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

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

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

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

$\operatorname{sign}(A)$ 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).

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.)

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,

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

Conditioning

From the theorems on the Fréchet derivative, for a diagonalizable A

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

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

Condition number

Theorem

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

where $N = (A^2)^{1/2}$.

Proof (sketch): let $L = L_{\mathrm{sign},A}(E)$. 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, $\mathrm{sep}(N, -N)$ plays a role.

Remark: if all eigenvalues of A are in the RHP, then the formula gives $\kappa_{abs}(\mathrm{sign}, A) = 0$.
Makes sense, since $\mathrm{sign}(A) = \mathrm{sign}(A + E) = I$ for all E for which eigenvalues do not cross the imaginary axis...

Schur-Parlett method

We can compute $\mathrm{sign}(A)$ with a Schur decomposition. It makes sense to reorder it so that eigenvalues in the LHP come first: $\Lambda(T_{11}) \subseteq LHP$, $\Lambda(T_{22}) \subseteq RHP$.

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

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 $\mathrm{sep}(T_{11}, T_{22})$.

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

(Matlab example)

Newton for the matrix sign

Most popular algorithm:

Newton for the matrix sign

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

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

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).

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 $y_{k+1} = -y_k^2$ (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., $\lim x_k = 1$).

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:

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

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 $X_k \to S$.

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

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

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.
(Once, or at each step.)

Scaling possibilities

Possibility 1: (determinantal scaling): choose $\mu = (\det A)^{-1/n}$, so that $\det A = 1$. 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 $\sigma_{\min}$.)

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...

Other iterations

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

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

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

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

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

Stability of the sign iterations

The stability analysis is complicated. [Bai Demmel '98 and Byers Mehrmann He '97]

While it works well in practice, the Newton iteration is not backward stable.

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

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

Nevertheless, the invariant subspaces it produces are: $A+E$ has a stable invariant subspace of the form $\begin{bmatrix} I \\ X \end{bmatrix}$, with

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

(Cfr. 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 $\operatorname{sign}(A)$ 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 $M_2, N_2, M_3, N_3, \dots$

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.,

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

All this is a sort of 'linear algebra on pencils': we map $N - xM$ to $N_1 - xM_1$ (one final project on this).