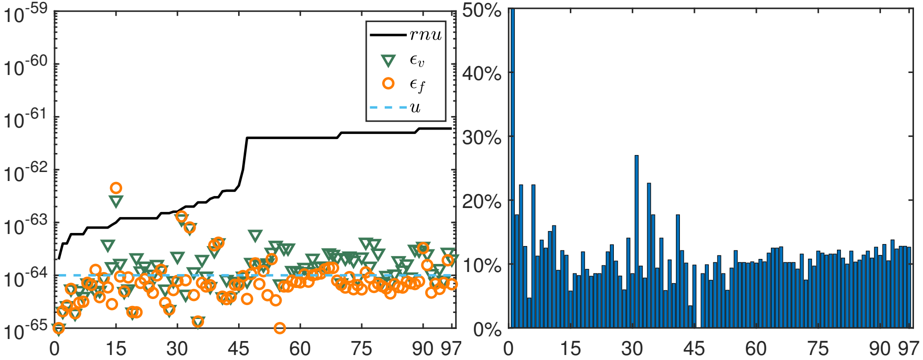


97 non-Hermitian matrices from **(Fasi and Higham, 2018)**, $2 \le n \le 100$. The degree m and scaling ℓ are from $\mathrm{e}^{A} \equiv \mathrm{e}^{2^{\ell}X} \approx p_m(X)^{2^{\ell}}$. $u = 10^{-256}$.

Left: $\epsilon = \|\widehat{p}_m - p_m(X)\|_1 / \|p_m(X)\|_1$. Right: the approximate percentages of complexity reduction.

Numerical Experiments

p_m from Padé Approximant of $\exp$ (Numerator)

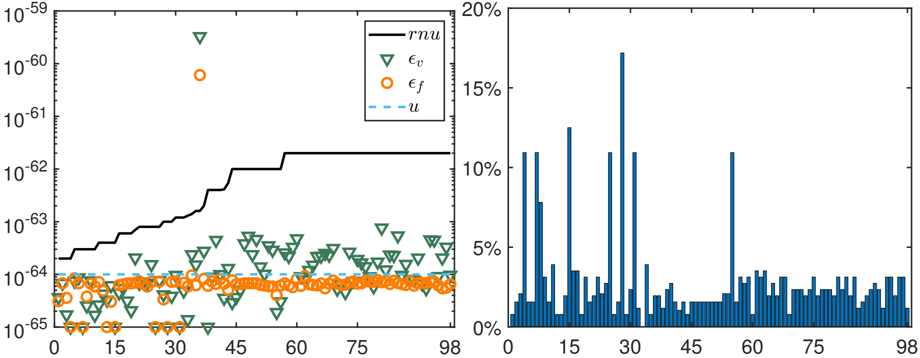


97 non-Hermitian matrices from **(Fasi and Higham, 2018)**, $2 \leq n \leq 100$. The degree m and scaling ℓ are from $\mathrm{e}^{A} \equiv \mathrm{e}^{2^{\ell}X} \approx r_{mm}(X)^{2^{\ell}}$. $u = 10^{-64}$.

- Scalar coefficients from Padé decay faster than from Taylor and smaller degree m is chosen!

Numerical Experiments

p_m from Taylor Approximant of $\cos$



98 non-Hermitian matrices from **(Al-Mohy, Higham and L, 2022)**, $4 \le n \le 100$. The degree m and scaling ℓ are from $\mathrm{e}^A \equiv \mathrm{e}^{2^\ell X} \approx p_m(X)^{2^\ell}$. $u = 10^{-64}$.

- Scalar coefficients for $\cos$ decay faster than for $\exp$ and smaller degree m is chosen (plus $p_m(X^2)$ is actually evaluated via Paterson–Stockmeyer).

Conclusions

- Lower(-than-working) precisions can be exploited in the Paterson–Stockmeyer method, if $\|X\|$ is “small” (which (I think) is satisfied in most of the practical cases) and modulus of the scalar coefficients decays quickly.
- The key idea is to perform computations on data of small magnitude (norm) in low precision.
- Better understanding of the method is desired (e.g., for $\exp$ the algorithm works well and the bound appears pessimistic).

▶ X. Liu. Mixed-precision Paterson–Stockmeyer method for evaluating polynomials of matrices. preprint, https://arxiv.org/abs/2312.17396.

Thank you for your attention!

Bibliographies I

Advanpix.
Multiprecision Computing Toolbox.
Advanpix, Tokyo, Version 5.1.1.15444.

Awad H. Al-Mohy and Nicholas J. Higham and Xiaobo Liu.
Arbitrary Precision Algorithms for Computing the Matrix Cosine and its Fréchet Derivative.
SIAM. J. Matrix Anal. Appl., 43(1):233–256, 2022.

Massimiliano Fasi.
Optimality of the Paterson–Stockmeyer method for evaluating matrix polynomials and rational matrix functions.
Linear Algebra Appl., 574:182–200, 2019.

Bibliographies II

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.

Gareth Hargreaves.
Topics in matrix computations: Stability and efficiency of algorithms.
PhD thesis, University of Manchester, Manchester, England, August 2005, 204 pp.

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.