Computing the matrix exponential

Note Title 2025-04-03

$$\exp(A) = I + A + \frac{1}{2}A^2 + \frac{1}{3!}A^3 + \dots$$

⚠ $\exp(A+B) \neq \exp(A)\exp(B)$ in general.

It holds only if A, B commute.

When $B = \beta I$, $(\beta \in \mathbb{C})$

$$\exp(A+\beta I) = \exp(A)\, e^{\beta}$$

($\exp(A)$: matrix; e^{β}: scalar)

Theorem: Suppose $A \in \mathbb{C}^{n\times n}$ such that $A_{ij} \geq 0$ for all $i \neq j$

Then, $[\exp(A)]_{ij} \geq 0$ for all i, j.

Proof:

$$\exp(A) = e^{-\beta}\exp(A+\beta I) = e^{-\beta}\sum_{k\geq 0}\frac{1}{k!}(A+\beta I)^k$$

If β is large enough, $A+\beta I$ has all positive entries, so all the summands are nonnegative. □

For matrices with $A_{ij} \geq 0$ $i \neq j$ (Metzler matrices, $-\Sigma$-matrices, essentially nonnegative matrices), this is a very stable method.

Lemma: the solution of the ODE $\begin{cases} \dot{v} = Av \\ v(0) = v_0 \end{cases}$ where $A \in \mathbb{C}^{n\times n}$, $v : [0,\infty) \to \mathbb{C}^n$

is $v(t) = \exp(tA)v_0$

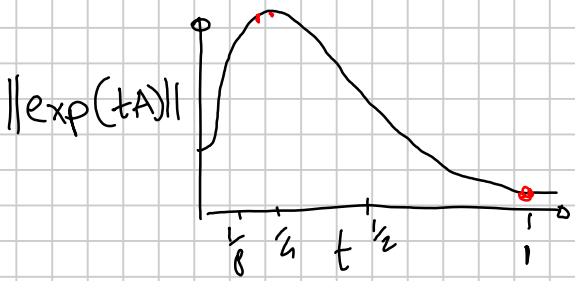
And the solution of $\begin{cases} \dot{M} = AM \\ M(0) = I \end{cases}$ is $\exp(tA)$

A method to compute $\exp(A)$ is solving this ODE on $[0,1]$

$M(1) = \exp(A)$.

E.g. Euler gives $\exp(A) \approx \left(I + \frac{1}{n}A\right)^n$

This method could be unstable, unfortunately, because of intermediate growth:

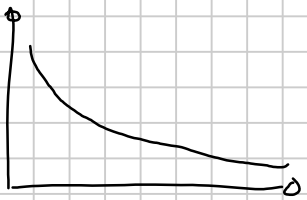


This for non-normal matrices: for <u>normal</u> matrices,

$$A = QDQ^* \qquad \exp(tA) = Q \begin{bmatrix} \exp(t\lambda_1) & & \\ & \ddots & \\ & & \exp(t\lambda_n) \end{bmatrix} Q^*$$

$$\|\exp(tA)\| = \max_{i=1..n} |e^{t\lambda_i}| = e^{t \max \mathrm{Re}\,\lambda_i}$$

so the plot is just a scalar exponential



If we wish to compute $\exp(tA)$ for A non-normal and t in the "tail" using a method based on solving ODEs, you have to compute also all the intermediate values on the "hump" → cancellation!

"19 dubious methods to compute matrix exponentials", Moler-Van Loan '78 '03

State-of-the-art method: • Padé approximants
• Scaling and squaring

A <u>Padé approximant</u> to $f(x)$ in $x=0$ with degrees (p,q) a rational function $r(x) = \frac{N(x)}{D(x)}$ with $\deg N = p$, $\deg D = q$

such that $f(x) - r(x) = O(x^{p+q+1})$ for $x \to 0$

N has $p+1$ coefficients

D has $q+1$ coefficients

-1 for scaling

$p+q+1$ degrees of freedom $\Rightarrow$ can match first $p+q+1$ coefficients of a Taylor series

To compute it: set normalization $D(0) = 1$

$$f(x) - \frac{N(x)}{D(x)} = O(x^{p+q+1}) \iff f(x)D(x) - N(x) = O(x^{p+q+1})$$

$\uparrow$ linear system in the coefficients.

Idea: $\exp(A) \approx D(A)^{-1} N(A) = N(A) D(A)^{-1}$

$\uparrow$ they commute

We hope this to be accurate when $\|A\|$ is small

We would like to give a sort of backward error bound on this approximation:

$$D(A)^{-1} N(A) = \exp(A + \Delta)$$

for a small matrix Δ. If we can show $\|\Delta\| / \|A\| \approx u = 2.2 \cdot 10^{-16}$, then the error in this approximation is of the same order as the error incurred when replacing A with its floating point approximation $fl(A)$, i.e.

$$\| \exp(A+\Delta) - \exp(A) \| \leq \| L_{f,A} \| \cdot \| \Delta \| + O(\|\Delta\|^2)$$

(Note that one can show $\|\exp(A)\| \leq \|L_{\exp, A}\| \leq e^{\|A\|}$.

For normal matrices, $\|\exp(A)\| = \|L_{\exp, A}\|$)

We shall see a first result of this kind:

Theorem: for $p=q=2$, if $\|A\| \leq 0.1$, then $\|\Delta\| \leq 1.4 \cdot 10^{-7}$

To prove this, let us start from the scalar version

there exists δ such that

$$\frac{N(x)}{D(x)} = e^{x+\delta} \qquad \delta = \log\left(e^{-x}\frac{N(x)}{D(x)}\right) = \delta(x)$$

$$\frac{N(x)}{D(x)} - e^x = O\left(x^{p+q+1}\right) \Rightarrow \quad e^{-x}\frac{N(x)}{D(x)} - 1 = O\left(x^{p+q+1}\right)$$

$$\log(1+y) = y + \frac{y^2}{2} + \cdots \Rightarrow$$

$$\delta(x) = \log\left(e^{-x}\frac{N(x)}{D(x)}\right) = O\left(x^{p+q+1}\right)$$

A plot shows that $|\delta(x)| \leq 1.5 \cdot 10^{-8}$ when $|x| \leq 0.1$

Now for matrices:

$$\Delta = \delta(A) = \log\left(\exp(-A)D(A)^{-1}N(A)\right)$$

$$\exp(\Delta) = \exp(-A)D(A)^{-1}N(A)$$

$$\exp(A)\exp(\Delta) = D(A)^{-1}N(A)$$

Δ is a function of A, so A and Δ commute, and

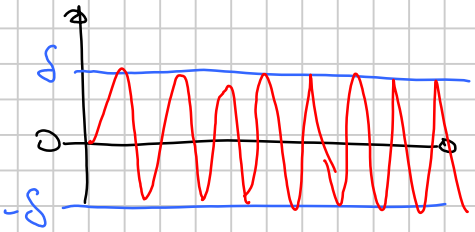
$$\exp(A)\exp(\Delta) = \exp(A+\Delta)$$

$$\Rightarrow \quad \exp(A+\Delta) = D(A)^{-1}N(A)$$

However, is it true that $\|\Delta\| = \|\delta(A)\|$ is small?

In general, the fact that a scalar function is small does <u>not</u> imply that the corresponding matrix function is.

Example: take $f(x)$ such that $|f(x)|$ is small but $|f'(x)|$ is large:



$$f\left(\begin{bmatrix} x & 1 \\ 0 & x \end{bmatrix}\right) = \begin{bmatrix} f(x) & f'(x) \\ 0 & f(x) \end{bmatrix}$$ can be arbitrarily large!

However, we can use brute force and Taylor expansions:

Let $\delta(x) = \sum_{i=p+q+1}^{\infty} c_i x^i$ be a Taylor expansion

$$\|\delta(A)\| = \left\| \sum_{i=p+q+1}^{\infty} c_i A^i \right\| \leq \sum_{i=p+q+1}^{\infty} |c_i| \cdot \|A\|^i$$

We can evaluate this computationally, truncating the series:

If $\|A\| < 0.1$, $\|\delta(A)\| \leq \sum_{i=p+q+1}^{\infty} |c_i| (0.1)^i$

$$\|\delta(A)\| \leq 1.4 \cdot 10^{-8} \qquad \|\delta(A)\| / \|A\| \leq 1.4 \cdot 10^{-7}$$

(The "19 dubious ways" paper contains a rigorous estimate for the tail of this series)

Theorem: (Higham book / table):

With $p = q = 13$, $\|A\| \leq 5.4$, then $\frac{\|\Delta\|}{\|A\|} \leq u = 2.2 \cdot 10^{-16}$

Also, $\|D(A)^{-1}\| \leq 17$ (with similar techniques)

$D(A) = 1 + d_1 x + \dots + d_9 x^9$

This shows that $D(A)$ is well conditioned for small A

(Indeed, $D(A) \to I$ when $A \to 0$).

(Note that one needs only 6 mat-vec multiplications to compute $N(A)$ and $D(A)$ at the same time for $p=q=13$)

Scaling and squaring:

Idea: $\exp(A) = \exp\left(\frac{1}{2}A\right)^2$

1) Compute k such that $\left\| \frac{1}{2^k} A \right\| = \frac{1}{2^k}\|A\| \leq 5.4$

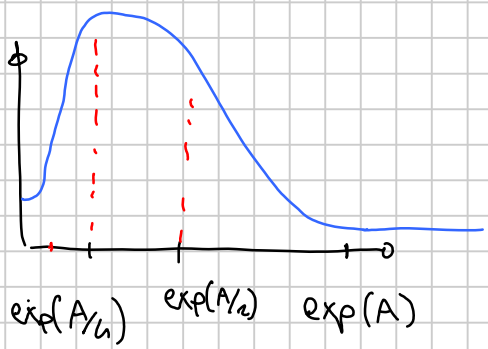
2) Compute $E = \exp\left(\frac{1}{2^k}A\right)$ using Padé approximants

3) Compute $\exp(A) = E^{2^k}$ by squaring k times.

The cost (number of products) depends on $\|A\|$!

This is Matlab's expm

Scaling and squaring can still suffer "hump phenomene":



large $\exp(A/4)$ small $\exp(A)$

Note that $\|L_{\exp,A}\|$ is going to be large in the "tail" anyway, so this is an ill-conditioned problem.

Scaling and squaring is quite robust in experiments, but

it is <u>not</u> backward stable in general.

$$\max_{x \in I} |f(x) - r(x)| \sim C^{d}$$ for some $C < 1$

(for fixed I)