True Newton vs. Modified Newton for sqrtm

Note Title 2023-04-20

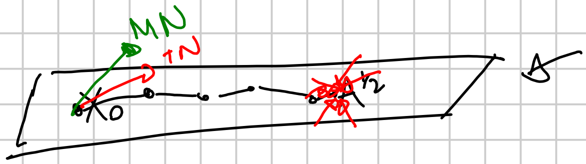
TN: $X_{k+1} = X_k - E$, E solves $EX_k + X_kE = X_k^2 - A$

MN: $X_{k+1} = \frac{1}{2}\left(X_k + X_k^{-1}A\right)$

X_0 commutes with A $\Rightarrow$ TN, MN coincide in exact arithmetic...

but <u>not</u> in machine arithmetic

Geometric picture:

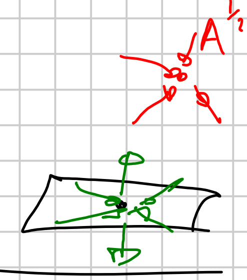


manifold of matrices that commute with A

TN, MN coincide on the manifold, but not on all $\mathbb{C}^{n\times n}$

When one allows for errors, two different iterations arise

TN is quadratically convergent to $A^{1/2}$ on all $\mathbb{C}^{n\times n}$



MN is convergent only on the manifold!

<u>Dynamical system</u> (in discrete time)

is defined by a map $F: \mathbb{C}^m \to \mathbb{C}^m$

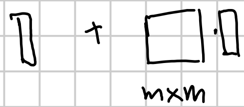
We are interested in studying the iteration $\begin{cases} x_0 \text{ given} \\ x_{k+1} = F(x_k) \end{cases}$

If $x_* = F(x_*)$, x_* is said to be a <u>fixed point</u>

We are interested in studying the iteration in a neighborhood of x_*, using its Jacobian J_{F,x_*} (which we assume exists)

$x_0 = x_* + e$ for a small perturbation $e \in \mathbb{C}^m$ $x_* \in \mathbb{C}^m$

$$x_1 = F(x_0 + e) = \underset{\substack{\| \\ x_*}}{F(x_*)} + J_{F,x_*} \cdot e + o(\|e\|)$$



$$x_2 = F(x_1) = F\left(x_* + J_{F,x_*} e + o(\|e\|)\right) = x_* + J_{F,x_*}^2 e + o(\|e\|)$$

$\vdots$

$$x_k = x_* + J_{F,x_*}^k e + o(\|e\|)$$

If $\rho(J_{F,x_*}) < 1$, $J_{F,x_*}^k \to 0$, hence the iteration started from $x_* + e$ will converge to x_*

x_* is <u>stable</u>

If $\rho(J_{F,x_*}) > 1$, $\|J_{F,x_*}^k\| \to \infty$, and at least for some choices of e the iteration will <u>not</u> converge to x_*

x_* is <u>unstable</u>

For TN, $J_{TN,A^{1/2}} = 0$, and this is what makes the Newton method converge quadratically.

We would like to see if MN has a stable or unstable fixed point in $A^{1/2}$, i.e. if $\rho(J_{MN,A^{1/2}})$ is smaller or <u>larger than 1.</u>

MN: $F: X \to \frac{1}{2}(X + X^{-1}A)$

The Jacobian of F (after vectorization) coincides with the Fréchet derivative of F.

To do this, we need to find the Fréchet derivative of the map $X \to X^{-1}$

$(I-M)^{-1} = I+M+M^2+M^3+\dots$ if $\rho(M)<1$.

$$(X+E)^{-1} = \left(X\left(I+X^{-1}E\right)\right)^{-1} = \left(I+X^{-1}E\right)^{-1}X^{-1} = \left(I - X^{-1}E + X^{-1}EX^{-1}E - X^{-1}EX^{-1}EX^{-1}E + \dots\right)X^{-1}$$

$$= X^{-1} - \underbrace{X^{-1}EX^{-1}}_{L_{inv,X}[E]} + O\left(\|E\|^2\right)$$

So

$$F(X+E) = \frac{1}{2}\left(X+E+(X+E)^{-1}A\right) = \frac{1}{2}\left(X+X^{-1}A\right) + \overset{L_{F,X}[E]}{\boxed{\frac{1}{2}\left(E - X^{-1}EX^{-1}A\right)}} + O\left(\|E\|^2\right)$$

$$\hat{L}_{F,X} = \frac{1}{2}\left(I_{n^2} - \left(X^{-1}A\right)^T \otimes X^{-1}\right)$$

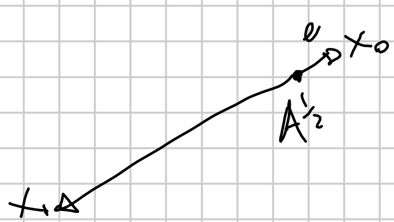
$$\hat{L}_{F,A^{1/2}} = \frac{1}{2}\left(I_{n^2} - \left(A^{1/2}\right)^T \otimes A^{-1/2}\right)$$

Assuming that $A^{-1/2}, A^{1/2}$ are put in Schur form, we obtain that if $\Lambda(A) = \{\lambda_1, \lambda_2, \dots \lambda_n\}$, then

$$\Lambda\left(\hat{L}_{F,A^{1/2}}\right) = \left\{ \frac{1}{2}\left(1 - \lambda_i^{1/2}\cdot\lambda_j^{-1/2}\right) : i,j=1,2,\dots n\right\}$$

When A is ill-conditioned, $|\lambda_{max}|/|\lambda_{min}|$ is large, and $\frac{1}{2}\left(1-\left(\frac{\lambda_{max}}{\lambda_{min}}\right)^{1/2}\right)$ can be large, too.

So MN is unstable: even if started very close to $X_* = A^{1/2}$,



the machine arithmetic error in the iterations can be amplified and make the method diverge.

So MN $\to$ can't be used, unstable

TN $\to$ can't be used, too expensive

But variants of MN have a very different stability behavior

<u>Variant</u>: Denman-Beavers iteration (DB)

Start from $X_{k+1} = \frac{1}{2}\left(X_k + \underbrace{X_k^{-1} A}_{Y_k^{-1} := X_k^{-1} A}\right)$

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

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

$$\boxed{\begin{cases} X_{k+1} = \frac{1}{2}\left(X_k + Y_k^{-1}\right) & X_0 = A \\ Y_{k+1} = \frac{1}{2}\left(Y_k + X_k^{-1}\right) & Y_0 = I \end{cases}}$$

DB iteration

<u>Ex</u>: show that DB is equivalent to applying the Newton iteration for the sign to compute

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

<u>Sol</u>: Newton for the sign reads

$$Z_k = \begin{bmatrix} 0 & X_k \\ Y_k & 0 \end{bmatrix} \qquad Z_{k+1} = \frac{1}{2}\left(Z_k + Z_k^{-1}\right) = \frac{1}{2}\left(\begin{bmatrix} 0 & X_k \\ Y_k & 0 \end{bmatrix} + \begin{bmatrix} 0 & Y_k^{-1} \\ X_k^{-1} & 0 \end{bmatrix}\right)$$

$$= \begin{bmatrix} 0 & \frac{1}{2}\left(X_k + Y_k^{-1}\right) \\ \frac{1}{2}\left(Y_k + X_k^{-1}\right) & 0 \end{bmatrix} \quad \checkmark$$

Let us study the stability of the DB iteration.

We are interested in the fixed point $(X_*, Y_*) = \left(A^{\frac{1}{2}}, A^{-\frac{1}{2}}\right)$

$$F\left(\begin{bmatrix} X \\ Y \end{bmatrix}\right) = \begin{bmatrix} \frac{1}{2}(X + Y^{-1}) \\ \frac{1}{2}(Y + X^{-1}) \end{bmatrix} \qquad \begin{aligned} X &\to X+E \\ Y &\to Y+F \end{aligned}$$

$$L_{DB,\begin{bmatrix} X \\ Y \end{bmatrix}} \begin{bmatrix} E \\ F \end{bmatrix} = \begin{bmatrix} \frac{1}{2}(E - Y^{-1}FY^{-1}) \\ \frac{1}{2}(F - X^{-1}EX^{-1}) \end{bmatrix} \qquad \hat{L}_{DB} \in \mathbb{C}^{2n^2 \times 2n^2}$$

Instead of computing eigenvalues, we note that $\hat{L}_{DB,\begin{bmatrix} A^{1/2} \\ A^{-1/2} \end{bmatrix}}$ is <u>idempotent</u>: $L_{DB}^2 = I$

$$L^2_{DB,\begin{bmatrix} A^{1/2} \\ A^{-1/2} \end{bmatrix}} \begin{bmatrix} E \\ F \end{bmatrix} = \begin{bmatrix} \frac{1}{2}\left(\frac{1}{2}(E - Y^{-1}FY^{-1}) - Y^{-1}\left(\frac{1}{2}(F - X^{-1}EX^{-1})\right)Y^{-1}\right) \\ \frac{1}{2}\left(\frac{1}{2}(F - X^{-1}EX^{-1}) - X^{-1}\left(\frac{1}{2}(E - Y^{-1}FY^{-1})\right)X^{-1}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{4}E - \frac{1}{4}Y^{-1}FY^{-1} - \frac{1}{4}Y^{-1}FY^{-1} + \frac{1}{4}\overbrace{\boxed{Y^{-1}X^{-1}}}^{=I}E\overbrace{\boxed{X^{-1}Y^{-1}}}^{=I} \\ 〃 \end{bmatrix}$$

$$= \begin{bmatrix} \left(\frac{1}{4}+\frac{1}{4}\right)E - \left(\frac{1}{4}+\frac{1}{4}\right)Y^{-1}FY^{-1} \\ 〃 \end{bmatrix} = L_{DB\begin{bmatrix} A^{1/2} \\ A^{-1/2} \end{bmatrix}} \begin{bmatrix} E \\ F \end{bmatrix}.$$

$\lambda = \lambda^2 \to \lambda \in \{0, 1\}$, or, more directly, the powers $\hat{L}_{DB,\begin{bmatrix} A^{1/2} \\ A^{-1/2} \end{bmatrix}}$ are all equal to $\hat{L}_{DB,\begin{bmatrix} A^{1/2} \\ A^{-1/2} \end{bmatrix}}$ and hence are bounded.

$$X_0 = X_* + e$$

$$X_k = X_* + (J_{F,X_*})^k e + o(\|e\|)$$

So DB does not amplify errors due to machine arithmetic $\Rightarrow$ the iteration is (locally) stable.

DB is indeed an effective iteration although it has 50%

higher computational cost.

$$\underbrace{\tfrac{1}{2}\big(X+\underbrace{X^{-1}A}_{\frac{8}{3}n^3}\big)}_{O(n^2)} \qquad \frac{8}{3}n^3+O(n^2) \qquad \text{vs.} \qquad \begin{cases} \overbrace{\tfrac{1}{2}\big(X+\overbrace{Y^{-1}}^{2n^3}\big)}^{O(n^2)} \\ \underbrace{\tfrac{1}{2}\big(Y+\underbrace{X^{-1}}_{2n^3}\big)}_{O(n^2)} \end{cases} \qquad 4n^3+O(n^2)$$

EX Show that Newton for the matrix sign has a stable fixed point.

(Matlab experiments)

Function of large-scale matrices.

Let A be large and sparse.

Problem: $f(A)$ is often full, e.g., $f(x)=x^{-1}$

Solution: we look for methods that can compute $f(A)\cdot b$ directly given $A\in\mathbb{C}^{n\times n}$ $b\in\mathbb{C}^n$, analogously to linear systems.

1) polynomial or rational approximation: $f(A)b \approx q^{-1}(A)p(A)b$

$$b, Ab, A^2b, \ldots \to \deg(p)\cdot \mathrm{nnz}(A) \text{ cost}$$

$q(A)^{-1}v$, with iterative methods, can still be computed with an iteration whose cost is $n_{steps}\cdot(\text{cost of evaluating } q(A)v)$

$$= n_{steps}\cdot \deg(q)\cdot \mathrm{nnz}(A).$$

Problem: need to find good rational approximant, and these require knowledge of $\Lambda(A)$.

2) Contour integration:

$$f(A) = \frac{1}{2\pi i}\int_{\Gamma} f(z)(zI-A)^{-1}dz \approx \sum_{k=1}^{N} w_k f(x_k)(x_k I - A)^{-1}.$$

for certain nodes $x_1, \dots x_N$ weights $w_1, \dots w_N$

$$f(A)b \approx \sum_{k=1}^{N} w_k f(x_k) \underbrace{(x_k I - A)^{-1} b}_{\text{linear system sol.}}$$

One can use methods for sparse linear systems, e.g. sparse LU.

Cost: $N \cdot (\text{cost of solving } (x_k I - A)y = b)$

This is actually also a rational approximation:

$$f(z) \approx \sum_{k=1}^{N} w_k f(x_k)(x_k - z)^{-1}$$

So it is a special case of method 1, with the function in a particular form.

3) Methods based on the Arnoldi iteration.

Also rational approximations, but include an almost-optimal way to choose them and a framework for their evaluation.

$f(A)b$