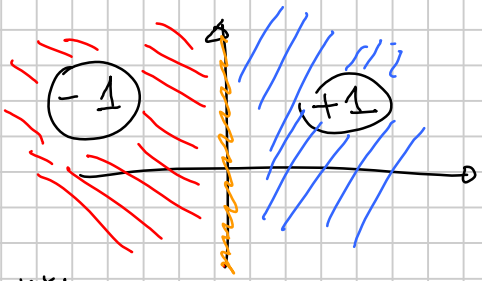
The matrix sign function

Note Title 2025-04-04

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

RHP (right half-plane)

LHP (left " ")



We shall assume that $M \in \mathbb{C}^{n\times n}$ has no purely imaginary eigenvalues $\operatorname{sign}(M)$ is defined, then.

Suppose

$$M = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} J_1 & \\ \hline & J_2 \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1}$$

is a Jordan form, reblocked such that $\Lambda(J_1) \subset$ LHP, $\Lambda(J_2) \subset$ RHP. Then,

$$S = \operatorname{sign}(M) = \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} -I & \\ \hline & I \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1}$$

So $\Lambda(\operatorname{sign}(M)) \subset \{-1, +1\}$

If $\Lambda(M) \subset$ RHP (no eigenvalues in LHP), $\operatorname{sign}(M) = I$.

Recall that $\operatorname{Im} V_1$, $\operatorname{Im} V_2$ are the invariant subspaces of M associated to the LHP, RHP respectively.

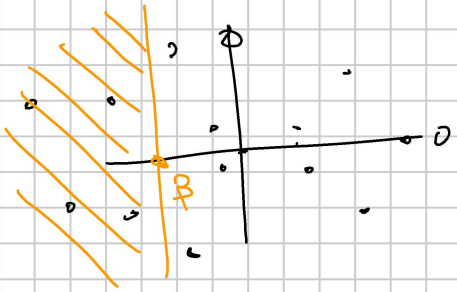
In particular, once we have computed $S = \operatorname{sign}(M)$, then we can compute $\operatorname{Im} V_1$, $\operatorname{Im} V_2$:

$$\ker(S+I) = \ker \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right] \left[\begin{array}{c|c} 0 & \\ \hline & 2I \end{array}\right] \left[\begin{array}{c|c} V_1 & V_2 \end{array}\right]^{-1} = \operatorname{Im} V_1$$

$$\ker(S-I) = \operatorname{Im} V_2$$

The sign can be used as a method to compute certain invariant subspaces.

Application in physics: computing the invariant subspace associated to the leftmost (in the complex plane) eigenvalues of a certain matrix.

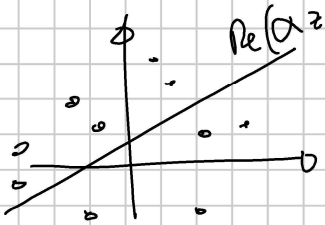


invariant subspace associated to the orange region: $\ker\left(\text{sign}(M+\beta I)+I\right)$

Computing eigenvalues by bisection with the sign:

Choose $\alpha, \beta \in \mathbb{C}$, compute $S = \text{sign}(\alpha M + \beta I)$

$\ker(S+I)$ = inv. subspace associated to the half-plane $\{z : \text{Re}(\alpha z+\beta) < 0\}$



$$Q = qr(S+I)$$

$$\Rightarrow \quad Q^* M Q = \left[\begin{array}{c|c} A & C \\ \hline 0 & B \end{array}\right]$$

$$\Lambda(A) = \Lambda(M) \cap \{z : \text{Re}(\alpha z+\beta) < 0\}$$
$$\Lambda(B) = \Lambda(M) \cap \{z : \text{Re}(\alpha z+\beta) > 0\}$$

This splits the spectrum of M into two; one can proceed recursively.

We see two methods to compute it.

Method 1: Schur-Parlett.

Compute a Schur form

$$M = Q T Q^* = Q \left[\begin{array}{c|c} A & C \\ \hline 0 & B \end{array}\right] Q^*$$

$\Lambda(A) \subset$ LHP
$\Lambda(B) \subset$ RHP

$$\text{sign}(M) = Q\,\text{sign}\left(\left[\begin{array}{c|c} A & C \\ \hline 0 & B \end{array}\right]\right)Q^* = Q\left[\begin{array}{c|c} -I & Z \\ \hline 0 & I \end{array}\right]Q^* \quad \text{for some } Z$$

(because $\text{sign}(A) = -I$, $\text{sign}(B) = +I$).

To compute Z, we use commutativity:

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

because $\begin{bmatrix} -1 & z \\ 0 & 1 \end{bmatrix}$ is a function of $\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}$

$$AZ + C = -C + ZB \iff AZ - ZB = -2C$$

This is a Sylvester equation that we can solve.

Algorithm:

1. Compute a Schur form, reordered as

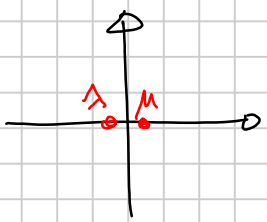
$$M = Q\begin{bmatrix} A & C \\ 0 & B \end{bmatrix}Q^*, \qquad \begin{array}{l} \Lambda(A) \subset \text{LHP} \\ \Lambda(B) \subset \text{RHP} \end{array}$$

2. Compute Z from $AZ - ZB = -2C$.

3. $S = Q\begin{bmatrix} -I & Z \\ 0 & I \end{bmatrix}Q^*$.

Cost: $O(n^3)$

The conditioning of the Sylvester equation is related to $\text{sep}(A, B)$



We might have trouble if there are eigenvalues close to the imaginary axis.

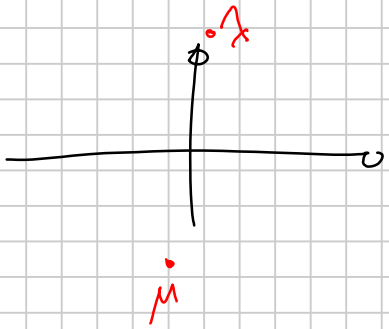
$$\text{sep}(A, B) \le \min\{|\lambda - \mu| : \lambda \in \Lambda(A), \mu \in \Lambda(B)\}$$

On the other hand,

$$\Lambda\left(\mathcal{L}_{\text{sign},M}\right) = \left\{ \text{sign}\left[\lambda,\mu\right] \;:\; \lambda,\mu \in \Lambda(M) \right\}$$

$$\text{sign}\left[\lambda,\mu\right] = \begin{cases} 0 & \text{if } \lambda = \mu \quad \text{(derivative)} \\ \dfrac{\text{sign}(\lambda) - \text{sign}(\mu)}{\lambda - \mu} & \text{if } \lambda \neq \mu \end{cases}$$

will also have a large eigenvalue if there are λ, μ close but on opposite sides of the imaginary axis.



Note: if $\text{sep}(A,B)$ is small, $\|Z\|$ is large $\Rightarrow \|S\|$ is large.

It is more useful to have an algorithm to compute $\text{sign}(M)$ that does not involve the Schur form!

Result: the matrix iteration

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

converges to $\lim_{k\to\infty} X_k = \text{sign}(M)$

Let us see why this works, starting from the scalar version of the iteration.

$$z_0 \in \mathbb{C} \qquad z_{k+1} = \frac{1}{2}\left(z_k + \frac{1}{z_k}\right) = \frac{z_k^2 + 1}{2z_k}$$

This is what you get if you apply the Newton method to $f(z) = z^2 - 1$

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

z^2-1 has simple zeros $-1, +1$, so we expect the sequence to converge to one of ± 1 quadratically fast.
We can tell more

Lemma: the sequence $z_{k+1} = \frac{1}{2}\left(z_k + \frac{1}{z_k}\right)$ converges quadratically to $\text{sign}(z_0)$ (if z_0 is not on the imaginary axis)

Proof: define

$$y_k = \frac{z_k - 1}{z_k + 1} \qquad (\text{if } z_k = -1,\ y_k = \infty)$$

$$y_{k+1} = \frac{z_{k+1} - 1}{z_{k+1} + 1} = \frac{\frac{1}{2}\left(z_k + \frac{1}{z_k}\right) - 1}{\frac{1}{2}\left(z_k + \frac{1}{z_k}\right) + 1} = \frac{z_k^2 + 1 - 2z_k}{z_k^2 + 1 + 2z_k}$$

$$= \left(\frac{z_k - 1}{z_k + 1}\right)^2 = y_k^2$$

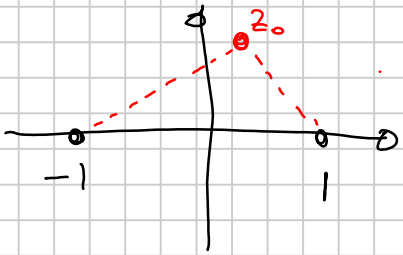
recursion $z_{k+1} = \frac{1}{2}\left(z_k + \frac{1}{z_k}\right)$

$$\begin{array}{ccccccc} z_0 & \to & z_1 & \to & z_2 & \to & z_3 \\ \downarrow & & \downarrow & & \downarrow & & \downarrow \\ y_0 & \to & y_1 & \to & y_2 & \to & y_3 \end{array}$$

squaring

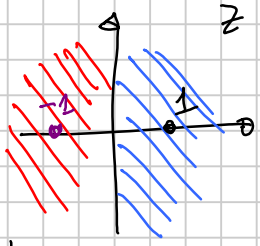
So $y_k = (y_0)^{2^k}$ and $\lim_{k\to\infty} y_k = \begin{cases} 0 & \text{if } |y_0| < 1 \\ \infty & \text{if } |y_0| > 1 \end{cases}$

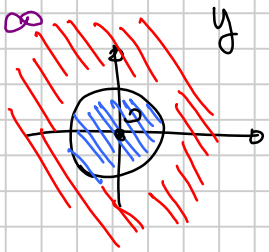
$$y_0 = \frac{z_0 - 1}{z_0 + 1} \qquad |y_0| = \frac{|z_0 - 1|}{|z_0 + 1|} = \frac{\text{dist}(z_0, 1)}{\text{dist}(z_0, -1)}$$



If $z_0 \in$ RHP, $|z_0 - 1| < |z_0 + 1| \Rightarrow |y_0| < 1$

If $z_0 \in$ LHP, $\Rightarrow |y_0| > 1$





If $z_0 \in$ RHP, $\lim_{k\to\infty} y_k = 0$ $\lim z_k = 1$.

If $z_0 \in$ LHP, $\lim_{k\to\infty} y_k = \infty$ $\lim z_k = -1$.

$$z_1 = \frac{1}{2}\left(z_0 + \frac{1}{z_0}\right) \qquad z_2 = \frac{1}{2}\left(z_1 + \frac{1}{z_1}\right) = \frac{p(z_0)}{q(z_0)}$$

The matrix version of this iteration:

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

$$X_0 = M \qquad X_1 = \frac{1}{2}(M + M^{-1}) = V \begin{bmatrix} \frac{1}{2}(\lambda_1 + \frac{1}{\lambda_1}) & & \\ & \ddots & \\ & & \frac{1}{2}(\lambda_n + \frac{1}{\lambda_n}) \end{bmatrix} V^{-1}$$

Similarly,

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

$f(z) = \frac{1}{2}(z + \frac{1}{z})$ $\quad f^k(z) =$ f composed with itself k times

$$\text{so} \lim_{k\to\infty} X_k = V \begin{bmatrix} \text{sign}(\lambda_1) & & \\ & \ddots & \\ & & \text{sign}(\lambda_n) \end{bmatrix} V^{-1}$$

<u>Proof</u> (that $\lim X_k \to \text{sign}(M)$):

We can assume (via a Schur form) that $M = X_0$ is triangular.

Then, all X_k are rational functions in X_0, so they are triangular and they commute with X_0.

In particular, $\text{diag}(X_k) = \left(f^k(\lambda_1), \ldots f^k(\lambda_n)\right)$

Define $Y_k = (X_k - S)(X_k + S)^{-1}$ with $S = \text{sign}(M)$

$\quad = (X_k + S)^{-1}(X_k - S)$

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

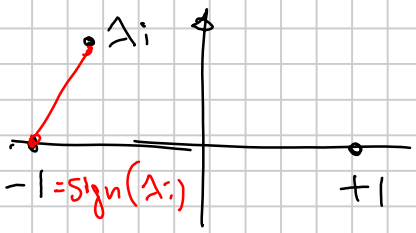
$$= \left(\left(\frac{1}{2}\left(X_k+X_k^{-1}\right)+S\right)\cdot 2X_k\right)^{-1}\left(\left(\frac{1}{2}\left(X_k+X_k^{-1}\right)-S\right)\left(2X_k\right)\right)$$

$$= \left(X_k^2+2X_kS+\underbrace{I}_{S^2}\right)^{-1}\left(X_k^2-2X_kS+\underbrace{I}_{S^2}\right)$$

$$= \left[\left(X_k+S\right)^{-1}\left(X_k-S\right)\right]^2 = Y_k^2$$

$Y_k^2 = Y_0^{2^k}$ $\quad Y_0$ is upper triangular and lies on the diagonal $y_{ii} = \dfrac{\lambda_i - \mathrm{sign}(\lambda_i)}{\lambda_i + \mathrm{sign}(\lambda_i)}$ with modulus $|y_{ii}| < 1$.

λ_i is closer to the one among ± 1 in its same half-plane $\rightarrow$



$\rho(Y_0) < 1 \Rightarrow Y_0^{2^k} \to 0 \quad Y_k \to 0 \quad X_k \to S$ □

The iteration

$$\begin{cases} X_0 = M \\ X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}\right) \end{cases}$$

converges quadratically to $S = \mathrm{sign}(M)$

It is known as "Newton iteration for the matrix sign"