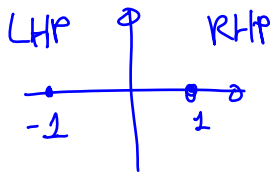


The matrix sign function



$$\text{sign}(x) = \begin{cases} 1 & \text{Re } x > 0, \\ -1 & \text{Re } x < 0, \\ \text{undefined} & \text{Re } x = 0. \end{cases}$$

Suppose the Jordan form of A is reblocked as

$\Lambda(J_2) \subseteq \text{LHP}$

$$A = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1}, \quad \begin{matrix} \mathcal{f}(J_1) = \begin{bmatrix} -1 & & \\ & \ddots & \\ & & -1 \end{bmatrix} \\ \mathcal{f}(J_2) = I \end{matrix}$$

$\Lambda(J_2) \subseteq \text{RHP}$

where J_1 contains all eigenvalues in the LHP (left half-plane) and J_2 in the RHP. Then,

$$\text{sign}(A) = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} -I & \\ & I \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1} \quad \text{sign}(A)^2 = I$$

$\text{sign}(A)$ is always diagonalizable with eigenvalues ± 1 . $\text{sign}(A) \pm I$ gives the projections on the span of the eigenvectors in the RHP/LHP (unstable/stable invariant subspace).

$$\text{sign}(A) = [v_1 \ v_2] \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} [v_1 \ v_2]^{-1}$$

$$\frac{1}{2}(\text{sign}(A) + \mathbf{I}) = [v_1 \ v_2] \begin{bmatrix} \frac{1}{2}(-\mathbf{I} + \mathbf{I}) & \\ & \frac{1}{2}(\mathbf{I} + \mathbf{I}) \end{bmatrix} [v_1 \ v_2]^{-1} =$$

$$= [v_1 \ v_2] \begin{bmatrix} 0 & 0 \\ 0 & \mathbf{I} \end{bmatrix} [v_1 \ v_2]^{-1} = \text{Proiettore su } \text{span}(v_2) \\ \text{immagine } \hat{V}_2, \text{ kernel } v_1.$$

$$-\frac{1}{2}(\text{sign}(A) - \mathbf{I}) = \text{Proiettore su } \text{span}(v_1) \rightarrow \text{immagine } v_1, \\ \text{kernel } v_2$$

$V_1 =$ "spazio stabile" di $A = \text{Span}(\text{autovettori con autovalori negativi (e loro catene di Jordan)})$

$V_2 =$ "spazio instabile"

Se $b \in$ spazio stabile, la soluzione di

$$\begin{cases} x(0) = b \\ \dot{x} = Ax \end{cases}$$

$$x(t) = \exp(tA)b$$

è tale che $\lim_{t \rightarrow \infty} x(t) = 0$

$$\begin{aligned} \exp(tA) \cdot b &= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \exp(tJ_1) & 0 \\ 0 & \exp(tJ_2) \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1} b = \\ &= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \exp(tJ_1) & \\ & \exp(tJ_2) \end{bmatrix} \begin{bmatrix} c \\ 0 \end{bmatrix} = V_1 \exp(tJ_1) c \end{aligned}$$

se J_1 ha solo autoval. negativi,
 $\lim_{t \rightarrow \infty} \exp(tJ_1) = 0$ e il limite è 0.

Similmente, se $b \in \text{span}(V_e) = \text{spazio instabile}$,

~~$$\lim_{t \rightarrow \infty} x(t) = \infty$$~~

$$\lim_{t \rightarrow -\infty} x(t) = 0$$

⚡

Comportamento generico,
non è dire molto

Sign and square root

Useful formula: $\text{sign}(A) = A(A^2)^{-1/2}$, where $A^{1/2}$ is the principal square root of A (all eigenvalues in the right half-plane), and $A^{-1/2}$ is its inverse.

Proof: consider eigenvalues, $\text{sign}(x) = \frac{x}{(x^2)^{1/2}}$. (Care with signs.)

⇒ se so calcolare sqrt, so calcolare sign

Theorem

If AB has no eigenvalues on $\mathbb{R}_{\leq 0}$ (hence neither does BA), then

$$\text{sign} \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix}, \quad C = A(BA)^{-1/2}.$$

Proof (sketch) Use $\text{sign}(A) = A(A^2)^{-1/2}$ (and then $\text{sign}(A)^2 = I$).

For instance,

$$\text{sign} \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix}.$$

B=I

⇒ se so calcolare sign, so calcolare sqrt

$$\text{sign}(x) = \frac{x}{(x^2)^{1/2}} =$$

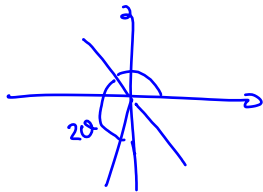
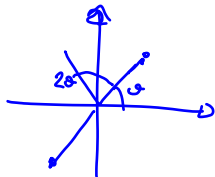
$$= \frac{re^{i\theta}}{\pm re^{i\theta}} = \pm 1$$

$$x = re^{i\theta}$$

$$x^2 = r^2 e^{i2\theta}$$

$$(x^2)^{1/2} = \begin{cases} re^{i\theta} & \text{se } \theta \in \left[\frac{\pi}{2}, \frac{3\pi}{2}\right] \\ -re^{i\theta} & \text{se } \theta \notin \left[\frac{\pi}{2}, \frac{3\pi}{2}\right] \end{cases}$$

Se gli autovalori di A sono
 tutti distinti, $\text{sign}(A) = A(A^2)^{-1/2}$,
 e per continuità vale $\forall A$.



$$\text{Proof} \quad \text{Sign} \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} = \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \left(\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}^2 \right)^{-1/2} =$$

$$= \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} AB & 0 \\ 0 & BA \end{bmatrix}^{-1/2} = \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} \begin{bmatrix} (AB)^{-1/2} & 0 \\ 0 & (BA)^{-1/2} \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & \underline{A(BA)^{-1/2}} \\ B(AB)^{-1/2} & 0 \end{bmatrix} = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}. \quad \text{È vero che } D = B(AB)^{-1/2} = C^{-1}?$$

Sì: segue da $\text{sign} \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}^2 = I$

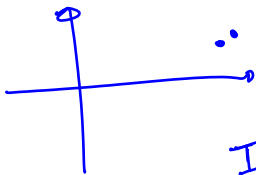
$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} = \begin{bmatrix} CD & 0 \\ 0 & DC \end{bmatrix} \Rightarrow D = C^{-1}$$

Conditioning

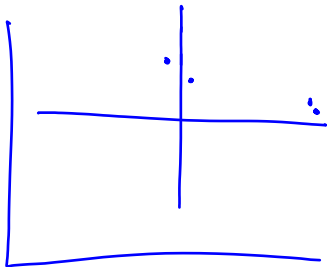
From the theorems on the Fréchet derivative, for a diagonalizable $A = V\Lambda V^{-1}$

$$\kappa_{abs}(\text{sign}(A)) \leq \kappa_2(V) \frac{2}{\min_{\text{Re } \lambda_i < 0, \text{Re } \lambda_j > 0} |\lambda_i - \lambda_j|}$$

Tells only part of the truth: computing $\text{sign}(A)$ is “better” than a full diagonalization: it is not sensitive to close eigenvalues that are far from the imaginary axis



$$\text{sign} \begin{bmatrix} 1.001 & 100 \\ 0 & 1 \end{bmatrix} = I \quad \text{ben conditioned}$$
$$\text{sign} \begin{bmatrix} 0.001 & 100 \\ 0 & -0.001 \end{bmatrix} \quad \text{wel conditioned}$$



Condition number

Theorem

$$\kappa_{abs}(\text{sign}, A) = \|(I \otimes N + N^T \otimes I)^{-1}(I - S^T \otimes S)\|,$$

where $N = (A^2)^{1/2}$, $S = \text{sign}(A)$ $A = SN$

Proof (sketch): let $L = L_{\text{sign}, A}(E)$. Then, up to second-order factors, $(A + E)(S + L) = (S + L)(A + E)$ and $(S + L)^2 = I$. Some manipulations give $NA + AN = E - SES$.

In particular, $\text{sep}(N, -N)$ plays a role.

Remark: if all eigenvalues of A are in the RHP, then the formula gives $\kappa_{abs}(\text{sign}, A) = 0$.

Makes sense, since $\text{sign}(A) = \text{sign}(A + E) = I$ for all E for which eigenvalues do not cross the imaginary axis...

Proof: prendiamo perturbazione $A+E$

$$\text{Sign}(A+E) = S+L + o(\|E\|^2) \quad L_{\text{sign}, A}(E)$$

$$1) (A+E)(S+L) = (S+L)(A+E) + o(\|E\|)$$

$$2) (S+L)^2 + o(\|E\|) = I = S^2 \quad \rightsquigarrow \overline{SL+LS} = o(\|E\|)$$

~~$$AS+AL+ES+EL = SA+SE+LA+LE \quad A=SN$$~~

$$ES - SE = LA - AL \quad \text{multiplico per } S^{-1} = S \text{ a sx.}$$

$$S^{-1}ES - E = \underbrace{S^{-1}LA}_{-LS^{-1}} - S^{-1}AL = -LN - NL$$

$$\underline{NL} + \underline{LN} = E - S^{-1}ES = E - SES$$

$\Rightarrow L$ è la sol. dell'equaz. di Sylvester

$$NL + LN = E - SES$$

$$\text{vec } NL = (I \otimes N) \text{vec } L$$

$$\text{vec } E = (I \otimes I) \cdot \text{vec } E$$

$$\text{vec } LN = (N^T \otimes I) \text{vec } L$$

$$\text{vec } SES = (S^T \otimes S) \text{vec } E$$

$$\Rightarrow (I \otimes N + N^T \otimes I) \text{vec } L = (I - S^T \otimes S) \text{vec } E$$

$$\text{vec } L = (I \otimes N + N^T \otimes I)^{-1} (I - S^T \otimes S) \text{vec } E$$

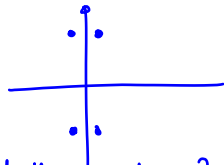
\Rightarrow la forma di Kronecker di $L_{\text{sign}, A}$ è

$$\hat{L} = \underbrace{(I \otimes N + N^T \otimes I)^{-1}} (I - S^T \otimes S)$$

$N = (A^2)^{1/2}$ ha autovalori nel RHP

$\Rightarrow \|I \otimes N + N^T \otimes I\|$ è invertibile perché N e $-N$ non hanno autovalori in comune ($\Lambda(N) \subseteq \text{RHP}$, $\Lambda(-N) \subseteq \text{LHP}$)

$$\|I \otimes N + N^T \otimes I\| = \underline{\text{sep}(N, -N)}.$$



$$\|I - S^T \otimes S\| \leq \|I\| + \|S^T \otimes S\| = 1 + \|S^T\| \cdot \|S\| = 1 + \|S\|^2$$

(se A ha tutti autoval. nel RHP, $\text{sign}(A) = I$)

$$I - S^T \otimes S = I - I \otimes I = 0 \Rightarrow L_{\text{sign} A}(E) = 0$$

(se perturbo A di poco, $\text{sign}(A+E) = I$, quindi $L = 0$)

Schur-Parlett method $Q^*AQ = T$
 $f(A) = Q f(T) Q^*$

We can compute $\text{sign}(A)$ with a Schur decomposition. Simplest case: the decomposition is ordered so that eigenvalues in the LHP come first: $\Lambda(T_{11}) \subseteq LHP$, $\Lambda(T_{22}) \subseteq RHP$. $f(T_{11}) = -I$ $f(T_{22}) = I$

$$Q^*AQ = \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}, \quad Q^*f(A)Q = \begin{bmatrix} -I & X \\ 0 & I \end{bmatrix}$$

where X solves $T_{11}X - XT_{22} = -f(T_{11})T_{12} + T_{12}f(T_{22}) = 2T_{12}$.

Condition number of this Sylvester equation: depends on $\text{sep}(T_{11}, T_{22})$.

$$\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix} \begin{bmatrix} -I & X \\ 0 & I \end{bmatrix} = \begin{bmatrix} -I & X \\ 0 & I \end{bmatrix} \begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}$$

$T \quad f(T) \quad f(T) \quad T$

$$T_{11}X - XT_{22} = T_{12} + T_{12} = 2T_{12}$$

Schur-Parlett for the sign

1. Compute $A = QTQ^T$.

2. Reorder Schur decomposition so that eigenvalues in the LHP come first.

3. Solve Sylvester equation for X .

4. $\text{sign}(A) = Q \begin{bmatrix} -I & X \\ 0 & I \end{bmatrix} Q^T$.

→ $\text{sep}(T_{11}, T_{22})$

Newton for the matrix sign

Most popular algorithm:

Newton for the matrix sign

$\text{sign}(A) = \lim_{k \rightarrow \infty} X_k$, where

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}), \quad X_0 = A.$$

Suppose A diagonalizable: then we may consider the scalar version of the iteration on each eigenvalue λ :

$$x_{k+1} = \frac{1}{2} \left(x_k + \frac{1}{x_k} \right) = \frac{x_k^2 + 1}{2x_k}, \quad x_0 = \lambda.$$

Fixed points: ± 1 (with local quadratic convergence). Eigenvalues in the RHP stay in the RHP (and same for LHP).

(It's Newton's method on $f(x) = x^2 - 1$, which justifies the name).

$$\frac{1}{2}(VDV^{-1} + (VDV^{-1})^{-1}) = \frac{1}{2}(VDV^{-1} + VD^{-1}V^{-1}) =$$

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

↔ se A diagonalizzabile,
equivale ad applicare la
versione scalare agli autovalori

$$X_{k+1} = \frac{1}{2} \left(X_k + \frac{1}{X_k} \right)$$

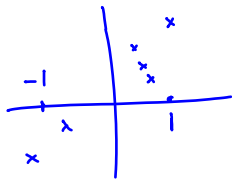
Se $X_k \in \text{RHP}$, $\frac{1}{X_k} \in \text{RHP}$,

⇒ $X_{k+1} \in \text{RHP}$

Punti fissi: $X = \frac{1}{2} \left(X + \frac{1}{X} \right)$

$$\frac{1}{2}X = \frac{1}{2} \frac{1}{X} \quad X = \frac{1}{X}$$

$$X = \pm 1$$



Questo mappa è il metodo di Newton
applicato a $f(x) = x^2 - 1$

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

(converge quadraticamente.)

$$g(x) = \frac{x^2 + 1}{2x} \quad \text{da } \mathbb{C} \cup \{\infty\} \text{ in sé}$$

Convergence analysis of the scalar iteration

Trick: change of variables (Cayley transform)

$$y = \frac{1+x}{1-x}, \text{ with inverse } x = \frac{y-1}{y+1}.$$

If $x \in \text{RHP}$, then $|x+1| > |x-1| \implies y$ outside the unit disk.

If $x \in \text{LHP}$, then $|x-1| > |x+1| \implies y$ inside the unit disk.

("Poor man's exponential")

$x_{k+1} = \frac{1}{2} \left(x_k + \frac{1}{x_k} \right)$ corresponds to $y_{k+1} = y_k^2$ (check it!).

If we start from $x_0 \in \text{LHP}$, then $|y_0| < 1$, then $\lim y_k = 0$ (i.e., $\lim x_k = -1$). ✓

If we start from $x_0 \in \text{RHP}$, then $|y_0| > 1$, the squares diverge, and $\lim y_k = \infty$ (i.e., $\lim x_k = 1$). ✓

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



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

$$x_0, x_1, x_2, \dots$$

$$y_k = \frac{1+x_k}{1-x_k}$$

$$y_0, y_1, y_2, \dots$$

$$x_{k+1} = g(x_k) \text{ diventa } y_{k+1} = y_k^2$$

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

$$= -y_k^2$$

\Rightarrow genera la successione $y_0, y_0^2, y_0^4, y_0^8, \dots$

Se $y_0 \in$ disco unitario, $y_k \rightarrow 0$

se $y_0 \notin$ disco unitario, $y_k \rightarrow \infty$

Nota che la mappa $g(x) = \frac{1+x}{1-x}$ è tale che:

se $x \in \text{RHP}$,
a+ib $|1+x| > |1-x| \Rightarrow y = g(x)$ fuori dal disco unitario

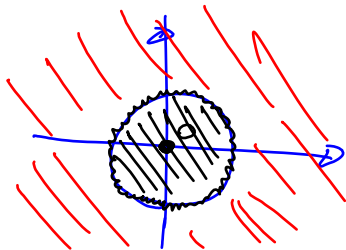
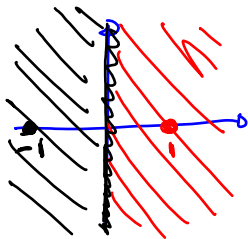
$$\frac{|a+1+ib|}{\sqrt{(a+1)^2+b^2}} > \frac{|1-a-ib|}{\sqrt{(a-1)^2+b^2}}$$

Se $x \in \text{LHP}$,
 $|1+x| < |1-x| \Rightarrow y$ dentro il disco unitario

$$g(-1) = 0$$

g

$$g(i) = \infty$$



$$X_0 \in \text{LHP} \Rightarrow y_0 \in \text{disco} \Rightarrow y_k \rightarrow 0 \Rightarrow X_k \rightarrow -1$$

$$X_0 \in \text{RHP} \Rightarrow y_0 \in \text{esterno disco} \Rightarrow y_k \rightarrow \infty \Rightarrow X_k \rightarrow 1$$

$$\mathbb{C} \cup \{\infty\} \xrightarrow{g} \mathbb{C} \cup \{\infty\}$$

$$\varphi(x) = \frac{1+x}{1-x}$$



$$\downarrow \frac{1+x}{1-x}$$

$$\mathbb{C} \cup \{\infty\} \xrightarrow{\text{quadrato}} \mathbb{C} \cup \{\infty\}$$

$$\begin{array}{ccc} X_k & \xrightarrow{g} & X_{k+1} = \left(X_k + \frac{1}{X_k}\right)^{\frac{1}{2}} \\ \varphi \downarrow & & \downarrow \varphi \\ y_k & \xrightarrow{\text{quadrato}} & y_{k+1} = -y_k^2 \end{array}$$

Anche $\exp(x)$ mappa LHP \rightarrow cerchio unitario
RHP \rightarrow esterno

$$\frac{1 + \frac{1}{2}x}{1 - \frac{1}{2}x} \approx 1 + x \approx \exp(x) \text{ per } x \text{ piccolo}$$

Convergence analysis of the matrix iteration

The same proof works, as long as A does not have the eigenvalue 1 (invertibility). Small modification to fix this case, too:

Change of variables: $S = \text{sign}(A)$ $A = X_0$

$$Y_k = \boxed{(X_k - S)(X_k + S)^{-1}} \quad \text{with inverse } X_k = (I - Y_k)^{-1}(I + Y_k)S.$$

All the X_k are rational functions of A , so they commute with it and with S .

Analyzing eigenvalues: the inverse exists and $\rho(Y_k) < 1$.

$$Y_{k+1} = (X_k^{-1}(X_k^2 + I - 2SX_k))X_k(X_k^2 + I + 2SX_k)^{-1} = Y_k^2.$$

$Y_k \rightarrow 0$, hence $X_k \rightarrow S$.

$$X_k = f^k(A), \quad \text{dove } f(x) = \frac{1}{x}$$

Se A diagonalizzabile,

$$V^{-1}AV = \begin{bmatrix} \lambda_1 & & \\ & \lambda_2 & \\ & & \ddots \\ & & & \lambda_n \end{bmatrix}$$

$$V^{-1}SV = \begin{bmatrix} \pm 1 & & \\ & \pm 1 & \\ & & \ddots \\ & & & \pm 1 \end{bmatrix} = \begin{bmatrix} \operatorname{sgn}(\lambda_1) & & \\ & \ddots & \\ & & \operatorname{sgn}(\lambda_n) \end{bmatrix}$$

$$Y_0 = (X_0 - S)(X_0 + S)^{-1} = \left(V \begin{bmatrix} \lambda_1 - \operatorname{sgn}(\lambda_1) & & \\ & \ddots & \\ & & \lambda_n - \operatorname{sgn}(\lambda_n) \end{bmatrix} V^{-1} \right) \left(V \begin{bmatrix} \lambda_1 + \operatorname{sgn}(\lambda_1) & & \\ & \ddots & \\ & & \lambda_n + \operatorname{sgn}(\lambda_n) \end{bmatrix} V^{-1} \right)^{-1} =$$

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

$\lambda_i + \text{sgn}(\lambda_i) \neq 0$ perché $\lambda_i, \text{sgn}(\lambda_i)$ stanno
tutti e due in LHP o RHP \Rightarrow quota trasformat.
è sempre ben definita, $(A+S)^{-1}$ esiste sempre.

Anche se A è non diag. zobile, facendo una
forma di Jordan, $A+S \sim \left(\begin{array}{c} \parallel \\ \parallel \end{array} \right) + \left(\begin{array}{c} \backslash \\ \backslash \end{array} \right)$
non ho mai autoval. zero, perché sulla diagonale
ho $\lambda_i + \text{sgn}(\lambda_i)$

Se A non ha autoval. sull'asse immaginario

$$\left| \frac{\lambda_i - \operatorname{sgn}(\operatorname{Re}(\lambda_i))}{\lambda_i + \operatorname{sgn}(\operatorname{Re}(\lambda_i))} \right| < 1$$

$$\text{se } \lambda_i = a + bi$$

$$\frac{\sqrt{(a - \operatorname{sgn}(a))^2 + b^2}}{\sqrt{(a + \operatorname{sgn}(a))^2 + b^2}} < 1$$

perché $a - \operatorname{sgn}(a) < a + \operatorname{sgn}(a)$.

$$Y_{k+1} = (X_{k+1} - S)(X_{k+1} + S)^{-1}$$

$$Y_k = (X_k - S)^{-1}(X_k + S)^{-1} \quad X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$$

Oss:

The algorithm

1. $X_0 = A$.
2. Repeat $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$, until convergence.

We really need to compute that matrix inverse (unusual in numerical linear algebra...)

Scaling

If $x_k \gg 1$, then

$$x_0 = 10^4 \quad x_1 = \frac{1}{2} \left(10^4 + \frac{1}{10^4} \right) \approx \frac{1}{2} 10^4$$
$$x_2 = \frac{1}{2} \left(\frac{2}{10^4} + \frac{10^4}{2} \right) \approx \frac{10^4}{4} \dots$$

$$x_{k+1} = \frac{1}{2} \left(x_k + \frac{1}{x_k} \right) \approx \frac{1}{2} x_k,$$

and “the iteration is an expensive way to divide by 2” [Higham].

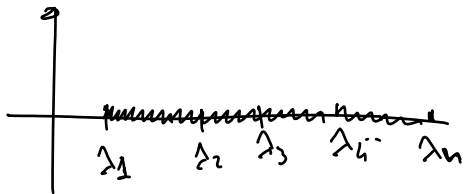
Same if $x_k \ll 1$ — the iteration just multiplies by 2.

Similarly, for matrices, convergence cannot occur until each eigenvalue has converged to ± 1 .

Trick: replace A with μA for a scalar $\mu > 0$ — they have the same sign. Choose this μ so that “eigenvalues ≈ 1 .” *non sempre possibile*
(Once, or at each step.)

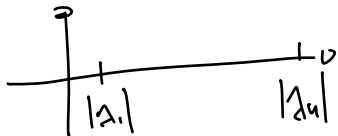
$$\lambda_i(\mu A) = \mu \lambda_i(A) \quad \text{sgn}(\lambda_i(\mu A)) = \text{sgn}(\lambda_i(A))$$

Ad es., se lo spettro di A è reale



Scelando, cambio punto intervallo

Intervallo ideale: quando $|\lambda_1| \cdot |\lambda_n| = 1$



Se punto da $\hat{X}_0 > X_0 > 1$, allora ad ogni passo
 $\hat{X}_k > X_k > 1$

Le iterazioni che converge più piano sono

quella che partono da λ_1 e λ_n

(se $|\lambda_1| < 1$, $|\lambda_n| > 1$)

→ convergenza più veloce $\max(|\lambda_n|, \frac{1}{|\lambda_1|})$ è più piccolo

Ottenute quando $|\lambda_1| \cdot |\lambda_n| = 1$

(se $|\lambda_1| \cdot |\lambda_n| > 1$, allora λ_n converge più piano di λ_1 , e posso spostare gli autovalori a sx;

se $|\lambda_1| \cdot |\lambda_n| < 1$, allora λ_1 converge più piano, e posso spostare a dx.

Scaling possibilities *(geometric)*

Possibility 1: (determinantal scaling): choose $\mu = (\det A)^{-1/n}$, so that $\det \mu A = 1$. Reduces "mean distance" from 1. Cheap to compute, since we already need to invert A .

Possibility 2: (spectral scaling): choose μ so that

$|\lambda_{\min}(\mu A)| |\lambda_{\max}(\mu A)| = 1$. (We can use the power method to estimate them.)

Possibility 3: (norm scaling): choose μ so that

$\sigma_{\min}(\mu A) \sigma_{\max}(\mu A) = 1$. (Again via the power method for σ_{\min} .)

Surprisingly, on a matrix with real eigenvalues Possibility 2 gives convergence in a finite number of iterations, if done at each step: the first iteration maps $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ to eigenvalues with the same modulus; then the second iteration adds a third eigenvalue with the same modulus...

when done at each step

$$\|A\| \cdot \|A^{-1}\|$$

Situations in cui det. Scaling funzione male:



Other iterations

There is an elegant framework to determine other iterations locally convergent to $\text{sign}(x)$ (in a neighbourhood of ± 1): start from

$$\text{sign}(z) = \frac{z}{(z^2)^{1/2}}$$

and replace the square root using a Padé approximant of $(1-x)^{1/2}$.

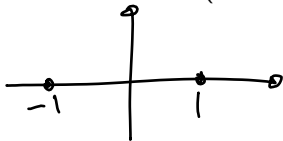
In the end, they produce iteration functions of the form

$$f_r(z) = \frac{(1+z)^r + (1-z)^r}{(1+z)^r - (1-z)^r}$$

es. ($r=2$)

$$\frac{z^2+1}{2z}$$

Advantage of using the Newton-sign iteration: it has the correct basins of attraction (convergence is global and not only local).



$$f_r(z) - 1 = o(z^r) \quad \text{per } z \rightarrow 1$$

Alcune iterazioni iniziali convergono a A ed ogni passo [ad es. Newton per $A^{1/2}$, $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A)$]

Il punto iniziale spesso non importa: è un Newton, converge (localmente) sempre. È "self-correcting":

errori in un passo (ad es. X_2^{-1} mal condizionata) vengono corretti.

Altre dipendono da A solo tramite i valori iniziali: ad es. Newton per il segno. Errori

ad ogni passo (ad es. X_2^{-1} mal condizionata) si accumulano.

Stability of the sign iterations

The stability analysis is complicated (and not even done completely in articles). [Bai Demmel '98 and Byers Mehrmann He '97]

While it works well in practice, the Newton iteration is **not** backward stable. (non posso scrivere $f(x_i) = \frac{1}{2}(\hat{x}_i + \hat{x}_{i-1})$).

The sign is not even stable under small perturbations: assuming (up to a change of basis) $A = \begin{bmatrix} A_{11} & A_{12} \\ 0 & A_{22} \end{bmatrix}$, then $\Lambda(A_{11}) \subseteq \text{RHP}$

$\Lambda(A_{22}) \subseteq \text{LHP}$

$\ker \text{sign}(A) \pm I$

$$\|\text{sign}(A + E) - \text{sign}(A)\| \lesssim \frac{\|E\|}{\text{sep}(A_{11}, A_{22})^3}$$

Nevertheless, the invariant subspaces it produces are: $A + E$ has a stable invariant subspace of the form $\begin{bmatrix} X \\ 0 \end{bmatrix}$, with (A has an inv. subsp. $\begin{bmatrix} X \\ 0 \end{bmatrix}$)

$$\|X\| \lesssim \frac{\|E\|}{\text{sep}(A_{11}, A_{22})}$$

(The invariant subspace stability bound from the first lectures.)

Inversion-free sign

Suppose that we are given M, N such that $A = M^{-1}N$. Can we compute $\text{sign}(A)$ without inverting M ? **Yes.**

$$\begin{aligned}X_1 &= \frac{1}{2}(A + A^{-1}) = \frac{1}{2}(M^{-1}N + N^{-1}M) \\&= \frac{1}{2}M^{-1}(N + MN^{-1}M) \\&= \frac{1}{2}M^{-1}(N + \hat{M}^{-1}\hat{N}M) \\&= \frac{1}{2}M^{-1}\hat{M}^{-1}(\hat{M}N + \hat{N}M) \\&= (\hat{M}M) \frac{1}{2}(\hat{M}N + \hat{N}M) =: M_1^{-1} N_1.\end{aligned}$$

assuming we can find \hat{M}, \hat{N} such that $MN^{-1} = \hat{M}^{-1}\hat{N}$.

Then the same computations produce $M_2, N_2, M_3, N_3, \dots$

Dati M, N tali che $A = M^{-1}N$

(stessa situazione di $\text{eig}(N, M)$), posso fare

l'iteraz. sopra senza usare M^{-1} : $X_0 = M^{-1}N$

$$\begin{aligned} X_1 &= \frac{1}{2} (X_0 + X_0^{-1}) = \frac{1}{2} (M^{-1}N + N^{-1}M) = \\ &= \frac{1}{2} M^{-1} (N + \underbrace{MN^{-1}M}) = \frac{1}{2} M^{-1} (N + \hat{M}^{-1} \hat{N} M) = \\ &= \frac{1}{2} M^{-1} \hat{M}^{-1} (\hat{M}N + \hat{N}M) = (2\hat{M}M) \end{aligned}$$

(se trovo \hat{M}, \hat{N} tali che $MN^{-1} = \hat{M}^{-1}\hat{N}$)

Inversion-free sign

How to find \hat{M}, \hat{N} such that $MN^{-1} = \hat{M}^{-1}\hat{N}$?

$\hat{M}M = \hat{N}N$, or $\begin{bmatrix} \hat{M} & \hat{N} \end{bmatrix} \begin{bmatrix} M \\ -N \end{bmatrix} = 0$. We can obtain \hat{M}, \hat{N} from a kernel.

Computing this kernel can be much more accurate than inverting M and/or N , e.g.,

$$\begin{bmatrix} M \\ -N \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \varepsilon \\ \varepsilon & 0 \\ 0 & 1 \end{bmatrix}.$$