Matrix sign function

Note Title 2023-04-04

Recap: Schur-Parlett method

1) Reduce to triangular case $\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$ $\Lambda(A) \subset$ LHP, $\Lambda(B) \subset$ RHP

2) solve $AZ - ZB = -2C$ $(*)$

3) $\operatorname{sign}\left(\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}\right) = \begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix}$

Conditioning of $(*)$ related to $\operatorname{sep}(A,B)$.

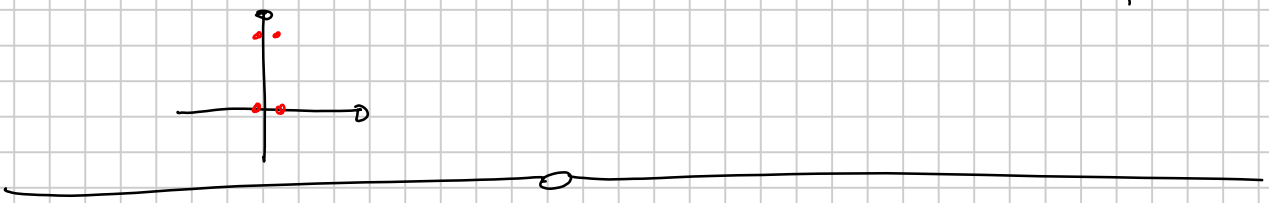
Perturbation theory

From generic results on matrix functions, if $M = VDV^{-1}$

$$\| \hat{L}_{\operatorname{sign}, M} \| \leq \kappa(V)^2 \max_{\lambda,\mu \in \Lambda(M)} |f[\lambda,\mu]|$$

$$f[\lambda,\mu] = \begin{cases} f'(\lambda) & \text{if } \lambda = \mu \\ \dfrac{f(\lambda) - f(\mu)}{\lambda - \mu} & \text{otherwise} \end{cases}$$

For the sign, $f[\lambda,\mu] = \begin{cases} 0 & \lambda = \mu \text{ or in same half-plane} \\ \dfrac{\pm 2}{\lambda - \mu} & \text{if } \lambda, \mu \text{ on opposite sides of im. axis} \end{cases}$



Theorem: If $M = Q \begin{bmatrix} A & C \\ 0 & B \end{bmatrix} Q^*$, where Q orthogonal, $\Lambda(A) \subset$ LHP, $\Lambda(B) \subset$ RHP. Let $\delta = \operatorname{sep}(A,B)$

and consider a perturbation $\tilde{M} = M + E$ of norm $\|E\|_F \leq \varepsilon$ sufficiently small.

Then,

1) $\|\text{sign}(M)\|_F = O\left(\frac{1}{\delta}\right)$

2) $\dfrac{\|\text{sign}(M+E) - \text{sign}(M)\|_F}{\|\text{sign}(M)\|_F} = O\left(\frac{\varepsilon}{\delta^2}\right)$ ($O\left(\frac{\varepsilon}{\delta^3}\right)$, $O\left(\frac{1}{\delta}\right)$)

3) the Hurwitz stable invariant subspace of $M+E$, i.e. its invariant subspace relative to the LHP, has a basis of the form $Q\begin{bmatrix} I \\ X \end{bmatrix}$, where $\|X\|_F = O\left(\frac{\varepsilon}{\delta}\right)$

(so it is an $O\left(\frac{\varepsilon}{\delta}\right)$ perturbation of the invariant subspace $\text{Im } Q\begin{bmatrix} I \\ 0 \end{bmatrix}$ of M)

Proof:

1) $\text{sign}(M) = Q\begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix}Q^*$ $\quad \|\text{sign}(M)\|_F = O\left(\|Z\|_F\right)$

Z solves $AZ - ZB = -2C$ hence $\|Z\|_F \leq \|(I \otimes A - B^T \otimes I)^{-1}\| \cdot \|2C\|_F$

$$= \frac{1}{\text{sep}(A,B)} \cdot 2\|C\|_F = O\left(\frac{1}{\delta}\right)$$

3) From a result in the beginning of the course:

$$Q^*(M+E)Q = \begin{bmatrix} I & 0 \\ X & I \end{bmatrix}\begin{bmatrix} \hat{A} & \hat{C} \\ 0 & \hat{B} \end{bmatrix}\begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} = \begin{bmatrix} \tilde{A}+\tilde{C}X & \tilde{C} \\ 0 & \tilde{B}-X\tilde{C} \end{bmatrix}$$

for a matrix X with norm $\|X\|_F = O\left(\frac{\varepsilon}{\delta}\right)$

and hence $Q\begin{bmatrix} I \\ X \end{bmatrix}$ spans the Hurwitz stable inv. subspace of $M+E$.

(For sufficiently small ε, $\Lambda(\tilde{A}+\tilde{C}X) \subset$ LHP, $\Lambda(\tilde{B}-X\tilde{C}) \subset$ RHP

$$Q^*(M+E)Q = \begin{bmatrix} \tilde{A} & \tilde{C} \\ \tilde{D} & \tilde{B} \end{bmatrix} \qquad Q^* M Q = \begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$$

2) Continue using the previous decomposition to compute $\operatorname{sign}(M+E)$:

$$\operatorname{sign}(M+E) = Q \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \operatorname{sign}\left(\begin{bmatrix} \hat{A} & \hat{C} \\ 0 & \hat{B} \end{bmatrix} \right) \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} Q^*$$

$$= Q \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} -I & \hat{Z} \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} Q^*$$

where $\hat{Z}$ solves $\hat{A}\hat{Z} - \hat{Z}\hat{B} = -2\hat{C}$

i.e. $(\tilde{A} + \tilde{C}X)\hat{Z} - \hat{Z}(\tilde{B} - X\tilde{C}) = -2\tilde{C}$

This is a small perturbation of $\operatorname{sign}(M) = Q \begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix} Q^*$

$\|\hat{Z} - Z\|_F = ?$ To bound $\|\hat{Z} - Z\|_F$, we use perturbation theory for linear systems.

$$AZ - ZB = -2C \iff (I \otimes A - B^T \otimes I)\operatorname{vec}(Z) = \operatorname{vec}(-2C)$$

$$(\tilde{A} + \tilde{C}X)\hat{Z} - \hat{Z}(\tilde{B} - X\tilde{C}) = -2\tilde{C} \iff (I \otimes (\tilde{A} - \tilde{C}X) - (\tilde{B} - X\tilde{C})^T \otimes I)\operatorname{vec}(\hat{Z}) = \operatorname{vec}(-2\tilde{C})$$

Lemma: if x solves $Tx = c$, and $\tilde{x}$ is solves $(T + \delta T)\tilde{x} = (c + \delta c)$ then

$$\frac{\|\tilde{x} - x\|}{\|x\|} \le \kappa(T)\left(\frac{\|\delta c\|}{\|c\|} + \frac{\|\delta T\|}{\|T\|} \right) + O\left(\left(\frac{\|\delta T\|}{\|T\|} \right)^2 \right)$$

$$\frac{\|\hat{Z} - Z\|_F}{\|Z\|_F} \le \kappa(T)\left(\frac{\|(\tilde{C} - C)\|_F}{\|C\|_F} + \frac{\|\tilde{A} - \tilde{C}X - A\|_F + \|\tilde{B} - X\tilde{C} - B\|_F}{\|T\|} \right) + O(\cdots)$$

$$O\left(\frac{1}{\delta}\right) \cdot \left(O(\varepsilon) + O\left(\frac{\varepsilon}{\delta}\right) \right)$$

$$= O\left(\frac{\varepsilon}{\delta^2}\right)$$

$$\operatorname{sign}(M+E) - \operatorname{sign}(M) = Q \begin{bmatrix} 1 & 0 \\ X & 1 \end{bmatrix} \begin{bmatrix} -1 & \hat{Z} \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ -X & 1 \end{bmatrix} Q^* - Q \begin{bmatrix} -1 & \hat{Z} \\ 0 & 1 \end{bmatrix} Q^*$$

$O\left(\frac{\varepsilon}{\delta}\right)$ part. of I — $O\left(\frac{\varepsilon}{\delta}\right)$ part. of I — $O\left(\frac{\varepsilon}{\delta^2}\right)\|Z\|$ part. of $\begin{bmatrix} -1 & Z \\ 0 & 1 \end{bmatrix}$

$$= O\left(\frac{\varepsilon}{\delta^2}\right)\|Z\|_F = O\left(\frac{\varepsilon}{\delta^2}\right) \cdot \|\operatorname{sign}(M)\|_F \qquad \square$$

[Byers, Mehrmann, He 97].

Newton method for the matrix sign

Let us consider the matrix iteration

$$X_0 = M \qquad X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}\right) \qquad k = 0, 1, 2, \ldots$$

We wish to see that $\lim_{k\to\infty} X_k = \operatorname{sign}(M)$

Let us first study the case $M = VDV^{-1}$ diagonalizable

(V: eigenvectors, D: diag(eigenvalues))

$$X_0 = VDV^{-1} \qquad X_1 = \frac{1}{2}\left(X_0 + X_0^{-1}\right) = V\left(\frac{1}{2}\left(D + D^{-1}\right)\right)V^{-1}$$

$$= V \begin{bmatrix} \frac{\lambda_1 + \frac{1}{\lambda_1}}{2} & & \\ & \frac{\lambda_2 + \frac{1}{\lambda_2}}{2} & \\ & & \ddots \\ & & & \frac{\lambda_n + \frac{1}{\lambda_n}}{2} \end{bmatrix} V^{-1} = V \begin{bmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_n) \end{bmatrix} V^{-1}$$

set $f(x) = \frac{1}{2}\left(x + \frac{1}{x}\right)$

and analogously $X_k = V \begin{bmatrix} f^k(\lambda_1) & & \\ & \ddots & \\ & & f^k(\lambda_n) \end{bmatrix} V^{-1}$ (by induction).

f^k: f composed k times

$$f(x) = \frac{1}{2}\left(x + \frac{1}{x}\right) = \frac{x^2+1}{2x}$$

$f(x)$ is the iterative method obtained by applying Newton's method to search a root of $x^2 - 1 = 0$: indeed,

$$x_{k+1} = x_k - \frac{x_k^2 - 1}{2x_k} = \frac{2x_k^2 - (x_k^2 - 1)}{2x_k} = \frac{x_k^2 + 1}{2x_k}$$

Fixed points of f: $f(a) = a$ $\quad \frac{a^2+1}{2a} = a \iff a^2 + 1 = 2a^2 \iff a = \pm 1$

Lemma: the limit of the scalar iteration $x_{k+1} = f(x_k) = \frac{1}{2}\left(x_k + \frac{1}{x_k}\right)$ is

$$\lim_{k \to \infty} f^k(x_0) = \begin{cases} 1 & \text{if } \mathrm{Re}(x_0) > 0 \\ -1 & \text{if } \mathrm{Re}(x_0) < 0 \\ \text{complicated} & \text{if } \mathrm{Re}(x_0) = 0 \end{cases}$$

Proof Change of variables (Cayley transform)

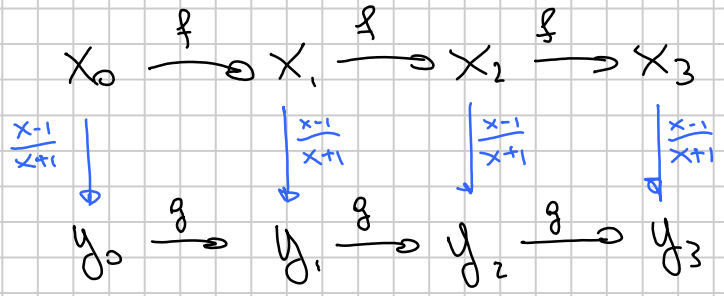
$$y = \frac{x-1}{x+1}$$

Let us first compute the inverse map:

$$y = \frac{x-1}{x+1} \iff xy + y = x - 1 \iff x(y-1) = -1 - y \iff x = \frac{1+y}{1-y}.$$

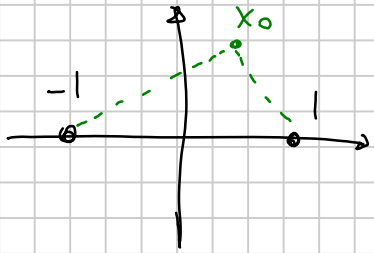
if $y_k = \frac{x_k - 1}{x_k + 1}$ and $x_{k+1} = \frac{1}{2}\left(x_k + \frac{1}{x_k}\right)$, then

$$y_{k+1} = \frac{x_{k+1} - 1}{x_{k+1} + 1} = \frac{\frac{1}{2}\left(x_k + \frac{1}{x_k}\right) - 1}{\frac{1}{2}\left(x_k + \frac{1}{x_k}\right) + 1} = \frac{x_k^2 + 1 - 2x_k}{x_k^2 + 1 + 2x_k} = \left(\frac{x_k - 1}{x_k + 1}\right)^2 = y_k^2$$



$g(y) = y^2$

So the iteration is, in the y variables, repeated squaring.



If $x_0 \in RHP$, $dist(x_0, 1) < dist(x_0, -1)$

$$|y_0| = \left| \frac{x_0 - 1}{x_0 + 1} \right| < 1$$

Then, $y_k = (y_0)^{2^k}$ converges to 0, i.e. $\lim_{k \to \infty} y_k = 0$

and by the reverse change of variables

$$\lim_{k \to \infty} x_k = \lim_{k \to \infty} \frac{1 + y_k}{1 - y_k} = 1$$

If $x_0 \in LHP$ $dist(x_0, -1) < dist(x_0, 1)$

$$|y_0| = \left| \frac{x_0 - 1}{x_0 + 1} \right| > 1$$

Then, $y_k = (y_0)^{2^k}$ converges to ∞ and

$$\lim_{k \to \infty} x_k = \lim_{k \to \infty} \frac{1 + y_k}{1 - y_k} = -1$$