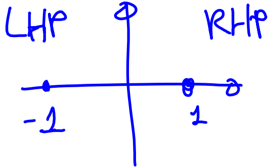
The matrix sign function



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

Suppose the Jordan form of A is reblocked as

$\Lambda(J_1) \subseteq \text{LHP}$

$f(J_1) = \begin{bmatrix} -1 & & \\ & \ddots & \\ & & -1 \end{bmatrix}$

$$A = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} J_1 & \\ & J_2 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1},$$

$f(J_2) = I$

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

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

$$\boxed{\operatorname{sign}(A) = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} -I & \\ & I \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1}.}$$

$\operatorname{sign}(A)^2 = I$

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

$$\operatorname{sign}(A) = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} -1 & \\ & 1 \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1}$$

$$\frac{1}{2}\left(\operatorname{sign}(A) + I\right) = \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \frac{1}{2}(-I+I) & \\ & \frac{1}{2}(I+I) \end{bmatrix} \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1} =$$

$$= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \left[\begin{array}{c|c} 0 & 0 \\ \hline 0 & I \end{array}\right] \begin{bmatrix} V_1 & V_2 \end{bmatrix}^{-1} = \text{Proiettore su span}(V_2)$$

immagine V_2, kernel V_1

$$-\frac{1}{2}\left(\operatorname{sign}(A) - I\right) = \text{Proiettore su span}(V_1) \to \text{immagine } V_1,$$

kernel V_2

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

$V_2 =$ "spazio instabile"

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

$$\boxed{\begin{cases} x(0) = b \\ \dot{x} = Ax \end{cases}} \qquad x(t) = \exp(tA)\, b$$

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

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

$$= \begin{bmatrix} V_1 & V_2 \end{bmatrix} \begin{bmatrix} \exp(tJ_1) & \\ & \exp(tJ_2) \end{bmatrix} \begin{bmatrix} c \\ 0 \end{bmatrix} = V_1 \exp(tJ_1)\, c \to 0$$

se J_1 ha solo autoval. negativi,

$$\lim_{t\to\infty} \exp(tJ_1) = 0$$ e il limite fa 0.

Similmente, se $b \in \mathrm{span}(V_2) =$ spazio instabile,

$$\cancel{\lim_{t\to\infty} x(t) = \infty}$$

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

$\Uparrow$

comportamento generico,

non è dire molto

Sign and square root

Useful formula: $\operatorname{sign}(A) = A(A^2)^{-1/2}$, 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, $\operatorname{sign}(x) = \frac{x}{(x^2)^{1/2}}$. (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

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

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

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

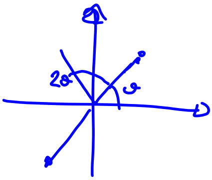
B=I => se so calcolare sign, so calcolare sqrt

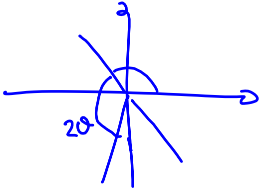
$$\mathrm{sign}(x) = \frac{x}{(x^2)^{1/2}} =$$

$$x = re^{i\vartheta} \qquad x^2 = r^2 e^{i2\vartheta}$$

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

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





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

Proof

$$\operatorname{sign}\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} = \begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}\left(\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}^2\right)^{-1/2} =$$

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

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

È vero che $D = B(AB)^{-1/2} = C^{-1}$?

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

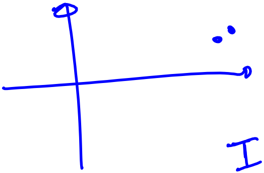
$$\begin{bmatrix} I & 0 \\ 0 & I \end{bmatrix} = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} = \begin{bmatrix} CD & 0 \\ 0 & DC \end{bmatrix} \Rightarrow D = C^{-1}$$

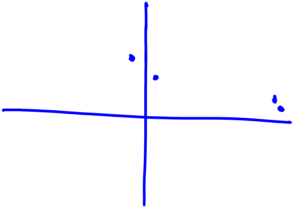
Conditioning

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

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

Tells only part of the truth: computing sign(A) is "better" than a full diagonalization: it is not sensitive to close eigenvalues that are far from the imaginary axis.





$$\operatorname{sign}\begin{bmatrix} 1.001 & 100 \\ 0 & 1 \end{bmatrix} = I$$

ben condizionato

$$\operatorname{sign}\begin{bmatrix} 0.001 & 100 \\ 0 & -0.001 \end{bmatrix}$$

mal condizionato

Condition number

Theorem

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

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

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 $A+E$

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

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

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

$$\cancel{AS} + AL + ES + \cancel{EL} = \cancel{SA} + SE + LA + \cancel{LE} \qquad A = SN$$

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

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

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

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

$$NL + LN = E - SES$$

$$\mathrm{vec}\, NL = (I \otimes N)\, \mathrm{vec}\, L \qquad \mathrm{vec}\, E = (I \otimes I) \cdot \mathrm{vec}\, E$$

$$\mathrm{vec}\, LN = (N^T \otimes I)\, \mathrm{vec}\, L \qquad \mathrm{vec}\, SES = (S^T \otimes S)\, \mathrm{vec}\, E$$

$$\Rightarrow \left(I \otimes N + N^T \otimes I\right) \mathrm{vec}\, L = \left(I - S^T \otimes S\right) \mathrm{vec}\, E$$

$$\mathrm{vec}\, L = \left(I \otimes N + N^T \otimes I\right)^{-1} \left(I - S^T \otimes S\right) \mathrm{vec}\, E$$

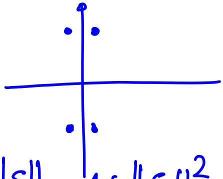
$\Rightarrow$ la forma di Kronecker di $L_{sign,A}$ è

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

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

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

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



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

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

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

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

Schur-Parlett method

$Q^*AQ = T$
$f(A) = Q f(T) Q^*$

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

$f(T_{11}) = -I \quad f(T_{22}) = I$

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

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 sep(T_{11}, T_{22}).

$$\underbrace{\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}}_{T} \underbrace{\begin{bmatrix} -1 & X \\ 0 & 1 \end{bmatrix}}_{f(T)} = \underbrace{\begin{bmatrix} -1 & X \\ 0 & 1 \end{bmatrix}}_{f(T)} \underbrace{\begin{bmatrix} T_{11} & T_{12} \\ 0 & T_{22} \end{bmatrix}}_{T}$$

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

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

$\operatorname{sep}(T_{11}, T_{22})$

Newton for the matrix sign

Most popular algorithm:

Newton for the matrix sign

$\mathrm{sign}(A) = \lim_{k\to\infty} X_k$, 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:

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

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).

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

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

$\Leftrightarrow$ se A diagonalizzabile, equivale ad applicare la versione scalare agli autovalori

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

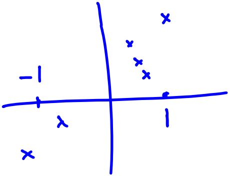
Punti fissi: $x = \frac{1}{2}\left(x + \frac{1}{x}\right)$

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

$x = \pm 1$

Se $x_k \in RHP$, $\frac{1}{x_k} \in RHP$,

$\Rightarrow x_{k+1} \in RHP$



Questa 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.)

$g(x) = \dfrac{x^2+1}{2x}$ da $\mathbb{C} \cup \{\infty\}$ 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 $x \in$ RHP, then $|x+1| > |x-1| \implies y$ outside the unit disk.

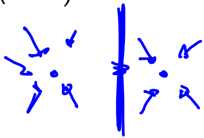
If $x \in$ LHP, then $|x-1| > |x+1| \implies y$ inside the unit disk.
("Poor man's exponential")

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

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

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

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



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

$x_0, x_1, x_2 \ldots\ldots$

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

$y_0, y_1, y_2, \ldots$

$x_{k+1} = g(x_k)$ 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 $y_0, y_0^2, y_0^4, y_0^8, \ldots$

Se parto da $y_0 \in$ disco unitario, $y_k \to 0$

se parto da $y_0 \notin$ disco unitario, $y_k \to \infty$

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

Se $x \in RHP$, $|1+x| > |1-x|$ $\Rightarrow$ $y = g(x)$ fuori del disco unitario

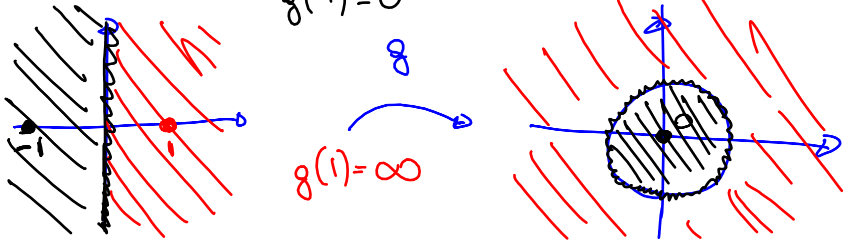
$a+ib$

$|a+1+ib|$ $|1-a-ib|$

$\sqrt{(a+1)^2+b^2}$ $\sqrt{(a-1)^2+b^2}$

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

$g(-1) = 0$



$x_0 \in LHP \Rightarrow y_0 \in$ disco $\Rightarrow y_k \to 0 \Rightarrow x_k \to -1$

$x_0 \in RHP \Rightarrow y_0 \in$ esterno disco $\Rightarrow y_k \to \infty \Rightarrow x_k \to 1$

$$\begin{tikzcd}
\mathbb{C}\cup\{\infty\} \arrow[r, "g"] \arrow[d, "\varphi(x)=\frac{1+x}{1-x}"'] & \mathbb{C}\cup\{\infty\} \arrow[d, "\frac{1+x}{1-x}"] \\
\mathbb{C}\cup\{\infty\} \arrow[r, "\text{quadrato}"] & \mathbb{C}\cup\{\infty\}
\end{tikzcd}$$

$$\begin{tikzcd}
x_k \arrow[r, "g"] \arrow[d, "\varphi"'] & x_{k+1}=\left(x_k+\frac{1}{x_k}\right)\frac{1}{2} \arrow[d, "\varphi"] \\
y_k \arrow[r, "\text{quadrato}"] & y_{k+1}=-y_k^2
\end{tikzcd}$$

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

RHP $\rightarrow$ esterno

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

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: $S = \operatorname{sign}(A)$ $A = X_0$

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

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

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

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

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

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

Se A diagonalizzabile,

$$V^{-1}AV = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix} \qquad V^{-1}SV = \begin{bmatrix} \pm 1 & & & \\ & \pm 1 & & \\ & & \ddots & \\ & & & \pm 1 \end{bmatrix} = \begin{bmatrix} \operatorname{sgn}(\lambda_1) & & \\ & \ddots & \\ & & \operatorname{sgn}(\lambda_n) \end{bmatrix}$$

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

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

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

Anche se A è non diag.zabile, facendo una forma di Jordan, $A+S \sim \begin{pmatrix} \ddots & \ddots & \\ & \ddots & \ddots \end{pmatrix} + \begin{pmatrix} \ddots & \\ & \ddots \end{pmatrix}$

non ho mai autoval. zero, perché sulla diagonale ho $\lambda_i + \mathrm{sgn}(\lambda_i)$

Se A non ha autoval. sull'asse immaginario

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

se $\lambda_i = a + bi$

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

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

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

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

<u>Oss</u>:

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

$X_0 = 10^4 \quad X_1 = \frac{1}{2}\left(10^4 + \frac{1}{10^4}\right) \approx \frac{1}{2}10^4$

$X_2 = \frac{1}{2}\left(\frac{2}{10^4} + \frac{10^4}{2}\right) \approx \frac{10^4}{4} \dots$

If $x_k \gg 1$, then

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

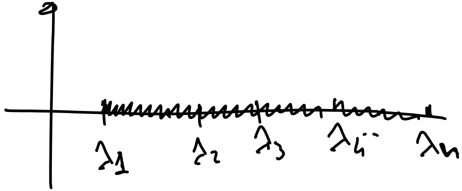
and "the iteration is an expensive way to divide by 2" [Higham].
Same if $x_k \ll 1$ — the iteration just multiplies by 2.

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." $\rightsquigarrow$ non sempre possibile
(Once, or at each step.)

$\lambda_i(\mu A) = \mu \lambda_i(A) \qquad \mathrm{sgn}(\lambda_i(\mu A)) = \mathrm{sgn}(\lambda_i(A))$

Ad es., se lo spettro di A è reale



Scalando, cambio questo intervallo

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

Se parto da $\hat{x}_0 > x_0 > 1$, allora ad ogni passo

$$\hat{x}_k > x_k > 1$$

Le iterazioni che convergono più piano sono quelle che partono da λ_1 e λ_n (se $|\lambda_1| < 1$, $|\lambda_n| > 1$)

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

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

(se $|\lambda_1||\lambda_n| > 1$, 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 $\boxed{\mu} = (\det A)^{-1/n}$, 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.)

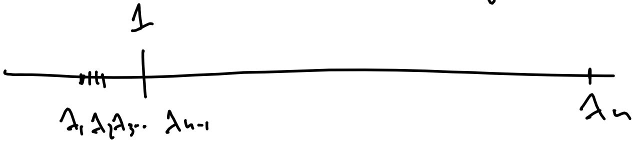
Possibility 3: (norm scaling): choose μ so that $\sigma_{\min}(\mu A)\sigma_{\max}(\mu A) = 1$. (Again via the power method for $\sigma_{\min}$.)

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

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

situazione in cui det. scaling funziona male:



Other iterations

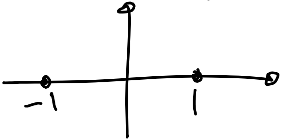
There is an elegant framework to determine other iterations locally convergent to $\operatorname{sign}(x)$ (in a neighbourhood of ± 1): start from

$$\boxed{\operatorname{sign}(z) = \frac{z}{(z^2)^{1/2}},}$$

and replace the square root using a Padé approximant of $(1-x)^{1/2}$.
In the end, they produce iteration functions of the form

$$\boxed{f_r(z) = \frac{(1+z)^r + (1-z)^r}{(1+z)^r - (1-z)^r}.} \quad \text{es. } (r=2) \qquad \frac{z^2+1}{2z}$$

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



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

Alcune iterazioni matriciali contengono A ad ogni passo [ad es. Newton per $A^{1/2}$, $X_{k+1}=\frac{1}{2}\left(X_k+X_k^{-1}A\right)$

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 $fl(X_1) = \frac{1}{2}(\hat{X}_0 + \hat{X}_0^{-1})$).

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 $\Lambda(A_{11}) \subseteq RHP$, $\Lambda(A_{22}) \subseteq LHP$

Ker $\operatorname{sgn}(A) \pm I$

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

Nevertheless, the invariant subspaces it produces are: $A + E$ has a stable invariant subspace of the form $\begin{bmatrix} I \\ X \end{bmatrix}$, with (A ha un sottosp. inv. $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$)

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

(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 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 $M_2, N_2, M_3, N_3, \ldots$

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

(stessa situazione di eig(N,M)), posso fare

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

$$X_1 = \frac{1}{2}\left(X_0 + X_0^{-1}\right) = \frac{1}{2}\left(M^{-1}N + N^{-1}M\right) =$$

$$= \frac{1}{2}M^{-1}\left(N + \underbrace{MN^{-1}}M\right) = \frac{1}{2}M^{-1}\left(N + \hat{M}^{-1}\hat{N}\,M\right) =$$

$$= \frac{1}{2}M^{-1}\hat{M}^{-1}\left(\hat{M}N + \hat{N}M\right) = \underbrace{\left(2\hat{M}M\right)^{-1}}_{M_1}\underbrace{\left(\hat{M}N + \hat{N}M\right)}_{N_1}$$

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

e posso continuare calcolando M_2, N_2 da M_1, N_1, eccetera.

Come trovare $\hat{M}, \hat{N}$:

$$MN^{-1} = \hat{M}^{-1}\hat{N} \iff \hat{M}M = \hat{N}N \iff \begin{bmatrix} \hat{M} & \hat{N} \end{bmatrix} \begin{bmatrix} M \\ -N \end{bmatrix} = 0$$

$\begin{bmatrix} \hat{M} & \hat{N} \end{bmatrix} \in \mathbb{C}^{n \times 2n}$ è il kernel sinistro di $\begin{bmatrix} M \\ -N \end{bmatrix} \in \mathbb{C}^{2n \times n}$

$\begin{bmatrix} M \\ -N \end{bmatrix}$ ha rango pieno (se M invertibile), quindi esisterà un sottospazio n-dim. di vettori riga lunghi $2n$ tali che
$v \begin{bmatrix} M \\ -N \end{bmatrix} = 0$

Per es., $\begin{bmatrix} M \\ -N \end{bmatrix} = \begin{bmatrix} Q \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix}$

$$\begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} = 0 \quad \Rightarrow \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} Q^T \end{bmatrix} \begin{bmatrix} Q \end{bmatrix} \begin{bmatrix} R \\ 0 \end{bmatrix} =$$

$$= \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} Q^T \end{bmatrix} \begin{bmatrix} M \\ -N \end{bmatrix} = 0$$

$\Rightarrow$ se $Q = \begin{bmatrix} Q_{11} & Q_{12} \\ Q_{21} & Q_{22} \end{bmatrix}$ $\quad \begin{bmatrix} 0 & I \end{bmatrix} \begin{bmatrix} Q_{11}^T & Q_{21}^T \\ Q_{12}^T & Q_{22}^T \end{bmatrix} = \begin{bmatrix} Q_{12}^T & Q_{22}^T \end{bmatrix}$

$\Rightarrow \begin{bmatrix} \hat{M} & \hat{N} \end{bmatrix} = \begin{bmatrix} Q_{12}^T & Q_{22}^T \end{bmatrix}$ è una possibile scelta

(e l'ho trovata senza invertire M).

Posso rimpiazzare $\hat{M}, \hat{N}$ con $K\hat{M}, K\hat{N}$ per una $K \in \mathbb{C}^{n \times n}$ invertibile qualunque, e viene un'altra soluzione valida di $\begin{bmatrix} \hat{M} & \hat{N} \end{bmatrix} \begin{bmatrix} M \\ -N \end{bmatrix} = 0$

Risostituendo, produce (KM_1, KN_1) al posto di (M_1, N_1) che vanno altrettanto bene perché mi interessava che $M_1^{-1} N_1 = X_1$

(in un certo senso, vuol dire "applicare l'iterazione sopra a un pencil": parto da $N - \lambda M$, produco $N_1 - \lambda M_1$

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.,

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