Matrix exponential

Note Title 2023-03-30

Backward error of Padé approximation:

can we write

$$r(A) = d(A)^{-1} n(A) = \exp(A+H)$$

for a (hopefully small) matrix H?

$$\frac{n(x)}{d(x)} - e^{x} = O(x^{p+q+1}) \qquad p = \deg n(x)$$

$$x \to 0 \qquad\qquad q = \deg d(x)$$

scalar version of the problem: find $h \in \mathbb{C}$ s.t.

$$\frac{n(x)}{d(x)} = e^{x+h} \qquad h = \log e^{-x} \frac{n(x)}{d(x)}$$

For a Maclaurin series in 0,

$$h(x) = c_{p+q+1} x^{p+q+1} + c_{p+q+2} x^{p+q+2} + \dots$$

↑ all previous coefficients are 0.

Indeed, $\frac{n(x)}{d(x)} - e^{-x} = O(x^{p+q+1}) \Rightarrow e^{-x}\frac{n(x)}{d(x)} = 1 + O(x^{p+q+1})$

$$\Rightarrow \log\left(e^{-x}\frac{n(x)}{d(x)}\right) = O(x^{p+q+1}).$$

This motivates our expectation that $h(x)$ is very small around 0.

<u>In the matrix case:</u> <u>idea:</u> define $H = h(A)$ $\qquad h(x) = \log e^{-x}\frac{n(x)}{d(x)}$

$$H = \log(\underbrace{\exp(-A)\, n(A)\, d(A)^{-1}}_{\text{any order works, since they commute}})$$

Then,

$$\exp(H) = \exp(-A)\, n(A)\, d(A)^{-1}$$

$$\exp(A+H) \underset{\uparrow}{=} \exp(A)\exp(H) = n(A)\, d(A)^{-1}$$

since A and H commute

Note that $\quad |h(x)| \leq \varepsilon \quad$ for $x \in [-\delta, \delta]$

$$\not\Rightarrow \quad \|H\|_2 \leq \varepsilon \quad \text{for} \quad \|A\| \in [-\delta, \delta]$$

Indeed, $A = VJV^{-1} \Rightarrow f(A) = V f(J) V^{-1}$, but $f(J)$ involves also derivatives $h'(x), h''(x), \ldots$ for which we do not have bounds

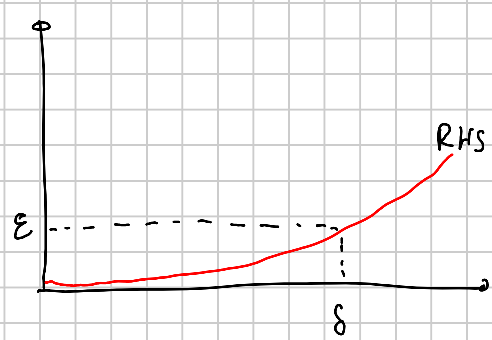
However, we can write

$$\|H\| = \|h(A)\| = \| c_{p+q+1} A^{p+q+1} + c_{p+q+2} A^{p+q+2} + \ldots \|$$

$$\leq |c_{p+q+1}| \cdot \|A\|^{p+q+1} + |c_{p+q+2}| \cdot \|A\|^{p+q+2} + \ldots$$

Let δ be the positive sol. of $\quad \varepsilon = \underbrace{|c_{p+q+1}|\, \delta^{p+q+1} + |c_{p+q+2}|\, \delta^{p+q+2} + \ldots}_{\text{strictly increasing}}$

Then, $\|A\| \leq \delta \Rightarrow \|H\| \leq \varepsilon$



Matlab solution: for $\varepsilon = 2.2 \cdot 10^{-16}$,

$\delta \approx 0.028 \qquad$ for $p = q = 2$

People found that for $p = q = 13$, $\varepsilon = 2.2 \cdot 10^{-16}$, $\delta = 5.4$

So for $\|A\| \leq 5.4$, the $(13, 13)$ Padé approximant has a backward error of at most u (machine precision)

What happens if A has larger norm?

Another idea: <u>scaling and squaring</u>

The exponential satisfies $e^x = \left(e^{\frac{x}{s}}\right)^s$ for all $s>0$

hence $\exp(A) = \left[\exp\left(\frac{1}{s}A\right)\right]^s$

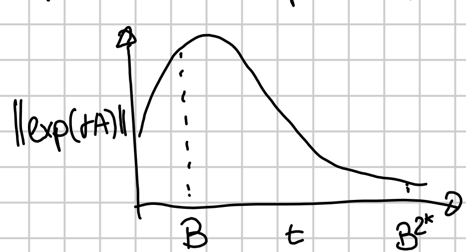
This suggests an algorithm:

1. Find $s = 2^k$ s.t. $\left\|\frac{1}{2^k}A\right\| \leq 5.4$
2. compute $B = \exp\left(\frac{1}{2^k}A\right)$ using a Padé approximant
3. Compute $\exp(A) \approx B^{2^k}$ with k successive squarings.

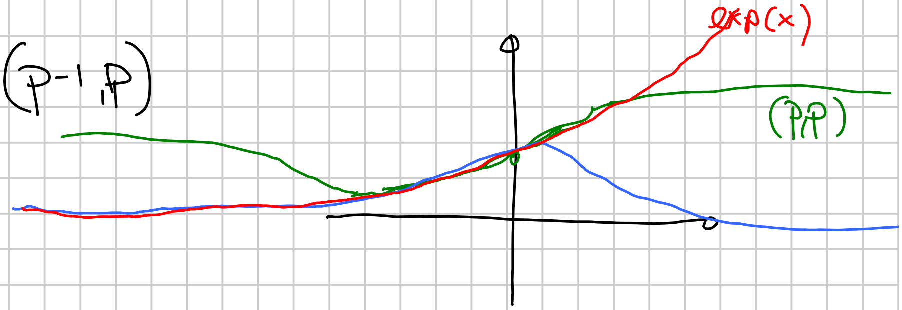
This is essentially Matlab's expm. Actually it does this

1. Compute $QTQ^* = A$ (Schur decomposition)
2. Compute $S = \text{expm}(T)$ by scaling and squaring (previous algorithm)
3. $S = \begin{bmatrix} s_{11} & \cdots & s_{1n} \\ & \ddots & \vdots \\ 0 & & s_{nn} \end{bmatrix}$ has $s_{ii} \approx \exp(t_{ii})$, but not exactly equal, so we re-compute it: $s_{ii} \leftarrow \exp(t_{ii})$
4. $\exp(A) = QSQ^*$

Is this stable? Yes in practice, but no full proof yet.



(p,p) Padé approx $\to$ $(p-1,p)$



Argument reduction: for certain matrices A and for suitably chosen $\tau \in \mathbb{R}$, $\|\underbrace{A-\tau I}_{B}\| << \|A\|$.

So one can compute $\exp(A) = \exp(B+\tau I) \underset{\uparrow}{=} \exp(B)\exp(\tau I)$

B and τI commute

$= \exp(B)\cdot e^{\tau}\cdot I$

Since $\|B\| << \|A\|$, I need fewer scaling and squaring steps.

e.g. $\exp\begin{bmatrix} 2 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 2 \end{bmatrix} = e^2 \cdot \exp\begin{bmatrix} 0 & -1 & & \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ & & -1 & 0 \end{bmatrix}.$

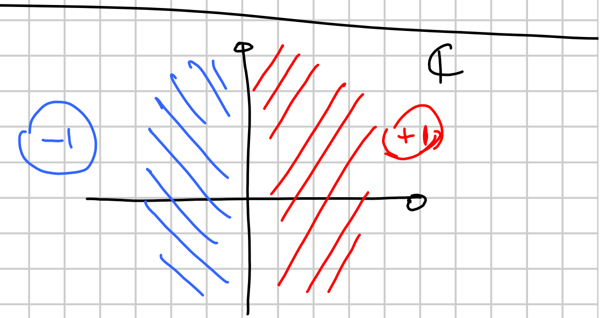
If B is a matrix such that $A = B+\tau I$ has all non-negative entries for some $\tau \in \mathbb{R}$, then

$$\underbrace{\sum_{k=0}^{\infty} \frac{1}{k!} A^k}_{\text{has all non-negative entries}} = \exp(A) = \exp(B+\tau I) = \underbrace{\exp(B)}_{\text{all non-negative entries, too!}} \cdot e^{\tau}$$

Theorem: suppose B has $B_{ij} \geq 0$ for all $i \neq j$.
Then, $[\exp(B)]_{ij} \geq 0$ for all i,j.

The sign function:

$$f(x) = \begin{cases} 1 & Re(x) > 0 \quad \text{RHP} \\ -1 & Re(x) < 0 \quad \text{LHP} \\ \text{undef.} & Re(x) = 0 \end{cases}$$



If $A = VJV^{-1}$ is a Jordan form, we can partition it (up to reordering) as

$$J = \left[\begin{array}{c|c} J_1 & 0 \\ \hline 0 & J_2 \end{array}\right] \quad \text{where} \quad \begin{array}{l} \Lambda(J_1) \subset LHP \\ \Lambda(J_2) \subset RHP \end{array}$$

and $V = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]$

Then,
$$\operatorname{sign}(A) = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} f(J_1) & 0 \\ \hline 0 & f(J_2) \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1}$$
$$= \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} -I & 0 \\ \hline 0 & I \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]$$

In particular,
$$\operatorname{Im}\left(\operatorname{sign}(A) + I\right) = \operatorname{Im} \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left(\begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix} + 1 \right) \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1}$$
$$= \operatorname{Im} \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & 1 \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1} = \operatorname{Im} V_2$$

and analogously $\operatorname{Im}\left(\operatorname{sign}(A) - I\right) = \operatorname{Im} V_1$.

$\operatorname{Im} V_1$, $\operatorname{Im} V_2$ are the two invariant subspaces associated to eigenvalues in the LHP and RHP respectively.

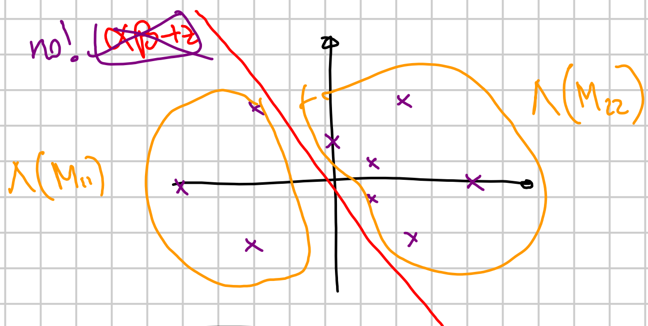
Application: computing eigenvalues by bisection: suppose you have an efficient algorithm to compute $\operatorname{sign}(\cdot)$. Then, given a starting matrix M, one can compute $\operatorname{sign}(\alpha M + \beta I)$ for given $\alpha, \beta \in \mathbb{C}$. Then Q = orthogonal qr factor of

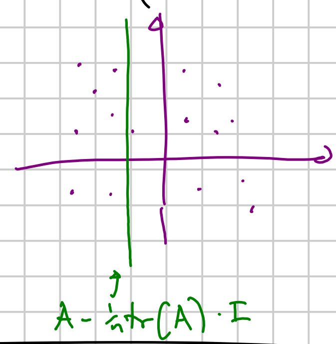
$\operatorname{sign}(\alpha M + \beta I)$ is a matrix such that

$$Q^* M Q = \left[\begin{array}{c|c} M_{11} & M_{12} \\ \hline O & M_{22} \end{array}\right]$$

$\Lambda(M_{11})$ contains the eigenvalues of M s.t. $\mathrm{Re}(\alpha\lambda + \beta) < 0$,

$\Lambda(M_{22})$ contains the eigenvalues of M s.t. $\mathrm{Re}(\alpha\lambda + \beta) > 0$.





First method to compute $\operatorname{sign}(A)$: <u>Schur-Parlett</u>:

1. Compute $M = QUQ^*$, with Q orthogonal, U upper triangular
2. Reorder the factorization to $M = \hat{Q}\hat{U}\hat{Q}^*$, where

$$\hat{U} = \left[\begin{array}{c|c} A & C \\ \hline 0 & B \end{array}\right] \quad \text{with } \Lambda(A) \subset LHP, \quad \Lambda(B) \subset RHP$$

3. Then, $\operatorname{sign}(\hat{U}) = \left[\begin{array}{c|c} -1 & Z \\ \hline 0 & 1 \end{array}\right]$ and we can determine Z

by using the relation $\hat{U} \cdot \operatorname{sign}(\hat{U}) = \operatorname{sign}(\hat{U}) \cdot \hat{U}$:

$$\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\begin{bmatrix} -1 & Z \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -1 & Z \\ 0 & 1 \end{bmatrix}\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

(1,2) block: $AZ + C = -C + ZB \iff AZ - ZB = -2C$.

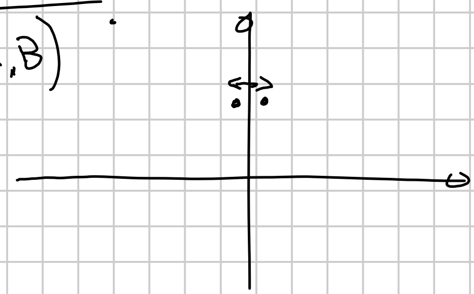
This is a Sylvester equation, it has unique solution because
$\Lambda(A) \subset LHP \quad \Lambda(B) \subset RHP$ hence $\Lambda(A) \cap \Lambda(B) = \emptyset$

and it is also easy to solve in practice because A, B
are both triangular (since they come from a Schur form).

4. $\operatorname{sign}(M) = \hat{Q}\operatorname{sign}(\hat{U})\hat{Q}^* = \hat{Q}\begin{bmatrix} -1 & Z \\ 0 & 1 \end{bmatrix}\hat{Q}^*.$

In particular, note that

$$\|Z\|_F \leq \left\|\left(I\otimes A - B^T\otimes I\right)^{-1}\right\|_2 \|2C\|_F = \frac{2\|C\|_F}{\operatorname{sep}(A,B)}.$$



The main reason why $\operatorname{sep}(A,B)$ is small is the presence of two close-by eigenvalues in $\Lambda(A)$, $\Lambda(B)$.