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.
How can we stabilize Newton’s method for the matrix square root?
Group Background

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

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

Define automorphism group

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

Recall adjoint $A^*$ of $A \in K^{n \times n}$ wrt $\langle \cdot, \cdot \rangle_M$ defined by

$$\langle Ax, y \rangle_M = \langle x, A^* y \rangle_M \quad \forall x, y \in K^n.$$ 

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

$$G = \{ A \in K^{n \times n} : A^* = A^{-1} \}.$$
Some Automorphism Groups

<table>
<thead>
<tr>
<th>Space</th>
<th>$M$</th>
<th>$A^*$</th>
<th>Automorphism group, $G$</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Groups corresponding to a bilinear form</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$I$</td>
<td>$A^T$</td>
<td>Real orthogonals</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$I$</td>
<td>$A^T$</td>
<td>Complex orthogonals</td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$\Sigma_{p,q}$</td>
<td>$\Sigma_{p,q}A^T \Sigma_{p,q}$</td>
<td>Pseudo-orthogonals</td>
</tr>
<tr>
<td>$\mathbb{R}^n$</td>
<td>$R$</td>
<td>$RA^T R$</td>
<td>Real perplectics</td>
</tr>
<tr>
<td>$\mathbb{R}^{2n}$</td>
<td>$J$</td>
<td>$-JA^T J$</td>
<td>Real symplectics</td>
</tr>
<tr>
<td>$\mathbb{C}^{2n}$</td>
<td>$J$</td>
<td>$-JA^T J$</td>
<td>Complex symplectics</td>
</tr>
<tr>
<td><strong>Groups corresponding to a sesquilinear form</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$I$</td>
<td>$A^*$</td>
<td>Unitaries</td>
</tr>
<tr>
<td>$\mathbb{C}^n$</td>
<td>$\Sigma_{p,q}$</td>
<td>$\Sigma_{p,q}A^* \Sigma_{p,q}$</td>
<td>Pseudo-unitaries</td>
</tr>
<tr>
<td>$\mathbb{C}^{2n}$</td>
<td>$J$</td>
<td>$-JA^* J$</td>
<td>Conjugate symplectics</td>
</tr>
</tbody>
</table>

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

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

If $A \in \mathcal{G}$

► For which $f$ does $f(A) \in \mathcal{G}$?

► How can we exploit the (nonlinear) group structure when computing $f(A)$?
Questions

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

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):

\[ A \in \mathbb{G} \text{ implies } \text{sign}(A) \in \mathbb{G} \text{ for any } \mathbb{G} \, .\]

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 \( A \in \mathbb{K}^{n \times n} \), \( f(A^*) = f(A)^* \).
(b) For \( A \in \mathbb{G} \), \( f(A) \in \mathbb{G} \) iff \( f(A^{-1}) = f(A)^{-1} \).

Proof. (a) We have

\[
f(A^*) = f(M^{-1} A^T M) = M^{-1} f(A^T) M = M^{-1} f(A)^T M = f(A)^*.
\]
(b) For \( A \in \mathbb{G} \), consider

\[
f(A)^* = f(A^*) \quad \| \quad f(A^{-1})
\]
Bilinear Forms

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

Proof. (a) We have

$$f(A^*) = f(M^{-1}A^T M) = M^{-1}f(A^T)M = M^{-1}f(A)^T M = f(A)^*.$$ 

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

$$f(A)^* = f(A^*)$$

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

Theorem 2  For all \( A \in K^{n \times n} \), 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 \( f(A) = \overline{f(A)} \).
Implications

For bilinear forms, \( f \) preserves group structure of \( A \) when
\[
f(A^{-1}) = f(A)^{-1}.
\]
This condition holds \textit{for all} \( A \) for

- **Matrix sign function**, \( \text{sign}(A) = A(A^2)^{-1/2} \).

- Any matrix power \( A^\alpha \), subject to suitable choice of branches. In particular, the

  - **principal matrix \( p \)th 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 \( A \in G \) and all \( G \).

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

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

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

\[
f(z) = \pm z^k p(z)/\text{rev}p(z),
\]

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

\[ f(z) = \pm z^k p(z)/\text{rev} p(z) , \]

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

\[
f(A)f(A^{-1}) = \pm A^k p(A)[\text{rev} p(A)]^{-1} \times \pm A^{-k} p(A^{-1})[\text{rev} p(A^{-1})]^{-1}
\]
\[
= A^k p(A)[A^n p(A^{-1})]^{-1} \times A^{-k} p(A^{-1})[A^{-n} p(A)]^{-1}
\]
\[
= A^k p(A)p(A^{-1})^{-1} A^{-n} \times A^{-k} p(A^{-1})p(A)^{-1} A^n
\]
\[
= I .
\]
Tool 1: Matrix Sign Relation

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

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

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

Proof. Use $\text{sign}(P) = P(P^2)^{-1/2}$. □
**Tool 1: Matrix Sign Relation**

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

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

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

**Proof.** Use $\text{sign}(P) = P(P^2)^{-1/2}$. \qed

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

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ I & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & A^{1/2} \\ A^{-1/2} & 0 \end{bmatrix},$$
Class of Square Root Iterations

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

\[
Y_{k+1} = Y_k T(Z_k Y_k), \quad Y_0 = A, \\
Z_{k+1} = Z_k T(Y_k Z_k), \quad Z_0 = I,
\]

then \( Y_k \to A^{1/2} \) and \( Z_k \to A^{-1/2} \) as \( k \to \infty \), both with order \( m \), and \( Y_k = AZ_k \) 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 \( k \).
Theorem 6 Assume $A \in \mathbb{G}$ and $\Lambda(A) \cap \mathbb{R}^- = \emptyset$. Consider

\[
Y_{k+1} = Y_k p_m(I - Z_k Y_k) [\text{rev} p_m(I - Z_k Y_k)]^{-1}, \quad Y_0 = A,
\]

\[
Z_{k+1} = Z_k p_m(I - Y_k Z_k) [\text{rev} p_m(I - Y_k Z_k)]^{-1}, \quad Z_0 = I,
\]

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 $Y_k \rightarrow A^{1/2}$, $Z_k \rightarrow A^{-1/2}$, both with order $2m + 1$. 
Theorem 6: Assume $A \in \mathbb{G}$ and $\Lambda(A) \cap \mathbb{R}^- = \emptyset$. Consider $$Y_{k+1} = Y_k p_m(I - Z_k Y_k) [\text{rev} p_m(I - Z_k Y_k)]^{-1}, \quad Y_0 = A,$$
$$Z_{k+1} = Z_k p_m(I - Y_k Z_k) [\text{rev} p_m(I - Y_k Z_k)]^{-1}, \quad Z_0 = I,$$

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 ($m = 1$):**

$$Y_{k+1} = Y_k(3I + Z_k Y_k)(I + 3Z_k Y_k)^{-1}, \quad Y_0 = A,$$
$$Z_{k+1} = Z_k(3I + Y_k Z_k)(I + 3Y_k Z_k)^{-1}, \quad Z_0 = I.$$
Theorem 7  Let \( G \) be a group. Any \( A \in \mathbb{K}^{n \times n} \) such that \((A^*)^* = A\) and \( \Lambda(A^*A) \cap \mathbb{R}^- = \emptyset \) has a unique decomposition \( A = WS \), where

\[
W \in G \quad (i.e., \ W^* = W^{-1}),
\]
\[
S^* = S,
\]

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

Note

- \((A^*)^* = A\) holds for all \( G \) in the earlier table.
- Both conditions are necessary for the existence.
- Other gpd’s exist with different conditions on \( \Lambda(S) \) (Rodman & co-authors).
Corollary 2  Let $A \in \mathbb{K}^{n \times n}$ have a generalized polar decomposition $A = WS$. Then

$$\text{sign} \left( \begin{bmatrix} 0 & A \\ A^* & 0 \end{bmatrix} \right) = \begin{bmatrix} 0 & W \\ W^* & 0 \end{bmatrix}.$$
Theorem 8  Suppose the iteration $X_{k+1} = X_k T(X_k^2)$, $X_0 = A$ converges to $\text{sign}(A)$ with order $m$. If $A$ has the generalized polar decomposition $A = WS$ w.r.t. a bilinear form then

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

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 $\text{sign}(A)$ with order $m$. If $A$ has the generalized polar decomposition $A = WS$ w.r.t. a bilinear form then

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

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

**Theorem 9** Let $G$ be any automorphism group and $A \in G$. 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

$$Y_{k+1} = \frac{1}{2} (Y_k + Y_k^-)$$

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

Then $Y_k \rightarrow A^{1/2}$ 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 $I + A$.
Newton Iteration

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

\[
Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-\star})
\]

\[
= \frac{1}{2}(Y_k + M^{-1}Y_k^{-T}M), \quad Y_1 = \frac{1}{2}(I + A).
\]

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 \( I + A \). □

Cf.

● Cardoso, Kenney & Silva Leite (2003, App. Num. Math.)—bilinear forms with \( M^T = \pm M \), \( M^T M = I \).

● H (2003, SIREV)—\( M = \sum_{p,q} \).
Newton Iteration

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

$$Y_{k+1} = \frac{1}{2}(Y_k + Y_k^{-\ast})$$

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

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 $I + A$.  

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

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

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

\[
\text{err}(X) = \frac{\|X - A^{1/2}\|_2}{\|A^{1/2}\|_2},
\]

\[
\mu_G(X) = \frac{\|X^* X - I\|_2}{\|X\|_2^2}.
\]
## Results

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

★ $f$ preserves group structure if $f(A^{-1}) = f(A)^{-1}$
(and if $f(A) = f(A)$ in sesquilinear case).

★ Rational functions mapping $\mathbb{G}$ into itself $\forall G$ 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 $A \in \mathbb{G}$.