Newton for the matrix sign

Note Title 2023-04-06

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

Scalar version:

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

$$c(x) = \frac{x-1}{x+1} \qquad y_k = c(x_k)$$

$$\begin{tikzcd}
x_0 \arrow[r, "f"] \arrow[d, "c"] & x_1 \arrow[r, "f"] \arrow[d, "c"] & x_2 \arrow[r, "f"] \arrow[d, "c"] & x_3 \arrow[d, "c"] \\
y_0 \arrow[r, "\text{square}"] & y_1 \arrow[r, "\text{square}"] & y_2 \arrow[r, "\text{square}"] & y_3
\end{tikzcd}$$

$y_k = y_0^{2^k}$ converges to 0 or ∞ quadratically

x_k converges to ± 1 quadratically

Remark: $f^{\circ k}$ (f composed with itself k times) are rational approximations of $\operatorname{sign}(x)$ with increasing degrees.

One can obtain $f^{\circ k}$ with a Padé-like construction: construct an approximant s.t. $f^{\circ k}(x) - \operatorname{sign}(x) = O\left(x^{2^k}\right)$ for both $x \to 1$ and $x \to -1$.

Let us prove convergence $X_k \to \operatorname{sign}(M)$ without the assumption that f is diagonalizable.

Theorem: Let M have no purely imaginary eigenvalues.

Then, the sequence $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$, $X_0 = M$,
converges to $\mathrm{sign}(M)$ (quadratically).

Let $Y_k = (X_k - S)(X_k + S)^{-1}$, where $S = \mathrm{sign}(M)$.

Note that M, S, X_k, Y_k all commute with each other, since they can be expressed as rational functions in M.

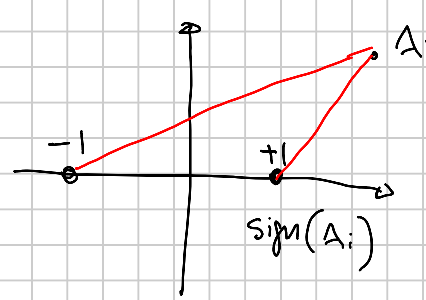
Up to a change of basis, we can assume M is upper triangular (Schur form). Then, also S, X_k, Y_k are upper triangular. In particular, this can be used to show that $(X_k+S)^{-1}$ exist: the upper triangular matrix $X_k + S$ has $f^{\circ k}(\lambda_i) + \mathrm{sign}(\lambda_i)$ on its diagonal, for $i = 1, 2, \dots n$,

$f^{\circ k}(\lambda_i)$ and $\mathrm{sign}(\lambda_i)$ are both in the same half-plane as λ_i (RHP or LHP), so they do not sum to 0.

$Y_0 = (M-S)(M+S)^{-1}$ has diagonal elements

$$\frac{\lambda_i - \mathrm{sign}(\lambda_i)}{\lambda_i + \mathrm{sign}(\lambda_i)},$$ and these are all smaller than 1

in modulus: $\mathrm{dist}(\lambda_i, \mathrm{sign}(\lambda_i)) < \mathrm{dist}(\lambda_i, -\mathrm{sign}(\lambda_i))$.



So $\rho(Y_0) < 1$

spectral radius

The squaring property continues to hold:

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

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

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

$$= (X_k - S)^2 (X_k + S)^{-2} = Y_k^2$$

we have used commutativity

$Y_k = Y_0^{2^k}$ converges to 0 quadratically, indeed

$$\|Y_k\|^{\frac{1}{2^k}} = \|Y_0^{2^k}\|^{\frac{1}{2^k}} \to \rho(Y_0).$$

Take the inverse transformation:

$$Y_k = (X_k - S)(X_k + S)^{-1} \iff Y_kX_k + Y_kS = X_k - S$$

$$\iff (I + Y_k)S = (I - Y_k)X_k \iff X_k = S(I+Y_k)(I-Y_k)^{-1}$$

If $Y_k \to 0$, $X_k \to S$ (both quadratically) $\square$

<u>Algorithm</u>:

1. set $X_0 = M$
2. for $k = 0, 1, 2, 3, \ldots$

   $X_{k+1} = \frac{1}{2}(X_k + \underbrace{X_k^{-1}}_{})$

   end

$\to$ $\mathrm{inv}(X_k)$, we do need the full inverse.

Stopping criterion: $\|X_{k+1} - X_k\| < \varepsilon$ works

<u>Problem</u>: slow convergence when $\|M\| >> 1$, or $\|M\| << 1$

(Matlab example)

Ideally, we would like to start from a matrix that has $|\lambda_i| \approx 1$ for all i.

<u>Scaling</u>: for each $\alpha > 0$, $\text{sign}(M) = \text{sign}(\alpha M)$

So we would like to start the iteration from a carefully chosen α.

Note that the matrix might have eigenvalues $\lambda_1, \dots \lambda_n$ with very different scales, so there is in general no hope to have <u>all</u> of them too close to 1,

$\alpha\lambda_1, \dots \alpha\lambda_n$ are still going to have different scales.

One reasonable choice is to ensure that the eigenvalues are "centered around 1", i.e. $|\alpha\lambda_{min}| \approx 10^{-k}$, $|\alpha\lambda_{max}| \approx 10^{+k}$ (logarithmically)

$$1 = |\alpha\lambda_{min}| \cdot |\alpha\lambda_{max}| \Rightarrow \alpha = \left(|\lambda_{min}| \cdot |\lambda_{max}|\right)^{-\frac{1}{2}}$$

Reasonable choice, if I know $\lambda_{min}, \lambda_{max}$ approximately.

To estimate λ_{min}, we can run a few iterations of the power method
$$\begin{cases} v_0 = \text{random} \\ v_{k+1} = M^{-1} v_k \end{cases} \qquad \lambda_{min} \approx \frac{v_k^T M v_k}{v_k^T v_k}.$$

(Note that the method needs M^{-1} anyway)

This is called <u>spectral scaling</u>.

<u>Variant</u>: <u>determinantal scaling</u>:

choose α to ensure $\det(\alpha M) = 1$.

So $\left(|\lambda_1 \lambda_2 \cdots \lambda_n|\right)^{1/n} = 1$, the geometric mean of the eigenvalues is 1.

$$1 = \det(\alpha M) = \alpha^n \det(M), \text{ so } \alpha = \det(M)^{-1/n}$$

Usually methods to compute M^{-1} produce the determinant for free (e.g. PLU factorization), so $\det(M)$ is cheap to obtain.

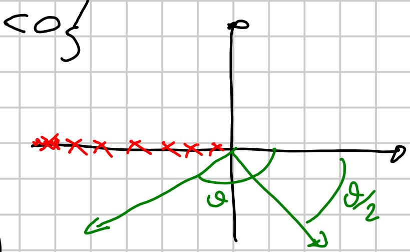
Variant: scale before every step, not only the first.

Stability of Newton for the sign: complicated, but the method seems to produce stable invariant subspaces.

Last function that we study in detail: matrix square root

$f(x)$ = principal square root: defined for all x outside of the negative real axis $\{a+ib : b=0, a<0\}$

$f(x)$ = unique square root in RHP



$$f(\rho e^{i\vartheta}) = \rho^{1/2} e^{i\vartheta/2} \qquad \vartheta \in (-\pi, \pi) \text{ (open)}$$
$$\rho \geq 0$$

$A^{1/2}$

$f(A)$ is ill-conditioned when A has
- one eigenvalue close to 0
- two eigenvalues close to the branch cut on opposite sides.

$f(A)$ defined if A has no eigenvalues on the negative real axis.

and all eigenvalues in $\mathcal{O}$ are simple (Jordan blocks have size 1).

Relation between sign and square root

Theorem: 1. for $A \in \mathbb{C}^{n \times n}$ without eigenvalues in $\mathcal{O}$ or in the negative real axis,

$$\text{sign}(A) = A(A^2)^{-\frac{1}{2}}$$

2. for A, B such that AB has no eigenvalues in $\mathcal{O}$ or on the negative real axis,

$$\text{sign}\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix} = \begin{bmatrix} 0 & C \\ C^{-1} & 0 \end{bmatrix} \qquad C = A(BA)^{-\frac{1}{2}}$$

In particular, $(B=I)$ $\text{sign}\begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} = \begin{bmatrix} 0 & A^{\frac{1}{2}} \\ A^{-\frac{1}{2}} & 0 \end{bmatrix}$

1. It is sufficient to prove the corresponding scalar identity $\text{sign}(z) = z(z^2)^{-\frac{1}{2}}$, because these identities extend to matrix functions.

z^2 has two square roots, z and $-z$.

$$(z^2)^{\frac{1}{2}} = \begin{cases} z & \text{if } z \in RHP \\ -z & \text{if } z \in LHP \end{cases} \Rightarrow z(z^2)^{-\frac{1}{2}} = \begin{cases} \frac{z}{z} = 1 & z \in RHP \\ \frac{z}{-z} = -1 & z \in LHP. \end{cases}$$

2. By the first part of the theorem,

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

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

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

We defined $C = A(BA)^{-1/2}$, we just need to prove that $D = B(AB)^{-1/2} = C^{-1}$

This follows from

$$I = \left(\operatorname{sign}\begin{bmatrix} 0 & A \\ B & 0 \end{bmatrix}\right)^2 = \begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix}\begin{bmatrix} 0 & C \\ D & 0 \end{bmatrix} = \begin{bmatrix} CD & 0 \\ 0 & DC \end{bmatrix}$$

hence $CD = I$.

Again, we will see two families of methods.

1. Method based on Schur-Parlett (Schur form + recursion).
2. Methods based on matrix iterations similar to Newton for the matrix sign.

2. can beat 1. when the convergence is sufficiently fast