Matrix Functions Preserving Group Structure and Iterations for the Matrix Square Root

Nick Higham
Department of Mathematics
University of Manchester

higham@ma.man.ac.uk
http://www.ma.man.ac.uk/~higham/

Joint work with Niloufer Mackey, D. Steven Mackey, and Françoise Tisseur.

THE UNIVERSITY of MANCHESTER

Key Question to be Answered

How can we stabilize Newton's method for the matrix square root?

Group Background

Given nonsingular M and $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$,

$$\langle x, y \rangle_{\mathrm{M}} = \begin{cases} x^T M y, & \text{real or complex bilinear forms,} \\ x^* M y, & \text{sesquilinear forms.} \end{cases}$$

Define automorphism group

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : \langle Ax, Ay \rangle_{\mathrm{M}} = \langle x, y \rangle_{\mathrm{M}},\ \forall x, y \in \mathbb{K}^n \,\}.$$

Recall adjoint $A^\star$ of $A \in \mathbb{K}^{n\times n}$ wrt $\langle \cdot, \cdot \rangle_{\mathrm{M}}$ defined by

$$\langle Ax, y \rangle_{\mathrm{M}} = \langle x, A^\star y \rangle_{\mathrm{M}} \quad \forall x, y \in \mathbb{K}^n.$$

Can show: $$A^\star = \begin{cases} M^{-1} A^T M, & \text{for bilinear forms,} \\ M^{-1} A^* M, & \text{for sesquilinear forms.} \end{cases}$$

$$\mathbb{G} = \{\, A \in \mathbb{K}^{n\times n} : A^\star = A^{-1} \,\}.$$

Some Automorphism Groups

Space	M	$A^\star$	**Automorphism group,** $\mathbb{G}$
Groups corresponding to a bilinear form			
$\mathbb{R}^n$	I	A^T	Real orthogonals
$\mathbb{C}^n$	I	A^T	Complex orthogonals
$\mathbb{R}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q} A^T \Sigma_{p,q}$	Pseudo-orthogonals
$\mathbb{R}^n$	R	$R A^T R$	Real perplectics
$\mathbb{R}^{2n}$	J	$-J A^T J$	Real symplectics
$\mathbb{C}^{2n}$	J	$-J A^T J$	Complex symplectics
Groups corresponding to a sesquilinear form			
$\mathbb{C}^n$	I	A^*	Unitaries
$\mathbb{C}^n$	$\Sigma_{p,q}$	$\Sigma_{p,q} A^* \Sigma_{p,q}$	Pseudo-unitaries
$\mathbb{C}^{2n}$	J	$-J A^* J$	Conjugate symplectics

$$R = \begin{bmatrix} & & 1 \\ & \cdot^{\cdot^{\cdot}} & \\ 1 & & \end{bmatrix}, \qquad J = \begin{bmatrix} 0 & I_n \\ -I_n & 0 \end{bmatrix}, \qquad \Sigma_{p,q} = \begin{bmatrix} I_p & 0 \\ 0 & -I_q \end{bmatrix}$$

Questions

Consider $f : \mathbb{K}^{n\times n} \to \mathbb{K}^{n\times n}$.

If $A \in \mathbb{G}$

- For which f does $f(A) \in \mathbb{G}$?
- How can we exploit the (nonlinear) group structure when computing $f(A)$?

Questions

Consider $f : \mathbb{K}^{n\times n} \to \mathbb{K}^{n\times n}$.

If $A \in \mathbb{G}$

- For which f does $f(A) \in \mathbb{G}$?
- How can we exploit the (nonlinear) group structure when computing $f(A)$?

- Known (H, Mackey, Mackey & Tisseur, 2003): $A \in \mathbb{G}$ implies $\mathrm{sign}(A) \in \mathbb{G}$ for any $\mathbb{G}$.
This talk will provide a third proof.

Applications

- Many applications of matrix functions, $f(A)$, including in eigenproblem.
- A often has group structure.

Applications

- Many applications of matrix functions, $f(A)$, including in eigenproblem.
- A often has group structure.

Functions of interest include:

Applications

- Many applications of matrix functions, $f(A)$, including in eigenproblem.
- A often has group structure.

Functions of interest include:

- $f(A) = A$

Applications

- Many applications of matrix functions, $f(A)$, including in eigenproblem.
- A often has group structure.

Functions of interest include:

- $f(A) = A$
- $f(A) = A^{-1}$

Bilinear Forms

Theorem 1
(a) *For any* f *and* $A \in \mathbb{K}^{n\times n}$, $f(A^\star) = f(A)^\star$.
(b) *For* $A \in \mathbb{G}$, $f(A) \in \mathbb{G}$ *iff* $f(A^{-1}) = f(A)^{-1}$.

Proof. (a) We have

$$f(A^\star) = f(M^{-1}A^TM) = M^{-1}f(A^T)M = M^{-1}f(A)^TM = f(A)^\star.$$

(b) For $A \in \mathbb{G}$, consider

$$\begin{array}{ccc} f(A)^\star & = & f(A^\star) \\ & & \| \\ & & f(A^{-1}) \end{array}$$

Bilinear Forms

Theorem 1
(a) *For any* f *and* $A \in \mathbb{K}^{n\times n}$, $f(A^\star) = f(A)^\star$.
(b) *For* $A \in \mathbb{G}$, $f(A) \in \mathbb{G}$ *iff* $f(A^{-1}) = f(A)^{-1}$.

Proof. (a) We have

$$f(A^\star) = f(M^{-1}A^TM) = M^{-1}f(A^T)M = M^{-1}f(A)^TM = f(A)^\star.$$

(b) For $A \in \mathbb{G}$, consider

$$\begin{array}{ccc} f(A)^\star & = & f(A^\star) \\ \| & & \| \\ f(A)^{-1} & = & f(A^{-1}) \end{array}$$

Sesquilinear Forms

Theorem 2 *For all $A \in \mathbb{K}^{n\times n}$, any two of the properties*

- $f(A^*) = f(A)^*$
- $f(A^{-1}) = f(A)^{-1}$,
- $f(A) \in \mathbb{G}$

imply the third.
The first condition is equivalent to $f(\overline{A}) = \overline{f(A)}$.

Implications

For bilinear forms, f preserves group structure of A when $f(A^{-1}) = f(A)^{-1}$.

This condition holds *for all A* for

- Matrix sign function, $\mathrm{sign}(A) = A(A^2)^{-1/2}$.
- Any matrix power A^α, subject to suitable choice of branches. In particular, the
  - principal matrix pth root $A^{1/p}$ ($p \in \mathbb{Z}^+$, $\Lambda(A) \cap \mathbb{R}^- = \emptyset$): unique X such that
    1. $X^p = A$.
    2. $-\pi/p < \arg(\lambda(X)) < \pi/p$.

Rational Functions (1)

When is $f(A^{-1}) = f(A)^{-1}$ for *all* $A \in \mathbb{G}$ and *all* $\mathbb{G}$.

By taking A diagonal see that $f(x)f(1/x) \equiv 1$ is necessary.

If p has degree m then $\mathrm{rev}\,p(x) := x^m p(1/x)$.

Theorem 3 *For bilinear forms, a rational f satisfies $f(\mathbb{G}) \subseteq \mathbb{G}$ for all $\mathbb{G}$ iff*

$$f(z) = \pm z^k p(z)/\mathrm{rev}\,p(z),$$

for some $k \in \mathbb{Z}$ and some monic p with $p(0) \neq 0$, where p is real (complex) if bilinear form is real (complex).

Rational Functions (2)

Proof. Sufficiency: show

$$\boxed{f(z) = \pm z^k p(z)/\mathrm{rev}\, p(z)},$$

where $\mathrm{rev}\, p(x) = x^n p(1/x)$, implies $f(A^{-1}) = f(A)^{-1}$.
We have

$$\begin{aligned} f(A)f(A^{-1}) &= \pm A^k p(A)[\mathrm{rev}\, p(A)]^{-1} \times \pm A^{-k} p(A^{-1})[\mathrm{rev}\, p(A^{-1})]^{-1} \\ &= A^k p(A)[A^n p(A^{-1})]^{-1} \times A^{-k} p(A^{-1})[A^{-n} p(A)]^{-1} \\ &= A^k p(A) p(A^{-1})^{-1} A^{-n} \times A^{-k} p(A^{-1}) p(A)^{-1} A^n \\ &= I. \end{aligned}$$

Tool 1: Matrix Sign Relation

Theorem 4 *Let* $A, B \in \mathbb{C}^{n\times n}$ *and* $\Lambda(AB) \cap \mathbb{R}^- = \emptyset$. *Then*

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

where $C = A(BA)^{-1/2}$.

Proof. Use $\operatorname{sign}(P) = P(P^2)^{-1/2}$. $\square$

Tool 1: Matrix Sign Relation

Theorem 4 *Let* $A, B \in \mathbb{C}^{n\times n}$ *and* $\Lambda(AB) \cap \mathbb{R}^- = \emptyset$*. Then*

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

where $C = A(BA)^{-1/2}$*.*

Proof. Use $\operatorname{sign}(P) = P(P^2)^{-1/2}$. $\square$

Corollary 1 *Let* $A \in \mathbb{C}^{n\times n}$ *and* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$*. Then*

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

Class of Square Root Iterations

Theorem 5 *Suppose the iteration* $X_{k+1} = X_k T(X_k^2)$, $X_0 = A$ *converges to* $\operatorname{sign}(A)$ *with order* m. *If* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$ *and*

$$\boxed{\begin{array}{ll} Y_{k+1} = Y_k T(Z_k Y_k), & Y_0 = A, \\ Z_{k+1} = Z_k T(Y_k Z_k), & Z_0 = I, \end{array}}$$

then $Y_k \to A^{1/2}$ *and* $Z_k \to A^{-1/2}$ *as* $k \to \infty$, *both with order* m, *and* $Y_k = A Z_k$ *for all* k. *Moreover, if* $X \in \mathbb{G}$ *implies* $XT(X^2) \in \mathbb{G}$ *then* $A \in \mathbb{G}$ implies $Y_k \in \mathbb{G}$ and $Z_k \in \mathbb{G}$ *for all* k.

Padé Square Root Iterations

Theorem 6 *Assume $A \in \mathbb{G}$ and $\Lambda(A) \cap \mathbb{R}^- = \emptyset$. Consider*

$$\begin{aligned} Y_{k+1} &= Y_k\, p_m(I - Z_k Y_k)\, [\mathrm{rev}\, p_m(I - Z_k Y_k)]^{-1}, & Y_0 &= A, \\ Z_{k+1} &= Z_k\, p_m(I - Y_k Z_k)\, [\mathrm{rev}\, p_m(I - Y_k Z_k)]^{-1}, & Z_0 &= I, \end{aligned}$$

where $p_m(t)$ ($m \geq 1$) is numerator in $[m/m]$ Padé approximant to $(1-t)^{-1/2}$. Then $Y_k \in \mathbb{G}$, $Z_k \in \mathbb{G}$ and $Y_k = AZ_k$, for all k, and $Y_k \to A^{1/2}$, $Z_k \to A^{-1/2}$, both with order $2m+1$.

Padé Square Root Iterations

Theorem 6 *Assume* $A \in \mathbb{G}$ *and* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$*. Consider*

$$\begin{aligned} Y_{k+1} &= Y_k\, p_m(I - Z_k Y_k)\, [\mathrm{rev}\, p_m(I - Z_k Y_k)]^{-1}, \qquad & Y_0 &= A, \\ Z_{k+1} &= Z_k\, p_m(I - Y_k Z_k)\, [\mathrm{rev}\, p_m(I - Y_k Z_k)]^{-1}, \qquad & Z_0 &= I, \end{aligned}$$

where $p_m(t)$ $(m \geq 1)$ *is numerator in* $[m/m]$ *Padé approximant to* $(1-t)^{-1/2}$*. Then* $Y_k \in \mathbb{G}$*,* $Z_k \in \mathbb{G}$ *and* $Y_k = AZ_k$*, for all* k*, and* $Y_k \to A^{1/2}$*,* $Z_k \to A^{-1/2}$*, both with order* $2m+1$*.*

Structure-preserving cubic ($m = 1$):

$$\begin{aligned} Y_{k+1} &= Y_k(3I + Z_k Y_k)(I + 3Z_k Y_k)^{-1}, \qquad & Y_0 &= A, \\ Z_{k+1} &= Z_k(3I + Y_k Z_k)(I + 3Y_k Z_k)^{-1}, \qquad & Z_0 &= I. \end{aligned}$$

Tool 2: Generalized Polar Decomposition

Theorem 7 *Let* $\mathbb{G}$ *be a group. Any* $A \in \mathbb{K}^{n\times n}$ *such that* $(A^\star)^\star = A$ *and* $\Lambda(A^\star A) \cap \mathbb{R}^- = \emptyset$ *has a unique decomposition* $A = WS$, *where*

$$\begin{aligned} &W \in \mathbb{G} \qquad (\textit{i.e.}, W^\star = W^{-1}), \\ &S^\star = S, \end{aligned}$$

and $\Lambda(S) \in$ *open right half-plane (i.e.,* $\operatorname{sign}(S) = I$*).*

Note

- $(A^\star)^\star = A$ holds for all $\mathbb{G}$ in the earlier table.
- Both conditions are *necessary* for the existence.
- Other gpd's exist with different conditions on $\Lambda(S)$ (Rodman & co-authors).

Tool 3: Matrix Sign Relation

Corollary 2 *Let* $A \in \mathbb{K}^{n\times n}$ *have a generalized polar decomposition* $A = WS$. *Then*

$$\operatorname{sign}\left(\begin{bmatrix} 0 & A \\ A^{\star} & 0 \end{bmatrix}\right) = \begin{bmatrix} 0 & W \\ W^{\star} & 0 \end{bmatrix}.$$

Generalized Polar Iteration

Theorem 8 *Suppose the iteration* $X_{k+1} = X_k T(X_k^2)$*,* $X_0 = A$ *converges to* $\mathrm{sign}(A)$ *with order* m*. If* A *has the generalized polar decomposition* $A = WS$ *w.r.t. a bilinear form then*

$$\boxed{Y_{k+1} = Y_k T(Y_k^\star Y_k), \qquad Y_0 = A}$$

converges to W *with order of convergence* m*.*

Generalized Polar Iteration

Theorem 8 *Suppose the iteration* $X_{k+1} = X_k T(X_k^2)$*,* $X_0 = A$ *converges to* $\operatorname{sign}(A)$ *with order* m*. If* A *has the generalized polar decomposition* $A = WS$ *w.r.t. a bilinear form then*

$$Y_{k+1} = Y_k T(Y_k^\star Y_k), \qquad Y_0 = A$$

converges to W *with order of convergence* m*.*

Theorem 9 *Let* $\mathbb{G}$ *be any automorphism group and* $A \in \mathbb{G}$*. If* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$ *then* $I + A = WS$ *is a generalized polar decomposition with* $W = A^{1/2}$ *and* $S = A^{-1/2} + A^{1/2}$*.*

Newton Iteration

Theorem 10 *Let* $A \in \mathbb{G}$ (*any group*), $\Lambda(A) \cap \mathbb{R}^- = \emptyset$, *and*

$$\begin{aligned} Y_{k+1} &= \frac{1}{2}(Y_k + Y_k^{-\star}) \\ &= \frac{1}{2}(Y_k + M^{-1}Y_k^{-T}M), \qquad Y_1 = \frac{1}{2}(I + A). \end{aligned}$$

Then $Y_k \to A^{1/2}$ *quadratically.*

Proof. Apply theorems above to adapt Newton sign $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$ to gen polar decomp of $I + A$. $\square$

Newton Iteration

Theorem 10 *Let* $A \in \mathbb{G}$ *(any group),* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$*, and*

$$\begin{aligned} Y_{k+1} &= \frac{1}{2}(Y_k + Y_k^{-\star}) \\ &= \frac{1}{2}(Y_k + M^{-1}Y_k^{-T}M), \qquad Y_1 = \frac{1}{2}(I + A). \end{aligned}$$

Then $Y_k \to A^{1/2}$ *quadratically.*

Proof. Apply theorems above to adapt Newton sign $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$ to gen polar decomp of $I + A$. $\square$

Cf.

- Cardoso, Kenney & Silva Leite (2003, App. Num. Math.)—bilinear forms with $M^T = \pm M$, $M^T M = I$.
- H (2003, SIREV)—$M = \Sigma_{p,q}$.

Newton Iteration

Theorem 10 *Let* $A \in \mathbb{G}$ *(any group),* $\Lambda(A) \cap \mathbb{R}^- = \emptyset$*, and*

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-\star})$$
$$= \frac{1}{2}(Y_k + M^{-1}Y_k^{-T}M), \qquad Y_1 = \frac{1}{2}(I + A).$$

Then $Y_k \to A^{1/2}$ *quadratically.*

Proof. Apply theorems above to adapt Newton sign $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$ to gen polar decomp of $I + A$. $\square$

Usual Newton for $A^{1/2}$:

$$X_{k+1} = \frac{1}{2}(X_k + X_k^{-1}A), \qquad X_0 = A.$$

Can show $Y_k \equiv X_k \ (k \geq 1)$!

Experiment

Random pseudo-orthogonal $A \in \mathbb{R}^{10\times 10}$, $M = \mathrm{diag}(I_6, -I_4)$, ($A^T M A = M$) and $\|A\|_2 = 10^5 = \|A^{-1}\|_2$, generated using alg of H (2003) and chosen to be symmetric positive definite.

$$\mathrm{err}(X) = \frac{\|X - A^{1/2}\|_2}{\|A^{1/2}\|_2},$$

$$\mu_{\mathbb{G}}(X) = \frac{\|X^\star X - I\|_2}{\|X\|_2^2}.$$

Results

k	Newton	Group Newton		Cubic, struc. pres.	
	$\mathrm{err}(X_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$	$\mathrm{err}(Y_k)$	$\mu_{\mathbb{G}}(Y_k)$
0	3.2e+2			3.2e+2	1.4e-15
1	1.6e+2	1.6e+2	1.0e-5	1.0e+2	7.2e-15
2	7.8e+1	7.8e+1	1.0e-5	3.4e+1	6.1e-14
3	3.9e+1	3.9e+1	1.0e-5	1.1e+1	5.1e-13
4	1.9e+1	1.9e+1	1.0e-5	3.0e+0	2.9e-12
5	8.9e+0	8.9e+0	9.9e-6	5.5e-1	4.4e-12
6	4.0e+0	4.0e+0	9.6e-6	2.0e-2	4.3e-12
7	3.2e+1	1.6e+0	8.5e-6	2.0e-6	4.5e-12
8	2.3e+5	4.9e-1	5.5e-6	2.1e-11	4.8e-12
9	4.6e+9	8.2e-2	1.5e-6		
10	2.3e+9	3.1e-3	6.1e-8		
11	1.1e+9	4.7e-6	9.5e-11		
12	5.6e+8	2.1e-11	2.4e-16		

Conclusions

- ★ f preserves group structure if $f(A^{-1}) = f(A)^{-1}$ (and if $f(\overline{A}) = \overline{f(A)}$ in sesquilinear case).
- ★ Rational functions mapping $\mathbb{G}$ into itself $\forall \mathbb{G}$ characterized.
- ★ Derived new family of coupled iterations for $A^{1/2}$ that is structure preserving for matrix groups.
- ★ Using gen polar decomp, derived numerically stable form of Newton for $A^{1/2}$ when $A \in \mathbb{G}$.

www.google.com